In Silico Signature Prediction Modeling in Cytolethal Distending Toxin-Producing Escherichia coli Strains

Maryam Javadi, Mana Oloomi*, Saeid Bouzari

Department of Molecular Biology, Pasteur Institute of Iran, Tehran 13164, Iran

In this study, cytolethal distending toxin (CDT) producer isolates genome were compared with genome of pathogenic and commensal Escherichia coli strains. Conserved genomic signatures among different types of CDT producer E. coli strains were assessed. It was shown that they could be used as biomarkers for research purposes and clinical diagnosis by polymerase chain reaction, or in vaccine development. cdt genes and several other genetic biomarkers were identified as signature sequences in CDT producer strains. The identified signatures include several individual phage proteins (holins, nucleases, and terminases, and transferases) and multiple members of different protein families (the lambda family, phage-integrase family, phage-tail tape protein family, putative membrane proteins, regulatory proteins, restriction-modification system proteins, tail fiber-assembly proteins, base plate-assembly proteins, and other prophage tail-related proteins). In this study, a sporadic phylogenetic pattern was demonstrated in the CDT-producing strains. In conclusion, conserved signature proteins in a wide range of pathogenic bacterial strains can potentially be used in modern vaccine-design strategies.

Keywords: biomarkers, cytolethal distending toxin, genomic signature, multiple alignments, pathogenic Escherichia coli

Introduction

The co-evolution of pathogenic bacteria and their hosts leads to the generation of functional pathogen-host interfaces. Well-adapted pathogens have evolved a variety of strategies for manipulating host cell functions to guarantee their successive colonization and survival. For instance, a group of gram-negative bacterial pathogens produces a toxin, known as cytolethal distending toxin (CDT) [1]. Among the vast majority of CDT producers are Escherichia coli, which is commonly found in the intestines of humans and other mammals. Most E. coli strains are harmless commensals; however, some isolates can cause severe diseases and are designated as pathogenic E. coli. Among the various pathogenic E. coli strains, some have acquired virulence determinants through the horizontal transfer of genes, such as the cdt genes encoding CDTs. CDTs were the first bacterial toxins identified that block the eukaryotic cell cycle and suppress cell proliferation, eventually resulting in cell death. The active subunits of CDT toxins exhibit features of type I deoxyribonuclease-like activity [2, 3].

In this study, comparative genome analysis of CDT-producer E. coli isolates with other pathogenic and commensal strains was performed. Alignments between multiple genomes led to the identification of a set of distinct ("signature") sequence motifs. These signature sequences could be used to delineate single genomes or a specified group of associated genomes within a desired group, such as the CDT-producing E. coli (the target group in this study). While genomic signatures were conserved in the target group, which they were not conserved or were absent in other related or unrelated genomes (i.e., the background group). From a clinical point of view, conserved signature sequences could offer advantages in predicting and further designing novel CDT inhibitors to vaccine candidates [4].

On the other hand, phylogenetic trees can be constructed based on multiple sequence alignments. It is important that phylogeny based on an immense number of genes and whole-genome sequences are more reliable than those based on a single gene or a few selected loci [5]. Phylogenetic analysis can provide an overall classification of the target group among the background group. Alignment of whole-genome sequences yields detailed information on specific differences...
between genomes and, consequently, has shed new insights into phylogenetic relationships in recent years [6-9].

In this study, phylogenetic relationships of CDT+ strains with other pathogenic and commensal E. coli strains were assessed, and conserved signature genomic regions in the target group (CDT-producers) were annotated. This information could be used for developing molecular diagnostics assays, polymerase chain reaction primer and probe design in modern vaccines.

Methods

CDT+ strains

Several databases were used to identify bacterial strains harboring cdt genes. Data was extracted from the following resources: NCBI, National Center for Biotechnology Information GenBank; EMBL, European Molecular Biology Laboratory; DDBJ, DNA Data Bank of Japan; PDB, Protein Data Bank; RefSeq, NCBI Reference Sequence Database; and UniProtKB, Swiss-Prot Database.

Whole-genome sequences

All genomes analyzed in this study were downloaded from the NCBI file transfer protocol (FTP) site at: ftp://ftp.ncbi.nih.gov/genomes.

Reordering of draft genomes

Ordering and orienting contigs in draft genomes facilitates comparative genome analysis. Contig ordering can be predicted by comparison of a reference genome that is expected to have a conserved genome organization [10]. ProgressiveMauve (version 2.3.1) was used for ordering contigs in draft genomes. Mauve contig mover (MCM) offers advantages over methods that rely on matches in limited regions near the ends of contigs [11, 12]. The E. coli K-12 MG1655 strain (accession No. NC_000913.3) was used as a reference genome.

The MCM optional parameters were used in this study including default seed weight, use seed families: 15 determine Locally Collinear Blocks (LCBs); LCBs, full alignment, iterative refinement, sum-of-pairs LCB scoring, and min LCB weight: 200.

Multiple genome alignments

In this study, Gegenees software (version 2.2.1) was used for multiple-genome alignments. The software is written in JAVA, and making it compatible with several platforms. Limitations were not observed in the speed calculation, number and memory of the genomes that could be aligned. Gegenees software is also capable of performing fragmented alignments [4]. Multiple alignments of E. coli genomes were created using a fragment size of 200 nucleotides, a step size of 100 parameters, and BLASTN, which was optimized for highly similar sequences.

Phylogenetic tree construction

A phylogram was produced in SplitsTree 4, using the neighbor-joining method and a distance matrix Nexus file exported from Gegenees software [13]. E. albertii TW07627 and E. fergusonii ATCC 35469 strains were set as the out-groups.

Identifying conserved signatures

CDT-producing isolates were set as the target group, and all other strains were used as the background group by using the in-group setting tab in Gegenees software. Because of the genomic diversity in CDT-producer E. coli, we repeated this procedure with five different strains, including E. coli 53638, E. coli IHE3034, E. coli RN587/1, E. coli STEC B2F1, and E. coli STEC C165-02, which were defined as separate reference strains.

The biomarker score (max/average) setting was also used. Biomarker scores were drawn graphically and loaded into the tabular view for further data analysis. In the tabular view, a score of 1.0 is the maximum biomarker score and is considered as a signature.

Assembling signature fragments

Several overlapping fragments were obtained, based on the sequences of each reference strain. To facilitate subsequent analysis steps, the overlapping fragments were assembled using DNA Dragon software, version 1.6.0 (http://www.dna-dragon.com/).

The settings were designed with minimum overlaps (100 bases) along the diagonal length, a minimum %-identity of complete overlapping fragments, and 100% full-search parameters.

BLAST

BLAST was done with sequences for each of the five reference strains by using NCBI BLASTX (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to identify the putative protein domains. Furthermore, putative conserved domains were also detected. The results were confirmed using the UniProtKB Bank BLASTX program (http://www.uniprot.org/blast/).

Results

Strains

The sequences of 76 strains were downloaded from the NCBI site. Details regarding genome sizes, %GC content,
Table 1. Strains characteristics

| Strain               | DNA length (Mb) | cat gene | GC%  | Protein count | Gene count | Genome type, No. of subsequences/contigs | Pathotype, serotype, other characteristic | Accession No. |
|----------------------|----------------|----------|------|---------------|------------|------------------------------------------|------------------------------------------|---------------|
| Escherichia coli     | 96.0497        | +        | 50.80| 4,862         | 5,026      | Draft, 13                                | Host: homo sapiens, O91:H21              | NZ_AEZQ00000000.2 |
| Escherichia coli     | 3003           | +        | 50.70| 4,825         | 4,982      | Draft, 8                                 | I.S: water, O157:H45                    | NZ_AFAF00000000.2 |
| Escherichia coli     | 5412           | +        | 50.20| 5,670         | 5,761      | Draft, 373                               | Host: homo sapiens, SFO157               | NZ_AMUJ00000000.1 |
| Escherichia coli     | 53638          | +        | 50.99| 4,803         | 5,218      | Draft, 2                                 | EIEC, O144                              | NZ_AAKB00000000.2 |
| Escherichia coli     | ARS4:2123      | +        | 50.80| 5,105         | 5,194      | Draft, 209                               | I.S: water, O157:H45                    | NZ_AMUL00000000.1 |
| Escherichia coli     | DECE3F         | +        | 50.60| 5,541         | 5,692      | Draft, 93                                | Host: homo sapiens, SF EHEC O157:H     | NZ_AIFJ00000000.1 |
| Escherichia coli     | KTE11          | +        | 50.50| 4,109         | 4,214      | Draft, 7                                 | No published information                | NZ_ANSR00000000.1 |
| Escherichia coli     | KTE28          | +        | 50.40| 4,673         | 4,760      | Draft, 12                                | No published information                | NZ_ANSY00000000.1 |
| Escherichia coli     | KTE47          | +        | 50.60| 4,694         | 4,798      | Draft, 11                                | No published information                | NZ_ANUB00000000.1 |
| Escherichia coli     | KTE60          | +        | 50.50| 4,664         | 4,756      | Draft, 20                                | No published information                | NZ_ANUJ00000000.1 |
| Escherichia coli     | KTE137         | +        | 50.50| 4,702         | 4,789      | Draft, 99                                | No published information                | NZ_ANYA00000000.1 |
| Escherichia coli     | KTE178         | +        | 50.60| 4,973         | 5,050      | Draft, 11                                | No published information                | NZ_ANTB00000000.1 |
| Escherichia coli     | KTE180         | +        | 50.60| 4,883         | 4,966      | Draft, 112                               | No published information                | NZ_ANYR00000000.1 |
| Escherichia coli     | KTE209         | +        | 50.50| 4,702         | 4,791      | Draft, 3                                 | No published information                | NZ_ANXD00000000.1 |
| Escherichia coli     | MS 21-1        | +        | 50.40| 5,744         | 5,860      | Draft, 206                               | No published information                | NZ_ADTR00000000.1 |
| Escherichia coli     | O157:H-493-89  | +        | 50.50| 4,838         | 4,946      | Draft, 204                               | Host: homo sapiens, O157:H-             | NZ_AETY00000000.1 |
| Escherichia coli     | O157:H43T22    | +        | 50.80| 4,859         | 4,935      | Draft, 64                                | I.S: milk from healthy cattle, O157:H43 | NZ_AHZD00000000.2 |
| Escherichia coli     | RN587/1        | +        | 50.60| 4,999         | 5,108      | Draft, 73                                | EPEC, O157:H8                          | NZ_ADUS00000000.1 |
| Escherichia coli     | STEC B2F1      | +        | 50.90| 4,875         | 5,006      | Draft, 37                                | STEC, O91:H21                          | NZ_AFDQ00000000.1 |
| Escherichia coli     | STEC C165-02   | +        | 50.60| 4,891         | 5,019      | Draft, 30                                | STEC, O73:H16                          | NZ_AFDQ00000000.1 |
| Escherichia coli     | TA271          | +        | 50.70| 5,081         | 5,197      | Draft, 83                                | Host: some mammal                       | NZ_ADAZ00000000.1 |
| Escherichia coli     | TW65971        | +        | 50.30| 5,521         | 5,650      | Draft, 45                                | Host: homo sapiens, O157:H-            | NZ_AKLT00000000.1 |
| Escherichia coli     | W26            | +        | 50.60| 4,852         | 4,920      | Draft, 165                               | Host: cow, I.S: feces                   | NZ_AGIA00000000.1 |
| Escherichia albertii | TW07627        | +        | 49.90| 4,386         | 4,889      | Draft, 43                                | Diarrhea genic                          | NZ_ABKX00000000.1 |
| Escherichia coli     | APEC O1        | +        | 50.29| 4,853         | 4,968      | Complete, 3                              | ExPEC, O1:K1:H7, avian pathogenic      | NC_008563.1    |
| Escherichia coli     | IHE3034        | +        | 50.70| 4,966         | 4,753      | Complete, 1                              | ExPEC, O18:K1:H7, meningitis           | NC_017628.1    |
| Strain                  | DNA length (Mb) | cdt gene length (Mb) | cdt GC% | Protein count | Gene count | Genome type, No. of subsequences/contigs | Pathotype, serotype, other characteristic | Accession No. |
|------------------------|----------------|----------------------|---------|---------------|------------|------------------------------------------|------------------------------------------|---------------|
| Escherichia coli 042   | 5.35332        | 50.58 4,920 5,036    |         |               |            | Complete, 2                              | EAE, O44:H18                             | NC_017626.1   |
| Escherichia coli 536   | 4.93892        | 50.50 4,619 4,779    |         |               |            | Complete, 1                              | UPEC, O6:K15:H31                         | NC_008253.1   |
| Escherichia coli 55989 | 5.15486        | 50.70 4,755 5,136    |         |               |            | Complete, 1                              | EAE                                      | NC_011748.1   |
| Escherichia coli 83972 | 5.13296        | 50.60 4,795 4,905    |         |               |            | Complete, 2                              | ExPEC, UTI, OR:K5:H-                      | NC_017631.1   |
| Escherichia coli O78   | 4.79843        | 50.70 4,588 4,695    |         |               |            | Complete, 1                              | ExPEC                                   | NC_020163.1   |
| Escherichia coli ATCC 8739 | 4.74622      | 50.90 4,199 4,408    |         |               |            | Complete, 1                              | K12 derivative                           | NC_010468.1   |
| Escherichia coli B REL606 | 4.62981      | 50.80 4,200 4,361    |         |               |            | Complete, 1                              | Commensal, strain B                      | NC_012967.1   |
| Escherichia coli BL21 DE3 | 4.55895      | 50.80 4,153 4,330    |         |               |            | Complete, 1                              | Commensal, strain B                      | NC_012971.2   |
| Escherichia coli BW2952 | 4.57816        | 50.80 4,079 4,262    |         |               |            | Complete, 1                              | K12 derivative                           | NC_012759.1   |
| Escherichia coli CFT073 | 5.23143        | 50.50 5,364 5,574    |         |               |            | Complete, 1                              | ExPEC, UPEC, O6:K2:H1                    | NC_004431.1   |
| Escherichia coli DH1    | 4.63071        | 50.80 4,160 4,375    |         |               |            | Complete, 1                              | K12 derivative                           | NC_017625.1   |
| Escherichia coli E24377A | 5.24929       | 50.54 4,991 5,258    |         |               |            | Complete, 7                              | ETEC, O139:H28                           | NC_009801.1   |
| Escherichia coli ED1a   | 5.20955        | 50.70 4,911 5,321    |         |               |            | Complete, 1                              | Commensal, O81                           | NC_011745.1   |
| Escherichia coli H10407 | 5.32859        | 50.73 4,872 5,084    |         |               |            | Complete, 5                              | ETEC, O7:H11                             | NC_017633.1   |
| Escherichia coli HS     | 4.64354        | 50.80 4,374 4,626    |         |               |            | Complete, 1                              | Commensal, O9                            | NC_009800.1   |
| Escherichia coli IA11   | 4.70056        | 50.80 4,345 4,629    |         |               |            | Complete, 1                              | Commensal                               | NC_011741.1   |
| Escherichia coli IA39   | 5.13207        | 50.60 4,725 5,092    |         |               |            | Complete, 1                              | ExPEC, UPEC, O7:K1                       | NC_011750.1   |
| Escherichia coli IB1886 | 5.30828        | 50.77 5,049 5,213    |         |               |            | Complete, 6                              | ExPEC, UPEC                             | NC_022648.1   |
| Escherichia coli K-12 DH10B | 4.68614       | 50.80 4,124 4,352    |         |               |            | Complete, 1                              | K12 derivative                           | NC_010473.1   |
| Escherichia coli K-12   | 4.64165        | 50.80 4,140 4,497    |         |               |            | Complete, 1                              | Commensal, K12                           | NC_000913.3   |
| Escherichia coli MG1655 | 5.11320        | 50.80 4,213 4,436    |         |               |            | Complete, 1                              | Commensal, K12                           | NC_007779.1   |
| Escherichia coli KO11FL | 5.02717        | 50.79 4,705 4,821    |         |               |            | Complete, 2                              | Commensal                               | NC_017660.1   |
| Escherichia coli LF82   | 4.77311        | 50.70 4,376 4,545    |         |               |            | Complete, 1                              | AEIE                                    | NC_011993.1   |
| Escherichia coli LY180  | 4.8356         | 50.90 4,463 4,624    |         |               |            | Complete, 1                              | Ethanologenic E. coli                     | NC_022364.1   |
| Escherichia coli NA114  | 4.97146        | 51.20 4,873 4,975    |         |               |            | Complete, 1                              | ExPEC, UPEC                              | NC_017644.1   |
| Escherichia coli O7:K1 CE10 | 5.37873       | 50.58 5,080 5,269    |         |               |            | Complete, 5                              | ExPec, Neonatal meningitis, O7:K1        | NC_017646.1   |
| Escherichia coli O26:H11 | 5.85553        | 50.66 5,155 5,985    |         |               |            | Complete, 5                              | EHEC, O26:H11                            | NC_013361.1   |
| Escherichia coli O55:H7 | 5.45235        | 50.48 5,117 5,367    |         |               |            | Complete, 2                              | EPEC, O55:H7                             | NC_013941.1   |
| Escherichia coli O83:H1 | 4.89488        | 50.71 4,582 4,690    |         |               |            | Complete, 2                              | AIEC                                    | NC_017634.1   |
| Escherichia coli O11128 | 4.90512        | 50.69 4,603 4,821    |         |               |            | Complete, 2                              | EPEC                                    | NC_017633.1   |
| Escherichia coli P12b   | 4.93329        | 50.90 4,379 4,567    |         |               |            | Complete, 1                              | O15:H17                                 | NC_017663.1   |
| Escherichia coli PMV 1  | 5.21093        | 50.67 4,979 5,257    |         |               |            | Complete, 2                              | ExPEC, O18:K1                            | NC_022370.1   |
| Strain                        | DNA length (Mb) | cdt gene | GC%  | Protein count | Gene count | Genome type, No. of subsequences/contigs | Pathotype, serotype, other characteristic | Accession No. |
|------------------------------|-----------------|----------|------|---------------|------------|------------------------------------------|-------------------------------------------|--------------|
| *Escherichia coli* S88       | 5.16612         | -        | 50.66| 4,823         | 5,187      | Complete, 2                              | ExPEC, Neonatal Meningitis, O45:K1:H7     | NC_011742.1  |
| *Escherichia coli* SE11      | 5.15563         | -        | 50.75| 4,996         | 5,103      | Complete, 7                              | Commensal, O152:H28                       | NC_011415.1  |
| *Escherichia coli* SE15      | 4.83968         | -        | 50.71| 4,486         | 4,592      | Complete, 2                              | Commensal, O150:H5                        | NC_013654.1  |
| *Escherichia coli* SMS-3-5   | 5.21538         | -        | 50.50| 4,912         | 5,127      | Complete, 5                              | Environmental isolate                     | NC_010498.1  |
| *Escherichia coli* UM146     | 5.10756         | -        | 50.61| 4,783         | 4,891      | Complete, 2                              | AIEC (adherent invasive)                  | NC_017632.1  |
| *Escherichia coli* UMND26    | 5.3582          | -        | 50.64| 5,010         | 5,294      | Complete, 3                              | ExPEC, UPEC, O7:K1                        | NC_011751.1  |
| *Escherichia coli* UMNK88    | 5.66676         | -        | 50.74| 5,607         | 5,754      | Complete, 6                              | Porcine ETEC, O149                        | NC_017641.1  |
| *Escherichia coli* UTI89     | 5.17997         | -        | 50.61| 5,162         | 5,272      | Complete, 2                              | ExPEC, UPEC, O18:K1:H7                    | NC_007946.1  |
| *Escherichia coli* W         | 5.00886         | -        | 50.78| 4,602         | 4,876      | Complete, 3                              | Commensal, ATCC 9637                      | NC_017635.1  |
| *Escherichia coli* Xuzhou21  | 5.51674         | -        | 50.38| 5,179         | 5,294      | Complete, 3                              | EHEC, O157:H7                            | NC_017906.1  |
| *Escherichia fergusonii* ATCC 35469 | 4.64386   | -    | 49.88| 4,314         | 4,543      | Complete, 2                              | I.S: Feces, human                        | NC_011740.1  |

I.S., isolation source; EIEC, enteroinvasive *E. coli*; EHEC, enterohemorrhagic *E. coli*; EPEC, enteropathogenic *E. coli*; STEC, Shiga toxin-producing *E. coli*; ExPEC, extraintestinal pathogenic *E. coli*; EAEC, enteroaggregative *E. coli*; UPEC, uropathogenic *E. coli*; ETEC, enterotoxigenic *E. coli*; AIEC, adherent invasive *E. coli*.

Fig. 1. Phylogenetic heat-plot overview of multiple-genome alignments. A heat plot based on a 200/100 BLASTN fragmented alignment was performed with Gegenees software. Six distinct genomic groups (T1–T6) recognized in cytolethal distending toxin (CDT)+ strains were observed sporadically among the strains that were studied, revealing the heterogeneous genomic nature of CDT-producing *Escherichia coli*.  

www.genominfo.org
Fig. 2. Phylogram overview. A phylogram was generated using SplitsTree 4 software, using the neighbor-joining method and a distance-matrix Nexus file exported from Gegenees software. The Escherichia albertii TW07627 and Escherichia fergusonii ATCC 35469 strains were set as out-groups. In addition, six unique groups (T1–T6) were analyzed. In the phylogenetic overview, a sporadic pattern of cytolethal distending toxin (CDT)–producing strains was observed, as were specific clades. These strains were related and their similarities were shown. CDT+ strains are shown in boxes. The Escherichia coli strains that were set as reference strains for biomarker-detection studies are indicated with red arrows.
the number of encoded proteins, encoded genes, genome type, pathotype, serotype, other characteristics, and accession numbers are summarized in Table 1. Most data presented were extracted from NCBI GenBank and UniProt Bank and some information was extracted from original articles [14, 15]. The genomes of 24 strains were drafted, and a reordering process of the draft genomes was performed. Twenty-five CDT+ E. coli strains were analyzed, including E. albertii TW07627.

**Phylogenetic analysis**

A heat-plot based on a 200/100 BLASTN fragmented alignment drawn with GeneNees software is shown in Fig 1. A phylogenetic overview is also shown in the heat-plot. A more detailed phylogram was constructed with SplitsTree 4 software, as shown in Fig. 2.

CDT-producer E. coli strains were displayed a sporadic, phylogenomic pattern in the heat-plot, with a lack of a consensus pattern. Six distinct genomic groups of CDT+ strains (T1 to T6 in Fig. 2) were shown in the phylogram, all of which were sporadic among the strains in Fig 1. As a sporadic pattern of CDT-producing strains was observed in the bacterial population in the phylogram for specific clades, these strains were related and some degrees of similarity were also found.

**Signature sequences in the target group**

In total, 1,527 fragments representing 3.0% of the E. coli 53638-strain genome were identified as signature sequences. Biomarkers were restricted to 21 highly significant regions, designated A to U. When E. coli IHE3034 was set as the reference strain, 220 signature sequences (0.4%) were detected. Biomarkers were identified in six regions, designated A to F. However, 1,512 (2.9%) signature fragments were obtained, which were restricted to 18 regions (A to R) in the genome of E. coli RN5871 when it was regarded as the reference strain. Moreover, 620 biomarker fragments (1.2%) were detected in the genome of E. coli STEC B2F1 when it was set as the reference strain, 16 biomarker regions (A to P) were recognized. In addition, when E. coli STEC C165-02 was used as the reference strain, 593 signature fragments (1.1%) were identified, which were restricted to eight regions (A to H). The signature regions for all reference strains are shown in Fig. 3, separately. In addition, the biomarker designation, domain description, BLASTX results and related putative conserved domains for each reference strain are provided in Supplementary Tables 1-6.

**Conserved signature proteins**

The most common biomarker proteins were distinguished by comparing BLASTX results for all reference strains fragments (Table 2). The signature proteins identified included: CDT, holin, lambda-family proteins, nuclease, phage integrase family proteins, phage tail tape measure family proteins, putative membrane proteins, regulatory proteins, restriction-modification system proteins, tail fiber assembly proteins, baseplate assembly proteins, tail fiber protein and other prophage tail related proteins, terminases and transferases. The nucleotide sequences of some proteins including anti-termination proteins, prophage DNA packaging and binding proteins, transposase and DNA transposition proteins, scaffold proteins, recombination-related domains, putative phage-replication proteins, hemolysin, helicase, glycol transferase, and glycohydrolase superfamilies, were detected as biomarkers in the target group, although these BLASTX results were not observed in all reference strains. Presumably, CDT-producer E. coli strains possess several hypothetical proteins whose functions are not yet defined and might be conserved proteins. The existence of these DNA biomarker sequences in reference strains is clear; however, the related proteins in some strains have not been determined.

**Significant putative conserved domains and superfamilies**

In the era of modern vaccines, finding conserved domains or epitopes has a great therapeutic value. Putative conserved domains were described as non-specific hits (NH), specific hits (SH), and multi-domains (MD), and it was shown in Supplementary Tables 1-6. The putative conserved domains and superfamilies that were associated with some signature proteins are shown below.

- NH: PRK15251, DUF4102, CdtB, CdtxinA, INT_P4, HP1_INT_C, Phage_integrase, INT_Lambda_C, Phage_integ_N, Methylase_S, Caudo_TAP, phage_tail_N, Tail_P2_I, gpl, phage_term_2, Terminase_3, Terminase_5, M, Phage_term_smal, COG5525, Terminase_GpA, Phage_Nu1, dexA, Phage_holin_2, DUF3751, Phage_attach, dcm, DNA_methylase, Cyt_C5_DNA_methylase, Dcm, Glycos_transf_2, and CESA_like
- SH: INT_REC_C, PhageMin_Tail, COG4220, Phage_fiber_2, HSDR_N, Glycos_transf_2, GT_2_like_d, PRK10018, and PLN02726
- MD: PRK09692, int, recomb_XeR_C, XerD, xerC, HsdS, N6_Mtase, HsdM, hsdM, rumA, P, Terminase_6, COG5484, PLN03114, COG5301, COG0610, hsdr, PRK10458, PRK10073, Glyco_transf_2_3, WcaA, PRK10073, and PTZ00240
- Superfamilies: RICIN superfamily, EEP superfamily, DNA_BRE_C superfamily, DUF4102 superfamily, Phage_integ_N superfamily, MCP_signal superfamily,
Fig. 3. Biomarker regions. Biomarker regions were illustrated in the whole-genome sequences of five different reference strains including *Escherichia coli* 53638, *E. coli* IHE3034, *E. coli* RN5871, *E. coli* STEC B2F1, and *E. coli* STEC C165-02. The biomarker score (max/average) setting was used. A score of 1.0 is the maximum biomarker score, which was considered to represent a signature sequence, as indicated in green. STEC, Shiga toxin-producing *E. coli*. 
| Signature protein                  | Reference strain | Reference strain | Reference strain | Reference strain | Reference strain |
|-----------------------------------|------------------|------------------|------------------|------------------|------------------|
| Cytolethal distending toxin       | 53638 Escherichia coli | IHE3034 Escherichia coli | RN587/1 Escherichia coli | STEC_B2F1 Escherichia coli | STEC_C165_02 Escherichia coli |
| Cytolethal distending toxin A     |                  |                  |                  |                  |                  |
| Cytolethal distending toxin B     |                  |                  |                  |                  |                  |
| Cytolethal distending toxin C     |                  |                  |                  |                  |                  |
| Cytolethal distending toxin subunit C |                |                  |                  |                  |                  |
| Holin                             | Phage holin, lambda family | Holin, lambda family | Holing | Phage holin, lambda family |
| Nuclease                          | Exodeoxyribonuclease B | Exonuclease family protein | Exonuclease family protein | Endonuclease/Exonuclease family protein | Restriction endonuclease family protein |
| Phage integrase                   | Phage integrase | Integrate for prophage CP-933T | Integrate Phage integrase family | Integrate Phage integrase family protein | Integrate Prophage CP4-57 integrase |
| Putative membrane protein         | Putative membrane protein | Hypothetical protein Ec53638_1156, [membrane protein] | Outer membrane autotransporter barrel domain protein | Putative membrane protein | Putative membrane protein |
| Regulatory proteins               | Phage regulatory protein Cro | Putative transcriptional regulator DicA157 | Regulatory protein CII Prophage CP4-57 regulatory protein family protein | Transcriptional regulator, AraC family | 4-Hydroxyphenylacetate catabolism regulatory protein HpaA Prophage CP4-57 regulatory protein family protein |
| Restriction-modification system   | Putative type I restriction-modification system, S subunit Type I restriction-modification system specificity subunit Type I restriction-modification enzyme, R subunit Type I restriction-modification system, M subunit | Type II restriction enzyme EcoRII Modification methylase EcoRII Type I restriction enzyme specificity protein Type I restriction-modification system, M subunit | Type I restriction-modification DNA specificity domain protein | Type I restriction-modification system specificity determinant Type III restriction enzyme, res subunit |
M Javadi, et al. Signature Prediction of CDT-Producing E. coli Strains

Table 2. Continued

| Signature protein | Reference strain |
|-------------------|------------------|
|                   | *Escherichia coli* | *Escherichia coli* | *Escherichia coli* | *Escherichia coli* | *Escherichia coli* |
|                   | 53638             | IHE3034           | RN587/1            | STEC_B2F1         | STEC_C16S_02      |
| Tail fiber assembly family, baseplate assembly proteins, Tail fiber protein and Tail tape measure protein | Tail fiber assembly protein | Tail fiber protein | Tail fiber family | Tail fiber assembly family protein, Hypothetical protein | Tail fiber assembly family protein |
|                   | Phage P2 baseplate assembly protein | Phage tail tape measure protein | ECNRN5871_3204, [tail fiber assembly protein, baseplate assembly protein] V, W | Phage tail fiber assembly protein | Phage tail fiber assembly protein |
|                   | gpV                |                    | ECNRN5871_3204, [tail fiber assembly protein, baseplate assembly protein] V, W | Tail fiber domain protein | Phage tail fiber repeat family protein |
|                   | Putative tail fiber protein | Tail fiber protein | Tail fiber family | Phage tail fiber family protein | Phage tail fiber family protein |
|                   | Tail fiber         | Tail fiber protein | Tail fiber family | Phage tail fiber family protein | Phage tail fiber family protein |
|                   | Phage tail tape measure protein | Phage tail tape measure protein | Tail fiber domain protein | Phage tail measure protein, lambda family | Phage tail measure protein, lambda family |
|                   | Tail fiber measure protein family | Tail fiber measure protein family | Tail fiber measure protein family | Phage tail measure protein, lambda family | Phage tail measure protein, lambda family |
|                   |                     |                    |                     |                     |                     |
| Terminase          | Phage terminase    | Phage terminase    | Phage terminase     | Terminase B protein | Terminase B protein |
|                   | large subunit      | subunit            | large subunit       | domain protein      | domain protein      |
|                   | Terminator         | Terminator         | Terminator          | Terminator B protein | Terminator B protein |
|                   |                    |                    |                    |                    |                    |
| Transferase        | Pyruvyl transferase | Glycosyl transferase domain protein, group 2 family | Hypothetical protein ECNRN5871_3204, [nucleotidyl transferase, PF08843 family] | Putative teichuronic acid biosynthesis glycosyltransferase tuaG | Acetylated transferase family protein |
|                   | Glycosyl transferase, sugar-binding region containing DXD motif | Glycosyl transferase, sugar-binding region containing DXD motif | D12 class N6 adenine-specific DNA methyltransferase family protein | Glucose-1-phosphate thymidylyltransferase | Hypothetical protein ECSTEC16S02_1295, [acytylated transferase] |
|                   |                    |                    | Hypothetical protein ECNRN5871_0025, [N-acytylated transferase, CNS] | RTX toxin acyltransferase family protein | Acetylated transferase family protein |
|                   |                    |                    |                    | Acetyl-CoA acetyltransferase | Acetylated transferase family protein |
|                   |                    |                    |                    | Hypothetical protein | Acetylated transferase family protein |
|                   |                    |                    |                    | ECSTEC16S02_1295, [acytylated transferase] | Acetylated transferase family protein |

*There are lots of hypothetical proteins with unknown function in desired genome which they have mentioned but their roles have not been defined yet.

Methylase_Superfamily, Caudovirales_TAP_superfamily, Phage_tail_Nsuperfamily, Tail_P2_1superfamily, Terminase_3superfamily, Terminase_5superfamily, Phage_term_smal superfamly, Terminase_GpA superfamily, Phage_Nu1 superfamily, DnaG-like-exosuperfamily, Phage_holin_2superfamily, DUF3751 superfamily, Phage_fiber_2superfamily, Gifsy-2 superfamily, HSDR_Nsuperfamily, Cyt_C5 DNA methylase superfamily, MethyltransferaseD12superfamily, Glyco TRANSF_GTA type superfamily, and Glyco TRANSF_GTA typesuperfamily

**Discussion**

The synchronic evolution of bacterial pathogens and virulence-associated determinants encoded by horizontally transferred genetic elements has been observed in several species. However, *E. coli* is a normal member of the intestinal microflora of humans and animals. *E. coli* strains have acquired virulence factors by the attainment of particular genetic loci through horizontal gene transfer, transposons, or phages. These elements frequently encode multiple factors that enable bacteria to colonize the host and initiate disease development [16]. CDTs belong to one such class of virulence-associated factors. CDT was first identified in *E. coli* by Johnson and Lior in 1988 [17]; since then several studies have been reported that CDTs can be produced by intestinal and extra-intestinal pathogenic bacteria [18].

In this study, the genomes of 25 CDT+ *E. coli* strains were acquired from several gene banks. Multiple genome comparisons with 49 CDT+ *E. coli* strains, including EPEC (enteropathogenic *E. coli*), ETEC (enterotoxigenic *E. coli*), STEC (Shiga toxin-producing *E. coli*), EAEC (enteroagg-
regenerative E. coli), EIEC (enteroinvasive E. coli), AIEC (adherent invasive E. coli), UPEC (uropathogenic E. coli), ExPEC (extraintestinal pathogenic E. coli), EHEC (entero-hemorrhagic E. coli), environmental strains and commensal strains were performed.

In fact, phylogenetic analysis based on whole-genome information is more accurate than those based on one gene or a set of limited genes. In this study, CDT-producing strains were not shown a phylogenetic relationship or pattern. Indeed, while they might carry the same or similar virulence gene sets, they also possess their own divergent genomic structures. This is probably because of their complex and distinct evolutionary pathways, indicating an independent acquisition of mobile genetic elements during their evolution.

The sporadic pattern in the phylogenomic dendrogram confirmed previous findings that CDT+ strains are heterogeneous. The heterogeneous nature of CDT-producing strains might arise from horizontal gene transfer through mobile genetic elements. These genetic exchanges that occur in bacteria provide genomic diversity and versatility [19].

A significant challenge in comparative genomics is the utilization of large datasets to identify specific sequence signatures that are biologically important or are useful in diagnosis [4, 20]. In this study, we define CDT-producing E. coli as the target group and found regions that were conserved that could serve as genomic signatures for the target group. Because of the heterogeneous genomic nature of CDT+ E. coli, five reference strains were selected instead of one, including EIEC, ExPEC, EPEC, STEC B2F1, and STEC C165-02. Moreover, in the phylogenomic overview, these five reference strains were selected from different clades of the phylogenetic tree, representing the T1–T6 groups.

The findings were presented in this study indicate that the major conserved biomarkers beyond CDT were exonucleasease, phage integrase, putative membrane, and tail-fiber proteins. Furthermore, with signature proteins of a targeted group, it was shown that phage-related proteins and virulence-associated factors could be commonly transferred by phages. Moreover, in the putative conserved domains of biomarker proteins, phage-related superfamilies were frequently observed. As a result, cdt genes were used as a signature sequences in CDT-producing E. coli strains, and it was shown that they can be used as a powerful biomarker.

In this study, the most significant signature proteins in the five E. coli strains were identified using in-silico whole-genome sequences. It was demonstrated that conserved signature proteins were expressed in a wide range of pathogenic bacterial strains, which could be used in future studies in a broad range of research applications and in modern vaccine-design strategies.

Supplementary materials

Supplementary data including six tables can be found with this article online at http://www.genominfo.org/src/sm/gni-15-69-s001.pdf.

Acknowledgments

This work was supported financially by the Pasteur Institute of Iran. We would like to thank Editage (http://www.editage.com) for English language editing.

References

1. Lara-Tejero M, Galan JE. Cytolethal distending toxin: limited damage as a strategy to modulate cellular functions. Trends Microbiol 2002;10:147-152.
2. Tóth I, Nougyérdei JP, Dobrindt U, Ledger TN, Boury M, Morabito S, et al. Cytolethal distending toxin type I and type IV genes are framed with lambloid prophage genes in extra-intestinal pathogenic Escherichia coli. Infect Immun 2009;77:492-500.
3. Lara-Tejero M, Galán JE. A bacterial toxin that controls cell cycle progression as a deoxyribonuclease I-like protein. Science 2000;290:354-357.
4. Agren J, Sundström A, Häfström T, Segerman B. Gegenes: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. PLoS One 2012;7:e39107.
5. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 2003;425:798-804.
6. Dubchak I, Poliakov A, Kislyuk A, Brudno M. Multiple whole-genome alignments without a reference organism. Genome Res 2009;19:682-689.
7. Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. Cactus: algorithms for genome multiple sequence alignment. Genome Res 2011;21:1512-1528.
8. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 2004;14:708-715.
9. Rausch T, Emde AK, Weese D, Döring A, Notredame C, Reinert K. Segment-based multiple sequence alignment. Bioinformatics 2008;24:i187-i192.
10. Rissman AI, Mau B, Biehl BS, Darling AE, Glaser JD, Perna NT. Reordering contigs of draft genomes using the Mauve aligner. Bioinformatics 2009;25:2071-2073.
11. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 2010;5:e11147.
12. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 2004;14:1394-1403.
13. Kloepfer TH, Huson DH. Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evol Biol 2008;8:22.
14. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced *Escherichia coli* genomes. *Microb Ecol* 2010; 60:708-720.

15. Gardner SN, Hall BG. When whole-genome alignments just won’t work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. *PLoS One* 2013;8:e81760.

16. Asakura M, Hinenoya A, Alam MS, Shima K, Zahid SH, Shi L, *et al*. An inducible lambdoid prophage encoding cytolethal distending toxin (Cdt-I) and a type III effector protein in enteropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A* 2007; 104:14483-14488.

17. Johnson WM, Lior H. A new heat-labile cytolethal distending toxin (CLDT) produced by *Escherichia coli* isolates from clinical material. *Microb Pathog* 1988;4:103-113.

18. Kim JH, Kim JC, Choo YA, Jang HC, Choi YH, Chung JK, *et al*. Detection of cytolethal distending toxin and other virulence characteristics of enteropathogenic *Escherichia coli* isolates from diarrheal patients in Republic of Korea. *J Microbiol Biotechnol* 2009;19:525-529.

19. Oloomi M, Bouzari S. Molecular profile and genetic diversity of cytolethal distending toxin (CDT)-producing *Escherichia coli* isolates from diarrheal patients. *APMIS* 2008;116:125-132.

20. Edwards DJ, Holt KE. Beginner’s guide to comparative bacterial genome analysis using next-generation sequence data. *Microb Inform Exp* 2013;3:2.
SUPPLEMENTARY INFORMATION

In Silico Signature Prediction Modeling in Cytolethal Distending Toxin-Producing Escherichia coli Strains

Maryam Javadi, Mana Oloomi*, Saeid Bouzari

Department of Molecular Biology, Pasteur Institute of Iran, Tehran 13164, Iran

http://www.genominfo.org/src/sm/gni-15-69-s001.pdf
### Supplementary Table 1. Signature details based on *Escherichia coli* 53638 reference

| Region | Biomarker range in *Escherichia coli* 53638 genome between | Protein obtained by NCBI Blastx [identical protein in other *Escherichia coli* strains] | Blast score | Blast Identity (%) | Blast E-value | Putative conserved domains non-specific hits (NH), specific hits (SH), multi domains (MD) |
|--------|-----------------------------------------------------------|---------------------------------------------------------------|-------------|--------------------|---------------|---------------------------------------------------------------------------------------|
| A      | 0 Mbp to 1 Mbp                                           | Invasion plasmid antigen                                       | 829         | 73                 | 0.0           | SH: NEL, NEL superfamily, MD: PRK15387, COG4886, PLN00113                              |
|        |                                                           | Replicative DNA helicase homolog                               | 885         | 100                | 0.0           | SH: DnaB, DnaB superfamily, NH: DnaB_C, RecA-like_NTPases superfamily, MD: phage_DnaB, PRK06904, DnaB, DnaB_C |
|        |                                                           | Replication protein                                            | 199         | 100                | 6e-60         | SH: HTH, XRE, NH: HTH_19, HTH_XRE superfamily, MD: PRK09706, HipB                     |
|        |                                                           | Hypothetical protein Ec53638_0018                               | 562         | 100                | 0.0           | NH: PHage_attach, Gifsy-2 superfamily                                                |
|        |                                                           | Hypothetical protein Ec53638_0019                               | 94.4        | 98                 | 1e-24         | NH: Packaging_Fl, Packaging_Fi superfamily                                           |
|        |                                                           | Conserved hypothetical protein                                 | 249         | 100                | 1e-82         | NH: Phage_CI_repr, Phage_CI_repr superfamily                                         |
|        |                                                           | Putative bacteriophage protein                                 | 745         | 100                | 0.0           | SH: ParE, NH: Plasmid_stabil, Plasmid_stabil superfamily                           |
|        |                                                           | Addiction module antitoxin, Axe family                         | 154         | 100                | 1e-47         | NH: Plasmid_stabil, Plasmid_stabil superfamily                                      |
|        |                                                           | Plasmid stabilization system protein, RelE/ParE family         | 216         | 100                | 3e-71         | MD: PRK09709                                                                         |
|        |                                                           | Phage Head-Tail Attachment                                     | 252         | 99                 | 2e-85         | NH: Phage_attach, Gifsy-2 superfamily                                                |
|        |                                                           | Putative DNA-packaging protein                                 | 223         | 100                | 5e-74         | NH: Packaging_Fl, Packaging_Fi superfamily                                           |
|        |                                                           | Bacteriophage CI repressor protein                             | 267         | 100                | 1e-91         | NH: Phage_CI_repr, Phage_CI_repr superfamily                                         |
|        |                                                           | Putative tail fiber protein                                    | 481         | 100                | 6e-171        | NH: Phage_tail_N, Phage_tail_N superfamily                                          |
|        |                                                           | Conserved domain protein                                        | 77.4        | 100                | 1e-19         | SH: ParE, NH: Plasmid_stabil, Plasmid_stabil superfamily                           |
|        |                                                           | Hypothetical bacteriophage protein                             | 205         | 99                 | 7e-68         | NH: Plasmid_stabil, Plasmid_stabil superfamily                                      |
|        |                                                           | Putative tail component of prophage                            | 203         | 98                 | 6e-67         | NH: Plasmid_stabil, Plasmid_stabil superfamily                                      |
|        |                                                           | Putative tail fiber protein                                    | 99.4        | 100                | 3e-26         | MD: PRK09709                                                                         |
|        |                                                           | Exodeoxyribonuclease 8                                         | 153         | 100                | 2e-44         | NH: Gam, Gam superfamily                                                            |
|        |                                                           | Host-nuclease inhibitor protein Gam                             | 153         | 100                | 2e-48         | NH: Gam, Gam superfamily                                                            |
|        |                                                           | Gifsy-1 prophage VmTH                                          | 226         | 100                | 3e-70         | MD: PRK09709                                                                         |
|        |                                                           | Hypothetical protein Ec53638_0122, [gp41 domain protein]       | 278         | 100                | 3e-97         | SH: Plasmid_stabil, Plasmid_stabil superfamily                                      |
|        |                                                           | Putative phage protein                                         | 202         | 100                | 2e-68         | NH: Plasmid_stabil, Plasmid_stabil superfamily                                      |
|        |                                                           | Conserved hypothetical protein                                 | 197         | 100                | 5e-67         | NH: Plasmid_stabil, Plasmid_stabil superfamily                                      |
|        |                                                           | Putative Prophage Qin DNA packaging protein NU1 homolog        | 204         | 100                | 8e-71         | NH: Plasmid_stabil, Plasmid_stabil superfamily                                      |
|        |                                                           | Prophage minor tail protein Z (GP2)                            | 195         | 100                | 8e-65         | NH: Plasmid_stabil, Plasmid_stabil superfamily                                      |
|        |                                                           | Putative bacteriophage protein                                 | 142         | 100                | 5e-44         | NH: Plasmid_stabil, Plasmid_stabil superfamily                                      |
|        |                                                           | Phage antitermination Q type 1 family                          | 135         | 100                | 5e-43         | NH: Plasmid_stabil, Plasmid_stabil superfamily                                      |
|        |                                                           | Conserved hypothetical protein                                 | 107         | 100                | 2e-31         | NH: DUF1133, DUF1133 superfamily                                                   |
| B      | 0 Mbp to 1 Mbp                                           | No significant results                                         |             |                    |               |                                                                                      |
| C      | 0 Mbp to 1 Mbp                                           | Hypothetical protein Ec53638_3910                               | 130         | 100                | 1e-41         |                                                                                      |
| D      | 0 Mbp to 1 Mbp                                           | No significant results                                         |             |                    |               |                                                                                      |
| E      | 1 Mbp to 2 Mbp                                           | gp27                                                           | 343         | 100                | 9e-119        | NH: DUF3486, DUF3486 superfamily                                                   |
|        |                                                           | Putative DNA packaging protein Gp17                            | 405         | 100                | 3e-139        | NH: COG4373, Terminase_1 superfamily, MD: 17                                       |
|        |                                                           | Tail fiber                                                    | 271         | 100                | 3e-91         | NH: Tail_P2_L, gpl, Tail_P2_L superfamily                                           |
|        |                                                           | Phage protein                                                 | 282         | 100                | 5e-92         | NH: COG4373, Terminase_1 superfamily, MD: P, Terminase_6                             |
|        |                                                           | Putative DNA packaging protein Gp17                            | 481         | 100                | 2e-171        | NH: COG4373, Terminase_1 superfamily, MD: P, Terminase_6                             |
|        |                                                           | Hypothetical protein Ec53638_1156, [membrane protein]          | 275         | 100                | 1e-93         | NH: Mor, Mor superfamily                                                           |
|        |                                                           | DNA-binding protein RdgB                                       | 115         | 100                | 1e-32         |                                                                                      |
|        |                                                           | Hypothetical protein Ec53638_1176                              | 177         | 100                | 1e-57         |                                                                                      |
|        |                                                           | Phage tail sheath protein                                     | 340         | 100                | 1e-116        | NH: Fl, COG3497, Phage sheath_1 superfamily, NH: GATase1_DI-1, GAT_1 superfamily  |
|        |                                                           | Baseplate                                                     | 217         | 100                | 1e-72         | NH: plasmid shear_v, gplV, Phage_base_v superfamily                               |
|        |                                                           | Bcv gene product                                              | 274         | 100                | 2e-93         |                                                                                      |
|        |                                                           | Conserved hypothetical protein                                 | 179         | 100                | 2e-57         |                                                                                      |
| F      | 1 Mbp to 2 Mbp                                           | Putative type I restriction-modification system, S subunit     | 352         | 100                | 9e-120        |                                                                                      |
**Supplementary Table 1.** Signature details based on *Escherichia coli* 53638 reference

| | 1 Mbp to 2 Mbp | 2 Mbp to 3 Mbp | 3 Mbp to 4 Mbp |
|---|---|---|---|
| **G** | **H** | **I** | **J** |
| iSCps8, transposase | Cytolethal distending toxin A | Transposase, iSCps8, transposase | Transposase, iSCps8, transposase |
| | Cytolethal distending toxin B | | |
| | Cytolethal distending toxin subunit C | | |
| | Hypothetical protein Ec53638_1905 | | |
| | Putative phage protein | | |
| **K** | **L** | **M** | **N** |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
### Supplementary Table 1. Signature details based on *Escherichia coli* 53638 reference

|   | 3 Mbp to 4 Mbp | 4 Mbp to 5 Mbp |
|---|----------------|----------------|
| P | Invasion plasmid antigen | 836 | 83 | 0.0 | SH: NEL, NEL superfamily, MD: PRK15370 |
|   | Putative bacteriophage protein | 528 | 100 | 0.0 | NH: phage_term_2, Terminase_3, Terminase_3 superfamily |
|   | Phage terminase large subunit | 937 | 100 | 0.0 | NH: V, V superfamily |
|   | Putative bacteriophage protein | 317 | 100 | 3e-100 | |
|   | Putative tail fiber protein | 574 | 100 | 0.0 | |
|   | Hypothetical protein Ec53638_3420 | 77.0 | 100 | 6e-18 | |
|   | Putative tail component of prophage | 239 | 99 | 3e-77 | |
|   | Putative bacteriophage protein | 688 | 99 | 0.0 | |
|   | Hypothetical protein Ec53638_3782 | 478 | 100 | 1e-174 | |
|   | Conserved hypothetical protein | 337 | 100 | 2e-118 | NH: DedA, PRK10847, SNARE_assoc superfamily |
|   | Bacteriophage lysis protein | 52.8 | 100 | 2e-09 | MD: PRK09709 |
|   | Exodeoxyribonuclease 8 | 162 | 100 | 7e-48 | NH: Gam, Gam superfamily |
|   | Host-nuclease inhibitor protein Gam | 143 | 100 | 9e-45 | |
|   | gpH | 180 | 100 | 6e-56 | |
|   | Bcv gene product | 98.6 | 57 | 5e-26 | |
|   | Tail fiber assembly protein | 102 | 100 | 3e-28 | |
|   | Hypothetical protein Ec53638_3785 | 277 | 100 | 1e-98 | NH: DUF1627, DUF1627 superfamily |
|   | Phage Mu protein F like protein | 266 | 100 | 5e-93 | NH: Phage_Mu_F, Phage_Mu_F superfamily, MD: COG2369 |
|   | Putative tail fiber protein | 146 | 100 | 5e-46 | |
|   | Tail fiber assembly protein | 137 | 100 | 5e-43 | NH: Caudo_TAP, Caudo_TAP superfamily |
|   | Conserved hypothetical protein | 130 | 100 | 2e-40 | NH: DUF1133, DUF1133 superfamily |
|   | Putative phage protein | 141 | 100 | 2e-45 | |

| Q | 3 Mbp to 4 Mbp | 4 Mbp to 5 Mbp |
|---|----------------|----------------|
|   | Transcriptional regulator, AlpA family | 114 | 98 | 1e-32 | NH: Phage_AlpA, Phage_AlpA superfamily, MD: PRK09692 |
|   | Hypothetical protein Ec53638_3914 | 108 | 100 | 1e-30 | SH: DUF4102, DUF4102 superfamily, NH: INT_P4, DNA_BRE_C superfamily, MD: PRK09692 |
|   | Prophage integrase | 256 | 100 | 4e-86 | |
|   | Conserved hypothetical protein | 379 | 100 | 2e-131 | |
|   | Hypothetical protein Ec53638_3910 | 119 | 100 | 4e-37 | |

| R | 3 Mbp to 4 Mbp | 4 Mbp to 5 Mbp |
|---|----------------|----------------|
|   | Putative tail component of prophage | 80.1 | 100 | 1e-17 | SH: NEL, NEL superfamily, MD: PRK15370, COG4886 |
|   | Invasion plasmid antigen | 821 | 74 | 0.0 | |
|   | Hypothetical bacteriophage protein | 412 | 100 | 2e-148 | |
|   | Putative tail fiber protein | 666 | 100 | 0.0 | |
|   | Putative tail component of prophage | 195 | 99 | 6e-60 | |
|   | Phage antitermination Q type 1 family | 141 | 100 | 5e-44 | |
|   | Hypothetical protein Ec53638_4012 | 81.6 | 100 | 8e-22 | NH: Phage_antitermQ, Phage_antitermQ superfamily |
|   | IS CPS8, transposase | 809 | 100 | 0.0 | SH: Transposase_20, Transposase_20 superfamily, MD: COG3547 |
|   | Hypothetical bacteriophage protein | 393 | 100 | 6e-137 | MD: PRK05643, DnaN |
|   | Conserved hypothetical protein | 496 | 100 | 3e-176 | SH: ORF6N, ORF6N superfamily |
|   | BRO family, N- domain protein | 516 | 100 | 0.0 | SH: Bro-N, NH: Bro-N, Bro-N superfamily, MD: COG3617 |
|   | Conserved domain protein | 129 | 100 | 1e-20 | NH: Phage_NinH, Phage_NinH superfamily |
|   | Hypothetical protein Ec53638_4059 | 83.6 | 100 | 4e-32 | |
|   | Putative tail fiber protein | 113 | 100 | 2e-23 | |
|   | Putative Prophage Qin DNA packaging protein NU1 homolog | 134 | 100 | 6e-42 | NH: COG4220, Phage_Nu1, Phage_Nu1 superfamily |
|   | Hypothetical bacteriophage protein | 96.3 | 100 | 2e-28 | |

| S | 4 Mbp to 5 Mbp | 4 Mbp to 5 Mbp |
|---|----------------|----------------|
|   | Putative membrane protein | 944 | 100 | 0.0 | NH: MATE_Wxz_like, MATE_Wxz_like superfamily, NH: Polysacc_synt, Polysacc_synt superfamily, MD: RfbX, spore_V_B |
|   | Conserved hypothetical protein | 909 | 100 | 0.0 | NH: FrhB_FdhB_C, FrhB_FdhB_C superfamily, MD: PRK09326, FrhB |
|   | Pyruvyl transferase | 774 | 100 | 0.0 | SH: PS_pyruv_trans, PS_pyruv_trans superfamily |
### Supplementary Table 1. Signature details based on *Escherichia coli* 53638 reference

| Glycosyl transferase domain protein, group 2 family | 202 | 100 | 1e-57 | NH: Glyco_transf_GTA_type, NH: Glycos_transf_2, PLN02726, PRK10018, Glyco_transf_GTA_type superfamily, MD: WcaA, PRK10073, Glyco_transf_2_PgaC_icaA, PTZ00260 |
| Glycosyltransferase, sugar-binding region containing DXD motif | 539 | 99 | 1e-177 | MD: Caps_synth, OCH1 |
| Putative membrane protein Cps23Fh | 680 | 100 | 0.0 | SH: EpsG, T7MR_DISM_7TM superfamily |
| UDP-galactopyranosyl mutase WfbU | 695 | 100 | 0.0 | SH: Glyco_transf_GTA_type, NH: Glycos_transf_2, PRK10018, PLN02726, Glyco_transf_GTA_type superfamily, MD: WcaA, PRK10073, glyC2_xrt_Gpos1, PTZ00260, Glyco_transf_2_PgaC_icaA, PTZ00260 |
| ISCps8, transposase | 758 | 99 | 0.0 | SH: GLF, GLF superfamily, SH: NAD_binding_8, NAD_binding_8 superfamily, MD: GLF, UDP-GALP_mutase, PRK07208 |
| gpH | 620 | 99 | 0.0 | NH: GT_2_like_b, Glycos_transf_2, Glyco_transf_GTA_type superfamily, MD: COG1216, WcaA |

| T | 4 Mbp to 5 Mbp |
| gpH | 705 | 100 | 0.0 | SH: DUF3751, DUF3751 superfamily |
| Putative phage gene | 811 | 100 | 0.0 | NH: DUF2590, DUF2590 superfamily |
| Conserved hypothetical protein | 228 | 100 | 4e-69 | NH: DUF2590, DUF2590 superfamily |
| Phage tail tape measure protein, family | 1,292 | 100 | 0.0 | SH: PhageMin_Tail, MCP_signal superfamily |
| Conserved hypothetical protein | 185 | 100 | 2e-54 | SH: DUF2765, DUF2765 superfamily |
| Bacteriophage lysis protein | 238 | 100 | 5e-72 | SH: Phage_lysis, Phage_lysis superfamily |
| Phage lysozyme | 306 | 100 | 7e-96 | NH: endolysin_autolysin, COG3772, Phage_lysozyme, Lysozyme_like superfamily |
| Phage holin, lambda family | 171 | 100 | 5e-49 | NH: DUF2597, DUF2597 superfamily |
| Tail tube | 311 | 100 | 2e-97 | NH: DUF2586, DUF2586 superfamily |
| Tail sheath | 774 | 100 | 0.0 | NH: P2_Plage_GpR, P2_Plage_GpR superfamily |
| Putative phage gene | 498 | 100 | 2e-162 | NH: P2_Plage_GpR, P2_Plage_GpR superfamily |
| Conserved hypothetical protein | 324 | 100 | 9e-102 | NH: Phage_GPL, Phage_GPL superfamily |
| Phage head completion protein (GPL) | 321 | 100 | 9e-101 | NH: Phage_GPL, Phage_GPL superfamily |
| Putative repressor protein CI | 373 | 100 | 1e-118 | NH: Nlp, HTH_35, PRK10344, HTH_35 superfamily |
| Putative DNA-binding protein Ner | 170 | 100 | 3e-49 | NH: HTH_Tnp_Mu_1, HTH_Tnp_Mu_1 superfamily, |
| Phage transposase | 579 | 100 | 0.0 | NH: HTH_Tnp_Mu_1, HTH_Tnp_Mu_1 superfamily, |
| Packaging protein | 450 | 100 | 7e-145 | NH: M, Phage_term_smal, Phage_term_smal superfamily |
| Phage major capsid protein, P2 family | 730 | 100 | 0.0 | NH: N, major_capsid_P2, Phage_cap_P2, Phage_cap_P2 superfamily |
| Scaffold | 666 | 100 | 0.0 | NH: O, Phage_GPO, Phage_GPO superfamily |
| Terminase | 1,232 | 100 | 0.0 | NH: Terminase_5, Terminase_5 superfamily, MD: P, Terminase_6, COGS484 |
| Phage portal protein, pbsx family | 674 | 100 | 0.0 | NH: Q, portal_PBSX, COGS5518, Phage_portal, Phage_portal superfamily |
| Phage transcriptional activator, Ogr/delta | 186 | 100 | 1e-54 | NH: DUF4406, DUF4406 superfamily |
| Hypothetical protein Ec53638_4365 | 176 | 100 | 6e-51 | NH: dam, PRK10904, Dam, MethyltransfD12, MethyltransfD12 superfamily |
| Hypothetical protein Ec53638_4366 | 162 | 100 | 3e-46 | NH: dam, PRK10904, Dam, MethyltransfD12, MethyltransfD12 superfamily |
| Putative phage replication protein | 254 | 100 | 4e-70 | NH: dam, PRK10904, Dam, MethyltransfD12, MethyltransfD12 superfamily |
| Hypothetical protein Ec53638_4288 | 102 | 98 | 3e-26 | NH: dam, PRK10904, Dam, MethyltransfD12, MethyltransfD12 superfamily |
| Putative phage gene | 1,133 | 100 | 0.0 | NH: dam, PRK10904, Dam, MethyltransfD12, MethyltransfD12 superfamily |
| Conserved hypothetical protein | 217 | 100 | 9e-86 | NH: dam, PRK10904, Dam, MethyltransfD12, MethyltransfD12 superfamily |
| DNA adenine methylase | 641 | 100 | 0.0 | NH: dam, PRK10904, Dam, MethyltransfD12, MethyltransfD12 superfamily |
### Supplementary Table 1. Signature details based on *Escherichia coli* 53638 reference

| Signature | Description | Score | E-value | Domain | Tag | Notes |
|-----------|-------------|-------|---------|--------|-----|-------|
| Hypothetical protein Ec53638_4377 | 712 | 100 | 3e-15 | NH: HP1_INT_C, Phage_integrase, DNA_BRE_C superfamily, MD: int, recomb_XerC, XerD, xerC |
| Hypothetical protein Ec53638_4380 | 133 | 100 | 8e-37 | DM: PRK05643, DnaN |
| Integrase for prophage CP-933T | 711 | 100 | 0.0 | NH: phage_tail_N, phage_tail_N superfamily |
| Hypothetical bacteriophage protein | 393 | 100 | 6e-137 | SH: FliC, FliC superfamily, MD: PRK08026 |
| Conserved hypothetical protein | 496 | 100 | 2e-176 | NH: ORF6N, ORF6N superfamily |
| Putative tail fiber protein | 586 | 100 | 0.0 | NH: Packaging_FI, Packaging_FI superfamily |
| Putative tail component of prophage CP-933K | 202 | 100 | 3e-63 | NH: PhageAttach, Gifsy-2 superfamily |
| Flagellin FliC | 469 | 100 | 8e-164 | NH: phage_tail_N, phage_tail_N superfamily |
| Putative DNA-packaging protein | 223 | 100 | 5e-74 | SH: HTH_36, HTH_36 superfamily, MD: PaaX_trns_reg |
| Phage Head-Tail Attachment | 252 | 99 | 2e-85 | |
| Hypothetical protein Ec53638_4317 | 387 | 100 | 1e-137 | |
| Phage transposase | 488 | 100 | 3e-170 | NH: rve, rve superfamily |
| PaaX family protein | 266 | 99 | 3e-90 | SH: HTH_36, HTH_36 superfamily, MD: PaaX_trns_reg |
| Tail fiber protein | 273 | 100 | 4e-91 | |
| Bcv gene product | 262 | 96 | 7e-89 | |
| Putative phage replication protein | 149 | 100 | 2e-43 | NH: tape_meas_lam_C, tape_meas_lam_C superfamily, MD: COG5281 |
| Gifsy-1 prophage VmtH | 229 | 100 | 1e-71 | |
| Putative phage regulatory protein, Rha family | 204 | 100 | 4e-66 | NH: Phage-MuB_C, Phage-MuB_C superfamily |
| DNA transposition protein | 203 | 100 | 2e-66 | |
| Gifsy-1 prophage VmtH | 163 | 100 | 3e-48 | SH: TMP_2, TMP_2 superfamily, MD: COG5281 |
| Prophage minor tail protein Z (GPZ) | 195 | 100 | 7e-65 | NH: Minor_tail_Z, Minor_tail_Z superfamily |
| Putative Prophage Qin DNA packaging protein NU1 homolog | 134 | 100 | 6e-42 | NH: COG4220, Phage_Nu1, Phage_Nu1 superfamily |
| Hypothetical protein Ec53638_4368, [inositol] | 139 | 100 | 2e-44 | |
| Monophosphatase 1 (IMPase 1) (IMP 1)(or 4)) | 112 | 100 | 5e-41 | |
| Protein gp42 | 140 | 100 | 4e-37 | NH: COG4396, Phage_Mu_Gam, Phage_Mu_Gam superfamily |
| Bacteriophage Mu Gam like protein | 122 | 100 | 4e-37 | |

**U**

| Signature | Description | Score | E-value | Domain | Tag | Notes |
|-----------|-------------|-------|---------|--------|-----|-------|
| Putative bacteriophage protein | 761 | 100 | 0.0 | NH: DUF2612, DUF2612 superfamily |
| Putative tail fiber protein | 575 | 100 | 0.0 | |
| Hypothetical bacteriophage protein | 410 | 100 | 6e-135 | |
| Phage P2 baseplate assembly protein gpV | 456 | 100 | 5e-152 | |
| Invasion plasmid antigen | 1,126 | 100 | 0.0 | |
| Putative tail component of prophage CP-933K | 246 | 99 | 6e-80 | NH: DUF4376, DUF4376 superfamily |
| Tail fiber assembly protein | 137 | 100 | 7e-43 | NH: Caudo_TAP, Caudo_TAP superfamily |
| gp33 TerL | 142 | 100 | 3e-42 | |
| Putative tail fiber protein | 146 | 100 | 4e-44 | NH: phage_tail_N, phage_tail_N superfamily |
| Conserved hypothetical protein | 236 | 100 | 3e-70 | |

**5 Mbp to end of the genome**
## Supplementary Table 2. Signature details based on *Escherichia coli* IHE3034 reference

| Region | Biomarker range in *Escherichia coli* IHE3034 genome between | Protein obtained by NCBI Blastx [identical protein in other *Escherichia coli* strains] | Blast score | Blast identity (%) | Blast E-value | Putative conserved domains non-specific hits (NH), specific hits (SH), multi domains (MD) |
|--------|-------------------------------------------------------------|----------------------------------------------------------------------------------------|-------------|-------------------|--------------|-------------------------------------------------------------------------------------|
| A      | 0 Mbp to 1 Mbp                                             | Tail fiber protein                                                                    | 559         | 100               | 0.0          | NH: CdtToxinA, RICIN superfamily                                                     |
|        |                                                             | Phage tail tape measure protein                                                       | 135         | 100               | 4e-39        |                                                                                     |
| B      | 2 Mbp to 3 Mbp                                             | Cytolethal distending toxin, subunit C                                               | 382         | 100               | 1e-127       | NH: PRK15251, CdtB, EEP superfamily                                                 |
|        |                                                             | Cytolethal distending toxin, subunit B                                               | 520         | 100               | 1e-179       | NH: CdtToxinA, RICIN superfamily                                                     |
|        |                                                             | Cytolethal distending toxin, subunit A                                               | 483         | 99                | 5e-166       | NH: dexA, DnaQ-like-exo superfamily                                                  |
|        |                                                             | Exonuclease family protein                                                            | 350         | 100               | 3e-116       |                                                                                     |
|        |                                                             | Hypothetical protein ECOK1_2135, [conserved domain protein]                           | 120         | 100               | 4e-36        |                                                                                     |
|        |                                                             | Hypothetical protein ECOK1_2134                                                      | 147         | 100               | 3e-47        |                                                                                     |
|        |                                                             | Hypothetical protein ECOK1_2122, [membrane protein]                                  | 164         | 100               | 2e-52        |                                                                                     |
|        |                                                             | Putative transcriptional regulator DicA157                                            | 370         | 100               | 3e-106       |                                                                                     |
|        |                                                             | Antitermination protein                                                              | 98.6        | 100               | 2e-27        |                                                                                     |
| C      | 2 Mbp to 3 Mbp                                             | Integrase/recombinase, phage integrase family                                        | 693         | 100               | 0.0          | NH: HP1_INT_C, Phage_integrase, DNA_BRE_C superfamily                                |
|        |                                                             | Putative regulatory protein Cox                                                      | 196         | 100               | 1e-59        | NH: Phage_Cox, Phage_Cox superfamily                                               |
|        |                                                             | Hypothetical protein ECOK1_2557, [membrane protein]                                 | 131         | 100               | 1e-36        |                                                                                     |
|        |                                                             | Hypothetical protein ECOK1_2558                                                     | 83.2        | 98                | 9e-20        |                                                                                     |
|        |                                                             | Hypothetical protein ECOK1_2602, [type VI secretion protein]                         | 585         | 100               | 0.0          | NH: VI_minor_1, DUF 3121 superfamily                                                |
|        |                                                             | Hypothetical protein ECOK1_2601, [double zinc ribbon family protein]                 | 194         | 100               | 4e-60        |                                                                                     |
|        |                                                             | Hypothetical protein ECOK1_2600                                                    | 675         | 100               | 0.0          |                                                                                     |
|        |                                                             | Holin, lambda family                                                                | 58.9        | 100               | 3e-12        |                                                                                     |
|        |                                                             | Hypothetical protein ECOK1_2581                                                    | 375         | 100               | 1e-133       | NH: Glyco_hydro_108, Glyco_hydro_108 superfamily                                   |
| D      | 2 Mbp to 3 Mbp                                             | Hypothetical protein ECOK1_2812                                                    | 519         | 100               | 8e-169       | NH: INT_P4, DNA_BRE_C superfamily, MD: int, XerD, recomb_XerD, xerD                |
|        |                                                             | Hypothetical protein ECOK1_2814                                                    | 2,219       | 100               | 0.0          | MD: PRK03918, COG1340, SMC_N, SMC_prok_B                                          |
|        |                                                             | Hypothetical protein ECOK1_2815                                                    | 1,837       | 99                | 0.0          |                                                                                     |
|        |                                                             | Hypothetical protein ECOK1_2816                                                    | 103         | 100               | 6e-25        |                                                                                     |
|        |                                                             | Hypothetical protein ECOK1_2809                                                    | 222         | 100               | 2e-72        |                                                                                     |
|        |                                                             | Enterohemolysin 1                                                                  | 317         | 100               | 2e-110       |                                                                                     |
| E      | 4 Mbp to 5 Mbp                                             | Site-specific recombinase, phage integrase family                                   | 223         | 100               | 6e-65        | NH: INT_P4, DNA_BRE_C superfamily, MD: PRK09692                                   |
|        |                                                             | Protein cII                                                                        | 531         | 100               | 0.0          | NH: Beta_protein, Beta_protein superfamily                                         |
|        |                                                             | Hypothetical protein ECOK1_4790, [beta family protein, Enterobacteria phage P4]    | 700         | 100               | 0.0          |                                                                                     |
| F      | 5 Mbp to end of the genome                                 | Hypothetical protein ECOK1_4914                                                    | 133         | 100               | 9e-42        |                                                                                     |
Supplementary Table 3. Signature details based on *Escherichia coli* RN587/1 reference

| Regio n | Biomarker range in *Escherichia coli* RN587/1 genome between | Protein obtained by NCBI Blastx [identical protein in other *Escherichia coli* strains] | Blast score | Blast Identity (%) | Blast E-value | Putative conserved domains non-specific hits (NH), specific hits (SH), multi domains (MD) |
|---------|---------------------------------------------------------------|-----------------------------------------------------------------|------------|-------------------|---------------|---------------------------------------------------------------------------------|
| A       | 0 Mbp to 1 Mbp                                               | Cytotolethal distending toxin A/C family protein                 | 205        | 100               | 2e-66         | NH: CdtoxinA, RICIN superfamily                                                  |
| B       | 0 Mbp to 1 Mbp                                               | Type II restriction enzyme EcoRII                                | 833        | 100               | 0.0           | SH: EcoRII-C, EcoRII-C superfamily, NH: EcoRII_N, BfiI_C_EcoRII_N_B3 superfamily |
|         |                                                               | Modification methylase EcoRII                                   | 949        | 100               | 0.0           | NH: dcm, DNA_methylase, Cyt_C5_DNA_methylase, Cyt_C5_DNA_methylase superfamily, MD: PRK10458 |
|         |                                                               | Outer membrane autotransporter barrel domain protein            | 214        | 100               | 3e-66         | NH: PL1_Passenger_AT, PL1_Passenger_AT superfamily, NH: DUF4353, DUF4353 superfamily, MD: PHA03255 |
| C       | Around 1 Mbp                                                 | Hypothetical protein ECRN5871_0833                              | 710        | 100               | 0.0           | NH: DUF3751, DUF3751 superfamily                                                |
|         |                                                               | Hypothetical protein ECRN5871_0834                              | 600        | 100               | 0.0           | SH: Caudo_TAP, Caudo_TAP superfamily                                            |
|         |                                                               | Hypothetical protein ECRN5871_0832                              | 79.3       | 97                | 6e-18         | NH: DUF2586, DUF2586 superfamily                                                |
|         |                                                               | Phage Tail Collar domain protein                                | 528        | 100               | 0.0           | NH: DUF2586, DUF2586 superfamily                                                |
|         |                                                               | Caudovirales tail fiber assembly family protein                 | 400        | 100               | 3e-136        | NH: SopA, SopA superfamily, NH: SopA_C, SopA_C superfamily, MD: Pentapeptide_4, PRK15377, COG1357 |
|         |                                                               | Hypothetical protein ECRN5871_0827                              | 442        | 100               | 6e-152        | NH: DUF2597, DUF2597 superfamily                                                |
|         |                                                               | Phage protein                                                  | 206        | 100               | 4e-67         | NH: DUF2586, DUF2586 superfamily                                                |
|         |                                                               | Pentapeptide repeats family protein                             | 1,107      | 100               | 0.0           | NH: P2_Phae_GpR, P2_Phae_GpR superfamily                                         |
|         |                                                               | hdmD                                                           | 475        | 100               | 3e-167        | SH: RT_Bac_retron_II, RT_like superfamily, MD: RVT_1                           |
|         |                                                               | Hypothetical protein ECRN5871_0812, [phage tail protein, P2   | 316        | 100               | 1e-110        | NH: XkdT, Baseplate_J superfamily                                               |
|         |                                                               | Phage tail completion protein R (GpR)]                         | 273        | 100               | 1e-95         | NH: DUF2590, DUF2590 superfamily                                               |
|         |                                                               | Phage protein                                                  | 69.7       | 100               | 1e-16         | NH: DUF2586, DUF2586 superfamily                                                |
|         |                                                               | Phage protein                                                  | 375        | 100               | 2e-131        | NH: DUF2586, DUF2586 superfamily                                                |
|         |                                                               | Phage small terminase subunit                                   | 163        | 100               | 1e-50         | NH: M, Phage_small_term superfamily                                             |
|         |                                                               | Long tail fiber protein p37 domain protein                     | 119        | 100               | 2e-35         | NH: M, Phage_small_term superfamily                                             |
|         |                                                               | Retron EC67 protein domain protein                             | 850        | 100               | 0.0           | SH: RT_Bac_retron_II, RT_like superfamily, MD: RVT_1                           |
|         |                                                               | Hypothetical protein ECRN5871_0823, [baseplate J-like protein] | 814        | 100               | 0.0           | NH: XkdT, Baseplate_J superfamily                                               |
|         |                                                               | Hypothetical protein ECRN5871_0822, [PF10761 family protein]    | 77.8       | 100               | 6e-18         | NH: DUF2590, DUF2590 superfamily                                               |
|         |                                                               | Phage tail tape measure protein, TP901 family, core region     | 775        | 100               | 0.0           | NH: Q, portal_PBSX, Phage_portal, Phage_portal superfamily                      |
|         |                                                               | Hypothetical protein ECRN5871_0798                             | 296        | 100               | 7e-101        | SH: DUF1311, DUF1311 superfamily, MD: PHA02067, Lpri                          |
|         |                                                               | Hypothetical protein ECRN5871_0799                             | 276        | 100               | 4e-93         | NH: Q, portal_PBSX, Phage_portal, Phage_portal superfamily                      |
|         |                                                               | Regulatory protein CII                                         | 38.1       | 100               | 3e-04         | NH: DUF1311, DUF1311 superfamily, MD: PHA02067, Lpri                          |
|         |                                                               | Hypothetical protein ECRN5871_0797                             | 119        | 100               | 6e-34         | NH: DUF1311, DUF1311 superfamily, MD: PHA02067, Lpri                          |
|         |                                                               | Hypothetical protein ECRN5871_0829                             | 702        | 100               | 0.0           | NH: DUF1311, DUF1311 superfamily, MD: PHA02067, Lpri                          |
|         |                                                               | Phage portal protein, PBSX family                              | 342        | 100               | 3e-119        | NH: Q, portal_PBSX, Phage_portal, Phage_portal superfamily                      |
|         |                                                               | Hypothetical protein ECRN5871_0691                             | 501        | 100               | 0.0           | SH: DUF1311, DUF1311 superfamily, MD: PHA02067, Lpri                          |
|         |                                                               | Terminase, ATPase subunit                                      | 605        | 100               | 0.0           | MD: P, Terminase_6                                                              |
|         |                                                               | Protein rhsB                                                  | 355        | 100               | 2e-113        | NH: Rhs_assc_core, Rhs_assc_core superfamily                                    |
|         |                                                               | Replication protein A                                          | 350        | 100               | 2e-118        | NH: Phage_GPA, Phage_GPA superfamily                                           |
|         |                                                               | Hypothetical protein ECRN5871_0824                             | 273        | 100               | 5e-95         | NH: HP1_INT_C, DNA_BRE_C superfamily, MD: int, recombin XerD, XerD,            |
|         |                                                               | Integrase                                                      | 495        | 100               | 1e-177        | NH: Q, Phage_GPO, Phage_GPO superfamily, MD: IncA                               |
|         |                                                               | Scaffold domain protein                                        | 473        | 100               | 6e-171        | NH: Q, Phage_GPO, Phage_GPO superfamily, MD: IncA                               |
|         |                                                               | Repressor protein CII                                         | 338        | 100               | 4e-120        | NH: Phage_CI_repr, Phage_CI_repr superfamily                                    |
|         |                                                               | Hypothetical protein ECRN5871_0819                             | 129        | 98                | 3e-40         | NH: Phage_CI_repr, Phage_CI_repr superfamily                                    |
### Supplementary Table 3. Signature details based on *Escherichia coli* RN587/1 reference

| C | Around 1 Mbp | Hypothetical protein ECRN5871_0801 | 202 | 100 | 2e-68 | NH: DUF3850, DUF3850 superfamily |
|   |              | Retron EC67 DNA adenine methylase | 204 | 100 | 3e-57 | NH: Dam, PRK10904, MethyltransfD12, dam, MethyltransfD12 superfamily |
|   |              | Phage tail tape measure protein, TP901 family, core region | 200 | 100 | 2e-62 | NH: PhageMin_Tail, MCP_signal superfamily |
|   |              | Replication protein A | 201 | 100 | 2e-62 | MD: PRK14960 |
|   |              | Phage major capsid protein, P2 family | 48.9 | 100 | 1e-08 |   |
|   |              | Holin | 93.6 | 100 | 8e-27 | NH: Phage_holin_2, Phage_holin_2 superfamily |
|   |              | Phage head completion protein family protein | 98.2 | 100 | 4e-28 |   |
|   |              | Phage protein | 152 | 100 | 1e-48 | NH: DUF2765, DUF2765 superfamily |
|   |              | Hypothetical protein ECRN5871_0700 | 107 | 100 | 2e-32 |   |
| D | 1 Mbp to 2 Mbp | Hypothetical protein ECRN5871_1072, [type III secretion system protein] | 594 | 100 | 1e-174 | MD: PRK15386 |
|   |              | Hypothetical protein ECRN5871_4139, [type III secretion system protein] | 321 | 65 | 2e-89 |   |
|   |              | Bacteriophage CI repressor helix-turn-helix domain protein | 431 | 100 | 2e-134 | NH: Phage_CI_repr, Phage_CI_repr superfamily |
|   |              | Exonuclease family protein | 676 | 100 | 0.0 |   |
|   |              | Avirulence protein A domain protein | 551 | 100 | 2e-173 |   |
|   |              | Hypothetical protein ECRN5871_1040 | 185 | 100 | 8e-54 |   |
|   |              | Hypothetical protein ECRN5871_1039 | 237 | 100 | 1e-71 |   |
|   |              | Hypothetical protein ECRN5871_1038 | 155 | 100 | 2e-45 |   |
|   |              | Eae-like protein | 223 | 100 | 2e-66 |   |
|   |              | Hypothetical protein ECRN5871_1027 | 224 | 100 | 2e-72 | NH: DUF1482, DUF1482 superfamily |
|   |              | Phage integrase family protein | 227 | 100 | 2e-67 | NH: INT_Lambda_C, DNA_BRE_C superfamily |
| E | 1 Mbp to 2 Mbp | ea59 protein | 1,043 | 100 | 0.0 | SH: AAA_21, NH: COG3910, ABC_ATPase superfamily, MD: AAA_15 |
|   |              | Hypothetical protein ECRN5871_4153, [HNH endonuclease family protein] | 576 | 100 | 0.0 |   |
|   |              | Prophage lambda integrase | 234 | 100 | 3e-68 | NH: INT_Lambda_C, DNA_BRE_C superfamily |
|   |              | Hypothetical protein ECRN5871_4172 | 829 | 100 | 0.0 | NH: INT_Lambda_C, DNA_BRE_C superfamily |
|   |              | Hypothetical protein ECRN5871_4173 | 301 | 100 | 5e-97 | NH: NA37, PRK00378, NA37 superfamily |
|   |              | Hypothetical protein ECRN5871_4177, [Rz1 lytic protein] | 91.3 | 98 | 3e-24 | NH: Phage_integ_N, Phage_integ_N superfamily |
|   |              | Prophage lambda integrase | 104 | 100 | 1e-28 | NH: Phage_integ_N, Phage_integ_N superfamily |
| F | 1 Mbp to 2 Mbp | Reverse transcriptase | 331 | 100 | 6e-107 | NH: RT_Bac_retron_II, RT_like superfamily |
|   |              | recF/RecN/SMC N terminal domain protein | 1,100 | 100 | 0.0 | SH: AAA_23, NH: ABC_RecF, COG3910, AAA_21, ABC_ATPase superfamily, MD: COG3950, recF, recF, AAA_15 |
|   |              | Hypothetical protein ECRN5871_4523, [TIGR02646 family protein] | 424 | 100 | 9e-142 | NH: TIGR02646, TIGR02646 superfamily |
|   |              | Hypothetical protein ECRN5871_4556 | 775 | 100 | 0.0 | NH: COG4688, COG4688 superfamily |
|   |              | Hypothetical protein ECRN5871_4549 | 91.7 | 98 | 1e-24 | NH: PRK15131, PLM02128, Abed superfamily, MD: ManA |
|   |              | Transposase for insertion sequence element IS1111A | 336 | 100 | 5e-118 | SH: Transposase_20, Transposase_20 superfamily, MD: COG3547 |
|   |              | Transposase IS116/IS110/IS902 family protein | 338 | 100 | 1e-117 | SH: Transposase_20, Transposase_20 superfamily, MD: COG3547 |
|   |              | Hypothetical protein ECRN5871_4552, [UvrD/REP helicase domain protein] | 273 | 100 | 3e-91 | MD: pcrA, UvrD, uvrD, UvrD-helicase |
| F | 1 Mbp to 2 Mbp | tnpA | 71.2 | 100 | 6e-18 | NH: ABC_SMC_barmotin, AAA_23, ABC_ATPase superfamily, |
|   |              | Hypothetical protein ECRN5871_4551, [chromosome segregation | 206 | 100 | 6e-65 |   |
**Supplementary Table 3.** Signature details based on *Escherichia coli* RN587/1 reference

| Protein Family | Mbp Range | Signature Details | EValue | Smith Waterman Score |
|----------------|-----------|-------------------|--------|----------------------|
| Phage protein, PBSX family | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 284 | 0.00 | 1e-86 |
| Terminase, ATPase subunit | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 1,238 | 1.00 | 0.00 |
| Presumed capsid-scaffolding protein | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 439 | 0.00 | 7e-144 |
| Phage major capsid protein, P2 family | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 632 | 0.00 | 0.00 |
| Terminase, endonuclease subunit | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 420 | 0.00 | 4e-138 |
| Head completion/stabilization protein | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 290 | 0.00 | 2e-91 |
| Hypothetical protein ECRN871_3522 | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 417 | 0.99 | 3e-137 |
| Hypothetical protein ECRN871_3519 | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 215 | 1.00 | 3e-65 |
| Hypothetical protein ECRN871_3516, [PF05449 family protein] | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 209 | 1.00 | 1e-62 |
| Hypothetical protein ECRN871_3515, [PF11860 family protein] | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 408 | 1.00 | 2e-132 |
| P2 phage tail completion protein R family protein | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 318 | 1.00 | 9e-101 |
| Phage virion morphogenesis protein | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 300 | 1.00 | 3e-94 |
| hibC family protein | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 230 | 1.00 | 5e-70 |
| Baseplate assembly protein V | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 377 | 1.00 | 5e-121 |
| Baseplate assembly protein W | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 236 | 1.00 | 3e-72 |
| Baseplate J-like family protein | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 330 | 1.00 | 5e-102 |
| Phage tail protein I | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 397 | 1.00 | 2e-127 |
| Phage tail fiber repeat family protein | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 1,291 | 1.00 | 0.00 |
| Tail fiber domain protein | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 219 | 0.68 | 7e-62 |
| Phage Tail Collar domain protein | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 244 | 1.00 | 1e-69 |
| Hypothetical protein ECRN871_3504, [tail fiber assembly protein] | 2 Mbp to 3 Mbp | phage portal protein, PBSX family | 470 | 1.00 | 6e-153 |

**Note:** Values in the table represent the signature details based on *Escherichia coli* RN587/1 reference, with EValue and Smith Waterman Score indicating the significance of the match.
## Supplementary Table 3. Signature details based on *Escherichia coli* RN587/1 reference

| L  | 3 Mbp to 4 Mbp |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |
Table 3. Signature details based on *Escherichia coli* RN587/1 reference

| O | 4 Mbp to 5 Mbp | No significant results | 4 Mbp to 5 Mbp | 4 Mbp to 5 Mbp |
|---|----------------|------------------------|----------------|----------------|
| P | 4 Mbp to 5 Mbp | Anaerobic C4-dicarboxylate transporter dcuA 721 100 0.0 | Fumarylacetoacetate (FAA) hydrolase family protein 610 100 0.0 | NH: Dcu, PRK09412, DcuB, DcuA_DcuB, DcuA_DcuB superfamily |
|   |                | Dihydrdipicolinate synthetase family protein 600 100 0.0 | SH: FAA_hydrolase, NH: PRK10691, HpaG-N-term, FAA_Hydrolase superfamily, MD: MhpD, HpaG-C-term, PRK15203 |
|   |                | Transcriptional regulator, LacI family 466 100 2e-152 | SH: HTH_LacI, NH: HTH_LacI, HTH_LacI superfamily, NH: PBPI_LacI_like_7, Periplasmic_Binding_Protein_Type_I superfamily, NH: Peripla_BP_3, MD: PurR, PRK10703, ccpA, Peripla_BP_1 |
|   |                | Sugar (and other) transporter family protein 941 100 0.0 | SH: MFS, MFS superfamily, NH: 2_A_01_02, PRK11195 superfamily, MD: 2A0115, PRK11551, Sugar_tr, PropP, ?nap?SV2 |
|   |                | Porin B 837 100 0.0 | NH: OppB, OppB superfamily |
| Q | 4 Mbp to 5 Mbp | Type I restriction enzyme specificity protein 213 100 2e-69 | Type I restriction-modification system, M subunit 256 100 7e-82 | MD: HsdM, hsdM, N6_Mtase |
|   |                | Type I site-specific deoxyribonuclease, HsdR family 139 100 2e-40 | NH: Methylase_S, Methylase_S superfamily, MD: HsdS, sufB |
| R | 4.5 Mbp to end of the genome | Hypothetical protein ECRN5871_0098 1,119 100 0.0 | Hypothetical protein ECRN5871_0099 1,435 100 0.0 | NH: DLP_2, Dynamin_N, Ras_like_GTPase superfamily |
|   |                | Hypothetical protein ECRN5871_0100 814 99 0.0 | sell repeat family protein 1,603 100 0.0 | NH: DLP_2, PRK09866, Ras_like_GTPase superfamily |
|   |                | Hypothetical protein ECRN5871_0138, [secretion protein EspT] 406 100 4e-136 | Hypothetical protein ECRN5871_0139 546 99 0.0 | NH: DLP_2, Ras_like_GTPase superfamily |
|   |                | Hypothetical protein ECRN5871_0129 435 100 1e-149 | NH: IpaB_EvC, IpaB_EvC superfamily |
|   |                | Hypothetical protein ECRN5871_0128, [secretion protein EspM] 103 100 7e-27 | NH: DUF1076, DUF1076 superfamily |
|   |                | Transposase IS116/IS110/IS902 family protein 233 100 2e-72 | SH: Transposase_20, Transposase_20 superfamily, MD: COG3547 |
|   |                | Hypothetical protein ECRN5871_0087 390 100 4e-134 | NH: DUF3491, DUF3491 superfamily |
|   |                | Cysteine protease domain, YopT-type domain protein 1,030 100 0.0 | NH: DUF1076, DUF1076 superfamily |
|   |                | Hypothetical protein ECRN5871_0025,[N-acetyltransferase GCN5] 332 100 2e-114 | NH: DUF1076, DUF1076 superfamily |
|   |                | Hypothetical protein ECRN5871_0026 205 100 1e-65 | SH: COG4453, DUF 1778 |
|   |                | Hypothetical protein ECRN5871_0137 399 100 4e-142 | NH: DUF1076, DUF1076 superfamily |
|   |                | Superoxide dismutase [Cu-Zn] 1 344 100 2e-121 | NH: PRK15388, SodC, Cu-Zn_Superoxide_Dismutase, Sod_Cu, PLN02386, Cu-Zn_Superoxide_Dismutase superfamily |
|   |                | Transposase for insertion sequence element IS1111A 161 100 8e-51 | NH: DEDD_Tnp_IS110, DEDD_Tnp_IS110 superfamily, SH: Transposase_20, Transposase_20 superfamily, MD: COG3547 |
|   |                | Retron EC67 protein domain protein 257 100 4e-88 | NH: DEDD_Tnp_IS110, DEDD_Tnp_IS110 superfamily, SH: Transposase_20, Transposase_20 superfamily, MD: COG3547 |
|   |                | Phage portal protein, PBSX family 95 100 3e-24 | NH: Q, Phage_portal superfamily |
|   |                | YadA 316 100 3e-111 | MD: PT200102 |
|   |                | Hypothetical protein ECRN5871_0131 114 100 4e-33 | NH: COG5525, Terminase_GpA, Terminase_GpA superfamily |
| R | 4.5 Mbp to end of the genome | Terminate large subunit 226 100 5e-71 | Terminate small subunit 206 100 7e-68 | SH: COG44220, NH: Phage_Nu1, Phage_Nu1 superfamily, MD: PLN03114 |
|   |                | Serine protease eatA 345 100 8e-112 | MD: Peptidase_S6 |
**Supplementary Table 3.** Signature details based on *Escherichia coli* RN587/1 reference

| ST44 protein | 133 | 100 | 4e-42 | NH: Transposase_mut, Transposase_mut superfamily, MD: COG3328 |
|--------------|-----|-----|-------|-------------------------------------------------------------|
| ygeA         | 162 | 100 | 5e-53 | NH: gpV, phage_P2_V, Phage_base_V, Phage_base_V superfamily  |
| Baseplate assembly protein V | 140 | 100 | 8e-44 |                                                      |
| Transposase, IS605 OrfB family | 62.4 | 100 | 8e-15 |                                                      |
| Putative transposase | 136 | 100 | 7e-44 |                                                      |
| Cysteine protease domain, YopT-type domain protein | 92.0 | 100 | 1e-23 |                                                      |
**Supplementary Table 4. Signature details based on *Escherichia coli* STEC_B2F1 reference**

| Region | Biomarker range in *Escherichia coli* STEC_B2F1 genome between | Protein obtained by NCBI Blastx [identical protein in other *Escherichia coli* strains] | Blast score | Blast Identity (%) | Blast E-value | Putative conserved domains |
|--------|---------------------------------------------------------------|-----------------------------------------------------------------------------------|------------|-------------------|---------------|---------------------------|
| A      | 0 Mbp to 1 Mbp                                               | Hypothetical protein ECSTECB2F1_0150                                              | 80.1       | 100               | 1e-21         | SH: collagen, collagen superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_0149, [transposase]                               | 82.4       | 100               | 5e-22         | SH: HTH_36, HTH_36 superfamily, MD: PaaX_trns_reg |
| B      | 0 Mbp to 1 Mbp                                               | Protein 40A                                                                         | 582        | 100               | 0.0           | SH: collagen superfamily |
| C      | 0 Mbp to 1 Mbp                                               | Collagen triple helix repeat family protein                                         | 487        | 100               | 3e-166        | SH: collagen, collagen superfamily |
|        |                                                                | Tail fiber assembly                                                                | 144        | 100               | 6e-42         | SH: HTH_36, HTH_36 superfamily, MD: PaaX_trns_reg |
|        |                                                                | Hypothetical protein ECSTECB2F1_0901, [tail fiber assembly protein, caudovirales tail fiber assembly protein] | 251        | 100               | 1e-79         | SH: collagen superfamily |
|        |                                                                | Outer membrane protein lom                                                          | 128        | 100               | 1e-39         | SH: AiL_Lom, PRK15240, COG3637, OMP_b-brl superfamily |
|        |                                                                | Phage integrase family protein                                                     | 56.2       | 100               | 1e-11         | SH: phage_tail_N, phage_tail_N superfamily |
|        |                                                                | Prophage tail fiber family protein                                                 | 74.3       | 96                | 3e-19         | SH: phage_tail_N, phage_tail_N superfamily |
| D      | 1 Mbp to 2 Mbp                                               | Putative endopeptidase                                                              | 63.9       | 100               | 2e-12         | SH: collagen superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_1098                                              | 78.2       | 100               | 3e-18         | SH: collagen superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_1099                                              | 86.7       | 98                | 3e-21         | SH: collagen superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_1078, [PF03235 family protein]                      | 726        | 100               | 0.0           | SH: DUF 262, DUF 262 superfamily |
|        |                                                                | Caudovirales tail fiber assembly family protein                                    | 125        | 100               | 5e-37         | SH: phage_tail_N, phage_tail_N superfamily |
|        |                                                                | Prophage lambda integrase                                                           | 275        | 100               | 2e-93         | SH: DUF 262, DUF 262 superfamily |
|        |                                                                | Prophage lambda integrase                                                           | 139        | 100               | 6e-42         | SH: INT_Lambda_C, DNA_BRE_C superfamily |
|        |                                                                | Prophage lambda integrase                                                           | 110        | 100               | 2e-31         | SH: Phage_integ_N, Phage_integ_N superfamily |
| E      | 1 Mbp to 2 Mbp                                               | Transcriptional regulator, AraC family                                              | 497        | 100               | 3e-175        | MD: PRK09940, COG4753 |
|        |                                                                | Collagen triple helix repeat family protein                                         | 228        | 100               | 2e-70         | SH: DUF3672, DUF3672 superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_1516, [tail fiber assembly protein, caudovirales tail fiber assembly protein] | 221        | 100               | 4e-70         | SH: DUF3672, DUF3672 superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_1255                                              | 89.4       | 100               | 1e-23         | MD: PRK10597, DinI superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_1256                                              | 188        | 100               | 5e-62         | MD: PRK10597, DinI superfamily |
|        |                                                                | Host specificity protein J                                                          | 308        | 100               | 1e-99         | SH: DUF3672, DUF3672 superfamily |
|        |                                                                | Antitermination protein Q                                                           | 28.5       | 100               | 0.13          | MD: PRK10597, DinI superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_1251                                              | 207        | 100               | 6e-69         | MD: PRK10597, DinI superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_1263                                              | 197        | 100               | 2e-61         | MD: PRK10597, DinI superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_1515                                              | 60.1       | 100               | 9e-14         | SH: ORF6N, ORF6N superfamily |
|        |                                                                | Helix-turn-helix family protein                                                    | 130        | 100               | 6e-41         | SH: ORF6N, ORF6N superfamily |
|        |                                                                | Outer membrane protein lom                                                          | 127        | 100               | 6e-40         | SH: ORF6N, ORF6N superfamily |
|        |                                                                | Prophage tail fiber family protein                                                 | 82.8       | 96                | 2e-22         | SH: ORF6N, ORF6N superfamily |
|        |                                                                | Collagen triple helix repeat family protein                                        | 85.9       | 100               | 4e-22         | SH: ORF6N, ORF6N superfamily |
| F      | 1 Mbp to 2 Mbp                                               | Phage integrase family protein                                                     | 212        | 100               | 1e-69         | SH: phage_tail_N, phage_tail_N superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_1296                                              | 91.3       | 100               | 7e-26         | SH: phage_tail_N, phage_tail_N superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_1297                                              | 56.2       | 100               | 6e-12         | SH: phage_tail_N, phage_tail_N superfamily |
|        |                                                                | Phage integrase family protein                                                     | 141        | 100               | 5e-43         | SH: phage_tail_N, phage_tail_N superfamily |
| G      | 1 Mbp to 2 Mbp                                               | Hypothetical protein ECSTECB2F1_1685                                              | 902        | 100               | 0.0           | SH: phage_tail_N, phage_tail_N superfamily |
| H      | 1 Mbp to 2 Mbp                                               | BRO family, N-terminal domain protein                                             | 537        | 100               | 0.0           | SH: Bro-N, NH: Bro-N, Bro-N superfamily, MD: COG3617 |
|        |                                                                | paaX-like family protein                                                           | 271        | 100               | 3e-92         | SH: HTH_36, HTH_36 superfamily, MD: Paax_trns_reg |
|        |                                                                | Phage tail fiber repeat family protein                                            | 355        | 100               | 2e-120        | SH: Phage_fiber_2, Phage_fiber_2 superfamily |
|        |                                                                | DNA-binding protein Roi                                                            | 245        | 100               | 2e-83         | SH: Phage_fiber_2, Phage_fiber_2 superfamily |
|        |                                                                | DNA-damage-inducible protein I                                                    | 59.7       | 100               | 6e-14         | SH: Phage_fiber_2, Phage_fiber_2 superfamily |
|        |                                                                | Hypothetical protein ECSTECB2F1_2003                                              | 69.7       | 100               | 4e-18         | SH: Phage_fiber_2, Phage_fiber_2 superfamily |
|   |   |   |   |   |   |
|---|---|---|---|---|---|
|   |   |   |   |   |   |
| I | 2 Mbp to 3 Mbp | Flagellin |   | 136 | 100 |
| J | 2 Mbp to 3 Mbp | Putative membrane protein |   | 632 | 100 |
|   |   | Glycosyl transferase family 2 family protein |   | 572 | 100 |
|   |   | Hypothetical protein ECSTECB2F1_2214 |   | 328 | 99 |
|   |   | Polysaccharide biosynthesis family protein |   | 764 | 100 |
|   |   | Erythromycin biosynthesis sensory transduction protein eryC1 |   | 753 | 100 |
|   |   | wbtB |   | 164 | 99 |
|   |   | Putative teichuronic acid biosynthesis glycosyltransferase tuaG |   | 525 | 100 |
|   |   | Glucose-1-phosphate thymidyllytransferase |   | 556 | 100 |
| K | 2 Mbp to 3 Mbp | Hypothetical protein ECSTECB2F1_2378 |   | 99.8 | 98 |
|   |   | Hypothetical protein ECSTECB2F1_2379 |   | 94.7 | 100 |
|   |   | clp protease family protein |   | 1,288 | 100 |
|   |   | Bacteriophage P4 DNA primease |   | 203 | 100 |
|   |   | Hypothetical protein ECSTECB2F1_2371 |   | 132 | 100 |
| L | 2 Mbp to 3 Mbp | Hemagglutination activity domain protein |   | 1,036 | 99 |
|   |   | Hypothetical protein ECSTECB2F1_3213 |   | 105 | 99 |
|   |   | Sulfatase family protein |   | 769 | 99 |
|   |   | Hypothetical protein ECSTECB2F1_3190, [arylsulfatase] |   | 261 | 100 |
|   |   | Outer membrane porin protein LC |   | 611 | 100 |
|   |   | RTX toxin acyltransferase family protein |   | 361 | 100 |
|   |   | Hemolysin secretion/activation protein ShiB/FhaC/HecB family protein |   | 746 | 100 |
|   |   | Hypothetical protein ECSTECB2F1_3193 |   | 130 | 100 |
|   |   | Transcriptional regulator, AraC family |   | 377 | 100 |
|   |   | Serine protease eata |   | 1,121 | 100 |
|   |   | Hypothetical protein ECSTECB2F1_3192, [membrane protein] |   | 60.5 | 100 |
|   |   | Hypothetical protein ECSTECB2F1_3199 |   | 89.4 | 100 |
|   |   | Hypothetical protein ECSTECB2F1_3200 |   | 133 | 100 |
|   |   | Neurotensin receptor R8 |   | 133 | 100 |
|   |   | tonB-dependent vitamin B12 receptor |   | 268 | 100 |
|   |   | Acetyl-CoA acetyltransferase |   | 44.7 | 100 |
**Supplementary Table 4.** Signature details based on *Escherichia coli* STEC_B2F1 reference

| Region | Gene Details | Identity | E-value | Actrans | Description |
|--------|--------------|----------|---------|---------|-------------|
| M 2 Mbp to 3 Mbp | Hypothetical protein ECSTECB2F1_3178 ompA-like transmembrane domain protein | 76.6 | 97 | 9e-20 | Actrans NH: COG3637, OMP_b-brl, OMP_b-brl superfamily |
| N 3 Mbp to 4 Mbp | Replication gene A protein | 69.7 | 100 | 6e-13 | NH: CdtxinA, RICIN superfamily |
| O 4 Mbp to end of the genome | Hypothetical protein ECSTECB2F1_4680 Putative membrane protein | 564 | 100 | 0.0 | NH: ResIII, DEXDc superfamily, MD: hsdR, COG0610 |
| P 4 Mbp to end of the genome | Host specificity protein J | 374 | 100 | 7e-119 | SH: DUF3672, DUF3672 superfamily |
| | Hypothetical protein ECSTECB2F1_1326, [hok/gef family protein] | 63.2 | 100 | 2e-13 | NH: phage_tail_N, phage_tail_N superfamily |
| | Prophage tail fiber family protein | 89.7 | 100 | 3e-22 | SH: BID_2, Big_2 superfamily, MD: COG5492 |
| | Major tail protein V | 337 | 100 | 6e-118 | |
| | Hypothetical protein ECSTECB2F1_1281 | 130 | 100 | 9e-41 | |
| | Hemin receptor domain protein | 40.0 | 100 | 9e-06 | |
| | Minor tail protein M | 144 | 99 | 4e-46 | NH: COG4718, Phage_min_tail, Phage_min_tail superfamily |
| | Phage tail tape measure protein, lambda family | 67.0 | 100 | 1e-14 | |
| | Phage terminase large subunit family protein | 140 | 100 | 6e-42 | NH: COG5525, Terminase_GpA superfamily |
## Supplementary Table 5. Signature details based on *Escherichia coli* STEC_C165_02 reference

| Region | Biomarker range in *Escherichia coli* STEC_C165_02 genome between | Protein obtained by NCBI Blastx [identical protein in other *Escherichia coli* strains] | Blast score | Blast Identity (%) | Blast E-value | Putative conserved domains non-specific hits (NH), specific hits (SH), multi domains (MD) |
|--------|---------------------------------------------------------------|-----------------------------------------------------------------|-------------|-------------------|--------------|---------------------------------------------------------------|
| A      | 0 Mbp to 1 Mbp                                              | Restriction endonuclease family protein                          | 723         | 100               | 0.0          | SH: COG4127, HsdM_N, NH: Mrr_cat, UPF0020, Mrr_cat superfamily, HsdM_N superfamily, MD: Mrr, HsdM, N6_Mtase |
|        |                                                               | TIR protein                                                     | 411         | 100               | 2e-131       | SH: Methyltransf_26, AdoMet_Mtase superfamily, HsdM_N, HsdM_N superfamily, MD: N6_Mtase, hsdM |
|        |                                                               | N-6 DNA Methylase family protein                                 | 974         | 100               | 0.0          | SH: HsdM_N, NH: UPF0020, HsdM_N superfamily, MD: HsdM, N6_Mtase, COG0610, hsdR, DEXDc |
|        |                                                               | Type I site-specific deoxyribonuclease, HsdR family protein     | 1,409       | 100               | 0.0          | NH: Methylation_S, Methylation_S superfamily, MD: N6_Mtase, HsdM, HsdS, PRK09737 |
|        |                                                               | Hypothetical protein ECSTECC16502_0289, [ABC transporter ATP-binding protein] | 1,107       | 100               | 0.0          | SH: Hnhc, Hnhc superfamily, MD: AAA_15, COG3950 |
|        |                                                               | Hypothetical protein ECSTECC16502_0290                          | 106         | 98                | 4e-27        | NH: PRK14512, S14_ClpP_1, ClpP, ClpP_protease_like superfamily, MD: int |
|        |                                                               | Hypothetical protein ECSTECC16502_0291                          | 74.3        | 100               | 4e-16        | NH: SR_ResInv, Resolvase, Ser_Recombinase superfamily, MD: PinR |
|        |                                                               | Type I restriction-modification system specificity determinant   | 1,176       | 100               | 0.0          | SH: HNH_2, HNHc superfamily, MD: COG3440 |
|        |                                                               | Putative membrane protein                                       | 364         | 100               | 6e-115       | MD: HpaA, AraC, HTH_ARAC, HTH_18, PRK09685 |
|        |                                                               | Type III restriction enzyme, res subunit                        | 2,261       | 100               | 0.0          | MH: F5s superfamily, MD: HpaX, F5s_1, PRK11551, NarK |
|        |                                                               | Hypothetical protein ECSTECC16502_0339                          | 1,450       | 100               | 0.0          | NH: HNH_2, HNHc superfamily, MD: COG3440 |
|        |                                                               | Metallo-beta-lactamase superfamily protein                     | 569         | 100               | 0.0          | SH: ElaC, NH: RNase_Z, PRK00055, Lactamase_B_2, Lactamase_B superfamily |
|        |                                                               | 4-Hydroxyphenylacetate catabolism regulatory protein HpaA       | 618         | 100               | 0.0          | MD: HpaA, AraC, HTH_ARAC, HTH_18, PRK09685 |
|        |                                                               | 4-Hydroxyphenylacetate permease                                 | 850         | 100               | 0.0          | NH: HsdR_N, MD: COG0610, hsdR, hsdR |
|        |                                                               | Hypothetical protein ECSTECC16502_0280, [HNH endonuclease]      | 516         | 99                | 0.0          | NH: PRK11551, NarK |
|        |                                                               | Filamentation induced by CAMP protein Fic                       | 138         | 100               | 2e-42        | MH: F5s superfamily, MD: HpaX, F5s_1, PRK11551, NarK |
| B      | 1 Mbp to 2 Mbp                                              | Phage virion morphogenesis protein                              | 70.5        | 100               | 1e-14        | NH: GP4d_helicase, RecA_like superfamily, MD: AAA_15, COG3950 |
|        |                                                               | Hypothetical protein ECSTECC16502_1311                          | 133         | 98                | 1e-37        | NH: PRK14512, S14_ClpP_1, ClpP, clpP, ClpP_protease_like superfamily, MD: int |
|        |                                                               | AAA ATPase                                                     | 795         | 100               | 0.0          | NH: PRK14512, S14_ClpP_1, ClpP, clpP, ClpP_protease_like superfamily, MD: int |
|        |                                                               | Putative membrane protein                                       | 852         | 100               | 0.0          | NH: PRK14512, S14_ClpP_1, ClpP, clpP, ClpP_protease_like superfamily, MD: int |
|        |                                                               | Hypothetical protein ECSTECC16502_4950, [Rz1 lytic protein]     | 62.0        | 100               | 1e-12        | NH: PRK14512, S14_ClpP_1, ClpP, clpP, ClpP_protease_like superfamily, MD: int |
|        |                                                               | Hypothetical protein ECSTECC16502_4968                          | 442         | 100               | 4e-159       | NH: PRK14512, S14_ClpP_1, ClpP, clpP, ClpP_protease_like superfamily, MD: int |
|        |                                                               | Acetyltransferase family protein                                | 394         | 100               | 4e-141       | NH: PRK14512, S14_ClpP_1, ClpP, clpP, ClpP_protease_like superfamily, MD: int |
|        |                                                               | Hypothetical protein ECSTECC16502_4986                          | 51.2        | 100               | 2e-10        | NH: DUF2829, DUF2829 superfamily |
|        |                                                               | isA                                                            | 43.9        | 95                | 6e-06        | MH: Int |
|        |                                                               | Hypothetical protein ECSTECC16502_1295, [acetyltransferase]     | 120         | 100               | 2e-35        | MH: Int |
|        |                                                               | Integrase                                                      | 208         | 100               | 3e-68        | MH: Int |
|        |                                                               | Tail fiber                                                     | 176         | 100               | 4e-55        | MH: int |
|        |                                                               | DNA-invertase                                                  | 92.8        | 100               | 2e-23        | NH: SR_ResInv, Resolvase, Ser_Recombinase superfamily, MD: PinR |
|        |                                                               | Caulovirales tail fiber assembly family protein                 | 338         | 100               | 6e-118       | NH: Caulo_TAP, Caulo_TAP superfamily |
|        |                                                               | Tail fiber domain protein                                      | 201         | 100               | 4e-65        | NH: Caulo_TAP, Caulo_TAP superfamily |
### Supplementary Table 5. Signature details based on *Escherichia coli STEC_C165_02* reference

| Region | Signature Details | E Value | Length (AA) | Notes |
|--------|-------------------|---------|-------------|-------|
| Recombination enhancement | 298 | 99 | 3e-104 | NH: Collar, Collar superfamily |
| Phage Tail Collar domain protein | 134 | 95 | 2e-41 | |
| DNA-Invertase | 28.9 | 100 | 0.061 | |
| Hypothetical protein ECSTECC16502_4984, [lysogeny establishment protein] | 120 | 100 | 6e-37 | |
| Phage tail fiber repeat family protein | 138 | 100 | 1e-41 | SH: Phage_fiber_2, Phage_fiber_2 superfamily, MD: COG5301 |
| Hypothetical protein ECSTECC16502_1393 | 103 | 100 | 3e-31 | |
| C | 1 Mbp to 2 Mbp | Hypothetical protein ECSTECC16502_1803 | 264 | 100 | 4e-87 |
| Hypothetical protein ECSTECC16502_1804 | 122 | 100 | 3e-34 | |
| Hypothetical protein ECSTECC16502_1805 | 213 | 100 | 1e-72 | |
| Hypothetical protein ECSTECC16502_1809 | 156 | 100 | 6e-49 | |
| Hypothetical protein ECSTECC16502_1807 | 138 | 100 | 7e-41 | |
| D | 2 Mbp to 3 Mbp | Hypothetical protein ECSTECC16502_2561 | 194 | 100 | 3e-62 |
| Peptidoglycan domain protein | 68.9 | 100 | 8e-16 | |
| Hypothetical protein ECSTECC16502_2576 | 188 | 100 | 5e-62 | |
| Hypothetical protein ECSTECC16502_2577 | 74.7 | 100 | 3e-18 | |
| gp41 domain protein | 312 | 99 | 2e-110 | |
| Hypothetical protein ECSTECC16502_2586 | 32.0 | 100 | 0.006 | |
| Hypothetical protein ECSTECC16502_2559 | 75.9 | 97 | 1e-19 | |
| Hypothetical protein ECSTECC16502_2587, [DNA-binding protein] | 168 | 100 | 1e-55 | NH: PHA00675, PHA00675 superfamily |
| Hypothetical protein ECSTECC16502_2588 | 133 | 100 | 8e-41 | NH: DUF1627, DUF1627 superfamily |
| E | 2 Mbp to 3 Mbp | Prophage CP4-57 integrase | 108 | 100 | 1e-27 | NH: INT_P4,DNA_BRE_C superfamily, MD: PRK09692 |
| Hypothetical protein ECSTECC16502_2827 | 168 | 100 | 6e-53 | |
| Hypothetical protein ECSTECC16502_2828 | 99.4 | 100 | 5e-27 | |
| Prophage CP4-57 regulatory protein family protein | 136 | 99 | 1e-40 | NH: Phage_AlpA, AlpA, Phage_AlpA superfamily |
| Terminase small subunit | 374 | 100 | 2e-132 | SH: COG4220, NH: Phage_Nu1, Phage_Nu1 superfamily |
| Hypothetical protein ECSTECC16502_2842 | 101 | 100 | 5e-28 | |
| Major head protein | 286 | 100 | 8e-97 | NH: Phage_cap_E, Phage_cap_E superfamily |
| Prophage CP4-57 integrase | 237 | 100 | 6e-77 | NH: INT_P4,DNA_BRE_C superfamily, SH: DUF4102, DUF4102 superfamily, MD: PRK09692 |
| Hypothetical protein ECSTECC16502_2843 | 128 | 100 | 2e-40 | |
| F | 3 Mbp to 4 Mbp | DNA topoisomerase IV, A subunit | 177 | 100 | 6e-54 | NH: TOP4c, TOP4c superfamily, MD: parc_Gneg, PRK05561, GyrA, TOP4c, DNA_topoisoIV, S2, PLN03128 |
| G | 4 Mbp to 5 Mbp | Cytolethal distending toxin A/C family protein | 456 | 100 | 3e-154 | NH: CDToxinA, CDToxinA superfamily |
| Endonuclease/Exonuclease/phosphatase family protein | 546 | 100 | 0.0 | NH: PRK15251, CdtB, EEP superfamily |
| Hypothetical protein ECSTECC16502_4757 | 189 | 99 | 3e-62 | |
| H | 4.5 Mbp to end of the genome | upf89.5 | 194 | 100 | 4e-59 | NH: PRK10276, peptidase_S24_S26 superfamily, MD: LexA |
| Putative lipoprotein | 185 | 100 | 2e-57 | |
| humD | 122 | 100 | 3e-34 | |
| Hot protein | 87.0 | 98 | 1e-21 | |
| Hypothetical protein ECSTECC16502_1327 | 96.7 | 98 | 3e-25 | |
| Terminase B protein domain protein | 565 | 100 | 0.0 | |
**Supplementary Table 5.** Signature details based on *Escherichia coli STEC_C165_02* reference

| Protein Type                     | Identity | Coverage | Score   |
|----------------------------------|----------|----------|---------|
| Putative membrane protein        | 80.9     | 100      | 6e-21   |
| Hypothetical protein ECSTECC16502_1333 | 155     | 100      | 3e-49   |
| Terminase B protein              | 283      | 100      | 1e-94   |
| VRR-NUC domain protein          | 165      | 100      | 2e-53   |
| Hypothetical protein ECSTECC16502_4927 | 57.4   | 100      | 2e-13   |

NH: Phage_ASH, Phage_ASH superfamily
## Supplementary Table 6. A alphabetic abbreviation and description of putative conserved domains

| Alphabetic Abbreviation | Description |
|------------------------|-------------|
| 17                     | Large terminase protein |
| 2_A_01_02              | Multidrug resistance protein |
| 2A0115                 | Benzoate transport; [Transport and binding proteins, Carbohydrates, organic alcohols] |
| S2                    | DNA topoisomerase II medium subunit; Provisional |
| AAA_13                 | AAA domain; This family of domains contain a P-loop motif |
| AAA_15                 | AAA ATPase domain; This family of domains contain a P-loop motif |
| AAA_21                 | AAA domain |
| AAA_23                 | AAA domain |
| ABC_RecF               | ATP-binding cassette domain of RecF; RecF is a recombinational DNA repair ATPase |
| ABC_SMC_barmotin       | ATP-binding cassette domain of barmotin, a member of the SMC protein family |
| AcCoA-C-Actrans        | Acetyl-CoA acetyltransferases |
| AHBA_syn               | 3-Amino-5-hydroxybenzoic acid synthase family (AHBA_syn) |
| AidA                   | Type V secretory pathway, adhesin AidA [Cell envelope biogenesis] |
| Ail_Lom                | Enterobacterial Ail/Lom protein; This family consists of several bacterial and phage Ail_Lom proteins |
| AIP3                   | Actin interacting protein 3; Aip3p/Bud6p is a regulator of cell and cytoskeletal polarity |
| Aldose_epim_Ec_YphB    | Aldose 1-epimerase, similar to Escherichia coli YphB |
| AlpA                   | Predicted transcriptional regulator [Transcription] |
| AntA                   | AntA/AntB antirepressor |
| AraC                   | AraC-type DNA-binding domain-containing proteins [Transcription] |
| AsLA                   | Arylsulfatase A and related enzymes [Inorganic ion transport and metabolism] |
| Baseplate_J            | Baseplate J-like protein; The P2 bacteriophage J protein lies at the edge of the baseplate |
| Beta_protein           | Beta protein; This family includes the beta protein from Bacteriophage T4 |
| BID_2                  | Bacterial Ig-like domain 2 |
| Bro-N                  | BRO family, N-terminal domain; This family includes the N-terminus of baculovirus BRO |
| btuB                   | Vitamin B12/cobalamin outer membrane transporter; Provisional |
| BtuB                   | Outer membrane cobalamin receptor protein [Coenzyme metabolism] |
| Caps_syn               | Capsular polysaccharide synthesis protein |
| Caudo_TAP              | Caudovirales tail fibre assembly protein |
| ccpA                   | Catabolite control protein A |
| CdtB                   | CdtB, the catalytic DNase I-like subunit of cytolethal distending toxin (CDT) protein |
| CDtoxA                 | Cytolethal distending toxin A/C family |
| Cep57_CLD              | Centrosome localisation domain of Cep57 |
| CESA_like              | CESA_like is the cellulose synthase superfamily; The cellulose synthase (CESA) superfamily |
| chol_sulfatase         | Choline-sulfatase; |
| clpP                   | ATP-dependent Clp endopeptidase, proteolytic subunit ClpP |
| ClpP                   | Protease subunit of ATP-dependent Clp proteases |
| CLP_protease           | Clp protease; The Clp protease has an active site catalytic triad |
| COG0436                | Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport and metabolism] |
| COG0610                | Type I site-specific restriction-modification system, R (restriction) subunit and related proteins |
| COG1216                | Predicted glycosyltransferases [General function prediction only] |
| COG1340                | Uncharacterized archaeal coiled-coil protein [Function unknown] |
| COG1357                | Pentapeptide repeats containing protein [Function unknown] |
| COG1451                | Predicted metal-dependent hydrolase [General function prediction only] |
| COG1479                | Uncharacterized conserved protein [Function unknown] |
**Supplementary Table 6. Alphabetical abbreviation and description of putative conserved domains**

| COG2253 | Uncharacterized conserved protein [Function unknown] |
|---------|------------------------------------------------------|
| COG2369 | Uncharacterized protein, homolog of phage Mu protein gp30 [Function unknown] |
| COG3328 | Transposase and inactivated derivatives [DNA replication, recombination, and repair] |
| COG3440 | Predicted restriction endonuclease [Defense mechanisms] |
| COG3497 | Phage tail sheath protein Fl [General function prediction only] |
| COG3498 | Phage tail tube protein FlI [General function prediction only] |
| COG3499 | Phage protein U [General function prediction only] |
| COG3500 | Phage protein D [General function prediction only] |
| COG3547 | Transposase and inactivated derivatives [DNA replication, recombination, and repair] |
| COG3561 | Phage anti-repressor protein [Transcription] |
| COG3566 | Uncharacterized protein conserved in bacteria [Function unknown] |
| COG3567 | Uncharacterized protein conserved in bacteria [Function unknown] |
| COG3586 | Uncharacterized conserved protein [Function unknown] |
| COG3617 | Prophage antirepressor [Transcription] |
| COG3628 | Phage baseplate assembly protein W [General function prediction only] |
| COG3637 | Opacity protein and related surface antigens [Cell envelope biogenesis, outer membrane] |
| COG3772 | Phage-related lysozyme (muraminidase) [General function prediction only] |
| COG3910 | Predicted ATPase [General function prediction only] |
| COG3948 | Phage-related baseplate assembly protein [General function prediction only] |
| COG3950 | Predicted ATP-binding protein involved in virulence [General function prediction only] |
| COG4127 | Uncharacterized conserved protein [Function unknown] |
| COG4220 | Phage DNA packaging protein, Nu1 subunit of terminase |
| COG4372 | Phage-related lysozyme [General function prediction only] |
| COG4396 | Mu-like prophage host-nuclease inhibitor protein Gam [General function prediction only] |
| COG4453 | Uncharacterized protein conserved in bacteria [Function unknown] |
| COG4643 | Uncharacterized protein conserved in bacteria [Function unknown] |
| COG4688 | Uncharacterized protein conserved in bacteria [Function unknown] |
| COG4694 | Uncharacterized protein conserved in bacteria [Function unknown] |
| COG4718 | Phage-related protein [Function unknown] |
| COG4753 | Response regulator containing CheY-like receiver domain and AraC-type DNA-binding domain |
| COG4886 | Leucine-rich repeat (LRR) protein [Function unknown] |
| COG5281 | Phage-related minor tail protein [Function unknown] |
| COG5283 | Phage-related tail protein [Function unknown] |
| COG5301 | Phage-related tail fibre protein [General function prediction only] |
| COG5484 | Uncharacterized conserved protein [Function unknown] |
| COG5492 | Bacterial surface proteins containing Ig-like domains [Cell motility and secretion] |
| COG5518 | Bacteriophage capsid portal protein [General function prediction only] |
| COG5525 | Phage terminase, large subunit GpA [Replication, recombination and repair] |
| COG5613 | Uncharacterized conserved protein [Function unknown] |
| Collagen | Collagen triple helix repeat (20 copies) |
| Collar | Phage Tail Collar Domain |
| Cu-Zn_Superoxide_Dismutase | Copper/zinc superoxide dismutase (SOD) |
| Cyt_CS_DNA_methylase | Cytosine-CS specific DNA methylases |
| D | tail protein; Provisional |
| dam | DNA adenine methylase (dam) |
### Supplementary Table 6. Alphabetical abbreviation and description of putative conserved domains

| Abbreviation | Description |
|--------------|-------------|
| Dam          | Site-specific DNA methylase [DNA replication, recombination, and repair] |
| dapA         | Dihydriodipicolinate synthase; Dihydriodipicolinate synthase is a homotrameric enzyme |
| DapA         | Dihydriodipicolinate synthase/N-acetylneuraminic lyase |
| Dcm          | Site-specific DNA methylase [DNA replication, recombination, and repair] |
| Dcu A        | Anaerobic c4-dicarboxylate membrane transporter family protein |
| Dcu A Dcu B  | Anaerobic C4-dicarboxylate transporter [General function prediction only] |
| DEAD         | DEAD/DEAH box helicase; Members of this family include the DEAD and DEAH box helicases |
| DEADc        | DEAD-box helicases. A diverse family of proteins involved in ATP-dependent RNA unwinding |
| DedA         | Uncharacterized membrane-associated protein [Function unknown] |
| DEED_Tnp_\_S110 | Transposase; Transposase proteins are necessary for efficient DNA transposition |
| DegT_Dnrl_EryC1 | DegT/Dnrl/EryC1/StrS aminotransferase family |
| dexA         | Exonuclease |
| DEXDc        | DEAD-like helicases superfamily |
| DEXH_lig_assoc | DEAH box helicase, DNA ligase-associated |
| DHDPS        | Dihydriodipicolinate synthetase family; This family has a TIM barrel structure |
| DHDPS-like   | Dihydriodipicolinate synthase family; Dihydriodipicolinate synthase family |
| DLP_2        | Dynamin-like protein including dynamins, mitofusins, and guanylate-binding proteins |
| DnaB_C       | DnaB helicase C terminal domain |
| DNA_methylase | C-5 Cytosine-specific DNA methylase |
| DnaN         | DNA polymerase sliding clamp subunit [PCNA homolog] [DNA replication, recombination] |
| DNA_pol3_theta | DNA polymerase III, theta subunit |
| DNA_topoisIV | DNA gyrase/topoisomerase IV, subunit A |
| Doc          | Prophage maintenance system killer protein [General function prediction only] |
| DOC_P1       | Death-on-curing family protein |
| DUF1073      | Protein of unknown function (DUF1073) |
| DUF1076      | Protein of unknown function (DUF1076); This family consists of several hypothetical bacterial proteins |
| DUF1133      | Protein of unknown function (DUF1133) |
| DUF1187      | Protein of unknown function (DUF1187) |
| DUF1311      | Protein of unknown function (DUF1311) |
| DUF1482      | Protein of unknown function (DUF1482) |
| DUF1524      | Protein of unknown function (DUF1524) |
| DUF1627      | Protein of unknown function (DUF1627) |
| DUF2213      | Uncharacterized protein conserved in bacteria (DUF2213) |
| DUF2544      | Protein of unknown function (DUF2544) |
| DUF2586      | Protein of unknown function (DUF2586) |
| DUF2590      | Protein of unknown function (DUF2590); This family of proteins has no known function |
| DUF2597      | Protein of unknown function (DUF2597) |
| DUF2612      | Protein of unknown function (DUF2612); This is a phage protein family |
| DUF262       | Protein of unknown function DUF262 |
| DUF2765      | Protein of unknown function (DUF2765); This family of proteins has no known function |
| DUF2791      | P-loop Domain of unknown function (DUF2791); This is a family of proteins found in archaea |
| DUF2829      | Protein of unknown function (DUF2829) |
### Supplementary Table 6. Alphabetic abbreviation and description of putative conserved domains

| Abbreviation   | Description                                                                 |
|----------------|-----------------------------------------------------------------------------|
| DUF3380        | Protein of unknown function (DUF3380)                                        |
| DUF3383        | Protein of unknown function (DUF3383)                                        |
| DUF3486        | Protein of unknown function (DUF3486)                                        |
| DUF3491        | Protein of unknown function (DUF3491); This family of proteins is functionally uncharacterized |
| DUF3584        | Protein of unknown function (DUF3584); This protein is found in bacteria and eukaryotes |
| DUF3672        | Fibronectin type III protein; This domain family is found in bacteria and viruses |
| DUF3751        | Phage tail-collar fibre protein; This domain family is found in bacteria and viruses |
| DUF3850        | Domain of Unknown Function with PDB structure (DUF3850)                      |
| DUF4102        | Domain of unknown function (DUF4102)                                         |
| DUF4353        | Domain of unknown function (DUF4353)                                         |
| DUF4376        | Domain of unknown function (DUF4376)                                         |
| DUF4406        | Protein of unknown function (DUF4406)                                        |
| DUF45          | Protein of unknown function DUF45                                            |
| DUF754         | Protein of unknown function (DUF754); This domain appears to be found in a group of prophage |
| Dynamin_N     | Dynamin family                                                               |
| EcoRII-C       | EcoRII C terminal; The C-terminal catalytic domain of the Restriction Endonuclease EcoRII |
| EcoRII-N       | Restriction endonuclease EcoRII, N-terminal                                   |
| ElaC           | Metal-dependent hydrolases of the beta-lactamase superfamily III [General function prediction] |
| endolysin_autolysin | Endolysins and autolysins are found in viruses and bacteria, respectively     |
| EpsG           | EpsG family                                                                  |
| EspA           | EspA-like secreted protein; EspA is the prototypical member of this family   |
| FAA_hydrolase  | Fumarylacetoacetate (FAA) hydrolase family                                   |
| FhaB           | Large exoproteins involved in heme utilization or adhesion                    |
| FhaC           | Hemolysin activation/secretion protein [Intracellular trafficking and secretion] |
| FI             | Major tail sheath protein; Provisional                                        |
| FII            | Major tail tube protein; Provisional                                         |
| Fil_haemagg_2 | Haemagglutinin repeat                                                         |
| fil_hemag_20aa | Adhesin HecA family 20-residue repeat (two copies)                            |
| Flavodoxin_2   | Flavodoxin-like fold; This family consists of a domain with a flavodoxin-like fold |
| FlIC           | Flagellin protein; This domain family is found in bacteria                   |
| FrhB           | Coenzyme F420-reducing hydrogenase, beta subunit [Energy production and conversion] |
| FrhB_FdhB_C    | Coenzyme F420 hydrogenase/dehydrogenase, beta subunit C terminus             |
| G1P_TT_short   | G1P_TT_short is the short form of glucose-1-phosphate thymidylyltransferase    |
| GalM           | Galactose mutarotase and related enzymes [Carbohydrate transport and metabolism] |
| galU           | UTP-glucose-1-phosphate uridylyltransferase                                    |
| GalU86         | UDP-glucose pyrophosphorylase [Cell envelope biogenesis, outer membrane]       |
| Gam            | Host-nuclease inhibitor protein Gam; The Gam protein inhibits RecBCD nuclease   |
| GATase1_DJ-1   | Type 1 glutamine amidotransferase (GATase1)-like domain found in Human DJ-1   |
| Glf            | UDP-galactopyranose mutase [Cell envelope biogenesis, outer membrane]          |
| GLF            | UDP-galactopyranose mutase                                                    |
| glyc2_xrt_Gpos1 | putative glycosyltransferase, exosortase G-associated                          |
| Glyco_hydro_108 | Glycosyl hydrolase 108; This family acts as a lysozyme (N-acetylmuramidase)    |
| Glycos_transf_2 | Glycosyl transferase family 2; Diverse family, transferring sugar from UDP-glucose |
| Glyco_transf_2_3 | Glycosyltransferase like family 2                                              |
| Glyco_transf_GTA_type | Glycosyltransferase family A (GT-A) includes diverse families of glycosyl transferases |
**Supplementary Table 6. Aalphabetic abbreviation and description of putative conserved domains**

| Abbreviation       | Description                                                                 |
|--------------------|-----------------------------------------------------------------------------|
| Golgin_A5          | Golgin subfamily A member 5                                                |
| GP4d_helicase      | GP4d_helicase is a homoheaxemeric 5'-3' helicases                          |
| gpl                | Bacteriophage P2-related tail formation protein [General function prediction only] |
| gpV                | Phage P2 baseplate assembly protein gpV [General function prediction only]  |
| GPW_gp25           | Gene 25-like lysozyme; This family includes the phage protein Gene 25 from T4 |
| gram_neg_porins    | Porins form aqueous channels for the diffusion of small hydrophilic molecules |
| GT_2_like_b        | Subfamily of Glycosyltransferase Family GT2 of unknown function            |
| GT_2_like_d        | Subfamily of Glycosyltransferase Family GT2 of unknown function            |
| GyrA               | Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV)              |
| Haemagg_act        | Haemagglutination activity domain                                           |
| Helicase_C         | Helicase conserved C-terminal domain; The Prosite family is restricted to DEAD/H helicases |
| HELICc             | Helicase superfamly C-terminal domain; associated with DEXDc-, DEAD-, and DEAH-box proteins |
| Hia                | Autotransporter adhesin [Intracellular trafficking and secretion / Extracellular structures] |
| HicB               | Predicted nuclease of the RNase H fold, HicB family [General function prediction only] |
| HipB               | Predicted transcriptional regulators [Transcription]                       |
| HlyC               | RTX toxin acyltransferase family; (hemolysin-activating protein)            |
| HNH_2              | HNH endonuclease                                                           |
| HP1_INT_C          | Phage HP1 integrase, C-terminal catalytic domain. Bacteriophage HP1 and related integrases |
| HpaA               | 4-Hydroxyphenylacetate catabolism regulatory protein HpaA, putative transcriptional protein |
| HpaG_C-term        | 4-Hydroxyphenylacetate degradation bifunctional isomerase/decarboxylase, C-terminal subunit |
| HpaG-N-term        | 4-Hydroxyphenylacetate degradation bifunctional isomerase/decarboxylase, N-terminal subunit |
| HpaX               | 4-Hydroxyphenylacetate permease                                            |
| HsdM               | Type I restriction-modification system methyltransferase subunit [Defense mechanisms] |
| HsdM_N             | HsdM N-terminal domain; This domain is found at the N-terminus of the methylase subunit |
| hsdR               | Type I site-specific deoxyribonuclease, HsdR family                        |
| HSDR_N             | Type I restriction enzyme R protein N terminus (HSDR_N)                    |
| HsdS               | Restriction endonuclease S subunits [Defense mechanisms]                   |
| HTH_18             | Helix-turn-helix domain                                                    |
| HTH_19             | Helix-turn-helix domain; Members of this family contains a DNA-binding helix-turn-helix domain |
| HTH_35             | Winged helix-turn-helix DNA-binding                                         |
| HTH_36             | Helix-turn-helix domain                                                    |
| HTH_ARAC           | helix_turn_helix, arabinose operon control protein                          |
| HTH_LacI           | Helix-turn-helix (HTH) DNA binding domain of the LacI family of transcriptional regulators |
| HTH_LACI           | Helix turn_helix lactose operon repressor                                   |
| HTH_Tnp_Mu_1       | Mu DNA-binding domain; This family consists of MuA-transposase and repressor protein Cl |
| HTH_XRE            | Helix-turn-helix XRE-family like proteins                                   |
| IncA               | IncA protein                                                               |
| int                | Integrase; Provisional                                                      |
| Int                | Integrase                                                                  |
| Integrate_1        | Integrate; This is a family of DNA-binding prophage integrases found in Proteobacteria. |
| INT_Lambda_C       | Lambda integrase, C-terminal catalytic domain                             |
| INT_P4             | Bacteriophage P4 integrase. P4-like integrases are found in temperate bacteriophages |
| INT_REC_C          | DNA breaking-rejoining enzymes, Integrate/recombinases, C-terminal catalytic domain |
| IpaB_EvcA          | IpaB/EvcA family; This family includes IpaB, which is an invasion plasmid antigen |
| ISH2_PI3K_IA_R     | Inter-Src homology 2 (ISH2) helical domain of Class IA Phosphoinositide 3-kinase Regulatory protein |
Supplementary Table 6. Aalphabetic abbreviation and description of putative conserved domains

| Abbreviation | Description |
|--------------|-------------|
| J | Baseplate assembly protein; Provisional |
| Lactamase_B_2 | Beta-lactamase superfamily domain; This family is part of the beta-lactamase superfamily |
| LexA | SOS-response transcriptional repressors (RecA-mediated autopeptidases) |
| Lhr | Lhr-like helicases [General function prediction only] |
| ligand_gated_channel | ToB dependent/Ligand-Gated channels |
| Lpr | Uncharacterized protein conserved in bacteria, putative lipoprotein [Function unknown] |
| LT_GEWL | Lytic Transglycosylase (LT) and Goose Egg White Lysozyme (GEWL) domain |
| M | Terminase endonuclease subunit; Provisional |
| major_capsid_P2 | Phage major capsid protein, P2 family |
| ManA | Phosphomannose isomerase [Carbohydrate transport and metabolism] |
| MATE_like_10 | Uncharacterized subfamily of the multidrug and toxic compound extrusion (MATE) proteins |
| MATE_Wzx_like | Wzx, a subfamily of the multidrug and toxic compound extrusion (MATE)-like proteins |
| Methylase_S | Type I restriction modification DNA specificity domain |
| Methyltransferase_D12 | Methyltransferase domain; This family contains methyltransferase domains |
| MFS | The Major Facilitator Superfamily (MFS) is a large and diverse group of secondary transporters |
| MFS_1 | Major Facilitator Superfamily |
| MhpD | 2-Keto-4-pentenoate hydratase/2-oxohepta-3-ene-1,7-dioic acid hydratase (catechol pathway) |
| Minor_tail_Z | Prophage minor tail protein Z (GPZ); This family consists of several prophage minor tail |
| MitD | Membrane-bound lytic murein transglycosylase D; Provisional |
| MitE | Soluble lytic murein transglycosylase and related regulatory proteins |
| Mor | Mor transcription activator family; Mor (Middle operon regulator) |
| Mrr | Restriction endonuclease [Defense mechanisms] |
| Mrr_cat | Restriction endonuclease; Prokaryotic family found in type II restriction enzymes |
| N | Capsid protein; Provisional |
| N6_Mtase | N-6 DNA Methylase; Restriction-modification (R-M) systems |
| NA37 | 37-kD nucleoid-associated bacterial protein |
| NAD_binding_8 | NAD(P)-binding Rossmann-like domain |
| NarK | Nitrate/nitrite transporter [Inorganic ion transport and metabolism] |
| NEL | C-terminal novel E3 ligase, LRR-interacting |
| NHT_00031 | Aminotransferase, LLPSF_NHT_00031 family |
| Nip | Predicted transcriptional regulator [Transcription] |
| NTP_transferase | Nucleotidyl transferase |
| O | Capsid-scaffoldin protein; Provisional |
| OCH1 | Mannosyltransferase OCH1 and related enzymes [Cell envelope biogenesis, outer membrane] |
| Ogr_Delta | Ogr/Delta-like zinc finger; This is a viral family of phage zinc-binding transcriptional proteins |
| OMP_b-brl | Outer membrane protein beta-barrel domain |
| OmpC | Outer membrane protein [porin] [Cell envelope biogenesis, outer membrane] |
| OprB | Carbohydrate-selective porin [Cell envelope biogenesis, outer membrane]; OprB family |
| ORF6N | ORF6N domain; This domain was identified by Iyer and colleagues |
| P | Terminase ATPase subunit; Provisional |
| P2_Phage_GpR | P2 phage tail completion protein R (GpR) |
| PaaJ | Acetyl-CoA acetyltransferase [Lipid metabolism] |
| PaaX_trns_reg | Phenylacetic acid degradation operon negative regulatory protein PaaX |
| Packaging_FI | DNA packaging protein F1; This family includes the lambda phage DNA-packaging protein F1 |
**Supplementary Table 6. Aalphabetic abbreviation and description of putative conserved domains**

| Abbreviation | Description |
|--------------|-------------|
| parC_Gneg    | DNA topoisomerase IV, A subunit, proteobacterial; Operationally |
| ParE         | Plasmid stabilization system protein [General function prediction only] |
| PAT1         | Topoisomerase II-associated protein PAT1 |
| PBP1_Lacl_like_7 | Ligand-binding domain of uncharacterized DNA-binding regulatory proteins |
| pcrA         | ATP-dependent DNA helicase PcrA |
| Pentapeptide_4 | Pentapeptide repeats (9 copies) |
| Peptidase_S6 | Immunoglobulin A1 protease; This family consists of immunoglobulin A1 protease proteins |
| Peripla_BP_1 | Periplasmic binding proteins and sugar binding domain of LacI family |
| Peripla_BP_3 | Periplasmic binding protein-like domain; Thi domain is found in a variety of transcriptional proteins |
| PgaC_IcaA    | Poly-beta-1,6 N-acetyl-D-glucosamine synthase |
| PG_binding_3 | Predicted Peptidoglycan domain; This family contains a potential peptidoglycan binding domain |
| PHA00368    | Internal virion protein D |
| PHA00675    | Hypothetical protein |
| PHA01399    | Membrane protein P6 |
| PHA02067    | Hypothetical protein |
| PHA03247    | Large tegument protein UL36; Provisional |
| PHA03255    | BDLF3; Provisional |
| Phage_AlpA  | Prophage CP4-57 regulatory protein (AlpA) |
| Phage_antitermQ | Phage antitermination protein Q; This family consists of several phage antitermination proteins |
| Phage_ASH   | Ash protein family; This family was identified by Iyer and colleagues |
| Phage_attach | Phage Head-Tail Attachment |
| Phage_base_V | Phage-related baseplate assembly protein |
| Phage_cap_E | Phage major capsid protein E |
| Phage_cap_P2 | Phage major capsid protein, P2 family |
| Phage_CI_repr | Bacteriophage CI repressor helix-turn-helix domain |
| Phage_Cox   | Regulatory phage protein cox |
| phage_DnaB  | Phage replicative helicase, DnaB family, HK022 subfamily |
| Phage_fiber_2 | Phage tail fibre repeat; This repeat is found in the tail fibres of phage |
| Phage_GPA   | Bacteriophage replication gene A protein (GPA) |
| Phage_GPD   | Phage late control gene D protein (GPD) |
| Phage_GPL   | Phage head completion protein (GPL) |
| Phage_GPO   | Phage capsid scaffolding protein (GPO) serine peptidase |
| Phage_holin_2 | Phage holin family 2; Holins are a diverse family of proteins |
| Phage_int_N | Bacteriophage lambda integrase, N-terminal domain |
| Phage_integrase | Phage integrase family |
| Phage_int_SAM_2 | Phage integrase, N-terminal; This is a family of DNA-binding prophage integrases |
| Phage_lys      | Bacteriophage Rz lysis protein; This protein is involved in host lysis |
| Phage_lysozyme | Phage lysozyme; This family includes lambda phage lysozyme and Escherichia coli endolysin |
| Phage_min_tail | Phage minor tail protein; This family consists of a series of phage minor tail proteins |
| PhageMin_Tail | Phage-related minor tail protein |
| Phage-MuB_C   | Mu B transposition protein, C terminal; The C terminal domain of the B transposition protein |
| Phage_Mu_F    | Phage Mu protein F like protein; Members of this family are found in double-stranded DNA |
| Phage_Mu_Gam  | Bacteriophage Mu Gam like protein; This family consists of bacterial and phage Gam proteins |
| Phage_Ninh    | Phage NinH protein; This family consists of several phage NinH proteins |
| Phage_Nu1     | Phage DNA packaging protein Nu1; Terminase, the DNA packaging enzyme of bacteriophage lambda |
**Supplementary Table 6. Alphabetical abbreviation and description of putative conserved domains**

| Abbreviation | Description |
|--------------|-------------|
| Phage_P2_GpE | Phage P2 GpE; This family consists of several phage and bacterial proteins |
| Phage_P2_GpU | Phage P2 GpU; This family consists of several bacterial and phage proteins |
| phage_P2_V  | Phage baseplate assembly protein V |
| Phage_portal | Phage portal protein; Bacteriophage portal proteins form a dodecamer |
| Phage_pRha  | Phage regulatory protein Rha (Phage_pRha) |
| phge_rel_HI1409 | Phage-related protein, HI1409 family |
| phage_tail_N | Prophage tail fibre N-terminial; This domain is found at the N-terminus of prophage tail fibre |
| Phage_tail_S | Phage virion morphogenesis family; Protein S of phage P2 |
| phage_term_2 | Phage terminase, large subunit, PBSX family |
| Phage_term_smal | Phage small terminase subunit; This family consists of several phage small terminase subunit |
| Phage_tube  | Phage tail tube protein FII; The major structural components of the contractile tail |
| PinR        | Site-specific recombinases, DNA invertase Pin homologs [DNA replication, recombination] |
| PL1_Passenger_AT | Pertactin-like passenger domains (virulence factors) |
| Plasmid_stabil | Plasmid stabilisation system protein |
| PLN00113    | Leucine-rich repeat receptor-like protein kinase |
| PLN00206    | DEAD-box ATP-dependent RNA helicase; Provisional |
| PLN02288    | Mannose-6-phosphate isomerase |
| PLN02386    | Superoxide dismutase [Cu-Zn] |
| PLN02417    | Dihydrolipidolinate synthase |
| PLN02726    | Dolichyl-phosphate beta-O-mannosyltransferase |
| PLN03114    | ADP-ribosylation factor GTPro-activating protein AGD10; Provisional |
| PLN03128    | DNA topoisomerase 2; Provisional |
| Plug        | TonB-dependent Receptor Plug Domain |
| PM1_typeI   | Phosphomannose isomerase type I |
| Polysacc_synt | Polysaccharide biosynthesis protein; Members of this family are integral membrane proteins |
| Porin_1     | Gram-negative porin |
| portal_PBSX | Phage portal protein, PBSX family |
| Prim_Zn_Ribbon | Zinc-binding domain of primase-helicase |
| PRK00055    | Ribonuclease Z; Reviewed |
| PRK00378    | Nucleoid-associated protein NdpA; Validated |
| PRK00871    | Glutathione-regulated potassium-efflux system ancillary protein KefF; Provisional |
| PRK03170    | Dihydrolipidolinate synthase; Provisional |
| PRK03918    | Chromosome segregation protein; Provisional |
| PRK05561    | DNA topoisomerase IV subunit A; Validated |
| PRK05643    | DNA polymerase III subunit beta; Validated |
| PRK06904    | Replicative DNA helicase |
| PRK07208    | Hypothetical protein; Provisional |
| PRK08026    | Flagellin; Validated |
| PRK09326    | F420H2 dehydrogenase subunit F; Provisional |
| PRK09412    | Anaerobic C4-dicarboxylate transporter; Reviewed |
| PRK09678    | DNA-binding transcriptional regulator; Provisional |
| PRK09685    | DNA-binding transcriptional activator FeaR; Provisional |
| PRK09692    | Integrase; Provisional |
| PRK09706    | Transcriptional repressor DicA; Reviewed |
| PRK09709    | Exonuclease VIII; Reviewed |
**Supplementary Table 6.** Aalpabetic abbreviation and description of putative conserved domains

| Accession  | Description                                                                 |
|------------|-----------------------------------------------------------------------------|
| PRK09737   | EcoKI restriction-modification system protein HsdS; Provisional              |
| PRK09750   | Hypothetical protein; Provisional                                           |
| PRK09866   | Hypothetical protein; Provisional                                           |
| PRK09940   | Transcriptional regulator YdeO; Provisional                                 |
| PRK10018   | Putative glycosyl transferase; Provisional                                   |
| PRK10073   | putative glycosyl transferase; Provisional                                   |
| PRK10122   | GalU regulator GalF; Provisional                                            |
| PRK10159   | Outer membrane phosphoprotein protein E; Provisional                        |
| PRK10276   | DNA polymerase V subunit UmuD; Provisional                                  |
| PRK10344   | DNA-binding transcriptional regulator Nlp; Provisional                      |
| PRK10458   | DNA cytosine methylase; Provisional                                         |
| PRK10554   | Outer membrane porin protein C; Provisional                                 |
| PRK05790   | Putative acyltransferase; Provisional                                       |
| PRK10597   | DNA damage-inducible protein I; Provisional                                 |
| PRK10691   | Hypothetical protein; Provisional                                           |
| PRK10703   | DNA-binding transcriptional repressor PurR; Provisional                     |
| PRK10847   | Hypothetical protein; Provisional                                           |
| PRK10904   | DNA adenine methylase; Provisional                                          |
| PRK10969   | DNA polymerase III subunit theta; Reviewed                                  |
| PRK11551   | Putative 3-hydroxyphenylpropionic transporter MhpT; Provisional             |
| PRK11658   | UDP-4-amino-4-deoxy-L-arabinose–oxoglutarate aminotransferase; Provisional |
| PRK12688   | Flagellin; Reviewed                                                        |
| PRK13759   | Arylsulfatase; Provisional                                                  |
| PRK13767   | ATP-dependent helicase; Provisional                                         |
| PRK14272   | Phosphate ABC transporter ATP-binding protein; Provisional                  |
| PRK14512   | ATP-dependent Clp protease proteolytic subunit; Provisional                 |
| PRK14960   | DNA polymerase III subunits gamma and tau; Provisional                      |
| PRK15099   | O-Antigen translocase; Provisional                                          |
| PRK15131   | Mannose-6-phosphate isomerase; Provisional                                  |
| PRK15203   | 4-Hydroxyphenylacetate degradation bifunctional isomerase/dcarboxylase; Provisional |
| PRK15240   | Resistance to complement killing; Provisional                               |
| PRK15251   | Cytolethal distending toxin subunit CdtB                                    |
| PRK15316   | RatA-like protein; Provisional                                               |
| PRK15319   | AIDA autotransporter-like protein ShdA; Provisional                         |
| PRK15370   | E3 ubiquitin-protein ligase SirP; Provisional                               |
| PRK15377   | E3 ubiquitin-protein ligase SopA; Provisional                               |
| PRK15386   | Type III secretion protein GogB; Provisional                               |
| PRK15387   | E3 ubiquitin-protein ligase SspH2                                           |
| PRK15388   | Cu/Zn superoxide dismutase; Provisional                                     |
| PRK15407   | Lipopolysaccharide biosynthesis protein RfbH; Provisional                   |
| PRK15480   | Glucose-1-phosphate thymidyltransferase RfbA; Provisional                   |
| ProP       | Permeases of the major facilitator superfamily                             |
| PTZ00102   | Polysaccharide pyruvyl transferase                                           |
| PTZ00110   | Helicase; Provisioned                                                       |
Supplementary Table 6. Alphabetic abbreviation and description of putative conserved domains

| Abbreviation             | Description                                                                 |
|--------------------------|-----------------------------------------------------------------------------|
| PTZ00260                | Dolichyl-phosphate beta-glucosyltransferase; Provisional                    |
| PurR                    | Transcriptional regulators [Transcription]                                  |
| Q                       | Portal vertex protein; Provisional                                          |
| recF                    | Recombination protein F; Reviewed                                          |
| recomb_XerC             | Tyrosine recombinase XerC; The phage integrase family describes a number of recombinases |
| recomb_XerD             | Tyrosine recombinase XerD (The phage integrase family)                     |
| Resolvase               | Resolvase, N terminal domain; The N-terminal domain of the resolvase family |
| ResIII                  | Type III restriction enzyme, res subunit                                    |
| RfbA                    | dTDP-glucose pyrophosphorylase [Cell envelope biogenesis, outer membrane]   |
| RfbX                    | Membrane protein involved in the export of O-antigen and teichoic acid     |
| Rhs_assc_core           | RHS repeat-associated core domain; This model represents a conserved unique core sequence |
| rmlA                    | Glucose-1-phosphate thymidylyltransferase                                   |
| RNase_Z                 | Ribonuclease Z                                                             |
| RT_Bac_retron_II        | RT_Bac_retron_II: Reverse transcriptases (RTs) in bacterial retrotransposons or retrons |
| rveA                    | 23S rRNA m(5)U1939 methyltransferase; Reviewed                            |
| rve                      | Integrase core domain                                                      |
| RV1_1                    | Reverse transcriptase (RNA-dependent DNA polymerase)                       |
| S14_ClpP_1              | Caseinolytic protease (ClpP) is an ATP-dependent, highly conserved serine protease |
| SDH_sah                 | Serine dehydrogenase proteinase; This family of archaeobacterial proteins |
| ShIB                    | Haemolysin secretion/activation protein ShIB/FhaC/HecB                     |
| sifB                    | Secreted effector protein SifB; Provisional                                |
| SLT                     | Transglycosylase SLT domain; This family is distantly related to pfam00062 |
| Smc                     | Chromosome segregation ATPases [Cell division and chromosome partitioning] |
| SMC_N                   | RecF/RecN/SMC N terminal domain; This domain is found at the N terminus of SMC proteins |
| SMC_prok_B              | Chromosome segregation protein SMC, common bacterial type                  |
| SopA_C                  | SopA-like catalytic domain; This domain is found in the Escherichia coli Type III secretion system |
| SodC                    | Cu/Zn superoxide dismutase [Inorganic ion transport and metabolism]        |
| Sod_Cu                  | Copper/zinc superoxide dismutase (SODC)                                    |
| SPEC                    | Spectrin repeats, found in several proteins involved in cytoskeletal structure |
| spore_V_B               | Stage V sporulation protein B; SpoVB is the stage V sporulation protein B  |
| SppA                    | Periplasmic serine proteases (Ctip class) [Posttranslational modification, protein turnover] |
| SR_ResInv               | Serine Recombinase (SR) family, Resolvase and Invertase subfamily, catalytic domain |
| sufB                    | FeS assembly protein SufB; This protein, SufB, forms a cytosolic complex SuFBCD |
| Sugar_tr                | Sugar (and other) transporter                                              |
| Sulfatase               | Sulfatase                                                                  |
| synapt_SV2              | Synaptic vesicle protein SV2                                               |
| tail_comp_S             | Phage virion morphogenesis (putative tail completion) protein              |
| Tail_P2_I               | Phage tail protein [Tail_P2_I]; These sequences represent the family of phage P2 protein I |
| tail_tube               | Phage contractile tail tube protein, P2 family; The tails of some phage are contractile |
| tape_meas_lam_C         | Phage tail tape measure protein, lambda family                             |
| Tape_meas_lam_C         | Lambda phage tail tape-measure protein (Tape_meas_lam_C)                  |
| tape_meas_TP901         | Phage tail tape measure protein, TP901 family, core region                |
| terB                    | Tellurite resistance protein terB; This family contains uncharacterized bacterial proteins |
| TerB-N                  | TerB-N; The TerB-N domain is found N terminus to TerB, and TerB-C containing proteins |
**Supplementary Table 6.** Alphabetic abbreviation and description of putative conserved domains

| Abbreviation | Description |
|--------------|-------------|
| TerB-C       | TerB-C domain; TerB-C occurs C terminal of TerB in TerB-N containing proteins |
| Terminase_3  | Phage terminase large subunit; Initiation of packaging of double-stranded viral DNA |
| Terminase_5  | Putative ATPase subunit of terminase (gpP-like) |
| Terminase_6  | Terminase-like family; This family represents a group of terminase proteins |
| Terminase_gP | Phage terminase large subunit (gpA) |
| thiolase     | Thiolase are ubiquitous enzymes |
| Tigr02646    | Tigr02646 family protein (uncharacterized protein family) |
| TMP_2        | Prophage tail length tape measure protein; This family represents a conserved region |
| TonB-B12     | TonB-dependent vitamin B12 receptor |
| Top4c        | DNA Topoisomerase, subtype IIA; domain A'; bacterial DNA topoisomerase IV, GyrA, ParC |
| Toprim_3     | Toprim domain; The toprim domain is found in a wide variety of enzymes; toprim primase |
| Transposase_20| Transposase IS116/IS110/IS902 family |
| Transposase_mut| Transposase, Mutator family |
| TTSSLRR      | Type III secretion system leucine rich repeat protein |
| UDP-GALP_mut | UDP-galactopyranose mutase |
| UPF0020      | Putative RNA methylase family UPF0020; This domain is probably a methylase |
| UvrD         | DNA-dependent helicase II; Provisional |
| UvrD-helicase| UvrD/REP helicase N-terminal domain |
| V            | Virion protein; Provisional |
| VapI         | Plasmid maintenance system antidote protein [General function prediction only] |
| VI_minor_1   | Type VI secretion-associated protein, VC_A0118 family |
| W            | Baseplate subunit; Provisional |
| WCaA         | Glycosyltransferases involved in cell wall biogenesis |
| WeCE        | Predicted pyridoxal phosphate-dependent enzyme |
| xerC         | Site-specific tyrosine recombinase XerC; Reviewed |
| XerC         | Integrase [DNA replication, recombination, and repair] |
| XerD         | Site-specific recombinase XerD [DNA replication, recombination, and repair] |
| XkdT         | Uncharacterized homolog of phage Mu protein gp47 [Function unknown] |
| zlS          | Lysozyme family protein [General function prediction only] |