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 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]