G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2017;15(2):69-80
https://doi.org/10.5808/GI.2017.15.2.69

ORIGINAL ARTICLE

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 phylogenic 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, phylogenic 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]. Phylogenic 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

Received April 28, 2017; Revised May 9, 2017; Accepted May 9, 2017

***Corresponding author:** Tel: +98-21-66953311-20, Fax: +98-21-66492619, E-mail: manaoloomi@yahoo.com

between genomes and, consequently, has shed new insights into phylogenetic relationships in recent years [6-9].

In this study, phylogenic 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.

Phylogenic 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)	*cdt* gene	GC%	Protein count	Gene count	Genome type, No. of subsequences/contigs	Pathotype, serotype, other characteristic	Accession No.
Escherichia coli 96.0497	5.01426	+	50.80	4,862	5,026	Draft, 13	Host: homo sapiens, O91:H21	NZ_AEZQ00000000.2
Escherichia coli 3003	4.91733	+	50.7	4,825	4,982	Draft, 8	I.S: water, O157:H45	NZ_AFAF00000000.2
Escherichia coli 5412	5.38651	+	50.20	5,670	5,761	Draft, 373	Host: homo sapiens, SFO157	NZ_AMUJ00000000.1
Escherichia coli 53638	5.37179	+	50.99	4,803	5,218	Draft, 2	EIEC, O144	NZ_AAKB00000000.2
Escherichia coli ARS4.2123	4.98276	+	50.50	5,105	5,194	Draft, 209	I.S: water, O157:H16	NZ_AMUL00000000.1
Escherichia coli DEC3F	5.4079	+	50.30	5,541	5,692	Draft, 93	Host: homo sapiens, SF EHEC O157:H	NZ_AIFJ00000000.1
Escherichia coli KTE11	4.52715	+	50.50	4,109	4,214	Draft, 7	No published information	NZ_ANSR00000000.1
Escherichia coli KTE28	5.0544	+	50.40	4,673	4,760	Draft, 12	No published information	NZ_ANSY00000000.1
Escherichia coli KTE47	4.98747	+	50.60	4,694	4,798	Draft, 11	No published information	NZ_ANUB00000000.1
Escherichia coli KTE60	5.07079	+	50.50	4,664	4,756	Draft, 20	No published information	NZ_ANUJ00000000.1
Escherichia coli KTE137	5.00154	+	50.50	4,702	4,789	Draft, 99	No published information	NZ_ANYA00000000.1
Escherichia coli KTE178	5.30789	+	50.60	4,973	5,050	Draft, 11	No published information	NZ_ANTB00000000.1
Escherichia coli KTE180	5.12548	+	50.60	4,883	4,966	Draft, 112	No published information	NZ_ANYR00000000.1
Escherichia coli KTE209	5.11008	+	50.50	4,702	4,791	Draft, 3	No published information	NZ_ANXD00000000.1
Escherichia coli MS 21-1	5.30899	+	50.40	5,744	5,860	Draft, 206	No published information	NZ_ADTR00000000.1
Escherichia coli O157:H-493-89	5.05482	+	50.50	4,838	4,946	Draft, 204	Host: homo sapiens, O157:H-	NZ_AETY00000000.1
Escherichia coli O157:H43 T22	4.95898	+	50.80	4,859	4,935	Draft, 64	I.S: milk from healthy cattle, O157:H43	NZ_AHZD00000000.2
Escherichia coli RN587/1	5.06158	+	50.60	4,999	5,108	Draft, 73	EPEC, O157:H8	NZ_ADUS00000000.1
Escherichia coli STEC B2F1	4.98941	+	50.90	4,875	5,006	Draft, 37	STEC, O91:H21	NZ_AFDQ00000000.1
Escherichia coli STEC C165-02	5.00927	+	50.60	4,891	5,019	Draft, 30	STEC, O73:H16	NZ_AFDR00000000.1
Escherichia coli TA271	5.07582	+	50.70	5,081	5,197	Draft, 83	Host: some mammal	NZ_ADAZ00000000.1
Escherichia coli TW06591	5.47546	+	50.30	5,521	5,650	Draft, 45	Host: homo sapiens, O157:H-	NZ_AKLT00000000.1
Escherichia coli W26	5.11853	+	50.60	4,852	4,920	Draft, 165	Host: cow, I.S: feces	NZ_AGIA00000000.1
Escherichia albertii TW07627	4.74659	+	49.90	4,386	4,889	Draft, 43	Diarrhea genic	NZ_ABKX00000000.1
Escherichia coli APEC O1	5.49765	+	50.29	4,853	4,968	Complete, 3	ExPEC, O1:K1:H7, avian pathogenic	NC_008563.1
Escherichia coli IHE3034	5.10838	+	50.70	4,966	4,753	Complete, 1	ExPEC, O18:K1:H7, meningitis	NC_017628.1

Table 1. Continued

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 042	5.35532	-	50.58	4,920	5,036	Complete, 2	EAEC, 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	EAEC	NC_011748.1
Escherichia coli ABU 83972	5.13296	-	50.60	4,795	4,905	Complete, 2	ExPEC UTI, OR:K5:H-	NC_017631.1
Escherichia coli APEC 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 ETEC H10407	5.32589	-	50.73	4,872	5,084	Complete, 5	ETEC, O78:H11	NC_017633.1
Escherichia coli HS	4.64354	-	50.80	4,374	4,626	Complete, 1	Commensal, O9	NC_009800.1
Escherichia coli IAI1	4.70056	-	50.80	4,345	4,629	Complete, 1	Commensal	NC_011741.1
Escherichia coli IAI39	5.13207	-	50.60	4,725	5,092	Complete, 1	ExPEC, UPEC, O7:K1	NC_011750.1
Escherichia coli JJ1886	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 MG1655	4.64165	-	50.80	4,140	4,497	Complete, 1	Commensal, K12	NC_000913.3
Escherichia coli K-12 W3110	4.64633	-	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	AIEC	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 11368	5.85553	-	50.66	5,515	5,985	Complete, 5	EHEC, O26:H11	NC_013361.1
Escherichia coli O55:H7 CB9615	5.45235	-	50.48	5,117	5,367	Complete, 2	EPEC, O55:H7	NC_013941.1
Escherichia coli O83:H1 NRG 857C	4.89488	-	50.71	4,582	4,690	Complete, 2	AIEC, O83:H1	NC_017634.1
Escherichia coli O103:H2 12009	5.52486	-	50.68	5,117	5,541	Complete, 2	EHEC, O103:H2	NC_013353.1
Escherichia coli O104:H4 2011C-3493	5.43741	-	50.63	5,149	5,269	Complete, 4	EAEC/STEC, O104:H4	NC_018658.1
Escherichia coli O111:H- 11128	5.76608	-	50.42	5,403	5,931	Complete, 6	EHEC, O111:H	NC_013364.1
Escherichia coli O127:H6 E2348 69	5.06968	-	50.55	4,647	5,011	Complete, 3	EPEC, O127:H6	NC_011601.1
Escherichia coli O157:H7 EC4115	5.70417	-	50.39	5,477	6,066	Complete, 3	EHEC, O157:H7	NC_011353.1
Escherichia coli O157:H7 EDL933	5.6394	-	50.45	5,772	5,920	Complete, 2	EHEC, O157:H7	NC_002655.2
Escherichia coli O157:H7 Sakai	5.59448	-	50.45	5,292	5,448	Complete, 3	EHEC, O157:H7	NC_002695.1
Escherichia coli O157:H7 TW14359	5.62274	-	50.46	5,363	5,586	Complete, 2	EHEC, O157:H7	NC_013008.1
Escherichia coli P12b	4.93529	-	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

Table 1. Continued

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 UMN026	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*.

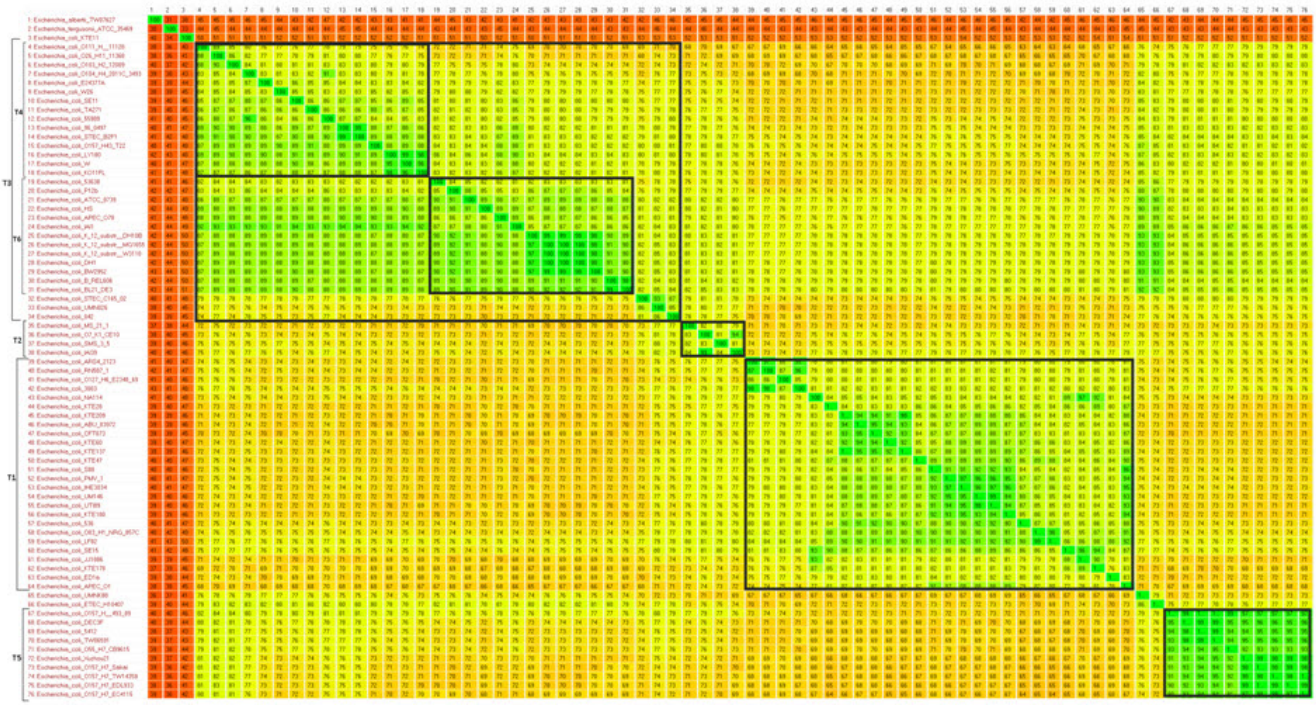

Fig. 1. Phylogenetic heat-plot overview of multiple-genome alignments. A heat plot based on a 200/100 BLASTN fragmented alignment was performed with Gegenees software. Six distinct genomic groups (T1–T6) recognized in cytolethal distending toxin $(CDT)^+$ strains were observed sporadically among the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

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.

Phylogenic analysis

A heat-plot based on a 200/100 BLASTN fragmented alignment drawn with Gegenees software is shown in Fig 1. A phylogenic 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* RN587/1 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, terminuses 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, CDtoxinA, INT_P4, HP1_INT_C, Phage_integrase, INT_Lambda_C, Phage_integ_N, Methylase_S, Caudo_TAP, phage_tail_N, Tail_P2_I, gpI, 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, PRK-10018, and PLN02726
- MD: PRK09692, int, recomb_XerC, XerD, xerC, HsdS, N6_Mtase, HsdM, hsdM, rumA, P, Terminase_6, COG-5484, PLN03114, COG5301, COG0610, hsdR, PRK-10458, PRK10073, Glyco_tranf_2_3, WcaA, PRK10073, and PTZ00260
- Superfamilies: RICIN superfamily, EEP superfamily, DNA_BRE_C superfamily, DUF4102 superfamily, Phage_integ_N superfamily, MCP_signal superfamily,

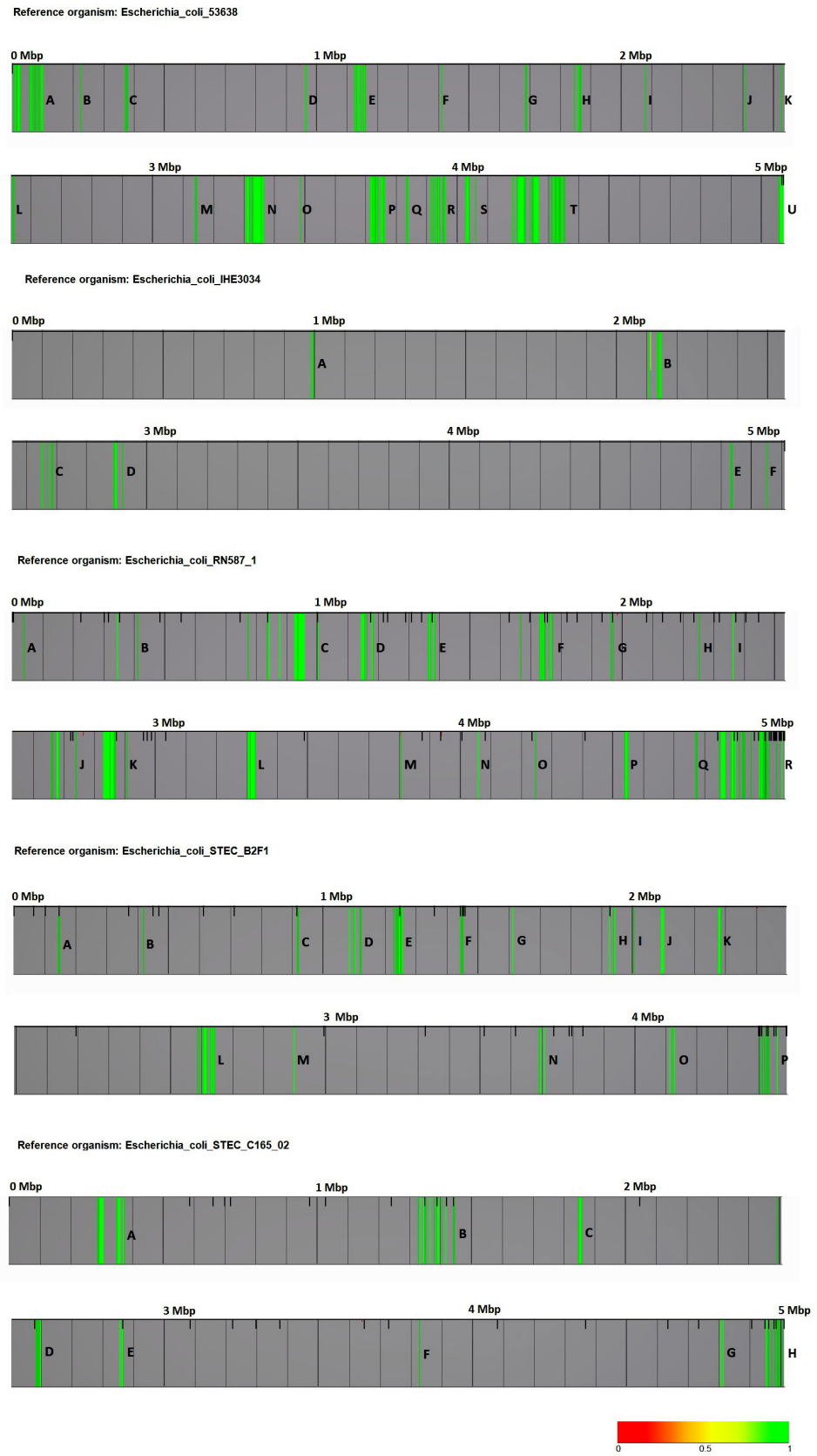

Fig. 3. Biomarker regions. Biomarker regions were illustrated in the whole-genome sequences of five different reference strains including *Escherichia coli* 53638, *E. coli* IHE3034, *E. coli* RN587/1, *E. coli* STEC B2F1, and *E. coli* STEC C165-02. The biomarker score (max/average) setting was used. 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*.

Table 2. Significant signature proteins in five reference *Escherichia coli* strains

Signature protein	Reference strain				
	Escherichia coli 53638	*Escherichia coli* IHE3034	*Escherichia coli* RN587/1	*Escherichia coli* STEC_B2F1	*Escherichia coli* STEC_C165_02
Cytolethal distending toxin	Cytolethal distending toxin A Cytolethal distending toxin B Cytolethal distending toxin subunit C	Cytolethal distending toxin, subunit C Cytolethal distending toxin, subunit B Cytolethal distending toxin, subunit A	Cytolethal distending toxin A/C family protein	Cytolethal distending toxin C Cytolethal distending toxin A/C family protein	Cytolethal distending toxin A/C family protein
Holin	Phage holin, lambda family	Holin, lambda family	Holing	-[a]	Phage holin, lambda family
Nuclease	Exodeoxyribonuclease 8	Exonuclease family protein	Exonuclease family protein Hypothetical protein ECRN5871_4153, [HNH endonuclease family protein]	Endonuclease/Exonuclease/phosphatase family protein Type I site-specific deoxyribonuclease, HsdR family	Restriction endonuclease family protein Type I site-specific deoxyribonuclease, HsdR family protein Hypothetical protein ECSTECC16502_0280, [HNH endonuclease] Endonuclease/Exonuclease/phosphatase family protein
Phage integrase	Phage integrase Prophage integrase Integrase for prophage CP-933T	Integrase/recombinase, phage integrase family Site-specific recombinase, phage integrase family	Integrase Phage integrase family protein Prophage lambda integrase Integrase domain protein	Phage integrase family protein Prophage lambda integrase	Integrase Prophage CP4-57 integrase
Putative membrane protein	Putative membrane protein Hypothetical protein Ec53638_1156, [membrane protein]	Hypothetical protein ECOK1_2122, [membrane protein] Hypothetical protein ECOK1_2557, [membrane protein]	Outer membrane autotransporter barrel domain protein	Putative membrane protein Hypothetical protein ECSTECB2F1_3192, [membrane protein] OmpA-like transmembrane domain protein Outer membrane porin protein LC Outer membrane protein lom	Putative membrane protein
Regulatory proteins	Phage regulatory protein Cro Transcriptional regulator, AlpA family Putative phage regulatory protein, Rha family	Putative transcriptional regulator DicA157 Putative regulatory protein Cox	Regulatory protein CII Prophage CP4-57 regulatory protein family protein Transcriptional regulator, LacI family	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	-[a]	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

Table 2. Continued

Signature protein	Reference strain				
	Escherichia coli 53638	*Escherichia coli* IHE3034	*Escherichia coli* RN587/1	*Escherichia coli* STEC_B2F1	*Escherichia coli* STEC_C165_02
Tail fiber assembly family, baseplate assembly proteins, Tail fiber protein and Tail tape measure protein	Tail fiber assembly protein Phage P2 baseplate assembly protein gpV Putative tail fiber protein Tail fiber Phage tail tape measure protein family	Tail fiber protein Phage tail tape measure protein	Caudovirales tail fiber Assembly family protein Hypothetical protein ECRN5871_3504,[tail fiber assembly protein] Baseplate assembly protein V, W Long tail fiber protein p37 domain protein Tail fiber domain protein Phage tail tape measure protein, TP901 family, core region	Tail fiber assembly Hypothetical protein ECSTECB2F1_0901, [tail fiber assembly protein, caudovirales tail fiber assembly protein] Caudovirales tail fiber assembly family protein Prophage tail fiber family protein Phage tail fiber repeat family protein Phage tail tape measure protein, lambda family	Caudovirales tail fiber assembly family protein Tail fiber Tail fiber domain protein Phage tail fiber repeat family protein
Terminase	Phage terminase large subunit Terminase	-[a]	Phage small terminase subunit Terminase, ATPase subunit Terminase, endonuclease subunit Terminase large subunit Terminase small subunit	Phage terminase large subunit family protein	Terminase small subunit Terminase B protein domain protein Terminase B protein
Transferase	Pyruvyl transferase Glycosyl transferase domain protein, group 2 family Glycosyltransferase, sugar-binding region containing DXD motif	-[a]	Hypothetical protein ECRN5871_3051, [nucleotidyl transferase, PF08843 family] D12 class N6 adenine-specific DNA methyl-transferase family protein Hypothetical protein ECRN5871_0025, [N-acetyltransferase CN5]	Putative teichuronic acid biosynthesis glycosyltransferase tuaG Glucose-1-phosphate thymidylyltransferase RTX toxin acyltransferase family protein Acetyl-CoA acetyltransferase	Acetyltransferase family protein Hypothetical protein ECSTECC16502_1295, [acetyltransferase]

[a]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_Ssuperfamily, Caudo_TAPsuperfamily, phage_tail_Nsuperfamily, Tail_P2_Isuperfamily, Terminase_3superfamily, Terminase_5superfamily, Phage_term_smalsuperfamily, Terminase_GpAsuperfamily, Phage_Nu1superfamily, DnaQ-like-exosuperfamily, Phage_holin_2superfamily, DUF3751 superfamily, Phage_fiber_2superfamily, Gifsy-2 superfamily, HSDR_Nsuperfamily, Cyt_C5_DNA_methylase superfamily, MethyltransfD12superfamily, 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-

regative *E. coli*), EIEC (enteroinvasive *E. coli*), AIEC (adherent invasive *E. coli*), UPEC (uropathogenic *E. coli*), ExPEC (extraintestinal pathogenic *E. coli*), EHEC (enterohemorrhagic *E. coli*), environmental strains and commensal strains were performed.

In fact, phylogenic 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 phylogenomic 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 genetic 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 phylogenic tree, representing the T1–T6 groups.

The findings was presented in this study indicate that the major conserved biomarkers beyond CDT were exonuclease, 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, Nougayrède JP, Dobrindt U, Ledger TN, Boury M, Morabito S, *et al*. Cytolethal distending toxin type I and type IV genes are framed with lambdoid prophage genes in extraintestinal 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, Håfström T, Segerman B. Gegenees: 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, Glasner 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. Kloepper 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	
		Hypothetical protein Ec53638_0018	562	100	0.0	
		Hypothetical protein Ec53638_0019	94.4	98	1e-24	
		Conserved hypothetical protein	249	100	1e-82	SH: HTH_XRE, NH: HTH_19, HTH_XRE superfamily, MD: PRK09706, HipB
		Putative bacteriophage protein	745	100	0.0	
		Addiction module antitoxin, Axe family	154	100	1e-47	
		Plasmid stabilization system protein, RelE/ParE family	216	100	3e-71	
		Phage Head-Tail Attachment	252	99	2e-85	NH: Phage_attach, Gifsy-2 superfamily
		Putative DNA-packaging protein	223	100	5e-74	NH: Packaging_FI, Packaging_FI superfamily
		Bacteriophage CI repressor protein	267	100	1e-91	NH: Phage_CI_repr, Phage_CI_repr superfmily
		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	
		Putative tail component of prophage	203	98	6e-67	
		Putative tail fiber protein	99.4	100	3e-26	
		Exodeoxyribonuclease 8	153	100	2e-44	MD: PRK09709
		Host-nuclease inhibitor protein Gam	153	100	2e-48	NH: Gam, Gam superfamily
		Gifsy-1 prophage VmtH	226	100	3e-70	NH: tape_meas_lam_C, tape_meas_lam_C superfamily, NH: TMP_2, TMP_2 superfamily, MD: COG5281
		Hypothetical protein Ec53638_0122, [gp41 domain protein]	278	100	3e-97	
		Putative phage protein	202	100	2e-68	
		Conserved hypothetical protein	197	100	5e-67	
		Putative Prophage Qin DNA packaging protein NU1 homolog	204	100	8e-71	NH: COG4220, Phage_Nu1, Phage_Nu1 superfamily
		Prophage minor tail protein Z (GPZ)	195	100	8e-65	NH: Minor_tail_Z, Minor_tail_Z superfamily
		Putative bacteriophage protein	142	100	5e-44	
		Phage antitermination Q type 1 family	135	100	5e-43	NH: Phage_antitermQ, Phage_antitermQ 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_I, gpI, Tail_P2_I superfamily
		Phage protein	282	100	5e-92	
		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	
		DNA-binding protein RdgB	115	100	1e-32	NH: Mor, Mor superfamily
		Hypothetical protein Ec53638_1176	177	100	1e-57	
		Phage tail sheath protein	340	100	1e-116	NH: FI, COG3497, Phage_sheath_1 superfamily, NH: GATase1_DJ-1, GAT_1 superfamily
		Baseplate	217	100	1e-72	NH: phage_P2_V, gpV, 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

G	1 Mbp to 2 Mbp	ISCps8, transposase	809	100	0.0	SH: Transposase_20, Transposase_20 superfamily, MD: COG3547
		Transposase	200	100	9e-68	
H	1 Mbp to 2 Mbp	Cytolethal distending toxin A	486	100	9e-166	NH: CDtoxinA, RICIN superfamily
		Cytolethal distending toxin B	545	100	0.0	NH: PRK15251, CdtB, EEP superfamily
		Cytolethal distending toxin subunit C	370	100	3e-122	NH: CDtoxinA, RICIN superfamily
		Hypothetical protein Ec53638_1905	158	100	2e-50	
		Putative phage protein	253	100	2e-81	
I	2 Mbp to 3 Mbp	ISCps8, transposase	809	100	0.0	SH: Transposase_20, Transposase_20 superfamily, MD: COG3547
J	2 Mbp to 3 Mbp	Transposase, IS1111 family	599	100	0.0	SH: Transposase_20, Transposase_20 superfamily, NH:
		ISAfe1, transposase	594	99	0.0	DEDD_Tnp_IS110, DEDD_Tnp_IS110 superfamily, MD: COG3547
K	2 Mbp to 3 Mbp	No significant results				
L	2 Mbp to 3 Mbp	Transposase, IS1111 family	580	100	0.0	SH: Transposase_20, Transposase_20 superfamily, MD: COG3547
		ISAfe1, transposase	576	99	0.0	SH: Transposase_20, Transposase_20 superfamily, MD: COG3547
M	3 Mbp to 4 Mbp	Transposase, IS1111 family	604	100	0.0	SH: Transposase_20, Transposase_20 superfamily, MD: COG3547
		ISAfe1, transposase	600	99	0.0	SH: Transposase_20, Transposase_20 superfamily, MD: COG3547
N	3 Mbp to 4 Mbp	Gp33 TerL	736	100	0.0	
		Putative phage-associated protein, HI1409 family	1,016	100	0.0	NH: DUF1073, COG3567, Phage_portal superfamily, NH: phge_rel_HI1409, phge_rel_HI1409 superfamily, NH: COG3566, DUF2213, DUF2213 superfamily
		Phage Mu protein F domain protein	530	100	1e-176	MD: COG2369
		Phage protein gp13	288	100	5e-89	
		Phage protein gp12	711	100	0.0	
		Conserved hypothetical protein	221	100	7e-66	NH: DUF3383, DUF3383 superfamily
		Transglycosylase SLT domain protein	888	100	0.0	SH: LT_GEWL, NH: SLT, Lysozyme_like superfamily, MD: MltE, mltD, PHA00368,
		Conserved hypothetical protein	451	100	9e-146	
		Phage P2 baseplate assembly protein gpV	456	100	8e-148	
		Putative bacteriophage protein	761	100	0.0	
		Putative tail fiber protein	586	100	0.0	
		Hypothetical bacteriophage protein	410	100	1e-131	NH: DUF2612, DUF2612 superfamily
		Phage integrase	1,079	99	0.0	SH: INT_REC_C, DNA_BRE_C superfamily
		Conserved hypothetical protein	1,628	99	0.0	NH: COG4688, COG4688 superfamily
		Putative tail component of prophage	247	100	2e-80	
		Invasion plasmid antigen, probably secreted by the Mxi-Spa machinery	690	64	0.0	NH: NEL, NEL superfamily, MD: PRK15387, COG4886, Golgin_A5
		Dead box helicase	65.5	100	4e-12	
		Hypothetical protein Ec53638_3354	557	100	0.0	NH: Phage_pRha, Phage_pRha superfamily
		Hypothetical protein Ec53638_3355	162	100	2e-48	
		Arc DNA binding domain protein	198	100	4e-61	
		Protein of unknown function	471	100	5e-168	SH: DUF45, NH: COG1451, superfamily: DUF45
		Phage anti-repressor protein AntB	534	100	0.0	NH: COG3561, AntA, AntA superfamily
		Type I restriction-modification system specificity subunit	348	100	6e-121	NH: Methylase_S, Methylase_S superfamily, MD: HsdS
		Conserved hypothetical protein	102	100	2e-29	
		Phage regulatory protein Cro	95.5	100	1e-26	NH: VapI, THE_XRE superfamily, MD: PRK09706
		gpH	163	100	1e-49	
		Bcv gene product	101	100	5e-27	
		Tail fiber assembly protein	77.4	100	5e-19	
		Type I restriction-modification enzyme, R subunit	210	100	7e-65	
		Type I restriction-modification system, M subunit	148	100	1e-44	MD: N6_Mtase, HsdM, hsdM, rumA
O	3 Mbp to 4 Mbp	ISCps8, transposase	809	100	0.0	SH: Transposase_20, Transposase_20 superfamily, MD: COG3547

Supplementary Table 1. Signature details based on *Escherichia coli* 53638 reference

P	3 Mbp to 4 Mbp	Invasion plasmid antigen	836	83	0.0	SH: NEL, NEL superfamily, MD: PRK15370
		Putative bacteriophage protein	528	100	0.0	
		Phage terminase large subunit	937	100	0.0	NH: phage_term_2, Terminase_3, Terminase_3 superfamily
		Putative bacteriophage protein	317	100	3e-100	NH: V, V superfamily
		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	
		Exodeoxyribonuclease 8	162	100	7e-48	MD: PRK09709
		Host-nuclease inhibitor protein Gam	143	100	9e-45	NH: Gam, Gam superfamily
		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	NH: phage_tail_N, phage_tail_N superfamily
		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	Transcriptional regulator, AlpA family	114	98	1e-32	NH: Phage_AlpA, Phage_AlpA superfamily, MD: PRK09692
		Hypothetical protein Ec53638_3914	108	100	1e-30	
		Prophage integrase	256	100	4e-86	SH: DUF4102, DUF4102 superfamily, NH: INT_P4, DNA_BRE_C superfamily, MD: PRK09692
		Conserved hypothetical protein	379	100	2e-131	
		Hypothetical protein Ec53638_3910	119	100	4e-37	
R	3 Mbp to 4 Mbp	Putative tail component of prophage	80.1	100	1e-17	
		Invasion plasmid antigen	821	74	0.0	SH: NEL, NEL superfamily, MD: PRK15370, COG4886
		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	NH: Phage_antitermQ, Phage_antitermQ superfamily
		Hypothetical protein Ec53638_4012	81.6	100	8e-22	
		ISCps8, transposase	809	100	0.0	SH: Transposase_20, Transposase_20 superfamily, MD: COG3547
		Hypothetical bacteriophage protein	393	100	6e-137	MD: PRK05643, DnaN
		Vonserved 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-40	NH: Phage_NinH, Phage_NinH superfamily
		Hypothetical protein Ec53638_4059	83.6	100	2e-23	
		Putative tail fiber protein	113	100	4e-32	NH: phage_tail_N, phage_tail_N superfamily
		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	Putative membrane protein	944	100	0.0	NH: MATE_Wzx_like, MATE_Wzx_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	SH: Glyco_tranf_GTA_type, NH: Glycos_transf_2, PLN02726, PRK10018, Glyco_tranf_GTA_type superfamily, MD: WcaA, PRK10073, Glyco_tranf_2_3, PgaC_IcaA, PTZ00260
		Glycosyltransferase, sugar-binding region containing DXD motif	539	99	1e-177	MD: Caps_synth, OCH1
			680	100	0.0	SH: EpsG, 7TMR_DISM_7TM superfamily
		Putative membrane protein	695	100	0.0	SH: Glyco_tranf_GTA_type, NH: Glycos_transf_2, PRK10018, PLN02726, Glyco_tranf_GTA_type superfamily, MD: WcaA, PRK10073, glyc2_xrt_Gpos1, PTZ00260, Glyco_tranf_2_3
		Cps23Fh	758	99	0.0	SH: GLF, GLF superfamily, SH: NAD_binding_8, NAD_binding_8 superfamily, MD: Glf, UDP-GALP_mutase, PRK07208
		UDP-galactopyranose mutase	620	99	0.0	NH: GT_2_like_b, Glycos_transf_2, Glyco_tranf_GTA_type superfamily, MD: COG1216, WcaA
		WfbU	810	100	0.0	SH: Transposase_20, Transposase_20 superfamily, MD: COG3547
		ISCps8, transposase				
T	4 Mbp to 5 Mbp	gpH	705	100	0.0	SH: DUF3751, DUF3751 superfamily
		Putative phage gene	811	100	0.0	
		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	
		Tail tube	311	100	2e-97	NH: DUF2597, DUF2597 superfamily
		Tail sheath	774	100	0.0	NH: DUF2586, DUF2586 superfamily
		Putative phage gene	498	100	2e-162	NH: P2_Phage_GpR, P2_Phage_GpR superfamily
		Conserved hypothetical protein	324	100	9e-102	
		Phage head completion protein (GPL)	321	100	9e-101	NH: Phage_GPL, Phage_GPL superfamily
		Putative repressor protein CI	373	100	1e-118	
		Putative DNA-binding protein Ner	170	100	3e-49	NH: Nlp, HTH_35, PRK10344, HTH_35 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, COG5484
		Phage portal protein, pbsx family	674	100	0.0	NH: Q, portal_PBSX, COG5518, Phage_portal, Phage_portal superfamily
		Phage transcriptional activator, Ogr/delta	186	100	1e-54	
		Hypothetical protein Ec53638_4365	176	100	6e-51	
		Hypothetical protein Ec53638_4366	162	100	3e-46	
		Putative phage replication protein	254	100	4e-70	
		Hypothetical protein Ec53638_4288	102	98	3e-26	
		Putative phage gene	1,133	100	0.0	
		Conserved hypothetical protein	217	100	9e-66	NH: DUF4406, DUF4406 superfamily
		DNA adenine methylase	641	100	0.0	NH: dam, PRK10904, Dam, MethyltransfD12, MethyltransfD12 superfamily
T	4 Mbp to 5 Mbp	Hypothetical protein Ec53638_4374	161	100	2e-46	
		Hypothetical protein Ec53638_4375	372	100	6e-121	
		Hypothetical protein Ec53638_4376	293	100	2e-92	
		Hypothetical protein Ec53638_4378	171	100	7e-50	

Supplementary Table 1. Signature details based on *Escherichia coli* 53638 reference

		Hypothetical protein Ec53638_4377	71.2	100	3e-15	
		Hypothetical protein Ec53638_4380	133	100	8e-37	
		Integrase for prophage CP-933T	711	100	0.0	NH: HP1_INT_C, Phage_integrase, DNA_BRE_C superfamily, MD: int, recomb_XerC, XerD, xerC
		Hypothetical bacteriophage protein	393	100	6e-137	DM: PRK05643, DnaN
		Conserved hypothetical protein	496	100	2e-176	NH: ORF6N, ORF6N superfamily
		Putative tail fiber protein	586	100	0.0	NH: phage_tail_N, phage_tail_N superfamily
		Putative tail component of prophage CP-933K	202	100	3e-63	
		Flagellin FliC	469	100	8e-164	SH: FliC, FliC superfamily, MD: PRK08026
		Putative DNA-packaging protein	223	100	5e-74	NH: Packaging_FI, Packaging_FI superfamily
		Phage Head-Tail Attachment	252	99	2e-85	NH: Phage_attach, Gifsy-2 superfamily
		Hypothetical protein Ec53638_4317	387	100	1e-137	
		Phage transposase	488	100	3e-170	SH: 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	
		Gifsy-1 prophage VmtH	229	100	1e-71	NH: tape_meas_lam_C, tape_meas_lam_C superfamily, MD: COG5281
		Putative phage regulatory protein, Rha family	204	100	4e-66	
		DNA transposition protein	203	100	2e-66	NH: Phage-MuB_C, Phage-MuB_C superfamily
		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 Monophosphatase 1 (IMPase 1) (IMP 1)(or 4)]	139	100	2e-44	
		Protein gp42	140	100	5e-41	
		Bacteriophage Mu Gam like protein	122	100	4e-37	NH: COG4396, Phage_Mu_Gam, Phage_Mu_Gam superfamily
U	5 Mbp to end of the genome	Putative bacteriophage protein	761	100	0.0	
		Putative tail fiber protein	575	100	0.0	
		Hypothetical bacteriophage protein	410	100	6e-135	NH: DUF2612, DUF2612 superfamily
		Phage P2 baseplate assembly protein gpV	456	100	5e-152	
		Invasion plasmid antigen	1,126	100	0.0	SH: TTSSLRR, TTSSLRR superfamily, SH: NEL, NEL superfamily, MD: PRK15370, COG4886
		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	

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	
		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: CDtoxinA, RICIN superfamily
		Cytolethal distending toxin, subunit B	520	100	1e-179	NH: PRK15251, CdtB, EEP superfamily
		Cytolethal distending toxin, subunit A	483	99	5e-166	NH: CDtoxinA, RICIN superfamily
		Exonuclease family protein	350	100	1e-116	NH: dexA, DnaQ-like-exo superfamily
		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 MD: int, XerD, recomb_XerD, xerD
		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 NH: PG_binding_3, PG_binding_3 superfamily MD: zliS
D	2 Mbp to 3 Mbp	Hypothetical protein ECOK1_2812	519	100	8e-169	NH: V, V superfamily
		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	
		Hypothetical protein ECOK1_4790, [beta family protein, Enterobacteria phage P4]	700	100	0.0	NH: Beta_protein, Beta_protein superfamily
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

Region	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	Cytolethal 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, Dcm, 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	
		Hypothetical protein ECRN5871_0834	600	100	0.0	
		Hypothetical protein ECRN5871_0832	79.3	97	6e-18	
		Phage Tail Collar domain protein	528	100	0.0	NH: DUF3751, DUF3751 superfamily
		Caudovirales tail fiber assembly family protein	400	100	3e-136	SH: Caudo_TAP, Caudo_TAP superfamily
		Hypothetical protein ECRN5871_0827	442	100	6e-152	
		Phage protein	206	100	4e-67	NH: DUF2586, DUF2586 superfamily
		Pentapeptide repeats family protein	1,107	100	0.0	NH: SopA, SopA superfamily, NH: SopA_C, SopA_C superfamily, MD: Pentapeptide_4, PRK15377, COG1357
		hdmD	475	100	3e-167	
		Hypothetical protein ECRN5871_0812, [phage tail protein, P2 Phage tail completion protein R (GpR)]	316	100	1e-110	NH: P2_Phage_GpR, P2_Phage_GpR superfamily
		Phage protein	273	100	1e-95	NH: DUF2597, DUF2597 superfamily
		Hypothetical protein ECRN5871_0702	69.7	100	1e-16	
		Phage protein	375	100	2e-131	NH: DUF2586, DUF2586 superfamily, MD: PAT1, PHA03247
		Phage small terminase subunit	163	100	1e-50	NH: M, Phage_smal_term superfamily
		Long tail fiber protein p37 domain protein	119	100	2e-35	
		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	
		Hypothetical protein ECRN5871_0798	296	100	7e-101	
		Hypothetical protein ECRN5871_0799	276	100	4e-93	
		Regulatory protein CII	38.1	100	3e-04	
		Hypothetical protein ECRN5871_0797	119	100	6e-34	
		Hypothetical protein ECRN5871_0829	702	100	0.0	
		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	
		Integrase	495	100	1e-177	NH: HP1_INT_C, DNA_BRE_C superfamily, MD: int, recomb_XerD, XerD,
		Scaffold domain protein	473	100	6e-171	NH: O, Phage_GPO, Phage_GPO superfamily, MD: IncA
		Repressor protein CI	338	100	4e-120	NH: Phage_CI_repr, Phage_CI_repr superfamily
		Hypothetical protein ECRN5871_0819	129	98	3e-40	

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-67	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	
		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	
		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, PMI_typeI, PLN02288, ABD 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	
		Hypothetical protein ECRN5871_4551, [chromosome segregation	206	100	6e-65	NH: ABC_SMC_barmotin, AAA_23, ABC_ATPase superfamily,

Supplementary Table 3. Signature details based on *Escherichia coli* RN587/1 reference

		protein SMC]				MD: SMC_N, Smc, PRK14272, recf
		mdaB domain protein	172	100	4e-57	NH: Flavodoxin_2, PRK00871, FMN_red superfamily
		Hypothetical protein ECRN5871_4526	137	100	1e-40	
		Hypothetical protein ECRN5871_4525	103	100	3e-30	
		Hypothetical protein ECRN5871_4553	58.9	61	1e-13	
G	1 Mbp to 2 Mbp	Peptidoglycan domain protein	374	100	2e-133	NH: Glyco_hydro_108, Glyco_hydro_108 superfamily, NH: PG_binding_3, PG_binding_3 superfamily, MD: zliS
		Tail fiber domain protein	131	100	9e-41	
H	2 Mbp to 3 Mbp	Exonuclease family protein	349	100	3e-116	NH: dexA, DnaQ_like_exo superfamily, MD: PRK09709
I	2 Mbp to 3 Mbp	Non-LEE-encoded effector EspJ	111	100	6e-32	
J	2 Mbp to 3 Mbp	Phage protein	268	100	3e-78	NH: Phage_ASH, Phage_ASH superfamily
		Zinc-binding domain of primase-helicase family protein	1,586	100	0.0	SH: Prim_Zn_Ribbon, NH: Prim_Zn_Ribbon, Prim_Zn_Ribbon superfamily, NH: Toprim_3, Toprim superfamily, MD: COG4643
		Hypothetical protein ECRN5871_3674	432	100	2e-140	
		Hypothetical protein ECRN5871_3673	271	100	2e-84	
		Hypothetical protein ECRN5871_3672	216	99	5e-65	
		Prophage CP4-57 regulatory protein family protein	151	100	2e-43	NH: Phage_AlpA, AlpA, Phage_AlpA superfamily
		Integrase	833	100	0.0	SH: INT_P4, DNA_BRE_C superfamily, NH: DUF4102, DUF4102 superfamily, MD: PRK09692, XerC
		Hypothetical protein ECRN5871_3699	343	100	1e-120	NH: DUF2544, DUF2544 superfamily
K	2 Mbp to 3 Mbp	Phage portal protein, PBSX family	284	100	1e-86	NH: Q, COG5518, portal_PBSX, Phage_portal superfamily
		Terminase, ATPase subunit	1,238	100	0.0	SH: Terminase_5, Terminase_5 superfamily, MD: P, Terminase_6, COG5484
		Presumed capsid-scaffolding protein	439	100	7e-144	NH: Phage_GPO, O, Phage_GPO superfamily
		Phage major capsid protein, P2 family	632	100	0.0	NH: N, Phage_cap_P2, major_capsid_P2, Phage_cap_P2 superfamily
		Terminase, endonuclease subunit	420	100	4e-138	NH: M, Phage_term_smal, Phage_term_smal superfamily
		Head completion/stabilization protein	290	100	2e-91	NH: Phage_GPL, Phage_GPL superfamily
		Hypothetical protein ECRN5871_3522	417	99	3e-137	
		Hypothetical protein ECRN5871_3519	215	100	3e-65	
		Hypothetical protein ECRN5871_3516, [PF05449 family protein]	209	100	1e-62	NH: DUF754, DUF754 superfamily
		Hypothetical protein ECRN5871_3515, [PF11860 family protein]	408	100	2e-132	SH: DUF3380, DUF3380 superfamily
		P2 phage tail completion protein R family protein	318	100	9e-101	NH: P2_Phage_GpR, P2_Phage_GpR superfamily
		Phage virion morphogenesis protein	300	100	3e-94	SH: Phage_tail_S, NH: tail_comp_S, Phage_tail_S superfamily
		hicB family protein	230	100	5e-70	NH: HicB, HicB superfamily, MD: HicB
		Baseplate assembly protein V	377	100	5e-121	NH: phage_P2_V, gpV, Phage_base_V, Phage_base_V superfamily
		Baseplate assembly protein W	236	100	3e-72	NH: W, COG3628, GPW_gp25, GPW_gp25 superfamily
		Baseplate J-like family protein	330	100	5e-102	NH: J, COG3948, Baseplate_J, Baseplate_J superfamily
		Phage tail protein I	397	100	2e-127	NH: gpI, tail_P2_I, tail_P2_I superfamily
		Phage tail fiber repeat family protein	1,291	100	0.0	NH: DUF3751, DUF3751 superfamily, SH: Phage_fiber_2, Phage_fiber_2 superfamily, MD: COG5301
		Tail fiber domain protein	219	68	7e-62	NH: DUF3751, DUF3751 superfamily
		Phage Tail Collar domain protein	244	58	1e-69	NH: DUF3751, DUF3751 superfamily
K	2 Mbp to 3 Mbp	Hypothetical protein ECRN5871_3504,[tail fiber assembly protein]	470	100	6e-153	NH: DUF4376, DUF4376 superfamily, NH: Caudo_TAP, Caudo_TAP superfamily

Supplementary Table 3. Signature details based on *Escherichia coli* RN587/1 reference

		Enoyl-CoA hydratase/carnithine racemase-like protein	321	100	4e-101	SH: Dam, NH: dam, PRK10904, MethyltransfD12,
		D12 class N6 adenine-specific DNA methyltransferase family protein	528	100	7e-174	MethyltransfD12 superfamily
		Major tail sheath protein	546	100	2e-178	NH: FI, COG3497, Phage_sheath_1 superfamily
		Phage major tail tube protein	328	100	5e-103	NH: FII, Phage_tube, COG3498, tail_tube, Phage_tube superfamily
		Phage tail tape measure protein, TP901 family, core region	1,687	100	0.0	NH: tape_meas_TP901, PhageMin_Tail, MCP_signal superfamily, MD: COG5283, NH: PHA01399, PHA01399 superfamily
		Phage P2 GpU family protein	287	100	7e-89	NH: COG3499, Phage_P2_GpU, Phage_P2_GpU superfamily
		Phage late control gene D family protein	727	100	0.0	NH: D, COG3500, Phage_GPD, Phage_GPD superfamily
		Caspase domain protein	671	100	0.0	
		Hypothetical protein ECRN5871_3490	1,759	100	0.0	
		Hypothetical protein ECRN5871_3489	259	100	0.0	
		Integrase	511	100	2e-169	SH: INT_P4, DNA_BRE_C superfamily, NH: DUF4102, DUF4102 superfamily, MD: PRK09692
		Phage portal protein, PBSX family	320	100	2e-106	NH: Q, portal_PBSX, COG5518, Phage_portal superfamily
		Hypothetical protein ECRN5871_3481, [RatA]	202	100	5e-63	MD: PRK15316
		Aldose 1-epimerase family protein	139	100	1e-42	NH: Aldose_epim_Ec_YphB, GalM, Aldose_epim superfamily
L	3 Mbp to 4 Mbp	Hypothetical protein ECRN5871_3053	1,435	100	0.0	MD: AAA_13, COG4694, SMC_prok_B
		Hypothetical protein ECRN5871_3052	380	100	1e-123	
		Hypothetical protein ECRN5871_3051, [nucleotidyl transferase, PF08843 family]	679	100	0.0	NH: COG2253, DUF1814 superfamily
		Hypothetical protein ECRN5871_3050, [ATPase, tellurite resistance protein TerB]	1,535	100	0.0	SH: TerB-N, TerB-N superfamily, SH: TerB-C, TerB-C superfamily, NH: terB, terB_like superfamily
		Biotin carboxylase	619	100	0.0	NH: DUF2791, AAA superfamily
		DEAD/DEAH box helicase family protein	988	100	0.0	SH: DEAD, NH: DEADc, DEXDc superfamily, NH: HELICc, HELICc superfamily, NH: HELICc, Helicase_C, Helicase_C superfamily, MD: Lhr, PRK13767, DEXH_lig_assoc, DEXDc, PTZ00110, PLN00206
		Hypothetical protein ECRN5871_3043, [ATP-dependent helicase]	407	100	2e-130	
		Hypothetical protein ECRN5871_3042, [PF09983 domain protein]	607	100	0.0	
		Hypothetical protein ECRN5871_3041	777	100	0.0	
		Hypothetical protein ECRN5871_3040	352	100	6e-111	
		Hypothetical protein ECRN5871_3039, [PF12128 family protein]	2,302	100	0.0	SH: Tropomyosin_1, Cep57_CLD, ApoLP_III_like superfamily, NH: SPEC, SPEC superfamily, NH: iSH2_PI3K_IA_R, iSH2_PI3K_IA_R superfamily, MD: DUF3584, Smc, PRK03918, SMC_prok_B, AIP3
		DNA-binding prophage protein	597	100	0.0	SH: Integrase_1, Integrase_1 superfamily, NH: Phage_int_SAM_2, Phage_int_SAM_2 superfamily
		DNA-binding protein Ner	145	100	5e-42	NH: Nlp, HTH_35, PRK10344, HTH_35 superfamily
		Integrase domain protein	450	100	6e-161	NH: INT_P4, Phage_integrase, DNA_BRE_C superfamily, MD: PRK09692, XerC, SH: DUF4102, DUF4102 superfamily
M	3 Mbp to 4 Mbp	Death on curing protein	255	100	4e-87	NH: Doc, DOC_P1, Fic superfamily
N	4 Mbp to 5 Mbp	espA-like secreted family protein	303	100	3e-104	NH: EspA, EspA superfamily
		Secretion system effector C (SseC) like family protein	147	100	1e-41	NH: COG5613, COG5613 superfamily

Supplementary Table 3. Signature details based on *Escherichia coli* RN587/1 reference

O	4 Mbp to 5 Mbp	No significant results				
P	4 Mbp to 5 Mbp	Anaerobic C4-dicarboxylate transporter dcuA	721	100	0.0	NH: Dcu, PRK09412, DcuB, DcuA_DcuB, DcuA_DcuB superfamily
		Fumarylacetoacetate (FAA) hydrolase family protein	610	100	0.0	SH: FAA_hydrolase, NH: PRK10691, HpaG-N-term, FAA_hydrolase superfamily, MD: MhpD, HpaG-C-term, PRK15203
		Dihydrodipicolinate synthetase family protein	600	100	0.0	SH: DHDPS-like, NH: DapA, PRK03170, DHDPS, dapA, PLN02417, Aldolase_Class_I superfamily
		Transcriptional regulator, LacI family	466	100	2e-152	SH: HTH_LacI, NH: HTH_LacI, HTH_LacI superfamily, NH: PBP1_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, ProP, synapt_SV2
		Porin B	837	100	0.0	NH: OprB, OprB superfamily
Q	4 Mbp to 5 Mbp	Type I restriction enzyme specificity protein	213	100	2e-69	NH: Methylase_S, Methylase_S superfamily, MD: HsdS, sufB
		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	SH: HSDR_N, HSDR_N superfamily, MD: COG0610, hsdR
R	4.5 Mbp to end of the genome	Hypothetical protein ECRN5871_0098	1,119	100	0.0	NH: DLP_2, Dynamin_N, Ras_like_GTPase superfamily
		Hypothetical protein ECRN5871_0099	1,435	100	0.0	NH: DLP_2, PRK09866, Ras_like_GTPase superfamily
		Hypothetical protein ECRN5871_0100	814	99	0.0	NH: DLP_2, Ras_like_GTPase superfamily
		sel1 repeat family protein	1,603	100	0.0	
		Hypothetical protein ECRN5871_0138, [secretion protein EspT]	406	100	4e-136	
		Hypothetical protein ECRN5871_0139	546	99	0.0	NH: IpaB_EvcA, IpaB_EvcA superfamily
		Hypothetical protein ECRN5871_0129	435	100	1e-149	NH: DUF1076, DUF1076 superfamily
		Hypothetical protein ECRN5871_0128, [secretion protein EspM]	103	100	7e-27	SH: IpaB_EvcA, IpaB_EvcA superfamily, NH: sifB, sif 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	
		Cysteine protease domain, YopT-type domain protein	1,030	100	0.0	NH: DUF3491, DUF3491 superfamily
		Hypothetical protein ECRN5871_0025,[N-acetyltransferase GCN5]	332	100	2e-114	
		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
		Hypothetical protein ECRN5871_0110	161	100	8e-51	
		Transposase for insertion sequence element IS1111A	292	86	3e-101	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	
		Phage portal protein, PBSX family	95.9	100	3e-24	NH: Q, Phage_portal superfamily
		yadA	316	100	3e-111	MD: PTZ00102
		Hypothetical protein ECRN5871_0131	114	100	4e-33	
R	4.5 Mbp to end of the genome	Terminase large subunit	226	100	5e-71	NH: COG5525, Terminase_GpA, Terminase_GpA superfamily
		Terminase small subunit	206	100	7e-68	SH: COG4220, 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	
		Baseplate assembly protein V	140	100	8e-44	NH: gpV, phage_P2_V, Phage_base_V, Phage_base_V superfamily
		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 non-specific hits (NH), specific hits (SH), multi domains (MD)
A	0 Mbp to 1 Mbp	Hypothetical protein ECSTECB2F1_0150	80.1	100	1e-21	
		Hypothetical protein ECSTECB2F1_0149, [transposase]	82.4	100	5e-22	
B	0 Mbp to 1 Mbp	Protein 40A	582	100	0.0	
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	
		Hypothetical protein ECSTECB2F1_0901, [tail fiber assembly protein, caudovirales tail fiber assembly protein]	251	100	1e-79	
		Outer membrane protein lom	128	100	1e-39	NH: Ail_Lom, PRK15240, COG3637, OMP_b-brl superfamily
		Phage integrase family protein	56.2	100	1e-11	
		Prophage tail fiber family protein	74.3	96	3e-19	NH: phage_tail_N, phage_tail_N superfamily
D	1 Mbp to 2 Mbp	Putative endopeptidase	63.9	100	2e-12	
		Hypothetical protein ECSTECB2F1_1098	78.2	100	3e-18	
		Hypothetical protein ECSTECB2F1_1099	86.7	98	3e-21	
		Hypothetical protein ECSTECB2F1_1078, [PF03235 family protein]	726	100	0.0	NH: DUF 262, DUF 262 superfamily
		Caudovirales tail fiber assembly family protein	125	100	5e-37	
		Prophage lambda integrase	275	100	2e-93	
		Prophage lambda integrase	139	100	6e-42	NH: INT_Lambda_C, DNA_BRE_C superfamily
		Prophage lambda integrase	110	100	2e-31	NH: 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	
		Hypothetical protein ECSTECB2F1_1516, [tail fiber assembly protein, caudovirales tail fiber assembly protein]	221	100	4e-70	
		Hypothetical protein ECSTECB2F1_1255	89.4	100	1e-23	
		Hypothetical protein ECSTECB2F1_1256	188	100	5e-62	
		Host specificity protein J	308	100	1e-99	SH: DUF3672, DUF3672 superfamily
		Antitermination protein Q	28.5	100	0.13	
		Hypothetical protein ECSTECB2F1_1251	207	100	6e-69	
		Hypothetical protein ECSTECB2F1_1263	197	100	2e-61	
		Hypothetical protein ECSTECB2F1_1515	60.1	100	9e-14	NH: ORF6N, ORF6N superfamily
		Helix-turn-helix family protein	130	100	6e-41	MD: PRK09706
		Outer membrane protein lom	127	100	6e-40	NH: : Ail_Lom, PRK15240, COG3637, OMP_b-brl superfamily
		Prophage tail fiber family protein	82.8	96	2e-22	
		Collagen triple helix repeat family protein	85.9	100	4e-22	NH: phage_tail_N, phage_tail_N superfamily
F	1 Mbp to 2 Mbp	Phage integrase family protein	212	100	1e-69	
		Hypothetical protein ECSTECB2F1_1296	91.3	100	7e-26	
		Hypothetical protein ECSTECB2F1_1297	56.2	100	6e-12	
		Phage integrase family protein	141	100	5e-43	NH: INT_Lambda_C, DNA_BRE_C superfamily
G	1 Mbp to 2 Mbp	Hypothetical protein ECSTECB2F1_1685	902	100	0.0	
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	
		DNA-damage-inducible protein I	59.7	100	6e-14	NH: PRK10597, DinI superfamily
		Hypothetical protein ECSTECB2F1_2003	69.7	100	4e-18	

Supplementary Table 4. Signature details based on *Escherichia coli* STEC_B2F1 reference

		outer membrane protein lom	136	100	2e-42	NH: Ail_Lom, PRK15240, COG3637, OMP_b-brl superfamily
I	2 Mbp to 3 Mbp	Flagellin	229	100	4e-74	
J	2 Mbp to 3 Mbp	Putative membrane protein	632	100	0.0	Glyco_transf_GTA type superfamily
		Glycosyl transferase family 2 family protein	572	100	0.0	NH: Glycos_transf_2, CESA_like, Glyco_transf_GTA type superfamily, MD: PRK10073
		Hypothetical protein ECSTECB2F1_2214	328	99	8e-103	
		Polysaccharide biosynthesis family protein	764	100	0.0	SH: MATE_like_10, NH: PRK15099, MATE_like superfamily, MD: RfbX
		Erythromycin biosynthesis sensory transduction protein eryC1	753	100	0.0	NH: AHBA_syn, DegT_DnrJ_EryC1, WecE, PRK11658, NHT_00031, AAT_I superfamily, MD: PRK15407, c
		wbtB	164	99	4e-47	
		Putative teichuronic acid biosynthesis glycosyltransferase tuaG	525	100	2e-170	SH: Glycos_transf_2, GT_2_like_d, PRK10018, PLN02726, Glyco_transf_GTA type superfamily, MD: Glyco_tranf_2_3, WcaA, PRK10073, PTZ00260
		Glucose-1-phosphate thymidylyltransferase	556	100	0.0	NH: G1P_TT_short, NTP_transferase, GalU, galU, PRK10122, Glyco_transf_GTA type superfamily , MD: rmlA , PRK15480, RfbA
K	2 Mbp to 3 Mbp	Hypothetical protein ECSTECB2F1_2378	99.8	98	1e-25	
		Hypothetical protein ECSTECB2F1_2379	94.7	100	6e-24	
		clp protease family protein	1,288	100	0.0	SH: S14_ClpP_1, NH: ClpP, CLP_protease, clpP, Clp_protease_like superfamily, MD: SDH_sah, SppA
		Bacteriophage P4 DNA primease	203	100	8e-65	
		Hypothetical protein ECSTECB2F1_2371	132	100	3e-40	
L	2 Mbp to 3 Mbp	Hemagglutination activity domain protein	1,036	99	0.0	SH: Fil_haemagg_2, fil_hemag_20aa, Fil_haemagg_2 superfamily, MD: FhaB, PRK15319, Hia,PRK12688
		Hypothetical protein ECSTECB2F1_3213	105	99	3e-26	
		Sulfatase family protein	769	99	0.0	NH: Sulfatase, Sulfatase superfamily, MD: AslA, PRK13759, chol_sulfatase
		Hypothetical protein ECSTECB2F1_3190, [arylsulfatase]	261	100	3e-81	NH: Sulfatase, Sulfatase superfamily, MD: AslA
		Outer membrane porin protein LC	611	100	0.0	NH: PRK10554, Porin_1, OmpC, gram_neg_porins, OM_channels superfamily
		RTX toxin acyltransferase family protein	361	100	1e-118	NH: Haemagg_act, SH: Haemagg_act, Haemagg_act superfamily
		Hemolysin secretion/activation protein ShlB/FhaC/HecB family pr	746	100	0.0	NH: HlyC, SH: HlyC, HlyC superfamily, MD: FhaC, ShlB
		Hypothetical protein ECSTECB2F1_3193	130	100	9e-37	NH: PRK09750, DUF1187, DUF1187 superfamily
		Transcriptional regulator, AraC family	377	100	9e-128	MD: HTH_ARAC, PRK09940, AraC
		Serine protease eatA	1,121	100	0.0	MD: Peptidase_S6, AidA
		Hypothetical protein ECSTECB2F1_3192, [membrane protein]	60.5	100	9e-13	
		Hypothetical protein ECSTECB2F1_3199	89.4	100	5e-24	
		Hypothetical protein ECSTECB2F1_3200	133	100	1e-40	
		Neurotensin receptor R8	133	100	2e-40	
		tonB-dependent vitamin B12 receptor	268	100	9e-88	NH: ligand_gated_channel, Plug, Plug superfamily, OM_channels superfamily, MD: BtuB, TonB-B12
		Acetyl-CoA acetyltransferase	44.7	100	6e-07	NH: Thiolase_C, thiolase, Thiolase_C superfamily, cond-enzymes superfamily, MD: PRK05790, PaaJ, AcCoA-C-

Supplementary Table 4. Signature details based on *Escherichia coli* STEC_B2F1 reference

		Hypothetical protein ECSTECB2F1_3178	76.6	97	9e-20	Actrans
		ompA-like transmembrane domain protein	135	100	2e-41	NH: COG3637, OMP_b-brl, OMP_b-brl superfamily
M	2 Mbp to 3 Mbp	Hypothetical protein ECSTECB2F1_3480	453	100	6e-161	
N	3 Mbp to 4 Mbp	Replication gene A protein	69.7	100	6e-13	
		Cytolethal distending toxin C	158	100	6e-46	NH: CDtoxinA, RICIN superfamily
		Endonuclease/Exonuclease/phosphatase family protein	545	100	0.0	NH: PRK15251, CdtB, EEP superfamily
		Cytolethal distending toxin A/C family protein	486	100	9e-166	NH: CDtoxinA, RICIN superfamily
		Caudovirales tail fiber assembly family protein	46.6	100	6e-07	
		Hypothetical protein ECSTECB2F1_4300	696	100	0.0	
		Phage tail fiber repeat family protein	257	100	5e-83	
O	4 Mbp to end of the genome	Hypothetical protein ECSTECB2F1_4680	564	100	0.0	
		Putative membrane protein	267	100	1e-85	
		Hypothetical protein ECSTECB2F1_4682	70.9	100	1e-15	
		Type I restriction modification DNA specificity domain protein	279	100	5e-96	
		Type I site-specific deoxyribonuclease, HsdR family	142	100	2e-41	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	
		Prophage tail fiber family protein	89.7	100	3e-22	NH: phage_tail_N, phage_tail_N superfamily
		Major tail protein V	337	100	6e-118	SH: BID_2, Big_2 superfamily, MD: COG5492
		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	
		N-6 DNA Methylase family protein	974	100	0.0	SH: Methyltransf_26, AdoMet_Mtase superfamily, HsdM_N, HsdM_N superfamily, MD: N6_Mtase, hsdM
		Type I site-specific deoxyribonuclease, HsdR family protein	1,409	100	0.0	SH: HsdM_N, NH: UPF0020, HsdM_N superfamily, MD: HsdM, N6_Mtase, COG0610, hsdR, DEXDc
		Hypothetical protein ECSTECC16502_0289, [ABC transporter ATP-binding protein]	1,107	100	0.0	
		Hypothetical protein ECSTECC16502_0290	106	98	4e-27	
		Hypothetical protein ECSTECC16502_0291	74.3	100	4e-16	
		Type I restriction-modification system specificity determinant	1,176	100	0.0	NH: Methylase_S, Methylase_S superfamily, MD: N6_Mtase, HsdM, HsdS, PRK09737
		Putative membrane protein	364	100	6e-115	
		Type III restriction enzyme, res subunit	2,261	100	0.0	SH: HSDR_N, MD: COG0610, hsdR, hsdR
		Hypothetical protein ECSTECC16502_0339	1,450	100	0.0	SH: COG1479, DUF1524, DUF1524 superfamily, NH: DUF262, DUF262 superfamily, COG3586, COG3586 superfamily
		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	MFS superfamily, MD: HpaX, MFS_1, PRK11551, NarK
		Hypothetical protein ECSTECC16502_0280, [HNH endonuclease]	516	99	0.0	SH: HNH_2, HNHc superfamily, MD: COG3440
		Filamentation induced by cAMP protein Fic	138	100	2e-42	
B	1 Mbp to 2 Mbp	Phage virion morphogenesis protein	70.5	100	1e-14	
		Hypothetical protein ECSTECC16502_1311	133	98	1e-37	
		AAA ATPase	795	100	0.0	NH: GP4d_helicase, RecA_like superfamily, MD: AAA_15, COG3950
		Putative membrane protein	852	100	0.0	
		Hypothetical protein ECSTECC16502_4950, [Rz1 lytic protein]	62.0	100	1e-12	NH: PRK14512, S14_ClpP_1, ClpP, clpP, ClpP_protease_like superfamily
		Hypothetical protein ECSTECC16502_4968	442	100	4e-159	
		Acetyltransferase family protein	394	100	4e-141	
		Hypothetical protein ECSTECC16502_4986	51.2	100	2e-10	
		Hypothetical protein ECSTECC16502_4996	466	100	2e-167	NH: DUF2829, DUF2829 superfamily
		isaA	43.9	95	6e-06	
		Hypothetical protein ECSTECC16502_1295, [acetyltransferase]	120	100	2e-35	
		Integrase	208	100	3e-68	MD: int
		Tail fiber	176	100	4e-55	
		DNA-invertase	92.8	100	2e-23	NH: SR_ResInv, Resolvase, Ser_Recombinase superfamily, MD: PinR
		Caudovirales tail fiber assembly family protein	338	100	6e-118	SH: Caudo_TAP, Caudo_TAP superfamily
		Tail fiber domain protein	201	100	4e-65	

Supplementary Table 5. Signature details based on *Escherichia coli* STEC_C165_02 reference

		Recombination enhancement	298	99	3e-104	
		Phage Tail Collar domain protein	134	95	2e-41	NH: Collar, Collar superfamily
		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	
		Resolvase, N terminal domain protein	138	100	7e-41	
		Hypothetical protein ECSTECC16502_1807	126	100	5e-40	
D	2 Mbp to 3 Mbp	Hypothetical protein ECSTECC16502_2561	194	100	3e-62	
		Phage holin, lambda family	68.9	100	8e-16	
		Peptidoglycan domain protein	374	100	3e-133	NH: Glyco_hydro_108, Glyco_hydro_108 superfamily, NH: PG_binding_3, PG_binding_3 superfamily, MD: zliS
		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	288	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, 52, PLN03128
G	4 Mbp to 5 Mbp	Cytolethal distending toxin A/C family protein	456	100	3e-154	NH: CDtoxinA, RICIN 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	
		Putative lipoprotein	185	100	2e-57	
		humD	122	100	3e-34	NH: PRK10276, peptidase_S24_S26 superfamily, MD: LexA
		Hot protein	87.0	98	1e-21	SH: DNA_pol3_theta, NH: PRK10969, DNA_pol3_theta superfamily
		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

		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. Aalphabetic 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]
52	DNA topisomerase 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_synth	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
CDtoxinA	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. Aalphabetic 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 FI [General function prediction only]
COG3498	Phage tail tube protein FII [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
COG4373	Mu-like prophage FluMu protein gp28 [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_C5_DNA_methylase	Cytosine-C5 specific DNA methylases
D	tail protein; Provisional
dam	DNA adenine methylase (dam)

Supplementary Table 6. Aalphabetic abbreviation and description of putative conserved domains

Dam	Site-specific DNA methylase [DNA replication, recombination, and repair]
dapA	Dihydrodipicolinate synthase; Dihydrodipicolinate synthase is a homotetrameric enzyme
DapA	Dihydrodipicolinate synthase/N-acetylneuraminate lyase
dcm	DNA-methyltransferase (dcm)
Dcm	Site-specific DNA methylase [DNA replication, recombination, and repair]
Dcu	Anaerobic c4-dicarboxylate membrane transporter family protein
DcuA_DcuB	Anaerobic c4-dicarboxylate membrane transporter
DcuB	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]
DEDD_Tnp_IS110	Transposase; Transposase proteins are necessary for efficient DNA transposition
DegT_DnrJ_EryC1	DegT/DnrJ/EryC1/StrS aminotransferase family
dexA	Exonuclease
DEXDc	DEAD-like helicases superfamily
DEXH_lig_assoc	DEXH box helicase, DNA ligase-associated
DHDPS	Dihydrodipicolinate synthetase family; This family has a TIM barrel structure
DHDPS-like	Dihydrodipicolinate synthase family; Dihydrodipicolinate synthase family
DLP_2	Dynamin-like protein including dynamins, mitofusins, and guanylate-binding proteins
DnaB	Replicative DNA helicase [DNA replication, recombination, and repair]
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_topoisoIV	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. Aalphabetic abbreviation and description of putative conserved domains

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	Haemagluttinin 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
GalU	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_tranf_2_3	Glycosyltransferase like family 2
Glyco_tranf_GTA_type	Glycosyltransferase family A (GT-A) includes diverse families of glycosyl transferases

Supplementary Table 6. Aalphabetic abbreviation and description of putative conserved domains

Golgin_A5	Golgin subfamily A member 5
GP4d_helicase	GP4d_helicase is a homohexameric 5'-3' helicases
gpI	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 hydrophillic 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 superfamily 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 CI
HTH_XRE	Helix-turn-helix XRE-family like proteins
IncA	IncA protein
int	Integrase; Provisional
Int	Integrase
Integrase_1	Integrase; 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, intergrase/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

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	TonB dependent/Ligand-Gated channels
LprI	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
Methyltransf_26	Methyltransferase domain; This family contains methyltransferase domains
MethyltransfD12	D12 class N6 adenine-specific DNA methyltransferase
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
mltD	Membrane-bound lytic murein transglycosylase D; Provisional
MltE	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
Nlp	Predicted transcriptional regulator [Transcription]
NTP_transferase	Nucleotidyl transferase
O	Capsid-scaffolding 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 FI; This family includes the lambda phage DNA-packaging protein FI

Supplementary Table 6. Aalphabetic abbreviation and description of putative conserved domains

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_LacI_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_integ_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_lysis	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. Aalphabetic abbreviation and description of putative conserved domains

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-terminal; 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	Dihydrodipicolinate synthase
PLN02726	Dolichyl-phosphate beta-D-mannosyltransferase
PLN03114	ADP-ribosylation factor GTPase-activating protein AGD10; Provisional
PLN03128	DNA topoisomerase 2; Provisional
Plug	TonB-dependent Receptor Plug Domain
PMI_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	Dihydrodipicolinate 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. Aalphabetic abbreviation and description of putative conserved domains

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 phosphoporin 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/decarboxylase; 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 SlrP; 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 thymidylyltransferase RfbA; Provisional
ProP	Permeases of the major facilitator superfamily
PS_pyruv_trans	Polysaccharide pyruvyl transferase
PTZ00102	Disulphide isomerase; Provisional
PTZ00110	Helicase; Provisional

Supplementary Table 6. Aalphabetic abbreviation and description of putative conserved domains

PTZ00260	Dolichyl-phosphate beta-glucosyltransferase; Provisional
PurR	Transcriptional regulators [Transcription]
Q	Portal vertex protein; Provisional
recf	recF protein
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
rumA	23S rRNA m(5)U1939 methyltransferase; Reviewed
rve	Integrase core domain
RVT_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 archaebacterial proteins
ShlB	Haemolysin secretion/activation protein ShlB/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 (ClpP 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. Aalphabetic abbreviation and description of putative conserved domains

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_GpA	Phage terminase large subunit (GpA)
thiolase	Thiolase are ubiquitous enzymes
Thiolase_C	Thiolase, C-terminal domain
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_mutase	UDP-galactopyranose mutase
UPF0020	Putative RNA methylase family UPF0020; This domain is probably a methylase
uvrD	DNA-dependent helicase II; Provisional
UvrD	Superfamily I DNA and RNA helicases [DNA replication, recombination, and repair]
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 wedge 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]
zliS	Lysozyme family protein [General function prediction only]