TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Mikrobielle Ökologie

Overlapping genes in *E. coli* EDL933 (EHEC) - Phylostratigraphy of alternative reading frames and functional analysis of the candidate gene *asa*

Sonja Vanderhaeghen

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. R. F. Vogel

Prüfer der Dissertation:

1. Prof. Dr. S. Scherer
2. Prof. Dr. H. Schäfer

Die Dissertation wurde am 26.11.2018 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 28.02.2019 angenommen.

Publications

Vanderhaeghen, S., B. Zehentner, S. Scherer, K. Neuhaus and Z. Ardern (2018) **The novel EHEC gene *asa* overlaps the TEGT transporter gene in antisense and is regulated by NaCl and growth phase**, *Scientific Reports - accepted,* DOI 10.1038/s41598-018-35756-y

Hücker, S.M., S. Vanderhaeghen, I. Abellan-Schneyder, S. Scherer and K. Neuhaus (2018) **The Novel Anaerobiosis-Responsive Overlapping Gene *ano* Is Overlapping Antisense to the Annotated Gene ECs2385 of *Escherichia coli* O157:H7 Sakai**, *Frontiers in Microbiology 9:931*.

Hücker, S.M., S. Vanderhaeghen I. Abellan-Schneyder, R. Wecko, S. Simon, S. Scherer and K. Neuhaus (2018) **A novel short L-arginine responsive protein-coding gene (*laoB*) antiparallel overlapping to a CadC-like transcriptional regulator in *Escherichia coli* O157:H7 Sakai originated by overprinting**, *BMC Evolutionary Genomics 18:21.*

Congresses

Sonja Vanderhaeghen, B. Zehentner, S. Hücker, R. Landstorfer, S. Scherer and K. Neuhaus (2018) **A novel antisense overlapping gene completely embedded in a TEGT transporter gene of *Escherichia coli* O157:H7 str. EDL933 is regulated growth phase and condition dependent,** *VAAM Tagung, Wolfsburg*

Sonja Vanderhaeghen, B. Zehentner, S. Hücker, C. Nelson, R. Landstorfer, S. Scherer and K. Neuhaus (2017) **Characterization of an antisense, overlapping gene provides insight in the evolutionary constraints of completely embedded genes**, *EMBL Concepts in Microbiology, Heidelberg.*

Summary

Bacterial genomes are string of DNA densely packed with genes. In the last years, more and more novel unknown small genes were discovered in intergenic regions. However, genes in alternative open reading frames overlapping to annotated genes are usually excluded by genome annotation programs. Overlapping genes have been described in viruses and may be associated with *de novo* gene emergence by a hypothetical process termed 'overprinting'. This dissertation deals with three major questions: (i) Do bacterial overlapping reading frames encode proteins? If this is the case, (ii) when did such overlapping genes emerge during evolution? (iii) How do overlapping genes emerge?

The genome of *Escherichia coli* O157:H7 strain EDL933 harbors thousands of 'alternative open reading frames' overlapping to annotated genes. Protein homology search (blastp), comparative genomics and the protein feature prediction tool PredictProtein were used to collect information about protein sequences encoded by 2,180 ORFs overlapping annotated mother genes in antisense in all three possible reading frames. The gene age was estimated by phylostratigraphy, i.e., a taxonomic approach searching for the furthermost related species in which a homologue to a candidate gene can be found. Blastp was used for all possible overlapping genes of *E. coli* (alternative open reading frames), and tblastn for those with experimental evidence for functionality. The latter technique also reveals the potential transition from non-intact to intact genes and identifies putative mechanisms for sequence evolution.

Using these bioinformatics approaches, promising putative functional overlapping gene candidates were identified. Most of them are embedded in the mother gene, taxonomically restricted and structurally similar to small proteins discovered in previous studies. Gene emergence may occur either by gradual or by discontinuous evolution, which is strongly influenced by the reading frame in which such putative genes are located relative to the mother gene, and by the overlap type. Apparently, overlapping genes are frequently lost during evolution, although the mother gene is kept intact. The phylostratigraphy of bacterial overlapping genes appears to be comparable to eukaryotic putative novel genes and may help flesh out speculation on gene emergence mechanisms at the sequence level. However, inferred gene ages have to be interpreted carefully and need experimental validation.

As a showcase, the novel overlapping gene *asa,* discovered by RNAseq and RIBOseq, was functionally characterized in this study. Its expression is growth phase and NaCl dependent, which was detected by RT-qPCR as well as phenotypic characterizations. The promoter is highly active in pyridoxine hydrochloride and a phenotype was identified in the presence of L-arginine. Western Blot of the tagged overexpressed gene product confirmed the presence of the protein

Asa. Bioinformatics reveals putative functions as a membrane associated protein or as an enzyme important for terpenoid biosynthesis. The comparison to *asa* homologues with respect to their expression activity indicates that *asa* may have evolved from an antisense, putatively non-coding RNA to a protein coding gene.

Zusammenfassung

Bakterielle Gene liegen im Genom hintereinander auf dem DNA Strang. Während kleine, unbekannte Gene in intergenischen Regionen Berücksichtigung finden, werden Gene in alternativen offenen Leserahmen, die zu annotierten Genen überlappen, von Annotationsprogrammen ausgeschlossen. Dennoch sind überlappende Gene in Viren beschrieben und können dort mit der *de novo* Genentstehung, durch einen hypothetischen Prozess namens ‚overprinting' zusammenhängen. Daher beschäftigt sich diese Dissertation hauptsächlich mit den folgenden Fragestellungen: (1) Kodieren bakterielle überlappende Gene für Proteine? Wenn das der Fall ist, (2) wann entstanden solche überlappenden Gene in der Evolution? (3) Welche Mechanismen führen zur Entstehung von überlappenden Genen?

Das Genom von *E. coli* O157:H7 Stamm EDL933 hat tausende von offenen Leserahmen, die in alternativen Leserahmen zu annotierten Genen überlappen. Eine Suche in Datenbanken nach homologen Proteinen (blastp), vergleichende Genomik und das Programm ‚PredictProtein', das Proteineigenschaften vorhersagt, wurden genutzt, um Informationen über mögliche funktionelle überlappende Gene und die Proteine, die sie kodieren, herauszufinden. Das Alter der Gene wurde mittels Phylostratigraphie bestimmt, einer taxonomischen Identifikation des am weitesten entfernten Verwandten mit einem Homologen zu einem Kandidaten-Gen. Blastp wurde für alle überlappenden Gene von *E. coli* verwendet und tblastn für solche, die einen experimentellen Nachweis für eine Funktionalität haben. Letztere Methode wurde auch dafür genutzt den Übergang von intakter zu nicht-intakter Sequenz im Sinne einer Funktionalität darzustellen. Auf diese Weise konnten auch mögliche Mechanismen für die Sequenz-Evolution identifiziert werden.

Mittels Bioinformatik konnten vielversprechende überlappende Gen-Kandidaten gefunden werden. Die meisten davon sind vollständig in das Muttergen eingebettet, kommen nur in wenigen Taxa vor und sind strukturell ähnlich zu kleinen Proteinen, die in früheren Studien entdeckt wurden. Die Genentstehung kann durch graduelle oder diskontinuierliche Evolution stattfinden, was stark vom relativen Leserahmen des überlappenden Gens zu seinem Muttergen, sowie vom Überlappungs-Typ abhängt. Die Phylostratigraphie von bakteriellen überlappenden Genen ist vergleichbar zu der von neuen Genen in Eukaryoten und erlaubt es über die Mechanismen der Genentstehung auf Sequenzebene zu spekulieren. Allerdings müssen die ermittelten Genalter vorsichtig interpretiert werden und brauchen experimentelle Bestätigungen für die Funktionalität des Gens.

Beispielhaft für weitere wurde das Gen *asa* in dieser Studie funktionell charakterisiert. Es wurde durch RNAseq und RIBOseq entdeckt. Seine Expression ist abhängig von der Wachstumsphase und der Menge an NaCl, was durch RT-qPCR und durch phänotypische Charakterisierung detektiert wurde. Der Promoter ist in Pyridoxin Hydrochlorid sehr aktiv und ein weiterer Phänotyp wurde in L-arginin identifiziert. Ein Western Blot des überexprimierten Gens bestätigte das Vorhandensein des Proteins Asa. Potentielle Funktionen wurden durch Bioinformatische Analysen nahegelegt. Asa könnte ein Membran-assoziiertes Protein sein oder ein Enzym, das in der Terpenoid Biosynthese eine wichtige Rolle spielt. Der Vergleich der Expression von *asa* Homologen zeigt auf, dass *asa* aus in antisense liegender und möglicherweise nicht kodierender RNA zu einem Protein kodierenden Gen evolvierte.

Abbreviation Index

+1 site	transcriptional start site, 5’ end of a messenger ribonucleic acid
aa	amino acids
aORF	open reading frame of an annotated gene
*asa*CF, *asa*SM, *asa*SE, *asa*HA	homologues of *asa* in *Citrobacter freundii, Serratia marcescens, Salmonella enterica, Hafnia alvei*
BP	biological process
CD database	conserved domain database
cDNA	single stranded copy deoxyribonucleic acid
cq value	value at the quantification or threshold cycle of RT-qPCR
DNAP	deoxyribonucleic acid polymerase
dN and dS	rates of non-synonymous and synonymous mutations
GFP	green fluorescent protein
GST	glutathione-S-transferase
HUS	hemolytic uremic syndrome
kDa	kilo dalton (protein weight)
LDF	linear discriminant function (score for σ^{70} promoter prediction with BProm)
MECP synthase	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
MF	molecular function
MLSA	multilocus sequence analysis
mORF	mother open reading frame or mother gene - the annotated gene of the overlapping gene pair
ncRNA	non-coding ribonucleic acid
nr database	non-redundant database of NCBI, blast
OD_{600}	optical density measured at a wavelength of 600 nm
OGC	overlapping gene candidate
ORF	open reading frame
PBS	phosphate buffered saline
Pfam database	protein family database
PVDF	polyvinylidene difluoride (membrane for Western Blot)
RACE	rapid amplification of cDNA ends
RNA	ribonucleic acid

RPKM	reads per kilobase per million sequenced reads
RT-qPCR	reverse transcription quantitative polymerase chain reaction
SDS PAGE	sodium dodecyl sulfate polyacrylamide electrophoresis
sORF	shadow open reading frame - novel gene; lies in the shadow of the mother gene
SPA tag	sequential peptide affinity tag
SRA database	sequence reads archive database
T3SS	type 3 secretion system
tar	translationally arrested mutant
TCA	trichloroacetic acid
TEGT family	testis enhanced gene transfer family
TRG	taxonomically restricted genes
U	units; measure for enzyme activity

1. Introduction

1.1 Overlapping genes

1.1.1 The bacterial genetic code in light on identifying novel genes

The information necessary for cellular processes is stored in the deoxyribonucleic acid (DNA), a double stranded, complementary string of nucleotides which consists of sugars (nitrogenous bases) and phosphate groups, and are structured as a helix. The nucleotides adenine (A), guanidine (G), thymidine (T) and cytosine (C) constitute the letters, the triplet code - encoding one amino acid - a syllable of the word, which is called an “open reading frame” (ORF). The triplet code is written in one of six reading frames (Figure 1.1). Protein coding ORFs are generally known as genes. Bacterial genes are located one after the other forming a densely packed genome. Regulatory elements, like promoters, Shine Dalgarno sequences and terminators are required for a cell to read the genetic code as proteins.

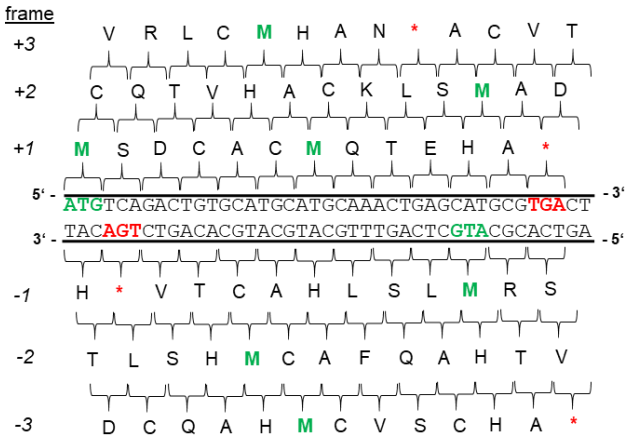


Figure 1.1: The genetic code written as six-frame translation; The canonical start codon is highlighted in green, the stop codons in red.

The identification of novel genes is accompanied by many aspects which have to be considered. A bacterial genome is traversed by many ORFs which can be detected by their start and stop codon. Not every ORF has a function, many are present for statistical rather than functional reasons (Mir, et al. 2012). The stop codon (TAG, TGA, TAA) is, with few exceptions (Heider, et al. 1992), always the translational stop. However, start codons are more complex. The canonical start codon is ATG and further alternative start codons are possible in bacteria (GTG, TTG, CTG, ATT, ATC and ATA; Hecht, et al. 2017). The ability of all codons to function as start codon was tested by Hecht, et al. (2017) in *Escherichia coli (E. coli)*. ATG (81,8%), GTG (13,8%) and TTG (4.3%) were most active and CTG (0.02%), ATT (0.02%), ATC (0.006%) and ATA (0.004%) were significantly weaker. All remaining 57 codons were also tested and their percentage to be active as start codon was, surprisingly, always higher than zero. Further, the amino acids encoded by start codons are frequently used in proteins making the identification of the start codon naturally used by a cell nearly impossible when identifying novel genes from DNA sequence alone. The question arises which codon is used to initiate translation when ORFs are identified to be putative functional genes. Usually, ORFs are annotated by using their first

canonical or alternative start codon downstream of a stop codon resulting in the longest possible ORF. Machine learning can be further used to identify a more probable start (Besemer, et al. 2001). In the end, the true start codon can only be identified by laboratory experiments (for example conducted by Hücker, Vanderhaeghen, Abellan-Schneyder, Scherer, et al. 2018).

1.1.2 Definition and classification of overlapping genes

Genome annotation programs usually decide for one of the six open reading frames, the one judged most probable to encode a gene (Delcher, et al. 2007). However, there is the possibility that more than one open reading frame is present at the same locus and two genes are overlapping. Per definition, the first discovered gene is called “mother ORF” (mORF); its reading frame is defined to be the +1 frame. The overlapping gene is called a “shadow ORF” (sORF), because it is located in the “shadow” of the mother gene (Yooseph, et al. 2007).

Possible overlapping gene classes are visualized in Figure 1.2. Shadow ORFs overlapping in *sense* (Cheng, et al. 2010) are in the alternative frames +2 or +3, those overlapping in *antisense* are in -1, -2 or -3 frame (Figure 1.1). Shadow ORFs can be either completely embedded in their mORF or can overlap terminally. Terminal *sense* overlaps can be head-to-tail or tail-to-head and *antisense* overlaps can be head-to-head or tail-to-tail.

The size of the overlapping region is also important. Short, ‘trivial’ overlaps (<90 bp, Mir, et al. 2012) are frequently observed in bacterial operons and translationally coupled (Johnson and Chisholm 2004; Lillo and Krakauer 2007). We are only interested in non-trivial overlaps, because there are experimental indications for the presence of more such bacterial overlapping genes than presumed so far.

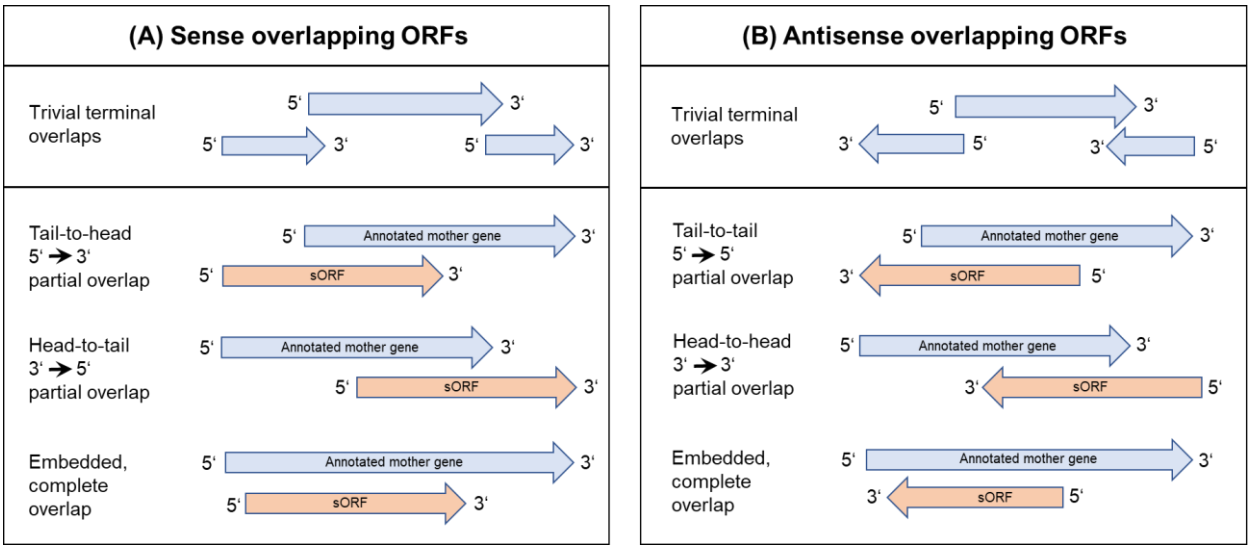


Figure 1.2: Overlap classes depending on the relative location of mORF and sORF.

1.1.3 Why do genes overlap?

The first overlapping gene pair was already discovered during the 1970s when modern molecular biology was still young (Barrell, et al. 1976). The *E. coli* bacteriophage ϕX174 has the gene E that is fully embedded in gene D, in *sense*. The proteins produced by the two OLG-pairs present in ϕX174 are not related based on their protein sequence and their magnitude of expression is different during infection of bacteria. Today, it is generally assumed that overlapping gene pairs are widespread in viruses (Chirico, et al. 2010). In some viruses, like hepatitis B, the majority of genes are overlapping (Mizokami, et al. 1997).

The first speculations about why overlapping gene pairs exist turned around the hypothesis that overlapping genes are an energy-effective way to increase the informational content of the genome (Krakauer 2000). Further, a viral genome compression could be caused by the physical size constraint of icosahedral capsids (Chirico, et al. 2010, and references within). Thus, viruses can gain novel functions more easily by obtaining overlapping genes than by increasing the genome size. This has to be relativized, because the correlation is not true for viruses with a more flexible capsid shape. Further, the genome of ϕX174 has been fully decompressed (Jaschke, et al. 2012). One virus type obtains a particular benefit from gene overlap - RNA viruses. Overlapping genes form a constraint against the fitness disadvantage which is caused by the high mutation rate of RNA. However, an argument against this hypothesis is that RNA viruses have a lower percentage of overlapping genes relative to their genome size compared to DNA viruses (Chirico, et al. 2010).

The existence of overlapping gene pairs should impair fitness effects on the organism, because one single point mutation does affect two genes and all biological processes they are involved in (Krakauer 2000). However, Hughes, et al. (2001) successfully showed that the evolutionary constraint on two genes could be different, even in overlapping regions. This gives indications about an individual evolution of both genes. Fernandes, et al. (2016) scanned all genes in HIV and found out that overlapping regions are not more conserved than non-overlapping regions. The evolutionary constraint is further dependent on the reading frame (Smith and Waterman 1980; Rogozin, et al. 2002; Lèbre and Gascuel 2017) and gene age (Sabath, et al. 2012). Simulation of the embedded, overlapping gene pairs showed that they can be artificially designed (Opuu, et al. 2017).

1.1.4 Distribution of overlapping genes in bacteria

Unfortunately, most authors do not distinguish between trivial and non-trivially overlapping genes (Saha, et al. 2015). The former ones are common, while the latter are thought to be a rare phenomenon. In the best studied bacterial organism *E. coli*, currently only nine pairs are described in the literature (McVeigh, et al. 2000; Behrens, et al. 2002; Delaye, et al. 2008; Balabanov, et al. 2012; Kurata, et al. 2013; Fellner, et al. 2014; Fellner, et al. 2015; Haycocks and Grainger 2016; Hücker, Vanderhaeghen, Abellan-Schneyder, Scherer, et al. 2018; Hücker, Vanderhaeghen, Abellan-Schneyder, Wecko, et al. 2018).

There are more bacterial species with published overlapping gene pairs. In the following a few examples will be described. *Bacillus subtilis* has a gene pair *dnaE/antE* (Wang, et al. 1999), in which both gene products are involved in sporulation. Their expression was shown to be regulated independently from each other. While *dnaE* is taxonomically widely distributed, *antE* was discovered by Wang, et al. (1999). Although the study was published 19 years ago, *antE* is annotated only in *Bacillus subtilis* so far (checked using blastp, refseq database), although the sequence is present in further *Bacillus* species, without annotation (checked using tblastn, nr database). Some shadow ORFs have a good chance to be “real” genes, but they have not been annotated in other genomes after their discovery. One example is *nov01,* which was discovered in *Pseudomonas flourescens* by mass spectrometry and RT-qPCR (Kim, et al. 2009). The novel gene was detected, but not more closely characterized in detail, thus its function is still unknown. The overlapping gene pair *rpmH/rnpA* discovered in *Thermus thermophilus* is a good example of a widely distributed and annotated overlapping gene pair (Feltens, et al. 2003). The mother gene *rpmH* encodes the ribosomal protein L34 and it is embedded in the newly described *rnpA* gene encoding a RNAse P with verified enzyme activity. Finally, there are some cases, in which the

discovery of the shadow ORF resulted in the removal of the annotated mother gene, such as in *Pseudomonas syringae* (Filiatrault, et al. 2010) or in *Desulfovibrio vulgaris* (Price, et al. 2011). Although the number of verified overlapping genes is low, high-throughput expression detection methods, like RNAseq and RIBOseq, suggest that there are many more active bacterial overlapping gene pairs than known so far.

1.2 The identification of genes

1.2.1 Genome annotation

Next generation sequencing technologies enable the sequencing of many genomes. Automatic annotation programs like GLIMMER (Gene Locator and Interpolated Markov ModelER, Delcher, et al. 2007) or RAST (Rapid Annotation using Subsystem Technology, Aziz, et al. 2008) enable identification of putative genes based on sequence homology and, thus, on a comparison to known genes (Delcher, et al. 2007). The quality of genome annotation increases by increasing number of sequenced genomes.

Homology based annotation is sometimes problematic, as extensively discussed in Richardson and Watson (2013). Paradoxically, new strains are sequenced to find genetic differences to close relatives, but similarity based methods are used for annotation. This can lead to conflicting data. Programmed translational frameshifts disrupt the sequence and lead to lack of recognition by annotation programs (Danchin, et al. 2018). Further, there are a multitude of ORFs annotated which cannot be assigned to a function. Not all functional genes look like well characterized genes, and some simply do not have any homologues because they are species or even strain specific. They are also potentially excluded by usual automatic annotation. However, algorithms steadily improve and today often include further indicators like regulatory elements (e.g. GeneMarkS-2, Lomsadze, et al. 2017).

1.2.2 Comparative genomics

The information available on all known genomic sequences is stored in databases such as NCBI RefSeq (Pruitt, et al. 2012) which can be searched for homologies with blast (Johnson, et al. 2008). The linking with further databases facilitates comparative genomics. There are protein structure databases like the RCSB Protein Data bank (Rose, et al. 2017) or such with function linked information as conserved domains (CD, Marchler-Bauer, et al. 2014) or protein family domains (Pfam domain, Finn, et al. 2013). Information about cellular processes are for example

stored as GO terms (Gene Ontology, Gene Ontology Consortium 2016) or in the KEGG database (Kyoto Encyclopedia of Genes and Genomes, Du, et al. 2014).

Genes without any homologous sequences or domains can be analyzed in respect to pattern or motives which can be found in known genes or proteins and give indications about processes they are involved in (PROSITE, Sigrist, et al. 2013). Structural protein features are transmembrane helices, secondary structure, topography or localization of a protein. There are many prediction tools available and listed on the ExPaSy, Bioinformatics Resource Portal (Artimo, et al. 2012). A tool which combines most of these protein features is PredictProtein (Rost, et al. 2004) which will also be used in this thesis.

1.2.3 Experimental approaches

The experimental discovery of novel (unannotated) genes encompasses high-throughput methods which are able to detect the expression of hundreds to thousands of novel genes. Microarrays scan active gene regions by hybridization of a fixed probe to the cDNA of interest. Microarrays can be combined with the identification of gene specific phenotypes (Mukherjee, et al. 2008). However, limitations of this method, like high background signal and low reproducibility (Leimena, et al. 2012), mean it has largely been replaced by the more recently developed next generation based approaches RNAseq (Flaherty, et al. 2011) and RIBOseq (Ingolia, et al. 2012) which detect signals of transcription and translation across whole genomes. RNAseq and RIBOseq have a relatively low cost and high sequencing depth (Leimena, et al. 2012). After discovery, active novel genes can be verified and more closely analyzed in respect to regulatory elements or biological processes they are involved in (Hücker, et al. 2016; Hücker, et al. 2017). RNAseq and RIBOseq are also used for the initial discovery of overlapping genes in our lab with subsequent characterization of the genes (e.g. Fellner 2015). Proteins encoded by novel genes can be detected by Western Blot (Baek, et al. 2017a), GFP fusion proteins (Hücker, Vanderhaeghen, Abellan-Schneyder, Scherer, et al. 2018) or proteomics (Omasits, et al. 2013). The latter one is often considered the ‘gold standard’, but it is one of the least sensitive high-throughput methods for the detection of proteins (Pandey and Mann 2000; Baek, et al. 2017a). Further, it needs a pre-assembled protein database in the background with a high specificity for the proteins to be detected. Thus, 6-frame translations of whole genomes are possible, but the sensitivity decreases substantially (Nesvizhskii 2014).

1.3 The evolution of novelty

1.3.1 Taxonomically restricted genes

Each organism contains genes that are species specific and do not have any homologues. These genes were formerly called orphans (sometimes ORFans). However, this phenomenon occurs on all taxonomic levels, thus, such restricted genes are better called “taxonomically restricted genes” (TRG). The abundance of TRGs ranges between 10% to 30% in Eukaryotes (Khalturin, et al. 2009) and 1% to 50% in bacteria and archaea (Satoshi and Nishikawa 2004). After the discovery of the first TRGs in the 1990s, it was thought that the gap would be closed over time and there simply had not been enough genomes sequenced and stored in the database to find the homologies between the genes of all organisms at that point in time. However, the number did not significantly decrease over subsequent years (Khalturin, et al. 2009).

Today, there are two hypotheses about the nature of TRGs. First of all, some seem to be genes which are necessary for the species specific ecological niche (Khalturin, et al. 2009). They result from the evolutionary novelty each species has in comparison to other species. Such TRGs used for adaptation can be both, old or young. Old genes often have more general functions (Pál, et al. 2006), are highly conserved (Reedy, et al. 2000) and can have a high complexity (Milde, et al. 2009). However, most TRGs have a high mutation rate (Daubin, et al. 2003), low complexity, low expression level (Carvunis, et al. 2012) and are indicative of recently emerged genes (Tautz and Domazet-Loso 2011). Until now, most studies on TRGs and gene emergence have been conducted in eukaryotes.

1.3.2 Mechanisms of gene emergence: Duplication-divergence

For a long time, there was the dogma that genes and the produced protein folds can only emerge from existing structures (Tautz 2014). Functional novelty would be predominantly gained by duplication and subsequent mutation of a gene copy while the “parental” gene is kept conserved (Tautz 2014). Sufficient evolutionary change over time can lead to a loss of similarity to the original gene (Domazet-Loso and Tautz 2003) or the loss of the parental gene during evolution (Tautz 2014). Duplication has the advantage that all necessities for the functionality of a gene are already present. This might include regulatory elements, like ribosome binding sites or promoters, and the open reading frame already has a “gene like” structure. Re-functionalization and reorganization in cellular networks can occur stepwise. However, there will be an adaptive conflict between old and duplicated gene when changing the function

(McLysaght and Guerzoni 2015 and references within). Absolute novelty cannot easily be gained from existing genes. The evolution of an already functional protein has constraints, particularly functional domains cannot be easily changed by mutations to bridge the gap of states, which are not functional in sequence space (Albà and Castresana 2007).

1.3.3 Mechanisms of gene emergence: emergence by gradual evolution

The massive number of TRGs identified by comparative genomics lead to a greater realization that genomes experience a frequent gain and loss of genes, including *de novo* gene emergence from non-coding DNA (Tautz and Domazet-Loso 2011; Tautz 2014). Carvunis, et al. (2012) developed a gene evolution model, a kind of “life cycles of genes”, which merges all fundamental thoughts about gene evolution. The gradual evolution of genes from non-coding sequences processes via an intermediate state - a “proto-gene”, not yet fixed in a cellular network and weakly expressed. Proto-genes form a reservoir for a cell to develop novel functions. A high mutation rate let them adapt to functional genes if they are required, but also leads to a quick loss of their coding potential (Tautz and Domazet-Loso 2011). If *de novo* genes are expressed, they more and more adapt to their function, become more conserved, longer and increase their expression rate. Sometimes, even highly conserved genes lose their function by pseudogenization, they “die” and ‘evolve’ back to non-genic sequences. Carvunis, et al. (2012) found that many more genes emerged *de novo* than by duplication and divergence. This model was supported by experimental data and protein feature predictions all showing a continuous evolution process. Support for this instable “life cycle” hypothesis of genes was provided by Palmieri, et al. (2014) in *Drosophila*. Further studies provide indications for a continuous evolution of genes (for example Neme 2014; McLysaght and Guerzoni 2015; Durand, et al. 2018).

1.3.4 Mechanisms of gene emergence: emergence by preadaptation

It hard to imagine that protein coding genes emerge continuously, because evolving proteins have to prevent folding states that form toxic, amyloid-like aggregations (Monsellier and Chiti 2007). Thus, non-genic sequences may preadapt to a gene-like sequence, before it is expressed to a stable protein. Preadaptation facilitates a non-genic sequence to become functional at one go, which results in the fixation of the gene. This hypothesis is in strong contrast to the continuum hypothesis where continuous processes let non-genic sequences evolve to a gene. The two models can be distinguished by the difference between young genes and non-genic

sequences. If this difference is high, there is evidence for an abrupt process by preadaptation (Wilson, et al. 2017). A low difference, as shown by Carvunis, et al. (2012), indicates continuous evolution. The preadaptation hypothesis is questioned by many scientists, because a gene-like cryptic sequence requires a 'foresight' of the in a future emerging gene. Masel (2006) found that non-genic sequences can be enriched for potential adaptations by 'self-evidently deleterious variations' which do not require a 'foresight'. However, the following questions remain still unanswered: (1) why and how non-genic sequences should evolve to a gene-like sequence and (2) why should gene-like non-genic sequences abruptly obtain a protein function ready for subsequent selection processes.

1.3.5 Mechanisms of gene emergence: Overprinting

Overprinting is a hypothetical mechanism for the emergence of overlapping gene pairs by preadaptation. The sequence of amino acids in which the novel gene is written, "is overprinted" on top of the non-coding sequence while the already existing gene has to be kept functional. Overprinting can be tested by phylogenetic analysis of both genes. If the mother gene is older than the overlapping gene, this is a strong indication for overprinting (Keese and Gibbs 1992). Keese and Gibbs (1992) hypothesized that overprinted genes have the advantage of emerging easier than genes from intergenic sequences due to the constraint given by the "mother gene". The hypothesis is in fact as old as the idea of overlapping genes (Grassé 1977). Since most overlapping genes were predominantly known from viruses and bacteriophages, most cases of overprinting have been described in these. There are only a few studies regarding eukaryotes (Klemke, et al. 2001; Nekrutenko, et al. 2005; Sherr 2006; Chung, et al. 2007; Neme and Tautz 2013) or prokaryotes (Delaye, et al. 2008; Fellner, et al. 2015; Hücker, Vanderhaeghen, Abellan-Schneyder, Scherer, et al. 2018; Hücker, Vanderhaeghen, Abellan-Schneyder, Wecko, et al. 2018) and their results generally confirm the current hypothesis of overprinting obtained from viruses.

1.3.6 Discovery of recently emerged genes by phylostratigraphy

The method to classify genes according to their age is called phylostratigraphy. The approach was originally developed by Domazet-Lošo, et al. (2007) to trace back the macro-evolutionary path of genes which emerged by duplication-divergence. In principle, all genes of one particular genome are blasted and the farthermost related organism, in which the gene has a hit, determines the gene age. The phylostrata (singular: phylostratum) are the taxonomic classes to

which each gene can be assigned to a particular age. For example, genes having homologues only in *E. coli* are very young and TRGs. Those found in *Salmonella* as more distantly related genus are older. Genes with more distant homologues in *Salmonella* are present in the genus *Escherichia* and in non-*Escherichia* enterobacteriaceae, and can thus assigned to the phylostratum 'enterobacteriaceae'. Homologues found in *Bacillus* are assigned to the phylostratum 'bacteria' by being present in non-proteobacteria, in proteobacteria, in the class γ-proteobacteria and so on. Up until now, phylostratigraphy has only been published in eukaryotes to determine expression patterns associated with ontogeny (Domazet-Lošo and Tautz 2010a), multicellularity and disease (Domazet-Lošo and Tautz 2010b) or the identification of *de novo* emerged genes (for example Carvunis, et al. 2012; Neme 2014).

1.4 The model organism EHEC and its genome

1.4.1 The EHEC pathogenicity

Enterohaemorrhagic *Escherichia coli* O157:H7 (EHEC) is a gastrointestinal pathogen which causes a severe form of gastroenteritis with hemorrhagic diarrhea. The natural host of EHEC are believed to be ruminants where a broad range of prevalence between 0.2% and 48.8% was measured in the feces. The EHEC colonization is asymptomatic in these animals (Pennington 2010).

The first EHEC outbreak in human was reported in 1982 when people had bloody diarrhea after eating hamburgers (Riley, et al. 1983). The transmission route is either foodborne or environmental (Pennington 2010). A study analyzed 90 outbreaks all over the world between 1982 and 2006 and found out that 42% of EHEC was transmitted by food, 12% by dairy products, 8% by animal contacts, 7% by contaminated water, 2% via the environment (such where animal contact is unlikely or which might be contaminated with animal feces) and 29% by unclear sources (Snedeker, et al. 2009). Most strains are cold- and acid tolerant which enables them to survive in fresh food products, like sprouts, in final products, like beef products, or in frozen food (Castro, et al. 2017). The main cause for the transmission of EHEC via meat products is undercooked or raw meat (Kintz, et al. 2017). Transmissions are also possible from person to person. According to Snedeker, et al. (2009), 20% of the outbreaks come from secondary spread. However, EHEC does persist only few days in the human gut which restrict the prevalence of person to person transmission (Pennington 2010).

Typical symptoms are abdominal pain and non-bloody diarrhea which become bloody after 1-4 days (Pennington 2010). In rare cases, people develop a hemolytic uremic syndrome (HUS)

leading to a renal failure (Tarr, et al. 2005), damage of the central nervous system (Eriksson, et al. 2001) and even death (Uchida, et al. 1999). Children younger than 5 years predominantly develop HUS and EHEC infections are the most common reason for HUS in the world (Pennington 2010). EHEC outbreaks must be registered in Germany. According to the Robert-Koch institute, 1816 infections were reported in 2016, and 29% of them were children < 5 years old. Among all patients, HUS appeared in 44 cases and 4 people died of it (Robert-Koch-Institut 2017).

The EHEC virulence is caused by the production of at least one Shiga toxin (Pennington 2010), a type 3 secretion system (T3SS) (Gally and Stevens 2017) and further virulence genes located on a pathogenicity island which is called *locus of enterocyte effacement*, LEE, organized in five operons (Castro, et al. 2017). When EHEC infects a human host, it forms “attaching and effacing lesions” to adhere to intestinal epithelial cells (Gally and Stevens 2017). The adherence is mediated by two proteins - the outer membrane protein intimin and its *translocated intimin receptor*, TIR. The T3SS forms a channel in the host cell membrane which enables EHEC to inject proteins into the host cell (Castro, et al. 2017), which rearrange the intestinal epithelial cell architecture and change the cell physiology, for instance loss of microvilli or accumulation of cytoskeletal proteins beneath the adherent bacteria (Gyles 2007).

There aren’t any effective therapies for EHEC treatment up to now. Antibiotics are not used due to the increased development of HUS, the dissemination of pathogenicity genes and the risk of the emergence of new serotypes (Castro, et al. 2017). Further medications, like non-steroidal anti-inflammatory agents also have side effects that can aggravate the illness (Tarr, et al. 2005). Thus, efforts are mainly made to prevent EHEC contaminations in foods. There are further studies that work on preventing infections with anti-EHEC phage therapies for ruminants (Sabouri, et al. 2017) which do not seem to be very promising (Arthur, et al. 2017), probiotics (Forano, et al. 2015; Bertin, et al. 2017) or vaccines for ruminants against proteins injected by the T3SS (Smith, et al. 2009) or the Shiga-toxin (Albanese, et al. 2018). Interestingly, the EHEC infection rate of cows strongly depend on the cow diet. Forage-fed ruminants have much lower EHEC than grain-fed ones. After sudden switch from grain-food to forage-food, the EHEC populations decline 1000-fold within 5 days in the cows (Callaway, et al. 2009).

1.4.2 The genome of *Escherichia coli* O157:H7 strain EDL933

Escherichia coli O157:H7 strain EDL933 is the reference strain isolated at the first outbreak in 1982 (Riley, et al. 1983). Actually, three fully sequenced genomes are in the NCBI database: Perna, et al. (2001), Latif, et al. (2014) and Fellner, et al. (2016). When starting this study in

2014, genome sequences published by Perna, et al. (2001) or by Latif, et al. (2014) were available. The ‘Latif genome’ was selected, because it is a revision of Perna, et al. (2001) omitting most ambiguities. The EHEC genome consists of 5.55 Mbp large chromosome and a 92 kbp plasmid. There are 5,675 annotated proteins encoded on the chromosome and 97 on the plasmid (Latif, et al. 2014). Strain EDL933 contains a further, smaller plasmid pOSAK1 (Fellner, et al. 2016) and other small plasmids may be present in EHEC (Gannon, et al. 2011). EHEC has an unusually high number of bacteriophages in its genome - 18 regions were found in EDL933. Two of these encode the infectious Shiga toxins Stx1AB and Stx2. Most of the prophages are defect and do not form lytic phages (Hayashi, et al. 2001). When comparing *E. coli* K12 with EDL933, there are regions, which are either K12 or EDL933 specific. Unexpectedly, several of the EDL933 specific regions are not necessarily associated with its virulence (Perna, et al. 2001), but may be important for environmental survival (Gannon, et al. 2011).

1.5 Aim of this thesis

Genes overlapping in alternative reading frames to annotated genes have been described in viruses. While bacterial overlapping genes are usually excluded from annotation, a few have been reported during the last years. This gives rise to the hypothesis that there may be more protein coding overlapping genes in prokaryotes than assumed so far. The genome of the organism of interest, *Escherichia coli* O157:H7 strain EDL933, has a multitude of overlapping open reading frames. Do these have the potential to be protein-coding genes? If they are protein-coding and not random DNA sequences, such overlapping open reading frames should have homologues to annotated protein-coding genes in databases. The biochemical features of the encoded amino acid sequence should be structured and not too different from known proteins. Thus, the biochemical protein features of coding genes should be similar to already characterized proteins and homologous proteins in the database would give indications about putative cellular functions. These contradicting hypotheses, alternative ORFs are present purely due to statistical reasons versus alternative ORFs may encode novel proteins, were tested by a search for annotated homologous proteins using blastp and by PredictProtein, a bioinformatic tool to predict protein features *ab initio* from amino acid sequences. An experimental proof of function was provided by the characterization of one candidate gene, *asa*. Since viral overlapping genes are associated with *de novo* gene emergence, are bacterial overlapping genes also young or are they highly conserved across the tree of life? This question can be answered by using phylostratigraphy, a taxonomic approach to identify the more distantly related

species having a protein-coding gene homologous to an overlapping gene. Finally, the mechanism of gene emergence, either by gradual (continuum hypothesis) or discontinuous evolution (preadaptation hypothesis), was tested by aligning homolog overlapping gene candidates at the transition from non-intact (perhaps “non-functional”) to intact (and, thus, perhaps “functional”) open reading frames.

2. Material and Methods

2.1 Tools and databases used for bioinformatics

Table 2.1: List of tools and databases used for bioinformatics on web pages.

database	application	web link
NCBI, Genbank	genomic nucleotide and protein sequence data	https://www.ncbi.nlm.nih.gov/nucleotide/ https://www.ncbi.nlm.nih.gov/protein/
EMBOSS, Getorf	extraction of all ORFs of a genome	http://www.bioinformatics.nl/cgi-bin/emboss/getorf
NCBI, blast	protein (blastp) or nucleotide (tblastn) sequence homology search using amino acids sequences as query	https://blast.ncbi.nlm.nih.gov/Blast.cgi
EMBOSS Needle	Global pairwise protein or nucleotide sequence alignment	https://www.ebi.ac.uk/Tools/psa/
Reverse Complement	conversion of a DNA sequence in reverse, complement or reverse complement counterpart	https://www.bioinformatics.org/sms/rev_comp.html
ExPaSy, Translate	translate DNA or RNA sequences to a protein sequence	https://web.expasy.org/translate/
Arb SILVA	database for 16S rRNA	https://www.arb-silva.de/
PredictProtein	Prediction of protein features	https://www.predictprotein.org/
ExPaSy, UniProt/Swiss-Prot	Protein database used to test different secretion prediction tools	https://www.uniprot.org/downloads
NCBI, conserved domain database	conserved domain search with protein or nucleotide sequences as query	https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
Promega, BioMath Calculator	Calculation of amount of vector and insert needed for cloning	http://www.promega.com/a/apps/biomath/?calc=m
NCBI, Taxonomy	Check for phylostrata for the single gene phylostratigraphy	https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Tree&id=2&lvl=3&srchmode=1&keep=1&unlock
HHblits	3D-homology model using profile-profile and HMM-HMM alignments; more sensitive than sequence homology alignment	https://toolkit.tuebingen.mpg.de/#/tools/hhblits
NCBI, SRA	database for RNAseq and RIBOseq data	https://www.ncbi.nlm.nih.gov/sra
RCSB database	protein structure database	https://www.rcsb.org/

2.2 Shadow ORF identification and protein homology

2.2.1 Downloading sORFs from the EHEC genome

All ORFs of the genomic sequence of *E. coli* O157:H7 str. EDL933 (Genbank CP008957.1, Latif, et al. 2014) were identified with Getorf (EMBOSS, parameters - code to use: bacterial; minimum nucleotide size: 93; output nucleotide sequences: between start and stop; circular: yes; number of flanking nucleotides: 0). Shadow ORFs were identified as described in Simon, et al. (2011). The minimal ORF length is 93 bp (including stop codon, 30 amino acids) with an overlap to annotated genes of ≥ 90 bp. The start codon was defined as the farthermost upstream possible start codon of an ORF. The start codons used were “NTG” with “N” to be any nucleotide under omission of the rare start codons ATT and ATC (Mir, et al. 2012). The sORF identification numbers appearing in this thesis are named according to the Getorf output. The identification numbers consist of a composite of the Genbank genome ID and a number originated from the consecutively numbered ORFs in the genome (“CP008957_number”, shortly written as #number).

2.2.2 Protein homology search

The complete dataset of overlapping ORFs meeting our criteria comprises 49,650 ORFs. Only those overlapping in *antisense* were aligned against the refseq database (NCBI, in September 2015) using blastp. An E-value cutoff of 10^{-3} or lower was chosen as threshold, in which protein sequences are considered homologues as used elsewhere (Neme and Tautz 2013; Kuchibhatla, et al. 2014). The above analysis was conducted by Svenja Simon, our cooperation partner in bioinformatics (Prof. Keim, Department of Data Analysis and Visualization, University of Konstanz). Mobile genetic elements, corresponding to phages or transposases, were manually excluded. For subsequent analyses, the hit with the highest E-value, but below or equal to the threshold of 10^{-3}, was defined as the furthest relative to *E. coli* (referred to as the “last hit”). All 2,180 resulting blastp hits are listed in Supplementary table S1.

2.2.3 Computational analysis of full-length match of sORFs to blastp hits

Shadow ORFs were analyzed in respect to the length of the sequence, which matches to a homologous gene in the NCBI database (‘coverage’). For this, the amino acid sequences were compared to their homologous protein counterparts after a blastp search against *E. coli* or enterobacteriaceae (nr database, February 2016, E-value cut-off ≤ 10^{-3}). In this case, the nr

database was chosen to get access to genome accession numbers of organisms in which the homologue is present. In contrast to the phylostratigraphic analyses, not only the farthest hit but all hits were included in the analysis. A blast hit was defined to match with its full-length in case of a query coverage of ≥ 80%. This allows variations in homologues by varying sequence length or terminal mutations, which are not shown by blast analysis. It was hypothesized that sORFs with blastp hit have non-intact annotated genes. Thus, fifty randomly selected sORFs with a complete match to proteins in refseq were analyzed in detail for being intact in overlapping gene pairs. The localization and sequences of respective mORF homologues were downloaded from NCBI. The mORF sequence was aligned with EMBOSS Needle pairwise alignment (Li, et al. 2015) and viewed with Artemis 17.0 (Carver, et al. 2011).

2.3 Phylostratigraphy and Predict Protein of sORFs and aORFs

2.3.1 Gene age classification

Shadow ORFs, which had a hit in the refseq database using blastp, were classified into seven phylostratigraphic levels. For this, the 16S rRNA sequence of the organism with the last blastp hit was downloaded from the ARB SILVA database (Quast, et al. 2013). The 16S rRNA sequences were aligned to the *E. coli* 16S rRNA sequence using EMBOSS Needle pairwise alignment (Li, et al. 2015). Gene age classification according to sequence identity cut-offs were defined after Yarza, et al. (2014) (Figure 2.1*)*.

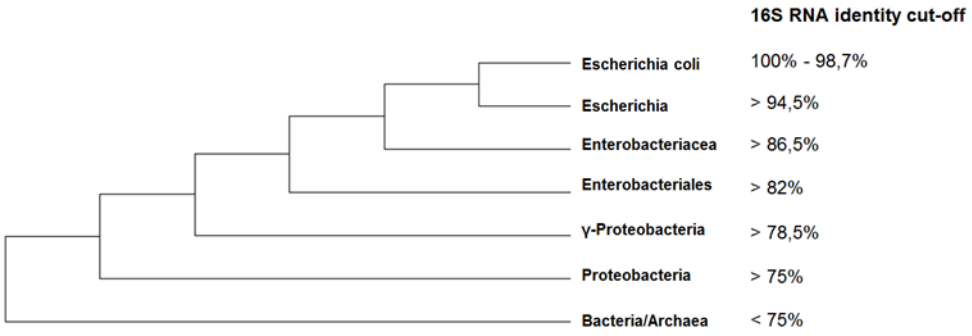


Figure 2.1: **Schematic view on Phylostrata of ORFs in EHEC EDL933 based on 16S rRNA sequence identities of the organisms in which the protein of the last blastp hit was found.** Protein blast was conducted against refseq database, E-value cutoff 10^{-3}. 16S rRNA sequence identity cut-offs were used as defined by Yarza, et al. (2014).

2.3.2 Protein feature prediction

A number of protein features of the sORFs were predicted with PredictProtein (Yachdav, et al. 2014) in cooperation with the RostLab (Prof. Rost, Department of Bioinformatics and Computational Biology, Garching, Germany) and analyzed in dependence on their phylostratigraphic level, using PROFphd (secondary structure, solvent accessibility, Rost and Sander 1994), DISULFIND (disulfide bonds, Ceroni, et al. 2006), METADISORDER (disordered regions, Schlessinger, et al. 2006) and PHDhtm (transmembrane helices, accuracy >86%, Rost, et al. 1996). To evaluate the best method for our purpose, three secretion prediction programs were tested in cooperation with Tatyana Goldberg (RostLab): Secretome2.0 (Bendtsen, et al. 2005, version for Gram-negative bacteria), SignalP4.1 (Petersen, et al. 2011, version for Gram-negative bacteria) and LocTree3 (Goldberg, et al. 2014, version for bacteria). A dataset of 77 bacterial proteins, which are experimentally verified to be secreted, was downloaded from Swiss-Prot (Boutet, et al. 2007). All of them were added to the database after development of the programs tested. Forty-seven proteins are non-redundant, because some of them were homologues (cutoff 20% identity). In the redundant set, LocTree3 predicted 68 proteins correctly, SignalP4.1 identified only 31 proteins as secreted (Supplementary figure S1). Finally, Secretome2.0 predicted 51 proteins as non-classically secreted proteins, i.e. those without signal peptides (Bendtsen, et al. 2005). About 10% of the proteins were not predicted to be secreted by any of the programs. All proteins predicted by Secretome2.0 were also predicted by both, SignalP4.1 and LocTree3. Only one additional protein was predicted by SignalP in comparison to LocTree3. For this, all subsequent analyses were performed with LocTree3. Disordered proteins were analyzed by using two parameters. One was the average percentage of disordered regions per ORF which was defined as ≥ 30 subsequent amino acids (Uversky 2011). The second was the average percentage of disordered amino acids over all ORFs independ from its length. The number of disulfide bonds was normalized to a length of 100 amino acids and plotted using “Vioplot” (CRAN, Hintze and Nelson 1998). The annotated ORFs of EHEC EDL933 were analyzed likewise with all prediction programs as control set.

2.3.3 Prosite pattern and conserved domains (CD) in shadow ORFs

Prosite pattern (SCANProsite, de Castro, et al. 2006) and Pfam domains (HMMER, Finn, Clements, et al. 2015) were identified via PredictProtein. Conserved domains were identified using the conserved domain database, NCBI (Wheeler, et al. 2007).

2.4 Evolutionary analysis of selected genes with phenotype

For the phylostratigraphic analysis of single overlapping genes, the gene age was indirectly identified by using the mother gene which was always older than or of an equal age as the shadow ORF. First, mother gene homologues were identified by tblastn (E-value cutoff 10^{-10}, nr database, identity ≥ 30%). A more stringent E-value cutoff was chosen, because an E-value cutoff of 10^{-3} could be questionable in some cases (see discussion). Further, the non-redundant database was selected, because the entries contain genome sequence information in contrast to the refseq database. Exemplary nucleotide sequences with a broad range of sequence identities were downloaded from NCBI. To identify intact and non-intact sORF homologues, the mORF sequences were aligned in reverse complement. In earlier phylostratigraphic trees (*laoB, ano, slyC*, Hücker 2017), the sORF homologue was identified by pairwise alignments (EMBOSS needle) of the respective reverse complement mORF sequence against the sORF of EHEC Sakai (which is the strain they are present) by using the nucleotide sequence. The matching overlapping region was translated in amino acids (ExPaSy, translate tool) and aligned using MUSCLE in Mega6 (Tamura, et al. 2013). This approach is able to distinguish between intact and non-intact sequences (see below), but it is inexact when indel mutations cause position effects or for short sequences as discussed in section 4.6. 'Intactness' is defined here as open reading frame, which do not have any internal stop codon within the region matching to the sORF in an alignment. An approach more precise than that used of *laoB, ano and slyC,* was implemented for the trees of the OGCs (in EHEC EDL933) to get a better resolution of the intact to non-intact transition. For this, all reverse complement mORF sequences were aligned in a multiple alignment using MUSCLE in Mega7 and the position of the EHEC overlapping gene was used to identify the start and stop positions of the homologous overlapping genes. The exact start and stop codons were manually denoted in each case. The identified nucleotide sequences were translated in amino acids and aligned with MUSCLE in Mega7 (Kumar, et al. 2016). Genes homologous to the sORF were defined by using a sequence identity cutoff of ≥30% (Bolten, et al. 2001). The sequence identity was determined by a pairwise alignment of the sORF homologue to the sORF while considering only the matching region. However, it must be kept in mind that the identity cutoff of 30% is just a first attempt, because the sORF sequence can be present because of the mORF anyway. Both sORF and mORF were additionally searched using blast (blastp: refseq, tblastn: nr, E-value cutoff ≤ 10^{-10}) to compare them with gene ages obtained by phylostratigraphy.

The phylogenetic species trees were constructed after Fellner, et al. (2015). For the trees of *laoB, ano* and *slyC*, a MLSA tree (concatenated sequence of the following housekeeping genes:

16S rDNA, *atpD, adk, gyrB, purA* and *recA*) was used. However, in far related species those genes are not always present. Consequently, a combination of 16S rRNA gene tree and MLSA tree was used for the OGC trees. 16S rRNA sequences were downloaded from the Arb SILVA database. The 16S rRNA gene tree was constructed for all selected species with mother gene homologue, the MLSA tree of species within the family *Enterobacteriales* to increase the resolution. Finally, both trees were merged. The 16S rRNA or the multilocus sequences were aligned using ClustalW in Mega6 (*laoB, ano, slyC*) or Mega7 (OGCs). The alignments were manually checked and columns with gaps or ambiguities were removed. The best nucleotide substitution model was calculated for each tree and the model with the highest Bayesian information criterion (BIC) was used. The final species tree was a Maximum-Likelihood tree using Neighbor Joining, which was bootstrapped 1000-times. The phylostratum of the farthermost related species with homologue to the sORF was determined by using the Taxonomy Browser, NCBI. The tree features specific for each tree can be found in Table 2.2. The sORF homologues and their locations can be found in Supplementary table S2.1 - 2.19.

A negative control of the overlapping gene pairs was constructed as follows. The methods described here were implemented as a BASH script written by Zachary Ardern. The positions of sORFs and the annotated mother gene were used to determine the relative reading frame of the homologous overlapping gene. The nucleotide sequence of overlapping gene in the genome in which it was detected (the 'comparison genome' - either EHEC EDL933 or EHEC Sakai) was then extracted and translated into the corresponding amino acid sequence, and likewise the nucleotide and amino acid sequences of the annotated mother gene. For each sORF, a selection of genomes in which blast hits for the mother had been found was previously selected, as described earlier. This selection was downloaded using the edirect E-Utilities program, and all ORFs in each genome (longest possible for each stop codon, at least 93 nucleotides long) were extracted using EMBOSS getorf. For each 'search' genome, these ORFs were constructed into a database for use with the DIAMOND blast algorithm (Buchfink, et al. 2014). Homologs of the sORF were then searched for in each 'search' genome database constructed, using the DIAMOND algorithm. The best match was taken to be the homolog previously detected with blastp, and the nucleotide sequence of the 'mother gene homolog ORF' extracted using the positions from the file created with 'getorf', which matched the DIAMOND result. This homolog was then aligned against the original annotated gene, using the EMBOSS needle aligner, and the sequence identity of the amino acid sequences calculated (number of matches in alignment divided by original mother gene length). The antisense sequence in the correct reading frame for the OLG was extracted and translated, this was similarly aligned against the original comparison genome (e.g. EHEC EDL933) overlapping gene sequence, and the sequence identity similarly

calculated. To create a negative control set, this whole process was repeated for 100 randomly selected mother genes from the comparison genome. A fully embedded control 'overlapping gene' sequence of the same length of the sORF was created by extraction of a randomly selected portion of the appropriate relative reading frame of the 'control mother gene'. As not all sORFs in EHEC EDL933 or Sakai are embedded in the mother gene, the negative control is more applicable for embedded genes (*laoB, sylC, asa,* OGC15, OGC57, OGC85, OGC121, OGC198) than for terminal overlaps.

Table 2.2: Features of each species tree. Abbreviations: BIC = Bayesian information criterion, GTR = General time reversible, TN93 = Tamura-Nei, T92 = Tamura 3-parameter, K2 = Kimura 2-parameter model, HYK = Hasegawa-Kishino-Yano, G = discrete Gamma distribution with five rate categories, I = the model allows some sites to be evolutionary invariable.

Overlapping gene	Tree	BIC	Nucleotide substitution model	Number of species per tree	Number of positions
asa	MLSA	174,245	GTR + G + I	45	8,189
	16S	38,428	K2 + G + I	93	1,320
laoB	MLSA	123,336	GTR + G	31	7,484
anoG	MLSA	125,016	GTR + G + I	40	8,025
sylC	MLSA	107,061	GTR + G + I	53	7,240
OGC15	MLSA	167,344	TN93 + G + I	28	7,930
	16S	33,505	GTR + G + I	56	1,353
OGC23	MLSA	136,793	TN93 + G + I	27	8,153
OGC51	MLSA	26,533	TN93 + G + I	11	8,313
OGC57	MLSA	204,228	TN93 + G + I	37	8,072
	16S	19,958	T92 + G + I	44	1,342
OGC75	MLSA	52,785	TN93 + G	10	8,298
OGC85	MLSA	143,986	TN93 + G + I	32	8,227
	16S	13,462	TN93 + G + I	35	1,439
OGC106	MLSA	136,793	TN93 + G + I	23	8,252
OGC121	MLSA	93,108	TN93 + G + I	17	8,192
	16S	18,582	TN93 + G + I	31	1,382
OGC167	MLSA	51,930	GTR + G + I	16	8,153
OGC194	MLSA	144,438	TN93 + G + I	29	8,275
OGC198	MLSA	119,837	TN93 + G + I	23	7,7761
	16S	19,957	T92 + G + I	43	1,404
OGC226	MLSA	179,904	TN93 + G + I	36	8,165
	16S	18,839	HKY + G + I	48	1,320
OGC231	MLSA	199,260	TN93 + G + I	39	7,846
	16S	24,580	GTR + G + I	57	1,263
OGC241	MLSA	128,986	TN93 + G + I	25	8,277

2.5 Media and buffers used

Table 2.3: List of media, substances and buffers used. All solutions were prepared in milliQ H_2O. LB medium, TE buffer, CTAB/NaCl and PBS were autoclaved for sterilization (120°C, 20 min).

LB medium	10 g	tryptone
	5 g	yeast extract
	5 g	NaCl
	16 g	agar (optional)
	ad 1 l H_2O, pH 7.4	
TE buffer (Tris/EDTA)	10 mM	Tris/HCl
pH 8.0	1 mM	EDTA
CTAB/NaCl	10%	CTAB (Sigma Aldrich)
	0.7 M	NaCl
Phosphate buffer (PBS)	58 mM	$Na_2HPO4{\cdot}2H_2O$ (Fluka)
	17 mM	$NaH_2PO_4{\cdot}H_2O$ (Carl Roth)
	68 mM	NaCl
1x Cathode buffer for SDS-Tris Tricine PAGE, pH 8.25	0.1M	Tris (Carl Roth)
	0.1 M	Tricine (Carl Roth)
	0.1%	SDS (Serva)
1x Anode buffer for SDS-Tris Tricine PAGE, pH 8.8	0.1 M	Tris
	22.5 mM	HCl (Merck)
1x Blotting buffer for Western Blot	50 mM	Glycin
	0.4 M	Tris
	1%	SDS
	20%	methanol
1x TBS-T for Western Blot	10 mM	Tris, pH8
	15 mM	NaCl
	0.05%	Tween20 (Sigma Aldrich)
reaction buffer for Western Blot, pH 9.5	0.1 M	Tris
	4 mM	$MgCl_2$
NBT for Western Blot	50 mg	nitro blue tetrazolium (Applichem) in
	1 ml	70% dimethyl formamide (Sigma Aldrich)
BCIP for Western Blot	20 mg	5-bromo-4-chloro-3-indolyl phosphate (Carl Roth) in
	1 ml	70% dimethyl formamide

2.6 Evolutionary and functional characterization of *asa*

The OGC59 (‘*asa*’) was experimentally characterized. The following chapters 2.6.1 - 2.6.7 describe the general procedure of cloning. This method was used for competitive growth experiments (2.6.8), promoter activity tests (2.6.10) and Western Blot (2.6.12). The primers used for cloning are listed in Supplementary tables S3 and S4. The vector used for competitive growth experiments and Western Blot was pBAD *myc*/His C (Thermo Fisher Scientific). The antibiotic used for selection of clones containing the vector was 120 µg/ml ampicillin (stock solution: 120 mg/ml in 50% ethanol, sterile filtered). Sterile filtration was conducted using 0.2 µm pore filters (Berrytec). As *Citrobacter freundii* CFNIH1 and *Serratia marcescens* WS1359 have natural ampicillin resistance, a streptomycin or kanamycin resistance gene was cloned in the ampicillin coding region of pBAD *myc*/His C. Positive clones of *Citrobacter freundii* were selected on LB + 30 µg/ml streptomycin (stock solution: 30 mg/ml in H_2O, sterile filtered) and of *Serratia marcescens* with 30 µg/ml kanamycin (stock solution: 100 mg/ml in H_2O, sterile filtered). The kanamycin resistance gene was originated from pProbeNT (Miller, et al. 2000, Genbank AF286453.1, nucleotide 5556-6350) and the streptomycin resistance gene from pMRs101 (Sarker and Cornelis 1997, vector used for genomic knockout mutants in our lab, Genbank LT727367.1, nucleotide 696-1499). The vector used to test promoter activities was pProbeNT having a kanamycin resistance gene.

2.6.1 Cultivation conditions

All bacteria used can be found in Table 2.4. If not stated otherwise, all precultures were grown by shaking (150 rpm) in Luria Broth medium (LB) overnight for 12-16 h. EHEC EDL933, *E. coli* TOP10, *C. freundii* CFNIH1, *S. enterica* serovar Gallinarum 287/91, and *Hafnia alvei* DSM30097 were incubated at 37°C, *S. marcescens* WS1359 at 30°C. Long-term storage of bacterial cultures occurred in 40% glycerol (culture:glycerol ratio: 1:1) at -80°C.

Table 2.4: List of bacteria and strains used for laboratory experiments.

Organism	Genbank accession number	Obtained from
Escherichia coli O157:H7 (EHEC) strain EDL933	CP008957.1	Collection de l'Institute Pasteur: CIP106327, also WS4202
Escherichia coli TOP 10	-	Invitrogen, Paisley, UK
Citrobacter freundii CFNIH1	CP007557.1	National Institute of Health, Rockville, USA
Serratia marcescens WS1359	Not available*	Weihenstephan strain collection
Salmonella enterica serovar Gallinarum 287/91	AM933173.1	Weihenstephan strain collection, WS4570
Hafnia alvei DSM30097	Not available**	Leibniz Institute DSMZ - German collection of Microorganisms and Cell cultures

* alternatively used genome: *S. marcescens* DB11 (Genbank: HG326223.1);
** alternatively used genome: *H. alvei* FB1 (Genbank: CP009706.1)

2.6.2 Genomic DNA (gDNA) Isolation

Genomic DNA was isolated from 5 ml of an overnight culture, which was centrifuged (10 min, 5000×*g*). The cell pellet was resuspended in 567 µl 1x TE buffer, mechanically lysed using 100 µl of 0.1 mm zirconia beads (Carl Roth) in a Ribolyzer (MP FastPrep, 2x 45s, 6,5 m/s) and centrifuged (5 min, 13,000×*g*). The cellular proteins were denaturized and degraded with 30 µl SDS and 3 µl proteinase K (20 mg/ml, Sigma-Aldrich) for 3 h at 37°C. The DNA was complexed with 100 µl of 5 M NaCl and 80 µl CTAB/NaCl for 30 min at 65 °C. The DNA extraction was conducted with 1 vol. Roti®Phenol (Sigma Aldrich) and twice with 1 vol. Roti®Phenol / Chloroform / Isoamylalcohol (Carl Roth). Between each step the reaction mix was centrifuged (5min, 15,000×*g*, room temperature) and the upper phase was used. The DNA was precipitated with 0.6 vol. of ice-cold isopropylalcohol (Carl Roth) for >15 min at 4°C. After centrifugation (5 min, 15,000×*g*, 4°C), the pellet was washed twice with 500 µl of ice-cold 70% ethanol (VWR) and centrifuged (5 min, 15,000×*g*, 4°C). The pellet was dried for 20 min at room temperature and resuspended overnight in 100 µl H_2O.

The RNA was digested by addition of 1 µl RNAse A (20mg/ml, Sigma-Aldrich) and incubated for 30 min at 37°C. After addition of 300 µl H_2O, the RNAse was separated from the gDNA with 1 vol. Roti®Phenol / Chloroform / Isoamylalcohol, centrifuged (5 min, 15000×*g*, room temperature) and the upper phase was mixed with 0.1 vol. 3 M sodium acetate (pH 5.2, Sigma Aldrich). After addition of 2.5 vol. of 100% ice-cold ethanol, the DNA was precipitated for 20 min at -20°C and centrifuged (5 min, 15,000×*g*, 4°C). The pellet was washed twice with 500 µl ice-cold 70% ethanol (centrifugation 5 min, 15,000×*g*, 4°C) and dried for 20 min at room temperature. The gDNA was resuspended overnight in 50 µl H_2O at 4°C and long-term stored at -20°C. The concentration was measured with a NanoDrop 1000 Spectrometer (Thermo Fisher Scientific,

Germany) and its purity was controlled with 1% agarose (Bioline) gel electrophoresis (110 mV for 30-45 min).

2.6.3 Polymerase chain reaction (PCR)

The amplification of DNA was conducted by PCR using the High-Fidelity Q5 polymerase (NewEngland Biolabs). For cloning steps, restriction enzyme cutting sites were added with primers. All primers were obtained from Eurofins (Supplemental table S3, *asa* characterization; or Supplemental table S4, *asa* homologues). The conditions used for PCR are listed in Table 2.5. The annealing temperature was variable according to the melting temperature of the primer. The elongation time varied according to the amplicon length (1 min / 1 kbp). The success of each PCR was checked with a 2% agarose gel. The DNA ladders used were either the 50 bp ladder (NewEngland Biolabs, range 50 - 1,350 bp), the 100 bp ladder (NewEngland Biolabs, range 100 - 1,517 bp) or the 1 kbp ladder (NewEngland Biolabs, range 0.5 - 10 kbp) PCR products were purified with the GenElute^{TM} PCR Clean-Up Kit (Sigma Aldrich).

Competitive growth experiments required the construction of translationally arrested knockout mutants using a modified protocol of Patel, et al. (2009) and An, et al. (2005). The point mutations were inserted in the first third of the sequence of *asa* (Figure 2.2) and *asa* homologues (Figure 2.3) introducing either the stop codon or synonymous mutations. The latter one was used for validation after competitive growth in sequencing (section 2.4.10). A scheme of all steps of the mutation PCR can be found in Figure 2.4. For the introduction of a mutation, two mutation primers were necessary which are reverse complement to each other and overlap 100%. The PCR of the translationally arrested insert was conducted in two steps with pBAD *myc*/His C *asa* as template. In the first step, two fragments were amplified in two PCRs - from the start to the mutation primer and from the mutation primer to the end. In a second step, both fragments were merged using the following protocol: All reagents were mixed except the primers and the PCR was performed for 15 cycles to form double stranded fragments. After the addition of the primers, the DNA was amplified for further 20 PCR cycles.

Table 2.5: PCR for Q5 Polymerase (DNAP). The annealing temperature of the primers and elongation time are variable (var.).

Reaction approach:			Reaction conditions (25 cycles):		
5x PCR buffer	5	µl	**Initial Denaturation**	98°C	30s
10 mM dNTP	0.5	µl	**Denaturation**	98°C	30s
10 µM forward Primer	1.25	µl	**Annealing**	var.	20s
10 µM reverse Primer	1.25	µl	**Elongation**	72°C	var.
Taq DNAP	0.25	µl	**Final Elongation**	72°C	2 min
Template	gDNA: 100 ng plasmid: 1 ng colony PCR: part of one colony				
H_2O	ad 25	µl			

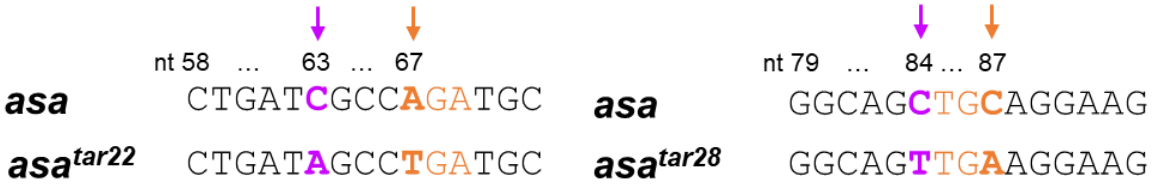


Figure 2.2: Nucleotides changed to construct translationally arrested mutants of *asa*. Two constructs (*asa*^{tar22} and *asa*^{tar28}) were tested in two independent experiments. For each construct, one nucleotide was mutated to introduce a stop codon at amino acid position 22 or 28 (orange bold, *asa*^{tar22}: A→T, Arg22 → stop; *asa*^{tar28}: C→A, Cys28 → stop). Respectively one synonymous mutation (violet bold, *asa*^{tar22}: C→A, *asa*^{tar28}: C→T) validated the peak height ratio of wild type : mutant obtained from Sanger sequencing.

Figure 2.3: Nucleotides changed to construct translationally arrested mutants of *asa* homologues (asa^{tar}). Sequences from the following organisms were tested: (**A**) *Citrobacter freundii* CFNIH1 (CF), (**B**) *Serratia marcescens* WS1359 (SM), (**C**) *Salmonella enterica* serovar Gallinarum 287/91 (SE), (**D**) *Hafnia alvei* DSM30097 (HA). Nucleotides changed are all within the same region (amino acid 22-23) and either introduce a stop codon (orange bold) or a synonymous mutation (violet bold) for a better validation of the peak height ratio of wild type: mutant obtained from Sanger sequencing.

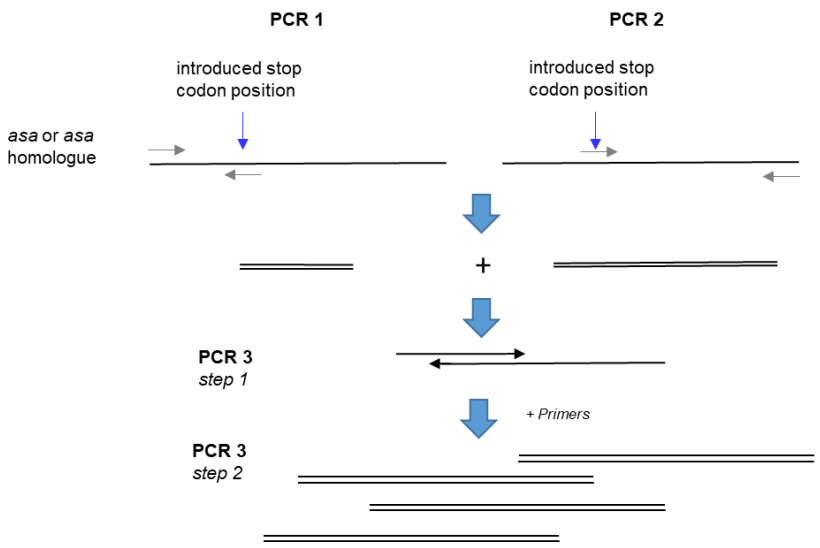


Figure 2.4: Scheme showing all steps required for the construction of translational arrested mutants. Grey arrows show the primers.

2.6.4 Restriction enzyme digestion and ligation

The ligation was prepared by double restriction enzyme digestion of all ORFs and vectors. The reaction (Table 2.6) was incubated for 3.5 h at 37°C. All restriction enzymes used in all cloning procedures were obtained from Thermo Fisher Scientific and can be found in Supplementary table S5. The restriction enzymes were inactivated by immediate purification of the inserts with the GenElute™ PCR Clean-Up Kit (Sigma Aldrich). The linearized vector was cleaned from the cut out fragment by 1% agarose gel electrophoresis and purified with the GenElute™ Gel Extraction Kit (Sigma Aldrich). The ligation of the insert into a vector (vector:insert ratio 1:3, calculated with Promega BioMath Calculator) was conducted with T4 ligase (Thermo Fisher Scientific) for 1 h at room temperature or overnight at 8°C. The reaction mixture can be found in Table 2.7. Heat inactivation of the ligase occurred for 20 min at 65 °C.

Table 2.6: Restriction enzyme digestion.

	PCR-Product	Vector
Enzyme 1	10 U	10 U
Enzyme 2	10 U	10 U
DNA	10 µl	1 µg
10x buffer	2 µl	2 µl
H_2O	ad 32 µl	ad 20 µl

Table 2.7: Ligation with T4 ligase.

DNA	variable
Vector	20-50 ng
10x buffer	2 µl
T4 DNA ligase (5 U/µl)	0.2 µl
H_2O	ad 20 µl

2.6.5 Electrocompetent bacteria

All bacteria tested were prepared for electroporation. First, 2 ml of a preculture was inoculated in 100 ml fresh LB medium and grown at 37°C, 150 rpm up to an OD_{600} of 0.4 - 0.6. The OD_{600} was measured in a Ultrospec 2000 UV / Visible spectrophotometer (Pharmacia Biotech, Sweden). The whole culture was harvested in two 50 ml tubes and cooled on ice for 10 min, centrifuged and the supernatant discarded. All centrifugation steps were conducted at 6000 rpm, 10 min and 4°C. For all subsequent steps, it was important to maintain a continuous cooling chain of the cells. Thus, all reagents and vessels were precooled. The pellets were resuspended in 50 ml sterile ice-cold water and centrifuged. After discarding the supernatant, the pellets were resuspended in 25 ml sterile ice-cold water, merged in one 50 ml tube and centrifuged. The pellet was washed once in 20 ml sterile ice-cold glycerol (10%, Merck) and centrifuged. Lastly, the pellet was resuspended in 10% sterile ice-cold glycerol, and aliquoted as 40 µl in 1.5 ml Eppendorf reaction tubes and immediately frozen in liquid nitrogen. The cells were long-term stored at -80°C.

2.6.6 Transformation by electroporation

For transformation, SOC medium was pre-warmed at 37°C, cuvettes were disinfected using UV-light for 10 min and pre-cooled on ice. The sample was desalted on a microfilter (0.025 µm, Merck Millipore) in MilliQ for 15-30 min. The competent cells were slowly thawed on ice. For electroporation, 1-5 µl ligation mixture or 1 µl plasmid was added to 40 µl electrocompetent cells, gently mixed and transferred air bubble free into electroporation cuvettes. After wiping the cuvette, it was pulsed in the Micropulser (BioRad, program: EC2, 2.5 kV, 1 - 4 ms, 10 µF). Immediately, 960 µl SOC medium was added and the cells were transferred into a fresh 1,5 ml reaction tube and incubated at 37°C, 150 rpm for 1 h. The cells were plated on LB with ampicillin (120 µg/ml) and incubated overnight at 37°C.

2.6.7 Verification of the correct insert

The correct insert length was checked by a colony PCR and subsequent 2% agarose gel electrophoresis. The colony PCR was conducted as described in section 2.4.3 with the following changes. The time of the initial denaturation was increased to 15 min to lyse the cells. The primers used here were vector specific (Supplementary table S3). Plasmids with inserts of the expected length were isolated from a 4 - 6 ml overnight culture with the GenElute™ Plasmid Miniprep Kit (Sigma-Aldrich). The plasmid was eluted with 30 µl H_2O. The concentration was

measured with a NanoDrop and the insert was sequenced by Sanger sequencing (Eurofins) by using one of the vector primers. After verification, the plasmids were transformed in the respective organism and stored in 40% glycerin at -80 °C.

2.6.8 Competitive growth experiments

The overexpression phenotype of *asa* and its homologues was tested in competitive growth assays. Bacteria having either pBAD *myc*/His C with intact insert or with translationally arrested mutant as insert, respectively, were grown competitively against the suitable counterpart. The combinations tested can be found in Table 2.8. Precultures of the bacteria containing an intact or translationally arrested insert were diluted to an OD_{600} = 1 with a final volume of 4 ml. The OD_{600} was checked and adjusted if necessary. The bacteria containing intact or translationally arrested insert were mixed in equal amounts. One part (5 µl) was diluted in LB-medium (1:300) and used for inoculation of the main culture (100 µl in 10 ml LB + antibiotics + 0.002% L-arabinose (w/v, Carl Roth, sterile filtered) + stress substance). The remaining part was centrifuged (14,000×*g*, 3 min) and the pellet stored at -20 °C (t_0). The tested stress substances were: 450 mM NaCl (Carl Rot, autoclaved) for all organisms and, additionally for EHEC, 20 mM L-arginine (Sigma Aldrich, sterile filtered) or 20 mM pyridoxine hydrochloride (Serva, sterile filtered). Plain LB medium was used as negative control. The main culture was incubated at 37°C, 150 rpm. After 6.5 h, 0.002% L-arabinose (w/v) was added for a second activation of the expression, because EHEC is able to digest L-arabinose. The plasmid was isolated from the cultures harvested at the time point t_0 and at the time point 22 h after induction with the GenElute™ Plasmid Miniprep Kit (Sigma Aldrich). The insert was sequenced with Sanger sequencing (Eurofins). The ratio of wild type and mutant determined by the peak height of the mutated bases. The peak height of the mutated bases was normalized to the peak height of the 23 nucleotides surrounding the mutated bases to exclude biases of the sequencing run. Each experiment was conducted in biological triplicate.

Table 2.8: Combinations of bacteria and plasmids tested in competitive growth experiments. The inserts tested are *asa* in EHEC, its intact homologues in *Serratia marcescens* (*asa*SM), in *Citrobacter freundii* (*asa*CF) and those homologues having a natural premature stop in *Salmonella enterica* (*asa*SE) or in *Hafnia alvei* (*asa*HA).

EHEC + pBAD-*asa*	vs. EHEC + pBAD-Δ*asa*	EHEC + pBAD-*asa*	vs. EHEC + pBAD-Δ*asa* alt ko
EHEC + pBAD-*asa*SM	vs. EHEC + pBAD-Δ*asa*SM	*S. marcescens* + pBAD-*asa*SM	vs. *S. marcescens* + pBAD-Δ*asa*SM
EHEC + pBAD-*asa*CF	vs. EHEC + pBAD-Δ*asa*CF	*C. freundii* + pBAD-*asa*CF	vs. *C. freundii* + pBAD-Δ*asa*CF
EHEC + pBAD-*asa*SE	vs. EHEC + pBAD-Δ*asa*SE	*S. enterica* + pBAD-*asa*SE	vs. *S. enterica* + pBAD-Δ*asa*SE
EHEC + pBAD-*asa*HA	vs. EHEC + pBAD-Δ*asa*HA	*H. alvei* + pBAD-*asa*HA	vs. *H. alvei* + pBAD-Δ*asa*HA

2.6.9 Quantitative PCR of reversely transcribed *asa* mRNA and of *asa* homologues

2.6.9.1 Growth conditions for RT-qPCR

To analyze the regulation of *asa* under sodium chloride stress conditions, EHEC was grown under stress-adapted conditions. Due to the slow growth in sodium chloride, overnight cultures were not grown in plain LB, as performed for phenotypic experiments, but were pre-adapted to the stressor. For this, the overnight cultures were grown in plain LB medium for two hours before addition of 450 mM sodium chloride. The next day, 500 µl of the overnight cultures were transferred into 40 ml fresh LB medium + 450 mM sodium chloride (control: cells grown in LB). An aliquot was harvested for RNA extraction at OD_{600} = 0.2 - 0.3 (early exponential phase) and at OD_{600} = 0.7 - 0.8 (exponential phase). The volume was adjusted to OD_{600} = 1 in 1 ml cell culture (≈ 8 x 10^8 cells). The regulation under stress shocked conditions was performed in EHEC precultures grown in plain LB medium, which are used for inoculation of the main culture (500 µl preculture in 40 ml fresh LB medium). The NaCl shock (450 mM) was conducted at an OD_{600} = 0.8. Aliquots were taken before induction and after 30 min, 60 min and 120 min. Plain LB medium was used as negative control. mRNA of *asa* homologues present in *Salmonella enterica*, *Hafnia alvei, Citrobacter freundii,* and *Serratia marcesces* was detected in cultures grown in plain LB medium and harvested at OD_{600} = 0.8. Aliquots were centrifuged in all cases (12,000×*g*, 3 - 5 min), the pellet was shock-frozen in liquid nitrogen and stored at -80°C until RNA isolation.

2.6.9.2 RNA extraction

The cell pellets were thawed on ice and resuspended in 1 ml RNAprotect (Qiagen). After vortexing (5 s) and incubation (room temperature, 5 min), the mixture was centrifuged for 10 min, 8,000×*g*. The resulting pellet was resuspended in 100 µl TE buffer supplemented with 0.4 mg/ml lysozyme. Subsequent steps were conducted with the SV total RNA Isolation System (Promega) according to the manufacturer's protocol. Instead of the DNAse provided by the kit, the TURBO™ DNase (Thermo Fisher Scientific) was used for DNA digestion for 1 h. The RNA concentration was measured with NanoDrop (Thermo Fisher). The absence of DNA was checked with PCR (Primer EHEC RNA: 8220+1F-*Nco*I, 8220+245R-*Hind*III; Primer homologues: 16S rRNA, Supplemental table S3) with Q5 High Fidelity DNA Polymerase (NEB) and by a 2% agarose gel. All samples were additionally tested to be DNA-free by qPCR on not reversely transcribed RNA using the 16S rRNA primers.

2.6.9.3 Reverse transcription and quantitative PCR

The regulation of the *asa* gene expression was detected by RT-qPCR. The bacterial RNA (1.6 µg) was reverse transcribed in cDNA with Supercript III (Thermo Fisher Scientific) according to the manufacturers protocol. The reaction was started using a random nanomer primer (50 µM, Sigma Aldrich) in presence of the Superase In RNAse Inhibitor (20 U, Invitrogen).

The amplicon of the qPCR was detected using Sybr Green as fluorophore implemented in the Sybr™ Select Master Mix (Thermo Fisher scientific). The reaction mix and conditions can be found in Table 2.9. A melting curve was measured at the end of each qPCR run starting at the annealing temperature to 95°C. The primers of *asa* (annealing temperature 61°C), *asa* homologues (annealing temperature 60°C), the normalizer 16S rRNA gene and the negative control (annealing temperature 58°C) are listed in Supplementary table S3 (*asa, control*) and S4 (homologues). The primer efficiency was tested with respective bacterial genomic DNA in the following concentrations: 400 ng, 40 ng, 4 ng and 0.4 ng DNA. The primer efficiencies can be found in Supplementary table S6 and the standard curves in Supplementary figure S2. All sequences were amplified in three technical replicates each qPCR run. Each condition was measured as biological triplicate. The negative control was tested in three technical replicates and one biological sample (RNA after growth in LB at exponential phase). The gene regulation was calculated with the ∆cq method (Pfaffl 2001).

Table 2.9: Reaction mix and reaction conditions used in qPCR.

Reaction approach:			Reaction conditions (40 cycles):		
Sybr™ Select Master Mix	12.5	µl	**Initial Denaturation**	95°C	5 min
50 µM forward Primer	0.5	µl	**Denaturation**	95°C	15 s
50 µM reverse Primer	0.5	µl	**Annealing**	var.	30 s
Template (gDNA or cDNA)	2	µl	**Elongation**	72°C	30 s
			Final Elongation	72°C	5 min
H_2O	9.5	µl			

2.6.10 Promoter activity with pProbe-NT

A region 160 bp upstream of the start codon with a length of 92 bp was chosen to test the activity of a putative promoter upstream of the +1 site. The tested fragment cloned in pProbe-NT (Miller, et al. 2000) following the steps described in section 2.6.1 - 2.6.7. A 76 bp long fragment in the terminator region of EDL933_1236 was cloned likewise as negative control of the promoter region (NC I). The correctly cloned insert sequences were confirmed by Sanger sequencing (Eurofins). The promoter activity tests were conducted from overnight cultures of *E. coli* Top10 + pProbeNT-insert (negative control of the induction of GFP expression, NC II: pProbeNT without insert). These were inoculated (1:100 dilution) in 10 ml medium + 50 µg/ml kanamycin and grown at 37°C, 150 rpm. Media: plain LB, LB + 450 mM NaCl, LB + 10 mM L-arginine, LB + 10 mM pyridoxine hydrochloride. The concentrations of L-arginine and pyridoxine hydrochloride were reduced in comparison to those in competitive growth experiments due to the slow growth of the cells at higher stressor concentrations. The cells were harvested at OD_{600} = 0.8. The pellet was washed with phosphate buffer (PBS) and resuspended in 1 ml PBS. The OD_{600} was adjusted to 0.3 and 0.6. The fluorescence was measured in black mitrotiter plates (four technical replicates, 200 µl volume) in Wallac Victor^3 (Perkin Elmer Life Science, excitation 485 nm, emission 535 nm, measuring time 1s). The background activity (Top10 without vector) was subtracted from measured values. Each condition was measured in biological triplicate.

2.6.11 Transcriptional start site (+1 site) by 5' RACE and Cappable seq

The 5' end of *asa* mRNA was determined by 5' RACE and Cappable seq. The 5' RACE was conducted in two experiments. In the first one, intrinsic EHEC mRNA was detected after growth in LB medium. Due to the detection of the high promoter activity in NaCl, the 5' RACE experiment was repeated with *E. coli* TOP10 expressing the *asa* promoter-GFP in pProbe-NT (section 2.6.10) while growing in LB + 450 mM NaCl. Both experiments were performed using the following protocol. Precultures grown in plain LB were inoculated 1:300 in fresh medium and

incubated up to an OD_{600} of 0.8. An aliquot of 500 µl was taken for total RNA extraction as described in section 2.4.9.2. The +1 site was detected using the 5'/3' RACE Kit, 2nd Generation (Roche). The dominant PCR products were excised from the agarose gel, purified with the GenElute™ Gel Extraction Kit (Sigma Aldrich) and Sanger sequenced. All primers used can be found in Supplementary table S3. The Cappable Seq experiments were conducted in three biological replicates by Barbara Zehentner following Ettwiller, et al. (2016).

2.6.12 Western Blot

The *asa* encoded protein, Asa (expected size: 10 kDa), was detected by Western Blot using an antibody which binds a C-terminal SPA-tag (Zeghouf, et al. 2004, expected size: 7.5 kDa). The experiment was planned by Barbara Zehentner in the context of testing further OGCs, but conducted by me. The composition of all buffers and solutions used can be found in Table 2.3. First, the sequences encoding *asa* without stop codon and the SPA tag were cloned in pBAD-*myc*/His C as described in section 2.6.1 - 2.6.7. The SPA encoding insert was cloned by merging two fragments (synthesized by Eurofins) by PCR to 'PCR3' of the construction of translationally arrested mutants (Figure 2.4). Asa-SPA was overexpressed in *E. coli* TOP10. For this, 100 µl of a preculture grown in LB + 120 ng/µl ampicillin was inoculated in 10 ml fresh LB + 120 ng/µl ampicillin (25 ml flask) and grown up to an OD_{600} of 0.3. Aliquots were taken 0.5 h, 1 h, 1.5 h, 2 h, 3 h and 4 h after induction with 0.002% arabinose (w/v). The volume was adjusted to receive the same OD_{600} as at time point 0.5 h. After centrifugation (16,000×*g*, room temperature, 2 min), the cell pellet was boiled in 50 µl 3x sample buffer (compounded after Tricine Sample Buffer, BioRad) at 95°C for 10 min.

A Tris-Tricine SDS PAGE (modified after Schägger 2006) was chosen due to the higher resolution at separating smaller proteins in comparison to a classical SDS-PAGE. The PAGE was poured as 16% running gel and 4% stacking gel (Table 2.10). The electrophoresis was performed in the Mini-Protean® Tetra Vertical Electrophoresis Cell (BioRad) using a 1x cathode buffer and a 1x anode buffer at 35 mA per gel for 2-3 h. The protein marker used was the Spectra Multicolor Low Range Marker (Thermo Fisher Scientific, range 1.7 - 40 kDa). The positive-control was pBAD-*gst*-SPA expressing the glutathione S-transferase of 22 kDa.

The gel was prepared for Western Blot (modified after Baek, et al. 2017a) by incubation in 1x blotting buffer for 10 min. The polyvinylidene difluoride (PVDF) membrane (Immobilon-PSQ, 0.22 µm, Merck) was incubated in 100% methanol for 15s, in H_2O for 5 min and in 1× blotting buffer for 10 min. Blotting occurred in the SemiDry Blotter Pegasus (Phase GmbH) at 12 V for 20 min. After washing the PVDF membrane with 3% trichloroacetic acid (TCA, Carl Roth) in H_2O

for 5 min, it was incubated overnight (4°C, shaking) in skim milk powder (1.25 g in 25 ml 1× TBS-T). A further three washing steps (in 1× TBS-T, respectively 10 min) prepared the membrane for the antibody reaction, which was as follows: 10 ml TBS-T + 10 µl Anti-SPA antibody (Anti-FLAG® MP2-Alkaline Phosphatase Clone M2, Sigma Aldrich) were mixed and shaken for 1 h at room temperature. The membrane was washed six times with TBS-T, each 5 min. After a washing step with reaction buffer for 5 min, the protein bands were visualized using a solution that consists of 100 µl NBT, 125 µl BCIP, and 10 ml reaction buffer. When bands appear after 10-60 s, NBT/BCIP can be removed and the reaction is stopped with 3% TCA.

Table 2.10: Composition of SDS-Tris Tricine polyacrylamide gels.

16% running gel:		4% stacking gel:	
4.8 ml	acrylamide bisacrylamide 40% (Carl Roth)	0.6 ml	acrylamide bisacrylamide 40%
4.05 ml	3× gel buffer	1.5 ml	3× gel buffer
1.2 ml	glycerin (85%, Merck)	3.9 ml	H_2O
1.95 ml	H_2O	9 µl	TEMED
8 µl	TEMED (Carl Roth)	90 µl	APS 10%
80 µl	APS (10%, Carl Roth)		

2.6.13 Transcriptomes and translatomes

All RNAseq and RIBOseq data were visualized in Artemis 17.0 (Carver, et al. 2011, sum signal over one to three biological replicates) to identify expression profiles of *asa* and homologues in Enterobacteria. *asa* was originally discovered to be transcribed and translated in EHEC EDL933 by Landstorfer (2014). Further data from our lab were added (EHEC Sakai, Hücker, et al. 2017) and the pathogen *E. coli* LF82 (obtained by Michaela Kreitmeier and Franziska Giehren, unpublished). All remaining data were downloaded from the Sequence Read Archive, SRA, NCBI. All data are listed in Supplementary table S7. The data were processed with the help of Zachary Ardern. For this, reads were trimmed using Fastq with poly-x trimming and without a quality threshold (Chen, et al. 2018), and aligned using Bowtie 2 (Langmead and Salzberg 2012) using a seed length of 19 and zero mismatches in the seed. The homologs of *asa* in each genome examined here were detected using the DIAMOND search algorithm (Buchfink, et al. 2014). Normalized reads per kilobase per million sequenced reads (RPKM) were calculated using the Bedtools coverage tool (Quinlan and Hall 2010).

2.6.14 Databases and bioinformatics tools

Homologous proteins were searched using blastp (NCBI, refseq database, E-value cutoff ≤ 10^{-10}) and HHblits (default parameters, database uniclust30_2017_10). Protein family domain search was implemented in PredictProtein (Rost and Liu 2003) using the Pfam database (Finn, Coggill, et al. 2015) and conserved domains were searched for in the CD database (Marchler-Bauer, et al. 2016). Protein features of putative *asa* encoded proteins were predicted with PredictProtein using the programs listed in section 2.2.2.

The sequence 300 bp upstream of the transcriptional start site was searched for a σ^{70} factor binding sites with BPROM (Softberry, Solovyev and Tatarinova 2011). The region 600 bp downstream of the stop codon was searched for ρ independent terminators with FindTerm (Softberry, Solovyev and Salamov 2011). Shine-Dalgarno sequences were calculated within a range of 30 bp upstream of the start codon after Ma, et al. (2002). The Gibbs free energy cutoff used was - 18.42 kJ/mol.

2.6.15 dN/dS

The dN/dS analysis was performed in cooperation with Chase Nelson (American Museum of Natural History, New York). Nucleotide sequences of *asa* and its mother gene, along with 40 other homologues (*n* = 41 sequences), were obtained from NCBI. As these data were obtained from a preliminary species tree, the sequences analysed were less than those in the phylostratigraphic analysis (section 3.3). We decided not to repeat the analysis with the later complete tree, because the results obtained here were sufficiently clear enough. It can be safely presumed that they are not significantly better when using more sequences. For the mother gene, sequences were translated and aligned at the amino acid level using the ClustalW algorithm in Mega6, with default settings. The amino acid alignment was then back-imposed on the nucleotide sequences to maintain intact codons. Approximately half of the sequences terminated at codon 88, therefore, the remaining 3'-proximal codons were excluded from analysis. The region spanning the sORF sequence contained no indels. The overlapping genes homologous to *asa* were then aligned in the same way by taking the reverse complement of the mother gene sequences and extracting the sORF. Out of 41 sequences, 27 contained a premature stop at alignment codon 62 (nucleotide sites 184-186). Sequences thus constituted two segments for the sORF, one preceding and including the premature stop (nucleotides 1-186, 62 codons), and the remainder (nucleotides 187-264, 26 codons).

To determine the evolutionary constraint of *asa* and its mother gene, nonsynonymous (d_N) and synonymous (d_S) substitution rates were estimated among the 41 homologous sequences as follows. Separately for codon alignments of the mother and sORFs, numbers of nonsynonymous and synonymous nucleotide differences and sites were estimated with the Nei-Gojobori method (Nei and Gojobori 1986) using the `snpgenie_within_group.pl` script of SNPGenie (Nelson, et al. 2015, https://github.com/chasewnelson/snpgenie>). Codons were numbered to maximize codon overlap between mother and sORF, the first two nucleotides of codon 1 of the sORF overlapped the first two nucleotides of codon 190 of the mother gene (opposite strand). Henceforth, codon positions refer to the mother alignment. Sliding windows of d_N and d_S for the mother gene and sORF were performed with a window size of 10 codons, step size of 1. Differences between d_N, d_S, and their ratio were tested with a *Z*-test, where the standard error was calculated using 1,000 codon-based bootstrap replicates (Nei and Kumar 2000).

3 Results

3.1 Selection of potentially functional shadow ORFs (sORFs)

3.1.1 Evidence for functionality: homology to annotated genes

The procedure applied to identify potentially functional overlapping genes is depicted in Figure 3.1.1. In a first step, all open reading frames of *E. coli* O157:H7 strain EDL933 (Genbank CP008957.1) with a minimum length of 93 base pairs (bp) were extracted from the genome. From these, the 5,675 annotated ORFs (aORFs) were separated. Then, all non-annotated overlapping ORFs (minimum length of the overlapping part ≤93 bp) were kept. Annotated genes were used in the analyses as a positive control for a gene-like comparison set.

The EHEC genome has 49,649 overlapping ORFs according to the above definition. Their average length with standard deviation (190 ± 131 bp ≡ 63 ± 44 amino acids, aa) is significantly lower than that of annotated genes (888 ± 705 bp ≡ 296 ± 235 aa). Gene overlap can be either partially or embedded (Figure 1.2). Most of these overlapping ORFs are completely embedded (85% of all overlapping ORFs) in their mother gene (mORF, the annotated overlapping gene). For 320 overlapping gene pairs, the mother gene is embedded in the overlapping ORF. The remaining gene pairs do partially overlap. Overlaps in antisense are 67% (33,460 ORFs) of all overlapping ORFs in EHEC EDL933.

Probably, most of these overlapping ORFs do not encode proteins. However, comparative genomics may provide information about the presence of proteins in the largest contemporarily available sequence database, Genbank, which are homologous to some overlapping ORFs of EHEC EDL933. Such putative overlapping genes are more likely to encode proteins. Putative functional overlapping ORFs were detected by using blastp to probe the NCBI refseq database. The refseq database contains well-annotated protein coding sequences and is frequently updated and corrected (O'Leary, et al. 2015).

The blastp homology search resulted in 4,479 overlapping ORFs which had homologous proteins in the database. This number was reduced by removing mobile elements from the dataset since transposable elements may blur the picture of evolution (Cerveau, et al. 2011). Bacteriophage sequences, which are known (or at least accepted) to harbor overlapping genes (Chirico, et al. 2010), were also removed. Next, *sense* overlapping ORFs were removed, because these overlapping ORFs cannot be distinguished from their mother gene in RIBOseq data. In total, 2,180 antisense overlapping ORFs were identified (Supplementary table S1), which are called shadow ORFs (sORFs, Yooseph, et al. 2007). Of those, 10% are lying in -3 frame to the mother gene, 28% in -2 frame and 62% in -1 frame. The removal of mobile

elements from annotated genes results in 5,010 aORFs. Shadow ORFs and aORFs were used for the subsequent analyses presented in the sections 3.1 - 3.3.

There are 172 sORF with putative predicted function (all listed in Supplementary table S8). Their homologues are enzymes (90), membrane proteins and transporters (44), transcriptional regulators (17) and other proteins (21) including chaperones, cold-shock protein or elongation factor TU. However, the majority of sORFs (92%) are homologous to 'hypothetical proteins', which have not yet been further characterized. The number of annotated genes characterized as hypothetical is considerably lower (30%).

Next, it was investigated which sORFs are matching over their complete length to their homologues found by the blast analysis. As complete matches are expected in closely related species, the blastp search was restricted to hits within enterobacteriaceae. The nr database (February 2016) of blastp was used, because the refseq database does not contain any information about genome sequences in which the hit is found. A complete match to a protein, obtained by blastp, was defined to have a query coverage of ≥80%. Out of 2,180 sORFs, 280 had complete matches to at least one protein entry in the database (Supplementary table S9). In September 2018, a re-blast of these 280 sORF proteins confirmed 225 sORFs with at least one full-length hit (Supplementary table S10). The remaining 1,900 shadow ORFs matched only partly and, thus, always to longer proteins in the database. Of the 280 sORFs, 88% of the homologues are annotated as hypothetical protein. The remaining 17 proteins match to membrane or cell surface proteins, enzymes, chaperones, heat-shock protein GrpE or the T3SS protein SepZ.

It is surprising that sequences encoded by a sORF are homologous to other annotated proteins, and, thus, assumedly functional, proteins. What happened to the respective mORF homologue? Is it still intact? Are there genomes in which both genes are annotated? A small test set to examine such questions of 50 sORFs with a match of the complete sequence length were randomly selected, the overlapping gene pairs downloaded and checked for annotation and intactness by visualization in Artemis and pairwise sequence alignment. Models for representative cases are visualized in Figure 3.1.2.

In case of an intact mORF, both overlapping gene pairs can be annotated (Figure 3.1.2 A). Particularly, the genome of *E. coli* CFT073 (AE014075.1) has an unusually high-cumulated number of overlapping genes. A further case observed was that the mORF sequence is intact, but not annotated (Figure 3.1.2 B) or a start codon was used which prevents non-trivial overlaps (Figure 3.1.2 C). Many badly sequenced genomes contain matches to full-length sORFs. In those cases, the mORFs are more often than not cut by an assembly gap (e.g., contig border). Therefore, it is most likely that the mORF is also intact but has eluded annotation due to the

contig border. This was observed in *Enterobacter cloacae* (Figure 3.1.2 D), but also in *Shigella* genomes. In other cases, the mORF is somehow disrupted, it was observed that parts of the sequence were "exchanged" (Figure 3.1.2 E, F) or disintegrated by frameshifts (Figure 3.1.2 F, G). Some mORFs had internal stops in an otherwise highly conserved sequence (Figure 3.1.2 H). Significant sequence differences in homologous overlapping gene pairs in comparison to the gene pair in EHEC were observed not only in species far related to EHEC, but also in *E. coli* strains different from EDL933 and in closely related species.

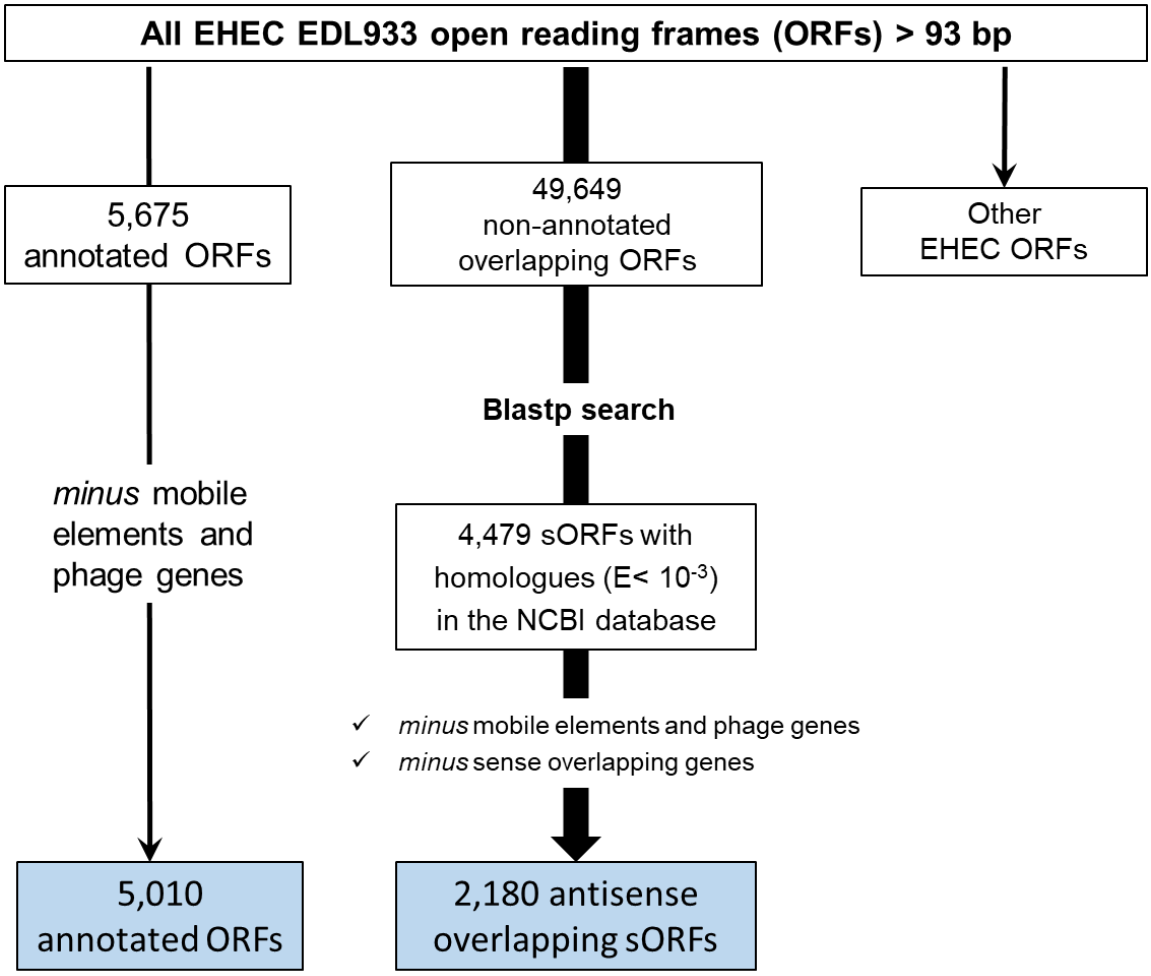


Figure 3.1.1: Overview of all steps applied to identify potentially functional overlapping genes in EHEC EDL933. This unpublished analysis has been performed by Svenja Simon (University of Konstanz).

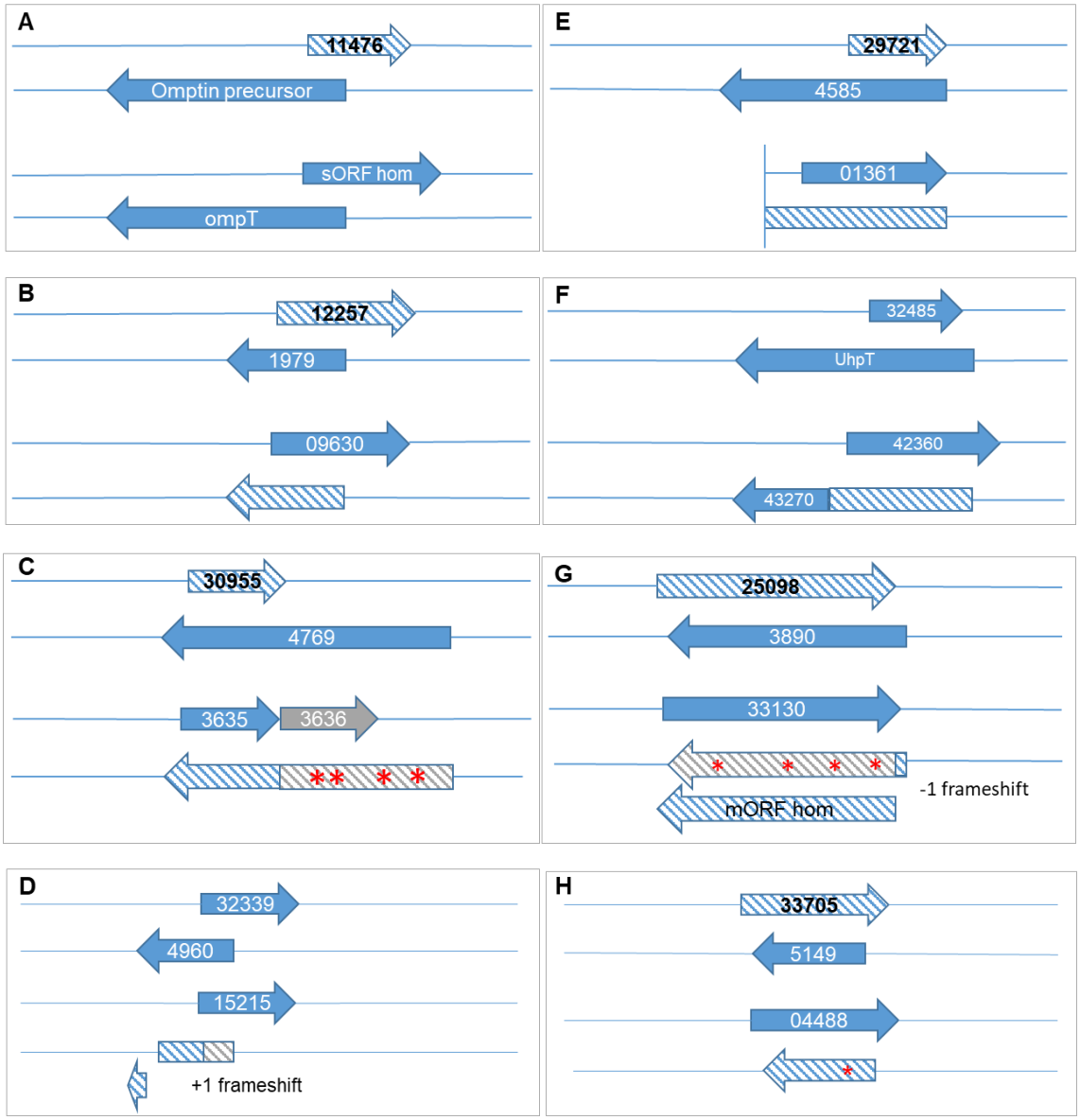


Figure 3.1.2: Model showing the annotation status and conservation of selected overlapping gene pairs in which the sORF completely matches to a protein homologue. Homologues were identified by blastp against *Enterobacteriaceae* (nr, NCBI). Legend: Each box represents one overlapping gene pair in EHEC EDL933 (top) and its farthermost related homologous counterpart to with the sORF sequence completely matches (below). Genes are marked with an arrow and named after the locus tag number or the gene name in case of annotation; annotated or homologous genes are marked in blue, non-annotated homologous regions are blue-hatched, annotated non-homologous regions are grey and non-annotated non-homologous regions are grey-hatched; stop codons are red asterisks. (A) OLG #11476 with homologue in a not further specified *E. coli* sequence (Genbank U82598.1), (B) OLG #12257 with homologue in *E. coli* OLC1128 (Genbank NWSA01000208.1), (C) OLG #32485 with homologue in *Klebsiella pneumonia* DSM30104 (Genbank AJJI01000016.1), (D) OLG #29721 with homologue in *Enterobacter cloaceae* (Genbank FKGQ01000677.1), (E) OLG #30955 with homologue in *Shigella flexneri* ATCC12022 (Genbank JPPN01000167.1), (F) OLG #32339 with homologue in *E. coli* VL2604 (Genbank MIXF01000051.1), (G) OLG #25098 with homologue in *Klebsiella pneumonia* DSM30104 (Genbank AJJI01000014.1), (H) OLG #33705 with homologue in *E. coli* H588 (Genbank ADIQ01000012.1).

3.1.2 Evidence for functionality: Conserved domains and ribosomal footprints

Prosite patterns are conserved amino acid patterns found in protein domains. They are used to identify functional proteins independent of information on sequence homology about those proteins (Sigrist, et al. 2002). In both proteins encoded by annotated ORFs and those suspected to be encoded by shadow ORFs, the abundance of proteins with prosite patterns is very high (95%). However, the variety of different pattern found was much higher in annotated ORFs (689 different patterns) compared to shadow ORFs (25 different patterns). Conserved domains (CDs) were identified in 44 shadow ORF-proteins and they are listed in Supplementary table S11. The majority of CDs are domains of unknown function. Eight specified CDs correspond to the function the protein found in the blast hit. They are transcriptional regulators (#11769), enzymes (#14354: restriction endonuclease, #23719 and #76450: glutamate dehydrogenase, #46888: DNA topoisomerase, #59154: prolyl-tRNA synthetase, #63696: trehalase) and transporter/membrane associated proteins (#32339: T3SS protein SepZ, #32417: sugar transporter). Interestingly, six hypothetical proteins can be specified by their CD (#18076: lipoprotein, #23725: glutamate dehydrogenase, #33739: synthetase, #57774: signal transduction protein, #8702: GnaA/GnaB family, #37939: Topoisomerase DNA C4 zinc finger). The finding of a CD might be a first hint for the function of these six proteins.

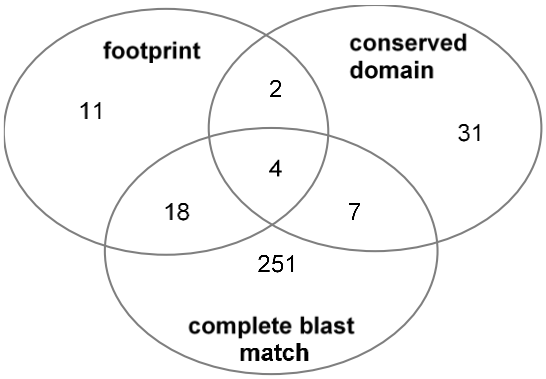


Figure 3.1.3: Venn diagram of shadow ORFs with ribosomal footprints, conserved domains and blast hits in which the sORF completely matches to the homologue.

The sORFs found in *E. coli* O157:H7 str. EDL933 having a blastp hit were checked for ribosomal footprints in the very closely related strain *E. coli* O157:H7 str. Sakai. RIBOseq was conducted by Sarah Hücker with EHEC grown in three different conditions (LB medium, BHI medium, BHI in 4% NaCl and grown at 14°C) as described in Hücker, et al. (2017). The sum signal of all replicates and conditions was used. Interestingly, 35 shadow ORFs had conspicuous patterns of ribosomal footprints. Fifteen of these were annotated in Sakai (but not in EDL933). The presence of ribosomal footprints strongly suggests the mRNA is translated and, therefore, the sORF codes for a protein.

Shadow ORFs with ribosomal footprints, conserved domains or a complete blast match are strong candidates for novel protein coding genes. However, such candidates should ideally be phenotypically tested to ensure the proteinaceous nature of the gene product. The number of

ORFs matching all three features is surprisingly low (Figure 3.1.3) and is given in Table 3.1.1, the respective ribosomal footprints are visualized in Figure 3.1.4.

Shadow ORF #11769 putatively encodes a hypothetical protein annotated in several *E. coli* strains including strain Sakai (ECs1706; Figure 3.1.4 A). Predicted protein features, like secretion, 47% disordered amino acids and a DNA-binding transcriptional regulator domain are further characteristics of the results obtained by the blast search. The shadow ORF was assigned to phylostratigraphic level “Enterobacteriaceae” and the mORF (EDL933_1905) is a high conserved gene involved in adhesion and penetration.

Shadow ORF #30413 putatively encodes a highly conserved hypothetical protein. The shadow ORF is embedded in the also highly conserved mother ORF (EDL933_4676), an ATP-binding protein FtsE, which is involved in cell division. The shadow ORF is not annotated in Sakai, but it has a clear signal for ribosomal footprints (Figure 3.1.4 B).

Shadow ORF #39433 possibly codes for a taxonomically restricted hypothetical protein only found in *E. coli* with an unspecified conserved domain (PRK09719). The shadow ORF overlaps with a UPF0131 protein YtfP encoding ORF (EDL933_5570) assigned to phylostratigraphic level “Bacteria/Archaea”. The shadow ORF of EDL933 is annotated in Sakai (ECs4757) as well as the mother ORF. The mother ORF has a strong signal for ribosomal footprints in Sakai and the shadow ORF has a weak signal (Figure 3.1.4 C).

Shadow ORF #74629 encodes a taxonomically restricted hypothetical protein. The shadow ORF is embedded in a core protein (EDL933_0244) which is highly conserved among bacteria and archaea. Both shadow ORF and mother ORF have a weak signal of ribosomal footprints in Sakai (mORF: ECs0242, sORF not annotated, Figure 3.1.4 D).

Table 3.1.1: Examples of shadow ORFs with conserved domain and ribosomal footprints.

shadow ORF ID	11769	30413	39433	74629
Phylostratum	enterobacteriaceae	Bacteria/Archaea	E. coli	E. coli
Blast hit (farthest related ancestor)	ref| WP_010917801.1| hypothetical protein [Escherichia coli]]	ref|WP_051876519.1| hypothetical protein [Cellulosimicrobium sp. MM]	ref|WP_063502666.1| hypothetical protein [Escherichia coli]	ref|WP_044710606.1| hypothetical protein [Escherichia coli]
mother ORF	AIG68093.1| Putative adhesion and penetration protein	AIG70821.1| Cell division transporter, ATP-binding protein FtsE	AIG71709.1| UPF0131 protein YtfP	AIG66450.1| core protein
Pfam/CD	PRK11388 DNA-binding transcriptional regulator DhaR; Provisional	PF07673.9 - Protein of unknown function (DUF1602);	PRK09719 hypothetical protein; Provisional	-
Length [bp]	361	186	435	261
Transmembrane helices	0	0	0	0
Disulfide bonds	1	0	0	1
% Disorder	47	100	66	100
Helices/ Beta sheet/ Loop [%]	45.6/ 5.7/ 48.7	0/ 29.5/ 70.5	1.4/ 25.7/ 72.9	0/ 2.33/ 97.7
Buried/ Exposed/ Intermediate [%]	34.8/ 43.7/ 21.5	11.5/ 57.4/ 31.2	25.7/ 46.5/ 27.8	1.2/ 86.0/ 12.8
Protein binding sites	27	29	55	40
Localization	secreted	secreted	secreted	secreted
MF/BP	yes/ yes	no/no	no/no	yes/no

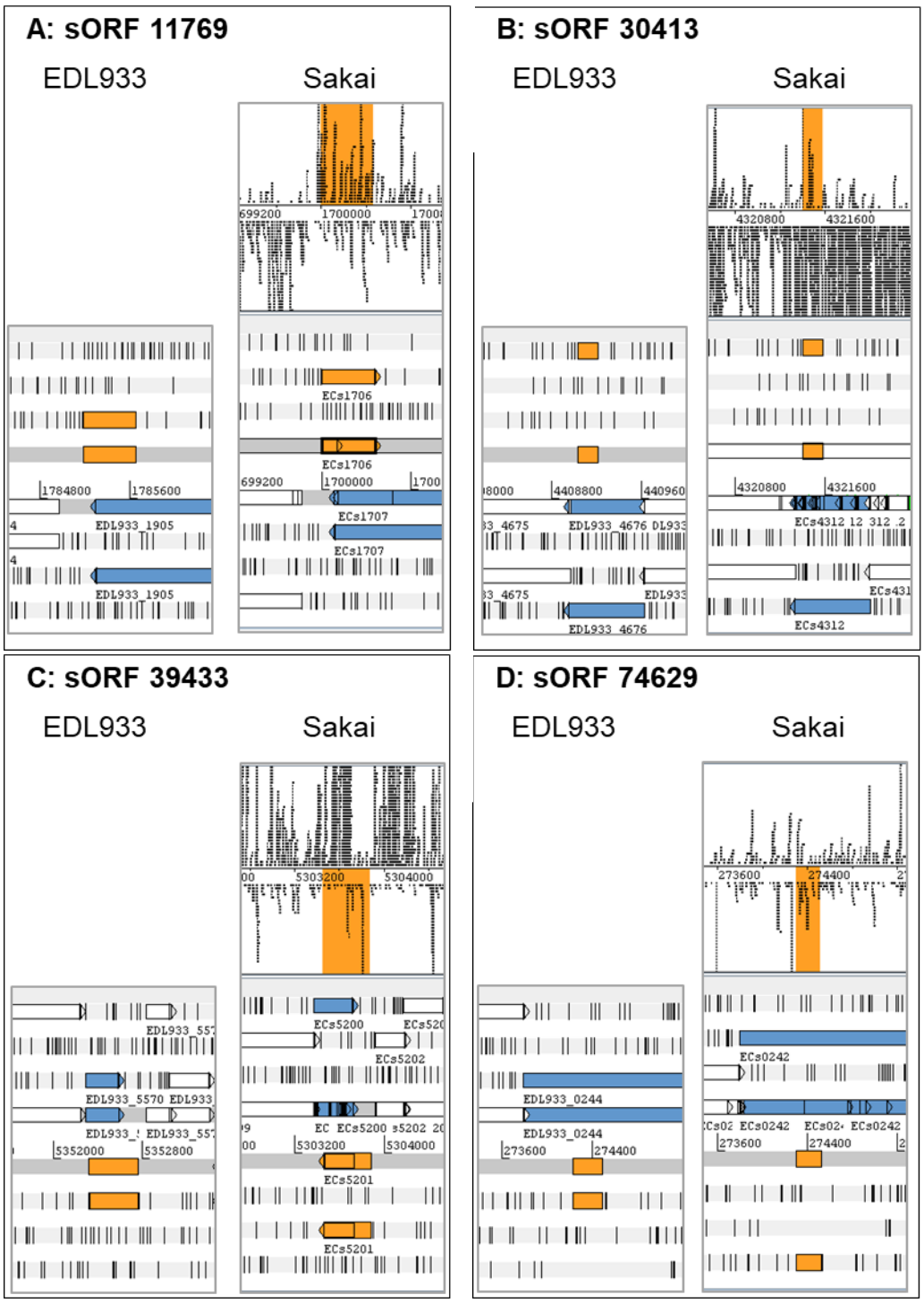


Figure 3.1.4: RIBOseq signals of the the sORFs shown in table 3.1.1 and the respective mORF sequences from E. coli EDL933 (CP008957.1) and Sakai (NC_002695). The sORFs and sORF homologues are highlighted in orange, mORF and mORF homologues in blue. Sequences were visualized with Artemis 17.0 (Carver, et al. 2011).

3.1.3 Taxonomic distribution of antisense sORFs

The sORFs and aORFs were classified according to their phylostratigraphic level. It is hypothesized that sORFs are younger than aORFs, because annotation programs typically detect the more conserved gene (Delcher, et al. 2007). In this analysis, only the blastp hit was used, which identifies a homologous protein in the species farthermost related to EHEC EDL933. An E-value cutoff was used as protein homology cutoff. Figure 3.1.5 shows the number of sORFs per E-value. The distribution of sORFs and aORFs are very similar. The aORF dataset is slightly shifted in direction to very low E-values indicating highly significant hits. Consequently, the average E-value of the sORF dataset is higher (7.0xE-05 ± 9.2xE-05) than that of the aORF dataset (7.7xE-06 ± 1.3xE-05), but both are of similar magnitude.

The phylostrata estimating the gene age were defined according to the similarity of the 16S rRNA gene between EHEC and the organism in which the farthest blast hit of the gene of interest was found (Figure 2.1). The first phylostratum ('E. coli') contains all *E. coli* strains, hence all genes are taxonomically restricted to the species (TRGs). In this phylostratum, 47% of the sORF homologues, but only 4% of aORF homologues are TRGs (Figure 3.1.6). In major contrast, most aORFs (78%) were assigned to the phylostratum 'Bacteria/Archaea', which contains all highly conserved genes that can also be found in bacteria and archaea. Of the sORF homologues, only 25% were highly conserved. Quite a few genes from each category were assigned to the phylostratum 'Enterobacteriaceae' (14% sORFs, 8% aORFs). The remaining phylostrata 'Enterobacteriales' and 'γ-proteobacteria' contain only low numbers of ORFs. Thus, the majority of ORFs is either young or very old with missing genes in the middle-aged phylostratigraphic levels.

The majority of sORFs assigned to a function (not 'hypothetical') are taxonomically restricted to *E. coli* (60%, Table 3.1.2), 17% are highly conserved and assigned to the level 'bacteria/archaea'. Of those having only hits to hypothetical proteins, 46% were TRGs, but a surprisingly high number of highly conserved genes are still uncharacterized (26%). A re-blast of sORFs with hits to predicted functions, conducted three years later, confirmed the overall phylostratigraphic trends with a small shift towards older phylostrata (Supplementary table S12). The annotated genes predicted to have a known function are predominantly highly conserved (90% are in the phylostratum "Bacteria/Archaea"), while about half of the uncharacterized aORFs were assigned to this phylostratum.

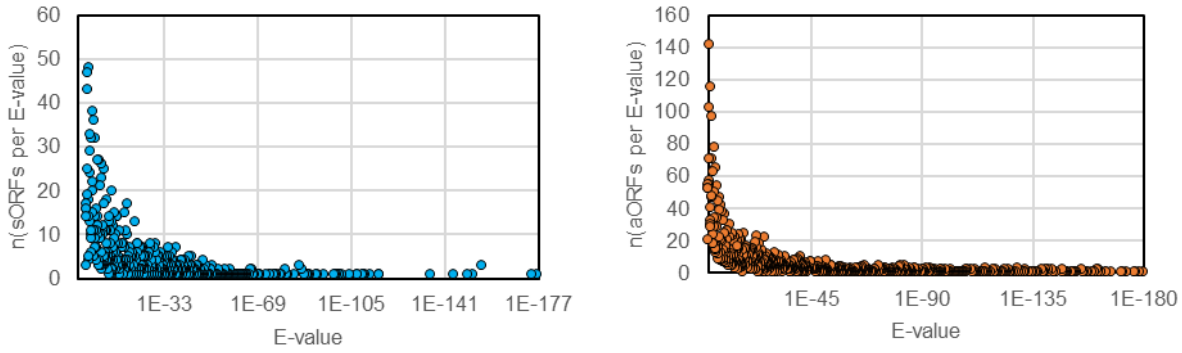


Figure 3.1.5: Number (n) of open reading frames with one particular E-value. The E-value results from all blastp hit of sORFs (blue; 2,180) and aORFs (orange; 5,010) in which the homologue of the farthermost related species was found.

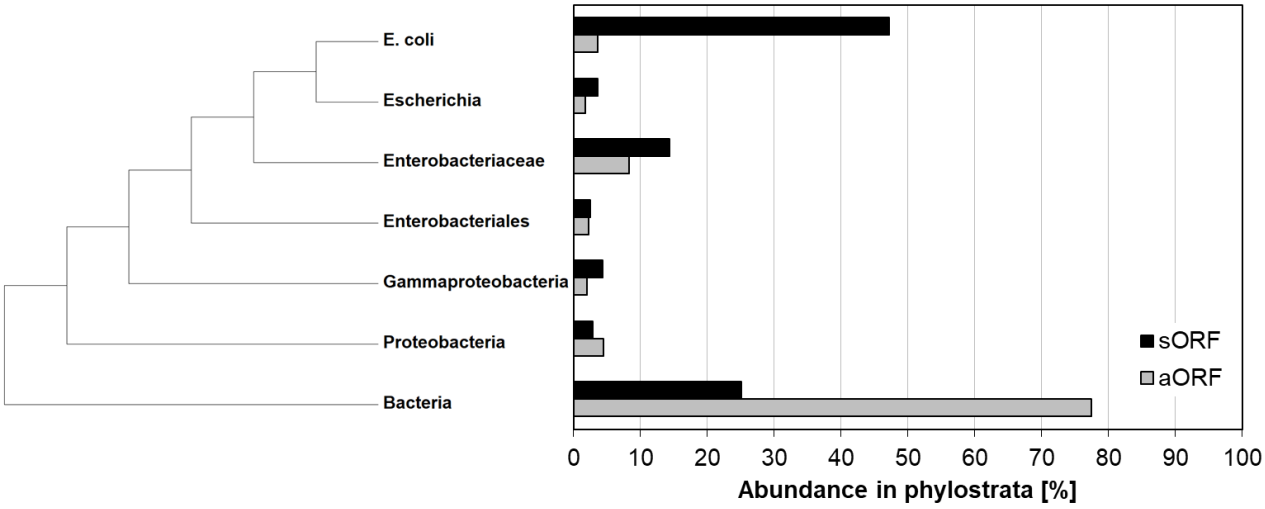


Figure 3.1.6: Phylostratigraphic distribution of all sORFs (■) and aORFs (□) with blastp hit.

Table 3.1.2: Distribution of shadow ORFs and annotated ORFs within phylostratigraphic levels. Left columns: genes with blastp hit of a predicted function; right columns: genes with blastp hit named as 'hypothetical protein'. The distribution of hypothetical proteins in a particular phylostratigraphic level is shown as percentage of all hypotheticals.

	predicted function				"Hypothetical"			
	shadow ORFs		annotated ORFs		shadow ORFs		annotated ORFs	
E. coli	**101**	**(58.7%)**	23	(0.7%)	**928**	**(46.2%)**	156	(10.3%)
Escherichia	4	(2.3%)	17	(4.9%)	75	(3.73%)	73	(4.8%)
Enterobacteriaceae	29	(16.9%)	128	(3.7%)	285	(14.2%)	290	(19.2%)
Enterobacteriales	3	(1.7%)	36	(1.0%)	50	(2.5%)	80	(5.3%)
γ-Proteobacteria	1	(0.6%)	32	(0.9%)	94	(4.7%)	68	(4.5%)
Proteobacteria	4	(2.3%)	103	(2.9%)	58	(2.9%)	121	(8.0%)
Bacteria/Archaea	**30**	**(17.4%)**	**3160**	**(90.3%)**	**502**	**(25.7%)**	**724**	**(47.8%)**
total	**172**	**(7.9%)**	**3499**	**(69.8%)**	**2008**	**(92.1%)**	**1511**	**(30.2%)**

3.2 Comparison of structural features of sORFs and annotated proteins

Protein features of sORF and aORF proteins were predicted with PredictProtein. These features were analyzed in dependence on the phylostratigraphic level of the ORFs. As protein features can be dependent on the sequence length, which would falsify the results of the phylostratigraphy, in which proteins were pooled together regardless of length, it was first analyzed how the protein features differ according to protein length. First, the number of ORFs having a particular length were plotted (Figure 3.2.1). It is evident that the overall length distribution of all sORF proteins is significantly shifted to lower values in comparison to the aORF proteins, which is also reflected by the average sequence length (sORF proteins: 120 ± 94 aa, aORF proteins: 296 ± 235 aa = 888 ± 705 bp). The amino acid composition of each sORF or aORF was determined with PredictProtein. There are only a few amino acids that may have different abundances in sORFs in comparison to aORFs. Shadow ORFs may have more methionine and lysine, but less arginine or leucine, but the differences are not significant. The composition of all remaining amino acids certainly is comparable in both datasets (Figure 3.2.2), which is not surprising as the sORFs had homologous proteins identified by blast searches.

The proteins encoded by sORFs and aORFs were assigned to one of eight length classes. The aORF proteins have a peak at a length of 101-300 aa (Figure 3.2.3 A). While most sORF-encoded proteins are shorter than 200 aa, the longer classes contain only a few proteins per class. The lowest abundance is present in proteins longer than 500 aa, with 23 putative sORF

encoded proteins in this class. The overall length distribution is more balanced for the aORF proteins than for sORF proteins. The dependence on the E-value of the hit, in which the farthermost related species was found, was compared to the length class. As expected, the significance of a hit increases (decreasing E-value) with increasing length (Figure 3.2.3 B). Interestingly, the significance of the sORF hits is higher than that of the aORF hits, which may be explained by their lower gene age as already described in section 3.2.1. Older sequences have more hits that are less significant, while younger ones had not enough time to diverge and, therefore, are much more similar.

The number of protein and nucleotide binding sites was analyzed in dependence on the length of the sORF and aORF protein sequence. It was observed that the number of DNA (Figure 3.2.4 A) and RNA binding sites (Figure 3.2.4 B) per protein increases with a peak at 200 aa. Longer proteins have a decreasing number of nucleotide binding sites with increasing length. The number of protein binding sites in aORF encoded proteins is independent of their length (Figure 3.2.4 C). In contrast, the sORF-encoded proteins show an increasing number of protein binding sites. The percentage of aORF-encoded proteins with known molecular function or biological processes is only low for small peptides and increases greatly with length (Figure 3.2.5). At a length of 300 aa, the function of nearly all proteins is known. This result was expected, because databases do not accept small molecules. For example, NCBI refseq only accepts sequences >200 bp = 67 aa (Storz, et al. 2014; O'Leary, et al. 2015).

The percentage of sORF-encoded proteins predicted to be secreted linearly decreases by increasing length, the opposite trend as for cytoplasmic proteins. The majority of aORF-encoded proteins is predicted to be cytoplasmic and only very small peptides (30-60 aa) are more frequently predicted to be secreted (Figure 3.2.6). The parameter ‘structuredness’ is a composite of secondary structure (i.e. helices, beta sheet or loop; Figure 3.2.7), solvent accessibility (i.e. buried, exposed or intermediate amino acids; Figure 3.2.8), and disordered amino acids (Figure 3.2.9). With higher structuredness, more amino acids are buried, fewer exposed and the number of disordered amino acids and of disordered regions per ORF decreases. The secondary structure of aORF-encoded proteins is length independent, while the sORF-encoded proteins have shown a decreasing trend of helices and increasing tendency for beta sheets by increasing length (Figure 3.2.7). Nonetheless, the magnitude of amino acids involved in a secondary structure and that of amino acids forming a loop are comparable in both groups. Both solvent accessibility and percentage of disorder are strongly increasing by length. The values in the sORF dataset are comparable to those in the aORF dataset. The percentage of ORFs with at least one transmembrane helix is length independent (Figure 3.2.10 A). Shorter proteins of below 200 aa have significantly more disulfide bonds in comparison to longer ones (Figure

3.2.10 B). In conclusion, sORFs clearly show similar structural overall trends compared to aORFs, which suggests functionality for the set overall (see Discussion section 4.1).

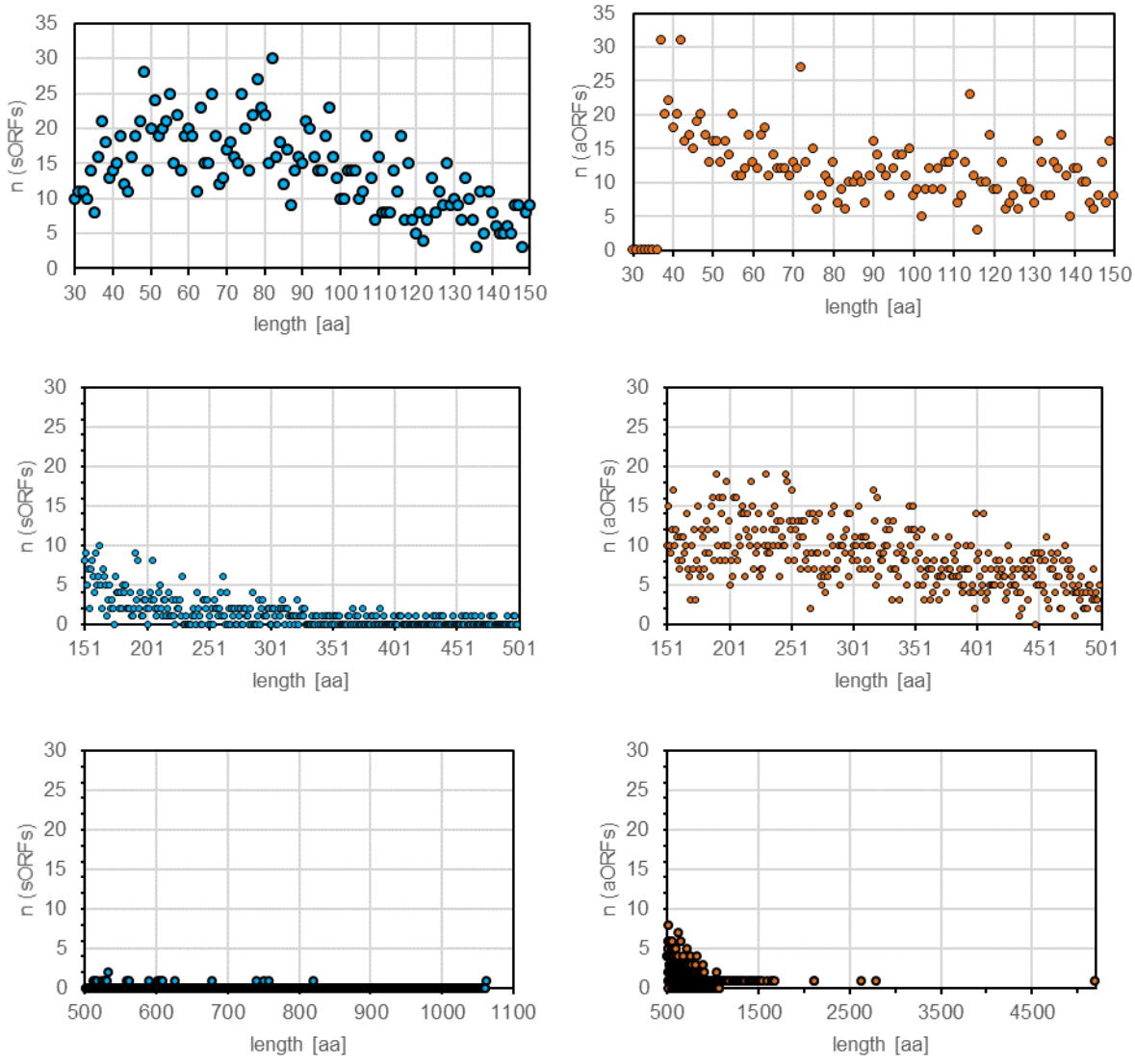


Figure 3.2.1: Number (n) of open reading frames with one particular amino acids [aa] length.
Legend: left, blue: 2,180 shadow ORFs (sORFs) with blastp hit; right, orange: 5,010 annotated genes (aORFs).

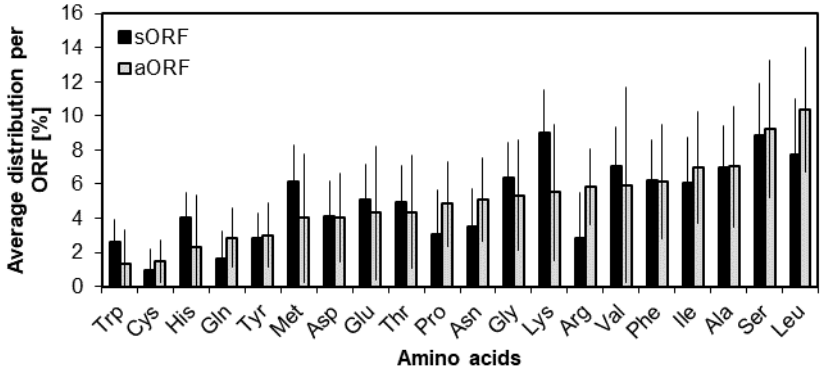


Figure 3.2.2: Average percentage of one particular amino acid per ORF. Shadow ORFs (■) are compared with annotated ORFs (□). The amino acid composition of each ORF was determined with PredictProtein. The mean values and standard deviations are shown in the plot.

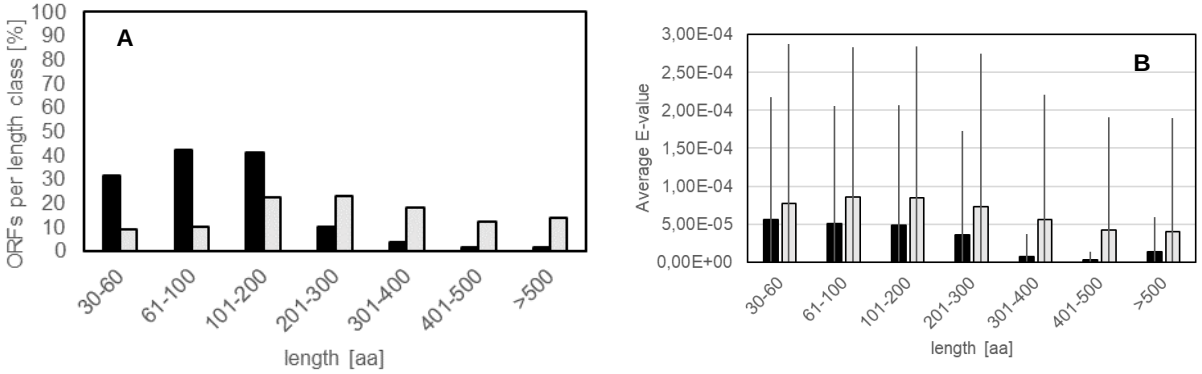


Figure 3.2.3: Distribution of ORFs per length class (A) and average E-values (B) in dependence of the length of the ORF encoded protein in amino acids (aa). Shadow ORFs (■) were compared to annotated ORFs (■). The distribution of ORFs per length class is shown as percentage of all sORFs or aORFs in the dataset. The E-value results from the hit in which the furthest related species was found. The mean values and standard deviations are shown in the plot.

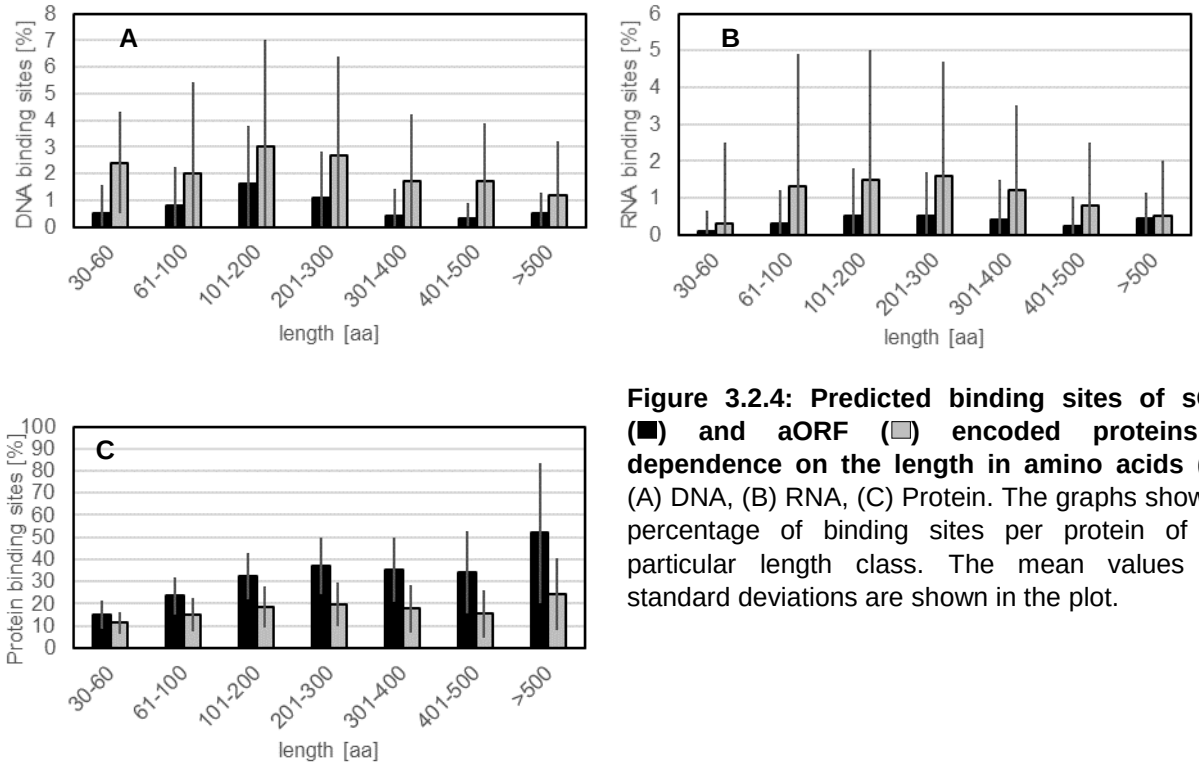


Figure 3.2.4: Predicted binding sites of sORF (■) and aORF (□) encoded proteins in dependence on the length in amino acids (aa). (A) DNA, (B) RNA, (C) Protein. The graphs show the percentage of binding sites per protein of one particular length class. The mean values and standard deviations are shown in the plot.

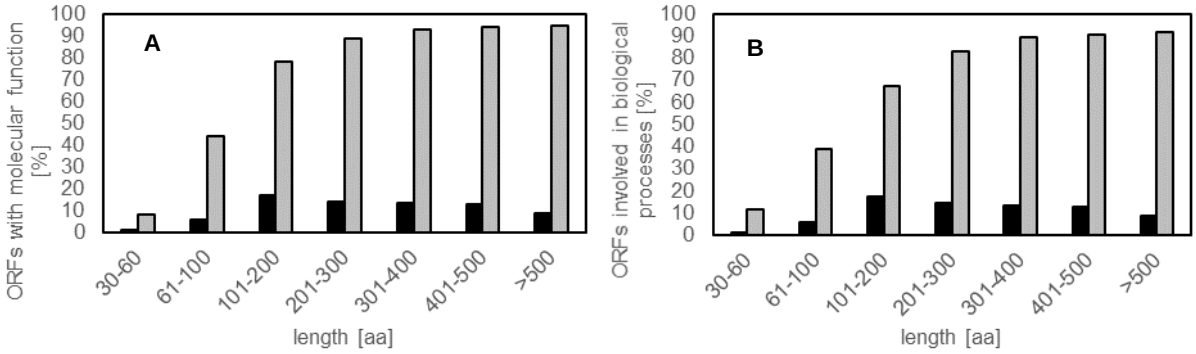


Figure 3.2.5: Percentage of sORFs (■) and aORFs (□) with predicted molecular function (A) and biological processes (B) in dependence on the length class of ORF encoded proteins in amino acids (aa). The graph is shown as percentage of ORFs in a particular class. The mean values and standard deviations are shown in the plot.

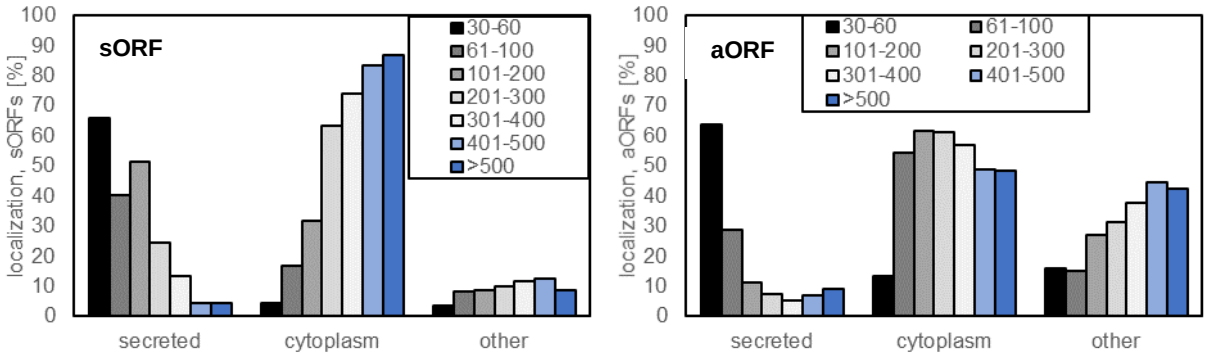


Figure 3.2.6: Predicted localization of sORFs and aORFs in dependence on the length class of ORF encoded proteins in amino acids. The localization is shown as percentage of all ORFs in a particular class. The group 'other' contains proteins predicted to be localized in the inner or outer membrane, the periplasm or the fimbrium.

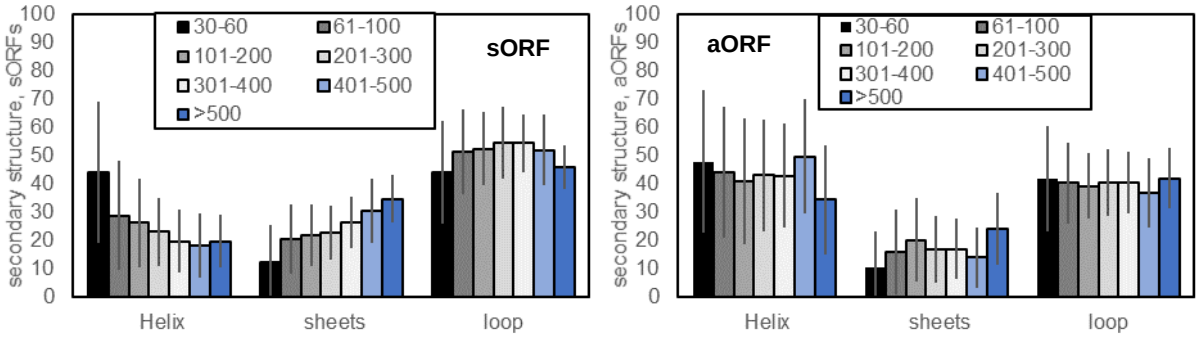


Figure 3.2.7: Predicted secondary structure of sORF and aORF encoded Proteins in dependence on the amino acid length. The secondary structure is shown as percentage of ORFs in a particular class. The mean values and standard deviations are shown in the plot.

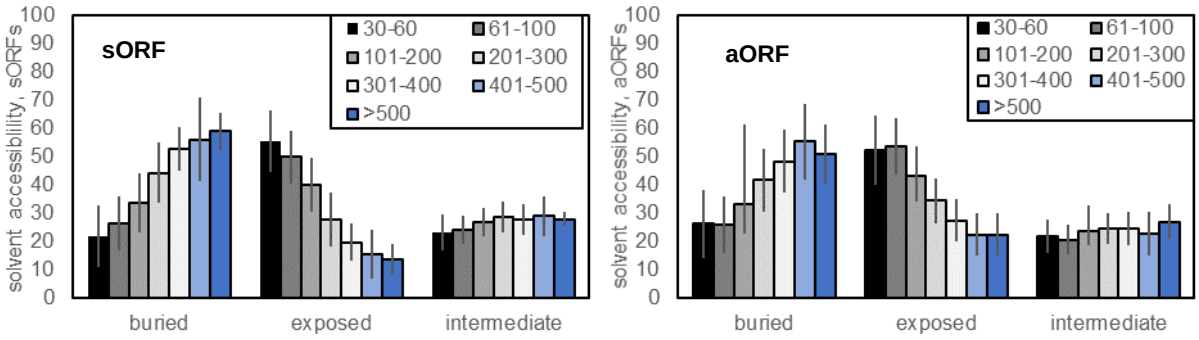


Figure 3.2.8: Predicted solvent accessibility of sORF and aORF encoded Proteins and in dependence on the amino acid length. The solvent accessibility is shown as percentage of ORFs in a particular class. The mean values and standard deviations are shown in the plot.

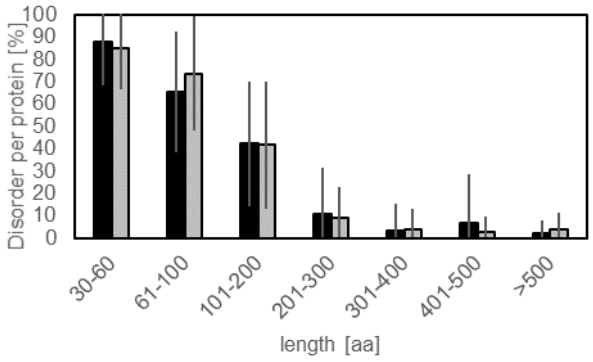


Figure 3.2.9: Predicted disorder [%] per sORF (■) or aORF (□) encoded protein in dependence on the amino acid length. The mean values and standard deviations are shown in the plot.

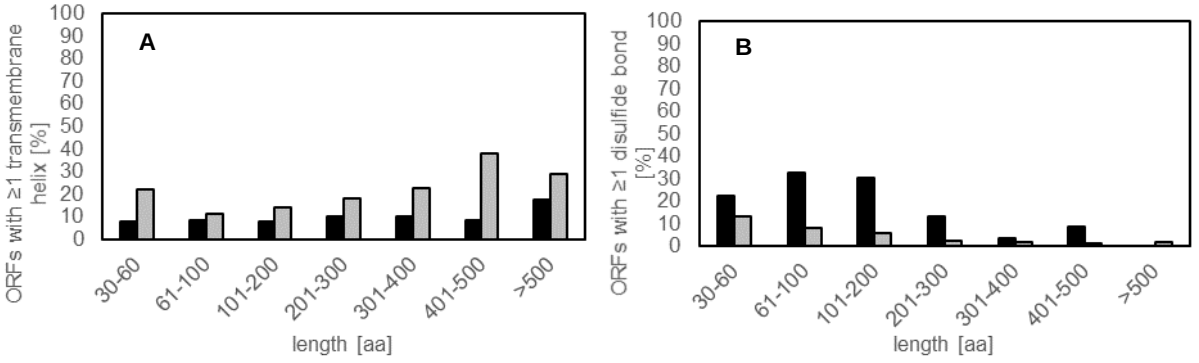


Figure 3.2.10: Percentage of sORF (■) and aORF (□) encoded proteins with more than one predicted transmembrane helix (A) or disulfide bond (B). All parameters are shown in dependence on the amino acid (aa) length. The mean values and standard deviations are shown in the plot.

3.3 Correlation of structural features of sORFs and aORFs with their phylostratigraphic level

The predicted features of proteins encoded by sORFs and aORF were analyzed in dependence on their relative gene age. Shadow ORFs and aORFs were classified in phylostratigraphic levels as described in section 3.1.3. First, a short overview is given and the values are discussed in more detail below. Absolute values for the features 'length' (Figure 3.3.1 A) and 'nucleotide binding sites' (Figure 3.3.1 B, C) are increasing when compared to the age level. In contrast, the number of amino acids involved in protein-protein interactions is more or less age independent (Figure 3.3.1 D). Cellular functions of a predicted protein are characterized by three parameters, 'localization' (Figure 3.3.2 A, B*),* 'molecular function' (MF) and 'biological process' (BP, Figure 3.3.2 C). While young proteins are mainly secreted and only a few are cytoplasmic, this reverses by increasing age (Figure 3.3.3 A, B). MF and BPs were only predictable for annotated proteins; the trend is increasing by age. In contrast, only very few shadow-ORF encoded proteins were predicted to have a molecular function or are involved in biological processes. However, this can be explained by the limits of this method. The prediction is based on gene ontology classification after PSI-BLAST. Most shadow ORFs have only a few homologous proteins in the database and many of them are annotated as 'hypothetical' or 'unknown' (92%), excluding any certain prediction. In contrast, for only 30% of the annotated ORFs the biological function is uncharacterized to date (Table 3.1.2). It was further tested, whether proteins involved in biological processes necessarily have a known molecular function. There are four sORFs with predicted BP and no MF, 192 vice versa and only four which have hits in both. The aORF distribution is as follows: 201 BP only, 456 MF only and 3,433 with both. A significant difference between BP and MF was already observed by Radivojac, et al. (2013) and can be explained by different informational contents in the database. For some annotated genes, only the function of the molecule is known, others have indications about the networks in which they are acting in.

Looking at the secondary structure, it can be seen that the percentage of amino acids involved in helix formation is decreasing with greater gene age, while those in loops is increasing (Figure 3.3.3 A, B). The percentage of amino acids forming beta sheets is age independent. The solvent accessibility is age independent (Figure 3.3.3 C, D), as well as the percentage of disordered regions (Figure 3.3.4). The number of disulfide bonds decreases by age (Figure 3.3.5 A, B). The same trend can be observed for the average number of cysteines per ORF (Figure 3.3.5 C). Both parameters only weakly correlate (Supplementary figure S3*)* which was also observed in a similar study with intergenic ORFs in EHEC by Neuhaus, et al. (2016). The number of sORFs

with more than one transmembrane helix is significantly lower than the number of aORFs with transmembrane helices (8% versus 21%).

The trends of shadow-ORF protein features obviously differ from those of annotated ORFs encoded proteins when plotted against the phylostratigraphic age. While taxonomically restricted proteins encoded by shadow ORFs are slightly longer than those encoded by annotated ORFs, the length of annotated proteins does conspicuously increase by age (from 64 aa to 344 aa). The length of sORFs only slightly increases (from 102 aa to 162 aa, Figure 3.3.1 A). Young sORF encoded proteins have more predicted nucleotide binding sites (Figure 3.3.1 B, C) (first three levels) in comparison to annotated proteins and for conserved proteins (last four levels) it is the other way round. The number of predicted protein binding sites of sORF is higher than that of annotated proteins (Figure 3.3.1 D). The trend for the prediction of cell localization of both all ORF groups is comparable, but more shadow-ORF encoded proteins are predicted to be secreted (Figure 3.3.2, A: 48% shadow ORFs and B: 38% annotated ORFs). The group ‘others’ contains proteins to be predicted for fimbrium, inner membrane, periplasm or the outer membrane. These proteins have been binned together due to low numbers in each group and have no clear overall trend.

Not all predicted parameters showed age dependent trends. The age dependent profiles of sORF and aORF encoded proteins with respect to the feature “transmembrane helices” (Figure 3.3.6) have no clear trend. Further, no shadow ORF was predicted to contain a β-barrel protein and only a few putative proteins were predicted to have coiled-coil structure (data not shown). This is easily explained by the short length of the shadow-ORF encoded proteins and is corroborated by the low predicted numbers of transmembrane proteins in this group. The comparison of predicted proteins either encoded in sORFs or aORFs shows that the overall age dependent trends are comparable, but less clear for most sORF protein features.

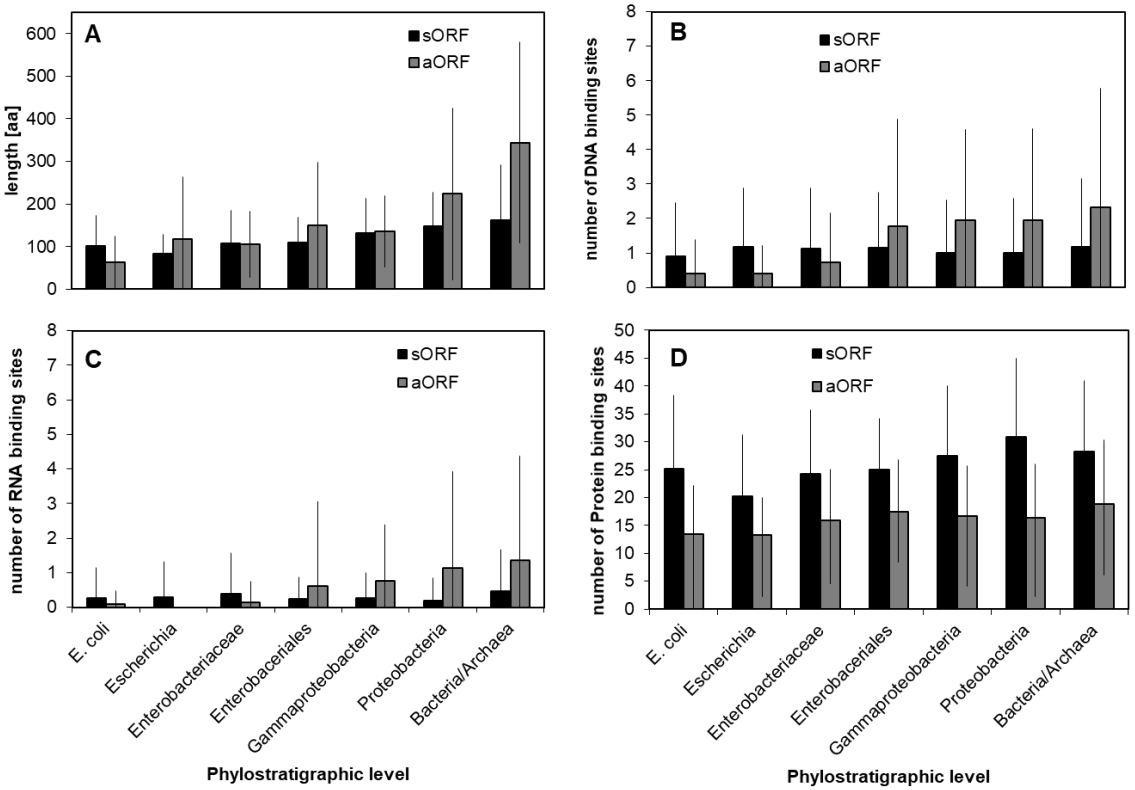


Figure 3.3.1: Protein features of proteins encoded by shadow ORFs (■) and annotated ORFs (□) in dependence on their phylostratigraphic level. Features were predicted with Predict Protein; (A) amino acid sequence length, (B) deoxyribonucleic acid - and (C) ribonucleic acid binding sites per amino acid sequence (D) protein binding sites per amino acid sequence. The mean values and standard deviations are shown in the plot.

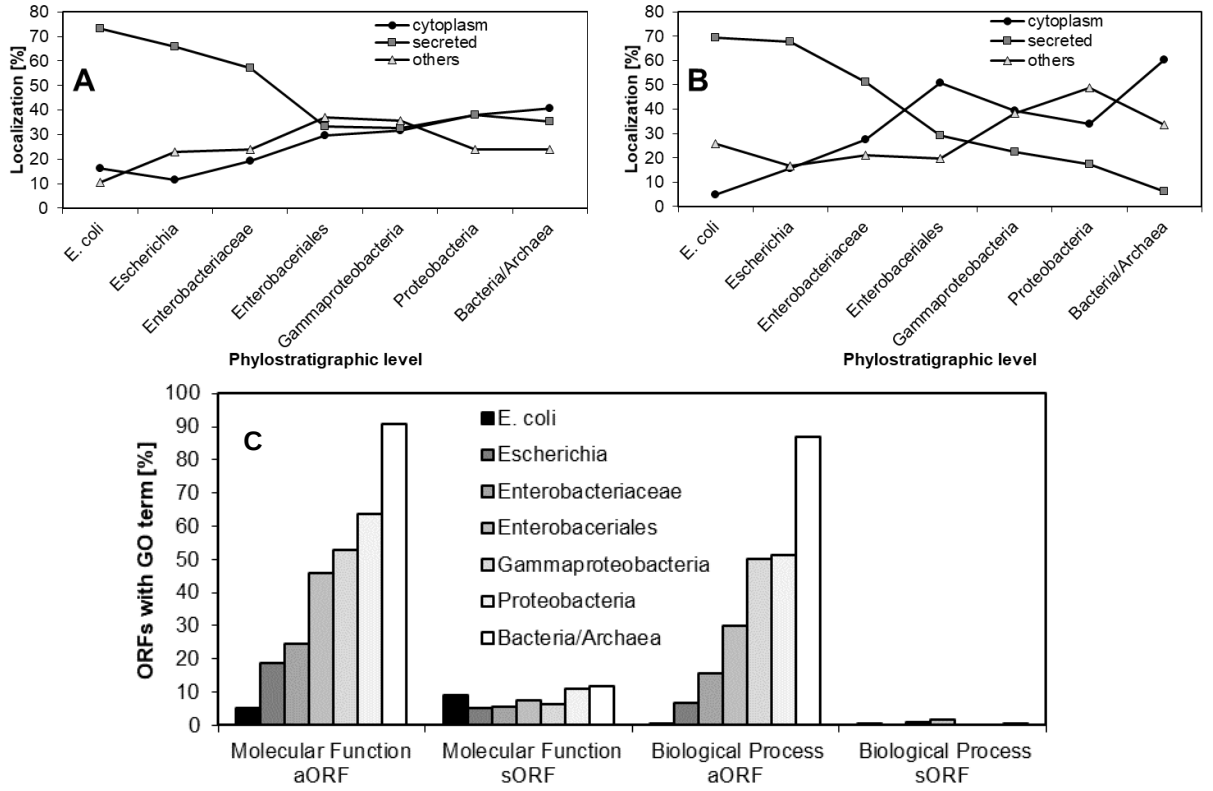


Figure 3.3.2: Predicted localization and function in dependence on phylostratigraphic level. (A) localization of proteins encoded by shadow ORFs and (B) localization of proteins encoded by annotated ORFs; (C) prediction of molecular function and biological process.

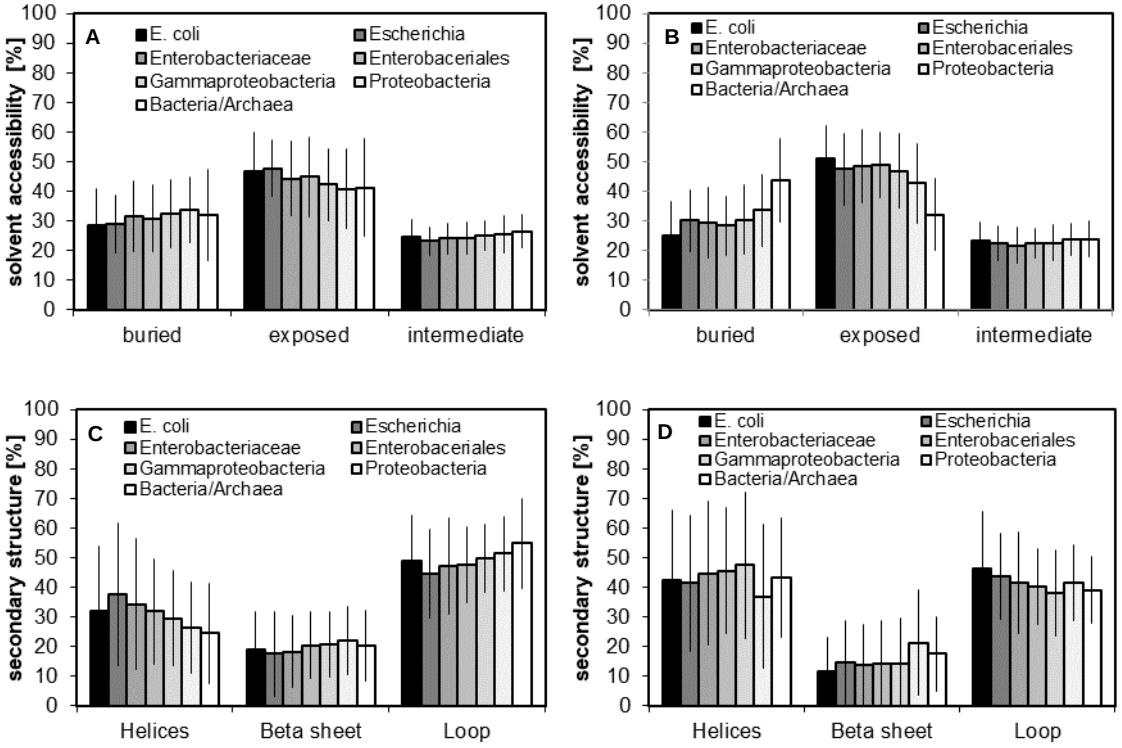


Figure 3.3.3: Predicted solvent accessibility and secondary structure in dependence on phylostratigraphic level. (A) Solvent accessibility of proteins encoded by shadow ORFs and (B) solvent accessibility of proteins encoded by annotated ORFs; (C) secondary structure of proteins encoded by shadow ORFs and (D) secondary structure of proteins encoded by annotated ORFs. The mean values and standard deviations are shown in the plot.

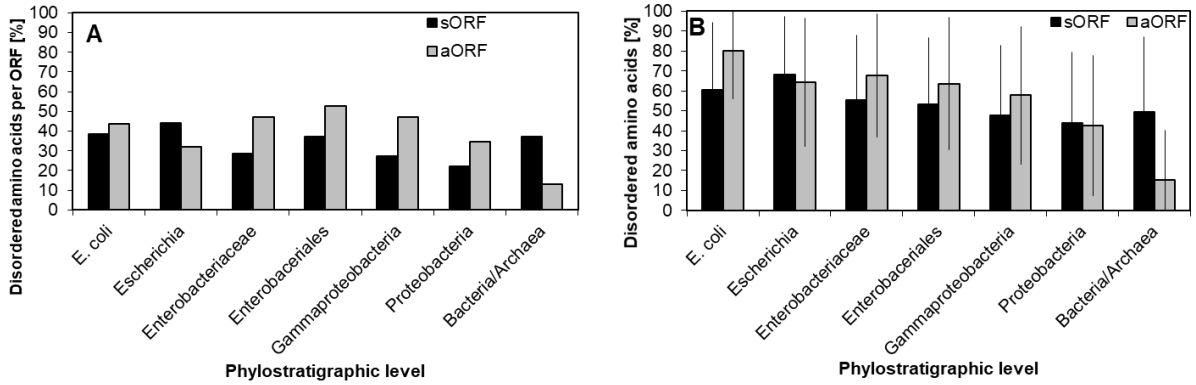


Figure 3.3.4: Predicted disordered amino acids of shadow ORFs (■) and annotated ORFs (□) in dependence on phylostratigraphic level. (A) per cent of amino acids in disordered regions of at least 30 consecutive disordered amino acids (Uversky 2011; Peng, et al. 2015); (B) total number of disordered amino acids in the dataset. The mean values and standard deviations are shown in the plot.

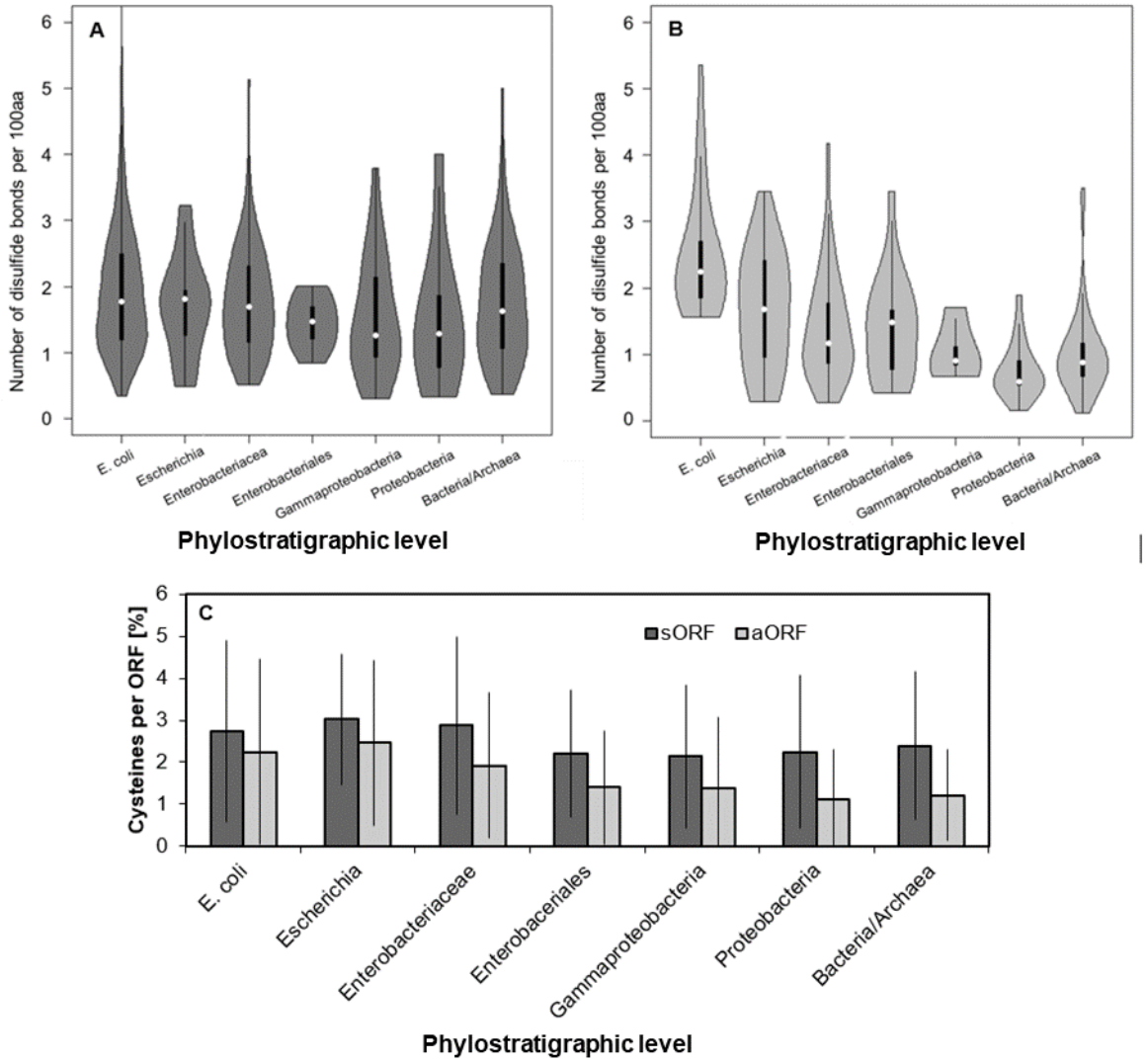


Figure 3.3.5: Predicted disulfide bonds and cysteines in dependence on phylostratigraphic level. (A) disulfide bonds of proteins encoded by shadow ORFs and (B) disulfide bonds of proteins encoded by annotated ORFs; c) Percentage of cysteines of proteins encoded by shadow ORFs (■) and by annotated ORFs (□).Predicted disulfide bonds were normalized to a length of 100 amino acids. A few outliers were found in the phylostratum E. coli of the shadow ORFs having up to 11 disulfide bonds per 100aa. ORFs predicted to have no disulfide bonds were excluded (74% in sORFs and 96% in aORFs). The violin plot (A and B) is a boxplot with a rotated kernel density plot on each site. It shows the median (white dot) and the range between both quantiles as black box. The violin plots were constructed with the CRAN-R package 'vioplot'. The bar plots (C) show the mean values and standard deviations.

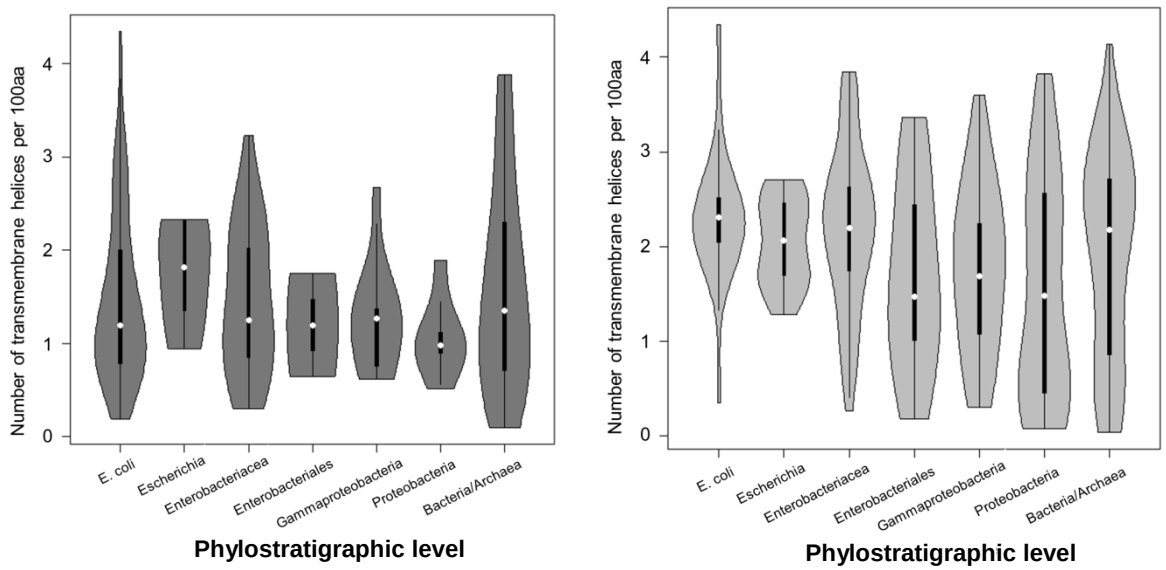


Figure 3.3.6: Predicted transmembrane helices of sORF (■) or aORF (□) proteins in dependence on phylostratigraphic level. The predicted transmembrane helices were normalized to a length of 100 amino acids. ORFs predicted to have no transmembrane helices were excluded (92% in sORFs, 79% in aORFs). The violin plot (A and B) is a boxplot with a rotated kernel density plot on each site. It shows the median (white dot) and the range between both quantiles as black box. The violin plots were constructed with the CRAN-R package 'vioplot'.

3.4 Functional characterization of the overlapping gene *asa*

The novel gene asa is transcribed and translated in EHEC

One overlapping gene was functionally (Section 3.4) and evolutionarily (Section 3.5.5) characterized more in detail. The novel gene *asa* was discovered by transcriptome and translatome analyses (Landstorfer 2014). The reads of RNAseq and RIBOseq were re-mapped to the genome published later by Latif, et al. (2014). The panels taken from Artemis 17.0 show the sum signal of both biological replicates. There is a clear signal for a larger transcribed region around the overlapping ORF and, for RIBOseq data, at the beginning of the open reading frame (Figure 3.4.1). However, the coverage is low for both RNAseq and RIBOseq (0.29 and 0.30, respectively). The RPKM values of transcription and translation are 14.6 and 13.6, respectively.

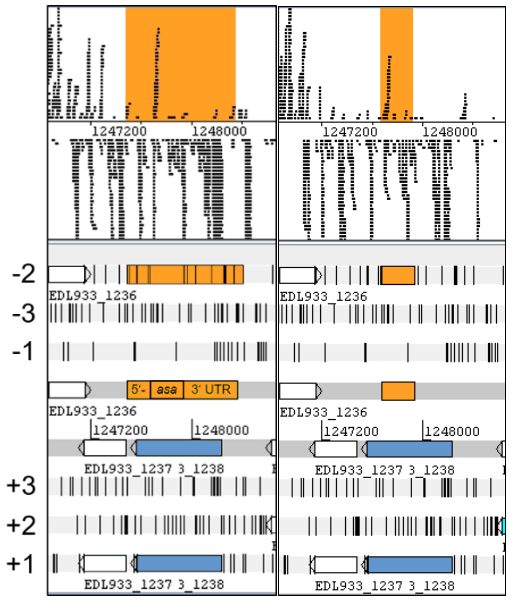


Figure 3.4.1: Transcription (A) and translation (B) of *asa* shown as RNAseq and RIBOseq reads. Transcribed or translated regions of *asa* are highlighted in orange, its mother gene (#1238) in blue.

The overlapping gene pair asa/EDL933_1238 is flanked by enzyme encoding genes and a prophage

The gene *asa* (start / stop position 1247671 / 1247934, respectively) has a length of 264 bp. It is completely embedded in antisense (-2 frame) to a Ca^{2+} regulating transporter gene (TEGT family, “Testis Enhanced Gene Transfer”, locus tag: EDL933_1238, *yccA* in the genome published by Perna, et al. 2001). The gene pair *yccA/asa* is flanked by two enzymes upstream of *yccA* (EDL933_1236: tRNA 2-thiouridine synthesizing protein E, EDL933_1238: acyl-phosphohydrolase gene) and downstream by the O-island #44 which encodes the prophage CP-EDL933M (Figure 3.4.2).

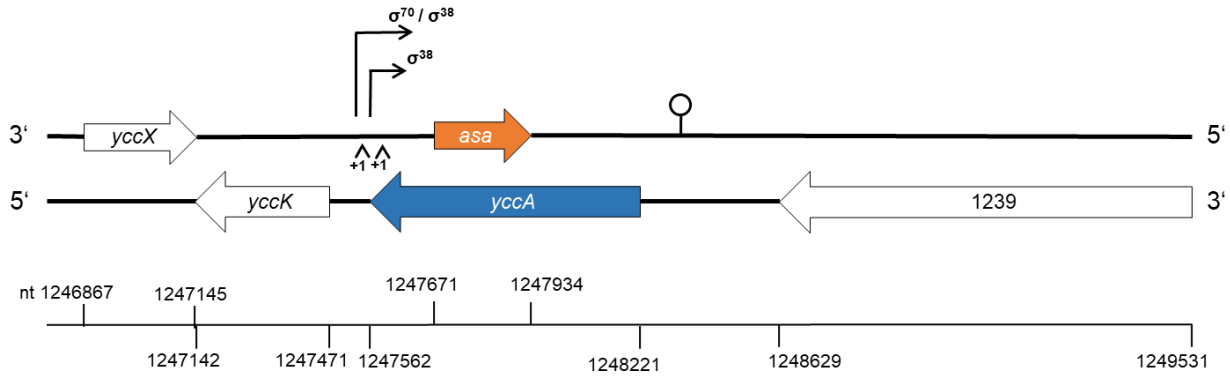


Figure 3.4.2: Genome organization of the overlapping gene pair *asa*/EDL933_1238 in EHEC EDL933. The transcription of the shadow ORF *asa* (orange) is initiated by three putative promoters (σ^{70} or σ^{38}) at three experimentally determined transcriptional start sites (+1) 178 bp, 186 bp and 188 bp upstream of *asa*. There is a putative terminator (ϒ) 430 bp downstream of *asa*. Downstream of the mother gene *yccA* (EDL933_1238, blue) are two enzymes, a tRNA 2-thiouridine synthesizing protein E gene (*yccX*) and putative acyl-phosphate phosphohydrolase gene (*yccK*). The O-island #44 is upstream of the mother gene and encodes the prophage CP-EDL933M including an integrase gene (1239). The genome localizations are shown on the number line.

Overexpression phenotypes in NaCl and L-arginine stress were confirmed

Preliminary experiments (Zehentner 2015) revealed three putative phenotypes of *asa* in NaCl, L-arginine and in pyridoxine hydrochloride (= vitamin B6). The overexpression phenotypes were confirmed by competitive growth of EHEC with overexpressed *asa* against an overexpressed translationally arrested mutant of *asa*. This experimental setting was chosen to maintain the same cellular environment, which can influence the fitness (EHEC + vector + insert of the same length).

The *asa* mutant has a stop codon, which leads to translational arrest (Figure 2.2). The stop codon was introduced at the first third of the sequence to prevent the expression of a stable shortened protein. The full-length Asa protein has a length of 88 aa and the shortened protein a length of 22 aa (arginine → stop, *asa*^{tar22}). To exclude position effects, a second translational arrested mutant with a stop codon at an alternative position was tested likewise (stop codon after 28 aa, cysteine → stop, *asa*^{tar28}). A synonymous mutation was additionally introduced (nucleotide 62 or 83) to validate the peak ratio obtained by Sanger sequencing, which was the parameter measured to determine the wild type-mutant ratio.

There is a growth disadvantage of EHEC overexpressing the full-length Asa protein in NaCl and L-arginine stress. The phenotype of pyridoxine hydrochloride disappeared. The initial wild type-mutant ratio of 1:1 changed to 1:4 after growth in LB supplemented with NaCl and 1:1.9 in LB

supplemented with L-arginine (Figure 3.4.3). The competitive growth of EHEC with intact *asa* against the mutant at an alternative position confirmed these results (ratio in both stressors: 1:4). The phenotype was stressor specific as growth in LB medium did not show any preference in a specific direction. Non-competitive growth tested in NaCl did not show any differences in growth behavior (Supplementary figure S4).

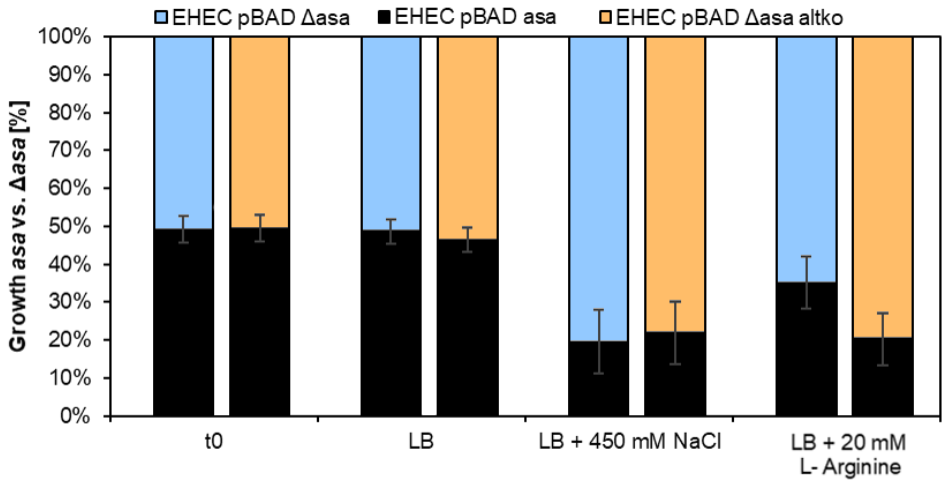


Figure 3.4.3: Over-expression phenotypes of *asa* shown as ratio of wild type and mutant [%]. EHEC overexpressing *asa* (black) grew competitively against EHEC overexpressing a translational arrested mutant Δ*asa* (blue) or a mutant with an alternative stop codon position Δ*asa* altko (salmon) in LB supplemented with 450 mM NaCl or 20 mM L-arginine. Cells grown in LB were used as negative control. For overexpression, *asa* or the mutant were cloned in pBAD-myc/His C. The initial wildtype:mutant ratio was 1:1 (t0). The gene expression was induced with 0.002% arabinose (w/V). EHEC was harvested after 22 h, the plasmid was isolated and sequenced by Sanger sequencing. The phenotype was determined by the peak height ratio of wild type and mutant at the positions mutated. The plot shows the mean values and standard deviations of three biological replicates.

Three putative transcriptional start sites were identified

The 5' end of *asa* mRNA was determined with 5' RACE ("Rapid Amplification of cDNA Ends") and Cappable seq. The 5' RACE experiments were conducted in two experimental settings. In the first experiment, EHEC grew in LB medium. There were some difficulties in getting a full-length sequence and it was assumed that there was not enough mRNA sufficient for the sensitivity of the method. This is the reason why the experiment was repeated after growth in NaCl, where a high promoter activity was observed (see section 3.4.1.5), with pProbeNT-GFP +

asa promoter region as insert. The 5' end of this construct was determined by using 5' RACE with parameters used in the first experiment.

The *asa* mRNA was first reverse transcribed in cDNA with *asa* specific primers. A poly-A tail was added with a terminal transferase. In a first PCR the single stranded DNA became double stranded and was amplified by using an oligo-dT primer and a nested *asa* primer. An anchor was added and used as primer binding site in a second PCR, which was conducted to increase the product yield. The PCR product size was checked with a 2% agarose gel; the product was cut out and sequenced. The first nucleotide downstream of the poly-A tail is the +1 site. The agarose gel shows a product at a length of ~ 400 bp after growth in LB and of 200-300 bp after growth in LB + NaCl (Figure 3.4.4) which corresponds to the expected size of 388 bp respectively 192 bp (sequenced *asa* + 5' UTR) plus poly-A tail and anchor (39 bp). The +1 sites are 186 bp upstream of the start codon (growth in LB) and 178 bp upstream of the start codon (growth in LB + NaCl).

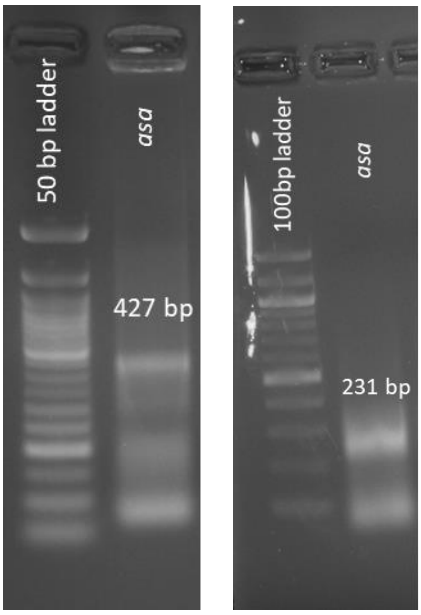


Figure 3.4.4: Agarose-Gel (2%) after the second PCR of 5' RACE to determine the transcriptional start site (+1 SITE). Legend: (left) *asa* mRNA after growth of EHEC in LB medium. (right) GFP with the 5' UTR of *asa* cloned in pProbe-NT for promoter activity experiments (section 3.4.1.5). In this experiment EHEC was grown in LB + 450 mM NaCl. The product size was compared with the respective DNA ladder (50 bp or 100 bp, *New England Biolabs*).

As both experiments do not result in the same +1 site, Cappable seq was consulted (conducted by Barbara Zehentner after Ettwiller, et al. (2016), unpublished data). Cappable seq is a next-generation sequencing based method which determines the transcriptional start sites of all mRNA in a particular genome. The experiment was implemented under four growth conditions (LB, M9 medium, LB + 500 mM NaCl, LB + 4 mM malic acid). The cells were harvested at two growth phases (early exponential phase OD_{600} = 0.3 and early stationary phase OD_{600} = 1 - 3.5). The +1 site was determined in three biological replicates. The +1 site detected here was 188 bp upstream of the start codon, which is close to the start site detected by 5'RACE after growth in LB medium, which was 186 bp upstream of the translation start.

The asa expression is regulated by three putative promoters

There are three promoters, which putatively regulate the *asa* expression (Figure 3.4.5). The expression of the mRNA with a 5' end 188 bp or 186 bp upstream of the start codon can be regulated by a σ^{70} promoter or a σ^{38} promoter. The σ^{70} promoter was identified by BProm 9 bp upstream of the +1 site. The linear discriminant function score (LDF, BProm) of 2.10 is significantly higher than the threshold of 0.2. The accuracy of the promoter prediction is 80% (Solovyev and Salamov 2011). The σ^{38} promoter was manually identified 7 bp upstream of the +1 site. The σ^{38} promoter has the consensus sequence "CTACACT" at -10 site (Lee and Gralla 2001). The -35 site of the σ^{38} promoter is variable and cannot be identified by a consensus sequence. The -10 site sequence found here "ATAATTA" appears to have an average conservation frequency of 36%, calculated after Lee and Gralla (2001). The mRNA with a 5' end 178 bp upstream of the start might be regulated by a σ^{38} promoter 7 bp upstream of the +1 site. The average conservation frequency of the sequence "CTACCTT" is 59% (Lee and Gralla 2001).

The promoter activity was measured by cloning the promoter region in a promoter-less pProbeNT-GFP vector (Figure 3.4.5, shown in black letters). In case of an active *asa* promoter, GFP would be expressed and could be measured by fluorescence. The intensity is proportional to the promoter activity. The cells were tested in LB and LB supplemented with NaCl, L-arginine or pyridoxine hydrochloride and harvested at exponential phase (OD_{600} = 0.8). The highest activities were measured after growth in NaCl and in pyridoxine hydrochloride (Figure 3.4.6). The activity in LB or in LB supplemented with L-arginine was significantly lower and comparable in their intensity. The negative controls (empty pProbeNT and a fragment at the terminator region of EDL933_1236) showed an insignificant low fluorescence.

TTGGCAGGCCAGCAATTTTGGTGGCTTGCTTAGCCGGACCTTTCGGAAACAGTCGGTAT
AAATAGCGGCTGTTACCTTTTTCTTCGCCAAATTTATTCGCCATCGCTTTTACCAGCAT
ACGAATCGCCGGAGAAGTATTGAATTCCAGATAGAAATCACGCACAAAACGCACCACTT
CCCAGTGTTCTGGCGACAGCGAAATCCCTTCGTTCTCTGCAATCACCACTGCCAGCGGC
TCACTCCACTGGCTGCTTTCTTTGAGATAGCCTTCGGTATCCGTTtctATCTCTTTACC
TTCGAAGATcAgcAtaaTTATTACTACCTTAATCAGACTGCCGGCAGTGTAACAAAAAA
CAAAGCCCCGCATAAAGCGAGGCTATGAAAGTGTTAGCGGGTGAGATTAATCGCGGCTG
GCGAAGCCCAGAATGCTCAGCAGGCTGACGAAGATGTTGTACAGCGAAACATACAGGCT
AACCGTGGCACGAATATAGTTCGTTTCACCGCCATGAATGATGTT

CAGCaTAattaTTACTACCTTAATCAGACTGCCGGCAGTGTAACAAAAAA

CAGCATAATTATTACTACCtTAATCAGACTGCCGGCAGTGTAACAAAAAA

Figure 3.4.5: Regulatory regions of *asa* expression. The +1 site determined with Cappable seq is highlighted in green, those determined by 5'RACE are highlighted in grey. The σ^{70} promoter (A) is highlighted in yellow, σ^{38} promoters (B, C) are highlighted in blue. The putative promoter region (93 bp) was cloned in pProbeNT to test the promoter activity (black letters). The negative control, the terminator region of gene 1236 (76 bp), is shown as blue letters.

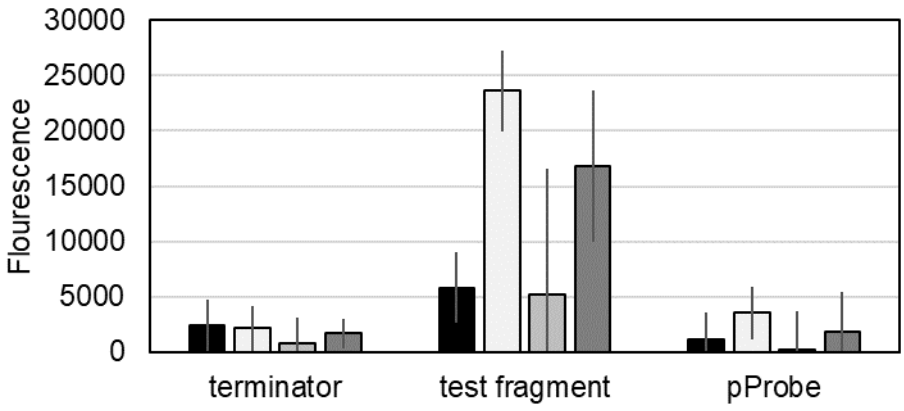


Figure 3.4.6: Fluorescence intensity as measure for GFP expression activated by the *asa* promoter. The promoter activity was tested with *E. coli* TOP10 with the pProbeNT-asa promoter construct ("test fragment"). *E. coli* was grown under the following conditions: LB (■), LB + 450 mM NaCl (□), LB + 10 mM L-arginine (▯) or LB + 10 mM pyridoxine hydrochloride(▯). The intensity at a cell density of OD_{600} = 0.8. Two negative controls were tested: The terminator region of EDL933_1236 ("terminator") and the promoter-less pProbeNT vector. The plot shows the mean values and standard deviations of three biological and four technical replicates.

The gene expression is regulated growth phase and condition dependent

NaCl was the stressor with the strongest effect on *asa* and was chosen as the condition in which to test for gene regulation. Growth in plain LB medium was used as control. The experiment was conducted in three biological replicates and three technical replicates per qPCR run. The quantification cycle (cq) of *asa* was normalized to that of the housekeeping gene, 16S rRNA (Δcq = cq[asa] - cq[16S]). The cq value negatively correlates with the mRNA titer.

Two negative controls were used: 1) a qPCR of each 16S rRNA sample which was not reverse transcribed and 2) a 59 bp region which has any RNAseq signal (one sample measured: LB, exponential phase). The first control excludes contamination with DNA (cq ≥ 29; Δcq ≥ 16; ratio DNA:RNA = $1:2^{16}$ = 1:65536). The second control ensures that the signal comes from *asa* mRNA and not by any mRNA. The 16S rDNA primer at this run was slightly contaminated with gDNA (cq [without reverse transcription] - cq [with reverse transcription] = 1; ratio DNA:RNA = $1:2^{1}$). However, the cq of the untranslated region itself was sufficiently high to show a sufficiently low mRNA titer (cq = 32, Supplementary table S13).

The influence of NaCl stress was tested in two experimental approaches. The *asa* expression of EHEC adapted overnight to NaCl was measured after harvesting the cells during early exponential phase (OD_{600} = 0.2 - 0.3) and during exponential phase (OD_{600} = 0.7 - 0.8). Stress shocked cells were tested after addition of NaCl at OD_{600} = 0.8. Samples were taken before the stress shock and 30 min, 60 min and 120 min thereafter. The growth curves of both experiments can be found in Supplementary figure S5. All cq-values of stress adaptation experiments can be found in Supplementary table S13 and those of stress shock experiments in Supplementary table S14.

The stress-adapted cells clearly show a growth phase and growth condition dependent regulation of *asa* expression. There is an upregulation from early exponential phase to exponential phase which is stronger in salt stress (2-fold in LB, 6-fold in NaCl, Figure 3.4.7 A). The mRNA titer of *asa* is lower in LB than in NaCl (1.7 fold lower) during early exponential phase, but higher in LB than in salt during exponential phase (2.1 fold higher).

When the cells were shocked with NaCl during exponential phase, the RNA titer decreases during the first 30 min thereafter (Figure 3.4.7 B). This effect is caused by NaCl, because the titer of the asa mRNA remains constant in LB. Sixty minutes thereafter (time point 60 min), the *asa* titer increases and reaches the magnitude of that in LB. At time point 60 min, four biological replicates were measured - two were of the magnitude of time point 30 min, two of time point 120 min. The highest RNA titer was found 120 min after stress shock and is equal to the titer in LB at that time point.

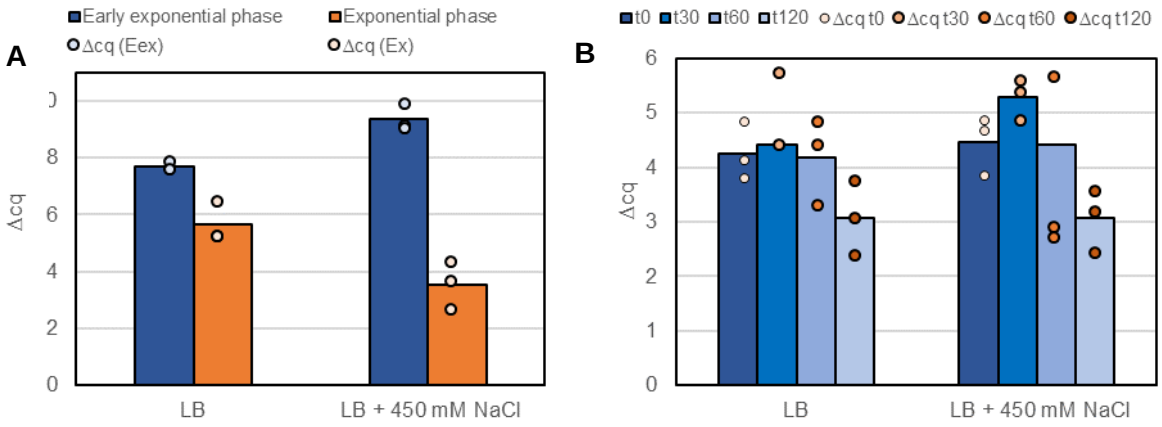


Figure 3.4.7: Gene regulation of *asa* in stress adapted (A) and stress shocked (B) EHEC. The stress condition used was LB supplemented with 450 mM NaCl; LB without stressor was used as control. The relative *asa* mRNA concentration was measured as quantification cycle (cq) by RT-qPCR. The quantification of *asa* was normalized to that of the 16S rRNA gene (Δcq = cq[asa] - cq[16S]). The graph shows the average Δcq of three biological replicates (bars) and each individual biological replicate (dots). Due to its high variation, the sample "NaCl, t60" was measured in four replicates. (A) Overnight cultures adapted to NaCl were transferred into fresh medium, incubated and harvested at early exponential phase (Eex, blue, OD_{600} = 0.2 - 0.3) and at exponential phase (Ex, orange, OD_{600} = 0.7 - 0.8). (B) EHEC grown in LB was shocked with NaCl at an OD_{600} = 0.8. Aliquots were taken before stress-shock and 30 min (t30), 60 min (t60) and 120 min (t120) thereafter.

Western Blot reveals an Asa protein

The *asa* encoded protein was linked to a C-terminal SPA tag (Sequential Peptide Affinity) which consists of a calmodulin binding peptide, a TEV protease cleavage site and three modified FLAG sequences (Zeghouf, et al. 2004). Overexpression of the 17.5 kDa fusion protein (10 kDa Asa + 7.5 kDa SPA) was detected in a Western Blot using a monoclonal ANTI-FLAG antibody, which was linked to an alkaline phosphatase. The fusion protein Asa::SPA was detected at six time points. The positive control (GST::SPA, *gst* encoding a glutathione S transferase) showed a band caused by a protein with the expected size of 22 kDa (Figure 3.4.8 lane 1). The lanes 3 - 8 contain Asa::SPA and show two bands stained on the blot. The upper and more prominent band migrated at approximately 20.6 kDa and the lower band at 17.8 kDa. The upper band is presumed to be the full-length product and the smaller product may be caused by translation from an alternative start codon, located downstream. The protein product of the alternative start codon would be, theoretically, 1.65 kDa smaller. The nucleotide sequence of this region (the first 75 bp of *asa*) were checked for ribosome-binding sites. Indeed, there is a strong Shine Dalgarno sequence (AGAGGAGAT, $\Delta G°$ = -20.9 kJ/mol) 5 bp upstream of an internal ATG codon at nucleotide position +46 (Figure 3.4.9). There is no predicted Shine Dalgarno sequence upstream

of the start codon of the full-length gene. Which start codon is naturally preferred in the cell is unknown, but the Western Blot indicates that both may be expressed. The band of the full-length product is putatively stronger, because its expression is regulated by a ribosome-binding site located on the vector and the shortened product by the natural ribosome-binding site.

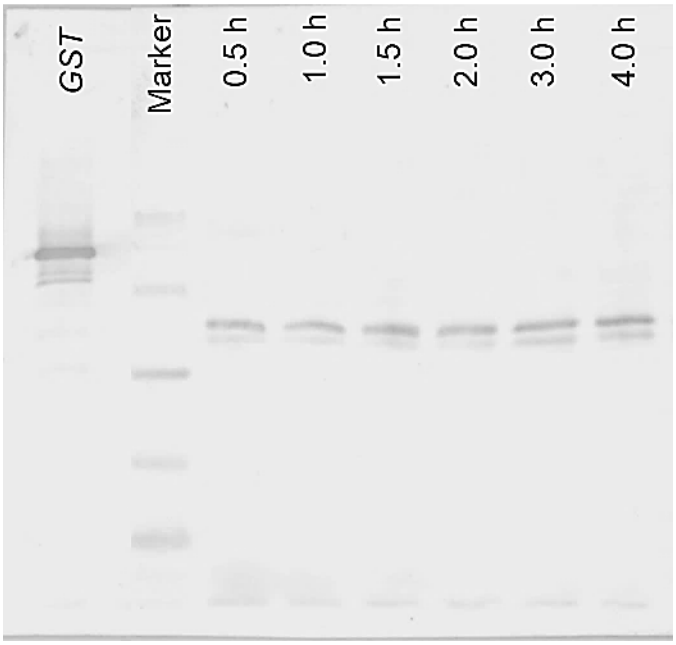


Figure 3.4.8: Western Blot of Asa linked to a C-terminal SPA-tag. Asa::SPA was detected with a monoclonal ANTI-FLAG® M2-Alkaline Phosphatase antibody, and visualized with NBT/BCIP. Legend from left to right: positive control (GST, 30 kDa), Spectra Low Range Marker (Thermo Fisher Scientific; time points taken as indicated after induction. The aliquot volume was adjusted to the OD_{600} measured 0.5 h after induction.

ATGATGTTGCTGGTTTCAAACAAA**ATA**GCGCCagAGGAGaTCAAA**ATG**
AAGACCGCG**CTG**ATCGCCAGATGCAGA

Figure 3.4.9: The first 75 bp of *asa* which correspond to the upstream region of the shortened Asa product detected by Western Blot. Legend: start codons in green letters, Shine Dalgarno sequence (ΔG° = -20.9 kJ/mol) in dark violet. Small letters in that region show mismatches to the consensus sequence (TAGGAGGT). The Shine Dalgarno sequence was calculated after Ma, et al. (2002).

Asa is a disordered protein with indications of a function as enzyme

Unfortunately, there is not any information in pBLAST nr database, Pfam domain database or conserved domain database that gives a hint for a putative function of *asa*. However, HHblits has two hits with low significance to an uncharacterized protein in *Glossina austeni* (savannah tsetse fly, UniProt entry A0A1A9UKK0) and *Glossina pallidipes* (tsetse fly, UniProt entry A0A1A9Z0V2). The GO term of this protein is a 2-C-methyl-D-erythritol 2,4-cyclodiphosphate (MECP) synthase involved in terpenoid biosynthesis process. The proteins have both a protein sequence identity of ~ 30% to the EHEC protein within the matching region. The coverage of the matching region was 77 and 68, respectively, of 87 amino acids and the e-values were 0.049 and 0, respectively. These hits could be a highly diverged version of Asa, originated from the EHEC synthase or from a synthase ancestral to the tsetse fly protein.

The prediction program PredictProtein is able to obtain information about protein features based on amino acid sequence motives (Figure 3.4.10). The *asa* encoding protein is predicted to be predominantly disordered (91% disordered amino acids), but there is one larger α-helix from amino acid 11-26 which is absent in the shortened version of Asa. There is one predicted disulfide bond (amino acid 23-29) and one transmembrane region (amino acid 26-42) which are both present in the shortened protein. Further, Asa is predicted to be secreted.

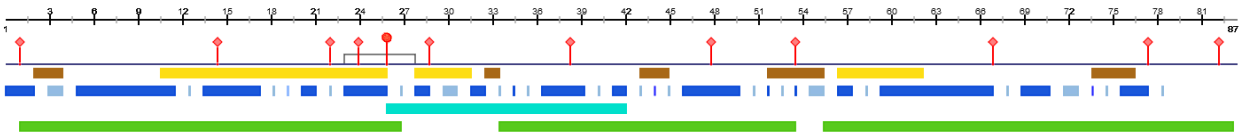


Figure 3.4.10: Predicted features of Asa. From top to bottom: The amino acid position is shown by the number line. Protein binding sites (red hash), DNA binding sites (red square), disulfide bond (transparent box), helices (yellow box) and beta sheets (brown box) show the secondary structure, buried (blue) and exposed (bright blue) amino acids show the solvent accessibility, transmembrane helices (turquoise), disordered regions (green).

3.5 Evolutionary analysis of the novel overlapping gene *asa*

Homology of an overlapping ORF to an annotated protein in the database is evidence for functionality (section 3.1) and a structural analysis of a large number of such sORFs showed, that, overall, structural features of sORFs are compatible with a potential protein coding capacity (section 3.2). However, homology and structural features provide both only circumstantial bioinformatics evidence for functionality. Therefore, as a case study, an individual sORF with a known phenotype was experimentally characterized in detail (section 3.4). An overall phylostratigraphic analysis yielded evidence that many sORFs may have originated only after the separation of *Escherichia* clade from other enterobacteriaceae (section 3.1). However, in order to learn more about the evolution of shadow ORFs, a phylostratigraphic analysis of sORFs is necessary whose functionality has been firmly established by experimental evidence. Therefore, as a first example, the evolution of the overlapping gene *asa* whose functionality has been demonstrated in the previous chapter was investigated in depth.

3.5.1 Bioinformatics

Sequence evolution

The gene *asa*, which was experimentally characterized (section 3.4), is completely embedded in EDL933_1238, a TEGT transporter gene (Figure 3.4.2). The *asa* evolution was analysed by bioinformatics and laboratory experiments of *asa* homologues. The sequence homology search with tblastn reveals 1211 (E^{-10}) hits for *asa* and 3487 hits for the mother gene. The farthermost hits of *asa* belong to the order enterobacteriales. Altogether, 69 species have a homolog of *asa*. Of those, species with the highest number of strains are *Escherichia coli* (423 with *asa*), *Salmonella enterica* (428 with *asa*) and *Klebsiella pneumoniae* (140 with *asa*).

The mother gene is more conserved and the taxonomic distribution is more balanced: a few hits (0.1%) belong to Eukaryotes, with the remaining hits to bacteria in the following composition: 37% enterobacteria, 28% γ-proteobacteria, 34% proteobacteria, 1% in bacteria more distantly related to EHEC than proteobacteria. All bacteria with putative *asa* homologues are Gram-negative and there are no hits to Archaea.

A phylostratigraphic tree shows that most of the sequences (75%) are within the phylostratigraphic level ‘γ-proteobacteria’, which is the phylostratigraphic level of *asa,* as the farthermost related species with full-length *asa* is *Moritella viscosa* (Figure 3.5.1).

The sequence evolution is gradual and even eukaryote sequences have an identity to *asa* higher than 30% (tblastn) despite being not intact (Figure 3.5.1). This gradual evolution is also reflected

in the mother gene sequence (Figure 3.5.2) and the identities of the sORF homologues to EHEC (Figure 3.5.3). The mORF constraint is higher than average of the negative control and the sequence identities of *asa* homologues to *asa* in EHEC is higher than that of the negative control in all homologues from EHEC up to γ-proteobacteria and significantly decreases in farther related species (Supplementary figure S6). There are a few positions in *asa* (for example amino acid 45, proline) which are conserved throughout the whole tree (Figure 3.5.1). The start codon position is highly conserved even in sequences containing internal stop codons. Only few species distantly related to EHEC have altered start codon positions. The stop codon position found in EHEC is more variable in the homologues. It is conserved in all *Escherichia* and most more distantly related enterobacteriaceae, but not beyond.

Stop codon at amino acid position 62

Interestingly, tblastn reveals that most sequences have a stop codon at amino acid position 62. This stop codon is widely distributed independent from the phylogenetic clade. When looking on the species level, 69,6% of all species with a tblastn hit to *asa* have this stop codon. Of those species, all strains have the stop codon. There are further species in which some strains have an intact and some a non-intact sequence (8,7%, *Cedecea neteri*, *Citrobacter freundii, Enterobacter sp., Serratia sp.*). The remaining species (21,7%) only have intact sequences: all species and strains of the genera *Escherichia* and *Shigella*, as well as all strains of *Kluyvera intermedia, Kosakonia sacchari, Citrobacter rodentium, Citrobacter werkmanii* and *Sodalis praecaptivus.* Interestingly, *Salmonella enterica*, the closest relative of *Escherichia,* has this internal stop and it is present in all strains of *Salmonella*. Homologues having more than this stop codon at position 62 are found first in the phylostratigraphic level γ-proteobacteria (Figure 3.5.1).

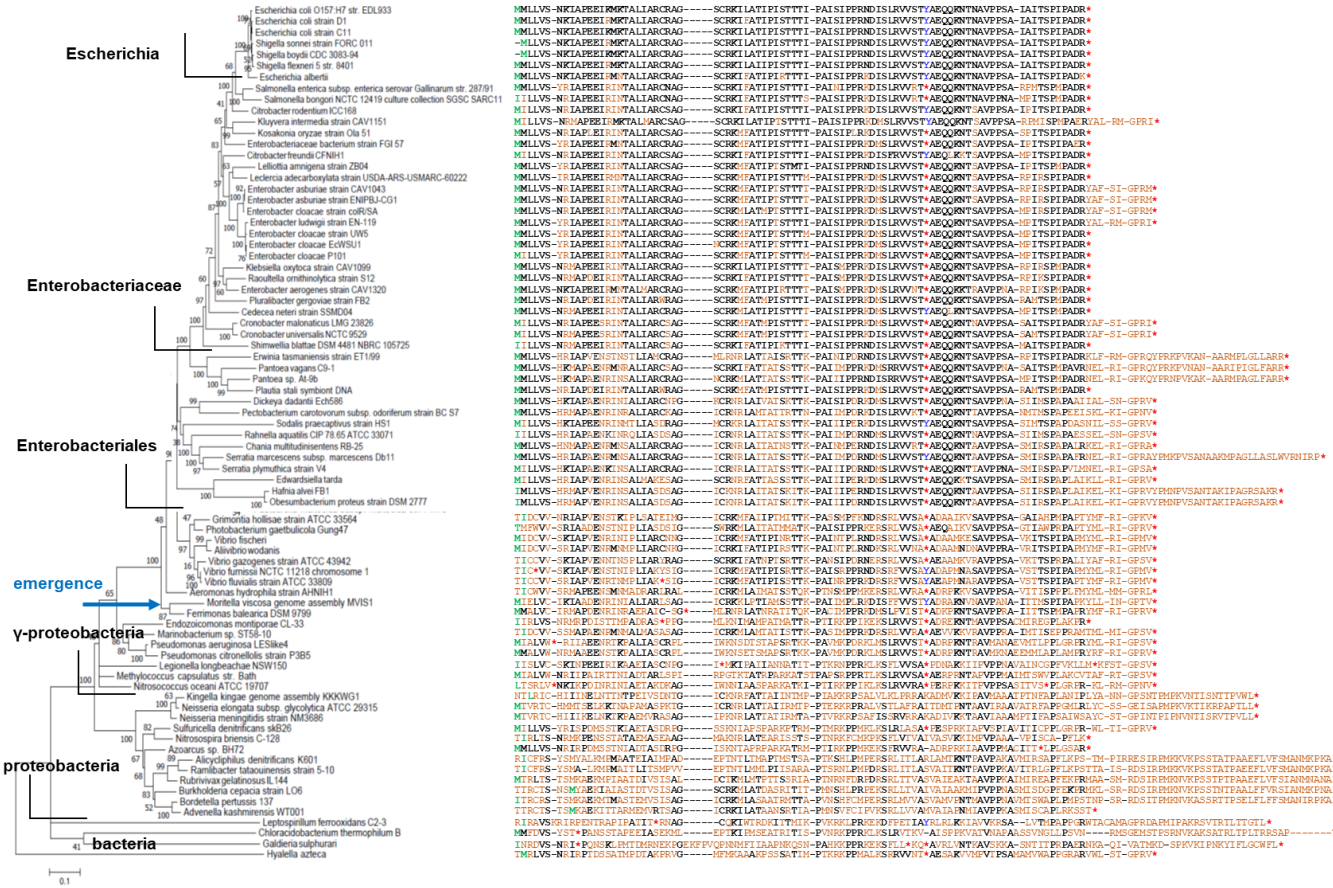


Figure 3.5.1: Phylostratigraphy of *asa*. The species tree is constructed from organisms in which the mORF homologue was found. The respective *asa* homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The “Y” (tyrosine) at amino acid position 62 present in intact homologues is shown as blue letter. The species tree was constructed from MLSA sequences within enterobacteriales and 16S rRNA sequences in more distantly related species. It is a neighbor joining tree calculated with Mega7 (Kumar, et al. 2016).

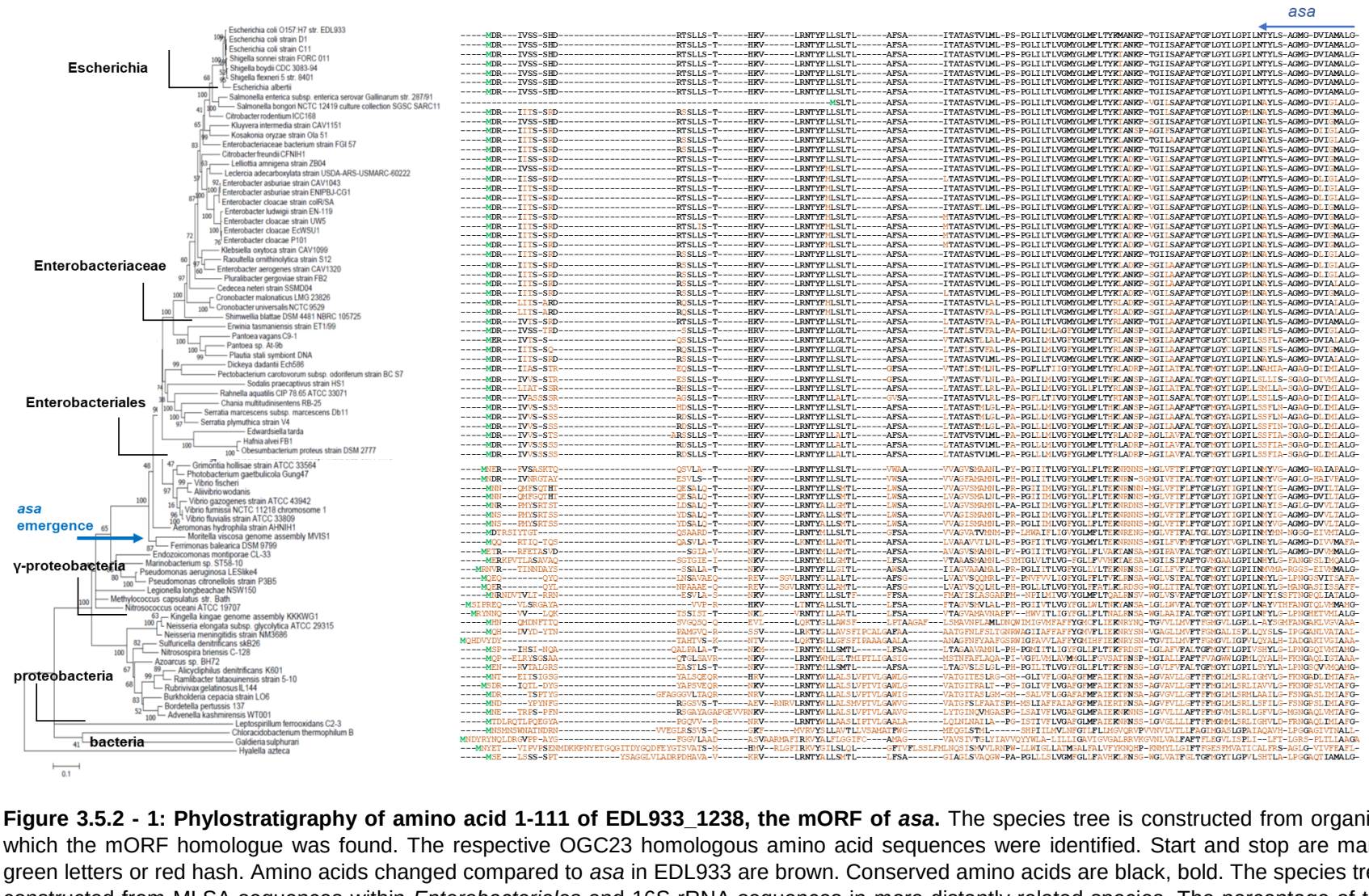


Figure 3.5.2 - 1: Phylostratigraphy of amino acid 1-111 of EDL933_1238, the mORF of *asa*. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales* and 16S rRNA sequences in more distantly related species. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7 (Kumar, et al. 2016).

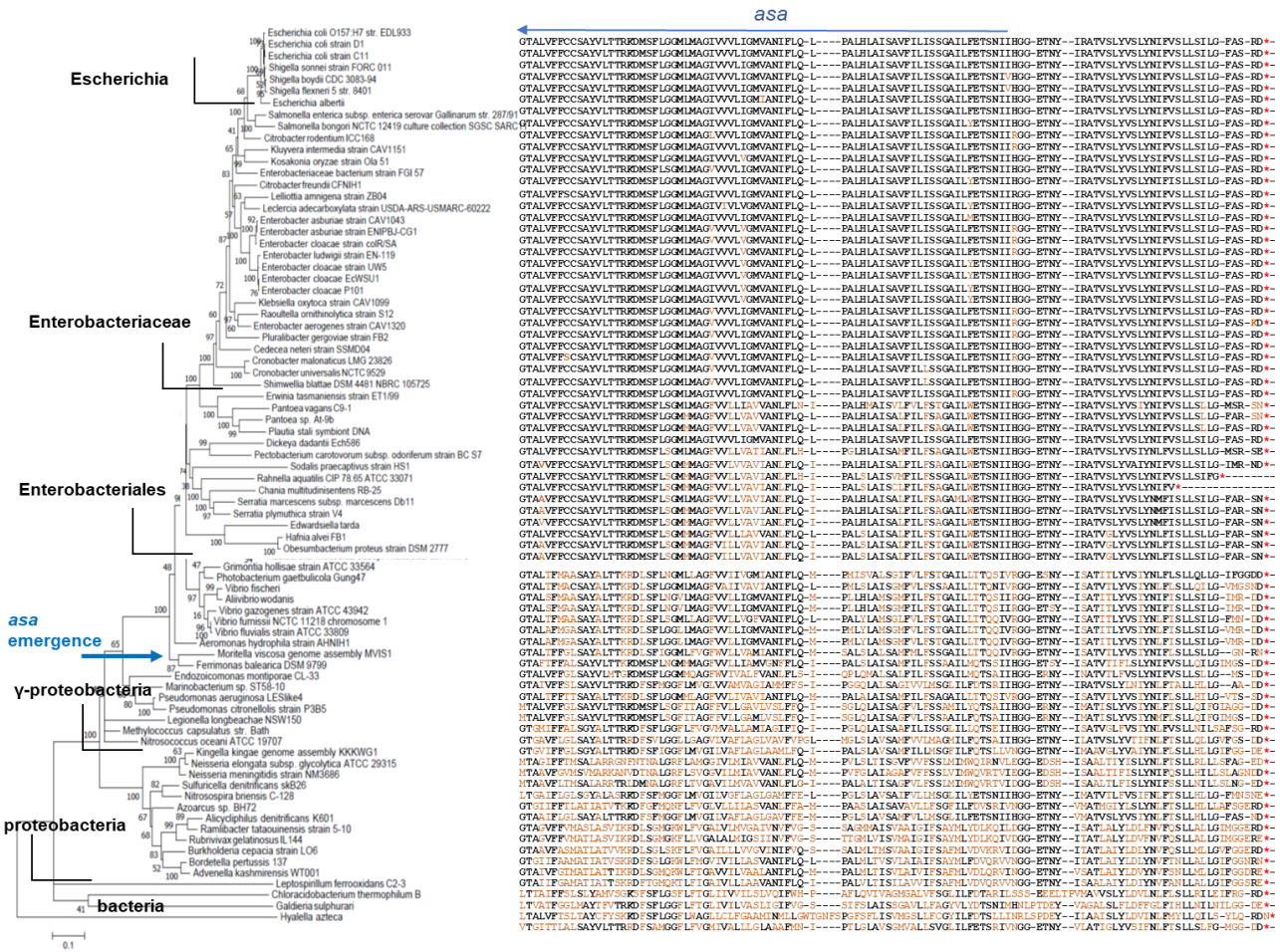


Figure 3.5.2 - 2: Phylostratigraphy of amino acid 112-220 of EDL933_1238, the mORF of *asa*. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales* and 16S rRNA sequences in more distantly related species. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7 (Kumar, et al. 2016).

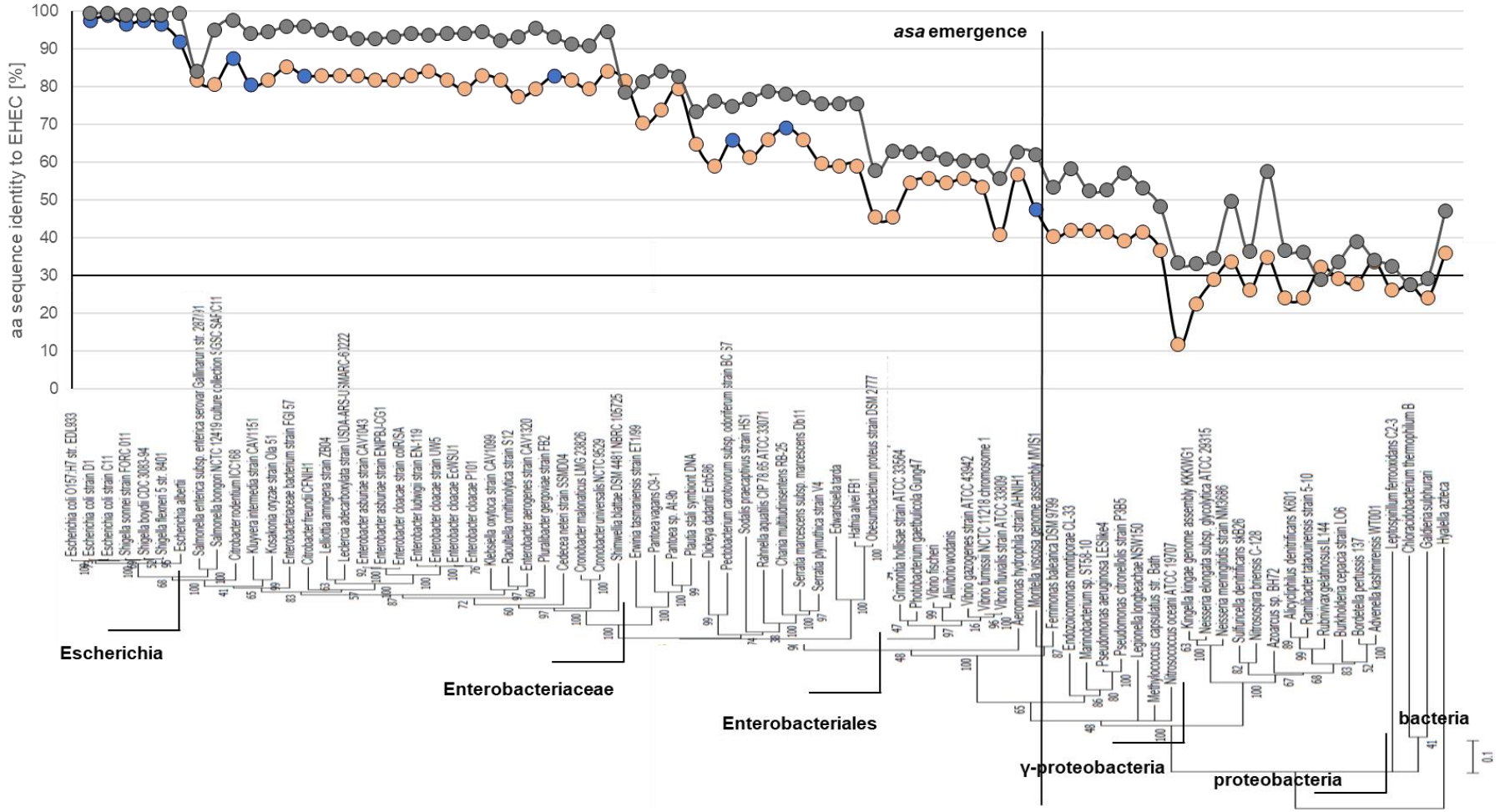


Figure 3.5.3: Amino acid (aa) sequence identities to EHEC [%] of the overlapping gene pair *asa*/EDL933_1238 during species evolution. Legend: graph on top: Sequence identity to EHEC of each homologue shown in the tree below. grey: EDL933_1238 homologues, blue: intact *asa* homologues, salmon: *asa* homologues with internal stop codons; the sequence homology cutoff of 30% is shown as black line. Graph below: Maximum likelihood tree of species containing EDL933_1238 homologues. The tree was constructed as combination of MLSA tree (Escherichia to Enterobacteriales) and a 16S rRNA tree (γ-proteobacteria and further related). The tree was calculated using MEGA 7 (Kumar, et al. 2016).

Evolutionary constraint

The gradual evolution of *asa* and EDL933_1238 was verified by d_N / d_S analysis in cooperation with Chase Nelson (American Museum of Natural History, New York). The ratio of nonsynonymous (d_N) and synonymous (d_S) mutations is a measure for the evolutionary constraint of a sequence. A pattern of $d_N < d_S$ is expected when purifying selection acts to prevent the accumulation of nonsynonymous mutations which are presumably more likely to have deleterious phenotypic effects than synonymous changes. The evolution of overlapping genes is thought to be further constrained by the interdependence of mutations occurring in both genes. For these analyses, a preliminary tree version was used containing 41 instead of later 81 sequences (list of species), whereas 66% of the sequences had the internal stop at position 62. To find out more about the evolution of the highly conserved stop codon, d_N / d_S of *asa* was calculated separately for homologues with and without premature terminations and for segments upstream and downstream internal stop codon position 62 (Table 3.5.1). It was hypothesized that evolutionary constraints would be higher for intact ("without internal stop") sequences in comparison to non-intact ("with internal stop") sequences in the second (downstream) segment. All segments of both types exhibited significant purifying selection, with $d_N < d_S$ in each case, and an overall d_N / d_S ratio of 0.342 ($P < 0.001$, Bonferroni correction; Table 3.5.1). Although the d_N / d_S ratio was indeed higher (less constraint) for both segments of the genes with an internal stop, the difference was not significant in either segment ($p > 0.6$). A 10-codon sliding window along *asa* and its antisense mother gene revealed similar profiles of nonsynonymous and synonymous substitutions, indicating that the two genes do not exhibit differing patterns of selective constraint (Figure 3.5.4).

Table 3.5.1: dN / dS ratio of genes with internal stop, without internal stop and the segments upstream (segment 1: nucleotide 1-186) and downstream (segment 2: nucleotide 187-264) of the internal stop. Lower ratios show a higher evolutionary constraint indicating a functionality.

Group	Number of sequences	dN / dS		
		segment 1	segment 2	complete sequence
With internal stop	27	0.327***	0.299***	0.317***
Without internal stop	14	0.319***	0.273***	0.302***

***$p < 0.001$ after a Bonferroni correction for multiple comparisons for a test of the null hypothesis d_N=d_S.

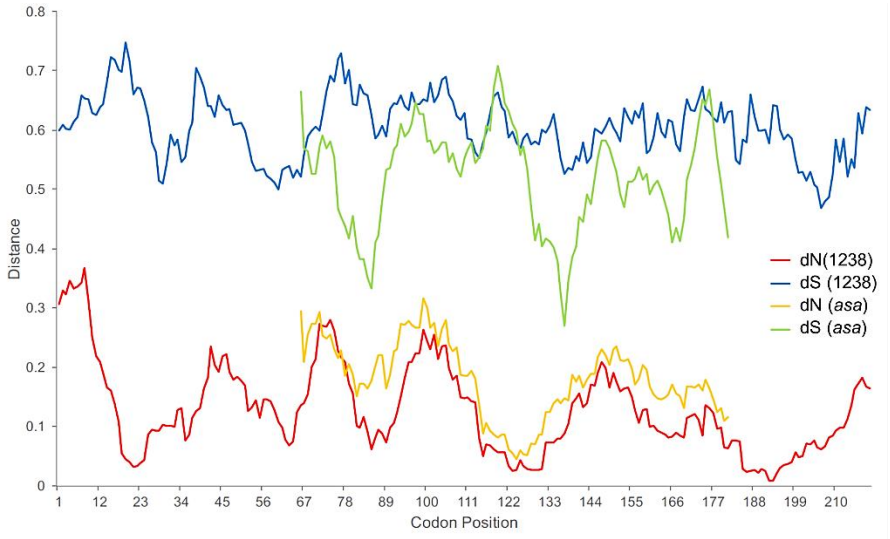


Figure 3.5.4: Mean d_N and d_S ('distance') in a sliding window of 10 codons (step size 1) for homologues of *asa* and its mother ORF (1238). Codon positions refer to the mother ORF. For *asa* comparisons, d_N and d_S are shown as orange and green, respectively, while for the mother gene comparisons, they are shown as red and blue, respectively. The *asa* sequence starts in codon 190 of the mother gene.

3.5.2 Experiments

Homologues with RNAseq and RIBOseq signal

To get information about the functional evolution of *asa,* RNAseq and RIBOseq data of *asa,* homologues were obtained from our lab (*E. coli* O157:H7 strain Sakai - EHEC Sakai, *E. coli* LF82 = AIEC, Adherent / Invasive *E. coli*) and from the SRA database of NCBI (Leinonen, et al. 2011). The raw data were processed and mapped by Zachary Ardern and visualized with Artemis 17.0. The homologues downloaded having both RNAseq and RIBOseq data were *E. coli* K12 MG1655, *E. coli* K12 MC4100, *Salmonella enterica* strain 14028S and *Salmonella enterica* strain SL1344. Homologues with RNAseq data only were *E. coli* E2348 (= EPEC / enteropathogenic *E. coli)*, *Shigella flexneri* 5a M90T, *Citrobacter rodentium* ICC168*, Sodalis praecaptivus* HS1, *Serratia marcescens* WW4, *Enterobacter aerogenes* KCTC2190, *Klebsiella pneumoniae* subsp. *pneumoniae* MGH78578 and *Cronobacter sakazakii* ATCC BAA-894. All bacteria are taxonomically located in the order enterobacteriales. The list of all bacteria and their SRA accession numbers can be found in Supplementary table S7.

The genome region around the overlapping gene pair *yccA* / *asa* of all homologues was schematically visualized (Figure 3.5.5). The region downstream of the mother gene is conserved in all homologues: the same strand gene *yccK* encodes a sulfurtransferase (TusE) which is only absent in *S. flexneri*. Downstream of *yccK* is the gene *yccX,* encoding an acyl-phosphate

phosphohydrolase in *antisense*. The upstream region is more variable. Most species have a serine tRNA in *sense* (AIEC LF82, *E. coli* K12 MG1655, *E. coli* K12 MC4100, EPEC E2348, *S. flexneri*) or in *antisense* (*S. enterica, C. rodentium, S. praecaptivus* and *S. marcescens*). The closest relative to EHEC EDL933, EHEC Sakai, also has a phage integrase gene (ECs1055). All other *E. coli* strains and *Shigella* have a gene encoding hydrogenase 1 (*hyaA*) in *sense* in the upstream region. *Salmonella enterica* has a *pipA* gene in the upstream region encoding the pathogenicity island protein A. *Citrobacter* and *Serratia* have genes not-further-specified and *Enterobacter, Klebsiella* and *Cronobacter* have *oppA*, an oligopeptide ABC transporter in the upstream region.

The bacteria with a full-length ORF of *asa* were all *E. coli* strains, *Shigella, Citrobacter, Sodalis* or *Serratia* (Supplementary figure S7). The remaining strains have a stop codon at amino acid position 62 (*Salmonella, Enterobacter, Klebsiella and Cronobacter*). The start codon position of all *asa* homologues is highly conserved (Supplementary figure S7). The sequences found in *Citrobacter, Enterobacter* and *Cronobacter* could be extended upstream by six amino acids. Those open reading frames in *Sodalis* and *Serratia* were even nine or 37 amino acids longer, respectively, than *asa* present in EHEC caused by a downstream extension in both cases.

The Artemis screenshots of RNAseq and RIBOseq data of EHEC Sakai, AIEC LF82, *E. coli* MG1655, *E. coli* MC4100, *S. enterica* 14028S and *S. enterica* SL1344 can be found in Figure 3.5.6 and those of species with available RNAseq data only in Figure 3.5.7 (EPEC E2348*,* Shigella flexneri 5a str. M90T, Citrobacter rodentium ICC168, Sodalis praecaptivus HS1, Serratia marcescens WW4; down: Enterobacter aerogenes KCTC2190, Klebsiella pneumoniae subsp. pneumoniae MGH78578, Cronobacter sakazakii ATCC BAA-894). The corresponding RPKM values can be found in Supplementary table S15. The signals of *E. coli* strains were varying. Expression occurs in pathogenic (EHEC Sakai, AIEC LF82 and EPEC E2348), but not in apathogenic strains (K12 substrain MG1655 or substrain MC4100). The *asa*-encoded protein Asa was detected in EHEC Sakai and AIEC LF82. RIBOseq data of EPEC are not available so far. Interestingly, the RNAseq expression pattern of EPEC is significantly different from that of EHEC EDL933, EHEC Sakai or AIEC LF82. The expression is much stronger and encompasses a region as large as the mother gene. This pattern is similar to that in *Shigella, Citrobacter* and *Sodalis*. Truncated homologues partly show *antisense* transcription (*S. enterica* SL1344, *Cronobacter*) and some have no signal for expression (*S. enterica* 14028S, *Enterobacter* and *Klebsiella*). As expected, the RIBOseq data of *S. enterica* SL1344 and 14028S show that a truncated ORF is not translated (Figure 3.5.6), which can be taken to be strong confirmation of our account.

Those enterobacteria having a RNAseq or RIBOseq signal, were checked for sequence conservation in the promoter region of *asa*. The σ^{70} promoter, belonging to the -186/-188 bp +1 site and predicted by BProm, is only conserved in *E. coli* and in *S. flexneri* (Supplementary figure S8 A). The average frequencies of σ^{38} promoter sequence conservation were calculated following Figure 1 in Lee and Gralla (2001) and can be found in Table 3.5.2. Lee and Gralla (2001) identified the consensus sequence of σ^{38} dependent promoters in an electrophoretic mobility shift assay. A broader range of species have nucleotides matching to the consensus sequences of the σ^{38} promoter, which may indicate functionality (Supplementary figure S8 B and C). The majority of the average conservation frequencies was lower than 50% which is rather low in comparison to the maximum conservation of a σ^{38} promoter consensus sequence (70%, consensus sequence CTACACT). There are only very few putative promoter regions with an average conservation of 0% and are, thus, unlikely be a functional σ^{38} promoter. Interestingly, the sequence at the site in which the σ^{38} promoter of -186/-188 bp +1 site in EHEC EDL933 was found, AIEC LF82 cannot have such a promoter due to an insertion of the nucleotide “C” (which could also be a sequencing error). The presence or absence of a promoter is independent from sequence intactness or from expression signals observed in the section above.

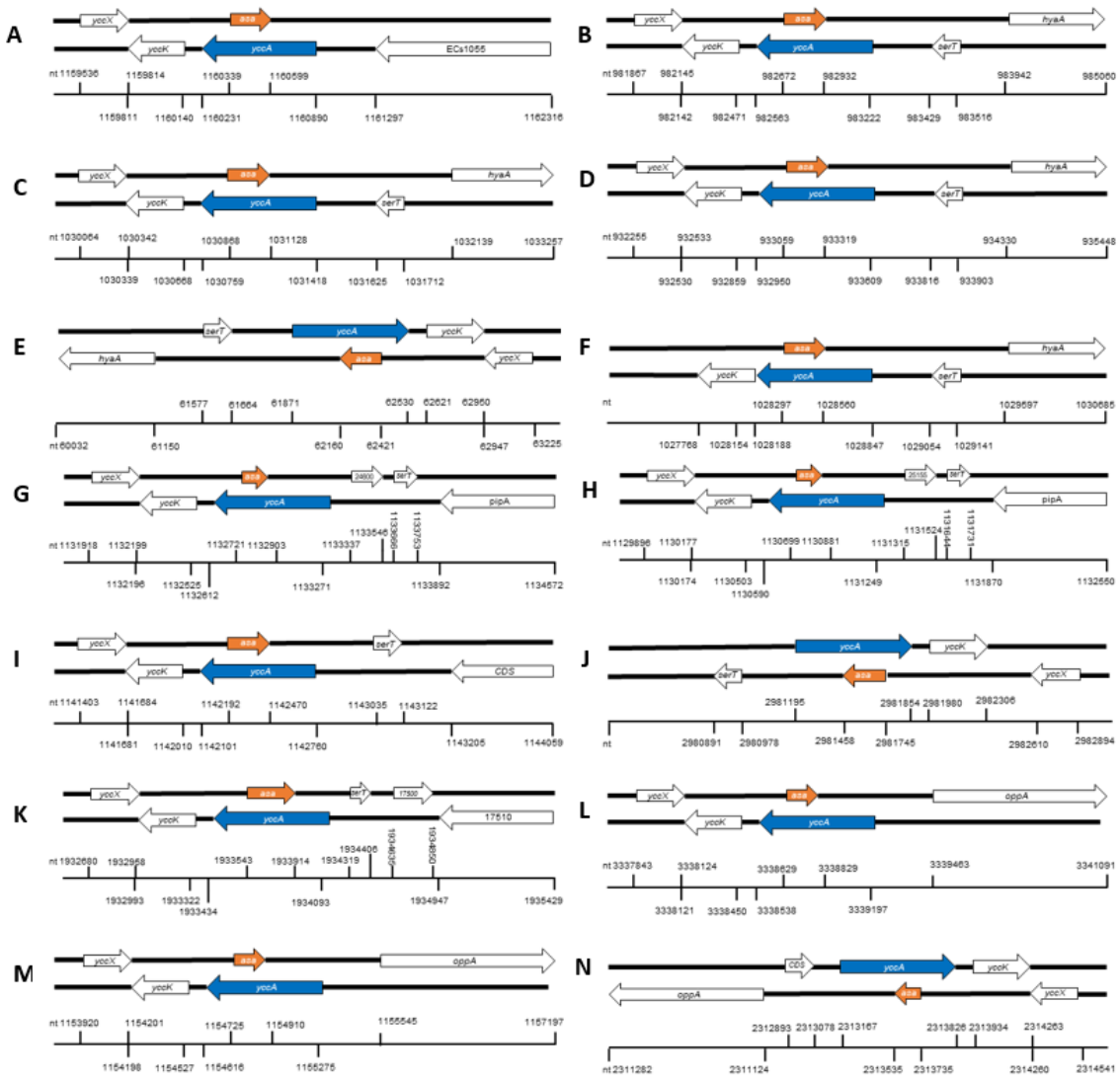


Figure 3.5.5: Gene region around *asa* homologues in species analyzed for RNAseq and/or RIBOseq signals. The following organisms are shown: (A) EHEC Sakai, (B) *E. coli* LF82, (C) *E. coli* K12 MG1655, (D) *E. coli* K12 MC4100, (E) EPEC, (F) Shigella flexneri M90T, (G) *S. enterica* 14028S, (H) *S. enterica* SL1344, (I) *Citrobacter rodentium* ICC168, (J) *Sodalis praecaptivus* HS1, (K) *Serratia marcescens* WW4, (L) *E. aerogenes* KCTC2190, (M) *K. pneumonia subsp. pneumoniae* MGH78578, (N) *Cronobacter Sakazakii* ATCC BAA-894. Same gene names encode the same product, but they are not necessarily the names used in the respective genome of the organism.

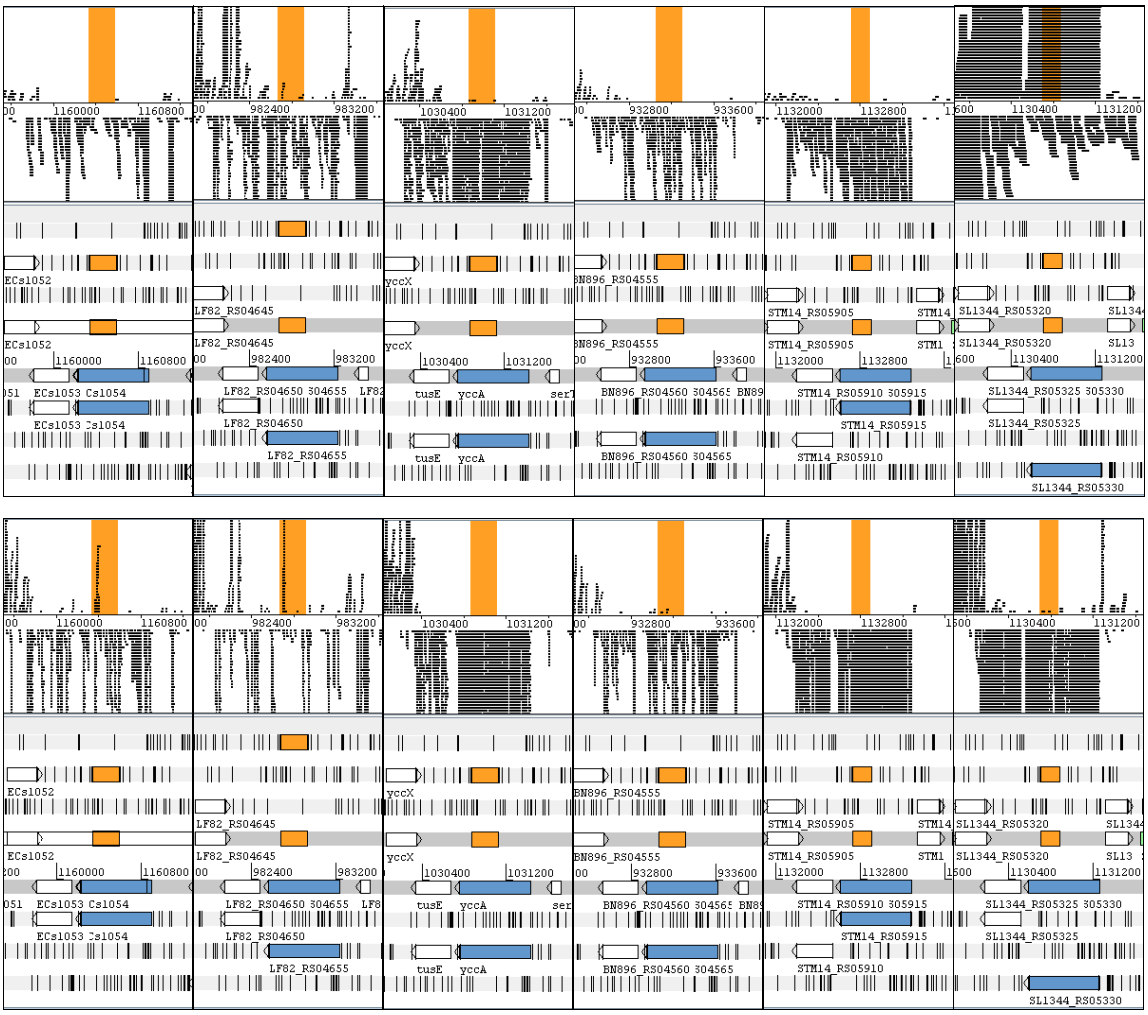


Figure 3.5.6: RNAseq (top) and RIBOseq (down) signals of *asa* homologues. From left to right: EHEC Sakai, *E. coli* LF82, *E. coli* K12 subsp. MG1655, *E. coli* K12 subsp. MC4100, *S. enterica* 14028S and *S. enterica* SL1344. The *asa* homologue is highlighted in orange, the mother gene in blue. All pictures are visualized with Artemis 17.0 (Carver, et al. 2011).

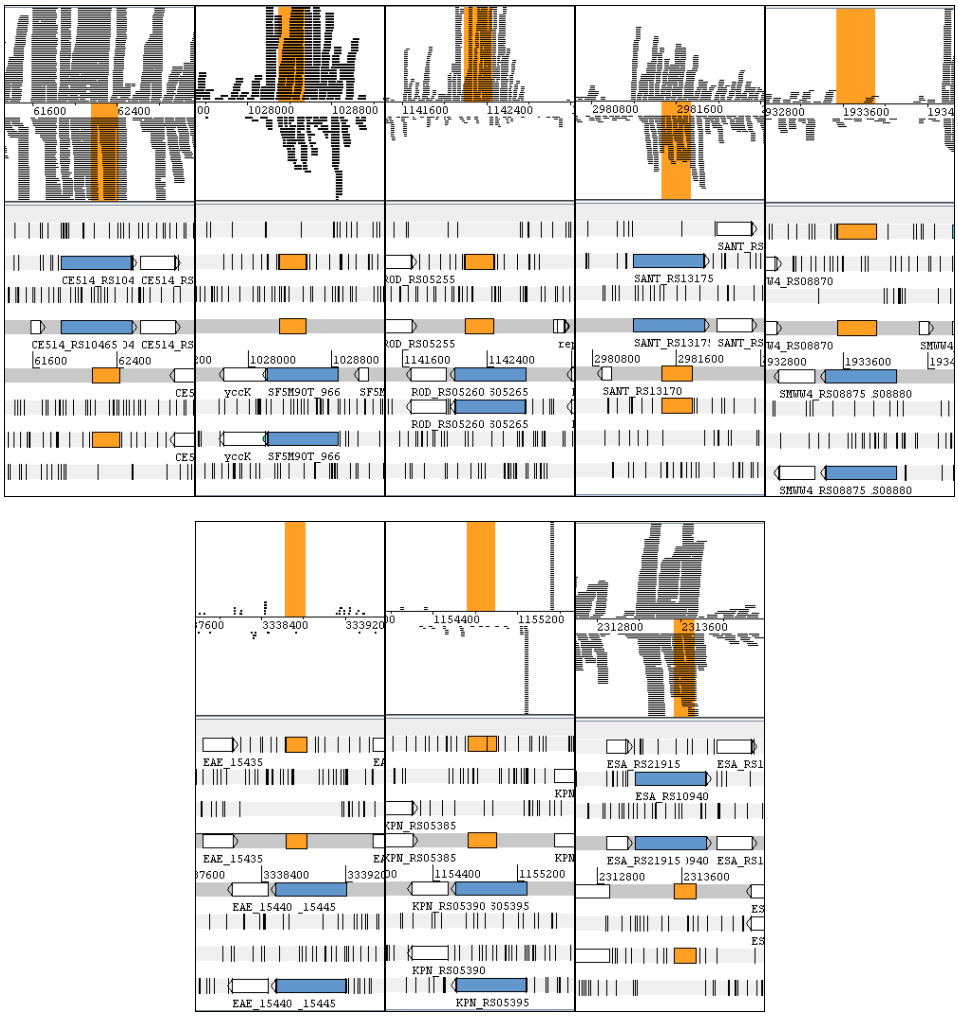


Figure 3.5.7: RNAseq signals of *asa* homologues. Top (from left to right, all intact): EPEC, *Shigella flexneri* 5a str. M90T, *Citrobacter rodentium* ICC168, *Sodalis praecaptivus* HS1, *Serratia marcescens* WW4; down (all not intact): *Enterobacter aerogenes* KCTC2190, *Klebsiella pneumoniae* subsp. pneumoniae MGH78578, *Cronobacter sakazakii* ATCC BAA-894. The asa homologue is highlighted in orange, the mother gene in blue. All pictures are visualized with Artemis 17.0 (Carver, Harris et al. 2011).

Table 3.5.2: Average frequencies of σ^{38} sequence conservation of *asa* homologues. All frequencies were calculated after Figure 1, Lee and Gralla (2001).

Organism	σ^{38} -188/-186 bp +1 site	σ^{38} -178 bp +1 site
EHEC EDL933	38%	59%
EHEC Sakai	38%	59%
AIEC LF82	0%	59%
E. coli MG1655	38%	59%
E. coli MC4100	38%	59%
S. enterica 14028S	40%	47%
S. enterica SL1344	40%	47%
EPEC E2348	38%	59%
S. flexneri M90T	38%	59%
C. rodentium ICC168	0%	27%
S. praecaptivus HS1	38%	0%
S. marcescens WW4	23%	0%
E. aerogenes KCTC2190	47%	50%
K. pneumoniae MGH78578	43%	32%
C. sakazakii ATCC BAA-894	0%	0%

Characterization of further homologues

Homologs of *asa* for four species within the order enterobacteriales were characterized using experiments and settings to compare them with *asa* of EHEC. Two of the strains (*Citrobacter freundii* CFNIH1, and *Serratia marcescens* WS1359) had a full-length *asa* homologue (Supplementary figure S9), but two other strains (*Salmonella enterica* serovar Gallinarum 287/91, *Hafnia alvei* DSM30097) had truncated *asa* homologues. As 'intactness' was broadly distributed in the phylostratigraphic tree (section 3.3 Figure 3.5.1), one closely related species (*Citrobacter freundii*, *Salmonella enterica*) and one more distantly related species (*Serratia marcescens*, *Hafnia alvei*) were selected with each one intact and one truncated *asa* sequence. The sequence similarities to *asa* are significant and only slightly decreasing with distance to EHEC (Supplementary table S16).

The genome organization of experimentally characterized *asa* homologues is schematically visualized in Figure 3.5.8. As there are no genome sequence data available in Genbank for *Serratia marcescens* WS1359 and *Hafnia alvei* DSM30097, the genomes of strain *Serratia marcescens* Db11 *and Hafnia alvei* FB1 were selected as surrogates. However, the sequences shown in Supplementary figure S9 correspond to the *asa* homologue in the original organism as the open reading frame was cloned in pBAD-*myc*/His C and sequenced, as described in the following section. The genome region of the species with RNAseq or RIBOseq signal is comparable. The downstream region of the transcribed or translated homologues is always conserved; the upstream region varies. *Salmonella enterica* 287/91 has an upstream region

homologous to that in *Salmonella enterica* 14028S or SL1344. As in all other species, *Hafnia alvei* has a gene upstream of *asa* encoding for a serine tRNA and upstream of the serine tRNA gene lays the gene *kdgM* encoding a oligogalacturonate-specific porin protein. Interestingly, *Citrobacter freundii* CFNIH1 has an upstream region different from *Citrobacter rodentium,* but similar to *Enterobacter, Klebsiella or Cronobacter*. Further, *C. freundii* CFNIH1 is the only strain of *Citrobacter freundii* with full-length *asa*. To ensure that the analyzed organism was in fact *Citrobacter freundii*, it was confirmed by 16S rRNA Sanger sequencing. The σ^{70} promoter region of all species tested in this analysis is not conserved (Supplementary figure S10). This is in correspondence to observations of those homologues with available RNAseq and RIBOseq data (Supplementary figure S8). The average conservation frequencies of the σ^{38} promoters can be found in Table 3.5.3.

Competitive growth experiments were repeated with the above introduced homologues in analogy to the experiments conducted for *asa* from EHEC EDL933. As the clearest phenotype of *asa* was observed in NaCl (Figure 3.4.3), this condition was used for all experiments. The pBAD-myc/His C *asa* homologue construct and a translational arrested mutant (stop codon sites see Figure 2.3) were tested both in EHEC and in the respective organism.
As the phenotype was caused by translational arrest of the protein, it was hypothesized that the truncated homologues should not cause a phenotype, whereas the full-length homologues should show a phenotype with decreasing intensity with increasing distance to EHEC. Unfortunately, most of the results were conflicting and sometimes not reproducible, and I could not find any plausible cause for this (Figure 3.5.9). Homologues of *C. freundii* (*asa*CF), *S. enterica* (*asa*SE) and *H. alvei* (*asa*HA) tested in EHEC and that of *S. marcescens* (*asa*SM) tested in *Serratia* showed all possibilities (growth advantages, disadvantages and no phenotypes) in different biological replicates without any clear pattern. Therefore, *asa*CF was tested in eighteen biological replicates, which resulted in a nearly equal and continuous distribution of all possibilities, ranging from strong growth advantage of the translationally arrested mutant to a strong growth disadvantage (Figure 3.5.8 9). Interestingly, there was a stable growth disadvantage of the wild type *asa*SM tested in EHEC, which was the expected phenotype according to the results found in EHEC before (Figure 3.5.9 D). Stable 'non-phenotypes' were found for *asa*CF tested in *C. freundii, asa*SE tested in *S. enterica* and *asa*HA tested in *H. alvei*. Results obtained for *asa*SE and *asa*HA confirmed the hypothesis that truncated *asa* would not cause a phenotype. However, in the light all results, this reproducible 'non-phenotype' cannot be considered as significant.

The four strains with *asa* homologues tested here have no available RNAseq or RIBOseq data. Thus, gene expression was detected by RT-qPCR. Cultures of the bacteria, grown in LB-medium, were harvested during exponential phase (OD_{600} = 0.8 - 0.9). The growth curves can be found in Supplementary figure S11. The experiment was implemented analogously to the RT-qPCR of *asa*. All standard curves for the primers used can be found in Supplementary figure S2 and were conducted with genomic DNA of each organism. The primer efficiencies can be found in Supplementary table S6. The cq values of the control for absence of DNA (qPCR of 16S rRNA without reverse transcription) were ≥26 which was ≥10 cq more than those with reverse transcription (Supplementary table S17, ratio DNA:RNA = $1:2^{10}$ = 1014).

All homologues are expressing the *asa* homologous region (Figure 3.5.10). The mRNA titer of *asa*CF, *asa*SM and *asa*SE were comparable to that of *asa* (Supplementary table S17 and S14). The high cq of *asa*HA was caused by insufficient primers, which is visible in the standard curve. The y-intercept is much higher than that of others (Supplementary figure S2).

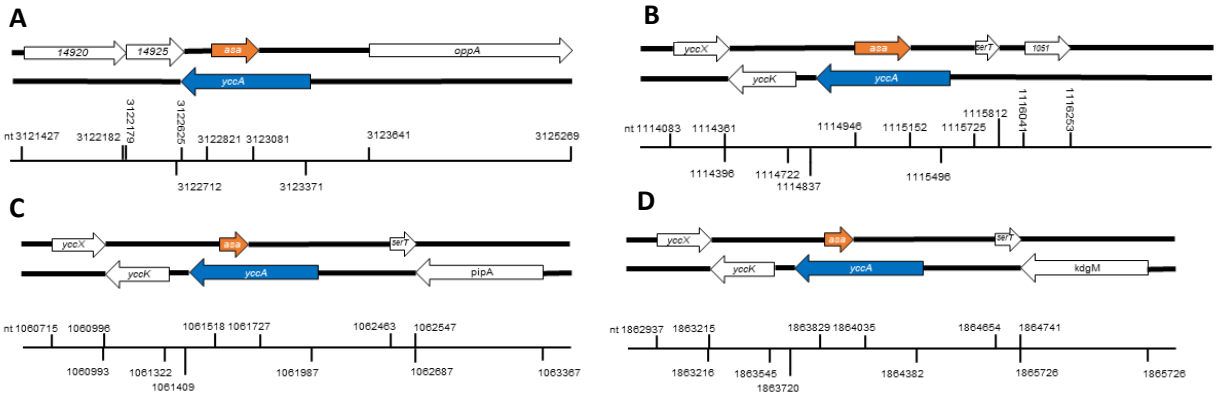


Figure 3.5.8: Synteny of experimentally characterized *asa* homologues. Coordinates below each panel are given in respect to the respective genome. The following organisms are shown: (A) *Salmonella enterica* serovar Gallinarum 287/91, (B) *Citrobacter freundii* CFNIH1, (C) *Serratia marcescens* DB11, (D) *H. alvei* FB1. For those organisms without sequenced genome (*Serratia marcescens* WS1359 and *H. alvei* DSM30097) the genome of a strain of the same species was used. Same gene names encode the same product, but are not necessarily the names used in the respective genome of the organism.

Table 3.5.3: Average frequencies of σ^{38} sequence conservation of experimentally characterized *asa* homologues. All frequencies were calculated after Figure 1, Lee and Gralla (2001).

Organism	σ^{38} -188/-186 bp +1 site	σ^{38} -178 bp +1 site
EHEC EDL933	38%	59%
C. freundii CFNIH1	0%	46%
S. marcescens Db11	0%	0%
S. enterica 287/91	40%	47%
H. alvei FB1	0%	0%

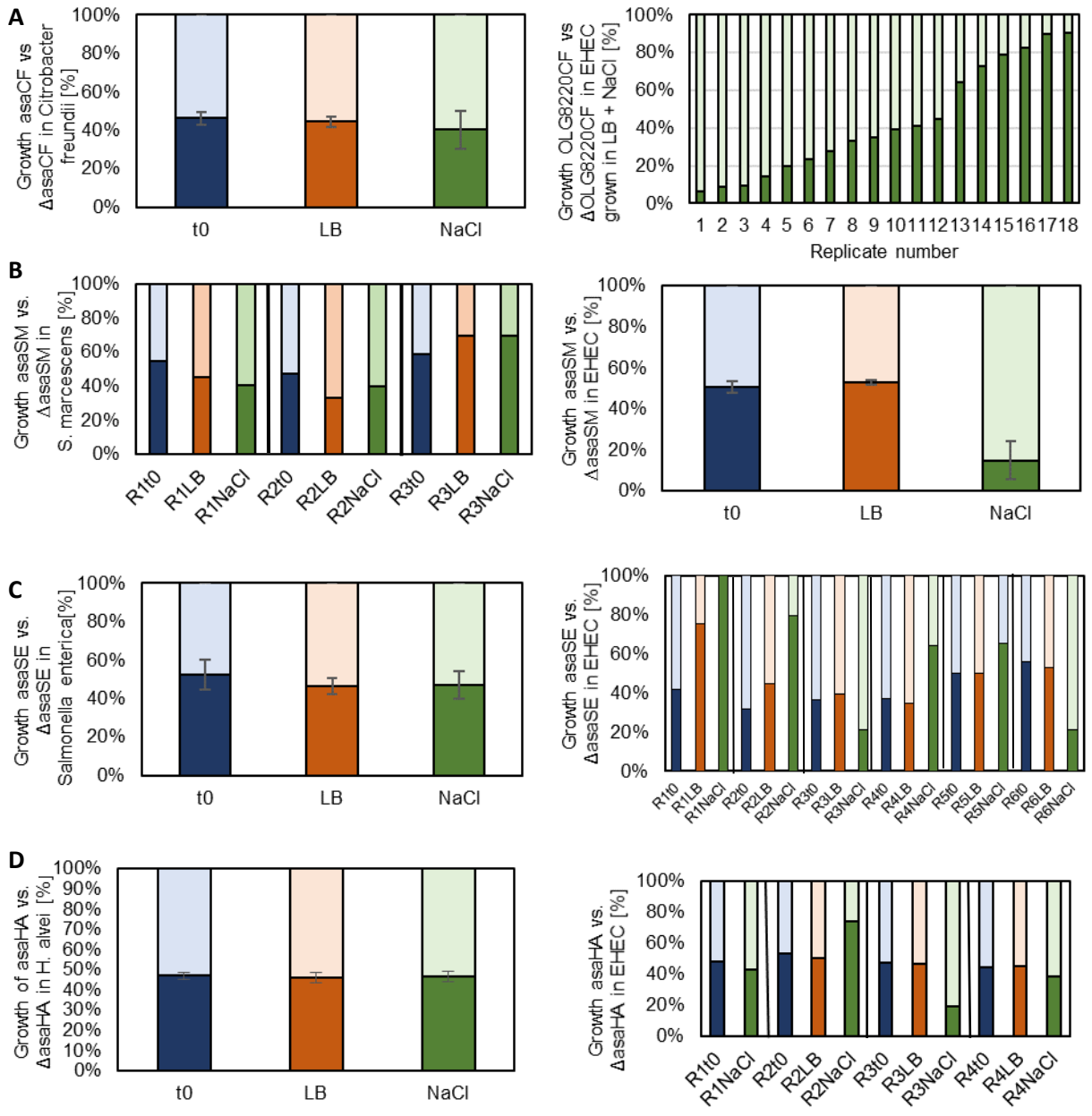


Figure 3.5.9: Overexpression phenotypes of *asa* homologues shown as ratio of wild type:mutant [%]. The respective organism (left) or EHEC (right) overexpressing the *asa* homologue (dark color) grew competitively against the respective organism overexpressing a translational arrested mutant of the homologue Δ*asa* (bright color). The following organisms were tested: (**A**) *Citrobacter freundii* CFNIH1, (**B**) *Serratia marcescens* WS1359, (**C**) *Salmonella enterica* serovar Gallinarum 287/91, (**D**) *Hafnia alvei* DSM30097. The sequences in *Citrobacter* and *Serratia* are intact, those in *Salmonella* and *Hafnia* are truncated. The condition tested was LB supplemented with 450 mM NaCl (green). LB were used as negative control (orange). For overexpression, *asa* or the mutant were cloned in pBAD-myc/His C. The initial wildtype:mutant ratio was 1:1 (t0, blue). The gene expression was induced with 0.002% arabinose (w/v); the cells were harvested after 22 h. The plasmid was sequenced by sanger sequencing and the phenotype was determined by the peak height ratio of wild type and mutant at the position where the stop codon was introduced. The plots show the mean value and standard deviation of three biological replicates (R). The competitive growth of the *C. freundii* homologue in EHEC was conducted altogether 18-times to see a trend of this unstable phenotype. The controls of this experiment (LB and t0, not shown) were negative.

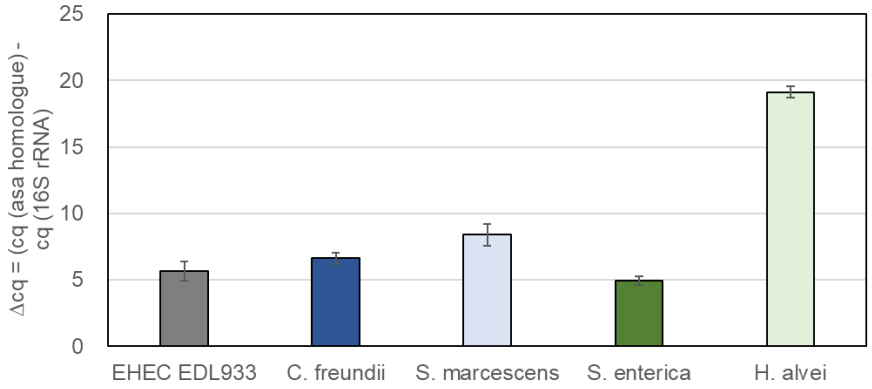


Figure 3.5.10: Messenger RNA titer of *asa* and of *asa* homologues shown as threshold cycles (Δcq) obtained by RT-qPCR. The threshold cycles of the ORFs were normalized to the respective 16S rRNA of the species. Lower Δcq values indicate higher amounts of mRNA and *vice versa*. The organisms tested (EHEC EDL933, *C. freundii*, *S. marcescens*, *S. enterica* and *H. alvei*) are grown in LB medium and harvested during exponential phase (OD_{600} = 0.8). The plot shows the mean values and standard deviations of three biological replicates.

3.6 Phylostratigraphic analysis of the overlapping genes *laoB, ano, slyC* and OGC106

A couple of phenotypes of sORFs have been discovered by Richard Landstorfer (Landstorfer 2014), Sarah Hücker (Hücker 2017) and Barbara Zehentner (unpublished data, dissertation in preparation). Based on their work, several functional sORFs have been selected from two data sets to be analysed phylogenetically.

The first dataset contains three genes of EHEC strain Sakai which were identified by RNAseq and RIBOseq and were functionally characterized by Hücker (2017). The genes selected for phylostratigraphic analysis are *laoB* with a deletion phenotype (genomic knockout mutant) in L-arginine (Hücker, Vanderhaeghen, Abellan-Schneyder, Wecko, et al. 2018), *ano* with a deletion phenotype under anaerobic stress conditions (Hücker, Vanderhaeghen, Abellan-Schneyder, Scherer, et al. 2018) and *slyC* with a deletion phenotype in L-arginine (Hücker 2017). The second dataset based on originally 242 EHEC EDL933 sORFs which display RNAseq (Landstorfer, et al. 2014) and RIBOseq signals (Landstorfer 2014). Out of these, a total of 216 “overlapping gene candidates” (OGCs) were analysed for overexpression phenotypes by Barbara Zehentner in her master thesis (Zehentner 2015) and her PhD thesis (Zehentner,

unpublished). Sixteen overexpression phenotypes were identified by Zehentner (genes listed in Supplementary table S18).

In the following sections, the phylostratigraphy of four representative overlapping genes, namely *laoB, ano, slyC* and OGC106, will be described in some detail, while 14 further sORFs with phenotypes will be discussed briefly.

3.6.1 Phylostratigraphic analysis of 14 overlapping genes with phenotypes

Fourteen further overlapping genes taken from the dataset of Barbara Zehentner were analysed with respect to the presence of homologues in species different from EHEC EDL933 in order to estimate the gene age (Table 3.6.1). The details of all analyses are given in the appendix. The genome organizations are visualized in Supplementary figure S12, the phylostratigraphic trees with the alignment are in Supplementary figure S13.1 - S13.14 and the trends of sequence identities can be found in Supplementary figure S14.1 - S14.7. A negative control was constructed from sequence identities of 100 randomly selected genes and a respective overlapping sequence of the length of the sORF to the EHEC genes (Supplementary figure S15.1 - S15.11). The random genes are originated from those organisms that are shown in the respective phylostrata. The graphs show that there is an evolutionary constraint on the sORFs. There are only five exceptions, which have a faster decreasing sequence identity in comparison to the negative controls (*laoB, ano,* OGC75, OGC85, OGC167).

The youngest gene was OGC167, which was taxonomically restricted to EHEC and few further *Escherichia coli* strains, two OGCs were taxonomically restricted to *Escherichia coli/Shigella*, three to the genera *Escherichia/Shigella*, three to enterobacteriaceae, six to enterobacteriales and three to γ-proteobacteria. OGC15 was classified to the phylostratum ‘proteobacteria’ and OGC121 was the oldest gene with the furthermost related homologue in a micrococcaceae species. With one exception (OGC51), the sORF sequences were always younger or of equal age and showed a quicker loss of the sequence identities by growing distance to EHEC than the mother gene. OGC51 is assigned to the same phylostratum as the mother gene, but the quicker loss of the mORF sequence identity by growing distance to EHEC in comparison to the OGC51 identities indicates a younger age. Most genes (11) are in -1 frame, one is in -3 frame, seven in -2 frame. Most sORFs are produced by discontinuous evolution or by a combination of discontinuous and gradual evolution.

Table 3.6.1: Summary of phylostratigraphic analysis of overlapping genes with phenotype. The representative genes are described more in detail in the following sections: *asa* in section 3.5, *laoB* in section 3.6.2, *ano* in section 3.6.3, *slyC* in section 3.6.4 and OGC106 in section 3.6.5. The data of the other 14 overlapping genes with phenotype can be found in the appendix (Supplementary figures S13-S15).

Overlapping gene	Phylostratigraphic level	Evolutionary processes	Reading frame	Overlap type
asa	γ-proteobacteria	gradual	-2	embedded
laoB	*Escherichia/Shigella*	discontinuous	-2	embedded
ano	*Escherichia/Shigella*	discontinuous	-3	head-to-head
slyC	enterobacteriales	discontinuous	-2	embedded
OGC106	enterobacteriales	discontinuous (2 steps), HGT to Gram-positive bacteria	-1	tail-to-tail
OGC15	proteobacteria	discontinuous in enterobacteriales, gradual	-2	embedded
OGC23	enterobacteriales	gradual	-1	tail-to-tail
OGC51	*Escherichia coli/Shigella*	discontinuous	-2	head-to-head
OGC57	enterobacteriales	gradual	-1	embedded
OGC75	*Escherichia coli/Shigella*	discontinuous	-1	tail-to-tail
OGC85	enterobacteriaceae	discontinuous	-1	embedded
OGC121	bacteria	discontinuous in *Escherichia/Shigella*, gradual	-1	embedded
OGC167	EHEC or *Escherichia*	discontinuous, putative HGT in *Escherichia*	-2	tail-to-tail
OGC174	*Escherichia/Shigella*	discontinuous (2 steps)	-1	tail-to-tail
OGC194	enterobacteriaceae	discontinuous (2 steps)	-1	head-to-head
OGC198	γ-proteobacteria	gradual	-2	embedded
OGC226	enterobacteriales	discontinuous (*Escherichia/Shigella*), gradual	-1	tail-to-tail
OGC231	γ-proteobacteria	gradual	-1	tail-to-tail
OGC241	enterobacteriaceae	discontinuous (*Escherichia/Shigella*), gradual	-1	tail-to-tail

3.6.2 The arginine responsive overlapping gene *laoB*

The small gene *laoB,* with a length of only 123 bp, is embedded in -2 frame of ECs5115 (1,539 bp), a CadC family transcriptional regulator (Figure 3.6.1). There is a putative gene upstream of *laoB* - *laoA* - which was newly discovered and is situated in an operon with *laoB* (Hücker, Vanderhaeghen, Abellan-Schneyder, Wecko, et al. 2018). The protein LaoB is present in 497 *Escherichia* and 18 *Shigella* strains and species (tblastn). The mother gene has 1121 tblastn hits, predominantly to bacteria within the family γ-proteobacteria. The phylostratigraphic tree shows that the *laoB* sequence is quickly disintegrated in more distantly related species (Figure 3.6.2). This is also visible by comparing the sequence identities of homologues with the negative control (Supplementary figure S15.1). The sequence identities of *laoB* homologues are significantly lower than the identities of the negative controls in species farthermost related from EHEC than the genus *Escherichia.* Sequence disruptions are caused either by indel mutations or by internal stop codons (Figure 3.6.2). The highest sequence variability in ORFs with intact *laoB* can be found in *Escherichia fergusonii*. The low conservation is also visible in the mORF sequence, but it remains intact in all species shown in the phylostratigraphic tree (Figure 3.6.3). The sequence identities of both genes quickly decrease at the split from *Escherichia/Shigella* to further Enterobacteria indicating the discontinuous overprinting of *laoB* (Figure 3.6.4).

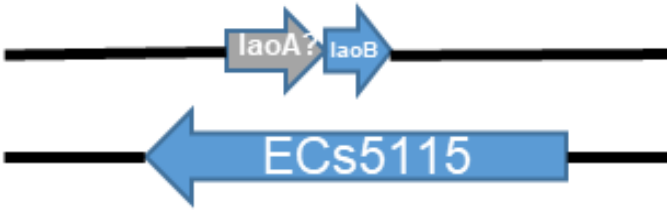


Figure 3.6.1: Genome organization of the arginine responsive overlapping gene *laoB* in EHEC Sakai. *LaoB* and its mother gene (ECs5115) are shown as blue arrows, the ORF encoding the putative gene *laoA* is shown as grey arrow.

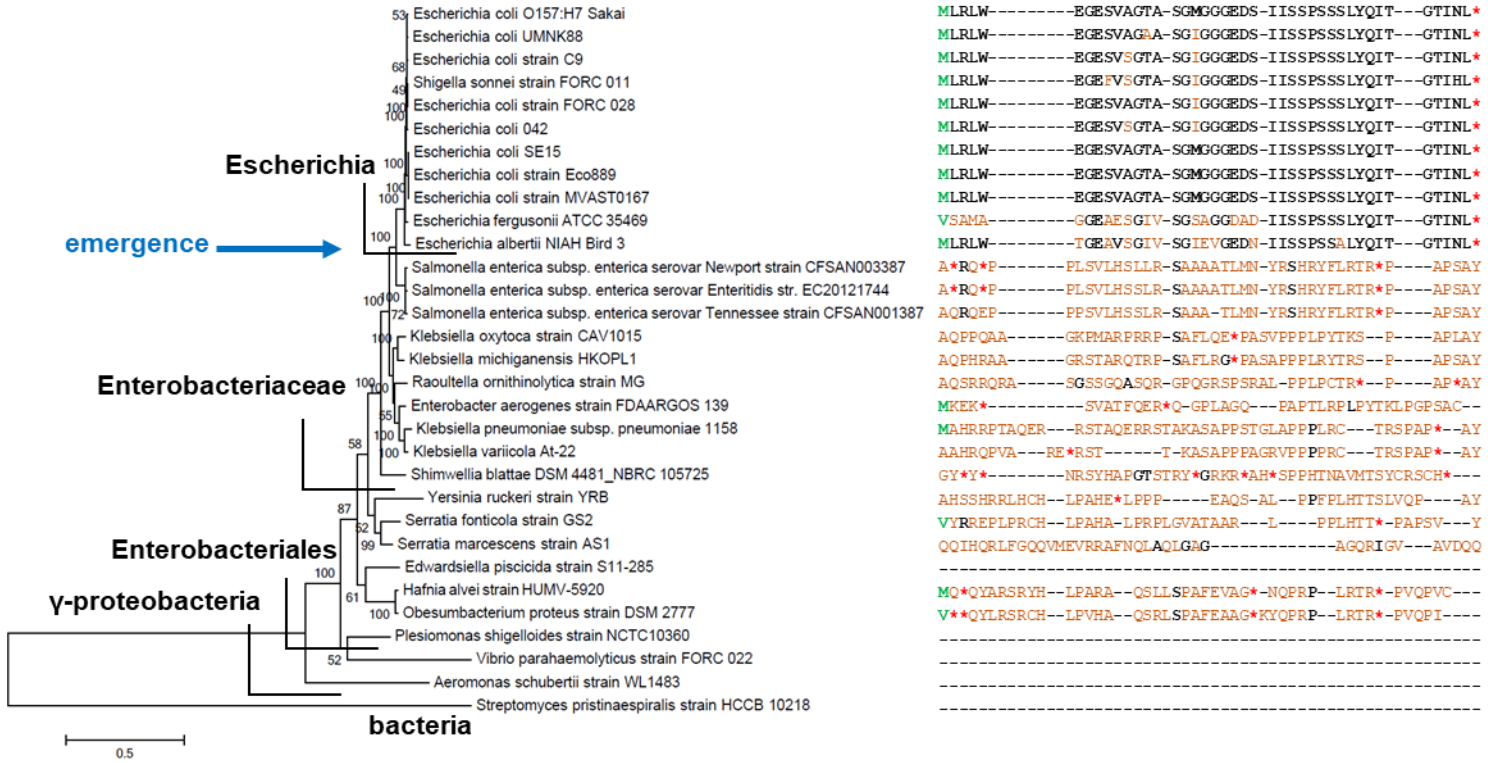


Figure 3.6.2: Phylostratigraphy of *laoB*. The species tree is constructed from organisms in which the mORF homologue was found. The respective *asa* homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EHEC Sakai are brown. The *laoB* emergence is marked as blue arrow. The species tree was constructed from 16S rRNA sequences. It is a neighbor joining tree calculated with Mega6 (Tamura, et al. 2013).

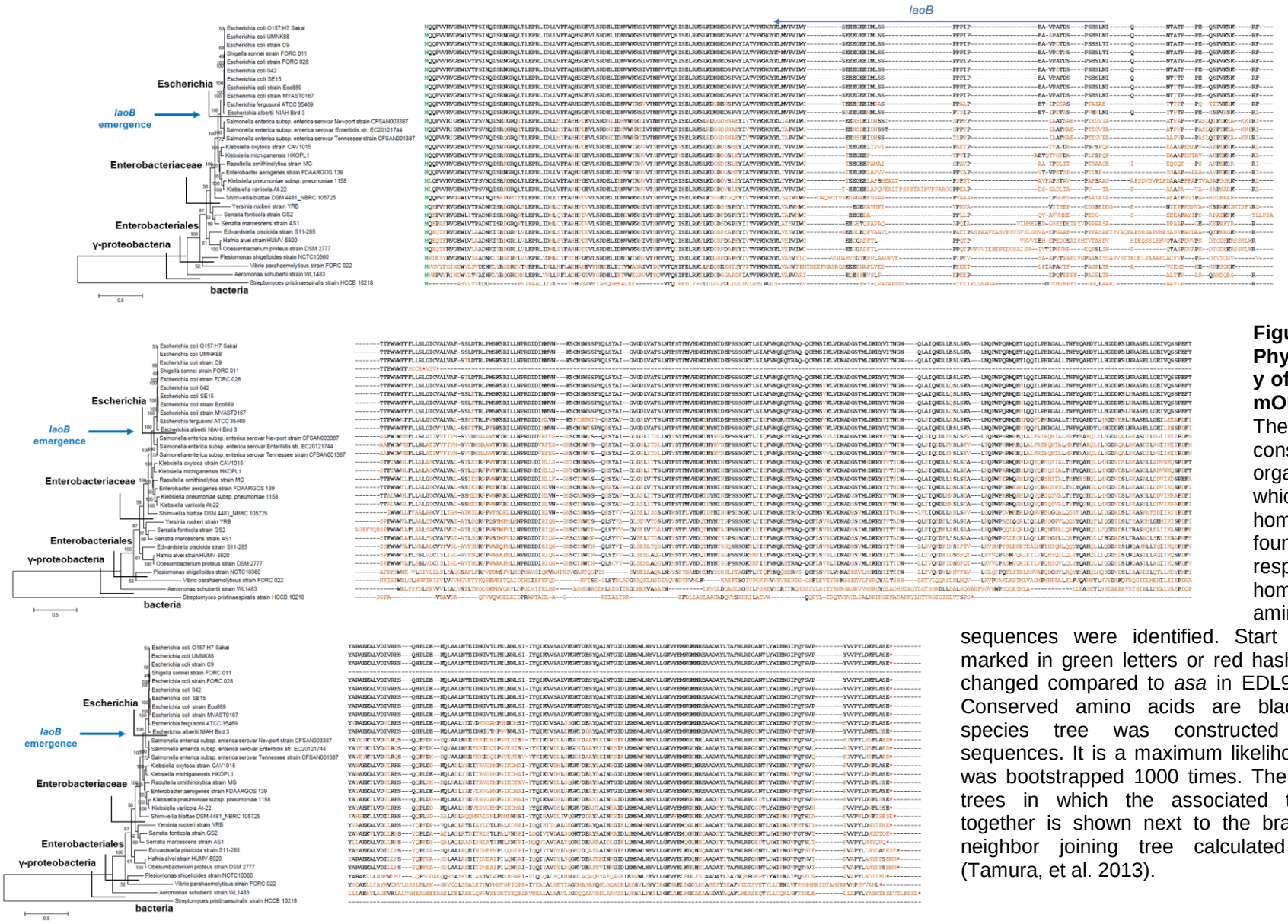


Figure 3.6.3: Phylostratigraphy of ECs5115, the mORF of *IaoB*. The species tree is constructed from organisms in which ECs5115 homologues were found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences. It is a maximum likelihood tree, which was bootstrapped 1000 times. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega6 (Tamura, et al. 2013).

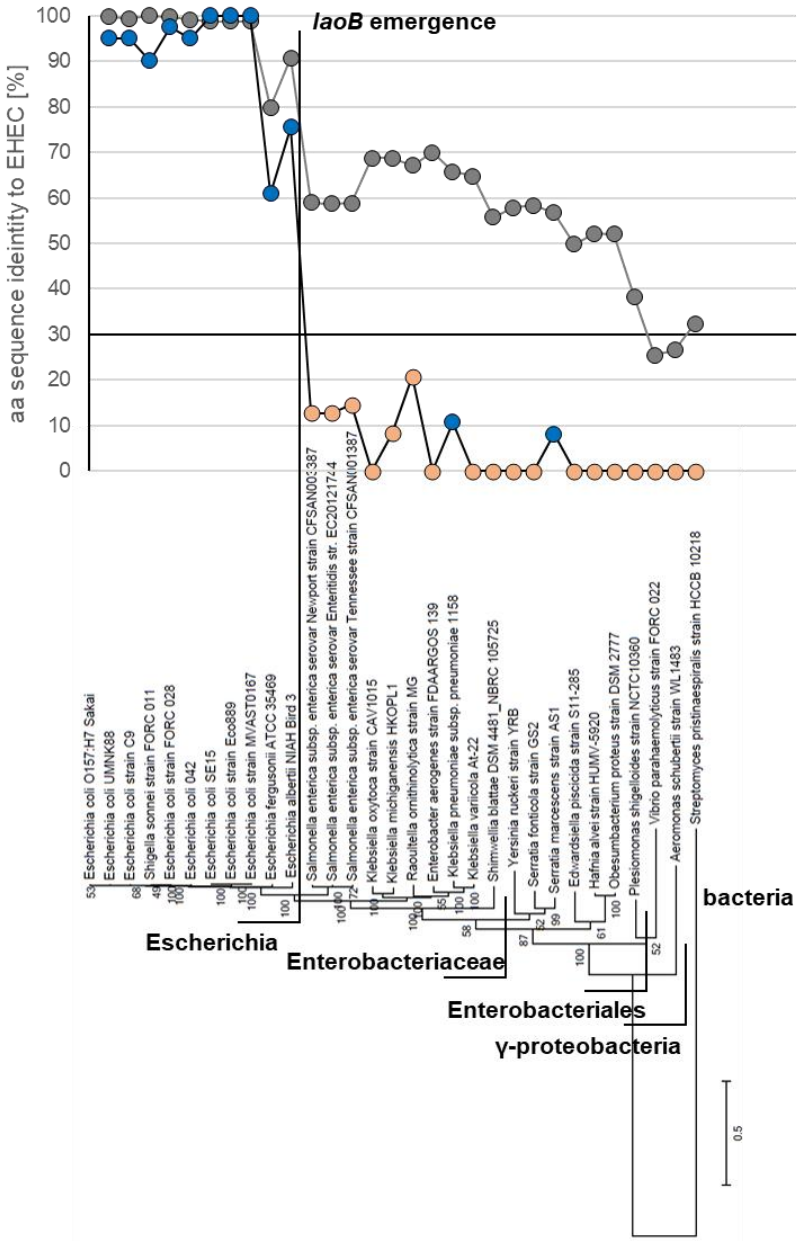


Figure 3.6.4: Amino acid (aa) sequence identities to EHEC [%] of the overlapping gene pair *laoB*/ECs5115 during species evolution. Legend: graph on top: Sequence identities to EHEC of each homologue shown in the tree below. Grey: ECs5115 homologues; blue: intact sORF homologues, salmon: sORF homologues with internal stop codons in the sORF matching region; the sequence homology cutoff of 30% is shown as black line. Graph below: Maximum likelihood tree of species containing mORF homologues. The tree was constructed as combination of MLSA tree (Escherichia to Enterobacteriales) and a 16S rRNA tree (γ-proteobacteria and further related). The tree was calculated using Mega6 (Tamura, et al. 2013).

3.6.3 The anaerobiosis-responsive overlapping gene *ano*

With a length of only 186 bp, *ano* is also a small gene which is terminal overlapping head-to-head to ECs2385, a L, D-transpeptidase gene. The reading frame of *ano* is -3 relative to its mother gene (Figure 3.6.5). There are 474 hits to genomes of the genera *Escherichia* and *Shigella*. The evolution of *ano* is gradual with one exception. *E. coli* FHI92 (Genbank LM997172.1) has an internal stop in the *ano* homologous ORF (Figure 3.6.6). The sequences present in *Escherichia albertii* or *Escherichia fergusonii* have significantly lower sequence identities (53.2% and 60.3% respectively), which is also visible in the negative control (Supplementary figure S15.1). In species more distantly related to *E. fergusonii*, the sORF identities are lower than the negative control. While ORFs of *E. albertii* are intact, those originated from *E. fergusonii* have a highly variable region with internal stop codons (Figure 3.6.6). The mother gene can be found in many bacteria (2630 hits) and has even one hit to a predicted eukaryotic mRNA in the parasitic wasp *Diachasma alloeum* (E-value 6×10^{-99}, Identity 53,3%). Similar to *laoB*, the mother gene has a low conservation in the *ano* overlapping region (Figure 3.6.7). The low conservation in the non-overlapping region is also visible in the sequence identities (Figure 3.6.8). The origination of *ano* by discontinuous overprinting at the split of *Escherichia/Shigella* from further enterobacteria confirmed the discovery of another taxonomically restricted overlapping gene.

Figure 3.6.5: Genome organization of the anaerobiosis responsive overlapping gene *ano* in EHEC Sakai. *Ano* and its mother gene (ECs2385) are shown as blue arrows, the gene upstream of *ano* (ECs2384) shown as grey arrow.

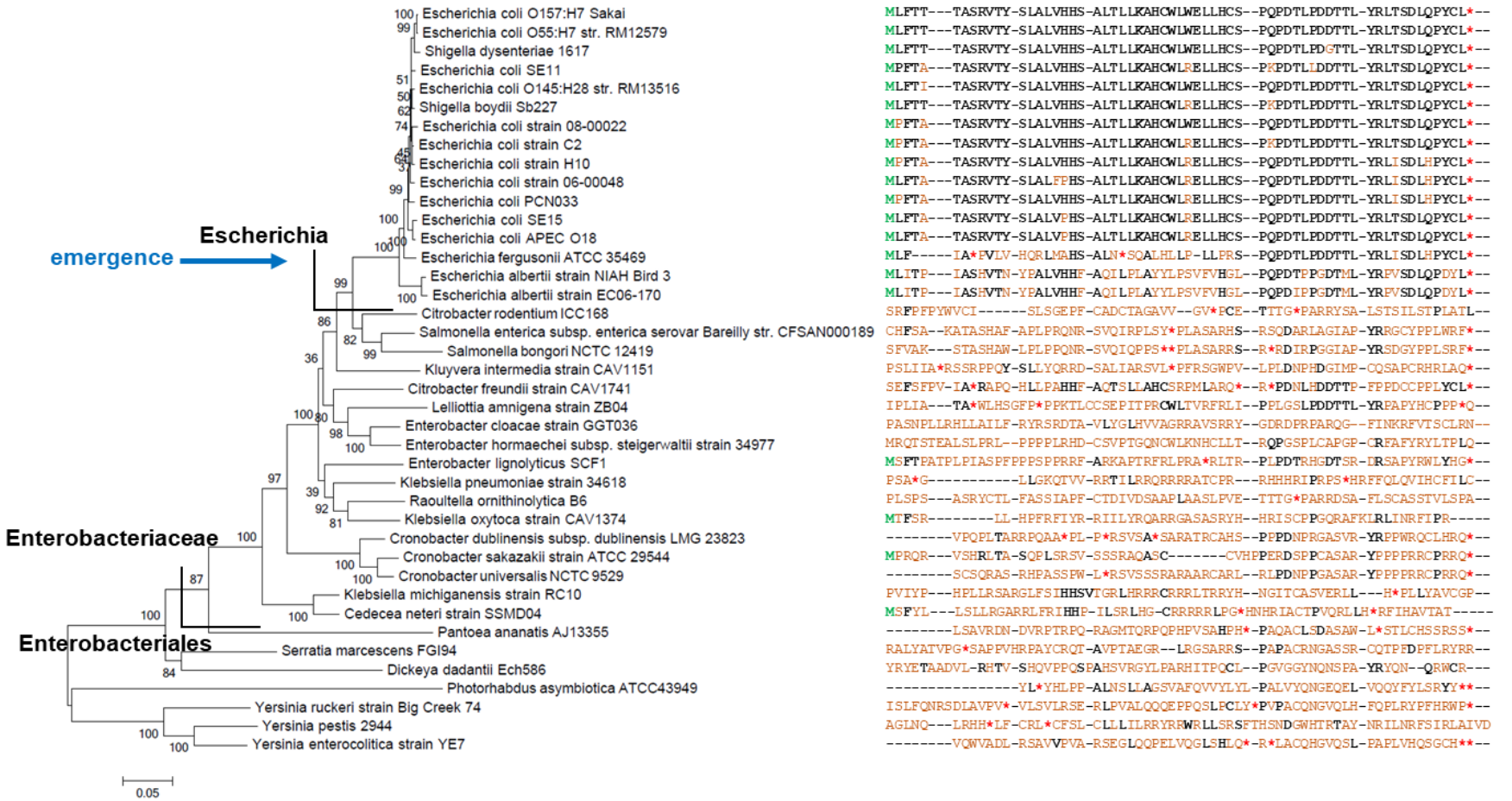


Figure 3.6.6: Phylostratigraphy of *ano*. The species tree is constructed from organisms in which the mORF homologue was found. The respective *asa* homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EHEC Sakai are brown. The *ano* emergence is marked as blue arrow. The species tree was constructed as MLSA tree. It is a neighbor joining tree calculated with Mega6 (Tamura, et al. 2013).

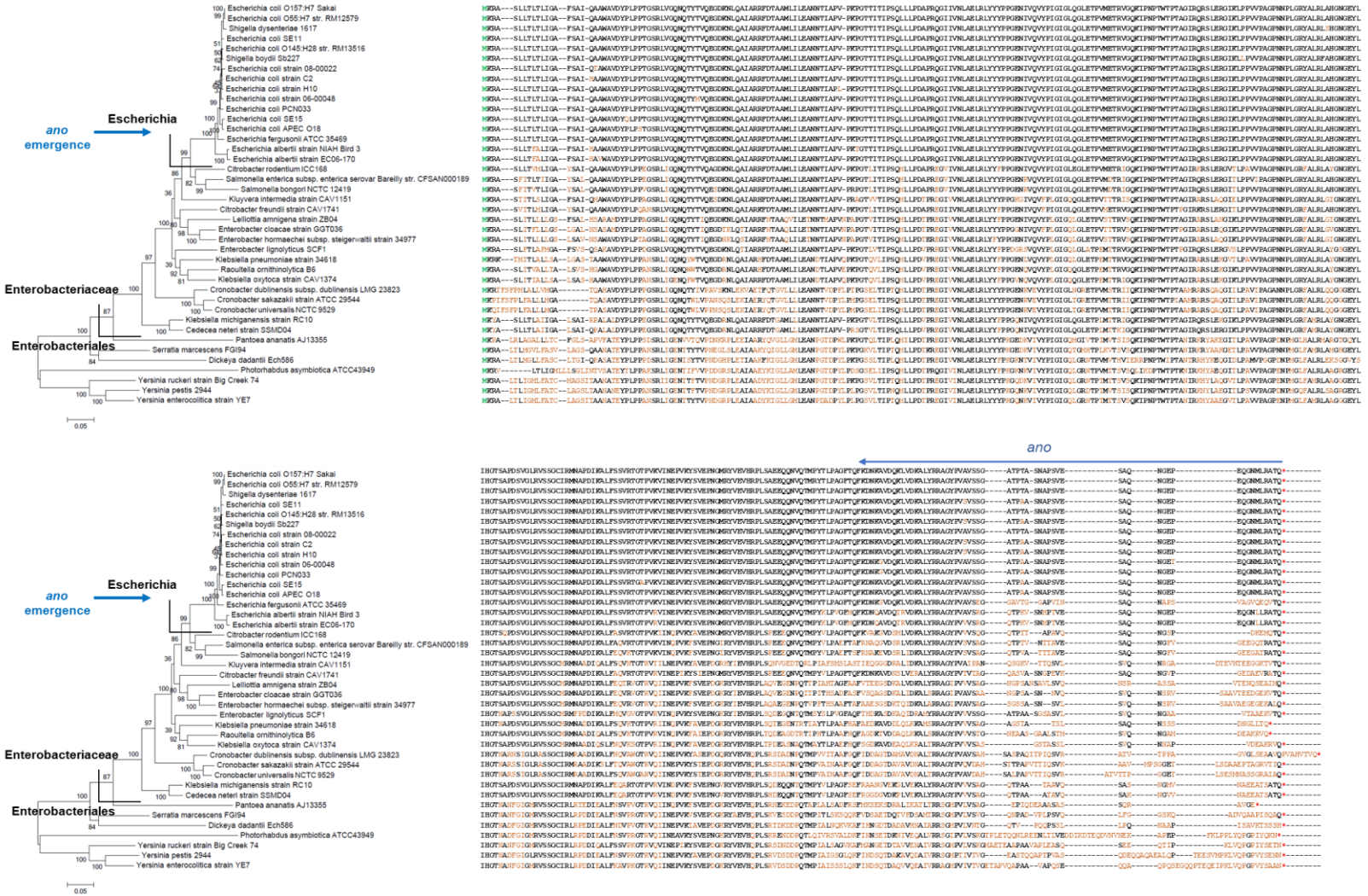


Figure 3.6.7: Phylostratigraphy of ECs2385, the mORF of *ano*. The species tree is constructed from organisms in which ECs2385 homologues were found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences. It is a maximum likelihood tree, which was bootstrapped 1000 times. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega6 (Tamura, et al. 2013).

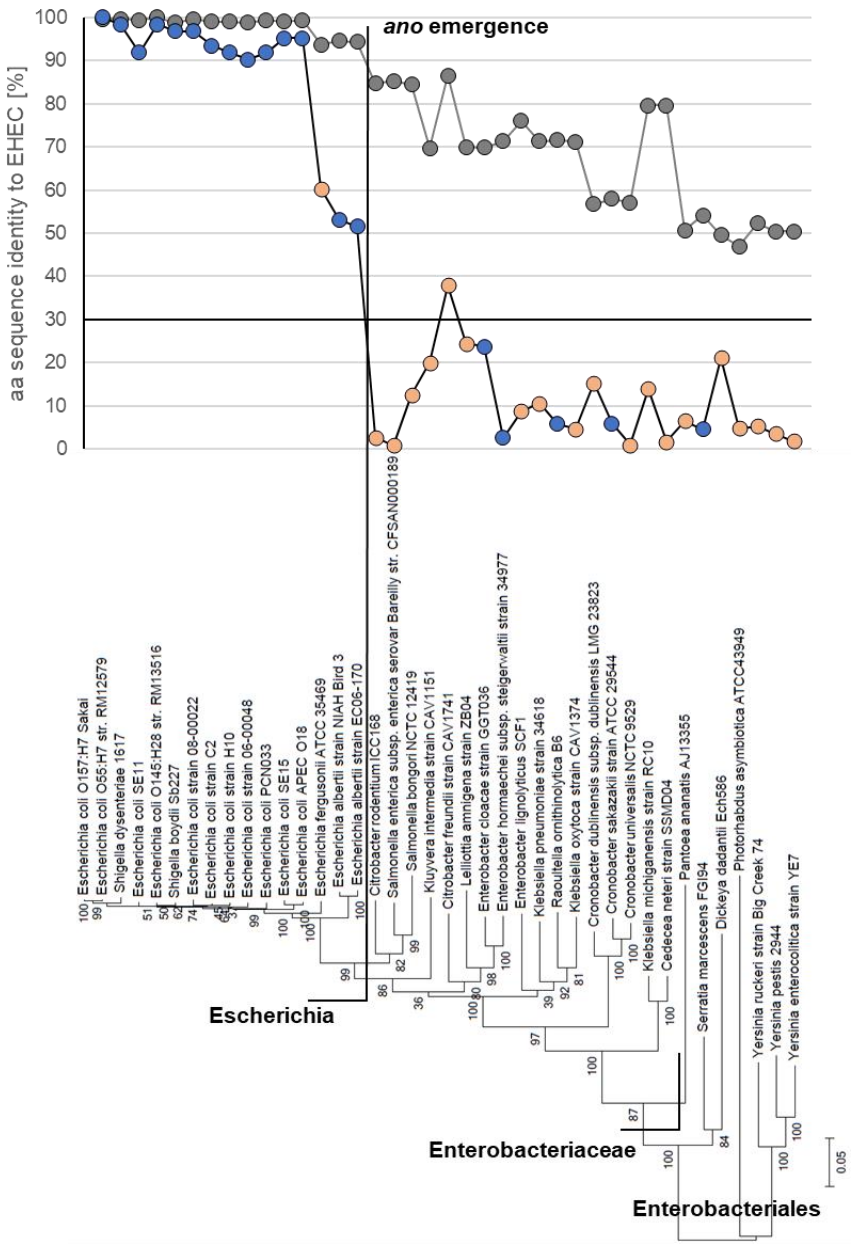


Figure 3.6.8: Amino acid (aa) sequence identities to EHEC [%] of the overlapping gene pair *ano*/ECs2385 during species evolution.
Legend: graph on top: Sequence identities to EHEC of each homologue shown in the tree below. Grey: ECs2385 homologues; blue: intact sORF homologues, salmon: sORF homologues with internal stop codons in the sORF matching region; the sequence homology cutoff of 30% is shown as black line. Graph below: Maximum likelihood tree of species containing mORF homologues. The tree was constructed as combination of MLSA tree (Escherichia to Entero-bacteriales) and a 16S rRNA tree (γ-proteobacteria and further related). The tree was calculated using Mega6 (Tamura, et al. 2013).

2.6.4 The ARG-box regulated overlapping gene *slyC*

The novel gene *slyC* (192 bp) is expressed in an operon with *slyB* (ECs2350). Their expression is regulated by an ARG box (Hücker 2017). *SlyC* is embedded in -2 frame of a transcriptional regulator *slyA* (ECs2351, Figure 3.6.9). Blasting with tblastn reveals a phylostratigraphic level of *slyC* in the order *enterobacteriales,* which was confirmed by the phylostratigraphic tree (1031 hits, Figure 3.6.10). Sequences of *slyC* homologues beyond and partly within this family could not be shown due to disruption. The mother gene *slyA* (441 bp) is highly conserved within multitude bacteria (2774 tblastn hits), particularly within the order enterobacteriales (Figure 3.6.11). The sequence identities to *slyC* decrease at the split of *Escherichia* from other enterobacteriaceae and remains constant within this family (Figure 3.6.12). The sequence identities of the mother gene do not change within the enterobacteriaceae. The negative control shows that homologues of the gene pair lay within the range of randomly selected sequences (Supplementary figure S15.2). The quick decrease of sequence identities in species more distantly related than enterobacteriaceae is also visible in the randomly selected genes. All homologues have a full-length or longer ORF within the family, but there are two exceptions: Those homologues in *Salmonella* have an internal stop at amino acid position 31 (calculated from the first putative start codon 'ATA') with an immediately following 'ATG' after the stop, which may indicate a split into two ORFs. Those from *Klebsiella pneumonia* subsp. *pneumonaie* KPNIH29 or *Enterobacter coloacae* subsp. *dissolvens* SDM have respectively one internal stop at the beginning with following start codons indicating a shortened ORF in comparison to *sylC*. Elongation occurs more frequently at the 5' end than at the 3' end of the sORF. It is proposed that *slyC* originated by discontinuous overprinting at the split to the clade enterobacteriales.

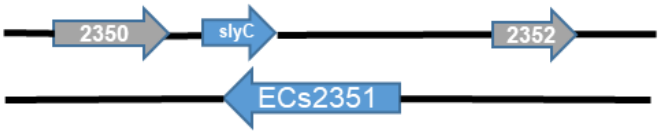


Figure 3.6.9: Genome organization of the ARG box regulated overlapping gene *slyC* in EHEC Sakai. *SlyC* and its mother gene (ECs2351) are shown as blue arrows, the genes upstream (ECs2350) and downstream (ECs2352) of *slyC* shown as grey arrow.

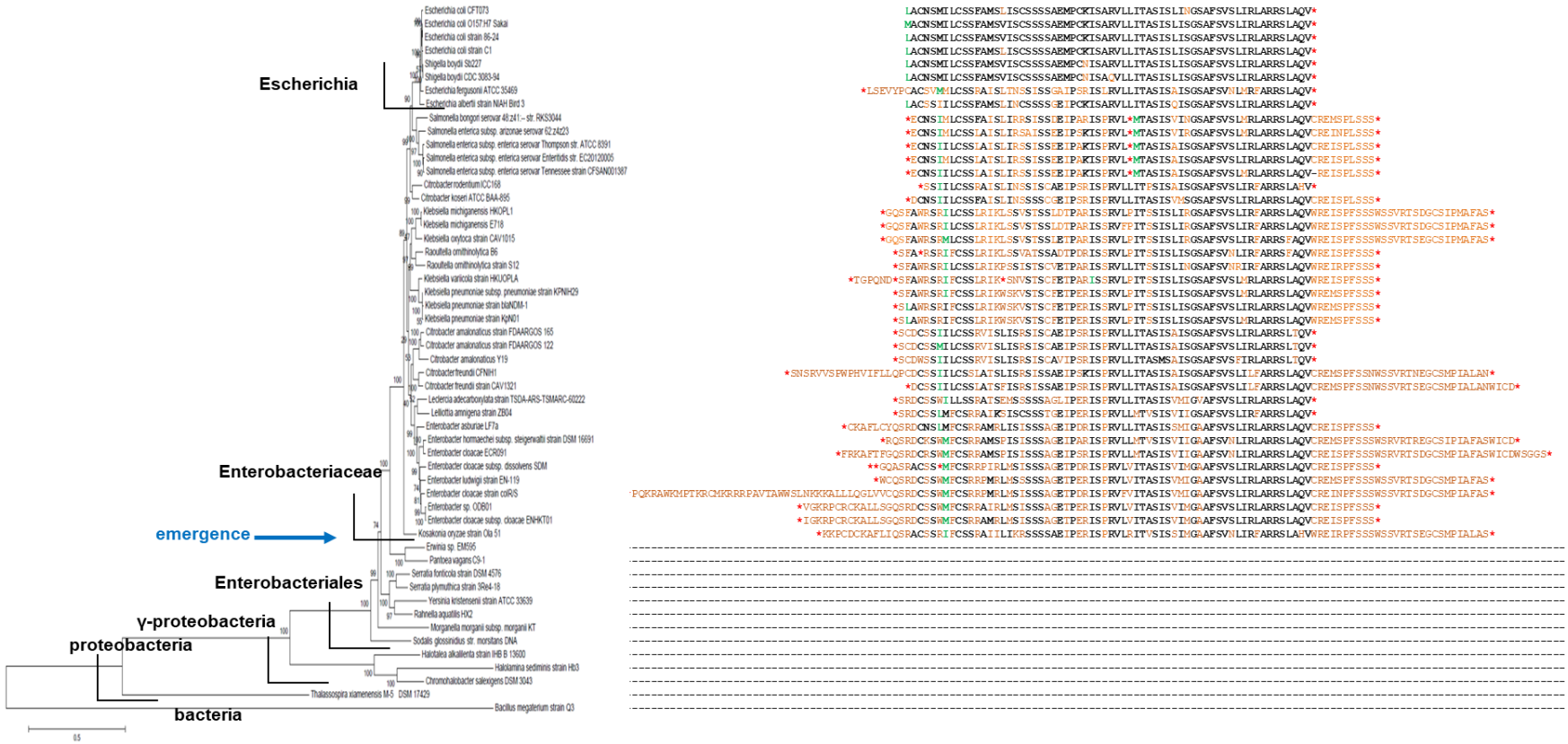


Figure 3.6.10: Phylostratigraphy of *slyC*. The species tree is constructed from organisms in which the mORF homologue was found. The respective *asa* homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EHEC Sakai are brown. The *slyC* emergence is marked as blue arrow. The species tree was constructed from 16S rRNA sequences. It is a neighbor joining tree calculated with Mega6 (Tamura, et al. 2013).

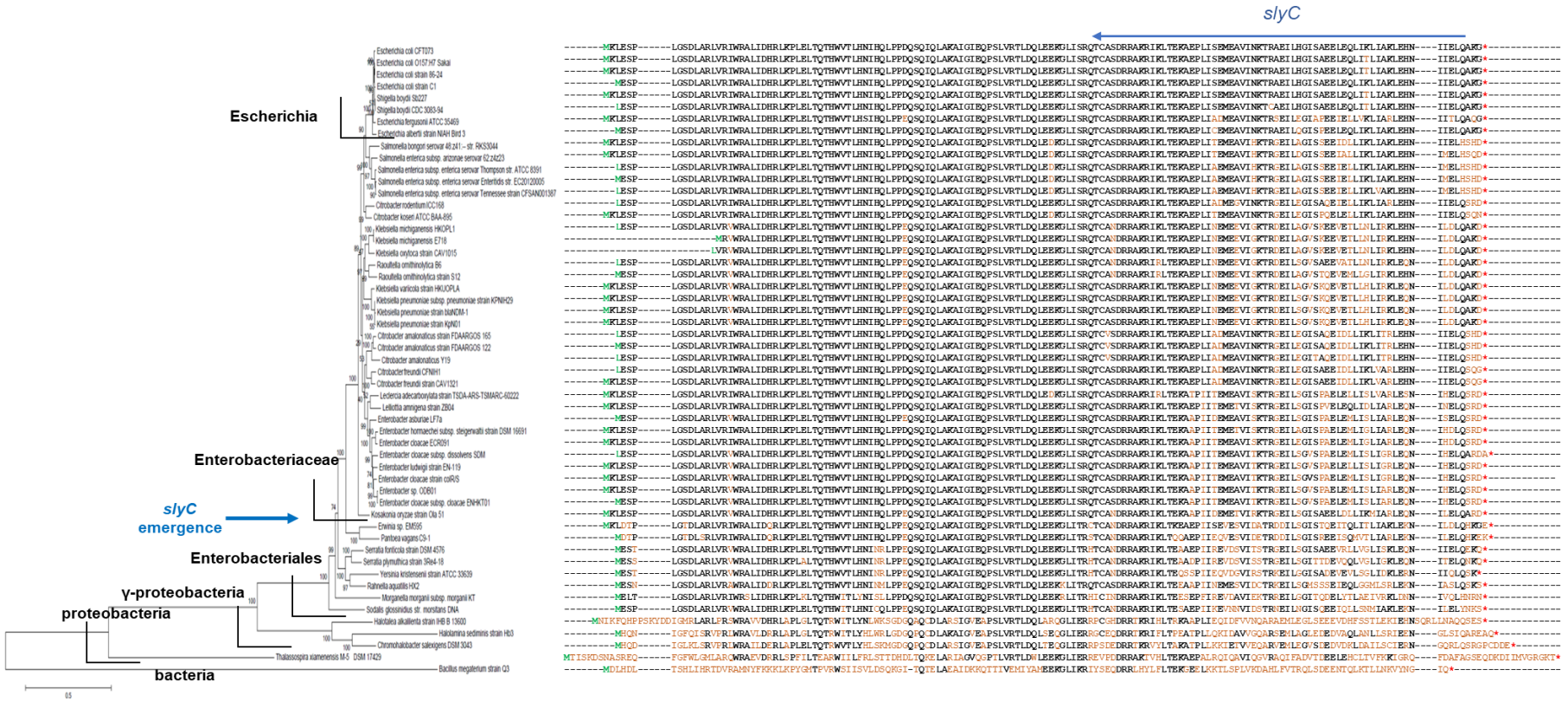


Figure 3.6.11: Phylostratigraphy of ECs2351, the mORF of *slyC*. The species tree is constructed from organisms in which ECs2351 homologues were found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences. It is a maximum likelihood tree, which was bootstrapped 1000 times. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega6 (Tamura, et al. 2013).

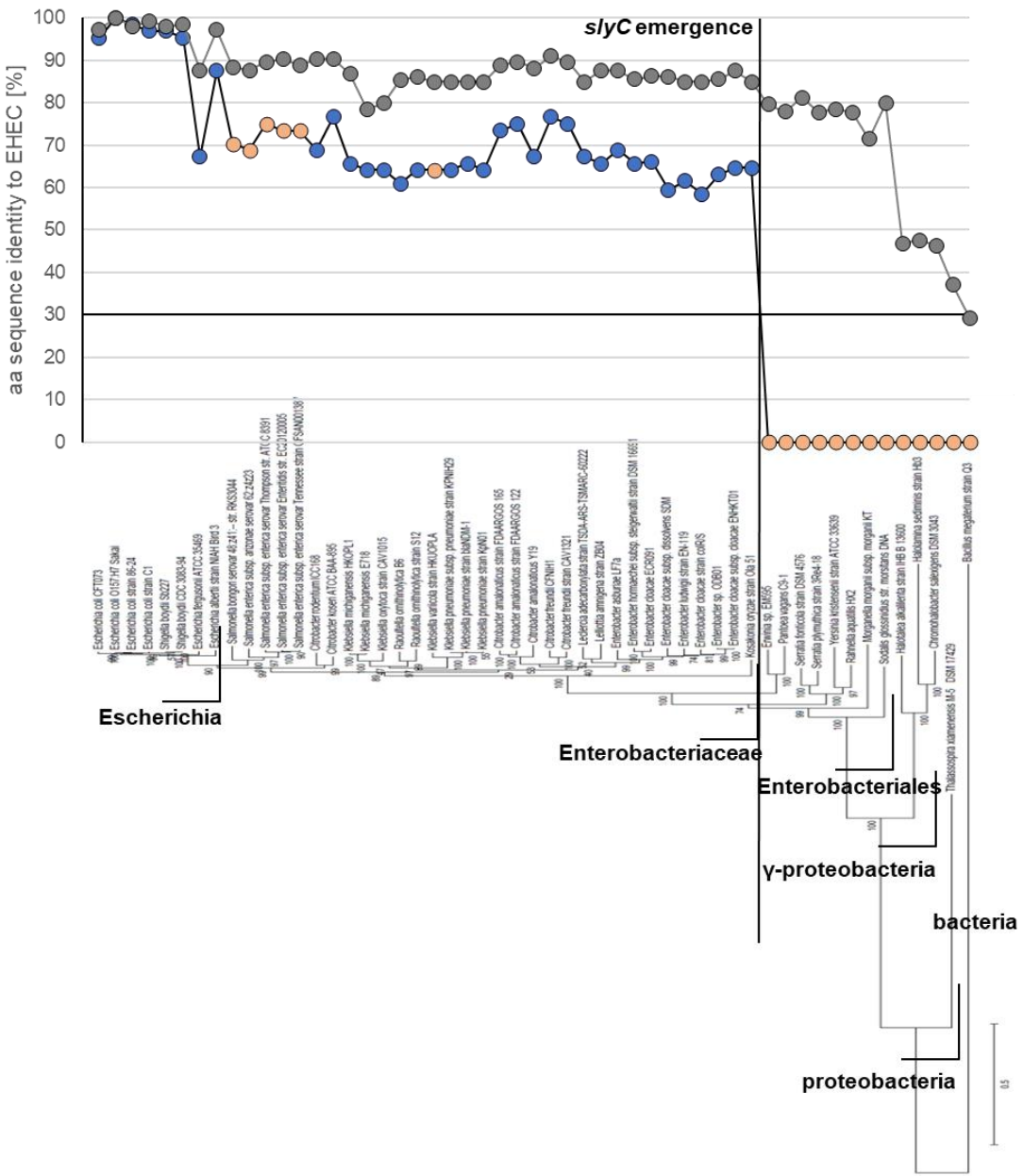


Figure 3.6.12: Amino acid (aa) sequence identities to EHEC [%] of the overlapping gene pair slyC/ECs2351 during species evolution. Legend: graph on top: Sequence identities to EHEC of each homologue shown in the tree below.; Grey: ECs5115 homologues, blue: intact sORF homologues, salmon: sORF homologues with internal stop codons; the sequence homology cutoff of 30% is shown as black line. Graph below: Maximum likelihood tree of species containing ECs2351 homologues. The tree was constructed as a combination of MLSA tree (Escherichia to Enterobacteriales) and a 16S rRNA tree (γ-proteobacteria and further related). The tree was calculated using Mega6 (Tamura, et al. 2013).

3.6.5 The overlapping gene candidate OGC106

OGC106 has a nucleotide length of 279 bp and is overlapping tail-to-tail in -1 frame to EDL933_2699, an ammonia-dependent NAD(+) synthase (828 bp, Figure 3.6.13; tblastn: 7798 hits). OGC106 is nearly completely embedded in its mother gene (Figure 3.6.14). The OGC106 homologues are present in many enterobacteriaceae genomes (1422 tblastn hits). The mORF evolution is highly constrained; the identities of the mORF homologues only slightly decrease with growing evolutionary distance to EHEC. The sORF homologues quickly decreases their sequence identity to OGC106 at the split of *E. coli/Shigella* from the genus *Escherichia* (Figure 3.6.15). A second downshift and clear sequence degradation is visible at the split of pectobacteriaceae from erwiniaceae suggesting a gene emergence at that time point. The evolution between both downshifts is gradual. There are no obvious differences of the evolutionary constraint visible when comparing the sequences in the overlapping and non-overlapping part of OGC106 (Figure 3.6.14). However, similar evolutionary patterns of OGC106 and its mother gene indicate an evolutionary dependency of both genes. The identities to EHEC of non-*Escherichia* enterobacteriaceae and pectobacteriaceae are stably between 30% and 50% and most sequences are intact. This pattern is also visible in the negative control (Supplementary figure S15.6). Both, OGC106 and mORF, lie within the sequence identity range of 100 randomly selected genes. The constraint on OGC106 may be caused by the reading frame and the mother gene rather than by an independent constraint.

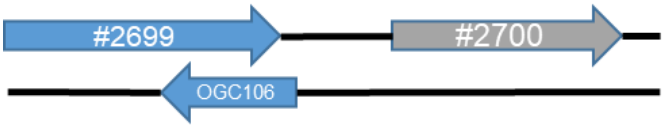


Figure 3.6.13: Genome organization of OGC106 in EHEC EDL933. OGC106 and its mother gene (EDL933_2699) are shown as blue arrows, the gene upstream of #2699 (#2700) is shown as grey arrow.

Interestingly, the sORF has tblastn hits in species of the genus *Lactobacillus* (*L. casei, L. rhamnosus, L coryniformis*). As the Gram-positive lactobacilli are only distantly related to the Gram-negative enterobacteria, it can be hypothesized that the gene was obtained by horizontal gene transfer. The mother gene has many hits to NAD(+) synthases in Gram-positive bacteria (for example *Bacillus, Listeria, Streptococcus, Lactobacillus, Lysinibacillus*) with a high sequence identity of ~70% and a coverage of 99%, thus the horizontal gene transfer did not only occur in lactobacilli. The gene pair OGC106/mORF is overlapping in *Lactobacillus* similarly to EHEC (tested exemplary for few strains). It can be hypothesized that the HGT occurred between Gram-positive bacteria to enterobacteriales, because genes in Gram-positive bacteria have a higher sequence identity to EHEC than those in non-enterobacteriales proteobacteria (≤ 50%, obtained by scanning the tblastn hits).

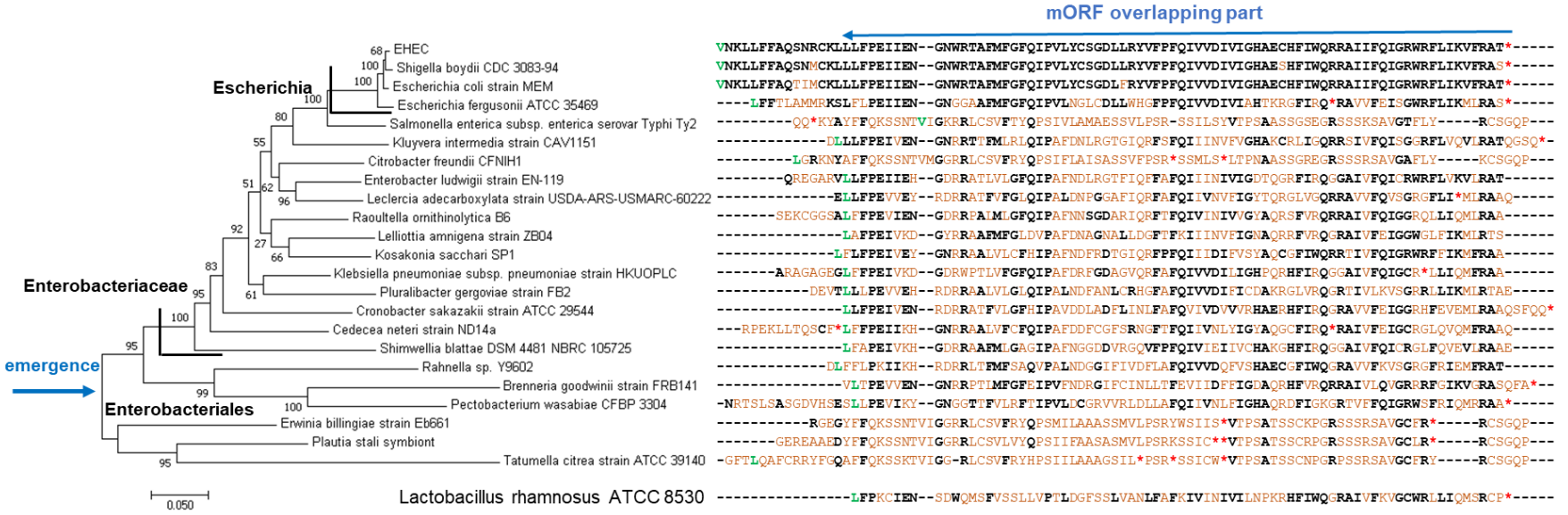


Figure 3.6.14: Phylostratigraphy of OGC106. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales*. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7 (Kumar, et al. 2016).

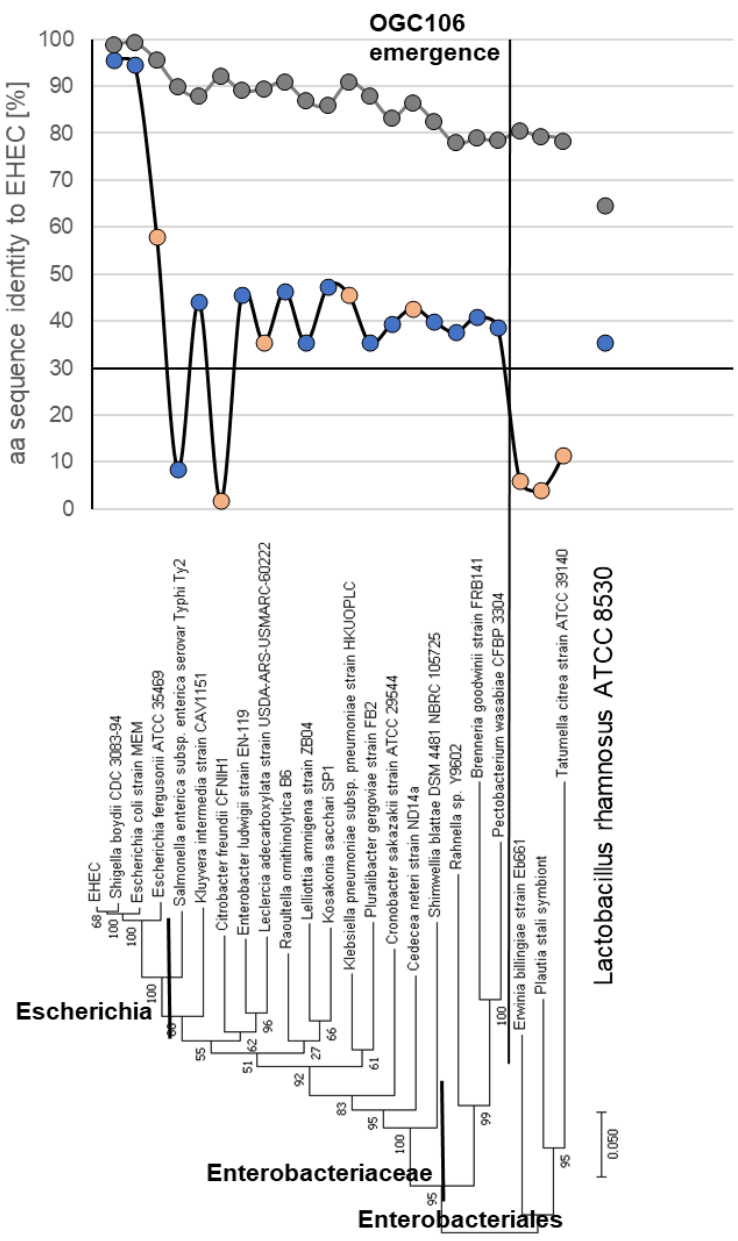


Figure 3.6.15: Amino acid (aa) sequence identities to EHEC [%] of the overlapping gene pair and OGC106/EDL933_2699 during species evolution. Legend: graph on top: Sequence identity to EHEC of each homologue shown in the tree below. grey: mORF homologues, blue: intact sORF homologues, salmon: sORF homologues with internal stop codons; the sequence homology cutoff of 30% is shown as black line. Graph below: Maximum likelihood tree of species containing the respective mORF homologues. The tree was constructed as MLSA tree and calculated using Mega7 (Kumar, et al. 2016).

4 Discussion

4.1 Shadow ORFs are putative genes with features similar to novel intergenic ORFs

4.1.1 Shadow ORFs form a reservoir of uncharacterized genes

In the past, it was thought that the bacterial genome is already densely packed with genes (Patthy 1999). Indeed, EHEC EDL933 has a genome size of 5.55 Mbp harboring 5,746 annotated genes (Latif, et al. 2014). This means that 86% of its genome is used for coding sequences and only 14% are intergenic. However, even in the intergenic sequences, many small protein coding genes appear to be located (Neuhaus, et al. 2016; Hücker, et al. 2017). According to this view, the EHEC genome has only a very small non-coding reserve. On the other hand, there is a huge number of ORFs (49,649) overlapping to annotated genes. These overlapping ORFs would have the potential to form a reservoir of so far unknown putative functional genes. Thus, the first aim of this dissertation was finding evidence whether some overlapping ORFs may encode novel proteins, a hypothesis which was found to be supported by a number of arguments.

Firstly, blast sequence homology search was used to find out if there is any kind of information about sORF-like proteins in the databases. In the blastp search, an E-value of 10^{-3} was chosen as a cutoff for significant homology, because this E-value was identified to be a good compromise between sensitivity and accuracy (Neme and Tautz 2013; Kuchibhatla, et al. 2014). A high sequence identity is evidence for functionality since it is unlikely to arise by chance. A surprisingly high number of sORFs (2,180) were found with homologous proteins in other species. However, 92% of the homologues are hypothetical. A “hypothetical protein” was predicted by annotation programs to be a coding sequence, but is not characterized experimentally. However, many hypothetical proteins have late been shown to be functional and “hypothetical” does not imply that features associated with function are completely unknown (Galperin 2001). Bioinformatics reveals hints for putative functions based on homologies, protein structures or DNA and protein binding sites (Nimrod, et al. 2008; Prava, et al. 2018). Many hypothetical genes are shown to be expressed (Landstorfer 2014) or experimentally validated to have a function (Li, et al. 2018; Tian, et al. 2018).

4.1.2 Information extraction for sORFs need multiple combined approaches

Comparative genomics was used extracting more information about the nature of hypothetical proteins (Nimrod, et al. 2008; Ijaq, et al. 2015). For example, the conserved domain database implements 3D structures to uses conserved protein domains (Marchler-Bauer, et al. 2016). This approach is more sensitive than a blast search, because protein structures are often better conserved than DNA or amino acid sequences (Watson, et al. 2005). However, only six of 33 sORFs, which had blastp matches to homologous hypothetical proteins, could be assigned to a putative function by consulting conserved domain search (section 3.1.2). Surprisingly, there was a low intersection between sORFs with conserved domains, sORFs with ribosomal footprints and sORFs with complete blast matches. Thus, only a combination of several analyses will find all pieces of the puzzle.

4.1.3 The protein features are similar to previously analyzed uncharacterized proteins

PredictProtein was used to obtain *ab initio* structure based features of proteins encoded by sORFs (Sleator and Walsh 2010). It has the advantage not to rely on pure homology searches. There are three studies published, which similarly analyzed so far unknown ORF encoded proteins. Neuhaus, et al. (2016) predicted protein features of 72 intergenic ORFs with RNAseq and RIBOseq data in the same organism (EHEC EDL933). They used a four-time larger number of length-matching annotated genes as positive control set to the intergenic ORFs. Hücker, et al. (2017) discovered multitude transcribed and translated short intergenic ORFs in EHEC Sakai. About half of them had blastp hits, which were compared to the other half without hit and to annotated proteins of comparable length using PredictProtein. The third study was published by Perdigão, et al. (2015), who predicted bacterial, eukaryotic and archaeal proteins using different protein prediction programs. Their dataset was the “dark proteome” extracted from all proteins stored in the Swiss-Prot protein database. The “darkness” of a protein was defined in terms of sequences, which do not match any known protein (Protein database), structure (Aquaria database) or finds residues, which are similar to protein structure models (Protein Model Portal). The trends and features of the three above mentioned studies is compared to sORF proteins classified according to their length (see section 3.2).

The mean length of sORFs with blastp hit was 120 aa. This is comparable to results obtained by Neuhaus, et al. (2016, 95% of the ORFs were smaller than 100 aa), Hücker, et al. (2017, average length of ORFs with annotated homologues: 172 aa) or Perdigão, et al. (2015, average length of bacterial dark proteins: 193 aa). Shadow ORFs of EHEC EDL933 have an overall increased number of protein binding sites in comparison to aORFs, which was confirmed by

Neuhaus, et al. (2016), but not by Perdigão, et al. (2015) who observed fewer protein interactions of dark proteins. The localization of sORF proteins is strongly length dependent, an effect, which was not observed for aORF proteins (Figure 3.2.6). The absolute number of secreted sORF proteins (57%) was significantly higher than that of aORF proteins (15%), which is accompanied by an increased abundance of disulfide bonds also observed in all three publications. Disulfide bonds are formed in the periplasm for which the proteins have to be transported through the membrane (Hatahet, et al. 2014) and are more frequently secreted into the extracellular medium (Kadokura, et al. 2003). Only a few sORF proteins are predicted to be membrane associated (7%), which is reflected by the low percentage of sORF proteins which are predicted to have transmembrane helices (8%) in comparison to aORF proteins (21%, Figure 3.2.10 A). This was also observed by Perdigão, et al. (2015). Interestingly, according to Hücker, et al. (2017), novel proteins with annotated homologues have a comparable abundance of transmembrane helices than annotated proteins of EHEC, but novel proteins without any homologues have fewer transmembrane helices. The ‘structuredness’ (Figure 3.2.7-3.2.9) and the amino acid composition is comparable between sORF and of aORF encoded proteins (Figure 3.2.2), which is a good indication that sORFs are not just arbitrary sequences.

Overall, sORF encoded proteins are comparable to the so far unknown intergenic proteins and to the ‘dark proteome’ of each genome, which are themselves comparable to annotated genes. Bioinformatics has its limitations in the feature prediction of such ‘unknowns’, because the algorithms are trained on known structures. It is suggested that it can reasonably assumed that there is a high number of functional sORFs, but the structural features, interactions, expression conditions or the cellular process in which they are involved in are not discovered so far. This leaves a large potential for future research.

4.2 *Asa* is a protein coding sORF with multiple phenotypes and gene regulation

One sORF, *asa*, was experimentally characterized in detail. Putative functional sORFs are usually first detected by RNAseq and RIBOseq. The many signals distributed all over a genome, which were first thought to be pervasive transcription and translation, find more and more evidence for functionality as non-coding RNA or proteins (Storz, et al. 2014; Neuhaus, et al. 2017; Hör, et al. 2018). The clear peak of *asa,* expressed after growth in LB medium, alerted us to choose this ORF for further characterization.

4.2.1 The *asa* expression responds to NaCl, L-arginine and pyridoxine hydrochloride

Proof for 'gene functionality' can be obtained by the identification of phenotypes and by understanding its regulation. Phenotypes are changes in the organism's fitness in case of an overexpressed or deleted gene (Prelich 2012). They are often only visible under particular conditions, for example by addition of stress substances or by temperature shift (Prelich 2012). Preliminary experiments revealed a putative overexpression phenotype in NaCl, L-arginine and pyridoxine hydrochloride (Zehentner 2015). The growth disadvantages of the wild type in comparison to the translationally arrested mutant was confirmed in NaCl and in L-arginine. The phenotype in pyridoxine hydrochloride disappeared, although there was a high promoter activity measured during growth in pyridoxine hydrochloride in exponential phase. The latter one, nonetheless, shows that pyridoxine hydrochloride influences *asa* expression. A growth disadvantage in NaCl seems to contradict the results obtained from the promoter activity test, for which a high promoter activity was detected (Figure 3.4.6). However, the amount of *asa* product present after artificial overexpression surely exceeds natural levels, which, in turn, decreased fitness. This becomes visible in the Western Blot (Figure 3.4.8) when comparing it with the natural *asa* expression detected by RIBOseq (Figure 3.4.1). The promoter activity in L-arginine was equal to the activity in LB, although it showed a strong growth disadvantage in competitive growth experiments (Figure 3.4.3). One possible explanation is the time point at which the experiments were measured (promoter activity: exponential phase, competitive growth: stationary phase). There is a possibility that *asa* is downregulated in stationary phase during growth in LB + L-arginine, which was not measured here. Further, the fitness effect causing a phenotype is amplified by competitive growth. When growing solely in LB + NaCl, EHEC with pBAD-*asa* shows no difference in growth behavior in comparison to EHEC with pBAD-*asa*^{tar22} (Supplementary figure S4).

Because of a phenotype and high promoter activity in NaCl, the expression of *asa* was analyzed after growth under salt stress in comparison to LB using RT-qPCR. EHEC aliquots were harvested under two growth conditions. There was an *asa* up-regulation from early exponential to exponential phase. The mRNA titer was lower in LB + NaCl than in plain LB during early exponential phase. The strongly increased mRNA titer under salt stress at exponential phase exceeds the titer in LB, which was visible using RT-qPCR and promoter activity tests. The *asa* expression is quickly (30 min) downregulated after addition of NaCl. One hour after shock, EHEC is adapted to the stressor and the mRNA titer is equal to that in plain LB. The down-regulation of *asa* after salt shock may indicate that the gene is non-essential, corresponding to similar observations by Fellner, et al. (2015) who analyzed the sORF *nog1* which is embedded in *citC* in EHEC EDL933.

4.2.2 Three putative transcriptional start sites and three putative promoters of *asa* were identified

Two methods were used to determine the +1 site of *asa.* After growth in LB, 5' RACE reveals a +1 site 186 bp upstream of the *asa* start codon. The comparison with Cappable seq data revealed a mRNA 5' end 188 bp upstream of the start under all conditions (LB, minimal medium, LB + malic acid, LB + NaCl) and growth phases (early exponential, exponential, early stationary phase) tested. Although this is very close to the +1 site determined with 5' RACE, the exact nucleotide could not be determined. The precision of Cappable seq is estimated to be in a range of five base pairs (Ettwiller, et al. 2016). In correspondence to our findings, Ettwiller, et al. (2016) found that the RNA polymerase does not start transcription at a specific single nucleotide in 60% of the promoters. Two putative *asa* promoters (a σ^{70} and a σ^{38} promoter) have the potential to initiate transcription at this site. The σ^{70} promoter is used for most genes in *E. coli* (Feklistov, et al. 2014) and has the following consensus sequence at -10 / -35 site: TATAAT / TTGAAT (Shultzaberger, et al. 2006). The σ^{38} promoters is recruited at general stress responses, like salt stress, or at the entry to the stationary phase as alternative to the σ^{70} promoter (Weber, et al. 2005). A putative σ^{38} promoter of *asa* was identified by manual search of the consensus sequence, which is CTACACT at the -10 site. Sigma 38 promoters have no conserved -35 site (Lee and Gralla 2001). Only little differences in the nucleotide sequence determines the recruitment of a σ^{70} or σ^{38} promoter for transcription and may lead to a change of one into the other due to few mutations in the consensus sequence (Becker and Hengge-Aronis 2001). A second transcriptional start site was determined 178 bp upstream of the *asa* start codon after

growth in LB + NaCl using 5' RACE. A consensus sequence for a σ^{38} promoter was identified in the upstream region of this +1 site.

It is known that genes can be regulated condition-dependent by more than one promoter, as shown for the sporulation associated genes in *Bacillus subtilis* (Wang, et al. 1999). Consequently, *asa* may switch its promoter when stressed with salt. Mechanisms driving a switch require the downregulation of σ^{70} factors, which are usually present in much higher amounts in the cell and have a higher affinity to bind the RNA polymerase, than alternative σ factors (Typas, et al. 2007). The σ^{70} factor is actively prevented from RNA polymerase binding by the recruitment of factors like DskA, ppGpp or 6S RNA and the σ^{38} factor is activated by molecules like the Crl protein (Typas, et al. 2007). This is in correspondence to the observations of stress-shocked EHEC gained by RT-qPCR. The *asa* expression is putatively upregulated by an on-off-switch. An indication for such a regulation mechanism is given by the fact that the mRNA titer of two biological replicates at time point 60 min were of the magnitude of that at time point 30 min and two of that at time point 120 min. EHEC is already in the stationary phase sixty minutes after stress shock and it continues to grow slowly (Supplementary figure S5). Possibly, *asa* is regulated by a σ^{70} factor during exponential phase, stress shock may lead to a switch to the σ^{38} binding promoter at the same +1 site. After entry into the stationary phase, the +1 site 178 bp upstream of the start may be used, which seems to be stronger, as the mRNA titer is higher during stationary phase. This model, however, needs to be experimentally tested. The *asa* expression seems to be affected by salt and increased levels of nutrients (amino acids or vitamins), which might be important when *E. coli* enters the intestinal tract. This would lead to osmotic stress and an increase of nutrient concentrations, necessitating internal changes in the gene expression profiles (Sévin and Sauer 2014; Litsios, et al. 2018).

4.2.3 The protein Asa was validated by three experiments

The *asa* gene is regulated and functional in EHEC EDL933, but this could also be caused by a non-coding RNA (ncRNA). However, there are several indications that *asa* is translated in to a functional peptide of 87 amino acids. Translation signals of RIBOseq experiments provide strong evidence that *asa* is a protein coding gene (Wade and Grainger 2014; Baek, et al. 2017b). Competitive growth experiments of overexpressed *asa* against two different overexpressed, translationally arrested *asa* mutants revealed clear phenotypes. Translationally arrested mutants express the full-length mRNA, but produce only a truncated protein. The phenotype is most probably caused by protein and not by ncRNA, since minor changes (each 2 bp) in the arrested mutants in comparison to full-length *asa* showed these effects. It is unlikely that ncRNA

functionality is affected by the point mutations, also because we tested two locations in independent experiments (Bobrovskyy and Vanderpool 2013). Western Blot verified a protein product, which appears already 0.5 h after induction of overexpression and is still stable 4 h after induction (Figure 3.4.8). The strong band on the blot shows a significant amount of Asa after overexpression, but the natural concentrations in the cell appear to be relatively low (see RIBOseq, Figure 3.4.1).

4.2.4 Asa is a putatively membrane-associated, disordered protein

There are no significant blastp results for Asa, which would give indications for a putative cellular function. PredictProtein was used to predict functional elements. The putative *asa* encoded protein may be predominantly unstructured. It is predicted to have long disordered regions and one α-helix (at least in the full-length version of Asa). Although it has been assumed that structuredness is a requirement for functionality (Koshland Jr 1995), in recent years many cellular processes in which disordered proteins are involved have been revealed (Habchi, et al. 2014). The features of *asa* correspond to those of proteins which are known to be difficult to be detected (see previous section). It is predicted to be secreted, has transmembrane regions and one disulfide bridge. The prediction of a transmembrane helix, secretion and its association with sodium chloride, which changes the membrane integrity (Poolman, et al. 2004), indicate a putative function as membrane protein. However, this hypothesis needs further experimental validation. Additionally, ‘disorder’ is often associated with the presence of transmembrane helices and both were observed in so-far unknown proteins (Perdigão, et al. 2015).

4.2.5 Asa is homologous to a MECP synthase in the far related tsetse fly

A hint for a putative function is given by HHblits results. There are two hits to uncharacterized proteins with a GO-term to a synthase gene (MECP synthase, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase). This synthase is involved in the terpenoid biosynthesis (Herz, et al. 2000), which is an essential cellular process found in all three domains of life (Yamada, et al. 2015). Both hits match to tsetse fly proteins and there are no hits to further organisms. As the tsetse fly has an enterobacteriales symbiont - *Wigglesworthia glossinida* (Genbank NC_016893.1) - one may speculate that the gene has originally been transferred from *Wigglesworthia* to its host. Although a sequence homologue of *asa* is not present in *Wigglesworthia* (no tblastn hits), it has a gene encoding the MECP synthase (no HHblits hit in *asa*). Interestingly, the MECP synthase gene of *Wigglesworthia* (WP_014354005.1, location

216779 / 217258) is located in antisense, next to, but not overlapping to the mother gene homologue of *asa* (WP_014354006.1, location 27424 / 218077). As this seems to be impossible by chance, it can be speculated that a long ago duplication of the overlapping gene pair occurred, which evolved to the loss of respectively one gene. Consequently, the overlapping gene pair was uncoupled. This needs to be further examined in future. However, there are some arguments against the hit to be reliable. While Asa is predominantly disordered, the MECP synthase is a globular protein and has a different PredictProtein pattern (RCSB entry: 1JY8, PredictProtein see Supplementary figure S16). EHEC EDL933 also has an annotated MECP synthase (locus tag EDL933_3916), which has no sequence similarity to *asa*. The enzyme is zinc dependent and *asa* did not show any phenotype in $ZnCl_2$ (Zehentner 2015). Finally, the significance of the hit is not strong enough to be sure that *asa* is a homologue of the MECP synthase encoding gene, but it might be a hint for a function, which needs be experimentally tested.

4.3 Shadow ORFs are evolutionarily young genes

There are only few studies on the age of bacterial overlapping genes published so far and those analyzed single genes (Delaye, et al. 2008; Fellner 2015; Hücker, Vanderhaeghen, Abellan-Schneyder, Scherer, et al. 2018; Hücker, Vanderhaeghen, Abellan-Schneyder, Wecko, et al. 2018). Here, we used phylostratigraphy, which is a large-scale approach established in eukaryotes (Domazet-Lošo, et al. 2007). Blastp on all sORFs was used to identify their taxonomic distribution. The key result is that nearly half of the sORFs are taxonomically restricted to *E. coli*, whereas 25% are highly conserved and found in many bacteria. Thus, many sORFs are extremely young, but some are also quite old (Figure 3.1.6). This is different to the gene age of the aORFs (4% restricted to *E. coli*, 78% highly conserved). Surprisingly, many sORFs and aORFs, matching to hypothetical proteins, are highly conserved, which was already detected elsewhere (Galperin and Koonin 2004).

Next, it was predicted, whether features of sORF proteins correspond to those of young genes. Taxonomically restricted genes have been suggested to carry protein features that are less ‘mature’ and less complex in comparison to conserved genes (Carvunis, et al. 2012). Carvunis, et al. (2012) reported on observations, which support a continuum in presumed evolutionary protein maturation.

First, the sORF and aORF length increases with increasing phylostratigraphic level (Figure 3.3.1 A). This was also observed elsewhere (Lipman, et al. 2002; Carvunis, et al. 2012; Schmitz, et al. 2018). Second, similar to Abrusán (2013), there is an increasing number of nucleotide interactions with increasing phylostratigraphic level (Figure 3.3.1 B, C). Nucleotide binding proteins are often ATP or GTP dependent proteins involved in many processes like membrane transport, cell division, haemolysin export or enzyme dependent, and energy demanding processes (Higgins, et al. 1986). Further, DNA or RNA binding proteins are involved in gene expression or regulation (Ren, et al. 2000). The latter ones belong to a class of proteins, which are essential in all three domains of life. They were even putatively present in LUCA, the last universal common ancestor (Weiss, et al. 2016). Consequently, conserved proteins should be involved in nucleotide binding more than non-conserved proteins.

Third, both sORFs and aORFs show a trade-off for proteins predicted to be secreted in comparison to proteins predicted to be cytoplasmic (Figure 3.3.2). The number of predicted secreted proteins is high in taxonomically restricted genes and strongly increasing by phylostratigraphic level. Secreted proteins can be exported into the media or anchored to the outer membrane (Desvaux, et al. 2009). They are effector molecules to change the environmental niche, find access to resources or supply interactions with other organisms (Nogueira, et al. 2012). They can further be involved in pathogenicity (Henderson, et al. 2004), for example as proteases (Henderson, et al. 1999). Genes encoding secreted proteins are known to be evolutionary young and have high mutation rates in contrast to cytoplasmic proteins (Nogueira, et al. 2012). The number of disulfide bonds of annotated genes is also decreasing by phylostratigraphic level, which is in correspondence to the cell localization (Figure 3.3.3 B). This is in correspondence to the slightly decreasing number of cysteines by phylostratigraphic level (Figure 3.3.3 C). The sORF proteins do not show any trend (Figure 3.3.3 A).

Forth, a higher protein complexity would be equivalent with a higher ‘structuredness’ and more functional elements. However, sORFs or aORFs do not show any changes in the number of protein binding sites (Figure 3.3.1 D), in secondary structure (Figure 3.3.5 A, B) or solvent accessibility (C, D) by increasing or decreasing phylostratigraphic level. The percentage of disordered regions per ORF is first increasing, later decreasing by age (Figure 3.3.4 E) in correspondence to an observation of Carvunis, et al. (2012) in yeast; however, these authors did not give any explanations. After Perdigão, et al. (2015), disorder is a feature of proteins, which is independent from a gene being an orphan or not. Carvunis, et al. (2012) observed a decreasing number of transmembrane helices with increasing age, but EHEC EDL933 sORFs and aORFs do not show any significant trend (Figure 3.3.6).

The predicted protein features in dependence on gene age, observed for aORFs, were also visible for the sORFs, to a less clear extent. This is an indication that sORFs are putatively functional genes. However, the trends are often less clear or disappear for reasons which will be discussed in more detail in section 4.6. The bacterial genes, analyzed here, can be compared to eukaryotic genes analyzed in the studies previously discussed regarding protein properties at different phylostratigraphic gene ages. The trends are often less extensive, and the standard deviations are high, which was also the case in comparable studies (particularly in Abrusán 2013; Neme and Tautz 2013; Neuhaus, et al. 2016). The reason is that protein features are only indirectly dependent on the functionality of a protein and, thus, protein features often do not clearly correlate. For example, proteins of the same length can have different numbers of disulfide bonds. Rather, the disulfide bonds depend on the localization and on its ability to form tertiary structures of proteins (Creighton 1997; Dutton, et al. 2008), which is a feature of many different proteins.

4.4 Evolution of sORF sequences

Results from an overall blastp analysis revealed that sORFs in tendency are either very young or very old (see section 4.3). However, a more detailed analysis of the evolution of sORFs must consider individual sequences. Therefore, phylostratigraphic trees were constructed from a total of 19 sORFs with phenotypes to understand their sequence evolution. The phylostratigraphic levels of sORFs analyzed in single gene phylostratigraphy are distributed in a broader range than those determined with blastp (Table 3.6.1). Only four sORFs are assigned to the phylostratum '*E. coli/Shigella*' and one to 'bacteria/archaea'.

An individual analysis of functional sORFs was used to visualize the transition from non-intact to intact ORFs. Evolutionary events visible in the trees are *de novo* gene emergence, gene loss and horizontal gene transfer, whereas a clear differentiation between emergence and loss cannot be determined from sequence alone. Indications on gene emergence are given by the farthest organism in which an intact sORF homologue is present. Gene loss is determined by non-intact, i.e. disrupted sequences, which is usually accompanied by a high mutation rate due to a loss of evolutionary constraint (Tautz and Domazet-Loso 2011). Today, it is generally accepted that gene gain and loss are balanced over time, which can be explained by the relatively constant number of genes per genome (Tautz and Domazet-Loso 2011). Horizontal gene transfer may blur the picture of orthologous evolution and can lead to misinterpretations concerning the time of gene emergence in case horizontal gene transfer was an early event

(Tautz and Domazet-Loso 2011). I observed one apparently clear horizontal gene transfer event in OGC106 between Gram-positive bacteria and enterobacteriaceae.

4.4.1 Phylostratigraphic trees reveal mechanisms for gene gain

The emergence of overlapping genes occurs by overprinting, either by a gradual or by discontinuous evolution (Keese and Gibbs 1992). A discontinuous emergence is clearly present in some genes taxonomically restricted to *Escherichia/Shigella*, like OGC75 (Supplementary figure S13.5) or OGC85 (Supplementary figure S13.6). A gradual emergence is visible by conserved amino acid positions and amino acids gradually mutated during evolution. The loss of stop codons can lead to continuous open reading frames as described by Delaye, et al. (2008). In this dissertation, a gradual emergence is observed for example in OGC198 (Supplementary figure S13.11). Species far related to EHEC even have some stop codon positions conserved, which are lost in some genes. Genes also emerge by the shift of an upstream initiation codon like in OGC167, which is taxonomically restricted to EHEC and few *Escherichia coli* (Supplementary figure S13.8 and S14.4). An upstream extension of the non-overlapping region, which has a low evolutionary constraint, leads to the gain of the gene sequence (OGC167). This mechanism can also lead to the extension of the mother gene (Figure 3.1.2).

4.4.2 Shadow ORFs are frequently lost during evolution

All mechanisms described for gene emergence can also lead to gene loss, which was observed frequently and randomly distributed throughout the species trees (for example *asa*, Figure 3.5.3). The sequence evolution is gradual in OGC57, OGC121, OGC194, OGC198, OGC231, OGC241 and *asa*). The majority of non-intact sequences, which were found in bacteria related farther away from EHEC, clearly have a lower sequence identity to the original or other intact sequences (for example in OGC57) indicating that a putative loss of functionality quickly leads to the sequence degradation. Surprisingly, other non-intact sequences have equal or higher sequence identities to the original or other intact sequences. The high abundance of non-intact sequences in OGC homologues indicate a rapid evolution, which supports the continuum hypothesis (Carvunis, et al. 2012). Interestingly, a frequent loss of intactness was not only observed in young genes (Schlötterer 2015), but also in taxonomically widely distributed OGCs.

4.4.3 Overlapping gene pairs show dynamic sequence evolutions

Sequence disruption may also occur by internal sequence substitutions. In case of an in-frame substitution the functionality can be maintained, like in the *ano* homologue present in *Escherichia fergusonii.* The nucleotide substitutions in the sORF sequences are also visible in the respective mother gene (Figure 3.6.7). The protein structure prediction of the mother genes reveals any significant structural changes caused by this substitution. Both proteins are disordered in this region (Supplementary figure S17). Out-of-frame substitutions can lead to frameshifts and length variations, which disrupt the sequence as observed for some mother genes of annotated sORF homologous sequences (Figure 3.1.2). Different sequence lengths, split, partial disruption and even protein domain fusions are effects of sequence changes observed in the phylostratigraphic trees. These were already described in a previous study for the overlapping gene pair *htgA/yaaW* (Delaye, et al. 2008). While *yaaW* is widely conserved throughout the phylogenetic tree, the *htgA* gene underwent many sequence changes. As the sequence is conserved only in *Escherichia coli*/*Shigella*, the authors conclude that *htgA* was overprinted at the split of *Escherichia coli/Shigella* from other Enterobacteria. The overlapping genes, phylostratigraphically analysed here, partially show massive changes leading to alternating intact and non-intact sequences.

4.4.4 Internal stop codons split ORFs or lead to pseudogenization

Internal stop codons can split an ORF into two ORFs, which occurred presumably in the *Salmonella enterica* homologue of *slyC* (Figure 3.6.9). When comparing *slyC* in *Salmonella* to other homologues, the gene split or pseudogenization is a *Salmonella* specific phenomenon. Both ORFs are very short (93 bp and 117 bp). The mutation causing a stop codon in *Salmonella* is silent in the mother gene. The amino acid ‘valine’ (position 108) in the mother gene has a T → C shift at the codon position three leading to a ‘TGA’ stop codon instead of a ‘TTA’ (leucine, L) in the shadow ORF. Without any additional experimental data (e.g. RNAseq data), it is unclear whether the gene is pseudogenized or forms two novel independent genes (Richardson and Watson 2013). The same phenomenon, a stop codon mutation being silent in the mother gene, was also observed in *asa*. In this case, there is no subsequent start codon. The introduction of a stop can be assumed to cause a pseudogenization for this gene. The introduction of a stop codon is the easiest possibility for a gene loss, or, if proceeding in the other way around, emergence.

4.4.5 The evolutionary constraint is determined by reading frame and overlap type

Both reading frame and overlap type affect, which case of emergence occurs and both interrelate simultaneously (Table 3.6.1). High constraints promote gradual evolution and low constraints allow discontinuous emergence events. The evolutionary constraint caused by the relative reading frame is caused by different overlapping codon-positions (Lèbre and Gascuel 2017). The mutation of nucleotide position 2 has the highest constraint, because a substitution always causes an amino acid change. The constraint on position 1 is higher than that on position 3. Gene pairs overlapping in frame -2 have the highest constraint with a 123/213 pairing (position 1 of the mORF is position 2 in the sORF). Reading frame -1 has a 123/321 and frame -3 a 123/132 pairing. Consequently, the reading frame order by increasing constraint is -1 ≤ -3 ≤ -2. This theory was confirmed in the following overlapping genes. OGC106, OGC174 and OGC194 are lying in -1 frame to their mother gene and emerge in two discontinuous events; OGC75, OGC85, OGC226 and OGC241 in one event. The high constraint of -2 frame promotes gradual evolution, like in OGC198 or *asa*. One exception was OGC51, which emerges discontinuously and lays in -2 frame. The ORF emergence often occurred as a combination of discontinuous and gradual evolution (for example in OGC121 or OGC226).

Overlapping parts often have a higher evolutionary constraint than non-overlapping parts like in OGC23, OGC167, OGC174, OGC226 and OGC231. One exception is OGC51, which shows no difference between the overlapping and non-overlapping part. Non-overlapping regions allow the disruption of the genes by indel mutations, which can cause frameshifts (mORF see Figure 3.1.2 or sORF like for OGC75). Tail-to-tail overlapping genes were observed to be young, because the first part of the sequence, including the start codon, is not conserved (OGC167, OGC226, OGC231). The naturally used start codons in EHEC sORFs are unknown. There is the possibility that start codons in the overlapping part are more conserved and used in sORF homologues (e.g. in OGC231). The evolution of embedded genes is often similar to that of the mother gene (e.g. for OGC15, OGC198 and *asa*).

4.4.6 The independent evolution of an overlapping gene pair is possible

The big question is whether an independent evolution of sORFs and mORF is possible. Keese and Gibbs (1992) could only imagine that independent evolution of overlapping genes can occur via decoupling by duplication and divergence. The phylostratigraphic trees revealed an uncoupled evolution of in OGC23, OGC75, OGC106 and OGC226. In OGC231 and OGC241, the sORF degradation is quicker than that of the mother gene, but the overall trend is

comparable in both genes. This reveals a dependency of both genes with a freedom in the evolutionary constraint, which enables individual evolution of both genes.

4.5 Transcription pattern of *asa* support the continuum hypothesis

The presence of a sequence is a requirement for a functional gene, but the gene can be inactive or not yet be transcribed. Further, *de novo* emerged genes can first be transcribed only and be protein-coding in a more matured state (Carvunis, et al. 2012). Here, we find clear indications for this hypothesis for *asa*. The small gene was analyzed in respect to transcription (RNAseq, RT-qPCR) and translation (RIBOseq) of homologues present in bacteria of the order enterobacteriales after growth in LB medium. Those with intact sequence were compared to those with internal stop at amino acid position 62. Two datasets were used. The first had next generation sequencing data available. Since RIBOseq is a relatively new method (first published in 2014 by Ingolia, et al.), and as most studies on Enterobacteria are motivated by clinical studies, there are combined RNAseq and RIBOseq data available only for *Escherichia coli* and *Salmonella enterica*. Both organisms are closely related and all *Escherichia coli* have intact and all *Salmonella enterica* have non-intact *asa* homologues. RNAseq signals were discovered in further organisms within the order enterobacteriales. The second dataset implements species without transcriptome or translatome data, but the organisms were experimentally characterized in respect to a phenotype and *asa* homologue expression using RT-qPCR. The four enterobacteriales species had either an intact (*C. freundii* CFNIH1, *S. marcescens* WS1359) or non-intact (*S. enterica* 287/91, *H. alvei* DSM30097) *asa* homologues.

When looking at the transcription of *asa* homologues, there are two patterns. In case of a protein expression, there is one single peak at the *asa* start. Western Blot shows that the full-length *asa* protein is expressed in EHEC EDL933 (Figure 3.4.8). This RIBOseq signal was observed in the pathogenic *E. coli* strains EHEC Sakai and AIEC LF82, two close relatives of EHEC EDL933. The second pattern is a strong signal for a transcript of the length of the mother gene indicating antisense RNA (Thomason and Storz 2010). Antisense RNA is situated in antisense to an annotated gene and regulates the dosage of its mRNA by antisense binding (Lybecker, et al. 2014). Nearly all organisms show a comparably strong pattern of sORF and mORF transcription (Figure 3.5.6 and Figure 3.5.7). This could indicate a regulation by stabilization of a mRNA by forming stable secondary structures and release of the ribosomal binding sites (Lybecker, et al. 2014). This was observed in close relatives (EPEC E2348, *S. flexneri, S. enterica*) and in more distantly related species (*C. rodentium, S. praecaptivus, C. sakazakii*). Further, no signals were

observed for either, closely related species or distantly related species (*E. coli* K12, *E. aerogenes, K. pneumoniae, S. marcescens*).

As expected, Asa production was validated by RIBOseq in strains with intact sequences. Putative antisense RNA was observed also in non-intact sequences, which is also expected, because stop codons do not necessarily affect transcription *per se*. However, it cannot be excluded that intact genes with putative antisense RNA expression pattern do not produce a protein. There are only RIBOseq data available for *Salmonella enterica* SL1344, which has a truncated homologue and, indeed, shows no protein product - an important negative control (Figure 3.5.6). The more distantly related organism with putative antisense RNA signal was *C. sakazakii.* In contrast, no antisense RNA signal was seen in *S. marcescens* strain WW4. RT-qPCR revealed *asa* homologue expression in all species tested; hence, *asa* is transcribed in *S. marcescens* strain WS1349. If *asa* is in fact protein coding in *E. coli*, but it encodes a functional RNA in more distantly related organisms, this would explain the unclear phenotype of *asa* homologues. It is possible that translationally arrested mutants do not correctly fold to the RNA structure required to be functional (Bobrovskyy and Vanderpool 2013). Due to the low number of expression datasets available, the age of functional *asa* cannot be exactly determined. It can just be assumed that it correlates with the emergence of an ORF, but it is not active in all organisms. Interestingly, Peer and Margalit (2014) found out that many regulatory small RNAs emerged at the split of enterobacteria with other γ-proteobacteria, which more or less correlates with the *asa* gene age. Further, the homologues were not further characterized or have unclear data and it is known that transcribed antisense RNA also can be spurious or byproducts (Lybecker, et al. 2014). Even if this is the case in some homologues, we seem to be able to observe a gradual evolution in *asa*.

A further hint for functionality during evolution is given by the conservation of the promoter region. There were three putative promoters at two putative transcriptional start sites identified in *asa*. All three promoters are conserved only in *E. coli* and *Shigella*. The σ^{70} promoter is not present in all other organisms and the σ^{38} promoters could be present with a changed sequence in further organisms. Unfortunately, this does not correspond to the transcription pattern, which is an indication that the conservation in *E. coli* comes from the close species relations rather than from a functional constraint. Similar to these observations, previous studies about antisense RNA found out that antisense regulatory RNA transcription is regulated by σ^{70} promoters (Lybecker, et al. 2014), their regulating promoters are not conserved within Enterobacteria (Raghavan, et al. 2012) and that regulatory elements can be very species specific (Browning and Busby 2016). However, the activity of the homologous promoter sites needs further experiments for a conclusive statement about the regulation of *asa* homologues.

4.6 Is phylostratigraphy suitable for gene age determination?

The phylostratigraphic method first used by Domazet-Lošo, et al. (2007) has the following steps: 1) finding all ORFs encoded proteins of a genome with homologues in other organisms using blastp, 2) classifying the corresponding ORFs into gene ages according to the furthermost related organism in which a homologue appears and 3) to identify gene expression profiles, which indicate functionality of the gene. The first who criticized the phylostratigraphic method were Moyers and Zhang (2014). Thereafter, it was extensively discussed in the literature (Moyers and Zhang 2014; Moyers and Zhang 2016; Domazet-Lošo, et al. 2017; Moyers and Zhang 2018). The most critical point was that using blastp introduces a bias which leads to different results for different types of genes (Moyers and Zhang 2014). Those will be discussed in the following.

4.6.1 The phylostratigraphic level is distorted by the dependency of the E-value on the sequence length and the number of hits

The most distantly related organism is determined by an E-value cutoff in the blastp analysis. The E-value indicates the significance of a hit, which strongly depends on sequence length (Wolf, et al. 2009; Moyers and Zhang 2014). Short sequences are less significant, because they have a higher probability to match by chance to a sequence in the database. This becomes particularly problematic at the limit of allowed sequence length (~200 bp = 67 aa, FAQ, blast, NCBI 2018). For this reason, the average amino acid sequence length of sORFs with blastp hit (120 aa) is twice of that of all sORFs found in the genome (63 aa). The aORFs are more than twofold longer (296 aa) than the sORFs with blastp hit. It was observed that short ORFs have relatively high E-values (i.e. tendency to be judged non-significant) even when the match appeared in the organism they originated from (not analyzed in detail).

Further, short sequences are underrepresented in protein databases. The mRNA or proteins of short genes are often less stable in a cell in comparison to those of longer genes and are harder to be detected (Baboo and Cook 2014). Small gene products are often at the technical detection limit of high-throughput methods like mass spectrometry and many are expressed only under particular conditions (Hemm, et al. 2010; Landstorfer, et al. 2014; Storz, et al. 2014; Hücker, et al. 2017). Underrepresentation consequently leads to a predicted younger age of the sORFs in comparison to aORFs. By fixing an E-value cut-off, short ORFs are treated equal to long ORFs, which they are not. As consequence, non-significant hits may be included after using blast on

long ORFs and significant hits will be missed for the short ORFs. Thus, a global E-value cut-off may not be a suitable criterion for gene age determination, but using length dependent E-value cutoffs could be a solution, which has to be tested.

Further, the error caused by the E-value cutoffs, particularly occurs in a gradual sequence evolution. For *asa*, an underestimation of the gene age was observed, when using a stringent cutoff in comparison to a less stringent cutoff. At an E-value of 10^{-3}, some γ-proteobacteria with intact *asa* homologues were discovered, which were excluded at an E-value cutoff of E^{-10}. The phylostratigraphic tree however shows that those sequences are significant and reveal the *asa* age (Figure 3.5.1).

Additionally, taxonomically restricted genes have a high evolution rate and therefore lower sequence similarities to homologues, which makes them harder to be detected by blast. False negative rates of blast hits can be up to 100% in such cases (Elhaik, et al. 2006). This further has the danger to overlook organisms, in which the gene is lost leading to an underestimation of the gene age.

4.6.2 Blastp is suitable for age determination of non-overlapping genes

However, there are statements that this problem seems only to be significant in minor cases. Moyers and Zhang (2014) found a non-detection of distant relative sequences of 13.85% of blast in their gene evolution simulation. Carvunis, et al. (2012) found a gene-age underestimation rate of 5% using a combination of blastp, tblastn and tblastx. Neme and Tautz (2013) concluded that such an error rate is acceptable, because only overall trends would be of interest. Moyers and Zhang (2018) tried to improve phylostratigraphy. They first tested the robustness of blastp in comparison to other protein search methods and concluded that blastp is contemporarily the most reliable protein search method. An improvement was obtained by the exclusion of error-prone, taxonomically restricted genes, which show simulated evolution pattern different from real data obtained using phylostratigraphy. In an earlier study, Moyers and Zhang (2014) found out that the phylostratigraphic gene age does not significantly change, when using different cutoffs between 10^{-1} and 10^{-10}. Thus, the authors suggested using phylostratigraphy only after removal of error-prone genes.

4.6.3 Blastp is not suitable age determination of overlapping genes

Overlapping genes are short (Figure 3.3.1), underrepresented in databases (section 3.1.1 and 3.1.2), taxonomically restricted (Figure 3.1.6) and 'not allowed' to be annotated in case of an annotated mother gene at the same locus (Delcher, et al. 2007). These features indicate that sORFs belong to the error-prone genes described by Moyers and Zhang (2018). Thus, there is the question of whether phylostratigraphy can be used for sORFs. When comparing the E-values of sORFs and aORFs, there is no significant difference (Figure 3.2.3) demonstrating some robustness of the method. However, it was observed that there is a low abundance of blastp hits per sORF encoded protein (14.3 ± 233 hits) in comparison to annotated proteins (26,896 ± 107,315 hits). In comparison to the number of hits in tblastn (section 3.6), many organisms containing a sORF homologue do not have an entry in blastp refseq. As consequence, many hits in organisms between the identified 'farthermost related species' and EHEC EDL933 are missing. In worst case, the blastp hits are not representative and the organisms, which are shown to be the farthermost related ones, are wrong. Thus, the results from the aORF dataset are more reliable than that of the sORF dataset and the gene age of the sORFs may be underestimated.

4.6.4 Tblastn of the mORF can identify the furthermost species with intact sORF

In a second phylostratigraphic approach, the gene age determination of sORFs was uncoupled from blast by using the blast hits of the mother gene. The reverse complement sequence of the mother gene homologues was scanned for a sORF homologue. Two assumptions were deployed to do this: 1) Shadow ORFs are taxonomically restricted in contrast to annotated genes, hence, most of them should be younger or of equal age than the mother gene. 2) The gene age is determined by the presence of an intact, homologous sequence instead of blast search alone. As an intact sORF sequence does not guarantee functionality, the phylostratigraphy was only implemented for sORFs with a phenotype.

In this approach, tblastn was used. The main advantage of tblastn in comparison to blastp is that the database is bigger, because all genomes were fully checked for sequence identity to the EHEC EDL933 sORF. This makes the search independent from annotation and from the fact that sORFs are not allowed. Further, the model organism used (*E. coli*) has many sequenced strains and taxonomically restricted genes have many hits in this species, which increases the significance. Nonetheless, hits are less significant in case of short sORFs. The genome sequence data enable to go more in detail and show the events, which lead to sequence gain or loss as discussed in section 4.4.

4.6.5 Blast needs further validation

The blast algorithm reports only on the matching region between query and subject, which might be the full query or only parts thereof. Partly matching ORFs decrease the length of the matching part and, hence, the significance of a hit. This has a significant impact on underestimation of the gene age (Moyers and Zhang 2014). Not matching regions can be caused by a varying evolutionary rate at different sites of the sequence causing point mutations or indel mutations. Only 280 of 2,180 blastp hits fully matched to the sORF proteins. However, using hits with a high sequence coverage does significantly decrease the dataset. Tblastn additionally shows non-intact sequences including those indel mutations leading to a loss of function and including internal stop codons. Thus, blast needs a further check for the intactness of the hits.

4.6.6 MLSA trees can be used to trace back the species evolution

EHEC EDL933 is the first bacterial species for which phylostratigraphy was conducted. The gene age determination strongly depends on the species tree used. Distances based on 16S rRNA sequence identity cutoffs were used for taxonomic classifications of phylostratigraphy of all EHEC sORFs (section 3.3) as obtained by Yarza, et al. (2014). The resolution of 16S rRNA trees is good on species level, but not at the strain level (Yarza, et al. 2014). The phylostratigraphy of single genes is more precise, because it uses MLSA trees for close relatives (section 3.6). The precision could be increased by using whole genome approaches like ANI, but MLSA shows a good correlation to ANI (Average Nucleotide Identity, Böhm, et al. 2015). However, most mother genes have a massive number of blast hits. Thus, not all species and strains, having a mother gene homologue, were considered. Consequently, an underestimation of the gene age cannot be excluded. It has to be kept in mind that bacterial species evolution is always blurred by horizontal gene transfer, which predominantly occurs between closely related organisms (Kloesges, et al. 2011). It was recently reported that horizontal gene transfer is also possible from bacteria to eukaryotes (Husnik and McCutcheon 2017). Although this is a rare event, it can cloud the picture of vertical evolution and leads to gene age overestimation. Nonetheless, horizontal gene transfer increases the error rate, but it does not significantly change the overall-trend.

4.6.7 An improved phylostratigraphy includes blastp, tblastn and experimental data

Based on these problems mentioned above, I suggest an improved procedure to analyze sORF phylostratigraphies. First, sORFs of interest should be analyzed by blastp, thus providing evidence whether sORFs have annotated homologues and, therefore, might be functional. If available, RNAseq and RIBOseq data can be used to validate expression and putative functionality of the sORFs. Shadow ORFs with blastp hits are analyzed using tblastn to identify the farthermost related hits, which are checked for intactness. In case of being not intact, the next, more closely related organism is checked and so on. However, a compelling phylostratigraphic analysis of sORFs requires using single genes, which are functionally characterized, as has been performed in this thesis for *laoB* (Hücker, Vanderhaeghen, Abellan-Schneyder, Wecko, et al. 2018), *ano* (Hücker, Vanderhaeghen, Abellan-Schneyder, Scherer, et al. 2018), *asa* (Vanderhaeghen et al., accepted), and *slyC* (Hücker 2017), as well as the other OGCs taken (Zehentner, unpublished). It can be stated that phylostratigraphy is by all means an important method to determine sORF ages, but any blast analysis needs thorough checks for validation. In coming years, the increasing surprising evidence for functionality of OLGs should be complemented by more in-depth evolutionary studies of how this functionality has arisen and evolved over time. This thesis provides a starting point and framework for this future work.

5 References

Abrusán G. 2013. Integration of new genes into cellular networks, and their structural maturation. Genetics:genetics. 113.152256.

Albà MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. BMC Evolutionary Biology 7:53.

Albanese A, Sacerdoti F, Seyahian EA, Amaral MM, Fiorentino G, Brando RF, Vilte DA, Mercado EC, Palermo MS, Cataldi A. 2018. Immunization of pregnant cows with Shiga toxin-2 induces high levels of specific colostral antibodies and lactoferrin able to neutralize E. coli O157: H7 pathogenicity. Vaccine 36:1728-1735.

An Y, Ji J, Wu W, Lv A, Huang R, Wei Y. 2005. A rapid and efficient method for multiple-site mutagenesis with a modified overlap extension PCR. Applied Microbiology and Biotechnology 68:774-778.

Arthur TM, Kalchayanand N, Agga GE, Wheeler TL, Koohmaraie M. 2017. Evaluation of bacteriophage application to cattle in lairage at beef processing plants to reduce Escherichia coli O157: H7 prevalence on hides and carcasses. Foodborne pathogens and disease 14:17-22.

Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, De Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E. 2012. ExPASy: SIB bioinformatics resource portal. Nucleic Acids Research 40:W597-W603.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, et al. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Baboo S, Cook PR. 2014. “Dark matter” worlds of unstable RNA and protein. Nucleus 5:281-286.

Baek J, Lee J, Yoon K, Lee H. 2017a. Identification of unannotated small genes in Salmonella. G3: Genes, Genomes, Genetics:g3. 116.036939.

Baek J, Lee J, Yoon K, Lee H. 2017b. Identification of Unannotated Small Genes in *Salmonella*. G3 (Bethesda) 7:983-989.

Balabanov VP, Kotova VY, Kholodii GY, Mindlin SZ, Zavilgelsky GB. 2012. A novel gene, *ardD*, determines antirestriction activity of the non-conjugative transposon Tn5053 and is located antisense within the *tniA* gene. FEMS Microbiol Lett 337:55-60.

Barrell BG, Air GM, Hutchison CA, 3rd. 1976. Overlapping genes in bacteriophage ΦX174. Nature 264:34-41.

Becker G, Hengge-Aronis R. 2001. What makes an Escherichia coli promoter σS dependent? Role of the−13/− 14 nucleotide promoter positions and region 2.5 of σS. Molecular Microbiology 39:1153-1165.

Behrens M, Sheikh J, Nataro JP. 2002. Regulation of the overlapping *pic*/*set* locus in *Shigella flexneri* and enteroaggregative *Escherichia coli*. Infection and Immunity 70:2915-2925.

Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. 2005. Non-classical protein secretion in bacteria. BMC Microbiology 5:58.

Bertin Y, Habouzit C, Dunière L, Laurier M, Durand A, Duchez D, Segura A, Thévenot-Sergentet D, Baruzzi F, Chaucheyras-Durand F. 2017. Lactobacillus reuteri suppresses E. coli O157: H7 in bovine ruminal fluid: Toward a pre-slaughter strategy to improve food safety? PLoS One 12:e0187229.

Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Research 29:2607-2618.

Bobrovskyy M, Vanderpool CK. 2013. Regulation of bacterial metabolism by small RNAs using diverse mechanisms. Annu Rev Genet 47:209-232.

Böhm M-E, Huptas C, Krey VM, Scherer S. 2015. Massive horizontal gene transfer, strictly vertical inheritance and ancient duplications differentially shape the evolution of Bacillus cereus enterotoxin operons hbl, cytK and nhe. BMC Evolutionary Biology 15:246.

Bolten E, Schliep A, Schneckener S, Schomburg D, Schrader R. 2001. Clustering protein sequences—structure prediction by transitive homology. Bioinformatics 17:935-941.

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. 2007. UniProtKB/Swiss-Prot - The manually annotated section of the UniProt KnowledgeBase. Methods Mol Biol 406:89-112.

Browning DF, Busby SJ. 2016. Local and global regulation of transcription initiation in bacteria. Nature Reviews Microbiology 14:638.

Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. Nat Methods.

Callaway TR, Carr M, Edrington T, Anderson RC, Nisbet DJ. 2009. Diet, Escherichia coli O157: H7, and cattle: a review after 10 years. Current Issues in Molecular Biology 11:67.

Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2011. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28:464-469.

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and *de novo* gene birth. Nature 487:370–374.

Castro VS, Carvalho RCT, Conte-Junior CA, Figuiredo EES. 2017. Shiga-toxin Producing Escherichia coli: Pathogenicity, Supershedding, Diagnostic Methods, Occurrence, and Foodborne Outbreaks. Comprehensive Reviews in Food Science and Food Safety 16:1269-1280.

Ceroni A, Passerini A, Vullo A, Frasconi P. 2006. DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. Nucleic Acids Research 34:W177-W181.

Cerveau N, Leclercq S, Bouchon D, Cordaux R. 2011. Evolutionary dynamics and genomic impact of prokaryote transposable elements. In. Evolutionary biology–concepts, biodiversity, macroevolution and genome evolution: Springer. p. 291-312.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. bioRxiv:274100.

Cheng CH, Yang CH, Chiu HT, Lu CL. 2010. Reconstructing genome trees of prokaryotes using overlapping genes. BMC Bioinformatics 11:102.

Chirico N, Vianelli A, Belshaw R. 2010. Why genes overlap in viruses. Proc Royal Soc B: Biol Sci 277:3809-3817.

Chung WY, Wadhawan S, Szklarczyk R, Pond SK, Nekrutenko A. 2007. A first look at ARFome: dual-coding genes in mammalian genomes. PLoS Comput Biol 3:e91.

Creighton TE. 1997. Protein folding coupled to disulphide bond formation.

Danchin A, Ouzounis C, Tokuyasu T, Zucker JD. 2018. No wisdom in the crowd: genome annotation in the era of big data–current status and future prospects. Microb Biotechnol.

Daubin V, Moran NA, Ochman H. 2003. Phylogenetics and the cohesion of bacterial genomes. Science 301:829-832.

de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34:W362-365.

Delaye L, Deluna A, Lazcano A, Becerra A. 2008. The origin of a novel gene through overprinting in *Escherichia coli*. BMC Evol Biol 8:31.

Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23:673-679.

Desvaux M, Hébraud M, Talon R, Henderson IR. 2009. Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. Trends in Microbiology 17:139-145.

Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends in Genetics 23:533-539.

Domazet-Lošo T, Carvunis A-R, Albà M, Šestak MS, Bakarić R, Neme R, Tautz D. 2017. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. Molecular Biology and Evolution 34:843-856.

Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in Drosophila. Genome Research 13:2213-2219.

Domazet-Lošo T, Tautz D. 2010a. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature 468:815-818.

Domazet-Lošo T, Tautz D. 2010b. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. BMC Biology 8:66.

Du J, Yuan Z, Ma Z, Song J, Xie X, Chen Y. 2014. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. Molecular bioSystems 10:2441-2447.

Durand É, Gagnon-Arsenault I, Hatin I, Nielly-Thibault L, Namy O, Landry CR. 2018. The high turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. bioRxiv:329730.

Dutton RJ, Boyd D, Berkmen M, Beckwith J. 2008. Bacterial species exhibit diversity in their mechanisms and capacity for protein disulfide bond formation. Proceedings of the National Academy of Sciences 105:11933-11938.

Elhaik E, Sabath N, Graur D. 2006. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. Mol Biol Evol 23:1-3.

Eriksson K, Boyd S, Tasker R. 2001. Acute neurology and neurophysiology of haemolytic–uraemic syndrome. Archives of Disease in Childhood 84:434-435.

Ettwiller L, Buswell J, Yigit E, Schildkraut I. 2016. A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. BMC Genomics 17:199.

Feklistov A, Sharon BD, Darst SA, Gross CA. 2014. Bacterial sigma factors: a historical, structural, and genomic perspective. Annual Review of Microbiology 68:357-376.

Fellner L. 2015. Fuctional characterization of overlapping genes in the food-borne pathogen Escherichia coli O157:H7. [PhD]. [Freising]: TU München.

Fellner L, Bechtel N, Witting MA, Simon S, Schmitt-Kopplin P, Keim D, Scherer S, Neuhaus K. 2014. Phenotype of *htgA* (*mbiA*), a recently evolved orphan gene of *Escherichia coli* and *Shigella*, completely overlapping in antisense to *yaaW*. FEMS Microbiology Letters 350:57-64.

Fellner L, Huptas C, Simon S, Mühlig A, Scherer S, Neuhaus K. 2016. Draft genome sequence of three European lab-derivates from the enterohemorrhagic *Escherichia coli* O157:H7 strain EDL933, including two plasmids. Genome Announcements 4:e01331-01315.

Fellner L, Simon S, Scherling C, Witting M, Schober S, Polte C, Schmitt-Kopplin P, Keim DA, Scherer S, Neuhaus K. 2015. Evidence for the recent origin of a bacterial protein-coding, overlapping orphan gene by evolutionary overprinting. BMC Evolutionary Biology 15:1.

Feltens R, Gößringer M, Willkomm DK, Urlaub H, Hartmann RK. 2003. An unusual mechanism of bacterial gene expression revealed for the RNase P protein of Thermus strains. Proceedings of the National Academy of Sciences 100:5724-5729.

Fernandes JD, Faust TB, Strauli NB, Smith C, Crosby DC, Nakamura RL, Hernandez RD, Frankel AD. 2016. Functional segregation of overlapping genes in hiv. Cell 167:1762-1773. e1712.

Filiatrault MJ, Stodghill PV, Bronstein PA, Moll S, Lindeberg M, Grills G, Schweitzer P, Wang W, Schroth GP, Luo S, et al. 2010. Transcriptome analysis of *Pseudomonas syringae* identifies new genes, noncoding RNAs, and antisense activity. J Bacteriol 192:2359-2372.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J. 2013. Pfam: the protein families database. Nucleic Acids Research:gkt1223.

Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR. 2015. HMMER web server: 2015 update. Nucleic Acids Research:gkv397.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A. 2015. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research 44:D279-D285.

Flaherty BL, Van Nieuwerburgh F, Head SR, Golden JW. 2011. Directional RNA deep sequencing sheds new light on the transcriptional response of *Anabaena* sp. strain PCC 7120 to combined-nitrogen deprivation. BMC Genomics 12:332.

Forano E, Bertin Y, Montel M-c, Chaucheyras-durand F. 2015. Use of hafnia alvei for reducing the carriage of escherichia coli producing shiga toxins (stec) in ruminants. In: Google Patents.

Gally DL, Stevens MP. 2017. Microbe Profile: Escherichia coli O157: H7–notorious relative of the microbiologist's workhorse. Microbiology 163:1-3.

Galperin MY. 2001. Conserved 'hypothetical'proteins: new hints and new puzzles. International Journal of Genomics 2:14-18.

Galperin MY, Koonin EV. 2004. 'Conserved hypothetical'proteins: prioritization of targets for experimental study. Nucleic Acids Research 32:5452-5463.

Gannon VP, Laing CR, Zhang Y. 2011. Insights from Genomic Studies of the Foodborne and Waterborne Pathogen Escherichia coli O157: H7. In. Genomes of Foodborne and Waterborne Pathogens: American Society of Microbiology. p. 1-21.

Gene Ontology Consortium. 2016. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Research 45:D331-D338.

Goldberg T, Hecht M, Hamp T, Karl T, Yachdav G, Ahmed N, Altermann U, Angerer P, Ansorge S, Balasz K, et al. 2014. LocTree3 prediction of localization. Nucleic Acids Res 42:W350-355.

Grassé PP. 1977. Evolution of living organisms: evidence for a new theory of transformation: Academic Press.

Gyles C. 2007. Shiga toxin-producing Escherichia coli: An overview 1. Journal of Animal Science 85:E45-E62.

Habchi J, Tompa P, Longhi S, Uversky VN. 2014. Introducing protein intrinsic disorder. Chem Rev 114:6561-6588.

Hatahet F, Boyd D, Beckwith J. 2014. Disulfide bond formation in prokaryotes: History, diversity and design. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1844:1402-1414.

Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, et al. 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 8:11-22.

Haycocks JR, Grainger DC. 2016. Unusually Situated Binding Sites for Bacterial Transcription Factors Can Have Hidden Functionality. PLoS One 11:e0157016.

Hecht A, Glasgow J, Jaschke PR, Bawazer LA, Munson MS, Cochran JR, Endy D, Salit M. 2017. Measurements of translation initiation from all 64 codons in E. coli. Nucleic Acids Research 45:3615-3626.

Heider J, Baron C, Böck A. 1992. Coding from a distance: dissection of the mRNA determinants required for the incorporation of selenocysteine into protein. The EMBO journal 11:3759-3766.

Hemm MR, Paul BJ, Miranda-Rios J, Zhang A, Soltanzad N, Storz G. 2010. Small stress response proteins in *Escherichia coli*: proteins missed by classical proteomic studies. J Bacteriol 192:46-58.

Henderson IR, Czeczulin J, Eslava C, Noriega F, Nataro JP. 1999. Characterization of *pic*, a secreted protease of *Shigella flexneri* and enteroaggregative *Escherichia coli*. Infect Immun 67:5587-5596.

Henderson IR, Navarro-Garcia F, Desvaux M, Fernandez RC, Ala'Aldeen D. 2004. Type V protein secretion pathway: the autotransporter story. Microbiology and Molecular Biology Reviews 68:692-744.

Herz S, Wungsintaweekul J, Schuhr CA, Hecht S, Lüttgen H, Sagner S, Fellermeier M, Eisenreich W, Zenk MH, Bacher A. 2000. Biosynthesis of terpenoids: YgbB protein converts 4-diphosphocytidyl-2C-methyl-D-erythritol 2-phosphate to 2C-methyl-D-erythritol 2, 4-cyclodiphosphate. Proceedings of the National Academy of Sciences 97:2486-2490.

Higgins CF, Hiles ID, Salmond GP, Gill DR, Downie JA, Evans IJ, Holland IB, Gray L, Buckel SD, Bell AW. 1986. A family of related ATP-binding subunits coupled to many distinct biological processes in bacteria. Nature 323:448.

Hintze JL, Nelson RD. 1998. Violin plots: a box plot-density trace synergism. The American Statistician 52:181-184.

Hör J, Gorski SA, Vogel J. 2018. Bacterial RNA Biology on a Genome Scale. Molecular Cell.

Hücker S. 2017. RIBOseq-based discovery of non-annotated genes in *Escherichia coli* O157:H7 Sakai and their functional characterization. [Technical University of Munich.

Hücker SM, Ardern Z, Goldberg T, Schafferhans A, Bernhofer M, Vestergaard G, Nelson CW, Schloter M, Rost B, Scherer S. 2017. Discovery of numerous novel small genes in the intergenic regions of the Escherichia coli O157: H7 Sakai genome. PLoS One 12:e0184119.

Hücker SM, Simon S, Scherer S, Neuhaus K. 2016. Transcriptional and translational regulation by RNA thermometers, riboswitches and the sRNA DsrA in Escherichia coli O157:H7 Sakai under combined cold and osmotic stress adaptation. FEMS Microbiol Lett.

Hücker SM, Vanderhaeghen S, Abellan-Schneyder I, Scherer S, Neuhaus K. 2018. The novel anaerobiosis-responsive overlapping gene ano is overlapping antisense to the annotated gene ECs2385 of Escherichia coli O157: H7 Sakai. Frontiers in microbiology 9:931.

Hücker SM, Vanderhaeghen S, Abellan-Schneyder I, Wecko R, Simon S, Scherer S, Neuhaus K. 2018. A novel short L-arginine responsive protein-coding gene (laoB) antiparallel overlapping to a CadC-like transcriptional regulator in Escherichia coli O157: H7 Sakai originated by overprinting. BMC Evolutionary Biology 18:21.

Hughes AL, Westover K, da Silva J, O'Connor DH, Watkins DI. 2001. Simultaneous positive and purifying selection on overlapping reading frames of the tat and vpr genes of simian immunodeficiency virus. J Virol 75:7966-7972.

Husnik F, McCutcheon JP. 2017. Functional horizontal gene transfer from bacteria to eukaryotes. Nature Reviews Microbiology.

Ijaq J, Chandrasekharan M, Poddar R, Bethi N, Sundararajan VS. 2015. Annotation and curation of uncharacterized proteins-challenges. Frontiers in Genetics 6:119.

Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. 2012. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc 7:1534-1550.

Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, Wills MR, Weissman JS. 2014. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Reports 8:1365-1379.

Jaschke PR, Lieberman EK, Rodriguez J, Sierra A, Endy D. 2012. A fully decompressed synthetic bacteriophage øX174 genome assembled and archived in yeast. Virology 434:278-284.

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. 2008. NCBI BLAST: a better web interface. Nucleic Acids Research 36:W5.

Johnson ZI, Chisholm SW. 2004. Properties of overlapping genes are conserved across microbial genomes. Genome Res 14:2268-2272.

Kadokura H, Katzen F, Beckwith J. 2003. Protein disulfide bond formation in prokaryotes. Annual Review of Biochemistry 72:111-135.

Keese PK, Gibbs A. 1992. Origins of genes: "big bang" or continuous creation? Proc Natl Acad Sci U S A 89:9489-9493.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends in Genetics 25:404-413.

Kim W, Silby MW, Purvine SO, Nicoll JS, Hixson KK, Monroe M, Nicora CD, Lipton MS, Levy SB. 2009. Proteomic detection of non-annotated protein-coding genes in *Pseudomonas fluorescens* Pf0-1. PLoS One 4:e8455.

Kintz E, Brainard J, Hooper L, Hunter P. 2017. Transmission pathways for sporadic Shiga-toxin producing E. coli infections: A systematic review and meta-analysis. International Journal of Hygiene and Environmental Health 220:57-67.

Klemke M, Kehlenbach RH, Huttner WB. 2001. Two overlapping reading frames in a single exon encode interacting proteins--a novel way of gene usage. The EMBO journal 20:3849-3860.

Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Mol Biol Evol 28:1057-1074.

Koshland Jr DE. 1995. The key–lock theory and the induced fit theory. Angewandte Chemie International Edition in English 33:2375-2378.

Krakauer DC. 2000. Stability and evolution of overlapping genes. Evolution 54:731-739.

Kuchibhatla DB, Sherman WA, Chung BY, Cook S, Schneider G, Eisenhaber B, Karlin DG. 2014. Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently “orphan” viral proteins. Journal of Virology 88:10-20.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Molecular Biology and Evolution 33:1870-1874.

Kurata T, Katayama A, Hiramatsu M, Kiguchi Y, Takeuchi M, Watanabe T, Ogasawara H, Ishihama A, Yamamoto K. 2013. Identification of the set of genes, including nonannotated *morA*, under the direct control of ModE in *Escherichia coli*. J Bacteriol 195:4496-4505.

Landstorfer R, Simon S, Schober S, Keim D, Scherer S, Neuhaus K. 2014. Comparison of strand-specific transcriptomes of enterohemorrhagic *Escherichia coli* O157:H7 EDL933 (EHEC) under eleven different environmental conditions including radish sprouts and cattle feces. BMC Genomics 15:353.

Landstorfer RB. 2014. Comparative transcriptomics and translatomics to identify novel overlapping genes, active hypothetical genes, and ncRNAs in *Escherichia coli* O157:H7 EDL933. [[München]: Technische Universität München.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357-359.

Latif H, Li HJ, Charusanti P, Palsson BØ, Aziz RK. 2014. A gapless, unambiguous genome sequence of the enterohemorrhagic *Escherichia coli* O157: H7 strain EDL933. Genome Announcements 2:e00821-00814.

Lèbre S, Gascuel O. 2017. The combinatorics of overlapping genes. Journal of Theoretical Biology 415:90-101.

Lee SJ, Gralla JD. 2001. Sigma38 (rpoS) RNA polymerase promoter engagement via− 10 region nucleotides. Journal of Biological Chemistry 276:30064-30071.

Leimena MM, Wels MW, Bongers RS, Smid EJ, Zoetendal EG, Kleerebezem M. 2012. Comparative analysis of Lactobacillus plantarum WCFS1 transcriptomes using DNA microarray and next generation sequencing technologies. Applied and Environmental Microbiology:AEM. 00470-00412.

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, et al. 2011. The European Nucleotide Archive. Nucleic Acids Res 39:D28-31.

Li G, Huang J, Yang J, He D, Wang C, Qi X, Taylor IA, Liu J, Peng Y-L. 2018. Structure based function-annotation of hypothetical protein MGG_01005 from Magnaporthe oryzae reveals it is the dynein light chain orthologue of dynlt1/3. Sci Rep 8:3952.

Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park YM, Buso N, Lopez R. 2015. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Research 43:W580-W584.

Lillo F, Krakauer DC. 2007. A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes. Biol Direct 2:22.

Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. 2002. The relationship of protein conservation and sequence length. BMC Evolutionary Biology 2:20.

Litsios A, Ortega ÁD, Wit EC, Heinemann M. 2018. Metabolic-flux dependent regulation of microbial physiology. Current Opinion in Microbiology 42:71-78.

Lomsadze A, Gemayel K, Tang S, Borodovsky M. 2017. Improved Prokaryotic Gene Prediction Yields Insights into Transcription and Translation Mechanisms on Whole Genome Scale. bioRxiv:193490.

Lybecker M, Bilusic I, Raghavan R. 2014. Pervasive transcription: detecting functional RNAs in bacteria. Transcription 5:e944039.

Ma J, Campbell A, Karlin S. 2002. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J Bacteriol 184:5733-5745.

Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR. 2016. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Research 45:D200-D203.

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI. 2014. CDD: NCBI's conserved domain database. Nucleic Acids Research:gku1221.

Masel J. 2006. Cryptic genetic variation is enriched for potential adaptations. Genetics 172:1985-1991.

McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Phil. Trans. R. Soc. B 370:20140332.

McVeigh A, Fasano A, Scott DA, Jelacic S, Moseley SL, Robertson DC, Savarino SJ. 2000. IS1414, an *Escherichia coli* insertion sequence with a heat-stable enterotoxin gene embedded in a transposase-like gene. Infect Immun 68:5710-5715.

Milde S, Hemmrich G, Anton-Erxleben F, Khalturin K, Wittlieb J, Bosch TC. 2009. Characterization of taxonomically restricted genes in a phylum-restricted cell type. Genome Biology 10:R8.

Miller WG, Leveau JH, Lindow SE. 2000. Improved *gfp* and *inaZ* broad-host-range promoter-probe vectors. Mol Plant Microbe Interact. 13:1243-1250.

Mir K, Neuhaus K, Scherer S, Bossert M, Schober S. 2012. Predicting statistical properties of open reading frames in bacterial genomes. PLoS One 7:e45103.

Mizokami M, Orito E, Ohba K, Ikeo K, Lau JY, Gojobori T. 1997. Constrained evolution with respect to gene overlap of hepatitis B virus. J Mol Evol 44 Suppl 1:S83-90.

Monsellier E, Chiti F. 2007. Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO reports 8:737-742.

Moyers BA, Zhang J. 2016. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution. Mol Biol Evol 33:1245-1256.

Moyers BA, Zhang J. 2014. Phylostratigraphic bias creates spurious patterns of genome evolution. Molecular Biology and Evolution 32:258-267.

Moyers BA, Zhang J. 2018. Toward reducing phylostratigraphic errors and biases. Genome Biol Evol 10:2037-2048.

Mukherjee A, LeClerc JE, Cebula TA. 2008. Phenotypic Microarray Approaches to the Study of Prokaryotes.

Frequently Asked Questions, blast handbook [Internet]. 2018. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ#expect

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3:418-426.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics: Oxford University Press.

Nekrutenko A, Wadhawan S, Goetting-Minesky P, Makova KD. 2005. Oscillating evolution of a mammalian locus with overlapping reading frames: an XLalphas/ALEX relay. PLoS Genet 1:e18.

Nelson CW, Moncla LH, Hughes AL. 2015. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. Bioinformatics 31:3709-3711.

Neme R. 2014. Evolutionary analyses of orphan genes in mouse lineages in the context of de novo gene birth. [Christian-Albrechts-Universität Kiel.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution. BMC Genomics 14:117.

Nesvizhskii AI. 2014. Proteogenomics: concepts, applications and computational strategies. Nature Methods 11:1114.

Neuhaus K, Landstorfer R, Fellner L, Simon S, Marx H, Ozoline O, Schafferhans A, Goldberg T, Rost B, Küster B, et al. 2016. Translatomics combined with transcriptomics and proteomics reveals novel functional, recently evolved orphan genes in *Escherichia coli* O157:H7 (EHEC). BMC Genomics 17:133.

Neuhaus K, Landstorfer R, Simon S, Schober S, Wright PR, Smith C, Backofen R, Wecko R, Keim DA, Scherer S. 2017. Differentiation of ncRNAs from small mRNAs in *Escherichia coli* O157:H7 EDL933 (EHEC) by combined RNAseq and RIBOseq – *ryhB* encodes the regulatory RNA RyhB and a peptide, RyhP. BMC Genomics.

Nimrod G, Schushan M, Steinberg DM, Ben-Tal N. 2008. Detection of functionally important regions in "hypothetical proteins" of known structure. Structure 16:1755-1763.

Nogueira T, Touchon M, Rocha EP. 2012. Rapid evolution of the sequences and gene repertoires of secreted proteins in bacteria. PLoS One 7:e49403.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D. 2015. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research 44:D733-D745.

Omasits U, Ahrens CH, Müller S, Wollscheid B. 2013. Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinformatics 30:884-886.

Opuu V, Silvert M, Simonson T. 2017. Computational design of fully overlapping coding schemes for protein pairs and triplets. Sci Rep 7:15873.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nature Reviews Genetics 7:337-348.

Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of Drosophila orphan genes. Elife (Cambridge) 3:e01311.

Pandey A, Mann M. 2000. Proteomics to study genes and genomes. Nature 405:837-846.

Patel DH, Wi SG, Bae HJ. 2009. Modification of overlap extension PCR: A mutagenic approach.

Patthy L. 1999. Genome evolution and the evolution of exon-shuffling - a review. Gene 238:103-114.

Peer A, Margalit H. 2014. Evolutionary patterns of Escherichia coli small RNAs and their regulatory interactions. RNA.

Peng Z, Yan J, Fan X, Mizianty MJ, Xue B, Wang K, Hu G, Uversky VN, Kurgan L. 2015. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cellular and Molecular Life Sciences 72:137-151.

Pennington H. 2010. Escherichia coli O157. The Lancet 376:1428-1435.

Perdigão N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, Signal B, Gloss BS, Hammang CJ, Rost B. 2015. Unexpected features of the dark proteome. Proceedings of the National Academy of Sciences USA 112:15898-15903.

Perna NT, Plunkett G, 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, et al. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409:529-533.

Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785-786.

Pfaffl MW. 2001. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29:e45.

Poolman B, Spitzer JJ, Wood JM. 2004. Bacterial osmosensing: roles of membrane structure and electrostatics in lipid–protein and protein–protein interactions. Biochimica et Biophysica Acta (BBA)-Biomembranes 1666:88-104.

Prava J, Pranavathiyani G, Pan A. 2018. Functional assignment for essential hypothetical proteins of Staphylococcus aureus N315. Int J Biol Macromol 108:765-774.

Prelich G. 2012. Gene overexpression: uses, mechanisms, and interpretation. Genetics 190:841-854.

Price MN, Deutschbauer AM, Kuehl JV, Liu H, Witkowska HE, Arkin AP. 2011. Evidence-based annotation of transcripts and proteins in the sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. J Bacteriol 193:5716-5727.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Research 40:D130-D135.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41:D590-D596.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A. 2013. A large-scale evaluation of computational protein function prediction. Nature Methods 10:221.

Raghavan R, Sloan DB, Ochman H. 2012. Antisense transcription is pervasive but rarely conserved in enteric bacteria. MBio 3.

Reedy MC, Bullard B, Vigoreaux JO. 2000. Flightin is essential for thick filament assembly and sarcomere stability in Drosophila flight muscles. J Cell Biol 151:1483-1500.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E. 2000. Genome-wide location and function of DNA binding proteins. Science 290:2306-2309.

Richardson EJ, Watson M. 2013. The automatic annotation of bacterial genomes. Briefings in bioinformatics 14:1-12.

Riley LW, Remis RS, Helgerson SD, McGee HB, Wells JG, Davis BR, Hebert RJ, Olcott ES, Johnson LM, Hargrett NT. 1983. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. New England Journal of Medicine 308:681-685.

Robert-Koch-Institut. 2017. Infektionsepidemiologisches Jahrbuch Meldepflichtiger Krankheiten für das Jahr 2016. Berlin: Robert Koch-Institut.

Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV. 2002. Purifying and directional selection in overlapping prokaryotic genes. Trends in Genetics 18:228-232.

Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, Costanzo LD, Duarte JM, Dutta S, Feng Z, et al. 2017. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Research 45:D271-D281.

Rost B, Fariselli P, Casadio R. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy–Topology prediction at 86% accuracy. Protein Science 5:1704-1718.

Rost B, Liu J. 2003. The predictprotein server. Nucleic Acids Research 31:3300-3304.

Rost B, Sander C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins: Structure, Function, and Bioinformatics 19:55-72.

Rost B, Yachdav G, Liu J. 2004. The predictprotein server. Nucleic Acids Research 32:W321-W326.

Sabath N, Wagner A, Karlin D. 2012. Evolution of viral proteins originated de novo by overprinting. Mol Biol Evol 29:3767-3780.

Sabouri S, Sepehrizadeh Z, Amirpour-Rostami S, Skurnik M. 2017. A minireview on the in vitro and in vivo experiments with anti-Escherichia coli O157: H7 phages as potential biocontrol and phage therapy agents. International Journal of Food Microbiology 243:52-57.

Saha D, Panda A, Podder S, Ghosh TC. 2015. Overlapping genes: a new strategy of thermophilic stress tolerance in prokaryotes. Extremophiles 19:345-353.

Sarker MR, Cornelis GR. 1997. An improved version of suicide vector pKNG101 for gene replacement in gram-negative bacteria. Mol Microbiol 23:410-411.

Satoshi F, Nishikawa K. 2004. Estimation of the number of authentic orphan genes in bacterial genomes. DNA Research 11:219-231.

Schägger H. 2006. Tricine–sds-page. Nat Protoc 1:16.

Schlessinger A, Yachdav G, Rost B. 2006. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 22:891-893.

Schlötterer C. 2015. Genes from scratch–the evolutionary fate of de novo genes. Trends in Genetics 31:215-219.

Schmitz JF, Ullrich KK, Bornberg-Bauer E. 2018. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. Nature ecology & evolution:1.

Sévin DC, Sauer U. 2014. Ubiquinone accumulation improves osmotic-stress tolerance in Escherichia coli. Nature Chemical Biology 10:266.

Sherr CJ. 2006. Divorcing ARF and p53: an unsettled case. Nature Reviews Cancer 6:663.

Shultzaberger RK, Chen Z, Lewis KA, Schneider TD. 2006. Anatomy of Escherichia coli σ 70 promoters. Nucleic Acids Research 35:771-788.

Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Briefings in bioinformatics 3:265-274.

Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. 2013. New and continuing developments at PROSITE. Nucleic Acids Res 41:D344-347.

Simon S, Oelke D, Landstorfer R, Neuhaus K, Keim D. 2011. Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes. IEEE Symp Biol Data Vis 1:47 - 54.

Sleator RD, Walsh P. 2010. An overview of in silico protein function prediction. Archives of Microbiology 192:151-155.

Smith DR, Moxley RA, Peterson RE, Klopfenstein TJ, Erickson GE, Bretschneider G, Berberov EM, Clowser S. 2009. A two-dose regimen of a vaccine against type III secreted proteins reduced Escherichia coli O157: H7 colonization of the terminal rectum in beef cattle in commercial feedlots. Foodborne pathogens and disease 6:155-161.

Smith TF, Waterman MS. 1980. Protein constraints induced by multiframe encoding. Mathematical Biosciences 49:17-26.

Snedeker KG, Shaw DJ, Locking ME, Prescott RJ. 2009. Primary and secondary cases in Escherichia coli O157 outbreaks: a statistical analysis. BMC Infectious Diseases 9:144.

Solovyev V, Salamov A. 2011. Automatic annotation of microbial genomes and metagenomic sequences. Metagenomics and its applications in agriculture, biomedicine and environmental studies:61-78.

Solovyev VV, Tatarinova TV. 2011. Towards the integration of genomics, epidemiological and clinical data. Genome Med 3:48.

Storz G, Wolf YI, Ramamurthi KS. 2014. Small proteins can no longer be ignored. Annu Rev Biochem 83:753-777.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 30:2725-2729.

Tarr PI, Gordon CA, Chandler WL. 2005. Shiga-toxin-producing Escherichia coli and haemolytic uraemic syndrome. The Lancet 365:1073-1086.

Tautz D. 2014. The discovery of de novo gene evolution. Perspectives in Biology and Medicine 57:149-161.

Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. Nat Rev Genet 12:692-702.

Thomason MK, Storz G. 2010. Bacterial antisense RNAs: how many are there, and what are they doing? Annu Rev Genet 44:167-188.

Tian M, Lian Z, Bao Y, Bao S, Yin Y, Li P, Ding C, Wang S, Li T, Qi J. 2018. Identification of a novel, small, conserved hypothetical protein involved in Brucella abortus virulence by modifying the expression of multiple genes. Transboundary and emerging diseases.

Typas A, Barembruch C, Possling A, Hengge R. 2007. Stationary phase reorganisation of the Escherichia coli transcription machinery by Crl protein, a fine-tuner of σs activity and levels. The EMBO journal 26:1569-1578.

Uchida H, Kiyokawa N, Horie H, Fujimoto J, Takeda T. 1999. The detection of Shiga toxins in the kidney of a patient with hemolytic uremic syndrome. Pediatr Res 45:133.

Uversky VN. 2011. Intrinsically disordered proteins from A to Z. Int J Biochem Cell Biol 43:1090-1103.

Wade JT, Grainger DC. 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol 12:647-653.

Wang LF, Park SS, Doi RH. 1999. A novel *Bacillus subtilis* gene, *antE*, temporally regulated and convergent to and overlapping *dnaE*. J Bacteriol 181:353-356.

Watson JD, Laskowski RA, Thornton JM. 2005. Predicting protein function from sequence and structural data. Current Opinion in Structural Biology 15:275-284.

Weber H, Polen T, Heuveling J, Wendisch VF, Hengge R. 2005. Genome-wide analysis of the general stress response network in Escherichia coli: σS-dependent genes, promoters, and sigma factor selectivity. Journal of Bacteriology 187:1591-1603.

Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, Martin WF. 2016. The physiology and habitat of the last universal common ancestor. Nature Microbiology 1:16116.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S. 2007. Database resources of the national center for biotechnology information. Nucleic Acids Research 35:D5-D12.

Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nature ecology & evolution 1:0146.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proceedings of the National Academy of Sciences 106:7273-7280.

Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, Honigschmid P, Schafferhans A, Roos M, Bernhofer M, et al. 2014. PredictProtein--an open resource for online prediction of protein structural and functional features. Nucleic Acids Res 42:W337-343.

Yamada Y, Kuzuyama T, Komatsu M, Shin-ya K, Omura S, Cane DE, Ikeda H. 2015. Terpene synthases are widely distributed in bacteria. Proceedings of the National Academy of Sciences 112:857-862.

Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology 12:635-645.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, et al. 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5:e16.

Zeghouf M, Li J, Butland G, Borkowska A, Canadien V, Richards D, Beattie B, Emili A, Greenblatt JF. 2004. Sequential Peptide Affinity (SPA) system for the identification of mammalian and bacterial protein complexes. J Proteome Res 3:463-468.

Zehentner B. 2015. Expression und Funktion von überlappenden ORFs in EHEC. [Masterarbeit]: Technische Universität München.

6 Acknowledgements

I thank Prof. Dr. Scherer that I could perform my doctorate thesis at his Chair. I thank for his excellent supervision, for encouraging words, and for giving me the opportunity to conduct my research freely within the frame of my project.

I thank PD Dr. Klaus Neuhaus for helping me with certain problems I had.

I thank Barbara Zehentner a lot, being able to characterize one of ‘here’ genes in detail, but also for her support by checking her Cappable Seq Data and implementing the Western Blot at the end of my thesis.

I thank Dr. Zachary Ardern, Dr. Svenja Simon and Dr. Chase Nelson for conducting some of the bioinformatics analyses for me.

I thank Dr. Andrea Schafferhans, Dr. Tanya Goldberg and Michael Bernhofer for helping in obtaining the PredictProtein data and instructions on interpretations.

I thank Romy Wecko for the excellent technical support in the lab.

I thank all colleagues in the Chair, as well as Isabel Abellan-Schneyder and Franziska Giehren of the ZIEL - Core Facility Microbiome/NGS for a great atmosphere.

I particularly thank all past and current members of the AG Neuhaus (today AG OLG) for accompanying me through my PhD, for their inspiring technical discussions and support.

Ich danke meinen Freunden, meiner Familie und insbesondere meinen Eltern, die mich immer unterstützt haben.

Curriculum vitae

2014-today: Doctorate at the Chair of Microbial Ecology, Technical University of Munich

2012-2014: Master in “Biochemie und Molekulare Biologie”, University of Bayreuth

Master thesis at the Laboratory of Food Biotechnology, IFNH, ETH Zürich

Topic: Analysis of functional properties of bifidobacteria and lactobacilli strains from feces and breast milk of mother - neonate pairs

2009-1012: Bachelor „Biochemie“, University of Bayreuth

Bachelor thesis at the Chair of Microbiology, University of Bayreuth

Eidesstaatliche Erklärung

Ich erkläre an Eides statt, dass ich die bei der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der TUM zur Promotionsprüfung vorgelegte Arbeit mit dem Titel

Overlapping genes in *E. coli* EDL933 (EHEC) - Phylostratigraphy of alternative reading frames and functional analysis of the candidate gene *asa*

am Lehrstuhl für Mikrobielle Ökologie, ZIEL - Institute for Food & Health unter der Anleitung- und Betreuung durch Prof. Dr. Siegfried Scherer

ohne sonstige Hilfe erstellt und bei der Abfassung nur die gemäß §6 Abs. 6 und 7 Satz 2 angegebenen Hilfsmittel benutzt habe.

Ich habe keine Organisation eingeschaltet, die gegen Entgelt Betreuerinnen und Betreuer für die Anfertigung von Dissertationen sucht, oder die mir obliegenden Pflichten hinsichtlich der Prüfungsleistungen für mich ganz oder teilweise erledigt.

Ich habe die Dissertation in dieser oder ähnlicher Form in keinem anderen Prüfungsverfahren als Prüfungsleistung vorgelegt.

Ich habe den angestrebten Doktorgrad noch nicht erworben und bin nicht in einem früheren Promotionsverfahren für den angestrebten Doktorgrad endgültig gescheitert.

Die öffentlich zugängliche Promotionsordnung der TUM ist mir bekannt, insbesondere habe ich die Bedeutung von § 28 (Nichtigkeit der Promotion) und § 29 (Entzug des Doktorgrades) zur Kenntnis genommen. Ich bin mir der Konsequenzen einer falschen Eidesstattlichen Erklärung bewusst.

Mit der Aufnahme meiner personenbezogenen Daten in die Alumni-Datei bei der TUM bin ich einverstanden.

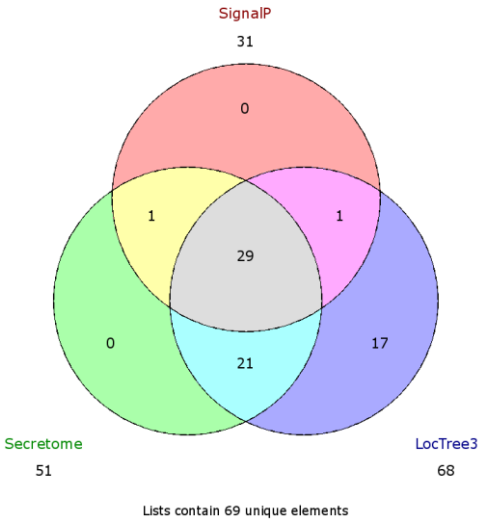
6 Supplementary material

6.1 Index

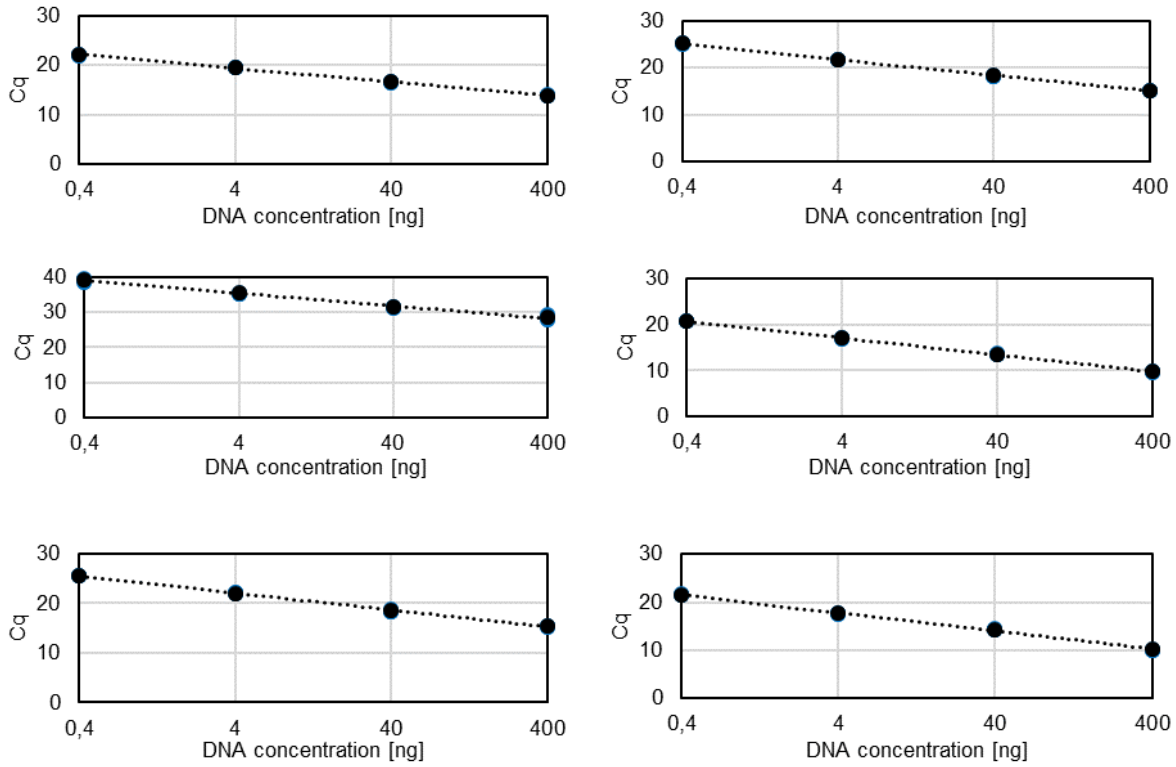
6.1.1 Figures

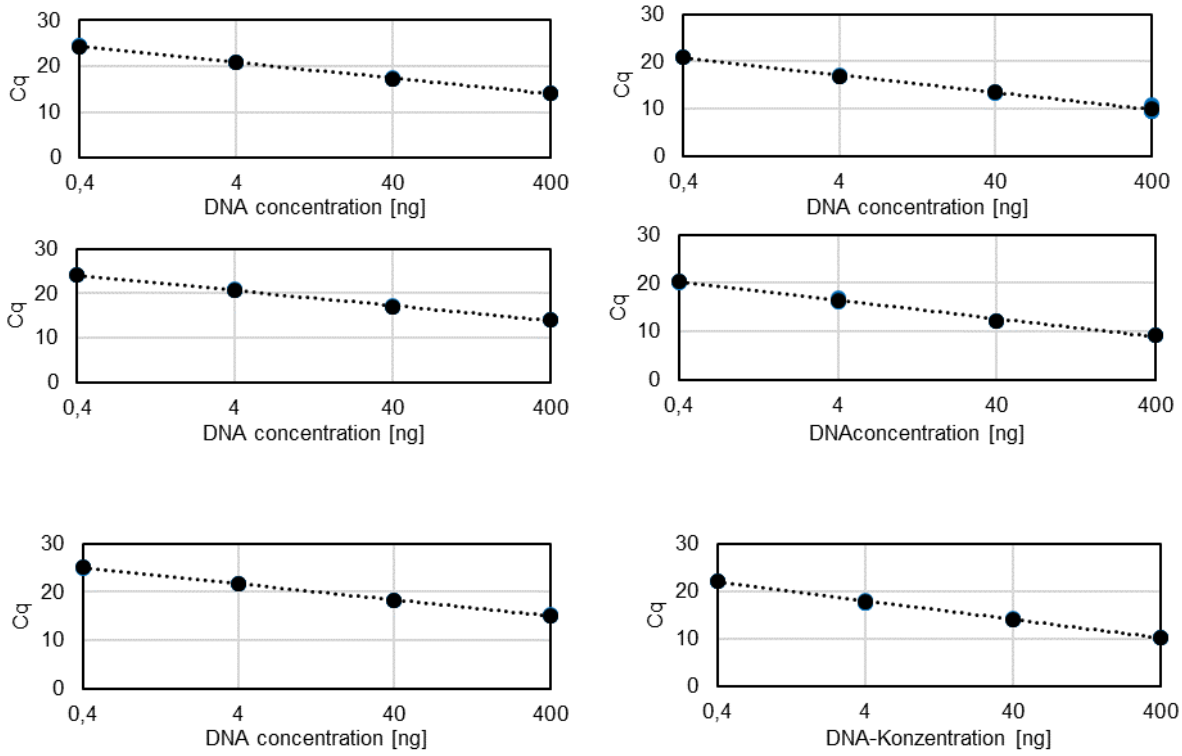
6.1.2 Tables

6.2 Supplementary Figures

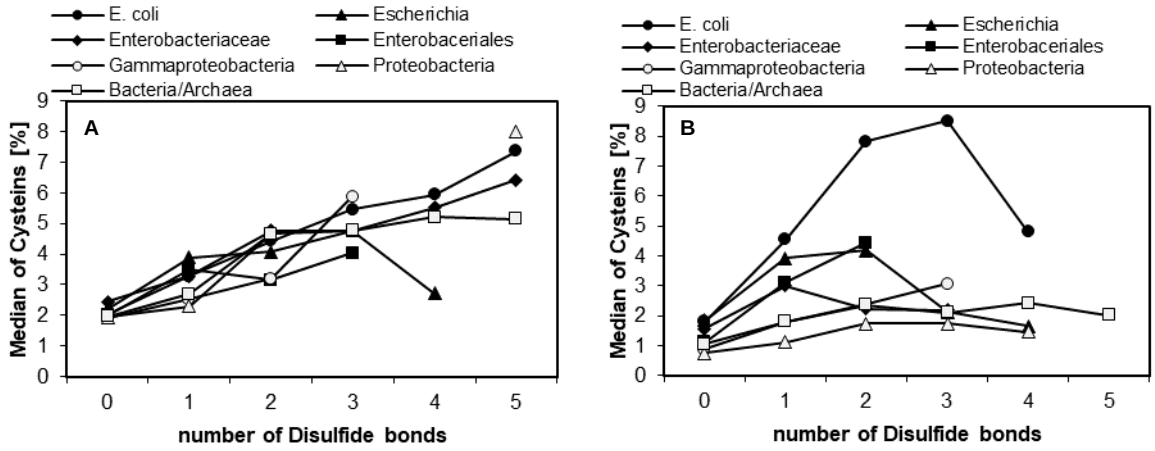


Supplementary Figure S1: Comparison of three programs used for the prediction of protein secretion from amino acid sequences (Secretome2.0, SignalP4.1 and LocTree3). A dataset of 77 annotated, bacterial proteins, which are experimentally verified to be secreted, was used to evaluate prediction precision of each program.

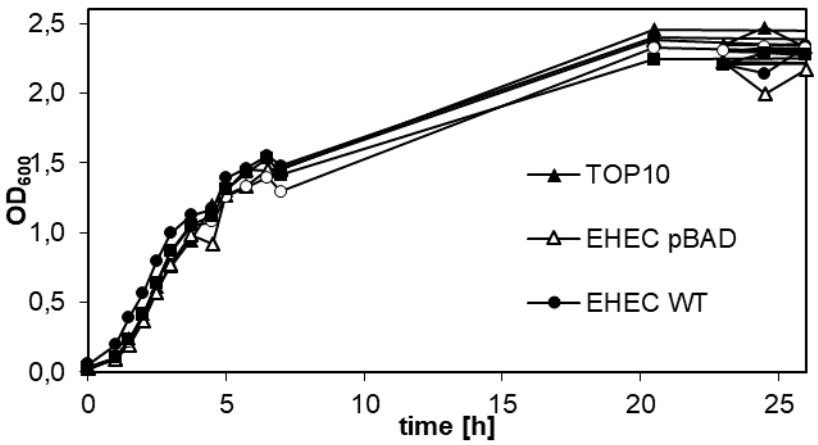




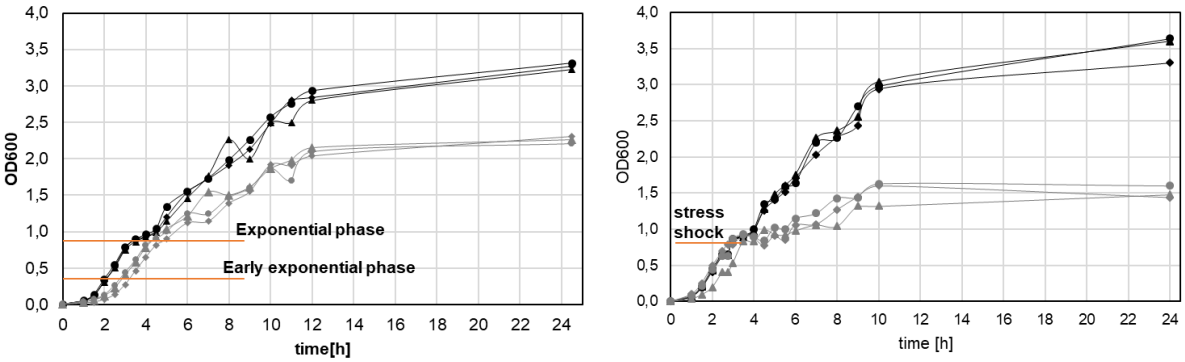
Supplementary figure S2: Standard curves of qPCR primers of *asa*, homologues and a negative control. Legend, left from top to bottom: *asa*, *H. alvei, C. freundii, S. marcescens, S. enterica,* negative control; right: respective 16S rRNA samples. EHEC DNA was used for *asa* and the negative control. DNA of the organisms was used for the respective primers. Annealing temperatures: 61°C (*asa*), 60°C (*asa* homologues), 58°C (negative control). The negative control is an untranscribed region identified by RNAseq.



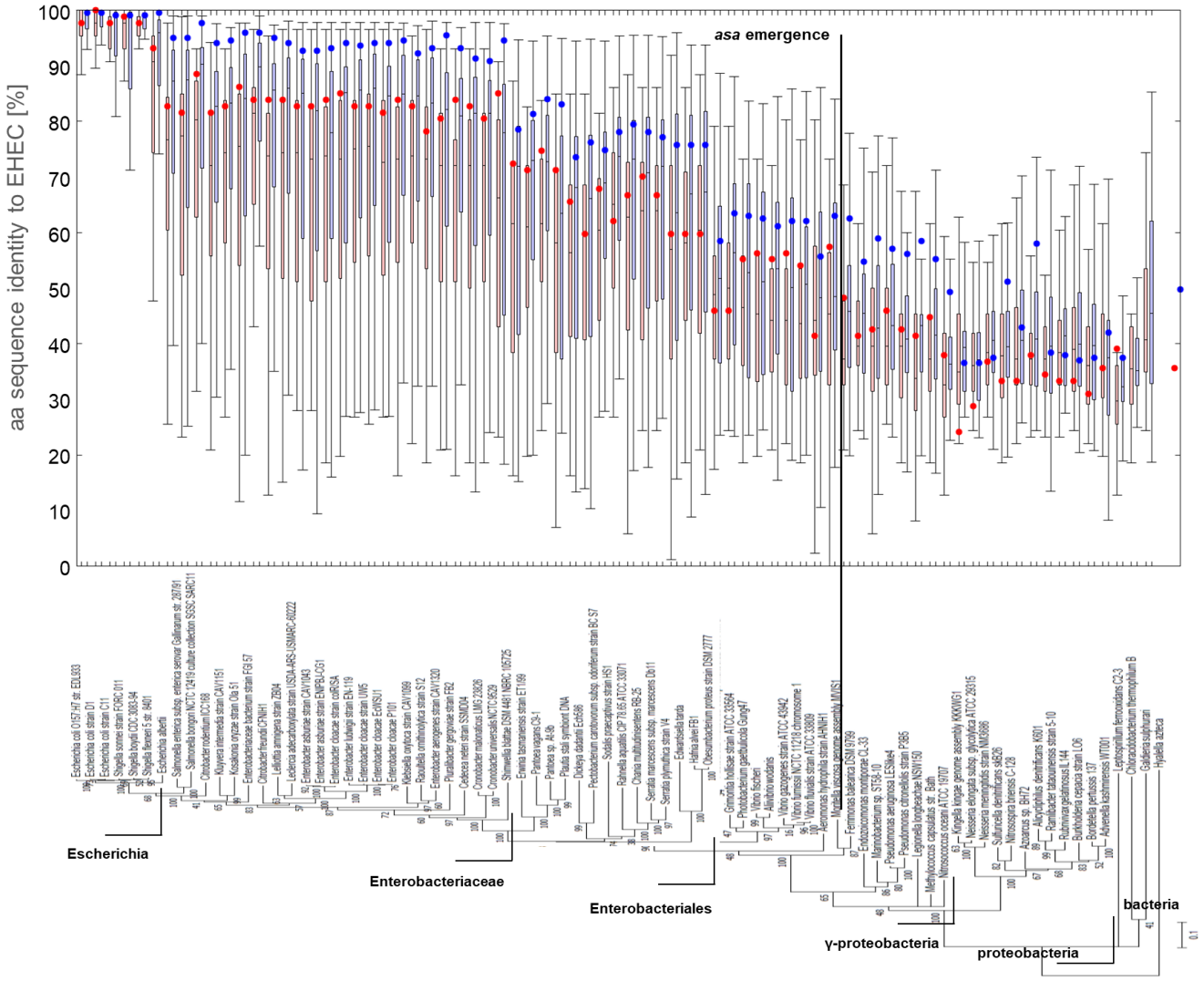
Supplementary figure S3: Average cysteines per ORF in dependence on the number of disulfide bonds per ORF of ORFs of different phylostratigraphic ages. (A) proteins encoded by shadow ORFs and (B) proteins encoded by annotated ORFs.



Supplementary figure S4: Non-competitive growth of EHEC pBAD-*myc*/His C-asa and EHEC pBAD-*myc*/HisC-asa^{tar22} in LB + 450 mM NaCl. *E. coli* TOP10, EHEC pBAD and EHEC without vector ("WT") were used as negative control. The gene expression was induced with 0.002% arabinose. The OD_{600} values shown here are the average of two biological replicates.



Supplementary figure S5: Growth curves for RT-qPCR of *asa*. (A) Stress-adapted conditions: EHEC grown in LB (black) or LB + 450 mM NaCl (grey). Overnight cultures were adapted to sodium chloride as follows: one colony was used for inoculation in liquid LB medium. EHEC was grown for 2 h before addition of 450 mM NaCl. The overnight cultures were used for inoculation in fresh medium and the OD_{600} was measured. Aliquots were taken at early exponential phase (OD_{600} = 0.2 - 0.3) and at exponential phase (OD_{600} = 0.7-0.8). (B) Stress shocked conditions: EHEC grown in LB (black) or shocked with sodium chloride at OD_{600} = 0.8 (grey). Aliquots were taken before stress shock and 30 min, 60 min and 120 min thereafter.

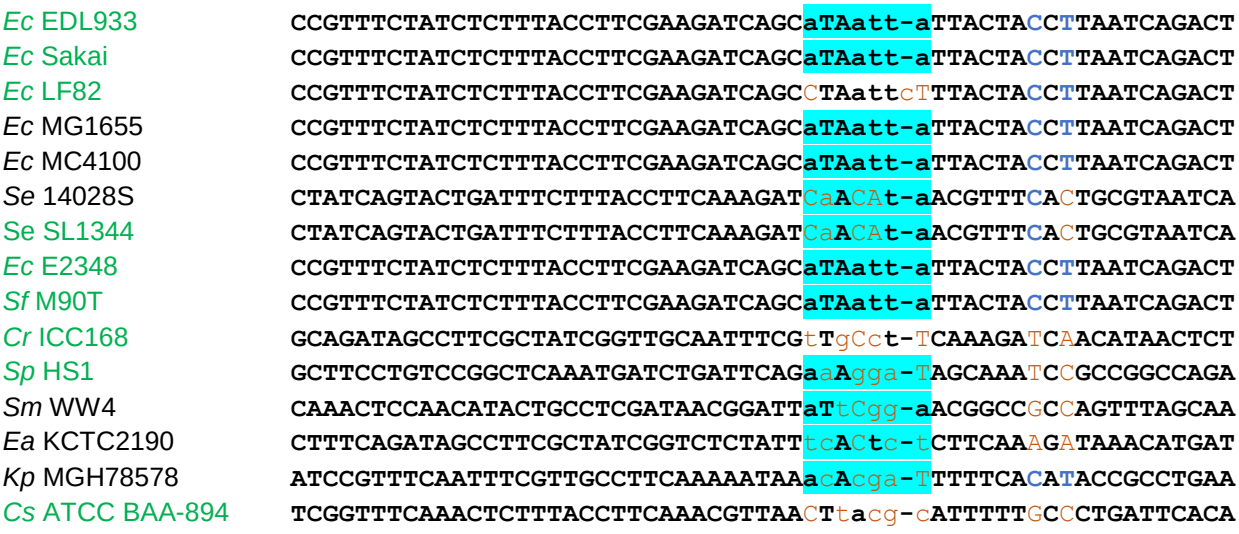


Supplementary figure S6: Amino acid (aa) sequence identities of the overlapping gene pair *asa* /EDL933_1238 and of a negative control. The negative control was calculated from 100 randomly selected genes and an overlapping sequence in the organisms in which an overlapping gene pair homologue was found. The overlapping sequence of the negative control is a random sequence of the length of the sORF, embedded in the selected random gene and in the reading frame in which the sORF is located.

Supplementary figure S7: Amino acid sequences of *asa* homologues for which only RNAseq data are available. Start and stop codon are highlighted in green and red. The tyrosine at the premature stop codon position is marked in blue. Amino acids which are different to those in EHEC EDL933 are highlighted in brown.

A

B



C

Ec EDL933	CCGTTTCTATCTCTTTACCTTCGAAGATCAGCATAATT-ATTACTACctTAATCAGACT
Ec Sakai	CCGTTTCTATCTCTTTACCTTCGAAGATCAGCATAATT-ATTACTACctTAATCAGACT
Ec LF82	CCGTTTCTATCTCTTTACCTTCGAAGATCAGCATAATTCTTTACTACctTAATCAGACT
Ec MG1655	CCGTTTCTATCTCTTTACCTTCGAAGATCAGCATAATT-ATTACTACctTAATCAGACT
Ec MC4100	CCGTTTCTATCTCTTTACCTTCGAAGATCAGCATAATT-ATTACTACctTAATCAGACT
Se 14028S	CTATCAGTACTGATTTCTTTACCTTCAAAGATCAACAT-AACGtTtCACTGCGTAATCA
Se SL1344	CTATCAGTACTGATTTCTTTACCTTCAAAGATCAACAT-AACGtTtCACTGCGTAATCA
Ec E2348	CCGTTTCTATCTCTTTACCTTCGAAGATCAGCATAATT-ATTACTACctTAATCAGACT
Sf M90T	CCGTTTCTATCTCTTTACCTTCGAAGATCAGCATAATT-ATTACTACctTAATCAGACT
Cr ICC168	GCAGATAGCCTTCGCTATCGGTTGCAATTTCGTTGCCT-TCAAagAtcaaCATAACTCT
Sp HS1	GCTTCCTGTCCGGCTCAAATGATCTGATTCAGAAAGGA-TAGCaaAtcCgCCGGCCAGA
Sm WW4	CAAACTCCAACATACTGCCTCGATAACGGATTATTCGG-AACGgccgcCaGTTTAGCAA
Ea KCTC2190	CTTTCAGATAGCCTTCGCTATCGGTCTCTATTTCACTG-CCTTCaAagcTAAACATGAT
Kp MGH78578	ATCCGTTTCAATTTCGTTGCCTTCAAAAATAAACATGA-TTTTtcACAtaCCGCCTGAA
Cs ATCC BAA-894	TCGGTTTCAAACTCTTTACCTTCAAACGTTAACTTACG-CATTtTtgcCcTGATTCACA

Supplementary Figure S8: Promoter region of *asa* and the homologues with RNAseq data. The sequences were implemented from *Escherichia coli* (Ec), *Salmonella enterica* (Se), *Shigella flexneri* (Sf), *Citrobacter rodentium* (Cr), *Sodalis praecaptivus* (Sp), *Serratia marcescens* (Sm), *Enterobacter aerogenes* (Ea), *Klebsiella pneumonia* (Kp) and *Cronobacter sakazakii* (Cs). (**A**) σ^{70} promoter of the TSS 188/186 bp upstream of the start codon, (**B**) σ^{38} promoter of the TSS 188/186 bp upstream of the start codon, (**C**) σ^{38} promoter 176 bp of the TSS upstream of the start codon. Organism with full-length *asa* are highlighted in green letters. σ^{70} promoters predicted by BProm are highlighted in yellow. σ^{38} promoters with an average conservation frequency higher than 0% are highlighted in blue. Mutated nucleotides are shown in brown letters, conserved nucleotides at the TSS position of EHEC are shown in blue letters. Small letters within the promoter region show mismatches to the consensus sequence (σ^{70}: ATAATTA, σ^{38}: CTACCTT).

```
Ec EDL933     MMLLVSNKIAPEEIKMKTALIARCRAGSCRKILATIPISTTTIPAISIPPRNDISLRVVSTYAEQQKNTNAVPPSAIAITSPIPADR*
Cf CFNIH1     MILLVSNRIAPEEIRINTALIARCSAGSCRKIFATIPISTTTIPAISIPPRKDISFRVVSTYAEQLKKTSAVPPSAMPITSPIPADR*
Sm WS1359     MMLLVSHRIAPAENRMNSALIARCRAGICRNRLAITATSSTTKPAIIMPDRNDMSLRVVSTYAEQQKKTAAVPPSASMIRSPAPAFRNELRIGPRAYPMKPVSANAARMPAGLLASLWVRNIRP*
Se 287/91     MMLLVSYRIAPEEIRINTALIARCNAGSCRKIFATIPISTTTIPAINIPPRKDISLRVVRT*AEQQKNTNAVPPSARPMTSPMPADR*
Ha DSM30097   IMLLVSHRMAPVENRINSALIASDSAGICKNRLAITATSKITKPAIIIPERNDISLRVVST*AEQQKNTAAVPPSASIIRSPAPLAIKELKIGPRVYPMNPVSANTAKIPAGRSAKR*
```

Supplementary Figure S9: Amino acid sequences of *asa* homologues which are experimentally characterized. Start and stop codon are highlighted in green and red. The tyrosine at the premature stop codon position is marked in blue. Amino acids which are different to those in EHEC EDL933 are highlighted in brown.

A

Ec EDL933 CCGTTtctaTCTCTTTACCTTCGAAGATCAGcATAATT-ATTACTACCTTAATCAGACT
Cf CFNIH1 AACTTtcAcCCGCCGTGTGGCTGATTCAGGAccggcgT-TATCCTACCTATATCTGTCG
Sm Db11 AAAcTccAaCATACTGCCTCGATAACGGATAatTcggA-ACGGCCGCCAGTTTAGCAAA
Se 287/91 CTATcagtaCTGATTTCTTTACCTTCAAAGATcaAcaT-AACGTTTCACTGCGTAATCA
Ha FB1 CGGcaGcgcGGCAAAATGAGCGGGAAGTACCgtgtAaA-CGTACCGTTCAGAAACAAGA

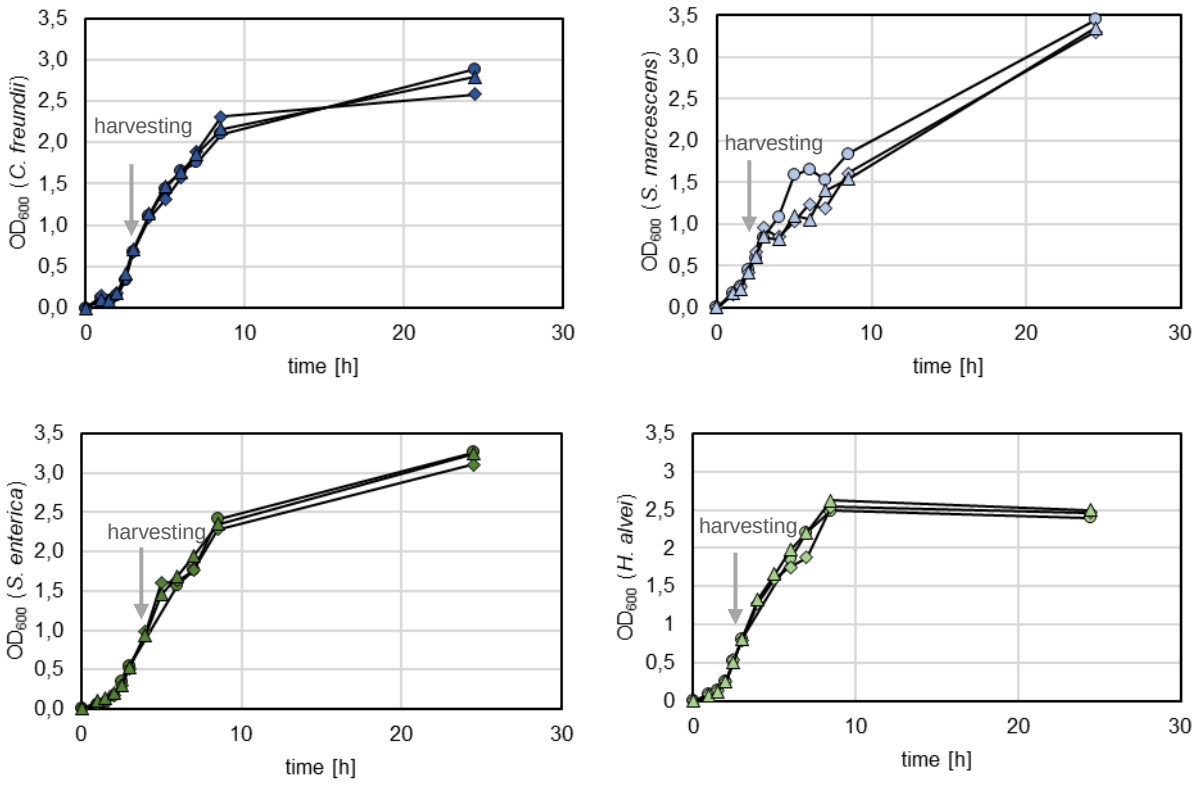
B

Ec EDL933 CCGTTTCTATCTCTTTACCTTCGAAGATCAGCaTAatt-aTTACTACCTTAATCAGACT
Cf CFNIH1 AACTTTCACCCGCCGTGTGGCTGATTCAGGACCggCgt-TATCCTACCTATATCTGTCG
Sm Db11 AAACTCCAACATACTGCCTCGATAACGGATAATtcgga-aCGGCCGCCAGTTTAGCAAA
Se 287/91 CTATCAGTACTGATTTCTTTACCTTCAAAGATCaACAt-aACGTTTCACTGCGTAATCA
Ha FB1 CGGCAGCGCGGCAAAATGAGCGGGAAGTACCGtgtaAa-cGTACCGTTCAGAAACAAGA

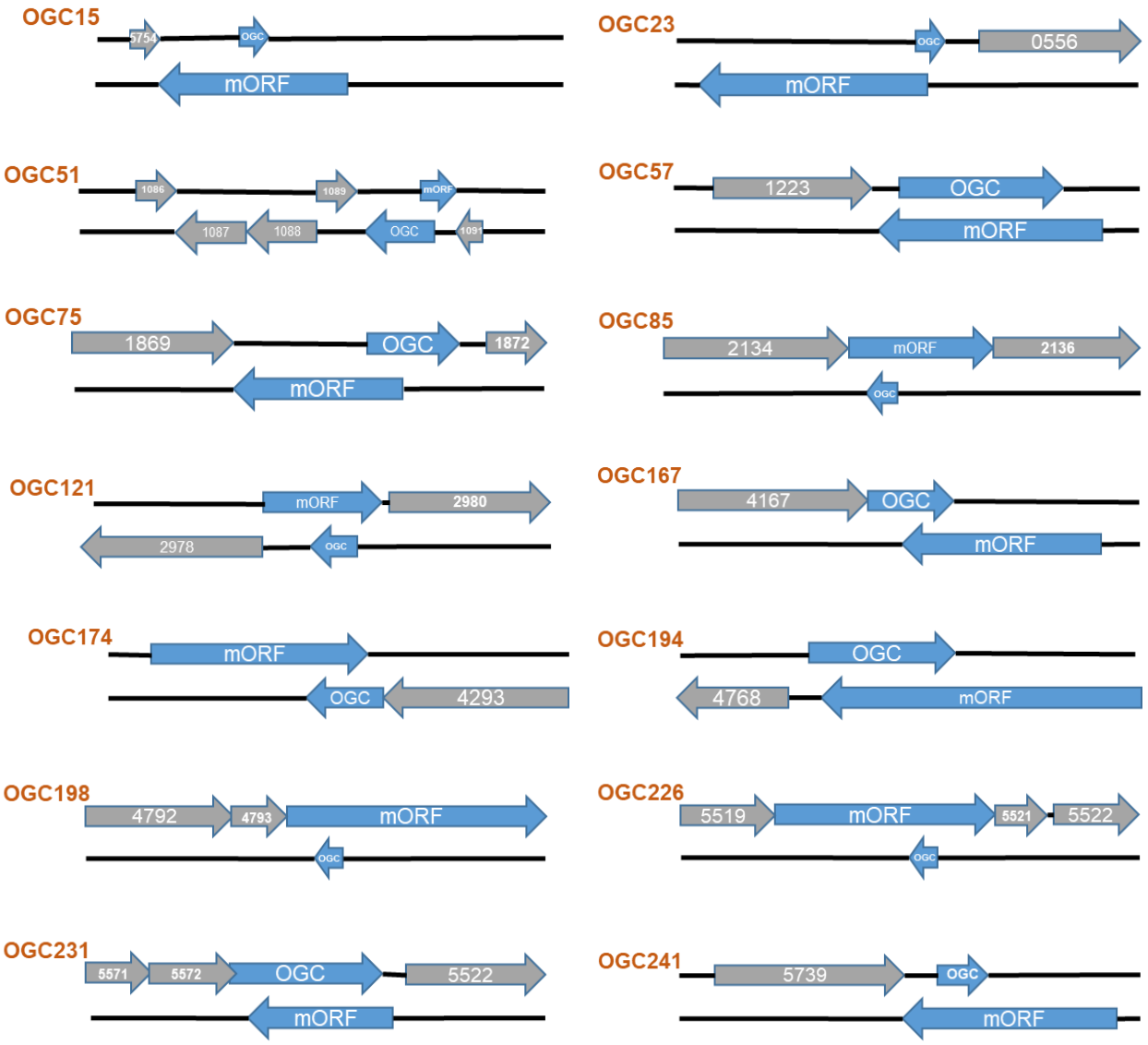
C

Ec EDL933 CCGTTTCTATCTCTTTACCTTCGAAGATCAGCATAATT-ATTACTACCtTAATCAGACT
Cf CFNIH1 AACTTTCACCCGCCGTGTGGCTGATTCAGGACCGGCGT-TATCCTACCtaTATCTGTCG
Sm Db11 AAACTCCAACATACTGCCTCGATAACGGATAATTCGGA-ACGGCcgCcagTTTAGCAAA
Se 287/91 CTATCAGTACTGATTTCTTTACCTTCAAAGATCAACAT-AACGtTtCACTGCGTAATCA
Ha FB1 CGGCAGCGCGGCAAAATGAGCGGGAAGTACCGTGTAAA-CGTACcgttCaGAAACAAGA

Supplementary Figure S10: Promoter region of *asa* and its experimentally analysed homologues. The sequences were implemented from EHEC (Ec EDL933), *Citrobacter freundii* (Cf), *Serratia marcescens* (Sm), *Salmonella enterica* (Se), *Hafnia alvei* (Ha). (**A**) σ^{70} promoter of the TSS 188/186 bp upstream of the start codon, (**B**) σ^{38} promoter of the TSS 188/186 bp upstream of the start codon, (**C**) σ^{38} promoter 176 bp of the TSS upstream of the start codon. σ^{70} promoters predicted by BProm are highlighted in yellow. σ^{38} promoters with an average conservation frequency higher than 0% are highlighted in blue. Mutated nucleotides are shown in brown letters, conserved nucleotides at the TSS position of EHEC are shown in blue letters. Small letters within the promoter region show mismatches to the consensus sequence (σ^{70}: ATAATTA, σ^{38}: CTACCTT).

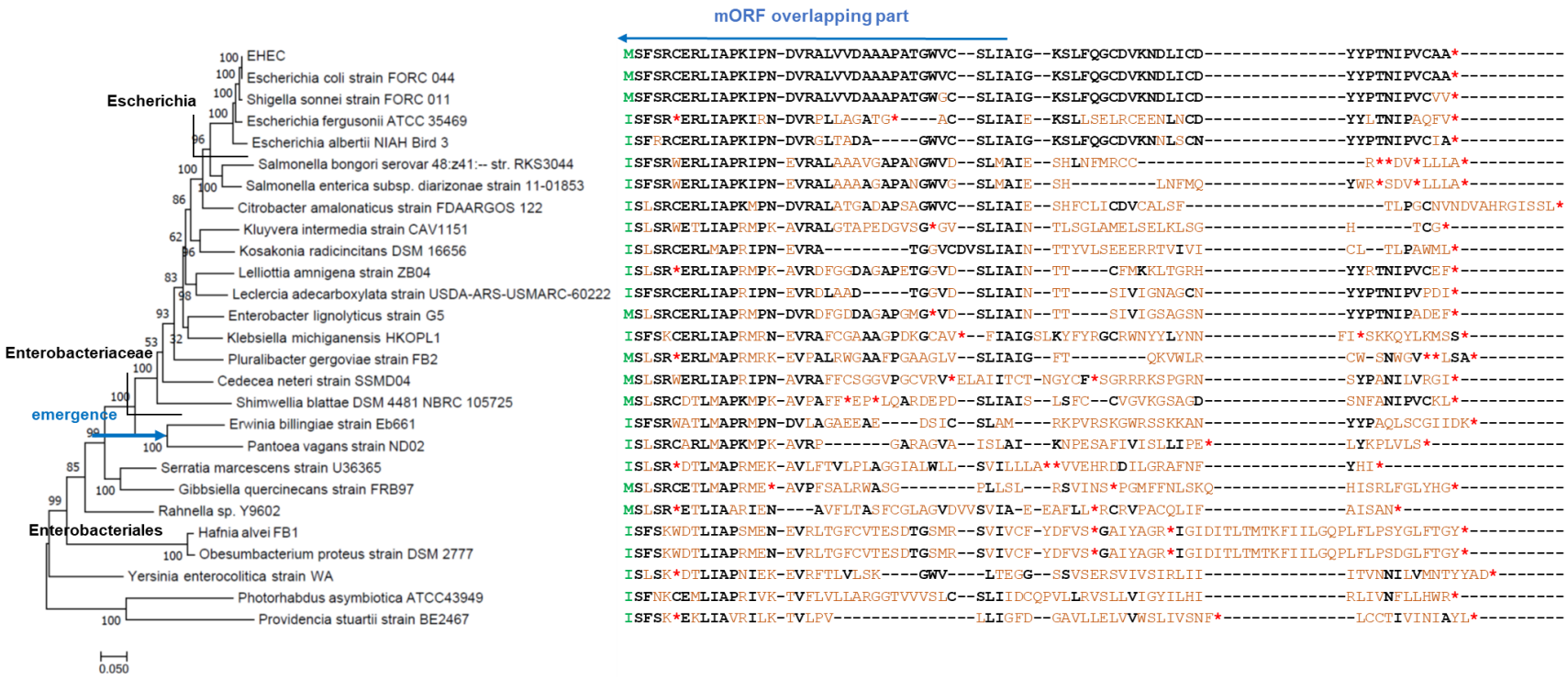


Supplementary Figure S11: Growth curves of Enterobacteria with *asa* homologue used for RT-qPCR. All organisms were grown in LB. Aliquots were taken at OD_{600} = 0.8-0.9 for total RNA extraction. (**A**) *Citrobacter freundii* CFNIH1, (**B**) *S. marcescens* WS1359, (**C**) *S. enterica* serovar Gallinarum 287/91, (**D**) *H. alvei* DSM30097.

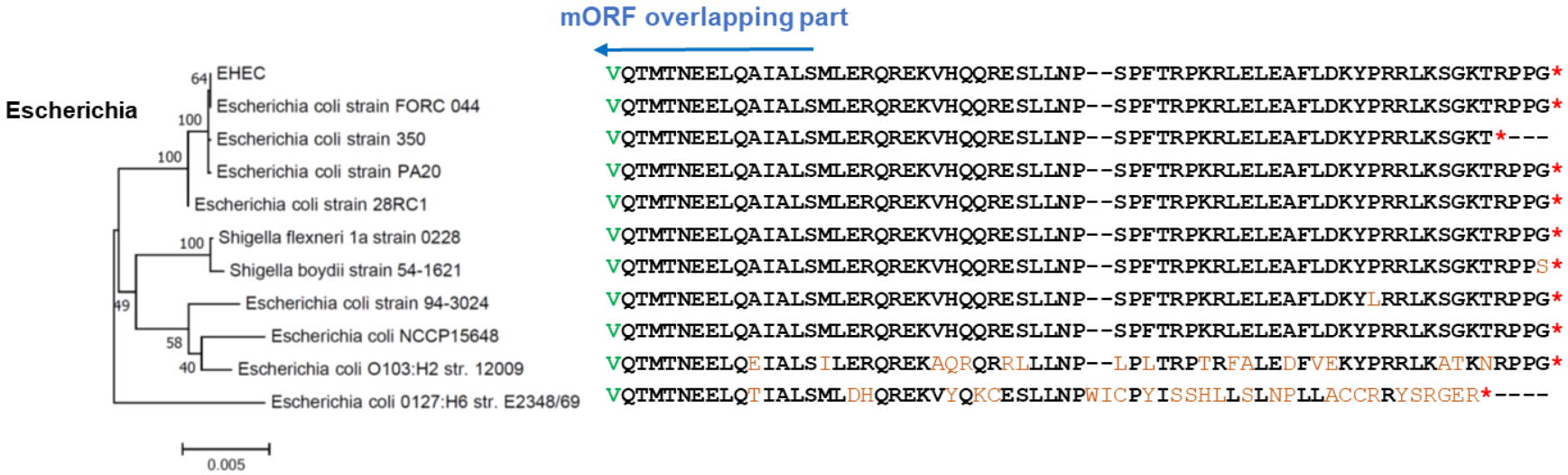


Supplementary figure S12: Genome organization of overlapping gene candidates (OGCs) in EHEC EDL933. OGC and its mother gene (mORF) are shown as blue arrows, genes in the neighborhood in grey arrows named after their location tag in the EHEC EDL933 genome.

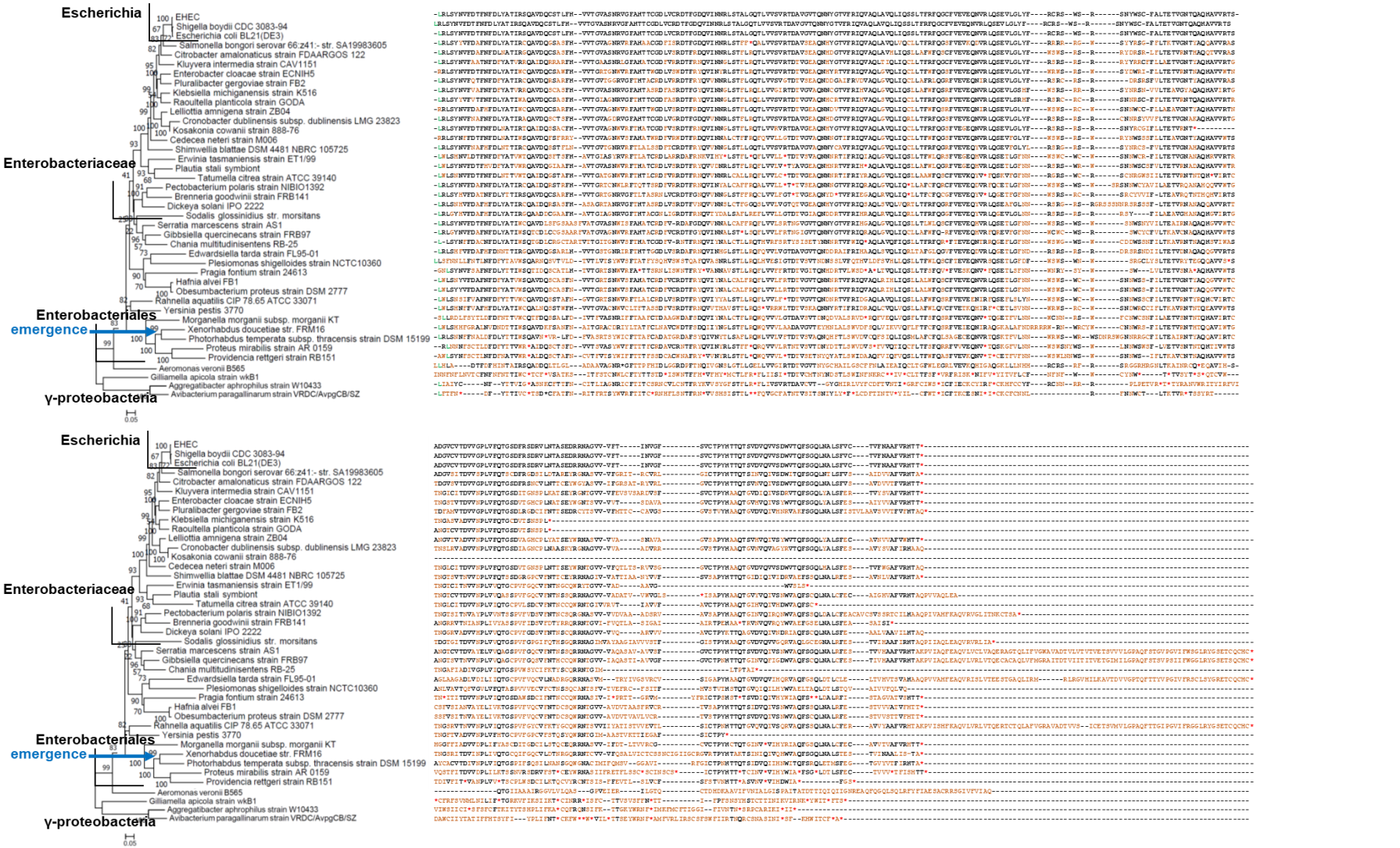
Supplementary figure S13.1: Phylostratigraphy of OGC15. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales*. It is a maximum likelihood tree, which was bootstrapped 1000 times. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).



Supplementary figure S13.2: Phylostratigraphy of OGC23. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales*. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).

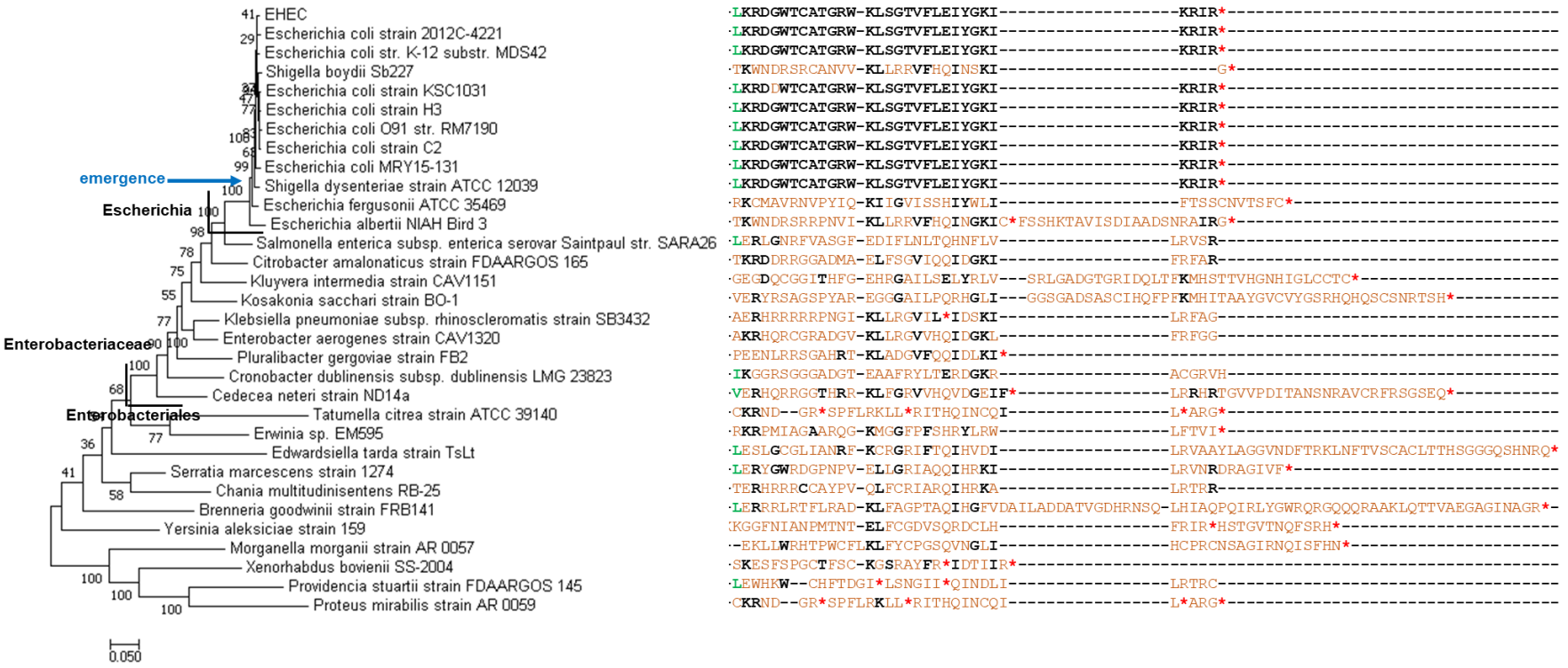


Supplementary figure S13.3: Phylostratigraphy of OGC51. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales*. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).



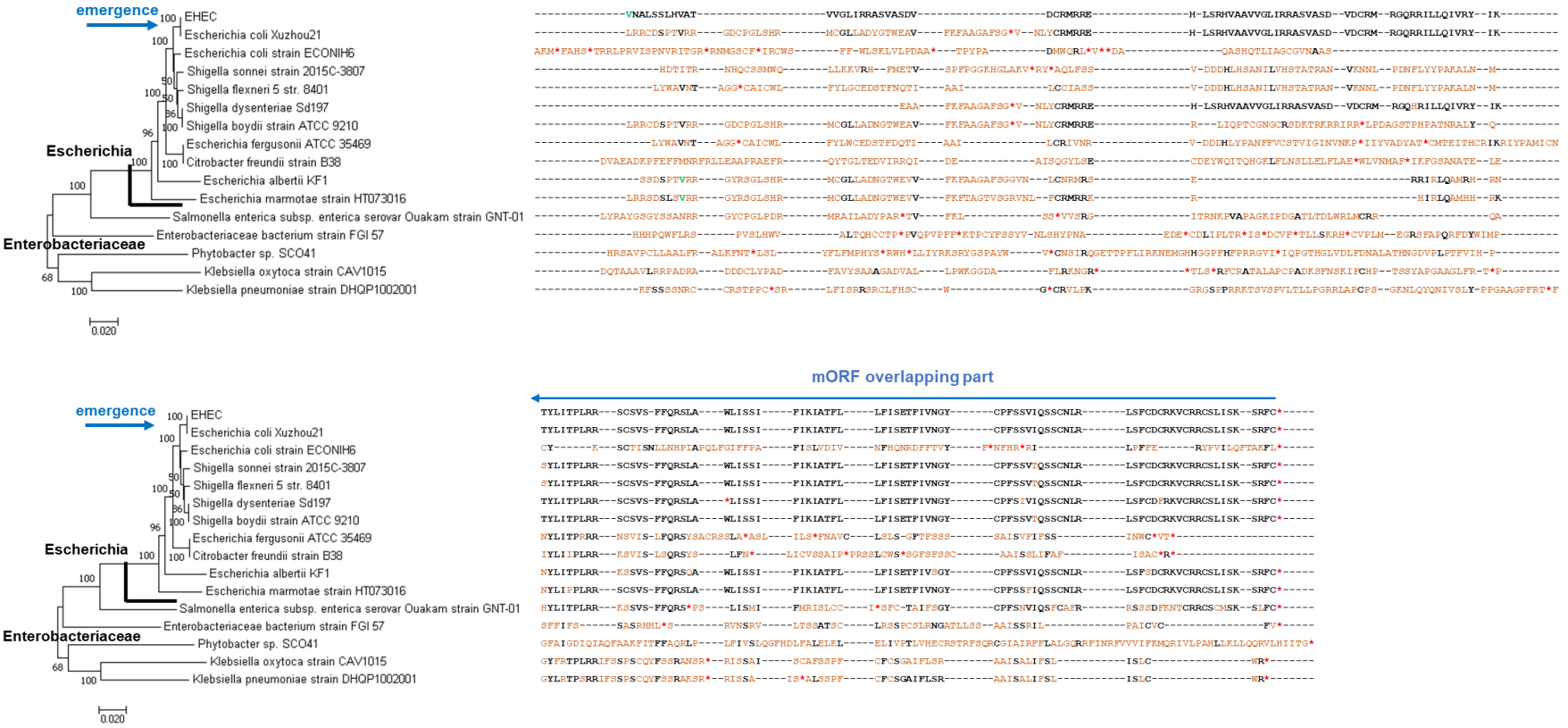
Supplementary figure S13.4: Phylostratigraphy of OGC57. Top: amino acid 1-170, bottom: amino acid 171-267. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales*. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).

Supplementary figure S13.5: Phylostratigraphy of OGC75. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales*. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).

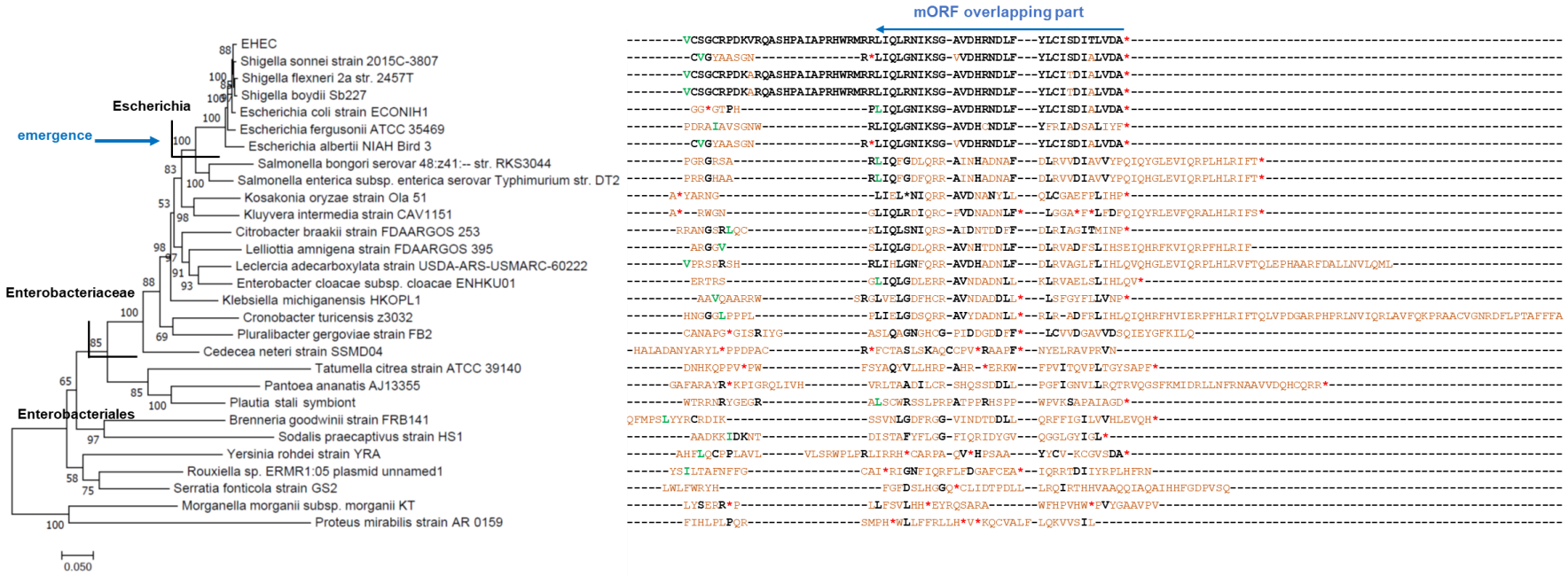


Supplementary figure S13.6: Phylostratigraphy of OGC85. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales*. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).

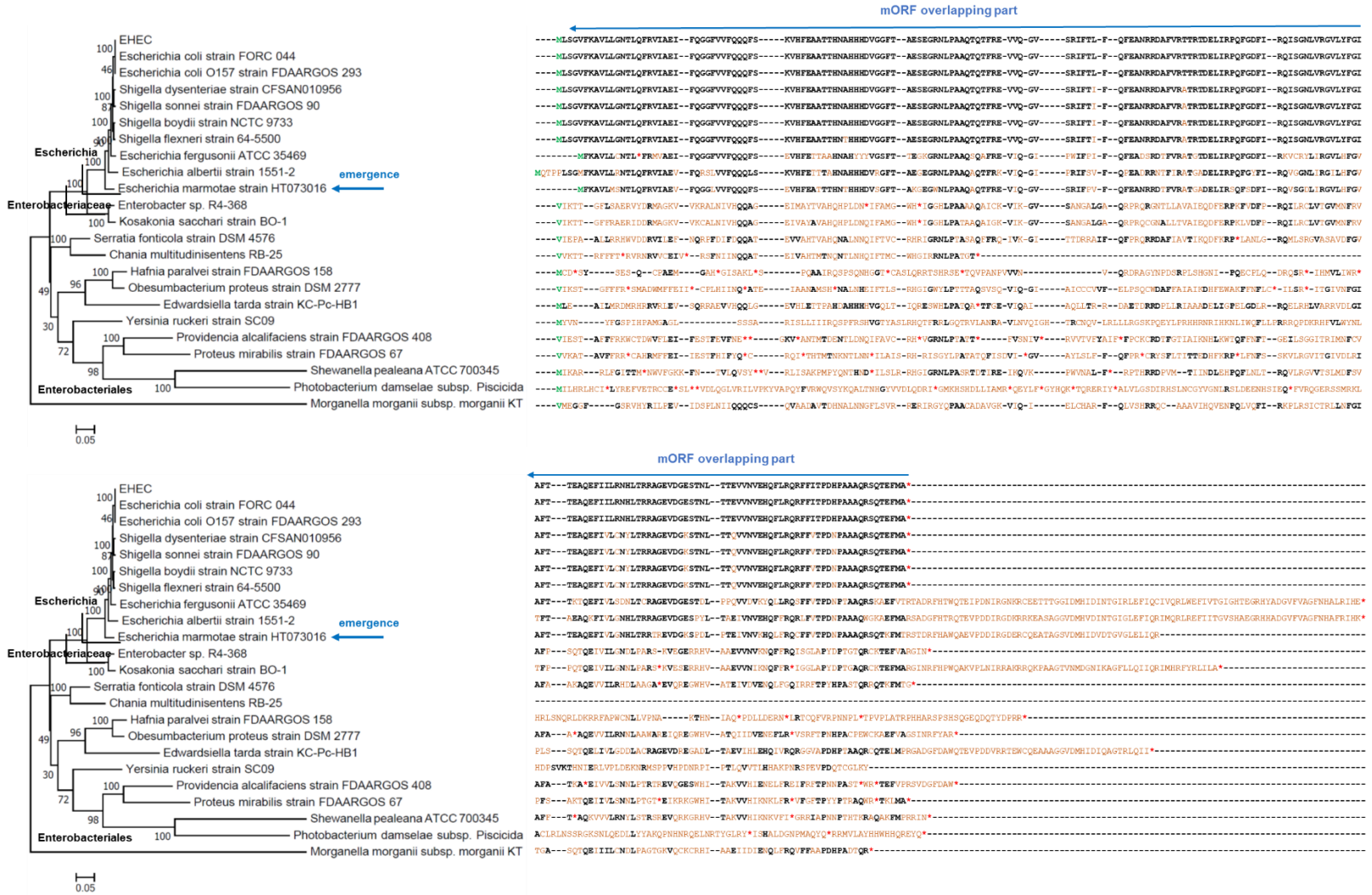
Supplementary figure S13.7: Phylostratigraphy of OGC121. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales*. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).



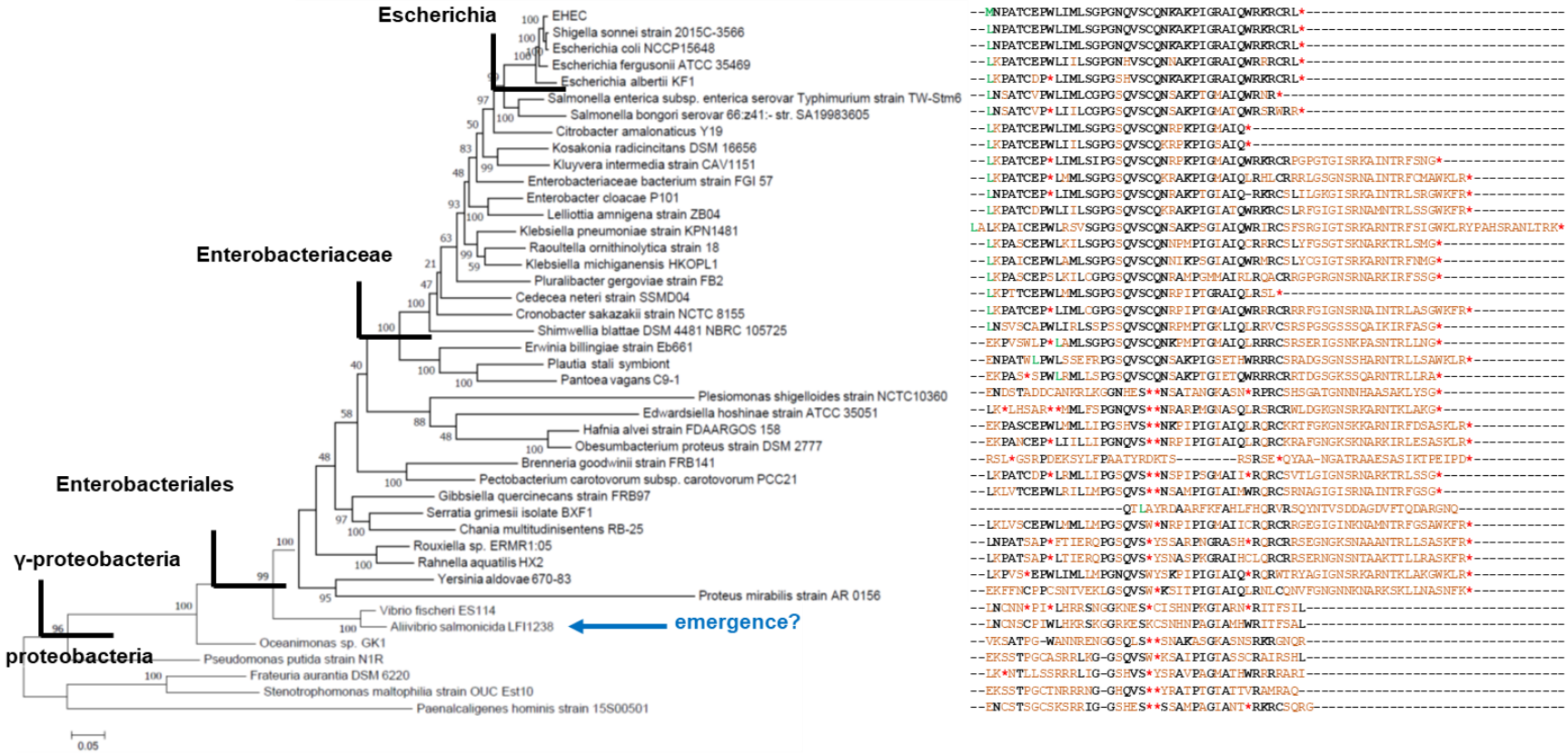
Supplementary figure S13.8: Phylostratigraphy of OGC167. Top: amino acid 1-73, bottom: 74-156. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales*. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).



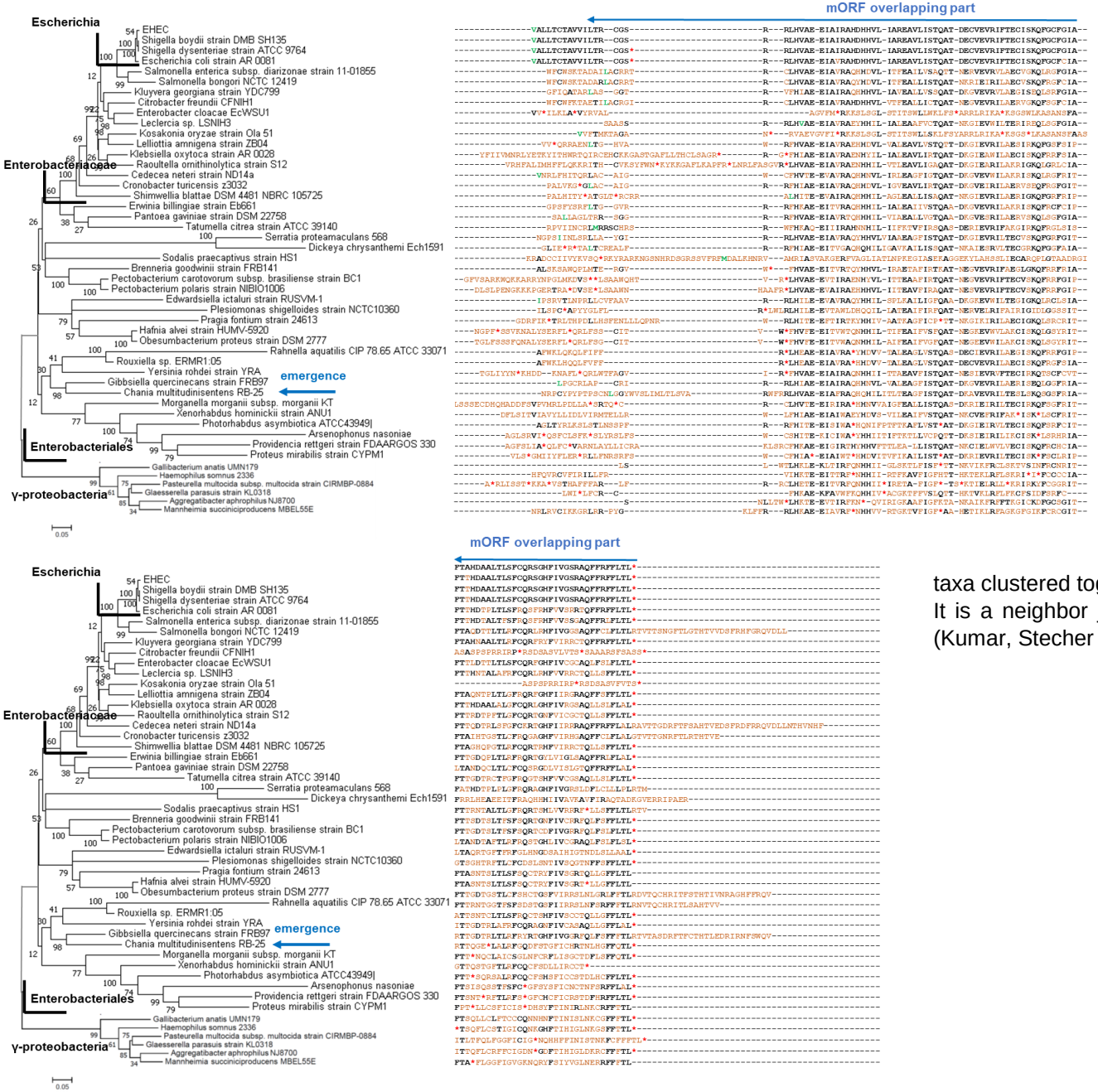
Supplementary figure S13.9: Phylostratigraphy of OGC174. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales*. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).



Supplementary figure S13.10: Phylostratigraphy of OGC194. Top: amino acid 1-131; bottom: 132-197. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales*. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).

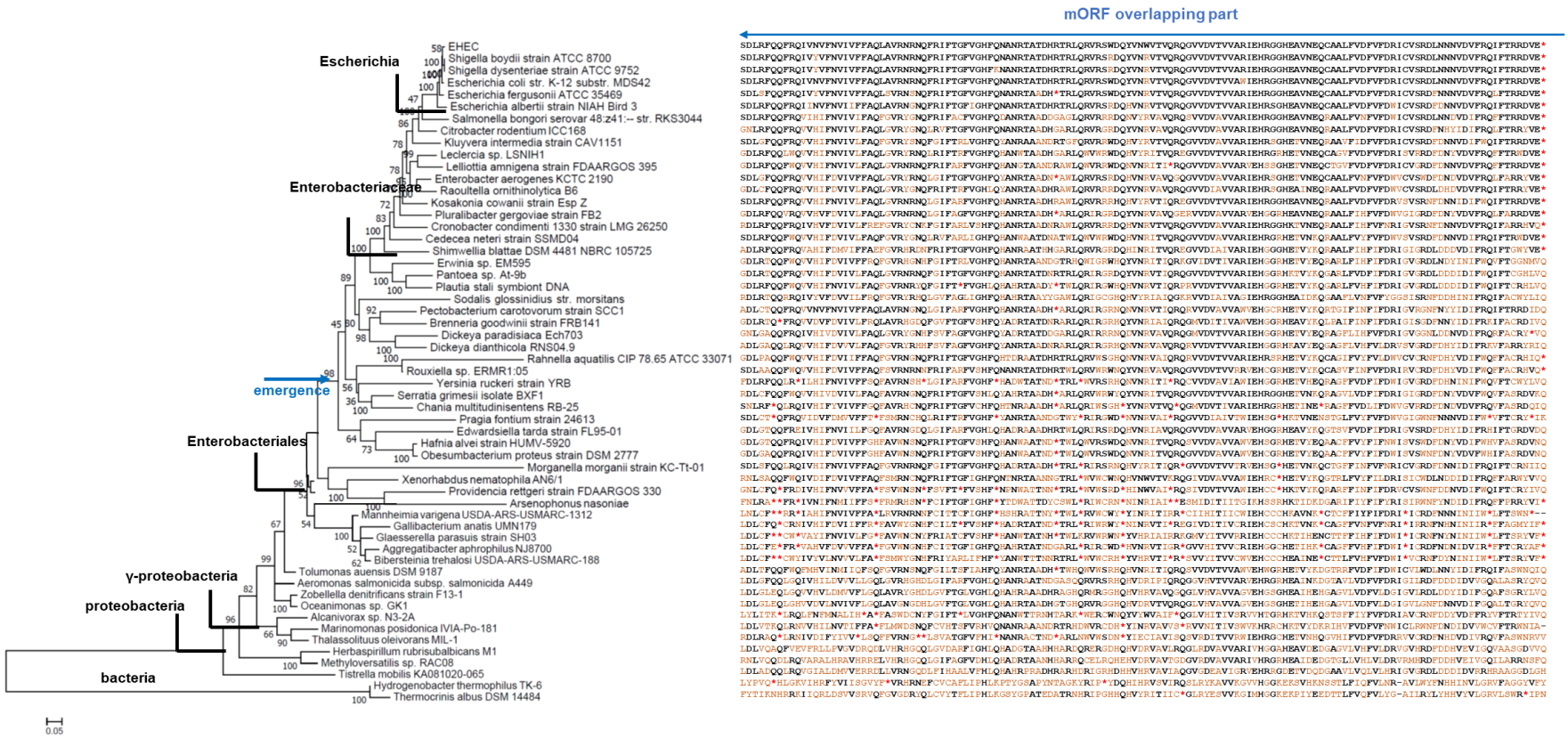


Supplementary figure S13.11: Phylostratigraphy of OGC198. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales* and 16S rRNA sequences in furthermore related species. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).



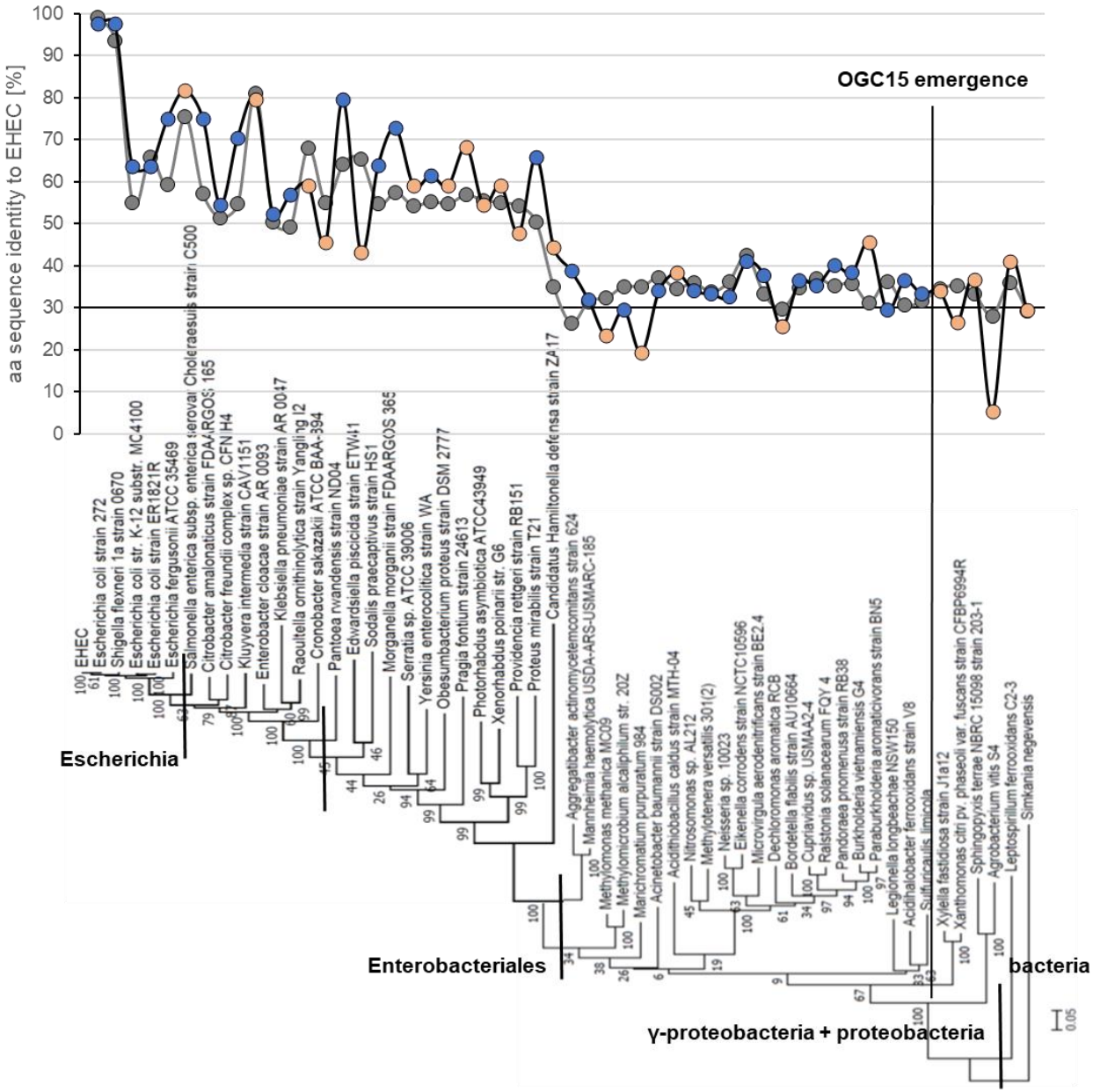
Supplementary figure S13.12: Phylostratigraphy of OGC226. Top: amino acid 1-72, bottom: amino acid: 73-106. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales* and 16S rRNA sequences in furthermore related species. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).


mORF overlapping part
Escherichia
EHEC
Shigella boydii strain ATCC 8700
Shigella dysenteriae strain ATCC 9752
Escherichia coli str. K-12 substr. MDS42
Escherichia fergusonii ATCC 35469
Escherichia albertii strain NIAH Bird 3
Salmonella bongori serovar 48:z41:-- str. RKS3044
Citrobacter rodentium ICC168
Kluyvera intermedia strain CAV1151
Leclercia sp. LSNIH1
Lelliottia amnigena strain FDAARGOS 395
Enterobacteriaceae
Enterobacter aerogenes KCTC 2190
Raoultella ornithinolytica B6
Kosakonia cowanii strain Esp Z
Pluralibacter gergoviae strain FB2
Cronobacter condimenti 1330 strain LMG 26250
Cedecea neteri strain SSMD04
Shimwellia blattae DSM 4481 NBRC 105725
Erwinia sp. EM595
Pantoea sp. At-9b
Plautia stali symbiont DNA
Sodalis glossinidius str. morsitans
Pectobacterium carotovorum strain SCC1
Brenneria goodwinii strain FRB141
Dickeya paradisiaca Ech703
Dickeya dianthicola RNS04.9
Rahnella aquatilis CIP 78.65 ATCC 33071
Rouxiella sp. ERMR1:05
emergence
Yersinia ruckeri strain YRB
Serratia grimesii isolate BXF1
Chania multitudinisentens RB-25
Pragia fontium strain 24613
Edwardsiella tarda strain FL95-01
Enterobacteriales
Hafnia alvei strain HUMV-5920
Obesumbacterium proteus strain DSM 2777
Morganella morganii strain KC-Tt-01
Xenorhabdus nematophila AN6/1
Providencia rettgeri strain FDAARGOS 330
Arsenophonus nasoniae
Mannheimia varigena USDA-ARS-USMARC-1312
Gallibacterium anatis UMN179
Glaesserella parasuis strain SH03
Aggregatibacter aphrophilus NJ8700
Bibersteinia trehalosi USDA-ARS-USMARC-188
γ-proteobacteria
Tolumonas auensis DSM 9187
Aeromonas salmonicida subsp. salmonicida A449
Zobellella denitrificans strain F13-1
proteobacteria
Oceanimonas sp. GK1
Alcanivorax sp. N3-2A
Marinomonas posidonica IVIA-Po-181
Thalassolituus oleivorans MIL-1
Herbaspirillum rubrisubalbicans M1
Methyloversatilis sp. RAC08
bacteria
Tistrella mobilis KA081020-065
Hydrogenobacter thermophilus TK-6
Thermocrinis albus DSM 14484
0.05

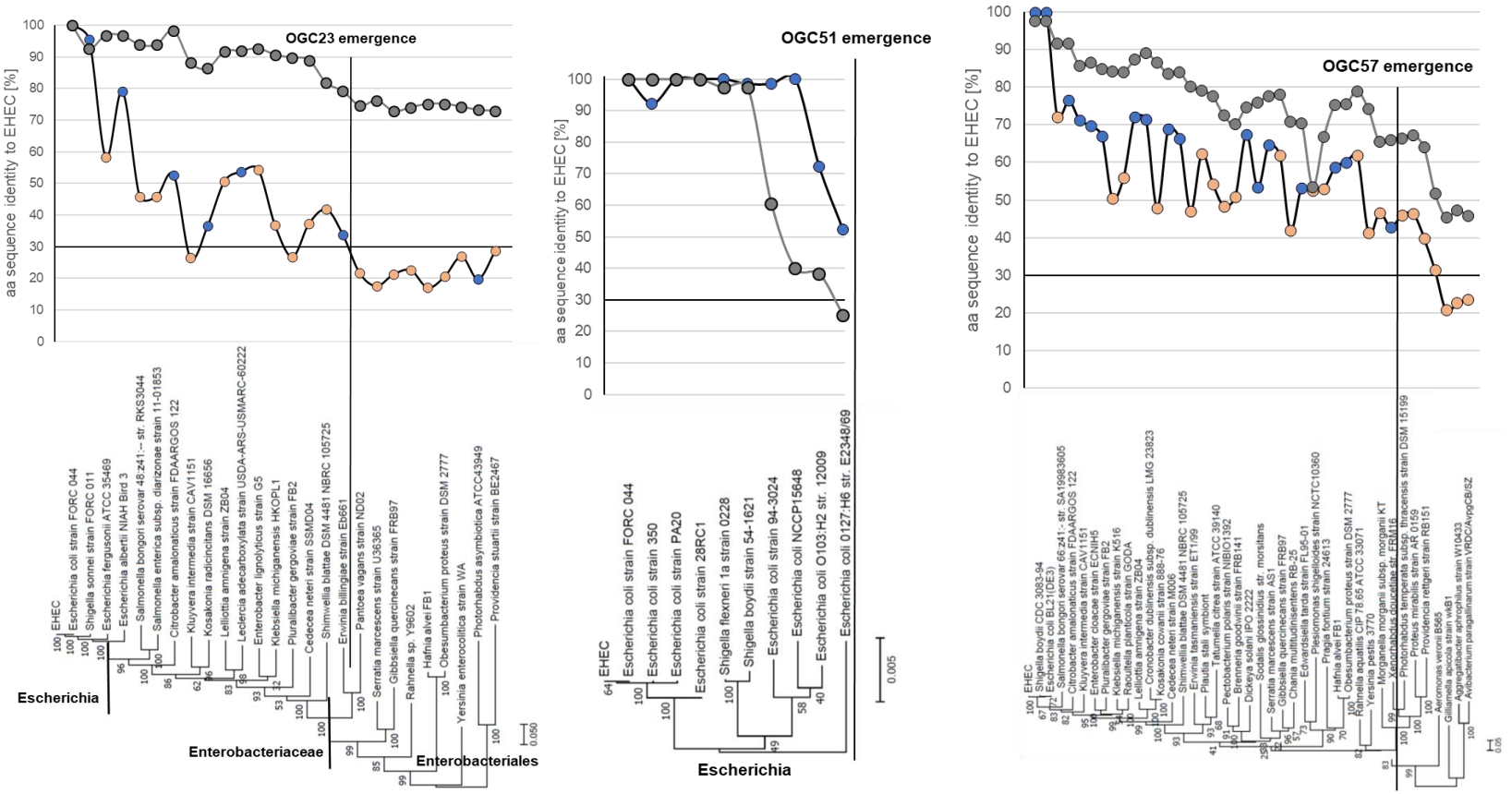


Supplementary figure S13.13: Phylostratigraphy of OGC231. Top: amino acid 1-89, bottom: amino acid: 90-222. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales* and 16S rRNA sequences in furthermore related species. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).

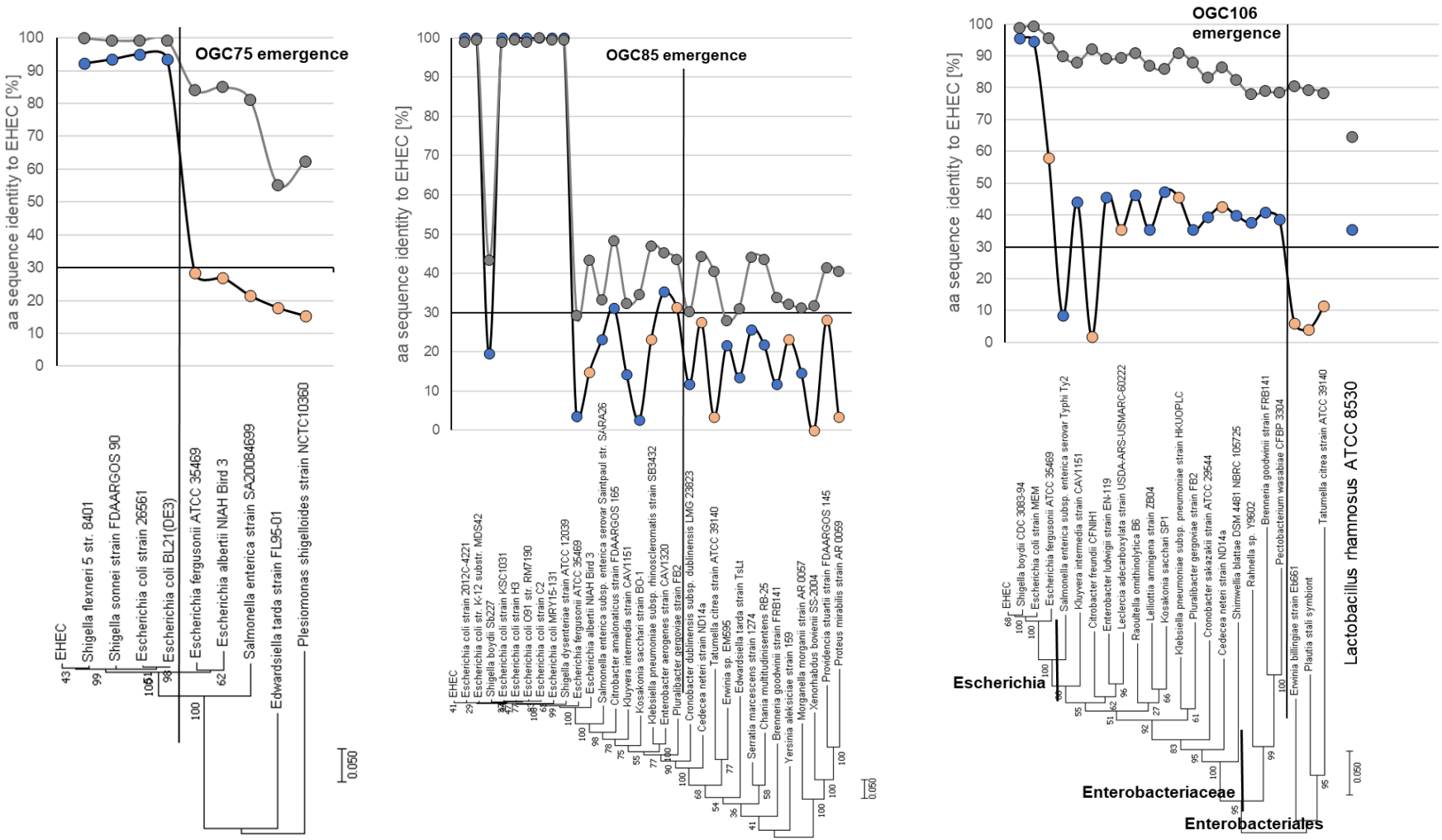
Supplementary figure S13.14: Phylostratigraphy of OGC241. The species tree is constructed from organisms in which the mORF homologue was found. The respective OGC23 homologous amino acid sequences were identified. Start and stop are marked in green letters or red hash. Amino acids changed compared to *asa* in EDL933 are brown. Conserved amino acids are black, bold. The species tree was constructed from MLSA sequences within *Enterobacteriales*. The percentage of trees in which the associated taxa clustered together is shown next to the branches. It is a neighbor joining tree calculated with Mega7.0 (Kumar, Stecher et al. 2016).



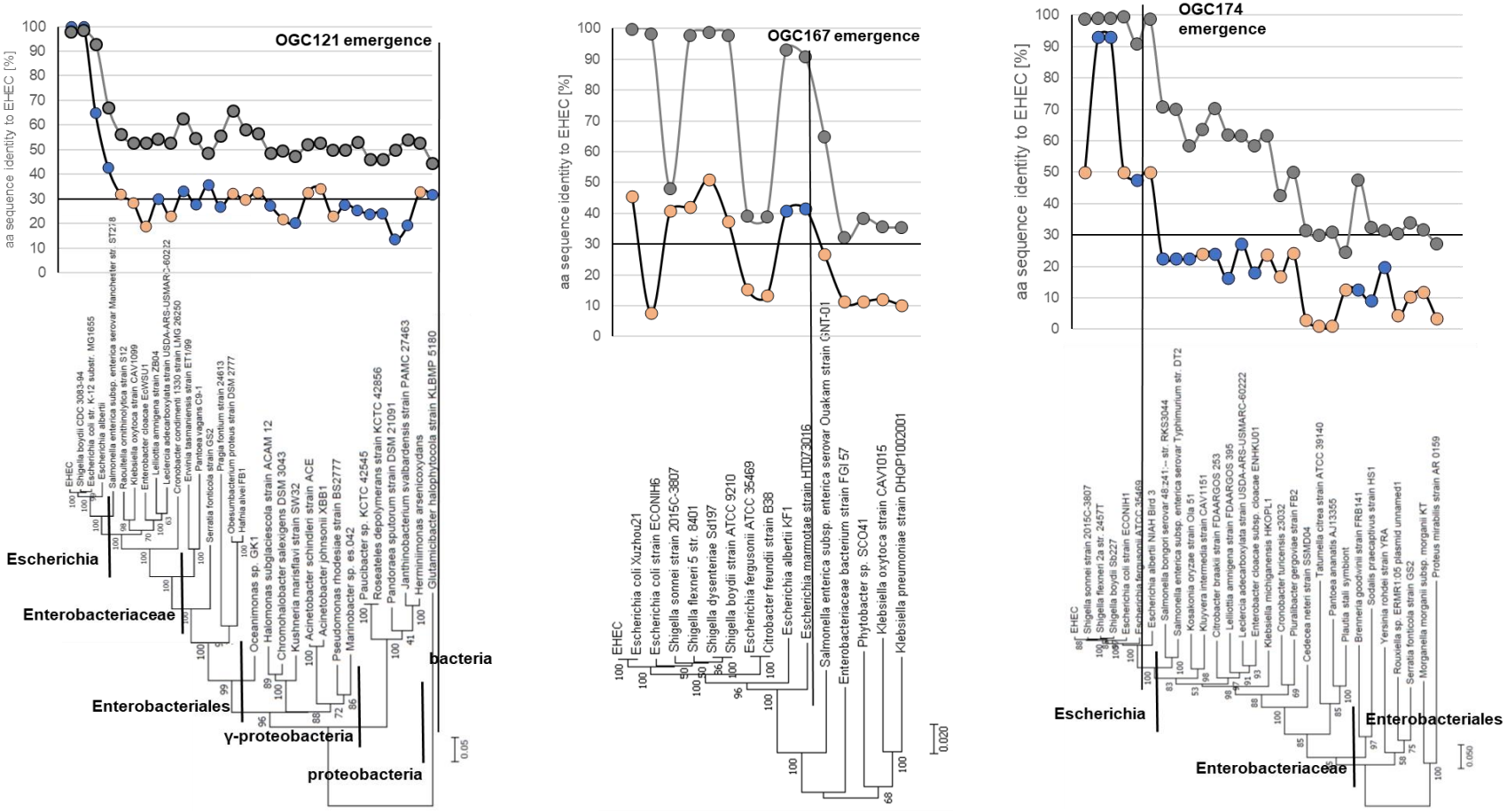
Supplementary figure S14.1: Amino acid (aa) sequence identities to EHEC [%] of the overlapping gene pair OGC15/EDL933_0277 during species evolution. Legend: graph on top: Sequence identity to EHEC of each homologue shown in the tree below. grey: EDL933_0277 homologues, blue: intact sORF homologues, salmon: sORF homologues with internal stop codons; the sequence homology cutoff of 30% is shown as black line. Graph below: Maximum likelihood tree of species containing EDL933_0277 homologues. The tree was constructed as combination of MLSA tree (Escherichia to Entero-bacteriales) and a 16S rRNA tree (γ-proteobacteria and further related). The tree was calculated using MEGA 7.0 (Kumar, Stecher et al. 2016).



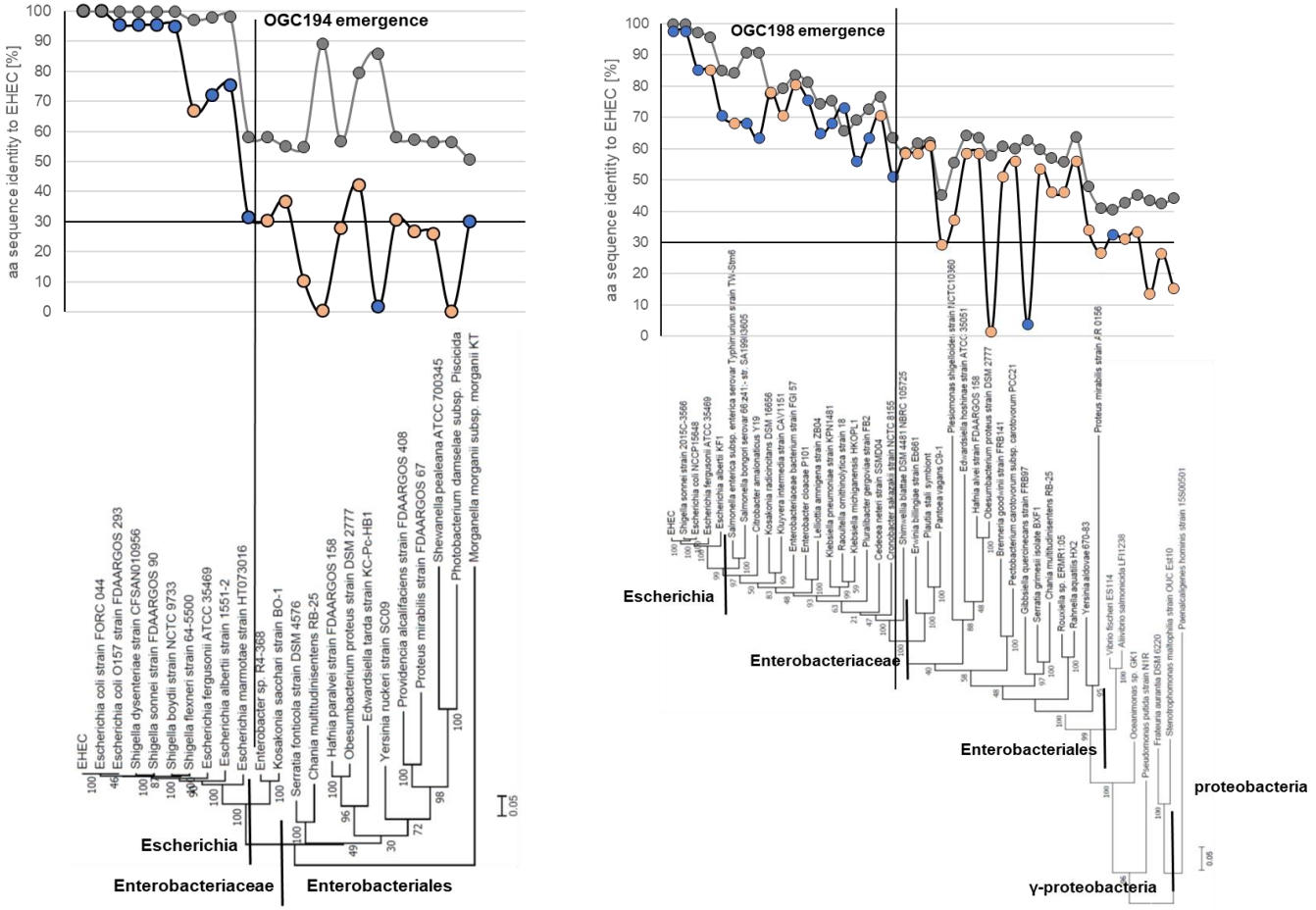
Supplementary figure S14.2: Amino acid (aa) sequence identities to EHEC [%] of the overlapping gene pairs OGC23/ EDL933_0555 (left) and OGC51/ EDL933_1089 (center) and OGC57/EDL933_1124 (right) during species evolution. Legend: graph on top: Sequence identity to EHEC of each homologue; grey: mORF homologues, blue: intact sORF homologues, salmon: sORF homologues with internal stop codons; the sequence homology cutoff of 30% is shown as black line. Graph below: Maximum likelihood tree of species containing the respective mORF homologues. The tree was constructed as a combination of MLSA tree (Escherichia to Enterobacteriales) and a 16S rRNA tree (γ-proteobacteria and further related) and calculated using MEGA 7.0 (Kumar, Stecher et al. 2016).



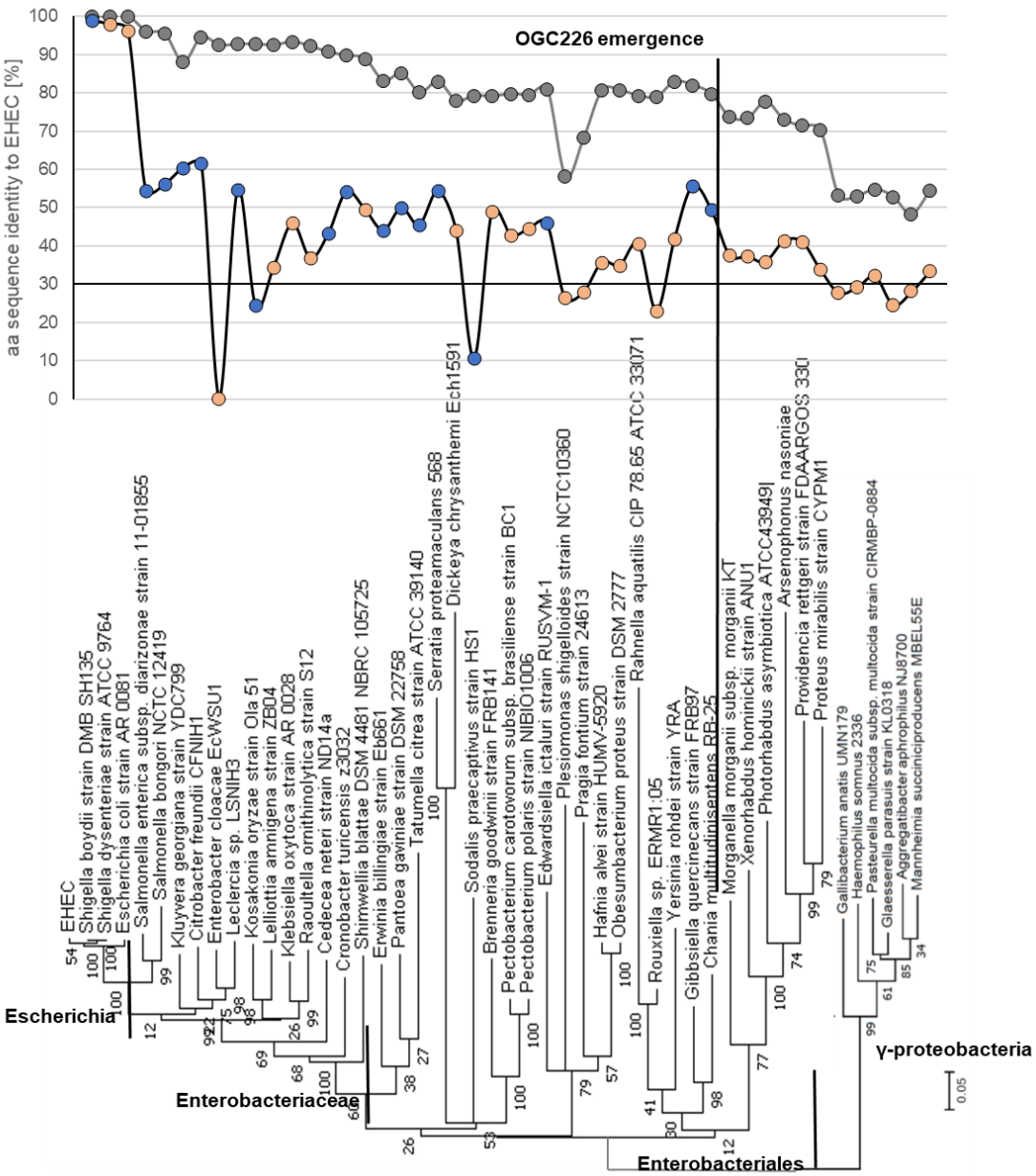
Supplementary figure S14.3: Sequence identities to EHEC [%] of the overlapping gene pairs OGC75/EDL933_1870 (left) and OGC85/EDL933_2135 (center) and OGC106/EDL933_2699 (right) during species evolution. Legend: graph on top: Sequence identity to EHEC of each homologue shown in the tree below. grey: mORF homologues, blue: intact sORF homologues, salmon: sORF homologues with internal stop codons; the sequence homology cutoff of 30% is shown as black line. Graph below: Maximum likelihood tree of species containing the respective mORF homologues. The tree was constructed as MLSA tree and calculated using MEGA 7.0 (Kumar, Stecher et al. 2016).



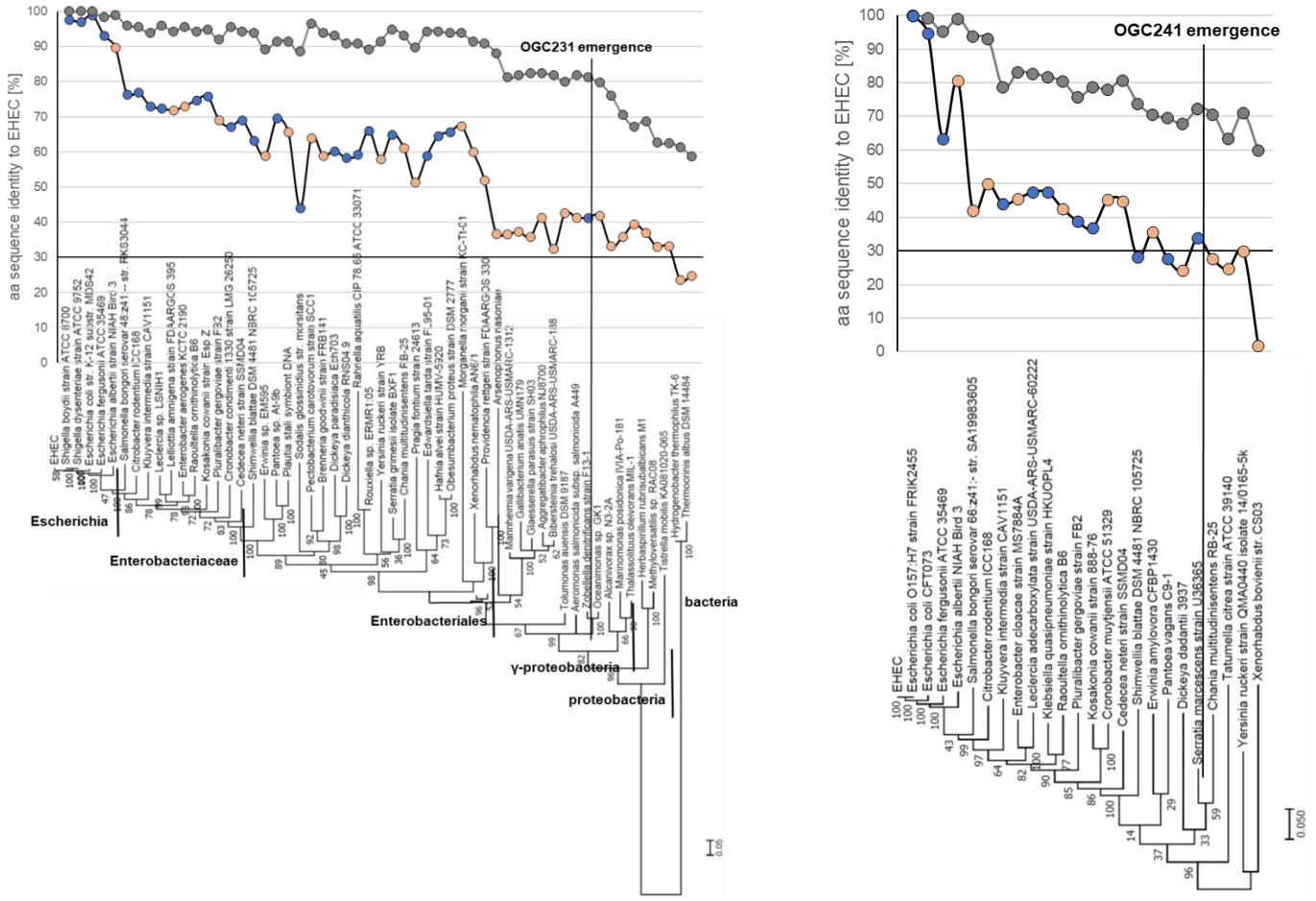
Supplementary figure S14.4: Amino acid (aa) sequence identities to EHEC [%] of the overlapping gene pairs OGC121/EDL933_2979 (left) and OGC167/EDL933_4168 (center) and OGC174/EDL933_4292 (right) during species evolution. Legend: graph on top: Sequence identity to EHEC of each homologue shown in the tree below. grey: mORF homologues, blue: intact sORF homologues, salmon: sORF homologues with internal stop codons; the sequence homology cutoff of 30% is shown as black line. Graph below: Maximum likelihood tree of species containing the respective mORF homologues. The tree was constructed as a combination of MLSA tree (Escherichia to Enterobacteriales) and a 16S rRNA tree (γ-proteobacteria and further related) and calculated using MEGA 7.0 (Kumar, Stecher et al. 2016).



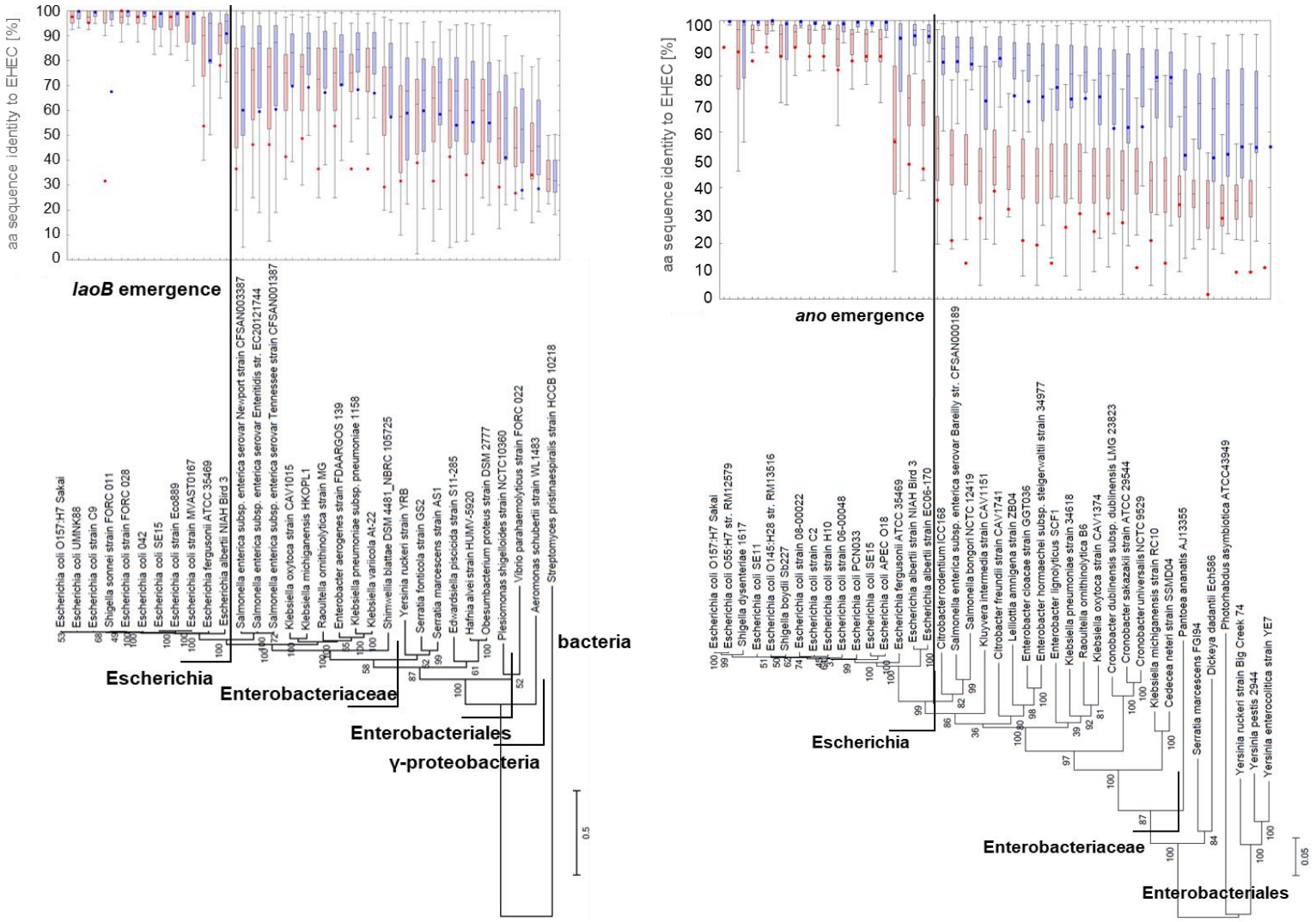
Supplementary figure S14.5: Amino acid (aa) sequence identities to EHEC [%] of the overlapping gene pairs OGC194/EDL933_4769 (left) and OGC198/EDL933_4794 (right) during species evolution. Legend: graph on top: Sequence identity to EHEC of each homologue shown in the tree below. grey: mORF homologues, blue: intact sORF homologues, salmon: sORF homologues with internal stop codons; the sequence homology cutoff of 30% is shown as black line. Graph below: Maximum likelihood tree of species containing the respective mORF homologues. The tree was constructed as a combination of MLSA tree (Escherichia to Enterobacteriales) and a 16S rRNA tree (γ-proteobacteria and further related) and calculated using MEGA 7.0 (Kumar, Stecher et al. 2016).



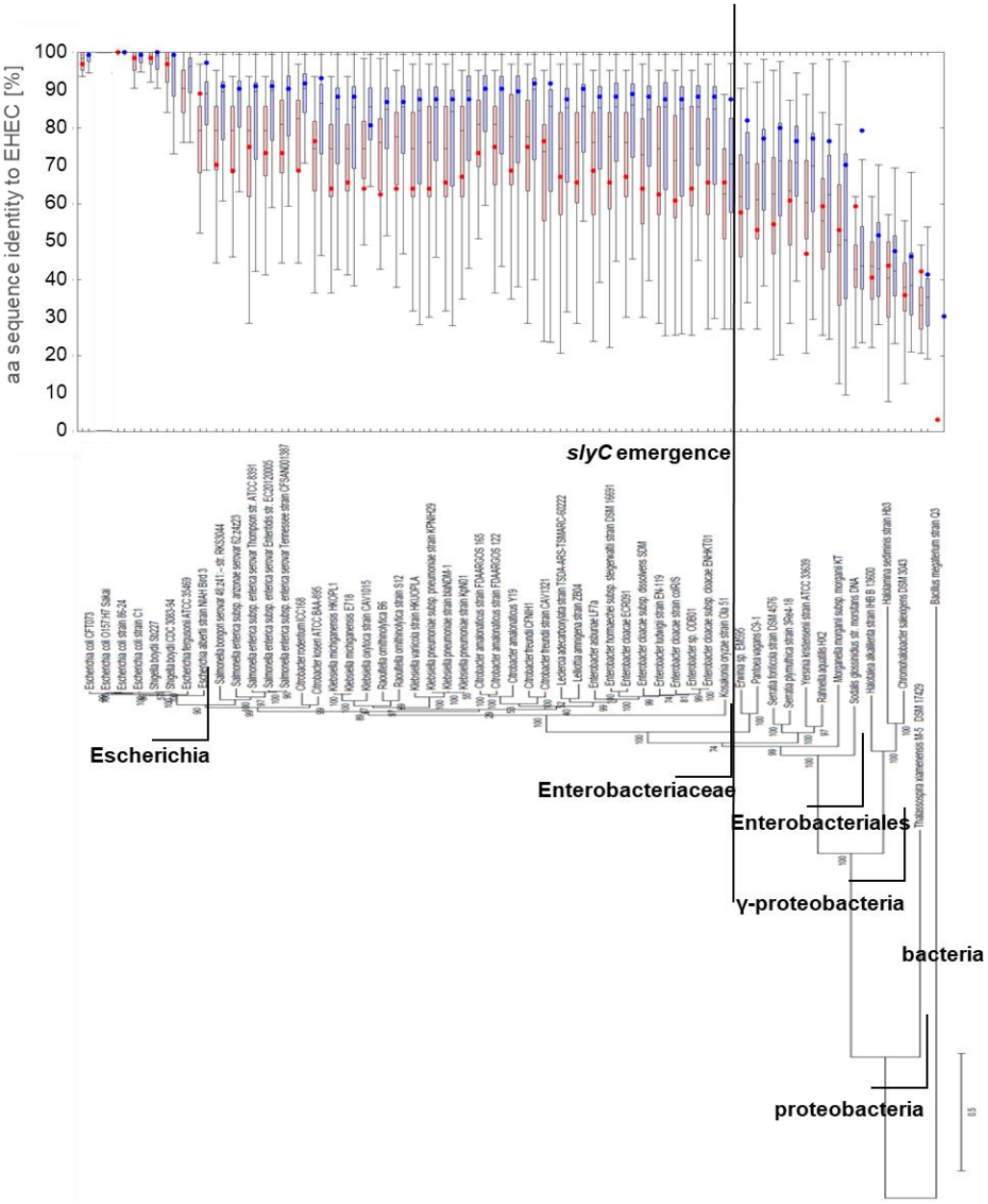
Supplementary figure S14.6: Amino acid (aa) sequence identities to EHEC [%] of the overlapping gene pair OGC226/EDL933_4769 during species evolution. Legend: graph on top: Sequence identity to EHEC of each homologue shown in the tree below. grey: EDL933_0555 homologues, blue: intact sORF homologues, salmon: sORF homologues with internal stop codons; the sequence homology cutoff of 30% is shown as black line. Graph below: Maximum likelihood tree of species containing the respective mORF homologues. The tree was constructed as a combination of MLSA tree (Escherichia to Enterobacteriales) and a 16S rRNA tree (γ-proteobacteria and further related) and calculated using MEGA 7.0 (Kumar, Stecher et al. 2016).



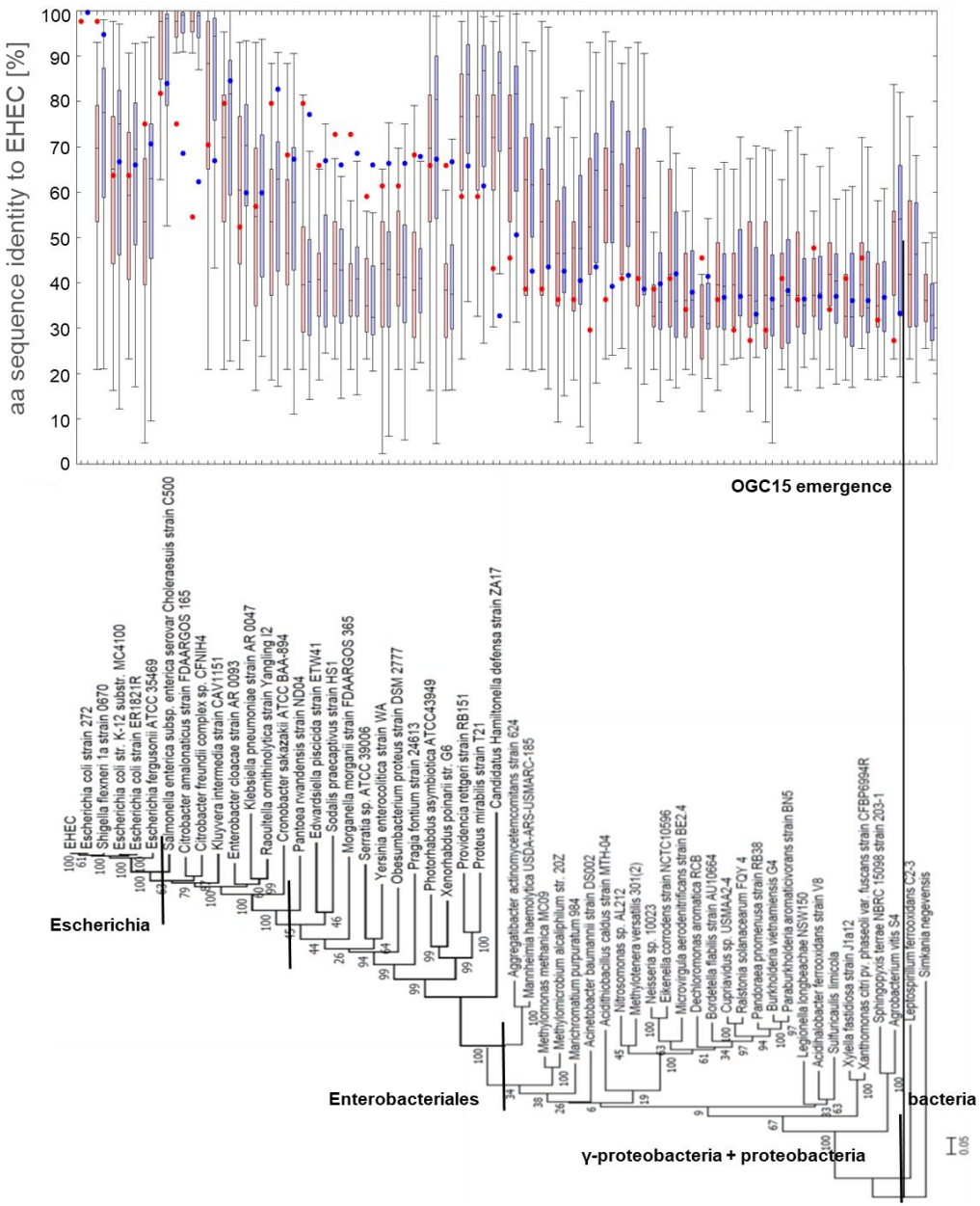
Supplementary figure S14.7: Amino acid (aa) sequence identities to EHEC [%] of the overlapping gene pairs OGC231/EDL933_5573 and OGC241/EDL933_5740 during species evolution. Legend: graph on top: Sequence identity to EHEC of each homologue shown in the tree below. grey: mORF homologues, blue: intact sORF homologues, salmon: sORF homologues with internal stop codons; the sequence homology cutoff of 30% is shown as black line. Graph below: Maximum likelihood tree of species containing the respective mORF homologues. The tree was constructed as MLSA tree (Escherichia to Enterobacteriales) and a 16S rRNA tree (γ-proteobacteria and further related) and calculated using MEGA 7.0 (Kumar, Stecher et al. 2016).



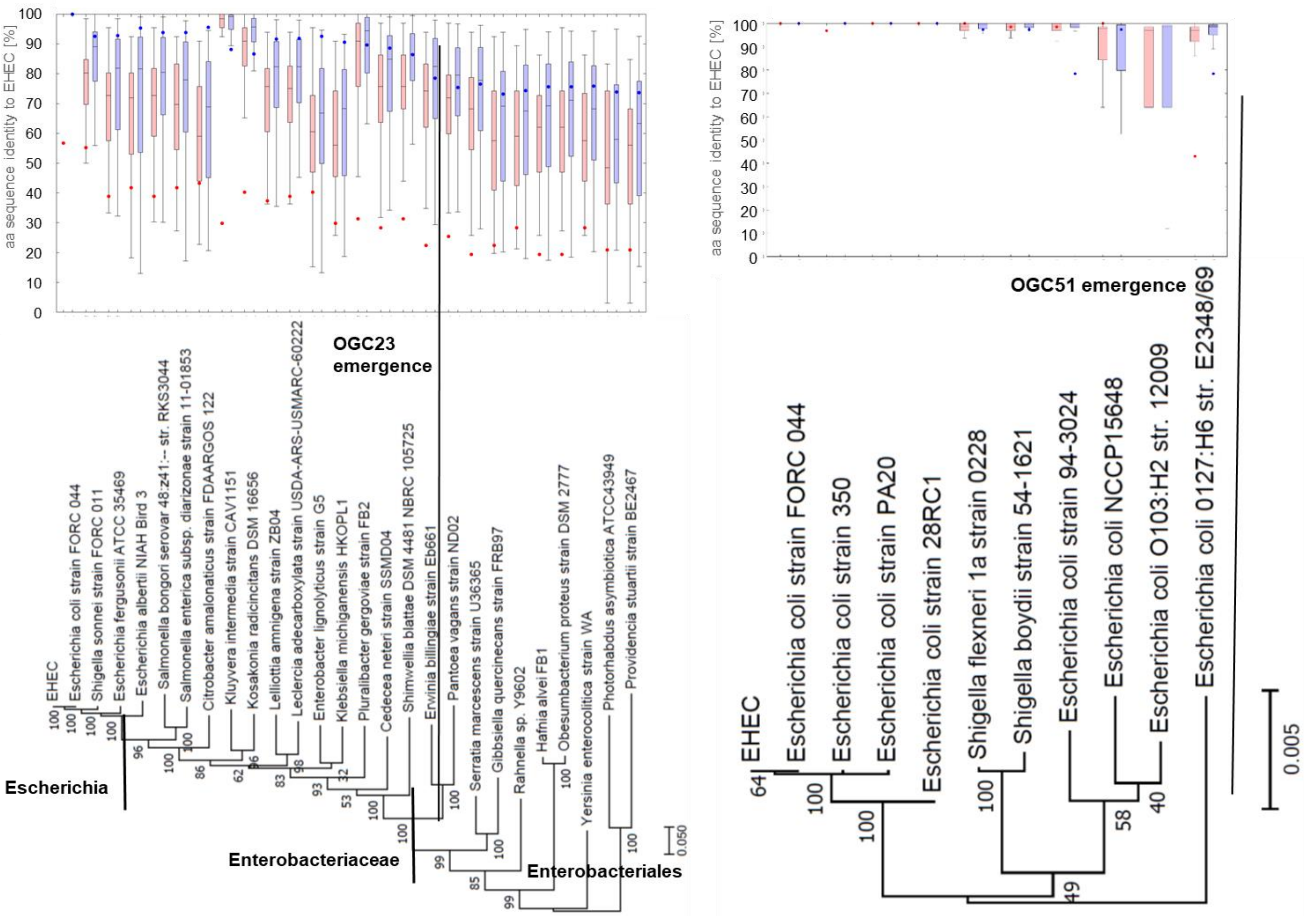
Supplementary figure S15.1: Amino acid (aa) sequence identities of the overlapping gene pairs *laoB*/ECs5115 and *ano*/ECs2385 and of a negative control. The negative control was calculated from 100 randomly selected genes and an overlapping sequence in the organisms in which an overlapping gene pair homologue was found. The overlapping sequence of the negative control is a random sequence of the length of the sORF, embedded in the selected random gene and in the reading frame in which the sORF is located.



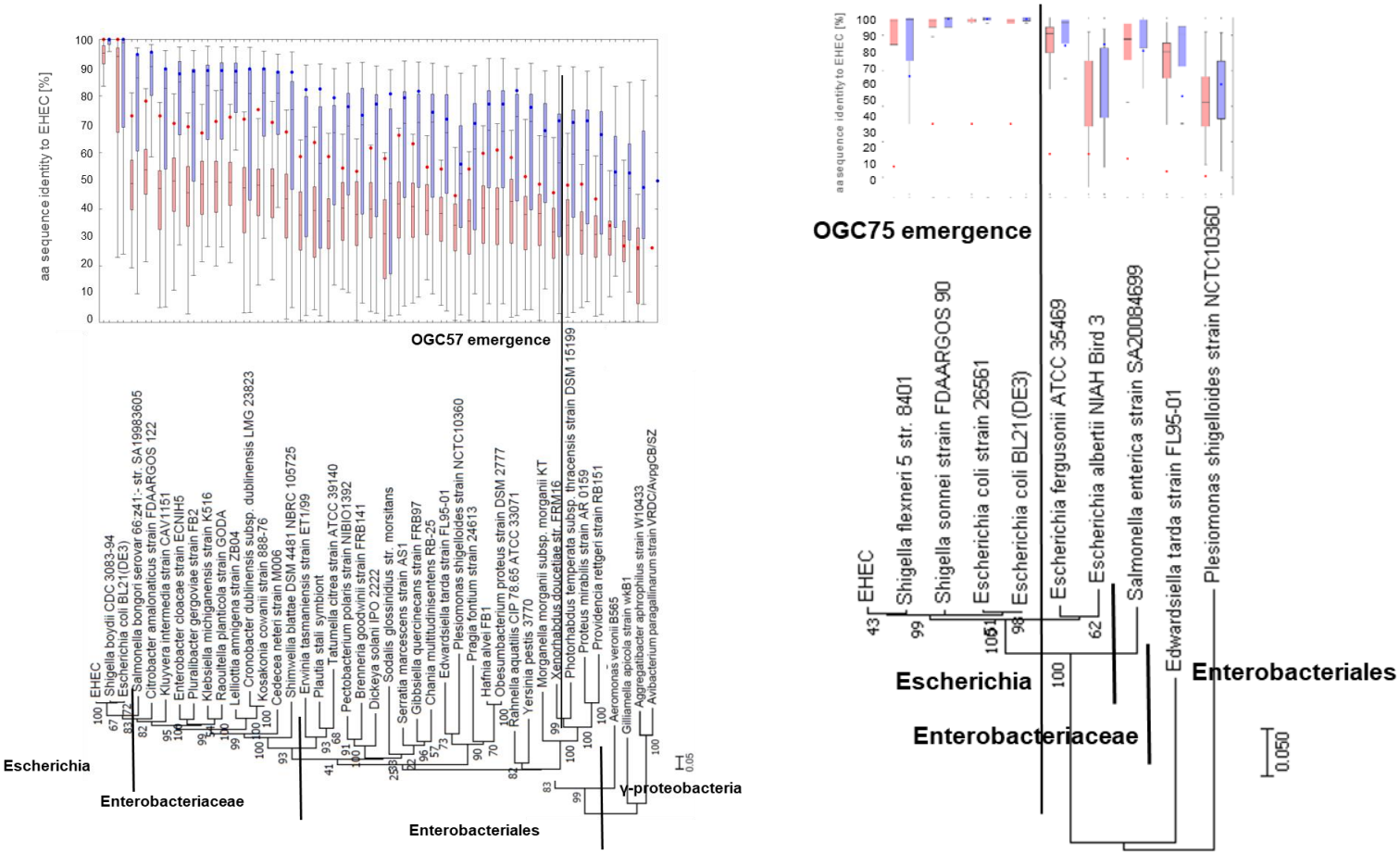
Supplementary figure S15.2: Amino acid (aa) sequence identities of the overlapping gene pair *slyC*/ECs2351 and of a negative control. The negative control was calculated from 100 randomly selected genes and an overlapping sequence in the organisms in which an overlapping gene pair homologue was found. The overlapping sequence of the negative control is a random sequence of the length of the sORF, embedded in the selected random gene and in the reading frame in which the sORF is located.



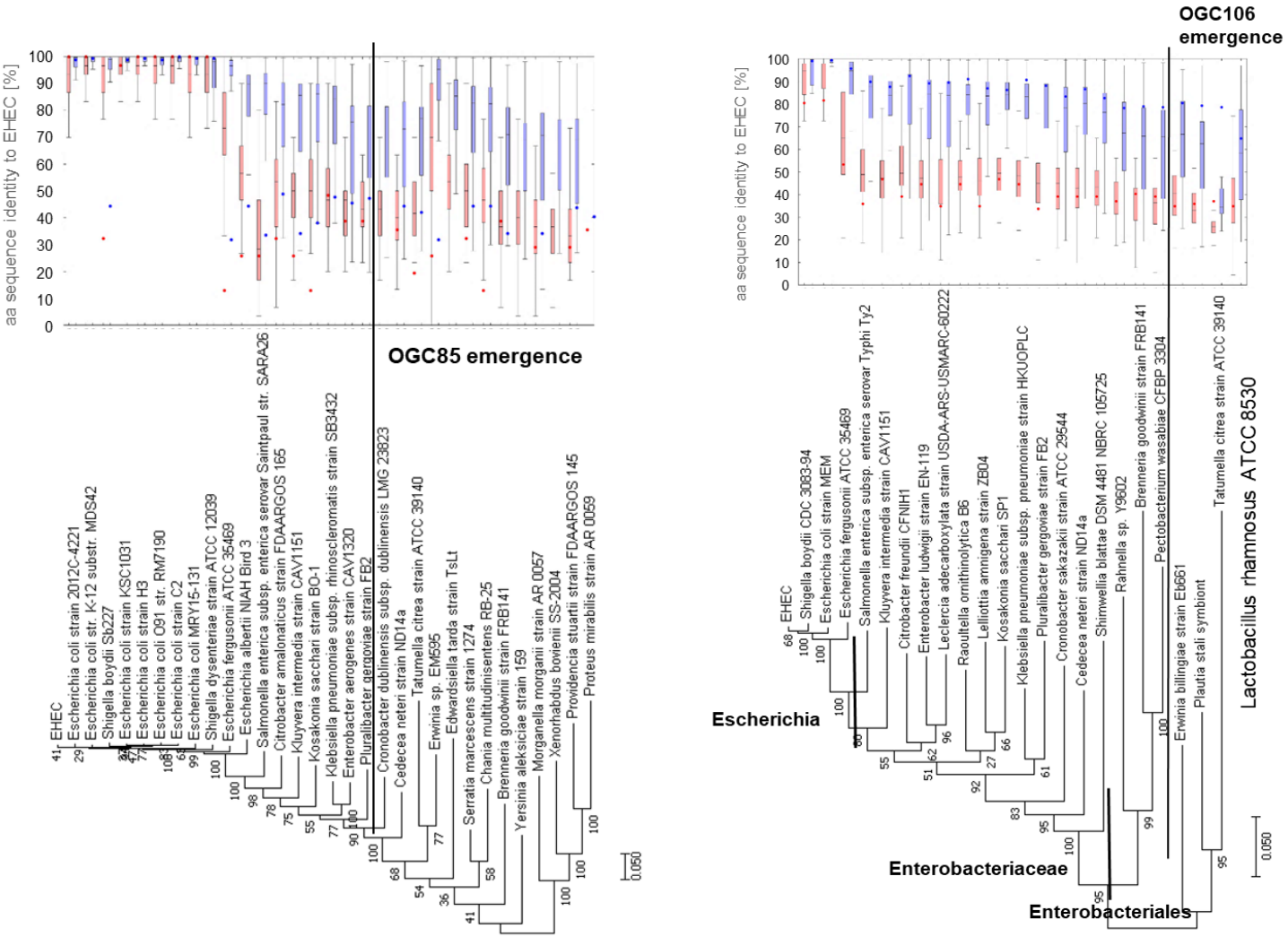
Supplementary figure S15.3: Amino acid (aa) sequence identities of the overlapping gene pair OGC15/ EDL933_0277 and of a negative control. The negative control was calculated from 100 randomly selected genes and an overlapping sequence in the organisms in which an overlapping gene pair homologue was found. The overlapping sequence of the negative control is a random sequence of the length of the sORF, embedded in the selected random gene and in the reading frame in which the sORF is located.



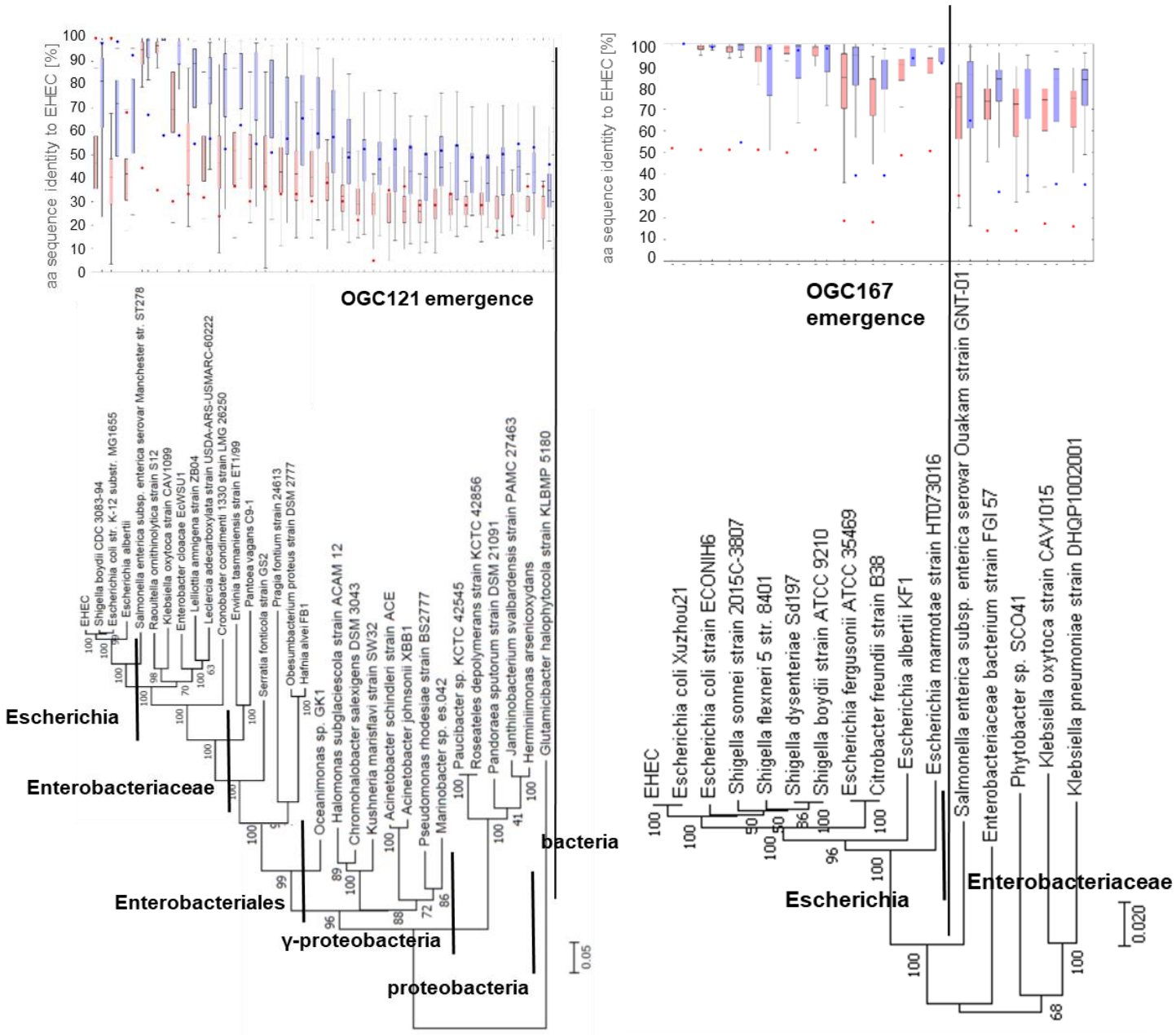
Supplementary figure S15.4: Amino acid (aa) sequence identities of the overlapping gene pair OGC23/ EDL933_0555 and OGC51/ EDL933_1089 and of a negative control. The negative control was calculated from 100 randomly selected genes and an overlapping sequence in the organisms in which an overlapping gene pair homologue was found. The overlapping sequence of the negative control is a random sequence of the length of the sORF, embedded in the selected random gene and in the reading frame in which the sORF is located.



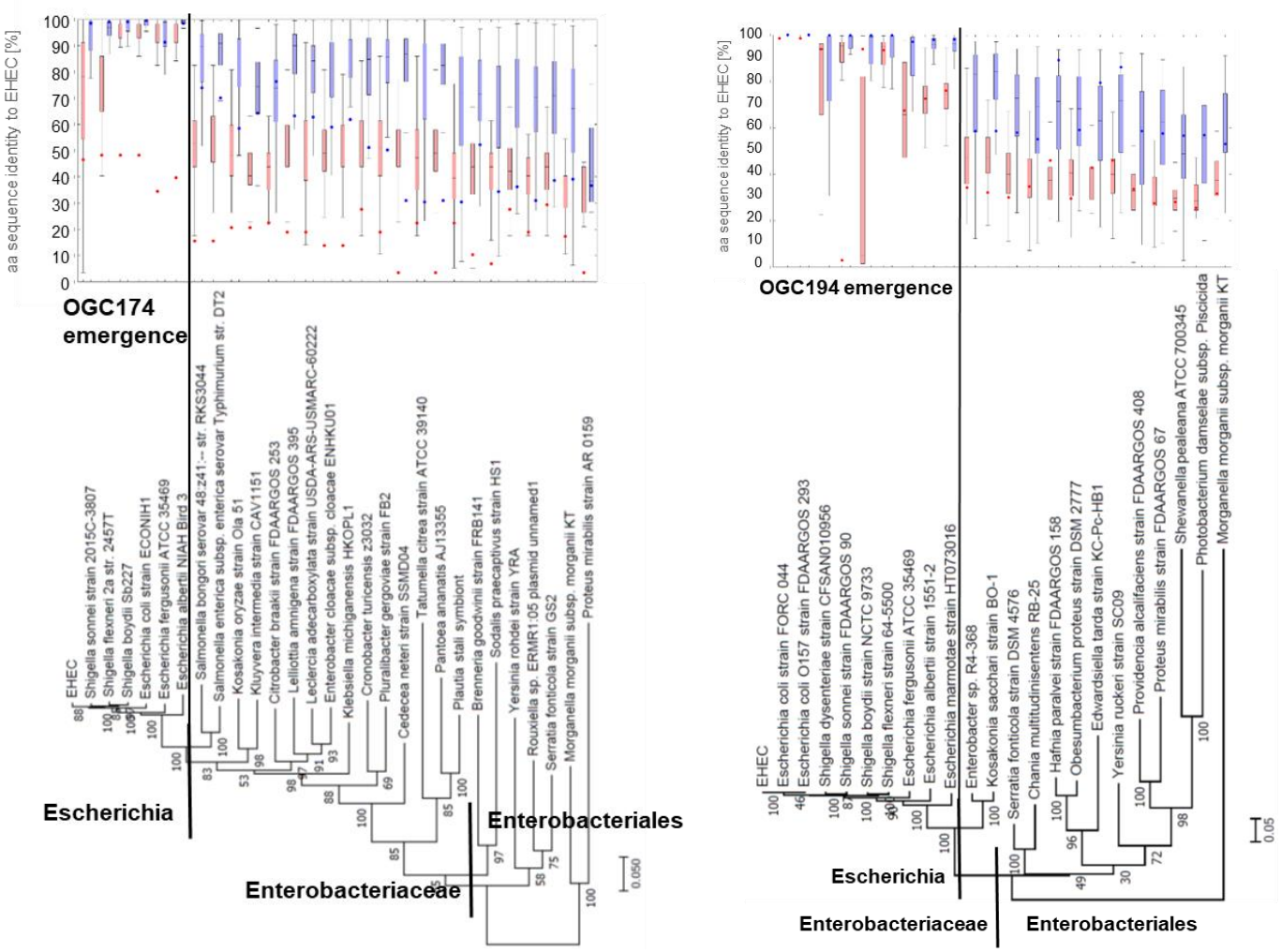
Supplementary figure S15.5: Amino acid (aa) sequence identities of the overlapping gene pairs OGC57/ EDL933_1224 and OGC75/ EDL933_1870 and of a negative control. The negative control was calculated from 100 randomly selected genes and an overlapping sequence in the organisms in which an overlapping gene pair homologue was found. The overlapping sequence of the negative control is a random sequence of the length of the sORF, embedded in the selected random gene and in the reading frame in which the sORF is located.



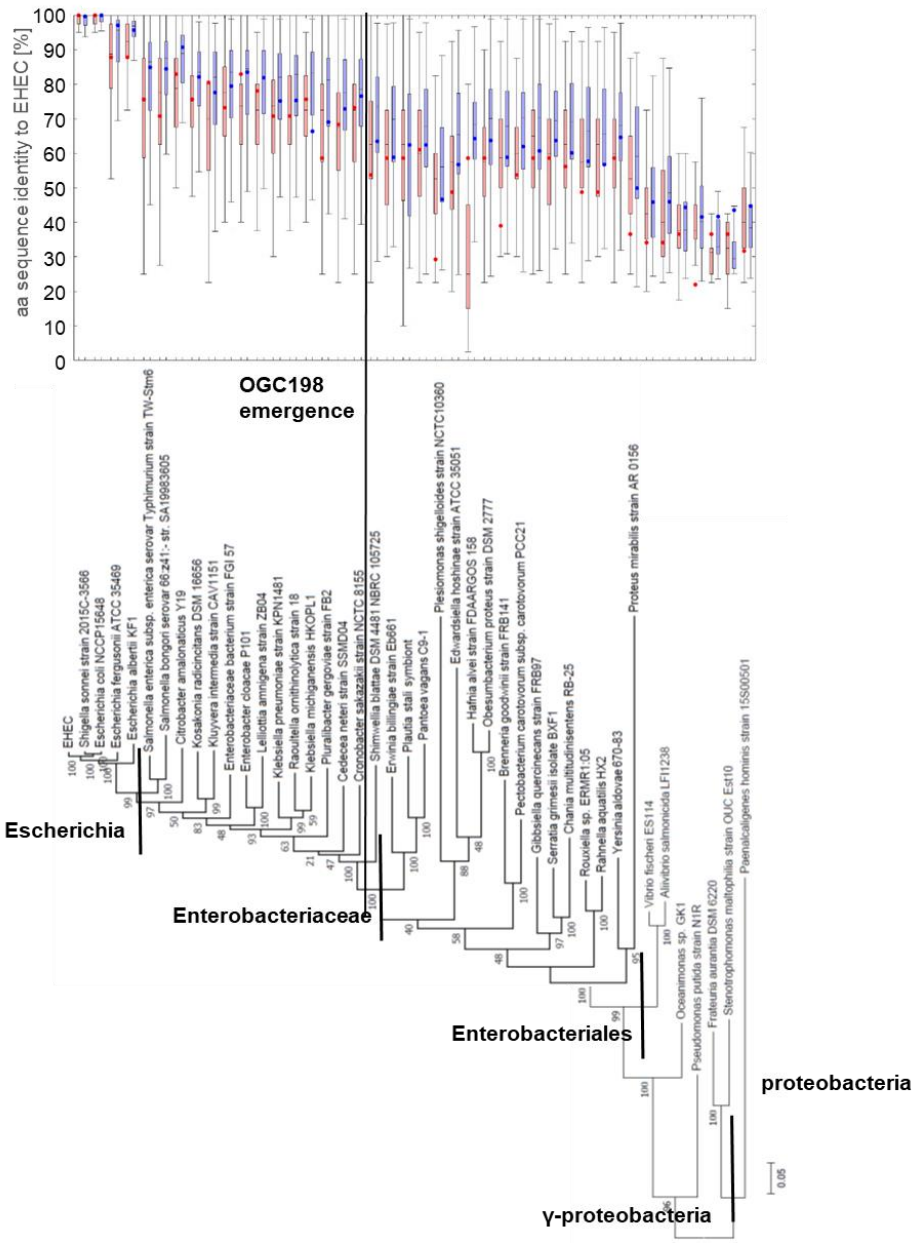
Supplementary figure S15.6: Amino acid (aa) sequence e identities of the overlapping gene pairs OGC85/ EDL933_2135 and OGC106/ EDL933_2699 and of a negative control. The negative control was calculated from 100 randomly selected genes and an overlapping sequence in the organisms in which an overlapping gene pair homologue was found. The overlapping sequence of the negative control is a random sequence of the length of the sORF, embedded in the selected random gene and in the reading frame in which the sORF is located.



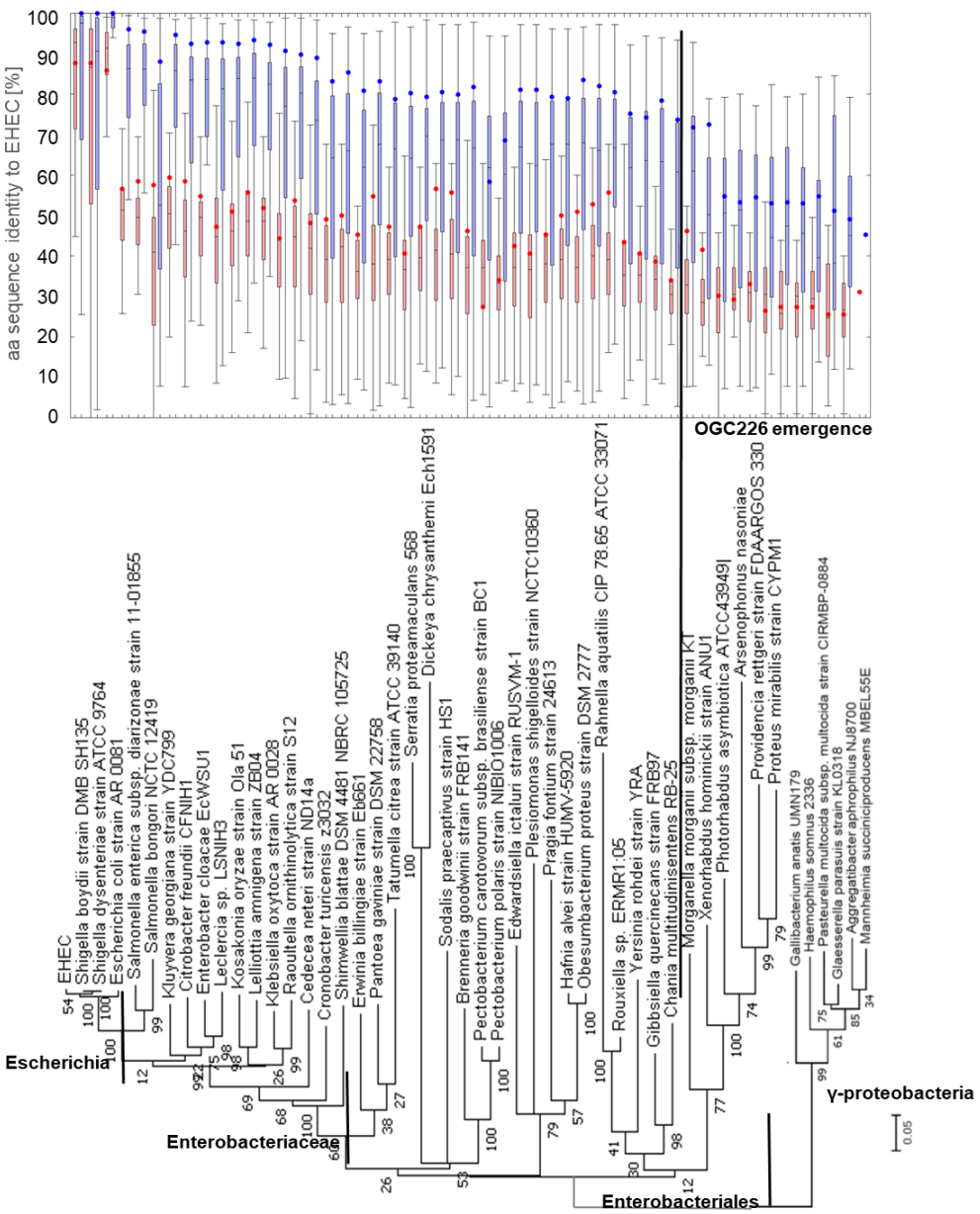
Supplementary figure S15.7: Amino acid (aa) sequence identities of the overlapping gene pairs OGC121/ EDL933_ 2979 and OGC167/ EDL933_4168 and of a negative control. The negative control was calculated from 100 randomly selected genes and an overlapping sequence in the organisms in which an overlapping gene pair homologue was found. The overlapping sequence of the negative control is a random sequence of the length of the sORF, embedded in the selected random gene and in the reading frame in which the sORF is located.



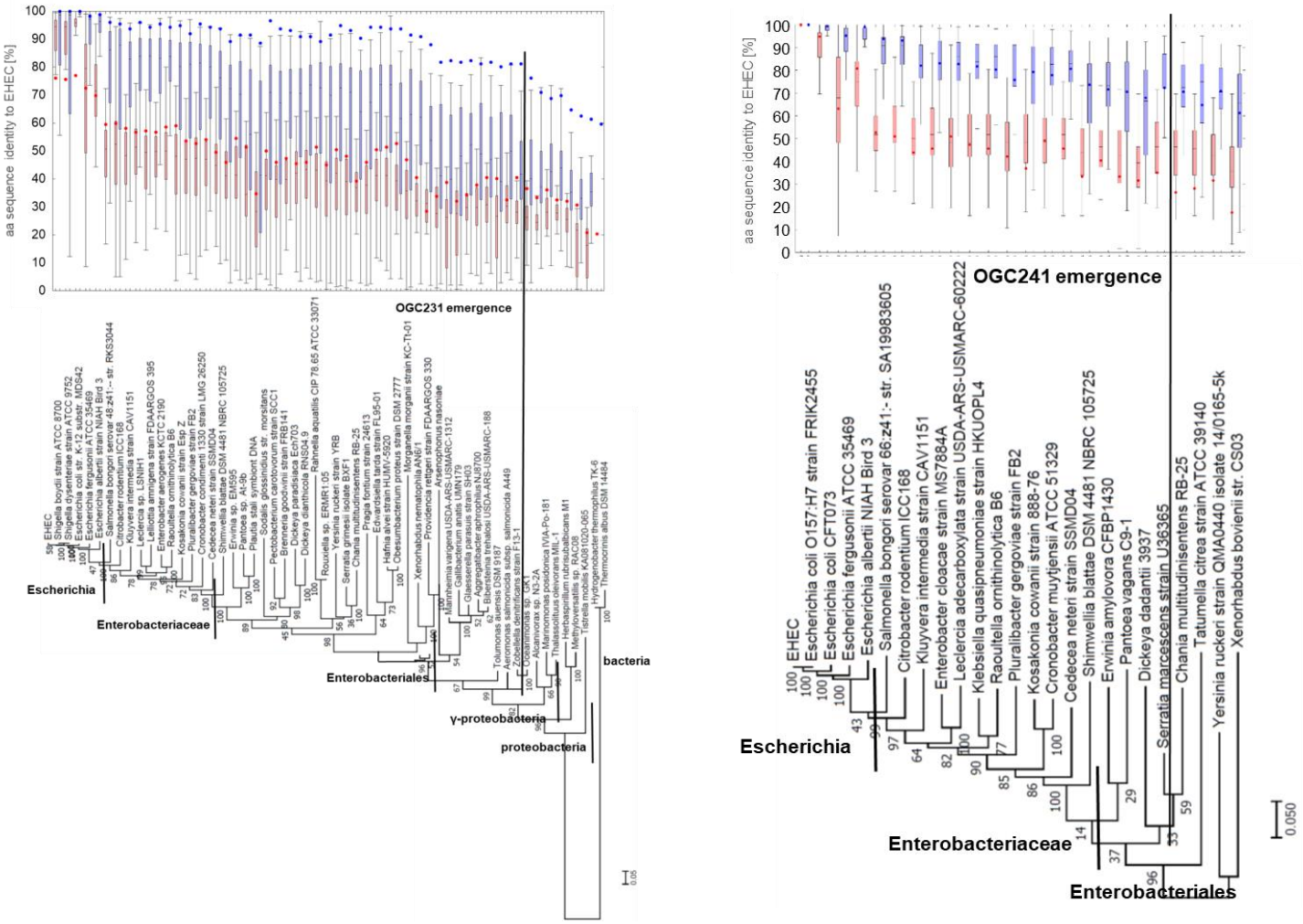
Supplementary figure S15.8: Amino acid (aa) sequence identities of the overlapping gene pairs OGC174/ EDL933_ 4292 and OGC194/ EDL933_4769 and of a negative control. The negative control was calculated from 100 randomly selected genes and an overlapping sequence in the organisms in which an overlapping gene pair homologue was found. The overlapping sequence of the negative control is a random sequence of the length of the sORF, embedded in the selected random gene and in the reading frame in which the sORF is located.



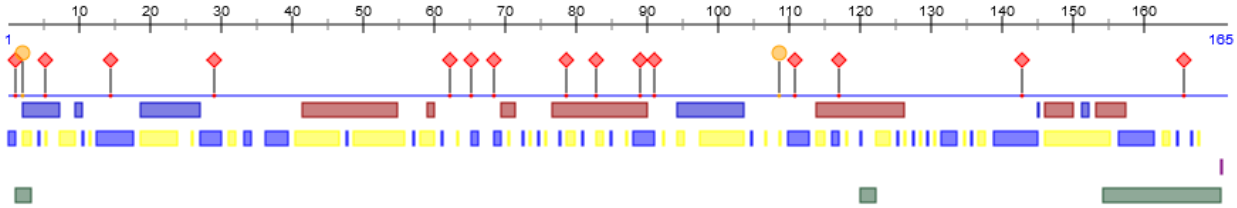
Supplementary figure S15.9: Amino acid (aa) sequence identities of the overlapping gene pair OGC198/ EDL933_ 4794 and of a negative control. The negative control was calculated from 100 randomly selected genes and an overlapping sequence in the organisms in which an overlapping gene pair homologue was found. The overlapping sequence of the negative control is a random sequence of the length of the sORF, embedded in the selected random gene and in the reading frame in which the sORF is located.



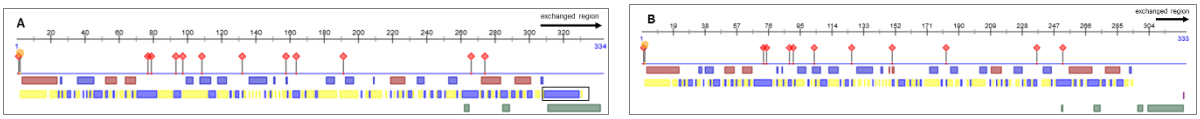
Supplementary figure S15.10: Amino acid (aa) sequence identities of the overlapping gene pair OGC226/ EDL933_5520 and of a negative control. The negative control was calculated from 100 randomly selected genes and an overlapping sequence in the organisms in which an overlapping gene pair homologue was found. The overlapping sequence of the negative control is a random sequence of the length of the sORF, embedded in the selected random gene and in the reading frame in which the sORF is located.



Supplementary figure S15.11: Amino acid (aa) sequence identities of the overlapping gene pairs OGC231/ EDL933_ 5573 and OGC241/ EDL933_ 5740 and of a negative control. The negative control was calculated from 100 randomly selected genes and an overlapping sequence in the organisms in which an overlapping gene pair homologue was found. The overlapping sequence of the negative control is a random sequence of the length of the sORF, embedded in the selected random gene and in the reading frame in which the sORF is located.



Supplementary figure S16: Protein structure of the MECP synthase originated from Tse-tse fly (UniProt A0A1A9Z0V2) determined with PredictProtein.



Supplementary figure S17: PredictProtein results showing pattern of protein features of ECs2385 (A) and its homologue in *E. fergusonii* (B). The last 28 amino acids are substituted leading to a partially substituted *ano* homologue (black arrow). Changes in amino acid composition are bordered by a black box.

Supplementary table S1: List of all 2,180 shadow ORFs (sORFs) with blastp hit and the respective mother gene.

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#52694	3382405	3383607	1203	227	hypothetical protein, partial [Klebsiella pneumoniae]	1,00E-36	-3	#3618	3381681	3382631	951	Transaldolase
#35101	5102250	5103068	819	112	hypothetical protein, partial [Cronobacter sakazakii]	5,00E-26	-3	#5337	5101072	5102361	1290	Phosphoribosylamine--glycine ligase
#46666	4290227	4290811	585	509	hypothetical protein [Escherichia coli]	7,00E-108	-3	#4556	4288822	4290735	1914	Glutathione-regulated potassium-efflux system ATP-binding protein
#28165	4090028	4090606	579	579	hypothetical protein [Escherichia coli]	2,00E-19	-3	#4336	4089708	4090916	1209	Acetate kinase
#65640	1524764	1525342	579	579	hypothetical protein [Escherichia coli]	7,00E-23	-3	#1574	1524226	1525458	1233	Co-activator of prophage gene expression IbrA
#68643	1124620	1125198	579	579	hypothetical protein [Escherichia coli]	7,00E-23	-3	#1113	1124082	1125314	1233	Co-activator of prophage gene expression IbrA
#64727	1642182	1642682	501	147	hypothetical protein [Escherichia coli]	5,00E-116	-3	#1723	1642223	1642369	147	hypothetical protein
#67649	1256691	1257182	492	317	hypothetical protein [Yokenella regensburgei]	8,00E-23	-3	#1253	1256450	1257007	558	hypothetical protein
#4685	724539	725000	462	112	hypothetical protein [Escherichia coli]	1,00E-107	-3	#0682	724462	724650	189	hypothetical protein
#76497	5327	5785	459	221	4'-phosphopantetheinyl transferase [Shigella dysenteriae]	5,00E-04	-3	#0005	5251	5547	297	hypothetical protein
#44284	4641176	4641631	456	125	hypothetical protein [Escherichia coli]	9,00E-18	-3	#4894	4640089	4641300	1212	3-deoxy-D-manno-octulosonic-acid transferase
#20708	3049878	3050324	447	447	hypothetical protein [Escherichia coli]	6,00E-08	-3	#3298	3049657	3050358	702	Fumarylacetoacetase
#39433	5352322	5352756	435	260	hypothetical protein [Escherichia coli]	2,00E-58	-3	#5570	5352294	5352581	288	UPF0131 protein YtfP
#62710	1911516	1911947	432	110	hypothetical protein, partial [Escherichia coli]	2,00E-10	-3	#2055	1911056	1911625	570	hypothetical protein
#64968	1610220	1610648	429	429	hypothetical protein, partial [Escherichia coli]	1,00E-23	-3	#1684	1610108	1610932	825	Thiamine kinase
#59154	2419868	2420278	411	120	prolyl-tRNA synthetase [Polaromonas naphthalenivorans]	4,00E-04	-3	#2604	2419879	2419998	120	hypothetical protein
#36795	5345722	5346114	393	130	hypothetical protein [Escherichia albertii]	1,00E-09	-3	#5565	5345714	5345851	138	hypothetical protein
#75414	166948	167334	387	387	hypothetical protein, partial [Staphylococcus aureus]	2,00E-48	-3	#0152	166383	168857	2475	ATP-dependent helicase HrpB
#25122	3660988	3661368	381	172	hypothetical protein [Shigella flexneri]	2,00E-67	-3	#3893	3660548	3661159	612	Formate hydrogenlyase subunit 2
#34642	5032706	5033083	378	378	Lom family protein [Escherichia coli]	8,00E-07	-3	#5281	5032440	5033543	1104	Glycerol dehydrogenase
#73665	410805	411179	375	203	hypothetical protein [Shigella flexneri]	7,00E-04	-3	#0393	409121	411007	1887	Propionate--CoA ligase
#13400	2046719	2047084	366	366	hypothetical protein [Escherichia coli]	3,00E-09	-3	#2189	2045670	2049872	4203	core protein
#37205	5408203	5408565	363	340	hypothetical protein [Escherichia coli]	4,00E-26	-3	#5629	5407691	5408542	852	hypothetical protein
#69828	961391	961753	363	257	hypothetical protein [Escherichia coli]	1,00E-13	-3	#0924	960562	961647	1086	Malate dehydrogenase
#75404	167887	168240	354	354	hypothetical protein, partial [Escherichia coli]	1,00E-28	-3	#0152	166383	168857	2475	ATP-dependent helicase HrpB
#64223	1705489	1705839	351	351	hypothetical protein [Escherichia coli]	1,00E-14	-3	#1807	1705260	1706084	825	Origin specific replication initiation factor
#16998	2537822	2538166	345	345	hypothetical protein [Escherichia coli]	2,00E-40	-3	#2722	2537805	2538845	1041	hypothetical protein
#37939	5521220	5521564	345	170	hypothetical protein [Escherichia albertii]	3,00E-14	-3	#5721	5521395	5522177	783	radical activating enzyme
#71252	757822	758163	342	316	hypothetical protein [Escherichia coli]	3,00E-24	-3	#0720	757848	758555	708	hypothetical protein
#63492	1813922	1814260	339	95	hypothetical protein [Escherichia coli]	8,00E-11	-3	#1931	1810273	1814016	3744	Respiratory nitrate reductase alpha chain
#74699	265652	265990	339	339	hypothetical protein, partial [Escherichia coli]	4,00E-09	-3	#0239	264862	267003	2142	VgrG protein
#16452	2461277	2461612	336	336	hypothetical protein [Escherichia coli]	2,00E-14	-3	#2644	2459802	2462858	3057	Fe-S protein, lactate dehydrogenase
#69616	993979	994311	333	333	hypothetical protein [Escherichia coli]	5,00E-13	-3	#0955	993516	995054	1539	Dipeptide-binding ABC transporter, periplasmic substrate-binding component
#64399	1682161	1682487	327	319	hypothetical protein [Escherichia coli]	2,00E-70	-3	#1775	1682169	1682912	744	hypothetical protein
#46292	4342627	4342944	318	140	histidine kinase, partial [Escherichia coli]	4,00E-17	-3	#4611	4342530	4342766	237	Ferrous iron-sensisng transcriptional regulator FeoC
#37825	5504880	5505191	312	312	hypothetical protein, partial [Escherichia coli]	3,00E-09	-3	#5702	5504380	5506671	2292	Phosphoglycerol transferase I

sORF with blastp hit							mother gene					
ID CP00895 7.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	rea-ding frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#47840	4112470	4112781	312	312	hypothetical protein [Escherichia coli]	5,00E-12	-3	#4360	4112310	4113170	861	Tagatose 1,6-bisphosphate aldolase
#31846	4624109	4624417	309	309	hypothetical protein, partial [Escherichia coli]	9,00E-29	-3	#4879	4623522	4624538	1017	Beta-1,3-galactosyltransferase
#73806	391617	391922	306	107	hypothetical protein [Escherichia albertii]	5,00E-04	-3	#0374	391538	391723	186	hypothetical protein
#59073	2428984	2429286	303	303	hypothetical protein [Rhodococcus qingshengii]	3,00E-04	-3	#2612	2428623	2429438	816	Putative lipoprotein
#12257	1855181	1855480	300	100	hypothetical protein [Plesiomonas shigelloides]	1,00E-04	-3	#1979	1855149	1855280	132	hypothetical protein
#17676	2635607	2635906	300	300	glycerol-3-phosphate acyltransferase [Escherichia coli]	2,00E-07	-3	#2829	2635515	2636486	972	Lipid A biosynthesis (KDO) 2-(lauroyl)-lipid IVA acyltransferase
#25226	3677395	3677694	300	274	hypothetical protein [Escherichia coli]	3,00E-18	-3	#3913	3677042	3677668	627	Protein-L-isoaspartate O-methyltransferase
#21730	3192144	3192440	297	297	hypothetical protein, partial [Bacillus cereus]	4,00E-37	-3	#3426	3191167	3192522	1356	O-succinylbenzoic acid--CoA ligase
#74126	347170	347466	297	297	hypothetical protein, partial [Escherichia coli]	9,00E-26	-3	#0332	346224	347762	1539	IS66 transposase
#57944	2594821	2595114	294	294	hypothetical protein, partial [Lactobacillus vaginalis]	3,00E-40	-3	#2785	2594718	2596316	1599	Rtn protein
#30132	4373099	4373389	291	291	hypothetical protein [Escherichia coli]	1,00E-49	-3	#4637	4372773	4374746	1974	Glycogen debranching enzyme
#22115	3244208	3244495	288	116	hypothetical protein [Escherichia albertii]	2,00E-07	-3	#3480	3244380	3245042	663	DedD protein
#40827	5141918	5142205	288	119	hypothetical protein [Escherichia albertii]	9,00E-59	-3	#5365	5141875	5142036	162	hypothetical protein
#29305	4255272	4255556	285	285	hypothetical protein [Escherichia coli]	5,00E-24	-3	#4503	4255075	4256199	1125	Rossmann fold nucleotide-binding protein Smf possibly involved in DNA uptake
#558	82714	82998	285	285	hypothetical protein, partial [Escherichia coli]	4,00E-05	-3	#0072	81470	83128	1659	SgrR, sugar-phosphate stress, transcriptional activator of SgrS small RNA
#62406	1948438	1948722	285	285	hypothetical protein [Escherichia coli]	4,00E-61	-3	#2101	1948419	1948814	396	Rhodanese-related sulfurtransferase
#63225	1850680	1850964	285	285	hypothetical protein, partial [Escherichia coli]	9,00E-12	-3	#1973	1850283	1851323	1041	Primosomal protein I
#42685	4883591	4883872	282	280	hypothetical protein [Escherichia coli]	2,00E-10	-3	#5136	4883593	4884543	951	Magnesium and cobalt transport protein CorA
#53470	3262428	3262706	279	98	hypothetical protein [Escherichia coli]	3,00E-04	-3	#3499	3261974	3262525	552	hypothetical protein
#58267	2549887	2550162	276	133	MFS transporter [Budvicia aquatica]	1,00E-10	-3	#2733	2550030	2550173	144	hypothetical protein
#14742	2224178	2224450	273	178	hypothetical protein [Escherichia coli]	5,00E-56	-3	#2381	2223075	2224355	1281	Gamma-glutamyl-putrescine oxidase
#788	119731	120003	273	273	hypothetical protein [Escherichia coli]	2,00E-19	-3	#0107	119012	120214	1203	Type IV fimbrial assembly protein PilC
#4359	674107	674376	270	270	hypothetical protein, partial [Escherichia coli]	1,00E-04	-3	#0632	673265	675166	1902	VgrG protein
#49207	3915511	3915780	270	113	membrane protein [Kluyvera ascorbata]	1,00E-21	-3	#4151	3915498	3915623	126	hypothetical protein
#68225	1181145	1181414	270	121	hypothetical protein, partial [Escherichia coli]	7,00E-11	-3	#1176	1181294	1183558	2265	DNA internalization-related competence protein ComEC/Rec2
#5175	787536	787802	267	173	hypothetical protein, partial [Escherichia albertii]	1,00E-08	-3	#0751	787630	787824	195	phosphopantetheinyltransferase component of enterobactin synthase multienzyme complex
#74629	274228	274488	261	261	hypothetical protein [Escherichia coli]	3,00E-06	-3	#0244	273792	275552	1761	core protein
#58449	2526862	2527116	255	255	hypothetical protein, partial [Escherichia coli]	4,00E-13	-3	#2709	2526402	2527208	807	Exodeoxyribonuclease III
#31434	4558959	4559210	252	154	hypothetical protein, partial [Bacillus cereus]	7,00E-36	-3	#4824	4558201	4559112	912	Glycyl-tRNA synthetase alpha chain
#69892	954482	954733	252	144	hypothetical protein, partial [Escherichia coli]	5,00E-05	-3	#0919	954529	954672	144	hypothetical protein
#13398	2046431	2046679	249	249	hypothetical protein, partial [Escherichia coli]	4,00E-15	-3	#2189	2045670	2049872	4203	core protein
#49412	3883510	3883758	249	123	hypothetical protein [Raoultella ornithinolytica]	3,00E-04	-3	#4116	3883587	3883709	123	hypothetical protein
#72175	619941	620189	249	249	hypothetical protein [Escherichia coli]	3,00E-28	-3	#0585	616748	620944	4197	core protein
#13579	2071570	2071815	246	246	hypothetical protein, partial [Bacillus cereus]	3,00E-09	-3	#2216	2071280	2073283	2004	putative collagenase
#65404	1553194	1553439	246	208	deoxyuridine 5'-triphosphate nucleotidohydrolase, partial [Escherichia coli]	3,00E-06	-3	#1621	1553232	1554653	1422	Cardiolipin synthetase
#41711	5019369	5019611	243	243	hypothetical protein, partial [Escherichia coli]	6,00E-08	-3	#5269	5016503	5020687	4185	core protein
#44584	4595374	4595616	243	243	hypothetical protein, partial [Escherichia coli]	6,00E-08	-3	#4854	4592508	4596737	4230	core protein
#70903	810652	810894	243	243	hypothetical protein, partial [Escherichia coli]	6,00E-08	-3	#0770	807786	811985	4200	core protein
#18112	2688357	2688596	240	235	hypothetical protein, partial [Escherichia coli]	2,00E-40	-3	#2886	2687590	2688591	1002	Integrase

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#70573	858907	859146	240	112	hypothetical protein, partial [Escherichia coli]	5,00E-42	-3	#0816	859035	859439	405	4-hydroxybenzoyl-CoA thioesterase family protein
#22336	3271658	3271894	237	237	hypothetical protein, partial [Escherichia coli]	3,00E-40	-3	#3509	3270435	3272579	2145	Enoyl-CoA hydratase
#29970	4351683	4351919	237	237	hypothetical protein [Escherichia coli]	1,00E-10	-3	#4619	4349815	4352208	2394	Maltodextrin phosphorylase
#13900	2111183	2111416	234	234	hypothetical protein [Escherichia coli]	5,00E-04	-3	#2247	2108631	2111666	3036	porin, autotransporter (AT) family
#27343	3974706	3974939	234	104	hypothetical protein [Escherichia coli]	1,00E-23	-3	#4222	3974836	3975723	888	Dienelactone hydrolase family
#4086	635654	635887	234	234	hypothetical protein [Escherichia coli]	2,00E-15	-3	#0600	635460	636245	786	Ureidoglycine aminohydrolase
#4694	725490	725723	234	234	hypothetical protein, partial [Escherichia coli]	1,00E-04	-3	#0684	725329	726135	807	Ribonuclease I precursor
#19397	2868678	2868908	231	231	hypothetical protein [Escherichia coli]	3,00E-47	-3	#3106	2868052	2869170	1119	GDP-mannose 4,6-dehydratase
#37492	5459402	5459632	231	231	hypothetical protein [Escherichia coli]	3,00E-20	-3	#5661	5459196	5460026	831	Uncharacterized protein YjiC
#22733	3327541	3327768	228	228	hypothetical protein [Escherichia coli]	4,00E-07	-3	#3560	3326606	3327781	1176	Manganese transport protein MntH
#67900	1221345	1221572	228	116	hypothetical protein [Cronobacter sakazakii]	4,00E-09	-3	#1210	1221338	1221460	123	putative fimbrial chaperone ycbF precursor
#27436	3987086	3987310	225	225	hypothetical protein, partial [Escherichia coli]	4,00E-11	-3	#4239	3986904	3989123	2220	putative Fe-S oxidoreductase family 2
#3695	565509	565730	222	222	hypothetical protein [Escherichia coli]	7,00E-31	-3	#0552	564805	565764	960	Acetyl esterase
#65756	1511495	1511716	222	95	hypothetical protein [Escherichia coli]	5,00E-14	-3	#1554	1511434	1511589	156	hypothetical protein
#68759	1111351	1111572	222	95	hypothetical protein [Escherichia coli]	5,00E-14	-3	#1093	1111290	1111445	156	hypothetical protein
#16356	2448481	2448699	219	169	hypothetical protein, partial [Escherichia coli]	1,00E-30	-3	#2630	2448023	2448649	627	putative oxidoreductase, Fe-S subunit
#42054	4966247	4966465	219	167	hypothetical protein [Escherichia albertii]	9,00E-11	-3	#5214	4966231	4966413	183	hypothetical protein
#941	143052	143270	219	133	hypothetical protein [Escherichia albertii]	2,00E-13	-3	#0126	143047	143184	138	phosphopantetheinyltransferase component of enterobactin synthase multienzyme complex
#34705	5042922	5043137	216	98	adhesin [Escherichia coli]	1,00E-05	-3	#5291	5043040	5044773	1734	UPF0141 membrane protein YijP possibly required for phosphoethanolamine modification of lipopolysaccharide
#68422	1152972	1153187	216	216	hypothetical protein [Escherichia coli]	3,00E-06	-3	#1154	1151633	1155661	4029	Cell division protein FtsK
#74999	219218	219433	216	95	hypothetical protein [Escherichia coli]	3,00E-41	-3	#0198	218602	219312	711	Copper homeostasis protein CutF precursor
#42323	4931070	4931282	213	213	hypothetical protein [Escherichia coli]	5,00E-10	-3	#5180	4930544	4933330	2787	DNA polymerase I
#29776	4321757	4321966	210	188	hypothetical protein [Escherichia coli]	5,00E-04	-3	#4594	4321779	4322573	795	Type IV pilus biogenesis protein PilM
#36275	5266484	5266693	210	142	hypothetical protein [Escherichia coli]	2,00E-09	-3	#5482	5266050	5266625	576	Transcriptional regulator, TetR family
#46303	4341248	4341457	210	210	hypothetical protein [Escherichia coli]	3,00E-32	-3	#4610	4340209	4342530	2322	Ferrous iron transport protein B
#42781	4872513	4872719	207	207	hypothetical protein, partial [Vibrio parahaemolyticus]	2,00E-04	-3	#5124	4872437	4874983	2547	Adenylate cyclase
#73226	470254	470460	207	207	hypothetical protein [Escherichia coli]	2,00E-24	-3	#0452	468498	470663	2166	Protein YkiA
#64800	1634258	1634461	204	204	hypothetical protein, partial [Escherichia coli]	7,00E-38	-3	#1708	1633807	1634844	1038	Primosomal protein I
#43474	4770155	4770355	201	142	hypothetical protein [Escherichia coli]	1,00E-27	-3	#5037	4770214	4770963	750	hypothetical protein
#54134	3161143	3161343	201	110	hypothetical protein, partial [Escherichia coli]	3,00E-20	-3	#3395	3158967	3161252	2286	Ribonucleotide reductase of class Ia (aerobic), alpha subunit
#72189	618135	618335	201	201	hypothetical protein [Escherichia coli]	6,00E-37	-3	#0585	616748	620944	4197	core protein
#22778	3333085	3333282	198	198	hypothetical protein [Escherichia coli]	2,00E-24	-3	#3565	3332918	3334066	1149	putative virulence protein
#72177	619536	619733	198	198	hypothetical protein [Escherichia coli]	2,00E-09	-3	#0585	616748	620944	4197	core protein
#8630	1308761	1308958	198	198	hypothetical protein [Escherichia coli]	9,00E-26	-3	#1319	1308633	1310813	2181	Tyrosine-protein kinase Wzc
#15744	2361268	2361462	195	195	hypothetical protein [Escherichia coli]	8,00E-24	-3	#2544	2360786	2361481	696	Dethiobiotin synthetase
#36428	5288736	5288930	195	195	hypothetical protein [Bacillus cereus]	8,00E-36	-3	#5507	5288638	5291961	3324	Potassium efflux system KefA protein
#40456	5195760	5195951	192	192	hypothetical protein, partial [Escherichia coli]	3,00E-11	-3	#5416	5195558	5196154	597	Cytochrome c-type heme lyase subunit nrfG, nitrite reductase complex assembly
#41387	5062253	5062444	192	122	hypothetical protein, partial [Salmonella enterica]	6,00E-25	-3	#5304	5060530	5062374	1845	Outer membrane vitamin B12 receptor BtuB
#41635	5029073	5029264	192	192	hypothetical protein, partial [Escherichia coli]	2,00E-12	-3	#5278	5028343	5030523	2181	Catalase

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#70074	927212	927403	192	151	hypothetical protein [Escherichia coli]	6,00E-29	-3	#0892	927253	928134	882	hypothetical protein
#52521	3405336	3405521	186	186	hypothetical protein, partial [Escherichia coli]	3,00E-09	-3	#3637	3404804	3406822	2019	Hydrogenase-4 component B
#39735	5309511	5309687	177	177	hypothetical protein [Rhodococcus qingshengii]	5,00E-15	-3	#5524	5309246	5311687	2442	3'-to-5' exoribonuclease RNase R
#8238	1249869	1250045	177	110	hypothetical protein [Escherichia coli]	2,00E-10	-3	#1241	1249936	1252407	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#10580	1607886	1608059	174	174	ABC transporter ATP-binding protein, partial [Escherichia coli]	2,00E-05	-3	#1679	1606192	1608381	2190	Putative OMR family iron-siderophore receptor precursor
#33970	4935445	4935618	174	113	high mobility group protein Z [Photorhabdus luminescens]	4,00E-04	-3	#5185	4935506	4935622	117	hypothetical protein
#49632	3849223	3849393	171	171	hypothetical protein, partial [Escherichia coli]	3,00E-04	-3	#4082	3847092	3849962	2871	putative hypoxanthine oxidase XdhD
#6865	1041518	1041688	171	171	hypothetical protein [Escherichia coli]	7,00E-05	-3	#1005	1040214	1041932	1719	Pyruvate oxidase [ubiquinone, cytochrome]
#15102	2272059	2272226	168	97	hypothetical protein [Shigella flexneri]	1,00E-16	-3	#2426	2271274	2272155	882	putative metal-dependent phosphoesterases (PHP family)
#48835	3967479	3967646	168	95	hypothetical protein, partial [Escherichia coli]	2,00E-06	-3	#4211	3966707	3967573	867	putative GST-like protein yghU associated with glutathionylspermidine synthetase/amidase
#14354	2172984	2173145	162	123	MULTISPECIES: restriction endonuclease [Enterobacteriaceae]	9,00E-25	-3	#2325	2173012	2173134	123	hypothetical protein
#32382	4706541	4706702	162	162	hypothetical protein [Citrobacter rodentium]	3,00E-13	-3	#4972	4706254	4706853	600	Orf4
#54548	3096595	3096756	162	162	hypothetical protein [Escherichia coli]	8,00E-08	-3	#3344	3095613	3097427	1815	ABC transporter, periplasmic substrate-binding protein
#21491	3157047	3157205	159	159	hypothetical protein [Escherichia coli]	1,00E-07	-3	#3394	3154519	3158223	3705	Type V secretory pathway, adhesin AidA
#37389	5442314	5442472	159	159	nitrite extrusion protein 2 [Escherichia coli]	2,00E-06	-3	#5643	5441451	5442557	1107	Sialic acid-induced transmembrane protein YjhT(NanM), possible mutarotase
#74627	274528	274686	159	159	hypothetical protein [Escherichia coli]	8,00E-16	-3	#0244	273792	275552	1761	core protein
#29358	4262245	4262400	156	146	hypothetical protein [Escherichia coli]	3,00E-07	-3	#4511	4262255	4262638	384	LSU ribosomal protein L17p
#10585	1608467	1608619	153	136	hypothetical protein [Escherichia coli]	9,00E-27	-3	#1680	1608414	1608602	189	hypothetical protein
#13982	2124851	2125003	153	97	hypothetical protein, partial [Escherichia coli]	3,00E-27	-3	#2260	2124381	2124947	567	hypothetical protein
#26984	3923110	3923262	153	115	hypothetical protein, partial [Escherichia coli]	1,00E-04	-3	#4161	3923093	3923224	132	hypothetical protein
#64632	1653062	1653214	153	153	hypothetical protein [Escherichia coli]	2,00E-07	-3	#1741	1652149	1654086	1938	hypothetical protein
#25079	3654227	3654376	150	127	hypothetical protein, partial [Escherichia coli]	2,00E-08	-3	#3886	3653883	3654353	471	Coenzyme F420 hydrogenase maturation protease
#29615	4299296	4299445	150	150	hypothetical protein, partial [Escherichia coli]	9,00E-10	-3	#4567	4299258	4299830	573	Peptidyl-prolyl cis-trans isomerase PpiA precursor
#30970	4498258	4498407	150	150	hypothetical protein [Escherichia coli]	4,00E-17	-3	#4770	4497551	4498948	1398	Cytochrome c551 peroxidase
#19543	2889970	2890116	147	147	hypothetical protein [Escherichia coli]	1,00E-04	-3	#3125	2889905	2890870	966	GDP-L-fucose synthetase
#8729	1320791	1320937	147	147	hypothetical protein [Escherichia coli]	2,00E-08	-3	#1331	1319259	1321973	2715	Sensor protein torS
#57043	2729894	2730037	144	128	hypothetical protein, partial [Escherichia coli]	3,00E-05	-3	#2942	2729722	2730021	300	hypothetical protein
#57612	2640308	2640451	144	144	aldehyde dehydrogenase, partial [Escherichia coli]	4,00E-10	-3	#2833	2639707	2640492	786	Zinc ABC transporter, inner membrane permease protein ZnuB
#59133	2422001	2422144	144	121	hypothetical protein [Escherichia coli]	5,00E-16	-3	#2608	2422024	2422431	408	Lactoylglutathione lyase
#56801	2764765	2764905	141	141	molecular chaperone Tir [Escherichia coli]	3,00E-23	-3	#2986	2764317	2765027	711	hypothetical protein
#56803	2764624	2764764	141	141	molecular chaperone Tir [Escherichia coli]	5,00E-23	-3	#2986	2764317	2765027	711	hypothetical protein
#56805	2764483	2764623	141	141	molecular chaperone Tir [Escherichia coli]	2,00E-22	-3	#2986	2764317	2765027	711	hypothetical protein
#56807	2764342	2764482	141	141	molecular chaperone Tir [Escherichia coli]	2,00E-22	-3	#2986	2764317	2765027	711	hypothetical protein
#8548	1296704	1296844	141	141	molecular chaperone Tir [Escherichia coli]	3,00E-23	-3	#1308	1296582	1297334	753	hypothetical protein
#8550	1296845	1296985	141	141	molecular chaperone Tir [Escherichia coli]	3,00E-23	-3	#1308	1296582	1297334	753	hypothetical protein
#8552	1296986	1297126	141	141	molecular chaperone Tir [Escherichia coli]	3,00E-23	-3	#1308	1296582	1297334	753	hypothetical protein
#8554	1297127	1297267	141	141	molecular chaperone Tir [Escherichia coli]	5,00E-23	-3	#1308	1296582	1297334	753	hypothetical protein
#18937	2799200	2799337	138	138	hypothetical protein, partial [Escherichia coli]	3,00E-15	-3	#3038	2798817	2799854	1038	Primosomal protein I
#45349	4480928	4481065	138	138	transcriptional regulator [Escherichia coli]	9,00E-05	-3	#4753	4480687	4482024	1338	Radical SAM family protein HutW, like coproporphyrinogen III oxidase, oxygen-independent, associated with heme uptake

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#47313	4190978	4191115	138	138	hypothetical protein [Shigella dysenteriae]	2,00E-23	-3	#4442	4190779	4191828	1050	Putative uncharacterized protein YhcG
#51757	3526309	3526446	138	136	hypothetical protein [Escherichia coli]	1,00E-10	-3	#3748	3526311	3527009	699	hypothetical protein
#15223	2291552	2291686	135	135	hypothetical protein [Escherichia coli]	2,00E-13	-3	#2449	2290791	2291702	912	hypothetical protein
#33036	4795210	4795344	135	135	hypothetical protein [Escherichia coli]	3,00E-06	-3	#5060	4794899	4796269	1371	N-acetylglucosamine-1-phosphate uridyltransferase
#9297	1405898	1406032	135	135	proline dehydrogenase [Escherichia coli]	2,00E-17	-3	#1435	1402167	1406129	3963	Transcriptional repressor of PutA and PutP
#13412	2048354	2048485	132	132	hypothetical protein [Escherichia coli]	3,00E-20	-3	#2189	2045670	2049872	4203	core protein
#34370	4996131	4996262	132	132	hypothetical protein [Escherichia albertii]	1,00E-10	-3	#5245	4996126	4996269	144	hypothetical protein
#70688	839862	839993	132	132	glycine dehydrogenase [Escherichia coli]	1,00E-08	-3	#0797	839255	840421	1167	Succinyl-CoA ligase [ADP-forming] beta chain
#5113	778253	778381	129	129	transcriptional regulator [Escherichia coli]	3,00E-04	-3	#0741	777453	779117	1665	Asparagine synthetase [glutamine-hydrolyzing]
#73051	493966	494094	129	129	hypothetical protein, partial [Bacillus cereus]	1,00E-22	-3	#0474	493134	494105	972	Protein-export membrane protein SecF
#27372	3978583	3978708	126	107	hypothetical protein, partial [Escherichia coli]	2,00E-13	-3	#4226	3978602	3978754	153	hypothetical protein
#18897	2794989	2795111	123	123	hypothetical protein [Escherichia coli]	1,00E-19	-3	#3029	2794789	2795148	360	hypothetical protein
#28848	4193487	4193609	123	123	hypothetical protein [Streptococcus anginosus]	2,00E-08	-3	#4445	4193221	4193910	690	N-acetylmannosamine-6-phosphate 2-epimerase
#60920	2170121	2170240	120	120	lysozyme [Escherichia coli]	8,00E-06	-3	#2323	2169538	2172138	2601	Exodeoxyribonuclease VIII
#6464	982272	982391	120	120	hypothetical protein [Escherichia coli]	8,00E-12	-3	#0944	981799	983064	1266	hypothetical protein
#68215	1182018	1182137	120	120	hypothetical protein, partial [Escherichia coli]	2,00E-16	-3	#1176	1181294	1183558	2265	DNA internalization-related competence protein ComEC/Rec2
#247	40533	40649	117	117	hypothetical protein, partial [Escherichia coli]	7,00E-07	-3	#0037	39784	40677	894	Carnitine racemase
#32381	4706403	4706519	117	117	hypothetical protein [Citrobacter rodentium]	1,00E-08	-3	#4972	4706254	4706853	600	Orf4
#56536	2805053	2805169	117	110	hypothetical protein [Escherichia coli]	3,00E-09	-3	#3045	2802691	2805162	2472	Exodeoxyribonuclease VIII
#18893	2794358	2794471	114	114	hypothetical protein, partial [Escherichia coli]	6,00E-07	-3	#3028	2793738	2794787	1050	Phage antitermination protein Q
#27298	3968668	3968781	114	112	hypothetical protein [Escherichia coli]	2,00E-06	-3	#4214	3968291	3968779	489	Hydrogenase-2 operon protein hybE
#27651	4016017	4016130	114	114	hypothetical protein [Escherichia coli]	3,00E-11	-3	#4264	4015340	4016155	816	Uncharacterized protein ygiD
#62102	1989354	1989467	114	114	hypothetical protein [Escherichia coli]	1,00E-07	-3	#2138	1988054	1990333	2280	Putative formate dehydrogenase oxidoreductase protein
#63181	1855798	1855911	114	114	hypothetical protein, partial [Escherichia coli]	3,00E-10	-3	#1980	1855482	1856528	1047	Putative cytoplasmic protein
#30946	4495008	4495118	111	111	hypothetical protein [Escherichia coli]	7,00E-11	-3	#4767	4494745	4495569	825	HTH-type transcriptional regulator gadX
#5290	803024	803134	111	111	hypothetical protein [Escherichia coli]	8,00E-07	-3	#0766	802611	803183	573	Potassium-transporting ATPase C chain
#58037	2581155	2581265	111	111	hypothetical protein [Escherichia coli]	4,00E-08	-3	#2770	2580956	2582080	1125	Putative dioxygenase, alpha subunit
#62835	1897883	1897993	111	111	hypothetical protein, partial [Escherichia coli]	4,00E-06	-3	#2031	1897432	1898475	1044	Primosomal protein I
#33169	4811795	4811902	108	108	hypothetical protein [Escherichia coli]	5,00E-05	-3	#5076	4810665	4812161	1497	Putative regulator protein
#17020	2540568	2540672	105	105	hypothetical protein, partial [Bacillus cereus]	8,00E-13	-3	#2723	2538973	2540934	1962	DNA topoisomerase III
#3744	573120	573224	105	105	hypothetical protein, partial [Escherichia coli]	6,00E-04	-3	#0558	573076	573870	795	hypothetical protein
#43833	4712600	4712704	105	105	hypothetical protein [Escherichia coli]	1,00E-05	-3	#4980	4712206	4713129	924	hypothetical protein
#1081	163625	163726	102	102	hypothetical protein [Escherichia coli]	2,00E-08	-3	#0148	163509	164435	927	glutamyl-Q-tRNA synthetase
#18561	2750789	2750890	102	102	hypothetical protein [Escherichia coli]	1,00E-06	-3	#2969	2749908	2751326	1419	DNA-cytosine methyltransferase
#68113	1192630	1192731	102	102	hypothetical protein, partial [Escherichia coli]	4,00E-07	-3	#1186	1192596	1193300	705	Chromosome partition protein MukE
#34094	4954748	4954846	99	99	hypothetical protein [Bacillus thuringiensis]	4,00E-05	-3	#5200	4954713	4955954	1242	Aldose-ketose isomerase YihS
#58412	2531271	2531369	99	99	hypothetical protein [Escherichia coli]	2,00E-08	-3	#2715	2531186	2532721	1536	ABC transporter, permease protein YnjC
#71948	650404	650499	96	96	amidohydrolase, partial [Escherichia coli]	2,00E-09	-3	#0615	650355	650897	543	type 1 fimbriae major subunit FimA
#30714	4459405	4459497	93	93	hypothetical protein [Escherichia coli]	2,00E-10	-3	#4729	4458317	4459519	1203	NAD(FAD)-utilizing dehydrogenases

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#33989	4939006	4939098	93	93	hypothetical protein, partial [Escherichia coli]	4,00E-09	-3	#5188	4938827	4939876	1050	Nitrogen regulation protein NtrB
#45653	4436182	4436274	93	93	hypothetical protein, partial [Escherichia coli]	6,00E-05	-3	#4707	4435896	4437470	1575	Nickel ABC transporter, periplasmic nickel-binding protein NikA
#57844	2606761	2606853	93	93	hypothetical protein [Escherichia coli]	3,00E-09	-3	#2801	2606568	2607941	1374	Putative transport protein
#45444	4467893	4469272	1380	271	hypothetical protein, partial [Lactobacillus johnsonii]	1,00E-33	-2	#4737	4467321	4468163	843	Protein involved in catabolism of external DNA
#76305	35081	36379	1299	94	hypothetical protein [Prevotella copri CAG:164]	2,00E-10	-2	#0032	34026	35174	1149	Carbamoyl-phosphate synthase small chain
#23856	3483235	3484110	876	876	hypothetical protein [Sorangium cellulosum]	5,00E-10	-2	#3710	3483219	3484217	999	putative sugar ABC transport system, permease protein YphD
#75642	130047	130892	846	124	hypothetical protein, partial [Burkholderia multivorans]	8,00E-08	-2	#0116	127507	130170	2664	Pyruvate dehydrogenase E1 component
#7161	1086168	1086959	792	149	hypothetical protein, partial [Niveispirillum irakense]	2,00E-05	-2	#1057	1086143	1086316	174	hypothetical protein
#9783	1486312	1487103	792	149	hypothetical protein, partial [Niveispirillum irakense]	2,00E-05	-2	#1518	1486287	1486460	174	hypothetical protein
#45627	4439177	4439959	783	716	hypothetical protein [Clostridium bolteae CAG:59]	0,001	-2	#4710	4439244	4440008	765	Nickel transport ATP-binding protein NikD
#34671	5036077	5036829	753	653	hypothetical protein [Escherichia coli]	5,00E-110	-2	#5283	5034228	5036729	2502	Phosphoenolpyruvate-protein phosphotransferase of PTS system
#75639	130442	131134	693	693	hypothetical protein [Bordetella bronchiseptica]	3,00E-05	-2	#0117	130185	132077	1893	Dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase complex
#51269	3601629	3602309	681	681	hypothetical protein, partial [Mycobacterium avium]	7,00E-07	-2	#3824	3601081	3602529	1449	Succinate-semialdehyde dehydrogenase [NADP+]
#58619	2500610	2501260	651	651	hypothetical protein [Nocardiopsis chromatogenes]	7,00E-04	-2	#2682	2500545	2501474	930	6-phosphofructokinase class II
#44501	4607194	4607835	642	642	DNA-directed RNA polymerase II [Burkholderia mallei]	1,00E-11	-2	#4866	4606733	4611499	4767	hypothetical protein
#19502	2883599	2884237	639	152	hypothetical protein [Azohydromonas australica]	2,00E-16	-2	#3118	2882272	2883750	1479	Lipopolysaccharide biosynthesis protein WzxC
#14229	2154535	2155161	627	168	hypothetical protein [Escherichia coli]	2,00E-89	-2	#2298	2154810	2154977	168	hypothetical protein
#76209	49727	50350	624	154	hypothetical protein, partial [Lactobacillus vaginalis]	2,00E-41	-2	#0044	48594	49880	1287	putative electron transfer flavoprotein-quinone oxidoreductase FixC
#5306	805204	805797	594	535	hypothetical protein, partial [Staphylococcus aureus]	1,00E-45	-2	#0768	805263	806936	1674	Potassium-transporting ATPase A chain
#6574	1000713	1001303	591	330	hypothetical protein [Escherichia coli]	9,00E-113	-2	#0960	1000838	1001167	330	DNA-binding protein
#13679	2084227	2084811	585	585	hypothetical protein [Rhizobium leguminosarum]	2,00E-08	-2	#2229	2084166	2085578	1413	Methyl-accepting chemotaxis protein III (ribose and galactose chemoreceptor protein)
#72359	593088	593672	585	585	hypothetical protein [Escherichia coli]	2,00E-110	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#17929	2667681	2668259	579	579	hypothetical protein, partial [Acidovorax avenae]	1,00E-19	-2	#2861	2667422	2669083	1662	Methyl-accepting chemotaxis protein II (aspartate chemoreceptor protein)
#6122	930211	930789	579	579	hypothetical protein [Dermatophilus congolensis]	5,00E-15	-2	#0895	930105	931337	1233	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase
#5299	804303	804869	567	567	hypothetical protein [Catenulispora acidiphila]	5,00E-44	-2	#0767	803192	805240	2049	Potassium-transporting ATPase B chain
#1662	252798	253352	555	555	hypothetical protein [Bacteroides clarus CAG:160]	1,00E-08	-2	#0225	251126	253897	2772	ClpB protein
#28508	4144362	4144913	552	552	hypothetical protein [Pyrococcus horikoshii]	5,00E-09	-2	#4397	4143467	4146139	2673	Translation initiation factor 2
#43057	4836476	4837024	549	549	hypothetical protein, partial [Burkholderia pseudomallei]	2,00E-24	-2	#5093	4836402	4837946	1545	Threonine dehydratase biosynthetic
#35324	5135140	5135685	546	221	hypothetical protein [Klebsiella oxytoca]	4,00E-20	-2	#5359	5134011	5135360	1350	Aspartokinase
#76444	13544	14089	546	546	hypothetical protein [Azospirillum brasilense]	3,00E-19	-2	#0013	12180	14096	1917	Chaperone protein DnaK
#38453	5499709	5500242	534	534	hypothetical protein, partial [Acidovorax avenae]	2,00E-20	-2	#5698	5498879	5500543	1665	Methyl-accepting chemotaxis protein I (serine chemoreceptor protein)
#75851	101701	102222	522	522	exodeoxyribonuclease V alpha chain [Salmonella enterica]	1,00E-05	-2	#0091	101693	103009	1317	UDP-N-acetylmuramoylalanine--D-glutamate ligase
#20726	3052107	3052622	516	516	hypothetical protein, partial [Bacillus cereus]	4,00E-04	-2	#3300	3051413	3052771	1359	4-hydroxybenzoate transporter
#22831	3340875	3341381	507	507	hypothetical protein [Bifidobacterium bifidum CAG:234]	2,00E-07	-2	#3574	3339431	3341446	2016	DNA ligase
#22243	3259982	3260485	504	504	hypothetical protein [Sulfolobus tokodaii]	4,00E-04	-2	#3497	3259756	3260841	1086	Chorismate synthase
#15556	2334024	2334524	501	168	hypothetical protein [Escherichia coli]	1,00E-110	-2	#2507	2334173	2334340	168	hypothetical protein
#18869	2791305	2791805	501	168	hypothetical protein [Escherichia coli]	3,00E-92	-2	#3024	2791454	2791621	168	hypothetical protein
#42325	4930771	4931262	492	492	hypothetical protein [Pseudomonas aeruginosa]	4,00E-28	-2	#5180	4930544	4933330	2787	DNA polymerase I
#65985	1477794	1478285	492	211	hypothetical protein, partial [Staphylococcus aureus]	3,00E-10	-2	#1503	1477702	1478004	303	Urease gamma subunit

sORF with blastp hit							mother gene					
ID CP00895 7.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	rea-ding frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#68988	1077650	1078141	492	211	hypothetical protein, partial [Staphylococcus aureus]	3,00E-10	-2	#1042	1077558	1077860	303	Urease gamma subunit
#55759	2922867	2923355	489	489	hypothetical protein [Akkermansia muciniphila CAG:154]	3,00E-04	-2	#3151	2922127	2925204	3078	Multidrug transporter MdtC
#52862	3356266	3356751	486	486	hypothetical protein [Streptomyces stelliscabiei]	9,00E-08	-2	#3591	3355949	3356845	897	N-acetylmuramic acid 6-phosphate etherase
#4543	701805	702287	483	483	hypothetical protein [Rhodococcus rhodnii]	9,00E-13	-2	#0659	701630	702622	993	Ferric enterobactin transport system permease protein FepG
#51630	3544110	3544592	483	342	hypothetical protein [Escherichia coli]	1,00E-22	-2	#3759	3544162	3544503	342	Ribosome hibernation protein YfiA
#11769	1785186	1785662	477	361	transcriptional regulator [Pantoea ananatis]	2,00E-05	-2	#1905	1785302	1786945	1644	Putative adhesion and penetration protein
#2798	426650	427126	477	383	hypothetical protein [Escherichia coli]	3,00E-28	-2	#0408	426187	427032	846	Mhp operon transcriptional activator
#48418	4031814	4032290	477	440	hypothetical protein [Neisseria elongata]	2,00E-14	-2	#4280	4031851	4032468	618	Acyl-phosphate:glycerol-3-phosphate O-acyltransferase PlsY
#27563	4003224	4003697	474	474	hypothetical protein [Klebsiella pneumoniae]	2,00E-11	-2	#4253	4002857	4003840	984	Putative iron compound permease protein of ABC transporter family
#27017	3927557	3928027	471	250	hypothetical protein [Escherichia albertii]	1,00E-05	-2	#4168	3927778	3928785	1008	Putative alpha helix chain
#24287	3540142	3540609	468	468	hypothetical protein [Bacteroides clarus CAG:160]	3,00E-10	-2	#3755	3538608	3541181	2574	ClpB protein
#50835	3662784	3663248	465	400	hypothetical protein [Saccharomonospora viridis]	1,00E-12	-2	#3897	3662311	3663183	873	[NiFe] hydrogenase nickel incorporation-associated protein HypB
#13718	2089210	2089665	456	456	hypothetical protein, partial [Mycobacterium avium]	2,00E-08	-2	#2233	2088606	2090045	1440	Aldehyde dehydrogenase A
#43487	4768388	4768843	456	456	hypothetical protein [Escherichia coli]	4,00E-04	-2	#5035	4767948	4769123	1176	Putative transport protein
#16638	2488942	2489385	444	386	hypothetical protein [Microlunatus phosphovorus]	2,00E-06	-2	#2670	2488347	2489327	981	Vitamin B12 ABC transporter, permease component BtuC
#72422	584508	584951	444	444	hypothetical protein [Escherichia coli]	1,00E-43	-2	#0565	581353	585738	4386	Large repetitive protein
#28385	4127122	4127562	441	340	hypothetical protein [Escherichia albertii]	8,00E-04	-2	#4378	4127223	4128263	1041	protein yraQ
#72498	574115	574555	441	235	hypothetical protein [Shigella flexneri]	2,00E-09	-2	#0559	573954	574349	396	HigA protein (antitoxin to HigB)
#17406	2596582	2597019	438	438	hypothetical protein, partial [Escherichia coli]	3,00E-05	-2	#2786	2596320	2597876	1557	Magnesium and cobalt efflux protein CorC
#65754	1511759	1512196	438	94	hypothetical protein [Escherichia coli]	9,00E-41	-2	#1555	1511655	1511852	198	hypothetical protein
#68757	1111615	1112052	438	94	hypothetical protein [Escherichia coli]	9,00E-41	-2	#1094	1111511	1111708	198	hypothetical protein
#13399	2046433	2046867	435	435	hypothetical protein [Escherichia coli]	1,00E-37	-2	#2189	2045670	2049872	4203	core protein
#35919	5221447	5221881	435	170	hypothetical protein [Escherichia coli]	1,00E-14	-2	#5442	5220936	5221616	681	Phosphonates transport ATP-binding protein PhnL
#72176	619753	620187	435	435	hypothetical protein [Escherichia coli]	2,00E-46	-2	#0585	616748	620944	4197	core protein
#20849	3069780	3070208	429	194	hypothetical protein, partial [Salmonella enterica]	2,00E-09	-2	#3319	3068090	3069973	1884	Colicin I receptor precursor
#5568	846022	846444	423	423	hypothetical protein [Burkholderia gladioli]	4,00E-12	-2	#0803	845268	846548	1281	Putative symport protein
#23965	3498168	3498587	420	188	hypothetical protein [Citrobacter koseri]	4,00E-32	-2	#3721	3496928	3498355	1428	Putative sensor-like histidine kinase YfhK
#4816	740982	741401	420	368	hypothetical protein [Methylobacterium radiotolerans]	1,00E-04	-2	#0700	740384	741349	966	Lipoate synthase
#74665	270090	270509	420	420	hypothetical protein [Escherichia coli]	3,00E-30	-2	#0240	267079	271293	4215	core protein
#76186	52282	52701	420	420	potassium transporter KefC, partial [Klebsiella pneumoniae]	2,00E-12	-2	#0048	52184	54046	1863	Glutathione-regulated potassium-efflux system protein KefC
#21067	3101838	3102251	414	414	hypothetical protein, partial [Escherichia coli]	1,00E-35	-2	#3349	3101819	3103009	1191	MFS family multidrug transport protein, bicyclomycin resistance protein
#58718	2482471	2482884	414	275	hypothetical protein [Escherichia coli]	1,00E-07	-2	#2663	2482610	2483656	1047	2-keto-3-deoxy-D-arabino-heptulosonate-7- phosphate synthase I alpha
#576	84757	85161	405	405	hypothetical protein [Alistipes finegoldii CAG:68]	4,00E-06	-2	#0074	84072	85472	1401	3-isopropylmalate dehydratase large subunit
#4360	673989	674390	402	402	hypothetical protein [Escherichia coli]	7,00E-18	-2	#0632	673265	675166	1902	VgrG protein
#72546	566497	566898	402	402	inosine/guanosine kinase, partial [Salmonella enterica]	6,00E-08	-2	#0553	565916	567220	1305	Inosine-guanosine kinase
#37844	5507073	5507471	399	350	hypothetical protein, partial [Escherichia coli]	2,00E-05	-2	#5703	5506925	5507422	498	Putative glycoprotein/receptor
#26109	3794601	3794996	396	396	hypothetical protein [Burkholderia gladioli]	4,00E-08	-2	#4025	3794438	3795619	1182	3-ketoacyl-CoA thiolase
#43529	4761308	4761703	396	396	hypothetical protein [Bacteroides coprocola CAG:162]	5,00E-10	-2	#5030	4760685	4762049	1365	GTPase and tRNA-U34 5-formylation enzyme TrmE
#53361	3279195	3279590	396	396	hypothetical protein [Salmonella enterica]	3,00E-63	-2	#3517	3278905	3279837	933	putative transport

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#58365	2536560	2536955	396	396	hypothetical protein [Streptomyces puniceus]	5,00E-11	-2	#2721	2536345	2537688	1344	NADP-specific glutamate dehydrogenase
#9224	1397192	1397587	396	396	FMN reductase [Klebsiella pneumoniae]	3,00E-04	-2	#1428	1397107	1397601	495	putative flavin reductase RutF in pyrimidine catabolism pathway
#71675	694833	695225	393	393	hypothetical protein [Stackebrandtia nassauensis]	7,00E-06	-2	#0654	694171	695373	1203	Enterobactin esterase
#17915	2665956	2666345	390	390	hypothetical protein, partial [Pimelobacter simplex]	6,00E-04	-2	#2860	2665775	2667376	1602	Methyl-accepting chemotaxis protein IV (dipeptide chemoreceptor protein)
#22825	3340071	3340460	390	390	hypothetical protein [Bifidobacterium bifidum CAG:234]	2,00E-10	-2	#3574	3339431	3341446	2016	DNA ligase
#12194	1846077	1846463	387	387	cell surface protein [Escherichia coli]	1,00E-28	-2	#1962	1844156	1846627	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#14306	2164996	2165382	387	140	hypothetical protein, partial [Escherichia coli]	5,00E-14	-2	#2312	2164389	2165135	747	DNA replication protein DnaC
#14824	2236822	2237208	387	387	hypothetical protein, partial [Escherichia coli]	2,00E-26	-2	#2395	2236686	2237849	1164	Putative membrane transport protein
#42665	4886279	4886665	387	117	hypothetical protein [Salmonella enterica]	6,00E-38	-2	#5140	4886334	4886450	117	hypothetical protein
#59703	2345990	2346376	387	387	cell surface protein [Escherichia coli]	7,00E-30	-2	#2528	2345826	2348297	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#65956	1481408	1481794	387	387	hypothetical protein [Delftia tsuruhatensis]	4,00E-07	-2	#1508	1481190	1481807	618	Urease accessory protein UreG
#68959	1081264	1081650	387	387	hypothetical protein [Delftia tsuruhatensis]	4,00E-07	-2	#1047	1081046	1081663	618	Urease accessory protein UreG
#46053	4378481	4378864	384	368	hypothetical protein [Vibrio parahaemolyticus]	5,00E-09	-2	#4640	4378497	4379090	594	Multiple antibiotic resistance protein marC
#20897	3077067	3077447	381	326	hypothetical protein [Psychromonas hadalis]	1,00E-05	-2	#3325	3076142	3077392	1251	putative pyrimidine nucleoside transport protein associated with pseudouridine catabolism
#34193	4970052	4970432	381	381	hypothetical protein [Escherichia coli]	4,00E-15	-2	#5218	4969034	4972084	3051	Formate dehydrogenase O alpha subunit, selenocysteine-containing
#40081	5255430	5255810	381	244	hypothetical protein, partial [Escherichia coli]	6,00E-04	-2	#5474	5255401	5255673	273	hypothetical protein YjdJ
#6857	1040131	1040511	381	298	hypothetical protein [Neorhizobium galegae]	1,00E-05	-2	#1005	1040214	1041932	1719	Pyruvate oxidase [ubiquinone, cytochrome]
#11825	1792687	1793064	378	257	electron transporter RsxA, partial [Escherichia coli]	7,00E-08	-2	#1911	1791996	1792943	948	Ribose-phosphate pyrophosphokinase
#49402	3884869	3885246	378	378	hypothetical protein [Desulfomicrobium baculatum]	3,00E-05	-2	#4119	3884855	3885748	894	Chromosome initiation inhibitor
#72392	588597	588974	378	378	hypothetical protein [Escherichia coli]	3,00E-21	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#28664	4165426	4165800	375	375	hypothetical protein, partial [Escherichia coli]	5,00E-10	-2	#4417	4165215	4166474	1260	UDP-N-acetylglucosamine 1-carboxyvinyltransferase
#72342	596001	596375	375	375	hypothetical protein [Escherichia coli]	3,00E-04	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#75282	185065	185439	375	356	hypothetical protein, partial [Vibrio cholerae]	4,00E-04	-2	#0166	185084	186508	1425	HtrA protease/chaperone protein
#38167	5538517	5538888	372	241	hypothetical protein [Escherichia coli]	3,00E-08	-2	#5737	5538431	5538757	327	Transcriptional repressor protein TrpR
#39624	5323967	5324338	372	306	hypothetical protein [Klebsiella pneumoniae]	3,00E-05	-2	#5539	5323986	5324291	306	Ascorbate-specific PTS system, EIIB component
#4556	703310	703681	372	314	hypothetical protein [Microlunatus phosphovorus]	8,00E-10	-2	#0660	702619	703623	1005	ABC-type Fe3+-siderophore transport system, permease component
#37996	5529388	5529756	369	369	hypothetical protein [Sodalis glossinidius]	3,00E-07	-2	#5730	5528772	5530460	1689	Lipoate-protein ligase A
#38276	5524321	5524689	369	369	hypothetical protein [Escherichia coli]	2,00E-23	-2	#5725	5524037	5524816	780	Deoxyribose-phosphate aldolase
#44778	4568317	4568682	366	366	hypothetical protein [Sorangium cellulosum]	2,00E-06	-2	#4833	4567514	4568695	1182	Xylose ABC transporter, permease protein XylH
#45174	4504949	4505314	366	366	hypothetical protein, partial [Mycobacterium avium]	2,00E-04	-2	#4776	4504548	4505870	1323	Inner membrane metabolite transport protein YhjE
#60462	2234815	2235180	366	347	hypothetical protein [Clostridium bolteae CAG:59]	7,00E-04	-2	#2393	2234834	2235826	993	Peptide transport system ATP-binding protein SapD
#32443	4715483	4715842	360	263	hypothetical protein [Escherichia coli]	8,00E-16	-2	#4983	4714507	4715745	1239	Putative transport protein
#3764	575315	575674	360	360	hypothetical protein [Bacteroides clarus CAG:160]	1,00E-11	-2	#0561	574564	577068	2505	Lead, cadmium, zinc and mercury transporting ATPase
#59823	2328170	2328526	357	158	hypothetical protein, partial [Escherichia coli]	4,00E-38	-2	#2497	2328369	2328596	228	hypothetical protein
#60208	2272503	2272859	357	357	hypothetical protein, partial [Vibrio parahaemolyticus]	1,00E-08	-2	#2427	2272429	2273991	1563	Anthranilate synthase, aminase component
#65589	1531446	1531802	357	357	hypothetical protein [Escherichia coli]	2,00E-28	-2	#1584	1530574	1532877	2304	Antigen 43 precursor
#68592	1131302	1131658	357	357	hypothetical protein [Escherichia coli]	2,00E-28	-2	#1123	1130430	1133111	2682	Antigen 43 precursor
#16290	2440061	2440414	354	239	hypothetical protein, partial [Escherichia coli]	2,00E-48	-2	#2622	2439043	2440299	1257	putative enzyme
#17927	2667294	2667647	354	226	hypothetical protein [Escherichia coli]	4,00E-21	-2	#2861	2667422	2669083	1662	Methyl-accepting chemotaxis protein II (aspartate chemoreceptor protein)

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#22819	3339272	3339622	351	158	hypothetical protein [Escherichia coli]	2,00E-11	-2	#3573	3339211	3339429	219	Putative cytoplasmic protein
#23693	3462282	3462632	351	197	hypothetical protein [Shigella flexneri]	1,00E-05	-2	#3684	3461702	3462478	777	Protein SseB
#5028	768331	768681	351	351	aminopeptidase, partial [Escherichia coli]	4,00E-05	-2	#0732	767976	768884	909	Glutamate Aspartate periplasmic binding protein precursor GltI
#56556	2802855	2803205	351	351	cell surface protein [Escherichia coli]	7,00E-33	-2	#3045	2802691	2805162	2472	Exodeoxyribonuclease VIII
#8258	1251893	1252243	351	351	cell surface protein [Escherichia coli]	7,00E-33	-2	#1241	1249936	1252407	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#9278	1403524	1403874	351	351	hypothetical protein, partial [Mycobacterium avium]	7,00E-12	-2	#1435	1402167	1406129	3963	Transcriptional repressor of PutA and PutP
#18059	2681453	2681800	348	290	hypothetical protein, partial [Escherichia coli]	7,00E-06	-2	#2874	2680753	2681742	990	L-arabinose-binding periplasmic protein precursor AraF
#38298	5521046	5521393	348	136	hypothetical protein [Escherichia coli]	8,00E-35	-2	#5720	5520399	5521181	783	Putative deoxyribonuclease YjjV
#44782	4567663	4568010	348	348	hypothetical protein, partial [Escherichia coli]	1,00E-25	-2	#4833	4567514	4568695	1182	Xylose ABC transporter, permease protein XylH
#1449	224495	224839	345	345	hypothetical protein [Actinomadura madurae]	4,00E-05	-2	#0204	224299	224952	654	Methionine ABC transporter permease protein
#14836	2238216	2238560	345	345	hypothetical protein [Xanthomonas vasicola]	1,00E-14	-2	#2396	2237858	2239231	1374	RND efflux system, outer membrane lipoprotein CmeC
#23405	3423443	3423787	345	345	hypothetical protein [Mycobacterium avium]	1,00E-04	-2	#3653	3423280	3423906	627	Uracil phosphoribosyltransferase
#23929	3494361	3494705	345	305	hypothetical protein [Streptomyces avermitilis]	4,00E-11	-2	#3718	3494327	3494665	339	Nitrogen regulatory protein P-II
#54026	3174399	3174743	345	126	hypothetical protein [Shigella flexneri]	3,00E-39	-2	#3408	3174427	3174552	126	hypothetical protein
#36968	5371547	5371888	342	104	hypothetical protein [Salmonella enterica]	3,00E-04	-2	#5591	5370703	5371650	948	Trehalose operon transcriptional repressor
#74517	291141	291482	342	116	hypothetical protein [Photorhabdus asymbiotica]	3,00E-06	-2	#0268	291367	291621	255	RtcB like protein
#22524	3297005	3297343	339	314	hypothetical protein [Xenorhabdus poinarii]	7,00E-11	-2	#3535	3295780	3297318	1539	Inner membrane component of tripartite multidrug resistance system
#50319	3740576	3740914	339	339	hypothetical protein, partial [Streptomyces europaeiscabiei]	1,00E-04	-2	#3978	3739602	3740969	1368	L-serine dehydratase
#70687	840007	840345	339	339	hypothetical protein [Mesorhizobium amorphae]	6,00E-17	-2	#0797	839255	840421	1167	Succinyl-CoA ligase [ADP-forming] beta chain
#10285	1561528	1561863	336	336	hypothetical protein [Pantoea agglomerans]	5,00E-15	-2	#1629	1561140	1562207	1068	Multidrug-efflux transporter, major facilitator superfamily (MFS)
#20182	2983483	2983818	336	336	histidine kinase, partial [Klebsiella pneumoniae]	6,00E-07	-2	#3203	2982270	2983955	1686	Autolysin sensor kinase
#3404	517332	517667	336	305	hypothetical protein [Escherichia coli]	6,00E-64	-2	#0502	516716	517636	921	Cytochrome O ubiquinol oxidase subunit II
#39720	5311510	5311845	336	178	hypothetical protein [Escherichia albertii]	8,00E-12	-2	#5524	5309246	5311687	2442	3'-to-5' exoribonuclease RNase R
#23482	3436016	3436348	333	333	hypothetical protein [Anaplasma phagocytophilum]	3,00E-08	-2	#3665	3435637	3437172	1536	Inosine-5'-monophosphate dehydrogenase
#32078	4660499	4660831	333	333	hypothetical protein [Escherichia coli]	2,00E-18	-2	#4917	4659631	4660836	1206	Sodium/glutamate symport protein
#6662	1012147	1012479	333	251	hypothetical protein, partial [Escherichia coli]	2,00E-57	-2	#0972	1011189	1012397	1209	Putative transport protein/putative regulator
#2264	342167	342496	330	269	hypothetical protein [Escherichia coli]	1,00E-09	-2	#0323	341767	342435	669	CFA/I fimbrial auxiliary subunit
#28129	4085560	4085889	330	330	hypothetical protein, partial [Streptomyces europaeiscabiei]	1,00E-10	-2	#4333	4085541	4086911	1371	L-serine dehydratase
#5562	845293	845622	330	330	hypothetical protein [Burkholderia gladioli]	0,001	-2	#0803	845268	846548	1281	Putative symport protein
#35279	5129942	5130268	327	244	hypothetical protein [Escherichia coli]	1,00E-21	-2	#5354	5130025	5130432	408	PTS system, sorbose-specific IIA component
#60798	2185530	2185856	327	327	hypothetical protein [Escherichia coli]	4,00E-30	-2	#2337	2185270	2185947	678	Aminobenzoyl-glutamate transport protein
#7374	1115881	1116207	327	117	hypothetical protein [Escherichia coli]	2,00E-23	-2	#1101	1116057	1116173	117	hypothetical protein
#9996	1516025	1516351	327	117	hypothetical protein [Escherichia coli]	2,00E-23	-2	#1562	1516201	1516317	117	hypothetical protein
#20471	3019388	3019711	324	324	hypothetical protein [Streptomyces vinaceus]	2,00E-67	-2	#3257	3019360	3020298	939	Origin specific replication initiation factor
#32441	4715120	4715443	324	324	hypothetical protein, partial [Escherichia coli]	3,00E-38	-2	#4983	4714507	4715745	1239	Putative transport protein
#64220	1705733	1706056	324	324	hypothetical protein [Streptomyces vinaceus]	2,00E-66	-2	#1807	1705260	1706084	825	Origin specific replication initiation factor
#65603	1529833	1530156	324	324	hypothetical protein [Escherichia coli]	3,00E-11	-2	#1582	1529330	1530202	873	NgrB
#68606	1129689	1130012	324	324	hypothetical protein [Escherichia coli]	3,00E-11	-2	#1121	1129186	1130058	873	NgrB
#74425	303970	304293	324	263	hypothetical protein [Streptomyces vinaceus]	3,00E-65	-2	#0282	304031	304321	291	Origin specific replication initiation factor

sORF with blastp hit							mother gene					
ID CP00895 7.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	rea-ding frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#42237	4941702	4942022	321	191	hypothetical protein [Salmonella enterica]	7,00E-28	-2	#5190	4941832	4943655	1824	GTP-binding protein TypA/BipA
#49067	3934171	3934488	318	318	hypothetical protein, partial [Escherichia coli]	8,00E-15	-2	#4177	3933653	3934732	1080	Membrane-bound lytic murein transglycosylase C precursor
#54047	3172405	3172722	318	318	hypothetical protein, partial [Escherichia coli]	3,00E-08	-2	#3406	3171956	3173215	1260	Anaerobic glycerol-3-phosphate dehydrogenase subunit B
#36100	5244822	5245136	315	122	hypothetical protein [Shigella dysenteriae]	2,00E-27	-2	#5464	5244035	5244943	909	Melibiose operon regulatory protein
#5302	804882	805196	315	315	hypothetical protein [Catenulispora acidiphila]	7,00E-08	-2	#0767	803192	805240	2049	Potassium-transporting ATPase B chain
#8163	1240667	1240981	315	277	hypothetical protein [Streptomyces ruber]	1,00E-06	-2	#1228	1240705	1241151	447	Inner membrane protein YccF
#31650	4589389	4589700	312	312	hypothetical protein [Xanthobacter autotrophicus]	7,00E-04	-2	#4850	4588341	4590185	1845	Selenocysteine-specific translation elongation factor
#41237	5082555	5082866	312	312	hypothetical protein, partial [Escherichia coli]	3,00E-33	-2	#5318	5081872	5086095	4224	DNA-directed RNA polymerase beta' subunit
#53058	3324687	3324998	312	230	hypothetical protein, partial [Salmonella enterica]	3,00E-07	-2	#3558	3324769	3326025	1257	Chloride channel protein
#54580	3092590	3092901	312	226	hypothetical protein, partial [Escherichia coli]	7,00E-31	-2	#3341	3092102	3092815	714	Putative membrane protein
#25835	3763316	3763624	309	309	hypothetical protein [Escherichia coli]	4,00E-40	-2	#3998	3762487	3766029	3543	Exodeoxyribonuclease V beta chain
#28593	4156318	4156626	309	309	hypothetical protein [Bacteroides finegoldii CAG:203]	3,00E-29	-2	#4406	4155126	4157060	1935	Cell division protein FtsH
#56324	2833844	2834152	309	309	hypothetical protein [Escherichia coli]	6,00E-42	-2	#3069	2833818	2835965	2148	Colicin I receptor precursor
#71781	678781	679089	309	309	hypothetical protein [Xanthomonas vasicola]	1,00E-07	-2	#0636	678038	679420	1383	Cation efflux system protein CusC precursor
#31226	4533564	4533869	306	241	hypothetical protein, partial [Bacillus thuringiensis]	3,00E-05	-2	#4800	4533629	4534531	903	Dipeptide transport system permease protein DppC
#41477	5051724	5052029	306	306	hypothetical protein [Salinispora pacifica]	2,00E-05	-2	#5296	5051113	5052486	1374	Argininosuccinate lyase
#65003	1604611	1604916	306	218	hypothetical protein [Vibrio mimicus]	3,00E-06	-2	#1678	1604699	1606132	1434	PTS system, glucose-specific IIB component
#67856	1227150	1227455	306	306	hypothetical protein [Haemophilus influenzae]	4,00E-04	-2	#1215	1226767	1228674	1908	ATPase components of ABC transporters with duplicated ATPase domains
#76123	62171	62476	306	137	hypothetical protein [Salmonella enterica]	4,00E-08	-2	#0058	62340	63155	816	DnaJ-like protein DjlA
#11457	1745188	1745490	303	303	hypothetical protein [Escherichia coli]	4,00E-11	-2	#1859	1745082	1746566	1485	Putative protease encoded within prophage CP-933X
#37380	5440827	5441129	303	303	hypothetical protein [Escherichia albertii]	7,00E-25	-2	#5642	5440406	5441386	981	hypothetical protein
#52425	3416448	3416750	303	303	hypothetical protein, partial [Escherichia coli]	2,00E-08	-2	#3646	3414934	3416946	2013	Formate hydrogenlyase transcriptional activator
#65241	1574654	1574956	303	239	hypothetical protein, partial [Escherichia coli]	3,00E-52	-2	#1645	1574718	1576253	1536	putative peptidoglycan lipid II flippase MurJ
#30208	4383338	4383637	300	300	hypothetical protein, partial [Escherichia coli]	2,00E-12	-2	#4646	4382602	4383942	1341	Low-affinity gluconate/H+ symporter GntU
#32339	4697204	4697503	300	122	type III secretion system protein SepZ [Escherichia albertii]	1,00E-39	-2	#4960	4697182	4697325	144	hypothetical protein
#40732	5155292	5155591	300	300	hypothetical protein [Escherichia coli]	2,00E-21	-2	#5377	5155065	5155937	873	4-hydroxybenzoate polyprenyltransferase
#48142	4066375	4066674	300	300	hypothetical protein [Streptomyces yeochonensis]	5,00E-04	-2	#4310	4065839	4066804	966	Integral membrane protein TerC
#61976	2005698	2005997	300	281	hypothetical protein [Escherichia albertii]	2,00E-39	-2	#2149	2005717	2007036	1320	redicted glycoside hydrolase
#72377	590394	590693	300	300	hypothetical protein [Escherichia coli]	5,00E-05	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#40657	5166963	5167259	297	297	hypothetical protein, partial [Prevotella disiens]	4,00E-13	-2	#5388	5166334	5167749	1416	Replicative DNA helicase
#43234	4815100	4815396	297	297	hypothetical protein [Streptomyces halstedii]	5,00E-06	-2	#5079	4814846	4816351	1506	Ribose ABC transport system, ATP-binding protein RbsA
#5005	765827	766123	297	297	hypothetical protein, partial [Mesoplasma photuris]	1,00E-04	-2	#0729	765667	766392	726	Glutamate Aspartate transport ATP-binding protein GltL
#51123	3620375	3620671	297	297	hypothetical protein [Xanthomonas oryzae]	3,00E-18	-2	#3847	3620088	3621272	1185	MFS permease protein
#72347	595182	595478	297	297	hypothetical protein [Escherichia coli]	6,00E-44	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72369	591591	591887	297	297	hypothetical protein [Escherichia coli]	9,00E-57	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72382	589797	590093	297	297	hypothetical protein [Escherichia coli]	2,00E-29	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72394	588297	588593	297	297	hypothetical protein [Escherichia coli]	1,00E-27	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#28027	4071531	4071824	294	294	hypothetical protein [Escherichia coli]	2,00E-21	-2	#4314	4070588	4072000	1413	Uronate isomerase
#37013	5377730	5378023	294	280	hypothetical protein [Neorhizobium galegae]	3,00E-04	-2	#5599	5377744	5378457	714	putative oxidoreductase YjgI

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#46501	4311678	4311971	294	294	hypothetical protein [Paenibacillus ehimensis]	3,00E-05	-2	#4581	4311529	4312326	798	Transcriptional regulator of fructoselysine utilization operon FrlR
#72345	595482	595775	294	294	hypothetical protein [Escherichia coli]	2,00E-14	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72402	587100	587393	294	294	hypothetical protein [Escherichia coli]	2,00E-06	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#73994	365574	365867	294	294	hypothetical protein [Streptomyces xinghaiensis]	6,00E-04	-2	#0348	365053	366480	1428	putative L-lactate dehydrogenase, Iron-sulfur cluster-binding subunit YkgF
#50487	3714138	3714428	291	178	hypothetical protein [Escherichia coli]	4,00E-54	-2	#3953	3713641	3714315	675	Uncharacterized protein YgcG
#72415	585309	585599	291	291	hypothetical protein [Escherichia coli]	1,00E-09	-2	#0565	581353	585738	4386	Large repetitive protein
#76200	50716	51006	291	291	hypothetical protein [Salmonella bongori]	5,00E-07	-2	#0046	50222	51553	1332	Putative metabolite transport protein yaaU
#36432	5289413	5289700	288	288	hypothetical protein [Escherichia coli]	6,00E-41	-2	#5507	5288638	5291961	3324	Potassium efflux system KefA protein
#38218	5532188	5532475	288	288	hypothetical protein [Dermatophilus congolensis]	3,00E-08	-2	#5732	5531583	5532965	1383	DNA repair protein RadA
#43241	4814388	4814675	288	257	hypothetical protein [Escherichia coli]	3,00E-18	-2	#5078	4814419	4814838	420	Ribose ABC transport system, high affinity permease RbsD
#51082	3625450	3625737	288	288	hypothetical protein, partial [Bacillus cereus]	4,00E-40	-2	#3853	3624395	3625933	1539	Multidrug resistance protein B
#72371	591291	591578	288	288	hypothetical protein [Escherichia coli]	1,00E-33	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72375	590694	590981	288	288	hypothetical protein [Escherichia coli]	7,00E-07	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#26793	3896216	3896500	285	131	hypothetical protein [Dickeya solani]	8,00E-10	-2	#4129	3895267	3896346	1080	Fructose-bisphosphate aldolase class II
#71004	794348	794632	285	285	hypothetical protein [Mycobacterium avium]	2,00E-24	-2	#0760	793359	794999	1641	Phosphoglucomutase
#34050	4948760	4949041	282	282	hypothetical protein [Escherichia coli]	4,00E-13	-2	#5195	4947787	4949190	1404	Glucuronide transport protein YihO
#3597	549177	549458	282	282	hypothetical protein [Akkermansia muciniphila CAG:154]	1,00E-22	-2	#0537	547319	550468	3150	RND efflux system, inner membrane transporter CmeB
#44545	4600740	4601021	282	170	hypothetical protein [Corynebacterium kutscheri]	7,00E-04	-2	#4860	4600852	4602765	1914	PTS system, mannitol-specific IIC component
#65851	1498987	1499268	282	282	hypothetical protein [Streptomyces ghanaensis]	6,00E-04	-2	#1534	1498922	1499503	582	Tellurium resistance protein TerD
#66438	1415149	1415430	282	282	hypothetical protein [Parabacteroides johnsonii CAG:246]	9,00E-11	-2	#1447	1414370	1415434	1065	Phosphate starvation-inducible protein PhoH, predicted ATPase
#68854	1098843	1099124	282	282	hypothetical protein [Streptomyces ghanaensis]	6,00E-04	-2	#1073	1098778	1099359	582	Tellurium resistance protein TerD
#72355	593688	593969	282	282	hypothetical protein [Escherichia coli]	1,00E-51	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#20858	3071118	3071396	279	279	hypothetical protein [Pseudomonas putida]	7,00E-05	-2	#3320	3070361	3071830	1470	Lysine-specific permease
#26879	3906945	3907223	279	152	hypothetical protein [Corynebacterium kutscheri]	3,00E-04	-2	#4141	3905708	3907096	1389	PTS system, mannitol-specific IIB component
#52626	3391437	3391715	279	279	hypothetical protein [Akkermansia muciniphila CAG:154]	1,00E-06	-2	#3624	3390610	3393723	3114	RND efflux system, inner membrane transporter CmeB
#72408	586503	586781	279	279	hypothetical protein [Escherichia coli]	2,00E-26	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#73662	411276	411554	279	125	hypothetical protein [Escherichia albertii]	2,00E-05	-2	#0395	411430	411624	195	hypothetical protein
#74700	265638	265916	279	279	hypothetical protein [Escherichia coli]	3,00E-19	-2	#0239	264862	267003	2142	VgrG protein
#75212	196057	196335	279	119	hypothetical protein [Escherichia coli]	6,00E-57	-2	#0176	196217	196774	558	Ribosome recycling factor
#14213	2153105	2153380	276	276	hypothetical protein, partial [Escherichia coli]	4,00E-38	-2	#2297	2152546	2154399	1854	Hypothetical protein
#15539	2332564	2332839	276	276	hypothetical protein, partial [Escherichia coli]	3,00E-38	-2	#2505	2332008	2333858	1851	Hypothetical protein
#18855	2789846	2790121	276	276	hypothetical protein, partial [Escherichia coli]	3,00E-38	-2	#3022	2789290	2791140	1851	Hypothetical protein
#32711	4750163	4750438	276	276	hypothetical protein [Escherichia coli]	2,00E-54	-2	#5020	4749997	4750809	813	Phosphatase YidA
#38023	5534380	5534655	276	130	hypothetical protein, partial [Escherichia coli]	4,00E-04	-2	#5735	5534526	5536193	1668	ABC transporter, ATP-binding protein
#62737	1908366	1908641	276	276	hypothetical protein, partial [Escherichia coli]	3,00E-38	-2	#2050	1907347	1909197	1851	Hypothetical protein
#63130	1860903	1861178	276	276	hypothetical protein, partial [Escherichia coli]	4,00E-37	-2	#1985	1859884	1861737	1854	Hypothetical protein
#67547	1268508	1268783	276	276	hypothetical protein, partial [Escherichia coli]	3,00E-38	-2	#1269	1267489	1269339	1851	Hypothetical protein
#24948	3635731	3636003	273	164	hypothetical protein [Cupriavidus basilensis]	0,001	-2	#3864	3635397	3635894	498	C-terminal domain of CinA type S
#30007	4357180	4357452	273	245	hypothetical protein [Sulfolobus tokodaii]	9,00E-07	-2	#4623	4356396	4357424	1029	RNA 3'-terminal phosphate cyclase

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#43051	4837082	4837354	273	273	hypothetical protein [Escherichia coli]	2,00E-26	-2	#5093	4836402	4837946	1545	Threonine dehydratase biosynthetic
#69701	981202	981474	273	273	hypothetical protein [Marinomonas posidonica]	7,00E-14	-2	#0943	980039	981631	1593	Putative ATPase component of ABC transporter with duplicated ATPase domain
#71702	690441	690713	273	95	hypothetical protein [Escherichia coli]	3,00E-32	-2	#0649	690619	690771	153	HokE protein
#17139	2556019	2556288	270	270	hypothetical protein, partial [Escherichia coli]	2,00E-16	-2	#2739	2554914	2556293	1380	Putative transport protein YdjK, MFS superfamily
#2475	375556	375825	270	270	hypothetical protein, partial [Mycobacterium avium]	6,00E-07	-2	#0361	374925	376397	1473	Betaine aldehyde dehydrogenase
#25795	3756978	3757247	270	197	hypothetical protein [Escherichia coli]	2,00E-22	-2	#3994	3756107	3757174	1068	Membrane-bound lytic murein transglycosylase A precursor
#11013	1674729	1674995	267	267	hypothetical protein [Escherichia coli]	3,00E-07	-2	#1767	1673726	1675087	1362	Leucine-rich repeat protein
#17967	2671422	2671688	267	267	hypothetical protein [Azospirillum brasilense]	2,00E-04	-2	#2863	2669750	2671708	1959	Signal transduction histidine kinase CheA
#35099	5102398	5102664	267	267	hypothetical protein, partial [Xanthomonas arboricola]	2,00E-10	-2	#5338	5102373	5103962	1590	IMP cyclohydrolase
#32427	4713338	4713601	264	264	hypothetical protein [Rhodococcus qingshengii]	3,00E-04	-2	#4981	4713133	4713951	819	Methionine ABC transporter substrate-binding protein
#40786	5148870	5149133	264	209	hypothetical protein [Nitratireductor pacificus]	5,00E-06	-2	#5372	5148925	5150040	1116	Maltose/maltodextrin transport ATP-binding protein MalK
#44244	4646426	4646689	264	148	reverse transcriptase [Escherichia coli]	3,00E-21	-2	#4902	4645977	4646573	597	Transcriptional regulator SlmA, TetR family
#44332	4631904	4632167	264	264	hypothetical protein [Escherichia coli]	7,00E-51	-2	#4886	4631425	4632633	1209	Oligosaccharide repeat unit polymerase Wzy
#50837	3662511	3662774	264	264	hypothetical protein [Saccharomonospora viridis]	7,00E-05	-2	#3897	3662311	3663183	873	[NiFe] hydrogenase nickel incorporation-associated protein HypB
#5499	834886	835146	261	107	MULTISPECIES: hypothetical protein [Enterobacteriaceae]	2,00E-54	-2	#0794	834861	834992	132	hypothetical protein
#55736	2925363	2925623	261	261	hypothetical protein [Kutzneria albida]	4,00E-11	-2	#3152	2925205	2926620	1416	Multidrug transporter MdtD
#14604	2207225	2207482	258	258	hypothetical protein [Nitratireductor pacificus]	3,00E-09	-2	#2361	2206678	2207760	1083	Multiple sugar ABC transporter, ATP-binding protein
#30119	4371575	4371832	258	258	hypothetical protein, partial [Escherichia coli]	2,00E-20	-2	#4636	4371460	4372755	1296	Glucose-1-phosphate adenylyltransferase
#31609	4584747	4585004	258	258	hypothetical protein [Mycobacterium avium]	2,00E-11	-2	#4847	4583930	4585468	1539	Aldehyde dehydrogenase B
#31807	4617324	4617581	258	258	hypothetical protein [Methylobacterium radiotolerans]	2,00E-10	-2	#4872	4617098	4618117	1020	Glycerol-3-phosphate dehydrogenase [NAD(P)+]
#14751	2225249	2225503	255	255	hypothetical protein [Pseudomonas syringae]	6,00E-06	-2	#2382	2224357	2225844	1488	Gamma-glutamyl-aminobutyraldehyde dehydrogenase
#22906	3352170	3352424	255	255	hypothetical protein [Dermatophilus congolensis]	2,00E-04	-2	#3587	3352001	3352834	834	Sulfate transport system permease protein CysT
#35306	5133619	5133873	255	140	hypothetical protein [Escherichia coli]	4,00E-09	-2	#5358	5133486	5133758	273	hypothetical protein
#51276	3600752	3601006	255	255	hypothetical protein [Escherichia coli]	1,00E-21	-2	#3823	3599790	3601058	1269	L-2-hydroxyglutarate oxidase
#51761	3525909	3526163	255	255	hypothetical protein [Xanthomonas hyacinthi]	5,00E-04	-2	#3747	3525823	3526242	420	Thioredoxin 2
#5293	803370	803624	255	255	hypothetical protein, partial [Streptomyces griseus]	7,00E-23	-2	#0767	803192	805240	2049	Potassium-transporting ATPase B chain
#74290	321028	321282	255	255	hypothetical protein [Salmonella enterica]	3,00E-04	-2	#0305	319559	321892	2334	Zinc binding domain protein
#74679	268236	268490	255	255	hypothetical protein, partial [Escherichia coli]	7,00E-23	-2	#0240	267079	271293	4215	core protein
#40857	5138028	5138279	252	233	hypothetical protein, partial [Shigella flexneri]	2,00E-06	-2	#5361	5138047	5138289	243	YjbE secreted protein
#5296	804039	804290	252	252	hypothetical protein, partial [Dactylosporangium aurantiacum]	8,00E-04	-2	#0767	803192	805240	2049	Potassium-transporting ATPase B chain
#58433	2528955	2529206	252	252	hypothetical protein [Heliobacterium modesticaldum]	8,00E-05	-2	#2712	2528782	2529489	708	Alkaline phosphatase like protein
#6630	1008247	1008498	252	200	hypothetical protein [Escherichia coli]	1,00E-08	-2	#0968	1007850	1008446	597	Putative permease
#73912	378008	378259	252	252	hypothetical protein [Propionibacterium acidifaciens]	0,001	-2	#0363	377127	379160	2034	High-affinity choline uptake protein BetT
#1068	162002	162250	249	199	hypothetical protein, partial [Escherichia coli]	1,00E-05	-2	#0147	162052	163416	1365	Poly(A) polymerase
#16557	2479051	2479299	249	249	hypothetical protein [Rubrivivax benzoatilyticus]	3,00E-07	-2	#2660	2478909	2481287	2379	Phosphoenolpyruvate synthase
#16671	2493256	2493504	249	135	hypothetical protein [Morganella morganii]	4,00E-05	-2	#2674	2493306	2493440	135	hypothetical protein
#25515	3716897	3717145	249	249	hypothetical protein [Bifidobacterium bifidum CAG:234]	5,00E-04	-2	#3957	3716668	3717966	1299	Enolase
#45276	4491183	4491431	249	249	hypothetical protein [Akkermansia muciniphila CAG:154]	6,00E-16	-2	#4765	4490173	4493286	3114	RND efflux system, inner membrane transporter CmeB
#49908	3804157	3804405	249	200	hypothetical protein [Escherichia coli]	1,00E-28	-2	#4038	3804206	3804709	504	IncI1 plasmid conjugative transfer putative membrane protein PilT

sORF with blastp hit							mother gene					
ID CP00895 7.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	rea-ding frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#51257	3603184	3603432	249	249	hypothetical protein [Nocardiopsis chromatogenes]	4,00E-14	-2	#3825	3602543	3603823	1281	Gamma-aminobutyrate:alpha-ketoglutarate aminotransferase
#69389	1023394	1023642	249	142	hypothetical protein [Escherichia coli]	2,00E-31	-2	#0987	1023047	1023535	489	Putative inner membrane protein
#69632	992470	992718	249	249	hypothetical protein [Bordetella parapertussis]	3,00E-08	-2	#0954	991625	993496	1872	Glutothione ABC transporter ATP-binding protein
#72352	594285	594533	249	249	hypothetical protein [Escherichia coli]	2,00E-10	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72399	587400	587648	249	249	hypothetical protein [Escherichia coli]	1,00E-06	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#75648	128181	128429	249	249	hypothetical protein [Pseudomonas knackmussii]	7,00E-15	-2	#0116	127507	130170	2664	Pyruvate dehydrogenase E1 component
#994	152411	152659	249	124	hypothetical protein [Vibrio cholerae]	4,00E-04	-2	#0137	152536	153387	852	Pantoate--beta-alanine ligase
#23633	3454377	3454622	246	246	citrate transporter [Escherichia coli]	5,00E-09	-2	#3680	3452876	3455188	2313	Penicillin-insensitive transglycosylase & transpeptidase PBP-1C
#26765	3892160	3892405	246	116	hypothetical protein, partial [Shigella dysenteriae]	6,00E-19	-2	#4125	3891697	3892275	579	HTH-type transcriptional regulator hdfR
#29422	4270451	4270696	246	246	50S ribosomal protein L14 [Staphylococcus epidermidis]	3,00E-14	-2	#4527	4270417	4270770	354	LSU ribosomal protein L14p (L23e)
#46630	4294500	4294745	246	246	hypothetical protein [Lactobacillus vaginalis]	5,00E-35	-2	#4562	4294336	4296426	2091	hypothetical protein
#65940	1483553	1483798	246	141	hypothetical protein [Escherichia coli]	1,00E-19	-2	#1513	1483584	1483724	141	hypothetical protein
#67577	1264550	1264795	246	227	hypothetical protein [Escherichia coli]	1,00E-49	-2	#1263	1264569	1265387	819	hypothetical protein
#68943	1083409	1083654	246	141	hypothetical protein [Escherichia coli]	1,00E-19	-2	#1052	1083440	1083580	141	hypothetical protein
#19360	2864414	2864656	243	243	phosphomannomutase, partial [Escherichia coli]	9,00E-39	-2	#3102	2863768	2865138	1371	Phosphomannomutase
#19512	2885847	2886089	243	243	phosphomannomutase, partial [Escherichia coli]	6,00E-38	-2	#3120	2885201	2886571	1371	Phosphomannomutase
#23127	3380473	3380715	243	243	hypothetical protein [Bordetella pertussis]	2,00E-10	-2	#3617	3379113	3381392	2280	NADP-dependent malic enzyme
#23133	3380980	3381222	243	243	hypothetical protein [Roseburia inulinivorans CAG:15]	1,00E-06	-2	#3617	3379113	3381392	2280	NADP-dependent malic enzyme
#30288	4394276	4394518	243	191	hypothetical protein [Nitratireductor pacificus]	5,00E-13	-2	#4657	4393396	4394466	1071	Glycerol-3-phosphate ABC transporter, ATP-binding protein UgpC
#31337	4546352	4546594	243	243	hypothetical protein [Erwinia amylovora]	7,00E-04	-2	#4812	4546003	4547205	1203	Putative resistance protein
#3572	545418	545660	243	229	acetyltransferase [Bacillus thuringiensis]	5,00E-09	-2	#0534	545432	545983	552	Maltose O-acetyltransferase
#60337	2252273	2252515	243	155	hypothetical protein, partial [Salmonella enterica]	3,00E-11	-2	#2406	2252361	2252579	219	Osmotically inducible lipoprotein B precursor
#1174	178682	178921	240	240	hypothetical protein [Acidovorax avenae]	7,00E-06	-2	#0159	177802	179082	1281	Glutamate-1-semialdehyde aminotransferase
#15745	2361288	2361527	240	194	hypothetical protein [Escherichia coli]	1,00E-49	-2	#2544	2360786	2361481	696	Dethiobiotin synthetase
#68434	1151542	1151781	240	149	hypothetical protein [Escherichia coli]	1,00E-08	-2	#1154	1151633	1155661	4029	Cell division protein FtsK
#71706	689998	690237	240	122	hypothetical protein [Escherichia coli]	2,00E-11	-2	#0647	690116	690268	153	HokE protein
#74628	274340	274579	240	240	hypothetical protein [Escherichia coli]	7,00E-38	-2	#0244	273792	275552	1761	core protein
#12504	1894015	1894251	237	98	hypothetical protein [Shigella dysenteriae]	2,00E-19	-2	#2025	1893924	1894112	189	Division inhibition protein dicB
#13413	2048479	2048715	237	237	hypothetical protein, partial [Escherichia coli]	4,00E-42	-2	#2189	2045670	2049872	4203	core protein
#13898	2111152	2111388	237	237	hypothetical protein [Streptococcus suis]	5,00E-06	-2	#2247	2108631	2111666	3036	porin, autotransporter (AT) family
#24088	3514860	3515096	237	218	hypothetical protein [Escherichia coli]	7,00E-35	-2	#3735	3514598	3515077	480	Sigma factor RpoE regulatory protein RseC
#28184	4092947	4093183	237	237	hypothetical protein [Cupriavidus basilensis]	1,00E-06	-2	#4338	4092295	4093284	990	Threonine dehydratase, catabolic
#2899	442047	442283	237	237	hypothetical protein [Vibrio nigripulchritudo]	4,00E-18	-2	#0422	441344	442375	1032	Ferric iron ABC transporter, iron-binding protein
#52625	3391770	3392006	237	237	hypothetical protein [Akkermansia muciniphila CAG:154]	5,00E-08	-2	#3624	3390610	3393723	3114	RND efflux system, inner membrane transporter CmeB
#72191	617905	618141	237	237	hypothetical protein, partial [Escherichia coli]	4,00E-42	-2	#0585	616748	620944	4197	core protein
#11814	1791657	1791890	234	215	hypothetical protein [Escherichia coli]	8,00E-47	-2	#1910	1790192	1791871	1680	Putative sulfate permease
#25009	3643198	3643431	234	234	hypothetical protein, partial [Escherichia coli]	5,00E-08	-2	#3876	3642324	3643838	1515	Anaerobic nitric oxide reductase transcription regulator NorR
#2760	422921	423154	234	161	hypothetical protein [Borrelia coriaceae]	3,00E-15	-2	#0404	420007	423081	3075	Beta-galactosidase
#29370	4263457	4263690	234	212	hypothetical protein [Methylobacterium radiotolerans]	8,00E-06	-2	#4512	4262679	4263668	990	DNA-directed RNA polymerase alpha subunit

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#32064	4658826	4659059	234	234	ATP-dependent DNA helicase RecG [Escherichia coli]	4,00E-44	-2	#4916	4658762	4659628	867	putative cytoplasmic protein
#34272	4982040	4982273	234	95	hypothetical protein [Escherichia coli]	3,00E-04	-2	#5228	4981310	4982134	825	Rhamnulose-1-phosphate aldolase
#52964	3338839	3339072	234	234	hypothetical protein [Aeromonas fluvialis]	3,00E-11	-2	#3572	3338216	3339214	999	Putative cytochrome oxidase
#53969	3183590	3183823	234	234	hypothetical protein [Bordetella holmesii]	6,00E-11	-2	#3418	3183429	3184586	1158	UDP-4-amino-4-deoxy-L-arabinose--oxoglutarate aminotransferase
#63346	1833305	1833538	234	234	hypothetical protein [Clostridium bolteae CAG:59]	4,00E-05	-2	#1950	1832784	1833797	1014	Oligopeptide transport ATP-binding protein OppD
#2266	342770	343000	231	231	hypothetical protein [Escherichia coli]	2,00E-23	-2	#0324	342493	343080	588	CFA/I fimbrial major subunit
#32499	4722866	4723096	231	231	acyl-CoA thioesterase [Escherichia coli]	5,00E-06	-2	#4991	4722061	4723380	1320	Hexose phosphate uptake regulatory protein UhpC
#48798	3974248	3974478	231	106	hypothetical protein [Salmonella enterica]	3,00E-14	-2	#4220	3974225	3974353	129	hypothetical protein
#76302	35737	35967	231	231	hypothetical protein [Akkermansia muciniphila CAG:154]	7,00E-19	-2	#0033	35192	38413	3222	Carbamoyl-phosphate synthase large chain
#12840	1948814	1949041	228	187	hypothetical protein [Escherichia coli]	1,00E-06	-2	#2102	1948855	1949955	1101	Putative transport protein
#24390	3554535	3554762	228	158	cytochrome c-type biogenesis heme exporter protein C, partial [Salmonella enterica]	9,00E-05	-2	#3771	3553331	3554692	1362	Signal recognition particle, subunit Ffh SRP54
#31438	4559373	4559600	228	134	hypothetical protein, partial [Escherichia coli]	3,00E-34	-2	#4825	4559207	4559506	300	putative lipoprotein YsaB precursor
#30147	4375464	4375688	225	225	hypothetical protein [Escherichia coli]	1,00E-05	-2	#4638	4374743	4376929	2187	1,4-alpha-glucan (glycogen) branching enzyme, GH-13-type
#33810	4911680	4911904	225	225	hypothetical protein [Streptomyces rimosus]	6,00E-04	-2	#5165	4911355	4912518	1164	3-ketoacyl-CoA thiolase
#35987	5230008	5230232	225	119	hypothetical protein [Escherichia coli]	9,00E-41	-2	#5453	5229713	5230126	414	Alkylphosphonate utilization operon protein PhnA
#45845	4412912	4413136	225	225	hypothetical protein [Escherichia coli]	2,00E-15	-2	#4681	4412625	4413251	627	putative enzyme yhhN
#56954	2741612	2741836	225	212	hypothetical protein [Achromobacter insuavis]	2,00E-07	-2	#2957	2741625	2741894	270	Flagellar biosynthesis protein FliQ
#67954	1214647	1214871	225	200	hypothetical protein [Shigella flexneri]	3,00E-09	-2	#1202	1214672	1215211	540	type 1 fimbriae major subunit FimA
#69593	996630	996854	225	225	hypothetical protein [Catenuloplanes japonicus]	8,00E-07	-2	#0957	995995	996906	912	Dipeptide transport system permease protein DppC
#72385	589494	589718	225	225	hypothetical protein [Escherichia coli]	2,00E-19	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#76445	13310	13534	225	225	hypothetical protein [Bifidobacterium bifidum CAG:234]	7,00E-09	-2	#0013	12180	14096	1917	Chaperone protein DnaK
#11569	1761230	1761451	222	161	hypothetical protein [Escherichia coli]	1,00E-34	-2	#1880	1759849	1761390	1542	Na+/H+ antiporter NhaB
#5295	803745	803966	222	222	hypothetical protein, partial [Dactylosporangium aurantiacum]	9,00E-11	-2	#0767	803192	805240	2049	Potassium-transporting ATPase B chain
#73484	433884	434105	222	222	hypothetical protein, partial [Bacillus cereus]	1,00E-42	-2	#0415	433657	434868	1212	4-hydroxybenzoate transporter
#946	143677	143898	222	222	hypothetical protein, partial [Escherichia coli]	3,00E-09	-2	#0127	143208	145598	2391	Glucose dehydrogenase, PQQ-dependent
#37680	5482235	5482453	219	219	hypothetical protein [Escherichia albertii]	5,00E-30	-2	#5684	5482174	5483586	1413	Transcriptional regulator, GntR family
#53685	3229692	3229910	219	219	hypothetical protein [Streptomyces xinghaiensis]	2,00E-13	-2	#3462	3227923	3230067	2145	Phosphate acetyltransferase
#59308	2400155	2400373	219	219	hypothetical protein, partial [Vibrio parahaemolyticus]	6,00E-06	-2	#2582	2399859	2400440	582	Electron transport complex protein RnfA
#66309	1436642	1436860	219	219	hypothetical protein, partial [Snodgrassella alvi]	8,00E-05	-2	#1464	1436361	1440173	3813	Hemolysin
#21622	3177022	3177237	216	216	hypothetical protein [Escherichia coli]	4,00E-06	-2	#3411	3176382	3177671	1290	L-rhamnonate transporter (predicted by chromosomal context)
#23544	3444352	3444567	216	216	hypothetical protein [Streptomyces iranensis]	4,00E-05	-2	#3672	3443718	3444836	1119	1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase
#28474	4139872	4140087	216	216	hypothetical protein [Thermocrinis albus]	2,00E-08	-2	#4393	4139157	4141292	2136	Polyribonucleotide nucleotidyltransferase
#36304	5270590	5270805	216	216	hypothetical protein [Streptomyces europaeiscabiei]	1,00E-20	-2	#5486	5270208	5271644	1437	Aspartate ammonia-lyase
#52152	3461436	3461651	216	107	hypothetical protein [Yersinia pseudotuberculosis]	8,00E-09	-2	#3683	3461545	3461670	126	hypothetical protein
#15545	2333248	2333460	213	213	hypothetical protein [Escherichia albertii]	2,00E-07	-2	#2505	2332008	2333858	1851	Hypothetical protein
#21626	3177556	3177768	213	116	hypothetical protein, partial [Escherichia coli]	4,00E-33	-2	#3411	3176382	3177671	1290	L-rhamnonate transporter (predicted by chromosomal context)
#29602	4297388	4297600	213	213	hypothetical protein, partial [Catenibacterium mitsuokai]	3,00E-05	-2	#4563	4296493	4297713	1221	Acetylornithine aminotransferase
#40206	5234170	5234382	213	213	hypothetical protein [Streptomyces globisporus]	5,00E-06	-2	#5456	5233817	5235319	1503	L-Proline/Glycine betaine transporter ProP
#47852	4111208	4111420	213	213	hypothetical protein, partial [Escherichia coli]	2,00E-14	-2	#4359	4111143	4112297	1155	Galactosamine-6-phosphate isomerase

sORF with blastp hit							mother gene					
ID CP00895 7.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	rea-ding frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#6373	969910	970122	213	213	hypothetical protein [Streptomyces bikiniensis]	3,00E-16	-2	#0932	969819	970541	723	Glutamate transport ATP-binding protein
#72349	594882	595094	213	213	hypothetical protein [Escherichia coli]	3,00E-23	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#17681	2636212	2636421	210	210	hypothetical protein, partial [Escherichia coli]	2,00E-07	-2	#2829	2635515	2636486	972	Lipid A biosynthesis (KDO) 2-(lauroyl)-lipid IVA acyltransferase
#20567	3032991	3033200	210	210	hypothetical protein [Escherichia coli]	2,00E-17	-2	#3281	3032936	3033667	732	Osmoprotectant ABC transporter inner membrane protein YehW
#25183	3672105	3672314	210	210	hypothetical protein, partial [Streptomyces afghaniensis]	5,00E-12	-2	#3907	3672014	3673441	1428	Hydroxyaromatic non-oxidative decarboxylase protein C
#31652	4589749	4589958	210	210	hypothetical protein [Xanthobacter autotrophicus]	3,00E-04	-2	#4850	4588341	4590185	1845	Selenocysteine-specific translation elongation factor
#76304	35503	35712	210	210	hypothetical protein [Akkermansia muciniphila CAG:154]	2,00E-09	-2	#0033	35192	38413	3222	Carbamoyl-phosphate synthase large chain
#839	125479	125688	210	210	hypothetical protein [Bifidobacterium adolescentis CAG:119]	2,00E-07	-2	#0113	124668	126038	1371	Aromatic amino acid transport protein AroP
#12056	1826908	1827114	207	207	hypothetical protein [Lactobacillus delbrueckii]	2,00E-04	-2	#1943	1824675	1827350	2676	Alcohol dehydrogenase
#70700	837904	838110	207	167	hypothetical protein [Cupriavidus basilensis]	6,00E-05	-2	#0796	837944	839161	1218	Dihydrolipoamide succinyltransferase component (E2) of 2-oxoglutarate dehydrogenase complex
#14732	2222814	2223017	204	204	hypothetical protein [Escherichia coli]	6,00E-32	-2	#2380	2221772	2223037	1266	Gamma-aminobutyrate:alpha-ketoglutarate aminotransferase
#1925	293252	293455	204	204	hypothetical protein [Parabacteroides merdae CAG:48]	3,00E-15	-2	#0270	292147	293604	1458	Aminoacyl-histidine dipeptidase (Peptidase D)
#24363	3551343	3551546	204	197	hypothetical protein [Bifidobacterium longum]	5,00E-04	-2	#3767	3551192	3551539	348	LSU ribosomal protein L19p
#34570	5022779	5022982	204	125	hypothetical protein [Escherichia coli]	3,00E-23	-2	#5273	5022586	5022903	318	Methionine repressor MetJ
#37938	5521321	5521524	204	130	hypothetical protein, partial [Escherichia coli]	2,00E-27	-2	#5721	5521395	5522177	783	radical activating enzyme
#48189	4060467	4060670	204	204	hypothetical protein, partial [Escherichia coli]	4,00E-21	-2	#4303	4058923	4060941	2019	2,4-dienoyl-CoA reductase [NADPH]
#65197	1580082	1580285	204	204	conjugal transfer protein, partial [Escherichia coli]	2,00E-06	-2	#1653	1579453	1580658	1206	Flagellar hook protein FlgE
#7973	1212461	1212664	204	204	hypothetical protein [Mycobacterium avium]	2,00E-04	-2	#1199	1211647	1212792	1146	Alkanesulfonate monooxygenase
#20654	3043847	3044047	201	201	hypothetical protein [Cronobacter malonaticus]	4,00E-19	-2	#3293	3043741	3044502	762	3-oxoacyl-[acyl-carrier protein] reductase
#21335	3139138	3139338	201	201	hypothetical protein [Geodermatophilus obscurus]	4,00E-04	-2	#3385	3138333	3141182	2850	Two-component sensor protein RcsC
#23115	3378600	3378800	201	201	glutamate decarboxylase isozyme [Shigella dysenteriae]	1,00E-05	-2	#3616	3378485	3378820	336	Ethanolamine utilization polyhedral-body-like protein EutS
#47006	4237762	4237962	201	201	hypothetical protein [Akkermansia muciniphila CAG:154]	5,00E-07	-2	#4488	4237619	4239505	1887	RND efflux system, inner membrane transporter CmeB
#59607	2357639	2357839	201	201	hypothetical protein, partial [Escherichia coli]	6,00E-09	-2	#2539	2357253	2357870	618	Anaerobic dimethyl sulfoxide reductase chain B
#60306	2256860	2257060	201	201	transaldolase, partial [Escherichia coli]	8,00E-11	-2	#2413	2256528	2257118	591	GTP cyclohydrolase II
#65060	1597949	1598149	201	201	hypothetical protein [Streptomyces scabiei]	1,00E-07	-2	#1670	1597500	1598234	735	3-oxoacyl-[acyl-carrier protein] reductase
#23880	3486662	3486859	198	98	hypothetical protein [Escherichia coli]	6,00E-18	-2	#3712	3485776	3486759	984	putative sugar ABC transport system, periplasmic binding protein YphF precursor
#26035	3785493	3785690	198	198	hypothetical protein [Escherichia coli]	5,00E-32	-2	#4016	3783881	3786040	2160	2-acylglycerophosphoethanolamine acyltransferase
#61722	2037647	2037844	198	149	hypothetical protein [Escherichia coli]	8,00E-15	-2	#2175	2037696	2039240	1545	Respiratory nitrate reductase beta chain
#6376	970381	970578	198	161	hypothetical protein [Streptomyces viridochromogenes]	1,00E-04	-2	#0932	969819	970541	723	Glutamate transport ATP-binding protein
#1196	181944	182138	195	195	lysE type translocator family protein [Escherichia coli]	0,001	-2	#0163	181862	182662	801	Vitamin B12 ABC transporter, B12-binding component BtuF
#14605	2207531	2207725	195	195	hypothetical protein, partial [Thermobrachium celere]	2,00E-04	-2	#2361	2206678	2207760	1083	Multiple sugar ABC transporter, ATP-binding protein
#23265	3400444	3400638	195	161	hypothetical protein, partial [Escherichia coli]	2,00E-21	-2	#3631	3399891	3400604	714	Phosphoribosylaminoimidazole-succinocarboxamide synthase
#39168	5391930	5392124	195	195	hypothetical protein, partial [Escherichia coli]	5,00E-14	-2	#5612	5391793	5392875	1083	putative Permease
#49617	3851137	3851331	195	195	hypothetical protein, partial [Bacillus thuringiensis]	1,00E-33	-2	#4083	3850127	3851527	1401	Putative purine permease YgfO
#7056	1070769	1070963	195	140	hypothetical protein [Escherichia coli]	2,00E-36	-2	#1032	1070597	1070908	312	Transposase
#46572	4301477	4301668	192	125	hypothetical protein [Klebsiella pneumoniae]	6,00E-07	-2	#4570	4301544	4304087	2544	Nitrite reductase [NAD(P)H] large subunit
#51552	3556046	3556237	192	192	porphobilinogen deaminase [Citrobacter amalonaticus]	1,00E-10	-2	#3773	3555696	3556958	1263	Hemolysin with CBS domain
#63335	1834570	1834761	192	192	hypothetical protein, partial [Escherichia coli]	2,00E-13	-2	#1951	1833794	1834798	1005	Oligopeptide transport ATP-binding protein OppF
#6795	1031477	1031668	192	192	hypothetical protein, partial [Kitasatospora cheerisanensis]	8,00E-04	-2	#0997	1031374	1032102	729	Arginine ABC transporter, ATP-binding protein ArtP

sORF with blastp hit							mother gene					
ID CP00895 7.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	rea-ding frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#71463	724299	724490	192	160	hypothetical protein [Cronobacter dublinensis]	6,00E-04	-2	#0681	723220	724458	1239	putative zinc-type alcohol dehydrogenase-like protein ybdR
#72341	596484	596675	192	192	hypothetical protein [Escherichia coli]	1,00E-04	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#11476	1747600	1747788	189	107	hypothetical protein [Escherichia coli]	7,00E-35	-2	#1860	1746753	1747706	954	Protease VII (Omptin) precursor
#43060	4836027	4836215	189	189	hypothetical protein [Mycobacterium obuense]	1,00E-04	-2	#5092	4834549	4836399	1851	Dihydroxy-acid dehydratase
#4815	740775	740963	189	189	hypothetical protein [Methylobacterium radiotolerans]	8,00E-09	-2	#0700	740384	741349	966	Lipoate synthase
#63854	1757888	1758076	189	189	hypothetical protein [Corynebacterium sputi]	2,00E-04	-2	#1878	1757859	1759127	1269	Error-prone, lesion bypass DNA polymerase V (UmuC)
#64397	1682381	1682569	189	189	hypothetical protein [Escherichia coli]	1,00E-36	-2	#1775	1682169	1682912	744	hypothetical protein
#65063	1597640	1597828	189	189	hypothetical protein, partial [Gilliamella apicola]	9,00E-19	-2	#1670	1597500	1598234	735	3-oxoacyl-[acyl-carrier protein] reductase
#18278	2709111	2709296	186	186	hypothetical protein [Pseudoxanthomonas spadix]	9,00E-07	-2	#2917	2709002	2709550	549	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase
#27383	3979633	3979818	186	186	hypothetical protein, partial [Escherichia coli]	1,00E-15	-2	#4228	3979233	3979967	735	MotA/TolQ/ExbB proton channel family protein
#30413	4409026	4409211	186	186	hypothetical protein, partial [Kitasatospora cheerisanensis]	2,00E-06	-2	#4676	4408950	4409618	669	Cell division transporter, ATP-binding protein FtsE
#45333	4483286	4483471	186	186	hypothetical protein [Microlunatus phosphovorus]	5,00E-05	-2	#4756	4483239	4484195	957	Hemin ABC transporter, permease protein
#53506	3256877	3257062	186	186	hypothetical protein [Escherichia coli]	4,00E-04	-2	#3491	3255132	3257138	2007	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase
#57946	2594386	2594571	186	186	hypothetical protein, partial [Streptomyces europaeiscabiei]	2,00E-09	-2	#2784	2593289	2594587	1299	L-serine dehydratase 1
#59306	2400616	2400801	186	186	hypothetical protein, partial [Vibrio parahaemolyticus]	4,00E-05	-2	#2583	2400440	2401018	579	Electron transport complex protein RnfB
#61884	2015980	2016165	186	186	hypothetical protein [Clostridium bolteae CAG:59]	3,00E-07	-2	#2157	2015525	2016511	987	Dipeptide transport system permease protein DppC
#13816	2101587	2101769	183	140	hypothetical protein, partial [Escherichia coli]	2,00E-19	-2	#2242	2100830	2101726	897	Phosphatidate cytidylyltransferase
#1794	271929	272111	183	128	hypothetical protein [Escherichia coli]	1,00E-31	-2	#0242	271910	272056	147	hypothetical protein
#18861	2790560	2790742	183	183	hypothetical protein [Escherichia albertii]	2,00E-04	-2	#3022	2789290	2791140	1851	Hypothetical protein
#32141	4670326	4670508	183	129	hypothetical protein [Escherichia coli]	2,00E-33	-2	#4924	4670337	4670465	129	hypothetical protein
#3759	574748	574930	183	183	hypothetical protein [Lactococcus lactis]	8,00E-05	-2	#0561	574564	577068	2505	Lead, cadmium, zinc and mercury transporting ATPase
#39577	5330576	5330758	183	183	hypothetical protein, partial [Photorhabdus temperata]	6,00E-05	-2	#5549	5330457	5330783	327	hypothetical protein
#40210	5233960	5234142	183	183	hypothetical protein [Streptomyces globisporus]	2,00E-06	-2	#5456	5233817	5235319	1503	L-Proline/Glycine betaine transporter ProP
#41864	4996928	4997110	183	183	hypothetical protein, partial [Escherichia coli]	8,00E-10	-2	#5246	4996419	4997408	990	Sulfate-binding protein Sbp
#42537	4903149	4903331	183	183	hypothetical protein [Corynebacterium xerosis]	9,00E-07	-2	#5155	4902790	4903545	756	Ubiquinone/menaquinone biosynthesis methyltransferase UbiE, 2-heptaprenyl-1,4-naphthoquinone methyltransferase
#42716	4879278	4879460	183	183	hypothetical protein [Mesorhizobium amorphae]	8,00E-04	-2	#5132	4879267	4881429	2163	ATP-dependent DNA helicase UvrD/PcrA
#58842	2465288	2465470	183	183	hypothetical protein [Escherichia coli]	3,00E-19	-2	#2647	2465244	2466458	1215	Putative transport system permease protein
#67633	1258268	1258450	183	128	hypothetical protein, partial [Escherichia coli]	6,00E-25	-2	#1255	1258323	1259069	747	DNA replication protein DnaC
#19652	2904878	2905057	180	180	hypothetical protein [Escherichia coli]	3,00E-21	-2	#3139	2903914	2905767	1854	AsmA protein
#22072	3239681	3239860	180	180	hypothetical protein [Burkholderia fungorum]	5,00E-05	-2	#3475	3239182	3239964	783	Histidine ABC transporter, histidine-binding periplasmic protein precursor HisJ
#22083	3240684	3240863	180	180	hypothetical protein [Burkholderia fungorum]	1,00E-07	-2	#3476	3240185	3240967	783	Lysine-arginine-ornithine-binding periplasmic protein precursor
#2217	336433	336612	180	117	hypothetical protein [Escherichia coli]	2,00E-18	-2	#0319	336447	336563	117	Ferredoxin
#3314	505755	505934	180	180	hypothetical protein [Bacteroides fragilis CAG:558]	4,00E-06	-2	#0489	504368	506230	1863	1-deoxy-D-xylulose 5-phosphate synthase
#48744	3981875	3982054	180	180	hypothetical protein [Escherichia coli]	1,00E-18	-2	#4230	3981546	3982205	660	DedA family inner membrane protein YghB
#70381	886844	887023	180	121	hypothetical protein [Escherichia fergusonii]	1,00E-04	-2	#0845	885531	886964	1434	2-oxoglutarate/malate translocator
#70570	859232	859411	180	180	hypothetical protein [Klebsiella pneumoniae]	3,00E-10	-2	#0816	859035	859439	405	4-hydroxybenzoyl-CoA thioesterase family protein
#75930	92043	92222	180	180	hypothetical protein [Escherichia coli]	1,00E-06	-2	#0081	91933	92454	522	Acetolactate synthase small subunit
#8166	1241018	1241197	180	134	hypothetical protein [Escherichia coli]	5,00E-10	-2	#1228	1240705	1241151	447	Inner membrane protein YccF
#36839	5353652	5353828	177	177	hypothetical protein [Rhodococcus qingshengii]	4,00E-08	-2	#5573	5353468	5353998	531	Inorganic pyrophosphatase

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#38753	5448036	5448212	177	134	membrane protein [Escherichia coli]	1,00E-06	-2	#5651	5448079	5448804	726	chaperone FimC
#47325	4189365	4189541	177	177	hypothetical protein [Salmonella enterica]	6,00E-05	-2	#4441	4189177	4190595	1419	Glutamate synthase [NADPH] small chain
#55300	2984234	2984410	177	121	hypothetical protein [Escherichia coli]	2,00E-21	-2	#3205	2984214	2984354	141	hypothetical protein
#59477	2376825	2377001	177	177	hypothetical protein [Escherichia coli]	3,00E-12	-2	#2560	2376421	2377140	720	Transcriptional regulatory protein RstA
#59761	2337589	2337765	177	170	hypothetical protein [Escherichia coli]	1,00E-06	-2	#2512	2337596	2337775	180	hypothetical protein
#8325	1262371	1262547	177	170	hypothetical protein [Escherichia coli]	1,00E-06	-2	#1259	1262361	1262540	180	hypothetical protein
#14863	2241007	2241180	174	174	hypothetical protein [Akkermansia muciniphila CAG:154]	6,00E-08	-2	#2397	2239254	2242337	3084	RND efflux system, inner membrane transporter CmeB
#15971	2393071	2393244	174	174	hypothetical protein [Cronobacter dublinensis]	9,00E-08	-2	#2574	2392242	2393270	1029	Maltose regulon regulatory protein MalI (repressor for malXY)
#3297	503961	504134	174	174	hypothetical protein [Streptomyces resistomycificus]	4,00E-09	-2	#0488	503339	504313	975	Putative oxidoreductase
#34998	5088440	5088613	174	174	hypothetical protein [Achromobacter xylosoxidans]	2,00E-04	-2	#5321	5088316	5089086	771	Thiazole biosynthesis protein ThiG
#41	6697	6870	174	174	hypothetical protein [Escherichia coli]	5,00E-09	-2	#0007	6546	7976	1431	Putative alanine/glycine transport protein
#46160	4359960	4360133	174	174	RNA 3'-terminal phosphate cyclase [Escherichia coli]	6,00E-07	-2	#4625	4358842	4360440	1599	Transcriptional regulatory protein RtcR
#61767	2032518	2032691	174	174	hypothetical protein [Salmonella enterica]	2,00E-19	-2	#2173	2032489	2033877	1389	Nitrate/nitrite transporter
#72227	613643	613816	174	174	hypothetical protein, partial [Kitasatospora cheerisanensis]	1,00E-04	-2	#0583	613221	613907	687	putative metabolite ABC transporter in Enterobacteriaceae, ATP-binding protein EC-YbbA
#16080	2411133	2411303	171	171	hypothetical protein [Azospirillum brasilense]	2,00E-07	-2	#2593	2410136	2411410	1275	Tyrosyl-tRNA synthetase
#5067	772661	772831	171	171	hypothetical protein [Parabacteroides johnsonii CAG:246]	2,00E-05	-2	#0737	772276	773316	1041	Phosphate starvation-inducible ATPase PhoH with RNA binding motif
#59074	2428961	2429131	171	171	hypothetical protein [Rhodococcus qingshengii]	1,00E-05	-2	#2612	2428623	2429438	816	Putative lipoprotein
#59077	2428565	2428735	171	113	hypothetical protein [Salmonella enterica]	4,00E-13	-2	#2612	2428623	2429438	816	Putative lipoprotein
#64881	1623904	1624074	171	171	hypothetical protein, partial [Microbacterium barkeri]	1,00E-05	-2	#1695	1623485	1624186	702	Lipoprotein releasing system ATP-binding protein LolD
#11683	1776282	1776449	168	168	hypothetical protein, partial [Escherichia coli]	3,00E-25	-2	#1895	1776074	1776718	645	Orf2
#22202	3254531	3254698	168	168	recombinase, partial [Escherichia coli]	2,00E-14	-2	#3490	3253753	3254973	1221	3-oxoacyl-[acyl-carrier-protein] synthase, KASI
#25217	3676760	3676927	168	143	hypothetical protein [Salmonella enterica]	7,00E-16	-2	#3912	3675763	3676902	1140	Lipoprotein NlpD
#26624	3873285	3873452	168	168	hypothetical protein [Escherichia coli]	5,00E-16	-2	#4105	3870923	3873796	2874	Glycine dehydrogenase [decarboxylating] (glycine cleavage system P protein)
#39544	5334581	5334748	168	168	hypothetical protein [Bifidobacterium adolescentis CAG:119]	7,00E-04	-2	#5554	5334204	5335616	1413	D-serine/D-alanine/glycine transporter
#49520	3866906	3867073	168	106	hypothetical protein [Escherichia albertii]	4,00E-06	-2	#4100	3866889	3867011	123	hypothetical protein
#61771	2031685	2031852	168	168	hypothetical protein, partial [Shigella dysenteriae]	9,00E-09	-2	#2172	2030924	2032183	1260	internalin, putative
#70018	934628	934795	168	168	hypothetical protein, partial [Escherichia coli]	2,00E-27	-2	#0899	934407	935084	678	Dethiobiotin synthetase
#73052	493832	493999	168	168	hypothetical protein [Escherichia coli]	1,00E-10	-2	#0474	493134	494105	972	Protein-export membrane protein SecF
#73383	449018	449185	168	168	hypothetical protein, partial [Escherichia coli]	3,00E-09	-2	#0429	448797	449648	852	Alpha-ketoglutarate-dependent taurine dioxygenase
#17071	2548188	2548352	165	165	hypothetical protein [Escherichia coli]	5,00E-09	-2	#2730	2547386	2548636	1251	Putative transport protein YdjK, MFS superfamily
#17460	2604733	2604897	165	107	hypothetical protein, partial [Escherichia coli]	2,00E-18	-2	#2797	2604552	2604839	288	hypothetical protein
#18280	2709315	2709479	165	165	hypothetical protein [Pseudoxanthomonas spadix]	7,00E-05	-2	#2917	2709002	2709550	549	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase
#3355	511455	511619	165	165	hypothetical protein [Escherichia coli]	5,00E-19	-2	#0497	511355	512719	1365	Putative transport protein
#46670	4289784	4289948	165	165	hypothetical protein [Escherichia coli]	1,00E-06	-2	#4556	4288822	4290735	1914	Glutathione-regulated potassium-efflux system ATP-binding protein
#54522	3100013	3100177	165	165	hypothetical protein [Clostridium bolteae CAG:59]	2,00E-04	-2	#3347	3099549	3101138	1590	Putative ABC transporter ATP-binding protein
#58058	2578193	2578357	165	165	hypothetical protein [Streptomyces sclerotialus]	2,00E-06	-2	#2768	2577978	2579063	1086	Tartrate dehydrogenase
#76424	16504	16668	165	165	hypothetical protein [Paraprevotella clara CAG:116]	3,00E-05	-2	#0015	16157	17323	1167	Na+/H+ antiporter NhaA type
#17609	2625768	2625929	162	162	hypothetical protein [Escherichia coli]	2,00E-13	-2	#2819	2625341	2626000	660	inner membrane protein YebE
#39333	5364736	5364897	162	162	hypothetical protein [Escherichia albertii]	5,00E-12	-2	#5585	5364725	5364919	195	hypothetical protein

sORF with blastp hit							mother gene					
ID CP00895 7.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	rea-ding frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#51261	3602608	3602769	162	162	hypothetical protein, partial [Catenibacterium mitsuokai]	3,00E-04	-2	#3825	3602543	3603823	1281	Gamma-aminobutyrate:alpha-ketoglutarate aminotransferase
#5149	783269	783430	162	143	hypothetical protein [Escherichia coli]	2,00E-12	-2	#0746	782611	783411	801	Glucosamine-6-phosphate deaminase
#63347	1833140	1833301	162	162	hypothetical protein, partial [Escherichia coli]	1,00E-17	-2	#1950	1832784	1833797	1014	Oligopeptide transport ATP-binding protein OppD
#72395	588057	588218	162	162	hypothetical protein [Escherichia coli]	2,00E-15	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#17653	2631400	2631558	159	159	hypothetical protein, partial [Escherichia coli]	2,00E-11	-2	#2825	2631132	2632607	1476	Glucose-6-phosphate 1-dehydrogenase
#40377	5208918	5209076	159	159	hypothetical protein [Escherichia coli]	3,00E-09	-2	#5428	5208439	5209146	708	Response regulator protein
#4742	730709	730867	159	154	hypothetical protein [Escherichia coli]	7,00E-07	-2	#0689	730714	731622	909	Citrate lyase beta chain
#48717	3985199	3985357	159	159	hydrogenase expression protein [Nocardiopsis alba]	1,00E-06	-2	#4234	3984606	3985433	828	Methylglyoxal reductase, acetol producing
#5176	787673	787831	159	152	hypothetical protein [Shigella dysenteriae]	1,00E-04	-2	#0751	787630	787824	195	phosphopantetheinyltransferase component of enterobactin synthase multienzyme complex
#52526	3404842	3405000	159	159	hypothetical protein [Escherichia coli]	2,00E-06	-2	#3637	3404804	3406822	2019	Hydrogenase-4 component B
#74300	319954	320112	159	159	hypothetical protein, partial [Escherichia coli]	6,00E-05	-2	#0305	319559	321892	2334	Zinc binding domain protein
#20719	3051504	3051659	156	156	hypothetical protein [Burkholderia contaminans]	4,00E-04	-2	#3300	3051413	3052771	1359	4-hydroxybenzoate transporter
#33502	4867170	4867325	156	104	hypothetical protein [Shigella flexneri]	5,00E-17	-2	#5118	4865618	4867273	1656	Arylsulfatase
#39962	5276386	5276541	156	94	hypothetical protein [Escherichia coli]	2,00E-05	-2	#5492	5276126	5276479	354	putative membrane protein yjel
#4174	648374	648529	156	156	hypothetical protein, partial [Escherichia coli]	8,00E-16	-2	#0612	648175	648696	522	Putative membrane-bound metal-dependent hydrolases
#7134	1082299	1082454	156	129	hypothetical protein, partial [Klebsiella pneumoniae]	3,00E-05	-2	#1050	1082322	1082450	129	IncF plasmid conjugative transfer surface exclusion protein TraT
#9756	1482443	1482598	156	129	hypothetical protein, partial [Klebsiella pneumoniae]	3,00E-05	-2	#1511	1482466	1482594	129	IncF plasmid conjugative transfer surface exclusion protein TraT
#19242	2846745	2846897	153	153	hypothetical protein [Leifsonia xyli]	8,00E-05	-2	#3087	2846519	2847877	1359	Putrescine importer
#32687	4746896	4747048	153	153	hypothetical protein [Escherichia coli]	6,00E-15	-2	#5017	4746670	4747920	1251	Uncharacterized protein YidR
#41710	5019481	5019633	153	153	hypothetical protein [Escherichia coli]	2,00E-15	-2	#5269	5016503	5020687	4185	core protein
#44583	4595486	4595638	153	153	hypothetical protein [Escherichia coli]	2,00E-15	-2	#4854	4592508	4596737	4230	core protein
#70513	866526	866678	153	153	hypothetical protein [Burkholderia glumae]	5,00E-06	-2	#0823	865735	866769	1035	Quinolinate synthetase
#70902	810764	810916	153	153	hypothetical protein [Escherichia coli]	2,00E-15	-2	#0770	807786	811985	4200	core protein
#72365	592032	592184	153	153	hypothetical protein [Escherichia coli]	9,00E-12	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72409	586263	586415	153	153	hypothetical protein [Escherichia coli]	4,00E-13	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#57200	2706004	2706153	150	115	hypothetical protein [Escherichia coli]	1,00E-23	-2	#2914	2705987	2706118	132	hypothetical protein
#20716	3051115	3051261	147	147	hypothetical protein [Acidovorax citrulli]	9,00E-04	-2	#3299	3050373	3051401	1029	Gentisate 1,2-dioxygenase
#31210	4531719	4531865	147	147	hypothetical protein, partial [Escherichia coli]	6,00E-05	-2	#4798	4531634	4532638	1005	Dipeptide transport ATP-binding protein DppF
#36293	5268796	5268942	147	147	hypothetical protein [Escherichia coli]	5,00E-24	-2	#5485	5268789	5269997	1209	C4-dicarboxylate transporter DcuA
#42559	4900659	4900805	147	147	type VI secretion protein, partial [Escherichia coli]	6,00E-10	-2	#5153	4900366	4901127	762	Uridine phosphorylase
#44067	4671929	4672075	147	147	hypothetical protein [Escherichia coli]	1,00E-10	-2	#4928	4671801	4672292	492	hypothetical protein
#49745	3835089	3835235	147	147	hypothetical protein [Escherichia coli]	4,00E-05	-2	#4073	3834721	3835932	1212	Putative deacetylase YgeY
#51270	3601482	3601628	147	147	hypothetical protein [Pseudomonas syringae]	4,00E-04	-2	#3824	3601081	3602529	1449	Succinate-semialdehyde dehydrogenase [NADP+]
#51717	3531359	3531505	147	147	hypothetical protein [Pantoea agglomerans]	2,00E-07	-2	#3751	3531267	3531539	273	Putative outer membrane lipoprotein
#52104	3471070	3471216	147	147	carboxymethylenebutenolidase, partial [Escherichia coli]	5,00E-05	-2	#3696	3470834	3471637	804	Inositol-1-monophosphatase
#60152	2278161	2278307	147	147	hypothetical protein [Escherichia coli]	2,00E-15	-2	#2431	2278153	2278959	807	Tryptophan synthase alpha chain
#63787	1766825	1766971	147	100	hypothetical protein, partial [Shigella flexneri]	2,00E-14	-2	#1885	1766781	1766924	144	hypothetical protein
#65489	1542791	1542937	147	147	hypothetical protein [Escherichia coli]	1,00E-07	-2	#1603	1542663	1543154	492	hypothetical protein
#67187	1317545	1317691	147	113	hypothetical protein [Enterobacter ludwigii]	4,00E-05	-2	#1328	1317579	1317698	120	hypothetical protein

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#68490	1142644	1142790	147	147	hypothetical protein [Escherichia coli]	1,00E-07	-2	#1144	1142516	1143007	492	hypothetical protein
#16564	2480047	2480190	144	144	hypothetical protein [Rubrivivax benzoatilyticus]	9,00E-08	-2	#2660	2478909	2481287	2379	Phosphoenolpyruvate synthase
#36935	5368494	5368637	144	144	hypothetical protein [Kitasatospora cheerisanensis]	2,00E-04	-2	#5589	5367458	5369113	1656	Trehalose-6-phosphate hydrolase
#40941	5123811	5123954	144	144	hypothetical protein, partial [Escherichia coli]	3,00E-19	-2	#5347	5123758	5125389	1632	Sodium-dependent phosphate transporter
#58312	2544457	2544600	144	144	hypothetical protein [Escherichia coli]	4,00E-20	-2	#2726	2542799	2544667	1869	Protease IV
#7962	1211013	1211156	144	144	transcriptional regulator, partial [Escherichia coli]	3,00E-05	-2	#1198	1210844	1211635	792	Alkanesulfonates transport system permease protein
#293	46205	46345	141	134	hypothetical protein, partial [Salmonella enterica]	3,00E-06	-2	#0041	44824	46338	1515	L-carnitine/gamma-butyrobetaine antiporter
#41241	5082072	5082212	141	141	hypothetical protein [Pararhodospirillum photometricum]	1,00E-15	-2	#5318	5081872	5086095	4224	DNA-directed RNA polymerase beta' subunit
#47037	4234007	4234147	141	141	hypothetical protein [Xanthomonas hyacinthi]	5,00E-06	-2	#4483	4233858	4234154	297	DNA-binding protein Fis
#72403	586944	587084	141	141	hypothetical protein [Escherichia coli]	8,00E-15	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#73675	409228	409368	141	141	hypothetical protein [Escherichia coli]	2,00E-06	-2	#0393	409121	411007	1887	Propionate--CoA ligase
#7370	1115133	1115273	141	141	hypothetical protein [Escherichia coli]	1,00E-24	-2	#1100	1114916	1115386	471	putative membrane protein
#9992	1515277	1515417	141	141	hypothetical protein [Escherichia coli]	1,00E-24	-2	#1561	1515060	1515530	471	putative membrane protein
#23484	3436514	3436651	138	138	hypothetical protein [Escherichia coli]	1,00E-19	-2	#3665	3435637	3437172	1536	Inosine-5'-monophosphate dehydrogenase
#27463	3991094	3991231	138	138	hypothetical protein, partial [Escherichia coli]	7,00E-24	-2	#4241	3990721	3991356	636	1-acyl-sn-glycerol-3-phosphate acyltransferase
#40760	5151353	5151490	138	100	hypothetical protein, partial [Escherichia coli]	2,00E-23	-2	#5373	5150112	5151452	1341	Maltoporin (maltose/maltodextrin high-affinity receptor, phage lambda receptor protein)
#41631	5029668	5029805	138	138	hypothetical protein [Escherichia coli]	3,00E-16	-2	#5278	5028343	5030523	2181	Catalase
#54173	3153637	3153774	138	119	hypothetical protein, partial [Escherichia coli]	1,00E-22	-2	#3393	3153656	3154378	723	3-demethylubiquinol 3-O-methyltransferase
#38107	5546720	5546854	135	135	hypothetical protein [Escherichia coli]	3,00E-13	-2	#5746	5546613	5547299	687	RNA methyltransferase, TrmH family, group 1
#46999	4238629	4238763	135	135	hypothetical protein [Akkermansia muciniphila CAG:154]	4,00E-04	-2	#4488	4237619	4239505	1887	RND efflux system, inner membrane transporter CmeB
#5065	772502	772636	135	135	hypothetical protein [Parabacteroides johnsonii CAG:246]	1,00E-04	-2	#0737	772276	773316	1041	Phosphate starvation-inducible ATPase PhoH with RNA binding motif
#58714	2483113	2483247	135	135	hypothetical protein [Erwinia amylovora]	1,00E-08	-2	#2663	2482610	2483656	1047	2-keto-3-deoxy-D-arabino-heptulosonate-7- phosphate synthase I alpha
#12812	1944565	1944696	132	97	hypothetical protein [Escherichia coli]	6,00E-12	-2	#2096	1944600	1945118	519	C-terminal domain of CinA type S
#15175	2284537	2284668	132	132	hypothetical protein [Escherichia coli]	2,00E-06	-2	#2442	2284128	2284718	591	Putative chaperone protein
#6903	1046244	1046375	132	132	hypothetical protein [Rhizobium gallicum]	4,00E-08	-2	#1009	1046198	1046893	696	Aquaporin Z
#72107	629209	629340	132	132	hypothetical protein [Streptomyces lavendulae]	8,00E-04	-2	#0594	628871	629749	879	2-hydroxy-3-oxopropionate reductase
#1877	286210	286338	129	94	hypothetical protein [Escherichia coli]	3,00E-15	-2	#0262	286245	287984	1740	Flagellar biosynthesis protein FlhA
#26710	3884111	3884239	129	129	hypothetical protein, partial [Escherichia coli]	5,00E-11	-2	#4117	3883813	3884472	660	Ribose 5-phosphate isomerase A
#27570	4004462	4004590	129	129	hypothetical protein [Escherichia coli]	2,00E-04	-2	#4254	4003837	4004646	810	transport
#3604	550197	550325	129	129	hypothetical protein [Akkermansia muciniphila CAG:154]	8,00E-04	-2	#0537	547319	550468	3150	RND efflux system, inner membrane transporter CmeB
#42327	4930609	4930737	129	129	hypothetical protein [Pseudomonas aeruginosa]	3,00E-06	-2	#5180	4930544	4933330	2787	DNA polymerase I
#57648	2635288	2635416	129	97	lipid A biosynthesis (KDO)2-(lauroyl)-lipid IVA acyltransferase [Klebsiella pneumoniae]	8,00E-06	-2	#2828	2633942	2635384	1443	Pyruvate kinase
#61860	2018602	2018730	129	115	hypothetical protein [Salmonella enterica]	7,00E-06	-2	#2161	2018579	2018716	138	Stationary-phase-induced ribosome-associated protein
#74551	286090	286218	129	101	hypothetical protein [Shigella dysenteriae]	2,00E-06	-2	#0261	286118	286237	120	hypothetical protein
#16376	2452197	2452322	126	126	hypothetical protein [Azospirillum brasilense]	1,00E-04	-2	#2635	2451893	2452897	1005	L,D-transpeptidase YnhG
#32558	4730558	4730683	126	126	hypothetical protein [Methanosphaera stadtmanae]	3,00E-04	-2	#4998	4729195	4730883	1689	Acetolactate synthase large subunit
#48354	4038852	4038977	126	126	membrane protein [Escherichia coli]	4,00E-08	-2	#4287	4038265	4040010	1746	DNA primase
#51	7669	7794	126	126	hypothetical protein [Pseudomonas fuscovaginae]	1,00E-04	-2	#0007	6546	7976	1431	Putative alanine/glycine transport protein
#55449	2967528	2967653	126	126	hypothetical protein, partial [Escherichia coli]	2,00E-08	-2	#3191	2967340	2970969	3630	Molybdate metabolism regulator

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#60392	2246227	2246352	126	126	hypothetical protein [Escherichia coli]	4,00E-06	-2	#2401	2245583	2246710	1128	hypothetical protein
#69868	957157	957282	126	126	swarming motility protein [Escherichia albertii]	5,00E-15	-2	#0922	957143	959431	2289	ATP-dependent helicase DinG/Rad3
#73674	409387	409512	126	126	hypothetical protein [Xanthomonas translucens]	2,00E-04	-2	#0393	409121	411007	1887	Propionate--CoA ligase
#75763	112352	112477	126	126	hypothetical protein [Salmonella enterica]	6,00E-09	-2	#0100	112236	112823	588	Secretion monitor precursor
#16916	2525730	2525852	123	123	hypothetical protein, partial [Catenibacterium mitsuokai]	6,00E-07	-2	#2707	2524736	2525956	1221	Succinylornithine transaminase
#41727	5017642	5017764	123	123	hypothetical protein, partial [Escherichia coli]	2,00E-09	-2	#5269	5016503	5020687	4185	core protein
#44600	4593647	4593769	123	123	hypothetical protein, partial [Escherichia coli]	2,00E-09	-2	#4854	4592508	4596737	4230	core protein
#46965	4243635	4243757	123	123	hypothetical protein, partial [Escherichia coli]	2,00E-12	-2	#4493	4242766	4243947	1182	Glutamate Aspartate transport system permease protein GltJ
#6270	951911	952033	123	123	hypothetical protein, partial [Escherichia coli]	6,00E-17	-2	#0916	951112	952107	996	putative membrane fusion protein (MFP) component of efflux pump, membrane anchor protein YbhG
#6909	1046793	1046915	123	101	hypothetical protein [Acinetobacter baumannii]	8,00E-05	-2	#1009	1046198	1046893	696	Aquaporin Z
#70919	808925	809047	123	123	hypothetical protein, partial [Escherichia coli]	2,00E-09	-2	#0770	807786	811985	4200	core protein
#72495	574390	574512	123	95	hypothetical protein [Shigella flexneri]	6,00E-05	-2	#0560	574418	574558	141	hypothetical protein
#14871	2241889	2242008	120	120	hypothetical protein [Klebsiella pneumoniae]	2,00E-09	-2	#2397	2239254	2242337	3084	RND efflux system, inner membrane transporter CmeB
#21749	3194516	3194635	120	120	hypothetical protein [Escherichia coli]	5,00E-05	-2	#3429	3194353	3195111	759	2-succinyl-6-hydroxy-2,4-cyclohexadiene-1- carboxylate synthase
#259	42381	42500	120	120	hypothetical protein [Escherichia coli]	7,00E-16	-2	#0039	42305	43522	1218	Crotonobetainyl-CoA:carnitine CoA-transferase
#448	68341	68460	120	100	hypothetical protein [Escherichia coli]	6,00E-09	-2	#0062	68361	70712	2352	DNA polymerase II
#52043	3478973	3479092	120	120	heme ABC transporter ATP-binding protein [Escherichia coli]	6,00E-05	-2	#3704	3478617	3479429	813	2,3-dihydroxy-2,3-dihydro-phenylpropionate dehydrogenase
#73505	431637	431756	120	120	hypothetical protein [Burkholderia phenoliruptrix]	5,00E-06	-2	#0413	431521	432471	951	Acetaldehyde dehydrogenase, acetylating, in cluster for degradation of phenols, cresols, catechol
#76314	33833	33952	120	104	hypothetical protein [Shimwellia blattae]	1,00E-04	-2	#0031	33849	33986	138	hypothetical protein
#18320	2713988	2714104	117	117	hypothetical protein [Streptomyces bikiniensis]	5,00E-05	-2	#2924	2713795	2714547	753	Cystine ABC transporter, ATP-binding protein
#21192	3119566	3119682	117	117	hypothetical protein [Salmonella enterica]	2,00E-04	-2	#3367	3119529	3120131	603	Cytochrome c-type protein NapC
#36239	5262515	5262631	117	117	hypothetical protein, partial [Escherichia coli]	1,00E-06	-2	#5480	5262013	5263347	1335	Lysine/cadaverine antiporter membrane protein CadB
#45177	4504667	4504783	117	117	hypothetical protein [Escherichia coli]	2,00E-08	-2	#4776	4504548	4505870	1323	Inner membrane metabolite transport protein YhjE
#56983	2738105	2738221	117	117	hypothetical protein [Escherichia coli]	6,00E-14	-2	#2951	2737395	2738522	1128	Flagellar hook-length control protein FliK
#58826	2467455	2467571	117	117	hypothetical protein [Escherichia coli]	2,00E-11	-2	#2648	2466685	2467950	1266	hypothetical protein
#22046	3237105	3237218	114	114	hypothetical protein [Streptomyces bikiniensis]	8,00E-05	-2	#3472	3236912	3237685	774	Histidine ABC transporter, ATP-binding protein HisP
#23179	3387633	3387746	114	114	hypothetical protein [Salmonella enterica]	3,00E-08	-2	#3622	3386561	3388540	1980	Glutamate synthase [NADPH] small chain
#43910	4699869	4699982	114	114	hypothetical protein [Escherichia coli]	2,00E-13	-2	#4963	4699000	4700538	1539	Type III secretion outermembrane pore forming protein (YscC,MxiD,HrcC, InvG)
#55029	3026923	3027036	114	114	regulator [Escherichia coli]	1,00E-08	-2	#3269	3026876	3027058	183	hypothetical protein
#5905	894290	894403	114	114	regulator [Escherichia coli]	3,00E-07	-2	#0854	894268	894450	183	hypothetical protein
#61632	2052329	2052442	114	114	membrane protein [Escherichia coli]	1,00E-12	-2	#2191	2052309	2052455	147	hypothetical protein
#8873	1339154	1339267	114	114	regulator [Escherichia coli]	1,00E-08	-2	#1357	1339132	1339314	183	hypothetical protein
#27513	3996520	3996630	111	111	hypothetical protein [Escherichia coli]	2,00E-04	-2	#4245	3996339	3996731	393	Protein ygiW precursor
#29053	4219337	4219447	111	111	hypothetical protein, partial [Escherichia coli]	3,00E-04	-2	#4468	4218913	4220382	1470	Cytoplasmic axial filament protein CafA and Ribonuclease G
#39943	5279684	5279794	111	95	hypothetical protein [Escherichia coli]	2,00E-14	-2	#5497	5279700	5279846	147	Entericidin B precursor
#42755	4875391	4875501	111	111	hypothetical protein [Shigella dysenteriae]	1,00E-06	-2	#5126	4875323	4875619	297	hypothetical protein
#48261	4050669	4050779	111	111	hypothetical protein, partial [Escherichia coli]	3,00E-11	-2	#4298	4049791	4052883	3093	Evolved beta-D-galactosidase, alpha subunit
#52676	3384267	3384377	111	111	hypothetical protein [Escherichia coli]	6,00E-07	-2	#3619	3382651	3384654	2004	Transketolase
#56001	2886638	2886748	111	111	hypothetical protein [Escherichia albertii]	8,00E-04	-2	#3121	2886597	2886755	159	hypothetical protein

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#61987	2004137	2004247	111	111	hypothetical protein [Escherichia coli]	2,00E-13	-2	#2148	2004051	2005586	1536	putative glutamate/gamma-aminobutyrate antiporter
#70288	900802	900912	111	111	hypothetical protein [Escherichia coli]	8,00E-14	-2	#0862	900797	900955	159	hypothetical protein
#72378	590238	590348	111	111	hypothetical protein [Escherichia coli]	4,00E-07	-2	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#74131	346697	346807	111	111	hypothetical protein, partial [Escherichia coli]	1,00E-09	-2	#0332	346224	347762	1539	IS66 transposase
#31222	4533056	4533163	108	108	hypothetical protein [Clostridium bolteae CAG:59]	1,00E-05	-2	#4799	4532635	4533618	984	Dipeptide transport ATP-binding protein DppD
#3391	516024	516131	108	108	hypothetical protein [Klebsiella pneumoniae]	1,00E-06	-2	#0501	514703	516694	1992	Cytochrome O ubiquinol oxidase subunit I
#46551	4304949	4305056	108	108	hypothetical protein [Escherichia coli]	1,00E-05	-2	#4572	4304536	4305342	807	Nitrite transporter NirC
#48056	4078078	4078185	108	108	sugar isomerase, partial [Escherichia coli]	1,00E-04	-2	#4324	4077731	4078213	483	Inner membrane protein YqjF
#65039	1600312	1600419	108	108	copper sensitivity supression protein, partial [Salmonella enterica]	1,00E-04	-2	#1673	1600130	1600939	810	Aminodeoxychorismate lyase
#11900	1803991	1804095	105	105	hypothetical protein [Escherichia coli]	7,00E-06	-2	#1925	1803777	1804130	354	Putative ACR protein
#1433	222989	223093	105	105	hypothetical protein, partial [Escherichia coli]	2,00E-08	-2	#0202	222922	223275	354	Protein RcsF
#37863	5509646	5509750	105	105	hypothetical protein [Enterobacter cloacae]	3,00E-06	-2	#5707	5509318	5510142	825	putative membrane protein
#49164	3920771	3920875	105	105	MULTISPECIES: hypothetical protein [Enterobacteriaceae]	1,00E-04	-2	#4157	3920259	3920990	732	Ribosomal RNA small subunit methyltransferase E
#50774	3668657	3668761	105	105	hypothetical protein [Xanthomonas vasicola]	8,00E-06	-2	#3904	3668403	3670964	2562	DNA mismatch repair protein MutS
#52463	3411943	3412047	105	105	NADH dehydrogenase [Escherichia coli]	3,00E-04	-2	#3642	3411473	3413188	1716	Hydrogenase-4 component G
#65235	1575125	1575229	105	105	hypothetical protein [Escherichia coli]	4,00E-08	-2	#1645	1574718	1576253	1536	putative peptidoglycan lipid II flippase MurJ
#69635	992125	992229	105	105	hypothetical protein [Clostridium bolteae CAG:59]	9,00E-05	-2	#0954	991625	993496	1872	Glutothione ABC transporter ATP-binding protein
#75510	149525	149629	105	105	hypothetical protein [Escherichia coli]	5,00E-12	-2	#0133	149454	150683	1230	Polysaccharide deacetylase
#7588	1147714	1147818	105	105	hypothetical protein [Escherichia coli]	1,00E-13	-2	#1150	1147605	1149371	1767	Transport ATP-binding protein CydD
#11019	1675965	1676066	102	102	hypothetical protein [Escherichia coli]	2,00E-07	-2	#1768	1675451	1676314	864	Spermidine Putrescine ABC transporter permease component PotB
#21647	3179966	3180067	102	102	hypothetical protein [Salmonella enterica]	2,00E-08	-2	#3414	3179950	3181152	1203	Molybdopterin binding motif, CinA N-terminal domain
#26867	3905569	3905670	102	102	SAM-dependent methyltransferase, partial [Escherichia coli]	1,00E-08	-2	#4140	3904416	3905693	1278	Putative oxidoreductase linked to yggC
#1277	194120	194218	99	99	hypothetical protein [Salmonella bongori]	9,00E-04	-2	#0173	194038	194325	288	hypothetical protein
#28499	4142911	4143009	99	99	methyltransferase, partial [Escherichia coli]	1,00E-08	-2	#4396	4142901	4143302	402	Ribosome-binding factor A
#63197	1854061	1854159	99	99	hypothetical protein [Escherichia coli]	8,00E-12	-2	#1977	1853417	1854163	747	Putative intestinal colonization factor encoded by prophage CP-933O
#71234	759814	759912	99	99	hypothetical protein [Escherichia coli]	1,00E-05	-2	#0721	758552	759979	1428	hypothetical protein
#26983	3923103	3923198	96	96	hypothetical protein [Shigella flexneri]	2,00E-05	-2	#4161	3923093	3923224	132	hypothetical protein
#2982	454964	455059	96	96	hypothetical protein [Escherichia coli]	1,00E-06	-2	#0433	454909	456066	1158	Penicillin-binding protein AmpH
#32490	4722257	4722352	96	96	hypothetical protein [Escherichia coli]	7,00E-07	-2	#4991	4722061	4723380	1320	Hexose phosphate uptake regulatory protein UhpC
#71703	690310	690405	96	96	hypothetical protein, partial [Escherichia coli]	5,00E-11	-2	#0648	690302	690436	135	hypothetical protein
#73329	457392	457487	96	96	hypothetical protein [Escherichia coli]	1,00E-11	-2	#0435	456418	457638	1221	SbmA protein
#73380	449198	449293	96	96	hypothetical protein, partial [Escherichia coli]	1,00E-05	-2	#0429	448797	449648	852	Alpha-ketoglutarate-dependent taurine dioxygenase
#8705	1318421	1318516	96	96	hypothetical protein [Escherichia coli]	1,00E-09	-2	#1330	1318114	1319187	1074	putative iron-sulfur cluster binding protein YccM
#37776	5495834	5495926	93	93	ubiquinone biosynthesis protein UbiB, partial [Escherichia coli]	4,00E-06	-2	#5696	5495665	5497779	2115	Carbon starvation protein A
#45512	4457747	4457839	93	93	hypothetical protein [Escherichia coli]	1,00E-10	-2	#4727	4456947	4457999	1053	hypothetical protein
#48543	4012468	4012560	93	93	hypothetical protein, partial [Escherichia coli]	7,00E-06	-2	#4261	4011836	4013317	1482	Type I secretion outer membrane protein, TolC precursor
#67952	1214875	1214967	93	93	hypothetical protein [Shigella flexneri]	2,00E-04	-2	#1202	1214672	1215211	540	type 1 fimbriae major subunit FimA
#41259	5078817	5082005	3189	2979	hypothetical protein, partial [Staphylococcus aureus]	3,00E-13	-1	#5317	5077767	5081795	4029	DNA-directed RNA polymerase beta subunit
#41233	5083558	5086020	2463	2463	hypothetical protein, partial [Staphylococcus aureus]	1,00E-04	-1	#5318	5081872	5086095	4224	DNA-directed RNA polymerase beta' subunit

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#75650	127756	130029	2274	2274	hypothetical protein [Streptococcus pneumoniae]	2,00E-04	-1	#0116	127507	130170	2664	Pyruvate dehydrogenase E1 component
#7707	1168256	1170508	2253	2253	hypothetical protein [Bifidobacterium adolescentis CAG:119]	1,00E-67	-1	#1166	1168247	1170529	2283	Pyruvate formate-lyase
#28483	4139094	4141316	2223	2136	hypothetical protein [Alistipes finegoldii CAG:68]	6,00E-08	-1	#4393	4139157	4141292	2136	Polyribonucleotide nucleotidyltransferase
#76450	12138	14174	2037	1917	NAD-specific glutamate dehydrogenase [Halorubrum coriense]	1,00E-41	-1	#0013	12180	14096	1917	Chaperone protein DnaK
#54145	3158877	3160754	1878	1788	ribonucleotide-diphosphate reductase subunit alpha, partial [Klebsiella pneumoniae]	9,00E-06	-1	#3395	3158967	3161252	2286	Ribonucleotide reductase of class Ia (aerobic), alpha subunit
#76363	26818	28644	1827	1827	hypothetical protein [Akkermansia muciniphila CAG:154]	2,00E-19	-1	#0025	26803	29619	2817	Isoleucyl-tRNA synthetase
#25098	3655946	3657766	1821	117	hypothetical protein, partial [Bacillus cereus]	1,00E-40	-1	#3889	3655520	3656062	543	Formate hydrogenlyase complex 3 iron-sulfur protein
#12055	1825182	1826999	1818	1818	hypothetical protein [Bifidobacterium adolescentis CAG:119]	4,00E-20	-1	#1943	1824675	1827350	2676	Alcohol dehydrogenase
#70708	836337	838142	1806	1593	hypothetical protein [Mycobacterium tuberculosis]	4,00E-15	-1	#0795	835128	837929	2802	2-oxoglutarate dehydrogenase E1 component
#26893	3908097	3909866	1770	1770	hypothetical protein, partial [Klebsiella pneumoniae]	2,00E-152	-1	#4144	3907881	3909872	1992	Transketolase
#31160	4522724	4524409	1686	1686	hypothetical protein, partial [Escherichia coli]	2,00E-57	-1	#4788	4522388	4525006	2619	Cellulose synthase catalytic subunit [UDP-forming]
#10469	1588035	1589717	1683	1683	hypothetical protein, partial [Escherichia coli]	4,00E-29	-1	#1661	1588023	1591208	3186	Ribonuclease E
#4924	753303	754982	1680	1647	hypothetical protein [Streptococcus salivarius CAG:79]	5,00E-12	-1	#0716	753336	755918	2583	Leucyl-tRNA synthetase
#38333	5516068	5517666	1599	174	hypothetical protein [Bacteroides uniformis CAG:3]	8,00E-25	-1	#5715	5515564	5516241	678	5'-nucleotidase YjjG
#39976	5274258	5275856	1599	1515	hypothetical protein, partial [Streptococcus anginosus]	8,00E-35	-1	#5491	5274342	5275988	1647	Heat shock protein 60 family chaperone GroEL
#72582	561324	562916	1593	1566	hypothetical protein, partial [Escherichia coli]	1,00E-16	-1	#0548	561015	562889	1875	Chaperone protein HtpG
#29487	4279497	4281083	1587	1551	hypothetical protein, partial [Streptococcus pyogenes]	5,00E-33	-1	#4543	4278933	4281047	2115	Translation elongation factor G
#68237	1179006	1180580	1575	1575	hypothetical protein [Parabacteroides johnsonii CAG:246]	1,00E-17	-1	#1174	1178970	1180643	1674	SSU ribosomal protein S1p
#37781	5495668	5497218	1551	1551	hypothetical protein [Escherichia coli]	3,00E-28	-1	#5696	5495665	5497779	2115	Carbon starvation protein A
#41242	5081965	5083506	1542	1542	hypothetical protein [Bifidobacterium bifidum CAG:234]	1,00E-29	-1	#5318	5081872	5086095	4224	DNA-directed RNA polymerase beta' subunit
#28594	4155297	4156832	1536	1536	hypothetical protein [Bifidobacterium longum CAG:69]	2,00E-42	-1	#4406	4155126	4157060	1935	Cell division protein FtsH
#36310	5270145	5271641	1497	1434	hypothetical protein [Halorubrum kocurii]	2,00E-11	-1	#5486	5270208	5271644	1437	Aspartate ammonia-lyase
#42233	4942111	4943604	1494	1494	hypothetical protein, partial [Elizabethkingia anophelis]	4,00E-16	-1	#5190	4941832	4943655	1824	GTP-binding protein TypA/BipA
#28526	4145987	4147465	1479	153	hypothetical protein, partial [Bacillus cereus]	3,00E-13	-1	#4397	4143467	4146139	2673	Translation initiation factor 2
#67131	1325174	1326619	1446	1446	hypothetical protein [Escherichia coli]	1,00E-66	-1	#1335	1325081	1327627	2547	Trimethylamine-N-oxide reductase
#21911	3215321	3216763	1443	1443	hypothetical protein [Mycobacterium intracellulare]	2,00E-06	-1	#3449	3215231	3217033	1803	NADH-ubiquinone oxidoreductase chain C
#59292	2402031	2403467	1437	1299	hypothetical protein, partial [Staphylococcus aureus]	3,00E-05	-1	#2584	2401011	2403329	2319	Electron transport complex protein RnfC
#26530	3859549	3860964	1416	1401	hypothetical protein [Coprococcus comes CAG:19]	3,00E-27	-1	#4091	3859432	3860949	1518	Lysyl-tRNA synthetase (class II)
#28505	4143383	4144783	1401	1317	hypothetical protein, partial [Bacillus cereus]	3,00E-20	-1	#4397	4143467	4146139	2673	Translation initiation factor 2
#72860	525222	526613	1392	1392	hypothetical protein, partial [Bacillus cereus]	6,00E-22	-1	#0511	524940	527294	2355	ATP-dependent protease La Type I
#37059	5385284	5386660	1377	1377	hypothetical protein [Escherichia coli]	3,00E-26	-1	#5607	5385263	5388118	2856	Valyl-tRNA synthetase
#55523	2958406	2959773	1368	1368	hypothetical protein, partial [Klebsiella pneumoniae]	7,00E-84	-1	#3186	2958349	2960382	2034	Methionyl-tRNA synthetase
#58930	2449928	2451274	1347	1347	hypothetical protein, partial [Bacteroides sartorii]	4,00E-05	-1	#2633	2449871	2451283	1413	Pyruvate kinase
#39446	5350375	5351697	1323	1323	hypothetical protein, partial [Staphylococcus aureus]	4,00E-32	-1	#5569	5348458	5352237	3780	Uncharacterized protein YtfN
#28456	4136311	4137630	1320	1320	hypothetical protein, partial [Streptomyces purpeofuscus]	2,00E-18	-1	#4391	4136095	4137984	1890	Cold-shock DEAD-box protein A
#63505	1812256	1813569	1314	1314	hypothetical protein, partial [Escherichia coli]	7,00E-48	-1	#1931	1810273	1814016	3744	Respiratory nitrate reductase alpha chain
#73137	483033	484337	1305	1305	hypothetical protein, partial [Streptococcus pyogenes]	2,00E-12	-1	#0465	483021	484340	1320	Proline-specific permease proY
#70583	857214	858494	1281	96	hypothetical protein [Xanthomonas oryzae]	4,00E-11	-1	#0812	855741	857309	1569	Cytochrome d ubiquinol oxidase subunit I
#48262	4050649	4051905	1257	1257	hypothetical protein [Bacteroides intestinalis CAG:564]	1,00E-15	-1	#4298	4049791	4052883	3093	Evolved beta-D-galactosidase, alpha subunit

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#29479	4277768	4279006	1239	1095	elongation factor TU [Bifidobacterium bifidum CAG:234]	2,00E-55	-1	#4542	4277678	4278862	1185	Translation elongation factor Tu
#63514	1811035	1812255	1221	1221	hypothetical protein, partial [Escherichia coli]	8,00E-175	-1	#1931	1810273	1814016	3744	Respiratory nitrate reductase alpha chain
#26803	3896525	3897742	1218	1164	phosphoglycerate kinase [Bifidobacterium bifidum CAG:234]	3,00E-16	-1	#4130	3896561	3897724	1164	Phosphoglycerate kinase
#25517	3716638	3717852	1215	1185	hypothetical protein, partial [Bacteroides sartorii]	3,00E-16	-1	#3957	3716668	3717966	1299	Enolase
#17763	2645568	2646746	1179	1179	hypothetical protein [Parabacteroides merdae CAG:48]	2,00E-18	-1	#2840	2645142	2646875	1734	Aspartyl-tRNA synthetase
#39749	5307450	5308619	1170	1128	hypothetical protein [Bifidobacterium longum CAG:69]	5,00E-17	-1	#5522	5307279	5308577	1299	Adenylosuccinate synthetase
#43543	4759239	4760408	1170	1170	hypothetical protein [Salmonella enterica]	2,00E-24	-1	#5029	4758984	4760579	1596	Inner membrane protein translocase component YidC, long form
#68354	1161798	1162958	1161	618	hypothetical protein, partial [Salmonella enterica]	2,00E-18	-1	#1159	1161858	1162475	618	Anaerobic dimethyl sulfoxide reductase chain B
#27274	3964757	3965908	1152	1152	hypothetical protein [Escherichia coli]	1,00E-41	-1	#4210	3964643	3966502	1860	Glutathionylspermidine synthase
#26792	3895294	3896442	1149	1053	hypothetical protein [Escherichia coli]	2,00E-14	-1	#4129	3895267	3896346	1080	Fructose-bisphosphate aldolase class II
#38697	5455387	5456523	1137	1137	hypothetical protein [Escherichia coli]	3,00E-68	-1	#5657	5455342	5456526	1185	Mannonate dehydratase
#64241	1702340	1703467	1128	1044	hypothetical protein [Parabacteroides merdae CAG:48]	3,00E-46	-1	#1805	1702424	1703674	1251	Isocitrate dehydrogenase [NADP]
#74632	273954	275081	1128	1128	hypothetical protein, partial [Escherichia coli]	1,00E-66	-1	#0244	273792	275552	1761	core protein
#32730	4752921	4754036	1116	1056	hypothetical protein [Bifidobacterium adolescentis CAG:119]	2,00E-12	-1	#5022	4751562	4753976	2415	DNA gyrase subunit B
#41300	5072572	5073681	1110	1095	elongation factor TU [Bifidobacterium bifidum CAG:234]	1,00E-53	-1	#5310	5072587	5073771	1185	Translation elongation factor Tu
#41264	5077692	5078795	1104	1029	hypothetical protein [Cronobacter dublinensis]	4,00E-24	-1	#5317	5077767	5081795	4029	DNA-directed RNA polymerase beta subunit
#76298	36452	37555	1104	1104	hypothetical protein [Prevotella copri CAG:164]	4,00E-19	-1	#0033	35192	38413	3222	Carbamoyl-phosphate synthase large chain
#36455	5291812	5292906	1095	150	hypothetical protein [Escherichia coli]	1,00E-46	-1	#5507	5288638	5291961	3324	Potassium efflux system KefA protein
#47592	4148735	4149826	1092	1065	hypothetical protein, partial [Piscirickettsia salmonis]	1,00E-46	-1	#4401	4148762	4150105	1344	Argininosuccinate synthase
#31422	4556341	4557411	1071	1071	hypothetical protein [Escherichia coli]	9,00E-12	-1	#4823	4556122	4558191	2070	Glycyl-tRNA synthetase beta chain
#21845	3207143	3208207	1065	1065	hypothetical protein [Bacteroides eggerthii CAG:109]	8,00E-06	-1	#3441	3206381	3208216	1836	NADH-ubiquinone oxidoreductase chain L
#41119	5099969	5101030	1062	1062	hypothetical protein, partial [Acidovorax avenae]	8,00E-13	-1	#5336	5099750	5101075	1326	Response regulator of zinc sigma-54-dependent two-component system
#29446	4273020	4274075	1056	822	hypothetical protein [Dermatophilus congolensis]	1,00E-13	-1	#4534	4273182	4274003	822	LSU ribosomal protein L2p (L8e)
#71009	793326	794372	1047	1014	hypothetical protein [Bifidobacterium bifidum CAG:234]	4,00E-07	-1	#0760	793359	794999	1641	Phosphoglucomutase
#26908	3910961	3912004	1044	891	hypothetical protein, partial [Bacillus cereus]	3,00E-23	-1	#4146	3911114	3912034	921	Agmatinase
#42446	4914997	4916034	1038	1038	hypothetical protein, partial [Catenibacterium mitsuokai]	5,00E-05	-1	#5167	4914907	4916238	1332	Xaa-Pro dipeptidase PepQ
#33077	4800044	4801072	1029	882	hypothetical protein [Bacteroides clarus CAG:160]	1,00E-15	-1	#5064	4799384	4800925	1542	ATP synthase alpha chain
#73800	392184	393203	1020	522	hypothetical protein [Escherichia coli]	1,00E-05	-1	#0375	391755	392705	951	Carbamate kinase
#63591	1799194	1800210	1017	855	hypothetical protein [Halomonas smyrnensis]	1,00E-25	-1	#1919	1799233	1800087	855	2-Keto-3-deoxy-D-manno-octulosonate-8-phosphate synthase
#36088	5242682	5243689	1008	1008	hypothetical protein, partial [[Haemophilus] parasuis]	1,00E-07	-1	#5463	5241569	5243839	2271	Arginine decarboxylase, catabolic
#5297	803438	804427	990	990	hypothetical protein, partial [Pseudomonas syringae]	5,00E-13	-1	#0767	803192	805240	2049	Potassium-transporting ATPase B chain
#75756	113245	114234	990	990	hypothetical protein [Clostridium clostridioforme CAG:132]	2,00E-08	-1	#0101	112885	115590	2706	Protein export cytoplasm protein SecA ATPase RNA helicase
#4364	674096	675082	987	987	hypothetical protein [Photorhabdus luminescens]	2,00E-12	-1	#0632	673265	675166	1902	VgrG protein
#16570	2480286	2481269	984	984	hypothetical protein [Rubrivivax benzoatilyticus]	5,00E-39	-1	#2660	2478909	2481287	2379	Phosphoenolpyruvate synthase
#46570	4302042	4303025	984	984	hypothetical protein, partial [Lysinimicrobium mangrovi]	2,00E-14	-1	#4570	4301544	4304087	2544	Nitrite reductase [NAD(P)H] large subunit
#73522	429758	430738	981	867	hypothetical protein [Pseudomonas nitroreducens]	2,00E-24	-1	#0411	429839	430705	867	2-hydroxy-6-ketonona-2,4-dienedioic acid hydrolase
#27313	3969944	3970921	978	978	hypothetical protein [Sutterella wadsworthensis CAG:135]	4,00E-10	-1	#4216	3969266	3970969	1704	Uptake hydrogenase large subunit
#21460	3152376	3153347	972	972	val start codon [Prevotella copri CAG:164]	2,00E-31	-1	#3392	3150882	3153509	2628	DNA gyrase subunit A
#33122	4805481	4806452	972	972	hypothetical protein [Acidaminococcus intestini CAG:325]	1,00E-26	-1	#5071	4804812	4806701	1890	tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#73669	410114	411085	972	894	hypothetical protein [Escherichia coli]	1,00E-08	-1	#0393	409121	411007	1887	Propionate--CoA ligase
#17657	2631408	2632376	969	969	hypothetical protein [Escherichia coli]	6,00E-06	-1	#2825	2631132	2632607	1476	Glucose-6-phosphate 1-dehydrogenase
#17147	2556497	2557459	963	900	hypothetical protein, partial [Rhodococcus qingshengii]	7,00E-89	-1	#2740	2556320	2557396	1077	Putative oxidoreductase YdjL
#2862	436704	437663	960	960	hypothetical protein [Mycobacterium marinum]	2,00E-10	-1	#0418	436560	437669	1110	S-(hydroxymethyl)glutathione dehydrogenase
#68370	1159619	1160575	957	957	hypothetical protein [Sodalis glossinidius]	7,00E-17	-1	#1158	1159403	1161847	2445	Anaerobic dimethyl sulfoxide reductase chain A
#75601	136717	137667	951	951	hypothetical protein, partial [Escherichia coli]	8,00E-05	-1	#0120	136066	138585	2520	Aconitate hydratase 2
#23919	3491814	3492761	948	948	hypothetical protein, partial [Pseudomonas aeruginosa]	3,00E-76	-1	#3715	3491511	3492764	1254	Serine hydroxymethyltransferase
#33705	4894964	4895911	948	756	membrane protein [Polaribacter irgensii]	1,00E-87	-1	#5149	4895018	4895773	756	hypothetical protein
#13257	2022994	2023938	945	885	hypothetical protein, partial [Piscirickettsia salmonis]	3,00E-18	-1	#2166	2023018	2023902	885	Formate dehydrogenase O beta subunit
#47080	4228669	4229613	945	945	choline dehydrogenase [Dialister invisus CAG:218]	2,00E-22	-1	#4478	4228504	4229853	1350	Biotin carboxylase of acetyl-CoA carboxylase
#4309	666278	667216	939	939	membrane protein, partial [Escherichia coli]	8,00E-81	-1	#0628	665891	667225	1335	Rhs-family protein
#70591	856149	857087	939	939	hypothetical protein, partial [Escherichia coli]	1,00E-36	-1	#0812	855741	857309	1569	Cytochrome d ubiquinol oxidase subunit I
#51804	3519947	3520882	936	723	hypothetical protein, partial [Streptomyces purpeofuscus]	2,00E-04	-1	#3741	3520160	3521494	1335	ATP-dependent RNA helicase SrmB
#68123	1191254	1192189	936	897	multidrug transporter, partial [Escherichia coli]	8,00E-05	-1	#1185	1191293	1192615	1323	Chromosome partition protein MukF
#75065	211896	212831	936	867	hypothetical protein [Ketogulonicigenium vulgare]	2,00E-24	-1	#0190	211965	212924	960	Acetyl-coenzyme A carboxyl transferase alpha chain
#22102	3241996	3242928	933	933	hypothetical protein, partial [Acinetobacter baumannii]	2,00E-32	-1	#3478	3241897	3243414	1518	Amidophosphoribosyltransferase
#44181	4655506	4656438	933	489	hypothetical protein [Enterobacter cloacae]	7,00E-08	-1	#4913	4653886	4655994	2109	GTP pyrophosphokinase, (p)ppGpp synthetase II
#69709	980042	980971	930	930	hypothetical protein [Ruminococcus obeum CAG:39]	4,00E-14	-1	#0943	980039	981631	1593	Putative ATPase component of ABC transporter with duplicated ATPase domain
#22956	3359625	3360551	927	900	hypothetical protein, partial [Piscirickettsia salmonis]	1,00E-63	-1	#3594	3359640	3360539	900	putative dye-decolorizing peroxidase (DyP), YfeX-like subgroup
#37067	5386715	5387641	927	927	hypothetical protein [Bacteroides caccae CAG:21]	1,00E-22	-1	#5607	5385263	5388118	2856	Valyl-tRNA synthetase
#72178	619517	620434	918	918	hypothetical protein [Delftia acidovorans]	8,00E-06	-1	#0585	616748	620944	4197	core protein
#23124	3379386	3380300	915	915	hypothetical protein [Achromobacter xylosoxidans]	2,00E-35	-1	#3617	3379113	3381392	2280	NADP-dependent malic enzyme
#35957	5226173	5227087	915	822	hypothetical protein, partial [Escherichia coli]	4,00E-14	-1	#5449	5226164	5226994	831	Phosphonate ABC transporter permease protein phnE
#65520	1539188	1540099	912	432	hypothetical protein [Rhodococcus qingshengii]	7,00E-37	-1	#1593	1539134	1539619	486	Antirestriction protein klcA
#68521	1139041	1139952	912	432	hypothetical protein [Rhodococcus qingshengii]	7,00E-37	-1	#1134	1138987	1139472	486	Antirestriction protein klcA
#48100	4072600	4073502	903	903	hypothetical protein, partial [Bacillus cereus]	1,00E-25	-1	#4315	4072483	4073781	1299	Hexuronate transporter
#74703	264946	265848	903	903	hypothetical protein [Photorhabdus luminescens]	4,00E-08	-1	#0239	264862	267003	2142	VgrG protein
#23972	3498892	3499791	900	861	hypothetical protein [Escherichia coli]	5,00E-100	-1	#3722	3498931	3502818	3888	Phosphoribosylformylglycinamidine synthase, synthetase subunit
#38268	5525417	5526316	900	849	hypothetical protein, partial [Bacillus cereus]	5,00E-25	-1	#5726	5524943	5526265	1323	Thymidine phosphorylase
#75595	137722	138618	897	864	hypothetical protein, partial [Pseudomonas syringae]	3,00E-05	-1	#0120	136066	138585	2520	Aconitate hydratase 2
#64636	1652692	1653585	894	894	hypothetical protein [Escherichia coli]	6,00E-95	-1	#1741	1652149	1654086	1938	hypothetical protein
#39725	5310740	5311630	891	891	amino acid permease, partial [Escherichia coli]	6,00E-05	-1	#5524	5309246	5311687	2442	3'-to-5' exoribonuclease RNase R
#31107	4516500	4517387	888	888	hypothetical protein, partial [Escherichia coli]	2,00E-37	-1	#4784	4515462	4518470	3009	Cellulose synthase operon protein C
#35588	5174747	5175634	888	888	hypothetical protein, partial [Vibrio parahaemolyticus]	1,00E-29	-1	#5396	5173061	5175883	2823	Excinuclease ABC subunit A
#37119	5394638	5395525	888	858	hypothetical protein [Hafnia alvei]	1,00E-04	-1	#5614	5394668	5395687	1020	Alcohol dehydrogenase
#41011	5113705	5114589	885	735	hypothetical protein [Shigella flexneri]	4,00E-33	-1	#5342	5113135	5114439	1305	Isocitrate lyase
#4322	668275	669159	885	885	membrane protein, partial [Escherichia coli]	8,00E-77	-1	#0630	667819	672756	4938	Rhs-family protein
#23750	3468193	3469071	879	879	hypothetical protein [Burkholderia sprentiae]	6,00E-16	-1	#3693	3467863	3469077	1215	Cysteine desulfurase, IscS subfamily
#36989	5375024	5375902	879	294	hypothetical protein [Escherichia coli]	4,00E-23	-1	#5593	5374931	5375317	387	Endoribonuclease L-PSP

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#58214	2558940	2559818	879	879	glyceraldehyde-3-phosphate dehydrogenase [Bacteroides stercoris CAG:120]	2,00E-18	-1	#2743	2558835	2559830	996	NAD-dependent glyceraldehyde-3-phosphate dehydrogenase
#23478	3434708	3435571	864	861	cytosine deaminase [Acidaminococcus intestini CAG:325]	4,00E-20	-1	#3664	3433991	3435568	1578	GMP synthase [glutamine-hydrolyzing], amidotransferase subunit
#23976	3499798	3500661	864	864	hypothetical protein, partial [Escherichia coli]	2,00E-74	-1	#3722	3498931	3502818	3888	Phosphoribosylformylglycinamidine synthase, synthetase subunit
#32721	4751376	4752239	864	678	hypothetical protein [Escherichia coli]	3,00E-08	-1	#5022	4751562	4753976	2415	DNA gyrase subunit B
#72961	507995	508858	864	864	hypothetical protein, partial [Staphylococcus epidermidis]	7,00E-41	-1	#0493	507602	509050	1449	tRNA S(4)U 4-thiouridine synthase (former ThiI)
#34427	5003187	5004044	858	651	hypothetical protein [Parabacteroides johnsonii CAG:246]	4,00E-12	-1	#5255	5003394	5004902	1509	Glycerol kinase
#75627	132879	133736	858	831	hypothetical protein, partial [Rhodococcus rhodochrous]	2,00E-11	-1	#0118	132222	133709	1488	Dihydrolipoamide dehydrogenase of pyruvate dehydrogenase complex
#47329	4189054	4189905	852	111	hypothetical protein, partial [Rhodococcus opacus]	2,00E-43	-1	#4440	4184611	4189164	4554	Glutamate synthase [NADPH] large chain
#5773	874712	875560	849	849	hypothetical protein, partial [Bacteroides sartorii]	2,00E-08	-1	#0833	874682	875698	1017	UDP-N-acetylglucosamine 4-epimerase
#13401	2046258	2047103	846	846	hypothetical protein [Escherichia coli]	2,00E-65	-1	#2189	2045670	2049872	4203	core protein
#5132	780567	781409	843	828	hypothetical protein [Escherichia coli]	3,00E-56	-1	#0744	780174	781394	1221	N-acetylglucosamine-6P-responsive transcriptional repressor NagC, ROK family
#71480	721530	722372	843	843	hypothetical protein [Escherichia coli]	2,00E-05	-1	#0678	720855	722450	1596	Alkyl hydroperoxide reductase protein F
#28026	4070804	4071637	834	834	hypothetical protein, partial [Bacillus cereus]	3,00E-04	-1	#4314	4070588	4072000	1413	Uronate isomerase
#70740	832299	833132	834	804	hypothetical protein [Klebsiella pneumoniae]	6,00E-22	-1	#0792	832329	834095	1767	Succinate dehydrogenase flavoprotein subunit
#75085	209571	210401	831	831	na /proline symporter [Clostridium clostridioforme CAG:511]	1,00E-19	-1	#0189	208470	211952	3483	DNA polymerase III alpha subunit
#1424	221166	221993	828	828	hypothetical protein [Sutterella wadsworthensis CAG:135]	5,00E-22	-1	#0200	220389	222107	1719	Prolyl-tRNA synthetase, bacterial type
#68105	1193324	1194151	828	828	cell division protein MukB, partial [Klebsiella pneumoniae]	9,00E-12	-1	#1187	1193300	1197760	4461	Chromosome partition protein MukB
#67998	1208388	1209206	819	819	hypothetical protein, partial [Escherichia coli]	4,00E-145	-1	#1196	1207425	1210037	2613	Membrane alanine aminopeptidase N
#19514	2885573	2886388	816	816	hypothetical protein, partial [Escherichia coli]	3,00E-49	-1	#3120	2885201	2886571	1371	Phosphomannomutase
#38259	5526593	5527405	813	813	hypothetical protein, partial [Staphylococcus hominis]	1,00E-18	-1	#5727	5526317	5527540	1224	Phosphopentomutase
#72838	527979	528791	813	813	hypothetical protein [Escherichia coli]	4,00E-50	-1	#0513	527967	529838	1872	Peptidyl-prolyl cis-trans isomerase PpiD
#52107	3470729	3471535	807	702	hypothetical protein, partial [Catenibacterium mitsuokai]	7,00E-17	-1	#3696	3470834	3471637	804	Inositol-1-monophosphatase
#42633	4890226	4891026	801	801	hypothetical protein [Megasphaera elsdenii CAG:570]	5,00E-11	-1	#5144	4889794	4891629	1836	ATP-dependent DNA helicase RecQ
#42929	4852774	4853574	801	780	hypothetical protein [Catenibacterium mitsuokai]	6,00E-14	-1	#5107	4852291	4853553	1263	UDP-glucose dehydrogenase
#44490	4608110	4608910	801	801	hypothetical protein [Escherichia coli]	1,00E-11	-1	#4866	4606733	4611499	4767	hypothetical protein
#8127	1235822	1236622	801	801	hypothetical protein, partial [Bacillus cereus]	4,00E-04	-1	#1224	1235816	1236880	1065	Outer membrane protein A precursor
#32930	4781958	4782752	795	741	hypothetical protein, partial [Streptococcus pyogenes]	2,00E-17	-1	#5048	4781925	4782698	774	Phosphate transport ATP-binding protein PstB
#13407	2047320	2048111	792	792	hypothetical protein [Escherichia coli]	5,00E-177	-1	#2189	2045670	2049872	4203	core protein
#13432	2050044	2050835	792	792	hypothetical protein, partial [Pseudomonas aeruginosa]	6,00E-04	-1	#2190	2049939	2052047	2109	VgrG protein
#4928	754986	755777	792	792	hypothetical protein [Dialister invisus CAG:218]	3,00E-14	-1	#0716	753336	755918	2583	Leucyl-tRNA synthetase
#65606	1529480	1530271	792	723	hypothetical protein, partial [Staphylococcus aureus]	1,00E-04	-1	#1582	1529330	1530202	873	NgrB
#68609	1129336	1130127	792	723	hypothetical protein, partial [Staphylococcus aureus]	1,00E-04	-1	#1121	1129186	1130058	873	NgrB
#72184	618509	619300	792	792	hypothetical protein [Escherichia coli]	0	-1	#0585	616748	620944	4197	core protein
#70690	839606	840394	789	789	hypothetical protein, partial [Staphylococcus aureus]	6,00E-06	-1	#0797	839255	840421	1167	Succinyl-CoA ligase [ADP-forming] beta chain
#3592	548258	549043	786	786	hypothetical protein, partial [Staphylococcus aureus]	4,00E-04	-1	#0537	547319	550468	3150	RND efflux system, inner membrane transporter CmeB
#41720	5018243	5019028	786	786	hypothetical protein [Escherichia coli]	8,00E-156	-1	#5269	5016503	5020687	4185	core protein
#44593	4594248	4595033	786	786	hypothetical protein [Escherichia coli]	8,00E-156	-1	#4854	4592508	4596737	4230	core protein
#70912	809526	810311	786	786	hypothetical protein [Escherichia coli]	8,00E-156	-1	#0770	807786	811985	4200	core protein
#35026	5091228	5092010	783	783	hypothetical protein [Megasphaera elsdenii CAG:570]	5,00E-05	-1	#5325	5090655	5092550	1896	Thiamin biosynthesis protein ThiC

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#71041	788927	789706	780	615	hypothetical protein [Escherichia coli]	9,00E-32	-1	#0753	788633	789541	909	N-acetylglucosamine-regulated outer membrane porin
#40865	5136707	5137483	777	777	hypothetical protein [Laribacter hongkongensis]	5,00E-08	-1	#5360	5135885	5137534	1650	Glucose-6-phosphate isomerase
#4733	729135	729911	777	741	NrfD protein [Salmonella enterica]	2,00E-05	-1	#0688	729171	730703	1533	Citrate lyase alpha chain
#24277	3538611	3539384	774	774	hypothetical protein [Klebsiella pneumoniae]	1,00E-17	-1	#3755	3538608	3541181	2574	ClpB protein
#4747	730705	731478	774	765	hypothetical protein [Escherichia coli]	2,00E-16	-1	#0689	730714	731622	909	Citrate lyase beta chain
#42854	4862101	4862871	771	771	hypothetical protein, partial [Streptococcus pyogenes]	1,00E-09	-1	#5116	4861918	4863303	1386	putative transport protein YifK
#72344	595837	596607	771	771	hypothetical protein [Escherichia coli]	9,00E-17	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#75229	193355	194125	771	726	hypothetical protein [Ketogulonicigenium vulgare]	3,00E-28	-1	#0172	193361	194086	726	SSU ribosomal protein S2p (SAe)
#40487	5191955	5192719	765	540	hypothetical protein, partial [Bacillus cereus]	3,00E-42	-1	#5412	5191823	5192494	672	NrfC protein
#63809	1764554	1765318	765	765	hypothetical protein [Herbaspirillum huttiense]	2,00E-35	-1	#1883	1764305	1765576	1272	D-amino acid dehydrogenase small subunit
#47320	4189987	4190745	759	609	hypothetical protein [Escherichia coli]	4,00E-59	-1	#4441	4189177	4190595	1419	Glutamate synthase [NADPH] small chain
#60404	2244630	2245388	759	738	hypothetical protein [Dickeya solani]	2,00E-51	-1	#2400	2244651	2245439	789	Enoyl-[acyl-carrier-protein] reductase [NADH]
#24410	3556980	3557732	753	594	inorganic polyphosphate kinase [Escherichia coli]	8,00E-24	-1	#3774	3557013	3557606	594	Heat shock protein GrpE
#30643	4449603	4450355	753	753	hypothetical protein, partial [Escherichia coli]	4,00E-09	-1	#4722	4448904	4451639	2736	ABC-type multidrug transport system, permease component
#13591	2072462	2073211	750	750	hypothetical protein [Bacteroides fragilis CAG:558]	1,00E-14	-1	#2216	2071280	2073283	2004	putative collagenase
#50287	3745625	3746374	750	750	hypothetical protein [Klebsiella pneumoniae]	9,00E-28	-1	#3983	3745610	3747385	1776	L-fucose isomerase
#26616	3871733	3872479	747	747	metal-dependent RNase [Bacteroides cellulosilyticus CAG:158]	1,00E-18	-1	#4105	3870923	3873796	2874	Glycine dehydrogenase [decarboxylating] (glycine cleavage system P protein)
#48333	4041354	4042100	747	693	hypothetical protein [Escherichia coli]	9,00E-39	-1	#4289	4040205	4042046	1842	RNA polymerase sigma factor RpoD
#70493	869451	870197	747	747	hypothetical protein, partial [Bacillus cereus]	1,00E-42	-1	#0827	869268	870320	1053	2-keto-3-deoxy-D-arabino-heptulosonate-7- phosphate synthase I alpha
#3613	550878	551621	744	699	hypothetical protein [Cronobacter sakazakii]	8,00E-26	-1	#0538	550491	551576	1086	Membrane fusion protein of RND family multidrug efflux pump
#73693	406850	407593	744	744	hypothetical protein [Escherichia coli]	6,00E-29	-1	#0391	406427	407596	1170	2-methylcitrate synthase
#74676	268531	269274	744	744	hypothetical protein [Escherichia coli]	7,00E-49	-1	#0240	267079	271293	4215	core protein
#35109	5103099	5103833	735	735	hypothetical protein [Bifidobacterium longum CAG:69]	4,00E-07	-1	#5338	5102373	5103962	1590	IMP cyclohydrolase
#59705	2345766	2346497	732	672	cell surface protein [Escherichia coli]	1,00E-60	-1	#2528	2345826	2348297	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#71759	681891	682622	732	732	hypothetical protein, partial [Bacillus cereus]	1,00E-04	-1	#0639	681027	684164	3138	Cobalt-zinc-cadmium resistance protein CzcA
#13415	2048310	2049038	729	729	hypothetical protein, partial [Escherichia coli]	2,00E-51	-1	#2189	2045670	2049872	4203	core protein
#22915	3352885	3353613	729	729	hypothetical protein [Cardiobacterium valvarum]	2,00E-14	-1	#3588	3352834	3353850	1017	Sulfate and thiosulfate binding protein CysP
#65968	1479937	1480665	729	93	hypothetical protein [Bradyrhizobium yuanmingense]	1,00E-06	-1	#1505	1478323	1480029	1707	Urease alpha subunit
#68971	1079793	1080521	729	93	hypothetical protein [Bradyrhizobium yuanmingense]	1,00E-06	-1	#1044	1078179	1079885	1707	Urease alpha subunit
#49071	3933524	3934246	723	594	hypothetical protein [Klebsiella pneumoniae]	4,00E-04	-1	#4177	3933653	3934732	1080	Membrane-bound lytic murein transglycosylase C precursor
#17745	2643842	2644561	720	702	hypothetical protein [Sutterella wadsworthensis CAG:135]	2,00E-11	-1	#2838	2643803	2644543	741	hypothetical protein YebC
#71551	711499	712218	720	720	hypothetical protein, partial [Alcaligenes faecalis]	8,00E-07	-1	#0668	711142	713247	2106	Carbon starvation protein A
#32943	4783882	4784598	717	717	hypothetical protein, partial [Bacillus cereus]	1,00E-05	-1	#5050	4783771	4784730	960	Phosphate transport system permease protein PstC
#29453	4274855	4275556	702	621	hypothetical protein [Bradyrhizobium japonicum]	8,00E-04	-1	#4537	4274936	4275565	630	LSU ribosomal protein L3p (L3e)
#21230	3123966	3124658	693	660	hypothetical protein, partial [Escherichia coli]	9,00E-10	-1	#3371	3122139	3124625	2487	Periplasmic nitrate reductase precursor
#22707	3323630	3324322	693	693	hypothetical protein [Escherichia coli]	1,00E-36	-1	#3557	3323600	3324565	966	Glucokinase
#35095	5101471	5102163	693	693	hypothetical protein [Megasphaera elsdenii CAG:570]	2,00E-11	-1	#5337	5101072	5102361	1290	Phosphoribosylamine--glycine ligase
#48200	4059112	4059804	693	693	hypothetical protein, partial [Catenibacterium mitsuokai]	2,00E-06	-1	#4303	4058923	4060941	2019	2,4-dienoyl-CoA reductase [NADPH]
#72354	594037	594729	693	693	hypothetical protein [Escherichia coli]	2,00E-150	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#27591	4007468	4008157	690	690	hypothetical protein, partial [Catenibacterium mitsuokai]	2,00E-09	-1	#4256	4007225	4009117	1893	Topoisomerase IV subunit B
#39557	5333251	5333937	687	621	hypothetical protein, partial [Escherichia coli]	1,00E-23	-1	#5553	5333275	5333895	621	FKBP-type peptidyl-prolyl cis-trans isomerase FklB
#4164	646074	646760	687	495	hypothetical protein [Rhodopirellula baltica]	1,00E-06	-1	#0610	646086	646580	495	Peptidyl-prolyl cis-trans isomerase PpiB
#51084	3625085	3625771	687	687	hypothetical protein [Escherichia coli]	8,00E-29	-1	#3853	3624395	3625933	1539	Multidrug resistance protein B
#71490	720160	720843	684	564	hypothetical protein [Escherichia coli]	2,00E-34	-1	#0677	720193	720756	564	Alkyl hydroperoxide reductase protein C
#72867	524054	524737	684	684	hypothetical protein [Escherichia coli]	2,00E-16	-1	#0510	523478	524752	1275	ATP-dependent Clp protease ATP-binding subunit ClpX
#44582	4595562	4596242	681	681	hypothetical protein [Delftia acidovorans]	4,00E-06	-1	#4854	4592508	4596737	4230	core protein
#1887	287289	287966	678	678	hypothetical protein [Escherichia coli]	6,00E-32	-1	#0262	286245	287984	1740	Flagellar biosynthesis protein FlhA
#25203	3674597	3675274	678	231	carbamoyl dehydratase HypE [Escherichia coli]	5,00E-06	-1	#3910	3674708	3674938	231	RNA polymerase sigma factor RpoS
#65464	1545532	1546209	678	168	hypothetical protein [Escherichia coli]	1,00E-27	-1	#1606	1544761	1545699	939	D-3-phosphoglycerate dehydrogenase
#53143	3311035	3311706	672	669	hypothetical protein [Escherichia coli]	4,00E-57	-1	#3546	3310783	3311703	921	Lipid A biosynthesis lauroyl acyltransferase
#36841	5353324	5353992	669	525	hypothetical protein, partial [Staphylococcus aureus]	8,00E-21	-1	#5573	5353468	5353998	531	Inorganic pyrophosphatase
#70685	840493	841161	669	669	hypothetical protein, partial [Staphylococcus aureus]	4,00E-04	-1	#0798	840421	841290	870	Succinyl-CoA ligase [ADP-forming] alpha chain
#71976	646757	647425	669	669	hypothetical protein, partial [Salmonella enterica]	2,00E-07	-1	#0611	646754	648139	1386	Cysteinyl-tRNA synthetase
#70901	810840	811505	666	666	hypothetical protein [Delftia acidovorans]	1,00E-05	-1	#0770	807786	811985	4200	core protein
#34181	4968146	4968808	663	663	hypothetical protein, partial [Piscirickettsia salmonis]	3,00E-14	-1	#5217	4968119	4969021	903	Formate dehydrogenase O beta subunit
#29076	4221669	4222328	660	660	hypothetical protein, partial [Escherichia coli]	6,00E-12	-1	#4471	4221462	4222565	1104	Rod shape-determining protein MreC
#29863	4334526	4335185	660	660	hypothetical protein, partial [Escherichia coli]	1,00E-05	-1	#4604	4334337	4335689	1353	Osmolarity sensory histidine kinase EnvZ
#51631	3544081	3544740	660	342	hypothetical protein [Cronobacter sakazakii]	5,00E-17	-1	#3759	3544162	3544503	342	Ribosome hibernation protein YfiA
#71060	786433	787092	660	660	hypothetical protein [Bacteroides faecis CAG:32]	4,00E-11	-1	#0750	785893	787557	1665	Glutaminyl-tRNA synthetase
#13468	2055987	2056643	657	657	hypothetical protein [Escherichia coli]	8,00E-08	-1	#2196	2055654	2056715	1062	hypothetical protein
#17930	2667614	2668270	657	657	hypothetical protein, partial [Staphylococcus aureus]	1,00E-05	-1	#2861	2667422	2669083	1662	Methyl-accepting chemotaxis protein II (aspartate chemoreceptor protein)
#22131	3245659	3246312	654	642	hypothetical protein [Sutterella wadsworthensis CAG:135]	8,00E-12	-1	#3481	3245032	3246300	1269	Dihydrofolate synthase
#5070	772828	773478	651	489	hypothetical protein, partial [Escherichia coli]	2,00E-87	-1	#0737	772276	773316	1041	Phosphate starvation-inducible ATPase PhoH with RNA binding motif
#76515	3529	4179	651	222	hypothetical protein [Escherichia coli]	7,00E-26	-1	#0003	2818	3750	933	Homoserine kinase
#3761	574555	575202	648	639	hypothetical protein [Bacteroides uniformis CAG:3]	3,00E-04	-1	#0561	574564	577068	2505	Lead, cadmium, zinc and mercury transporting ATPase
#42898	4856978	4857625	648	330	hypothetical protein, partial [Bacillus thuringiensis]	7,00E-44	-1	#5111	4856177	4857307	1131	4-keto-6-deoxy-N-Acetyl-D-hexosaminyl-(Lipid carrier) aminotransferase
#74680	267877	268524	648	648	hypothetical protein, partial [Escherichia coli]	2,00E-23	-1	#0240	267079	271293	4215	core protein
#74669	269704	270348	645	645	hypothetical protein, partial [Escherichia coli]	2,00E-43	-1	#0240	267079	271293	4215	core protein
#75771	111385	112026	642	642	hypothetical protein, partial [Catenibacterium mitsuokai]	4,00E-15	-1	#0099	111163	112080	918	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase
#13722	2089404	2090042	639	639	hypothetical protein [Escherichia coli]	3,00E-15	-1	#2233	2088606	2090045	1440	Aldehyde dehydrogenase A
#42902	4856324	4856962	639	639	hypothetical protein [Mycobacterium tuberculosis]	1,00E-07	-1	#5111	4856177	4857307	1131	4-keto-6-deoxy-N-Acetyl-D-hexosaminyl-(Lipid carrier) aminotransferase
#51956	3493161	3493799	639	639	hypothetical protein [Burkholderia cepacia]	2,00E-05	-1	#3717	3493092	3494282	1191	Flavohemoprotein, hemoglobin-like protein, flavohemoglobin, nitric oxide dioxygenase
#64949	1612885	1613523	639	591	hypothetical protein [Methylobacterium radiotolerans]	3,00E-13	-1	#1687	1612933	1614237	1305	NADH dehydrogenase
#21195	3119526	3120161	636	603	hypothetical protein [Escherichia coli]	6,00E-89	-1	#3367	3119529	3120131	603	Cytochrome c-type protein NapC
#25531	3719080	3719715	636	612	hypothetical protein, partial [Vibrio parahaemolyticus]	3,00E-15	-1	#3958	3718054	3719691	1638	CTP synthase
#26608	3870926	3871561	636	636	hypothetical protein [Salmonella enterica]	5,00E-43	-1	#4105	3870923	3873796	2874	Glycine dehydrogenase [decarboxylating] (glycine cleavage system P protein)
#29450	4274284	4274919	636	600	hypothetical protein [Pseudomonas fluorescens]	1,00E-27	-1	#4536	4274320	4274925	606	LSU ribosomal protein L4p (L1e)
#30215	4384000	4384632	633	435	hypothetical protein [Escherichia coli]	2,00E-52	-1	#4647	4383946	4384434	489	Gluconokinase

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#6518	988734	989366	633	501	hypothetical protein [Shewanella haliotis]	7,00E-27	-1	#0950	988485	989234	750	Molybdopterin biosynthesis protein MoeB
#72414	585592	586221	630	147	hypothetical protein, partial [Escherichia coli]	2,00E-78	-1	#0565	581353	585738	4386	Large repetitive protein
#68423	1152782	1153408	627	627	hypothetical protein, partial [Mycobacterium intracellulare]	1,00E-62	-1	#1154	1151633	1155661	4029	Cell division protein FtsK
#41867	4996608	4997231	624	624	hypothetical protein [Cardiobacterium valvarum]	2,00E-08	-1	#5246	4996419	4997408	990	Sulfate-binding protein Sbp
#62077	1992788	1993411	624	624	hypothetical protein [Rhodococcus qingshengii]	7,00E-136	-1	#2141	1992080	1993699	1620	N-acetylgalactosamine 6-sulfate sulfatase (GALNS)
#23565	3446434	3447054	621	621	hypothetical protein [Megasphaera elsdenii CAG:570]	6,00E-06	-1	#3674	3446161	3447315	1155	Ribosomal RNA large subunit methyltransferase N
#2463	373427	374047	621	621	hypothetical protein [Vibrio cholerae]	8,00E-09	-1	#0360	373223	374911	1689	Choline dehydrogenase
#25283	3683777	3684397	621	621	hypothetical protein [Bradyrhizobium elkanii]	6,00E-20	-1	#3923	3683570	3684478	909	Sulfate adenylyltransferase subunit 2
#30428	4410554	4411174	621	564	membrane protein [Shigella flexneri]	3,00E-13	-1	#4677	4409621	4411117	1497	Signal recognition particle receptor protein FtsY (alpha subunit)
#33054	4797034	4797654	621	594	aGAP012078-PA [Parabacteroides johnsonii CAG:246]	1,00E-15	-1	#5062	4797061	4798443	1383	ATP synthase beta chain
#49203	3915971	3916591	621	621	hypothetical protein, partial [Lactobacillus johnsonii]	4,00E-11	-1	#4152	3915818	3916972	1155	S-adenosylmethionine synthetase
#59611	2357157	2357777	621	525	hypothetical protein, partial [Salmonella enterica]	2,00E-15	-1	#2539	2357253	2357870	618	Anaerobic dimethyl sulfoxide reductase chain B
#8099	1232773	1233393	621	501	hypothetical protein, partial [Escherichia coli]	3,00E-07	-1	#1221	1232755	1233273	519	3-hydroxydecanoyl-[acyl-carrier-protein] dehydratase
#47017	4236612	4237229	618	618	hypothetical protein [Cronobacter sakazakii]	3,00E-43	-1	#4487	4236450	4237607	1158	RND efflux system, membrane fusion protein CmeA
#55866	2909432	2910049	618	618	hypothetical protein [Escherichia coli]	4,00E-101	-1	#3143	2907932	2910739	2808	hypothetical protein
#75810	106088	106705	618	618	hypothetical protein, partial [Rhodococcus opacus]	1,00E-45	-1	#0094	105371	106846	1476	UDP-N-acetylmuramate--alanine ligase
#22246	3260317	3260931	615	525	hypothetical protein [Ketogulonicigenium vulgare]	4,00E-26	-1	#3497	3259756	3260841	1086	Chorismate synthase
#29414	4268588	4269202	615	192	hypothetical protein, partial [Escherichia coli]	6,00E-39	-1	#4522	4268246	4268779	534	LSU ribosomal protein L6p (L9e)
#31472	4563354	4563968	615	615	hypothetical protein, partial [Rhodococcus qingshengii]	2,00E-30	-1	#4830	4563237	4564559	1323	Xylose isomerase
#65175	1582460	1583074	615	615	hypothetical protein, partial [Escherichia coli]	1,00E-35	-1	#1656	1582406	1583104	699	Flagellar L-ring protein FlgH
#16679	2494045	2494656	612	129	hypothetical protein [Salmonella enterica]	6,00E-28	-1	#2676	2493976	2494173	198	LSU ribosomal protein L35p
#68003	1207758	1208369	612	612	hypothetical protein, partial [Escherichia coli]	3,00E-24	-1	#1196	1207425	1210037	2613	Membrane alanine aminopeptidase N
#41709	5019557	5020162	606	606	hypothetical protein [Delftia acidovorans]	8,00E-06	-1	#5269	5016503	5020687	4185	core protein
#50963	3643995	3644600	606	576	hypothetical protein, partial [Escherichia coli]	3,00E-45	-1	#3877	3644025	3645260	1236	Anaerobic nitric oxide reductase flavorubredoxin
#24162	3525152	3525754	603	465	hypothetical protein, partial [Escherichia coli]	7,00E-13	-1	#3746	3524579	3525616	1038	putative tRNA/rRNA methyltransferase yfiF
#381	58944	59546	603	603	hypothetical protein [Escherichia coli]	4,00E-16	-1	#0056	58392	59678	1287	Survival protein SurA precursor (Peptidyl-prolyl cis-trans isomerase SurA)
#72435	582763	583365	603	603	hypothetical protein [Escherichia coli]	4,00E-08	-1	#0565	581353	585738	4386	Large repetitive protein
#16656	2491013	2491612	600	600	hypothetical protein, partial [Bacillus cereus]	5,00E-10	-1	#2672	2489732	2492119	2388	Phenylalanyl-tRNA synthetase beta chain
#29521	4285443	4286042	600	588	hypothetical protein [Escherichia coli]	2,00E-09	-1	#4552	4285440	4286030	591	FKBP-type peptidyl-prolyl cis-trans isomerase SlyD
#73078	489727	490326	600	561	hypothetical protein [Bacteroides stercoris CAG:120]	5,00E-10	-1	#0471	489766	490893	1128	tRNA-guanine transglycosylase
#65315	1563959	1564555	597	597	hypothetical protein, partial [Staphylococcus aureus]	2,00E-11	-1	#1631	1563683	1564735	1053	Rhodanese-related sulfurtransferase
#30955	4495934	4496524	591	585	hypothetical protein, partial [Escherichia coli]	2,00E-81	-1	#4769	4495940	4497340	1401	Glutamate decarboxylase
#50318	3740673	3741263	591	297	hypothetical protein, partial [Pseudomonas amygdali]	6,00E-44	-1	#3978	3739602	3740969	1368	L-serine dehydratase
#73911	378216	378806	591	591	hypothetical protein, partial [Bacillus cereus]	3,00E-38	-1	#0363	377127	379160	2034	High-affinity choline uptake protein BetT
#28649	4163050	4163637	588	312	hypothetical protein [Acidovorax radicis]	5,00E-04	-1	#4414	4163119	4163430	312	LSU ribosomal protein L21p
#71473	723052	723639	588	420	hypothetical protein [Salmonella enterica]	8,00E-13	-1	#0681	723220	724458	1239	putative zinc-type alcohol dehydrogenase-like protein ybdR
#17453	2603185	2603769	585	168	hypothetical protein [Laribacter hongkongensis]	4,00E-06	-1	#2794	2602543	2603352	810	Ribosomal RNA large subunit methyltransferase A
#19359	2863951	2864535	585	585	hypothetical protein, partial [Salmonella enterica]	4,00E-16	-1	#3102	2863768	2865138	1371	Phosphomannomutase
#35574	5173376	5173960	585	585	hypothetical protein, partial [Escherichia coli]	9,00E-06	-1	#5396	5173061	5175883	2823	Excinuclease ABC subunit A

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#64620	1654337	1654921	585	585	hypothetical protein, partial [Escherichia coli]	1,00E-76	-1	#1743	1654298	1656799	2502	Phage portal protein
#76352	28693	29277	585	585	hypothetical protein [Sutterella wadsworthensis CAG:135]	2,00E-11	-1	#0025	26803	29619	2817	Isoleucyl-tRNA synthetase
#35579	5174114	5174695	582	582	hypothetical protein [Bacteroides thetaiotaomicron CAG:40]	4,00E-05	-1	#5396	5173061	5175883	2823	Excinuclease ABC subunit A
#47342	4187809	4188390	582	582	hypothetical protein, partial [Pseudomonas aeruginosa]	9,00E-14	-1	#4440	4184611	4189164	4554	Glutamate synthase [NADPH] large chain
#61756	2033830	2034411	582	453	hypothetical protein [Escherichia coli]	5,00E-44	-1	#2174	2033959	2037699	3741	Respiratory nitrate reductase alpha chain
#23547	3444225	3444803	579	579	hypothetical protein [Clostridium bolteae CAG:59]	1,00E-17	-1	#3672	3443718	3444836	1119	1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase
#48623	3998980	3999558	579	561	hypothetical protein [Dickeya dadantii]	3,00E-10	-1	#4248	3998998	3999579	582	Modulator of drug activity B
#5740	870687	871265	579	552	hypothetical protein [Bordetella pertussis]	5,00E-06	-1	#0828	870486	871238	753	Phosphoglycerate mutase
#61992	2003311	2003889	579	579	hypothetical protein, partial [Escherichia coli]	6,00E-81	-1	#2147	2002495	2003895	1401	Glutamate decarboxylase
#23470	3434108	3434683	576	576	cytosine deaminase [Ruminococcus obeum CAG:39]	3,00E-13	-1	#3664	3433991	3435568	1578	GMP synthase [glutamine-hydrolyzing], amidotransferase subunit
#30961	4496618	4497193	576	576	hypothetical protein [Escherichia coli]	8,00E-30	-1	#4769	4495940	4497340	1401	Glutamate decarboxylase
#41925	4988360	4988932	573	537	hypothetical protein [Escherichia coli]	1,00E-06	-1	#5236	4988396	4989016	621	Manganese superoxide dismutase
#63520	1810195	1810764	570	492	hypothetical protein [Escherichia coli]	4,00E-105	-1	#1931	1810273	1814016	3744	Respiratory nitrate reductase alpha chain
#26711	3883750	3884316	567	504	hypothetical protein, partial [Bacillus cereus]	4,00E-25	-1	#4117	3883813	3884472	660	Ribose 5-phosphate isomerase A
#33073	4799393	4799959	567	567	hypothetical protein [Escherichia coli]	1,00E-37	-1	#5064	4799384	4800925	1542	ATP synthase alpha chain
#67121	1326626	1327192	567	567	hypothetical protein [Escherichia coli]	7,00E-14	-1	#1335	1325081	1327627	2547	Trimethylamine-N-oxide reductase
#2129	323666	324229	564	564	hypothetical protein, partial [Escherichia coli]	4,00E-26	-1	#0307	323657	324556	900	Transcriptional regulator
#44232	4648487	4649050	564	537	hypothetical protein, partial [Salmonella enterica]	8,00E-17	-1	#4905	4648160	4649023	864	Protein YicC
#65479	1543802	1544362	561	489	hypothetical protein [Xenorhabdus bovienii]	2,00E-09	-1	#1605	1543448	1544290	843	hypothetical protein
#68480	1143655	1144215	561	489	hypothetical protein [Xenorhabdus bovienii]	2,00E-09	-1	#1146	1143301	1144143	843	hypothetical protein
#34663	5035275	5035832	558	558	hypothetical protein [Escherichia coli]	1,00E-34	-1	#5283	5034228	5036729	2502	Phosphoenolpyruvate-protein phosphotransferase of PTS system
#38456	5499305	5499862	558	558	hypothetical protein, partial [Bacillus cereus]	0,001	-1	#5698	5498879	5500543	1665	Methyl-accepting chemotaxis protein I (serine chemoreceptor protein)
#1619	246401	246955	555	555	hypothetical protein [Escherichia coli]	7,00E-10	-1	#0221	245399	248833	3435	IcmF-related protein
#21121	3109985	3110539	555	555	hypothetical protein [Escherichia coli]	1,00E-35	-1	#3357	3109637	3112228	2592	Putative ATP-binding component of a transport system
#72389	588961	589515	555	555	hypothetical protein [Escherichia coli]	3,00E-33	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72571	563034	563588	555	555	hypothetical protein [Mycobacterium avium]	5,00E-08	-1	#0550	563010	563714	705	Adenylate kinase
#14645	2211555	2212106	552	183	hypothetical protein [Cronobacter malonaticus]	8,00E-07	-1	#2364	2210694	2211737	1044	Putative oxidoreductase YcjS, NADH-binding protein
#44678	4581399	4581950	552	525	hypothetical protein [Escherichia coli]	2,00E-80	-1	#4843	4580949	4581923	975	LysR family transcriptional regulator YiaU
#72381	589813	590364	552	552	hypothetical protein [Escherichia coli]	8,00E-07	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#31224	4533046	4533594	549	549	hypothetical protein [Nitratireductor aquibiodomus]	6,00E-19	-1	#4799	4532635	4533618	984	Dipeptide transport ATP-binding protein DppD
#33378	4845681	4846229	549	549	hypothetical protein, partial [Elizabethkingia anophelis]	3,00E-19	-1	#5100	4845396	4846661	1266	ATP-dependent RNA helicase RhlB
#4358	673541	674089	549	549	hypothetical protein, partial [Pseudomonas aeruginosa]	2,00E-08	-1	#0632	673265	675166	1902	VgrG protein
#50	7242	7790	549	549	hypothetical protein [Staphylococcus aureus]	2,00E-07	-1	#0007	6546	7976	1431	Putative alanine/glycine transport protein
#46888	4254059	4254601	543	237	DNA topoisomerase I [Halorubrum terrestre]	2,00E-13	-1	#4501	4254149	4254385	237	hypothetical protein
#5146	782716	783258	543	543	hypothetical protein [Bacteroides sartorii]	2,00E-13	-1	#0746	782611	783411	801	Glucosamine-6-phosphate deaminase
#65523	1538567	1539109	543	477	hypothetical protein [Escherichia coli]	2,00E-27	-1	#1592	1538225	1539043	819	hypothetical protein
#68524	1138420	1138962	543	477	hypothetical protein [Escherichia coli]	2,00E-27	-1	#1133	1138078	1138896	819	hypothetical protein
#2337	352869	353408	540	498	hypothetical protein [Bacteroides plebeius CAG:211]	3,00E-04	-1	#0338	352791	353366	576	oxidoreductase, aldo/keto reductase family
#72364	592246	592785	540	540	hypothetical protein [Escherichia coli]	2,00E-116	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#3384	514640	515176	537	474	hypothetical protein, partial [Staphylococcus aureus]	2,00E-09	-1	#0501	514703	516694	1992	Cytochrome O ubiquinol oxidase subunit I
#37943	5521704	5522240	537	474	hypothetical protein [Shigella sonnei]	2,00E-43	-1	#5721	5521395	5522177	783	radical activating enzyme
#72376	590452	590988	537	537	hypothetical protein [Escherichia coli]	7,00E-36	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#7729	1172642	1173178	537	537	hypothetical protein [Escherichia coli]	3,00E-04	-1	#1168	1171847	1173607	1761	hypothetical protein
#30142	4374276	4374809	534	471	hypothetical protein [Escherichia coli]	3,00E-61	-1	#4637	4372773	4374746	1974	Glycogen debranching enzyme
#428	65255	65788	534	498	hypothetical protein, partial [Rhodococcus qingshengii]	9,00E-89	-1	#0061	65291	68197	2907	RNA polymerase associated protein RapA
#69900	953538	954071	534	534	hypothetical protein, partial [Escherichia coli]	2,00E-31	-1	#0918	953010	954377	1368	ATP-dependent RNA helicase RhlE
#32948	4784874	4785404	531	531	hypothetical protein [Escherichia coli]	2,00E-28	-1	#5051	4784817	4785857	1041	Phosphate ABC transporter, periplasmic phosphate-binding protein PstS
#46565	4303029	4303559	531	531	hypothetical protein, partial [Streptomyces rimosus]	2,00E-15	-1	#4570	4301544	4304087	2544	Nitrite reductase [NAD(P)H] large subunit
#47752	4123996	4124526	531	531	hypothetical protein, partial [Escherichia coli]	7,00E-28	-1	#4374	4123525	4125561	2037	LppC putative lipoprotein
#63161	1858098	1858628	531	531	hypothetical protein [Escherichia coli]	3,00E-19	-1	#1984	1858029	1859087	1059	DNA methyl transferase, phage-associated
#65596	1530712	1531242	531	531	hypothetical protein [Escherichia coli]	2,00E-10	-1	#1584	1530574	1532877	2304	Antigen 43 precursor
#68599	1130568	1131098	531	531	hypothetical protein [Escherichia coli]	2,00E-10	-1	#1123	1130430	1133111	2682	Antigen 43 precursor
#15886	2381512	2382036	525	525	hypothetical protein [Alistipes finegoldii CAG:68]	4,00E-11	-1	#2564	2380993	2382639	1647	Fumarate hydratase class I, aerobic
#6886	1043564	1044088	525	525	hypothetical protein, partial [Escherichia coli]	4,00E-35	-1	#1007	1043045	1044697	1653	Hydroxylamine reductase
#46323	4338721	4339242	522	522	hypothetical protein, partial [Piscirickettsia salmonis]	4,00E-39	-1	#4607	4337206	4339527	2322	Transcription accessory protein (S1 RNA-binding domain)
#49170	3920253	3920774	522	516	hypothetical protein [Pseudomonas fuscovaginae]	1,00E-12	-1	#4157	3920259	3920990	732	Ribosomal RNA small subunit methyltransferase E
#27709	4024193	4024711	519	519	hypothetical protein, partial [Bacillus cereus]	1,00E-12	-1	#4273	4023956	4026796	2841	Glutamate-ammonia-ligase adenylyltransferase
#35917	5220957	5221475	519	519	hypothetical protein [Escherichia coli]	2,00E-34	-1	#5442	5220936	5221616	681	Phosphonates transport ATP-binding protein PhnL
#30206	4382989	4383504	516	516	transcriptional regulator MalT, partial [Escherichia coli]	4,00E-10	-1	#4646	4382602	4383942	1341	Low-affinity gluconate/H+ symporter GntU
#46954	4244716	4245231	516	345	hypothetical protein, partial [Piscirickettsia salmonis]	2,00E-06	-1	#4494	4243957	4245060	1104	Glutamate Aspartate transport system permease protein GltK
#56484	2811203	2811718	516	516	hypothetical protein [Escherichia coli]	1,00E-101	-1	#3049	2808089	2815951	7863	adherence and invasion outermembrane protein (Inv,enhances Peyer's patches colonization)
#69150	1055550	1056065	516	516	hypothetical protein [Klebsiella pneumoniae]	2,00E-15	-1	#1016	1054110	1056386	2277	ATP-dependent Clp protease ATP-binding subunit ClpA
#19494	2882317	2882829	513	513	hypothetical protein, partial [Escherichia coli]	2,00E-20	-1	#3118	2882272	2883750	1479	Lipopolysaccharide biosynthesis protein WzxC
#36131	5249392	5249904	513	513	hypothetical protein [Alistipes finegoldii CAG:68]	2,00E-05	-1	#5468	5248996	5250642	1647	Fumarate hydratase class I, anaerobic
#47842	4112337	4112849	513	513	hypothetical protein [Alistipes putredinis CAG:67]	4,00E-06	-1	#4360	4112310	4113170	861	Tagatose 1,6-bisphosphate aldolase
#75097	208179	208688	510	255	hypothetical protein [Escherichia coli]	1,00E-13	-1	#0188	207837	208433	597	Ribonuclease HII
#27020	3927973	3928479	507	507	hypothetical protein [Escherichia coli]	1,00E-99	-1	#4168	3927778	3928785	1008	Putative alpha helix chain
#34464	5008020	5008526	507	429	hypothetical protein, partial [Escherichia coli]	8,00E-33	-1	#5260	5007963	5008448	486	Ribonuclease E inhibitor RraA
#40779	5149285	5149791	507	507	phenol hydroxylase [Natronorubrum tibetense]	2,00E-10	-1	#5372	5148925	5150040	1116	Maltose/maltodextrin transport ATP-binding protein MalK
#76309	34455	34961	507	507	hypothetical protein [Neisseria bacilliformis]	3,00E-11	-1	#0032	34026	35174	1149	Carbamoyl-phosphate synthase small chain
#21449	3150579	3151082	504	201	hypothetical protein [Escherichia coli]	2,00E-25	-1	#3392	3150882	3153509	2628	DNA gyrase subunit A
#32723	4752276	4752779	504	504	hypothetical protein [Bifidobacterium adolescentis CAG:119]	2,00E-06	-1	#5022	4751562	4753976	2415	DNA gyrase subunit B
#35670	5188910	5189413	504	477	hypothetical protein [Haloferax elongans]	2,00E-05	-1	#5409	5187428	5189386	1959	Acetyl-coenzyme A synthetase
#42917	4854624	4855127	504	489	hypothetical protein [Alistipes finegoldii CAG:68]	4,00E-04	-1	#5109	4854639	4855520	882	Glucose-1-phosphate thymidylyltransferase
#75632	132366	132869	504	504	hypothetical protein, partial [Escherichia coli]	1,00E-13	-1	#0118	132222	133709	1488	Dihydrolipoamide dehydrogenase of pyruvate dehydrogenase complex
#29493	4281588	4282088	501	375	hypothetical protein [Halomonas stevensii]	6,00E-05	-1	#4545	4281711	4282085	375	SSU ribosomal protein S12p (S23e)
#31862	4626484	4626984	501	501	hypothetical protein, partial [Escherichia coli]	3,00E-23	-1	#4881	4625812	4627008	1197	2-amino-3-ketobutyrate coenzyme A ligase
#46573	4301415	4301915	501	372	hypothetical protein [Salmonella enterica]	2,00E-26	-1	#4570	4301544	4304087	2544	Nitrite reductase [NAD(P)H] large subunit

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#6224	946177	946677	501	459	hypothetical protein, partial [Anaerococcus lactolyticus]	2,00E-06	-1	#0911	945874	946635	762	Endonuclease/Exonuclease/phosphatase family protein
#41273	5076524	5077021	498	492	hypothetical protein [Serratia marcescens]	3,00E-34	-1	#5315	5076518	5077015	498	LSU ribosomal protein L10p (P0)
#49178	3918865	3919362	498	441	hypothetical protein [Photobacterium ganghwense]	9,00E-12	-1	#4155	3918922	3919377	456	Protein sprT
#16693	2495659	2496153	495	495	hypothetical protein [Klebsiella pneumoniae]	2,00E-12	-1	#2678	2494816	2496744	1929	Threonyl-tRNA synthetase
#44544	4600978	4601472	495	495	hypothetical protein, partial [Escherichia coli]	4,00E-22	-1	#4860	4600852	4602765	1914	PTS system, mannitol-specific IIC component
#5125	779683	780177	495	444	hypothetical protein, partial [Anaerococcus lactolyticus]	2,00E-15	-1	#0743	779374	780126	753	Phosphatase NagD predicted to act in N-acetylglucosamine utilization subsystem
#61743	2035342	2035836	495	495	hypothetical protein, partial [Escherichia coli]	2,00E-12	-1	#2174	2033959	2037699	3741	Respiratory nitrate reductase alpha chain
#1836	279738	280229	492	492	hypothetical protein [Escherichia coli]	1,00E-17	-1	#0251	278928	281372	2445	Butyryl-CoA dehydrogenase
#20848	3069509	3070000	492	465	hypothetical protein [Escherichia coli]	1,00E-30	-1	#3319	3068090	3069973	1884	Colicin I receptor precursor
#26485	3854186	3854677	492	378	hypothetical protein, partial [Lactobacillus fermentum]	4,00E-38	-1	#4086	3854300	3854695	396	putative oxidoreductase, Fe-S subunit
#30327	4398908	4399399	492	492	hypothetical protein [Rhizobium leguminosarum]	2,00E-10	-1	#4664	4398836	4399549	714	Branched-chain amino acid transport ATP-binding protein LivF
#4739	730038	730529	492	492	hypothetical protein [Acidaminococcus intestini CAG:325]	2,00E-10	-1	#0688	729171	730703	1533	Citrate lyase alpha chain
#61997	2002726	2003217	492	492	hypothetical protein [Escherichia coli]	1,00E-07	-1	#2147	2002495	2003895	1401	Glutamate decarboxylase
#19620	2899796	2900284	489	489	hypothetical protein [Escherichia coli]	2,00E-18	-1	#3135	2899787	2900833	1047	Polysaccharide export lipoprotein Wza
#29369	4262988	4263476	489	489	hypothetical protein [Aeromonas taiwanensis]	4,00E-37	-1	#4512	4262679	4263668	990	DNA-directed RNA polymerase alpha subunit
#40584	5176123	5176611	489	474	hypothetical protein [Polaromonas naphthalenivorans]	1,00E-22	-1	#5397	5176138	5176674	537	Single-stranded DNA-binding protein
#4240	658692	659180	489	138	hypothetical protein [Escherichia coli]	3,00E-70	-1	#0623	657939	658829	891	hypothetical protein
#578	84876	85364	489	489	Flp pilus assembly protein TadG [Haloferax denitrificans]	8,00E-06	-1	#0074	84072	85472	1401	3-isopropylmalate dehydratase large subunit
#29395	4266206	4266691	486	486	hypothetical protein, partial [Escherichia coli]	2,00E-86	-1	#4517	4265405	4266736	1332	Preprotein translocase secY subunit
#31605	4584032	4584517	486	486	hypothetical protein, partial [Staphylococcus aureus]	2,00E-05	-1	#4847	4583930	4585468	1539	Aldehyde dehydrogenase B
#3765	575275	575760	486	486	hypothetical protein, partial [Lactobacillus hilgardii]	2,00E-06	-1	#0561	574564	577068	2505	Lead, cadmium, zinc and mercury transporting ATPase
#65360	1558540	1559025	486	486	hypothetical protein, partial [Bacillus mycoides]	1,00E-09	-1	#1625	1557769	1560282	2514	Glucans biosynthesis glucosyltransferase H
#32870	4772610	4773092	483	333	hypothetical protein [Klebsiella pneumoniae]	2,00E-05	-1	#5039	4771605	4772942	1338	Xanthine/uracil/thiamine/ascorbate permease family protein
#40656	5167126	5167608	483	483	hypothetical protein [Yersinia pestis]	5,00E-08	-1	#5388	5166334	5167749	1416	Replicative DNA helicase
#41636	5028679	5029161	483	483	hypothetical protein, partial [Catenibacterium mitsuokai]	3,00E-19	-1	#5278	5028343	5030523	2181	Catalase
#61974	2005873	2006355	483	483	transcription termination factor Rho, partial [Escherichia coli]	9,00E-06	-1	#2149	2005717	2007036	1320	redicted glycoside hydrolase
#6386	971263	971745	483	411	L-asparaginase II, partial [Escherichia coli]	1,00E-14	-1	#0934	971335	972081	747	Glutamine ABC transporter, periplasmic glutamine-binding protein
#69903	953031	953513	483	483	hypothetical protein, partial [Streptomyces purpeofuscus]	1,00E-05	-1	#0918	953010	954377	1368	ATP-dependent RNA helicase RhlE
#8646	1310776	1311258	483	426	hypothetical protein [Yokenella regensburgei]	3,00E-04	-1	#1320	1310833	1311279	447	Low molecular weight protein-tyrosine-phosphatase Wzb
#15883	2381014	2381493	480	480	hypothetical protein [Alistipes finegoldii CAG:68]	4,00E-05	-1	#2564	2380993	2382639	1647	Fumarate hydratase class I, aerobic
#17899	2663859	2664338	480	480	hypothetical protein [Thioclava pacifica]	7,00E-04	-1	#2858	2663844	2664893	1050	Chemotaxis response regulator protein-glutamate methylesterase CheB
#41151	5096444	5096923	480	102	hypothetical protein [Aeromonas sanarellii]	1,00E-04	-1	#5330	5095955	5096545	591	hypothetical protein
#42324	4930778	4931257	480	480	hypothetical protein [Bacteroides coprophilus CAG:333]	3,00E-05	-1	#5180	4930544	4933330	2787	DNA polymerase I
#46354	4333683	4334159	477	477	hypothetical protein, partial [Staphylococcus hominis]	4,00E-41	-1	#4603	4332639	4334261	1623	Phosphoenolpyruvate carboxykinase [ATP]
#51092	3624220	3624696	477	159	hypothetical protein [Xenorhabdus poinarii]	4,00E-15	-1	#3852	3623398	3624378	981	Multidrug resistance protein A
#67304	1299819	1300295	477	450	hypothetical protein [Sutterella wadsworthensis CAG:135]	2,00E-14	-1	#1311	1299846	1301639	1794	Uptake hydrogenase large subunit
#1661	252839	253312	474	474	hypothetical protein, partial [Staphylococcus aureus]	3,00E-11	-1	#0225	251126	253897	2772	ClpB protein
#22238	3259336	3259809	474	417	hypothetical protein, partial [Bacillus mycoides]	4,00E-43	-1	#3496	3258928	3259752	825	Murein endopeptidase
#23862	3484428	3484901	474	474	hypothetical protein [Chlorobaculum tepidum]	3,00E-04	-1	#3711	3484242	3485753	1512	putative sugar ABC transport system, ATP-binding protein YphE

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#31654	4589667	4590140	474	474	hypothetical protein [Paracoccus pantotrophus]	2,00E-09	-1	#4850	4588341	4590185	1845	Selenocysteine-specific translation elongation factor
#31999	4647674	4648147	474	360	hypothetical protein [Ketogulonicigenium vulgare]	1,00E-08	-1	#4904	4647317	4648033	717	Ribonuclease PH
#58038	2580983	2581456	474	474	hypothetical protein [Escherichia coli]	5,00E-63	-1	#2770	2580956	2582080	1125	Putative dioxygenase, alpha subunit
#69152	1055025	1055498	474	474	hypothetical protein, partial [Streptococcus pyogenes]	2,00E-06	-1	#1016	1054110	1056386	2277	ATP-dependent Clp protease ATP-binding subunit ClpA
#19520	2886875	2887345	471	471	hypothetical protein [Chromohalobacter israelensis]	3,00E-08	-1	#3122	2886764	2888200	1437	Mannose-1-phosphate guanylyltransferase (GDP)
#22893	3350588	3351058	471	471	phenol hydroxylase [Natronorubrum tibetense]	3,00E-04	-1	#3585	3350039	3351136	1098	Sulfate and thiosulfate import ATP-binding protein CysA
#29611	4298361	4298831	471	438	hypothetical protein, partial [Vibrio parahaemolyticus]	1,00E-32	-1	#4565	4298394	4298996	603	Cell filamentation protein fic
#60912	2170921	2171391	471	471	hypothetical protein [Rhodococcus qingshengii]	1,00E-78	-1	#2323	2169538	2172138	2601	Exodeoxyribonuclease VIII
#70712	835830	836300	471	471	hypothetical protein [Mycobacterium tuberculosis]	7,00E-04	-1	#0795	835128	837929	2802	2-oxoglutarate dehydrogenase E1 component
#75090	208839	209309	471	471	hypothetical protein [Megasphaera elsdenii CAG:570]	1,00E-04	-1	#0189	208470	211952	3483	DNA polymerase III alpha subunit
#40647	5168030	5168497	468	468	hypothetical protein, partial [Escherichia coli]	3,00E-32	-1	#5389	5167802	5168881	1080	Alanine racemase
#35921	5221874	5222338	465	465	hypothetical protein [Pseudomonas aeruginosa]	5,00E-36	-1	#5443	5221727	5222485	759	Phosphonates transport ATP-binding protein PhnK
#39964	5276165	5276629	465	315	hypothetical protein [Escherichia albertii]	2,00E-11	-1	#5492	5276126	5276479	354	putative membrane protein yjel
#44466	4611996	4612460	465	465	tyrosine protein kinase, partial [Salmonella enterica]	2,00E-06	-1	#4867	4611867	4613522	1656	L-lactate permease
#64638	1652227	1652691	465	465	hypothetical protein [Escherichia coli]	7,00E-64	-1	#1741	1652149	1654086	1938	hypothetical protein
#65196	1580164	1580628	465	465	hypothetical protein, partial [Escherichia coli]	1,00E-32	-1	#1653	1579453	1580658	1206	Flagellar hook protein FlgE
#72417	585067	585531	465	465	hypothetical protein [Escherichia coli]	2,00E-06	-1	#0565	581353	585738	4386	Large repetitive protein
#26042	3785972	3786433	462	117	hypothetical protein [Klebsiella oxytoca]	4,00E-13	-1	#4017	3786119	3786235	117	hypothetical protein
#35816	5207537	5207998	462	462	hypothetical protein, partial [Escherichia coli]	1,00E-22	-1	#5427	5206163	5208148	1986	hypothetical protein
#51542	3557603	3558064	462	264	heat-shock protein GrpE [Kluyvera ascorbata]	1,00E-12	-1	#3776	3557801	3558607	807	NAD kinase
#72357	593542	594003	462	462	hypothetical protein [Escherichia coli]	2,00E-06	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#1091	164484	164942	459	444	hypothetical protein [Escherichia coli]	3,00E-14	-1	#0149	164472	164927	456	C4-type zinc finger protein, DksA/TraR family
#1651	251540	251998	459	459	hypothetical protein [Klebsiella pneumoniae]	1,00E-11	-1	#0225	251126	253897	2772	ClpB protein
#25007	3642912	3643370	459	459	hypothetical protein, partial [Acidovorax avenae]	5,00E-07	-1	#3876	3642324	3643838	1515	Anaerobic nitric oxide reductase transcription regulator NorR
#4185	649591	650049	459	294	hypothetical protein [Shewanella algae]	4,00E-16	-1	#0614	649018	649884	867	Methylenetetrahydrofolate dehydrogenase (NADP+)
#54341	3125956	3126414	459	309	hypothetical protein [Shigella dysenteriae]	3,00E-20	-1	#3375	3125776	3126264	489	Proteinase inhibitor I11, ecotin precursor
#5470	829658	830116	459	459	hypothetical protein, partial [Rathayibacter toxicus]	7,00E-09	-1	#0787	829607	830890	1284	Citrate synthase (si)
#58060	2578056	2578514	459	459	LysR family transcriptional regulator [Escherichia coli]	4,00E-25	-1	#2768	2577978	2579063	1086	Tartrate dehydrogenase
#74697	266074	266532	459	459	hypothetical protein, partial [Halomonas anticariensis]	1,00E-06	-1	#0239	264862	267003	2142	VgrG protein
#23699	3462951	3463406	456	456	hypothetical protein, partial [Brevibacterium album]	2,00E-08	-1	#3685	3462621	3463904	1284	Peptidase B
#24148	3523193	3523648	456	345	hypothetical protein [Escherichia coli]	3,00E-05	-1	#3744	3523154	3523537	384	Pyruvate formate-lyase
#24945	3635415	3635870	456	456	hypothetical protein, partial [Escherichia coli]	7,00E-20	-1	#3864	3635397	3635894	498	C-terminal domain of CinA type S
#44519	4605017	4605472	456	363	hypothetical protein [Escherichia coli]	5,00E-73	-1	#4864	4605101	4605463	363	hypothetical protein
#72996	500937	501392	456	423	hypothetical protein [Escherichia coli]	3,00E-08	-1	#0486	500853	501359	507	Phosphatidylglycerophosphatase A
#15890	2382049	2382501	453	453	hypothetical protein [Alistipes finegoldii CAG:68]	3,00E-09	-1	#2564	2380993	2382639	1647	Fumarate hydratase class I, aerobic
#23185	3388112	3388564	453	429	hypothetical protein [Enterobacter ludwigii]	1,00E-08	-1	#3622	3386561	3388540	1980	Glutamate synthase [NADPH] small chain
#35935	5223371	5223823	453	453	hypothetical protein [Nitratireductor aquibiodomus]	4,00E-18	-1	#5445	5223320	5224384	1065	PhnI protein
#45840	4413382	4413834	453	453	hypothetical protein, partial [Bacillus cereus]	8,00E-34	-1	#4682	4413325	4415523	2199	Lead, cadmium, zinc and mercury transporting ATPase
#49073	3933262	3933714	453	276	hypothetical protein [Bordetella holmesii]	5,00E-07	-1	#4176	3933313	3933588	276	putative Fe(2+)-trafficking protein YggX

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#51271	3601348	3601800	453	453	hypothetical protein, partial [Streptococcus pneumoniae]	7,00E-13	-1	#3824	3601081	3602529	1449	Succinate-semialdehyde dehydrogenase [NADP+]
#72127	626866	627318	453	453	hypothetical protein, partial [Bacillus cereus]	2,00E-05	-1	#0592	626200	627981	1782	Glyoxylate carboligase
#72193	617582	618034	453	453	hypothetical protein, partial [Escherichia coli]	4,00E-51	-1	#0585	616748	620944	4197	core protein
#74500	293997	294449	453	327	hypothetical protein [Rhodococcus qingshengii]	8,00E-99	-1	#0271	293865	294323	459	Xanthine-guanine phosphoribosyltransferase
#17983	2673068	2673517	450	450	hypothetical protein [Pseudomonas fuscovaginae]	3,00E-06	-1	#2865	2672642	2673529	888	Flagellar motor rotation protein MotA
#18363	2718776	2719225	450	384	hypothetical protein [Pseudomonas stutzeri]	6,00E-06	-1	#2931	2718842	2720599	1758	Flagellar biosynthesis protein FliC
#22048	3236972	3237421	450	450	hypothetical protein [Burkholderia mimosarum]	1,00E-15	-1	#3472	3236912	3237685	774	Histidine ABC transporter, ATP-binding protein HisP
#47780	4119877	4120326	450	450	hypothetical protein [Rhodococcus qingshengii]	6,00E-41	-1	#4368	4117906	4120497	2592	type 1 fimbriae anchoring protein FimD
#51729	3529881	3530330	450	450	hypothetical protein [Escherichia coli]	6,00E-60	-1	#3750	3529815	3531170	1356	CDP-diacylglycerol--serine O-phosphatidyltransferase
#51817	3517990	3518439	450	450	hypothetical protein [Erwinia amylovora]	9,00E-06	-1	#3739	3517684	3519306	1623	L-aspartate oxidase
#64434	1678092	1678541	450	450	hypothetical protein, partial [Escherichia coli]	5,00E-15	-1	#1771	1677786	1678910	1125	Tripeptide aminopeptidase
#74071	354667	355116	450	450	hypothetical protein [Salmonella enterica]	4,00E-10	-1	#0339	353932	358185	4254	Putative adhesin
#10201	1547723	1548169	447	444	hypothetical protein [Escherichia coli]	2,00E-35	-1	#1611	1547726	1548559	834	Curli production assembly/transport component CsgG
#48839	3967154	3967600	447	420	hypothetical protein [Delftia tsuruhatensis]	5,00E-15	-1	#4211	3966707	3967573	867	putative GST-like protein yghU associated with glutathionylspermidine synthetase/amidase
#13440	2051511	2051954	444	444	hypothetical protein, partial [Photorhabdus luminescens]	1,00E-08	-1	#2190	2049939	2052047	2109	VgrG protein
#14733	2222624	2223067	444	414	hypothetical protein [Megasphaera elsdenii CAG:570]	1,00E-05	-1	#2380	2221772	2223037	1266	Gamma-aminobutyrate:alpha-ketoglutarate aminotransferase
#19486	2881082	2881525	444	444	hypothetical protein [Escherichia coli]	2,00E-21	-1	#3117	2880812	2882092	1281	Colanic acid biosysnthesis protein WcaK
#44353	4629032	4629475	444	345	hypothetical protein, partial [Escherichia coli]	5,00E-14	-1	#4883	4628444	4629376	933	ADP-L-glycero-D-manno-heptose-6-epimerase
#60916	2170363	2170806	444	444	hypothetical protein [Rhodococcus qingshengii]	3,00E-62	-1	#2323	2169538	2172138	2601	Exodeoxyribonuclease VIII
#65644	1524340	1524783	444	444	hypothetical protein [Escherichia coli]	6,00E-37	-1	#1574	1524226	1525458	1233	Co-activator of prophage gene expression IbrA
#68647	1124196	1124639	444	444	hypothetical protein [Escherichia coli]	6,00E-37	-1	#1113	1124082	1125314	1233	Co-activator of prophage gene expression IbrA
#10796	1639073	1639513	441	180	hypothetical protein [Plesiomonas shigelloides]	3,00E-04	-1	#1719	1639253	1639432	180	hypothetical protein
#28895	4199184	4199624	441	441	hypothetical protein [Halomonas zincidurans]	3,00E-08	-1	#4450	4199142	4199639	498	ClpXP protease specificity-enhancing factor
#3390	515561	516001	441	441	hypothetical protein, partial [Piscirickettsia salmonis]	3,00E-10	-1	#0501	514703	516694	1992	Cytochrome O ubiquinol oxidase subunit I
#35952	5225751	5226191	441	393	hypothetical protein [Klebsiella pneumoniae]	4,00E-11	-1	#5448	5225418	5226143	726	Transcriptional regulator PhnF
#46337	4336495	4336935	441	303	hypothetical protein [Pseudomonas fuscovaginae]	3,00E-12	-1	#4606	4336633	4337109	477	Transcription elongation factor GreB
#7695	1167612	1168052	441	441	hypothetical protein [Rhodococcus qingshengii]	5,00E-101	-1	#1165	1167315	1168055	741	Pyruvate formate-lyase activating enzyme
#8326	1262181	1262621	441	180	hypothetical protein [Plesiomonas shigelloides]	3,00E-04	-1	#1259	1262361	1262540	180	hypothetical protein
#17631	2628585	2629022	438	438	hypothetical protein, partial [Staphylococcus aureus]	6,00E-15	-1	#2823	2628408	2629049	642	4-hydroxy-2-oxoglutarate aldolase
#75158	201171	201608	438	336	hypothetical protein [Escherichia coli]	1,00E-24	-1	#0182	201273	203705	2433	Outer membrane protein assembly factor YaeT precursor
#22153	3248970	3249404	435	411	hypothetical protein [Escherichia coli]	2,00E-08	-1	#3485	3248994	3250007	1014	Aspartate-semialdehyde dehydrogenase
#6276	952365	952799	435	417	hypothetical protein [Serratia marcescens]	7,00E-06	-1	#0917	952110	952781	672	Transcriptional regulator YbiH, TetR family
#28631	4161365	4161796	432	369	hypothetical protein, partial [Rhodococcus qingshengii]	9,00E-34	-1	#4411	4160561	4161733	1173	GTP-binding protein Obg
#37459	5453527	5453958	432	300	hypothetical protein [Escherichia coli]	3,00E-32	-1	#5656	5453659	5455002	1344	Fructuronate transporter GntP
#49065	3934349	3934780	432	384	hypothetical protein, partial [Shigella flexneri]	5,00E-34	-1	#4177	3933653	3934732	1080	Membrane-bound lytic murein transglycosylase C precursor
#54484	3105751	3106182	432	285	hypothetical protein [Aeromonas dhakensis]	3,00E-07	-1	#3352	3105766	3106050	285	LSU ribosomal protein L25p
#30287	4393924	4394352	429	429	phenol hydroxylase [Natronorubrum tibetense]	2,00E-09	-1	#4657	4393396	4394466	1071	Glycerol-3-phosphate ABC transporter, ATP-binding protein UgpC
#37069	5387735	5388163	429	384	hypothetical protein, partial [Bacillus cereus]	4,00E-41	-1	#5607	5385263	5388118	2856	Valyl-tRNA synthetase
#41178	5093519	5093947	429	429	hypothetical protein, partial [Vibrio parahaemolyticus]	1,00E-43	-1	#5327	5093354	5094127	774	NADH pyrophosphatase

sORF with blastp hit							mother gene					
ID CP00895 7.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	rea-ding frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#12183	1844759	1845184	426	426	hypothetical protein [Rhodococcus qingshengii]	6,00E-35	-1	#1962	1844156	1846627	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#38452	5499869	5500294	426	426	hypothetical protein, partial [Staphylococcus aureus]	3,00E-05	-1	#5698	5498879	5500543	1665	Methyl-accepting chemotaxis protein I (serine chemoreceptor protein)
#39626	5323635	5324060	426	336	hypothetical protein, partial [Escherichia coli]	2,00E-18	-1	#5538	5322573	5323970	1398	Ascorbate-specific PTS system, EIIC component
#43049	4837230	4837655	426	426	hypothetical protein, partial [Pseudomonas amygdali]	1,00E-11	-1	#5093	4836402	4837946	1545	Threonine dehydratase biosynthetic
#5283	801875	802300	426	426	hypothetical protein, partial [Bordetella holmesii]	4,00E-32	-1	#0765	799934	802618	2685	Osmosensitive K+ channel histidine kinase KdpD
#56481	2811809	2812234	426	426	hypothetical protein [Escherichia coli]	2,00E-86	-1	#3049	2808089	2815951	7863	adherence and invasion outermembrane protein (Inv,enhances Peyer's patches colonization)
#26058	3788276	3788698	423	423	hypothetical protein, partial [Acinetobacter haemolyticus]	2,00E-14	-1	#4019	3787664	3788926	1263	Diaminopimelate decarboxylase
#28643	4162631	4163053	423	213	hypothetical protein [Haloglycomyces albus]	8,00E-04	-1	#4413	4162841	4163098	258	LSU ribosomal protein L27p
#29481	4279041	4279463	423	423	hypothetical protein [Corynebacterium glutamicum]	4,00E-12	-1	#4543	4278933	4281047	2115	Translation elongation factor G
#34102	4955427	4955849	423	423	hypothetical protein [Escherichia coli]	3,00E-19	-1	#5200	4954713	4955954	1242	Aldose-ketose isomerase YihS
#44867	4555184	4555606	423	213	hypothetical protein [Alcaligenes faecalis]	4,00E-06	-1	#4821	4555247	4555459	213	Cold shock protein CspA
#72340	596611	597033	423	423	hypothetical protein [Escherichia coli]	3,00E-04	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#74673	269281	269703	423	423	hypothetical protein, partial [Escherichia coli]	3,00E-28	-1	#0240	267079	271293	4215	core protein
#11029	1676973	1677392	420	420	phenol hydroxylase [Natronorubrum tibetense]	1,00E-04	-1	#1769	1676298	1677434	1137	Putrescine transport ATP-binding protein PotA
#1845	280887	281306	420	420	hypothetical protein, partial [Bordetella bronchiseptica]	2,00E-07	-1	#0251	278928	281372	2445	Butyryl-CoA dehydrogenase
#41714	5019029	5019448	420	420	hypothetical protein [Escherichia coli]	6,00E-46	-1	#5269	5016503	5020687	4185	core protein
#44587	4595034	4595453	420	420	hypothetical protein [Escherichia coli]	6,00E-46	-1	#4854	4592508	4596737	4230	core protein
#51756	3526428	3526847	420	420	hypothetical protein, partial [Escherichia coli]	1,00E-40	-1	#3748	3526311	3527009	699	hypothetical protein
#58558	2508968	2509387	420	420	hypothetical protein, partial [Cecembia lonarensis]	2,00E-07	-1	#2691	2508041	2510302	2262	Catalase
#70906	810312	810731	420	420	hypothetical protein [Escherichia coli]	6,00E-46	-1	#0770	807786	811985	4200	core protein
#16688	2495122	2495538	417	417	protein of PilT N-term./Vapc superfamily [Bacteroides eggerthii CAG:109]	1,00E-07	-1	#2678	2494816	2496744	1929	Threonyl-tRNA synthetase
#17968	2671340	2671756	417	369	hypothetical protein [Escherichia coli]	3,00E-51	-1	#2863	2669750	2671708	1959	Signal transduction histidine kinase CheA
#23129	3380316	3380732	417	417	hypothetical protein [Halarchaeum acidiphilum]	7,00E-08	-1	#3617	3379113	3381392	2280	NADP-dependent malic enzyme
#51171	3614796	3615212	417	417	hypothetical protein, partial [Streptococcus pyogenes]	5,00E-08	-1	#3841	3613119	3615263	2145	Ribonucleotide reductase of class Ib (aerobic), alpha subunit
#58562	2508500	2508916	417	417	transcriptional regulator [Coxiella burnetii]	2,00E-04	-1	#2691	2508041	2510302	2262	Catalase
#15801	2369288	2369701	414	414	hypothetical protein [Pseudomonas oryzihabitans]	2,00E-07	-1	#2553	2369234	2370622	1389	NAD(P) transhydrogenase subunit beta
#19546	2889980	2890393	414	414	hypothetical protein [Bordetella hinzii]	2,00E-04	-1	#3125	2889905	2890870	966	GDP-L-fucose synthetase
#23535	3443044	3443457	414	414	hypothetical protein, partial [Bacillus cereus]	2,00E-48	-1	#3671	3442333	3443607	1275	Histidyl-tRNA synthetase
#35378	5143298	5143711	414	414	hypothetical protein, partial [Escherichia coli]	2,00E-16	-1	#5368	5142920	5144395	1476	D-xylose proton-symporter XylE
#35750	5200468	5200881	414	414	hypothetical protein [Vibrio parahaemolyticus]	5,00E-62	-1	#5420	5198770	5200917	2148	Formate dehydrogenase H, selenocysteine-containing
#36136	5249959	5250372	414	414	hypothetical protein [Alistipes finegoldii CAG:68]	2,00E-12	-1	#5468	5248996	5250642	1647	Fumarate hydratase class I, anaerobic
#444	67691	68104	414	414	hypothetical protein [Escherichia coli]	1,00E-10	-1	#0061	65291	68197	2907	RNA polymerase associated protein RapA
#63118	1862391	1862804	414	303	hypothetical protein, partial [Escherichia coli]	2,00E-15	-1	#1988	1862502	1863035	534	hypothetical protein
#64690	1646002	1646415	414	303	hypothetical protein, partial [Escherichia coli]	3,00E-15	-1	#1731	1646113	1646646	534	Phage lysin, 1,4-beta-N-acetylmuramidase or lysozyme
#36300	5269677	5270087	411	321	hypothetical protein [Escherichia coli]	2,00E-28	-1	#5485	5268789	5269997	1209	C4-dicarboxylate transporter DcuA
#68100	1194182	1194592	411	411	hypothetical protein [Escherichia coli]	3,00E-08	-1	#1187	1193300	1197760	4461	Chromosome partition protein MukB
#14606	2207338	2207745	408	408	phenol hydroxylase [Natronorubrum tibetense]	1,00E-04	-1	#2361	2206678	2207760	1083	Multiple sugar ABC transporter, ATP-binding protein
#23132	3380751	3381158	408	408	hypothetical protein [Halarchaeum acidiphilum]	3,00E-18	-1	#3617	3379113	3381392	2280	NADP-dependent malic enzyme
#484	73086	73493	408	99	hypothetical protein, partial [Staphylococcus aureus]	6,00E-23	-1	#0064	71682	73184	1503	L-arabinose isomerase

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#55740	2924863	2925270	408	342	hypothetical protein, partial [Acinetobacter haemolyticus]	2,00E-10	-1	#3151	2922127	2925204	3078	Multidrug transporter MdtC
#63155	1858656	1859063	408	408	hypothetical protein [Escherichia coli]	1,00E-08	-1	#1984	1858029	1859087	1059	DNA methyl transferase, phage-associated
#64622	1654062	1654469	408	162	hypothetical protein [Escherichia coli]	7,00E-12	-1	#1742	1654191	1654352	162	hypothetical protein
#69981	939405	939812	408	408	hypothetical protein [Cronobacter sakazakii]	7,00E-07	-1	#0902	939180	940169	990	Molybdenum cofactor biosynthesis protein MoaA
#23719	3465144	3465548	405	405	NAD-specific glutamate dehydrogenase [Haloferax denitrificans]	7,00E-04	-1	#3688	3464631	3466481	1851	Chaperone protein HscA
#23725	3465948	3466352	405	405	hypothetical protein [Escherichia coli]	8,00E-31	-1	#3688	3464631	3466481	1851	Chaperone protein HscA
#49531	3865753	3866157	405	399	hypothetical protein, partial [Salmonella enterica]	3,00E-09	-1	#4097	3865630	3866151	522	Flavodoxin 2
#51549	3556368	3556772	405	405	hypothetical protein, partial [Providencia rettgeri]	2,00E-06	-1	#3773	3555696	3556958	1263	Hemolysin with CBS domain
#53252	3294863	3295267	405	405	hypothetical protein [Actinomyces massiliensis]	2,00E-05	-1	#3534	3294359	3295672	1314	D-serine dehydratase
#56553	2803105	2803509	405	405	cell surface protein [Escherichia coli]	3,00E-16	-1	#3045	2802691	2805162	2472	Exodeoxyribonuclease VIII
#6881	1042911	1043315	405	123	hypothetical protein [Clostridium carboxidivorans]	0,001	-1	#1006	1042065	1043033	969	NADH oxidoreductase hcr
#73699	406223	406627	405	201	hypothetical protein, partial [Staphylococcus aureus]	2,00E-06	-1	#0391	406427	407596	1170	2-methylcitrate synthase
#8255	1251589	1251993	405	405	cell surface protein [Escherichia coli]	3,00E-16	-1	#1241	1249936	1252407	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#16740	2503043	2503444	402	366	hypothetical protein [Sodalis glossinidius]	5,00E-16	-1	#2685	2502872	2503408	537	Putative membrane protein
#31430	4558303	4558704	402	402	oxidoreductase, partial [Escherichia coli]	2,00E-07	-1	#4824	4558201	4559112	912	Glycyl-tRNA synthetase alpha chain
#34430	5004141	5004542	402	402	hypothetical protein [Parabacteroides johnsonii CAG:246]	1,00E-08	-1	#5255	5003394	5004902	1509	Glycerol kinase
#35971	5228078	5228479	402	390	hypothetical protein [Salinispora pacifica]	8,00E-05	-1	#5451	5228090	5228878	789	Phosphonate ABC transporter ATP-binding protein
#42920	4854192	4854593	402	402	hypothetical protein [Bradyrhizobium liaoningense]	1,00E-05	-1	#5108	4853553	4854620	1068	dTDP-glucose 4,6-dehydratase
#44357	4628348	4628749	402	306	hypothetical protein, partial [Escherichia coli]	4,00E-12	-1	#4883	4628444	4629376	933	ADP-L-glycero-D-manno-heptose-6-epimerase
#56864	2754900	2755301	402	402	hypothetical protein, partial [Escherichia coli]	5,00E-15	-1	#2976	2754837	2755688	852	Chaperone protein hchA
#57062	2726937	2727338	402	402	hypothetical protein, partial [Klebsiella pneumoniae]	1,00E-27	-1	#2939	2726805	2727473	669	Gifsy-2 prophage protein
#58806	2470134	2470535	402	402	hypothetical protein [Escherichia coli]	3,00E-54	-1	#2651	2469762	2471357	1596	Coenzyme A transferase
#73823	389190	389591	402	402	hypothetical protein, partial [Escherichia coli]	8,00E-13	-1	#0371	388275	389822	1548	Putative oxidoreductase subunit
#73837	387431	387832	402	402	hypothetical protein, partial [Escherichia coli]	1,00E-48	-1	#0370	387422	388285	864	hypothetical protein
#10327	1566653	1567051	399	399	lysine decarboxylase LdcC, partial [Escherichia coli]	3,00E-04	-1	#1635	1566344	1567462	1119	N-methyl-L-amino-acid oxidase
#12044	1824783	1825181	399	399	hypothetical protein [Salmonella enterica]	2,00E-31	-1	#1943	1824675	1827350	2676	Alcohol dehydrogenase
#20709	3049978	3050376	399	381	hypothetical protein [Ketogulonicigenium vulgare]	9,00E-09	-1	#3298	3049657	3050358	702	Fumarylacetoacetase
#22142	3247404	3247802	399	363	hypothetical protein [Escherichia coli]	1,00E-06	-1	#3483	3247440	3248099	660	DedA protein
#65160	1584378	1584776	399	399	hypothetical protein [Escherichia coli]	2,00E-04	-1	#1658	1584213	1585154	942	Flagellar protein FlgJ [peptidoglycan hydrolase]
#70284	901036	901434	399	147	hypothetical protein [Escherichia coli]	1,00E-92	-1	#0863	901060	901206	147	hypothetical protein
#21782	3198607	3199002	396	396	hypothetical protein [Rhodococcus qingshengii]	5,00E-36	-1	#3433	3198601	3199062	462	ElaA protein
#32712	4750294	4750689	396	396	hypothetical protein [Escherichia coli]	4,00E-52	-1	#5020	4749997	4750809	813	Phosphatase YidA
#3606	550094	550489	396	375	hypothetical protein, partial [Staphylococcus aureus]	7,00E-12	-1	#0537	547319	550468	3150	RND efflux system, inner membrane transporter CmeB
#14736	2223165	2223557	393	393	hypothetical protein, partial [Providencia rettgeri]	6,00E-34	-1	#2381	2223075	2224355	1281	Gamma-glutamyl-putrescine oxidase
#15758	2362485	2362877	393	342	hypothetical protein [Escherichia coli]	1,00E-06	-1	#2545	2361606	2362826	1221	Mlc, transcriptional repressor of MalT (the transcriptional activator of maltose regulon) and manXYZ operon
#31217	4532279	4532671	393	360	hypothetical protein [Escherichia coli]	1,00E-30	-1	#4798	4531634	4532638	1005	Dipeptide transport ATP-binding protein DppF
#41733	5017004	5017396	393	393	hypothetical protein [Escherichia coli]	1,00E-85	-1	#5269	5016503	5020687	4185	core protein
#44606	4593009	4593401	393	393	hypothetical protein [Escherichia coli]	1,00E-85	-1	#4854	4592508	4596737	4230	core protein
#6374	969843	970235	393	393	hypothetical protein [Streptomyces viridochromogenes]	2,00E-07	-1	#0932	969819	970541	723	Glutamate transport ATP-binding protein

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#66807	1367047	1367439	393	387	hypothetical protein, partial [Xanthomonas hyacinthi]	5,00E-68	-1	#1400	1367053	1368060	1008	hypothetical protein
#70925	808287	808679	393	393	hypothetical protein [Escherichia coli]	1,00E-85	-1	#0770	807786	811985	4200	core protein
#28396	4128717	4129106	390	255	hypothetical protein, partial [Escherichia coli]	6,00E-24	-1	#4379	4128336	4128971	636	Oxidoreductase
#30715	4459238	4459627	390	282	hypothetical protein [Salmonella enterica]	1,00E-20	-1	#4729	4458317	4459519	1203	NAD(FAD)-utilizing dehydrogenases
#56545	2804134	2804523	390	390	hypothetical protein [Rhodococcus qingshengii]	5,00E-25	-1	#3045	2802691	2805162	2472	Exodeoxyribonuclease VIII
#59693	2347269	2347658	390	390	hypothetical protein [Rhodococcus qingshengii]	4,00E-25	-1	#2528	2345826	2348297	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#71575	708230	708619	390	390	murein hydrolase transporter LrgA [Arenibacter latericius]	4,00E-04	-1	#0664	707318	708928	1611	2,3-dihydroxybenzoate-AMP ligase [enterobactin] siderophore
#72793	534482	534871	390	390	hypothetical protein [Campylobacter coli]	1,00E-33	-1	#0521	534446	534904	459	Putative HTH-type transcriptional regulator ybaO
#8247	1250575	1250964	390	390	hypothetical protein [Rhodococcus qingshengii]	5,00E-25	-1	#1241	1249936	1252407	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#1009	153786	154172	387	387	hypothetical protein [Cronobacter universalis]	2,00E-07	-1	#0138	153399	154193	795	3-methyl-2-oxobutanoate hydroxymethyltransferase
#14888	2243928	2244314	387	246	hypothetical protein [Escherichia coli]	3,00E-45	-1	#2399	2243607	2244173	567	Transcriptional regulator, TetR family
#22149	3248338	3248724	387	387	hypothetical protein, partial [Escherichia coli]	5,00E-33	-1	#3484	3248182	3248994	813	tRNA pseudouridine synthase A
#50186	3759196	3759582	387	315	amino acid acetyltransferase [Escherichia coli]	4,00E-35	-1	#3996	3759268	3760599	1332	N-acetylglutamate synthase
#5388	818038	818424	387	387	hypothetical protein [Escherichia coli]	1,00E-22	-1	#0775	817498	818979	1482	Di/tripeptide permease YbgH
#72346	595411	595797	387	387	hypothetical protein [Escherichia coli]	0,001	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72368	591820	592206	387	387	hypothetical protein [Escherichia coli]	7,00E-08	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#1926	293203	293586	384	384	hypothetical protein [Bacteroides ovatus CAG:22]	1,00E-03	-1	#0270	292147	293604	1458	Aminoacyl-histidine dipeptidase (Peptidase D)
#20653	3043633	3044016	384	276	hypothetical protein [Shigella dysenteriae]	3,00E-27	-1	#3293	3043741	3044502	762	3-oxoacyl-[acyl-carrier protein] reductase
#31213	4531862	4532245	384	384	hypothetical protein [Haloferax gibbonsii]	1,00E-13	-1	#4798	4531634	4532638	1005	Dipeptide transport ATP-binding protein DppF
#35868	5214180	5214563	384	384	hypothetical protein [Piscirickettsia salmonis]	4,00E-05	-1	#5434	5213685	5215205	1521	Ribose ABC transport system, ATP-binding protein RbsA
#40829	5141557	5141940	384	315	hypothetical protein [Escherichia albertii]	4,00E-11	-1	#5364	5139775	5141871	2097	YjbH outer membrane lipoprotein
#4847	744426	744809	384	384	hypothetical protein, partial [Escherichia coli]	7,00E-13	-1	#0705	743886	745097	1212	D-alanyl-D-alanine carboxypeptidase
#5197	790889	791272	384	378	hypothetical protein, partial [Providencia rettgeri]	9,00E-22	-1	#0756	790736	791266	531	Flavodoxin 1
#76192	51655	52038	384	378	hypothetical protein [Pseudomonas fluorescens]	3,00E-08	-1	#0047	51661	52191	531	Glutathione-regulated potassium-efflux system ancillary protein KefF
#11773	1786016	1786396	381	381	hypothetical protein [Escherichia coli]	1,00E-75	-1	#1905	1785302	1786945	1644	Putative adhesion and penetration protein
#12889	1955831	1956211	381	381	hypothetical protein, partial [Escherichia coli]	4,00E-06	-1	#2112	1955690	1956571	882	LysR family transcriptional regulator YneJ
#61262	2117742	2118122	381	381	hypothetical protein [Bacteroides stercoris CAG:120]	5,00E-11	-1	#2255	2117499	2121023	3525	Pyruvate-flavodoxin oxidoreductase
#71652	697364	697744	381	381	hypothetical protein, partial [Bacillus cereus]	2,00E-07	-1	#0656	695591	699472	3882	Enterobactin synthetase component F, serine activating enzyme
#72370	591349	591729	381	381	hypothetical protein [Escherichia coli]	6,00E-47	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72754	538718	539098	381	282	hypothetical protein, partial [Anaerococcus lactolyticus]	7,00E-07	-1	#0525	538661	538999	339	Nitrogen regulatory protein P-II, glnK
#10651	1618137	1618514	378	378	hypothetical protein [Clostridium clostridioforme CAG:132]	1,00E-06	-1	#1692	1617384	1620830	3447	Transcription-repair coupling factor
#12790	1941020	1941397	378	378	hypothetical protein, partial [Escherichia coli]	2,00E-13	-1	#2093	1940852	1941538	687	Transcriptional regulator, GntR family
#46672	4289323	4289700	378	378	hypothetical protein, partial [Escherichia coli]	2,00E-80	-1	#4556	4288822	4290735	1914	Glutathione-regulated potassium-efflux system ATP-binding protein
#54705	3073861	3074238	378	204	hypothetical protein, partial [Escherichia coli]	8,00E-19	-1	#3322	3073015	3074064	1050	Putative membrane protein YeiH
#13768	2095578	2095952	375	375	hypothetical protein [Escherichia coli]	1,00E-04	-1	#2238	2092992	2096894	3903	ATP-dependent helicase HrpA
#15032	2264143	2264517	375	375	hypothetical protein [Prevotella copri CAG:164]	2,00E-05	-1	#2418	2262043	2264640	2598	DNA topoisomerase I
#19244	2846690	2847064	375	375	hypothetical protein, partial [Escherichia coli]	4,00E-10	-1	#3087	2846519	2847877	1359	Putrescine importer
#22106	3242980	3243354	375	375	hypothetical protein [Cyclobacterium qasimii]	1,00E-07	-1	#3478	3241897	3243414	1518	Amidophosphoribosyltransferase
#24285	3539970	3540344	375	375	hypothetical protein, partial [Streptococcus pyogenes]	1,00E-04	-1	#3755	3538608	3541181	2574	ClpB protein

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#272	43552	43926	375	276	hypothetical protein, partial [Escherichia coli]	3,00E-28	-1	#0040	43651	44793	1143	Crotonobetainyl-CoA dehydrogenase
#44483	4609502	4609876	375	375	hypothetical protein [Escherichia coli]	1,00E-36	-1	#4866	4606733	4611499	4767	hypothetical protein
#44911	4548871	4549245	375	375	hypothetical protein, partial [Escherichia coli]	3,00E-05	-1	#4815	4548850	4549290	441	Acetyltransferase
#59200	2414135	2414509	375	375	hypothetical protein [Cobetia crustatorum]	2,00E-04	-1	#2597	2414063	2414530	468	Outer membrane lipoprotein pcp precursor
#39736	5309327	5309698	372	372	hypothetical protein [Rhodococcus qingshengii]	7,00E-48	-1	#5524	5309246	5311687	2442	3'-to-5' exoribonuclease RNase R
#566	83708	84079	372	354	hypothetical protein, partial [Escherichia coli]	2,00E-18	-1	#0073	83456	84061	606	3-isopropylmalate dehydratase small subunit
#65720	1516553	1516924	372	351	hypothetical protein [Enterobacter cancerogenus]	3,00E-06	-1	#1564	1516574	1517197	624	hypothetical protein
#68723	1116409	1116780	372	351	hypothetical protein [Enterobacter cancerogenus]	3,00E-06	-1	#1103	1116430	1117053	624	hypothetical protein
#37034	5382249	5382617	369	369	hypothetical protein, partial [Escherichia coli]	1,00E-06	-1	#5603	5381688	5382692	1005	Ornithine carbamoyltransferase
#14183	2149472	2149837	366	255	hypothetical protein, partial [Cronobacter universalis]	3,00E-09	-1	#2290	2149583	2150041	459	putative enzyme
#15064	2268003	2268368	366	366	hypothetical protein [Simplicispira psychrophila]	7,00E-04	-1	#2423	2267961	2268833	873	Ribosomal large subunit pseudouridine synthase B
#15509	2328848	2329213	366	255	hypothetical protein, partial [Cronobacter universalis]	1,00E-09	-1	#2498	2328959	2329426	468	putative endopeptidase
#27756	4030384	4030749	366	285	hypothetical protein, partial [Escherichia coli]	3,00E-26	-1	#4278	4030465	4031286	822	Undecaprenyl-diphosphatase
#21079	3103493	3103855	363	240	aldose-1-epimerase [Escherichia coli]	2,00E-17	-1	#3350	3103037	3103732	696	Ribosomal small subunit pseudouridine synthase A
#25090	3655538	3655900	363	363	hypothetical protein, partial [Vibrio parahaemolyticus]	2,00E-28	-1	#3889	3655520	3656062	543	Formate hydrogenlyase complex 3 iron-sulfur protein
#75364	172527	172889	363	363	hypothetical protein [Dickeya dadantii]	3,00E-05	-1	#0155	171807	174050	2244	Ferric hydroxamate outer membrane receptor FhuA
#2779	424572	424931	360	360	hypothetical protein [Streptomyces rimosus]	3,00E-05	-1	#0406	424323	425276	954	Transcriptional regulator, AraC family
#30165	4378141	4378500	360	165	hypothetical protein, partial [Escherichia coli]	2,00E-45	-1	#4639	4377202	4378305	1104	Aspartate-semialdehyde dehydrogenase
#34499	5012681	5013040	360	360	hypothetical protein, partial [Bacillus cereus]	1,00E-07	-1	#5265	5012549	5013541	993	Transcriptional (co)regulator CytR
#50342	3736445	3736804	360	360	hypothetical protein, partial [Escherichia coli]	3,00E-32	-1	#3976	3736334	3737698	1365	Decarboxylase family protein
#59762	2337515	2337874	360	180	hypothetical protein [Plesiomonas shigelloides]	1,00E-04	-1	#2512	2337596	2337775	180	hypothetical protein
#23501	3438934	3439290	357	357	hypothetical protein, partial [Bacillus cereus]	1,00E-04	-1	#3668	3438916	3440388	1473	GTP-binding protein EngA
#31968	4642568	4642924	357	168	hypothetical protein [Pectobacterium carotovorum]	2,00E-28	-1	#4897	4642733	4642900	168	LSU ribosomal protein L33p, zinc-independent
#35320	5134854	5135210	357	357	hypothetical protein [Cronobacter universalis]	2,00E-06	-1	#5359	5134011	5135360	1350	Aspartokinase
#5191	789977	790333	357	333	hypothetical protein [Rhodococcus qingshengii]	4,00E-64	-1	#0755	790001	790447	447	Ferric uptake regulation protein FUR
#63824	1761936	1762292	357	357	hypothetical protein, partial [Bacillus cereus]	3,00E-26	-1	#1881	1761612	1762331	720	Transcriptional regulator for fatty acid degradation FadR, GntR family
#65820	1502451	1502807	357	357	hypothetical protein [Pluralibacter gergoviae]	3,00E-25	-1	#1538	1502250	1502828	579	Tellurium resistance protein TerD
#68409	1154786	1155142	357	357	formate dehydrogenase, partial [Escherichia coli]	1,00E-04	-1	#1154	1151633	1155661	4029	Cell division protein FtsK
#68823	1102307	1102663	357	357	hypothetical protein [Pluralibacter gergoviae]	3,00E-25	-1	#1077	1102106	1102684	579	Tellurium resistance protein TerD
#71707	689939	690295	357	153	hypothetical protein [Escherichia coli]	2,00E-20	-1	#0647	690116	690268	153	HokE protein
#72348	594964	595320	357	357	hypothetical protein [Escherichia coli]	6,00E-40	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#75208	196433	196789	357	342	ribonucleotide-diphosphate reductase subunit beta, partial [Escherichia coli]	1,00E-09	-1	#0176	196217	196774	558	Ribosome recycling factor
#33146	4809166	4809519	354	300	hypothetical protein [Salmonella enterica]	3,00E-30	-1	#5075	4809220	4810671	1452	hypothetical protein
#38037	5535942	5536295	354	252	hypothetical protein [Escherichia coli]	2,00E-13	-1	#5735	5534526	5536193	1668	ABC transporter, ATP-binding protein
#53121	3315031	3315384	354	354	hypothetical protein, partial [Shigella flexneri]	8,00E-65	-1	#3549	3313828	3315507	1680	Autolysis histidine kinase LytS
#66465	1412009	1412362	354	354	hypothetical protein [Pseudomonas extremaustralis]	9,00E-11	-1	#1444	1411622	1412749	1128	Ferrous iron transport periplasmic protein EfeO, contains peptidase-M75 domain and (frequently) cupredoxin-like domain
#17514	2611296	2611646	351	351	hypothetical protein, partial [Klebsiella pneumoniae]	4,00E-53	-1	#2804	2611128	2611826	699	ProQ: influences osmotic activation of compatible solute ProP
#18938	2799024	2799374	351	351	ATP-dependent transporter SufC domain protein, partial [Escherichia coli]	2,00E-16	-1	#3038	2798817	2799854	1038	Primosomal protein I
#25632	3733268	3733618	351	351	hypothetical protein, partial [Staphylococcus aureus]	4,00E-09	-1	#3972	3733028	3733810	783	tRNA pseudouridine synthase C

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#34433	5004558	5004908	351	345	hypothetical protein [Parabacteroides johnsonii CAG:246]	5,00E-10	-1	#5255	5003394	5004902	1509	Glycerol kinase
#39731	5310128	5310478	351	351	hypothetical protein [Sutterella wadsworthensis CAG:135]	3,00E-05	-1	#5524	5309246	5311687	2442	3'-to-5' exoribonuclease RNase R
#43209	4818239	4818589	351	243	hypothetical protein, partial [Klebsiella pneumoniae]	5,00E-59	-1	#5082	4818347	4819291	945	Ribokinase
#44435	4615674	4616024	351	351	hypothetical protein [Burkholderia graminis]	1,00E-06	-1	#4870	4615671	4616144	474	tRNA (cytidine(34)-2'-O)-methyltransferase
#45293	4489291	4489641	351	351	hypothetical protein [Cronobacter sakazakii]	2,00E-05	-1	#4764	4489192	4490148	957	Membrane fusion protein of RND family multidrug efflux pump
#50838	3662272	3662622	351	312	hypothetical protein [Klebsiella pneumoniae]	2,00E-04	-1	#3897	3662311	3663183	873	[NiFe] hydrogenase nickel incorporation-associated protein HypB
#64799	1634287	1634637	351	351	ATP-dependent transporter SufC domain protein, partial [Escherichia coli]	2,00E-16	-1	#1708	1633807	1634844	1038	Primosomal protein I
#70593	855768	856118	351	351	hypothetical protein [Escherichia coli]	2,00E-19	-1	#0812	855741	857309	1569	Cytochrome d ubiquinol oxidase subunit I
#73739	399601	399951	351	351	hypothetical protein, partial [Staphylococcus aureus]	8,00E-06	-1	#0384	399580	400629	1050	Alcohol dehydrogenase
#20965	3086015	3086362	348	348	hypothetical protein, partial [Escherichia coli]	8,00E-25	-1	#3333	3085613	3086743	1131	Fructose-specific phosphocarrier protein HPr
#41035	5110486	5110833	348	348	hypothetical protein [Escherichia coli]	2,00E-35	-1	#5340	5110306	5111235	930	Homoserine O-succinyltransferase
#46842	4259829	4260176	348	348	propionyl-CoA synthetase, partial [Citrobacter amalonaticus]	4,00E-04	-1	#4507	4259157	4260533	1377	Trk system potassium uptake protein TrkA
#5747	871743	872090	348	348	hypothetical protein [Escherichia coli]	2,00E-25	-1	#0830	871440	872480	1041	Aldose 1-epimerase
#61277	2116031	2116378	348	348	hypothetical protein [Pseudomonas pseudoalcaligenes]	3,00E-05	-1	#2251	2115440	2116429	990	D-lactate dehydrogenase
#64886	1623428	1623775	348	291	hypothetical protein [Desulfovibrio hydrothermalis]	6,00E-04	-1	#1695	1623485	1624186	702	Lipoprotein releasing system ATP-binding protein LolD
#65561	1535092	1535439	348	348	hypothetical protein, partial [Escherichia coli]	2,00E-38	-1	#1588	1533739	1535916	2178	Putative vimentin
#73281	463803	464150	348	348	hypothetical protein [Escherichia coli]	5,00E-26	-1	#0445	463482	464597	1116	Protein YaiC
#31544	4574840	4575184	345	345	hypothetical protein [Enterobacter cloacae]	6,00E-04	-1	#4838	4574816	4575289	474	Electron transport protein HydN
#33109	4804386	4804730	345	345	hypothetical protein, partial [Pseudomonas syringae]	3,00E-46	-1	#5070	4804125	4804748	624	rRNA small subunit 7-methylguanosine (m7G) methyltransferase GidB
#42313	4931756	4932100	345	345	hypothetical protein [Pseudomonas aeruginosa]	2,00E-14	-1	#5180	4930544	4933330	2787	DNA polymerase I
#59650	2352893	2353237	345	345	hypothetical protein [Sodalis glossinidius]	9,00E-17	-1	#2537	2352332	2354758	2427	Anaerobic dimethyl sulfoxide reductase chain A
#62834	1897924	1898268	345	345	ATP-dependent transporter SufC domain protein, partial [Escherichia coli]	5,00E-16	-1	#2031	1897432	1898475	1044	Primosomal protein I
#71649	697823	698167	345	345	hypothetical protein, partial [Streptomyces afghaniensis]	5,00E-08	-1	#0656	695591	699472	3882	Enterobactin synthetase component F, serine activating enzyme
#72750	539062	539406	345	345	hypothetical protein, partial [Catenibacterium mitsuokai]	2,00E-09	-1	#0526	539029	540315	1287	Ammonium transporter
#13570	2070431	2070772	342	105	hypothetical protein [Escherichia coli]	6,00E-39	-1	#2212	2070098	2070535	438	hypothetical protein
#24082	3514068	3514409	342	333	hypothetical protein [Escherichia coli]	3,00E-74	-1	#3734	3512601	3514400	1800	Translation elongation factor LepA
#24289	3540519	3540860	342	342	hypothetical protein, partial [Staphylococcus aureus]	7,00E-08	-1	#3755	3538608	3541181	2574	ClpB protein
#4155	645325	645666	342	306	hypothetical protein, partial [Escherichia coli]	5,00E-06	-1	#0609	645361	646083	723	UDP-2,3-diacylglucosamine hydrolase
#52464	3411788	3412129	342	342	hypothetical protein, partial [Escherichia coli]	7,00E-07	-1	#3642	3411473	3413188	1716	Hydrogenase-4 component G
#26676	3878531	3878869	339	339	hypothetical protein, partial [Escherichia coli]	3,00E-21	-1	#4111	3878270	3879595	1326	Xaa-Pro aminopeptidase
#31	5625	5963	339	264	hypothetical protein, partial [Escherichia coli]	7,00E-04	-1	#0006	5700	6476	777	UPF0246 protein YaaA
#70743	832003	832341	339	327	hypothetical protein [Escherichia coli]	2,00E-19	-1	#0791	831982	832329	348	Succinate dehydrogenase hydrophobic membrane anchor protein
#3018	459830	460165	336	336	hypothetical protein, partial [Acinetobacter baumannii]	2,00E-11	-1	#0439	459608	460702	1095	D-alanine--D-alanine ligase A
#34288	4983802	4984137	336	336	hypothetical protein, partial [Escherichia coli]	3,00E-14	-1	#5230	4983565	4985034	1470	Rhamnulokinase
#61963	2007317	2007652	336	240	hypothetical protein [Shigella dysenteriae]	4,00E-52	-1	#2150	2007413	2008795	1383	Putative Heme-regulated two-component response regulator
#2488	376798	377130	333	201	hypothetical protein [Pseudomonas putida]	4,00E-06	-1	#0362	376411	376998	588	HTH-type transcriptional regulator BetI
#27096	3937836	3938168	333	333	hypothetical protein, partial [Escherichia coli]	6,00E-04	-1	#4179	3936240	3938375	2136	Ornithine decarboxylase
#3307	504833	505165	333	333	hypothetical protein, partial [Vibrio parahaemolyticus]	5,00E-11	-1	#0489	504368	506230	1863	1-deoxy-D-xylulose 5-phosphate synthase
#42580	4897636	4897968	333	333	hypothetical protein [Pseudomonas pseudoalcaligenes]	7,00E-08	-1	#5151	4896988	4899249	2262	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#42766	4874039	4874371	333	333	hypothetical protein [Escherichia coli]	6,00E-57	-1	#5124	4872437	4874983	2547	Adenylate cyclase
#42782	4872203	4872535	333	99	hypothetical protein [Citrobacter freundii]	2,00E-16	-1	#5124	4872437	4874983	2547	Adenylate cyclase
#43053	4836837	4837169	333	333	hypothetical protein, partial [Piscirickettsia salmonis]	2,00E-04	-1	#5093	4836402	4837946	1545	Threonine dehydratase biosynthetic
#49371	3888509	3888841	333	333	hypothetical protein, partial [Burkholderia glathei]	1,00E-46	-1	#4121	3888089	3889084	996	putative periplasmic protein kinase ArgK and related GTPases of G3E family
#49645	3847851	3848183	333	333	hypothetical protein, partial [Escherichia coli]	7,00E-31	-1	#4082	3847092	3849962	2871	putative hypoxanthine oxidase XdhD
#61883	2015996	2016328	333	333	hypothetical protein [Haloferax gibbonsii]	2,00E-04	-1	#2157	2015525	2016511	987	Dipeptide transport system permease protein DppC
#6513	988350	988682	333	198	hypothetical protein [Escherichia coli]	2,00E-05	-1	#0950	988485	989234	750	Molybdopterin biosynthesis protein MoeB
#70424	880669	881001	333	333	hypothetical protein, partial [Bordetella bronchiseptica]	6,00E-04	-1	#0839	880198	881256	1059	Molybdenum transport ATP-binding protein ModC
#25050	3649001	3649330	330	330	hypothetical protein [Klebsiella pneumoniae]	1,00E-22	-1	#3881	3648923	3649450	528	Fe-S-cluster-containing hydrogenase components 2
#28901	4199972	4200301	330	312	hypothetical protein [Leuconostoc mesenteroides]	6,00E-06	-1	#4451	4199645	4200283	639	Stringent starvation protein A
#52451	3413207	3413536	330	330	hypothetical protein, partial [Vibrio parahaemolyticus]	4,00E-05	-1	#3643	3413198	3413743	546	Hydrogenase-4 component H
#55727	2926288	2926617	330	330	hypothetical protein, partial [Staphylococcus aureus]	1,00E-19	-1	#3152	2925205	2926620	1416	Multidrug transporter MdtD
#16321	2444589	2444915	327	327	hypothetical protein [Morganella morganii]	5,00E-07	-1	#2627	2444517	2445071	555	putative ferredoxin-like protein YdhX
#28987	4211613	4211939	327	327	hypothetical protein, partial [Escherichia coli]	2,00E-30	-1	#4463	4211058	4211987	930	Fusaric acid resistance protein fusE
#33537	4871196	4871522	327	327	hypothetical protein, partial [Anaerococcus lactolyticus]	6,00E-07	-1	#5123	4871109	4872065	957	Porphobilinogen deaminase
#43009	4842269	4842595	327	327	hypothetical protein [Dorea longicatena CAG:42]	4,00E-08	-1	#5098	4841708	4843729	2022	ATP-dependent DNA helicase Rep
#4326	669340	669666	327	327	hypothetical protein [Escherichia coli]	1,00E-51	-1	#0630	667819	672756	4938	Rhs-family protein
#63848	1758381	1758707	327	327	hypothetical protein [Escherichia coli]	1,00E-13	-1	#1878	1757859	1759127	1269	Error-prone, lesion bypass DNA polymerase V (UmuC)
#65639	1525135	1525461	327	324	hypothetical protein [Escherichia coli]	2,00E-25	-1	#1574	1524226	1525458	1233	Co-activator of prophage gene expression IbrA
#68642	1124991	1125317	327	324	hypothetical protein [Escherichia coli]	2,00E-25	-1	#1113	1124082	1125314	1233	Co-activator of prophage gene expression IbrA
#12814	1944678	1945001	324	324	hypothetical protein [Escherichia coli]	4,00E-37	-1	#2096	1944600	1945118	519	C-terminal domain of CinA type S
#17338	2586380	2586703	324	324	hypothetical protein, partial [Escherichia coli]	4,00E-12	-1	#2774	2586242	2586823	582	Starvation lipoprotein Slp-like protein
#18076	2683836	2684159	324	114	hypothetical protein, partial [Neisseria flavescens]	3,00E-04	-1	#2879	2683863	2683976	114	hypothetical protein
#21284	3130605	3130928	324	282	hypothetical protein [Burkholderia lata]	9,00E-07	-1	#3379	3130647	3131711	1065	ADA regulatory protein
#29458	4275613	4275936	324	297	hypothetical protein [Pseudomonas mosselii]	3,00E-28	-1	#4538	4275598	4275909	312	SSU ribosomal protein S10p (S20e)
#31220	4532671	4532994	324	324	hypothetical protein, partial [Staphylococcus aureus]	1,00E-29	-1	#4799	4532635	4533618	984	Dipeptide transport ATP-binding protein DppD
#35021	5090835	5091158	324	324	hypothetical protein, partial [Staphylococcus aureus]	7,00E-07	-1	#5325	5090655	5092550	1896	Thiamin biosynthesis protein ThiC
#41469	5052711	5053034	324	282	hypothetical protein [Acidithiobacillus thiooxidans]	8,00E-04	-1	#5297	5052753	5053670	918	Hydrogen peroxide-inducible genes activator
#42710	4880020	4880343	324	324	hypothetical protein, partial [Bacillus cereus]	1,00E-14	-1	#5132	4879267	4881429	2163	ATP-dependent DNA helicase UvrD/PcrA
#48616	3999604	3999927	324	315	hypothetical protein [Escherichia coli]	6,00E-18	-1	#4249	3999610	3999924	315	hypothetical protein
#49740	3835940	3836263	324	291	hypothetical protein [Shigella flexneri]	1,00E-54	-1	#4074	3835973	3837370	1398	Dihydropyrimidinase
#65054	1598541	1598864	324	141	hypothetical protein [Escherichia coli]	5,00E-13	-1	#1671	1598445	1598681	237	Acyl carrier protein
#74683	267541	267864	324	324	hypothetical protein [Escherichia coli]	1,00E-13	-1	#0240	267079	271293	4215	core protein
#10163	1542047	1542367	321	246	hypothetical protein [Pseudomonas aeruginosa]	4,00E-05	-1	#1601	1542122	1542397	276	Transposase
#10668	1620090	1620410	321	321	hypothetical protein [Neisseria gonorrhoeae]	2,00E-07	-1	#1692	1617384	1620830	3447	Transcription-repair coupling factor
#11771	1785671	1785991	321	321	hypothetical protein [Escherichia coli]	2,00E-64	-1	#1905	1785302	1786945	1644	Putative adhesion and penetration protein
#29735	4316824	4317144	321	306	hypothetical protein [Cronobacter sakazakii]	1,00E-11	-1	#4586	4315843	4317129	1287	DamX, an inner membrane protein involved in bile resistance
#38031	5535423	5535743	321	321	hypothetical protein [Bacteroides intestinalis CAG:564]	6,00E-05	-1	#5735	5534526	5536193	1668	ABC transporter, ATP-binding protein
#51796	3520910	3521230	321	321	hypothetical protein [Escherichia coli]	9,00E-41	-1	#3741	3520160	3521494	1335	ATP-dependent RNA helicase SrmB

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#53007	3332495	3332815	321	321	hypothetical protein [Escherichia coli]	6,00E-04	-1	#3564	3332486	3332878	393	hypothetical protein
#59624	2355572	2355892	321	321	hypothetical protein [Sodalis glossinidius]	6,00E-13	-1	#2538	2354816	2357242	2427	Anaerobic dimethyl sulfoxide reductase chain A
#6343	965643	965963	321	111	hypothetical protein [Shigella dysenteriae]	4,00E-46	-1	#0928	963471	965753	2283	Ferrichrome-iron receptor
#69410	1020559	1020879	321	321	hypothetical protein [Rhizobium mongolense]	0,001	-1	#0984	1020049	1021182	1134	Putrescine transport ATP-binding protein PotG
#10782	1636651	1636968	318	183	hypothetical protein [Xenorhabdus poinarii]	1,00E-09	-1	#1713	1636750	1636932	183	hypothetical protein
#17913	2665802	2666119	318	318	hypothetical protein [Burkholderia cenocepacia]	8,00E-05	-1	#2860	2665775	2667376	1602	Methyl-accepting chemotaxis protein IV (dipeptide chemoreceptor protein)
#35661	5187356	5187673	318	246	hypothetical protein, partial [Staphylococcus aureus]	6,00E-05	-1	#5409	5187428	5189386	1959	Acetyl-coenzyme A synthetase
#37783	5497417	5497734	318	318	hypothetical protein [Salmonella enterica]	3,00E-16	-1	#5696	5495665	5497779	2115	Carbon starvation protein A
#40477	5193509	5193826	318	264	hypothetical protein [Escherichia coli]	2,00E-60	-1	#5414	5193563	5195185	1623	Cytochrome c-type heme lyase subunit nrfE, nitrite reductase complex assembly
#60767	2188671	2188988	318	279	hypothetical protein, partial [Salmonella enterica]	6,00E-31	-1	#2342	2188710	2189462	753	Fumarate and nitrate reduction regulatory protein
#64825	1631424	1631741	318	318	hypothetical protein [Edwardsiella tarda]	2,00E-19	-1	#1704	1631307	1631822	516	hypothetical protein
#2965	450406	450720	315	315	hypothetical protein, partial [Escherichia coli]	2,00E-04	-1	#0430	449755	450729	975	Porphobilinogen synthase
#29721	4315431	4315745	315	306	hypothetical protein, partial [Escherichia coli]	7,00E-58	-1	#4585	4314900	4315736	837	Methyl-directed repair DNA adenine methylase
#32109	4665275	4665589	315	315	hypothetical protein, partial [Staphylococcus aureus]	1,00E-19	-1	#4920	4664390	4666708	2319	Alpha-xylosidase
#47919	4102624	4102938	315	315	hypothetical protein [Escherichia coli]	3,00E-06	-1	#4348	4102144	4103715	1572	D-galactarate dehydratase
#48478	4020935	4021249	315	315	hypothetical protein [Escherichia coli]	5,00E-31	-1	#4271	4020020	4021681	1662	Inner membrane protein YqiK
#50057	3780896	3781210	315	264	hypothetical protein [Escherichia coli]	2,00E-17	-1	#4012	3780446	3781159	714	membrane protein
#51814	3518455	3518769	315	315	thiol:disulfide interchange protein DsbC [Vibrio parahaemolyticus]	5,00E-04	-1	#3739	3517684	3519306	1623	L-aspartate oxidase
#63178	1856070	1856384	315	315	hypothetical protein [Escherichia coli]	1,00E-23	-1	#1980	1855482	1856528	1047	Putative cytoplasmic protein
#65755	1511637	1511951	315	198	hypothetical protein [Escherichia coli]	1,00E-31	-1	#1555	1511655	1511852	198	hypothetical protein
#6794	1031323	1031637	315	264	6-pyruvoyl-tetrahydropterin synthase [Sutterella wadsworthensis CAG:135]	5,00E-08	-1	#0997	1031374	1032102	729	Arginine ABC transporter, ATP-binding protein ArtP
#68758	1111493	1111807	315	198	hypothetical protein [Escherichia coli]	1,00E-31	-1	#1094	1111511	1111708	198	hypothetical protein
#3108	474657	474968	312	312	hypothetical protein [Escherichia coli]	9,00E-12	-1	#0458	474591	477734	3144	Exonuclease SbcC
#5116	778653	778964	312	312	hypothetical protein [Dickeya dianthicola]	6,00E-26	-1	#0741	777453	779117	1665	Asparagine synthetase [glutamine-hydrolyzing]
#53733	3221553	3221864	312	312	hypothetical protein [Escherichia coli]	1,00E-29	-1	#3454	3220737	3221954	1218	Alanine transaminase
#53741	3220764	3221075	312	312	hypothetical protein [Sutterella wadsworthensis CAG:135]	2,00E-05	-1	#3454	3220737	3221954	1218	Alanine transaminase
#65528	1538231	1538542	312	312	hypothetical protein [Burkholderia sacchari]	2,00E-04	-1	#1592	1538225	1539043	819	hypothetical protein
#67301	1300317	1300628	312	312	hypothetical protein [Sutterella wadsworthensis CAG:135]	6,00E-11	-1	#1311	1299846	1301639	1794	Uptake hydrogenase large subunit
#68529	1138084	1138395	312	312	hypothetical protein [Burkholderia sacchari]	2,00E-04	-1	#1133	1138078	1138896	819	hypothetical protein
#75686	123373	123684	312	312	hypothetical protein [Aeromonas hydrophila]	4,00E-10	-1	#0111	123223	123774	552	N-acetylmuramoyl-L-alanine amidase AmpD
#32667	4744303	4744611	309	309	hypothetical protein [Escherichia coli]	2,00E-32	-1	#5011	4743220	4744905	1686	Mediator of hyperadherence YidE
#33739	4899688	4899996	309	309	hypothetical protein [Escherichia coli]	4,00E-70	-1	#5152	4899289	4900104	816	Putative carboxymethylenebutenolidase
#44192	4654003	4654311	309	309	hypothetical protein, partial [Vibrio parahaemolyticus]	8,00E-09	-1	#4913	4653886	4655994	2109	GTP pyrophosphokinase, (p)ppGpp synthetase II
#5006	765964	766272	309	309	hypothetical protein, partial [Escherichia coli]	3,00E-06	-1	#0729	765667	766392	726	Glutamate Aspartate transport ATP-binding protein GltL
#51759	3525979	3526287	309	264	hypothetical protein [Vibrio parahaemolyticus]	9,00E-08	-1	#3747	3525823	3526242	420	Thioredoxin 2
#52537	3403620	3403928	309	309	hypothetical protein, partial [Escherichia coli]	1,00E-24	-1	#3635	3403464	3403934	471	Thiol peroxidase, Bcp-type
#5315	806550	806858	309	309	hypothetical protein [Escherichia coli]	1,00E-47	-1	#0768	805263	806936	1674	Potassium-transporting ATPase A chain
#68226	1180812	1181120	309	276	hypothetical protein, partial [Pseudomonas chlororaphis]	5,00E-08	-1	#1175	1180803	1181087	285	Integration host factor beta subunit
#72400	587386	587694	309	309	hypothetical protein, partial [Escherichia coli]	5,00E-34	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#30726	4461198	4461503	306	183	hypothetical protein [Pectobacterium carotovorum]	1,00E-08	-1	#4731	4461321	4461602	282	Universal stress protein B
#4699	725902	726207	306	234	hypothetical protein [Rhodococcus qingshengii]	2,00E-28	-1	#0684	725329	726135	807	Ribonuclease I precursor
#4850	744843	745148	306	255	hypothetical protein, partial [Escherichia coli]	6,00E-22	-1	#0705	743886	745097	1212	D-alanyl-D-alanine carboxypeptidase
#58923	2451588	2451893	306	237	hypothetical protein [Serratia marcescens]	1,00E-41	-1	#2634	2451594	2451830	237	major outer membrane lipoprotein
#12060	1827141	1827443	303	210	hypothetical protein, partial [Klebsiella pneumoniae]	1,00E-38	-1	#1943	1824675	1827350	2676	Alcohol dehydrogenase
#17916	2666153	2666455	303	303	hypothetical protein, partial [Staphylococcus aureus]	1,00E-05	-1	#2860	2665775	2667376	1602	Methyl-accepting chemotaxis protein IV (dipeptide chemoreceptor protein)
#39605	5326302	5326604	303	303	hypothetical protein [Kushneria aurantia]	1,00E-05	-1	#5543	5326293	5326979	687	L-ribulose-5-phosphate 4-epimerase UlaF (L-ascorbate utilization protein F)
#45663	4435050	4435352	303	147	hypothetical protein [Escherichia coli]	3,00E-10	-1	#4705	4433967	4435196	1230	3-oxoacyl-[ACP] synthase
#75924	92610	92912	303	279	hypothetical protein, partial [Escherichia coli]	1,00E-34	-1	#0083	92634	93638	1005	Fructose repressor FruR, LacI family
#14171	2148275	2148574	300	270	hypothetical protein, partial [Salmonella enterica]	5,00E-06	-1	#2287	2148179	2148544	366	putative Dnase
#25344	3692169	3692468	300	300	hypothetical protein [Klebsiella pneumoniae]	1,00E-07	-1	#3933	3692016	3694715	2700	CRISPR-associated helicase Cas3, protein
#57774	2617560	2617859	300	300	hypothetical protein [Salmonella enterica]	3,00E-14	-1	#2808	2616435	2617955	1521	Ribosomal RNA small subunit methyltransferase F
#72428	583636	583935	300	300	hypothetical protein [Escherichia coli]	1,00E-08	-1	#0565	581353	585738	4386	Large repetitive protein
#2188	332075	332371	297	297	hypothetical protein [Duganella zoogloeoides]	2,00E-05	-1	#0315	331145	333343	2199	Periplasmic aromatic aldehyde oxidoreductase, molybdenum binding subunit YagR
#4182	649288	649584	297	297	hypothetical protein [Azotobacter chroococcum]	1,00E-12	-1	#0614	649018	649884	867	Methylenetetrahydrofolate dehydrogenase (NADP+)
#43551	4758100	4758396	297	141	hypothetical protein [Marinobacter lipolyticus]	7,00E-06	-1	#5027	4758193	4758333	141	LSU ribosomal protein L34p
#46985	4240619	4240915	297	150	hypothetical protein, partial [Shigella flexneri]	3,00E-58	-1	#4489	4239536	4240768	1233	RND efflux system, inner membrane transporter CmeB
#51182	3613578	3613874	297	297	hypothetical protein [Cronobacter sakazakii]	3,00E-08	-1	#3841	3613119	3615263	2145	Ribonucleotide reductase of class Ib (aerobic), alpha subunit
#58990	2440884	2441180	297	294	hypothetical protein [Escherichia coli]	3,00E-36	-1	#2623	2440872	2441177	306	Protein ydhR precursor
#18040	2679499	2679792	294	294	hypothetical protein [Escherichia coli]	1,00E-31	-1	#2873	2679169	2680683	1515	L-arabinose transport ATP-binding protein AraG
#22658	3317315	3317608	294	294	hypothetical protein [Escherichia coli]	6,00E-37	-1	#3552	3317129	3319624	2496	Phosphoenolpyruvate-protein phosphotransferase of PTS system
#25768	3752911	3753204	294	228	hypothetical protein [Escherichia coli]	4,00E-59	-1	#3990	3752944	3753171	228	putative lipoprotein
#27447	3988878	3989171	294	246	hypothetical protein [Enterobacter cloacae]	3,00E-07	-1	#4239	3986904	3989123	2220	putative Fe-S oxidoreductase family 2
#32035	4652571	4652864	294	204	hypothetical protein, partial [Escherichia coli]	4,00E-49	-1	#4910	4652517	4652774	258	hypothetical protein
#40964	5121226	5121519	294	294	hypothetical protein [Bacteroides clarus CAG:160]	3,00E-04	-1	#5346	5119855	5123538	3684	5-methyltetrahydrofolate--homocysteine methyltransferase
#41728	5017628	5017921	294	294	hypothetical protein, partial [Escherichia coli]	1,00E-18	-1	#5269	5016503	5020687	4185	core protein
#44601	4593633	4593926	294	294	hypothetical protein, partial [Escherichia coli]	1,00E-18	-1	#4854	4592508	4596737	4230	core protein
#55790	2919466	2919759	294	294	hypothetical protein, partial [Staphylococcus aureus]	5,00E-04	-1	#3150	2919004	2922126	3123	Multidrug transporter MdtB
#63610	1796377	1796670	294	294	hypothetical protein, partial [Staphylococcus aureus]	3,00E-08	-1	#1915	1796080	1797162	1083	Peptide chain release factor 1
#70920	808911	809204	294	294	hypothetical protein, partial [Escherichia coli]	1,00E-18	-1	#0770	807786	811985	4200	core protein
#72358	593113	593406	294	294	hypothetical protein [Escherichia coli]	9,00E-12	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72603	558683	558976	294	294	hypothetical protein [Cronobacter dublinensis]	5,00E-05	-1	#0545	558056	559918	1863	DNA polymerase III subunits gamma and tau
#75417	166527	166820	294	294	hypothetical protein [Pseudomonas putida]	1,00E-09	-1	#0152	166383	168857	2475	ATP-dependent helicase HrpB
#887	134452	134745	294	294	hypothetical protein [Escherichia coli]	3,00E-44	-1	#0119	133780	135522	1743	Putative exported protein
#15335	2309114	2309404	291	177	hypothetical protein [Escherichia coli]	5,00E-20	-1	#2466	2309228	2309572	345	hypothetical protein
#17463	2604908	2605198	291	120	hypothetical protein [Klebsiella pneumoniae]	1,00E-05	-1	#2798	2604914	2605033	120	Putative inner membrane protein
#19362	2864665	2864955	291	291	phosphomannomutase, partial [Escherichia coli]	1,00E-15	-1	#3102	2863768	2865138	1371	Phosphomannomutase
#21166	3115994	3116284	291	291	hypothetical protein [Escherichia coli]	1,00E-23	-1	#3361	3114833	3116776	1944	Cytochrome c heme lyase subunit CcmF
#45198	4502139	4502429	291	291	hypothetical protein, partial [Salmonella enterica]	4,00E-05	-1	#4774	4502103	4503074	972	LysR family transcriptional regulator YhjC

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#47680	4132836	4133126	291	291	hypothetical protein [Escherichia coli]	1,00E-09	-1	#4387	4132614	4133492	879	Putative protease
#53924	3189261	3189551	291	291	hypothetical protein, partial [Bacillus cereus]	7,00E-04	-1	#3422	3188427	3190079	1653	Polymyxin resistance protein ArnT, undecaprenyl phosphate-alpha-L-Ara4N transferase
#54587	3091734	3092024	291	291	hypothetical protein, partial [Klebsiella pneumoniae]	6,00E-11	-1	#3340	3091077	3092063	987	Putative metal chaperone, involved in Zn homeostasis, GTPase family
#55771	2921383	2921673	291	291	hypothetical protein, partial [Acinetobacter haemolyticus]	3,00E-26	-1	#3150	2919004	2922126	3123	Multidrug transporter MdtB
#5713	867541	867831	291	291	hypothetical protein, partial [Vibrio parahaemolyticus]	3,00E-05	-1	#0825	867523	868458	936	Zinc transporter ZitB
#586	85967	86257	291	291	hypothetical protein, partial [Rhodococcus opacus]	1,00E-33	-1	#0075	85475	86566	1092	3-isopropylmalate dehydrogenase
#6880	1043021	1043311	291	267	hypothetical protein [Escherichia coli]	4,00E-09	-1	#1007	1043045	1044697	1653	Hydroxylamine reductase
#74663	270367	270657	291	291	hypothetical protein, partial [Escherichia coli]	3,00E-30	-1	#0240	267079	271293	4215	core protein
#8392	1271770	1272060	291	174	hypothetical protein, partial [Escherichia coli]	7,00E-08	-1	#1273	1271476	1271943	468	putative endopeptidase
#17753	2645169	2645456	288	288	hypothetical protein [Escherichia coli]	1,00E-32	-1	#2840	2645142	2646875	1734	Aspartyl-tRNA synthetase
#27871	4045806	4046093	288	288	hypothetical protein [Rhodopseudomonas palustris]	2,00E-07	-1	#4293	4044714	4046234	1521	Aerotaxis sensor receptor protein
#28872	4196566	4196853	288	282	hypothetical protein [Escherichia coli]	1,00E-20	-1	#4448	4196572	4197354	783	Transcriptional regulator NanR
#5554	844132	844419	288	288	hypothetical protein, partial [Bacillus cereus]	2,00E-11	-1	#0802	843508	845160	1653	Fumarate hydratase class I
#55869	2908910	2909197	288	288	hypothetical protein [Escherichia coli]	1,00E-06	-1	#3143	2907932	2910739	2808	hypothetical protein
#63381	1829428	1829715	288	288	hypothetical protein [Escherichia coli]	2,00E-15	-1	#1947	1829167	1830843	1677	Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA
#13605	2075098	2075382	285	234	hypothetical protein [Shigella dysenteriae]	1,00E-43	-1	#2219	2075149	2075817	669	putative membrane lipoprotein clustered with tellurite resistance proteins TehA/TehB
#19458	2877283	2877567	285	285	hypothetical protein [Escherichia coli]	5,00E-27	-1	#3114	2877037	2878032	996	UDP-glucose 4-epimerase
#20811	3064714	3064998	285	267	hypothetical protein, partial [Vibrio parahaemolyticus]	4,00E-23	-1	#3314	3063940	3064980	1041	Mgl repressor and galactose ultrainduction factor GalS, HTH-type transcriptional regulator
#25190	3673103	3673387	285	285	hypothetical protein [Escherichia coli]	3,00E-26	-1	#3907	3672014	3673441	1428	Hydroxyaromatic non-oxidative decarboxylase protein C
#39545	5334483	5334767	285	285	hypothetical protein, partial [Providencia rettgeri]	9,00E-06	-1	#5554	5334204	5335616	1413	D-serine/D-alanine/glycine transporter
#59554	2365650	2365934	285	285	hypothetical protein [Escherichia vulneris]	3,00E-06	-1	#2548	2365611	2365946	336	Acid shock protein precursor
#62034	1998509	1998793	285	285	hypothetical protein [Escherichia coli]	1,00E-31	-1	#2145	1996922	1999294	2373	hypothetical protein
#65083	1595333	1595617	285	189	hypothetical protein [Salmonella enterica]	5,00E-34	-1	#1667	1594481	1595521	1041	Phosphate:acyl-ACP acyltransferase PlsX
#14628	2209972	2210253	282	282	hypothetical protein, partial [Vibrio parahaemolyticus]	9,00E-08	-1	#2363	2208430	2210697	2268	Maltose phosphorylase
#17872	2660873	2661154	282	282	hypothetical protein [Escherichia coli]	1,00E-05	-1	#2853	2659364	2661442	2079	Flagellar biosynthesis protein FlhA
#28149	4088577	4088858	282	282	hypothetical protein [Bifidobacterium adolescentis CAG:119]	1,00E-11	-1	#4335	4087380	4089674	2295	2-ketobutyrate formate-lyase
#35918	5221491	5221772	282	126	hypothetical protein, partial [Escherichia coli]	7,00E-31	-1	#5442	5220936	5221616	681	Phosphonates transport ATP-binding protein PhnL
#42236	4941784	4942065	282	234	hypothetical protein [Escherichia coli]	2,00E-59	-1	#5190	4941832	4943655	1824	GTP-binding protein TypA/BipA
#48955	3951294	3951575	282	282	hypothetical protein [Salmonella enterica]	7,00E-07	-1	#4194	3951240	3952229	990	hypothetical protein
#63493	1813603	1813884	282	282	hypothetical protein [Escherichia coli]	1,00E-09	-1	#1931	1810273	1814016	3744	Respiratory nitrate reductase alpha chain
#71105	779071	779352	282	165	hypothetical protein [Escherichia coli]	5,00E-44	-1	#0742	779086	779250	165	putative lipoprotein
#13509	2062553	2062831	279	129	hypothetical protein, partial [Escherichia coli]	5,00E-35	-1	#2204	2062460	2062681	222	hypothetical protein
#23177	3387305	3387583	279	279	hypothetical protein, partial [Rhodococcus opacus]	8,00E-16	-1	#3622	3386561	3388540	1980	Glutamate synthase [NADPH] small chain
#30134	4373256	4373534	279	279	hypothetical protein [Escherichia coli]	4,00E-41	-1	#4637	4372773	4374746	1974	Glycogen debranching enzyme
#34716	5044150	5044428	279	279	hypothetical protein [Escherichia coli]	1,00E-09	-1	#5291	5043040	5044773	1734	UPF0141 membrane protein YijP possibly required for phosphoethanolamine modification of lipopolysaccharide
#36746	5339719	5339997	279	279	hypothetical protein [Escherichia coli]	1,00E-21	-1	#5559	5338951	5340894	1944	2',3'-cyclic-nucleotide 2'-phosphodiesterase
#48194	4059898	4060176	279	279	hypothetical protein [Escherichia coli]	9,00E-11	-1	#4303	4058923	4060941	2019	2,4-dienoyl-CoA reductase [NADPH]
#50339	3736805	3737083	279	279	hypothetical protein [Pseudomonas fuscovaginae]	8,00E-10	-1	#3976	3736334	3737698	1365	Decarboxylase family protein
#51210	3610022	3610300	279	261	hypothetical protein [Escherichia coli]	2,00E-10	-1	#3836	3609953	3610282	330	hypothetical protein

sORF with blastp hit							mother gene					
ID CP00895 7.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	rea-ding frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#52064	3476553	3476831	279	279	hypothetical protein, partial [Bacillus thuringiensis]	3,00E-44	-1	#3701	3476424	3477785	1362	3-phenylpropionate dioxygenase, alpha subunit
#5257	799248	799526	279	267	hypothetical protein [Pectobacterium carotovorum]	3,00E-09	-1	#0764	799260	799937	678	DNA-binding response regulator KdpE
#59669	2350053	2350331	279	261	hypothetical protein [Escherichia coli]	6,00E-32	-1	#2533	2350071	2350412	342	putative secreted protein
#64858	1626548	1626826	279	279	hypothetical protein [Escherichia coli]	2,00E-29	-1	#1698	1626386	1627207	822	NAD-dependent protein deacetylase of SIR2 family
#10006	1517477	1517752	276	174	hypothetical protein [Citrobacter rodentium]	7,00E-35	-1	#1566	1517579	1517770	192	hypothetical protein
#10474	1590039	1590314	276	276	hypothetical protein [Sutterella wadsworthensis CAG:135]	2,00E-07	-1	#1661	1588023	1591208	3186	Ribonuclease E
#1451	224975	225250	276	276	hypothetical protein, partial [Bacillus cereus]	3,00E-19	-1	#0205	224945	225976	1032	Methionine ABC transporter ATP-binding protein
#1747	262346	262621	276	276	hypothetical protein [Escherichia coli]	3,00E-49	-1	#0233	261236	262627	1392	Uncharacterized protein ImpC
#19463	2878178	2878453	276	264	hypothetical protein [Escherichia coli]	7,00E-17	-1	#3115	2878190	2879584	1395	Colanic acid biosynthesis protein wcaM
#28571	4153032	4153307	276	276	hypothetical protein [Nocardiopsis halotolerans]	2,00E-09	-1	#4404	4152858	4154195	1338	Phosphoglucosamine mutase
#2865	437749	438024	276	231	hypothetical protein [Escherichia vulneris]	5,00E-06	-1	#0419	437704	437979	276	FrmR: Negative transcriptional regulator of formaldehyde detoxification operon
#32485	4721636	4721911	276	276	hypothetical protein, partial [Bacillus cereus]	8,00E-04	-1	#4990	4720532	4721923	1392	Hexose phosphate transport protein UhpT
#43280	4808160	4808435	276	213	hypothetical protein, partial [Klebsiella pneumoniae]	5,00E-04	-1	#5074	4808223	4809215	993	Aspartate--ammonia ligase
#56559	2802658	2802933	276	243	hypothetical protein [Escherichia coli]	4,00E-11	-1	#3045	2802691	2805162	2472	Exodeoxyribonuclease VIII
#7384	1117333	1117608	276	174	hypothetical protein [Citrobacter rodentium]	7,00E-35	-1	#1105	1117435	1117626	192	hypothetical protein
#76330	31674	31949	276	276	hypothetical protein [Escherichia coli]	3,00E-07	-1	#0029	31668	32582	915	Inosine-uridine preferring nucleoside hydrolase
#16400	2455232	2455504	273	273	hypothetical protein [Escherichia coli]	2,00E-33	-1	#2638	2454692	2455963	1272	Iron-sulfur cluster assembly protein SufD
#30417	4409352	4409624	273	267	hypothetical protein, partial [Pseudomonas mendocina]	2,00E-05	-1	#4676	4408950	4409618	669	Cell division transporter, ATP-binding protein FtsE
#42973	4847943	4848215	273	273	hypothetical protein, partial [Staphylococcus aureus]	7,00E-16	-1	#5102	4847448	4848707	1260	Transcription termination factor Rho
#50054	3781246	3781518	273	219	hypothetical protein [Cronobacter sakazakii]	1,00E-13	-1	#4013	3781297	3781515	219	putative lipoprotein ygdR precursor
#53114	3315837	3316109	273	273	hypothetical protein, partial [Escherichia coli]	5,00E-19	-1	#3550	3315522	3316256	735	putative response regulatory protein ypdB
#62791	1902443	1902715	273	258	hypothetical protein, partial [Escherichia coli]	2,00E-04	-1	#2040	1902458	1902730	273	hypothetical protein
#67935	1217078	1217350	273	273	hypothetical protein [Escherichia coli]	2,00E-15	-1	#1205	1216727	1218619	1893	type 1 fimbriae anchoring protein FimD
#33793	4908472	4908741	270	270	hypothetical protein [Escherichia coli]	5,00E-22	-1	#5162	4908268	4908756	489	Transcriptional activator RfaH
#34271	4981955	4982224	270	180	hypothetical protein [Klebsiella pneumoniae]	6,00E-18	-1	#5228	4981310	4982134	825	Rhamnulose-1-phosphate aldolase
#3958	612897	613166	270	270	hypothetical protein [Nitrococcus mobilis]	7,00E-04	-1	#0582	612627	613250	624	Arylesterase precursor
#39781	5302897	5303166	270	270	hypothetical protein [Acidovorax temperans]	4,00E-14	-1	#5517	5302879	5303187	309	RNA-binding protein Hfq
#52682	3383761	3384030	270	270	hypothetical protein, partial [Klebsiella pneumoniae]	6,00E-16	-1	#3619	3382651	3384654	2004	Transketolase
#57256	2696426	2696695	270	261	hypothetical protein, partial [Escherichia coli]	4,00E-32	-1	#2901	2693864	2696686	2823	Phage replication protein
#62070	1993733	1994002	270	252	hypothetical protein, partial [Escherichia coli]	8,00E-32	-1	#2142	1993751	1994908	1158	GALNS arylsulfatase regulator (Fe-S oxidoreductase)
#73913	377904	378173	270	270	hypothetical protein, partial [Clostridium botulinum]	9,00E-04	-1	#0363	377127	379160	2034	High-affinity choline uptake protein BetT
#8261	1252165	1252434	270	243	hypothetical protein [Escherichia coli]	3,00E-10	-1	#1241	1249936	1252407	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#9255	1400956	1401225	270	246	hypothetical protein [Escherichia coli]	2,00E-09	-1	#1433	1400110	1401201	1092	putative monooxygenase RutA in pyrimidine catabolism pathway
#11186	1700865	1701131	267	210	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase [Klebsiella pneumoniae]	8,00E-04	-1	#1802	1699968	1701074	1107	tRNA-specific 2-thiouridylase MnmA
#17014	2539696	2539962	267	267	hypothetical protein [Escherichia coli]	1,00E-25	-1	#2723	2538973	2540934	1962	DNA topoisomerase III
#21236	3125139	3125405	267	231	hypothetical protein [Citrobacter amalonaticus]	2,00E-10	-1	#3373	3124875	3125369	495	Ferredoxin-type protein NapF (periplasmic nitrate reductase)
#29869	4335378	4335644	267	267	hypothetical protein [Escherichia coli]	1,00E-55	-1	#4604	4334337	4335689	1353	Osmolarity sensory histidine kinase EnvZ
#33058	4797817	4798083	267	267	aGAP012078-PA [Parabacteroides johnsonii CAG:246]	2,00E-05	-1	#5062	4797061	4798443	1383	ATP synthase beta chain
#33983	4937844	4938110	267	267	hypothetical protein, partial [Acidovorax avenae]	1,00E-08	-1	#5187	4937406	4938824	1419	Nitrogen regulation protein NR(I)

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#46948	4245512	4245778	267	267	hypothetical protein [Streptomyces viridochromogenes]	4,00E-04	-1	#4495	4245068	4245826	759	amino acid ABC transporter, ATP-binding protein
#57738	2622386	2622652	267	240	hypothetical protein [Escherichia coli]	1,00E-35	-1	#2817	2622413	2623075	663	Exodeoxyribonuclease X
#71581	707510	707776	267	267	enterobactin synthase subunit E, partial [Klebsiella pneumoniae]	2,00E-06	-1	#0664	707318	708928	1611	2,3-dihydroxybenzoate-AMP ligase [enterobactin] siderophore
#8654	1311681	1311947	267	267	hypothetical protein [Piscirickettsia salmonis]	5,00E-25	-1	#1321	1311267	1312406	1140	Putative polysaccharide export protein YccZ precursor
#20774	3060548	3060811	264	264	hypothetical protein, partial [Piscirickettsia salmonis]	8,00E-19	-1	#3310	3060056	3061066	1011	Galactose/methyl galactoside ABC transport system, permease protein MglC
#35875	5214993	5215256	264	213	hypothetical protein [Streptomyces flavidovirens]	4,00E-05	-1	#5434	5213685	5215205	1521	Ribose ABC transport system, ATP-binding protein RbsA
#44690	4579767	4580030	264	189	hypothetical protein [Escherichia coli]	9,00E-05	-1	#4841	4577985	4579955	1971	Putative glycosyl hydrolase of unknown function (DUF1680)
#47366	4184563	4184826	264	216	hypothetical protein [Citrobacter werkmanii]	9,00E-09	-1	#4440	4184611	4189164	4554	Glutamate synthase [NADPH] large chain
#28096	4081177	4081437	261	261	hypothetical protein [Escherichia coli]	4,00E-55	-1	#4328	4080574	4081470	897	LysR-family transcriptional regulator YhaJ
#35502	5161765	5162025	261	261	hypothetical protein [Aeromonas salmonicida]	2,00E-07	-1	#5383	5161537	5162052	516	Zinc uptake regulation protein ZUR
#35570	5173091	5173351	261	261	hypothetical protein [Akkermansia muciniphila CAG:154]	2,00E-04	-1	#5396	5173061	5175883	2823	Excinuclease ABC subunit A
#43530	4761051	4761311	261	261	hypothetical protein [Comamonas aquatica]	2,00E-04	-1	#5030	4760685	4762049	1365	GTPase and tRNA-U34 5-formylation enzyme TrmE
#46992	4239347	4239607	261	159	hypothetical protein, partial [Enterobacter cloacae]	3,00E-05	-1	#4488	4237619	4239505	1887	RND efflux system, inner membrane transporter CmeB
#50806	3665708	3665968	261	261	hypothetical protein, partial [Escherichia coli]	2,00E-12	-1	#3901	3665648	3667726	2079	Formate hydrogenlyase transcriptional activator
#5247	798025	798285	261	261	hypothetical protein, partial [Escherichia coli]	1,00E-19	-1	#0762	796372	798579	2208	Ornithine decarboxylase
#56283	2837930	2838190	261	144	hypothetical protein, partial [Klebsiella oxytoca]	1,00E-15	-1	#3074	2838047	2838295	249	YeeU protein (antitoxin to YeeV)
#65509	1540561	1540821	261	261	hypothetical protein, partial [Klebsiella oxytoca]	2,00E-13	-1	#1598	1540558	1540932	375	YeeU protein (antitoxin to YeeV)
#68510	1140414	1140674	261	261	hypothetical protein, partial [Klebsiella oxytoca]	2,00E-13	-1	#1139	1140411	1140785	375	YeeU protein (antitoxin to YeeV)
#70805	821645	821905	261	261	hypothetical protein, partial [Escherichia coli]	2,00E-18	-1	#0779	821588	822322	735	Lactam utilization protein LamB
#72423	584497	584757	261	261	hypothetical protein [Escherichia coli]	9,00E-12	-1	#0565	581353	585738	4386	Large repetitive protein
#72551	565778	566038	261	123	hypothetical protein, partial [Escherichia coli]	6,00E-50	-1	#0553	565916	567220	1305	Inosine-guanosine kinase
#75308	180906	181166	261	249	hypothetical protein, partial [Bacillus cereus]	8,00E-04	-1	#0161	180810	181154	345	putative iron binding protein from the HesB_IscA_SufA family
#13033	1980505	1980762	258	129	hypothetical protein [Klebsiella pneumoniae]	7,00E-04	-1	#2131	1980556	1980684	129	hypothetical protein
#19712	2913277	2913534	258	258	hypothetical protein [Escherichia coli]	3,00E-42	-1	#3146	2913187	2915127	1941	Putative chaperonin
#23928	3494420	3494677	258	246	thiamine ABC transporter permease [Escherichia coli]	8,00E-07	-1	#3718	3494327	3494665	339	Nitrogen regulatory protein P-II
#37772	5495346	5495603	258	192	hypothetical protein [Serratia symbiotica]	3,00E-13	-1	#5695	5495412	5495615	204	putative small protein yjiX
#5749	872109	872366	258	258	hypothetical protein [Escherichia coli]	2,00E-20	-1	#0830	871440	872480	1041	Aldose 1-epimerase
#64882	1623872	1624129	258	258	hypothetical protein, partial [Escherichia coli]	2,00E-14	-1	#1695	1623485	1624186	702	Lipoprotein releasing system ATP-binding protein LolD
#6505	987286	987543	258	228	hypothetical protein, partial [Salmonella enterica]	2,00E-11	-1	#0948	986614	987513	900	Pyruvate formate-lyase activating enzyme
#6941	1053332	1053589	258	105	hypothetical protein, partial [Escherichia coli]	1,00E-10	-1	#1014	1053212	1053436	225	Cold shock protein CspD
#19549	2890520	2890774	255	255	hypothetical protein [Gordonia rhizosphera]	2,00E-04	-1	#3125	2889905	2890870	966	GDP-L-fucose synthetase
#24364	3551309	3551563	255	231	hypothetical protein [Ferrimonas futtsuensis]	2,00E-11	-1	#3767	3551192	3551539	348	LSU ribosomal protein L19p
#36196	5257086	5257340	255	255	hypothetical protein [Coprococcus comes CAG:19]	1,00E-10	-1	#5477	5256516	5258033	1518	Lysyl-tRNA synthetase (class II)
#52092	3472821	3473075	255	249	hypothetical protein [Salmonella enterica]	5,00E-15	-1	#3698	3472827	3474107	1281	Stationary phase inducible protein CsiE
#54170	3153818	3154072	255	255	hypothetical protein [Escherichia coli]	1,00E-09	-1	#3393	3153656	3154378	723	3-demethylubiquinol 3-O-methyltransferase
#70009	935771	936025	255	255	hypothetical protein [Halorubrum distributum]	7,00E-04	-1	#0900	935663	937684	2022	Excinuclease ABC subunit B
#71554	711217	711471	255	255	hypothetical protein [Salmonella enterica]	4,00E-06	-1	#0668	711142	713247	2106	Carbon starvation protein A
#71700	690544	690798	255	153	hypothetical protein [Escherichia coli]	1,00E-13	-1	#0649	690619	690771	153	HokE protein
#72393	588313	588567	255	255	hypothetical protein [Escherichia coli]	8,00E-04	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)

sORF with blastp hit							mother gene					
ID CP00895 7.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	rea-ding frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#19821	2933291	2933542	252	249	hypothetical protein [Escherichia coli]	2,00E-50	-1	#3159	2932805	2933539	735	Galactitol utilization operon repressor
#20793	3062810	3063061	252	252	hypothetical protein, partial [Escherichia coli]	7,00E-13	-1	#3312	3062663	3063661	999	Galactose/methyl galactoside ABC transport system, D-galactose-binding periplasmic protein MglB
#25872	3766913	3767164	252	252	hypothetical protein [Cronobacter malonaticus]	2,00E-04	-1	#3999	3766022	3768910	2889	Protease III precursor
#27598	4008674	4008925	252	252	hypothetical protein [Deinococcus radiodurans]	1,00E-04	-1	#4256	4007225	4009117	1893	Topoisomerase IV subunit B
#29275	4251580	4251831	252	123	hypothetical protein [Salmonella enterica]	1,00E-04	-1	#4496	4251682	4251804	123	hypothetical protein
#51790	3521294	3521545	252	201	LysR family transcriptional regulator [Salmonella enterica]	5,00E-08	-1	#3741	3520160	3521494	1335	ATP-dependent RNA helicase SrmB
#61875	2016792	2017043	252	252	hypothetical protein [Haloferax gibbonsii]	6,00E-05	-1	#2158	2016504	2017430	927	Dipeptide transport ATP-binding protein DppF
#68578	1133059	1133310	252	180	hypothetical protein [Escherichia coli]	5,00E-16	-1	#1124	1133131	1133373	243	Antigen 43 precursor
#72963	507701	507952	252	252	hypothetical protein [Yersinia pestis]	1,00E-08	-1	#0493	507602	509050	1449	tRNA S(4)U 4-thiouridine synthase (former ThiI)
#14227	2154735	2154983	249	168	hypothetical protein [Escherichia coli]	1,00E-13	-1	#2298	2154810	2154977	168	hypothetical protein
#22782	3333635	3333883	249	249	hypothetical protein, partial [Shigella flexneri]	2,00E-36	-1	#3565	3332918	3334066	1149	putative virulence protein
#27488	3993935	3994183	249	96	hypothetical protein [Escherichia coli]	8,00E-28	-1	#4243	3994088	3995695	1608	Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA
#29471	4277328	4277576	249	168	ferredoxin [Citrobacter freundii]	7,00E-15	-1	#4541	4277301	4277495	195	Bacterioferritin-associated ferredoxin
#33735	4899262	4899510	249	222	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase [Escherichia coli]	2,00E-48	-1	#5152	4899289	4900104	816	Putative carboxymethylenebutenolidase
#34796	5054850	5055098	249	204	hypothetical protein, partial [Klebsiella pneumoniae]	3,00E-14	-1	#5298	5053653	5055053	1401	Soluble pyridine nucleotide transhydrogenase
#3492	533203	533451	249	174	hypothetical protein [Shigella dysenteriae]	3,00E-35	-1	#0518	531676	533376	1701	Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA
#38210	5533088	5533336	249	249	hypothetical protein [Escherichia coli]	3,00E-30	-1	#5733	5532986	5534218	1233	NadR transcriptional regulator
#39570	5331319	5331567	249	249	hypothetical protein, partial [Shigella flexneri]	7,00E-37	-1	#5551	5331136	5332284	1149	putative virulence protein
#45626	4439436	4439684	249	249	hypothetical protein, partial [Escherichia coli]	5,00E-10	-1	#4710	4439244	4440008	765	Nickel transport ATP-binding protein NikD
#59664	2350762	2351010	249	246	hypothetical protein [Shigella dysenteriae]	5,00E-34	-1	#2534	2350447	2351007	561	Spermidine N1-acetyltransferase
#62751	1906860	1907108	249	168	hypothetical protein [Escherichia coli]	1,00E-13	-1	#2048	1906866	1907033	168	hypothetical protein
#7341	1110870	1111118	249	114	hypothetical protein [Escherichia coli]	4,00E-23	-1	#1091	1110951	1111064	114	hypothetical protein
#9963	1511014	1511262	249	114	hypothetical protein [Escherichia coli]	4,00E-23	-1	#1552	1511095	1511208	114	hypothetical protein
#27374	3978705	3978950	246	150	hypothetical protein [Escherichia coli]	4,00E-52	-1	#4227	3978801	3979226	426	Biopolymer transport protein ExbD/TolR
#28409	4130985	4131230	246	246	hypothetical protein [Escherichia coli]	4,00E-05	-1	#4384	4130877	4131401	525	Putative lipid carrier protein
#38494	5492617	5492862	246	138	hypothetical protein, partial [Shigella dysenteriae]	1,00E-29	-1	#5692	5492725	5493009	285	hypothetical protein
#45809	4416976	4417221	246	246	hypothetical protein [Klebsiella pneumoniae]	2,00E-18	-1	#4685	4416829	4417386	558	DcrB protein precursor
#56507	2808509	2808754	246	246	hypothetical protein [Serratia symbiotica]	3,00E-05	-1	#3049	2808089	2815951	7863	adherence and invasion outermembrane protein (Inv,enhances Peyer's patches colonization)
#64365	1685907	1686152	246	246	hypothetical protein [Escherichia coli]	1,00E-29	-1	#1781	1684038	1686170	2133	DNA primase, phage-associated
#25101	3658444	3658686	243	243	hypothetical protein [Cronobacter sakazakii]	4,00E-05	-1	#3891	3657799	3658722	924	Formate hydrogenlyase subunit 4
#26351	3830658	3830900	243	243	hypothetical protein, partial [Escherichia coli]	7,00E-19	-1	#4070	3829965	3831743	1779	putative sigma-54-dependent transcriptional regulator YgeV
#27091	3937425	3937667	243	243	hypothetical protein [Salmonella enterica]	2,00E-04	-1	#4179	3936240	3938375	2136	Ornithine decarboxylase
#30203	4382614	4382856	243	243	sulfur acceptor protein CsdL, partial [Escherichia coli]	2,00E-05	-1	#4646	4382602	4383942	1341	Low-affinity gluconate/H+ symporter GntU
#36962	5370991	5371233	243	243	hypothetical protein, partial [Escherichia coli]	2,00E-37	-1	#5591	5370703	5371650	948	Trehalose operon transcriptional repressor
#44172	4656861	4657103	243	243	hypothetical protein [Escherichia coli]	5,00E-25	-1	#4915	4656696	4658777	2082	ATP-dependent DNA helicase RecG
#56504	2808911	2809153	243	243	hypothetical protein [Salmonella enterica]	3,00E-05	-1	#3049	2808089	2815951	7863	adherence and invasion outermembrane protein (Inv,enhances Peyer's patches colonization)
#66748	1373871	1374113	243	114	outer membrane protein [Escherichia coli]	3,00E-50	-1	#1408	1373796	1373984	189	hypothetical protein
#72291	604317	604559	243	243	hypothetical protein, partial [Staphylococcus aureus]	3,00E-10	-1	#0568	603675	604850	1176	Putative membrane spanning export protein
#73883	381408	381650	243	243	hypothetical protein [Escherichia coli]	6,00E-26	-1	#0364	379668	383717	4050	AidA-I adhesin-like protein

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#8389	1271488	1271730	243	243	hypothetical protein, partial [Cronobacter universalis]	2,00E-10	-1	#1273	1271476	1271943	468	putative endopeptidase
#15515	2329676	2329915	240	240	hypothetical protein, partial [Xanthomonas pisi]	2,00E-30	-1	#2500	2329580	2330149	570	hypothetical protein
#16674	2493711	2493950	240	213	hypothetical protein, partial [Pseudomonas protegens]	1,00E-15	-1	#2675	2493567	2493923	357	LSU ribosomal protein L20p
#21291	3131424	3131663	240	240	inverted ada-Golga3 fusion protein [Escherichia coli]	7,00E-42	-1	#3379	3130647	3131711	1065	ADA regulatory protein
#25303	3687101	3687340	240	240	hypothetical protein [Shigella dysenteriae]	7,00E-10	-1	#3927	3686447	3687370	924	CRISPR-associated protein Cas1
#38774	5445808	5446047	240	219	hypothetical protein [Klebsiella oxytoca]	1,00E-04	-1	#5648	5445829	5446425	597	type 1 fimbriae regulatory protein FimE
#42740	4876861	4877100	240	234	hypothetical protein, partial [Rhodococcus opacus]	1,00E-42	-1	#5129	4876867	4877574	708	Protein of unknown function DUF484
#42940	4851737	4851976	240	240	hypothetical protein [Pseudomonas aeruginosa]	6,00E-05	-1	#5106	4851164	4852294	1131	UDP-N-acetylglucosamine 2-epimerase
#45268	4492291	4492530	240	240	hypothetical protein, partial [Staphylococcus aureus]	5,00E-04	-1	#4765	4490173	4493286	3114	RND efflux system, inner membrane transporter CmeB
#51138	3619066	3619305	240	240	hypothetical protein, partial [Escherichia coli]	2,00E-10	-1	#3845	3618904	3619896	993	L-proline glycine betaine binding ABC transporter protein ProX
#52637	3390072	3390311	240	240	hypothetical protein, partial [Staphylococcus aureus]	1,00E-38	-1	#3623	3388746	3390446	1701	Nitrate/nitrite sensor protein
#62712	1911290	1911529	240	240	hypothetical protein, partial [Xanthomonas pisi]	2,00E-30	-1	#2055	1911056	1911625	570	hypothetical protein
#63108	1863540	1863779	240	240	hypothetical protein, partial [Xanthomonas pisi]	2,00E-30	-1	#1989	1863306	1863875	570	hypothetical protein
#64680	1647151	1647390	240	240	hypothetical protein, partial [Xanthomonas pisi]	2,00E-30	-1	#1732	1646917	1647486	570	hypothetical protein
#66872	1360571	1360810	240	240	hypothetical protein, partial [Xanthomonas pisi]	2,00E-30	-1	#1391	1360337	1360906	570	hypothetical protein
#74513	291671	291910	240	240	hypothetical protein [Pseudomonas aeruginosa]	4,00E-06	-1	#0269	291590	292090	501	Peptide chain release factor-like protein
#74873	235549	235788	240	240	hypothetical protein [Escherichia coli]	3,00E-40	-1	#0210	235330	236100	771	SAM-dependent methyltransferase YafE (UbiE-like protein)
#15578	2336608	2336844	237	237	hypothetical protein [Escherichia coli]	2,00E-19	-1	#2511	2336464	2337513	1050	Putative cytoplasmic protein
#18296	2711173	2711409	237	237	hypothetical protein [Cronobacter sakazakii]	1,00E-17	-1	#2918	2709607	2711439	1833	Excinuclease ABC subunit C
#18823	2785873	2786109	237	174	hypothetical protein, partial [Escherichia coli]	3,00E-06	-1	#3014	2785579	2786046	468	putative endopeptidase
#21472	3154501	3154737	237	219	hypothetical protein, partial [Escherichia coli]	4,00E-29	-1	#3394	3154519	3158223	3705	Type V secretory pathway, adhesin AidA
#23154	3384647	3384883	237	135	hypothetical protein, partial [Escherichia coli]	3,00E-05	-1	#3620	3384749	3385792	1044	Putative exported protein
#24870	3626629	3626865	237	204	hypothetical protein, partial [Actinomyces urogenitalis]	5,00E-32	-1	#3855	3626662	3628218	1557	Glutamate--cysteine ligase
#27036	3930370	3930606	237	237	hypothetical protein [Escherichia coli]	2,00E-35	-1	#4170	3930124	3930843	720	Uncharacterized protein YggN
#30694	4456216	4456452	237	237	hypothetical protein, partial [Escherichia coli]	7,00E-30	-1	#4726	4454953	4456626	1674	hypothetical protein
#48351	4039267	4039503	237	237	hypothetical protein, partial [Escherichia coli]	9,00E-18	-1	#4287	4038265	4040010	1746	DNA primase
#530	78904	79140	237	237	hypothetical protein [Escherichia coli]	6,00E-44	-1	#0069	77731	79341	1611	Thiamin ABC transporter, transmembrane component
#61729	2036812	2037048	237	237	hypothetical protein, partial [Escherichia coli]	5,00E-14	-1	#2174	2033959	2037699	3741	Respiratory nitrate reductase alpha chain
#62780	1903401	1903637	237	237	hypothetical protein [Escherichia coli]	2,00E-19	-1	#2041	1902732	1903781	1050	Putative cytoplasmic protein
#64744	1640184	1640420	237	237	hypothetical protein [Escherichia coli]	1,00E-17	-1	#1720	1639515	1640564	1050	Putative cytoplasmic protein
#67588	1263292	1263528	237	237	hypothetical protein [Escherichia coli]	2,00E-19	-1	#1260	1262623	1263672	1050	Putative cytoplasmic protein
#71365	738332	738568	237	144	cold-shock protein [Erwinia tracheiphila]	1,00E-25	-1	#0696	738425	738634	210	Cold shock protein CspA
#8739	1321713	1321949	237	237	hypothetical protein [Escherichia coli]	1,00E-32	-1	#1331	1319259	1321973	2715	Sensor protein torS
#962	146654	146890	237	237	hypothetical protein [Hyphomonas jannaschiana]	2,00E-05	-1	#0129	146381	147043	663	Carbonic anhydrase
#23245	3398300	3398533	234	234	hypothetical protein, partial [Escherichia coli]	1,00E-11	-1	#3629	3396830	3398845	2016	putative P-loop ATPase fused to an acetyltransferase COG1444
#3582	547322	547555	234	234	hypothetical protein, partial [Shigella flexneri]	2,00E-07	-1	#0537	547319	550468	3150	RND efflux system, inner membrane transporter CmeB
#45144	4509601	4509834	234	234	hypothetical protein, partial [Escherichia coli]	3,00E-44	-1	#4779	4509181	4510110	930	2-dehydro-3-deoxygluconate kinase
#50828	3663611	3663844	234	234	hypothetical protein [Cronobacter sakazakii]	2,00E-19	-1	#3899	3663446	3664567	1122	[NiFe] hydrogenase metallocenter assembly protein HypD
#52292	3437268	3437501	234	234	hypothetical protein [Siccibacter turicensis]	2,00E-25	-1	#3666	3437265	3438635	1371	Exodeoxyribonuclease VII large subunit

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#59720	2344277	2344510	234	222	hypothetical protein [Curvibacter lanceolatus]	1,00E-06	-1	#2523	2344142	2344498	357	hypothetical protein
#69144	1056443	1056676	234	99	hypothetical protein [Shigella flexneri]	9,00E-19	-1	#1017	1056407	1056541	135	hypothetical protein
#74002	364677	364910	234	234	hypothetical protein [Escherichia coli]	6,00E-18	-1	#0347	364323	365042	720	putative L-lactate dehydrogenase, Fe-S oxidoreductase subunit YkgE
#75112	206956	207189	234	234	hypothetical protein, partial [Anaerococcus lactolyticus]	6,00E-08	-1	#0187	206692	207840	1149	Lipid-A-disaccharide synthase
#16072	2409901	2410131	231	174	alpha/beta hydrolase, partial [Vibrio parahaemolyticus]	4,00E-07	-1	#2592	2409214	2410074	861	Pyridoxal kinase
#25628	3732660	3732890	231	231	hypothetical protein [Cronobacter sakazakii]	2,00E-09	-1	#3971	3732561	3733010	450	putative flavoprotein YqcA (clustered with tRNA pseudouridine synthase C)
#26939	3915284	3915514	231	117	hypothetical protein [Shigella dysenteriae]	5,00E-43	-1	#4150	3915257	3915400	144	hypothetical protein
#32068	4659368	4659598	231	231	hypothetical protein [Escherichia coli]	4,00E-48	-1	#4916	4658762	4659628	867	putative cytoplasmic protein
#44195	4653586	4653816	231	225	hypothetical protein [Aeromonas bivalvium]	3,00E-09	-1	#4912	4653592	4653867	276	DNA-directed RNA polymerase omega subunit
#57790	2616154	2616384	231	231	hypothetical protein [Yersinia pestis]	6,00E-06	-1	#2807	2613796	2616435	2640	Paraquat-inducible protein B
#67549	1268140	1268370	231	231	hypothetical protein [Escherichia coli]	3,00E-22	-1	#1269	1267489	1269339	1851	Hypothetical protein
#76293	37592	37822	231	231	hypothetical protein [Ketogulonicigenium vulgare]	4,00E-04	-1	#0033	35192	38413	3222	Carbamoyl-phosphate synthase large chain
#10334	1567688	1567915	228	147	hypothetical protein, partial [Vibrio parahaemolyticus]	2,00E-12	-1	#1636	1567577	1567834	258	Putative cytoplasmic protein
#10405	1576960	1577187	228	105	hypothetical protein [Escherichia coli]	1,00E-15	-1	#1648	1577083	1577742	660	Flagellar basal-body P-ring formation protein FlgA
#21309	3134336	3134563	228	117	hypothetical protein [Klebsiella pneumoniae]	1,00E-12	-1	#3382	3134447	3134572	126	hypothetical protein
#27859	4045050	4045277	228	228	hypothetical protein [Achromobacter xylosoxidans]	6,00E-07	-1	#4293	4044714	4046234	1521	Aerotaxis sensor receptor protein
#32910	4779874	4780101	228	228	hypothetical protein, partial [Escherichia coli]	8,00E-39	-1	#5045	4778275	4780536	2262	hypothetical protein
#36655	5326988	5327215	228	108	hypothetical protein [Escherichia coli]	2,00E-08	-1	#5544	5327108	5327383	276	UPF0379 protein yjfY precursor
#36942	5369124	5369351	228	189	histidine kinase, partial [Escherichia coli]	1,00E-04	-1	#5590	5369163	5370584	1422	PTS system, trehalose-specific IIB component
#44341	4630802	4631029	228	228	hypothetical protein [Escherichia coli]	5,00E-35	-1	#4885	4630436	4631428	993	Lipopolysaccharide heptosyltransferase I
#45496	4460618	4460845	228	228	hypothetical protein, partial [Escherichia coli]	6,00E-16	-1	#4730	4459751	4461250	1500	Low-affinity inorganic phosphate transporter
#53278	3291756	3291983	228	228	hypothetical protein, partial [Escherichia coli]	3,00E-10	-1	#3531	3291537	3292862	1326	Sucrose-6-phosphate hydrolase
#58774	2474585	2474812	228	228	hypothetical protein, partial [Rhodococcus qingshengii]	6,00E-35	-1	#2656	2474576	2475514	939	Electron transfer flavoprotein, alpha subunit
#6605	1004964	1005191	228	228	hypothetical protein [Shigella dysenteriae]	8,00E-41	-1	#0965	1004913	1005539	627	putative glutathione S-transferase-like protein
#66680	1381233	1381460	228	228	hypothetical protein [Xanthomonas pisi]	1,00E-35	-1	#1420	1381212	1382477	1266	hypothetical protein
#74240	329185	329412	228	228	hypothetical protein, partial [Escherichia coli]	4,00E-08	-1	#0313	328873	330000	1128	Putative transcriptional regulator
#8749	1323339	1323566	228	228	hypothetical protein [Serratia liquefaciens]	2,00E-08	-1	#1333	1323087	1323779	693	TorCAD operon transcriptional regulatory protein TorR
#13417	2049084	2049308	225	225	hypothetical protein [Escherichia coli]	2,00E-43	-1	#2189	2045670	2049872	4203	core protein
#17961	2670896	2671120	225	225	hypothetical protein [Citrobacter freundii]	1,00E-04	-1	#2863	2669750	2671708	1959	Signal transduction histidine kinase CheA
#29875	4336256	4336480	225	150	transcription accessory protein (S1 RNA-binding domain) [Cronobacter malonaticus]	2,00E-16	-1	#4605	4335686	4336405	720	Two-component system response regulator OmpR
#34217	4974480	4974704	225	225	hypothetical protein [Rhodococcus qingshengii]	2,00E-41	-1	#5222	4974369	4976117	1749	Putative frv operon regulatory protein
#4348	672394	672618	225	225	hypothetical protein [Escherichia coli]	3,00E-30	-1	#0630	667819	672756	4938	Rhs-family protein
#49162	3920961	3921185	225	183	hypothetical protein, partial [Escherichia coli]	4,00E-10	-1	#4158	3921003	3921950	948	Glutathione synthetase
#5003	765670	765894	225	225	6-pyruvoyl-tetrahydropterin synthase [Sutterella wadsworthensis CAG:135]	5,00E-05	-1	#0729	765667	766392	726	Glutamate Aspartate transport ATP-binding protein GltL
#53980	3182220	3182444	225	225	hypothetical protein [Stenotrophomonas maltophilia]	4,00E-10	-1	#3416	3182100	3182498	399	Pyrimidine deoxynucleoside triphosphate (dYTP) pyrophosphohydrolase YfoO
#5474	830522	830746	225	225	hypothetical protein, partial [Bacillus cereus]	5,00E-04	-1	#0787	829607	830890	1284	Citrate synthase (si)
#59071	2429106	2429330	225	225	hypothetical protein [Rhodococcus qingshengii]	7,00E-29	-1	#2612	2428623	2429438	816	Putative lipoprotein
#6210	944385	944609	225	225	hypothetical protein [Enterococcus faecium]	2,00E-40	-1	#0909	943680	944636	957	Inner membrane protein YbhQ
#6675	1013864	1014088	225	225	hypothetical protein [Escherichia coli]	5,00E-04	-1	#0974	1013192	1014877	1686	TrkA, Potassium channel-family protein

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#70847	816959	817183	225	225	hypothetical protein [Kibdelosporangium aridum]	5,00E-10	-1	#0774	815930	817348	1419	Deoxyribodipyrimidine photolyase
#72195	617312	617536	225	225	hypothetical protein [Escherichia coli]	2,00E-43	-1	#0585	616748	620944	4197	core protein
#74202	335568	335792	225	225	hypothetical protein [Klebsiella pneumoniae]	1,00E-04	-1	#0318	335400	336014	615	DUF1440 domain-containing membrane protein
#19526	2887886	2888107	222	222	hypothetical protein, partial [Vibrio parahaemolyticus]	4,00E-17	-1	#3122	2886764	2888200	1437	Mannose-1-phosphate guanylyltransferase (GDP)
#22698	3322215	3322436	222	222	hypothetical protein [Escherichia coli]	9,00E-21	-1	#3555	3321786	3323033	1248	PTS system, fructose-specific IIBC component
#31964	4642201	4642422	222	222	hypothetical protein, partial [Escherichia coli]	2,00E-35	-1	#4896	4641826	4642635	810	Formamidopyrimidine-DNA glycosylase
#34292	4984486	4984707	222	222	hypothetical protein, partial [Escherichia coli]	4,00E-04	-1	#5230	4983565	4985034	1470	Rhamnulokinase
#3919	607769	607990	222	102	hypothetical protein [Shigella flexneri]	6,00E-06	-1	#0576	607889	608347	459	Putative activity regulator of membrane protease YbbK
#52524	3405098	3405319	222	222	hypothetical protein [Escherichia coli]	5,00E-44	-1	#3637	3404804	3406822	2019	Hydrogenase-4 component B
#54498	3104310	3104531	222	222	diguanylate cylase [Escherichia coli]	1,00E-04	-1	#3351	3103881	3105641	1761	ATP-dependent RNA helicase YejH
#58721	2481917	2482138	222	222	hypothetical protein, partial [Escherichia coli]	1,00E-29	-1	#2662	2481620	2482453	834	hypothetical protein
#20900	3077666	3077884	219	219	hypothetical protein, partial [Escherichia coli]	5,00E-17	-1	#3326	3077492	3078433	942	Inosine-uridine preferring nucleoside hydrolase
#40559	5179514	5179732	219	219	hypothetical protein [Advenella kashmirensis]	5,00E-07	-1	#5401	5179481	5179945	465	Redox-sensitive transcriptional activator SoxR
#47737	4125558	4125776	219	219	hypothetical protein [Cronobacter malonaticus]	4,00E-06	-1	#4375	4125519	4125914	396	putative endonuclease distantly related to archaeal Holliday junction resolvase
#51624	3545059	3545277	219	219	hypothetical protein [Escherichia coli]	1,00E-41	-1	#3760	3544753	3545913	1161	Chorismate mutase I
#5243	797485	797703	219	219	hypothetical protein [Salmonella enterica]	5,00E-10	-1	#0762	796372	798579	2208	Ornithine decarboxylase
#5261	799731	799949	219	207	hypothetical protein [Variovorax paradoxus]	7,00E-04	-1	#0764	799260	799937	678	DNA-binding response regulator KdpE
#6471	983198	983416	219	201	hypothetical protein [Salmonella enterica]	1,00E-10	-1	#0945	983216	984031	816	Hydrolase (HAD superfamily)
#69078	1065315	1065533	219	219	hypothetical protein [Pantoea stewartii]	3,00E-07	-1	#1024	1065087	1066880	1794	Type III restriction enzyme, res subunit:DEAD/DEAH box helicase, N- terminal
#71034	789750	789968	219	168	hypothetical protein [Citrobacter freundii]	4,00E-06	-1	#0754	789591	789917	327	putative lipoprotein ybfN precursor
#15074	2269304	2269519	216	216	hypothetical protein [Escherichia coli]	4,00E-34	-1	#2424	2269046	2270629	1584	hypothetical protein
#17659	2632494	2632709	216	114	hypothetical protein [Escherichia coli]	2,00E-42	-1	#2825	2631132	2632607	1476	Glucose-6-phosphate 1-dehydrogenase
#27571	4004467	4004682	216	180	hypothetical protein, partial [Escherichia coli]	3,00E-38	-1	#4254	4003837	4004646	810	transport
#28228	4099634	4099849	216	201	hypothetical protein [Rhodococcus qingshengii]	5,00E-44	-1	#4345	4099649	4100419	771	2-dehydro-3-deoxyglucarate aldolase
#2855	436126	436341	216	216	hypothetical protein [Aeromonas bivalvium]	1,00E-06	-1	#0417	435634	436467	834	S-formylglutathione hydrolase
#44484	4609205	4609420	216	216	hypothetical protein [Escherichia coli]	5,00E-08	-1	#4866	4606733	4611499	4767	hypothetical protein
#64707	1644174	1644389	216	216	hypothetical protein [Escherichia coli]	2,00E-04	-1	#1726	1642848	1644785	1938	Hypothetical protein
#65024	1602392	1602607	216	204	hypothetical protein [Cronobacter malonaticus]	1,00E-08	-1	#1675	1601954	1602595	642	Thymidylate kinase
#69293	1035635	1035850	216	165	hypothetical protein [Serratia odorifera]	2,00E-04	-1	#1000	1035476	1035799	324	hypothetical protein
#75395	168693	168908	216	165	hypothetical protein, partial [Catenibacterium mitsuokai]	8,00E-05	-1	#0152	166383	168857	2475	ATP-dependent helicase HrpB
#21933	3219635	3219847	213	183	hypothetical protein, partial [Salmonella enterica]	5,00E-09	-1	#3453	3218879	3219817	939	LysR family transcriptional regulator lrhA
#26101	3793954	3794166	213	198	hypothetical protein [Escherichia coli]	3,00E-41	-1	#4024	3793315	3794151	837	4-deoxy-L-threo-5-hexosulose-uronate ketol-isomerase
#38757	5447527	5447739	213	195	hypothetical protein [Klebsiella oxytoca]	8,00E-18	-1	#5650	5447545	5448042	498	type 1 fimbriae protein FimI, unknown function
#39434	5352309	5352521	213	213	hypothetical protein [Escherichia coli]	2,00E-26	-1	#5570	5352294	5352581	288	UPF0131 protein YtfP
#51546	3556773	3556985	213	186	hypothetical protein, partial [Salmonella enterica]	2,00E-10	-1	#3773	3555696	3556958	1263	Hemolysin with CBS domain
#56195	2850613	2850825	213	213	hypothetical protein [Escherichia coli]	1,00E-09	-1	#3091	2850565	2851464	900	ATP phosphoribosyltransferase
#59606	2357851	2358063	213	192	hypothetical protein [Shigella flexneri]	1,00E-44	-1	#2540	2357872	2358726	855	Anaerobic dimethyl sulfoxide reductase chain C
#72748	539470	539682	213	213	hypothetical protein, partial [Catenibacterium mitsuokai]	2,00E-10	-1	#0526	539029	540315	1287	Ammonium transporter
#72850	526863	527075	213	213	hypothetical protein [Escherichia coli]	1,00E-27	-1	#0511	524940	527294	2355	ATP-dependent protease La Type I

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#37690	5483284	5483493	210	210	hypothetical protein [Escherichia coli]	4,00E-15	-1	#5684	5482174	5483586	1413	Transcriptional regulator, GntR family
#574	84633	84842	210	210	TetR family transcriptional regulator [Escherichia coli]	2,00E-05	-1	#0074	84072	85472	1401	3-isopropylmalate dehydratase large subunit
#70065	928109	928318	210	135	hypothetical protein [Escherichia coli]	5,00E-14	-1	#0893	928169	928303	135	hypothetical protein
#73111	486401	486610	210	150	hypothetical protein, partial [Escherichia coli]	4,00E-35	-1	#0467	486383	486550	168	hypothetical protein
#14603	2207086	2207292	207	207	phenol hydroxylase [Halococcus hamelinensis]	2,00E-06	-1	#2361	2206678	2207760	1083	Multiple sugar ABC transporter, ATP-binding protein
#31939	4638331	4638537	207	207	hypothetical protein [Escherichia coli]	9,00E-34	-1	#4892	4637431	4638555	1125	UDP-glucose:(heptosyl) LPS alpha1,3-glucosyltransferase WaaG
#42768	4873820	4874026	207	207	hypothetical protein [Escherichia coli]	3,00E-40	-1	#5124	4872437	4874983	2547	Adenylate cyclase
#51774	3523854	3524060	207	207	hypothetical protein [Escherichia coli]	2,00E-34	-1	#3745	3523842	3524531	690	Uracil-DNA glycosylase, family 1
#52049	3478073	3478279	207	207	hypothetical protein, partial [Bacillus cereus]	2,00E-31	-1	#3702	3477782	3478300	519	3-phenylpropionate dioxygenase, beta subunit
#57079	2725264	2725470	207	207	hypothetical protein [Escherichia coli]	1,00E-39	-1	#2937	2725261	2726466	1206	Putative transport system permease protein
#69979	939855	940061	207	207	hypothetical protein, partial [Clostridium botulinum]	8,00E-31	-1	#0902	939180	940169	990	Molybdenum cofactor biosynthesis protein MoaA
#70287	900839	901045	207	117	hypothetical protein [Escherichia coli]	3,00E-28	-1	#0862	900797	900955	159	hypothetical protein
#25179	3671833	3672036	204	171	hypothetical protein [Klebsiella pneumoniae]	9,00E-24	-1	#3906	3671767	3672003	237	Hydroxyaromatic non-oxidative decarboxylase protein D
#30123	4372288	4372491	204	204	hypothetical protein [Acinetobacter baumannii]	5,00E-15	-1	#4636	4371460	4372755	1296	Glucose-1-phosphate adenylyltransferase
#32298	4691364	4691567	204	204	hypothetical protein [Escherichia coli]	1,00E-39	-1	#4954	4691112	4691618	507	hypothetical protein
#38538	5484657	5484860	204	204	hypothetical protein, partial [Escherichia coli]	7,00E-25	-1	#5686	5484024	5484905	882	hypothetical protein
#40665	5166235	5166438	204	105	hypothetical protein [Escherichia coli]	5,00E-15	-1	#5388	5166334	5167749	1416	Replicative DNA helicase
#48935	3953699	3953902	204	204	hypothetical protein [Pectobacterium wasabiae]	3,00E-06	-1	#4196	3953330	3954175	846	DDE endonuclease
#57956	2592683	2592886	204	204	hypothetical protein [Escherichia coli]	3,00E-05	-1	#2783	2592461	2593039	579	putative nudix hydrolase YeaB
#58981	2442215	2442418	204	204	hypothetical protein [Escherichia coli]	1,00E-36	-1	#2624	2441303	2442907	1605	Acyl-CoA dehydrogenases
#59507	2373385	2373588	204	204	hypothetical protein, partial [Serratia marcescens]	2,00E-11	-1	#2556	2372689	2373633	945	Protein ydgH precursor
#63224	1850910	1851113	204	204	ATP-dependent transporter SufC domain protein, partial [Escherichia coli]	2,00E-08	-1	#1973	1850283	1851323	1041	Primosomal protein I
#10211	1549415	1549615	201	195	hypothetical protein [Escherichia coli]	2,00E-40	-1	#1614	1549421	1550071	651	Transcriptional regulator CsgD for 2nd curli operon
#14392	2178094	2178294	201	201	hypothetical protein, partial [Corynebacterium striatum]	3,00E-34	-1	#2331	2177890	2178873	984	Zinc transport protein ZntB
#16396	2454519	2454719	201	177	hypothetical protein [Escherichia coli]	8,00E-08	-1	#2637	2453475	2454695	1221	Cysteine desulfurase, SufS subfamily
#23090	3375621	3375821	201	201	hypothetical protein, partial [Klebsiella pneumoniae]	4,00E-06	-1	#3612	3375576	3376592	1017	Phosphate acetyltransferase, ethanolamine utilization-specific
#24372	3552211	3552411	201	138	hypothetical protein, partial [Salmonella enterica]	6,00E-19	-1	#3768	3551581	3552348	768	tRNA (Guanine37-N1) - methyltransferase
#31842	4623594	4623794	201	201	hypothetical protein, partial [Escherichia coli]	2,00E-10	-1	#4879	4623522	4624538	1017	Beta-1,3-galactosyltransferase
#40493	5191515	5191715	201	201	hypothetical protein, partial [Escherichia coli]	8,00E-19	-1	#5411	5191260	5191826	567	Cytochrome c-type protein NrfB precursor
#47679	4133127	4133327	201	201	hypothetical protein [Escherichia coli]	1,00E-14	-1	#4387	4132614	4133492	879	Putative protease
#50762	3670404	3670604	201	201	hypothetical protein, partial [Bacteroides sartorii]	9,00E-06	-1	#3904	3668403	3670964	2562	DNA mismatch repair protein MutS
#56323	2833908	2834108	201	201	hypothetical protein, partial [Escherichia coli]	3,00E-39	-1	#3069	2833818	2835965	2148	Colicin I receptor precursor
#65053	1598925	1599125	201	201	hypothetical protein, partial [Bacillus cereus]	5,00E-29	-1	#1672	1598805	1600010	1206	3-oxoacyl-[acyl-carrier-protein] synthase, KASII
#72384	589522	589722	201	201	hypothetical protein [Escherichia coli]	1,00E-06	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72396	588022	588222	201	201	hypothetical protein [Escherichia coli]	5,00E-09	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#72410	586228	586428	201	201	hypothetical protein [Escherichia coli]	5,00E-09	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#73996	365344	365544	201	201	hypothetical protein [Rhodococcus qingshengii]	9,00E-37	-1	#0348	365053	366480	1428	putative L-lactate dehydrogenase, Iron-sulfur cluster-binding subunit YkgF
#1258	191421	191618	198	198	hypothetical protein [Escherichia coli]	3,00E-32	-1	#0170	189465	192137	2673	[Protein-PII] uridylyltransferase
#24103	3517010	3517207	198	198	hypothetical protein [Azotobacter chroococcum]	3,00E-06	-1	#3738	3516701	3517276	576	RNA polymerase sigma factor RpoE

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#29873	4336037	4336234	198	198	hypothetical protein [Acidovorax avenae]	2,00E-06	-1	#4605	4335686	4336405	720	Two-component system response regulator OmpR
#41004	5114641	5114838	198	198	hypothetical protein, partial [Escherichia coli]	3,00E-15	-1	#5343	5114623	5116359	1737	Isocitrate dehydrogenase phosphatase /kinase
#43550	4758386	4758583	198	198	hypothetical protein [Photobacterium halotolerans]	3,00E-09	-1	#5028	4758383	4758709	327	Ribonuclease P protein component
#46847	4259121	4259318	198	162	hypothetical protein [Citrobacter amalonaticus]	4,00E-10	-1	#4507	4259157	4260533	1377	Trk system potassium uptake protein TrkA
#71635	699206	699403	198	198	hypothetical protein, partial [Escherichia coli]	1,00E-22	-1	#0656	695591	699472	3882	Enterobactin synthetase component F, serine activating enzyme
#73152	480821	481018	198	198	hypothetical protein [Escherichia coli]	4,00E-07	-1	#0462	479870	481165	1296	Phosphate regulon sensor protein PhoR (SphS)
#36925	5367536	5367730	195	195	hypothetical protein [Escherichia coli]	4,00E-06	-1	#5589	5367458	5369113	1656	Trehalose-6-phosphate hydrolase
#41730	5017418	5017612	195	195	hypothetical protein, partial [Escherichia coli]	2,00E-07	-1	#5269	5016503	5020687	4185	core protein
#44603	4593423	4593617	195	195	hypothetical protein, partial [Escherichia coli]	2,00E-07	-1	#4854	4592508	4596737	4230	core protein
#46144	4362723	4362917	195	195	hypothetical protein, partial [Citrobacter amalonaticus]	3,00E-04	-1	#4629	4362588	4364093	1506	Aerobic glycerol-3-phosphate dehydrogenase
#48841	3966839	3967033	195	195	hypothetical protein [Delftia tsuruhatensis]	7,00E-12	-1	#4211	3966707	3967573	867	putative GST-like protein yghU associated with glutathionylspermidine synthetase/amidase
#49134	3924458	3924652	195	195	hypothetical protein [Vibrio parahaemolyticus]	8,00E-07	-1	#4163	3924230	3924934	705	Hypothetical protein YggS, proline synthase co-transcribed bacterial PROSC-like protein
#70922	808701	808895	195	195	hypothetical protein, partial [Escherichia coli]	2,00E-07	-1	#0770	807786	811985	4200	core protein
#11619	1769360	1769551	192	192	hypothetical protein [Klebsiella pneumoniae]	1,00E-10	-1	#1887	1768880	1769794	915	Muramoyltetrapeptide carboxypeptidase
#25346	3692490	3692681	192	192	hypothetical protein [Klebsiella pneumoniae]	1,00E-11	-1	#3933	3692016	3694715	2700	CRISPR-associated helicase Cas3, protein
#25966	3777243	3777434	192	192	hypothetical protein, partial [Bacillus cereus]	2,00E-13	-1	#4008	3776214	3778460	2247	Phosphocarrier protein kinase/phosphorylase, nitrogen regulation associated
#29496	4282223	4282414	192	192	hypothetical protein [Klebsiella oxytoca]	9,00E-07	-1	#4546	4282211	4282498	288	tRNA 5-methylaminomethyl-2-thiouridine synthase TusB
#41723	5018051	5018242	192	192	hypothetical protein [Escherichia coli]	1,00E-24	-1	#5269	5016503	5020687	4185	core protein
#44596	4594056	4594247	192	192	hypothetical protein [Escherichia coli]	1,00E-24	-1	#4854	4592508	4596737	4230	core protein
#5114	778449	778640	192	192	hypothetical protein [Dickeya dianthicola]	7,00E-08	-1	#0741	777453	779117	1665	Asparagine synthetase [glutamine-hydrolyzing]
#52	7815	8006	192	162	hypothetical protein, partial [Shigella flexneri]	6,00E-39	-1	#0007	6546	7976	1431	Putative alanine/glycine transport protein
#5209	792445	792636	192	159	hypothetical protein, partial [Staphylococcus aureus]	4,00E-06	-1	#0758	791839	792603	765	Esterase ybfF
#56564	2802234	2802425	192	177	hypothetical protein [Escherichia coli]	6,00E-29	-1	#3043	2802222	2802410	189	Division inhibition protein dicB
#64181	1711170	1711361	192	192	hypothetical protein, partial [Escherichia coli]	4,00E-05	-1	#1817	1711155	1711844	690	Putative Q antiterminator encoded by prophage CP-933P
#65092	1594131	1594322	192	126	hypothetical protein, partial [Aeromonas encheleia]	5,00E-05	-1	#1666	1594197	1594370	174	LSU ribosomal protein L32p
#70915	809334	809525	192	192	hypothetical protein [Escherichia coli]	1,00E-24	-1	#0770	807786	811985	4200	core protein
#13539	2066621	2066809	189	189	hypothetical protein [Mesorhizobium loti]	2,00E-04	-1	#2209	2066192	2067205	1014	Putrescine transport ATP-binding protein PotA
#17765	2646777	2646965	189	99	hypothetical protein, partial [Klebsiella pneumoniae]	2,00E-18	-1	#2840	2645142	2646875	1734	Aspartyl-tRNA synthetase
#24007	3504844	3505032	189	189	hypothetical protein [Klebsiella michiganensis]	3,00E-11	-1	#3724	3504628	3505131	504	tRNA-specific adenosine-34 deaminase
#48365	4037918	4038106	189	168	hypothetical protein [Halomonas halodenitrificans]	3,00E-11	-1	#4286	4037939	4038154	216	SSU ribosomal protein S21p
#63847	1758741	1758929	189	189	hypothetical protein [Escherichia coli]	9,00E-08	-1	#1878	1757859	1759127	1269	Error-prone, lesion bypass DNA polymerase V (UmuC)
#12149	1840715	1840900	186	186	hypothetical protein, partial [Escherichia coli]	3,00E-31	-1	#1958	1840382	1840921	540	Intracellular septation protein IspA
#13923	2114606	2114791	186	186	hypothetical protein [Burkholderia glathei]	2,00E-34	-1	#2250	2112593	2115232	2640	Putative uncharacterized protein ydbH
#3315	505931	506116	186	186	dehydrogenase [Bacteroides stercoris CAG:120]	6,00E-07	-1	#0489	504368	506230	1863	1-deoxy-D-xylulose 5-phosphate synthase
#34253	4978663	4978848	186	186	hypothetical protein [Salmonella enterica]	3,00E-16	-1	#5225	4978627	4979073	447	PTS system, fructose-specific IIA component
#4626	714491	714676	186	186	hypothetical protein, partial [Shigella flexneri]	2,00E-36	-1	#0671	713636	714724	1089	Glycerol dehydrogenase
#47455	4172364	4172549	186	186	3-deoxy-D-manno-octulosonate 8-phosphate phosphatase KdsC, partial [Escherichia coli]	1,00E-12	-1	#4426	4172253	4172819	567	3-deoxy-D-manno-octulosonate 8-phosphate phosphatase
#52617	3392740	3392925	186	186	hypothetical protein, partial [Staphylococcus aureus]	3,00E-04	-1	#3624	3390610	3393723	3114	RND efflux system, inner membrane transporter CmeB
#61592	2057917	2058102	186	186	permease, partial [Escherichia coli]	6,00E-10	-1	#2198	2057041	2059059	2019	putative tonB-dependent receptor yncD precursor

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#65608	1529249	1529434	186	105	hypothetical protein, partial [Escherichia coli]	1,00E-34	-1	#1582	1529330	1530202	873	NgrB
#68611	1129105	1129290	186	105	hypothetical protein, partial [Escherichia coli]	1,00E-34	-1	#1121	1129186	1130058	873	NgrB
#10485	1591014	1591196	183	183	hypothetical protein [Serratia marcescens]	2,00E-17	-1	#1661	1588023	1591208	3186	Ribonuclease E
#19448	2876186	2876368	183	183	hypothetical protein [Escherichia coli]	9,00E-23	-1	#3113	2875901	2876794	894	UTP--glucose-1-phosphate uridylyltransferase
#22469	3289718	3289900	183	183	hypothetical protein [Escherichia coli]	2,00E-35	-1	#3528	3288980	3290227	1248	Sucrose permease, major facilitator superfamily
#37586	5469849	5470031	183	183	hypothetical protein [Serratia symbiotica]	6,00E-05	-1	#5672	5465343	5470379	5037	adherence and invasion outermembrane protein (Inv,enhances Peyer's patches colonization)
#38135	5542403	5542585	183	183	hypothetical protein [Citrobacter amalonaticus]	1,00E-10	-1	#5743	5542367	5543791	1425	Two-component response regulator CreC
#40127	5247932	5248114	183	174	hypothetical protein [Escherichia coli]	3,00E-13	-1	#5466	5246771	5248105	1335	Melibiose carrier protein, Na+/melibiose symporter
#72401	587143	587325	183	183	hypothetical protein [Escherichia coli]	1,00E-04	-1	#0566	585946	601512	15567	putative cell-wall-anchored protein SasA (LPXTG motif)
#73008	499868	500050	183	162	hypothetical protein, partial [Escherichia coli]	2,00E-07	-1	#0485	499889	500866	978	Thiamine-monophosphate kinase
#15101	2272021	2272200	180	135	hypothetical protein [Escherichia coli]	5,00E-35	-1	#2426	2271274	2272155	882	putative metal-dependent phosphoesterases (PHP family)
#23757	3469546	3469725	180	132	hypothetical protein [Pseudomonas pseudoalcaligenes]	2,00E-08	-1	#3694	3469189	3469677	489	Iron-sulfur cluster regulator IscR
#25277	3683212	3683391	180	180	hypothetical protein, partial [Catenibacterium mitsuokai]	7,00E-08	-1	#3922	3682141	3683568	1428	Sulfate adenylyltransferase subunit 1
#33827	4913725	4913904	180	180	hypothetical protein, partial [Salmonella enterica]	1,00E-15	-1	#5166	4912528	4914717	2190	Enoyl-CoA hydratase
#41659	5026134	5026313	180	180	hypothetical protein [Escherichia coli]	6,00E-14	-1	#5275	5024343	5026775	2433	Aspartokinase
#42771	4873496	4873675	180	180	hypothetical protein [Escherichia coli]	6,00E-19	-1	#5124	4872437	4874983	2547	Adenylate cyclase
#53128	3313930	3314109	180	180	hypothetical protein, partial [Bacillus cereus]	1,00E-14	-1	#3549	3313828	3315507	1680	Autolysis histidine kinase LytS
#60520	2228862	2229041	180	111	hypothetical protein [Escherichia coli]	1,00E-32	-1	#2386	2227554	2228972	1419	Gamma-glutamyl-putrescine synthetase
#75929	92152	92331	180	180	hypothetical protein [Escherichia coli]	3,00E-30	-1	#0081	91933	92454	522	Acetolactate synthase small subunit
#43056	4836501	4836677	177	177	hypothetical protein [Escherichia coli]	6,00E-05	-1	#5093	4836402	4837946	1545	Threonine dehydratase biosynthetic
#44380	4623409	4623585	177	108	glycosyl transferase [Escherichia coli]	1,00E-08	-1	#4878	4622557	4623516	960	Putative periplasmic protein YibQ
#61510	2071006	2071182	177	165	collagenase [Shigella dysenteriae]	1,00E-22	-1	#2215	2071018	2071188	171	hypothetical protein
#6849	1039134	1039310	177	135	hypothetical protein [Escherichia coli]	1,00E-05	-1	#1004	1039176	1040177	1002	Low-specificity L-threonine aldolase
#7906	1202769	1202945	177	177	hypothetical protein, partial [Escherichia coli]	2,00E-04	-1	#1192	1202697	1203785	1089	Outer membrane protein F precursor
#11565	1760776	1760949	174	174	hypothetical protein, partial [Bacillus cereus]	4,00E-06	-1	#1880	1759849	1761390	1542	Na+/H+ antiporter NhaB
#21171	3116812	3116985	174	174	hypothetical protein [Escherichia coli]	2,00E-23	-1	#3362	3116773	3117252	480	Cytochrome c-type biogenesis protein CcmE, heme chaperone
#25354	3693291	3693464	174	174	hypothetical protein [Klebsiella pneumoniae]	2,00E-13	-1	#3933	3692016	3694715	2700	CRISPR-associated helicase Cas3, protein
#26056	3788099	3788272	174	174	hypothetical protein, partial [Acinetobacter haemolyticus]	8,00E-10	-1	#4019	3787664	3788926	1263	Diaminopimelate decarboxylase
#33029	4794504	4794677	174	174	hypothetical protein, partial [Escherichia coli]	2,00E-06	-1	#5059	4792908	4794737	1830	Glucosamine--fructose-6-phosphate aminotransferase [isomerizing]
#35648	5185186	5185359	174	93	hypothetical protein [Escherichia coli]	3,00E-07	-1	#5407	5185267	5186916	1650	Acetate permease ActP (cation/acetate symporter)
#36062	5240083	5240256	174	174	hypothetical protein, partial [Escherichia coli]	4,00E-09	-1	#5460	5239009	5240346	1338	Arginine/agmatine antiporter
#49123	3925599	3925772	174	174	hypothetical protein [Escherichia coli]	2,00E-25	-1	#4165	3925515	3925805	291	UPF0235 protein
#51878	3507009	3507182	174	174	hypothetical protein [Mannheimia varigena]	3,00E-09	-1	#3727	3506937	3507197	261	4Fe-4S ferredoxin, iron-sulfur binding protein
#56286	2837590	2837763	174	132	hypothetical protein [Escherichia coli]	3,00E-29	-1	#3073	2837632	2837853	222	Uncharacterized protein YeeT
#58364	2536642	2536815	174	174	hypothetical protein [Taylorella asinigenitalis]	5,00E-06	-1	#2721	2536345	2537688	1344	NADP-specific glutamate dehydrogenase
#71751	683610	683783	174	174	hypothetical protein, partial [Lactobacillus vaginalis]	2,00E-30	-1	#0639	681027	684164	3138	Cobalt-zinc-cadmium resistance protein CzcA
#10314	1565041	1565211	171	171	hypothetical protein, partial [Rhodococcus qingshengii]	8,00E-22	-1	#1632	1564777	1565352	576	Protein ycel precursor
#124	20301	20471	171	171	hypothetical protein, partial [Escherichia coli]	1,00E-21	-1	#0019	19299	21749	2451	Putative outer membrane protein
#26059	3788747	3788917	171	171	hypothetical protein, partial [Salmonella enterica]	2,00E-06	-1	#4019	3787664	3788926	1263	Diaminopimelate decarboxylase

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#30340	4400507	4400677	171	171	hypothetical protein [Escherichia coli]	1,00E-22	-1	#4666	4400315	4401592	1278	Branched-chain amino acid transport system permease protein LivM
#47462	4171261	4171431	171	171	hypothetical protein, partial [Escherichia coli]	3,00E-13	-1	#4425	4171246	4172232	987	Arabinose 5-phosphate isomerase
#59316	2399083	2399253	171	171	hypothetical protein [Pluralibacter gergoviae]	6,00E-10	-1	#2580	2399041	2399256	216	Cnu protein
#59355	2394819	2394989	171	171	hypothetical protein [Salmonella enterica]	4,00E-06	-1	#2575	2393445	2395037	1593	PTS system, maltose and glucose-specific IIC component
#73895	380157	380327	171	171	hypothetical protein [Escherichia coli]	2,00E-17	-1	#0364	379668	383717	4050	AidA-I adhesin-like protein
#10317	1565422	1565589	168	168	hypothetical protein [Escherichia coli]	1,00E-30	-1	#1633	1565356	1565922	567	Cytochrome B561
#14617	2208703	2208870	168	168	hypothetical protein [Escherichia coli]	1,00E-15	-1	#2363	2208430	2210697	2268	Maltose phosphorylase
#22760	3330647	3330814	168	168	hypothetical protein, partial [Escherichia coli]	5,00E-06	-1	#3562	3329432	3331621	2190	hypothetical protein
#31597	4583146	4583313	168	168	hypothetical protein [Aeromonas rivuli]	3,00E-04	-1	#4845	4583062	4583385	324	GTPase
#32428	4713571	4713738	168	168	hypothetical protein [Rhodococcus qingshengii]	6,00E-23	-1	#4981	4713133	4713951	819	Methionine ABC transporter substrate-binding protein
#32580	4733573	4733740	168	138	hypothetical protein, partial [Escherichia coli]	1,00E-18	-1	#5001	4733213	4733710	498	Putative transcriptional regulator
#33809	4911577	4911744	168	168	hypothetical protein, partial [Anaerococcus lactolyticus]	2,00E-04	-1	#5165	4911355	4912518	1164	3-ketoacyl-CoA thiolase
#47226	4203876	4204043	168	168	glucarate transporter, partial [Escherichia coli]	2,00E-06	-1	#4456	4203606	4204973	1368	Outer membrane stress sensor protease DegQ, serine protease
#53521	3255105	3255272	168	141	hypothetical protein [Klebsiella pneumoniae]	2,00E-18	-1	#3491	3255132	3257138	2007	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase
#5767	874373	874540	168	168	hypothetical protein [Escherichia coli]	3,00E-04	-1	#0832	873626	874672	1047	Galactose-1-phosphate uridylyltransferase
#64431	1678596	1678763	168	168	isopropylmalate isomerase, partial [Escherichia coli]	1,00E-15	-1	#1771	1677786	1678910	1125	Tripeptide aminopeptidase
#72285	604991	605158	168	168	hypothetical protein [Edwardsiella tarda]	6,00E-06	-1	#0569	604847	605254	408	HTH-type transcriptional regulator cueR
#73620	416001	416168	168	168	hypothetical protein [Dickeya dadantii]	4,00E-14	-1	#0399	415575	416234	660	Carbonic anhydrase
#14092	2139417	2139581	165	165	hypothetical protein [Escherichia coli]	4,00E-16	-1	#2277	2139405	2139749	345	hypothetical protein
#15554	2334182	2334346	165	159	hypothetical protein [Escherichia coli]	2,00E-13	-1	#2507	2334173	2334340	168	hypothetical protein
#17637	2629467	2629631	165	165	hypothetical protein [Escherichia coli]	8,00E-07	-1	#2824	2629086	2630897	1812	Phosphogluconate dehydratase
#18867	2791463	2791627	165	159	hypothetical protein [Escherichia coli]	2,00E-13	-1	#3024	2791454	2791621	168	hypothetical protein
#21758	3195816	3195980	165	165	hypothetical protein, partial [Klebsiella pneumoniae]	2,00E-06	-1	#3430	3195108	3196778	1671	2-succinyl-5-enolpyruvyl-6-hydroxy-3- cyclohexene-1-carboxylic-acid synthase
#21918	3217697	3217861	165	93	hypothetical protein [Bacillus cereus]	1,00E-05	-1	#3450	3217127	3217789	663	NADH-ubiquinone oxidoreductase chain B
#2622	403961	404125	165	165	hypothetical protein [Rhodococcus qingshengii]	4,00E-28	-1	#0389	403556	405142	1587	Propionate catabolism operon regulatory protein PrpR
#31610	4584995	4585159	165	165	hypothetical protein, partial [Vibrio parahaemolyticus]	3,00E-09	-1	#4847	4583930	4585468	1539	Aldehyde dehydrogenase B
#49136	3924275	3924439	165	165	hypothetical protein, partial [Klebsiella pneumoniae]	7,00E-06	-1	#4163	3924230	3924934	705	Hypothetical protein YggS, proline synthase co-transcribed bacterial PROSC-like protein
#54570	3093648	3093812	165	147	hypothetical protein [Acinetobacter baumannii]	3,00E-13	-1	#3342	3093228	3093794	567	Lipoprotein spr precursor
#55459	2966545	2966709	165	165	hypothetical protein [Escherichia coli]	2,00E-08	-1	#3190	2965609	2967330	1722	Molybdate metabolism regulator
#64725	1642360	1642524	165	159	hypothetical protein [Escherichia coli]	5,00E-13	-1	#1724	1642366	1642533	168	hypothetical protein
#75348	174509	174673	165	165	hypothetical protein [Stackebrandtia nassauensis]	5,00E-04	-1	#0156	174101	174898	798	Ferric hydroxamate ABC transporter, ATP-binding protein FhuC
#19559	2891722	2891883	162	162	hypothetical protein [Desulfobulbus alkaliphilus]	2,00E-04	-1	#3126	2890873	2891994	1122	GDP-mannose 4,6-dehydratase
#23686	3461687	3461848	162	147	hypothetical protein [Escherichia coli]	3,00E-09	-1	#3684	3461702	3462478	777	Protein SseB
#41740	5016113	5016274	162	144	hypothetical protein [Microbulbifer variabilis]	3,00E-04	-1	#5268	5016131	5016343	213	LSU ribosomal protein L31p, zinc-dependent
#51703	3532951	3533112	162	126	hypothetical protein, partial [Shigella dysenteriae]	3,00E-29	-1	#3753	3532513	3533076	564	hypothetical protein
#63653	1790000	1790161	162	138	hypothetical protein [Kosakonia radicincitans]	6,00E-07	-1	#1909	1789859	1790137	279	Putative membrane protein YchH
#72426	583936	584097	162	162	hypothetical protein, partial [Escherichia coli]	0,001	-1	#0565	581353	585738	4386	Large repetitive protein
#74124	347436	347597	162	162	hypothetical protein [Escherichia coli]	1,00E-26	-1	#0332	346224	347762	1539	IS66 transposase
#74660	270658	270819	162	162	hypothetical protein, partial [Escherichia coli]	3,00E-12	-1	#0240	267079	271293	4215	core protein

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#75100	207903	208064	162	162	hypothetical protein, partial [Escherichia coli]	2,00E-12	-1	#0188	207837	208433	597	Ribonuclease HII
#75203	196839	197000	162	135	hypothetical protein, partial [Escherichia coli]	3,00E-14	-1	#0177	196866	198062	1197	1-deoxy-D-xylulose 5-phosphate reductoisomerase
#75367	171840	172001	162	162	hypothetical protein [Klebsiella pneumoniae]	8,00E-04	-1	#0155	171807	174050	2244	Ferric hydroxamate outer membrane receptor FhuA
#10743	1630842	1631000	159	159	hypothetical protein, partial [Escherichia coli]	1,00E-20	-1	#1703	1630785	1631129	345	Exodeoxyribonuclease VIII
#12178	1844213	1844371	159	159	hypothetical protein, partial [Escherichia coli]	1,00E-20	-1	#1962	1844156	1846627	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#29421	4270209	4270367	159	159	propionate CoA-transferase [Escherichia coli]	1,00E-20	-1	#4526	4270092	4270406	315	LSU ribosomal protein L24p (L26e)
#34795	5054685	5054843	159	159	hypothetical protein, partial [Klebsiella pneumoniae]	2,00E-10	-1	#5298	5053653	5055053	1401	Soluble pyridine nucleotide transhydrogenase
#40737	5154570	5154728	159	159	hypothetical protein, partial [Escherichia coli]	4,00E-12	-1	#5376	5154555	5155052	498	Chorismate--pyruvate lyase
#56537	2804947	2805105	159	159	hypothetical protein, partial [Escherichia coli]	3,00E-09	-1	#3045	2802691	2805162	2472	Exodeoxyribonuclease VIII
#59687	2348082	2348240	159	159	hypothetical protein, partial [Escherichia coli]	2,00E-21	-1	#2528	2345826	2348297	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#60872	2175617	2175775	159	159	hypothetical protein [Escherichia coli]	1,00E-22	-1	#2329	2174975	2175910	936	tRNA(Cytosine32)-2-thiocytidine synthetase
#8239	1249993	1250151	159	159	hypothetical protein, partial [Escherichia coli]	4,00E-09	-1	#1241	1249936	1252407	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#12471	1890121	1890276	156	147	hypothetical protein [Klebsiella oxytoca]	1,00E-09	-1	#2019	1890106	1890267	162	hypothetical protein
#2033	308149	308304	156	129	hypothetical protein [Klebsiella pneumoniae]	2,00E-06	-1	#0289	307504	308277	774	hypothetical protein
#24491	3570479	3570634	156	147	hypothetical protein [Klebsiella oxytoca]	1,00E-09	-1	#3795	3570464	3570625	162	hypothetical protein
#2453	372387	372542	156	114	membrane protein [Escherichia coli]	5,00E-19	-1	#0359	372429	372587	159	hypothetical protein
#25635	3733658	3733813	156	153	hypothetical protein, partial [Staphylococcus aureus]	6,00E-14	-1	#3972	3733028	3733810	783	tRNA pseudouridine synthase C
#26410	3839912	3840067	156	105	hypothetical protein [Klebsiella oxytoca]	2,00E-04	-1	#4076	3838391	3840016	1626	Uncharacterized protein YqeB
#38054	5539285	5539440	156	141	hypothetical protein, partial [Escherichia coli]	1,00E-28	-1	#5738	5538904	5539425	522	Inosine/xanthosine triphosphatase
#48280	4048786	4048941	156	156	hypothetical protein [Escherichia coli]	9,00E-25	-1	#4297	4048624	4049607	984	Evolved beta-D-galactosidase transcriptional repressor
#52405	3419632	3419787	156	156	mobilization protein mbeA, partial [Escherichia coli]	5,00E-04	-1	#3649	3419128	3420591	1464	Exported zinc metalloprotease YfgC precursor
#56353	2830268	2830423	156	147	hypothetical protein [Klebsiella oxytoca]	1,00E-09	-1	#3065	2830277	2830438	162	hypothetical protein
#59132	2422141	2422296	156	156	hypothetical protein [Azotobacter chroococcum]	4,00E-05	-1	#2608	2422024	2422431	408	Lactoylglutathione lyase
#60079	2287088	2287243	156	147	hypothetical protein [Klebsiella oxytoca]	1,00E-09	-1	#2444	2287097	2287258	162	hypothetical protein
#63949	1742677	1742832	156	147	hypothetical protein [Klebsiella oxytoca]	1,00E-09	-1	#1856	1742686	1742847	162	hypothetical protein
#67188	1317308	1317463	156	156	hypothetical protein [Loktanella vestfoldensis]	6,00E-04	-1	#1327	1317287	1317499	213	Cold shock protein CspG
#7049	1070156	1070311	156	147	hypothetical protein [Klebsiella oxytoca]	1,00E-09	-1	#1031	1070141	1070302	162	hypothetical protein
#74140	345660	345815	156	147	hypothetical protein [Klebsiella oxytoca]	1,00E-09	-1	#0330	345669	345830	162	hypothetical protein
#9671	1470300	1470455	156	147	hypothetical protein [Klebsiella oxytoca]	1,00E-09	-1	#1492	1470285	1470446	162	hypothetical protein
#11568	1761139	1761291	153	153	hypothetical protein [Glaciecola pallidula]	8,00E-05	-1	#1880	1759849	1761390	1542	Na+/H+ antiporter NhaB
#20468	3019226	3019378	153	138	hypothetical protein, partial [Escherichia coli]	4,00E-13	-1	#3256	3018662	3019363	702	Origin specific replication binding factor
#37201	5407868	5408020	153	153	hypothetical protein [Escherichia coli]	1,00E-08	-1	#5629	5407691	5408542	852	hypothetical protein
#49373	3888335	3888487	153	153	hypothetical protein, partial [Escherichia coli]	8,00E-05	-1	#4121	3888089	3889084	996	putative periplasmic protein kinase ArgK and related GTPases of G3E family
#5110	777972	778124	153	153	hypothetical protein [Dickeya dianthicola]	2,00E-09	-1	#0741	777453	779117	1665	Asparagine synthetase [glutamine-hydrolyzing]
#55768	2921728	2921880	153	153	hypothetical protein, partial [Acinetobacter haemolyticus]	1,00E-06	-1	#3150	2919004	2922126	3123	Multidrug transporter MdtB
#64216	1706066	1706218	153	138	hypothetical protein, partial [Escherichia coli]	2,00E-13	-1	#1808	1706081	1706782	702	Origin specific replication binding factor
#66983	1347959	1348111	153	138	hypothetical protein, partial [Escherichia coli]	2,00E-13	-1	#1371	1347974	1348675	702	Replication protein P
#14414	2180816	2180965	150	129	hypothetical protein [Escherichia coli]	3,00E-07	-1	#2333	2180381	2180944	564	hypothetical protein
#26054	3787940	3788089	150	150	hypothetical protein, partial [Acinetobacter haemolyticus]	1,00E-09	-1	#4019	3787664	3788926	1263	Diaminopimelate decarboxylase

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#28533	4147961	4148110	150	141	hypothetical protein [Klebsiella pneumoniae]	5,00E-19	-1	#4399	4147679	4148101	423	protein clustered with transcription termination protein NusA
#45611	4441072	4441221	150	147	hypothetical protein [Escherichia coli]	1,00E-16	-1	#4712	4440817	4441218	402	Nickel responsive regulator NikR
#4727	728979	729128	150	150	hypothetical protein, partial [Escherichia coli]	1,00E-12	-1	#0687	728616	729167	552	Apo-citrate lyase phosphoribosyl-dephospho-CoA transferase
#531	79156	79305	150	150	hypothetical protein [Escherichia coli]	2,00E-25	-1	#0069	77731	79341	1611	Thiamin ABC transporter, transmembrane component
#63300	1839702	1839851	150	138	DNA topoisomerase I, partial [Escherichia coli]	1,00E-04	-1	#1956	1839120	1839839	720	Ferric siderophore transport system, periplasmic binding protein TonB
#70269	902592	902741	150	150	hypothetical protein, partial [Escherichia coli]	2,00E-19	-1	#0866	902286	902753	468	putative endopeptidase
#75816	105356	105505	150	135	hypothetical protein [Klebsiella pneumoniae]	1,00E-06	-1	#0094	105371	106846	1476	UDP-N-acetylmuramate--alanine ligase
#8278	1254257	1254406	150	96	hypothetical protein [Escherichia coli]	3,00E-18	-1	#1247	1254224	1254352	129	hypothetical protein
#13433	2050893	2051039	147	147	hypothetical protein, partial [Citrobacter rodentium]	4,00E-05	-1	#2190	2049939	2052047	2109	VgrG protein
#2759	422902	423048	147	147	hypothetical protein [Bacillus thuringiensis]	3,00E-21	-1	#0404	420007	423081	3075	Beta-galactosidase
#35904	5219306	5219452	147	147	hypothetical protein, partial [Escherichia coli]	4,00E-18	-1	#5440	5219246	5219803	558	ATP-binding protein PhnN
#44073	4671439	4671585	147	147	hypothetical protein [Klebsiella pneumoniae]	1,00E-05	-1	#4927	4671430	4671804	375	YeeV toxin protein
#55744	2924518	2924664	147	147	hypothetical protein, partial [Anaerococcus lactolyticus]	3,00E-06	-1	#3151	2922127	2925204	3078	Multidrug transporter MdtC
#57364	2683488	2683634	147	144	acid-inducible small membrane-associated protein [Escherichia coli]	7,00E-11	-1	#2877	2683485	2683631	147	hypothetical protein
#6236	947812	947958	147	147	hypothetical protein, partial [Elizabethkingia anophelis]	2,00E-04	-1	#0913	947140	948246	1107	ABC transport system, permease component YbhR
#65505	1540988	1541134	147	147	hypothetical protein [Klebsiella pneumoniae]	6,00E-05	-1	#1599	1540979	1541260	282	YeeV toxin protein
#67988	1209741	1209887	147	147	hypothetical protein [Escherichia coli]	1,00E-12	-1	#1196	1207425	1210037	2613	Membrane alanine aminopeptidase N
#68506	1140841	1140987	147	147	hypothetical protein [Klebsiella pneumoniae]	6,00E-05	-1	#1140	1140832	1141113	282	YeeV toxin protein
#73793	393228	393374	147	147	hypothetical protein, partial [Escherichia coli]	1,00E-12	-1	#0376	392715	394097	1383	Cytosine deaminase
#74698	265870	266016	147	147	hypothetical protein, partial [Escherichia coli]	5,00E-08	-1	#0239	264862	267003	2142	VgrG protein
#13410	2048160	2048303	144	144	hypothetical protein [Escherichia coli]	2,00E-04	-1	#2189	2045670	2049872	4203	core protein
#15085	2270429	2270572	144	144	hypothetical protein, partial [Escherichia coli]	2,00E-20	-1	#2424	2269046	2270629	1584	hypothetical protein
#16268	2436935	2437078	144	144	hypothetical protein [Pseudomonas aeruginosa]	3,00E-04	-1	#2620	2436773	2437414	642	Riboflavin synthase eubacterial/eukaryotic
#16607	2485632	2485775	144	144	hypothetical protein [Escherichia coli]	2,00E-22	-1	#2666	2485482	2486195	714	hypothetical protein
#45998	4388110	4388253	144	144	hypothetical protein, partial [Escherichia coli]	1,00E-20	-1	#4652	4388020	4388508	489	Putative acetyltransferase
#46199	4354410	4354553	144	144	hypothetical protein, partial [Klebsiella pneumoniae]	8,00E-06	-1	#4620	4352820	4355525	2706	Transcriptional activator of maltose regulon, MalT
#48079	4075188	4075331	144	144	hypothetical protein [Vibrio parahaemolyticus]	1,00E-11	-1	#4318	4075032	4075694	663	DedA family inner membrane protein YqjA
#56554	2802955	2803098	144	144	cell surface protein, partial [Escherichia coli]	3,00E-19	-1	#3045	2802691	2805162	2472	Exodeoxyribonuclease VIII
#65757	1511452	1511595	144	138	hypothetical protein [Escherichia coli]	6,00E-11	-1	#1554	1511434	1511589	156	hypothetical protein
#68760	1111308	1111451	144	138	hypothetical protein [Escherichia coli]	6,00E-11	-1	#1093	1111290	1111445	156	hypothetical protein
#72187	618317	618460	144	144	hypothetical protein [Escherichia coli]	8,00E-04	-1	#0585	616748	620944	4197	core protein
#8256	1252000	1252143	144	144	cell surface protein, partial [Escherichia coli]	3,00E-19	-1	#1241	1249936	1252407	2472	Exodeoxyribonuclease encoded by cryptic prophage CP-933P
#8696	1316903	1317046	144	99	hypothetical protein [Edwardsiella tarda]	1,00E-04	-1	#1326	1316789	1317001	213	Cold shock protein CspH
#31278	4539809	4539949	141	114	hypothetical protein, partial [Escherichia coli]	2,00E-08	-1	#4805	4538231	4539922	1692	Phosphoethanolamine transferase specific for the outer Kdo residue of lipopolysaccharide
#33716	4896447	4896587	141	141	hypothetical protein [Klebsiella oxytoca]	2,00E-06	-1	#5150	4895799	4896752	954	Transcriptional activator MetR
#45150	4509160	4509300	141	120	hypothetical protein [Salmonella enterica]	7,00E-09	-1	#4779	4509181	4510110	930	2-dehydro-3-deoxygluconate kinase
#72522	570740	570880	141	141	hypothetical protein [Escherichia coli]	2,00E-06	-1	#0556	570704	572356	1653	UDP-sugar hydrolase
#17990	2674315	2674452	138	138	hypothetical protein [Serratia marcescens]	2,00E-07	-1	#2867	2674237	2674587	351	Flagellar transcriptional activator FlhD
#21444	3150399	3150536	138	138	hypothetical protein [Escherichia coli]	2,00E-19	-1	#3391	3149046	3150644	1599	hypothetical protein

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#29441	4272917	4273054	138	120	hypothetical protein [Halomonas halodenitrificans]	1,00E-08	-1	#4533	4272887	4273036	150	SSU ribosomal protein S19p (S15e)
#35210	5119185	5119322	138	138	hypothetical protein, partial [Bacillus cereus]	2,00E-07	-1	#5345	5118831	5119655	825	Acetate operon repressor
#49959	3796411	3796548	138	138	hypothetical protein [Escherichia coli]	2,00E-18	-1	#4026	3795874	3797103	1230	Serine transporter
#56279	2838402	2838539	138	138	hypothetical protein [Klebsiella pneumoniae]	3,00E-04	-1	#3075	2838384	2838647	264	YeeV toxin protein
#61417	2087752	2087889	138	138	hypothetical protein, partial [Shigella flexneri]	3,00E-18	-1	#2232	2087563	2088564	1002	NAD-dependent glyceraldehyde-3-phosphate dehydrogenase
#13627	2077590	2077724	135	114	hypothetical protein [Escherichia coli]	3,00E-12	-1	#2221	2076711	2077703	993	Tellurite resistance protein TehA
#1737	261032	261166	135	114	hypothetical protein, partial [Escherichia coli]	8,00E-07	-1	#0232	260732	261145	414	Uncharacterized protein, VCA0109 like protein
#28111	4083386	4083520	135	135	hypothetical protein, partial [Escherichia coli]	9,00E-08	-1	#4331	4082597	4083907	1311	Inner membrane protein
#33345	4840730	4840864	135	114	hypothetical protein [Cronobacter sakazakii]	3,00E-05	-1	#5096	4840562	4840843	282	Peptidyl-prolyl cis-trans isomerase PpiC
#22780	3333344	3333475	132	132	hypothetical protein [Shigella flexneri]	1,00E-12	-1	#3565	3332918	3334066	1149	putative virulence protein
#29524	4286134	4286265	132	132	hypothetical protein [Erwinia amylovora]	5,00E-04	-1	#4553	4286125	4286325	201	Putative cytoplasmic protein ,probably associated with Glutathione-regulated potassium-efflux
#39568	5331727	5331858	132	132	hypothetical protein [Shigella flexneri]	1,00E-13	-1	#5551	5331136	5332284	1149	putative virulence protein
#40739	5154138	5154269	132	132	hypothetical protein, partial [Escherichia coli]	1,00E-19	-1	#5375	5152752	5154332	1581	YjbI protein
#5721	868306	868437	132	132	hypothetical protein [Salmonella enterica]	6,00E-04	-1	#0825	867523	868458	936	Zinc transporter ZitB
#13541	2066819	2066947	129	129	hypothetical protein, partial [Escherichia coli]	2,00E-08	-1	#2209	2066192	2067205	1014	Putrescine transport ATP-binding protein PotA
#14875	2242233	2242361	129	105	hypothetical protein, partial [Staphylococcus aureus]	7,00E-07	-1	#2397	2239254	2242337	3084	RND efflux system, inner membrane transporter CmeB
#55010	3028548	3028676	129	129	hypothetical protein, partial [Escherichia coli]	2,00E-17	-1	#3272	3027831	3028778	948	hypothetical protein
#64192	1709781	1709909	129	129	hypothetical protein, partial [Escherichia coli]	1,00E-14	-1	#1812	1709754	1710209	456	hypothetical protein
#69594	996445	996573	129	129	hypothetical protein [Escherichia coli]	2,00E-19	-1	#0957	995995	996906	912	Dipeptide transport system permease protein DppC
#7192	1091346	1091474	129	129	hypothetical protein [Klebsiella oxytoca]	2,00E-05	-1	#1064	1091148	1091477	330	hypothetical protein
#73660	411520	411648	129	105	hypothetical protein [Escherichia coli]	1,00E-04	-1	#0395	411430	411624	195	hypothetical protein
#75773	111244	111372	129	129	hypothetical protein [Klebsiella pneumoniae]	6,00E-17	-1	#0099	111163	112080	918	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase
#9814	1491490	1491618	129	129	hypothetical protein [Klebsiella oxytoca]	2,00E-05	-1	#1525	1491292	1491621	330	hypothetical protein
#29302	4255177	4255302	126	126	hypothetical protein [Cedecea neteri]	4,00E-10	-1	#4503	4255075	4256199	1125	Rossmann fold nucleotide-binding protein Smf possibly involved in DNA uptake
#40197	5235176	5235301	126	126	hypothetical protein [Escherichia coli]	1,00E-10	-1	#5456	5233817	5235319	1503	L-Proline/Glycine betaine transporter ProP
#47871	4109091	4109216	126	120	hypothetical protein [Enterobacter aerogenes]	4,00E-04	-1	#4355	4108332	4109210	879	PTS system, N-acetylgalactosamine-specific IID component
#51957	3492904	3493029	126	114	hypothetical protein [Salmonella enterica]	1,00E-04	-1	#3716	3492916	3493044	129	hypothetical protein
#71529	714818	714943	126	111	hypothetical protein, partial [Staphylococcus aureus]	3,00E-04	-1	#0672	714833	715993	1161	Methionine aminotransferase, PLP-dependent
#25260	3681259	3681381	123	123	hypothetical protein [Citrobacter amalonaticus]	1,00E-07	-1	#3920	3681163	3681486	324	Putative cytochrome oxidase subunit
#2749	421825	421947	123	123	hypothetical protein [Odoribacter splanchnicus CAG:14]	5,00E-04	-1	#0404	420007	423081	3075	Beta-galactosidase
#33710	4896132	4896254	123	123	hypothetical protein [Morganella morganii]	2,00E-05	-1	#5150	4895799	4896752	954	Transcriptional activator MetR
#42937	4852031	4852153	123	123	hypothetical protein, partial [Catenibacterium mitsuokai]	6,00E-08	-1	#5106	4851164	4852294	1131	UDP-N-acetylglucosamine 2-epimerase
#24296	3541671	3541790	120	120	hypothetical protein [Escherichia coli]	2,00E-04	-1	#3756	3541311	3542042	732	hypothetical protein
#33937	4929791	4929910	120	120	hypothetical protein [Cronobacter universalis]	1,00E-05	-1	#5178	4929248	4929991	744	putative acyltransferase yihG
#45423	4470558	4470677	120	120	hypothetical protein [Staphylococcus aureus]	5,00E-07	-1	#4741	4470468	4470821	354	Arsenical resistance operon repressor
#33706	4895928	4896044	117	117	hypothetical protein, partial [Salmonella enterica]	8,00E-10	-1	#5150	4895799	4896752	954	Transcriptional activator MetR
#41724	5017928	5018044	117	117	hypothetical protein [Escherichia coli]	0,001	-1	#5269	5016503	5020687	4185	core protein
#44597	4593933	4594049	117	117	hypothetical protein [Escherichia coli]	0,001	-1	#4854	4592508	4596737	4230	core protein
#46273	4345189	4345305	117	117	hypothetical protein [Raoultella planticola]	3,00E-04	-1	#4615	4344718	4345365	648	Competence protein F-like protein, phosphoribosyltransferase domain

sORF with blastp hit							mother gene					
ID CP008957.1_	start	stop	length (bp)	overlap length	blastp hit	e-value	reading frame	loc_tag EDL933_	start	stop	length (bp)	gene product
#49433	3880406	3880522	117	117	hypothetical protein [Serratia odorifera]	8,00E-05	-1	#4113	3880367	3880696	330	Z-ring-associated protein ZapA
#5105	777687	777803	117	117	hypothetical protein [Dickeya dianthicola]	3,00E-04	-1	#0741	777453	779117	1665	Asparagine synthetase [glutamine-hydrolyzing]
#51103	3622933	3623049	117	117	hypothetical protein, partial [Enterobacter asburiae]	2,00E-05	-1	#3850	3622549	3623079	531	Transcription repressor
#58369	2536369	2536485	117	117	hypothetical protein [Salmonella enterica]	5,00E-10	-1	#2721	2536345	2537688	1344	NADP-specific glutamate dehydrogenase
#70916	809211	809327	117	117	hypothetical protein [Escherichia coli]	0,001	-1	#0770	807786	811985	4200	core protein
#26248	3816676	3816789	114	114	transcriptional regulator [Escherichia coli]	2,00E-16	-1	#4058	3815533	3816825	1293	Flagellum-specific ATP synthase FliI
#5107	777825	777938	114	114	hypothetical protein [Dickeya dianthicola]	1,00E-08	-1	#0741	777453	779117	1665	Asparagine synthetase [glutamine-hydrolyzing]
#5111	778173	778286	114	114	hypothetical protein [Dickeya dianthicola]	6,00E-08	-1	#0741	777453	779117	1665	Asparagine synthetase [glutamine-hydrolyzing]
#51705	3532816	3532929	114	114	alpha-ketoglutarate permease [Escherichia coli]	3,00E-18	-1	#3753	3532513	3533076	564	hypothetical protein
#54936	3039720	3039833	114	114	hypothetical protein [Escherichia coli]	9,00E-05	-1	#3287	3039417	3041132	1716	D-Lactate dehydrogenase
#55035	3026334	3026447	114	114	hypothetical protein [Escherichia coli]	3,00E-14	-1	#3268	3026199	3026876	678	putative enzyme
#5910	894879	894992	114	114	hypothetical protein [Escherichia coli]	8,00E-14	-1	#0855	894447	895127	681	putative enzyme
#69863	957839	957952	114	114	hypothetical protein, partial [Escherichia coli]	1,00E-15	-1	#0922	957143	959431	2289	ATP-dependent helicase DinG/Rad3
#8878	1339743	1339856	114	114	hypothetical protein [Escherichia coli]	3,00E-14	-1	#1358	1339314	1339991	678	putative enzyme
#20583	3034837	3034947	111	111	hypothetical protein [Escherichia coli]	8,00E-07	-1	#3283	3034591	3035748	1158	Osmoprotectant ABC transporter permease protein YehY
#3956	612651	612758	108	108	hypothetical protein, partial [Escherichia coli]	8,00E-11	-1	#0582	612627	613250	624	Arylesterase precursor
#76202	50531	50638	108	108	hypothetical protein, partial [Escherichia coli]	2,00E-07	-1	#0046	50222	51553	1332	Putative metabolite transport protein yaaU
#27594	4008470	4008574	105	105	hypothetical protein, partial [Escherichia coli]	1,00E-09	-1	#4256	4007225	4009117	1893	Topoisomerase IV subunit B
#27404	3982971	3983072	102	102	hypothetical protein, partial [Escherichia coli]	1,00E-06	-1	#4231	3982245	3983144	900	Transcriptional regulator, AraC family
#47092	4227103	4227204	102	102	hypothetical protein [Shigella flexneri]	2,00E-14	-1	#4475	4227082	4227246	165	hypothetical protein
#49878	3809982	3810083	102	102	hypothetical protein, partial [Escherichia coli]	4,00E-08	-1	#4049	3809604	3810104	501	Transcriptional regulator, ArsR family
#8702	1317964	1318065	102	102	hypothetical protein, partial [Vibrio parahaemolyticus]	4,00E-05	-1	#1329	1317871	1318077	207	hypothetical protein
#19592	2896334	2896432	99	99	hypothetical protein, partial [Lactobacillus ultunensis]	7,00E-08	-1	#3132	2896241	2897080	840	Colanic acid biosynthesis glycosyl transferase WcaA
#33718	4896627	4896725	99	99	hypothetical protein [Klebsiella oxytoca]	3,00E-04	-1	#5150	4895799	4896752	954	Transcriptional activator MetR
#39945	5279458	5279556	99	93	hypothetical protein [Escherichia coli]	4,00E-13	-1	#5496	5279464	5279589	126	Entericidin A precursor
#52927	3345363	3345461	99	99	hypothetical protein [Salmonella enterica]	1,00E-10	-1	#3580	3345348	3347075	1728	Phosphoenolpyruvate-protein phosphotransferase of PTS system
#58455	2526216	2526314	99	96	hypothetical protein [Escherichia coli]	4,00E-14	-1	#2708	2526219	2526389	171	hypothetical protein
#32417	4712000	4712095	96	96	sugar transporter [Escherichia albertii]	9,00E-10	-1	#4979	4711958	4712143	186	hypothetical protein
#50791	3667196	3667291	96	96	hypothetical protein [Escherichia coli]	1,00E-06	-1	#3901	3665648	3667726	2079	Formate hydrogenlyase transcriptional activator
#59709	2345369	2345464	96	96	hypothetical protein [Escherichia coli]	3,00E-06	-1	#2526	2345357	2345545	189	Division inhibition protein dicB
#34293	4984753	4984845	93	93	hypothetical protein, partial [Escherichia coli]	8,00E-06	-1	#5230	4983565	4985034	1470	Rhamnulokinase
#35760	5201646	5201738	93	93	LysR family transcriptional regulator, partial [Escherichia coli]	2,00E-08	-1	#5422	5201115	5202581	1467	Outer membrane component of tripartite multidrug resistance system
#61578	2060821	2061216	396	175	hypothetical protein [Escherichia coli]	2,00E-33	2	#2200	2059958	2060995	1038	Putative oxidoreductase YncB
#63696	1781140	1781508	369	139	trehalase [Gramella echinicola]	2,00E-04	2	#1900	1779821	1781278	1458	Trehalase
#52317	3432188	3432478	291	118	membrane protein, partial [Vibrio parahaemolyticus]	1,00E-12	2	#3660	3432153	3432305	153	hypothetical protein
#36888	5361309	5361803	495	196	hypothetical protein [Escherichia coli]	2,00E-04	3	#5580	5360131	5361504	1374	UDP-N-acetylmuramate:L-alanyl-gamma-D-glutamyl- meso-diaminopimelate ligase
#17216	2570664	2570957	294	202	hypothetical protein [Escherichia coli]	1,00E-44	3	#2753	2570419	2570865	447	Putative inner membrane protein
#35611	5178946	5179218	273	124	hypothetical protein, partial [Escherichia coli]	9,00E-18	3	#5399	5177483	5179069	1587	hypothetical protein

Supplementary table S2.1: Localization and features of homologues used for the phylostratigraphic analysis of *laoB*. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to *laoB*	full-length sORF homologue	length sORF (aa)	Identity mORF to ECs5115	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC Sakai	-	-	41	-	512	5216097	5216219	5214977	5216512
CP002729.1|Escherichia coli UMNK88	95.1	yes	40	99.8	512	4884791	4884913	4883668	4885206
CP010129.1|Escherichia coli strain C9	95.1	yes	40	99.4	512	4202639	4202761	4202346	4203884
CP010829.1|Shigella sonnei strain FORC_011	90.2	yes	40	100	99	4589140	4589262	4588021	4589055
CP012693.1|Escherichia coli strain FORC_028	97.6	yes	40	99.8	512	5205839	5205961	5205546	5207084
FN554766.1|Escherichia coli 042	95.1	yes	41	99.2	512	4830629	4830751	4829506	4831044
AP009378.1|Escherichia coli SE15 DNA	100	yes	41	98.8	512	4393319	4393445	4392199	4393734
CP015159.1|Escherichia coli strain Eco889	100	yes	41	98.8	512	1179793	1179915	1178670	1180208
CP014492.1|Escherichia coli strain MVAST0167	100	yes	41	98.8	512	4482934	4483059	4481814	4483349
CU928158.2|Escherichia fergusonii ATCC 35469	61	yes	41	79.9	509	1949134	1949256	1948020	1949546
AP014855.1|Escherichia albertii NIAH_Bird_3	75.6	yes	41	90.8	512	3678930	3679055	3678640	3680178
CP014994.1|Salmonella enterica subsp. enterica serovar Tennessee strain CFSAN001387	12.7	no	43	59	514	1305167	1305286	1304040	1305578
CP007373.2|Salmonella enterica subsp. enterica serovar Enteritidis str. EC20121744	12.7	no	43	58.7	514	2697686	2697814	2697394	2698935
CP016014.1|Salmonella enterica subsp. enterica serovar Newport strain CFSAN003387	14.5	no	43	58.8	513	1331537	1331673	1330422	1331963
CP017928.1|Klebsiella oxytoca strain CAV1015	0	no	43	68.8	511	6114854	6114982	6114562	6116094
CP004887.1|Klebsiella michiganensis HKOPL1	8.3	no	45	68.8	510	5098135	5098263	5097026	5098555
CP017802.1|Raoultella ornithinolytica strain MG	20.7	no	41	67.1	510	1524958	1525088	1523848	1525377
CP014748.1|Enterobacter aerogenes strain FDAARGOS_139	0	no	40	70	508	3329043	3329162	3328735	3330258
CP006722.1|Klebsiella pneumoniae subsp. pneumoniae 1158	10.7	yes	48	65.8	529	3985391	3985534	3984240	3985826
CP001891.1|Klebsiella variicola At-22	0	no	41	64.8	526	4090268	4090390	4089105	4090682
CP001560.1|Shimwellia blattae DSM 4481 = NBRC 105725	0	no	48	55.8	515	3264164	3264298	3263382	3264917
CP009539.1|Yersinia ruckeri strain YRB	0	no	42	57.9	515	707391	707519	706279	707811
CP013913.1|Serratia fonticola strain GS2	0	no	42	58.4	516	5307965	5308090	5306847	5308382
CP010584.1|Serratia marcescens strain AS1	8.1	yes	40	56.8	517	3194910	3195029	859056	860708
CP016044.1|Edwardsiella piscicida strain S11-285	-	no	-	49.8	550	-	-	859056	860708
CP015379.1|Hafnia alvei strain HUMV-5920	0	no	45	52.2	535	2539646	2539780	2539354	2540940
CP014608.1|Obesumbacterium proteus strain DSM 2777	0	no	45	52	535	1712106	1712234	1711814	1713400
LT575468.1|Plesiomonas shigelloides strain NCTC10360	-	no	-	38.3	536	-	-	2712357	2713967
CP013248.1|Vibrio parahaemolyticus strain FORC_022	-	no	-	25.4	522	-	-	3187647	3189215
CP013067.1|Aeromonas schubertii strain WL1483	-	no	-	26.6	504	-	-	3027838	3029352
CP011340.1|Streptomyces pristinaespiralis strain HCCB 10218	-	no	-	32.4	230	-	-	1784111	1784803

Supplementary table S2.2: Localization and features of homologues used for the phylostratigraphic analysis of *ano*. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to *ano*	full-length sORF homologue	length sORF (aa)	Identity mORF to ECs2385	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC Sakai	-	-	62	-	334	2357439	2357741	2357572	2358573
CP003109.1|Escherichia coli O55:H7 str. RM12579	100	yes	62	99.7	334	2109364	2109549	2109377	2110381
CP006736.1|Shigella dysenteriae 1617	98.4	yes	62	99.7	334	1700323	1700508	1700336	1701340
AP009240.1|Escherichia coli SE11	91.9	yes	62	99.4	334	1891202	1891387	1891215	1892219
CP006262.1|Escherichia coli O145:H28 str. RM13516	98.4	yes	62	100	334	2067897	2068082	2067910	2068914
CP000036.1|Shigella boydii Sb227	96.8	yes	62	98.8	334	1430402	1430587	1429570	1430574
CP013662.1|Escherichia coli strain 08-00022	96.8	yes	62	99.7	334	1992641	1992826	1992654	1993658
CP010117.1|Escherichia coli strain C2	93.5	yes	62	99.1	334	983040	983225	982208	983212
CP010176.1|Escherichia coli strain H10	91.9	yes	62	99.1	334	1635049	1635234	1635062	1636066
CP015229.1|Escherichia coli strain 06-00048	90.3	yes	62	98.8	334	2269174	2269359	2269187	2270191
CP006632.1|Escherichia coli PCN033	91.9	yes	62	99.4	334	1852987	1853172	1853000	1854004
AP009378.1|Escherichia coli SE15	95.2	yes	62	99.1	335	1654413	1654598	1654426	1655430
CP006830.1|Escherichia coli APEC O18	95.2	yes	62	99.4	334	663713	663898	663726	664730
CU928158.2|Escherichia fergusonii ATCC 35469	60.3	no	59	93.7	333	1411625	1411801	1410793	1411794
AP014855.1|Escherichia albertii strain NIAH_Bird_3	53.2	yes	62	94.6	334	1608608	1608793	1608621	1609625
AP014857.1|Escherichia albertii strain EC06-170	51.6	yes	62	94.3	334	1692354	1692539	1692367	1693371
FN543502.1|Citrobacter rodentium ICC168	2.7	no	59	84.7	329	1473940	1474116	1473110	1474081
CP006053.1|Salmonella enterica subsp. enterica serovar Bareilly str. CFSAN000189	0.8	no	63	85.3	333	2570875	2571057	2570891	2571889
FR877557.1|Salmonella bongori NCTC 12419	12.6	no	62	84.4	333	1329764	1329949	1328932	1329930
CP011602.1|Kluyvera intermedia strain CAV1151	20	no	65	69.7	338	1456616	1456813	1455788	1456747
CP011657.1|Citrobacter freundii strain CAV1741	37.9	no	64	86.5	334	95692	95883	94860	95861
CP015774.1|Lelliottia amnigena strain ZB04	24.4	no	65	70	337	2001574	2001768	2000742	2001701
CP009756.1|Enterobacter cloacae strain GGT036	23.7	yes	66	69.8	337	1978149	1978343	1977315	1978274
CP010376.2|Enterobacter hormaechei subsp. steigerwaltii strain 34977	2.6	yes	65	71.4	337	2039297	2039479	2038460	2039359
CP002272.1|Enterobacter lignolyticus SCF1	8.7	no	66	76	333	2251722	2251925	2250894	2251859
CP010392.1|Klebsiella pneumoniae strain 34618, complete genome	10.6	no	56	71.3	328	2153876	2154043	2153042	2154013
CP004142.1|Raoultella ornithinolytica B6	5.8	yes	63	71.6	332	968753	968941	967920	968915
CP011636.1|Klebsiella oxytoca strain CAV1374	4.7	no	56	71.2	328	2933784	2933951	2932950	2933921
CP012266.1|Cronobacter dublinensis subsp. dublinensis LMG 23823	15.3	no	59	56.7	344	1900018	1900197	1899190	1900131
CP011047.1|Cronobacter sakazakii strain ATCC 29544	5.9	yes	58	58	344	3501388	3501498	3501470	3502366
CP012257.1|Cronobacter universalis NCTC 9529	0.9	no	58	57.1	344	1874662	1874793	1873815	1874702
CP011077.1|Klebsiella michiganensis strain RC10	14	no	63	79.6	332	1933926	1934114	1933095	1934090
CP009451.1|Cedecea neteri strain SSMD04	1.7	no	63	79.6	332	755104	755292	755128	756123
AP012032.2|Pantoea ananatis AJ13355 DNA	6.5	no	59	50.7	326	1280836	1280994	1280025	1280903
CP003942.1|Serratia marcescens FGI94	4.5	yes	65	54.1	338	2153846	2154046	2153012	2153968
CP001836.1|Dickeya dadantii Ech586	21.1	no	63	49.6	334	1873801	1874004	1873219	1874190
FM162591.1|Photorhabdus asymbiotica ATCC43949	4.8	no	52	47	358	2150793	2150951	2149965	2150879
CP011078.1|Yersinia ruckeri strain Big Creek 74	5.3	no	68	52.3	346	524541	524741	524572	525582
CP006792.1|Yersinia pestis 2944	3.6	no	67	50.4	354	844253	844453	843413	844420
CP016935.1|Yersinia enterocolitica strain YE7	1.8	no	59	50.4	355	2246146	2246322	2245308	2246237

Supplementary table S2.3: Localization and features of homologues used for the phylostratigraphic analysis of *slyC*. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to *slyC*	full-length sORF homologue	length sORF (aa)	Identity mORF to ECs2351	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
AE014075.1|Escherichia coli CFT073	95.3	yes	64	97.3	146	1880926	1881117	1880919	1881359
EHEC Sakai	-	-	64	-	144	2320260	2320451	2320256	2320693
CP017251.1|Escherichia coli strain NADC 5570/86-24/6564	98.4	yes	64	97.9	146	3089907	3090098	3089903	3090334
CP010116.1|Escherichia coli strain C1	96.9	yes	64	99.3	144	3010913	3011104	3010906	3011340
CP000036.1|Shigella boydii Sb227	96.9	yes	64	97.9	146	1470224	1470415	1469982	1470422
CP001063.1|Shigella boydii CDC 3083-94	95.3	yes	64	98.6	144	1677708	1677899	1677701	1678135
CT928158.2|Escherichia fergusonii ATCC 35469	67.2	yes	64	87.7	146	1441324	1441500	1441082	1441522
AP014855.1|Escherichia albertii strain NIAH_Bird_3	87.5	yes	64	97.2	144	1574496	1574687	1574489	1574923
CP006692.1|Salmonella bongori serovar 48:z41:-- str. RKS3044	70.3	no	63	88.4	147	1302980	1303168	1302738	1303178
CP000880.1|Salmonella enterica subsp. arizonae serovar 62:z4,z23	68.8	no	63	87.7	146	1496273	1496461	1496263	1496703
CP011396.1|Salmonella enterica subsp. enterica serovar Thompson str. ATCC 8391	75	no	63	89.6	145	2419306	2419494	2419296	2419730
CP007267.2|Salmonella enterica subsp. enterica serovar Enteritidis str. EC20120005	73.4	no	63	90.3	144	1708607	1708792	1708594	1709028
CP014994.1|Salmonella enterica subsp. enterica serovar Tennessee strain CFSAN001387	73.4	no	63	88.9	145	2590516	2590704	2590506	2590940
FN543502.1|Citrobacter rodentium ICC168	68.8	yes	61	90.3	145	1525108	1525284	1524872	1525306
CP000822.1|Citrobacter koseri ATCC BAA-895	76.6	yes	63	90.4	146	1592605	1592802	1592583	1593023
CP004887.1|Klebsiella michiganensis HKOPL1	65.6	yes	63	86.8	145	3268459	3268716	3268307	3268741
CP003683.1|Klebsiella michiganensis E718	64.1	yes	63	78.5	132	3463341	3463598	3511235	3511669
CP017928.1|Klebsiella oxytoca strain CAV1015	64.1	yes	63	79.9	133	1754360	1754545	1754335	1754736
CP004142.1|Raoultella ornithinolytica B6	60.9	yes	61	85.4	145	1144424	1144618	1144209	1144643
CP010557.1|Raoultella ornithinolytica strain S12	64.1	yes	63	86.1	144	3511260	3511463	3511235	3511669
CP012252.1|Klebsiella variicola strain HKUOPLA	64.1	no	63	84.9	146	554926	555120	554901	555341
CP009863.1|Klebsiella pneumoniae subsp. pneumoniae strain KPNIH29	64.1	yes	63	84.9	146	2978752	2978946	2978727	2979167
CP009114.1|Klebsiella pneumoniae strain blaNDM-1	65.6	yes	63	84.9	146	1201645	1201839	1201620	1202060
CP012987.1|Klebsiella pneumoniae strain KpN01	64.1	yes	64	84.9	146	5087698	5087892	5087477	5087917
CP014070.1|Citrobacter amalonaticus strain FDAARGOS_165	73.4	yes	63	88.9	145	3280094	3280270	3280072	3280506
CP014015.1|Citrobacter amalonaticus strain FDAARGOS_122	75	yes	63	89.6	144	4205021	4205197	4204785	4205219
CP011132.1|Citrobacter amalonaticus Y19	67.2	yes	63	88.2	144	1413575	1413757	1413339	1413773
CP007557.1|Citrobacter freundii CFNIH1	76.6	yes	63	91	144	3575296	3575556	3575144	3575578
CP011612.1|Citrobacter freundii strain CAV1321	75	yes	63	89.7	146	842616	842876	842594	843034
CP013990.1|Leclercia adecarboxylata strain TSDA-ARS-TSMARC-60222	67.2	yes	63	84.9	146	2706072	2706341	2706047	2706487
CP015774.1|Lelliottia amnigena strain ZB04	65.6	yes	63	87.7	146	2032368	2032544	2032126	2032566
CP003026.1|Enterobacter asburiae LF7a	68.8	yes	63	87.5	144	1969441	1969647	1969235	1969669
CP017179.1|Enterobacter hormaechei subsp. steigerwaltii strain DSM 16691	65.6	yes	63	85.6	147	1966789	1967058	1966643	1967083
CP008905.1|Enterobacter cloacae ECR091	66.2	yes	63	86.3	146	1975081	1975266	1974839	1975279
CP003678.1|Enterobacter cloacae subsp. dissolvens SDM	59.4	yes	58	86.2	146	2083904	2084188	2083779	2084216
CP017279.1|Enterobacter ludwigii strain EN-119	61.5	yes	65	84.9	146	1227378	1227635	1227220	1227660
CP010512.1|Enterobacter cloacae strain colR/S	58.5	yes	64	84.9	147	3771798	3772055	3771773	3772213
CP015227.1|Enterobacter sp. ODB01	63.1	yes	65	85.6	146	2996515	2996718	2996303	2996743
CP003737.1|Enterobacter cloacae subsp. cloacae ENHKT01	64.6	yes	65	87.5	144	2055357	2055560	2055151	2055585
CP014007.1|Kosakonia oryzae strain Ola 51	64.6	yes	65	84.9	147	3007063	3007320	3007038	3007478

LN907827.1|Erwinia sp. EM595	-	-	-	79.6	147	-	-	1832236	1832679
CP002206.1|Pantoea vagans C9-1	-	-	-	77,9	145	-	-	1253504	1253941
CP011254.1|Serratia fonticola strain DSM 4576	-	-	-	81,2	144	-	-	4568660	4569094
CP012097.1|Serratia plymuthica strain 3Re4-18	-	-	-	77,8	144	-	-	2335741	2336175
CP008955.1|Yersinia kristensenii strain ATCC 33639	-	-	-	78,5	143	-	-	1287746	1288177
CP003403.1|Rahnella aquatilis HX2	-	-	-	77,8	144	-	-	3045833	3046267
CP004345.1|Morganella morganii subsp. morganii KT	-	-	-	71,5	144	-	-	2620961	2621395
AP008232.1|Sodalis glossinidius str. 'morsitans' DNA	-	-	-	79,9	144	-	-	2418212	2418646
CP015243.1|Halotalea alkalilenta strain IHB B 13600	-	-	-	46,8	158	-	-	3168030	3168506
CP018139.1|Halolamina sediminis strain Hb3	-	-	-	47,6	147	-	-	587471	587914
CP000285.1|Chromohalobacter salexigens DSM 3043	-	-	-	46,3	149	-	-	2686893	2687342
CP004388.1|Thalassospira xiamenensis M-5 = DSM 17429	-	-	-	37,1	166	-	-	1445296	1445796
CP010586.1|Bacillus megaterium strain Q3	-	-	-	29,3	139	-	-	3531324	3531743

Supplementary table S2.4: Localization and features of homologues used for the phylostratigraphic analysis of *asa*. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to *asa* [%]	full-length sORF homologue	length sORF [aa]	Identity mORF to #1238	length mORF [aa]	genome localization sORF		genome localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	87	-	219	1247671	1247934	1247562	1248221
CP010134.1|Escherichia_coli_strain_D1	97.7	yes	87	99.5	219	797511	797771	797221	797880
CP010133.1|Escherichia_coli_strain_C11	98.9	yes	87	99.5	219	536629	536889	536339	536998
CP010829.1|Shigella_sonnei_strain_FORC_011	96.6	yes	86	99.1	219	1051824	1052081	1051712	1052371
CP001063.1|Shigella_boydii_CDC_3083-94	97.7	yes	86	99.1	219	2135202	2135459	2134912	2135571
CP000266.1|Shigella_flexneri_5_str._8401	96.6	yes	87	99.1	219	1027432	1027689	1027320	1027979
AP014855.1|Escherichia_albertii_DNA	92	yes	87	99.5	219	1019141	1019404	1019032	1019691
AM933173.1|Salmonella_enterica_subsp._enterica_serovar_Gallinarum_str._287/91	81.8	no	87	84	192	1061518	1061778	1061409	1061987
FR877557.1|Salmonella_bongori_NCTC_12419	80.7	no	87	95	219	1006073	1006333	1005964	1006623
FN543502.1|Citrobacter_rodentium_ICC168	87.5	yes	87	97.7	219	1142210	1142416	1142101	1142760
CP011602.1|Kluyvera_intermedia_strain_CAV1151	80.7	yes	96	94.1	219	2676565	2676855	2676305	2676964
CP014007.1|Kosakonia_oryzae_strain_Ola_51	81.8	no	87	94.5	219	3428447	3428650	3428100	3428759
CP003938.1|Enterobacteriaceae_bacterium_strain_FGI_57	85.2	no	87	95.9	219	2902806	2903012	2902462	2903121
CP007557.1|Citrobacter_freundii_CFNIH1	83	yes	87	95.9	219	3122821	3123081	3122712	3123371
CP015774.1|Lelliottia_amnigena_strain_ZB04	83	no	87	95	219	1723909	1724169	1723800	1724459
CP013990.1|Leclercia_adecarboxylata_strain_USDA-ARS-USMARC-60222	83	no	87	94.1	219	3137208	3137423	3136873	3137532
CP011591.1|Enterobacter_asburiae_strain_CAV1043	83	no	87	92.7	219	1480942	1481229	1480679	1481338
CP014993.1|Enterobacter_asburiae_strain_ENIPBJ-CG1	81.8	no	96	92.7	219	1536402	1536689	1536139	1536798
CP010512.1|Enterobacter_cloacae_strain_colR/S	81.8	no	96	93.2	219	4085220	4085507	4084957	4085616
CP017279.1|Enterobacter_ludwigii_strain_EN-119	83	no	96	94.1	219	831209	831496	831100	831759
CP011798.1|Enterobacter_cloacae_strain_UW5	84.1	no	96	93.6	219	2154597	2154857	2154488	2155147
CP002886.1|Enterobacter_cloacae_EcWSU1	81.8	no	87	94.1	219	1631203	1631463	1631094	1631753
CP006580.1|Enterobacter_cloacae_P101	79.5	no	87	94.1	219	3470951	3471211	3470661	3471320
CP011597.1|Klebsiella_oxytoca_strain_CAV1099	83	no	87	94.5	219	1377477	1377737	1377187	1377846

CP010557.1|Raoultella_ornithinolytica_strain_S12	81.8	no	87	92.2	219	2319932	2320192	2319823	2320482
CP011574.1|Enterobacter_aerogenes_strain_CAV1320	77.3	no	87	93.2	219	1505855	1506115	1505565	1506224
CP009450.1|Pluralibacter_gergoviae_strain_FB2	79.5	no	87	95.4	219	4202638	4202898	4202348	4203007
CP009451.1|Cedecea_neteri_strain_SSMD04	83	yes	87	93.2	219	1048791	1049054	1048504	1049163
CP013940.1|Cronobacter_malonaticus_LMG_23826	81.8	no	96	91.3	219	2544341	2544628	2544078	2544737
CP012257.1|Cronobacter_universalis_NCTC_9529	79.5	no	96	90.9	219	1599681	1599968	1599572	1600231
CP001560.1|Shimwellia_blattae_DSM_4481___NBRC_105725	84.1	no	87	94.5	219	2521588	2521848	2521298	2521957
CU468135.1|Erwinia_tasmaniensis_strain_ET1/99	81.6	no	117	78.5	219	2353982	2354332	2353785	2354441
CP002206.1|Pantoea_vagans_C9-1	70.5	no	117	81.3	219	839103	839456	838994	839647
CP002433.1|Pantoea_sp._At-9b	73.9	no	117	84	219	1555300	1555653	1555191	1555847
AP012551.1|Plautia_stali_symbiont_DNA	79.5	no	87	82.7	220	917133	917393	917024	917683
CP001836.1|Dickeya_dadantii_Ech586	64.8	no	96	73.5	219	3368659	3368949	3368399	3369058
CP009678.1|Pectobacterium_carotovorum_subsp._odoriferum_strain_BC_S7	59.1	no	96	76.3	219	2879181	2879471	2878921	2879580
CP006569.1|Sodalis_praecaptivus_strain_HS1	65.9	yes	96	74.9	219	2981455	2981745	2981195	2981854
CP003244.1|Rahnella_aquatilis_CIP_78.65___ATCC_33071	61.4	no	96	76.7	223	1716092	1716382	1715983	1716654
CP007044.2|Chania_multitudinisentens_RB-25	65.9	no	96	78.7	221	3110877	3111167	3110768	3111427
HG326223.1|Serratia_marcescens_subsp._marcescens_Db11	69.3	yes	124	78.1	219	1114946	1115320	1114837	1115496
CP007439.1|Serratia_plymuthica_strain_V4	65.9	no	96	77.2	219	1969527	1969817	1969418	1970077
CP006664.1|Edwardsiella_tarda	59.6	no	96	75.5	220	3267108	3267398	3266999	3267661
CP009706.1|Hafnia_alvei_FB1	59.1	no	117	75.5	220	1863829	1864182	1863720	1864382
CP014608.1|Obesumbacterium_proteus_strain_DSM_2777	59.1	no	117	75.5	220	2932736	2933089	2932627	2933289
CP014056.1|Grimontia_hollisae_strain_ATCC_33564	45.5	no	95	57.9	221	2541398	2541688	2541286	2541951
CP005974.1|Photobacterium_gaetbulicola_Gung47	45.5	no	95	62.9	221	2630456	2630746	2630193	2630858
CP000020.2|Vibrio_fischeri	54.5	no	96	62.7	220	1718481	1718771	1718218	1718880
LN554846.1|Aliivibrio_wodanis	55.7	no	96	62.3	220	1864801	1865091	1864538	1865200
CP018835.1|Vibrio_gazogenes_strain_ATCC_43942	54.5	no	95	60.9	220	147696	147986	147587	148249
CP002377.1|Vibrio_furnissii_NCTC_11218	55.7	no	95	60.5	223	2111414	2111704	2111151	2111813
CP014035.1|Vibrio_fluvialis_strain_ATCC_33809	53.4	no	95	60.5	223	1368996	1369286	1368733	1369395
CP016380.1|Aeromonas_hydrophila_strain_AHNIH1	40.9	no	95	55.7	219	1879468	1879758	1879362	1880018
LN554852.1|Moritella_viscosa_genome_assembly_MVIS1	56.8	yes	96	62.7	220	3827155	3827445	3826895	3827557
CP002209.1|Ferrimonas_balearica_DSM_9799	47.7	no	96	62	221	2043145	2043435	2043039	2043695
CP013251.1|Endozoicomonas_montiporae_CL-33	40.4	no	88	53.3	225	2882997	2883263	2882698	2883369
CP015839.1|Marinobacterium_sp._ST58-10	42	no	96	58.4	221	2142652	2142942	2142543	2143208
CP006985.1|Pseudomonas_aeruginosa_LESlike4	42.1	no	97	52.4	227	2673709	2674002	2673597	2674265
CP014158.1|Pseudomonas_citronellolis_strain_P3B5	41.6	no	97	52.8	233	4248963	4249256	4248703	4249368
FN650140.1|Legionella_longbeachae_NSW150	39.3	no	98	57.1	224	2290064	2290360	2289798	2290472
AE017282.2|Methylococcus_capsulatus_str._Bath	41.5	no	98	53.2	231	1398261	1398560	1397995	1398672
CP000127.1|Nitrosococcus_oceani_ATCC_19707	36.7	no	97	48.2	224	1848684	1848980	1848575	1849243
LN869922.1|Kingella_kingae_genome_assembly_KKKWG1	11.8	no	114	33.3	240	1299264	1299608	1299152	1299838
CP007726.1|Neisseria_elongata_subsp._glycolytica_ATCC_29315	22.4	no	114	33.1	245	278966	279310	278851	279543
CP009418.1|Neisseria_meningitidis_strain_NM3686	28.9	no	114	34.6	237	1248822	1249166	1248595	1249278
AP013066.1|Sulfuricella_denitrificans_skB26	33.7	no	96	49.6	226	2187095	2187388	2186983	2187651
CP012371.1|Nitrosospira_briensis_C-128	26.1	no	86	36.4	228	2505391	2505651	2505276	2505884
AM406670.1|Azoarcus_sp._BH72	34.8	no	87	57.5	221	975277	975543	974987	975652
CP002657.1|Alicycliphilus_denitrificans_K601	24.2	no	164	36.7	229	2968657	2969154	2968580	2969269

CP010951.1|Ramlibacter_tataouinensis_strain_5-10	24	no	163	36.2	229	4344220	4344717	4344105	4344794
AP012320.1|Rubrivivax_gelatinosus_IL144	32.2	no	181	29	224	56918	57460	56803	57330
CP011301.1|Burkholderia_cepacia_strain_LO6	29.2	no	191	33.6	244	2633871	2634446	2633756	2634454
CP010323.1|Bordetella_pertussis_137	27.8	no	198	39	236	2305120	2305713	2305005	2305715
CP003555.1|Advenella_kashmirensis_WT001	33.7	no	89	34.2	240	1727433	1727702	1727318	1728010
AP012342.1|Leptospirillum_ferrooxidans_C2-3	26.1	no	116	32.5	240	185236	185586	185000	185704
CP002514.1|Chloracidobacterium_thermophilum_B	27.7	no	131	27.6	246	116888	117286	116682	117401
XM_005708963.1|Galdieria_sulphurari	24.2	no	121	29.3	270	355	720	41	835
XM_018173156.1|PREDICTED:_Hyalella_azteca	36	no	97	47.2	231	294	587	1	696

Supplementary table S2.5: Localization and features of homologues used for the phylostratigraphic analysis of OGC106. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC106	full-length sORF homologue	length sORF (aa)	Identity mORF to #2699	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	92	-	275	2517026	2517304	2516432	2517259
CP001063.1|Shigella boydii CDC 3083-94	95.7	yes	92	98.9	275	1809398	1809673	1808801	1809628
CP012378.1|Escherichia coli strain MEM	94.6	yes	92	99.3	275	2201190	2201465	2201235	2202062
CU928158.2|Escherichia fergusonii ATCC 35469	58.1	no	88	95.6	275	1364220	1364483	1364253	1365080
AE014613.1|Salmonella enterica subsp. enterica serovar Typhi Ty2	8.5	yes	82	89.8	275	1257979	1258227	1257996	1258823
CP011602.1|Kluyvera intermedia strain CAV1151	44	yes	83	88	275	1402729	1402967	1402731	1403561
CP007557.1|Citrobacter freundii CFNIH1	1.8	no	80	92	276	3426533	3426781	3426547	3427377
CP017279.1|Enterobacter ludwigii strain EN-119	45.7	yes	84	89.1	275	1108753	1109004	1108774	1109601
CP013990.1|Leclercia adecarboxylata strain USDA-ARS-USMARC-60222	35.5	no	79	89.5	275	2812074	2812310	2811480	2812307
CP004142.1|Raoultella ornithinolytica B6	46.3	yes	85	90.9	275	1888339	1888593	1887742	1888569
CP015774.2|Lelliottia amnigena strain ZB04	35.5	yes	77	86.9	275	1918232	1918462	1918232	1919059
CP007215.3|Kosakonia sacchari SP1	47.3	yes	78	85.9	277	2788306	2788539	2787709	2788539
CP012300.1|Klebsiella pneumoniae subsp. pneumoniae strain HKUOPLC	45.7	no	85	90.9	275	3021961	3022215	3021364	3022191
CP009450.1|Pluralibacter gergoviae strain FB2	35.5	yes	82	88	275	3915541	3915786	3914947	3915774
CP011047.1|Cronobacter sakazakii strain ATCC 29544	39.4	yes	82	83.3	275	3560865	3561110	3560283	3561110
CP009459.1|Cedecea neteri strain ND14a	42.7	no	90	86.5	275	2870428	2870697	2870464	2871291
CP001560.1|Shimwellia blattae DSM 4481 = NBRC 105725	39.8	yes	78	82.5	275	1637732	1637965	2115378	2116205
CP002505.1|Rahnella sp. Y9602	37.6	yes	79	77.9	276	2193233	2193469	2193236	2194066
CP014137.1|Brenneria goodwinii strain FRB141	40.9	yes	78	78.9	275	1905465	1905704	1904877	1905701
CP015750.1|Pectobacterium wasabiae CFBP 3304	38.8	yes	91	78.5	275	4029125	4029397	4029170	4029994
FP236843.1|Erwinia billingiae strain Eb661	6	no	79	80.4	276	2115367	2115609	2115378	2116205
AP012551.1|Plautia stali symbiont	4.1	no	85	79.3	276	982408	982662	982431	983258
CP015581.1|Tatumella citrea strain ATCC 39140	11.4	no	90	78.3	277	1792968	1793240	1793006	1793836
CP003094.1|Lactobacillus rhamnosus ATCC 8530	35.5	**yes**	76	64.7	272	1780069	1780296	1780069	1780887

Supplementary table S2.6: Localization and features of homologues used for the phylostratigraphic analysis of OGC15. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC15	full-length sORF homologue	length sORF (aa)	Identity mORF to #0277	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	44	-	324	300575	300709	300073	301047
CP018239.1|Escherichia coli strain 272	97.7	yes	44	99	324	303142	303276	302640	303614
CP020086.1|Shigella flexneri 1a strain 0670	97.7	yes	44	93.6	312	4108048	4108182	4107746	4108684
HG738867.1|Escherichia coli str. K-12 substr. MC4100	63.6	longer	76	55	387	467300	467533	466798	467961
CP016018.1|Escherichia coli strain ER1821R	63.6	longer	76	65.7	305	3293766	3293999	3293584	3294501
CU928158.2|Escherichia fergusonii ATCC 35469	75	longer	106	59.2	387	591143	591463	590641	591804
CP007639.1|Salmonella enterica subsp. enterica serovar Choleraesuis strain C500	81.8	no	44	75.4	387	372839	372973	372337	373500
CP014070.2|Citrobacter amalonaticus strain FDAARGOS_165	75	longer	106	57.2	387	4208846	4209166	4208505	4209668
CP026231.1|Citrobacter freundii complex sp. CFNIH4	54.5	longer	180	51.3	387	341533	342075	341028	342191
CP011602.1|Kluyvera intermedia strain CAV1151	70.5	longer	99	54.7	387	3317167	3317467	3316666	3317829
CP027604.1|Enterobacter cloacae strain AR_0093	79.5	no	44	81.1	331	1854431	1854565	1853929	1854924
CP021539.1|Klebsiella pneumoniae strain AR_0047	52.3	longer	106	50.5	377	1736670	1736981	1736165	1737298
CP013338.1|Raoultella ornithinolytica strain Yangling I2	56.8	longer	72	49.2	387	4121628	4121846	4121188	4122351
CP000783.1|Cronobacter sakazakii ATCC BAA-894	59.1	no	33	68	387	3024780	3024878	3024217	3025380
CP009454.1|Pantoea rwandensis strain ND04	45.5	no	27	55.1	385	3633452	3633535	3632880	3634043
CP019440.1|Edwardsiella piscicida strain ETW41	79.5	longer	117	64.1	382	850345	850698	850037	851185
CP006569.1|Sodalis praecaptivus strain HS1	43.2	no	27	65.3	386	3371606	3371689	3371034	3372194
CP023505.1|Morganella morganii strain FDAARGOS_365	63.8	longer	115	54.7	387	810463	810810	809955	811118
CP025084.1|Serratia sp. ATCC 39006	72.7	longer	117	57.4	384	1678356	1678709	1677863	1679017
CP009367.1|Yersinia enterocolitica strain WA	59.1	no	43	54.2	388	2575250	2575381	2574723	2575889
CP014608.1|Obesumbacterium proteus strain DSM 2777	61.4	longer	99	55.3	388	4328818	4329117	4328456	4329622
CP010423.1|Pragia fontium strain 24613	59.1	no	82	54.7	385	1479572	1479820	1479070	1480227
FM162591.1|Photorhabdus asymbiotica ATCC43949	68.2	no	44	56.8	388	4172616	4172750	4172114	4173277
FO704551.1|Xenorhabdus poinarii str. G6	54.5	no	32	55.5	385	319260	319358	318764	319921
CP017671.1|Providencia rettgeri strain RB151	59.1	no	43	55	385	3507552	3507650	3506989	3508146
CP017082.1|Proteus mirabilis strain T21	47.7	no	32	54.2	385	2248670	2248768	2248174	2249331
CP017613.1|Candidatus Hamiltonella defensa strain ZA17	65.9	longer	84	50.4	390	1264526	1264780	1264119	1265291
CP012959.1|Aggregatibacter actinomycetemcomitans strain 624	44.4	no	45	34.9	348	1806559	1806693	1806108	1807154
CP004753.2|Mannheimia haemolytica USDA-ARS-USMARC-185	38.8	longer	60	26.4	351	865407	865589	864991	866046
CP002738.1|Methylomonas methanica MC09	31.8	yes	41	31.5	343	2116726	2116845	2116272	2117303
FO082060.1|Methylomicrobium alcaliphilum str. 20Z	23.4	no	71	32.3	233	1195409	1195691	1195083	1195784
CP007031.1|Marichromatium purpuratum 984	29.5	longer	52	35.1	359	1210442	1210597	1209946	1211025
CP027704.1|Acinetobacter baumannii strain DS002	19.3	no	45	35	346	2829570	2829704	2829113	2830150
CP026328.1|Acidithiobacillus caldus strain MTH-04	34	yes	45	37.2	281	1206596	1206730	1206390	1207232
CP002552.1|Nitrosomonas sp. AL212	38.3	no	45	34.6	337	198613	198747	198167	199180
CP002056.1|Methylotenera versatilis 301	34.1	longer	114	35.9	340	2429195	2429533	2428762	2429781
CP022278.1|Neisseria sp. 10023	33.3	longer	75	33.7	377	1304958	1305173	1304402	1305535
LT906482.1|Eikenella corrodens strain NCTC10596	32.6	longer	83	36.2	352	441613	441867	441135	442193
CP028519.1|Microvirgula aerodenitrificans strain BE2.4	40.9	longer	72	42.4	361	3959779	3959994	3959372	3960457
CP000089.1|Dechloromonas aromatica RCB	37.7	longer	73	33.2	338	1138303	1138518	1137870	1138886

CP016172.1|Bordetella flabilis strain AU10664	25.5	no	46	29.8	355	1466024	1466161	1465564	1466631
CP017749.1|Cupriavidus sp. USMAA2-4	36.4	longer	98	34.7	346	2772805	2773098	2772351	2773388
CP004012.1|Ralstonia solanacearum FQY_4	35.3	yes	45	37	360	3268537	3268798	3268161	3269243
CP007506.3|Pandoraea pnomenusa strain RB38	40	longer	73	35.2	350	1432158	1432376	1431707	1432759
CP000614.1|Burkholderia vietnamiensis G4	38.3	yes	45	35.8	310	209885	210019	209398	210327
CP022989.1|Paraburkholderia aromaticivorans strain BN5	45.5	no	44	31.2	336	1116285	1116416	1115861	1116871
FN650140.1|Legionella longbeachae NSW150	29.5	yes	44	36.3	334	2149593	2149724	2149172	2150173
CP019434.1|Acidihalobacter ferrooxidans strain V8	36.4	longer	72	30.7	296	2430906	2431121	2430754	2431644
AP014879.1|Sulfuricaulis limicola	33.3	longer	59	31.6	349	1103022	1103198	1102609	1103655
CP009823.1|Xylella fastidiosa strain J1a12	34	no	46	34.4	276	1196461	1196598	1196168	1196998
CP020987.1|Xanthomonas citri pv. phaseoli var. fuscans strain CFBP6994R	26.5	no	48	35.3	376	442521	442797	442386	443516
CP013342.1|Sphingopyxis terrae NBRC 15098 strain 203-1	36.7	no	47	33.3	367	3871553	3871690	3871033	3872133
CP000633.1|Agrobacterium vitis S4	5.4	no	44	28	343	611191	611322	610760	611791
AP012342.1|Leptospirillum ferrooxidans C2-3	40.9	no	43	35.9	292	517189	517324	516697	517572
FR872582.1|Simkania negevensis Z	29.4	no	51	29.3	369	1296188	1296340	1295746	1296855

Supplementary table S2.7: Localization and features of homologues used for the phylostratigraphic analysis of OGC23. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC23	full-length sORF homologue	length sORF (aa)	Identity mORF to #0555	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	67	-	406	570374	570574	569266	570486
CP016755.1|Escherichia coli strain FORC_044	100	yes	67	100	406	4123438	4123665	4123526	4124746
CP010829.1|Shigella sonnei strain FORC_011	95.5	yes	67	92.6	380	499993	500220	498990	500132
CU928158.2|Escherichia fergusonii ATCC 35469	58.2	no	63	96.8	402	532859	533074	531778	532986
AP014855.1|Escherichia albertii NIAH_Bird_3	79.1	yes	62	96.6	401	471581	471793	470500	471705
CP006692.1|Salmonella bongori serovar 48:z41:-- str. RKS3044	45.8	no	59	93.8	406	477037	477240	475956	477176
CP011289.1|Salmonella enterica subsp. diarizonae strain 11-01853	45.8	no	59	93.8	406	1444523	1444726	1443442	1444662
CP014015.2|Citrobacter amalonaticus strain FDAARGOS_122	52.6	yes	72	98.3	406	3192664	3192906	3191509	3192729
CP011602.1|Kluyvera intermedia strain CAV1151	26.6	no	60	88.2	406	3349892	3350263	3348812	3350032
CP018016.1|Kosakonia radicincitans DSM 16656	36.5	yes	58	86.3	399	4262667	4262867	4262749	4263948
CP015774.2|Lelliottia amnigena strain ZB04	50.7	no	63	91.6	406	1123656	1123871	1122575	1123795
CP013990.1|Leclercia adecarboxylata strain USDA-ARS-USMARC-60222	53.7	yes	58	91.9	401	3668562	3668762	3668638	3669843
CP012871.1|Enterobacter lignolyticus strain G5	54.4	no	63	92.6	406	1200736	1200952	1199656	1200876
CP004887.1|Klebsiella michiganensis HKOPL1	36.8	no	71	90.6	406	5148732	5148971	5148832	5150052
CP009450.1|Pluralibacter gergoviae strain FB2	26.8	no	60	89.7	406	4877403	4877609	4877470	4878690
CP009451.1|Cedecea neteri strain SSMD04	37.3	no	70	88.9	406	3867939	3868175	3866858	3868075
CP001560.1|Shimwellia blattae DSM 4481 = NBRC 105725	41.8	no	65	81.8	428	293809	294031	2927525	2928811
FP236843.1|Erwinia billingiae strain Eb661	33.8	yes	63	79.1	400	1228805	1229020	1227727	1228929
CP011427.1|Pantoea vagans strain ND02	21.6	no	57	74.5	398	2938194	2938392	2937117	2938313
CP016032.1|Serratia marcescens strain U36365	17.5	no	61	76.2	406	1115263	1115473	1114177	1115397
CP014136.1|Gibbsiella quercinecans strain FRB97	21.2	no	61	72.8	406	1365874	1366083	1365950	1367170
CP002505.1|Rahnella sp. Y9602	22.5	no	60	74	403	3571841	3572047	3571917	3573128
CP009706.1|Hafnia alvei FB1	17	no	84	75.1	407	3546811	3547089	3546956	3548179

CP014608.1|Obesumbacterium proteus strain DSM 2777	20.6	no	84	75.1	407	2513968	2514246	2512878	2514101
CP009367.1|Yersinia enterocolitica strain WA	26.9	no	91	74.2	410	2596926	2597147	2596996	2598228
FM162591.1|Photorhabdus asymbiotica ATCC43949	19.8	yes	69	73.2	405	1124797	1125030	1123713	1124930
CP017054.1|Providencia stuartii strain BE2467	28.6	no	57	72.9	407	1676556	1676753	1676611	1677834

Supplementary table S2.8: Localization and features of homologues used for the phylostratigraphic analysis of OGC51. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC51	full-length sORF homologue	length sORF (aa)	Identity mORF to #1089	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	65	-	37	1110685	1110682	1110833	1110946
CP016755.1|Escherichia coli strain FORC_044	100	yes	65	100	37	3268327	3268524	3268263	3268373
CP018243.1|Escherichia coli strain 350	92.3	yes	60	100	37	1376044	1376229	1376183	1376296
CP017669.1|Escherichia coli strain PA20	100	yes	65	100	37	2745120	2745317	2745271	2745384
CP015020.1|Escherichia coli strain 28RC1	100	yes	59	100	37	5129696	5129893	5129847	5129960
CP012735.1|Shigella flexneri 1a strain 0228	100	yes	65	97.3	37	1128562	1128759	1128713	1128826
CP026810.1|Shigella boydii strain 54-1621	98.5	yes	65	97.3	37	4492084	4492278	4492020	4492130
CP009106.2|Escherichia coli strain 94-3024	98.5	yes	65	60.4	48	4045748	4045945	4045648	4045794
CP009050.1|Escherichia coli NCCP15648	100	yes	65	40	90	1271357	1271554	1271349	1271621
AP010958.1|Escherichia coli O103:H2 str. 12009	72.3	yes	65	38.3	343	4659079	4659276	4658565	4659596
FM180568.1|Escherichia coli 0127:H6 E2348/69 strain E2348/69	52.2	yes	62	25.2	116	1149785	1149973	1149484	1149831

Supplementary table S2.9: Localization and features of homologues used for the phylostratigraphic analysis of OGC57. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC57	full-length sORF homologue	length sORF (aa)	Identity mORF to #1224	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	266	-	354	1235822	1236622	1235816	1236880
AM946981.2|Escherichia coli BL21(DE3)	100	yes	266	97.7	346	1024102	1024902	1024096	1025136
CP001063.1|Shigella boydii CDC 3083-94	100	yes	266	97.7	346	2147288	2148088	2147054	2148094
CP022120.1|Salmonella bongori serovar 66:z41:- str. SA19983605	72.1	no	269	91.6	349	2738876	2739685	2738642	2739691
CP014015.2|Citrobacter amalonaticus strain FDAARGOS_122	76.5	yes	270	91.6	351	3749450	3750262	3749444	3750496
CP011602.1|Kluyvera intermedia strain CAV1151	71.3	yes	272	85.8	352	2689443	2690261	2689209	2690267
CP009854.1|Enterobacter cloacae strain ECNIH5	69.8	yes	266	86.6	358	1637303	1638103	1637297	1638373
CP009450.1|Pluralibacter gergoviae strain FB2	66.9	yes	272	84.8	352	4213144	4213962	4212910	4213968
CP022348.1|Klebsiella michiganensis strain K516	50.4	no	197	84.3	356	5778868	5779377	5778862	5779932
CP019899.1|Raoultella planticola strain GODA	56	no	197	84.1	356	3636356	3636946	3671491	3672561
CP015774.2|Lelliottia amnigena strain ZB04	72	yes	266	87.4	358	1712053	1712853	1712047	1713123

CP012266.1|Cronobacter dublinensis subsp. dublinensis LMG 23823	71.5	yes	267	89	355	1632684	1633487	1632678	1633745
CP019445.1|Kosakonia cowanii strain 888-76	47.8	no	162	86.6	349	2903800	2904285	2903794	2904843
CP009458.1|Cedecea neteri strain M006	68.9	yes	271	83.6	358	2741102	2741917	2741096	2742172
CP001560.1|Shimwellia blattae DSM 4481 = NBRC 105725	66.3	yes	270	84	355	2534723	2535535	2534474	2535541
CU468135.1|Erwinia tasmaniensis strain ET1/99	47	no	222	80.2	355	2365712	2366380	2365319	2366386
AP012551.1|Plautia stali symbiont	62.4	no	270	79.1	356	905506	906345	905500	906570
CP015581.1|Tatumella citrea strain ATCC 39140	54.3	no	247	77.6	351	1490896	1491639	1490890	1491945
CP017482.1|Pectobacterium polaris strain NIBIO1392	48.3	no	299	72.5	361	1852656	1853555	1852476	1853561
CP014137.1|Brenneria goodwinii strain FRB141	50.9	no	262	70.1	358	1173843	1174631	1173837	1174913
CP015137.1|Dickeya solani IPO 2222	67.4	yes	272	74.7	354	705969	706784	705726	706790
AP008232.1|Sodalis glossinidius str. 'morsitans'	53.3	yes	286	76	356	1702512	1703372	1702506	1703576
CP010584.1|Serratia marcescens strain AS1	64.7	yes	349	77.6	351	4059432	4060484	4059411	4060490
CP014136.1|Gibbsiella quercinecans strain FRB97	61.9	no	348	78	358	4580831	4581875	4580810	4581886
CP007044.2|Chania multitudinisentens RB-25	42	no	217	70.8	360	3099254	3099910	3099251	3100333
CP011359.1|Edwardsiella tarda strain FL95-01	53.1	yes	341	70.5	351	2285532	2286557	2285511	2286563
LT575468.1|Plesiomonas shigelloides strain NCTC10360	52.6	no	266	53.5	354	2133143	2133871	2132858	2133922
CP010423.1|Pragia fontium strain 24613	52.9	no	244	66.7	359	2525310	2525993	2525000	2526076
CP009706.1|Hafnia alvei FB1	58.7	yes	270	75.3	351	1851659	1852435	1851653	1852708
CP014608.1|Obesumbacterium proteus strain DSM 2777	60	yes	269	75.6	350	2920407	2921177	2920401	2921453
CP003244.1|Rahnella aquatilis CIP 78.65 = ATCC 33071	61.9	no	349	79	358	1703849	1704898	1703843	1704919
CP006751.1|Yersinia pestis 3770	41.3	no	230	74.2	354	3044678	3045370	3044672	3045733
CP004345.1|Morganella morganii subsp. morganii KT	46.6	no	270	65.5	357	1789348	1790160	1789339	1790412
FO704550.1|Xenorhabdus doucetiae str. FRM16	42.8	yes	285	66	374	1613831	1614688	1613819	1614943
CP011104.1|Photorhabdus temperata subsp. thracensis strain DSM 15199	45.9	no	255	66.3	367	5060071	5060802	5059746	5060849
CP021550.1|Proteus mirabilis strain AR_0159	46.3	no	276	67.3	362	858559	859389	858310	859398
CP017671.1|Providencia rettgeri strain RB151	39.9	no	249	64.1	356	1544876	1545625	2132858	2133922
CP002607.1|Aeromonas veronii B565	31.4	no	266	51.8	354	613326	614123	613155	614117
CP007445.1|Gilliamella apicola strain wkB1	20.7	no	248	45.4	368	242376	243023	242283	243386
CP012067.1|Aggregatibacter aphrophilus strain W10433	22.6	no	243	47.3	367	1745490	1746194	1745487	1746587
KJ621078.1|Avibacterium paragallinarum strain VRDC/AvpgCB/SZ haemagglutinin (HagA) gene, complete cds	23.4	no	240	45.7	345	322	1032	1	1035

Supplementary table S2.10: Localization and features of homologues used for the phylostratigraphic analysis of OGC75. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region (up to the frameshift).

homologue	Identity sORF to OGC75	full-length sORF homologue	length sORF (aa)	Identity mORF to #1870	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	112	-	133	1754037	1754375	1753745	1754146
CP000266.1|Shigella flexneri 5 str. 8401	92.3	frameshift	81	100	72	1227786	1227986	1226194	1226412
CP014099.1|Shigella sonnei strain FDAARGOS_90	93.5	frameshift	80	99.2	133	1603279	1603518	1603368	1603769
CP027118.1|Escherichia coli strain 26561	95.1	frameshift	92	99.2	133	2642351	2642626	2642517	2642918
AM946981.2|Escherichia coli BL21(DE3)	93.5	frameshift	80	99.2	133	1215516	1215791	1215224	1215625
CU928158.2|Escherichia fergusonii ATCC 35469	28.3	no	100	84.2	133	1840380	1840670	1840567	1840968
AP014855.1|Escherichia albertii NIAH_Bird_3	26.9	no	116	85	133	1186841	1187188	1186549	1186950
CP022497.1|Salmonella enterica strain SA20084699	21.6	no	116	81.2	133	2716845	2717192	2716553	2716954
CP011359.1|Edwardsiella tarda strain FL95-01	17.9	no	115	55.2	133	1515824	1516168	1515529	1515933
LT575468.1|Plesiomonas shigelloides strain NCTC10360	15.3	no	99	62.4	133	1983052	1983315	1983206	1983607

Supplementary table S2.11: Localization and features of homologues used for the phylostratigraphic analysis of OGC85. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC85	full-length sORF homologue	length sORF (aa)	Identity mORF to #2135	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	31	-	176	1985820	1985915	1985700	1986230
CP027437.1|Escherichia coli strain 2012C-4221	100	yes	31	98.9	176	573669	573761	573546	574076
AP012306.1|Escherichia coli str. K-12 substr. MDS42	100	yes	31	99.4	176	1295451	1295543	1295136	1295666
CP000036.1|Shigella boydii Sb227	19.5	yes	28	43.3	176	4421655	4421738	4421523	4422053
CP019560.1|Escherichia coli strain KSC1031	100	yes	31	98.9	176	844665	844757	844350	844880
CP010167.1|Escherichia coli strain H3	100	yes	31	99.4	176	2754297	2754389	2753982	2754512
CP015244.1|Escherichia coli O91 str. RM7190	100	yes	31	98.9	176	1733081	1733173	1732766	1733296
CP010117.1|Escherichia coli strain C2	100	yes	31	100	176	1141473	1141565	1141350	1141880
AP017620.1|Escherichia coli MRY15-131	100	yes	31	99.4	176	1683874	1683966	1683751	1684281
CP026831.1|Shigella dysenteriae strain ATCC 12039	100	yes	31	99.4	176	2023674	2023766	2023359	2023889
CU928158.2|Escherichia fergusonii ATCC 35469	3.6	yes	38	29.3	190	3464785	3464898	3464452	3465024
AP014855.1|Escherichia albertii NIAH_Bird_3	14.8	no	51	43.3	176	4489778	4489930	4489715	4490245
CP017727.1|Salmonella enterica subsp. enterica serovar Saintpaul str. SARA26	23.1	yes	32	33.3	174	219257	219352	219134	219658
CP014070.1|Citrobacter amalonaticus strain FDAARGOS_165	31.2	yes	32	48.3	175	1524622	1524717	1524248	1524775
CP011602.1|Kluyvera intermedia strain CAV1151	14.3	yes	47	32.4	183	4695967	4696108	4695605	4696156
CP016337.1|Kosakonia sacchari strain BO-1	2.7	yes	47	34.5	181	1846860	1847072	1846542	1847087
FO203501.1|Klebsiella pneumoniae subsp. rhinoscleromatis strain SB3432	23.1	no	32	46.9	175	1830288	1830383	1829973	1830500
CP011574.1|Enterobacter aerogenes strain CAV1320	35.3	yes	32	45.2	176	4228180	4228275	4227865	4228395
CP009450.1|Pluralibacter gergoviae strain FB2	31.4	no	27	43.6	177	4750356	4750436	4750041	4750574
CP012266.1|Cronobacter dublinensis subsp. dublinensis LMG 23823	11.8	yes	33	30.2	178	1461103	1461201	1460794	1461330
CP009459.1|Cedecea neteri strain ND14a	27.5	no	58	44.3	174	2055252	2055425	2055213	2055737
CP015581.1|Tatumella citrea strain ATCC 39140	3.4	no	32	40.4	183	4375127	4375222	4374989	4375540

LN907827.1|Erwinia sp. EM595	21.6	yes	32	27.9	190	625265	625360	625091	625660
GQ337958.1|Edwardsiella tarda strain TsLt	13.4	yes	67	30.9	179	22	222	1	540
CP019927.2|Serratia marcescens strain 1274	25.6	yes	39	44.1	174	3274070	3274186	3273977	3274501
CP007044.2|Chania multitudinisentens RB-25	21.9	yes	32	43.6	175	1318760	1318855	1318445	1318972
CP014137.1|Brenneria goodwinii strain FRB141	11.7	yes	83	33.9	190	2537151	2537399	2536845	2537414
CP011975.1|Yersinia aleksiciae strain 159	23.1	no	45	32.1	174	1821971	1822105	1821671	1822195
CP027177.1|Morganella morganii strain AR_0057	14.6	yes	45	31.2	185	2340140	2340274	2340023	2340580
FN667741.1|Xenorhabdus bovienii SS-2004	0	no	28	31.7	175	1879256	1879342	1879130	1879654
CP014024.1|Providencia stuartii strain FDAARGOS_145	28.1	no	32	41.4	178	2614429	2614524	2614111	2614647
CP020052.1|Proteus mirabilis strain AR_0059	3.4	no	32	40.4	184	4375309	4375405	4374989	4375540

Supplementary table S2.12: Localization and features of homologues used for the phylostratigraphic analysis of OGC121. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC121	full-length sORF homologue	length sORF (aa)	Identity mORF to #2979	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	63	-	137	2758129	2758320	2757958	2758371
CP001063.1|Shigella boydii CDC 3083-94	100	yes	63	97.8	137	3345418	3345606	852683	853096
CP009685.1|Escherichia coli str. K-12 substr. MG1655	100	yes	63	98.5	137	1660833	1661021	1660782	1661195
AP014855.1|Escherichia albertii NIAH_Bird_3	65.1	yes	60	92.7	137	1871571	1871759	1871397	1871810
CP019414.1|Salmonella enterica subsp. enterica serovar Manchester str. ST278	42.9	yes	61	67.2	137	1030161	1030343	1029984	1030394
CP010557.1|Raoultella ornithinolytica strain S12	31.9	no	87	56.2	137	2325147	2325407	2325048	2325458
CP011597.1|Klebsiella oxytoca strain CAV1099	28.4	no	110	52.7	136	6157872	6158201	6157842	6158252
CP002886.1|Enterobacter cloacae EcWSU1	18.9	no	49	52.9	136	73715	73861	73664	74074
CP015774.2|Lelliottia amnigena strain ZB04	30	yes	93	54.5	136	105433	105711	105382	105792
CP013990.1|Leclercia adecarboxylata strain USDA-ARS-USMARC-60222	23.1	no	120	52.9	136	4719355	4719714	4719355	4719765
CP012264.1|Cronobacter condimenti 1330 strain LMG 26250	33.3	yes	119	62.8	135	4228723	4229079	4228723	4229130
CU468135.1|Erwinia tasmaniensis strain ET1/99	27.9	yes	107	54.6	140	197056	197376	197005	197427
CP002206.1|Pantoea vagans C9-1	35.8	yes	60	48.6	140	1406035	1406214	1405984	1406406
CP013913.1|Serratia fonticola strain GS2	26.9	yes	93	55.8	136	5967505	5967783	82525	82935
CP010423.1|Pragia fontium strain 24613	32.3	no	87	65.7	136	2142348	2142608	2142249	2142659
CP014608.1|Obesumbacterium proteus strain DSM 2777	29.7	no	61	58.3	137	2391746	2391925	2391569	2391982
CP009706.1|Hafnia alvei FB1	32.8	no	60	56.8	137	1356025	1356201	1355845	1356258
CP003171.1|Oceanimonas sp. GK1	27.1	yes	91	48.6	136	1082882	1083124	1082831	1083238
LT670847.1|Halomonas_subglaciescola_strain_ACAM_12	21.9	no	75	49.6	138	2248049	2248273	2247998	2248414
CP000285.1|Chromohalobacter_salexigens_DSM_3043	20.3	yes	117	47.5	135	126238	126516	126187	126594
CP021358.1|Kushneria_marisflavi_strain_SW32	32.8	no	65	52.2	136	1982633	1982827	1982582	1982992
CP015615.1|Acinetobacter_schindleri_strain_ACE	34.3	no	65	52.9	131	1883276	1883470	1883126	1883521
CP010350.1|Acinetobacter_johnsonii_XBB1	23	no	65	50	132	1435134	1435328	1435083	1435481
LT629801.1|Pseudomonas_rhodesiae_strain_BS2777	27.7	longer	65	50	136	5084431	5084625	5084266	5084676
LT897781.1|Marinobacter_sp._es.042	25.3	longer	102	53.2	136	2769002	2769292	2768951	2769361
CP013692.1|Paucibacter_sp._KCTC_42545	23.9	longer	119	46.2	142	110317	110673	110296	110724
CP013729.1|Roseateles_depolymerans_strain_KCTC_42856	24	longer	67	46.3	142	3176380	3176580	3176203	3176631
CP010431.2|Pandoraea_sputorum_strain_DSM_21091	13.7	longer	65	50	137	1094398	1094592	1094233	1094643

CP023422.1|Janthinobacterium svalbardensis strain PAMC 27463	19.2	yes	63	54	136	37218	37406	37047	37457
CU207211.1|Herminiimonas_arsenicoxydans	32.9	no	65	52.9	136	603835	604029	603784	604194
CP012750.1|Glutamicibacter_halophytocola_strain_KLBMP_5180	31.7	yes	57	44.5	118	3850084	3850254	3850033	3850389

Supplementary table S2.13: Localization and features of homologues used for the phylostratigraphic analysis of OGC167. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC167	full-length sORF homologue	length sORF (aa)	Identity mORF to #4168	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-		157		335	3927557	3928027	3927778	3928785
CP001925.1|Escherichia coli Xuzhou21	45.4	no	179	99.7	335	3767221	3767757	3767511	3768518
CP026199.1|Escherichia coli strain ECONIH6	7.6	no	161	98.2	335	3679307	3679792	3678494	3679501
CP022459.1|Shigella sonnei strain 2015C-3807	40.7	no	173	48.1	163	4371516	4372034	4371791	4372279
CP000266.1|Shigella flexneri 5 str. 8401	42	no	164	97.6	335	3078203	3078694	3078448	3079455
CP000034.1|Shigella dysenteriae Sd197	51	no	148	98.8	335	2890867	2891277	2890106	2891113
CP011511.1|Shigella boydii strain ATCC 9210	37.3	no	83	97.6	335	3081855	3082394	3081094	3082101
CU928158.2|Escherichia fergusonii ATCC 35469	15.3	no	66	39.1	335	2962799	2963287	2963092	2964096
CP016762.1|Citrobacter freundii strain B38	13.4	no	167	38.8	335	761509	761916	760610	761617
CP007025.1|Escherichia albertii KF1	40.7	yes	148	93.1	335	3190667	3191110	3189906	3190913
CP025979.1|Escherichia marmotae strain HT073016	41.5	yes	150	90.7	335	4064260	4064709	4063499	4064506
CP012038.1|Salmonella enterica subsp. enterica serovar Ouakam strain GNT-01	26.8	no	163	64.8	335	1880032	1880520	1880274	1881281
CP003938.1|Enterobacteriaceae bacterium strain FGI 57	11.3	no	166	32.2	334	758119	758619	757316	758320
CP027225.1|Phytobacter sp. SCO41	11.4	no	235	38.4	335	878553	879257	877953	878957
CP017928.1|Klebsiella oxytoca strain CAV1015	12.1	no	168	35.7	335	3630948	3631448	3631247	3632254
CP016811.1|Klebsiella pneumoniae strain DHQP1002001	10.2	no	164	35.4	335	4941926	4942426	4941120	4942127

Supplementary table S2.14: Localization and features of homologues used for the phylostratigraphic analysis of OGC174. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC174	full-length sORF homologue	length sORF (aa)	Identity mORF to #4292	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	58	-	207	4044465	4044641	4043937	4044560
CP022459.1|Shigella sonnei strain 2015C-3807	50	no	42	98.6	207	4271673	4271798	4271706	4272329
AE014073.1|Shigella flexneri 2a str. 2457T	93.1	yes	58	99	207	3196419	3196592	3195888	3196511
CP000036.1|Shigella boydii Sb227	93.1	yes	58	99	207	2933940	2934113	2933409	2934032
CP009859.1|Escherichia coli strain ECONIH1	50	no	39	99.5	207	3677732	3677848	3677201	3677824
CU928158.2|Escherichia fergusonii ATCC 35469	47.5	yes	43	90.9	208	3093495	3093623	3092961	3093587
AP014855.1|Escherichia albertii NIAH_Bird_3	50	no	42	98.6	207	3114416	3114532	3113885	3114508
CP006692.1|Salmonella bongori serovar 48:z41:-- str. RKS3044	22.4	yes	58	70.8	214	2990536	2990709	2990041	2990685

HG326213.1|Salmonella enterica subsp. enterica serovar Typhimurium str. DT2	22.4	yes	58	70	205	3338370	3338543	3337902	3338519
CP014007.1|Kosakonia oryzae strain Ola 51	22.4	yes	38	58.5	173	641738	641851	641759	642280
CP011602.1|Kluyvera intermedia strain CAV1151	24.1	no	57	63.6	199	5369643	5369814	5369362	5369961
CP020448.1|Citrobacter braakii strain FDAARGOS_253	23.9	yes	42	70.2	223	4190998	4191123	4190849	4191517
CP023529.1|Lelliottia amnigena strain FDAARGOS_395	16.2	yes	55	61.8	181	2770528	2770692	2770129	2770674
CP013990.1|Leclercia adecarboxylata strain USDA-ARS-USMARC-60222	27.1	yes	78	61.5	181	664640	664873	664667	665212
CP003737.1|Enterobacter cloacae subsp. cloacae ENHKU01	18	yes	39	58.5	165	4093142	4093258	4092746	4093240
CP004887.1|Klebsiella michiganensis HKOPL1	23.7	no	43	61.5	184	1402625	1402753	1402661	1403215
FN543093.2|Cronobacter turicensis z3032	16.9	no	104	42.5	145	521318	521581	521303	521740
CP009450.1|Pluralibacter gergoviae strain FB2	24.2	no	56	50	169	5564	5731	5180	5686
CP009451.1|Cedecea neteri strain SSMD04	2.9	no	53	31.4	172	189841	189999	189468	189983
CP015581.1|Tatumella citrea strain ATCC 39140	1	no	47	29.9	166	820654	820794	820260	820760
AP012032.2|Pantoea ananatis AJ13355	0.9	no	70	30.9	176	2973262	2973498	2973325	2973855
AP012551.1|Plautia stali symbiont	12.5	no	44	24.6	150	3131459	3131590	3131496	3131948
CP014137.1|Brenneria goodwinii strain FRB141	12.7	yes	50	47.4	220	1422632	1422781	1422677	1423336
CP006569.1|Sodalis praecaptivus strain HS1	9.2	yes	39	32.4	181	4039019	4039135	4039055	4039597
CP009787.1|Yersinia rohdei strain YRA	19.7	yes	51	31.4	221	1433067	1433219	1432568	1433233
CP019063.1|Rouxiella sp. ERMR1:05 plasmid unnamed1	4.4	no	50	30.4	160	381489	381638	381498	381977
CP013913.1|Serratia fonticola strain GS2	10.4	no	55	33.9	212	2435657	2435821	2435195	2435833
CP004345.1|Morganella morganii subsp. morganii KT	11.9	no	40	31.6	251	2358099	2358224	2357480	2358235
CP021550.1|Proteus mirabilis strain AR_0159	3.5	no	41	27.1	272	491532	491654	491533	492351

Supplementary table S2.15: Localization and features of homologues used for the phylostratigraphic analysis of OGC194. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC194	full-length sORF homologue	length sORF (aa)	Identity mORF to #4769	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	196	-	466	4495934	4496524	4495940	4497340
CP016755.1|Escherichia coli strain FORC_044	100	yes	196	100	466	261056	261643	260237	261637
CP022050.2|Escherichia coli O157 strain FDAARGOS_293	100	yes	196	100	466	2782848	2783435	2782029	2783429
CP026827.1|Shigella dysenteriae strain CFSAN010956	95.4	yes	196	99.8	466	2535890	2536477	3520925	3522325
CP014099.2|Shigella sonnei strain FDAARGOS_90	95.4	yes	196	99.8	466	3984221	3984808	3984227	3985627
CP026845.1|Shigella boydii strain NCTC 9733	95.4	yes	196	99.8	466	3184139	3184726	3183320	3184720
CP026811.1|Shigella flexneri strain 64-5500	94.9	yes	196	99.8	466	3476958	3477545	3476139	3477539
CU928158.2|Escherichia fergusonii ATCC 35469	66.8	no	279	97	466	1614228	1615064	1613670	1615070
CP025317.1|Escherichia albertii strain 1551-2	72	yes	287	97.9	466	3660371	3661231	3660389	3661789
CP025979.1|Escherichia marmotae strain HT073016	75.5	yes	415	98.1	466	2996987	2997823	2996981	2998381
CP005991.1|Enterobacter sp. R4-368	31.4	no	198	58	461	1826974	1827573	1826968	1828353
CP016337.1|Kosakonia sacchari strain BO-1	30.3	no	254	58	461	4263993	4264754	4263375	4264760
CP011254.1|Serratia fonticola strain DSM 4576	36.6	no	196	55	466	3118387	3118974	3117580	3118980
CP007044.2|Chania multitudinisentens RB-25	10.2	no	68	54.7	466	519465	519614	519405	520805
CP014031.2|Hafnia paralvei strain FDAARGOS_158	0.3	no	196	89.1	466	1274367	1274954	1273556	1274956
CP014608.1|Obesumbacterium proteus strain DSM 2777	27.8	no	204	56.7	426	1358657	1359268	1358651	1359931

CP023706.1|Edwardsiella tarda strain KC-Pc-HB1	42.1	no	238	79.4	464	2369052	2369765	2368377	2369771
CP025800.1|Yersinia ruckeri strain SC09	1.9	yes	187	85.9	466	2920115	2920702	2920113	2921516
CP023536.1|Providencia alcalifaciens strain FDAARGOS_408	30.5	no	204	58.2	466	1158304	1158915	1158298	1159698
CP026051.1|Proteus mirabilis strain FDAARGOS_67	26.7	no	194	57.3	463	523974	524555	523968	525359
CP000851.1|Shewanella pealeana ATCC 700345	26.1	no	196	56.4	464	3615446	3616033	3615440	3616834
AP018045.1|Photobacterium damselae subsp. Piscicida	0.2	no	225	56.4	466	2618358	2619038	2617658	2619058
CP004345.1|Morganella morganii subsp. morganii KT	30.2	yes	181	50.7	460	540098	540640	539264	540646

Supplementary table S2.16: Localization and features of homologues used for the phylostratigraphic analysis of OGC198. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC198	full-length sORF homologue	length sORF (aa)	Identity mORF to #4794	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	41	-	559	4528288	4528163	4527984	4529663
CP022457.1|Shigella sonnei strain 2015C-3566	97.6	yes	41	100	559	1255544	1255666	1254169	1255848
CP009050.1|Escherichia coli NCCP15648	97.6	yes	41	100	559	4325228	4325350	4325046	4326725
CU928158.2|Escherichia fergusonii ATCC 35469	85.4	yes	41	97.1	559	3609398	3609520	3609216	3610895
CP007025.1|Escherichia albertii KF1	85.4	no	41	95.7	559	2569916	2570038	3661023	3662702
CP019649.1|Salmonella enterica subsp. enterica serovar Typhimurium strain TW-Stm6	70.7	yes	38	85	559	3827850	3827963	3827659	3829338
CP022120.1|Salmonella bongori serovar 66:z41:- str. SA19983605	68.3	no	41	84.4	559	200394	200516	199019	200698
CP011132.1|Citrobacter amalonaticus Y19	68.3	yes	34	90.7	559	4160051	4160152	4159848	4161527
CP018016.1|Kosakonia radicincitans DSM 16656	63.4	yes	34	90.7	559	197097	197198	195722	197401
CP011602.1|Kluyvera intermedia strain CAV1151	78	no	59	77.6	559	4896830	4897007	4895456	4897135
CP003938.1|Enterobacteriaceae bacterium strain FGI 57	70.7	no	63	79.4	559	207829	208017	206454	208133
CP006580.1|Enterobacter cloacae P101	80.5	no	62	83.5	559	314055	314243	312680	314359
CP015774.2|Lelliottia amnigena strain ZB04	75.6	yes	63	81.4	558	4394709	4394897	4394602	4396272
CP020847.1|Klebsiella pneumoniae strain KPN1481	65.1	yes	77	74.5	559	2168488	2168718	2168411	2170090
CP012555.1|Raoultella ornithinolytica strain 18	68.3	yes	59	75.3	556	4223614	4223790	4222245	4223915
CP004887.1|Klebsiella michiganensis HKOPL1	73.2	yes	59	65.8	497	954644	954820	2168411	2170090
CP009450.1|Pluralibacter gergoviae strain FB2	56.1	yes	59	69.1	448	509971	510147	509846	511492
CP009451.1|Cedecea neteri strain SSMD04	63.4	yes	38	72.7	560	2733543	2733656	2732165	2733847
CP012253.1|Cronobacter sakazakii strain NCTC 8155	70.7	no	63	76.6	561	4143151	4143339	4141770	4143455
CP001560.1|Shimwellia blattae DSM 4481 = NBRC 105725	51.2	yes	59	63.5	546	156795	156971	155447	157087
FP236843.1|Erwinia billingiae strain Eb661	58.5	no	59	58.9	545	4878409	4878585	4878293	4879930
AP012551.1|Plautia stali symbiont	58.5	no	63	61.8	546	1688548	1688736	1687200	1688837
CP002206.1|Pantoea vagans C9-1	61	no	59	62	558	3558468	3558644	3557108	3558784
LT575468.1|Plesiomonas shigelloides strain NCTC10360	29.3	no	59	45.2	565	76941	77117	76819	78516
CP016043.1|Edwardsiella hoshinae strain ATCC 35051	37.3	no	59	55.5	561	112822	112998	111444	113126
CP014031.1|Hafnia alvei strain FDAARGOS_158	58.5	no	63	64.2	555	1238871	1239060	1237509	1239176
CP014608.1|Obesumbacterium proteus strain DSM 2777	58.5	no	63	63.5	554	865129	865317	863769	865433
CP014137.1|Brenneria goodwinii strain FRB141	1.4	no	54	57.8	550	4070667	4070831	4069318	4070970
CP003776.1|Pectobacterium carotovorum subsp. carotovorum PCC21	51.2	no	59	60.9	553	4682167	4682343	4680822	4682483
CP014136.1|Gibbsiella quercinecans strain FRB97	56.1	no	59	60.1	549	2397959	2398135	2397834	2399483
LT883155.1|Serratia grimesii isolate BXF1	3.8	yes	44	62.9	550	18505	18637	171495	173147

CP007044.2|Chania multitudinisentens RB-25	53.7	no	63	59.8	550	757506	757694	757399	759051
CP019062.1|Rouxiella sp. ERMR1:05	46.3	no	63	57.1	549	1416160	1416348	1416047	1417696
CP003403.1|Rahnella aquatilis HX2	46.3	no	63	55.8	549	2403389	2403577	2402044	2403690
CP009781.1|Yersinia aldovae 670-83	56.1	no	63	63.8	553	2829181	2829369	2827821	2829482
CP021852.1|Proteus mirabilis strain AR_0156	34.1	no	64	47.9	569	3463286	3463474	3463179	3464888
CP000021.2|Vibrio fischeri ES114	26.7	no	42	41.1	543	1005685	1005810	1005530	1007158
FM178380.1|Aliivibrio salmonicida LFI1238	32.6	yes	42	40.6	541	1024389	1024514	1024234	1025859
CP003171.1|Oceanimonas sp. GK1	31	no	41	42.8	510	3139523	3139712	3138294	3139823
LT707061.1|Pseudomonas putida strain N1R	33.3	no	41	45.2	538	3573210	3573332	3573043	3574656
CP015612.1|Stenotrophomonas maltophilia strain OUC_Est10	13.5	no	40	43.4	543	649698	649817	649531	651162
CP003350.1|Frateuria aurantia DSM 6220	26.5	no	41	42.4	537	2508928	2509041	2508770	2510380
CP019697.1|Paenalcaligenes hominis strain 15S00501	15.3	no	42	44.3	556	92525	92650	92343	94013

Supplementary table S2.17: Localization and features of homologues used for the phylostratigraphic analysis of OGC226. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC226	full-length sORF homologue	length sORF (aa)	Identity mORF to #5520	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	106	-	334	5306609	5306929	5305892	5306896
CP026867.1|Shigella boydii strain DMB SH135	99.1	yes	106	100	334	3340287	3340604	3339567	3340571
CP026872.1|Shigella dysenteriae strain ATCC 9764	98.1	no	107	100	334	2480571	2480888	2480604	2481608
CP027534.1|Escherichia coli strain AR_0081	96.3	no	107	100	334	322872	323189	322905	323909
CP011288.1|Salmonella enterica subsp. diarizonae strain 11-01855	54.4	yes	106	96.1	334	675510	675791	674754	675758
FR877557.1|Salmonella bongori NCTC 12419	56.1	yes	106	95.5	334	4255785	4256102	4255065	4256069
CP022114.1|Kluyvera georgiana strain YDC799	60.4	yes	133	88	334	4477352	4477750	4477376	4478380
CP007557.1|Citrobacter freundii CFNIH1	61.5	yes	106	94.6	334	1763827	1764144	1763107	1764111
CP002886.1|Enterobacter cloacae EcWSU1	0	no	102	92.5	334	400882	401058	400025	401029
CP026387.1|Leclercia sp. LSNIH3	54.7	yes	94	92.8	334	2007580	2007861	2007580	2008581
CP014007.1|Kosakonia oryzae strain Ola 51	24.4	yes	103	92.8	334	4895178	4895327	4895210	4896214
CP015774.2|Lelliottia amnigena strain ZB04	34.5	no	92	92.5	334	441022	441222	440191	441195
CP026715.1|Klebsiella oxytoca strain AR_0028	46	no	137	93.4	334	4344879	4345289	4345005	4346009
CP010557.1|Raoultella ornithinolytica strain S12	36.8	no	140	92.2	334	943345	943764	942625	943629
CP009459.1|Cedecea neteri strain ND14a	43.4	yes	105	90.7	334	1588113	1588427	1587393	1588397
FN543093.2|Cronobacter turicensis z3032	54.1	yes	103	89.8	334	3834306	3834725	3834330	3835334
CP001560.1|Shimwellia blattae DSM 4481 = NBRC 105725	49.5	no	123	88.9	334	3693577	3693945	3693610	3694614
FP236843.1|Erwinia billingiae strain Eb661	44.1	yes	103	83.2	334	544337	544645	543617	544621
CP026377.1|Pantoea gaviniae strain DSM 22758	50	yes	102	85.1	335	3973234	3973539	3973252	3974259
CP015581.1|Tatumella citrea strain ATCC 39140	45.5	yes	106	80.1	337	3643989	3644306	3644013	3645026
CP000826.1|Serratia proteamaculans 568	54.5	yes	106	83	335	486905	487222	486185	487192
CP001655.1|Dickeya chrysanthemi Ech1591	44	no	110	78	331	562610	562939	561917	562912
CP006569.1|Sodalis praecaptivus strain HS1	10.7	yes	151	79.3	338	4012448	4012906	4012745	4013761
CP014137.1|Brenneria goodwinii strain FRB141	49.1	no	108	79.1	329	5035752	5036018	5035002	5035991

CP009769.1|Pectobacterium carotovorum subsp. brasiliense strain BC1	42.9	no	124	79.8	331	4254063	4254374	4254081	4255076
CP017481.1|Pectobacterium polaris strain NIBIO1006	44.5	no	123	79.5	331	509520	509828	509538	510530
CP020466.1|Edwardsiella ictaluri strain RUSVM-1	46	yes	107	81	334	2754307	2754627	2753593	2754597
LT575468.1|Plesiomonas shigelloides strain NCTC10360	26.5	no	107	58.2	299	3004619	3004867	3004580	3005479
CP010423.1|Pragia fontium strain 24613	28	no	117	68.2	334	486394	486744	485677	486681
CP015379.1|Hafnia alvei strain HUMV-5920	35.7	no	118	80.7	333	984578	984931	984641	985642
CP014608.1|Obesumbacterium proteus strain DSM 2777	34.9	no	118	80.7	333	4935926	4936243	4935989	4936990
CP003244.1|Rahnella aquatilis CIP 78.65 = ATCC 33071	40.7	no	129	79.3	332	468561	468947	467925	468923
CP019062.1|Rouxiella sp. ERMR1:05	23.1	no	119	79	333	3169220	3169576	3169244	3170242
CP009787.1|Yersinia rohdei strain YRA	41.9	no	117	83	334	3348602	3348952	3347888	3348892
CP014136.1|Gibbsiella quercinecans strain FRB97	55.7	yes	100	82	334	2062713	2063012	2062728	2063732
CP007044.2|Chania multitudinisentens RB-25	49.6	yes	149	79.8	336	285590	286036	285665	286675
CP004345.1|Morganella morganii subsp. morganii KT	37.6	no	100	73.8	337	231972	232340	232041	233054
CP016176.1|Xenorhabdus hominickii strain ANU1	37.3	no	149	73.4	334	3026767	3027108	3026812	3027813
FM162591.1|Photorhabdus asymbiotica ATCC43949|	35.8	no	99	77.7	337	4742448	4742744	4742478	4743488
FN545161.1|Arsenophonus nasoniae	41.3	no	115	73	333	87464	87808	86756	87757
CP027418.1|Providencia rettgeri strain FDAARGOS_330	41.1	no	118	71.4	334	1122305	1122658	1122368	1123369
CP012674.1|Proteus mirabilis strain CYPM1	33.9	no	114	70.3	334	252243	252584	251535	252539
CP002667.1|Gallibacterium anatis UMN179	27.8	no	92	53.1	299	674498	674773	674492	675388
CP000947.1|Haemophilus somnus 2336	29.2	no	104	53	296	2028169	2028462	2028202	2029089
CP020345.1|Pasteurella multocida subsp. multocida strain CIRMBP-0884	32.1	no	73	54.8	296	1699626	1699949	1698999	1699886
CP009237.1|Glaesserella parasuis strain KL0318	24.6	no	98	52.8	296	876671	876961	876062	876949
CP001607.1|Aggregatibacter aphrophilus NJ8700	28.1	no	95	48.2	334	132925	133206	132931	133818
AE016827.1|Mannheimia succiniciproducens MBEL55E	33.3	no	110	54.5	296	1614673	1615002	1614061	1614948

Supplementary table S2.18: Localization and features of homologues used for the phylostratigraphic analysis of OGC231. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region.

homologue	Identity sORF to OGC231	full-length sORF homologue	length sORF (aa)	Identity mORF to #5573	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	174	-	176	5353324	5353992	5353468	5353998
CP026731.1|Shigella boydii strain ATCC 8700	97.7	yes	174	100	176	319256	319780	319256	319786
CP026839.1|Shigella dysenteriae strain ATCC 9752	97.1	yes	174	100	176	4026719	4027243	4026713	4027243
AP012306.1|Escherichia coli str. K-12 substr. MDS42	98.9	yes	174	100	176	3878224	3878748	3878224	3878754
CU928158.2|Escherichia fergusonii ATCC 35469	93.1	yes	174	98.3	176	4416564	4417088	4416564	4417094
AP014855.1|Escherichia albertii strain NIAH_Bird_3	89.7	no	174	98.9	176	4747692	4748216	4424919	4425449
CP006692.1|Salmonella bongori serovar 48:z41:-- str. RKS3044	76.4	yes	174	96	176	4177930	4178454	4177930	4178460
FN543502.1|Citrobacter rodentium ICC168	77	yes	174	95.5	176	3467808	3468332	3467808	3468338
CP011602.1|Kluyvera intermedia strain CAV1151	73	yes	174	93.8	176	4042263	4042784	4042254	4042784
CP026167.1|Leclercia sp. LSNIH1	72.4	yes	172	96	176	1410561	1411079	1410558	1411085
CP023529.1|Lelliottia amnigena strain FDAARGOS_395	71.8	no	174	94.3	176	3950212	3950736	3950212	3950742
CP002824.1|Enterobacter aerogenes KCTC 2190	73	no	175	95.5	176	2018674	2019198	2018674	2019204

CP004142.1|Raoultella ornithinolytica B6	74.7	yes	174	94.3	176	3620353	3620874	3620347	3620874
CP022690.1|Kosakonia cowanii strain Esp_Z	75.9	yes	174	94.9	176	599183	599707	599183	599713
CP009450.1|Pluralibacter gergoviae strain FB2	69	no	174	92	176	75228	75752	75222	75752
CP012264.1|Cronobacter condimenti 1330 strain LMG 26250	67.2	yes	173	95.5	175	3753089	3753610	3753083	3753610
CP009451.1|Cedecea neteri strain SSMD04	69	yes	173	94.3	175	2108563	2109084	2108557	2109084
CP001560.1|Shimwellia blattae DSM 4481 = NBRC 105725	63.2	yes	173	93.8	175	3663413	3663934	3663407	3663934
LN907827.1|Erwinia sp. EM595	58.9	no	175	89.2	176	3302695	3303219	3302689	3303219
CP002433.1|Pantoea sp. At-9b	69.7	yes	175	91.5	176	3866493	3867017	3866487	3867017
AP012551.1|Plautia stali symbiont DNA	65.7	no	175	91.5	176	2445548	2446072	2445548	2446078
AP008232.1|Sodalis glossinidius str. 'morsitans'	44	yes	174	88.6	175	648305	648826	648305	648832
CP021894.1|Pectobacterium carotovorum strain SCC1	64	no	175	96.6	176	937710	938234	937710	938240
CP014137.1|Brenneria goodwinii strain FRB141	58.9	no	175	93.8	176	111792	112316	111792	112322
CP001654.1|Dickeya paradisiaca Ech703	60.3	yes	172	93.2	176	935758	936282	935758	936288
CP017638.1|Dickeya dianthicola RNS04.9	58.3	yes	175	90.9	176	3441400	3441924	3441394	3441924
CP003244.1|Rahnella aquatilis CIP 78.65 = ATCC 33071	59.2	yes	175	90.9	175	496644	497168	496644	497174
CP019062.1|Rouxiella sp. ERMR1:05	66.1	yes	174	89.2	176	3140968	3141492	3140962	3141492
CP009539.1|Yersinia ruckeri strain YRB	58	no	174	91.5	176	2545409	2545933	2545409	2545939
LT883155.1|Serratia grimesii isolate BXF1	64.9	yes	174	94.9	176	438350	438874	438350	438880
CP007044.2|Chania multitudinisentens RB-25	61.1	no	174	93.2	176	257342	257863	257336	257863
CP010423.1|Pragia fontium strain 24613	51.4	no	173	89.8	177	520859	521386	520859	521392
CP011359.1|Edwardsiella tarda strain FL95-01	58.9	yes	176	94.3	176	3135566	3136090	3135560	3136090
CP015379.1|Hafnia alvei strain HUMV-5920	64.6	yes	176	94.3	176	955257	955781	955251	955781
CP014608.1|Obesumbacterium proteus strain DSM 2777	65.7	yes	176	93.8	176	4906532	4907056	4906526	4907056
CP025933.1|Morganella morganii strain KC-Tt-01	67.4	no	175	93.8	176	3424746	3425270	3424740	3425270
LN681227.1|Xenorhabdus nematophila AN6/1	60	no	175	91.5	176	414227	414751	414227	414757
CP027418.1|Providencia rettgeri strain FDAARGOS_330	52	no	176	90.9	176	1092388	1092912	1092382	1092912
FN545161.1|Arsenophonus nasoniae	36.8	no	175	88.1	176	107664	108188	107664	108194
CP006944.1|Mannheimia varigena USDA-ARS-USMARC-1312	36.6	no	193	81.2	176	1935949	1936524	1936012	1936539
CP002667.1|Gallibacterium anatis UMN179	37.2	no	255	81.8	176	1360047	1360709	1359939	1360466
CP009158.1|Glaesserella parasuis strain SH03	35.8	no	185	82.4	176	2108995	2109546	2108986	2109513
CP001607.1|Aggregatibacter aphrophilus NJ8700	41.3	no	220	82.4	176	1571319	1571975	1571310	1571837
CP006954.1|Bibersteinia trehalosi USDA-ARS-USMARC-188	32.4	no	226	81.8	176	91588	92382	91864	92391
CP001616.1|Tolumonas auensis DSM 9187	42.6	no	252	79.9	177	2770844	2771599	2770838	2771371
CP000644.1|Aeromonas salmonicida subsp. salmonicida A449	41.3	no	236	81.8	176	788444	789151	788627	789157
CP012621.1|Zobellella denitrificans strain F13-1	41.3	longer	200	81.2	176	285518	286117	285593	286123
CP003171.1|Oceanimonas sp. GK1	41.8	no	229	79.7	176	805645	806331	805807	806337
CP022307.1|Alcanivorax sp. N3-2A	33.1	no	221	76.1	176	209514	209987	209520	210050
CP002771.1|Marinomonas posidonica IVIA-Po-181	35.8	no	248	70.6	176	2819245	2819988	2819236	2819766
HF680312.1|Thalassolituus oleivorans MIL-1	39.4	no	223	67.2	179	694340	694966	694445	694984
CP013737.1|Herbaspirillum rubrisubalbicans M1	37	no	237	68.7	177	3419681	3420391	3419864	3420397
CP016448.1|Methyloversatilis sp. RAC08	32.9	no	246	62.6	182	3642944	3643681	3642938	3643486
CP003236.1|Tistrella mobilis KA081020-065	33.1	no	231	62.5	177	3786759	3787451	3786930	3787457
AP011112.1|Hydrogenobacter thermophilus TK-6	23.6	no	193	61.4	175	1474369	1474947	1474426	1474953
CP001931.1|Thermocrinis albus DSM 14484	24.7	no	246	58.7	178	357715	358452	357709	358245

Supplementary table S2.19: Localization and features of homologues used for the phylostratigraphic analysis of OGC241. Homologues were identified by tblastn of the mother gene. The identity was determinded by pairwise protein alignment within the matching region. The pairwise alignment of OGC241 started at the first 'L' in the overlapping region which was assumed to be the true start.

homologue	Identity sORF to OGC241	full-length sORF homologue	length sORF (aa)	Identity mORF to #5740	length mORF (aa)	localization sORF		localization mORF	
						start	stop	start	stop
EHEC EDL933	-	-	57	-	289	5540205	5540378	5540112	5540981
CP015843.2|Escherichia coli O157:H7 strain FRIK2455	100	yes	57	100	289	5638115	5638285	5638022	5638891
AE014075.1|Escherichia coli CFT073	94.7	yes	57	99.3	289	5224310	5224480	5224217	5225086
CU928158.2|Escherichia fergusonii ATCC 35469	63.2	yes	139	95.2	289	4581452	4581868	4581359	4582228
AP014855.1|Escherichia albertii NIAH_Bird_3	80.7	no	57	99	289	4553320	4553490	4553227	4554096
CP022120.1|Salmonella bongori serovar 66:z41:- str. SA19983605	42.1	no	41	93.8	289	3745827	3745949	3745173	3746042
FN543502.1|Citrobacter rodentium ICC168	50	no	40	93.1	289	5339492	5339668	5339399	5340268
CP011602.1|Kluyvera intermedia strain CAV1151	43.9	yes	57	78.7	290	3871739	3871909	3871133	3872002
CP022532.1|Enterobacter cloacae strain MS7884A	45.6	no	27	83	289	1135451	1135723	1134947	1135816
CP013990.1|Leclercia adecarboxylata strain USDA-ARS-USMARC-60222	47.5	yes	57	82.7	289	4122150	4122320	4121544	4122413
CP014156.1|Klebsiella quasipneumoniae strain HKUOPL4	47.4	yes	126	81.7	289	3530728	3531105	3530635	3531504
CP004142.1|Raoultella ornithinolytica B6	42.6	no	27	80.3	289	3344094	3344357	3343581	3344450
CP009450.1|Pluralibacter gergoviae strain FB2	38.7	longer	126	75.8	289	5358808	5359185	5358409	5359278
CP019445.1|Kosakonia cowanii strain 888-76	36.8	longer	91	78.7	290	488356	488628	487852	488724
CP012268.1|Cronobacter muytjensii ATCC 51329	45.2	no	75	77.9	289	3351180	3351404	3350628	3351497
CP009451.1|Cedecea neteri strain SSMD04	44.8	no	88	80.6	289	1969102	1969365	1968589	1969458
CP001560.1|Shimwellia blattae DSM 4481 = NBRC 105725	28.1	yes	75	73.7	289	3465003	3465227	3464451	3465320
FN434113.1|Erwinia amylovora CFBP1430	35.6	no	80	70.4	294	3056120	3056359	3055583	3056467
CP002206.1|Pantoea vagans C9-1	27.6	yes	106	69.4	293	59376	59690	59268	60149
CP002038.1|Dickeya dadantii 3937	24.2	no	57	67.8	289	4212913	4213155	4212379	4213245
CP016032.1|Serratia marcescens strain U36365	33.9	longer	98	72.3	289	643076	643369	642983	643852
CP007044.2|Chania multitudinisentens RB-25	27.6	no	55	70.6	289	5297223	5297447	5296671	5297540
CP015581.1|Tatumella citrea strain ATCC 39140	24.6	no	46	63.4	295	3507378	3507515	3506742	3507626
CP017236.1|Yersinia ruckeri strain QMA0440 isolate 14/0165-5k	30	no	81	70.9	289	3237302	3237544	3236768	3237634
FO818637.1|Xenorhabdus bovienii str. CS03	1.8	no	8	59.8	294	1056011	1056328	1055552	1056433

Supplementary table S3: List of primers used for phenotypic characterization, RT-qPCR, Promoter activity and 5' RACE of asa.

Phenotypic characterization	Asa Cloning forward	**8220+1F-NcoI**	GATCCATGGGGATGTTGCTGGTTTCAAACA
	Asa Cloning reverse	**8220+245R-HindIII**	GCCAAGCTTCTATCTGTCTGCCGGAATGG
	Asa translational arrested mutant forward	**8220KO+52F**	CCGCGCTGATAGCCTGATGCA
	Asa translational arrested mutant reverse	**8220KO+73R**	TGCATCAGGCTATCAGCGCGG
	Asa alternative translational arrested mutant forward	**8220alterKO+76F**	CAGGCAGTTGAAGGAAGATAT
	Asa alternative translational arrested mutant reverse	**8220alterKO+97R**	ATATCTTCCTTCAACTGCCTG
	Colony PCR primer forward	**pBAD-C+165F**	CAGAAAAGTCCACATTGATT
	Colony PCR primer reverse	**pBAD-C-R**	TGATTTAATCTGTATCAGGC
RT-qPCR	Asa forward	**qPCR-OLG8220+25F**	TAGCGCCAGAGGAGATCAAA
	Asa reverse	**qPCR-OLG8220+191R**	CGTTAGTGTTCTTCTGCTGC
	Negative control forward	**qPCR-neg8220F**	GTCATCCACTGCGACAAGAA
	Negative control reverse	**qPCR-neg8220R**	GTACACTTAGATTTGACAACCGC
	16S rRNA forward	**rrHF**	AATGTTGGGTTAAGTCCCGC
	16S rRNA reverse	**rrHR**	GGAGGTGATCCAACCGCAGG
Promoter activity	pProbeNT-asa cloning forward	**1Prom8220-261FHindIII**	AGCAAGCTTGCGTGAGATAGCCTTCGGTATCC
	pProbeNT-asa cloning reverse	**1Prom8220-173RSacI**	AATGAGCTCAGCTTGTTACACTGCCGGCAGTC
	pProbe negative forward	**4Prom8220-493FHindIII**	TAGAAGCTTGCGTTGGCAGGCCAGCAATTTTGG
	pProbe negative reverse	**4Prom8220-316RSacI**	CGCGAGCTCAACGGTAACAGCCGCTATTTATAC
5'RACE	Experiment 1 SP1	**OLOZ1238+235R**	GGAATGGGTGACGTAAT
	Experiment 1 SP2	**OLOZ1238+207R**	CACTGGGCGGAACGGCGTTA
	Experiment 1 SP3	**OLOZ1238+179R**	TCTTCTGCTGCTCTGCATATGTGC
	Experiment 1 Sequencing	**OLG8220+25R**	TTTGATCTCCTCTGGCGCTA
	Experiment 2 SP1	**RACEGFP+284R**	TCCTGTACATAACCTTCGG
	Experiment 2 SP2	**RACEGFP+192R**	AAAGTAGTGACAAGTGTTGGCC
	Experiment 2 SP3 (+ Sequencing)	**RACEGFP+117R**	GTATGTTGCATCACCTTCACCTC
Western blot	SPA tag forward	**SPA-tag-F-HindIII**	ATCAAGCTTACAAGAGAAAAAAGAATTTCATAGCCGTCT
	SPA tag reverse	**SPA-tag-R-SalI**	TTCGTCGACCTACTTGTCATCGTCATCCTTGTAGTCGATGTCATG
	asa forward	**8220+1F-NcoI**	GATCCATGGGGATGTTGCTGGTTTCAAACA
	asa reverse	**OGC243R-HindIII**	GCCAAGCTTTCTGTCTGCCGGAATGGGTG

Supplementary table S4: List of primers used for phenotypic characterization and RT-PCR of asa homologues.

Overexpression phenotype	Asa CF Cloning forward	**CF8220+1F-NcoI**	GATCCATGGGGATATTGCTGGTTTCAAACAGG
	Asa CF Cloning reverse	**CF8220+242R-HindIII**	GCCAAGCTTCTACCTGTCTGCAGGTATGGGTGACG
	Asa CF Knockout mutation forward	**CF8220KO+1F**	CCGCGCTGATCGCCTGATGCA
	Asa CF Knockout mutation reverse	**CF8220KO+52R**	TGCATCAGGCGATCAGCGCGG
	Streptomycin resistance forward	**pRMsStrep+696F-PvuI**	CATGCTCGATCGATGAATCGAACTAATATTT
	Streptomycin resistance reverse	**pRMsStrep+1499R-AseI**	CCATCGATTAATTCAACCCCAAGTCAGAGGG
	Asa SM Cloning forward	**SM8220+1F-NcoI**	GATCCATGGGGATGTTGCTGGTTTCCCACA
	Asa SM Cloning reverse	**SM8220+191R-HindIII**	GCCAAGCTTCTACGTCTTCTTCTGCTGCTCGG
	Asa SM Knockout mutation forward	**SM8220KO+1F**	GCGCGCTGATAGCCTGATGCA
	Asa SM Knockout mutation reverse	**SM8220KO+52R**	TGCATCAGGCTATCAGCGCGC
	Kanamycin resistance forward	**pprobekan+6350PvuI**	CATGCTCGATCGATGATTGAACAAGATGGA
	Kanamycin resistance reverse	**pprobekan+5556R-AseI**	CCATCGATTAATTCAGAAGAACTCGTCAAGA
	Asa SE Cloning forward	**8220SE+1F-NcoI**	GATCCATGGGGATGTTACTGGTTTCATACA
	Asa SE Cloning reverse	**8220SE+171R-PstI**	ATCTGCAGTTACGTTCTGACCACCCG
	Asa SE Knockout mutation forward	**8220SE+55Mut-F**	GCGCTAATCGCCAGATGAAATGCAGGCA
	Asa SE Knockout mutation reverse	**8220SE+82Mut-R**	TGCCTGCATTTCATCTGGCGATTAGCGC
	Asa HA Cloning forward	**8220HA+1F-NcoI**	GATCCATGGGGATGTTGCTGGTTTCCCACA
	Asa HA Cloning reverse	**8220HA+167R-PstI**	ATCTGCAGTTACGTGCTGACCACACGCA
	Asa HA Knockout mutation forward	**8220HA+55Mut-F**	GCGCTAATCGCCAGCTAAAGCGCAGGGA
	Asa HA Knockout mutation reverse	**8220HA+82Mut-R**	TCCCTGCGCTTTAGCTGGCGATTAGCGC
RT-PCR	Asa CF forward	**qPCR-CF+12F**	GTTTCAAACAGGATTGCGCCTG
	Asa CF reverse	**qPCR-CF+104R**	GTGGTTGTGCTGATCGGTATG
	Asa SM forward	**qPCR-SM+2F**	TGATGTTGCTGGTTTCCCACAG
	Asa SM reverse	**qPCR-SM+156R**	GACCACCCGTAAAGACATGTC
	Asa SE forward	**qPCR-SE+9F**	CTGGTTTCATACAGGATAGCGC
	Asa SE reverse	**qPCR-SE+140R**	GATATGTCTTTCCTGGGCGGTATG
	Asa HA forward	**qPCR-HA+1F**	ATGTTGCTGGTTTCCCACAGG
	Asa HA reverse	**qPCR-HA+155R**	GCTGACCACACGCAAAGATATG

Supplemental table S5: Restriction enzymes (RE) used for cloning. The respective cutting sites were introduced by PCR. All RE are obtained from Thermo Fisher Scientific.

Experiment	Cloned insert	RE pairs	Buffer
Overexpression phenotype (in pBAD*myc*-His C)	*asa*	*Nco*I / *Hind*III	Tango
	asa CF	*Nco*I / *Hind*III	Tango
	asa SM	*Nco*I / *Hind*III	Tango
	asa SE	*Nco*I / *Pst*I	Tango
	asa HA	*Nco*I / *Pst*I	Tango
	streptomycin or kanamycin resistance gene	*Pvu*I / *Ase*I	2x Tango
Promoter activity (in pProbe NT)	*asa*	*Hind*III / *Sac*I	Tango
	NC I	*Hind*III / *Sac*I	Tango
Western blot	*asa*	*Nco*I / *Hind*III	Tango
	SPA-tag	*Hind*III / *Sal*I	2x Tango

Supplementary table S6: Primer efficiencies of RT-qPCR of *asa* and *asa* homologues and the negative control.

Gene	Annealing temperature	Efficiency [%] gene primer	R^2 gene primer	Efficiency [%] 16S rRNA primer	R^2 16S rRNA primer
asa	61°C	130	0.9992	87	0.9995
asa CF	60°C	97	0.9990	85	0,9983
asa SM	60°C	96	0.9996	88	0,9989
asa SE	60°C	98	0.9988	85	0,9958
asa HA	60°C	90	0.9972	88	0,9997
Negative control	58°C	99.6	0.9993	80	0.9981

Supplementary table S7: List of bacteria having *asa* homologous sequences with available RNAseq and /or RIBOseq data.

Organism	Refseq accession number	SRA accession number		Publication
Escherichia coli O157:H7 (EHEC) strain EDL933	CP008957.1	RNAseq	Replicate 1: SRR5266617 Replicate 2: SRR5266619	(Landstorfer 2014)
		RIBOseq	Replicate 1: SRR5266618 Replicate 2: SRR5266620	
EHEC strain Sakai	NC_002695	RNAseq	Replicate 1: SRR5874481 Replicate 2: not uploaded	(Hücker, Ardern et al. 2017)
		RIBOseq	Replicate 1: SRR5874484 Replicate 2: not uploaded	
E. coli LF82	NC_011993.1	RNAseq	Not uploaded	Unpublished, provided by Michaela Kreitmeier and Franziska Giehren in 2018
		RIBOseq	Not uploaded	
E. coli K12 substrain MG1655	NC_000913.3	RNAseq	Replicate 1: SRR4023277 Replicate 2: SRR4023278 Replicate 3: SRR4023279	(Hwang and Buskirk 2017)
		RIBOseq	Replicate 1: SRR4023274 Replicate 2: SRR4023275 Replicate 3: SRR4023276	
E. coli K12 substrain MC4100	NZ_HG738867.1	RNAseq	Replicate 1: SRR2016456 Replicate 2: SRR2016464	(Bartholomaus, Fedyunin et al. 2016)
		RIBOseq	Replicate 1: SRR2016457 Replicate 2: SRR2016465	
S. enterica 14028S	NC_016856.1	RNAseq	SRR4417739	(Baek, Lee et al. 2017)
		RIBOseq	SRR4417735	
S.enterica SL1344	NC_016810.1	RNAseq	SRR5090710	(Ndah, Jonckheere et al. 2017)
		RIBOseq	SRR5090708	
Escherichia coli E2348	NZ_BDOY01000009.1	RNAseq	SRR2601721	(Hazen, Daugherty et al. 2015)
Shigella flexneri 5a M90T	CM001474.1	RNAseq	ERR364203	(Vergara-Irigaray, Fookes et al. 2014)
Citrobacter rodentium ICC168	NC_013716	RNAseq	ERR026449	(Petty, Feltwell et al. 2011)
Sodalis praecaptivus	NZ_CP006569.1	RNAseq	SRR5445356	(Enomoto, Chari et al. 2017)
Serratia marcescens WW4	NC_020211.1	RNAseq	SRR3744056	(Madison, Berg et al. 2017)
Enterobacter aerogenes KCTC2190	NC_015663.1	RNAseq	SRR3994393	(Prados, Linder et al. 2016)
Klebsiella pneumonia subsp. pneumonia MGH78578	NC_009648.1	RNAseq	SRR408498	(Kim, Hong et al. 2012)
Cronobacter sakazakii ATCC BAA-894	NC_009778.1	RNAseq	SRR2047124	(Jing, Du et al. 2016)

Supplementary table S8: Blastp (September 2015) and Re-blast (September 2018) of the 172 sORFs matching to proteins with predicted function. The re-blast shows the top hit (highest E-value) having a predicted function or, if the protein is still in the database, the protein detected at 2015. The re-blast found 146 sORFs (of 172 in 2015) with hit and confirmed 120 proteins with homologues of predicted function. The minority of sORFs (**29) find the same hit as in 2015**. Of those 10 are now named to be hypothetical proteins. **Ninety of the proteins with different accession numbers have the same predicted function**, which is a strong hind for a reliable hit.

ID CP008957.1_	blastp 2015			Hit still present?	blastp 2018		
	blast hit	E-value	coverage		Top hit 2018 (with same product)	E-value	coverage
#30287	WP_006090480.1|phenol hydroxylase [Natronorubrum tibetense]	2,00E-09	100	no	ABA49062.1|phenol hydroxylase, putative [Burkholderia pseudomallei 1710b]	5,00E-33	100
#32339	WP_000386948.1|type III secretion system protein SepZ [Escherichia albertii]	1,00E-39	100	**yes**	**WP_000386948.1|hypothetical protein [Escherichia albertii]**	**2,00E-40**	**100**
#33735	WP_032290238.1|5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase [Escherichia coli]	2,00E-48	100	no	ENG09504.1|5-methyltetrahydropteroyltriglutamate--homocysteine S-methyltransferase [Escherichia coli P0305260.5]	7,00E-49	100
#66748	**WP_001303605.1|outer membrane protein [Escherichia coli]**	3,00E-50	100	**yes**	**WP_001303605.1|outer membrane protein [Escherichia coli]**	3,00E-51	100
#32064	**WP_001513412.1|ATP-dependent DNA helicase RecG [Escherichia coli]**	4,00E-44	100	**yes**	**WP_001513412.1|ATP-dependent DNA helicase RecG [Escherichia coli]**	3,00E-44	99
#56554	WP_044167163.1|cell surface protein, partial [Escherichia coli]	3,00E-19	100	no	EGI36713.1|putative cell surface protein [Escherichia coli TA271]	1,00E-20	100
#8256	WP_044167163.1|cell surface protein, partial [Escherichia coli]	3,00E-19	100	no	EGI36713.1|putative cell surface protein [Escherichia coli TA271]	1,00E-20	87
#56801	WP_001113770.1|molecular chaperone Tir [Escherichia coli]	3,00E-23	100	no	EYW21033.1|molecular chaperone Tir [Escherichia coli O157:H7 str. 2011EL-2114]	6,00E-24	100
#56803	WP_001113768.1|molecular chaperone Tir [Escherichia coli]	5,00E-23	100	no	OVD69729.1|molecular chaperone Tir [Escherichia coli]	3,00E-23	98
#56805	WP_044707557.1|molecular chaperone Tir [Escherichia coli]	2,00E-22	100	no	EZQ41898.1|molecular chaperone Tir [Escherichia coli O157: str. 2010EL-2045]	3,00E-23	98
#56807	WP_044707557.1|molecular chaperone Tir [Escherichia coli]	2,00E-22	100	no	EZQ41898.1|molecular chaperone Tir [Escherichia coli O157: str. 2010EL-2045]	3,00E-23	98
#8548	WP_001113770.1|molecular chaperone Tir [Escherichia coli]	3,00E-23	100	no	EYW21033.1|molecular chaperone Tir [Escherichia coli O157:H7 str. 2011EL-2114]	6,00E-24	100
#8550	WP_001113770.1|molecular chaperone Tir [Escherichia coli]	3,00E-23	100	no	EYW21033.1|molecular chaperone Tir [Escherichia coli O157:H7 str. 2011EL-2114]	6,00E-24	100
#8552	WP_001113770.1|molecular chaperone Tir [Escherichia coli]	3,00E-23	100	no	EYW21033.1|molecular chaperone Tir [Escherichia coli O157:H7 str. 2011EL-2114]	6,00E-24	100
#8554	WP_001113768.1|molecular chaperone Tir [Escherichia coli]	5,00E-23	100	no	OVD69729.1|molecular chaperone Tir [Escherichia coli]	3,00E-23	100
#70688	WP_044813414.1|glycine dehydrogenase [Escherichia coli]	1,00E-08	100	no	no hit		
#51705	WP_046671466.1|alpha-ketoglutarate permease [Escherichia coli]	3,00E-18	100	**yes**	**WP_046671466.1|MULTISPECIES: hypothetical protein [Enterobacteriaceae]**	**8,00E-19**	**100**
#51790	WP_000465894.1|LysR family transcriptional regulator [Salmonella enterica]	5,00E-08	99	no	STI72899.1|transcriptional regulator YfiE [Escherichia coli]	4,00E-14	57
#23478	WP_022227540.1|cytosine deaminase [Acidaminococcus intestini CAG:325]	4,00E-20	98	no	KPX31290.1|Cytosine deaminase [Pseudomonas ficuserectae]	2,00E-58	42
#26867	WP_044817382.1|SAM-dependent methyltransferase, partial [Escherichia coli]	1,00E-08	97	no	no hit		
#32417	**WP_002460384.1|sugar transporter [Escherichia albertii]**	9,00E-10	97	**yes**	**WP_002460384.1|sugar efflux transporter [Escherichia albertii]**	7,00E-10	96
#29422	**WP_011082766.1|50S ribosomal protein L14 [Staphylococcus epidermidis]**	3,00E-14	96	**yes**	**WP_011082766.1|50S ribosomal protein L14 [Staphylococcus epidermidis]**	2,00E-14	96
#11029	WP_006090480.1|phenol hydroxylase [Natronorubrum tibetense]	1,00E-04	95	no; only hypothetical proteins	SAJ25528.1|Uncharacterised protein [Enterobacter cloacae]	4,00E-34	100
#19360	WP_032360975.1|phosphomannomutase, partial [Escherichia coli]	9,00E-39	95	no	no hit		
#19512	WP_032360975.1|phosphomannomutase, partial [Escherichia coli]	6,00E-38	95	no	no hit		
#33970	**WP_041379862.1|high mobility group protein Z [Photorhabdus luminescens]**	4,00E-04	95	**yes**	**WP_041379862.1|MULTISPECIES: hypothetical protein [Photorhabdus]**	4,00E-04	44
#29421	WP_044810876.1|propionate CoA-transferase [Escherichia coli]	1,00E-20	94	no	CDL33070.1|hypothetical protein [Enterobacter cloacae ISC8]	7,00E-07	53
#71948	WP_032282718.1|amidohydrolase, partial [Escherichia coli]	2,00E-09	94	no	WP_077756891.1|isochorismatase family protein [Escherichia coli]	3,00E-05	74
#5003	WP_021994652.1|6-pyruvoyl-tetrahydropterin synthase [Sutterella wadsworthensis CAG:135]	5,00E-05	93	no	EKS68785.1|6-pyruvoyl-tetrahydropterin synthase [Burkholderia sp. SJ98]	9,00E-17	100
#9297	WP_032175075.1|proline dehydrogenase [Escherichia coli]	2,00E-17	93	no	EHV16895.1|proline dehydrogenase domain protein [Escherichia coli DEC4E]	4,00E-18	86
#35760	WP_044813740.1|LysR family transcriptional regulator, partial [Escherichia coli]	2,00E-08	93	no	no hit		
#44501	WP_024900448.1|DNA-directed RNA polymerase II [Burkholderia mallei]	1,00E-11	92	no	CRY03927.1|DNA-directed RNA polymerase II%2C large subunit [Burkholderia pseudomallei]	2,00E-12	40
#14354	**WP_010989194.1|MULTISPECIES: restriction endonuclease [Enterobacteriaceae]**	9,00E-25	92	**yes**	**WP_010989194.1|MULTISPECIES: type I toxin-antitoxin system endodeoxyribonuclease toxin RalR [Enterobacteriaceae]**	2,00E-25	92
#26248	**WP_000362479.1|transcriptional regulator [Escherichia coli]**	2,00E-16	92	**yes**	**WP_000362479.1|transcriptional regulator [Escherichia coli]**	7,00E-17	91

ID CP008957.1_	blastp 2015			Hit still present?	blastp 2018		
	blast hit	E-value	coverage		Top hit 2018 (with same product)	E-value	coverage
#61632	WP_040079954.1|membrane protein [Escherichia coli]	1,00E-12	92	no	EHV26593.1|putative membrane protein [Escherichia coli DEC5A]	7,00E-16	100
#21291	WP_001462118.1|inverted ada-Golga3 fusion protein [Escherichia coli]	7,00E-42	91	no	EIP32594.1|inverted ada-Golga3 fusion protein [Escherichia coli EC4402]	1,00E-42	91
#65039	WP_023376426.1|copper sensitivity supression protein, partial [Salmonella enterica]	1,00E-04	91	no	ESO57391.1|copper sensitivity supression protein [Salmonella enterica subsp. enterica serovar Newport str. VA_R100506907]	2,00E-04	91
#28499	WP_044713641.1|methyltransferase, partial [Escherichia coli]	1,00E-08	91	no; only hypothetical proteins	CNU97973.1 |Uncharacterised protein [Salmonella enterica subsp. enterica serovar Bovismorbificans]	1,00E-08	100
#22893	WP_006090480.1|phenol hydroxylase [Natronorubrum tibetense]	3,00E-04	90	no	AAW76941.1|phenol hydroxylase [Xanthomonas oryzae pv. oryzae KACC 10331]	1,00E-28	100
#61510	**WP_011378736.1|collagenase [Shigella dysenteriae]**	1,00E-22	90	**yes**	**WP_011378736.1|peptidase [Shigella dysenteriae]**	1,00E-22	89
#49207	**WP_035892017.1|membrane protein [Kluyvera ascorbata]**	1,00E-21	89	**yes**	**WP_035892017.1|DUF2684 family protein [Kluyvera ascorbata]**	3,00E-22	59
#48056	WP_044708501.1|sugar isomerase, partial [Escherichia coli]	1,00E-04	89	no	no hit		
#14603	WP_007696036.1|phenol hydroxylase [Halococcus hamelinensis]	2,00E-06	88	no	CEI76402.1|Phenol hydroxylase [Pseudomonas aeruginosa]	1,00E-10	49
#33705	**WP_018941820.1|membrane protein [Polaribacter irgensii]**	1,00E-87	86	**yes**	**WP_018941820.1|membrane protein [Polaribacter irgensii]**	1,00E-88	48
#33058	WP_021862155.1|aGAP012078-PA [Parabacteroides johnsonii CAG:246]	2,00E-05	86	no	SKN50229.1|cellobiose phosphorylase [Mycobacteroides abscessus subsp. massiliense]	2,00E-24	96
#2453	**WP_001356433.1|membrane protein [Escherichia coli]**	5,00E-19	86	**yes**	**WP_001356433.1|hypothetical protein [Escherichia coli]**	1,00E-19	86
#26803	WP_022173511.1|phosphoglycerate kinase [Bifidobacterium bifidum CAG:234]	3,00E-16	85	no	CCK15843.1|Peptidyl-prolyl cis-trans isomerase [Cronobacter universalis NCTC 9529]	2,00E-271	68
#69868	**WP_001230998.1|swarming motility protein [Escherichia albertii]**	5,00E-15	85	**yes**	**WP_001230998.1|MULTISPECIES: DUF1768 domain-containing protein [Escherichia]**	2,00E-15	85
#52463	**WP_001511455.1|NADH dehydrogenase [Escherichia coli]**	3,00E-04	85	**yes**	**WP_001511455.1|NADH dehydrogenase [Escherichia coli]**	7,00E-04	85
#47080	WP_022026487.1|choline dehydrogenase [Dialister invisus CAG:218]	2,00E-22	84	no	CDA48120.1|choline dehydrogenase [Dialister sp. CAG:486]	2,00E-25	56
#16688	WP_021939429.1|protein of PilT N-term./Vapc superfamily [Bacteroides eggerthii CAG:109]	1,00E-07	82	no; only hypothetical proteins	CCY55369.1|uncharacterized protein of PilT N-term./Vapc superfamily [Bacteroides eggerthii CAG:109]	2,00E-07	81
#29875	WP_007778793.1|transcription accessory protein (S1 RNA-binding domain) [Cronobacter malonaticus]	2,00E-16	81	no	CCJ94451.1|Transcription accessory protein (S1 RNA-binding domain) [Cronobacter malonaticus 681]	4,00E-16	81
#3315	WP_022103919.1|dehydrogenase [Bacteroides stercoris CAG:120]	6,00E-07	80	no	CEL33240.1|Glycosyltransferase family 28 C-terminal domain protein (modular protein) [Xanthomonas citri pv. citri]	1,00E-09	95
#26616	WP_022210669.1|metal-dependent RNase [Bacteroides cellulosilyticus CAG:158]	1,00E-18	79	no	ENO95326.1|putative metal-dependent RNase [Thauera phenylacetica B4P]	2,00E-45	97
#51552	WP_044327615.1|porphobilinogen deaminase [Citrobacter amalonaticus]	1,00E-10	79	no	WP_080776275.1|hydroxymethylbilane synthase [Citrobacter amalonaticus]	4,00E-13	85
#23719	WP_004970659.1|NAD-specific glutamate dehydrogenase [Haloferax denitrificans]	7,00E-04	78	no	CSP54182.1|NAD-specific glutamate dehydrogenase [Shigella sonnei]	3,00E-82	92
#23470	WP_022388214.1|cytosine deaminase [Ruminococcus obeum CAG:39]	3,00E-13	76	no	EHS35468.1|cytosine deaminase [Pseudomonas aeruginosa MPAO1/P2]	1,00E-56	51
#71575	WP_026811502.1|murein hydrolase transporter LrgA [Arenibacter latericius]	4,00E-04	76	no; only hypothetical proteins	CSR69080.1|Uncharacterised protein [Shigella sonnei]	2,00E-79	100
#63224	WP_001511305.1|ATP-dependent transporter SufC domain protein, partial [Escherichia coli]	2,00E-08	75	no	EKW78207.1|putative ATP-dependent transporter SufC domain protein [Escherichia coli 97.1742]	1,00E-08	74
#76450	WP_006112961.1|NAD-specific glutamate dehydrogenase [Halorubrum coriense]	1,00E-41	74	no	ABE05525.1|putative glutamate dehydrogenase [Escherichia coli UTI89]	0	100
#71365	WP_046371832.1|cold-shock protein [Erwinia tracheiphila]	1,00E-25	74	no	AHA63536.1|CrcB family protein [Shigella dysenteriae 1617]	2,00E-38	79
#57648	**WP_032437443.1|lipid A biosynthesis (KDO)2-(lauroyl)-lipid IVA acyltransferase [Klebsiella pneumoniae]**	8,00E-06	74	**yes**	**WP_032437443.1|lauroyl-Kdo(2)-lipid IV(A) myristoyltransferase [Klebsiella pneumoniae]**	7,00E-06	73
#46888	**WP_007344964.1|DNA topoisomerase I [Halorubrum terrestre]**	2,00E-13	73	**yes**	**WP_007344964.1|DNA topoisomerase I [Halorubrum terrestre]**	1,00E-13	73
#48717	WP_014913562.1|hydrogenase expression protein [Nocardiopsis alba]	1,00E-06	73	no	AFR11110.1|MMPL family protein [Nocardiopsis alba ATCC BAA-2165]	6,00E-07	73
#37776	WP_044815295.1|ubiquinone biosynthesis protein UbiB, partial [Escherichia coli]	4,00E-06	73	no	no hit		
#6794	WP_021994652.1|6-pyruvoyl-tetrahydropterin synthase [Sutterella wadsworthensis CAG:135]	5,00E-08	72	no	CCZ17254.1|6-pyruvoyl-tetrahydropterin synthase [Sutterella wadsworthensis CAG:135]	5,00E-08	72
#40779	WP_006090480.1|phenol hydroxylase [Natronorubrum tibetense]	2,00E-10	70	no	CEI76402.1|Phenol hydroxylase [Pseudomonas aeruginosa]	1,00E-27	69
#55029	**WP_042099490.1|regulator [Escherichia coli] (2018: hypothetical protein)**	1,00E-08	70	**yes**	**WP_042099490.1|hypothetical protein [Escherichia coli]**	2,00E-08	96
#5905	**WP_042099490.1|regulator [Escherichia coli]**	3,00E-07	70	**yes**	**WP_042099490.1|hypothetical protein [Escherichia coli]**	5,00E-07	70
#8873	**WP_042099490.1|regulator [Escherichia coli]**	1,00E-08	70	**yes**	**WP_042099490.1|hypothetical protein [Escherichia coli]**	2,00E-08	70
#41300	WP_022173867.1|elongation factor TU [Bifidobacterium bifidum CAG:234]	1,00E-53	68	no	EST14915.1|elongation factor Tu domain protein [Pseudomonas putida S610]	2,00E-98	86
#48354	WP_044818234.1|membrane protein [Escherichia coli]	4,00E-08	68	no	CFN63089.1|Uncharacterised protein [Bordetella pertussis]	8,00E-10	95

ID CP008957.1_	blastp 2015			Hit still present?	blastp 2018		
	blast hit	E-value	coverage		Top hit 2018 (with same product)	E-value	coverage
#578	WP_004970766.1|Flp pilus assembly protein TadG [Haloferax denitrificans]	8,00E-06	67	no	CEE78955.1|Flp pilus assembly protein TadG [Xanthomonas citri pv. citri]	2,00E-17	100
#574	WP_044794817.1|TetR family transcriptional regulator [Escherichia coli]	2,00E-05	67	no	EIE47447.1|Flp pilus assembly protein TadG [Pseudomonas aeruginosa PADK2_CF510]	4,00E-06	97
#38753	WP_001297239.1|membrane protein [Escherichia coli]	1,00E-06	67	no	ENF07497.1|putative membrane protein [Escherichia coli P0304777.9]	9,00E-07	67
#52043	**WP_044863991.1|heme ABC transporter ATP-binding protein [Escherichia coli]**	6,00E-05	67	**yes**	**WP_044863991.1|sugar ABC transporter ATP-binding protein [Escherichia coli]**	4,00E-05	66
#64431	WP_044818628.1|isopropylmalate isomerase, partial [Escherichia coli]	1,00E-15	65	no; only hypothetical proteins	CGA43153.1|Uncharacterised protein [Salmonella enterica subsp. enterica serovar Typhi]	1,00E-16	100
#42559	WP_044806910.1|type VI secretion protein, partial [Escherichia coli]	6,00E-10	63	no	no hit		
#57612	WP_009624052.1|aldehyde dehydrogenase, partial [Escherichia coli]	4,00E-10	62	no	EKW85649.1|aldehyde dehydrogenase family protein [Escherichia coli 97.0007]	2,00E-10	61
#60920	WP_032334060.1|lysozyme [Escherichia coli]	8,00E-06	62	no	KHO57510.1|lysozyme [Escherichia coli]	2,00E-05	61
#14606	WP_006090480.1|phenol hydroxylase [Natronorubrum tibetense]	1,00E-04	61	no	ABA49062.1|phenol hydroxylase, putative [Burkholderia pseudomallei 1710b]	4,00E-18	99
#29479	WP_022173867.1|elongation factor TU [Bifidobacterium bifidum CAG:234]	2,00E-55	60	no	EAQ75389.1| ISxac3 transposase [Synechococcus sp. WH 5701]	6,00E-76	88
#21460	WP_022122124.1|val start codon [Prevotella copri CAG:164]	2,00E-31	60	no	BAM28612.1|gyrase subunit A [Escherichia coli]	7,00E-41	94
#3572	WP_000512627.1|acetyltransferase [Bacillus thuringiensis]	5,00E-09	60	no	WP_000512627.1|MULTISPECIES: hypothetical protein [Bacillus cereus group]	3,00E-09	67
#22202	WP_044818288.1|recombinase, partial [Escherichia coli]	2,00E-14	60	no	no hit		
#56556	WP_044722053.1|cell surface protein [Escherichia coli]	7,00E-33	58	no	WP_071536639.1|cell surface protein [Escherichia coli]	2,00E-53	75
#8258	WP_044722053.1|cell surface protein [Escherichia coli]	7,00E-33	58	no	WP_071536639.1|cell surface protein [Escherichia coli]	2,00E-53	100
#57364	**WP_010723106.1|acid-inducible small membrane-associated protein [Escherichia coli]**	7,00E-11	58	**yes**	**WP_010723106.1|MULTISPECIES: stress response protein AzuC [Proteobacteria]**	3,00E-11	58
#45349	WP_044716407.1|transcriptional regulator [Escherichia coli]	9,00E-05	58	no	no hit		
#5113	WP_044814954.1|transcriptional regulator [Escherichia coli]	3,00E-04	57	no	no hit		
#29471	**WP_043018543.1|ferredoxin [Citrobacter freundii]**	7,00E-15	56	**yes**	**WP_043018543.1|MULTISPECIES: hypothetical protein [Citrobacter]**	3,00E-15	78
#71581	WP_040209525.1|enterobactin synthase subunit E, partial [Klebsiella pneumoniae]	2,00E-06	55	no	WP_077257667.1|2,3-dihydroxybenzoate-AMP ligase [Klebsiella pneumoniae]	3,00E-06	54
#44244	**WP_012602583.1|reverse transcriptase [Escherichia coli]**	3,00E-21	54	**yes**	**WP_012602583.1|RNA-directed DNA polymerase [Escherichia coli]**	5,00E-21	54
#46160	**WP_023142319.1|RNA 3'-terminal phosphate cyclase [Escherichia coli]**	6,00E-07	54	**yes**	**WP_023142319.1|RNA 3'-terminal phosphate cyclase [Escherichia coli]**	1,00E-06	54
#11186	WP_040175544.1|tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase [Klebsiella pneumoniae]	8,00E-04	53	no; only hypothetical proteins	CSP98036.1|Uncharacterised protein [Shigella sonnei]	9,00E-58	100
#54498	WP_044817398.1|diguanylate cylase [Escherichia coli]	1,00E-04	53	no; only hypothetical proteins	SRN48511.1|Uncharacterised protein [Shigella flexneri]	8,00E-45	100
#58562	WP_040947702.1|transcriptional regulator [Coxiella burnetii]	2,00E-04	52	no	AIO69498.1|atalase-2 domain protein [Burkholderia oklahomensis]	3,00E-12	92
#12194	WP_044722053.1|cell surface protein [Escherichia coli]	1,00E-28	52	no	WP_001453714.1|hypothetical protein [Escherichia coli]	2,00E-53	67
#59703	WP_044722053.1|cell surface protein [Escherichia coli]	7,00E-30	52	no	OKV69418.1|cell surface protein [Escherichia coli]	2,00E-52	67
#63696	WP_035716401.1|trehalase [Gramella echinicola]	2,00E-04	52	no	ADD56055.1|trehalase [Escherichia coli O55:H7 str. CB9615]	8,00E-83	99
#20182	WP_040221869.1|histidine kinase, partial [Klebsiella pneumoniae]	6,00E-07	52	no	no hit		
#47455	WP_001675934.1|3-deoxy-D-manno-octulosonate 8-phosphate phosphatase KdsC, partial [Escherichia coli]	1,00E-12	52	no	ENG75893.1|3-deoxy-D-manno-octulosonate 8-phosphate phosphatas KdsC domain protein [Escherichia coli p0305293.9]	5,00E-13	52
#16072	WP_042782087.1|alpha/beta hydrolase, partial [Vibrio parahaemolyticus]	4,00E-07	51	no	WP_080154881.1|alpha/beta hydrolase [Salmonella enterica]	4,00E-12	64
#4322	WP_047667856.1|membrane protein, partial [Escherichia coli]	8,00E-77	50	no	ERE37650.1|putative membrane domain protein [Escherichia coli B90]	0	99
#11825	WP_044808772.1|electron transporter RsxA, partial [Escherichia coli]	7,00E-08	50	no; only hypothetical proteins	STU70850.1|Uncharacterised protein [Klebsiella pneumoniae subsp. ozaenae]	3,00E-56	92
#61592	WP_044815706.1|permease, partial [Escherichia coli]	6,00E-10	49	no; only hypothetical proteins	CSP98403.1|Uncharacterised protein [Shigella sonnei]	9,00E+35	100
#50186	**WP_032245923.1|amino acid acetyltransferase [Escherichia coli]**	4,00E-35	48	**yes**	**WP_032245923.1|hypothetical protein [Escherichia coli]**	**7,00E-36**	**48**
#60306	WP_044815515.1|transaldolase, partial [Escherichia coli]	8,00E-11	48	no	no hit		
#37389	WP_032292028.1|nitrite extrusion protein 2 [Escherichia coli]	2,00E-06	48	no; only hypothetical proteins	CSQ01281.1|Uncharacterised protein [Shigella sonnei]	2,00E-23	100
#4309	WP_047667856.1|membrane protein, partial [Escherichia coli]	8,00E-81	47	no	EHV17674.1|YD repeat domain protein [Escherichia coli DEC4E]	0	100
#51814	WP_005468326.1|thiol:disulfide interchange protein DsbC [Vibrio parahaemolyticus]	5,00E-04	47	no; only hypothetical proteins	SWC21767.1| Uncharacterised protein [Klebsiella pneumoniae]	1,00E-67	100
#75085	WP_022512348.1|na /proline symporter [Clostridium clostridioforme CAG:511]	1,00E-19	45	no	CDC85086.1|na /proline symporter [Escherichia coli CAG:4]	7,00E-167	90

ID CP008957.1_	blastp 2015			Hit still present?	blastp 2018		
	blast hit	E-value	coverage		Top hit 2018 (with same product)	E-value	coverage
#62834	WP_001511305.1|ATP-dependent transporter SufC domain protein, partial [Escherichia coli]	5,00E-16	45	no; only hypothetical proteins	WP_087893104.1|hypothetical protein [Escherichia coli]	6,00E-52	99
#44380	WP_032325984.1|glycosyl transferase [Escherichia coli]	1,00E-08	45	no; only hypothetical proteins	EFX09132.1|hypothetical protein ECO5101_20760 [Escherichia coli O157:H7 str. G5101]	1,00E-31	100
#47226	WP_044707562.1|glucarate transporter, partial [Escherichia coli]	2,00E-06	45	no	WP_077785119.1|MFS transporter [Escherichia coli]	1,00E-07	49
#18938	WP_001511305.1|ATP-dependent transporter SufC domain protein, partial [Escherichia coli]	2,00E-16	44	no; only hypothetical proteins	WP_075702026.1|hypothetical protein [Escherichia coli]	2,00E-57	100
#64799	WP_001511305.1|ATP-dependent transporter SufC domain protein, partial [Escherichia coli]	2,00E-16	44	no; only hypothetical proteins	WP_075702026.1|hypothetical protein [Escherichia coli]	9,00E-64	100
#52104	WP_044710525.1|carboxymethylenebutenolidase, partial [Escherichia coli]	5,00E-05	44	no	no hit		
#33054	WP_021862155.1|aGAP012078-PA [Parabacteroides johnsonii CAG:246]	1,00E-15	43	no	XP_320450.4|AGAP012078-PA [Anopheles gambiae str. PEST]	1,00E-31	88
#46292	WP_044817181.1|histidine kinase, partial [Escherichia coli]	4,00E-17	43	no	no hit		
#52405	WP_001420173.1|mobilization protein mbeA, partial [Escherichia coli]	5,00E-04	43	no	ENB46886.1|putative mobilization protein mbeA [Escherichia coli MP021561.3]	7,00E-04	43
#63300	WP_044726088.1|DNA topoisomerase I, partial [Escherichia coli]	1,00E-04	43	no	no hit		
#7962	WP_038427453.1|transcriptional regulator, partial [Escherichia coli]	3,00E-05	43	no	no hit		
#59705	WP_024181921.1|cell surface protein [Escherichia coli]	1,00E-60	42	no	EGI36713.1|putative cell surface protein [Escherichia coli TA271]	3,00E-106	74
#65197	WP_044697000.1|conjugal transfer protein, partial [Escherichia coli]	2,00E-06	42	no	no hit		
#52317	WP_042781765.1|membrane protein, partial [Vibrio parahaemolyticus]	1,00E-12	41	no	SQR09665.1|Permease, cytosine/purine, uracil, thiamine, allantoin family [Escherichia coli]	9,00E-35	60
#32499	WP_044714707.1|acyl-CoA thioesterase [Escherichia coli]	5,00E-06	41	no; only hypothetical proteins	CSS74283.1|Uncharacterised protein [Shigella sonnei]	7,00E-49	100
#5028	WP_044817630.1|aminopeptidase, partial [Escherichia coli]	4,00E-05	40	no; only hypothetical proteins	CRR62018.1|hypothetical protein PAERUG_P48_London_17_VIM_2_01_13_00931 [Pseudomonas aeruginosa]	2,00E-28	79
#19362	WP_032339336.1|phosphomannomutase, partial [Escherichia coli]	1,00E-15	39	no; only hypothetical proteins	CSP62188.1|Uncharacterised protein [Shigella sonnei]	5,00E-56	100
#24390	WP_001732663.1|cytochrome c-type biogenesis heme exporter protein C, partial [Salmonella enterica]	9,00E-05	39	no	ELP06542.1|cytochrome c-type biogenesis heme exporter protein C [Salmonella enterica subsp. enterica serovar Enteritidis str. 50-5646]	1,00E-04	38
#34705	WP_001408502.1|adhesin [Escherichia coli]	1,00E-05	39	no	OSL93369.1|adhesin [Escherichia coli T426]	3,00E-06	39
#58267	WP_029097005.1|MFS transporter [Budvicia aquatica]	1,00E-10	37	no	STQ29106.1|metabolite transport protein [Escherichia coli]	6,00E-33	61
#59154	WP_011799900.1|prolyl-tRNA synthetase [Polaromonas naphthalenivorans]	4,00E-04	36	no	CSP62758.1|Predicted Fe-S protein [Shigella sonnei]	3,00E-90	97
#75208	WP_044698966.1|ribonucleotide-diphosphate reductase subunit beta, partial [Escherichia coli]	1,00E-09	36	no	no hit		
#58214	WP_022104319.1|glyceraldehyde-3-phosphate dehydrogenase [Bacteroides stercoris CAG:120]	2,00E-18	35	no	AWR89628.1|glyceraldehyde-3-phosphate dehydrogenase A [Pectobacterium zantedeschiae]	2,00E-60	54
#58060	**WP_001467889.1|LysR family transcriptional regulator [Escherichia coli]**	4,00E-25	35	**yes**	**WP_001467889.1|LysR family transcriptional regulator [Escherichia coli]**	1,00E-25	34
#23633	WP_044688508.1|citrate transporter [Escherichia coli]	5,00E-09	35	no	no hit		
#44466	WP_045714027.1|tyrosine protein kinase, partial [Salmonella enterica]	2,00E-06	33	no	KJT45791.1|tyrosine kinase [Salmonella enterica subsp. enterica serovar Heidelberg str. 607310-1]	8,00E-06	33
#21079	WP_001658585.1|aldose-1-epimerase [Escherichia coli]	2,00E-17	33	no	RIL20340.1|aldose epimerase [Escherichia coli]	4,00E-17	33
#23115	WP_005017596.1|glutamate decarboxylase isozyme [Shigella dysenteriae]	1,00E-05	33	no	WP_005017596.1|hypothetical protein [Shigella dysenteriae]	1,00E-05	33
#51542	WP_035894939.1|heat-shock protein GrpE [Kluyvera ascorbata]	1,00E-12	32	no	CUU94843.1|Protein GrpE (fragment) [Escherichia coli]	2,00E-103	100
#10580	WP_044814876.1|ABC transporter ATP-binding protein, partial [Escherichia coli]	2,00E-05	32	no	no hit		
#68409	WP_044807518.1|formate dehydrogenase, partial [Escherichia coli]	1,00E-04	30	no; only hypothetical proteins	SAJ34031.1|Uncharacterised protein [Enterobacter cloacae]	2,00E-40	100
#1196	WP_001511318.1|lysE type translocator family protein [Escherichia coli]	0,001	30	no	EKW77792.1|lysE type translocator family protein [Escherichia coli 97.1742]	1,00E-03	29
#76186	WP_040188637.1|potassium transporter KefC, partial [Klebsiella pneumoniae]	2,00E-12	29	no	CCJ83841.1|cell wall surface anchor family protein [Cronobacter dublinensis 582]	7,00E-62	99
#34642	WP_044717907.1|Lom family protein [Escherichia coli]	8,00E-07	29	no	no hit		
#23928	WP_044814946.1|thiamine ABC transporter permease [Escherichia coli]	8,00E-07	29	no	CUW46219.1|Methionine aminotransferase [Komagataeibacter xylinus]	3,00E-11	95
#56553	WP_024181921.1|cell surface protein [Escherichia coli]	3,00E-16	28	no	OKU49113.1|cell surface protein [Escherichia coli]	9,00E-17	28
#8255	WP_024181921.1|cell surface protein [Escherichia coli]	3,00E-16	28	no	OKU49113.1|cell surface protein [Escherichia coli]	8,00E-17	28
#65404	WP_044819142.1|deoxyuridine 5'-triphosphate nucleotidohydrolase, partial [Escherichia coli]	3,00E-06	28	no	no hit		

ID CP008957.1_	blastp 2015			Hit still present?	blastp 2018		
	blast hit	E-value	coverage		Top hit 2018 (with same product)	E-value	coverage
#36942	WP_044728640.1|histidine kinase, partial [Escherichia coli]	1,00E-04	27	no	WP_072258598.1|PTS trehalose transporter subunit IIBC [Escherichia coli]	7,00E-05	28
#6386	WP_044712568.1|L-asparaginase II, partial [Escherichia coli]	1,00E-14	26	no; only hypothetical proteins	STG53894.1|Uncharacterised protein [Escherichia coli]	1,00E-47	55
#30203	WP_044812898.1|sulfur acceptor protein CsdL, partial [Escherichia coli]	2,00E-05	26	no	KGQ13425.1|inner membrane protein yhgN [Beauveria bassiana D1-5]	7,00E-22	58
#17676	WP_045176425.1|glycerol-3-phosphate acyltransferase [Escherichia coli]	2,00E-07	25	no	no hit		
#72546	WP_001747952.1|inosine/guanosine kinase, partial [Salmonella enterica]	6,00E-08	23	no	ELP14553.1|inosine/guanosine kinase [Salmonella enterica subsp. enterica serovar Agona str. SH08SF124]	1,00E-07	22
#46842	WP_044327237.1|propionyl-CoA synthetase, partial [Citrobacter amalonaticus]	4,00E-04	23	no; only hypothetical proteins	CSG40775.1|Uncharacterised protein [Shigella sonnei]	6,00E-41	60
#68105	WP_040174565.1|cell division protein MukB, partial [Klebsiella pneumoniae]	9,00E-12	22	no; only hypothetical proteins	CSP76618.1|Uncharacterised protein [Shigella sonnei]	0	100
#24410	**WP_001296311.1|inorganic polyphosphate kinase [Escherichia coli]**	8,00E-24	21	**yes**	**WP_001296311.1|hypothetical protein [Escherichia coli]**	3,00E-24	21
#11769	WP_041930714.1|transcriptional regulator [Pantoea ananatis]	2,00E-05	21	no	AFJ28690.1|hypothetical protein CDCO157_1633 [Escherichia coli Xuzhou21]	2,00E-108	100
#25203	WP_046424497.1|carbamoyl dehydratase HypE [Escherichia coli]	5,00E-06	20	**yes**	**WP_046424497.1|hydrogenase expression/formation protein HypE [Escherichia coli]**	**2,00E-05**	**20**
#76497	WP_038814293.1|4'-phosphopantetheinyl transferase [Shigella dysenteriae]	5,00E-04	19	no	STH34021.1|enterobactin synthase multienzyme complex phosphopantetheinyltransferase [Escherichia coli]	2,00E-95	93
#31430	WP_044815906.1|oxidoreductase, partial [Escherichia coli]	2,00E-07	19	no	WP_081987797.1|TraB/GumN family protein [Sphingomonas sp. 37zxx]	6,00E-11	78
#9224	WP_040237279.1|FMN reductase [Klebsiella pneumoniae]	3,00E-04	19	no; only hypothetical proteins	EJK91382.1|hypothetical protein UUU_14070 [Klebsiella pneumoniae subsp. pneumoniae DSM 30104]	6,00E-44	95
#30428	WP_001412659.1|membrane protein [Shigella flexneri]	3,00E-13	18	no	ENE03198.1|putative membrane protein [Escherichia coli P0304799.3]	5,00E-29	29
#75851	WP_001613530.1|exodeoxyribonuclease V alpha chain [Salmonella enterica]	1,00E-05	18	no	EHC46058.1|Exodeoxyribonuclease V alpha chain [Salmonella enterica subsp. enterica serovar Give str. S5-487]	3,00E-05	17
#30206	WP_044819364.1|transcriptional regulator MalT, partial [Escherichia coli]	4,00E-10	16	no	CSG42545.1|Uncharacterised protein [Shigella sonnei]	5,00E-91	88
#61974	WP_044816265.1|transcription termination factor Rho, partial [Escherichia coli]	9,00E-06	16	no; only hypothetical proteins	CSP86261.1|Uncharacterised protein [Shigella sonnei]	8,00E-95	86
#10327	WP_032273022.1|lysine decarboxylase LdcC, partial [Escherichia coli]	3,00E-04	16	no; only hypothetical proteins	CSR67463.1|Uncharacterised protein [Shigella sonnei]	1,00E-90	100
#4733	WP_001651897.1|NrfD protein [Salmonella enterica]	2,00E-05	14	no	CEE04850.1|citrate lyase alpha chain domain protein [Escherichia coli]	2,00E-79	94
#68123	WP_044816749.1|multidrug transporter, partial [Escherichia coli]	8,00E-05	9	no; only hypothetical proteins	KFZ99951.1|hypothetical protein DP20_3679 [Shigella flexneri]	0	94
#39725	WP_001511392.1|amino acid permease, partial [Escherichia coli]	6,00E-05	8	no	EKW75662.1|amino acid permease family protein [Escherichia coli 97.1742]	8,00E-05	7
#54145	WP_040222945.1|ribonucleotide-diphosphate reductase subunit alpha, partial [Klebsiella pneumoniae]	9,00E-06	7	no; only hypothetical proteins	SPW47680.1|Uncharacterised protein [Escherichia coli]	0	48

Supplementary table S9: Shadow ORFs matching with their full-length to the blastp hit. The blast search was conducted in February 2016. A re-blast confirmed 225 (in September 2018) of 280 (in 2016) full-length matching proteins. A full-length blast hit is defined to have a query coverage of ≥80%.

Shadow ORF						mother gene		
ID CP008957.1_	start position	stop position	sORF length (bp)	overlap length (bp)	blastp hit	locus tag EDL933_	mORF product	mORF length (bp)
#10006	1517477	1517752	276	174	hypothetical protein [E. coli]	#1566	hypothetical protein	192
#10163	1542047	1542367	321	246	hypothetical protein [Pseudomonas aeruginosa]	#1601	Transposase	276
#10211	1549415	1549615	201	195	hypothetical protein [Escherichia coli]	#1614	Transcriptional regulator CsgD for 2nd curli operon	651
#10317	1565422	1565589	168	168	hypothetical protein [Escherichia coli]	#1633	Cytochrome B561	567
#10782	1636651	1636968	318	183	membrane protein [E. coli]	#1713	hypothetical protein	183
#10796	1639073	1639513	441	180	hypothetical protein [Plesiomonas shigelloides]	#1719	hypothetical protein	180
#11476	1747600	1747788	189	107	hypothetical protein [Escherichia coli]	#1860	Protease VII (Omptin) precursor	954
#11769	1785186	1785662	477	361	PTS-dependent dihydroxyacetone kinase operon transcriptional regulator DhaR [E. coli]	#1905	Putative adhesion and penetration protein	1644
#12060	1827141	1827443	303	210	hypothetical protein [Plautia stali symbiont]	#1943	Alcohol dehydrogenase	2676
#12257	1855181	1855480	300	100	hypothetical protein [Plesiomonas shigelloides]	#1979	hypothetical protein	132
#1258	191421	191618	198	198	hypothetical protein [Escherichia coli]	#0170	[Protein-PII] uridylyltransferase	2673
#1277	194120	194218	99	99	hypothetical protein [Franconibacter helveticus]	#0173	hypothetical protein	288
#13401	2046258	2047103	846	846	type IV section protein [Trabulsiella odontotermitis]	#2189	core protein	4203
#13407	2047320	2048111	792	792	hypothetical protein [E. coli]	#2189	core protein	4203
#13417	2049084	2049308	225	225	hypothetical protein [Escherichia coli]	#2189	core protein	4203
#13440	2051511	2051954	444	444	hypothetical protein [Shigella sonnei]	#2190	VgrG protein	2109
#13509	2062553	2062831	279	129	hypothetical protein, partial [Escherichia coli]	#2204	hypothetical protein	222
#13570	2070431	2070772	342	105	hypothetical protein [Escherichia coli]	#2212	hypothetical protein	438
#13900	2111183	2111416	234	234	hypothetical protein [E. coli]	#2247	porin, autotransporter (AT) family	3036
#14171	2148275	2148574	300	270	hypothetical protein, partial [Salmonella enterica]	#2287	putative Dnase	366
#14229	2154535	2155161	627	168	hypothetical protein [E. coli]	#2298	hypothetical protein	168
#14888	2243928	2244314	387	246	hypothetical protein [Escherichia coli]	#2399	Transcriptional regulator, TetR family	567
#15102	2272059	2272226	168	97	hypothetical protein [Shigella flexneri]	#2426	putative metal-dependent phosphoesterases (PHP family)	882
#15556	2334024	2334524	501	168	hypothetical protein [E. coli]	#2507	hypothetical protein	168
#16693	2495659	2496153	495	495	hypothetical protein [Klebsiella pneumoniae]	#2678	Threonyl-tRNA synthetase	1929
#16998	2537822	2538166	345	345	hypothetical protein [Escherichia coli]	#2722	hypothetical protein	1041
#17453	2603185	2603769	585	168	hypothetical protein [E. coli]	#2794	Ribosomal RNA large subunit methyltransferase A	810
#1794	271929	272111	183	128	hypothetical protein [Escherichia coli]	#0242	hypothetical protein	147
#17968	2671340	2671756	417	369	hypothetical protein [Escherichia coli]	#2863	Signal transduction histidine kinase CheA	1959
#18076	2683836	2684159	324	114	hypothetical protein, partial [Neisseria flavescens]	#2879	hypothetical protein	114
#18869	2791305	2791805	501	168	hypothetical protein [E. coli]	#3024	hypothetical protein	168
#1887	287289	287966	678	678	hypothetical protein [E. coli]	#0262	Flagellar biosynthesis protein FlhA	1740
#19397	2868678	2868908	231	231	hypothetical protein [Escherichia coli]	#3106	GDP-mannose 4,6-dehydratase	1119
#19520	2886875	2887345	471	471	hypothetical protein [Enterobacter cloacae]	#3122	Mannose-1-phosphate guanylyltransferase (GDP)	1437

Shadow ORF						mother gene		
ID CP008957.1_	start position	stop position	sORF length (bp)	overlap length (bp)	blastp hit	locus tag EDL933_	mORF product	mORF length (bp)
#19712	2913277	2913534	258	258	hypothetical protein [E. coli]	#3146	Putative chaperonin	1941
#19821	2933291	2933542	252	249	hypothetical protein [Escherichia coli]	#3159	Galactitol utilization operon repressor	735
#20653	3043633	3044016	384	276	hypothetical protein [Shigella sp. SF-2015]	#3293	3-oxoacyl-[acyl-carrier protein] reductase	762
#21079	3103493	3103855	363	240	aldose-1-epimerase [E. coli]	#3350	Ribosomal small subunit pseudouridine synthase A	696
#21291	3131424	3131663	240	240	inverted ada-Golga3 fusion protein [Escherichia coli]	#3379	ADA regulatory protein	1065
#22524	3297005	3297343	339	314	hypothetical protein [Xenorhabdus poinarii]	#3535	Inner membrane component of tripartite multidrug resistance system	1539
#22698	3322215	3322436	222	222	hypothetical protein [Escherichia coli]	#3555	PTS system, fructose-specific IIBC component	1248
#22778	3333085	3333282	198	198	hypothetical protein [E. coli]	#3565	putative virulence protein	1149
#22782	3333635	3333883	249	249	hypothetical protein [Shigella flexneri]	#3565	putative virulence protein	1149
#23185	3388112	3388564	453	429	hypothetical protein [Enterobacter ludwigii]	#3622	Glutamate synthase [NADPH] small chain	1980
#23693	3462282	3462632	351	197	hypothetical protein [E. coli]	#3684	Protein SseB	777
#23965	3498168	3498587	420	188	hypothetical protein [Citrobacter koseri]	#3721	Putative sensor-like histidine kinase YfhK	1428
#23972	3498892	3499791	900	861	phosphoribosylformylglycinamidine synthase [E. coli]	#3722	Phosphoribosylformylglycinamidine synthase, synthetase subunit	3888
#25098	3655946	3657766	1821	117	hypothetical protein, partial [Bacillus cereus]	#3889	Formate hydrogenlyase complex 3 iron-sulfur protein	543
#25122	3660988	3661368	381	172	hypothetical protein [Shigella flexneri]	#3893	Formate hydrogenlyase subunit 2	612
#26101	3793954	3794166	213	198	hypothetical protein [Escherichia coli]	#4024	4-deoxy-L-threo-5-hexosulose-uronate ketol-isomerase	837
#26608	3870926	3871561	636	636	hypothetical protein [Salmonella enterica]	#4105	Glycine dehydrogenase [decarboxylating] (glycine cleavage system P protein)	2874
#26893	3908097	3909866	1770	1770	hypothetical protein [Klebsiella pneumoniae]	#4144	Transketolase	1992
#26908	3910961	3912004	1044	891	hypothetical protein, partial [Bacillus cereus]	#4146	Agmatinase	921
#26939	3915284	3915514	231	117	hypothetical protein [Shigella dysenteriae]	#4150	hypothetical protein	144
#27372	3978583	3978708	126	107	hypothetical protein [E. coli]	#4226	hypothetical protein	153
#27374	3978705	3978950	246	150	hypothetical protein [E. coli]	#4227	Biopolymer transport protein ExbD/TolR	426
#28096	4081177	4081437	261	261	hypothetical protein [Escherichia coli]	#4328	LysR-family transcriptional regulator YhaJ	897
#28643	4162631	4163053	423	213	hypothetical protein [Serratia sp. TEL]	#4413	LSU ribosomal protein L27p	258
#28649	4163050	4163637	588	312	hypothetical protein [Acidovorax radicis]	#4414	LSU ribosomal protein L21p	312
#29446	4273020	4274075	1056	822	hypothetical protein [Nephila clavipes]	#4534	LSU ribosomal protein L2p (L8e)	822
#29453	4274855	4275556	702	621	hypothetical protein [Klebsiella pneumoniae]	#4537	LSU ribosomal protein L3p (L3e)	630
#29471	4277328	4277576	249	168	ferredoxin [Cronobacter dubliensis]	#4541	Bacterioferritin-associated ferredoxin	195
#29479	4277768	4279006	1239	1095	Uncharacterised protein [Serratia marcescens]	#4542	Translation elongation factor Tu	1185
#29721	4315431	4315745	315	306	hypothetical protein, partial [Escherichia coli]	#4585	Methyl-directed repair DNA adenine methylase	837
#29735	4316824	4317144	321	306	hypothetical protein [Cronobacter sakazakii]	#4586	DamX, an inner membrane protein involved in bile resistance	1287
#29869	4335378	4335644	267	267	hypothetical protein [Enterobacter aerogenes]	#4604	Osmolarity sensory histidine kinase EnvZ	1353
#30165	4378141	4378500	360	165	hypothetical protein, partial [Escherichia coli]	#4639	Aspartate-semialdehyde dehydrogenase	1104
#30215	4384000	4384632	633	435	hypothetical protein [Escherichia coli]	#4647	Gluconokinase	489
#30643	4449603	4450355	753	753	hypothetical protein [Escherichia coli]	#4722	ABC-type multidrug transport system, permease component	2736
#30955	4495934	4496524	591	585	hypothetical protein, partial [Escherichia coli]	#4769	Glutamate decarboxylase	1401

Shadow ORF						mother gene		
ID CP008957.1_	start position	stop position	sORF length (bp)	overlap length (bp)	blastp hit	locus tag EDL933_	mORF product	mORF length (bp)
#30961	4496618	4497193	576	576	hypothetical protein [Escherichia coli]	#4769	Glutamate decarboxylase	1401
#31337	4546352	4546594	243	243	hypothetical protein [Erwinia amylovora]	#4812	Putative resistance protein	1203
#31654	4589667	4590140	474	474	hypothetical protein [Shigella sonnei]	#4850	Selenocysteine-specific translation elongation factor	1845
#31968	4642568	4642924	357	168	hypothetical protein [Pectobacterium carotovorum]	#4897	LSU ribosomal protein L33p, zinc-independent	168
#32035	4652571	4652864	294	204	hypothetical protein, partial [Escherichia coli]	#4910	hypothetical protein	258
#32068	4659368	4659598	231	231	hypothetical protein [Escherichia coli]	#4916	putative cytoplasmic protein	867
#32298	4691364	4691567	204	204	hypothetical protein [Escherichia coli]	#4954	hypothetical protein	507
#32339	4697204	4697503	300	122	type III secretion system protein SepZ [Escherichia albertii]	#4960	hypothetical protein	144
#32485	4721636	4721911	276	276	hypothetical protein, partial [Bacillus cereus]	#4990	Hexose phosphate transport protein UhpT	1392
#32667	4744303	4744611	309	309	hypothetical protein [Escherichia coli]	#5011	Mediator of hyperadherence YidE	1686
#32712	4750294	4750689	396	396	hypothetical protein [Escherichia coli]	#5020	Phosphatase YidA	813
#33146	4809166	4809519	354	300	hypothetical protein [Salmonella enterica]	#5075	hypothetical protein	1452
#33705	4894964	4895911	948	756	membrane protein [Polaribacter irgensii]	#5149	hypothetical protein	756
#33739	4899688	4899996	309	309	hypothetical protein [Escherichia coli]	#5152	Putative carboxymethylenebutenolidase	816
#33970	4935445	4935618	174	113	hypothetical protein [Photorhabdus luminescens]	#5185	hypothetical protein	117
#3404	517332	517667	336	305	hypothetical protein [Escherichia coli]	#0502	Cytochrome O ubiquinol oxidase subunit II	921
#34217	4974480	4974704	225	225	hypothetical protein [Rhodococcus qingshengii]	#5222	Putative frv operon regulatory protein	1749
#34271	4981955	4982224	270	180	hypothetical protein [Klebsiella pneumoniae]	#5228	Rhamnulose-1-phosphate aldolase	825
#34370	4996131	4996262	132	132	hypothetical protein [Escherichia coli]	#5245	hypothetical protein	144
#34671	5036077	5036829	753	653	hypothetical protein [Escherichia coli]	#5283	Phosphoenolpyruvate-protein phosphotransferase of PTS system	2502
#34795	5054685	5054843	159	159	hypothetical protein [Klebsiella pneumoniae]	#5298	Soluble pyridine nucleotide transhydrogenase	1401
#34796	5054850	5055098	249	204	hypothetical protein [Klebsiella pneumoniae]	#5298	Soluble pyridine nucleotide transhydrogenase	1401
#3492	533203	533451	249	174	hypothetical protein [Shigella dysenteriae]	#0518	Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA	1701
#35324	5135140	5135685	546	221	hypothetical protein [Cedecea davisae]	#5359	Aspartokinase	1350
#3582	547322	547555	234	234	hypothetical protein [Shigella flexneri]	#0537	RND efflux system, inner membrane transporter CmeB	3150
#35918	5221491	5221772	282	126	hypothetical protein, partial [Escherichia coli]	#5442	Phosphonates transport ATP-binding protein PhnL	681
#35987	5230008	5230232	225	119	hypothetical protein [Escherichia coli]	#5453	Alkylphosphonate utilization operon protein PhnA	414
#36100	5244822	5245136	315	122	hypothetical protein [Shigella dysenteriae]	#5464	Melibiose operon regulatory protein	909
#36310	5270145	5271641	1497	1434	hypothetical protein [Halorubrum kocurii]	#5486	Aspartate ammonia-lyase	1437
#36795	5345722	5346114	393	130	hypothetical protein [Escherichia albertii]	#5565	hypothetical protein	138
#37586	5469849	5470031	183	183	hypothetical protein [Serratia symbiotica]	#5672	adherence and invasion outermembrane protein (Inv,enhances Peyer's patches colonization)	5037
#37680	5482235	5482453	219	219	hypothetical protein [Escherichia albertii]	#5684	Transcriptional regulator, GntR family	1413
#37690	5483284	5483493	210	210	hypothetical protein [Escherichia coli]	#5684	Transcriptional regulator, GntR family	1413
#37783	5497417	5497734	318	318	hypothetical protein [Salmonella enterica]	#5696	Carbon starvation protein A	2115
#37863	5509646	5509750	105	105	hypothetical protein [Enterobacter cloacae]	#5707	putative membrane protein	825

Shadow ORF						mother gene		
ID CP008957.1_	start position	stop position	sORF length (bp)	overlap length (bp)	blastp hit	locus tag EDL933_	mORF product	mORF length (bp)
#37939	5521220	5521564	345	170	4 mannitol-1-phosphate 5-dehydrogenase [E. coli]	#5721	radical activating enzyme	783
#37943	5521704	5522240	537	474	hypothetical protein [Shigella sonnei]	#5721	radical activating enzyme	783
#38757	5447527	5447739	213	195	hypothetical protein [Klebsiella oxytoca]	#5650	type 1 fimbriae protein FimI, unknown function	498
#39433	5352322	5352756	435	260	hypothetical protein [Escherichia coli]	#5570	UPF0131 protein YtfP	288
#39570	5331319	5331567	249	249	hypothetical protein [Shigella flexneri]	#5551	putative virulence protein	1149
#39720	5311510	5311845	336	178	hypothetical protein [Escherichia coli]	#5524	3'-to-5' exoribonuclease RNase R	2442
#40477	5193509	5193826	318	264	hypothetical protein [Escherichia coli]	#5414	Cytochrome c-type heme lyase subunit nrfE, nitrite reductase complex assembly	1623
#40827	5141918	5142205	288	119	hypothetical protein [Escherichia albertii]	#5365	hypothetical protein	162
#40857	5138028	5138279	252	233	hypothetical protein [Escherichia coli]	#5361	YjbE secreted protein	243
#41233	5083558	5086020	2463	2463	maltose operon protein MalM [Escherichia coli]	#5318	DNA-directed RNA polymerase beta' subunit	4224
#41259	5078817	5082005	3189	2979	hypothetical protein [Klebsiella pneumoniae]	#5317	DNA-directed RNA polymerase beta subunit	4029
#41264	5077692	5078795	1104	1029	hypothetical protein [Cronobacter dublinensis]	#5317	DNA-directed RNA polymerase beta subunit	4029
#41273	5076524	5077021	498	492	hypothetical protein [Serratia marcescens]	#5315	LSU ribosomal protein L10p (P0)	498
#41300	5072572	5073681	1110	1095	Uncharacterised protein [Klebsiella pneumoniae]	#5310	Translation elongation factor Tu	1185
#41720	5018243	5019028	786	786	hypothetical protein [Escherichia coli]	#5269	core protein	4185
#4240	658692	659180	489	138	hypothetical protein [Escherichia coli]	#0623	hypothetical protein	891
#42665	4886279	4886665	387	117	hypothetical protein [Salmonella enterica]	#5140	hypothetical protein	117
#4309	666278	667216	939	939	hypothetical protein [Escherichia coli]	#0628	Rhs-family protein	1335
#4322	668275	669159	885	885	hypothetical protein [Escherichia coli]	#0630	Rhs-family protein	4938
#43543	4759239	4760408	1170	1170	hypothetical protein [Salmonella enterica]	#5029	Inner membrane protein translocase component YidC, long form	1596
#43550	4758386	4758583	198	198	RNase P [Edwardsiella ictaluri]	#5028	Ribonuclease P protein component	327
#4364	674096	675082	987	987	hypothetical protein [Shigella sonnei]	#0632	VgrG protein	1902
#44181	4655506	4656438	933	489	hypothetical protein [Escherichia coli]	#4913	GTP pyrophosphokinase, (p)ppGpp synthetase II	2109
#44332	4631904	4632167	264	264	hypothetical protein [Escherichia coli]	#4886	Oligosaccharide repeat unit polymerase Wzy	1209
#44483	4609502	4609876	375	375	hypothetical protein [Escherichia coli]	#4866	hypothetical protein	4767
#44484	4609205	4609420	216	216	hypothetical protein [Escherichia coli]	#4866	hypothetical protein	4767
#44519	4605017	4605472	456	363	hypothetical protein [Escherichia coli]	#4864	hypothetical protein	363
#44582	4595562	4596242	681	681	hypothetical protein [Delftia acidovorans]	#4854	core protein	4230
#44587	4595034	4595453	420	420	hypothetical protein [Escherichia coli]	#4854	core protein	4230
#44593	4594248	4595033	786	786	hypothetical protein [Escherichia coli]	#4854	core protein	4230
#44678	4581399	4581950	552	525	hypothetical protein [Escherichia coli]	#4843	LysR family transcriptional regulator YiaU	975
#45444	4467893	4469272	1380	271	hypothetical protein [Salmonella enterica]	#4737	Protein involved in catabolism of external DNA	843
#4556	703310	703681	372	314	hypothetical protein [Microlunatus phosphovorus]	#0660	ABC-type Fe3+-siderophore transport system, permease component	1005
#4685	724539	725000	462	112	hypothetical protein [Escherichia coli]	#0682	hypothetical protein	189
#46888	4254059	4254601	543	237	DNA topoisomerase I [Halorubrum terrestre]	#4501	hypothetical protein	237
#46985	4240619	4240915	297	150	hypothetical protein [Escherichia coli]	#4489	RND efflux system, inner membrane transporter CmeB	1233
#4699	725902	726207	306	234	hypothetical protein [Salmonella enterica]	#0684	Ribonuclease I precursor	807

Shadow ORF						mother gene		
ID CP008957.1_	start position	stop position	sORF length (bp)	overlap length (bp)	blastp hit	locus tag EDL933_	mORF product	mORF length (bp)
#47092	4227103	4227204	102	102	hypothetical protein [Shigella flexneri]	#4475	hypothetical protein	165
#47320	4189987	4190745	759	609	hypothetical protein [Escherichia coli]	#4441	Glutamate synthase [NADPH] small chain	1419
#47737	4125558	4125776	219	219	hypothetical protein [Siccibacter colletis]	#4375	putative endonuclease distantly related to archaeal Holliday junction resolvase	396
#47780	4119877	4120326	450	450	hypothetical protein [Rhodococcus qingshengii]	#4368	type 1 fimbriae anchoring protein FimD	2592
#47871	4109091	4109216	126	120	hypothetical protein [Enterobacter aerogenes]	#4355	PTS system, N-acetylgalactosamine-specific IID component	879
#4850	744843	745148	306	255	hypothetical protein, partial [Escherichia coli]	#0705	D-alanyl-D-alanine carboxypeptidase	1212
#48623	3998980	3999558	579	561	hypothteical protein [Dickeya dadantii]	#4248	Modulator of drug activity B	582
#48798	3974248	3974478	231	106	hypothetical protein [Salmonella enterica]	#4220	hypothetical protein	129
#48955	3951294	3951575	282	282	hypothetical protein [Salmonella enterica]	#4194	hypothetical protein	990
#49065	3934349	3934780	432	384	hypothetical protein [Shigella flexneri]	#4177	Membrane-bound lytic murein transglycosylase C precursor	1080
#49071	3933524	3934246	723	594	Uncharacterised protein [Klebsiella pneumoniae]	#4177	Membrane-bound lytic murein transglycosylase C precursor	1080
#49207	3915511	3915780	270	113	membrane protein [Kluyvera ascorbata]	#4151	hypothetical protein	126
#49908	3804157	3804405	249	200	hypothetical protein [Escherichia coli]	#4038	IncI1 plasmid conjugative transfer putative membrane protein PilT	504
#50186	3759196	3759582	387	315	amino acid acetyltransferase [Escherichia coli]	#3996	N-acetylglutamate synthase	1332
#50287	3745625	3746374	750	750	Uncharacterised protein [Klebsiella pneumoniae]	#3983	L-fucose isomerase	1776
#50318	3740673	3741263	591	297	hypothetical protein, partial [Pseudomonas amygdali]	#3978	L-serine dehydratase	1368
#50487	3714138	3714428	291	178	hypothetical protein [Escherichia coli]	#3953	Uncharacterized protein YgcG	675
#5070	772828	773478	651	489	hypothetical protein [Shigella sonnei]	#0737	Phosphate starvation-inducible ATPase PhoH with RNA binding motif	1041
#51092	3624220	3624696	477	159	hypothetical protein [Xenorhabdus poinarii]	#3852	Multidrug resistance protein A	981
#51542	3557603	3558064	462	264	heat-shock protein GrpE [Kluyvera ascorbata]	#3776	NAD kinase	807
#51705	3532816	3532929	114	114	hypothetical protein [Escherichia coli]	#3753	hypothetical protein	564
#51729	3529881	3530330	450	450	hypothetical protein [Escherichia coli]	#3750	CDP-diacylglycerol--serine O-phosphatidyltransferase	1356
#52107	3470729	3471535	807	702	hypothetical protein, partial [Catenibacterium mitsuokai]	#3696	Inositol-1-monophosphatase	804
#52152	3461436	3461651	216	107	hypothetical protein [Yersinia pseudotuberculosis]	#3683	hypothetical protein	126
#52292	3437268	3437501	234	234	hypothetical protein [Cronobacter turicensis]	#3666	Exodeoxyribonuclease VII large subunit	1371
#52317	3432188	3432478	291	118	membrane protein [Salmonella enterica]	#3660	hypothetical protein	153
#52524	3405098	3405319	222	222	hypothetical protein [Escherichia coli]	#3637	Hydrogenase-4 component B	2019
#52682	3383761	3384030	270	270	Uncharacterised protein [Klebsiella pneumoniae]	#3619	Transketolase	2004
#52694	3382405	3383607	1203	227	Uncharacterised protein [Klebsiella pneumoniae]	#3618	Transaldolase	951
#53470	3262428	3262706	279	98	mannitol-1-phosphate 5-dehydrogenase [Escherichia albertii]	#3499	hypothetical protein	552
#54026	3174399	3174743	345	126	hypothetical protein [Shigella flexneri]	#3408	hypothetical protein	126
#54145	3158877	3160754	1878	1788	ribonucleotide-diphosphate reductase subunit alpha, partial [Klebsiella pneumoniae]	#3395	Ribonucleotide reductase of class Ia (aerobic), alpha subunit	2286
#54484	3105751	3106182	432	285	hypothetical protein [Edwardsiella ictaluri]	#3352	LSU ribosomal protein L25p	285
#5499	834886	835146	261	107	MULTISPECIES: hypothetical protein [Enterobacteriaceae]	#0794	hypothetical protein	132
#55866	2909432	2910049	618	618	hypothetical protein [Escherichia coli]	#3143	hypothetical protein	2808
#56283	2837930	2838190	261	144	hypothetical protein [Escherichia coli]	#3074	YeeU protein (antitoxin to YeeV)	249

Shadow ORF						mother gene		
ID CP008957.1_	start position	stop position	sORF length (bp)	overlap length (bp)	blastp hit	locus tag EDL933_	mORF product	mORF length (bp)
#56324	2833844	2834152	309	309	hypothetical protein [Escherichia coli]	#3069	Colicin I receptor precursor	2148
#56507	2808509	2808754	246	246	hypothetical protein [Serratia symbiotica]	#3049	adherence and invasion outermembrane protein (Inv,enhances Peyer's patches colonization)	7863
#56564	2802234	2802425	192	177	hypothetical protein [Escherichia coli]	#3043	Division inhibition protein dicB	189
#56801	2764765	2764905	141	141	molecular chaperone Tir [Escherichia coli]	#2986	hypothetical protein	711
#56803	2764624	2764764	141	141	molecular chaperone Tir [Escherichia coli]	#2986	hypothetical protein	711
#56805	2764483	2764623	141	141	molecular chaperone Tir [Escherichia coli]	#2986	hypothetical protein	711
#56807	2764342	2764482	141	141	hypothetical protein [Escherichia coli]	#2986	hypothetical protein	711
#57079	2725264	2725470	207	207	hypothetical protein [Escherichia coli]	#2937	Putative transport system permease protein	1206
#5713	867541	867831	291	291	hypothetical protein [Salmonella enterica]	#0825	Zinc transporter ZitB	936
#57364	2683488	2683634	147	144	acid-inducible small membrane-associated protein [Escherichia coli]	#2877	hypothetical protein	147
#5740	870687	871265	579	552	hypothetical protein [Pantoea sp. SL1_M5]	#0828	Phosphoglycerate mutase	753
#5749	872109	872366	258	258	hypothetical protein [Escherichia coli]	#0830	Aldose 1-epimerase	1041
#57774	2617560	2617859	300	300	hypothetical protein [Salmonella enterica]	#2808	Ribosomal RNA small subunit methyltransferase F	1521
#58806	2470134	2470535	402	402	hypothetical protein [Escherichia coli]	#2651	Coenzyme A transferase	1596
#58981	2442215	2442418	204	204	hypothetical protein [Escherichia coli]	#2624	Acyl-CoA dehydrogenases	1605
#58990	2440884	2441180	297	294	hypothetical protein [Escherichia coli]	#2623	Protein ydhR precursor	306
#59154	2419868	2420278	411	120	prolyl-tRNA synthetase [Polaromonas naphthalenivorans]	#2604	hypothetical protein	120
#59606	2357851	2358063	213	192	hypothetical protein [Shigelly dysenteriae]	#2540	Anaerobic dimethyl sulfoxide reductase chain C	855
#59705	2345766	2346497	732	672	cell surface protein [Escherichia coli]	#2528	Exodeoxyribonuclease encoded by cryptic prophage CP-933P	2472
#59720	2344277	2344510	234	222	hypothetical protein [Curvibacter lanceolatus]	#2523	hypothetical protein	357
#59762	2337515	2337874	360	180	hypothetical protein [Plesiomonas shigelloides]	#2512	hypothetical protein	180
#60520	2228862	2229041	180	111	hypothetical protein [Shigella sonnei]	#2386	Gamma-glutamyl-putrescine synthetase	1419
#60767	2188671	2188988	318	279	hypothetical protein, partial [Salmonella enterica]	#2342	Fumarate and nitrate reduction regulatory protein	753
#61510	2071006	2071182	177	165	hypothetical protein [Shigelly dysenteriae]	#2215	hypothetical protein	171
#61578	2060821	2061216	396	175	hypothetical protein [Escherichia coli]	#2200	Putative oxidoreductase YncB	1038
#61743	2035342	2035836	495	495	hypothetical protein [Enterobacter cloacae]	#2174	Respiratory nitrate reductase alpha chain	3741
#61771	2031685	2031852	168	168	hypothetical protein [Shigella flexneri]	#2172	internalin, putative	1260
#61976	2005698	2005997	300	281	hypothetical protein [Escherichia albertii]	#2149	redicted glycoside hydrolase	1320
#61992	2003311	2003889	579	579	hypothetical protein, partial [Escherichia coli]	#2147	Glutamate decarboxylase	1401
#62034	1998509	1998793	285	285	hypothetical protein [Escherichia coli]	#2145	hypothetical protein	2373
#62406	1948438	1948722	285	285	hypothetical protein [Escherichia coli]	#2101	Rhodanese-related sulfurtransferase	396
#63161	1858098	1858628	531	531	hypothetical protein [Escherichia coli]	#1984	DNA methyl transferase, phage-associated	1059
#6343	965643	965963	321	111	hypothetical protein [Shigella dysenteriae]	#0928	Ferrichrome-iron receptor	2283
#63514	1811035	1812255	1221	1221	hypothetical protein [Enterobacter cloacae]	#1931	Respiratory nitrate reductase alpha chain	3744
#63696	1781140	1781508	369	139	trehalase [Gramella echinicola]	#1900	Trehalase	1458
#64399	1682161	1682487	327	319	hypothetical protein [Escherichia coli]	#1775	hypothetical protein	744
#64620	1654337	1654921	585	585	hypothetical protein [Escherichia coli]	#1743	Phage portal protein	2502

Shadow ORF						mother gene		
ID CP008957.1_	start position	stop position	sORF length (bp)	overlap length (bp)	blastp hit	locus tag EDL933_	mORF product	mORF length (bp)
#64638	1652227	1652691	465	465	hypothetical protein [Escherichia coli]	#1741	hypothetical protein	1938
#64725	1642360	1642524	165	159	hypothetical protein [Escherichia coli]	#1724	hypothetical protein	168
#64727	1642182	1642682	501	147	hypothetical protein [Escherichia coli]	#1723	hypothetical protein	147
#64825	1631424	1631741	318	318	hypothetical protein [Edwardsiella piscicida]	#1704	hypothetical protein	516
#65083	1595333	1595617	285	189	hypothetical protein [Salmonella enterica]	#1667	Phosphate:acyl-ACP acyltransferase PlsX	1041
#65241	1574654	1574956	303	239	hypothetical protein [Shigella sonnei]	#1645	putative peptidoglycan lipid II flippase MurJ	1536
#65479	1543802	1544362	561	489	hypothetical protein [Xenorhabdus bovienii]	#1605	hypothetical protein	843
#65509	1540561	1540821	261	261	hypothetical protein, partial [Klebsiella oxytoca]	#1598	YeeU protein (antitoxin to YeeV)	375
#65528	1538231	1538542	312	312	hypothetical protein [Burkholderia sacchari]	#1592	hypothetical protein	819
#65561	1535092	1535439	348	348	hypothetical protein, partial [Escherichia coli]	#1588	Putative vimentin	2178
#65606	1529480	1530271	792	723	hypothetical protein [Escherichia coli]	#1582	NgrB	873
#65639	1525135	1525461	327	324	hypothetical protein [Escherichia coli]	#1574	Co-activator of prophage gene expression IbrA	1233
#6574	1000713	1001303	591	330	hypothetical protein [Escherichia coli]	#0960	DNA-binding protein	330
#6662	1012147	1012479	333	251	hypothetical protein, partial [Escherichia coli]	#0972	Putative transport protein/putative regulator	1209
#66748	1373871	1374113	243	114	outer membrane protein [Escherichia coli]	#1408	hypothetical protein	189
#67577	1264550	1264795	246	227	hypothetical protein [Escherichia coli]	#1263	hypothetical protein	819
#67649	1256691	1257182	492	317	hypothetical protein [Enterobacter cloacae]	#1253	hypothetical protein	558
#68480	1143655	1144215	561	489	hypothetical protein [Xenorhabdus bovienii]	#1146	hypothetical protein	843
#68510	1140414	1140674	261	261	hypothetical protein, partial [Klebsiella oxytoca]	#1139	YeeU protein (antitoxin to YeeV)	375
#68529	1138084	1138395	312	312	hypothetical protein [Burkholderia sacchari]	#1133	hypothetical protein	819
#68609	1129336	1130127	792	723	hypothetical protein [Escherichia coli]	#1121	NgrB	873
#68642	1124991	1125317	327	324	hypothetical protein [Escherichia coli]	#1113	Co-activator of prophage gene expression IbrA	1233
#69144	1056443	1056676	234	99	hypothetical protein [Escherichia coli]	#1017	hypothetical protein	135
#70284	901036	901434	399	147	hypothetical protein [Escherichia coli]	#0863	hypothetical protein	147
#7056	1070769	1070963	195	140	hypothetical protein [Escherichia coli]	#1032	Transposase	312
#70740	832299	833132	834	804	hypothetical protein [Klebsiella pneumoniae]	#0792	Succinate dehydrogenase flavoprotein subunit	1767
#70901	810840	811505	666	666	hypothetical protein [Delftia acidovorans]	#0770	core protein	4200
#70906	810312	810731	420	420	hypothetical protein [Escherichia coli]	#0770	core protein	4200
#70912	809526	810311	786	786	hypothetical protein [Escherichia coli]	#0770	core protein	4200
#70925	808287	808679	393	393	hypothetical protein [Escherichia coli]	#0770	core protein	4200
#71041	788927	789706	780	615	hypothetical protein [Escherichia coli]	#0753	N-acetylglucosamine-regulated outer membrane porin	909
#71252	757822	758163	342	316	hypothetical protein [Escherichia coli]	#0720	hypothetical protein	708
#71473	723052	723639	588	420	hypothetical protein [Salmonella enterica]	#0681	putative zinc-type alcohol dehydrogenase-like protein ybdR	1239
#71490	720160	720843	684	564	hypothetical protein [Escherichia coli]	#0677	Alkyl hydroperoxide reductase protein C	564
#7161	1086168	1086959	792	149	hypothetical protein [Escherichia coli]	#1057	hypothetical protein	174
#72178	619517	620434	918	918	hypothetical protein [Escherichia coli]	#0585	core protein	4197
#72195	617312	617536	225	225	hypothetical protein [Escherichia coli]	#0585	core protein	4197
#72348	594964	595320	357	357	hypothetical protein [Escherichia coli]	#0566	putative cell-wall-anchored protein SasA (LPXTG motif)	15567
#72358	593113	593406	294	294	hypothetical protein [Escherichia sp. TW09308]	#0566	putative cell-wall-anchored protein SasA (LPXTG motif)	15567

Shadow ORF						mother gene		
ID CP008957.1_	start position	stop position	sORF length (bp)	overlap length (bp)	blastp hit	locus tag EDL933_	mORF product	mORF length (bp)
#72370	591349	591729	381	381	hypothetical protein [Escherichia coli]	#0566	putative cell-wall-anchored protein SasA (LPXTG motif)	15567
#72384	589522	589722	201	201	hypothetical protein [Escherichia coli]	#0566	putative cell-wall-anchored protein SasA (LPXTG motif)	15567
#72396	588022	588222	201	201	hypothetical protein [Escherichia coli]	#0566	putative cell-wall-anchored protein SasA (LPXTG motif)	15567
#72410	586228	586428	201	201	hypothetical protein [Escherichia coli]	#0566	putative cell-wall-anchored protein SasA (LPXTG motif)	15567
#72423	584497	584757	261	261	hypothetical protein [Escherichia coli]	#0565	Large repetitive protein	4386
#73837	387431	387832	402	402	hypothetical protein, partial [Escherichia coli]	#0370	hypothetical protein	864
#7384	1117333	1117608	276	174	hypothetical protein [Citrobacter rodentium]	#1105	hypothetical protein	192
#74632	273954	275081	1128	1128	hypothetical protein [Escherichia coli]	#0244	core protein	1761
#74663	270367	270657	291	291	hypothetical protein [Shigella sp.]	#0240	core protein	4215
#74683	267541	267864	324	324	hypothetical protein [Escherichia coli]	#0240	core protein	4215
#74703	264946	265848	903	903	hypothetical protein [Shigella sonnei]	#0239	VgrG protein	2142
#75212	196057	196335	279	119	hypothetical protein [Escherichia coli]	#0176	Ribosome recycling factor	558
#75282	185065	185439	375	356	ABC transporter permease [Salmonella enterica]	#0166	HtrA protease/chaperone protein	1425
#75364	172527	172889	363	363	hypothetical protein [Dickeya dadantii]	#0155	Ferric hydroxamate outer membrane receptor FhuA	2244
#75595	137722	138618	897	864	hypothetical protein [Dickeya sp.]	#0120	Aconitate hydratase 2	2520
#8326	1262181	1262621	441	180	hypothetical protein [Plesiomonas shigelloides]	#1259	hypothetical protein	180
#8389	1271488	1271730	243	243	hypothetical protein, partial [Cronobacter universalis]	#1273	putative endopeptidase	468
#8548	1296704	1296844	141	141	molecular chaperone Tir [Escherichia coli]	#1308	hypothetical protein	753
#8550	1296845	1296985	141	141	hypothetical protein [Escherichia coli]	#1308	hypothetical protein	753
#8552	1296986	1297126	141	141	hypothetical protein [Escherichia coli]	#1308	hypothetical protein	753
#8554	1297127	1297267	141	141	hypothetical protein [Escherichia coli]	#1308	hypothetical protein	753
#8646	1310776	1311258	483	426	hypothetical protein [Shigella sonnei]	#1320	Low molecular weight protein-tyrosine-phosphatase Wzb	447
#8749	1323339	1323566	228	228	hypothetical protein [Serratia liquefaciens]	#1333	TorCAD operon transcriptional regulatory protein TorR	693
#9297	1405898	1406032	135	135	proline dehydrogenase [Escherichia coli]	#1435	Transcriptional repressor of PutA and PutP	3963
#941	143052	143270	219	133	hypothetical protein [Escherichia albertii]	#0126	phosphopantetheinyltransferase component of enterobactin synthase multienzyme complex	138
#9783	1486312	1487103	792	149	hypothetical protein [Escherichia coli]	#1518	hypothetical protein	174

Supplementary table S10: Blastp (February 2016) and re-blast (September 2018) of 280 sORFs with full-length matches to proteins in the database. A full-length matching hit is defined to have a coverage ≥80%.

ID CP008957.1_	blastp 2016	blast 2018			
		hit	coverage [%]	E-value	identity [%]
#10006	hypothetical protein [E. coli]	CBJ03433.1|conserved hypothetical protein [Escherichia coli ETEC H10407]	100	2,00E-54	93
#10163	hypothetical protein [Pseudomonas aeruginosa]	WP_077738204.1| hypothetical protein [Escherichia coli]	100	1,00E-22	50
#10211	hypothetical protein [Escherichia coli]	CSQ84761.1|Uncharacterised protein [Shigella sonnei]	100	1,00E-40	98
#10317	hypothetical protein [Escherichia coli]	AEJ55850.1|hypothetical protein UMNF18_1243 [Escherichia coli UMNF18]	100	1,00E-32	100
#10782	membrane protein [E. coli]	WP_042110866.1|DUF4752 family protein [Escherichia coli]	100	2,00E-63	94
#10796	hypothetical protein [Plesiomonas shigelloides]	ETJ57615.1|hypothetical protein Q456_0217590 [Escherichia coli ATCC BAA-2193]	100	1,00E-83	84
#11476	hypothetical protein [Escherichia coli]	WP_045903741.1|hypothetical protein [Escherichia coli]	100	1,00E-35	100
#11769	PTS-dependent dihydroxyacetone kinase operon transcriptional regulator DhaR [E. coli]	AFJ28690.1| hypothetical protein CDCO157_1633 [Escherichia coli Xuzhou21]	100	2,00E-108	100
#12060	hypothetical protein [Plautia stali symbiont]	SRN40382.1|Uncharacterised protein [Shigella flexneri]	100	2,00E-42	75
#12257	hypothetical protein [Plesiomonas shigelloides]	WP_077826881.1|hypothetical protein [Escherichia coli]	100	7,00E-49	78
#1258	hypothetical protein [Escherichia coli]	SAJ28306.1|Uncharacterised protein [Enterobacter cloacae]	100	5,00E-15	54
#1277	hypothetical protein [Franconibacter helveticus]	KNS81849.1|hypothetical protein AEW32_17955 [Salmonella enterica subsp. enterica serovar Hadar]	100	2,00E-08	81
#13401	type IV section protein [Trabulsiella odontotermitis]	ABM54879.1|hypothetical protein ECf0004 [Escherichia coli]	100	4,00E-123	79
#13407	hypothetical protein [E. coli]	CTT92499.1|Uncharacterised protein [Escherichia coli]	99	2,00E-178	98
#13417	hypothetical protein [Escherichia coli]	STL43598.1|Uncharacterised protein [Escherichia coli]	70	2,00E-07	52
#13440	hypothetical protein [Shigella sonnei]	SVF54532.1|Uncharacterised protein [Escherichia coli]	100	2,00E-74	75
#13509	hypothetical protein, partial [Escherichia coli]	AER84266.1|hypothetical protein i02_1694 [Escherichia coli str. 'clone D i2']	100	1,00E-56	91
#13570	hypothetical protein [Escherichia coli]	AAG56338.1|orf, hypothetical protein [Escherichia coli O157:H7 str. EDL933]	57	1,00E-39	98
#13900	hypothetical protein [E. coli]	OWC39904.1|hypothetical protein A8F96_23255, partial [Escherichia coli]	87	3,00E-29	91
#14171	hypothetical protein, partial [Salmonella enterica]	OEN73507.1|hypothetical protein BHF53_19990 [Escherichia coli]	97	1,00E-60	95
#14229	hypothetical protein [E. coli]	KDV50401.1|hypothetical protein BU54_36115 [Escherichia coli O45:H2 str. 2010C-4211]	82	2,00E-117	97
#14888	hypothetical protein [Escherichia coli]	AAG56523.1|hypothetical protein Z2511 [Escherichia coli O157:H7 str. EDL933]	100	4,00E-88	100
#15102	hypothetical protein [Shigella flexneri]	EFI89311.1|hypothetical protein HMPREF9551_01685 [Escherichia coli MS 196-1]	90	1,00E-25	96
#15556	hypothetical protein [E. coli]	AFJ28520.1|hypothetical protein CDCO157_1459 [Escherichia coli Xuzhou21]	100	6,00E-112	96
#16693	hypothetical protein [Klebsiella pneumoniae]	CSI42537.1|Uncharacterised protein [Shigella sonnei]	100	1,00E-109	99
#16998	hypothetical protein [Escherichia coli]	no hits			
#17453	hypothetical protein [E. coli]	WP_012602292.1|hypothetical protein [Escherichia coli]	100	2,00E-139	99
#1794	hypothetical protein [Escherichia coli]	WP_001303800.1|MULTISPECIES: hypothetical protein [Enterobacteriaceae]	100	2,00E-32	100
#17968	hypothetical protein [Escherichia coli]	CSP91274.1|Uncharacterised protein [Shigella sonnei]	100	8,00E-96	99
#18076	hypothetical protein, partial [Neisseria flavescens]	WP_064554502.1|hypothetical protein [Buttiauxella noackiae]	100	6,00E-20	39
#18869	hypothetical protein [E. coli]	ADD56521.1|hypothetical protein G2583_1950 [Escherichia coli O55:H7 str. CB9615]	100	2,00E-92	94
#1887	hypothetical protein [E. coli]	EKH87216.1|hypothetical protein ECPA45_0409 [Escherichia coli PA45]	88	1,00E-137	99
#19397	hypothetical protein [Escherichia coli]	no hits			
#19520	hypothetical protein [Enterobacter cloacae]	CSQ63911.1|Uncharacterised protein [Shigella sonnei]	100	2,00E-98	93
#19712	hypothetical protein [E. coli]	WP_119177803.1|hypothetical protein [Shigella flexneri]	100	5,00E-50	92

ID CP008957.1_	blastp 2016	blast 2018			
		hit	coverage [%]	E-value	identity [%]
#19821	hypothetical protein [Escherichia coli]	SVQ40934.1|Uncharacterised protein [Klebsiella pneumoniae]	100	2,00E-51	96
#20653	hypothetical protein [Shigella sp. SF-2015]	WP_074157434.1|hypothetical protein [Escherichia coli]	45	9,00E-23	79
#21079	aldose-1-epimerase [E. coli]	P28247.2|PUTATIVE PSEUDOGENE: RecName: Full=Putative uncharacterized protein BicB	100	3,00E-81	98
#21291	inverted ada-Golga3 fusion protein [Escherichia coli]	EII08957.1|hypothetical protein EC50959_4065 [Escherichia coli 5.0959]	91	9,00E-36	85
#22524	hypothetical protein [Xenorhabdus poinarii]	ENG92689.1|hypothetical protein EC178850_2491 [Escherichia coli 178850]	100	9,00E-71	96
#22698	hypothetical protein [Escherichia coli]	CSI01492.1|Uncharacterised protein [Shigella sonnei]	100	1,00E-41	93
#22778	hypothetical protein [E. coli]	WP_074435495.1|hypothetical protein [Escherichia coli]	100	4,00E-33	88
#22782	hypothetical protein [Shigella flexneri]	EID65455.1|hypothetical protein ECW26_40280 [Escherichia coli W26]	100	4,00E-46	91
#23185	hypothetical protein [Enterobacter ludwigii]	EFJ58054.1|hypothetical protein HMPREF9549_00481 [Escherichia coli MS 185-1]	100	5,00E-95	93
#23693	hypothetical protein [E. coli]	CNV27459.1|Uncharacterised protein [Salmonella enterica subsp. enterica serovar Bovismorbificans]	61	4,00E-27	77
#23965	hypothetical protein [Citrobacter koseri]	OAF32095.1|hypothetical protein AXK30_22115 [Escherichia coli]	60	2,00E-56	100
#23972	phosphoribosylformylglycinamidine synthase [E. coli]	ABA48290.1|hypothetical protein BURPS1710b_2462 [Burkholderia pseudomallei 1710b]	93	2,00E-37	37
#25098	hypothetical protein, partial [Bacillus cereus]	EJK90193.1|hypothetical protein UUU_33130 [Klebsiella pneumoniae subsp. pneumoniae DSM 30104]	94	0,00E+00	68
#25122	hypothetical protein [Shigella flexneri]	OON35465.1|hypothetical protein BU230_33305 [Klebsiella pneumoniae]	90	2,00E-17	47
#26101	hypothetical protein [Escherichia coli]	OBX35909.1|hypothetical protein A8U91_00245 [Halomonas elongata]	100	9,00E-04	31
#26608	hypothetical protein [Salmonella enterica]	ENO95326.1|putative metal-dependent RNase [Thauera phenylacetica B4P]	99	4,00E-50	43
#26893	hypothetical protein [Klebsiella pneumoniae]	KFD09439.1|hypothetical protein GSMA_04821 [Serratia marcescens subsp. marcescens ATCC 13880]	100	2,00E-170	53
#26908	hypothetical protein, partial [Bacillus cereus]	AAA83910.1|Select seq AAA83910.1 ORF1; putative [Escherichia coli]	97	0,00E+00	93
#26939	hypothetical protein [Shigella dysenteriae]	CQR82378.1|hypothetical protein b2940 [Escherichia coli K-12]	100	5,00E-48	100
#27372	hypothetical protein [E. coli]	EFK68361.1|hypothetical protein HMPREF9347_02788 [Escherichia coli MS 124-1]	97	5,00E-21	98
#27374	hypothetical protein [E. coli]	AAG58140.1|orf, hypothetical protein [Escherichia coli O157:H7 str. EDL933]	100	4,00E-53	100
#28096	hypothetical protein [Escherichia coli]	CSP58355.1|Uncharacterised protein [Shigella sonnei]	100	4,00E-53	97
#28643	hypothetical protein [Serratia sp. TEL]	STF41748.1|Uncharacterised protein [Escherichia coli]	74	1,00E-64	100
#28649	hypothetical protein [Acidovorax radicis]	CBG36302.1|conserved hypothetical protein [Escherichia coli 042]	96	2,00E-136	99
#29446	hypothetical protein [Nephila clavipes]	ABX23982.1|hypothetical protein SARI_04193 [Salmonella enterica subsp. arizonae serovar 62:z4,z23:-]	100	0	95
#29453	hypothetical protein [Klebsiella pneumoniae]	CDN08676.1|conserved hypothetical protein [Klebsiella quasipneumoniae subsp. similipneumoniae]	100	2,00E-145	87
#29471	ferredoxin [Cronobacter dubliensis]	WP_071525009.1|MULTISPECIES: ferredoxin [Enterobacteriaceae]	58	7,00E-27	100
#29479	Uncharacterised protein [Serratia marcescens]	SQP93016.1|Uncharacterised protein [Escherichia coli]	100	0	99
#29721	hypothetical protein, partial [Escherichia coli]	SAE14040.1|Uncharacterised protein [Enterobacter cloacae]	100	3,00E-32	49
#29735	hypothetical protein [Cronobacter sakazakii]	ABU79532.1|hypothetical protein ESA_04353 [Cronobacter sakazakii ATCC BAA-894]	100	7,00E-13	53
#29869	hypothetical protein [Enterobacter aerogenes]	CSP83838.1|Uncharacterised protein [Shigella sonnei]	100	1,00E-56	100
#30165	hypothetical protein, partial [Escherichia coli]	PQO12922.1|hypothetical protein C5K19_01660 [Shigella flexneri]	73	5,00E-57	99
#30215	hypothetical protein [Escherichia coli]	AAN82664.1|Hypothetical protein c4226 [Escherichia coli CFT073]	100	2,00E-144	97
#30643	hypothetical protein [Escherichia coli]	CSP80332.1|Uncharacterised protein [Shigella sonnei]	80	3,00E-124	90
#30955	hypothetical protein, partial [Escherichia coli]	KFZ97440.1|hypothetical protein DP20_3635 [Shigella flexneri]	96	1,00E-120	91
#30961	hypothetical protein [Escherichia coli]	SQY60989.1|Uncharacterised protein [Escherichia coli]	100	6,00E-133	97

ID CP008957.1_	blastp 2016	blast 2018			
		hit	coverage [%]	E-value	identity [%]
#31337	hypothetical protein [Erwinia amylovora]	no hits			
#31654	hypothetical protein [Shigella sonnei]	CSP98099.1|Uncharacterised protein [Shigella sonnei]	91	8,00E-98	97
#31968	hypothetical protein [Pectobacterium carotovorum]	ABE09607.1|hypothetical protein UTI89_C4179 [Escherichia coli UTI89]	77	1,00E-59	100
#32035	hypothetical protein, partial [Escherichia coli]	WP_109545302.1|hypothetical protein [Escherichia coli]	98	3,00E-57	92
#32068	hypothetical protein [Escherichia coli]	WP_024203629.1|hypothetical protein [Escherichia coli]	100	5,00E-49	100
#32298	hypothetical protein [Escherichia coli]	WP_000274020.1|hypothetical protein [Escherichia coli]	100	2,00E-40	100
#32339	type III secretion system protein SepZ [Escherichia albertii]	WP_000386949.1|MULTISPECIES: type III secretion system LEE cytoprotective effector EspZ [Enterobacteriaceae]	100	3,00E-60	100
#32485	hypothetical protein, partial [Bacillus cereus]	KUW80460.1|hypothetical protein AWF71_22475 [Escherichia coli]	100	1,00E-57	99
#32667	hypothetical protein [Escherichia coli]	CSP34245.1|Uncharacterised protein [Shigella sonnei]	100	2,00E-68	97
#32712	hypothetical protein [Escherichia coli]	BAK13326.1|hypothetical protein PAJ_3246 [Pantoea ananatis AJ13355]	100	4,00E-16	33
#33146	hypothetical protein [Salmonella enterica]	STM79633.1|Uncharacterised protein [Escherichia coli]	100	1,00E-80	99
#33705	membrane protein [Polaribacter irgensii]	OSK47270.1|putative membrane protein [Escherichia coli H588]	100	0	99
#33739	hypothetical protein [Escherichia coli]	AAP18855.1|hypothetical protein S3847 [Shigella flexneri 2a str. 2457T]	100	4,00E-71	100
#33970	hypothetical protein [Photorhabdus luminescens]	WP_072171068.1|hypothetical protein [Trabulsiella odontotermitis]	100	3,00E-19	72
#3404	hypothetical protein [Escherichia coli]	SAD47617.1|Uncharacterised protein [Enterobacter cloacae]	90	8,00E-60	91
#34217	hypothetical protein [Rhodococcus qingshengii]	KDQ00142.1|hypothetical protein EN35_07805 [Rhodococcus qingshengii]	97	2,00E-42	97
#34271	hypothetical protein [Klebsiella pneumoniae]	EGJ83006.1|hypothetical protein SF274771_4232 [Shigella flexneri 2747-71]	67	2,00E-32	95
#34370	hypothetical protein [Escherichia coli]	EGB75972.1|hypothetical protein HMPREF9532_03573 [Escherichia coli MS 57-2]	97	2,00E-16	88
#34671	hypothetical protein [Escherichia coli]	SVW18104.1|Uncharacterised protein [Klebsiella pneumoniae]	80	2,00E-136	98
#34795	hypothetical protein [Klebsiella pneumoniae]	PAW15911.1|hypothetical protein CKJ89_11525 [Klebsiella pneumoniae]	100	3,00E-10	48
#34796	hypothetical protein [Klebsiella pneumoniae]	EFJ68094.1|hypothetical protein HMPREF9547_00654 [Escherichia coli MS 175-1]	100	1,00E-53	99
#3492	hypothetical protein [Shigella dysenteriae]	WP_001334127.1|hypothetical protein [Escherichia coli]	76	3,00E-37	97
#35324	hypothetical protein [Cedecea davisae]	SYQ84979.1|Uncharacterised protein [Klebsiella pneumoniae]	100	2,00E-128	99
#3582	hypothetical protein [Shigella flexneri]	SQD06307.1|Uncharacterised protein [Escherichia coli]	100	2,00E-35	86
#35918	hypothetical protein, partial [Escherichia coli]	CSF15287.1|Uncharacterised protein [Shigella sonnei]	100	2,00E-61	99
#35987	hypothetical protein [Escherichia coli]	WP_077899371.1|hypothetical protein [Escherichia coli]	100	2,00E-45	99
#36100	hypothetical protein [Shigella dysenteriae]	WP_117122804.1|hypothetical protein [Klebsiella variicola]	66	3,00E-38	90
#36310	hypothetical protein [Halorubrum kocurii]	AAN83643.1|Hypothetical protein c5221 [Escherichia coli CFT073]	96	0	98
#36795	hypothetical protein [Escherichia albertii]	ANK04062.1|Hypothetical protein WLH_02801 [Escherichia coli O25b:H4]	46	1,00E-09	67
#37586	hypothetical protein [Serratia symbiotica]	SQP55726.1|Uncharacterised protein [Escherichia coli]	100	5,00E-33	95
#37680	hypothetical protein [Escherichia albertii]	PHX52271.1|hypothetical protein AO354_27580 [Pseudomonas syringae pv. syringae]	68	1,00E-03	45
#37690	hypothetical protein [Escherichia coli]	ETJ23126.1|hypothetical protein Q609_ECAC01557G0002 [Escherichia coli DORA_A_5_14_21]	100	2,00E-36	94
#37783	hypothetical protein [Salmonella enterica]	SVJ61792.1|Uncharacterised protein [Klebsiella pneumoniae]	100	7,00E-24	51
#37863	hypothetical protein [Enterobacter cloacae]	WP_013095468.1|hypothetical protein [Enterobacter cloacae]	100	2,00E-06	68
#37939	4 mannitol-1-phosphate 5-dehydrogenase [E. coli]	SRN31005.1|Protein of uncharacterised function (DUF3521) [Shigella flexneri]	100	5,00E-69	92
#37943	hypothetical protein [Shigella sonnei]	ESA64667.1|hypothetical protein HMPREF1589_04335 [Escherichia coli 113290]	46	4,00E-51	100
#38757	hypothetical protein [Klebsiella oxytoca]	KEN49305.1|hypothetical protein AB81_4979 [Escherichia coli 6-537-08_S1_C3]	80	7,00E-32	96

ID CP008957.1_	blastp 2016	blast 2018			
		hit	coverage [%]	E-value	identity [%]
#39433	hypothetical protein [Escherichia coli]	AAG59421.1|orf, hypothetical protein [Escherichia coli O157:H7 str. EDL933]	61	2,00E-59	100
#39570	hypothetical protein [Shigella flexneri]	CSE40754.1|Uncharacterised protein [Shigella sonnei]	100	3,00E-49	95
#39720	hypothetical protein [Escherichia coli]	KDV18071.1|hypothetical protein BW72_02400 [Escherichia coli O78:H12 str. 00-3279]	100	2,00E-71	99
#40477	hypothetical protein [Escherichia coli]	STL90606.1|Uncharacterised protein [Escherichia coli]	59	2,00E-37	100
#40827	hypothetical protein [Escherichia albertii]	WP_001301827.1|hypothetical protein [Escherichia coli]	96	2,00E-61	100
#40857	hypothetical protein [Escherichia coli]	WP_001308199.1|hypothetical protein [Escherichia coli]	98	1,00E-46	99
#41233	maltose operon protein MalM [Escherichia coli]	CSH40721.1|Uncharacterised protein [Shigella sonnei]	97	0	99
#41259	hypothetical protein [Klebsiella pneumoniae]	ABX23332.1|hypothetical protein SARI_03507 [Salmonella enterica subsp. arizonae serovar 62:z4,z23:-]	100	0	84
#41264	hypothetical protein [Cronobacter dublinensis]	EMI37041.1|hypothetical protein MTE2_4528 [Klebsiella pneumoniae VA360]	99	5,00E-177	74
#41273	hypothetical protein [Serratia marcescens]	SAE40317.1|Uncharacterised protein [Enterobacter cloacae]	72	1,00E-69	91
#41300	Uncharacterised protein [Klebsiella pneumoniae]	CTZ96174.1|Uncharacterised protein [Escherichia coli]	100	0	99
#41720	hypothetical protein [Escherichia coli]	SQY53850.1|Uncharacterised protein [Escherichia coli]	91	9,00E-154	93
#4240	hypothetical protein [Escherichia coli]	KFZ99631.1|hypothetical protein DP20_3361 [Shigella flexneri]	70	6,00E-71	98
#42665	hypothetical protein [Salmonella enterica]	STJ31935.1|Uncharacterised protein [Escherichia coli]	100	2,00E-75	87
#4309	hypothetical protein [Escherichia coli]	ERE37650.1|putative membrane domain protein [Escherichia coli B90]	100	0	93
#4322	hypothetical protein [Escherichia coli]	EHV43549.1|YD repeat domain protein [Escherichia coli DEC5C]	99	3,00E-171	88
#43543	hypothetical protein [Salmonella enterica]	SSL95586.1|Uncharacterised protein [Klebsiella pneumoniae]	66	2,00E-100	61
#43550	RNase P [Edwardsiella ictaluri]	KMQ79224.1|RNase P [Edwardsiella ictaluri]	100	6,00E-15	51
#4364	hypothetical protein [Shigella sonnei]	EGB41200.1|hypothetical protein EREG_03224 [Escherichia coli H120]	100	0	92
#44181	hypothetical protein [Escherichia coli]	Q3ECS7|hypothetical 77K protein (spoT 3' region) - Escherichia coli	100	0	96
#44332	hypothetical protein [Escherichia coli]	EYV67893.1|hypothetical protein BX25_10350 [Escherichia coli O121:H19 str. 2009C-4659]	91	9,00E-52	100
#44483	hypothetical protein [Escherichia coli]	CSP85862.1|Uncharacterised protein [Shigella sonnei]	87	2,00E-67	99
#44484	hypothetical protein [Escherichia coli]	CSP85871.1|Uncharacterised protein [Shigella sonnei]	100	1,00E-42	100
#44519	hypothetical protein [Escherichia coli]	WP_012602578.1|hypothetical protein [Escherichia coli]	90	4,00E-74	89
#44582	hypothetical protein [Delftia acidovorans]	CSQ03185.1|Uncharacterised protein [Shigella sonnei]	83	9,00E-121	92
#44587	hypothetical protein [Escherichia coli]	EFU32062.1|hypothetical protein HMPREF9350_06135 [Escherichia coli MS 85-1]	100	3,00E-82	95
#44593	hypothetical protein [Escherichia coli]	CTX21150.1|Uncharacterised protein [Escherichia coli]	91	2,00E-162	96
#44678	hypothetical protein [Escherichia coli]	EZD90611.1|hypothetical protein BX05_04490 [Escherichia coli O157:NM str. 08-4540]	65	2,00E-81	99
#45444	hypothetical protein [Salmonella enterica]	SSW81605.1|Uncharacterised protein [Klebsiella pneumoniae]	45	4,00E-83	62
#4556	hypothetical protein [Microlunatus phosphovorus]	CSQ66332.1|Uncharacterised protein [Shigella sonnei]	100	1,00E-78	98
#4685	hypothetical protein [Escherichia coli]	AAP16044.1|hypothetical protein S0533 [Shigella flexneri 2a str. 2457T]	100	4,00E-107	99
#46888	DNA topoisomerase I [Halorubrum terrestre]	WP_032206968.1|hypothetical protein [Escherichia coli]	100	2,00E-128	99
#46985	hypothetical protein [Escherichia coli]	SRA75415.1|Uncharacterised protein [Escherichia coli]	100	6,00E-63	94
#4699	hypothetical protein [Salmonella enterica]	EGB79723.1|hypothetical protein HMPREF9533_05504 [Escherichia coli MS 60-1]	64	3,00E-36	97
#47092	hypothetical protein [Shigella flexneri]	OEL95601.1|hypothetical protein BHF16_21920 [Escherichia coli]	100	5,00E-14	91
#47320	hypothetical protein [Escherichia coli]	GAR62940.1|hypothetical protein NGUA15_04763 [Salmonella enterica]	54	5,00E-15	32
#47737	hypothetical protein [Siccibacter colletis]	CSS27840.1|Uncharacterised protein [Shigella sonnei]	100	4,00E-33	92
#47780	hypothetical protein [Rhodococcus qingshengii]	KDQ00078.1|hypothetical protein EN35_16580 [Rhodococcus qingshengii]	46	1,00E-41	97

ID CP008957.1_	blastp 2016	blast 2018			
		hit	coverage [%]	E-value	identity [%]
#47871	hypothetical protein [Enterobacter aerogenes]	WP_015703582.1|hypothetical protein [Klebsiella aerogenes]	95	3,00E-04	44
#4850	hypothetical protein, partial [Escherichia coli]	EDU83285.1|hypothetical protein ECH7EC4501_2313 [Escherichia coli O157:H7 str. EC4501]	43	1,00E-22	100
#48623	hypothteical protein [Dickeya dadantii]	EMD08296.1|Modulator of drug activity (mda66) [Escherichia coli SEPT362]	100	1,00E-128	93
#48798	hypothetical protein [Salmonella enterica]	STE81992.1|protein [Escherichia coli]	93	3,00E-44	99
#48955	hypothetical protein [Salmonella enterica]	EJA17391.1|hypothetical protein SEEN447_15883 [Salmonella enterica subsp. enterica serovar Newport str. CVM 19447]	65	2,00E-06	41
#49065	hypothetical protein [Shigella flexneri]	CST05653.1|Uncharacterised protein [Shigella sonnei]	72	2,00E-63	96
#49071	Uncharacterised protein [Klebsiella pneumoniae]	OAC23075.1|hypothetical protein EC2772a_47c02870 [Escherichia coli]	96	5,00E-165	98
#49207	membrane protein [Kluyvera ascorbata]	CUQ98185.1|Inner membrane protein [Escherichia coli]	98	2,00E-56	98
#49908	hypothetical protein [Escherichia coli]	WP_077769272.1|hypothetical protein [Escherichia coli]	89	1,00E-34	85
#50186	amino acid acetyltransferase [Escherichia coli]	CSE76570.1|Uncharacterised protein [Shigella sonnei]	100	2,00E-87	98
#50287	Uncharacterised protein [Klebsiella pneumoniae]	SRN31903.1|Uncharacterised protein [Shigella flexneri]	100	5,00E-180	99
#50318	hypothetical protein, partial [Pseudomonas amygdali]	CSF52589.1|Uncharacterised protein [Shigella sonnei]	72	2,00E-98	98
#50487	hypothetical protein [Escherichia coli]	ABE08596.1|hypothetical protein UTI89_C3144 [Escherichia coli UTI89]	100	2,00E-56	92
#5070	hypothetical protein [Shigella sonnei]	CSF04814.1|Uncharacterised protein [Shigella sonnei]	69	1,00E-103	97
#51092	hypothetical protein [Xenorhabdus poinarii]	CEU73156.1|Uncharacterised protein [Salmonella enterica subsp. enterica serovar Typhi]	99	2,00E-76	72
#51542	heat-shock protein GrpE [Kluyvera ascorbata]	CTQ82973.1|Protein GrpE (fragment) [Escherichia coli]	100	2,00E-103	96
#51705	hypothetical protein [Escherichia coli]	WP_101983432.1|alpha-ketoglutarate permease [Escherichia coli]	100	8,00E-16	92
#51729	hypothetical protein [Escherichia coli]	SAJ34150.1|Uncharacterised protein [Enterobacter cloacae]	100	3,00E-46	52
#52107	hypothetical protein, partial [Catenibacterium mitsuokai]	BAK12240.1|hypothetical protein PAJ_2160 [Pantoea ananatis AJ13355]	80	9,00E-71	55
#52152	hypothetical protein [Yersinia pseudotuberculosis]	WP_012606359.1|hypothetical protein [Yersinia pseudotuberculosis]	83	5,00E-09	56
#52292	hypothetical protein [Cronobacter turicensis]	CSR95369.1|Uncharacterised protein [Shigella sonnei]	96	8,00E-48	99
#52317	membrane protein [Salmonella enterica]	WP_001322717.1|hypothetical protein [Escherichia coli]	100	4,00E-62	97
#52524	hypothetical protein [Escherichia coli]	KUG89326.1|hypothetical protein ARC90_05285 [Escherichia coli]	100	1,00E-42	96
#52682	Uncharacterised protein [Klebsiella pneumoniae]	CSS26714.1|Uncharacterised protein [Shigella sonnei]	100	2,00E-13	45
#52694	Uncharacterised protein [Klebsiella pneumoniae]	CSH61930.1|Uncharacterised protein [Shigella sonnei]	99	0	94
#53470	mannitol-1-phosphate 5-dehydrogenase [Escherichia albertii]	WP_115724152.1|hypothetical protein [Escherichia coli]	82	3,00E-04	39
#54026	hypothetical protein [Shigella flexneri]	OWF22517.1|hypothetical protein A8M76_19890 [Escherichia coli]	57	1,00E-41	100
#54145	ribonucleotide-diphosphate reductase subunit alpha, partial [Klebsiella pneumoniae]	EJK89822.1|hypothetical protein UUU_29410 [Klebsiella pneumoniae subsp. pneumoniae DSM 30104]	50	7,00E-143	72
#54484	hypothetical protein [Edwardsiella ictaluri]	AAN81177.1|Hypothetical protein c2723 [Escherichia coli CFT073]	100	9,00E-93	94
#5499	MULTISPECIES: hypothetical protein [Enterobacteriaceae]	OZX79305.1|hypothetical protein CIJ90_09825 [Escherichia coli]	100	2,00E-39	81
#55866	hypothetical protein [Escherichia coli]	CSF32452.1|Uncharacterised protein [Shigella sonnei]	87	1,00E-127	98
#56283	hypothetical protein [Escherichia coli]	AAN82123.1|Hypothetical protein c3675 [Escherichia coli CFT073]	100	5,00E-36	79
#56324	hypothetical protein [Escherichia coli]	KNF59067.1|hypothetical protein WQ67_25275 [Escherichia coli]	70	4,00E-35	90
#56507	hypothetical protein [Serratia symbiotica]	AAG57042.1|hypothetical protein Z3136 [Escherichia coli O157:H7 str. EDL933]	100	7,00E-52	100
#56564	hypothetical protein [Escherichia coli]	EIP31799.1|hypothetical protein ECEC4013_2127 [Escherichia coli EC4013]	100	2,00E-22	71
#56801	molecular chaperone Tir [Escherichia coli]	EZD02136.1|molecular chaperone Tir [Escherichia coli O157:H7 str. K5852]	100	4,00E-18	85

ID CP008957.1_	blastp 2016	blast 2018			
		hit	coverage [%]	E-value	identity [%]
#56803	molecular chaperone Tir [Escherichia coli]	OVD47041.1|molecular chaperone Tir [Escherichia coli]	100	2,00E-18	85
#56805	molecular chaperone Tir [Escherichia coli]	OVD47041.1|molecular chaperone Tir [Escherichia coli]	100	8,00E-18	83
#56807	hypothetical protein [Escherichia coli]	OVC49966.1|molecular chaperone Tir [Escherichia coli]	100	5,00E-15	74
#57079	hypothetical protein [Escherichia coli]	SSM34220.1|Uncharacterised protein [Klebsiella pneumoniae]	100	3,00E-17	56
#5713	hypothetical protein [Salmonella enterica]	SUG84241.1|Uncharacterised protein [Salmonella enterica subsp. enterica]	76	2,00E-17	53
#57364	acid-inducible small membrane-associated protein [Escherichia coli]	EYD84814.1|hypothetical protein AC26_1812 [Escherichia coli 1-176-05_S3_C2]	100	9,00E-25	92
#5740	hypothetical protein [Pantoea sp. SL1_M5]	CUN74125.1|Uncharacterised protein [Collinsella aerofaciens]	93	1,00E-29	34
#5749	hypothetical protein [Escherichia coli]	no hits			
#57774	hypothetical protein [Salmonella enterica]	SUX64632.1|Uncharacterised protein [Citrobacter amalonaticus]	98	4,00E-24	48
#58806	hypothetical protein [Escherichia coli]	KXG70337.1|hypothetical protein LT30_03315 [Escherichia coli]	83	2,00E-58	95
#58981	hypothetical protein [Escherichia coli]	WP_032158850.1|hypothetical protein [Escherichia coli]	100	2,00E-37	97
#58990	hypothetical protein [Escherichia coli]	SAQ24885.1|Uncharacterised protein [Klebsiella oxytoca]	62	2,00E-11	48
#59154	prolyl-tRNA synthetase [Polaromonas naphthalenivorans]	ESE27032.1|hypothetical protein HMPREF1623_00405 [Escherichia coli 910096-2]	100	3,00E-89	96
#59606	hypothetical protein [Shigelly dysenteriae]	CSQ02940.1|Uncharacterised protein [Shigella sonnei]	100	9,00E-43	93
#59705	cell surface protein [Escherichia coli]	CTZ85300.1|Uncharacterised protein [Escherichia coli]	100	2,00E-161	96
#59720	hypothetical protein [Curvibacter lanceolatus]	WP_096217753.1|hypothetical protein [Enterobacter kobei]	85	1,00E-25	65
#59762	hypothetical protein [Plesiomonas shigelloides]	EDX39001.1|conserved hypothetical protein [Escherichia coli 101-1]	100	1,00E-63	82
#60520	hypothetical protein [Shigella sonnei]	WP_072047577.1|hypothetical protein [Klebsiella variicola]	59	9,00E-04	58
#60767	hypothetical protein, partial [Salmonella enterica]	EYZ42022.1|hypothetical protein BW91_12335 [Escherichia coli O91:H14 str. 06-3691]	100	2,00E-69	100
#61510	hypothetical protein [Shigelly dysenteriae]	WP_011378736.1|peptidase [Shigella dysenteriae]	89	1,00E-22	87
#61578	hypothetical protein [Escherichia coli]	KDV50094.1|hypothetical protein BU57_28745 [Escherichia coli O121:H19 str. 2011C-3609]	91	3,00E-79	97
#61743	hypothetical protein [Enterobacter cloacae]	KUQ90625.1|hypothetical protein AWI27_21855 [Enterobacter hormaechei subsp. steigerwaltii]	92	2,00E-32	59
#61771	hypothetical protein [Shigella flexneri]	no hits			
#61976	hypothetical protein [Escherichia albertii]	SRN42906.1|Uncharacterised protein [Shigella flexneri]	100	2,00E-62	97
#61992	hypothetical protein, partial [Escherichia coli]	KFZ97440.1|hypothetical protein DP20_3635 [Shigella flexneri]	98	1,00E-24	93
#62034	hypothetical protein [Escherichia coli]	CSR91768.1|Uncharacterised protein [Shigella sonnei]	100	7,00E-57	95
#62406	hypothetical protein [Escherichia coli]	no hits			
#63161	hypothetical protein [Escherichia coli]	AAM88318.1|unknown [Escherichia coli]	75	2,00E-73	84
#6343	hypothetical protein [Shigella dysenteriae]	ESA98022.1|hypothetical protein HMPREF1620_00995 [Escherichia coli 909945-2]	77	5,00E-35	93
#63514	hypothetical protein [Enterobacter cloacae]	CEE08775.1|hypothetical protein BN1008_3296 [Escherichia coli]	88	0	98
#63696	trehalase [Gramella echinicola]	ADD56055.1|trehalase [Escherichia coli O55:H7 str. CB9615]	99	8,00E-83	100
#64399	hypothetical protein [Escherichia coli]	SRN41932.1|Uncharacterised protein [Shigella flexneri]	100	7,00E-61	85
#64620	hypothetical protein [Escherichia coli]	CTW98764.1|Uncharacterised protein [Escherichia coli]	98	4,00E-112	84
#64638	hypothetical protein [Escherichia coli]	APL78653.1|hypothetical protein RG72_02145 [Escherichia coli]	57	3,00E-51	93
#64725	hypothetical protein [Escherichia coli]	EKH68938.1|hypothetical protein ECPA49_3883 [Escherichia coli PA49]	87	8,00E-25	100
#64727	hypothetical protein [Escherichia coli]	RFQ70460.1|hypothetical protein CRE03_05570 [Escherichia coli]	100	3,00E-94	94
#64825	hypothetical protein [Edwardsiella piscicida]	KKY92492.1|hypothetical protein OA48_18610 [Klebsiella pneumoniae]	90	7,00E-18	48

ID CP008957.1_	blastp 2016	blast 2018			
		hit	coverage [%]	E-value	identity [%]
#65083	hypothetical protein [Salmonella enterica]	EGA09167.1|hypothetical protein SEEM0055_17245 [Salmonella enterica subsp. enterica serovar Montevideo str. MB110209-0055]	90	9,00E-35	71
#65241	hypothetical protein [Shigella sonnei]	CSS38474.1|Uncharacterised protein [Shigella sonnei]	100	5,00E-66	100
#65479	hypothetical protein [Xenorhabdus bovienii]	EIF85570.1|hypothetical protein ESMG_02720 [Escherichia coli M919]	81	2,00E-88	85
#65509	hypothetical protein, partial [Klebsiella oxytoca]	ABJ03620.1|conserved hypothetical protein [Escherichia coli APEC O1]	100	3,00E-33	73
#65528	hypothetical protein [Burkholderia sacchari]	EFK44487.1|hypothetical protein HMPREF9346_03897 [Escherichia coli MS 119-7]	100	2,00E-35	63
#65561	hypothetical protein, partial [Escherichia coli]	OWB95046.1|hypothetical protein A8M80_24920 [Escherichia coli]	52	6,00E-33	95
#65606	hypothetical protein [Escherichia coli]	EOU49835.1|hypothetical protein WC3_02778 [Escherichia coli KTE35]	100	0	96
#65639	hypothetical protein [Escherichia coli]	EIG43084.1|hypothetical protein ESTG_03907 [Escherichia coli B799]	29	3,00E-01	66
#6574	hypothetical protein [Escherichia coli]	CTT54903.1|putative signaling protein [Escherichia coli]	97	1,00E-112	94
#6662	hypothetical protein, partial [Escherichia coli]	CSS45419.1|Uncharacterised protein [Shigella sonnei]	96	5,00E-70	100
#66748	outer membrane protein [Escherichia coli]	WP_001303605.1|outer membrane protein [Escherichia coli]	100	3,00E-51	100
#67577	hypothetical protein [Escherichia coli]	EIP02888.1|hypothetical protein ECTW09195_1228 [Escherichia coli TW09195]	51	1,00E-20	100
#67649	hypothetical protein [Enterobacter cloacae]	WP_108703644.1| hypothetical protein [Klebsiella michiganensis]	96	7,00E-22	37
#68480	hypothetical protein [Xenorhabdus bovienii]	EIF85570.1|hypothetical protein ESMG_02720 [Escherichia coli M919]	81	2,00E-88	85
#68510	hypothetical protein, partial [Klebsiella oxytoca]	ABJ03620.1|conserved hypothetical protein [Escherichia coli APEC O1]	100	3,00E-33	73
#68529	hypothetical protein [Burkholderia sacchari]	WP_077874033.1|DUF932 domain-containing protein [Escherichia coli]	100	3,00E-41	72
#68609	hypothetical protein [Escherichia coli]	EOU49835.1|hypothetical protein WC3_02778 [Escherichia coli KTE35]	100	0	96
#68642	hypothetical protein [Escherichia coli]	EIG43084.1|hypothetical protein ESTG_03907 [Escherichia coli B799]	29	1,00E-03	66
#69144	hypothetical protein [Escherichia coli]	WP_105467182.1|hypothetical protein [Escherichia coli]	100	8,00E-21	53
#70284	hypothetical protein [Escherichia coli]	WP_012578864.1|hypothetical protein [Escherichia coli]	100	7,00E-94	100
#7056	hypothetical protein [Escherichia coli]	WP_032247156.1|hypothetical protein [Escherichia coli]	100	2,00E-36	98
#70740	hypothetical protein [Klebsiella pneumoniae]	SLX25598.1|Uncharacterised protein [Klebsiella pneumoniae]	100	1,00E-135	69
#70901	hypothetical protein [Delftia acidovorans]	CSQ03185.1|Uncharacterised protein [Shigella sonnei]	85	6,00E-120	91
#70906	hypothetical protein [Escherichia coli]	EFU32062.1|hypothetical protein HMPREF9350_06135 [Escherichia coli MS 85-1]	100	3,00E-82	95
#70912	hypothetical protein [Escherichia coli]	SQY53850.1|Uncharacterised protein [Escherichia coli]	91	9,00E-154	93
#70925	hypothetical protein [Escherichia coli]	STL43598.1| Uncharacterised protein [Escherichia coli]	67	5,00E-50	93
#71041	hypothetical protein [Escherichia coli]	ABB60814.1|hypothetical protein SDY_0621 [Shigella dysenteriae Sd197]	100	0	97
#71252	hypothetical protein [Escherichia coli]	no hits			
#71473	hypothetical protein [Salmonella enterica]	CSP94314.1|Uncharacterised protein [Shigella sonnei]	95	2,00E-129	95
#71490	hypothetical protein [Escherichia coli]	AAN79168.1| Hypothetical protein c0693 [Escherichia coli CFT073]	75	7,00E-117	99
#7161	hypothetical protein [Escherichia coli]	EFZ61405.1|hypothetical protein ECOK1180_5553 [Escherichia coli OK1180]	100	0	99
#72178	hypothetical protein [Escherichia coli]	CCK46665.1|putative uncharacterized protein [Escherichia coli chi7122]	100	1,00E-60	86
#72195	hypothetical protein [Escherichia coli]	STL43598.1|Uncharacterised protein [Escherichia coli]	70	2,00E-07	52
#72348	hypothetical protein [Escherichia coli]	RDR86465.1|hypothetical protein C3999_02369 [Escherichia marmotae]	99	5,00E-47	76
#72358	hypothetical protein [Escherichia sp. TW09308]	CZW58126.1|Uncharacterised protein [Enterobacter cloacae]	100	4,00E-07	40
#72370	hypothetical protein [Escherichia coli]	CZV81019.1| Uncharacterised protein [Enterobacter cloacae]	98	4,00E-16	40
#72384	hypothetical protein [Escherichia coli]	SAD28616.1|Uncharacterised protein [Enterobacter cloacae]	100	2,00E-05	42
#72396	hypothetical protein [Escherichia coli]	CZW58126.1|Uncharacterised protein [Enterobacter cloacae]	100	5,00E-04	42

ID CP008957.1_	blastp 2016	blast 2018			
		hit	coverage [%]	E-value	identity [%]
#72410	hypothetical protein [Escherichia coli]	CZU80210.1|Uncharacterised protein [Enterobacter cloacae]	100	4,00E-04	34
#72423	hypothetical protein [Escherichia coli]	CZX07109.1|Uncharacterised protein [Enterobacter cloacae]	100	9,00E-04	35
#73837	hypothetical protein, partial [Escherichia coli]	CUX85737.1|conserved hypothetical protein [Escherichia coli]	86	2,00E-67	89
#7384	hypothetical protein [Citrobacter rodentium]	CBJ03433.1|conserved hypothetical protein [Escherichia coli ETEC H10407]	100	2,00E-54	93
#74632	hypothetical protein [Escherichia coli]	ABM54879.1|hypothetical protein ECf0004 [Escherichia coli]	100	0	99
#74663	hypothetical protein [Shigella sp.]	ABM54879.1|hypothetical protein ECf0004 [Escherichia coli]	100	6,00E-14	50
#74683	hypothetical protein [Escherichia coli]	CSQ01344.1|Uncharacterised protein [Shigella sonnei]	43	6,00E-04	53
#74703	hypothetical protein [Shigella sonnei]	EGB41200.1|hypothetical protein EREG_03224 [Escherichia coli H120]	100	2,00E-170	82
#75212	hypothetical protein [Escherichia coli]	WP_012602173.1|hypothetical protein [Escherichia coli]	100	6,00E-58	100
#75282	ABC transporter permease [Salmonella enterica]	SBV63744.1|conserved hypothetical protein [uncultured Citrobacter sp.]	100	8,00E-63	83
#75364	hypothetical protein [Dickeya dadantii]	EJK92626.1|hypothetical protein UUU_06120 [Klebsiella pneumoniae subsp. pneumoniae DSM 30104]	81	4,00E-14	43
#75595	hypothetical protein [Dickeya sp.]	CSO91250.1|Uncharacterised protein [Shigella sonnei]	84	8,00E-151	89
#8326	hypothetical protein [Plesiomonas shigelloides]	ETJ57615.1|hypothetical protein Q456_0217590 [Escherichia coli ATCC BAA-2193]	99	1,00E-83	84
#8389	hypothetical protein, partial [Cronobacter universalis]	EYY33043.1|hypothetical protein BX84_17025 [Escherichia coli O121:H19 str. 2010C-4989]	100	7,00E-22	64
#8548	molecular chaperone Tir [Escherichia coli]	OVC49966.1|molecular chaperone Tir [Escherichia coli]	100	4,00E-17	80
#8550	hypothetical protein [Escherichia coli]	OVC49966.1|molecular chaperone Tir [Escherichia coli]	100	4,00E-17	80
#8552	hypothetical protein [Escherichia coli]	OVC49966.1|molecular chaperone Tir [Escherichia coli]	100	4,00E-17	80
#8554	hypothetical protein [Escherichia coli]	OVC49966.1|molecular chaperone Tir [Escherichia coli]	100	4,00E-17	80
#8646	hypothetical protein [Shigella sonnei]	SAP80750.1|Uncharacterised protein [Klebsiella oxytoca]	44	2,00E-04	38
#8749	hypothetical protein [Serratia liquefaciens]	CDP76543.1|Putative uncharacterized protein [Escherichia coli D6-117.29]	100	3,00E-46	97
#9297	proline dehydrogenase [Escherichia coli]	WP_072018896.1|proline dehydrogenase [Escherichia coli]	90	1,00E-16	88
#941	hypothetical protein [Escherichia albertii]	OSM88468.1|hypothetical protein L317_01210 [Escherichia coli SHECO003]	98	1,00E-43	100
#9783	hypothetical protein [Escherichia coli]	EFZ61405.1|hypothetical protein ECOK1180_5553 [Escherichia coli OK1180]	100	0	99

Supplementary table S11: Overview of shadow ORFs with conserved domains and their blast hit and prosite pattern. Those shadow ORFs with 'hypothetical' blast hit and a more specified CD hit are bold.

shadow ORF ID	BLAST hit	CD	Prosite pattern	Protein family domain	footprint
CP008957.1_11769	transcriptional regulator [Pantoea ananatis]	Transcriptional regulator of acetoin/glycerol metabolism [Transcription]	PS00004 cAMP- and cGMP-dependent protein kinase phosphorylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;		no
CP008957.1_14354	MULTISPECIES: restriction endonuclease [Enterobacteriaceae]	Restriction alleviation protein Lar	PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site;	PF14354.1 Restriction alleviation protein Lar;	no
CP008957.1_14604	hypothetical protein [Nitratireductor pacificus]	Protein of unknown function (DUF1602)	PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site;	PF07673.9 Protein of unknown function (DUF1602);	no
CP008957.1_18076	**hypothetical protein, partial [Neisseria flavescens]**	**YecR-like lipoprotein**	**PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;**	**PF13992.1 YecR-like lipoprotein;**	**no**
CP008957.1_23719	NAD-specific glutamate dehydrogenase [Haloferax denitrificans]	NAD-specific glutamate dehydrogenase	PS00004 cAMP- and cGMP-dependent protein kinase phosphorylation site; PS00005 Protein kinase C phosphorylation site; PS00007 Tyrosine kinase phosphorylation site; PS00008 N-myristoylation site;	PF10712.4 NAD-specific glutamate dehydrogenase;	no
CP008957.1_23725	**hypothetical protein [Escherichia coli]**	**NAD-specific glutamate dehydrogenase**	**PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00008 N-myristoylation site;**		**no**
CP008957.1_28385	hypothetical protein [Escherichia albertii]	Protein of unknown function (DUF3521)	PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00008 N-myristoylation site; PS00342 Microbodies C-terminal targeting signal;	PF12035.3 Protein of unknown function (DUF3521);	yes
CP008957.1_30413	hypothetical protein, partial [Kitasatospora cheerisanensis]	Protein of unknown function (DUF1602)	PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;	PF07673.9 Protein of unknown function (DUF1602);	yes
CP008957.1_30417	hypothetical protein, partial [Pseudomonas mendocina]	Uncharacterized protein (COG4954)	PS00001 N-glycosylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;		no
CP008957.1_31222	hypothetical protein [Clostridium bolteae CAG:59]	Protein of unknown function (DUF1602)		PF07673.9 Protein of unknown function (DUF1602);	no
CP008957.1_32339	type III secretion system protein SepZ [Escherichia albertii]	SepZ	PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;	PF06066.6 SepZ;	no
CP008957.1_32417	sugar transporter [Escherichia albertii]	sugar efflux transporter (2A0120)			no
CP008957.1_33705	CFB group bacteria	RhaT (COG0697)	PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;	PF00892.15 EamA-like transporter family;	no
CP008957.1_33739	**hypothetical protein [Escherichia coli]**	**Glutamylglutaminyl-tRNA synthetase (PRK12410)**	**PS00005 Protein kinase C phosphorylation site; PS00008 N-myristoylation site; PS00016 Cell attachment sequence;**		**no**
CP008957.1_34370	hypothetical protein [Escherichia albertii]	Protein of unknown function (DUF3521)	PS00005 Protein kinase C phosphorylation site;	PF12035.3 Protein of unknown function (DUF3521);	yes
CP008957.1_36795	hypothetical protein [Escherichia albertii]	Protein of unknown function (DUF3521)	PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00008 N-myristoylation site; PS00016 Cell attachment sequence;	PF12035.3 Protein of unknown function (DUF3521);	no
CP008957.1_39433	hypothetical protein [Escherichia coli]	hypothetical protein (PRK09719)	PS00001 N-glycosylation site; PS00004 cAMP- and cGMP-dependent protein kinase phosphorylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site;		no
CP008957.1_39720	hypothetical protein [Escherichia albertii]	Protein of unknown function (DUF3521)	PS00005 Protein kinase C phosphorylation site; PS00008 N-myristoylation site;	PF12035.3 Protein of unknown function (DUF3521);	no
CP008957.1_46888	DNA topoisomerase I [Halorubrum terrestre]	DNA topoisomerase I (PRK06599)	PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;	PF01396.14 Topoisomerase DNA binding C4 zinc finger; PF13240.1 zinc-ribbon domain;	no
CP008957.1_49207	membrane protein [Kluyvera ascorbata]	Protein of unknown function (DUF2684)	PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;	PF10885.3 Protein of unknown function (DUF2684);	no
CP008957.1_5005	hypothetical protein, partial [Mesoplasma photuris]	Protein of unknown function (DUF1602)	PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;	PF07673.9 Protein of unknown function (DUF1602);	no
CP008957.1_5175	hypothetical protein, partial [Escherichia albertii]	Protein of unknown function (DUF3521)	PS00005 Protein kinase C phosphorylation site; PS00008 N-myristoylation site;	PF12035.3 Protein of unknown function (DUF3521);	yes
CP008957.1_57774	**hypothetical protein [Salmonella enterica]**	**Signal transduction protein containing GAF and PtsI domains (COG3605)**	**PS00008 N-myristoylation site;**		**no**
CP008957.1_59154	prolyl-tRNA synthetase [Polaromonas naphthalenivorans]	Protein of unknown function (DUF1289)	PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;	PF06945.8 Protein of unknown function (DUF1289);	no

CP008957.1_61884	hypothetical protein [Clostridium bolteae CAG:59]	Protein of unknown function (DUF1602)	PS00001 N-glycosylation site; PS00004 cAMP- and cGMP-dependent protein kinase phosphorylation site; PS00005 Protein kinase C phosphorylation site; PS00008 N-myristoylation site;	PF07673.9 Protein of unknown function (DUF1602);	no
CP008957.1_63696	trehalase [Gramella echinicola]	Trehalase	PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site; PS00342 Microbodies C-terminal targeting signal;		no
CP008957.1_6373	hypothetical protein [Streptomyces bikiniensis]	Protein of unknown function (DUF1602)	PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;	PF07673.9 Protein of unknown function (DUF1602);	no
CP008957.1_64881	hypothetical protein, partial [Microbacterium barkeri]	Protein of unknown function (DUF1602)	PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;	PF07673.9 Protein of unknown function (DUF1602);	no
CP008957.1_72227	hypothetical protein, partial [Kitasatospora cheerisanensis]	Protein of unknown function (DUF1602)	PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;	PF07673.9 Protein of unknown function (DUF1602);	no
CP008957.1_72414	hypothetical protein, partial [Escherichia coli]	cathepsin L protease (PTZ00203)	PS00005 Protein kinase C phosphorylation site; PS00008 N-myristoylation site; PS00139 Eukaryotic thiol (cysteine) proteases cysteine active site;		no
CP008957.1_74629	hypothetical protein [Escherichia coli]	hypothetical protein (PHA03375)	PS00001 N-glycosylation site; PS00004 cAMP- and cGMP-dependent protein kinase phosphorylation site; PS00005 Protein kinase C phosphorylation site; PS00008 N-myristoylation site;		yes
CP008957.1_76450	NAD-specific glutamate dehydrogenase [Halorubrum coriense]	NAD-specific glutamate dehydrogenase	PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00007 Tyrosine kinase phosphorylation site; PS00008 N-myristoylation site; PS00029 Leucine zipper pattern;	PF10712.4 NAD-specific glutamate dehydrogenase;	yes
CP008957.1_8702	**hypothetical protein, partial [Vibrio parahaemolyticus]**	**GnsA/GnsB family**	**PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site;**	**PF08178.6 GnsA/GnsB family;**	**yes**
CP008957.1_941	hypothetical protein [Escherichia albertii]	Protein of unknown function (DUF3521)	PS00005 Protein kinase C phosphorylation site;	PF12035.3 Protein of unknown function (DUF3521);	no
CP008957.1_18320	hypothetical protein [Streptomyces bikiniensis]	Protein of unknown function (DUF1602)	PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site;	PF07673.9 Protein of unknown function (DUF1602);	no
CP008957.1_22046	hypothetical protein [Streptomyces bikiniensis]	Protein of unknown function (DUF1602)	PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;	PF07673.9 Protein of unknown function (DUF1602);	no
CP008957.1_23693	hypothetical protein [Shigella flexneri]	Protein of unknown function (DUF3521)	PS00005 Protein kinase C phosphorylation site; PS00008 N-myristoylation site;	PF12035.3 Protein of unknown function (DUF3521);	no
CP008957.1_36888	hypothetical protein [Escherichia coli]	Protein of unknown function (DUF3521)	PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;		no
CP008957.1_37939	**hypothetical protein [Escherichia albertii]**	**Topoisomerase DNA binding C4 zinc finger; zinc-ribbon domain**	**PS00001 N-glycosylation site; PS00004 cAMP- and cGMP-dependent protein kinase phosphorylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site;**	**PF12035.3 Protein of unknown function (DUF3521);**	**yes**
CP008957.1_42054	hypothetical protein [Escherichia albertii]	Protein of unknown function (DUF3521)	PS00005 Protein kinase C phosphorylation site; PS00008 N-myristoylation site;	PF12035.3 Protein of unknown function (DUF3521);	yes
CP008957.1_54522	hypothetical protein [Clostridium bolteae CAG:59]	Protein of unknown function (DUF1602)	PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site;	PF07673.9 Protein of unknown function (DUF1602);	no
CP008957.1_6795	hypothetical protein, partial [Kitasatospora cheerisanensis]	Protein of unknown function (DUF1602)	PS00001 N-glycosylation site; PS00005 Protein kinase C phosphorylation site; PS00008 N-myristoylation site;	PF07673.9 Protein of unknown function (DUF1602);	no
CP008957.1_69635	hypothetical protein [Clostridium bolteae CAG:59]	Protein of unknown function (DUF1602)		PF07673.9 Protein of unknown function (DUF1602);	no
CP008957.1_73806	hypothetical protein [Escherichia albertii]	Protein of unknown function (DUF3521)	PS00005 Protein kinase C phosphorylation site; PS00006 Casein kinase II phosphorylation site; PS00008 N-myristoylation site;	PF12035.3 Protein of unknown function (DUF3521);	no

Supplementary table S12: Phylostratum of sORFs with blastp hit of a predicted function blasted in 2015 in comparison to those blasted in 2018. In 2018, 23 sORFs did not have any blastp hit. The blastp search was conducted against the nr database of NCBI, E-value cutoff 1E-03. The farthermost related species was selected and the phylostratum was determined as described in section 2.3.1.

Phylostratum	**blast 2015**	**blast 2018**
E. coli	**101 (58.7%)**	57 (39.0%)
Escherichia	4 (2.3%)	2 (1.3%)
Enterobacteriaceae	29 (16.9%)	32 (21.9%)
Enterobacteriales	3 (1.7%)	10 (6.8%)
γ-Proteobacteria	1 (0.6%)	4 (2.7%)
Proteobacteria	4 (2.3%)	6 (4.1%)
Bacteria/Archaea	**30 (17.4%)**	**38 (26.0%)**
total	**172**	**146**

Supplementary table S13: Quantification cycles (cq) as measure for *asa* gene expression of stress adapted EHEC using RT-qPCR. The cq values of *asa* were normalized to the 16S rDNA expression (∆cq). EHEC was grown in LB or in LB + 450 mM NaCl and harvested at early exponential phase (OD_{600} = 0.2 - 0.3) and at exponential phase (OD_{600} = 0.7 - 0.8). Two negative controls were used: qPCR of all samples without reverse transcription (RT) and an un-transcribed region determined by RNAseq (NC). The cq value is the average of three technical replicates. The experiment was conducted in three biological replicates (R1-R3). Cq values higher than the threshold are marked as being not available (NA). The cq value negatively correlates with the relative mRNA concentration.

growth phase	**growth condition**	**cq (gene)**	**cq (16S)**	**∆cq (cq gene - cq 16S)**	**average ∆cq biological replicates**	**cq (16S) without RT**
early exponential phase	LB (R1)	19.0	11.4	7.6	7.67 ± 0.15	32.4
	LB (R2)	19.4	11.6	7.8		28.7
	LB (R3)	19.3	11.7	7.6		29.0
	NaCl (R1)	18.7	9.6	9.1	9.35 ± 0.47	35.0
	NaCl (R2)	17.7	8.6	9.0		38.7
	NaCl (R3)	17.9	8.0	9.9		35.4
exponential phase	LB (R1)	19.5	14.2	5.3	5.67 ± 0.72	NA
	LB (R2)	18.3	13.1	5.2		38.2
	LB (R3)	19.5	13.0	6.5		NA
	NaCl (R1)	18.2	14.5	3.6	3.53 ± 0.85	39.9
	NaCl (R2)	18.9	16.2	2.7		NA
	NaCl (R3)	17.2	12.9	4.3		38.6
	LB (NC)	32.0	15.6	- 16.4	-	16.6

Supplementary table S14: Quantification cycles (cq) as measure for *asa* gene expression of stress shocked EHEC determined by RT-qPCR. The cq values of *asa* were normalized to the 16S rDNA expression (∆cq). The cells were grown in LB or in LB + 450 mM NaCl and harvested before (t0), 30 min (t30), 60 min (t60) or 120 min (t120) after induction. Two negative controls were used: qPCR of all samples without reverse transcription (RT) and an untranscribed region determined by RNAseq (NC, table xy). The cq value is the average of three technical replicates. The experiment was conducted in three biological replicates (R1-R3). Cq values lower than the threshold are marked as being not available (NA). The cq value negatively correlates with the relative mRNA concentration.

Time after induction	Growth condition	cq (gene)	cq (16S)	∆cq (cq gene - cq 16S)	average ∆cq biological replicates	cq (16S) without RT
t0	LB (R1)	19.2	15.4	3.8	4.3 ± 0.5	39.2
	LB (R2)	18.1	13.2	4.8		36.9
	LB (R3)	17.4	13.2	4.1		31.9
	NaCl (R1)	20.1	16.2	3.8	4.5 ± 0.5	32.2
	NaCl (R2)	19.4	14.6	4.8		38.2
	NaCl (R3)	17.4	12.8	4.7		NA
t30	LB (R1)	20.3	14.6	5.7	4.4 ± 1.0	31.5
	LB (R2)	18.1	13.7	4.4		39.4
	LB (R3)	19.8	13.5	6.3		31.3
	NaCl (R1)	17.1	14.3	4.9	5.3 ± 0.4	34.4
	NaCl (R2)	18.3	12.9	5.4		37.6
	NaCl (R3)	17.9	12.3	5.6		31.5
t60	LB (R1)	18.5	15.2	3.3	4.2 ± 1.0	38.4
	LB (R2)	18.4	13.6	4.8		37.4
	LB (R3)	18.4	14.0	4.4		32.8
	NaCl (R1)	18.0	11.6	6.4	4.4 ± 1.9	39.2
	NaCl (R2)	17.3	14.4	2.9		34.0
	NaCl (R3)	16.7	11.0	5.7		32.4
	NaCl (R4)	18.3	15.6	2.7		31.2
t120	LB (R1)	16.8	14.4	2.4	3.1 ± 0.7	38.5
	LB (R2)	17.3	14.3	3.1		38.1
	LB (R3)	17.4	13.6	3.8		37.1
	NaCl (R1)	17.5	15.1	2.4	3.1 ± 0.6	33.8
	NaCl (R2)	17.0	13.4	3.6		34.2
	NaCl (R3)	17.8	14.6	3.2		NA

Supplementary table S15: RPKM values and coverage of RNAseq and RIBOseq of *asa* and of those homologues with a signal for transcription or translation.

Organism	data	RPKM	coverage
EHEC EDL933	RNAseq RIBOseq	13.58 14.63	0.29 0.30
EHEC Sakai	RNAseq RIBOseq	Not significant 8.40	- 0.23
E. coli LF82	RNAseq RIBOseq	Not significant 18.84	- 0.21
E. coli E2348, EPEC	RNAseq	119.22	1.00
S. flexneri 5a M90T	RNAseq	1420.24	1.00
C. rodentium ICC168	RNAseq	1269.63	1.00
S. praecaptivus HS1	RNAseq	431.281	1.00
C. sakazakii ATCC BAA-894	RNAseq	876.81	1.00

Supplementary table S16: Sequence similarities [%] of all experimentally characterized *asa* homologues. The length was always equal and correspond to the length of *asa* in EHEC. The species are sorted according to their distance to EHEC shown in the phylostratigraphy (section 3.5, figure 3.5.1). All sequence similarities were obtained by pairwise alignment of the amino acid sequence (EMBOSS Needle).

	EHEC EDL933	*Salmonella enterica* 287/91	*Citrobacter freundii* CFNIH1	*Serratia marcescens* WS1359	*Hafnia alvei* DSM30097
EHEC EDL933	**100**	**90**	**91**	**77**	**69**
Salmonella enterica 287/91		**100**	**91**	**72**	**68**
Citrobacter freundii CFNIH1			**100**	**75**	**67**
Serratia marcescens WS1359				**100**	**69**
Hafnia alvei DSM30097					**100**

Supplementary table S17: Quantification cycles (cq) as measure for gene expression of *asa* homologues determined by RT-qPCR. The cq values of *asa* were normalized to the 16S rDNA expression (∆cq). The following organisms were tested: *Citrobacter freundii* CFNIH1 (CF), *S. marcescens* WS1359 (SM), *S. enterica* serovar Gallinarum 287/91 (SE) and *H. alvei* DSM30097 (HA). All bacteria were grown in LB and harvested at at exponential phase (OD_{600} = 0.8 - 0.9). Two negative controls were used: qPCR of all samples without reverse transcription (RT) and an untranscribed region determined by RNAseq (NC, data in Supplementary table S10). The cq value is the average of three technical replicates. The experiment was conducted in three biological replicates (R1-R3). The cq value negatively correlates with the relative mRNA concentration.

Organism	cq (gene)	cq (16S)	∆cq (cq gene - cq 16S)	average ∆cq biological replicates	cq (16S) without RT
CF (R1)	19.2	12.7	6.4	6.62 ± 0.44	33.9
CF (R2)	20.1	13.4	6.7		26.0
CF (R3)	19.5	12.8	6.7		29.6
SM (R1)	23.6	16.0	7.6	8.40 ± 0.83	25.9
SM (R2)	22.8	14.3	8.5		27.0
SM (R3)	21.7	12.6	9.0		28.8
SE (R1)	19.4	14.8	4.6	4.95 ± 0.32	30.9
SE (R2)	18.9	13.5	5.4		33.7
SE (R3)	18.8	14.0	4.9		31.0
HA (R1)	32.6	12.8	19.8	19.13 ± 0.44	32.1
HA (R2)	32.1	13.2	18.9		32.3
HA (R3)	31.7	13.0	18.7		26.3

Supplementary table S18: Overlapping genes with phenotypes used for phylostratigraphic analysis.

ID or gene name	Start/stop sORF in EHEC	Start/stop mORF in EHEC	Reading frame	Overlap type
laoB	5216097 / 5216222	5214974 / 5216512	-2	embedded
Ano	2357439 / 2357744	2357569 / 2358573	-3	head-to-head
SlyC	2320123 / 2320317	2320253 / 2320687	-2	embedded
OGC 15	300575 / 300709	300073 / 301047	-2	embedded
OGC 23	570371 / 570574	569266 / 570486	-1	tail-to-tail
OGC 51	1110879 / 1110682	1110833 / 1110946	-2	head-to-head
OGC 57	1235822 / 1236622	1235816 / 1236880	-1	embedded
OGC 59 = *asa*	1247671 /1247934	1247562 /1248221	-2	tail-to-tail
OGC 75	1754037 / 1754375	1753745 / 1754146	-1	embedded
OGC 85	1985820 / 1985915	1985700 / 1986230	-1	tail-to-tail
OGC 106	2517026 / 2517304	2516432 / 2517259	-1	embedded
OGC 121	2758129 / 2758320	2757958 / 2758371	-1	tail-to-tail
OGC 167	3927557 / 3928027	3927557 / 3928027	-2	tail-to-tail
OGC 174	4044465 / 4044641	4044465 / 4074561	-1	head-to-head
OGC 194	4495934 / 4496524	4495934 / 4496524	-1	embedded
OGC 198	4528288 / 4528163	4528288 / 4528163	-2	tail-to-tail
OGC 226	5306609 / 5306929	5306609 / 5306929	-1	tail-to-tail
OGC 231	5353324 / 5353992	5353324 / 5353992	-1	tail-to-tail
OGC 241	5540205 / 5540378	5540205 / 5540378	-1	embedded

References, Supplementary material

Baek, J., J. Lee, K. Yoon and H. Lee (2017). "Identification of unannotated small genes in Salmonella." G3: Genes, Genomes, Genetics: g3. 116.036939.

Bartholomaus, A., I. Fedyunin, P. Feist, C. Sin, G. Zhang, A. Valleriani and Z. Ignatova (2016). "Bacteria differently regulate mRNA abundance to specifically respond to various stresses." Philos Trans A Math Phys Eng Sci **374**(2063).

Enomoto, S., A. Chari, A. L. Clayton and C. Dale (2017). "Quorum sensing attenuates virulence in Sodalis praecaptivus." Cell host & microbe **21**(5): 629-636. e625.

Hazen, T. H., S. C. Daugherty, A. Shetty, A. A. Mahurkar, O. White, J. B. Kaper and D. A. Rasko (2015). "RNA-Seq analysis of isolate-and growth phase-specific differences in the global transcriptomes of enteropathogenic Escherichia coli prototype isolates." Frontiers in microbiology **6**: 569.

Hücker, S. M., Z. Ardern, T. Goldberg, A. Schafferhans, M. Bernhofer, G. Vestergaard, C. W. Nelson, M. Schloter, B. Rost and S. Scherer (2017). "Discovery of numerous novel small genes in the intergenic regions of the Escherichia coli O157: H7 Sakai genome." PloS one **12**(9): e0184119.

Hwang, J. Y. and A. R. Buskirk (2017). "A ribosome profiling study of mRNA cleavage by the endonuclease RelE." Nucleic Acids Res **45**(1): 327-336.

Jing, C.-e., X.-j. Du, P. Li and S. Wang (2016). "Transcriptome analysis of Cronobacter sakazakii ATCC BAA-894 after interaction with human intestinal epithelial cell line HCT-8." Applied microbiology and biotechnology **100**(1): 311-322.

Kim, D., J. S.-J. Hong, Y. Qiu, H. Nagarajan, J.-H. Seo, B.-K. Cho, S.-F. Tsai and B. Ø. Palsson (2012). "Comparative analysis of regulatory elements between Escherichia coli and Klebsiella pneumoniae by genome-wide transcription start site profiling." PLoS genetics **8**(8): e1002867.

Kumar, S., G. Stecher and K. Tamura (2016). "MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets." Molecular biology and evolution **33**(7): 1870-1874.

Landstorfer, R. B. (2014). Comparative transcriptomics and translatomics to identify novel overlapping genes, active hypothetical genes, and ncRNAs in *Escherichia coli* O157:H7 EDL933. Doctorate, Technische Universität München.

Madison, J. D., E. A. Berg, J. G. Abarca, S. M. Whitfield, O. Gorbatenko, A. Pinto and J. L. Kerby (2017). "Characterization of Batrachochytrium dendrobatidis inhibiting bacteria from amphibian populations in Costa Rica." Frontiers in microbiology **8**: 290.

Ndah, E., V. Jonckheere, A. Giess, E. Valen, G. Menschaert and P. Van Damme (2017). "REPARATION: ribosome profiling assisted (re-) annotation of bacterial genomes." Nucleic acids research **45**(20): e168-e168.

Petty, N. K., T. Feltwell, D. Pickard, S. Clare, A. L. Toribio, M. Fookes, K. Roberts, R. Monson, S. Nair and R. A. Kingsley (2011). "Citrobacter rodentium is an unstable pathogen showing evidence of significant genomic flux." PLoS pathogens **7**(4): e1002018.

Prados, J., P. Linder and P. Redder (2016). "TSS-EMOTE, a refined protocol for a more complete and less biased global mapping of transcription start sites in bacterial pathogens." BMC genomics **17**(1): 849.

Vergara-Irigaray, M., M. C. Fookes, N. R. Thomson and C. M. Tang (2014). "RNA-seq analysis of the influence of anaerobiosis and FNR on Shigella flexneri." BMC genomics **15**(1): 438.