Supplemental File 1: Contains additional figure S1 and tables S1-S6.

Figures:

S1. Cartoon explaining the SearchDOGS process for establishing orthology between genomic segments across species (relevant to paper section: "Overview of SearchDOGS search procedure" in Methods.

Tables:

S1: Genome annotations used in the SearchDOGS Bacterial analysis (relevant to paper section: "Test set of genomes used in this analysis" in Methods.
S2: Potential unannotated *S. boydii* genes identified (relevant to paper section: "Missing genes in the *Shigella boydii* genome annotation")
S3: Potential unannotated *S. boydii* requiring frameshift correction or stop codon readthrough.
S4: Orthologs of short (<60 codons) *E. coli* K-12 genes identified (see section: "Identification of short bacterial proteins using SearchDOGS").
S5: Examples of potential unannotated genes identified in model organism *E. coli* K-12 MG1655 (see section: "Potential missing genes in the *E. coli* K-12 MG1655 annotation?")
S6: Table of annotated genes likely to be pseudogenic (see section "Identification of pseudogenes")

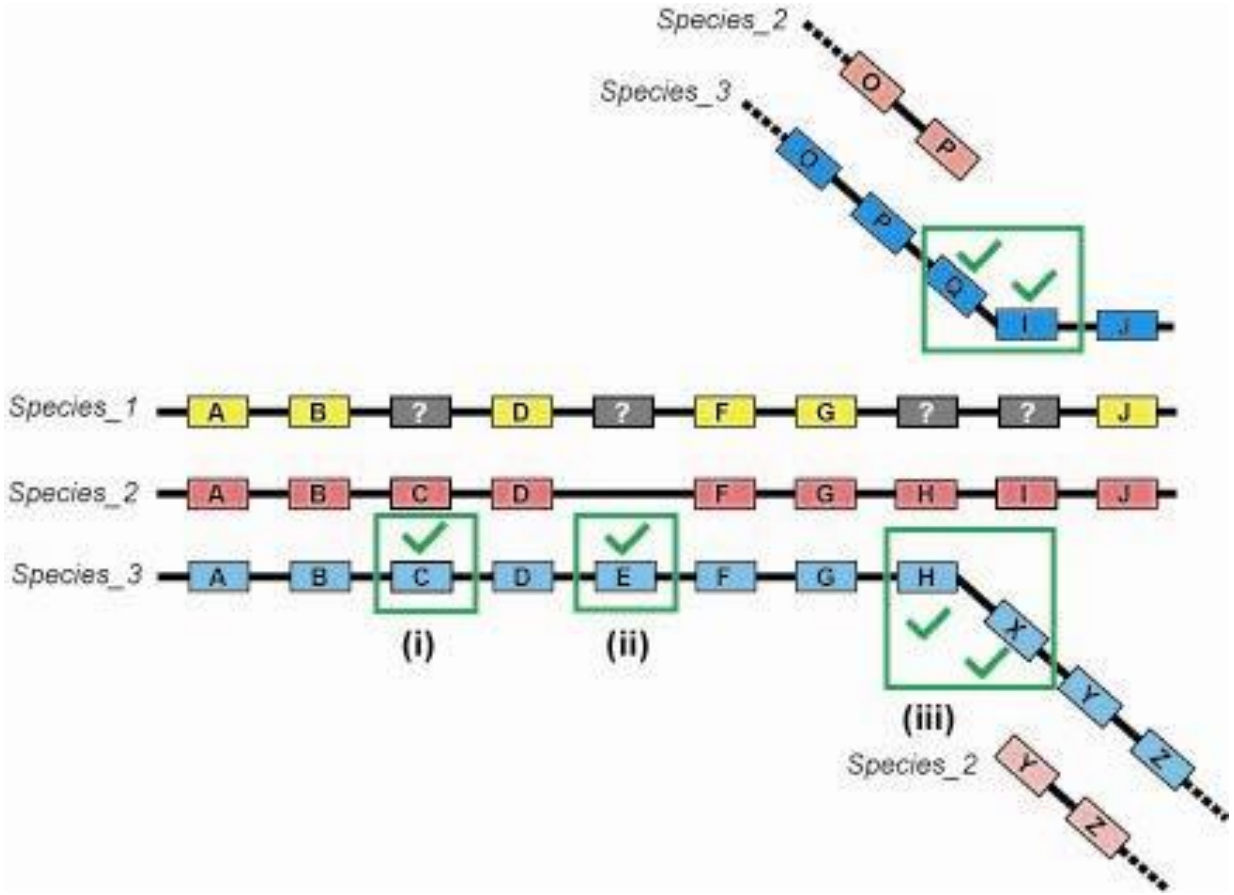

FIG S1: Cartoon illustrating the automated SearchDOGS method for establishing orthology between genomic segments. For ease of explanation, only three species are shown.

(i) Testing for a possible ortholog of gene *C* in Species_1. For the Species_1 genomic fragment *B-D*, ortholog pillars containing orthologs of genes B and D in Species _2 and Species_3 exist. The genes between B and D in each species are put into a database. The intergenic region between B and D in Species 1 is then searched against this database (here containing two copies of C) using BLASTX.

(ii) Testing for a possible ortholog of gene *E* in species 1. The process is the same as for (i), except that the database the Species_1 intergenic region is tested against now only contains on gene, Gene *E* in Species_3.

(iii). In this example, an interspecies rearrangement has occurred in Species_3 relative to Species_1 and Species_2, creating two new gene orders *G-H-X-Y-Z* and *O-P-Q-I-J* in Species_3. It is straightforward to identify genes in Species_2 to test against the G-J interval in Species 1. However, to include potentially orthologous genes from Species_3 in the database, we define two orthologous Species_3 genomic segments as follows. First, we consider the gene on the left end of the Species_1 segment, Species_1 *G*. We identify its Species_3 ortholog from the same pillar (Species_3 *G*), and walk rightwards from this gene until we reach the point where synteny is lost (Species_3 *Y*; we know that synteny is lost because it is in an ortholog pillar that maps to a different part of the Species_1/Species_2 genomes). We therefore put Species_3 genes encountered on this walk (Species_3 *H* and Species_3 *X*) into the database against which the Species_1 *G-J* intergenic interval will be searched. Second, we similarly consider the gene on the right end of the Species_1 segment, Species_1 *J*, find its Species_3 ortholog (Species_3 *J*) and walk leftwards in Species_3 until synteny is known to be lost (at Species_3 *P*). We add the Species_3 genes encountered on this walk (Species_3 *I* and *Q*) to the database. Thus the Species_1 *G-J* intergenic region will be used as a BLASTX query against a database containing *H*, *X*, *Q* and *I* from Species 3, as well as *H* and I from Species 2.

TABLE S1: Details on the genome annotation files used in this analysis, with some details on the original annotation methods and updates to the annotations obtained from the relevant papers.

Species	Most recent GenBank modify date (as of March 2014)	Version used in SearchDOGS analysis	Original annotation – paper+ method	Updated annotations – paper + method	Subjective estimate of confidence in annotation
Escherichia coli K-12 substr. MG1655	U00096.3 31-JAN-2014 (GI: 545778205)	U00096.2 26-FEB-2013 (GI: 48994873)	Blattner et al.(1) Quote from the paper: "*Postulation of genes in uncharacterized base sequences was surprisingly difficult. They were selected from among the numerous available*	Riley et al. (2) Quote from the paper: *"By comparing and re-sequencing regions of discrepancies between*	Very high – model organism

		open reading frames (ORFs) on the basis of codon usage statistics, sequence searches versus SWISS-PROT release 34, Link's database of NH_2-terminal peptide sequences from E. coli*, computer prediction of signal peptides, upstream matches to the Shine-Delgarno ribosome binding site, and other information including personal communications from colleagues "*	*MG1655 and W3110, highly accurate genomes have now been created for both strains"* Annotations for this model organism improved by large-scale community efforts. Workshops held to pool and reconciliate annotation data. Quote from the paper: "*Functional annotation was carried out by small groups of the Workshop participants incorporating extensive new experimental data from the literature, melding and reconciling collections of data from several sources (Table 1). When no experimental data beyond the sequence were available, these groups reached consensus after surveying predictions previously made by others with new predictions based on sequence similarity, domain content and other*

				predictive techniques and information"	
Escherichia coli O157:H7 str. Sakai	BA000007.2 17-JUL-2012 (GI: 47118301) (Plasmids: AB011548.2, AB011549.2)	BA000007.2 17-JUL-2012 (GI: 47118301)	Hayashi et al. (3) Strain-specific regions (relative to *E. coli* K-12 MG1655) were identified by comparing the whole chromosomal sequence using MUMmer. ORFs in the strain-specific regions and on the regions conserved were identified and annotated using Genome Gambler (v1.41), GLIMMER (2.01) and BLAST. ORFs larger than 150bp searched using automated means. Some manual identification of small genes.	Bergholz et al. (4) Global gene expression experiments for *E. coli* O157:H7 (no mention in paper of reannotation).	High
Escherichia coli S88	CU928161.2 19-JUL-2012 (GI: 218363708) (Plasmid: CU928146.1)	CU928161.2 19-JUL-2012 (GI: 218363708)	Touchon et al. (5) Gene prediction conducted using AMIGene software, predicted genes submitted to automatic functional annotation using MAGE. Final functional assignment was pased on transfer of E. coli K-12 MG1655 annotations between strong othologs. Manual validation of automatic annotations performed using MaGe. "Specific" regions (containing genes not orthologous to ones in E. coli K-12 MG1655) were manually annotated.		High
Shigella boydii Sb227	CP000036.1 (Plasmid: CP000037) 31-JAN-2014	CP000036.1 (Plasmid: CP000037) 19-MAY-	Yang et al (6) Annotations were performed as described in their previous paper:		Medium - high

		2012	Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K-12 and O157 (7) Gene prediction performed using GLIMMER 2.0 to identify ORFs with >30 consecutive codons. Overlapping and clustered ORFs manually examined. Predicted sequences were searched against the non-redundant (nr) protein databases using BLASTP. Mobile elements and predictive sequences were identified using pairwise comparisons, tRNA sequences using tRNAscan-SE. Whole genomic comparisons may have been performed using GenomeComp.	
Salmonella enterica subsp. enterica serovar Typhi str. Ty2	AE014613.1 31-JAN-2014 (GI: 29140506)	AE014613.1 21-JUL-2012 (GI: 29140506)	Deng et al (8) ORFs in strain Ty2 were identified using GeneMark.hmm and compared against strain CT18 ORFs using BLAST. Differences were checked in detail using Lasergene DNA analysis (DNASTAR). 450 “real” differences between the genomes remained after analysis.	Medium - high
Yersinia pestis antiqua	CP000308.1 (Plasmids: CP000309.1, CP000310.1, CP000311.1) 28-JAN-2014 (GI: 108777911)	CP000308.1 (Plasmids: CP000309.1, CP000310.1, CP000311.1) 14-JUL-2012 (GI: 108777911)	Chain et al (9). Genome was annotated by combining results from Critica, Generation and Glimmer. Translations were compared against nr database using BLASTP. Protein set also searched against other databases for function annotation. Sequence alignment and protein	Medium

			domain search tools (BLAST, CLUSTALW Pfam) were also used.	
Pseudomonas syringae pv. tomato str. DC3000	AE016853.1 (Plasmids: AE016854.1, AE016855.1) 31-JAN-2014 (GI: 28856110)	AE016853.1 (AE016854.1 AE016855.1) 20-JUL-2012 (GI: 28856110)	From GenBank file: "This represents the most recent version of the continually updated genome annotations for Pseudomonas syringae pv. tomato strain DC3000 by the Pseudomonas-Plant Interaction project. See http://www.pseudomonas-syringae.org for the latest updates and expanded annotations." Buell et al. (10) The complete genome sequence of the Arabidopsis and tomato pathogen *Pseudomonas syringae* pv. tomato DC3000. ORFs were identified using GLIMMER. Predicted proteins were searched against nr amino acid database and searched for domains using HMMR with the Pfam and TIGRfam databases. ORFs were manually curated and assigned to role catagories adapted from Riley (1993)	Medium
Vibrio cholerae O395	CP000626.1 (chromosome 1) CP000627.1 (chromosome	CP000626.1 (chromosome 1) CP000627.1 (chromosome	Direct submission, no publication attached.	Low

	2) 31-JAN-2014 (GI: 146313784, 146314918)	2) 30-JUN-2012 (GI: 146313784, 146314918)		
Xanthomonas campestris pv. campestris str. ATCC 33913	AE008922.1 31-JAN-2014 (GI: 2116637)	AE008922.1 17-JUL-2012 (GI: 2116637)	da Silva et al. (11) Putative protein-coding genes were identified using GeneMark and Glimmer. Curators assigned functions by comparing to sequences in public databases. RNA species identified using BLASTN and tRNAscan-SE. Metabolic pathways analysed using KEGG. Transporter proteins annotated based on BLAST comparison with a database of transporters.	Medium

TABLE S2 Coordinates of potential unannotated *S. boydii* genes identified. Annotated homologs and other species in which an unannotated candidate ORF exists are listed. Protein function is listed for the annotated ortholog(s). Species acronyms are as follows: ECK1: *Escherichia coli* K-12 *substr.* MG1655, ECO1: *Escherichia coli* O157:H7 *str. Sakai*, ECS8: *Escherichia coli* S88, SBOY: *Shigella boydii* Sb227, SETY: *Salmonella enterica* subsp. enterica serovar Typhi str. Ty2, YPAN: *Yersinia pestis* antiqua, PSYR: *Pseudomonas syringae* pv. tomato str. DC3000, VCHO: *Vibrio cholerae* O395, XCAM: *Xanthomonas campestris* pv. campestris str. ATCC 33913.

Neigh - bouring genes (protein id)	Coordinates (start to stop)	Length (codons)	Annotated homologs	Predicted orthologs	Non - consensus start / overlap?	Protein function (1, 2)
ABB64744.1	complement(7400..8734)	445	ECK1 ECO1		66bp	predicted transporter

ABB64745.1			ECS8 SETY PSYR VCHO		overlap	
ABB66209.1 ABB66210.1	complement (1591525..1592556)	344	ECK1 ECO1 ECS8 XCAM		6bp	predicted NADP-dependent, Zn-dependent oxidoreductase
ABB65003.1 ABB65004.1	299314..300223	304	ECK1 ECO1 ECS8 SETY YPAN VCHO		GTG start	putative regulator; Not classified; manno(fructo)kinase
ABB66623.1 ABB66624.1	2017589..2018389	267	ECK1 ECO1 ECS8 SETY			predicted DNA-binding transcriptional regulator for the rhm operon
ABB68237.1 ABB68238.1	complement (3788841..3789533)	231	ECK1 ECO1 ECS8 SETY Y PAN			predicted transcriptional regulator
ABB66162.1 ABB66163.1	complement (1529861..1530529)	223	ECK1 ECO1 ECS8		76bp overlap	oxidoreductase, Fe-S subunit
ABB68660.1 ABB68661.1	complement (4293624..4294244)	207	ECK1 ECO1 ECS8 SETY YPAN VCHO PSYR			FKBP-type peptidyl-prolyl cis-trans isomerase rotamase)
ABB65471.1 ABB65472.1	complement (806757..807272)	172	ECK1 ECO1 SETY YPAN	ECS8		lipoprotein
ABB68418.1 ABB68419.1	4000662..4001153	164	ECO1		26bp overlap	hypothetical protein
ABB68182.1 ABB68183.1	3723956..3724381	142	ECO1		GTG start	hypothetical protein
ABB66729.1 ABB66730.1	2129589..2129975	129	ECK1 ECS8 YPAN VCHO PSYR XCAM	ECO1		conserved protein, UPF0153 family
ABB65836.1 ABB65837.1	complement (1197541..1197882)	114	ECO1		GTG start 8bp overlap	putative minor tail protein
ABB65062.1 ABB65063.1	366566..366904	113	ECS8	ECK1 ECO1	97bp overlap	hypothetical protein

ABB65686.1 ABB65687.1	complement (1049076..1049414)	113	ECS8 VCHO PSYR	ECK1	GTG start	transposase ORF A, IS3 family
ABB68075.1 ABB68076.1	3608495..3608932	112	ECO1	ECK1		hypothetical protein
ABB65625.1 ABB65636.1	complement (997938..998264)	109	ECO1		4bp overlap	regulatory protein
ABB65776.1 ABB65777.1	complement (1142666..1142992)	109	SETY		TTG start	hypothetical protein
ABB66744.1 ABB66745.1	2146635..2146943	103	ECS8	ECK1 ECO1	GTG start 88bp overlap	hypothetical protein
ABB65836.1 ABB65837.1	complement (1197541..1197840)	100	ECO1 ECS8			putative minor tail protein
ABB68407.1 ABB68408.1	3985896..3986189	98	ECO1		72bp overlap	hypothetical protein
ABB65425.1 ABB65426.1	768960..769250	97	ECS8 SETY		62bp overlap	hypothetical protein
ABB67811.1 ABB67812.1	complement (3310628..3310915)	96	ECK1 ECO1 ECS8 SETY YPAN PSYR VCHO			protein required for 2-thiolation step of mnm(5)-s(2)U34-tRNA synthesis
ABB66416.1 ABB66417.1	complement (1795302..1795586)	95	ECK1 ECO1 ECS8	YPAN		hypothetical protein
ABB65839.1 ABB65840.1	complement (1203924..1204205)	94	ECO1			putative tail protein
ABB68644.1 ABB68645.1	4273749..4273895	93	XCAM		Start codon unclear	expressed protein
ABB67964.1 ABB67965.1	complement (3468842..3469117)	92	ECO1		4bp overlap	HicA[Haemophilus influenzae]-like protein
ABB68021.1	complement	92	ECS8 SETY	ECK1	GTG start	toxic polypeptide, small

ABB68022.1	(3545591..3545866)			ECO1	likely	
ABB68299.1	complement	92	ECS8		Unclear if	hypothetical protein
ABB68300.1	(3865726..3866001)				ATG or TTG start	
ABB66292.1	complement	89	ECK1 ECO1			predicted lipoprotein, DUF333 family
ABB66293.1	(1671558..1671824)		ECS8 SETY			
ABB65042.1	343570..343833	88	ECS8	ECK1	GTG start	hypothetical protein
ABB65043.1				ECO1	67bp overlap	
ABB65236.1	561576..561830	85	ECS8	ECK1	GTG start	hypothetical protein
ABB65237.1				ECO1		
ABB65512.1	complement	85	ECK1 ECS8		GTG start,	toxin of the YoeB-YefM toxin-antitoxin
ABB65513.1	(858760..859014)		PSYR		4bp overlap	system
ABB64792.1	64852..65103	84	ECK1 ECS8	ECO1	32bp	Inhibitor of glucose uptake
ABB64793.1					overlap	
ABB66612.1	1996943..1997194	84	ECO1			hypothetical protein
ABB66613.1						
ABB68021.1	complement	81	ECK1 ECS8	ECO1	GTG start	toxic polypeptide, small
ABB68022.1	(3545591..3545833)					
ABB65867.1	1234142..1234381	80	ECK1 ECO1	XCAM		predicted protein
ABB65868.1			ECS8 SETY YPAN			
ABB66108.1	complement	79	ECK1 ECO1		GTG start	predicted inner membrane protein
ABB66109.1	(1469551..1469787)		ECS8 PSYR SETY			
ABB66441.1	complement	73	ECO1 ECS8	ECK1		hypothetical protein
ABB66442.1	(1825330..1825548					
ABB66518.1	complement	72	ECO1			putative holin protein
ABB66519.1	(1904154..1904369)					
ABB67826.1	3324641..3324856	72	ECS8	ECK1		hypothetical protein
ABB67827.1				SBOY		

				PSYR ECO1		
ABB66391.1 ABB66392.1	Complement(1772778..1772990)	71	ECS8	ECK1 ECO1 SALE	TTG start	Hypothetical protein
ABB65857.1 ABB65858.1	complement (1226011..1226220)	70	ECO1	ECK1 SBOY SETY YPAN PSYR	13bp overlap	hypothetical protein
ABB67427.1 ABB67428.1	complement(2915628..2915834)	69	ECS8	ECK1 ECO1		hypothetical protein
ABB68601.1 ABB68602.1	4228826..4229032	69	ECS8	ECK1 ECO1	ATG start, TTG orthologs	hypothetical protein
ABB65639.1 ABB65640.1	1002031..1002234	68	ECO1			hypothetical protein
ABB66748.1 ABB66749.1	complement (2153586..2153789)	68	ECS8	ECK1 ECO1	11bp overlap	hypothetical protein
ABB67155.1 ABB67156.1	complement (2618200..2618403)	68	ECS8	ECK1 ECO1 SETY	11bp overlap	conserved hypothetical protein, putative Z3866 protein
ABB65250.1 ABB65251.1	complement (579612..579812)	67	ECS8	ECO1 ECO1	87bp overlap	potassium ion accessory transporter subunit
ABB66126.1 ABB66127.1	complement (1486137..1486337)	67	ECK1 ECO1 ECS8	SETY	TTG start	beta-lactam resistance membrane protein
ABB67817.1 ABB67818.1	complement(3314542..3314742)	67	ECK1 ECO1 ECS8 SETY YPAN VCHO PSYR			predicted protein
ABB65236.1	complement	66	ECK1	ECO1 ECS8	8bp overlap	ryhB-regulated fur leader peptide

ABB65237.1	(561524..561721)					
ABB64986.1	285596..285790	65	ECS8	ECK1		hypothetical protein
ABB64987.1				ECO1		
ABB67807.1	complement	65	ECK1 ECO1	YPAN	GTG start in	bacterioferritin-associated ferredoxin
ABB67808.1	(3305718..3305912)		ECS8 SETY		XCAM	
			VCHO PSYR			
			XCAM			
ABB68277.1	3842481..3842675	65	ECK1 ECO1			predicted lipoprotein
ABB68279.2			ECS8 SETY			
			YPAN PSYR			
			VCHO XCAM			
ABB66117.1	1066311..1066499	63	ECS8	ECK1		hypothetical protein
ABB66118.1				ECO1		
ABB65146.1	457628..457816	63	ECO1			hypothetical protein
ABB65147.1						
ABB65686.1	complement	63	PSYR		ATG start	ISPsy12, transposase OrfA
ABB65687.1	(1049076..1049264)				but TTG in	
					PSYR	
ABB65704.1	1066311..1066499	63	ECS8	ECK1		hypothetical protein
ABB65705.1				ECO1		
ABB66387.1	complement	63	ECS8 ECK1	ECO1		hypothetical protein
ABB66388.1	(1767468..1767656)			SETY		
ABB67638.1	complement	63	ECS8	ECK1	GTG start	hypothetical protein
ABB67639.2	(3140054..3140242)			ECO1		
				SETY		
				VCHO		
ABB66489.1	complement	62	ECK1	ECS8	GTG start	hypothetical protein
ABB66490.1	(1879455..1879640)					
ABB64835.1	complement	61	ECS8	ECK1	L start 19bp	hypothetical protein
ABB64836.1	(111513..111695)			ECO1	overlap	
				SETY		

ABB66379.1	1758463..1758645	61	ECK1 ECO1	SETY		hypothetical protein
ABB66380.1			ECS8			
ABB66519.1	complement	61	ECO1 ECS8			hypothetical protein
ABB66521.1	(1904717..1904899)					
ABB67983.1	3492770..3492949	60	ECS8	ECK1	GTG start	hypothetical protein
ABB67984.1				ECO1		
ABB68844.1	4489589..4489768	60	ECK1 ECS8	ECO1		predicted protein
ABB68845.1			SETY YPAN			
ABB66746.1	complement	59	ECS8	ECK1	27bp	hypothetical protein
ABB66747.1	(2151465..2151641)			ECO1	overlap	
				SETY		
ABB66489.1	1879741..1879914	58	ECK1	ECS8		hypothetical protein
ABB66490.1						
ABB66676.1	complement	58	ECK1 ECS8	ECO1	GTG start	hypothetical protein
ABB6677.1	(2081286..2081459)			SETY		
ABB66950.1	complement	57	ECS8	ECK1	TTG start	hypothetical protein
ABB66951.1	(2378919..2379089)			ECO1		
ABB65948.1	complement	55	ECS8	ECK1	TTG start	hypothetical protein
ABB65949.1	(1306818..1306982)			ECO1		
ABB66387.1	complement	55	ECK1 ECS8	ECO1	TTG start	hypothetical protein
ABB66388.1	(1767638..1767802)			SETY		
ABB66393.1	1774383..1774547	55	ECS8	ECK1	TTG start	hypothetical protein
ABB66394.1				ECO1		
				SETY		
ABB67133.1	complement	55	ECS8	ECK1	GTG start	hypothetical protein
ABB67134.1	(2597273..2597437)			ECO1		
ABB67678.1	complement	54	YPAN		GTG start	hypothetical protein
ABB67679.1	(3188677..3188838)					
ABB68021.1	complement	54	ECS8	ECK1	TTG start	hypothetical protein
ABB68022.1	(3545863..3546024)			ECO1		
ABB65794.1	1159658..1159816	53	ECS8	ECK1	38bp	hypothetical protein

ABB65795.1				ECO1 SETY	overlap	
ABB67929.1	3434516..3434671	52	ECK1		38bp	hypothetical protein
ABB67930.1					overlap	
ABB68476.1	4086890..4087045	52	ECO1	ECS8		hypothetical protein
ABB68477.1						
ABB65779.1	complement	51	ECS8 SETY	ECK1	TTG start	hypothetical protein
ABB65780.1	(1145707..1145859)			ECO1		
ABB65949.1	1307613..1307762	50	ECS8	ECK1	GTG start	hypothetical protein
ABB65950.1				ECO1 SETY		
ABB66271.1	1649980..1650129	50	ECK1 ECO1	ECS8		toxic polypeptide, small
ABB66272.1						
ABB67983.1	complement	50	ECK1 ECS8	ECO1		Damage inducible, function unknown
ABB67984.1	(3492665..3492814)					
ABB68187.1	3728689..3728835	49	ECS8	ECO1	TTG start	hypothetical protein
ABB68188.1						
ABB68725.1	complement	49	ECK1 ECO1			entericidin B membrane lipoprotein
ABB68726.1	(4357398..4357544)		ECS8 SETY			
ABB66604.1	1990246..1990389	48	ECS8	ECK1	TTG start	hypothetical protein
ABB66605.1				ECO1		
ABB67607.1	complement	48	ECK1	ECO1 ECS8	9bp overlap	expressed protein
ABB67608.1	(3105530..3105673					
ABB68791.1	4430829..4430972	48	ECS8	ECK1	44bp	hypothetical protein
ABB68792.1				ECO1	overlap	
ABB64942.1	236822..236962	47	ECK1 ECO1	ECS8	1bp overlap	rpmJ (L36) paralog
ABB64943.1			SETY YPAN VCHO XCAM			
ABB67193.1	2660831..2660971	47	ECK1			expressed protein
ABB67194.1						
ABB66963.1	complement	46	ECS8	ECK1	TTG start	hypothetical protein

ABB66964.1	(2396887..2397024)			ECO1	46bp overlap	
ABB67876.1 ABB67877.1	complement (3374611..3374742)	44	ECS8	ECK1 ECO1	GTG start	hypothetical protein
ABB65380.1 ABB65381.1	complement (722544..722672)	43	ECS8	ECK1 ECO1 SETY		hypothetical protein
ABB65366.1 ABB65367.1	complement (704716..704841)	42	SETY		ATG start (TTG in SETY)	hypothetical protein
ABB68725.1 ABB68726.1	complement (4357655..4357780	42	ECK1 ECS8 SETY	ECO1		entericidin A membrane
ABB65449.1 ABB65450.1	785649..785768	40	SETY		GTG start, 8bp overlap	hypothetical protein
ABB65925.1 ABB65926.1	complement (1285080..1285196)	40	ECK1	ECO1 ECS8 SETY		expressed protein, membrane-associated
ABB68733.1 ABB68734.1	4365618..4365734	39	SETY	ECK1 ECO1 ECS8	GTG start	hypothetical protein
ABB65280.1 ABB65281.1	618789..618902	38	ECK1 ECO1 SETY XCAM	ECS8 YPAN VCHO		conserved protein
ABB68330.1 ABB68331.1	complement (3907968..3908078)	37	ECK1	ECO1 ECS8 SETY		expressed protein
ABB65671.1 ABB65672.1	complement (1037014..1037121)	36	ECK1 ECS8 SETY YPAN	ECO1	20bp overlap	predicted protein
ABB66376.1 ABB66377.1	complement (1755424..1755528)	35	ECK1		3bp overlap	expressed protein
ABB64941.1 ABB64942.1	236094..236195	34	ECK1	ECS8		expressed protein
ABB66213.1 ABB66214.1	1594365..1594460	32	ECK1	ECO1 ECS8 SETY		hypothetical protein

ABB66091.1 ABB66091.1	1451447..1451536	30	ECK1	ECO1 ECS8 SETY		hypothetical protein
ABB68172.1 ABB68173.1	complement (3712419..3712508)	30	ECK1	ECO1 SETY		lexA-regulated toxic peptide
ABB65657.1 ABB65658.1	complement (1021634..1021717	28	ECK1	ECO1 ECS8		expressed protein
ABB66154.1 ABB66155.1	complement (1520342..1520425)	28	ECK1	ECO1 ECS8 YPAN XCAM		hypothetical protein
ABB67133.1 ABB67134.1	complement (2597196..2597276)	27	ECK1	ECO1 ECS8		toxic membrane protein
ABB68147.1 ABB68148.1	complement (3684407..3684487)	27	ECK1 ECS8 ECO1	PSYR	44bp overlap	tryptophanase leader peptide
ABB6965.1 ABB66966.1	2398850..2398921	24	ECK1	ECO1 ECS8 SETY	GTG start	expressed protein, membrane-associated
ABB65813.1 ABB65814.1	complement (1177889..1177954	22	ECK1	ECO1 ECS8		expressed protein
ABB65510.1 ABB65511.1	Complement (856601..856663)	21	ECK1	ECO1 ECS8 SETY		expressed protein
ABB67036.1 ABB67037.1	complement (2480851..2480910)	20	ECK1	ECO1 ECS8 SETY		hypothetical protein
ABB67424.1 ABB67425.1	2909787..2909846	20	ECK1	ECO1 ECS8 SETY		toxic membrane protein
ABB67424.1 ABB67425.1	2909411..2909470	20	ECK1	ECO1 ECS8 SETY		toxic membrane protein
ABB67585.1 ABB67586.1	3081134..3081193	20	ECK1	ECO1 ECS8		toxic membrane protein
ABB65566.1 ABB65567.1	complement (923722..923778)	19	ECK1	ECO1 ECS8 SETY		toxic membrane protein
ABB68711.1 ABB68713.1	complement (4341335..4341388)	18	ECK1	ECO1 ECS8 SETY	18bp overlap	expressed protein

ABB68241.1 ABB68242.1	3798714..3798764	16	ECK1	ECO1 ECS8 SETY YPAN	expressed protein
ABB66001.1 ABB66002.1	1354234..1354278	15	ECK1 ECS8 ECO1 SETY		phenylalanyl-tRNA synthetase operon leader peptide

TABLE S3 Coordinates of potential unannotated *S. boydii* genes in which a single frameshift correction or readthrough of a stop codon results in a full-length, conserved protein relative to its orthologs.

Neighbouring genes (protein id)	Coordinates (start to stop)	Length (codons)	Annotated homologs	Error corrected?	Protein function (2)
ABB65106.1 ABB65108.1	414237..416648	804	ECK1 ECO1 ECS8 SETY YPAN PSYR VCHO XCAM	STOP, 4bp overlap	predicted ABC transporter permease
ABB67455.1 ABB67456.1	2946115..2948466	784	ECK1 ECO1 ECS8	STOP	trehalose-6-phosphate phosphatase, biosynthetic
ABB68564.1 ABB68565.1	4176617..4178845	743	ECK1 ECO1 ECS8	STOP	conserved protein with nucleoside triphosphate hydrolase domain
ABB68315.1 ABB68316.1	complement(3882029..3884246)	739	ECK1 ECO1 ECS8 SETY YPAN PSYR VCHO	FRAMESHIFT	fadB fused 3-hydroxybutyryl-CoA epimerase/delta(3)-cis-delta(2)-trans-enoyl-CoA isomerase/enoyl-CoA hydratase/3-hydroxyacyl-CoA dehydrogenase
ABB65555.1 ABB65556.1	complement(904366..906528)	721	ECK1 ECO1 ECS8 SETY PSYR	STOP	protein-tyrosine kinase
ABB68540.1 ABB68541.1	complement(4153034..4155181)	716	ECK1 ECO1 ECS8 SETY YPAN VCHO XCAM	STOP	anaerobic respiration; formate dehydrogenase-H, selenopolypeptide subunit
ABB67582.1 ABB67583.1	complement(3075135..3077279)	715	ECK1 ECO1	STOP, 8bp overlap	methylmalonyl-CoA mutase

ABB67058.1 ABB67059.1	2505556..2507568	671	ECK1 ECO1 VCHO PSYR XCAM	STOP	putative 2-component regulator, interaction with sigma 54
ABB66172.1 ABB66173.1	1547559..1549244	562	ECK1 ECO1 ECS8 PSYR	STOP	fused predicted multidrug transporter subunits of ABC superfamily
ABB68026.1 ABB68027.1	complement(3551709..3553316)	536	ECK1 ECO1 ECS8 SETY YPAN PSYR	STOP	dipeptide transporter
ABB68642.1 ABB68643.1	complement(4269824..4271326	501	ECK1 ECO1 ECS8 YPAN	STOP	predicted sugar transporter subunit: ATP-binding component of ABC superfamily
ABB67465.1 ABB67466.1	complement(2959264..2960751)	496	ECK1 ECO1 ECS8 YPAN	STOP	altronate hydrolase
ABB67851.1 ABB67852.1	complement(3348078..3349382)	435	ECK1 ECS8	STOP	fused predicted acetyl-CoA:acetoacetyl-CoA transferase: alpha subunit/beta subunit
ABB65546.1 ABB65547.1	complement(895314..896537)	408	ECK1 ECO1 ECS8 SETY	STOP, 4bp overlap	predicted glycosyl transferase
ABB66102.1 ABB66104.1	complement(1463056..1464153)	366	ECK1 ECO1 ECS8 SETY YPAN PSYR VCHO XCAM	STOP	N-ethylmaleimide reductase, FMN-linked
ABB66190.1 ABB66191.1	1573041..1574081	347	ECK1 ECO1 ECS8 SETY	STOP	Energy metabolism, carbon: Anaerobic respiration; ethanol-active dehydrogenase/acetaldehyde-active reductase
ABB68101.1 ABB68102.1	complement(3637832..3638866)	345	ECK1 ECO1 ECS8 SETY YPAN PSYR	STOP, 14bp overlap	predicted glycosyl transferase
ABB65689.1 ABB65690.1	complement(1052698..1053702)	335	ECK1 ECO1 ECS8 SETY YPAN PSYR XCAM	STOP	membrane-anchored, periplasmic TMAO, DMSO reductase
ABB66572.1 ABB66573.1	complement(1958464..1959435)	324	ECK1 ECO1 ECS8 SETY YPAN PSYR VCHO XCAM	STOP 31bp overlap	flagellar biosynthesis
ABB65657.1	1020296..1021246	317	ECK1 ECO1 ECS8	STOP	tRNA-dihydrouridine synthase C

ABB65658.1			SETY YPAN PSYR VCHO		
ABB67031.1 ABB67032.1	2467183..2468133	317	ECK1 ECO1 ECS8 SETY	STOP	Central intermediary metabolism: Non-oxidative branch, pentose pathway;
ABB67649.1 ABB67650.1	3153349..3154281	311	ECK1 ECO1 ECS8 SETY YPAN PSYR XCAM	STOP, GTG start	aaeA p-hydroxybenzoic acid efflux system component
ABB65003.1 ABB65004.1	299314..300223	304	ECK1 ECO1 ECS8 SETY YPAN VCHO	FRAMESHIFT, GTG start	manno(fructo)kinase
ABB66058.1 ABB66059.1	complement(1415756..1416622)	289	ECK1 ECO1 ECS8 PSYR	STOP	quinate/shikimate 5-dehydrogenase
ABB67594.1 ABB67595.1	3092572..3093315	248	ECK1 PSYR	Stop readthrough	predicted NAD(P)-binding oxidoreductase with NAD(P)-binding Rossmann-fold domain
ABB66637.1 ABB66638.1	2030151..2030801	217	ECK1 ECO1 ECS8 PSYR	STOP	conserved protein, acid-induced
ABB68837.1 ABB68838.1	4481694..4482316	208	ECK1 ECO1 ECS8 SETY	Frameshift correction	DNA-binding transcriptional activator for silent bgl operon, requires the bglJ4 allele to function; LuxR family
ABB67678.1 ABB67679.1	complement(3188677..3188964)	96	ECK1 ECO1 ECS8 PSYR YPAN SETY VCHO XCAM	Stop readthrough	ribosome hibernation promoting factor HPF; stabilizes 70S dimers (100S)

TABLE S4 List of loci at which orthologs of short (<60 codon) *E. coli* K-12 MG1655 genes have been identified in the species studied. Species acronyms are as listed in Table S1.

Gene name	Length (codons)	Annotated in:	Unannotated ortholog found:	Protein function (2)
ydhU	59	ECK1 SBOY ECO1 ECS8	SETY	Putative membrane protein
yciY	57	ECO1 ECS8	SBOY SETY	Hypothetical protein
yciZ	57	ECK1 ECO1	SBOY SETY	Hypothetical protein
yjdO	57	ECK1	ECO1 ECS8	predicted protein
ymdF	57	ECK1 SETY	ECO1	conserved protein
ymgI	57	ECK1	ECS8 SBOY	hypothetical protein
gnsA	57	ECK1 ECS8 SETY	SBOY	Multicopy suppressor of secG(Cs) and fabA6(Ts); predicted regulator of phosphatidylethanolamine synthesis
YjdO	57	ECK1	ECO1 ECS8	Hypothetical membrane protein
ydaE	56	ECK1	ECO1	RAC prophage; conserved protein
ninE	56	ECK1	ECO1	Phage or prophage related
YHFg	55	ECK1	PSYR	Putative FIC-binding protein
yojO	54	ECK1	SBOY ECO1	hypothetical protein
ytjA	53	ECK1	ECO1 SBOY	Hypothetical protein
hokD	51	ECK1 ECO1 SBOY	ECS8	Qin prophage; small toxic polypeptide
yrhD	51	ECK1	ECS8 SBOY	Hypothetical protein
hokB	49	ECK1 ECO1	ECS8 SBOY	toxic polypeptide, small
ypdI	48	ECK1	ECO1 ECS8 SBOY	hypothetical protein
Yjjy	46	ECK1 ECO1 SBOY SETY	ECS8	predicted protein
ykgO	46	ECK1 ECO1 SETY YPAN VCHO XCAM	ECS8 SBOY	rpmJ (L36) paralog
ylcG	46	ECK1 ECO1	ECS8	expressed protein, DLP12 prophage
yqcG	46	ECK1	SBOY	expressed protein
Sra	45	ECK1 ECO1 SBOY SETY	ECS8	Stationary-phase-induced ribosome-associated protein
mntS	42	ECK1 ECS8	SBOY ECO1 SETY	Mn(2)-response protein, MntR-repressed
Blr	41	ECK1 ECO1 ECS8	SETY SBOY	beta-lactam resistance membrane protein
yqfG	41	ECK1	ECO1 ECS8 SBOY	Expressed protein
rpmJ	38	ECK1 ECO1 SBOY SETY	ECS8 YPAN	50S ribosomal subunit protein L36

		PSYR VCHO		
ybgT	37	ECK1 SETY XCAM	ECO1 ECS8 YPAN VCHO	Hypothetical protein
yshB	36	ECK1	ECO1 ECS8 SBOY SETY	expressed protein
ldrA	35	ECK1	ECO1	toxic polypeptide, small
ldrB	35	ECK1 ECS8 SBOY	ECO1	toxic polypeptide, small
ldrC	35	ECK1	ECO1 SBOY	toxic polypeptide, small
yniD	35	ECK1 ECS8 SBOY	ECO1	predicted protein
yohO	35	ECK1 ECS8	ECO1 SBOY	predicted protein
ymiB	34	ECK1	ECO1 ECS8 SBOY	expressed protein
yoaI	34	ECK1 SBOY ECO1 ECS8	SETY	predicted protein
ykgR	33	ECK1	ECS8 SBOY	expressed protein
ylcH	33	ECK1	ECO1	hypothetical protein, DLP12 prophage
yoaK	32	ECK1	ECO1 ECS8 SETY	Expressed protein, membrane-associated
yncL	31	ECK1	ECO1 ECS8 SBOY SETY	hypothetical protein
yneM	31	ECK1	ECO1 ECS8 SETY	expressed protein, membrane-associated
yccB	30	ECK1 SBOY	ECO1 ECS8 SETY	hypothetical protein
tisb	29	ECK1	ECO1 ECS8 SBOY SETY	lexA-regulated toxic peptide
ynhF	29	ECK1	ECO1 ECS8 SBOY SETY	hypothetical protein
kdpF	29	ECK1	SETY	Potassium ion accessory transporter subunit
azuC	28	ECK1	ECO1 SBOY	expressed protein
Uof	28	ECK1	ECO1 SBOY	ryhB-regulated fur leader peptide
ydgU	27	ECK1	ECO1 ECS8 SBOY YPAN	hypothetical protein
yohP	27	ECK1	ECO1 ECS8 SBOY	expressed protein
shoB	26	ECK1	ECO1 SBOY	toxic membrane protein
yqeL	26	ECK1	ECO1	expressed protein
yrbN	26	ECK1	ECO1 ECS8 SBOY SETY	expressed protein
yoaJ	24	ECK1	ECO1 ECS8 SBOY SETY	expressed protein, membrane-associated
ypdK	23	ECK1	ECO1 ECS8 SBOY SETY	expressed protein, membrane-associated

yobI	21	ECK1	ECO1 ECS8 SBOY	expressed protein
thrL	21	ECK1	YPES VCHO	thr operon leader peptide
yoeI	20	ECK1	ECO1 ECS8 SBOY SETY	expressed protein
ibsA	19	ECK1	ECO1 ECS8 SBOY SETY	toxic membrane protein
ibsC	19	ECK1	ECO1 ECS8 SBOY	toxic membrane protein
ibsD	19	ECK1	ECO1 ECS8 SBOY SETY	toxic membrane protein
ibsE	19	ECK1	ECO1 ECS8 SBOY SETY	toxic membrane protein
ypfM	19	ECK1	ECO1 ECS8 SBOY SETY	hypothetical protein
ibsB	18	ECK1	ECO1 ECS8 SBOY SETY	toxic membrane protein
yjeV	17	ECK1	ECO1 ECS8 SBOY SETY	expressed protein
mgtL	17	ECK1	ECO1 ECS8 SBOY SETY	Regulatory leader peptide for mgtA
hisL	16	ECK1	YPAN	his operon leader peptide
ilvX	16	ECK1	ECO1 ECS8 SBOY YPAN	expressed protein

TABLE S5 Coordinates of potential unannotated *E. coli* K-12 genes identified. Species containing annotated orthologs are listed using the species acronyms described for Table S4.

Neigh-bouring genes	Coordinates	Length	Annotated homologs	Nonconsensus start/overlap?
insH essD	complement(574981..576108)	376	ECS8	
yghD yghG	complement (3110076..3110942)	289	ECS8 YPAN VCHO XCAM	ATG start (TTG start in XCAM/VCHO)
cyaY yifL	3991873..3992358	162	ECO1 SBOY	210bp overlap
ybaZ ybaA	475499..475837	113	ECS8	97bp overlap
yahG yahI	338993..339313	107	ECO1	GTG start
ychE oppA	complement (1298626..1298940)	105	ECS8	TTG start
lysP yeiE	complement (2246538..2246846)	103	ECS8	GTG start 88bp overlap
yfgF yfgG	complement (2627177..2627467)	97	ECS8 SETY	156bp overlap
ydfU rem	complement (1642330..1642608)	93	ECO1	
glxR ybbW	536720..536998	93	ECO1	142bp overlap
narI tpr	1285932..1286207	93	ECO1	
ldrD yhjV	complement (3698006..3698278)	92	ECS8	GTG start 105bp overlap
bolA tig	453947..454210	88	ECS8	GTG start, 67 bp overlap
tolC ygiB	3177618..3177878	87	ECO1 SBOY	113bp overlap
sdhB sucA	757687..757947	87	SBOY ECO1	GTG start
yjiK yjiL	complement (4561691..4561948)	86	SBOY ECO1 ECS8	4bp overlap
uof fldA	709914..710168	85	ECS8	GTG start, 35bp overlap
yqgC metK	complement (3084421..3084672)	84	ECO1	GTG start
potA pepT	complement (1184796..1185047)	84	ECS8	GTG start, 22bp overlap
ldrD yhjV	complement (3698006..3698245)	81	ECS8	GTG start 105bp overlap
yodB mtfA	2040945..2041187	81	ECO1	GTG start
wrbA ymdF	1067135..1067371	79	ECS8	TTG start, 68bp overlap
narX narK	1276867..1277085	73	ECO1 ECS8	
ompA sulA	1019434..1019649	73	ECS8	TTG start, 17bp overlap
yhfA crp	3483920..3484135	72	ECS8	
ykgG ykgH	complement(323632..323844)	71	ECO1	46bp overlap
yciN topA	1328737..1328949	71	ECS8	TTG start
yoaE manX	complement (1899597..1899806)	70	ECO1	13bp overlap

ygiF ygiM	complement (3199004..3199210)	69	ECS8	
yjgB insC	complement (4494307..4494513)	69	ECS8	TTG start
bamD raiA	2734935..2735141	69	ECO1 SBOY	TTG start
yfiF trxC	complement (2716540..2716743)	68	ECS8	11bp overlap
gals yeiB	2239680..2239883	68	ECS8	11bp overlap
opgD ydcH	complement (1496456..1496659)	68	ECO1 ECS8 SBOY	GTG start, 80bp overlap
kdpF ybfA	complement(727958..728158)	67	ECS8	87bp overlap
ykgG ykgH	complement(323751..323948)	66	ECS8	29bp overlap
ampH sbmA	395649..395843	65	ECS8	
yhjH kdgK	3677164..3677358	65	ECS8	47bp overlap
zupT rib	complement (3181403..3181597)	65	ECS8	TTG start
yjhU yjhF	complement (4518447..4518638)	64	ECS8	GTG start
yedQ yodC	complement (2025962..2026150)	63	ECS8	80bp overlap
ygfT ygfU	3029256..3029444	63	ECS8	GTG start 56bp overlap
mreB csrD	3399217..3399405	63	ECS8	GTG start
betT yaA	complement (331090..331275)	62	ECS8	
dinQ arsR	3645833..3646012	60	ECS8	GTG start 24bp overlap
ydiP ydiQ	complement (1777428..1777604)	59	ECS8	
folE yeiG	2241828..2242004	59	ECS8	73bp overlap
xylH xylR	3732836..3733012	59	ECS8	11bp overlap
ydaN dbpA	1407332..1407505	58	ECS8	
mlaA yfdC	complement (2463055..2463225)	57	ECS8	TTG start
yciK sohB	complement (1327180..1327344)	55	ECS8	TTG start
yhaC garK	3267685..3267849	55	SBOY	GTG start
ydjA sppA	1846754..1846918	55	ECS8	TTG start, 58bp overlap
ldrD yhjV	complement (3698275..3698436)	54	ECS8	TTG start
coaA tufB	complement (4173236..4173391)	52	ECO1 SBOY SETY	
yciN topA	1328685..1328840	52	ECS8	TTG start, 8bp overlap
exbB metC	complement (3149999..3150154)	52	ECO1 SBOY SETY	GTG start 8bp overlap
yfgF yfgG	complement (2627142..2627294)	51	ECO1	
selD ydjA	complement (1845974..1846123)	50	ECS8	GTG start

yjiC iraD	4554907..4555050	48	ECS8	35bp overlap
opgC opgG	complement (1108209..1108352)	48	ECS8	TTG start
ypdI yfdY	complement (2492980..2493117)	46	ECS8	TTG start, 46bp overlap
yhgE pck	complement (3530537..3530668)	45	ECS8	GTG start
mutT yacG	complement(111564..111698)	45	SBOY	GTG start, 50bp overlap
acs nrfA	4285571..4285690	40	SETY	
aspA fxsA	complement(4366386..4366502)	39	SETY	GTG start
trxA rho	3964254..3964355	34	SBOY ECO1 ECS8	
ybdR rnk	complement(642553..642741)	30	ECS8	

TABLE S6 Set of genes annotated in the species studied that are likely to be pseudogenic based on length and sequence similarity to known pseudogenes in *E. coli* K-12 MG1655. Each gene in this list hit a syntenic region in *E. coli* K-12 containing a pseudogene. Pseudogene descriptions provided from the GenBank reference file (1, 2) or Ecogene (12)

E. coli K-12 pseudogene name	Annotated homologs likely to be pseudogenic	Description of *E. coli* K-12 pseudogene (2, 12)
yedS	*S. boydii* ABB65694.1	*Salmonella* OmpS1 homolog.
yhiL	*S. boydii* ABB67971.1	An intact version of YhiL is present in *E. coli* O157:H7 as Z4888. The yhiL gene can be transcribed in vitro with sigma28 (FliA) holoenzyme (Yu 2006)
yaiT	*S. boydii* ABB64984.1	First 27 aa predicted to be a signal peptide.
insZ	*E. coli 0157:H7* BAB38689.1 *S. boydii* ABB66429.1	Two frameshifts (at codons 62 and 111) and an internal deletion of about 150 codons have mutated this homolog of IS4 transposase InsG (442 aa)
ygeQ	*E. coli 0157:H7* BAB3_11764.1	Remnant of the type three secretion system (T3SS) pathogenicity island ETT2.
yghE	*P. syringiae* AA058138.1 *E. coli S88* CAR06003.1 *V. cholerae* ABQ18370.1	The yghFED operon appears to have suffered a deletion of the gspDEFGHIJK homologs (7403 bp) between the gspC-like (yghF) and the gspLM-like (yghED) genes. The stop codon of yghF was removed, fusing 12 C-terminal residues out-of-frame but overlapping part of the fused yghE gene. The N-terminal 74 residues of yghE were removed by the deletion event.
yejO	*S. boydii* ABB66695.1 *E. coli 0157:H7* BAB36504.1	IS5K inserted at codon 21 and made a 4 bp target site duplication TTAT. The first 29 aa are predicted to be a signal peptide
yjbI	*S. boydii* ABB68513.1	YjbI' and YjcF belong to COG1357. Apparent frameshifts at codons 62 and 86 were repaired to make a hypothetical reconstruction.
ydfJ	*E. coli 0157:H7* BAB35575.1	The first 28 codons of ydfJ were separated by the insertion of 20,460 bp of the Qin prophage; 28 aa (translated from 1650862 to 1650779 bp) have been added back to the YdfJ protein sequence presented. An intact version is present in *E. coli* 536 (UniProtKB: Q0THP5).
mdtQ	*E. coli S88* CAR03556.1	First 21 aa are predicted type II signal peptide. An apparent frameshift at codon 51 has been reconstructed.
yfdL	*E. coli 0157:H7* BAB33706.1 *E. coli S88* CAR04113.1	"pseudogene, CPS-53 (KpLE1) prophage; Phage or Prophage Related"
ylbH	*E. coli 0157:H7*	pseudogene, rhs-like

	BAB33986.1	
cybC	*S. boydii* ABB68636.1	pseudogene, truncated cytochrome b562
pinH	*S. boydii* ABB67380.1	pseudogene, predicted invertase fragment
ydeT	*E. coli 0157:H7* BAB35533.1	Outer membrane fimbrial subunit export usher protein FimD family.
yneO	*E. coli 0157:H7* BAB35540.1	pseudogene, AidA homolog
ycgH	*E. coli S88* CAR02558.2	Probable pseudogene; putative ATP-binding component of a transport system (.gbk)
yddK	*E. coli 0157:H7* BAB35498.1	A deletion has apparently removed the 5' end of yddK and the 3' 273 codons of yddL.
lfhA	*E. coli 0157:H7* BAB33679.1 *E. coli S88* CAR04588.1	Intact *E. coli* O42 allele: SP Q5DY37. The *E. coli* K-12 lfhA pseudogene is missing the first 127 codons.
yghO	*E. coli S88* CAR01619.1	pseudogene, DNA-binding transcriptional regulator homology
yegZ	*E. coli O157:H7* BAB36313.1	yegZ is adjacent to the ogrK copy of the P2 ogr gene, indicating the presence of a P2-like prophage remnant. Intact alleles are present in several *E. coli* strains and *Yersinia pestis* phage L-413C (UniProtKB: Q858U5).
ydfE	*E. coli S88* CAR02939.1	Qin prophage; pseudogene;Phage or Prophage Related
yibS	*E. coli O157:H7* BAB37895.1	Four stop codons (3, 11, 25, 27)
arpB	*E. coli S88* CAR03080.1 *E. coli O157:H7* BAB35850.1	A frameshift at codon 142 is translated as an X in the reconstructed protein sequence. An intact allele is present in O157:H7 EDL933 as Z2749, which has K142.
yjhZ	*E. coli S88* CAR06069.1	An inframe stop codon at position 44 was translated as an X for the reconstruction. An intact version of YjhZ is present in Escherichia sp. 3_2_53FAA as ESAG_039().
yhdW	*E. coli S88* CAR04880.1 *S. boydii* ABB67758.1	An apparent frameshift mutation at codon 23, as compared to other alleles and homologs of this gene, is translated as H23 in the reconstructed protein sequence since this position is a His residue in all the intact *E. coli* alleles.
ybfQ	*E. coli O157:H7* BAB34154.1	N-terminal domain fragment, matches first 79 residues of paralogs YhhI, YdcC, YbfD, pseudogene YbfL, and the more distant pseudogene paralog YncI
bcsQ	*E. coli O157:H7* BAB37837.1	Stop codon 6 is translated as an X in the reconstructed protein sequence; other *E. coli* strains have a Leu codon at this position
rhsE	*S. boydii* ABB65109.1	pseudogene, rhsE element core protein RhsE
ybfG	*S. boydii* ABB65243.1	An in-frame stop at codon 70 is replaced with an X in

		the reconstruction. An intact allele is found in *E. coli* 53638 as Ecol5_01004515 (GenBank gi:75511145).
yhiS	*S. boydii* ABB67989.1	IS5T inserted at codon 249 and made a 4 bp target site duplication TTAG. *E. coli* O157:H7 YhiS (Z4907) has no IS5 and has a frameshift near the C-terminus relative to K-12. the *S. flexneri* version (SF3539) has a similar C-terminus to the K-12 version.
ybeM	*S. boydii* ABB65191.1 *E. coli S88* CAR02007.2	1bp deletion at codon 66
yqfE	*E. coli S88* CAR04229.1	An inframe stop codon at position 19
ykiA	*S. boydii* ABB65002.1 *E. coli 0157:H7* BAB33865.1 *E. coli S88* CAR01736.1	An intact 759 aa version of YkiA is present in *E. coli* B185 (UniProt:D6I6K1)
ymdE	*E. coli O157:H7* BAB34720.1	Pseudogene

References:

1. **Blattner, F. R., G. Plunkett, 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao.** 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277:**1453-1462.
2. **Riley, M., T. Abe, M. B. Arnaud, M. K. Berlyn, F. R. Blattner, R. R. Chaudhuri, J. D. Glasner, T. Horiuchi, I. M. Keseler, T. Kosuge, H. Mori, N. T. Perna, G. Plunkett, 3rd, K. E. Rudd, M. H. Serres, G. H. Thomas, N. R. Thomson, D. Wishart, and B. L. Wanner.** 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot--2005. Nucleic Acids Res **34:**1-9.
3. **Hayashi, T., K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C. G. Han, E. Ohtsubo, K. Nakayama, T. Murata, M. Tanaka, T. Tobe, T. Iida, H. Takami, T. Honda, C. Sasakawa, N. Ogasawara, T. Yasunaga, S. Kuhara, T. Shiba, M. Hattori, and H. Shinagawa.** 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res **8:**11-22.
4. **Bergholz, T. M., L. M. Wick, W. Qi, J. T. Riordan, L. M. Ouellette, and T. S. Whittam.** 2007. Global transcriptional response of Escherichia coli O157:H7 to growth transitions in glucose minimal medium. BMC microbiology **7:**97.
5. **Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, A. Calteau, H. Chiapello, O. Clermont, S. Cruveiller, A. Danchin, M. Diard, C. Dossat, M. E. Karoui, E. Frapy, L. Garry, J. M. Ghigo, A. M. Gilles, J. Johnson, C. Le Bouguenec, M. Lescat, S. Mangenot, V. Martinez-Jehanne, I. Matic, X. Nassif, S. Oztas, M. A. Petit, C. Pichon, Z. Rouy, C. S. Ruf, D. Schneider, J. Tourret, B. Vacherie, D. Vallenet, C. Medigue, E. P. Rocha, and E. Denamur.** 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet **5:**e1000344.
6. **Yang, F., J. Yang, X. Zhang, L. Chen, Y. Jiang, Y. Yan, X. Tang, J. Wang, Z. Xiong, J. Dong, Y. Xue, Y. Zhu, X. Xu, L. Sun, S. Chen, H. Nie, J. Peng, J. Xu, Y. Wang, Z. Yuan, Y. Wen, Z. Yao, Y. Shen, B. Qiang, Y. Hou, J. Yu, and Q. Jin.** 2005. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. Nucleic Acids Res **33:**6445-6458.
7. **Jin, Q., Z. Yuan, J. Xu, Y. Wang, Y. Shen, W. Lu, J. Wang, H. Liu, J. Yang, F. Yang, X. Zhang, J. Zhang, G. Yang, H. Wu, D. Qu, J. Dong, L. Sun, Y. Xue, A. Zhao, Y. Gao, J. Zhu, B. Kan, K. Ding, S. Chen, H. Cheng, Z. Yao, B. He, R. Chen, D. Ma, B. Qiang, Y. Wen, Y. Hou, and J. Yu.** 2002. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. Nucleic Acids Res **30:**4432-4441.
8. **Deng, W., S. R. Liou, G. Plunkett, 3rd, G. F. Mayhew, D. J. Rose, V. Burland, V. Kodoyianni, D. C. Schwartz, and F. R. Blattner.** 2003. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. J Bacteriol **185:**2330-2337.

9. **Chain, P. S., P. Hu, S. A. Malfatti, L. Radnedge, F. Larimer, L. M. Vergez, P. Worsham, M. C. Chu, and G. L. Andersen.** 2006. Complete genome sequence of *Yersinia pestis* strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen. J Bacteriol **188:**4453-4463.
10. **Buell, C. R., V. Joardar, M. Lindeberg, J. Selengut, I. T. Paulsen, M. L. Gwinn, R. J. Dodson, R. T. Deboy, A. S. Durkin, J. F. Kolonay, R. Madupu, S. Daugherty, L. Brinkac, M. J. Beanan, D. H. Haft, W. C. Nelson, T. Davidsen, N. Zafar, L. Zhou, J. Liu, Q. Yuan, H. Khouri, N. Fedorova, B. Tran, D. Russell, K. Berry, T. Utterback, S. E. Van Aken, T. V. Feldblyum, M. D'Ascenzo, W. L. Deng, A. R. Ramos, J. R. Alfano, S. Cartinhour, A. K. Chatterjee, T. P. Delaney, S. G. Lazarowitz, G. B. Martin, D. J. Schneider, X. Tang, C. L. Bender, O. White, C. M. Fraser, and A. Collmer.** 2003. The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000. Proc Natl Acad Sci U S A **100:**10181-10186.
11. **da Silva, A. C., J. A. Ferro, F. C. Reinach, C. S. Farah, L. R. Furlan, R. B. Quaggio, C. B. Monteiro-Vitorello, M. A. Van Sluys, N. F. Almeida, L. M. Alves, A. M. do Amaral, M. C. Bertolini, L. E. Camargo, G. Camarotte, F. Cannavan, J. Cardozo, F. Chambergo, L. P. Ciapina, R. M. Cicarelli, L. L. Coutinho, J. R. Cursino-Santos, H. El-Dorry, J. B. Faria, A. J. Ferreira, R. C. Ferreira, M. I. Ferro, E. F. Formighieri, M. C. Franco, C. C. Greggio, A. Gruber, A. M. Katsuyama, L. T. Kishi, R. P. Leite, E. G. Lemos, M. V. Lemos, E. C. Locali, M. A. Machado, A. M. Madeira, N. M. Martinez-Rossi, E. C. Martins, J. Meidanis, C. F. Menck, C. Y. Miyaki, D. H. Moon, L. M. Moreira, M. T. Novo, V. K. Okura, M. C. Oliveira, V. R. Oliveira, H. A. Pereira, A. Rossi, J. A. Sena, C. Silva, R. F. de Souza, L. A. Spinola, M. A. Takita, R. E. Tamura, E. C. Teixeira, R. I. Tezza, M. Trindade dos Santos, D. Truffi, S. M. Tsai, F. F. White, J. C. Setubal, and J. P. Kitajima.** 2002. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. Nature **417:**459-463.
12. **Rudd, K. E.** 2000. EcoGene: a genome sequence database for *Escherichia coli* K-12. Nucleic Acids Res **28:**60-64.