TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Mikrobielle Ökologie

RIBOseq-based discovery of non-annotated genes in *Escherichia coli* O157:H7 Sakai and their functional characterization

SARAH MARIA MARGRET HÜCKER

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

genehmigten Dissertation.

Vorsitzender: Prof. Dr. B. Küster

Prüfer der Dissertation:
1. Prof. Dr. S. Scherer
2. Prof. Dr. W. Liebl
3. Hon.-Prof. Dr. M. Schloter

Die Dissertation wurde am 31.08.2017 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt 18.12.2017 angenommen.

Table of content

Table of content I

Abstract V

Zusammenfassung VII

Publications and personal contribution IX

Abbreviations XV

List of figures XVII

List of tables XIX

Content of attached CD XXI

1. Introduction 1

1.1 Model organism *Escherichia coli* O157:H7 strain Sakai 1

1.1.1 Pathogenicity 2

1.1.2 Reservoirs 3

1.2 Short genes overlooked at genome annotation 4

1.3 Antiparallel overlapping genes 5

1.3.1 Definition 5

1.3.2 Origin and evolution of OLGs 8

1.3.3 OLGs in bacteria 10

1.3.4 Functionally characterized OLGs in EHEC 10

1.4 High-throughput discovery of novel genes facilitated by next generation sequencing technologies 11

1.4.1 Transcriptome sequencing (RNAseq) 11

1.4.2 Translatome sequencing (RIBOseq) 13

1.4.3 Discovery of novel protein coding genes using combined RNAseq and RIBOseq 13

1.5 Perspectives of this study 16

2. Results and Discussion 22

Part I: Transcription and translation of annotated EHEC genes 22

2.1 Reproducibility of RNAseq and RIBOseq data 22

2.1.1 Technical replicates 22

2.1.2 Biological replicates 23

2.2 Correlation of RNAseq to RIBOseq data 26

2.3 Correlation of RIBOseq data to proteome data 27

2.4 **Publication 1**: 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 29

2.5 Differential expression of annotated genes in LB compared to BHI 37

Part II: High-throughput discovery of novel genes 39

2.6 **Publication 2**: Discovery of numerous novel small genes in the intergenic regions of the *Escherichia coli* O157:H7 Sakai genome 39

2.7 Detection of 380 transcribed and translated antiparallel overlapping novel genes 73

2.7.1 Three-hundred-and-eighty antiparallel overlapping ORFs show evidence of translation 73

2.7.2 Reproducibility of RNAseq and RIBOseq signals of overlapping ORFs 75

2.7.3 Correlation of transcription and translation for overlapping ORFs 76

2.7.4 Size and RCV distribution of the potential novel genes 77

2.7.5 Translated OLGs with annotated homologs 78

2.7.6 Translated OLGs are differentially regulated under three growth conditions 79

2.7.7 Prediction of σ^{70} promoters 81

2.7.8 Prediction of ρ-independent terminators 81

2.7.9 Prediction of Shine-Dalgarno sequences 82

2.7.10 Factors influencing RIBOseq signals in experiments 83

2.7.11 Comparison to published RIBOseq data 84

Part III: Functional characterization of selected antiparallel overlapping genes 87

2.8 **Publication 3**: 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 87

2.9 **Publication 4**: The novel, antiparallel overlapping gene pair *anoG*/ECs2385 in *Escherichia coli* O157:H7 Sakai 112

2.10 **Publication 5**: Discovery of the novel gene *slyC* antiparallel overlapping the transcriptional regulator *slyA* in *Escherichia coli* O157:H7 Sakai, and characterization of the influence of L-arginine on its gene expression 138

2.11 Overlapping gene OLGECs0016 163

2.11.1 Transcription and translation 163

2.11.2 Characterization of the promoter region 165

2.11.3 EGFP-fusion protein expression 166

2.12 Overlapping gene OLGECs4930 167

2.12.1 Transcription and translation 167

2.12.2 Promoter activity 169

2.12.3 Cloning of a strand-specific knock-out mutant 170

2.13 A novel operon antiparallel overlapping to ECs0535 172

2.13.1 Transcription and translation of a tri-cistronic operon 172

2.13.2 Characterization of the polycistronic mRNA and of the promoter region 176

2.13.3 EGFP-fusion protein expression 177

2.14 Overlapping gene OLGZ5561 of EHEC EDL933 178

2.14.1 Summary of previous results concerning OLGZ5561 178

2.14.2 Predicted properties of OLOZ5561 180

2.14.3 Competitive growth experiments 181

2.14.4 Complementation of the erythromycin phenotype 182

2.14.5 qRT-PCR 184

2.14.6 Failure to show protein expression 185

3. Conclusion and Outlook 187

3.1 Improvements required for accurate detection of novel genes using RIBOseq 187

3.2 High-throughput detection of proteins corresponding to the novel genes 189

3.3 Functional characterization of novel genes 189

3.4 Characterization of the novel genes at a single cell level 191

3.5 Origin, evolution and fate of *de novo* genes 191

4. References 194

5. Supplement 204

Abstract

The food-borne pathogen *Escherichia coli* O157:H7 strain Sakai (EHEC) is the causative agent of hemorrhagic colitis and hemolytic-uremic syndrome. The genome of EHEC Sakai was sequenced in 2001 and it contains 5,358 annotated genes. However, intergenic regions might harbor additional (small) protein-coding genes. Due to the DNA triplet code, it is also possible that the sequences of two genes overlap at a given locus. This study focuses on the detection of non-trivial, antiparallel overlapping genes (OLGs).

The transcriptome and the translatome of EHEC were determined using the high-throughput next generation sequencing methods RNAseq and RIBOseq. Three different growth conditions were tested, representing two optimal conditions (growth in LB and BHI medium), and one severe stress condition combining long-term cold and osmotic stress. The sequencing results show excellent reproducibility. The RIBOseq data of the annotated genes correlates to previously published proteome data. About one third of the annotated genes are differentially expressed at the transcriptional and/or translational level comparing either the two optimal conditions, or the stress condition to the optimal BHI condition. Riboswitches and the ncRNA DsrA were found to be involved in regulating gene expression after adaptation to cold and osmotic stress.

In addition, translation of non-annotated intergenic and antiparallel overlapping open reading frames (ORFs) was investigated. Amazingly, 465 intergenic ORFs and 380 OLGs show evidence of translation. The translatability was found to be similar to annotated genes, which supports the hypothesis these translated ORFs represent novel protein-coding genes. Further evidence for this claim includes the discovery of annotated homologs, differential regulation between growth conditions, presence of a reading frame in the sum signal of RIBOseq reads, and predicted regulatory elements (such as σ^{70} promoter, ρ-independent terminator, and a Shine-Dalgarno sequence).

Three of the novel OLG pairs discovered, namely *anoG*/ECs2385, *laoB*/ECs5115, and *slyC/slyA*, were functionally characterized. The transcription initiation sites were determined, and promoter activity of sequences upstream was detected. Presence of a protein was confirmed by expressing a C-terminally EGFP-fusion. Most importantly, a phenotype was observed in competitive growth experiments using EHEC wild type against a strand-specific, translationally arrested mutant of the respective gene. Conditions with high promoter activity or those causing a phenotype provide evidence for potential functions of the novel OLGs. Phylostratigraphic analyses of the annotated mother genes and the overlapping embedded genes indicate that the OLG originated *de novo* by overprinting in all three cases. Characterization of another four OLG candidates provides some evidence that they might be novel functional protein-coding genes as well. Interestingly, the first potential antiparallel overlapping operon consisting of three ORFs was discovered, which is encoded antisense to ECs0535.

All-in-all, this study shows that the genome of EHEC Sakai, and probably also other bacterial genomes, are under-annotated due to the systematical omission of small genes. Furthermore, genes encoded antiparallel to annotated genes seem to occur more frequently than previously presumed. The origin, evolution, and functions of these OLGs are interesting topics for future research.

Zusammenfassung

Das Lebensmittelpathogen *Escherichia coli* O157:H7 Stamm Sakai (EHEC) verursacht hämorrhagische Kolitis und das hämolytisch-urämische Syndrom. Das EHEC Sakai Genom wurde 2001 sequenziert und enthält 5.358 annotierte Gene. Jedoch könnten intergenische Bereiche weitere (kleine) protein-kodierende Gene beherbergen. Aufgrund der Triplet-Periodizität des genetischen Kodes ist es auch möglich, dass die Sequenzen zweier Gene überlappen. Der Schwerpunkt dieser Arbeit ist die Detektion nicht-trivialer, antiparallel überlappender Gene (OLGs).

Das Transkriptom und das Translatom von EHEC wurden mittels der Hochdurchsatz *next generation sequencing* Methoden RNAseq und RIBOseq ermittelt. Insgesamt wurden drei verschiedene Wachstumsbedingungen untersucht: Zwei optimale Bedingungen (Wachstum in LB und BHI Medium) und eine starke Stressbedingung, die Kälte- und osmotischen Stress kombiniert. Die Sequenzierergebnisse zeigen eine sehr gute Reproduzierbarkeit. Die RIBOseq Ergebnisse der annotierten Gene korrelieren mit bereits publizierten Proteom Daten. Bei einem Vergleich der optimalen Wachstumsbedingungen oder der Stressbedingung mit der optimalen BHI-Bedingung zeigt etwa ein Drittel der annotierten Gene differenzielle Regulation auf transkriptioneller und/oder translationaler Ebene. Riboswitches und die nicht-kodierende RNS DsrA sind an der Regulation der Genexpression nach Anpassung an Kälte- und osmotischen Stress beteiligt.

Außerdem wurde die Translation nicht-annotierter, intergenischer und antiparallel überlappender offener Leserahmen (ORFs) untersucht. Erstaunlicherweise zeigten 465 intergenische ORFs und 380 OLGs Hinweise auf Translation. Diese hatten eine vergleichbare Translationseffizienz wie annotierte Gene, was die Erkenntnis unterstützt, dass es sich bei diesen translatierten ORFs um protein-kodierende Gene handelt. Weitere Beweise dieser Hypothese sind die Entdeckung annotierter Homologe, die differenzielle Regulation zwischen den Wachstumsbedingungen, das Vorliegen eines Leserahmens im Summensignal der RIBOseq Daten und die Prädiktion regulatorischer Elemente (z.B. σ^{70} Promotoren, ρ-unabhängige Terminatoren und einer Shine-Dalgarno Sequenz).

Drei der entdeckten neuen OLG-Paare, nämlich *anoG*/ECs2385, *laoB*/ECs5115 und *slyC/slyA,* wurden funktional charakterisiert. Der Transkriptionsstart wurde bestimmt und es wurde Promotoraktivität der vor diesem liegenden Sequenz nachgewiesen. Die Expression eines C-terminalen EGFP-Fusionsproteins bestätigte das Vorhandensein eines Proteins. Das wichtigste Ergebnis bestand darin, dass Phänotypen in kompetitiven Wachstumsversuchen mit dem EHEC Wildtyp gegen eine strangspezifische, translational arretierte Mutante des jeweiligen Gens beobachtet wurden. Bedingungen, unter denen hohe Promotoraktivität und ein Phänotyp auftraten, geben Hinweise auf die möglichen Funktionen des neuen OLGs. Phylostratigrafische Analysen der annotierten und überlappenden Gene deuten darauf hin, dass das OLG in allen drei Fällen *de novo* durch *overprinting* entstanden ist. Die Charakterisierung von vier weiteren OLG Kandidaten liefert Hinweise, dass es sich ebenfalls um neue, funktionale, protein-kodierende Gene handeln könnte. Interessanterweise wurde das erste antiparallel überlappende Operon entdeckt, welches aus drei ORFs besteht und auf dem Gegenstrang von ECs0535 kodiert ist.

Zusammenfassend zeigt diese Arbeit, dass das Genom von EHEC Sakai und vermutlich auch andere bakterielle Genome unterannotiert sind, weil kleine Gene systematisch ausgeschlossen werden. Weiterhin scheinen Gene, die antiparallel zu annotierten Genen kodiert sind, häufiger aufzutreten als bisher angenommen. Die Entstehung, Evolution und Funktionen dieser OLGs sind spannende Fragen für zukünftige Untersuchungen.

Publications and personal contribution

1. **S. M. Hücker**, S. Simon, S. Scherer and K. Neuhaus, 2017. 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, Vol. 364, No. 2, doi: 10.1093/femsle/fnw262

Abstract

The enteric pathogen *Escherichia coli* O157:H7 Sakai (EHEC) is able to grow at lower temperatures compared to commensal *E. coli*. Growth at environmental conditions displays complex challenges different to those in a host. EHEC was grown at 37°C (control) and at 14°C with 4% NaCl, a combination of cold and osmotic stress as present in the food chain. Comparison of RNAseq and RIBOseq data provided a snap shot of ongoing transcription and translation, differentiating transcriptional and post-transcriptional gene regulation, respectively. Indeed, cold and osmotic stress related genes are simultaneously regulated at both levels, but translational regulation clearly dominates. Special emphasis was given to genes regulated by RNA secondary structures in their 5'UTRs, such as RNA thermometers and riboswitches, or genes controlled by small RNAs encoded *in trans*. The results reveal large differences in gene expression between short-time shock compared to adaptation in combined cold and osmotic stress. Whereas the majority of cold shock proteins, such as CspA, are translationally downregulated after adaptation, many osmotic stress genes are still significantly upregulated mainly translationally, but several also transcriptionally.

Personal contribution

The study was designed by S. M. Hücker, K. Neuhaus and S. Scherer. Experiments and data analysis were performed by S. M. Hücker. S. Simon provided her script to extract all ORFs ≥ 93 bp of the EHEC Sakai genome and her R script to calculate the RPKM values of all annotated genes. The manuscript was written by S. M. Hücker and edited by K. Neuhaus and S. Scherer.

2. **S. M. Hücker**, Z. Ardern, T. Goldberg, A. Schafferhans, M. Bernhofer, G. Vestergaard, C. W. Nelson, M. Schloter, B. Rost, S. Scherer and K. Neuhaus,

2017. Discovery of numerous novel small genes in the intergenic regions of the *Escherichia coli* O157:H7 Sakai genome. PLoS One, accepted.

Abstract

In the past, short protein-coding genes were often disregarded by genome annotation pipelines. Transcriptome sequencing (RNAseq) signals outside of annotated genes have usually been interpreted to indicate either ncRNA or pervasive transcription. Therefore, in addition to the transcriptome, the translatome (RIBOseq) of the enteric pathogen *Escherichia coli* O157:H7 strain Sakai was determined at two optimal growth conditions and a severe stress condition combining low temperature and high osmotic pressure. All intergenic open reading frames potentially encoding a protein of ≥ 30 amino acids were investigated with regard to coverage by transcription and translation signals and their translatability expressed by the ribosomal coverage value. This led to discovery of 465 unique, putative novel genes not yet annotated in this *E. coli* strain, which are evenly distributed over both DNA strands of the genome. For 255 of the novel genes, annotated homologs in other bacteria were found, and a machine-learning algorithm, trained on small protein-coding *E. coli* genes, predicted that 89% of these translated open reading frames represent *bona fide* genes. The remaining 210 putative novel genes without annotated homologs were compared to the 255 novel genes with homologs and to 250 short annotated genes of this *E. coli* strain. All three groups turned out to be similar with respect to their translatability distribution, fractions of differentially regulated genes, secondary structure composition, and the distribution of evolutionary constraint, suggesting that both novel groups represent legitimate genes. However, the machine-learning algorithm only recognized a small fraction of the 210 genes without annotated homologs. It is possible that these genes represent a novel group of genes, which have unusual features dissimilar to the genes of the machine-learning algorithm training set.

Personal contribution

The study was designed by S. M. Hücker, K. Neuhaus and S. Scherer. RNAseq and RIBOseq experiments were performed by S. M. Hücker. Additionally, S. M. Hücker identified the translated intergenic ORFs, searched annotated homologs, investigated differential regulation and predicted presence of promoters, terminators and a Shine-Dalgarno sequence. Z. Ardern performed tblastn analysis and created figures 1, 5 and

S2. A. Schafferhans and B. Rost performed PredictProtein analysis. T. Goldberg developed the machine-learning algorithm based on these predictions and with help of M. Bernhofer analysed the putative novel genes. G. Vestergaard and M. Schloter performed the reading frame determination. C. W. Nelson performed the k_A/k_S analysis. S. M. Hücker wrote the manuscript, which was edited by Z. Ardern, S. Scherer and K. Neuhaus.

3. **S. M. Hücker**, S. Vanderhaeghen, I. Abellan-Schneyder, R. Wecko, S. Simon, S. Scherer and K. Neuhaus, 2017. 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, under review.

Abstract

Background: Due to the DNA triplet code it is possible that the sequences of two or more protein-coding genes overlap to a large degree. However, such non-trivial overlaps are usually excluded by genome annotation pipelines and, thus, only a few overlapping gene pairs have been described in bacteria. In contrast, transcriptome and translatome sequencing showed many signals antisense to annotated genes, of which we analyzed an example gene pair in more detail. **Results:** A small open reading frame of *Escherichia coli* O157:H7 Sakai, designated *laoB* (L-arginine responsive overlapping gene), is embedded in reading frame -2 in the antisense strand of ECs5115, encoding a CadC-like transcriptional regulator. This overlapping gene shows evidence of transcription and translation in LB and BHI medium based on RNAseq and RIBOseq. The transcriptional start site is 289 bp upstream of the start codon and transcription termination is 155 bp downstream of the stop codon. Overexpression of LaoB fused to an EGFP reporter was possible. The sequence upstream of the transcriptional start site displayed strong promoter activity under different conditions, whereas promoter activity was significantly decreased in presence of L-arginine. A strand-specific translationally arrested mutant of *laoB* provided a significant growth advantage in competitive growth experiments in the presence of L-arginine compared to the wildtype, which returned to wildtype level after complementation of *laoB in trans*. A phylostratigraphic analysis indicated that the novel gene is restricted to the *Escherichia/Shigella* clade and might have originated recently by overprinting leading to the expression of part of the antisense strand of ECs5115. **Conclusions:** Here, we present evidence of a novel small protein-coding gene *laoB* encoded in the antisense frame -2 of the annotated gene ECs5115. Clearly, *LaoB* is evolutionary young and it originated in the *Shigella/Escherichia* clade by overprinting, which may be more important for the *de novo* evolution of novel bacterial genes than previously assumed.

Personal contribution

S. M. Hücker, S. Scherer and K. Neuhaus designed and planed the study. S. M. Hücker performed the 3' and 5' RACE experiments, the promoter activity assays, the competitive growth experiments and the complementation. S. Simon identified the optimal position for the strand-specific knock-out mutant and R. Wecko cloned the mutant. I. Abellan-Schneyder performed the expression of the EGFP-LaoB fusion protein. S. Vanderhaeghen did the phylostratigraphic analysis of the overlapping gene pair *laoB*/ECs5115. S. M. Hücker and S. Scherer wrote the manuscript, which was edited by K. Neuhaus.

4. **S. M. Hücker**, S. Vanderhaeghen, I. Abellan-Schneyder, R. Wecko, S. Scherer and K. Neuhaus, 2017. The novel, antiparallel overlapping gene pair *anoG*/ECs2385 in *Escherichia coli* O157:H7 Sakai. Submitted.

Abstract

Standard genome annotation presumes that only one protein is encoded at a given bacterial dsDNA locus. In contrast to this assumption, transcription and translation of an overlapping open reading frame of 186 bp length were discovered by RNAseq and RIBOseq experiments. This open reading frame is completely embedded in the annotated gene ECs2385 in *Escherichia coli* O157:H7 Sakai in the antiparallel reading frame -3. The open reading frame is transcribed as part of a polycistronic mRNA, which includes the annotated upstream gene ECs2384, encoding a murein lipoprotein. The transcriptional start site of the operon resides 38 bp upstream of the ECs2384 start codon, driven by a predicted σ^{70} promoter, which is constitutively active at different growth conditions. The polycistronic operon contains a ρ-independent terminator just upstream of the novel gene, significantly decreasing its transcription. The novel gene can be stably expressed as an EGFP-fusion protein and a translationally arrested

mutant shows a growth advantage under anaerobiosis in competitive growth compared to the wild type. Therefore, the novel antiparallel overlapping gene is named *anoG* – anaerobiosis responsive overlapping gene. A phylostratigraphic analysis indicates that *anoG* originated recently *de novo* by overprinting after the *Escherichia/Shigella* clade separated from other enterobacteria.

Personal contribution

S. M. Hücker, S. Scherer and K. Neuhaus designed and planned the study. S. M. Hücker performed the 5' RACE experiment, the promoter activity assays, the competitive growth experiments and the complementation. I. Abellan-Schneyder performed the expression of the AnoG-EGFP fusion proteins. R. Wecko cloned the mutant Δ*anoG*. S. Vanderhaeghen did the phylostratigraphic analysis of the overlapping gene pair *anoG/*ECs2385 and the annotated gene ECs2384. S. M. Hücker wrote the manuscript, which was edited by S. Scherer and K. Neuhaus.

5. **S. M. Hücker**, S. Vanderhaeghen, L. Dübbel, R. Wecko, S. Scherer and K. Neuhaus, 2017. Discovery of the novel gene *slyC* antiparallel overlapping the transcriptional regulator *slyA* in *Escherichia coli* O157:H7 Sakai, and characterization of the influence of L-arginine on its gene expression. Submitted.

Abstract

Transcription and translation of an open reading frame, named *slyC*, which overlaps antiparallel in reading frame -2 to the transcriptional regulator *slyA,* was detected in RNAseq and RIBOseq data of the enteric pathogen *Escherichia coli* O157:H7 Sakai. S*lyC* is annotated as an outer membrane lipoprotein in other *E. coli* strains, and the open reading frame is present in many *Enterobacteriales*, where it also overlaps to *slyA.* The transcriptional start site is located upstream of the annotated gene *slyB,* and RT-PCR confirmed polycistronic transcription of the operon *slyBC*. The sequence upstream of the transcriptional start contains the predicted consensus motif of two ARG boxes overlapping with the promoter, probably binding to the L-arginine dependent transcription factor ArgR. Promoter activity was decreased after L-arginine supplementation, and the strand-specific translationally arrested mutant Δ*slyC* shows a growth disadvantage in LB medium containing L-arginine compared to the wild type in competitive growth experiments. A SlyC-EGFP fusion protein could be expressed. Therefore, *slyC* represents an arginine regulated, novel antiparallel overlapping gene.

Personal contribution

S. M. Hücker, S. Scherer and K. Neuhaus designed and planned the study. S. M. Hücker performed the 3'/5' RACE experiments, the SlyC-EGFP fusion protein expression, the competitive growth experiments and the complementation. L. Dübbel cloned the mutant ∆*slyC*. R. Wecko performed the promoter activity assays. S. Vanderhaeghen did the phylostratigraphic analysis of the overlapping gene pair *slyC/slyA* and the annotated gene ECs2350. S. M. Hücker wrote the manuscript, which was edited by S. Scherer.

Abbreviations

µl	microliter
µM	micromolar
AA	amino acid
BHI	brain heart infusion broth
bp	basepair
cDNA	complementary deoxyribonucleic acid
cfu	colony forming unit
COS	combined cold and osmotic stress
CRISPR	clustered regulatory interspaced short palindromic repeats
DNA	deoxyribonucleic acid
EGFP	enhanced green fluorescent protein
EHEC	enterohemorrhagic *Escherichia coli*
ETEC	enterotoxigenic *Escherichia coli*
FDR	false discovery rate
fg	femtogram
Gb3	globotrianosylceramide receptor
h	hour
HUS	hemolytic-uremic syndrom
IPTG	Isopropyl-β-D-thiogalactopyranosid
Kb	kilobase
LB	lysogeny broth
LDF	linear discriminant function
LEE	locus of enterocyte effacement
lncRNA	long non-coding ribonucleic acid
M	molar
Mb	megabase
min	minute
ml	milliliter
mM	millimolar

mRNA	messenger ribonucleic acid
ncRNA	non-coding ribonucleic acid
NGS	next generation sequencing
nm	nanometer
OD	optical density
OLG	overlapping gene
ORF	open reading frame
PBS	phosphate buffered saline
PCR	polymerase chain reaction
qPCR	quantitative polymerase chain reaction
RACE	rapid amplification of cDNA ends
RIBOseq	translatome sequencing, ribosomal footprinting
RNA	ribonucleic acid
RNAseq	transcriptome sequencing
RPKM	reads per kilobase per million mapped reads
rpm	rounds per minute
rRNA	ribosomal ribonucleic acid
RT	reverse transcription
s	second
SD	Shine-Dalgarno
sRNA	small ribonucleic acid
tRNA	transfer ribonucleic acid
TSS	transcriptional start site
UTR	untranslated region

List of figures

Figure 1: The DNA double strand and the six possible reading frames 6

Figure 2: Possible orientations of overlapping genes 8

Figure 3: Workflow of the RNAseq and RIBOseq experiment 18

Figure 4: Workflow of the competitive growth experiment 21

Figure 5: Reproducibility of technical replicates 23

Figure 6: Reproducibility of biological RNAseq and RIBOseq replicates 24

Figure 7: Artemis view of the RIBOseq reads for the condition LB at 37°C mapped to the annotated genes ECs0036 and ECs0037 of the two biological replicates separately 25

Figure 8: Correlation of transcriptome to translatome data 26

Figure 9: Correlation of translatome to proteome data 27

Figure 10: Differentially regulated genes in LB at 37°C compared to BHI at 37°C 38

Figure 11: Venn-Diagram showing the number of translated OLGs for the three growth conditions 74

Figure 12: Correlation of the two RNAseq and RIBOseq biological replicates for the 380 putative OLGs at the condition LB at 37°C 75

Figure 13: Correlation of the RNAseq data to the RIBOseq data for the 380 translated OLGs 76

Figure 14: Size distribution of the 380 translated OLGs compared to the 250 shortest annotated genes of EHEC Sakai 77

Figure 15: RCV distribution of the 380 translated OLGs compared to 250 short annotated EHEC genes 78

Figure 16: Differentially regulated OLGs 80

Figure 17: Translation of ECs0016 and OLGECs0016 visualized in the Artemis genome browser 163

Figure 18: DNA sequence of OLGECs0016 including its upstream and downstream region 165

Figure 19: Expression of OLGECs0016-EGFP fusion proteins 166

Figure 20: Translation of ECs4930 and OLGECs4930 visualized in the Artemis genome browser 168
Figure 21: DNA sequence of OLGECs4930 including its upstream and downstream region 169
Figure 22: Promoter activity of the sequence upstream of the first putative start codon of OLGECs4930 170
Figure 23: Translation of ECs0535 and OLGECs0535 I-III visualized in the Artemis genome browser 174
Figure 24: Organization of the tri-cistronic operon OLGECs0535 I-III 175
Figure 25: Expression of EGFP-OLGECs0535 II and EGFP-OLGECs0535 III fusion proteins 177
Figure 26: Transcription of OLGZ5561 and *rpoC* visualized in the Artemis genome browser 179
Figure 27: DNA sequence of OLGZ5561 including its upstream and downstream region 180
Figure 28: Competitive growth of EHEC wild type against ∆OLGZ5561 182
Figure 29: Complementation of EHEC ∆OLGZ5561 *in trans* 184
Figure 30: Transcription of OLGZ5561 and ∆OLGZ5561 determined by qRT-PCR 185

List of tables

Table 1: Prediction of σ^{70} promoters for the translated OLGs and for the short annotated genes 81

Table 2: Prediction of ρ-independent terminators for the translated OLGs and for the short annotated genes 82

Table 3: Prediction of Shine-Dalgarno sequences for the translated OLGs and for the short annotated genes 83

Table 4: Expression of OLGECs0016 and the annotated gene ECs0016 164

Table 5: Oligonucleotides used for the characterization of the OLGECs0016 promoter region 165

Table 6: Oligonucleotides used for cloning of an OLGECs0016-EGFP fusion protein 166

Table 7: Expression of OLGECs4930 and the annotated mother gene ECs4930 168

Table 8: Oligonucleotides used for cloning of pProbe-OLGECs4930 170

Table 9: Oligonucleotides used for cloning of the strand-specific knock-out mutant ΔOLGECs4930 171

Table 10: Expression of OLGECs0535 I, OLGECs0535 II, OLGECs0535 III, and the annotated gene ECs0535 173

Table 11: Oligonucleotides used for the characterization of the OLGECs0535 I-III promoter region 176

Table 12: Oligonucleotides used for cloning of pEGFP-OLGECs0535 II and pEGFP-OLGECs0535 III 178

Table 13: Oligonucleotides used for 5’ RACE 181

Table 14: Oligonucleotides used for cloning of the pBAD complementation plasmids 183

Table 15: Oligonucleotides used for qRT-PCR of OLGZ5561 185

Supplementary Table S2: Genome location, expression level, translatability and ORF coverage of the novel OLGs 204

Supplementary Table S3: Properties of the novel OLGs 214

Supplementary Table S4: Differentially regulated OLGs in BHI stress compared to BHI control 225

Supplementary Table S5: Differentially regulated OLGs in LB compared to BHI control 227

Content of attached CD

Supplementary Table S1: Differential expressed annotated genes in LB at 37°C and OD_{600} = 0.4 compared to BHI at 37°C and OD_{600} = 0.1

Supplementary Data Publication I:

Supplementary Table S1: Read numbers for the transcriptome and translatome libraries

Supplementary Figure S1: Reproducibility of RNAseq and RIBOseq

Supplementary Data Publication II:

S1 Table: Summary of NGS results

S2 Table: RNAseq and RIBOseq results of three different growth conditions for the 465 novel genes and the 250 short annotated genes

S3 Table: Properties of the novel genes

S4 Table: Properties of the 250 short annotated genes

S5 Table: Conservation of the novel genes

S6 Table: Significant transcriptional and translational regulation in LB compared to BHI control of the novel genes and the short annotated genes

S7 Table: Transcriptional and translational regulation at BHI COS compared to BHI control of the novel genes and the short annotated genes

S8 Table: Summary of the Predict Protein results for the putative proteins encoded by the novel genes

S9 Table: Summary of the Predict Protein results for the short annotated genes

S10 Table: Classification into 'real' and 'pseudo' proteins by the machine-learning algorithm

S1 Figure: Distribution of RCV for the short annotated genes, novel genes with and without annotated homologs

S2 Figure: Conservation of intergenic sequences

S1 File: Custom script used for reading frame determination in the sum signal of annotated genes, novel genes with and without homologs

S2 File: Custom script used for detecting sequence conservation

S3 File: Custom script used for extracting intergenic sequences

Supplementary Data Publication III:

Additional file S1: Bacterial strains and plasmids used in this study

Additional file S2: Oligonucleotides used in this study

Additional file S3: Ratio in percent of EHEC wild type to EHEC Δ*laoB* after competitive growth at different growth conditions

Additional file S4: Phylogenetic analysis of ECs5115 by the Maximum Likelihood Method

Additional file S5: Phylogenetic analysis of *laoB* at the DNA level

Supplementary Data Publication IV:

Supplementary Table S1: Bacterial strains and plasmids used in this study

Supplementary Table S2: Oligonucleotides used in this study

Supplementary Figure S1: Competitive growth of EHEC wild type against EHEC Δ*anoG* at additional anaerobic conditions

Supplementary Figure S2: Phylogenetic tree of ECs2385

Supplementary Figure S3: Phylogenetic tree of ECs2384

Supplementary Data Publication V:

Supplementary Table S1: Bacterial strains and plasmids used in this study

Supplementary Table S2: Oligonucleotides used in this study

Supplementary Figure S1: Phylostratigraphic distribution of *slyA*

Supplementary Figure S2: Phylostratigraphic distribution of *slyB*

Supplementary Figure S3: Promoter activity of the region upstream of the minor TSS

Supplementary Figure S4: Ratio in percent of EHEC wild type over EHEC Δ*slyC* after competitive growth at different growth conditions

1. Introduction

1.1 Model organism *Escherichia coli* O157:H7 strain Sakai

Escherichia coli is a Gram-negative, rod-shaped, oxidase negative bacterium. The *E. coli* species is very diverse containing many commensals, which colonize the intestine of humans and other hosts. Since decades the strain K-12 is the laboratory workhorse (Croxen et al., 2013). Also, the metabolic capacity of *E. coli* strains differs: only 965 genes take part in forming the core metabolome, whereas 1,460 genes are able to form the pan metabolome. The ability to utilize different carbon or nitrogen sources can be used to distinguish between strains (Monk et al., 2013). The species *E. coli* contains a subset of pathogenic strains, which can cause enteric diseases or urinary tract infections. The enteric pathogenic strains are divided in five groups: EPEC (enteropathogenic *E. coli*), EIEC (enteroinvasive *E. coli*), EAEC (enteroaggregative *E. coli*), ETEC (enterotoxigenic *E. coli*) and STEC (Shiga-toxin producing *E. coli*). EHEC (enterohemorrhagic *E. coli*) is a subgroup of STEC, which are causing most severe illness of pathogenic *E. coli* strains (Croxen et al., 2013). After consumption of contaminated food, EHEC colonizes the colonic epithelium (Lewis et al., 2015), which may lead to diarrhea and the disease can progress to hemorrhagic colitis. In some cases, the disease gets systemic and infected persons develop the life-threatening hemolytic-uremic syndrome (HUS) or meningitis (Lim et al., 2010b).

In this study, the EHEC strain O157:H7 Sakai was used. O157:H7 defines the serotype: this strain possesses the somatic antigen variation (O) 157 and the flagellum antigen (H) 7. Overall, over 400 serotypes are described (Croxen et al., 2013), whereat O157:H7 is most frequently isolated from patients. Sakai is a city in Japan, where this strain was isolated during an EHEC outbreak in 1996, with 6,000 infected people due to consumption of contaminated radish sprouts (Hayashi et al., 2001). The genome of EHEC Sakai was sequenced in 2001. It has a GC-content of 50.5% and a size of 5.5 Mb encoding about 5,200 genes. Additionally, EHEC Sakai has two plasmids: the F-like plasmid pO157 (92 kb) harboring 100 protein-coding genes (Burland et al., 1998) and the cryptic plasmid pOSAK1 (3.3 kb) with only four genes (Hayashi et al., 2001). Due to many genomic regions acquired by integration of phages (in total 24 prophages) and horizontal gene transfer (O-islands), the EHEC Sakai genome is 20% larger than the K-12 genome (Sadiq et al., 2014). The integration sites of the prophages are homologous and allow recombination leading to frequent large scale genomic inversions (Iguchi et al., 2006).

1.1.1 Pathogenicity

The majority of EHEC outbreaks are linked to contaminated food, especially bovine products, like undercooked beef, milk, and dairy products, which caused 75% of outbreaks, where the infection source could be determined (Nguyen and Sperandio, 2012). But also person-to-person contagion is possible (Lim et al., 2010b). Because EHEC has a high acid tolerance, facilitating survival during stomach passage, a very low infection dose of only 100 cfu is enough to cause disease (Reiland et al., 2014). Usually, EHEC binds to the epithelial cells of the intestine, but it is not internalized, however, it can be taken up by M-cells of the Payers Patches and it even survives inside macrophages (Etienne-Mesmin et al., 2011). In total, EHEC contains 400 virulence associated genes, but here only the three major virulence factors are introduced. First, EHEC can produce Shiga-toxin(s). There are two types of Shiga-toxin, stx1 and stx2, respectively. Stx2 is 1000-fold more toxic and more often associated with HUS (Muniesa et al., 2012). Both types are encoded on prophages. Shiga-toxin is an AB_5 toxin: after secretion, it binds to the globotrianosylceramide receptor (Gb3) of vascular and renal endothelium cells. The B-subunits form a pore, whereas the A-subunit enters the mammalian cell and inhibits translation by destroying the 28S rRNA, which leads to apoptosis (Croxen et al., 2013; Etienne-Mesmin et al., 2011). Second, EHEC possesses the locus of enterocyte effacement (LEE), which is responsible for the typical attaching and effacing phenotype. On a DNA stretch of 35 kb, 41 LEE genes are encoded (Reiland et al., 2014). After adhesion to the host cells, a type-III-secretion system is expressed, which translocates the effector protein Tir. The latter is integrated into the host cell membrane and binds to intimin on the EHEC surface. This induces actin polymerization, pedestal formation, and ensures a close contact of EHEC to the colonic epithelial cells (Battle et al., 2014). The third virulence factor is the large virulence plasmid pO157. It encodes a hemolysin and is important for biofilm formation (Lim et al., 2010a).

Besides the low infection dose, EHEC is an important public health concern, because no targeted therapy is available. The enteric manifestation is self-limiting, but about 4% of patients develop systemic HUS, which has a mortality rate of 0.5-20% (Croxen et al., 2013). EHEC O157:H7 is responsible for 100,000 infections per year in the USA (Eppinger and Cebula, 2015). In 2015, the Robert Koch Institut registered 1,604 EHEC infections and 69 HUS cases in Germany leading in three cases to a fatal outcome (Gilsdorf, 2016). A major EHEC outbreak occurred in Germany in 2011: 3,842 people got infected, 855 HUS cases were reported and 54 people died. The fatality rate was higher than for previous outbreaks. The causative agent could be identified as contaminated fenugreek sprout seeds (Muniesa et al., 2012). The responsible *E. coli* strain was atypical, because it has the serotype O104:H4, which classifies it as an enteroaggregative *E. coli*, but this strain acquired the Shiga-toxin genes and an antibiotic resistance plasmid (Karch et al., 2012). Generally, treatment of EHEC with antibiotics is contra indicated, because then the phages, carrying the Shiga-toxin genes, enter the lytic cycle and even more toxin will be produced (Wong et al., 2000). Monoclonal antibodies against Shiga-toxin, Gb3 receptor analogs, probiotics, and vaccination are under investigation (Croxen et al., 2013). Additionally, research on EHEC is complicated by the fact that no animal model is available, which mirrors all parts of human infection. Gnotobiotic pigs and germ-free or streptomycin-treated mice are most frequently used (Mohawk and O'Brien, 2011). To prevent disease, it is important to get a better understanding of the biology of this enteric pathogen.

1.1.2 Reservoirs

The major reservoir of EHEC is believed to be cattle and other ruminants (Lim et al., 2010b). In contrast to humans, they do not express vascular Gb3 receptors leading to asymptomatic colonization (Nguyen and Sperandio, 2012). About one half of the cattle population is shedding EHEC in their stool at any time and especially super-shedders contribute to a massive environmental spread of bacteria (Stein and Katz, 2017). Also, EHEC colonizes the gastrointestinal tract of other mammals and birds (Persad and

LeJeune, 2014). Worldwide, 10^{20} cfu EHEC are shed to the environment every day (Karch et al., 2012). Insects and snails can serve as vectors carrying shed EHEC from animal dung to other hosts, causing a cycling between different habitats (Semenov et al., 2010; Wasala et al., 2013). Green-leave plants are another important reservoir of EHEC: not only the plant surface is colonized, but growth of EHEC was also verified inside stomata (Saldana et al., 2011), and EHEC can even internalize into roots and seedlings (Hou et al., 2013; Jayaraman et al., 2014). Additionally, EHEC tolerates many stress conditions like low temperature, high osmotic pressure (Hücker et al., 2017b), and low nutrient availability facilitating survival in the environment without any host. In soil, EHEC survives 30-110 days (Ma et al., 2011) and persistence in cold water renders EHEC even more resistant against antibiotics (Duffitt et al., 2011). Furthermore, EHEC can colonize protozoan hosts such as *Acantamoeba polyphaga* (Barker et al., 1999).

1.2 Short overlooked genes

Today, thousands of bacterial genomes have been sequenced. For genome annotation, usually bioinformatics pipelines like GLIMMER are used (Delcher et al., 2007). They identify open reading frames (ORFs) of a certain size and investigate several parameters of every ORF for prediction, i.e., if this ORF may represent a protein-coding gene or not. Usually, annotation algorithms search for homologs in other bacteria, known domains of the potential protein, presence of a Shine-Dalgarno sequence in the upstream region, canonical start codons, and low codon bias. The smaller the ORF gets, the higher is the possibility of false positive annotations. Therefore, many annotation algorithms use an arbitrary size threshold of 50-100 amino acids (AA). However, smaller proteins exist and are functional (Baek et al., 2017; Landry et al., 2015; Neuhaus et al., 2017). Therefore, this genome annotation practice led to a systematic omission of small proteins (Boekhorst et al., 2011; Storz et al., 2014; Warren et al., 2010). Additionally, small proteins are difficult to investigate by molecular biology techniques such as Western blot, because they are easily lost during protein purification, run off the SDS gel, or are blotted through the membrane. Global proteome studies also often miss small proteins, since they do not lead to enough tryptic peptides of a detectable size (Landry et al., 2015; Slavoff et al., 2013). Recently, small proteins got more into focus, but knowledge about their functions is still very limited (Storz et al., 2014). They seem to contain a high proportion of membrane or membrane-associated proteins (Kemp and Cymer, 2014). For example, Hemm et al. (2008) showed the expression of 18 small *E. coli* proteins, whereof nine possess a transmembrane helix. In addition, small proteins have properties different to regular proteins: their knock-out is not lethal and often no phenotype is detectable at all; they are more hydrophobic, contain more α-helices, and use non-ATG start codons (Brylinski, 2013; Storz et al., 2014).

A high proportion of the small proteins belongs to the 'dark' proteome. This means their molecular conformation is completely unknown, because neither they fit to any structural family of the protein universe, nor they show similarity to any PDB structure (Levitt, 2009; Perdigão et al., 2015). In prokaryotes, 5% of all proteins are completely 'dark' and additional 8% contain dark regions (Perdigão et al., 2015). These proteins seem to show several unusual features compared to known proteins: they are more often secreted, contain more disulfide bonds, are shorter, have a lower number of protein-protein interactions, lower evolutionary reuse, more hydrophobic AA topology, and higher folding energy (Bitard-Feildel and Callebaut, 2017). Regarding disordered regions and transmembrane helices, conflicting results have been published. Whereas Perdigão et al. (2015) argue that dark proteins are less disordered than known proteins, Bitard-Feildel and Callebaut (2017) claim that they are more disordered. Additionally, they report a higher number of transmembrane helices, which is also disagreed by Perdigão et al. (2015).

1.3 Antiparallel overlapping genes

1.3.1 Definition

Three DNA nucleotides, called codon, encode one AA. The DNA consists of the four different bases adenine, guanine, cytosine, and thymidine. Theoretically, it would be possible to encode 64 different AAs using unique codons. However, only 20 proteinogenic AAs exist and most AAs are encoded by more than one codon. This leads to functional redundancy of the genetic code. Additionally, the DNA triplet code offers the possibility that two or more protein-coding genes are encoded at the same DNA locus (Figure 1). There are three possible reading frames on the sense strand (+1, +2,

+3) and three additional reading frames on the antisense strand (-1, -2, -3). An overlapping gene (OLG) is defined such that at least one nucleotide of the coding region of a gene overlaps with the coding region of another gene (Simon et al., 2011). Sometimes also overlaps of promoter regions or alternative splicing in eukaryotes is called OLG (Normark et al., 1983), but in this work the term is only used for overlaps of coding regions. The frame of the annotated mother gene is defined as +1 and the frame of the OLG is determined relative to the annotated gene's frame.

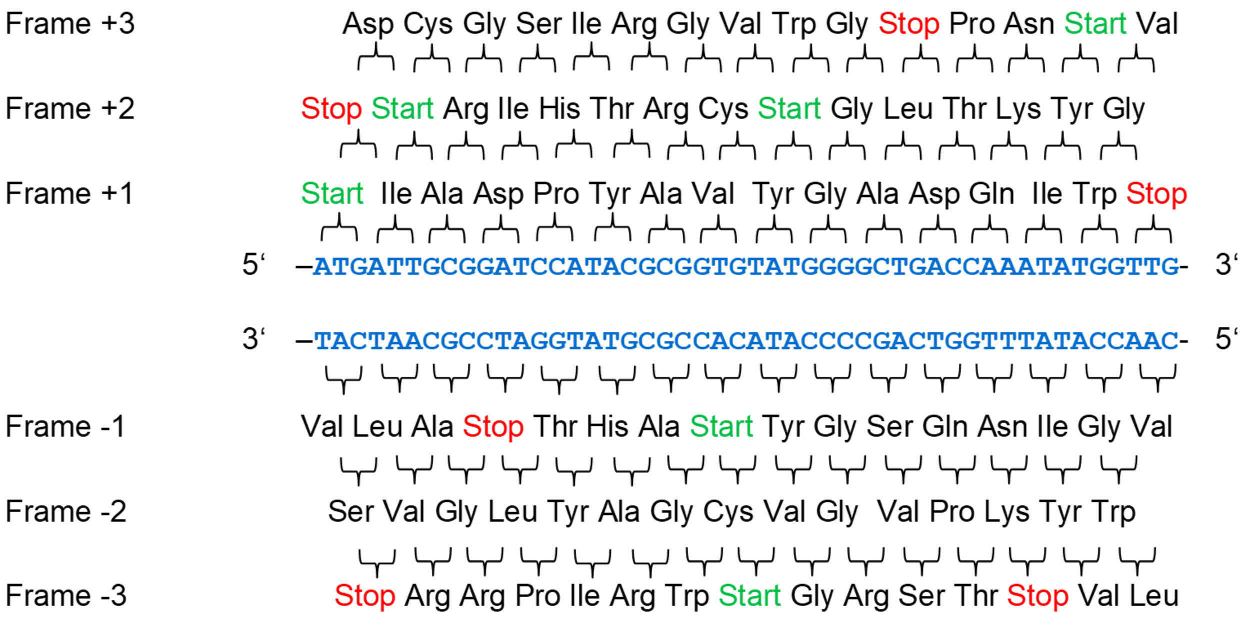

Frame +3
Asp Cys Gly Ser Ile Arg Gly Val Trp Gly Stop Pro Asn Start Val
Frame +2
Stop Start Arg Ile His Thr Arg Cys Start Gly Leu Thr Lys Tyr Gly
Frame +1
Start Ile Ala Asp Pro Tyr Ala Val Tyr Gly Ala Asp Gln Ile Trp Stop
5‘ –ATGATTGCGGATCCATACGCGGTGTATGGGGCTGACCAAATATGGTTG- 3‘
3‘ –TACTAACGCCTAGGTATGCGCCACATACCCCGACTGGTTTATACCAAC- 5‘
Frame -1
Val Leu Ala Stop Thr His Ala Start Tyr Gly Ser Gln Asn Ile Gly Val
Frame -2
Ser Val Gly Leu Tyr Ala Gly Cys Val Gly Val Pro Lys Tyr Trp
Frame -3
Stop Arg Arg Pro Ile Arg Trp Start Gly Arg Ser Thr Stop Val Leu

Figure 1: The DNA double strand and the six possible reading frames. An example section of a DNA double strand is shown in blue. The DNA sequence is translated into the corresponding AAs for every reading frame. Potential start codons are highlighted in green and stop codons in red. Reading frame +1 encodes a protein-coding gene and also the overlapping reading frames contain ORFs, which might encode proteins.

Diverse types of OLGs exist: first, trivial overlaps must be distinguished from non-trivial overlaps. In trivial overlaps, the overlapping region is smaller than 90 bp and, in many cases, the overlap has a size of only 1-4 bp. This type of overlap is very frequent in bacteria, because many genes are organized in operons (Johnson and Chisholm, 2004) and transcribed as a single polycistronic mRNA. Trivial same-strand overlaps allow translational coupling, meaning that the ribosomes do not dissociate from the mRNA after translation of the first gene, but stay on the mRNA and continue with the translation of the next gene, which does not require its own ribosome binding site (Rex et al., 1994).

In *E. coli,* 50% of all genes are overlapping, but the large majority are these trivial overlaps (Merino et al., 1994). Here, only non-trivial OLGs ≥ 90 bp are of interest. In viruses, many non-trivial OLGs are described and 38% of viral proteins are encoded overlapping (Rancurel et al., 2009). It is hypothesized that viral OLGs have originated due to the limited space inside the capsid, which would favor a small genome (Chirico et al., 2010). In contrast, in bacteria only a handful of non-trivial OLGs have been reported in literature. The main reason for this is that genome annotation algorithms do not allow for longer overlaps and only the gene with the better score will become annotated (Delcher et al., 2007). Genome size reduction seems not to be the driving force for the evolution of non-trivial OLGs in bacteria (Johnson and Chisholm, 2004; Lillo and Krakauer, 2007), whereas genome streamlining, i.e., translational coupling through trivial overlaps, occurs frequently in thermophilic bacteria (Sabath et al., 2013; Saha et al., 2015).

OLGs occur in different organizations (Figure 2). The overlap can lie on the sense DNA strand (same-strand overlapping gene) or on the antisense strand (antiparallel overlapping gene). This study focuses on antiparallel OLGs, because they are easier to detect in experiments. It is possible that the overlapping reading frame is completely embedded in the mother gene, or head-to-head/divergent, and tail-to-tail/convergent OLGs are feasible. The organization tail-to-tail occurs more often, because for head-to-head OLGs also the regulatory elements upstream of the start codon overlap with the coding sequence (Fonseca et al., 2014; Huvet and Stumpf, 2014).

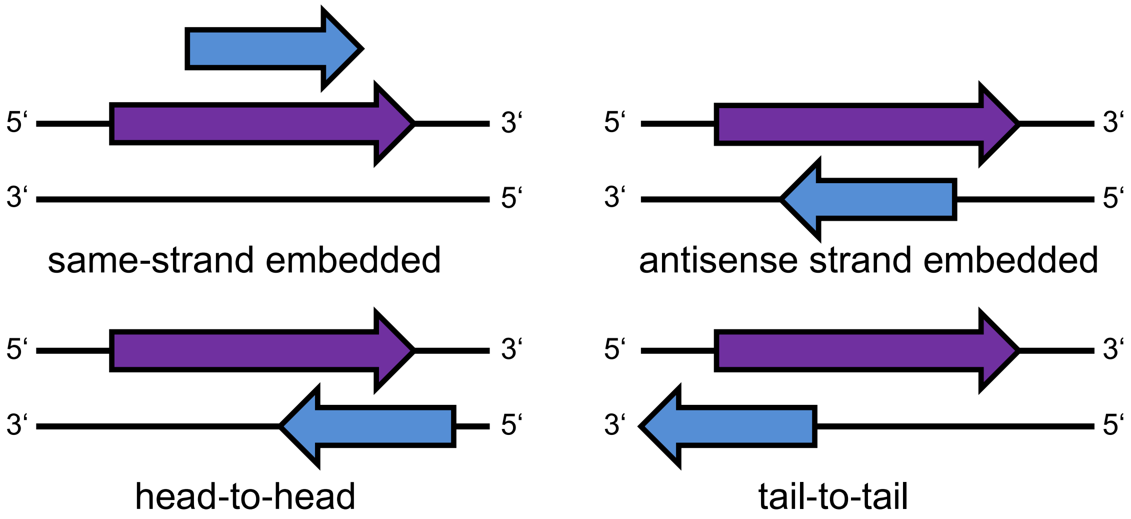


Figure 2: Possible orientations of overlapping genes. The mother gene is colored in purple and the novel overlapping gene in blue. The overlap can occur on the same strand or on the antisense strand. Furthermore, the 3' ends (head-to-head) or the 5' ends (tail-to-tail) of the genes can overlap. Alternatively, the sequence of one gene can be completely embedded into the other gene.

1.3.2 Origin and evolution of OLGs

The established hypothesis for the creation of novel genes is based on the assumption that an existing gene is duplicated and evolves over time to a new function by neofunctionalization, or that the encoded ancestor protein had more than one function and the copies will specialize on one function by subfunctionalization (Betran, 2015). Another reason, why OLGs have been ignored or even denied in the past, is that they must originate *de novo* (Keese and Gibbs, 1992). In addition, the evolutionary constraints on an overlapping gene pair are higher, because a mutation often influences both reading frames (Lèbre and Gascuel, 2017). However, the region antisense to annotated genes contains less stop codons than statistically expected (Mir et al., 2012). For frame -1 long ORFs are somewhat expected, because the third codon position of genes is enriched in GC, but all stop codons start with T. Thus, the specific codon usage of the annotated gene might cause long antisense ORFs as a byproduct (Boldogköirid, 2000; Veloso et al., 2005). Also, the reading frames -2 and -3 show less stop codons than expected statistically (Mir et al., 2012). Generally, organism with high GC-content are expected to contain more long OLGs (Merino et al., 1994).

Same-strand OLGs can be created by programed ribosomal frameshifting (Caliskan et al., 2014) or programed transcriptional realignment (Sharma et al., 2011). OLGs likely arise *de novo* by overprinting (Grassé, 1977). A point mutation leads to loss of a stop codon, what creates a novel ORF or a mutation can cause a novel start codon (Delaye et al., 2008; Keese and Gibbs, 1992). When an upstream promoter is present by chance, the ORF can be transcribed (Neme and Tautz, 2013). Next, this RNA can be used as a template for translation, forming a novel mRNA. If the novel peptide has no beneficial function or even is detrimental, the novel ORF will get lost again (Huvet and Stumpf, 2014). However, in some cases the novel peptide will indeed impart an advantage. Alternatively, the ORF can become fixed by neutral evolution (Lillo and Krakauer, 2007). Over time, the ORF will evolve towards a gene with a fully-fledged function, e.g., acquiring regulatory elements, the ORF will become longer, transcription and translation rates will increase. Maybe the overlapping gene pair will become decoupled by a duplication. This scenario fits well the proto-gene hypothesis published by Carvunis et al. (2012), discussing *de novo* gene birth in the intergenic regions of *S. cerevisiae*. The authors postulate a continuum from non-coding DNA over proto-genes to established genes. Presumably, OLGs represent evolutionary young genes and therefore, they will be taxonomically restricted or are even ORFans. ORFans do not have any homologs in closely related bacteria. They are shorter, have higher codon bias, and contain less domains (Neme and Tautz, 2013). In EHEC, ORFan genes have an AA composition more comparable to non-coding sequence than to established proteins (Yomtovian et al., 2010). Phylostratigraphic analysis of an overlapping gene pair can be used to identify the mother gene (i.e., preceding gene) and the novel gene (i.e., overlapping gene) originated by overprinting: the mother gene is expected to show a broader phylogenetic distribution (Pavesi et al., 2013).

Contradicting results have been published regarding the preferred reading frame for OLGs. The genetic code shows the highest degrees of freedom at position 3, position 1 is slightly degenerated, whereas position 2 is completely restricted. Therefore, an OLG in frame -3 would have the greatest degree of evolutionary independence: a point mutation in one frame would lead to a synonymous codon in the overlapping frame in many cases, which opens the possibility of sequence evolution. On the other hand, in frame -2 almost every mutation will change the AA sequence of the mother gene, thus, this high information cost is believed to render frame -2 unlikely (Krakauer, 2000). However, frame -2 contains longer ORFs than frame -3 (Mir et al., 2012) and frame -2 is conserved by the mother gene (Mir and Schober, 2014). The upper studies investigated the constraints on DNA level. If the constrains of OLGs at the AA level are considered, frame -3 shows the highest number of 'forbidden' dipeptides due to a stop codon in the mother frame and should be rare (Lèbre and Gascuel, 2017).

1.3.3 OLGs in bacteria

Until now, only about 70 OLGs are suspected in literature in prokaryotes. Most are just predicted bioinformatically (Jensen et al., 2006), for others only transcription was detected using In Vivo Expression Technology (Silby and Levy, 2004; Silby et al., 2004). Only a handful functionally characterized OLGs have been reported, and the precise protein function is often still not known. Some represent a toxin-antitoxin system like *pic/setAB* of *Shigella flexneri* (Behrens et al., 2002), the OLG pair *aatS/aatC* of ETEC was discovered because of an intragenic transcription factor binding site (Haycocks and Grainger, 2016), or *tpnA/astA* of *E. coli* (Sousa, 2003), and *tniA/ardD* in *Xanthomonas* are encoded on transposons (Balabanov et al., 2012), and yet for several other OLGs peptides mapping to potential proteins in antisense were detected in mass spectrometry in *Pseudomonas fluorescence* (Kim et al., 2009), *Shigella flexneri* (Zhao et al., 2011), and *Deinococcus desertii* (de Groot et al., 2014). *Streptomyces coelicolor* contains the OLG pair *dmdR1/adm.* The mother gene *dmdR1* is a regulator of iron metabolism. Strand-specific knock-out mutants were cloned and both showed a phenotype leading to overproduction of different antibiotics (Tunca et al., 2009).

1.3.4 Functionally characterized OLGs in EHEC

In the EHEC strain EDL933, which is very closely related to strain Sakai, two overlapping gene pairs have been functionally characterized. The first OLG pair is *htgA/yaaW*. *YaaW* represents the phylogenetically older gene, because it is found in diverse *γ-proteobacteria*, whereas *htgA* is restricted to the *Escherichia-Klebsiella* clade and it has originated by overprinting completely embedded into the *yaaW* coding region (Delaye et al., 2008; Fellner et al., 2014). The region upstream of the *htgA* start codon shows promoter activity. Y*aaW* is organized in an operon with the annotated gene *yaaI,* and the promoter is located upstream of *yaaI*. Translationally arrested mutants of both genes showed a phenotype in biofilm formation, and the metabolomic profiles were altered. Previously, HtgA was associated with heat shock, but the absence of a phenotype after an upshift of growth temperature contradicts this hypothesis. However, only YaaW could be overexpressed and detected by Western blot, whereas evidence for HtgA on protein level is still missing (Fellner et al., 2014).

The second characterized overlapping gene pair is *nog1/citC*. *CitC* represents the mother ORF, and encodes a citrate lyase ligase. Transcription of a completely embedded ORF in frame -2 was detected by RNAseq in LB medium and cow dung (Landstorfer et al., 2014). A strand-specific Δ*nog1* knock-out mutant showed a significant growth disadvantage in LB medium supplemented with $MgCl_2$ in competitive growth experiments against EHEC wild type. Accordingly, activity of the *nog1* promoter was increased under $MgCl_2$ supplementation compared to plain LB. In addition, the metabolome of the mutant was changed. Probably, *nog1* originated by overprinting as well, because the ORF is restricted to the *Escherichia-Shigella* clade, whereas *citC* is distributed over the *γ-proteobacteria* (Fellner et al., 2015).

Furthermore, Fellner (2015) cloned translationally arrested mutants of additional eleven OLG candidates and all of them showed a phenotype in competitive growth experiments at diverse stress conditions (e.g., menadione, NaCl, cycloheximide, $MgCl_2$, malonic acid, and $Cu(II)Cl_2$). However, those potential OLGs were not characterized further.

1.4 High-throughput discovery of novel genes facilitated by Next Generation sequencing technologies

1.4.1 Transcriptome sequencing (RNAseq)

In the past, differential gene expression of selected genes was studied using qRT-PCR (Pfaffl, 2001), and at the genome level microarrays were used (Duffitt et al., 2011; Kocharunchitt et al., 2012). Then, high-throughput sequencing technologies improved rapidly and the costs per base highly decreased (van Dijk et al., 2014). Today, the Illumina sequencing system dominates the market. In the first step of library preparation, adapters are ligated to the RNA sample. Their sequence is complementary to DNA sequences immobilized on the flow cell. Then, the sample is reverse transcribed and amplified. After quality control and concentration adjustment, the library is ready for sequencing. Next, the sample is loaded on a flow cell and the bound DNA is amplified by bridge amplification in solid phase resulting in clusters of about 1,000 identical molecules. Fluorescently labeled nucleotides are added sequentially to the clusters and the incorporated fluorescence dye is detected at every sequencing cycle for all clusters in parallel to read the sequence. A possible application of this next-generation sequencing (NGS) technology is the determination of complete transcriptomes (RNAseq), i.e., the whole RNA content of a bacterial culture at a certain time point is being read after reverse transcription to cDNA (van Dijk et al., 2014). Most cellular RNA is ribosomal RNA (He et al., 2010), but for gene expression analysis only mRNA is of interest. Therefore, rRNA depletion is advisable to increase the number of meaningful reads. RNAseq is very useful to study differential gene transcription between different growth conditions (Landstorfer et al., 2014). Improved RNAseq protocols even allow strand-specific mapping of reads (Flaherty et al., 2011; Perkins et al., 2009).

Surprisingly, RNAseq uncovered massive transcription outside of annotated protein-coding genes (Wade and Grainger, 2014). Also, transcription antisense to genes occurred frequently (Dornenburg et al., 2010). Unfortunately, with RNAseq data alone discrimination of protein-coding genes from ncRNA is impossible. Therefore, the RNAseq signals outside genes were interpreted as ncRNA or just pervasive transcription without any biological meaning (Lasa et al., 2011; Lin et al., 2013; Wade and Grainger, 2014). Indeed, ncRNA research was boosted by RNAseq due to the discovery and characterization of novel ncRNAs (Kröger et al., 2012; Raghavan et al., 2011). However, when RNAseq is combined with proteome analysis, novel genes can be identified. For instance, de Groot et al. (2014) detected five novel genes in *Deinococcus desertii*, of which two are even antiparallel overlapping to annotated genes.

1.4.2 Translatome sequencing (RIBOseq)

In 2009, (Ingolia et al.) reported a new NGS method, which determines the strand-specific translatome (RIBOseq). The idea of the method is to sequence only mRNA, which is used as template for translation. This is the case, if an actively translating ribosome binds to the mRNA and protects the incorporated stretch of mRNA against RNases. Thus, the first step is to obtain the cytosol containing the polysomes. A translational inhibitor can be added beforehand to prevent ribosome run off. Next, all mRNA not protected by ribosomes is digested using RNas(es), which do not destroy the integrity of the ribosome (Gerashchenko and Gladyshev, 2017; Miettinen and Bjorklund, 2015). The ribosomes are harvested by sucrose density gradient centrifugation. Then, the mRNA is isolated and DNA contamination is removed. After size selection – eukaryotic ribosomes protect an mRNA stretch of 28 bp and prokaryotic ribosomes protect 21 bp – rRNA needs to be depleted, and the sequencing library can be prepared. The RIBOseq protocol was originally published for yeast (Ingolia et al., 2009), but with a few adaptations the method was successfully applied for mammalian cell culture (Carlevaro-Fita et al., 2016), tissue (Fields et al., 2015), plants (Hsu et al., 2016), viruses (Stern-Ginossar et al., 2012), and bacteria (Li et al., 2012). Amongst others the results uncovered the importance of translational regulation (Wang et al., 2015; Zupanic et al., 2014) and allowed a deeper understanding of the translation process, e.g., initiation (Gao et al., 2015), translational start-site choice (Nakahigashi et al., 2016), elongation (Subramaniam et al., 2014), termination (Baggett et al., 2017), and ribosome conformation (Lareau et al., 2014; O'Connor et al., 2013).

1.4.3 Discovery of novel protein-coding genes using combined RNAseq and RIBOseq

In agreement to RNAseq, RIBOseq also showed many signals outside and antisense of annotated genes. In eukaryotes, particularly short ORFs upstream of annotated genes are clearly translated (Bazzini et al., 2014; Fields et al., 2015; Fritsch et al., 2012). In addition, translation of previously annotated ncRNA was observed frequently (Carlevaro-Fita et al., 2016; Ji et al., 2015; Landry et al., 2015; Ruiz-Orera et al., 2014). Eukaryotic RIBOseq data shows the triplet code caused by the codon-wise progression of the translating ribosome on the mRNA (Ingolia et al., 2009). For determination of the triplet periodicity, it is counted, how many RIBOseq reads have their 3’ or 5’ end on a given codon position. In contrast, false positive RIBOseq signals caused by co-purified RNA binding proteins or inactive ribosomes are not expected to reflect triplet periodicity. Usually, only a particular read length is investigated, whereupon some lengths show a better reading frame signal than others (Legendre et al., 2015). This is not only the case for the sum signal of annotated genes, but also translated single genes show a clear reading frame (Aspden et al., 2014; Calviello et al., 2016; Smith et al., 2014). For example, Fields et al. (2015) conducted RIBOseq of mouse dendritic cells and the authors detected translation of 317 intergenic ORFs, 1,379 ORFs upstream, 22 ORFs downstream, and 264 ORFs antiparallel overlapping to annotated genes, respectively. Interestingly, translation of many of these non-annotated ORFs is conserved in human fibroblasts. Therefore, non-annotated ORFs covered by RIBOseq reads were analyzed for the presence of a reading frame, which is a convincing evidence for a protein-coding gene. However, in prokaryotes the situation is different. RIBOseq data only show a poor triplet periodicity, which is detectable only in the sum signal of annotated genes but not for single genes (Landstorfer, 2014; Li et al., 2012). Very recently, a new protocol was published, where the endonuclease RelE was added during unprotected mRNA digestion leading to a reading frame with comparable resolution to eukaryotes. Even the analysis of a frame shift event was possible (Hwang and Buskirk, 2017).

Although discrimination of ncRNAs from protein-coding genes by reading frame determination is difficult, there is another possibility to distinguish them: when RIBOseq is performed in combination with RNAseq, the translatability of every ORF can be calculated (Neuhaus et al., 2017). This ribosomal coverage value (RCV) is the ratio of the reads per kilobase per million sequenced reads (RPKM) for the translatome over the RPKM for the transcriptome. Transfer RNAs are not expected to be translated and have RCVs of 0.01-0.1 (Hücker et al., 2017a); thus, ncRNAs should have RCVs in a similar range. Translated protein-coding genes on the other hand, have RCVs considerably higher than 0.1. Neuhaus et al. (2017) discovered that RCVs for several ORFs annotated as ncRNAs in *Escherichia coli* O157:H7 EDL933 are in the same range as for annotated genes. For *ryhB*, the translation into the peptide RyhB was proven. Similarly, Jeong et al. (2016) reported translation of 31 predicted ncRNAs of *Streptomyces coelicolor*.

RNAseq and RIBOseq data of EHEC EDL933 mentioned above, which was obtained in LB medium at 37°C, was also investigated regarding translated intergenic ORFs and novel antiparallel OLGs. Indeed, 72 novel protein-coding genes in intergenic sequences were found, and seven proteins could be confirmed by mass spectrometry (Neuhaus et al., 2016). An even higher number of 242 antiparallel overlapping ORFs show evidence of translation (Landstorfer, 2014). Baggett et al. (2017) report 43 recoding events preferentially occurring at TGA stop codons (stop codon read through and translational frameshifts) in *E. coli* K-12. Expression of three of these downstream proteins was further confirmed by Western blot. Nakahigashi et al. (2016) performed RIBOseq of *E. coli* using the antibiotic tetracycline to stall ribosomes at the translation initiation sites. They report 328 non-annotated initiation sites in the intergenic regions indicating translation of short ORFs. Additionally, Baek et al. (2017) discovered 130 small genes in *Salmonella* and confirmed 25 of them by Western blot. Therefore, combined RNAseq and RIBOseq is a powerful method to uncover short intergenic or antiparallel overlapping genes missed by genome annotation pipelines.

Another insight RIBOseq data provided, was that translation of non-annotated ORFs initiates frequently at rare start codons. In *E. coli,* ATG is the most common start codon followed by GTG and TTG (Kozak, 1983). An ATT start codon is reported only for the two genes *pcnB* (Binns and Masters, 2002), and *infC* (Liveris et al., 1993). Nakahigashi et al. (2016) detected three alternative translation initiation sites of *E. coli* with a CTG start codon. Additionally, the plasmid-borne gene *repA* is confirmed to use an CTG start codon (Spiers and Bergquist, 1992). However, testing the initiation efficiency of different start codons in yeast, using the gene of the alanyl-tRNA synthetase, indicates that even other codons are able to initiate translation (Chang et al., 2010). The codons CTG, ATT and ACG showed an initiation efficiency of 50% compared to ATG; ATA and ATC showed still an initiation efficiency of 20%, only the codons AAG and AGG were not able to initiate translation. Moreover, start codon context is important for non-ATG codons to stabilize imperfect codon-anticodon base pairing. Many intergenic ORFs with translation signals likely use a rare start codon (Hücker et al., 2017a; Neuhaus et al., 2016). Eukaryotic RIBOseq data confirms frequent usage of rare start codons, especially in case of the non-annotated, short ORFs upstream of annotated genes (Chu et al., 2015; Fields et al., 2015; Fritsch et al., 2012; Iwasaki and Ingolia, 2017; Ji et al., 2015). N-terminal proteomics also determines the start codon: according to the RIBOseq data, several N-terminally extended proteins were reported to use rare start codons (Van Damme et al., 2014; Willems et al., 2017). These near-cognate start codons are just a point mutation away from an optimal ATG codon.

1.5 Perspectives of this study

The aim of this study is the detection of novel genes with special emphasis on genes antiparallel overlapping to annotated genes in the food-born pathogen *Escherichia coli* O157:H7 Sakai. Combined RNAseq and RIBOseq were applied at three different growth conditions, whereupon two conditions reflect optimal growth and one condition severe cold and osmotic stress, to detect transcription and translation of non-annotated ORFs. Previous NGS studies of bacteria did not investigate OLGs or ignored signals outside of annotated genes completely. Next, bioinformatics analysis was used to confirm the protein-coding character of those expressed intergenic and antisense ORFs. In a second step, obviously translated OLGs were selected for functional characterization using the following methods: cloning of a translationally arrested mutant, search of a phenotype in competitive growth experiments using the EHEC wild type against the mutant, characterization of the promoter region, and showing translation into a protein by overexpression of an EGFP-fusion protein. The complete study was divided into the following working packages:

1) <u>Identification of interesting conditions for the NGS experiments.</u> A literature search was performed regarding proteome data of EHEC Sakai. A publication was found, which included proteome data for a condition, in which the expression of some genes is regulated by RNA thermometers or riboswitches (Giuliodori et al., 2010; Kocharunchitt et al., 2012; Kouse et al., 2013). Confirmation of known RNA thermometer/riboswitch gene regulation would represent a dedicated support for the method applied.

2) RNAseq and RIBOseq experiments. First, the RIBOseq protocol developed by Landstorfer (2014) was optimized. Unprotected mRNA was digested with a mixture of five RNases to avoid sequence specificity, which might improve the reading frame. Additionally, a more efficient rRNA depletion kit was introduced. Then, RNAseq and RIBOseq at three selected growth conditions were performed (workflow see Figure 3): LB medium at 37°C, harvest at OD_{600} 0.4; BHI medium at 37°C, harvest at OD_{600} 0.1, and BHI medium supplemented with 4% NaCl at 14°C, harvest at OD_{600} 0.1. The experiments were conducted in two biological replicates.

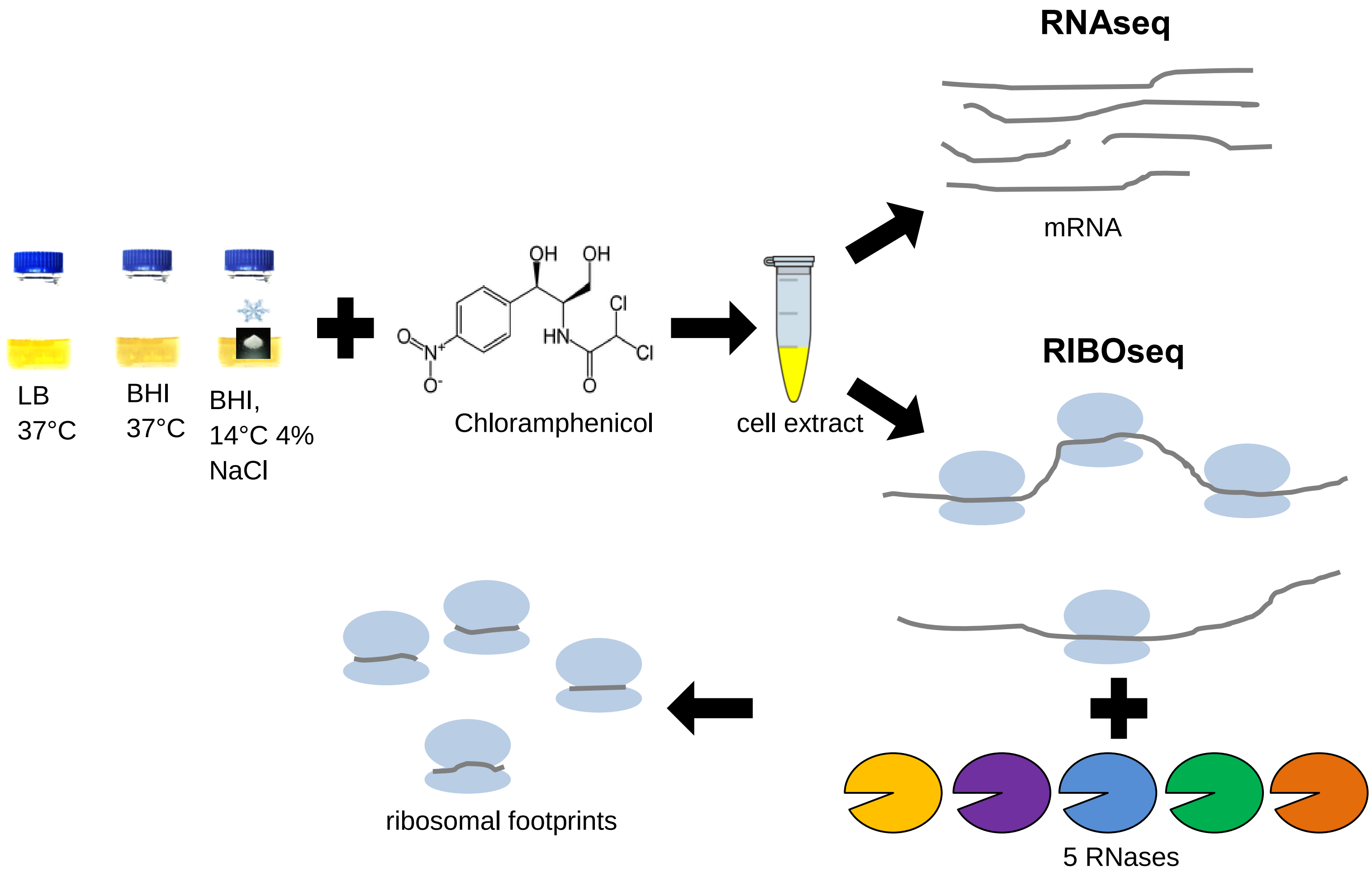


Figure 3: Workflow of the RNAseq and RIBOseq experiment. EHEC was grown at three different growth conditions and when the desired optical density was reached, translation was stalled using the antibiotic chloramphenicol. The cell extract was harvested and the sample was split. For RNAseq, total RNA was isolated. For RIBOseq, a mixture of five RNases was added to digest all mRNA not protected by ribosomes. Then, the ribosomes were harvested by sucrose density gradient centrifugation and ribosomal footprint mRNA was isolated.

3) Analysis of NGS data. The RNAseq and RIBOseq reads were mapped to the EHEC Sakai genome. RPKM values for the annotated genes and every ORF ≥ 93 bp were calculated in R (the R script was written by Svenja Simon, Chair for Data Analysis and Visualization, University of Konstanz). The ribosomal coverage value (RCV) was calculated by the ratio of RPKM value translatome over the RPKM value transcriptome. Differentially expressed genes on transcriptional and translational level were detected using the software *edgeR*.
4) Detection of novel intergenic genes. Novel gene candidates were selected regarding RPKM and RCV thresholds, and coverage with RIBOseq reads. Furthermore, annotated homologs were searched with BLASTP, non-annotated homologs with TBLASTN, and diverse properties of the novel gene candidates were compared to a selection of short annotated EHEC genes. Presence of a reading frame in the sum signal was investigated in cooperation with Gisle Vestergaard (Research Unit Comparative Microbiome Analysis, Helmholtz Zentrum München). Regulatory elements such as promoter, terminator, and Shine-Dalgarno sequence were predicted. The software PredictProtein was applied to predict properties of the potential novel proteins (in cooperation with Andrea Schafferhans, Department of Bioinformatics and Computational Biology, TU München), and based on PredictProtein data a machine-learning algorithm, trained with small EHEC proteins, was developed to distinguish novel proteins from scrambled sequences (in cooperation with Tatyana Goldberg and Michael Bernhofer, Department of Bioinformatics and Computational Biology, TU München). Moreover, the ratio of non-synonymous over synonymous substitutions was calculated, which determines if the sequence is under evolutionary pressure (in cooperation with Chase W. Nelson, American Museum of Natural History, New York).
5) Detection of novel antiparallel overlapping genes. Novel OLG candidates were identified analogous to the novel intergenic genes, and confirmed by visual inspection in Artemis. Additionally, differential expression between the three growth conditions, presence of annotated homologs, σ^{70} promoters, ρ-independent terminators, and Shine-Dalgarno sequences, compared to a group

of short annotated EHEC genes, were investigated. Moreover, the data was compared to previously published RIBOseq data of prokaryotes.

6) Functional characterization of selected OLGs. Six OLG candidates were selected. A new method for cloning of translationally arrested mutants was established (Kim et al., 2014). In competitive growth experiments, using diverse stress conditions, a phenotype was searched for (workflow see Figure 4). If a phenotype was detectable, a complementation with the intact ORF encoded on a plasmid was tried. The transcriptional start/stop sites were determined by 5'/3' RACE. Promoter activity of the upstream sequences was analyzed using a fluorescent dye. In addition, an OLG-EGFP fusion protein was overexpressed to prove translation into a protein. Using a phylostratigraphic analysis, the evolutionary age of the mother gene and the OLG was investigated (in cooperation with Sonja Vanderhaeghen, Chair for Microbial Ecology, TU München).

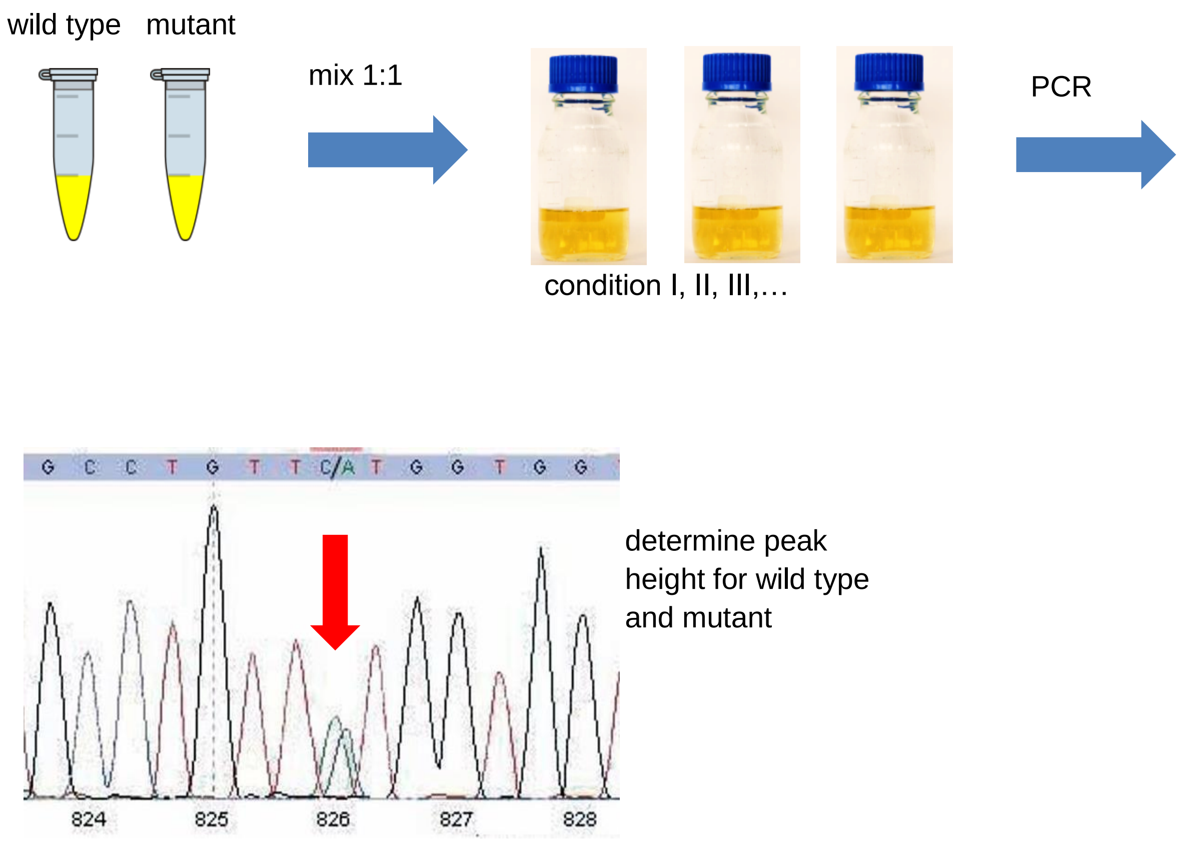

wild type
mutant
mix 1:1
condition I, II, III,…
PCR
G C C T G T T C/A T G G T G G
determine peak height for wild type and mutant
824
825
826
827
828

Figure 4: Workflow of the competitive growth experiment. Overnight cultures of EHEC wild type and translationally arrested mutant were mixed in equal ratio, and inoculated into 0.5 LB medium with different supplementations. After 18 h competitive growth, the genomic region containing the mutation(s) was amplified by PCR. The ratio of wild type over mutant was determined by comparing peak heights of the Sanger sequencing results, and absolute values were converted into percentage values.

2. Results and Discussion

Part I: Transcription and translation of annotated EHEC genes

2.1 Reproducibility of RNAseq and RIBOseq data

Global transcription and translation of *Escherichia coli* O157:H7 Sakai was investigated at three different growth conditions in two biological replicates: (1) LB medium at 37°C, OD_{600} = 0.4 (mid exponential growth phase), (2) BHI medium at 37°C, OD_{600} = 0.1 (early exponential growth phase), and (3) BHI medium at 14°C supplemented with 4% NaCl, OD_{600} = 0.1 (early exponential growth phase). The first two conditions represent optimal growth conditions, whereas the third is a severe stress condition of combined cold and osmotic stress (COS). Strand-specific RNAseq was performed as described in Landstorfer et al. (2014) for the Illumina system. For RIBOseq, the protocol of Ingolia et al. (2009), which was developed for yeast, was adapted for bacteria (Landstorfer, 2014), i.e., translation is stalled using the antibiotic chloramphenicol, and footprint size selection is performed at 22 ± 2 bp, because bacterial ribosomes incorporate a smaller stretch of mRNA. In this work, the protocol was further optimized regarding efficiency of rRNA depletion and precise digestion of all mRNA not protected by ribosomes using a mixture of five RNases. First, the quality of RNAseq and RIBOseq data was evaluated.

2.1.1 Technical replicates

For a technical replicate, the same RNAseq and RIBOseq library (workflow of library preparation see Figure 3) was sequenced twice on an Illumina HiSeq 2500 machine using different flow cells. Potential variation would be caused by the sequencing process itself.

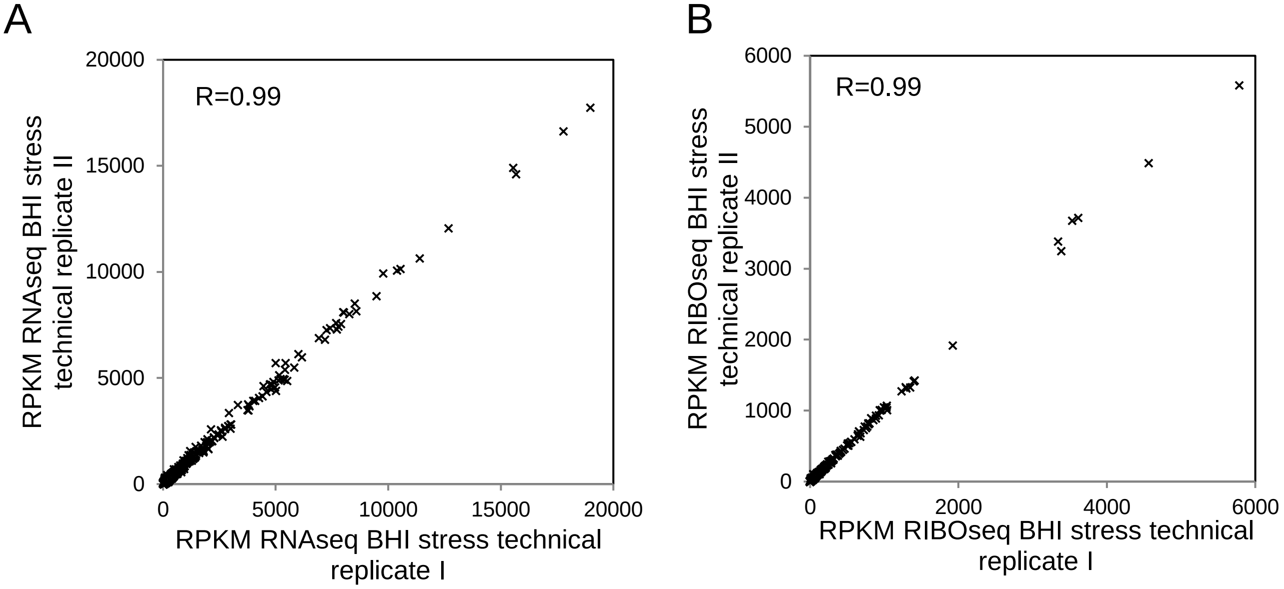


Figure 5: Reproducibility of technical replicates. The RPKM values of the annotated genes of the two technical replicates for the condition BHI COS were plotted against each other and Pearson correlation was calculated. **A** Reproducibility of technical RNAseq replicates. **B** Reproducibility of technical RIBOseq replicates.

Both, RNAseq and RIBOseq technical replicates of the condition BHI COS show an excellent reproducibility, demonstrated by a correlation of R=0.99 (Figure 5). The correlation between technical replicates is also very high for the other two investigated conditions (data not shown). This indicates that the sequencing process itself causes only negligible bias, even though batch effects for the sequencing reagents and different flow cells were reported previously (Buschmann et al., 2016).

2.1.2 Biological replicates

Next, the reproducibility of the biological replicates was investigated. At optimal growth conditions, the reproducibility of both RNAseq (R=0.99 in LB, and R=0.96 in BHI, respectively), and RIBOseq (R=0.96 in LB, and R=0.92 in BHI control) was very high (Figure 6). Accordingly, Ingolia et al. (2009) reported a Pearson correlation of R=0.98 of their yeast RIBOseq data, which was also obtained at an optimal growth condition. However, the correlation for the COS condition is lower with R=0.81 for the RNAseq experiment and R=0.79 for the RIBOseq experiment, respectively. This might be explained by an altered expression pattern with reduced overall gene expression at stress (see below). The overall decrease in expression diminishes the signal-to-noise ratio, hence, causing the lower correlation. Despite this fact, the reproducibility is still good enough for meaningful analysis.

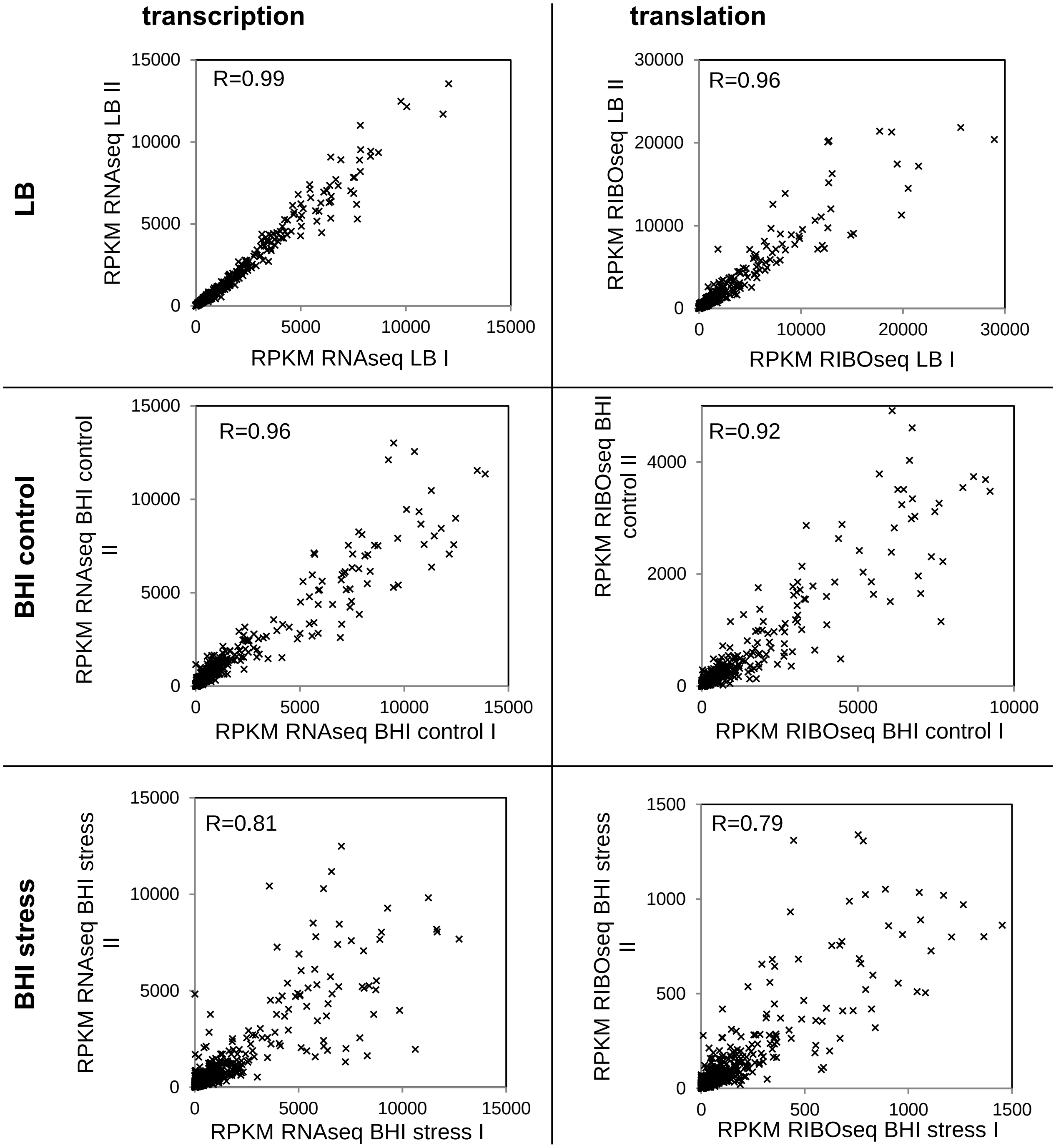


Figure 6: Reproducibility of biological RNAseq and RIBOseq replicates. The RPKM values of all annotated EHEC genes of the biological replicates were plotted against each other, and the Pearson correlation was calculated. The left column shows the results for the RNAseq experiments and the right column for the RIBOseq experiments. The upper panel shows the condition LB at 37°C, in the middle BHI at 37°C is depicted, and at the bottom BHI COS.

Usually, RNAseq and RIBOseq reads are not equally distributed over single protein-coding genes, but show an irregular pattern of many or few reads per locus. In contrast to reproducibility of RPKM values for individual genes, the reproducibility of read distribution on individual genes (comparable pattern of stacks of reads) is much lower, with many genes showing different patterns between biological replicates (Figure 7). This observation is confirmed by Diament and Tuller (2016), who investigated 15 RIBOseq data sets regarding their reproducibility on sub-codon level. The global translatome correlation was usually R≥0.85, whereas on sub-codon level, only correlations of R≤0.4 were detected. Only highly expressed genes have a slightly better reproducibility with R=0.6. Differences in sub-codon reproducibility might be explained by the choice of RNases (Gerashchenko and Gladyshev, 2017), by the translational inhibitor used, or by rRNA depletion, and ribosome purification methods (Diament and Tuller, 2016). Gene regions with a high number of RIBOseq reads at one position (stacks) might be caused by so called internal Shine-Dalgarno sequences (Li et al., 2012), but later publications rebut this finding, and report slow decoding rates for the codons of specific AAs instead (Martens et al., 2015; Mohammad et al., 2016; Nakahigashi et al., 2014; Woolstenhulme et al., 2015). Additionally, the library preparation depends on enzymes, i.e., ligase, reverse transcriptase and DNA polymerase, which have certain sequence specificities (Buschmann et al., 2016), and this might contribute to the lower reproducibility on sub-codon level.

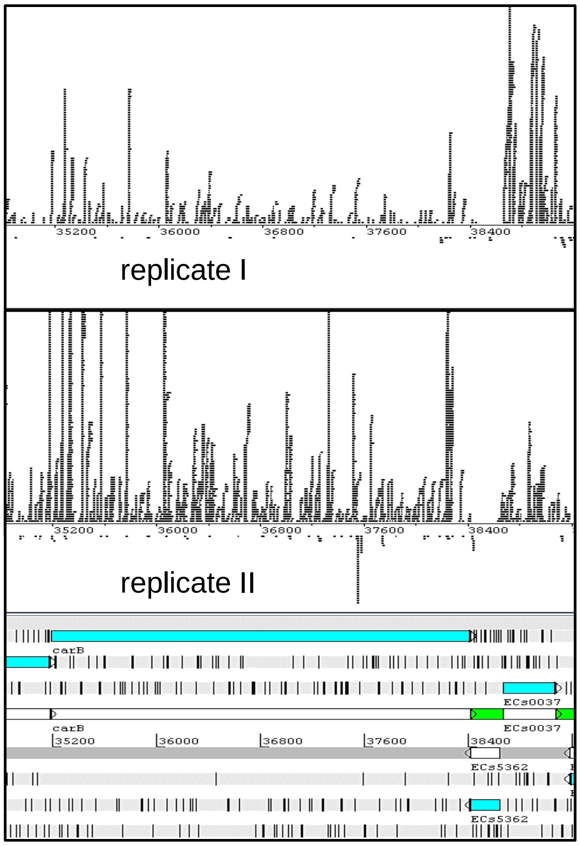


Figure 7: Artemis view of the RIBOseq reads for the condition LB at 37°C mapped to the annotated genes ECs0036 and ECs0037 of the two biological replicates separately. In the upper part, every black line represents a mapped read. Obviously, the total number of mapped reads is higher in replicate II for ECs0036 and in replicate I for ECs0037. Likewise, the distribution of the reads over the two genes is different between the replicates.

2.2 Correlation of RNAseq to RIBOseq data

RNAseq measures the abundance of mRNAs, which is dependent on the frequency of gene transcription and mRNA half-live. RIBOseq depicts mRNA stretches covered by ribosomes, obtaining a snapshot of ongoing translation. Since not every mRNA is translated to the same extent, post-transcriptional regulation can be detected (Kuersten et al., 2013). Therefore, a low correlation between RNAseq and RIBOseq data indicates abundant post-transcriptional regulation.

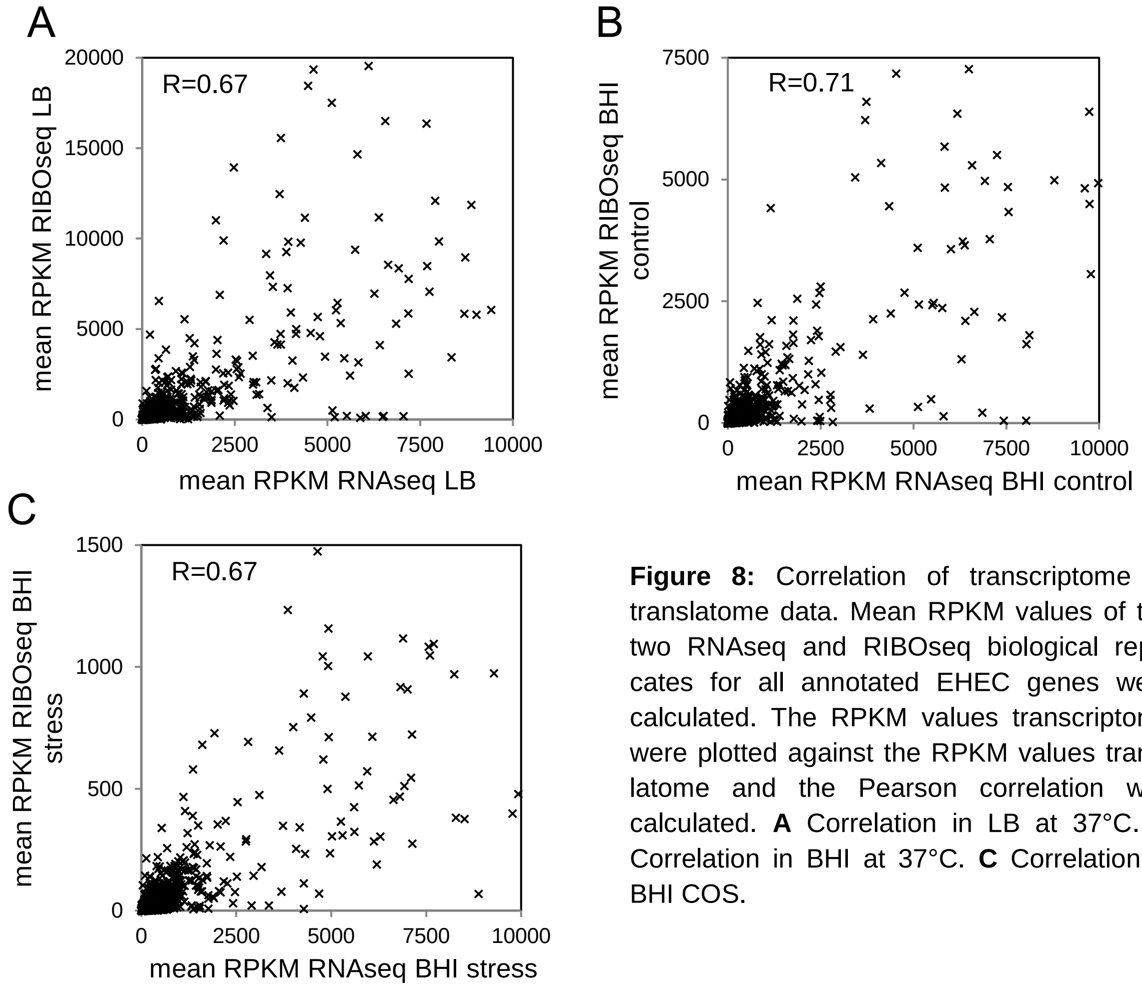


Figure 8: Correlation of transcriptome to translatome data. Mean RPKM values of the two RNAseq and RIBOseq biological replicates for all annotated EHEC genes were calculated. The RPKM values transcriptome were plotted against the RPKM values translatome and the Pearson correlation was calculated. **A** Correlation in LB at 37°C. **B** Correlation in BHI at 37°C. **C** Correlation in BHI COS.

In the data obtained in this study, the correlation between transcriptome and translatome is only moderate (Figure 8). With R=0.71, BHI control shows the best correlation, whereas the other two conditions have a correlation of R=0.67. This indicates that post-transcriptional regulation plays a large role at the investigated conditions. Other studies in bacteria report comparable correlations between RNAseq and RIBOseq data

(Bartholomaus et al., 2016; Jeong et al., 2016). Post-transcriptional regulation is especially important for genes encoded in an operon, if the co-transcribed genes are required in different stoichiometry: all genes are transcribed to the same extent, but protein abundance is regulated by translational efficiency (Li et al., 2014). Thus, this type of regulation cannot be detected in RNAseq data only.

2.3 Correlation of RIBOseq data to proteome data

Post-translational regulation can alter protein abundance by regulating protein degradation rates. Such changes are mainly detected by mass spectrometry to obtain the proteome. Kocharunchitt et al. (2012) determined the proteome of the EHEC strain used in this study at several BHI conditions, including two conditions used here. They used spectral counting to quantify protein abundance for soluble and membrane proteins separately. The correlation of the RIBOseq data to the proteome data of Kocharunchitt et al. (2012) was investigated.

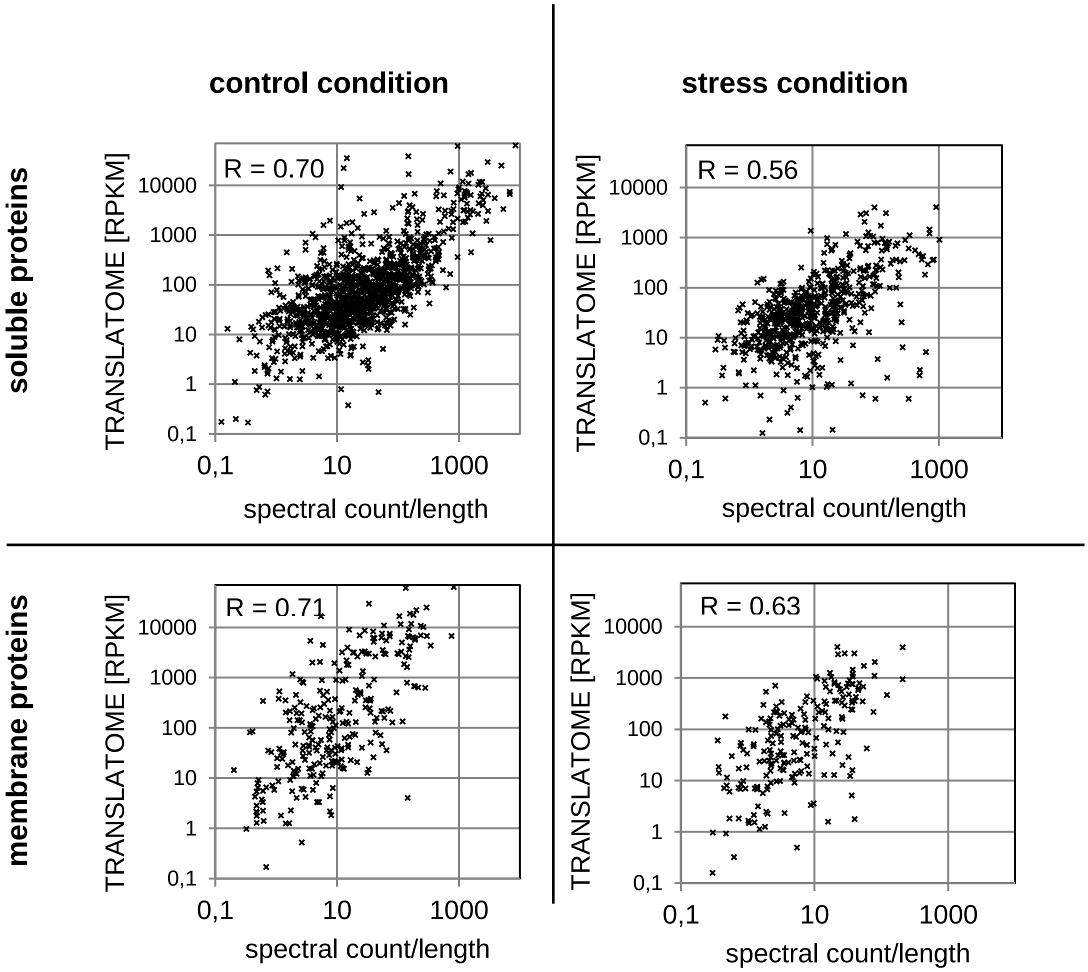
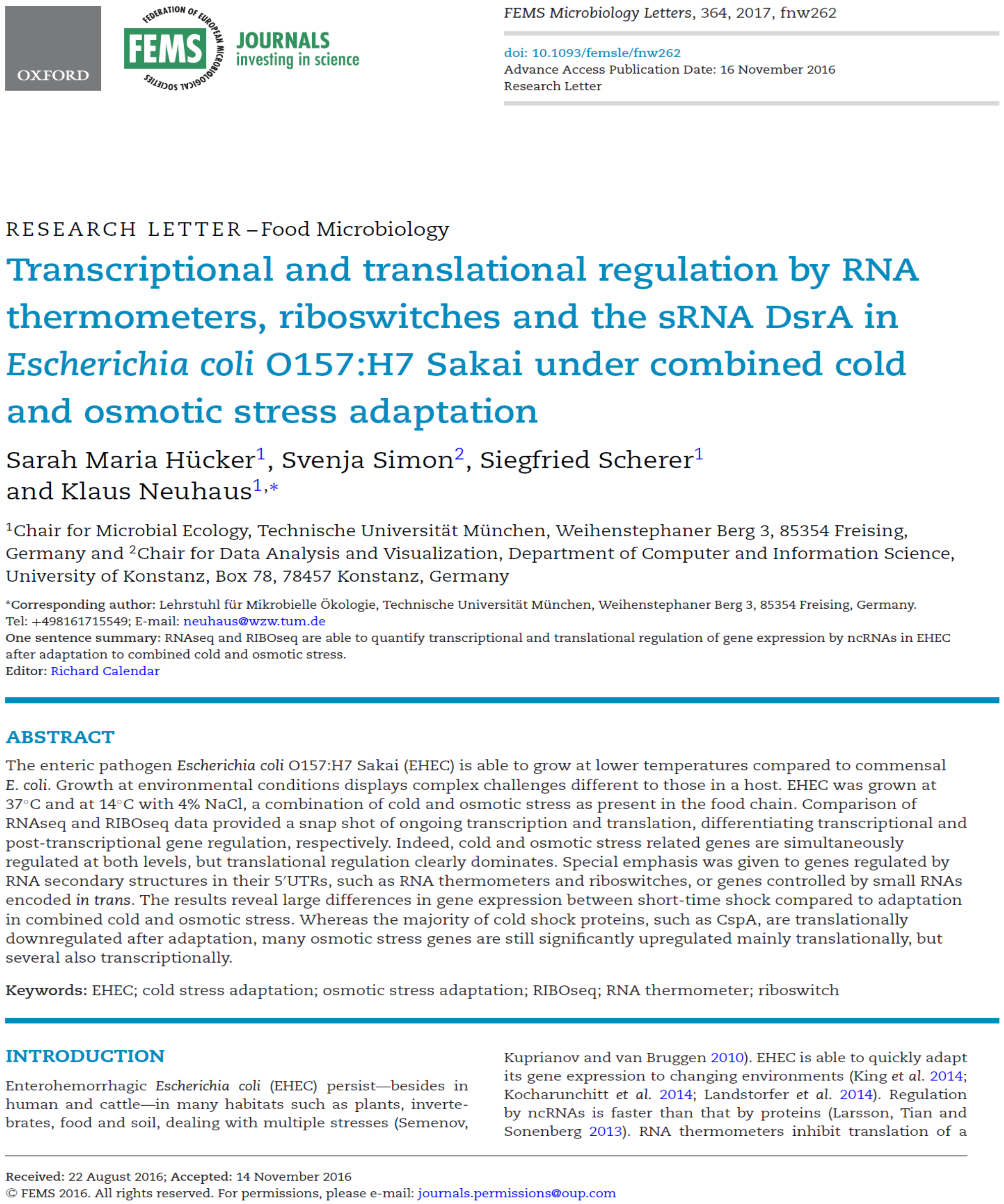


Figure 9: Correlation of translatome to proteome data. The number of detected spectra for annotated EHEC genes in the MS/MS experiment (Kocharunchitt et al., 2012) was normalized using the gene length and plotted against the RPKM values translatome (shown in logarithmic scale). The R-values in the figure were calculated by Pseudo-Pearson correlation. R-values for the Pearson correlation are: BHI control soluble proteins R=0.54, BHI control membrane proteins R=0.59, BHI stress soluble proteins R=0.41, and BHI stress membrane proteins R=0.5, respectively.

The condition BHI control shows a correlation of R=0.7 and the stress condition of R=0.56 between my translatome and the proteome of Kocharunchitt et al. (2012), respectively. Interestingly, the correlation of membrane proteins is slightly better compared to soluble proteins (Figure 9). Considering that the proteome experiments were conducted by a different group, and that spectral counting is a relatively inaccurate method to quantify proteins, the correlation is quite good. Moderate correlations between RNAseq and proteome data were reported in the past (Jayapal et al., 2008; Maier et al., 2011). Larsson et al. (2013) describe a correlation between *E. coli* RNAseq data and proteome data of R^2=0.29 to 0.59. This means that about 30-60% of the variance in protein abundance can be explained by variance in mRNA abundance, the other 40-70% have different causes like post-transcriptional regulation or experimental bias. Also, Guimaraes et al. (2014) claim for *E. coli* that the transcript level is the best predictor for protein abundance, followed by translational elongation and, eventually, initiation. A major difference between mRNA and proteins is the half-life: mRNAs have a half-life in the area of minutes, but proteins are stable in the range of hours (Taniguchi et al., 2010). Quite interestingly, on single cell level, mRNA abundance does not correlate at all with protein level in *E. coli* due to high fluctuations in the numbers of a certain mRNA over time. This is also the case in eukaryotes (Schwanhäusser et al., 2011). The different half-lifes also affect the correlation between RIBOseq and proteome data, because RIBOseq only measures the synthesis of new protein, but is not able to detect already present protein. Thus, translatomics and proteomics complement each other (Ingolia, 2014), and studies in eukaryotes observe a correlation in the same range between RIBOseq and proteome data as described here: R=0.62 in yeast (Wang et al., 2015), and R=0.7 in mammalian cells (Zur et al., 2016). A higher correlation of R=0.8 between proteome data to RIBOseq data was observed, only when newly synthesized peptides were investigated using pulsed-labeling SILAC proteomics in myeloma cells (Liu et al., 2017).

2.4 Publication 1: 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



OXFORD

FEMS JOURNALS investing in science

FEMS Microbiology Letters, 364, 2017, fnw262

doi: 10.1093/femsle/fnw262
Advance Access Publication Date: 16 November 2016
Research Letter

RESEARCH LETTER – Food Microbiology

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

Sarah Maria Hücker[1], Svenja Simon[2], Siegfried Scherer[1] and Klaus Neuhaus[1,*]

[1]Chair for Microbial Ecology, Technische Universität München, Weihenstephaner Berg 3, 85354 Freising, Germany and [2]Chair for Data Analysis and Visualization, Department of Computer and Information Science, University of Konstanz, Box 78, 78457 Konstanz, Germany

*Corresponding author: Lehrstuhl für Mikrobielle Ökologie, Technische Universität München, Weihenstephaner Berg 3, 85354 Freising, Germany. Tel: +498161715549; E-mail: neuhaus@wzw.tum.de

One sentence summary: RNAseq and RIBOseq are able to quantify transcriptional and translational regulation of gene expression by ncRNAs in EHEC after adaptation to combined cold and osmotic stress.

Editor: Richard Calendar

ABSTRACT

The enteric pathogen *Escherichia coli* O157:H7 Sakai (EHEC) is able to grow at lower temperatures compared to commensal *E. coli*. Growth at environmental conditions displays complex challenges different to those in a host. EHEC was grown at 37°C and at 14°C with 4% NaCl, a combination of cold and osmotic stress as present in the food chain. Comparison of RNAseq and RIBOseq data provided a snap shot of ongoing transcription and translation, differentiating transcriptional and post-transcriptional gene regulation, respectively. Indeed, cold and osmotic stress related genes are simultaneously regulated at both levels, but translational regulation clearly dominates. Special emphasis was given to genes regulated by RNA secondary structures in their 5′UTRs, such as RNA thermometers and riboswitches, or genes controlled by small RNAs encoded *in trans*. The results reveal large differences in gene expression between short-time shock compared to adaptation in combined cold and osmotic stress. Whereas the majority of cold shock proteins, such as CspA, are translationally downregulated after adaptation, many osmotic stress genes are still significantly upregulated mainly translationally, but several also transcriptionally.

Keywords: EHEC; cold stress adaptation; osmotic stress adaptation; RIBOseq; RNA thermometer; riboswitch

INTRODUCTION

Enterohemorrhagic *Escherichia coli* (EHEC) persist—besides in human and cattle—in many habitats such as plants, invertebrates, food and soil, dealing with multiple stresses (Semenov, Kuprianov and van Bruggen 2010). EHEC is able to quickly adapt its gene expression to changing environments (King *et al.* 2014; Kocharunchitt *et al.* 2014; Landstorfer *et al.* 2014). Regulation by ncRNAs is faster than that by proteins (Larsson, Tian and Sonenberg 2013). RNA thermometers inhibit translation of a

Received: 22 August 2016; **Accepted:** 14 November 2016

downstream protein coding open reading frame by secondary structures in the 5′UTR, sequestering the ribosomal binding site or the start codon. After a temperature shift, the secondary structure becomes unstable and translation occurs (reviewed in Narberhaus, Waldminghaus and Chowdhury 2006; Kortmann and Narberhaus 2012). Consequently, some genes induced by cold shock contain an RNA thermometer, i.e. *cspA, cspE, cspG, deaD* and *rbfA* (Giuliodori *et al.* 2010; Phadtare and Severinov 2010). In riboswitches, the conformational change of mRNA secondary structure is induced by ligand binding, which leads to transcriptional or translational activation or inhibition. Riboswitches acting on transcription can form a terminator, while translationally acting riboswitches usually sequester the Shine-Dalgarno sequence or the start codon (reviewed in Bastet *et al.* 2011). Additionally, small *trans*-encoded RNAs (sRNAs) control gene expression by base pairing to their target mRNA(s) (reviewed in Storz, Vogel and Wassarman 2011). Finally, some RNAs perform dual functions as riboswitches and as sRNAs regulating further target genes (Loh *et al.* 2009).

Although each cold and osmotic shock responses of *E. coli* have been extensively studied (reviewed in Wood 2007; Phadtare and Severinov 2010; Bartholomäus *et al.* 2016), not much is known about adaptation either to low temperature or to osmotic stress (Duffitt *et al.* 2011; Barria, Malecki and Arraiano 2013) and even less about growth under a combination reflecting an important environmental condition in food production (Kocharunchitt *et al.* 2012).

Cold shock reduces membrane fluidity and ribosomal function, alters protein folding and stabilizes RNA secondary structures. Therefore, *E. coli* immediately ceases growth after cold shock (Phadtare and Severinov 2010). The alternative σ-factor RpoS regulates the general stress response and is activated at multiple stress conditions, among others during cold and osmotic stress (Lange and Hengge-Aronis 1994; Barria, Malecki and Arraiano 2013). RpoS is required for the expression of cold shock proteins and after an acclimation phase cell growth continues at lower rate (Phadtare and Severinov 2010). The RNA chaperon CspA is highly increased after cold shock controlled by an RNA thermometer: At 37°C, the secondary structure in the 5′UTR is unstable, causing rapid mRNA degradation, whereas below 20°C the structure becomes stable with accessible ribosomal binding site (Giuliodori *et al.* 2010). Similarly, transcript and protein levels of CspE transiently increase after cold shock, but the protein is already expressed at 37°C (Uppal, Akkipeddi and Jawali 2008). Transcript levels of *cspG* and *deaD* are increased in EHEC when adapted to growth in soil (Duffitt *et al.* 2011). Concerning oxidative and osmotic stress, the transcription factors MarR, MarA and SoxS enhance the expression of efflux pumps (Duval and Lister 2013; Cohen 2014). Furthermore, the intracellular K^+ concentration is increased and organic osmolyte transporters are upregulated (Wood 2007).

RIBOseq detects mRNA in the process of translation. Combined with classic RNAseq (transcriptomics) it obtains a snapshot of transcription and translation at the time point of sampling (Ingolia *et al.* 2009; Neuhaus *et al.* 2016), thus allowing quantification and comparison of both cellular processes. Moreover, transcriptional regulation can be distinguished from translational regulation. In this work, the cellular response of combined cold and osmotic stress (COS) was investigated for genes regulated by RNA thermometers, riboswitches or the sRNA DsrA. The impact of transcriptional and translational regulation during growth under adverse conditions was quantified and many genes show regulation mainly at the translational level.

MATERIAL AND METHODS

Transcriptome and translatome sequencing

Escherichia coli O157:H7 Sakai was used in this study (Hayashi *et al.* 2001). An overnight culture was inoculated 1:100 in brain heart infusion broth (BHI). EHEC was incubated at 37°C and 150 rpm until an OD_{600} of 0.1 was reached (control), or 4% NaCl was added to the BHI and incubation was carried out at 14°C and 150 rpm until an OD_{600} of 0.1 was reached (COS).

RNA was isolated using Trizol, rRNA was depleted for RNAseq using Ribominus kit (Life Technologies Carlsbad, California, USA) and for RIBOseq using RiboZero for Gram-negative bacteria (Epicentre, Madison, Wisconsin, USA). RNAseq was conducted as described by Landstorfer *et al.* (2014). For RIBOseq, the method published by Ingolia *et al.* (2009) was adapted as described by Neuhaus *et al.* (2016) with following further changes: digestion of mRNA outside ribosomes was performed using 1 ml cell extract in buffer NEB 4 plus 1 mM $CaCl_2$ for 1 h at RT with a mixture of RNases minimizing sequence specificity: 250 U micrococcus nuclease (Roche, Penzberg, Germany), 5 U XRN-1 (NEB, Frankfurt am Main, Germany), 250 U RNase I (Ambion, Carlsbad, California, USA), 50 U RNase R (Biozym, Hessisch Oldendorf, Germany) and 12 U RNase T (NEB). For size selection, the crude footprint preparation was loaded to a 15% denaturing polyacrylamide gel. An oligonucleotide of 23 bp was used as a marker which is about the size of a bacterial ribosomal footprint. The region of 23 nt ± 3 nt was excised from the gel after SYBR Gold staining. For library preparation, the TruSeq Small RNA Sample Preparation Kit (Illumina, San Diego, California, USA) was used according to the manual. Each of the two biological replicates was sequenced on an Illumina HiSeq2500.

Bioinformatics

Results of the RNAseq and RIBOseq were analyzed for each gene using reads per kilobase gene length per million sequenced (RPKM) (Mortazavi *et al.* 2008). The ratio of RPKM RIBOseq to RPKM RNAseq gives the ribosomal coverage value (RCV), which is a measure for the 'translatability' of a given mRNA at the time point measured which is used for translation efficiency quantification of individual mRNAs. These values were determined as described (Neuhaus *et al.* 2016). Genes regulated by RNA thermometers, riboswitches and the sRNA DsrA were examined in further detail. Differential gene expression was analyzed using the *Bioconductor* package *edgeR* (version 3.2.4; Robinson, McCarthy and Smyth 2009). Read counts of the annotated genes were normalized to the smallest library. Dispersion of the number of counts in the two biological replicates was estimated genewise between the investigated conditions and differential expression was determined by the exact test, which is an analog to Fisher's test adapted to overdispersed data. Significantly changed genes had to fulfill the following criteria: fold change ≥ 2, p-value ≤ 0.05 and false discovery rate (FDR) ≤ 0.1.

RESULTS AND DISCUSSION

The biological replicates of control and COS conditions show excellent reproducibility for the transcriptome of $R = 0.96$ and $R = 0.8$ and the translatome of $R = 0.92$ and 0.79 (Fig. S1, Supporting Information). The numbers of total and mapped reads for each experiment and the amount of rRNA, tRNA and mRNA are listed in Table S1 (Supporting Information). When comparing the transcription of the control to the COS condition, the overall

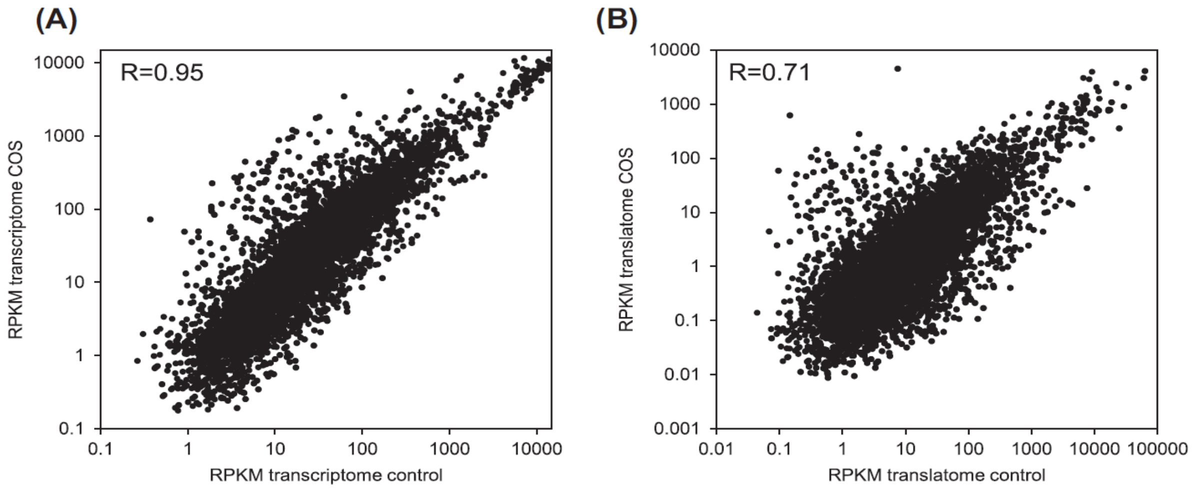


Figure 1. Comparison of the control condition (x-axis) to the COS condition (y-axis) for the transcriptome (**A**) and translatome (**B**) experiments. The Pearson correlation of the two conditions is markedly different between the transcriptome measurement and the translatome measurement (transcriptome R = 0.95 and translatome R = 0.71). The lower correlation coefficient value for the translatome experiment indicates that translational control surpasses that of transcriptional control on a global level. The global offset in the translational experiment between COS and control (about 7-fold) is caused by a lesser overall engagement of ribosomes in the COS compared to control condition.

correlation is still quite high (R = 0.95; Fig. 1A). However, when globally comparing the translatome data of the control with COS, the overall correlation drops (R = 0.71) and the cloud of data points is broadened (Fig. 1B). The general offset between control and COS is due to an overall reduced percentage of translated mRNA in the COS condition. However, the decrease in correlation between the two conditions for the translatome experiments (control vs COS) compared to the two transcriptome experiments indicates that translational regulation surpasses transcriptional regulation on a global level. This general finding was corroborated for the group of genes in which we were interested in this study (see below). For those genes, read counts, fold changes, RPKM and values expressing significance of the regulation are shown in Table 1.

Many major cold shock induced genes regulated by RNA thermometers show decreased expression at COS adaptation

Transcription of the major cold shock chaperone *cspA* (ECs4441) is slightly reduced and translation is 5-fold decreased at COS adaptation (Table 1). This decrease mainly seems to be the result of reduced translation efficiency, because the RCV is 15-fold decreased. This contradicts the well-known control by a RNA thermometer, which causes upregulation after cold shock (Giuliodori *et al.* 2010). However, transcription and translation of *cspA* are already very high at optimal growth conditions and CspA is also induced at early exponential phase (Brandi *et al.* 1999). CspA appears to be a stable protein (Ivancic, Jamnik and Stopar 2013); thus, synthesis of further protein after cold shock seems to be unnecessary for cold adaptation. Even more, bacteria control the *cspA* mRNA before growth commences, since translation initiation seems to be quite optimal for this mRNA, blocking other bulk mRNAs (Neuhaus *et al.* 2000). After adaptation to cold temperatures, *cspA* translation is repressed by autoregulatory pseudoknot formation (Kortmann and Narberhaus 2012).

The cold shock protein CspE (ECs0662) showed a significant increase of transcription (3.7-fold) and translation (2-fold) at COS (Table 1). In accordance with this finding, Duffitt *et al.* (2011) measured a 2-fold increase of *cspE* transcription of EHEC adapted to growth in soil.

Similarly, *cspG* (ECs1145) was 5-fold transcriptionally increased, when EHEC was adapted to growth in soil (Duffitt *et al.* 2011). In contrast to this, the transcription of *cspG* was slightly reduced and the translation was 5-fold decreased after COS due to RCV reduction of 11-fold. According to Uppal and Jawali (2015), CRP downregulates expression of CspG by an indirect mechanism. This regulation takes place at 37°C and 14°C, but *crp* levels were not changed in our conditions (data not shown). Anyway, adaptation to soil differs from COS and the observed downregulation of *cspG* under our stress condition could be due to the combined action of two stresses.

DeaD (ECs4043) is a helicase and during cold shock it assists translation initiation of structured mRNAs and ribosome assembly (Phadtare and Severinov 2010). In COS, translation of *deaD* is 5-fold increased. Duffitt *et al.* (2011) found a significant transcriptional increase of *deaD* in EHEC adapted to growth in soil at cold temperature. Transcription was also increased in our data, but not reaching significance (p-value 0.074; Table 1). This indicates that translational regulation of *deaD* dominates.

The heme transport protein ShuA does not belong to classical cold shock proteins, but its translation is regulated by a FourU-RNA thermometer in *Shigella dysenteriae*. At 25°C, the 5′UTR of *shuA* forms a stem loop blocking the Shine-Dalgarno sequence, at 37°C the secondary structure melts and translation occurs. EHEC possesses a homolog, *chuA* (ECs4380; Kouse *et al.* 2013). Despite Kouse *et al.* (2013) suggest a translational block below 25°C, we found transcription and translation of *chuA* highly increased (5.4-fold and 12-fold) in COS. Interestingly, high salinity causes iron starvation in *Bacillus subtilis*, followed by an upregulation of iron uptake proteins (Hoffmann *et al.* 2002). Likewise, iron uptake proteins were reported to be elevated in EHEC using the same growth conditions (Kocharunchitt *et al.* 2012). Thus, the role of the reported *chuA* RNA thermometer for upregulation in COS is unclear and the involvement of additional regulation mechanisms should be considered.

Table 1. Transcriptional and translational regulation of cold or osmotic stress-related genes. The mean number of the two biological replicates of transcriptome and translatome counts of the control and the stress condition, respectively, are shown. The fold change was calculated and differential gene expression was determined with the software edgeR. Transcriptional or translational changes were considered significant, when they showed a p-value of ≤ 0.05 and a FDR of ≤ 0.1. Significant changes in the COS adaptation compared to control are indicated with an asterisk. Mean RPKM values were used for RCV calculation.

Gene	Counts transcriptome control[a]	Counts transcriptome stress[a]	fold change (edgeR)	p-value (edgeR)	FDR (edgeR)	RPKM transcriptome control[b]	RPKM transcriptome stress[b]	Counts translatome control[a]	Counts translatome stress[a]	fold change (edgeR)	p-value (edgeR)	FDR (edgeR)	RPKM translatome control[b]	RPKM translatome stress[b]	Reference(s) about gene function and regulation
								1. Regulation by RNA thermometers							
ECs0662 (*cspE*)	438	1618	3.7*	7.77E-05	0.0013	1478	5729	889	1847	2.1*	0.023	0.077	1188	514	Uppal, Akkipeddi and Jawali (2008); Duffitt *et al.* (2011)
ECs1145 (*cspG*)	480	240	0.5	0.034	0.159	1605	839	1149	232	0.2*	2.05E-06	2.85E-05	1352	64	Phadtare and Severinov (2010)
ECs4043 (*deaD*)	1992	3538	1.8	0.074	0.261	750	1392	1591	7789	4.9*	1.95E-06	2.75E-06	234	239	Phadtare and Severinov (2010); Duffitt *et al.* (2011)
ECs4048 (*rbfA*)	325	333	1.0	0.944	1	574	616	474	379	0.8	0.487	0.68	334	55	Phadtare and Severinov (2010)
ECs4108 (*htrA*)	57	38	0.7	0.273	0.575	38	26	28	29	1.0	0.962	1	6	2	Klinkert *et al.* (2012)
ECs4380 (*chuA*)	32	174	5.4*	2.39E-06	6.86E-05	12	65	12	144	12*	3.15E-10	8.29E-09	1	4	Kouse *et al.* (2013)
ECs4441 (*cspA*)	4909	3076	0.6	0.144	0.399	16407	10737	16280	3243	0.2*	1.38E-06	2.01E-05	20603	881	Giuliodori *et al.* (2010)
ECs4627 (*ibpA*)	57	86	1.5	0.243	0.541	97	154	74	96	1.3	0.46	0.652	46	14	Krajewski, Nagel and Narberhaus (2013)
								2. Regulation by riboswitches							
ECs0002 (*thrA*)	111	99	0.9	0.752	0.954	32	30	78	149	1.9	0.053	0.143	8	4	Nawrocki *et al.* (2015)
ECs0072 (*tbpA*)	62	38	0.6	0.161	0.427	45	29	126	105	0.8	0.6	0.774	35	7	Nawrocki *et al.* (2015)
ECs0859 (*moaA*)	176	121	0.7	0.259	0.556	126	91	84	164	2.0	0.046	0.128	20	10	Regulski *et al.* (2008)
ECs1836 (*trpE*)	7	13	1.9	0.199	0.48	3	6	22	99	4.5*	3.80E-05	3.69E-04	3	4	Nawrocki *et al.* (2015)
ECs2421 (*pheS*)	154	179	1.2	0.646	0.897	111	136	217	278	1.3	0.446	0.641	58	16	Nawrocki *et al.* (2015)
ECs2820 (*hisG*)	16	83	5.2*	1.66E-05	0.0004	13	69	50	298	6.0*	3.91E-07	6.20E-06	14	20	Chan and Landick (1993)
ECs2907 (*thiM*)	21	20	1.0	0.949	1	19	19	52	34	0.7	0.24	0.424	18	3	Nawrocki *et al.* (2015)
ECs3462 (*pheA*)	48	69	1.4	0.292	0.595	29	45	21	41	2.0	0.07	0.175	5	2	Nawrocki *et al.* (2015)
ECs3929 (*ribB*)	346	258	0.7	0.367	0.675	375	294	346	422	1.2	0.541	0.722	146	38	Raghavan, Groisman and Ochman (2011)
ECs4897 (*btuB*)	314	368	1.2	0.626	0.891	121	149	133	305	2.3*	0.012	0.046	18	10	Vitreschak *et al.* (2003)
ECs4917 (*thiC*)	21	31	1.5	0.297	0.601	8	13	29	52	1.8	0.118	0.255	4	2	Raghavan, Groisman and Ochman (2011)
ECs5007 (*lysC*)	82	80	1.0	0.956	1	43	44	70	178	2.5*	0.006	0.026	12	8	Sudarsan *et al.* (2003)
ECs5219 (*mgtA*)	1378	667	0.5	0.025	0.126	15	20	1711	689	0.4*	0.005	0.023	43	2	Spinelli *et al.* (2008)

Table 1. – (*Continued*).

Gene	Counts transcriptome control[a]	Counts transcriptome stress[a]	fold change (edgeR)	p-value (edgeR)	FDR (edgeR)	RPKM transcriptome control[b]	RPKM transcriptome stress[b]	Counts translatome control[a]	Counts translatome stress[a]	fold change (edgeR)	p-value (edgeR)	FDR (edgeR)	RPKM translatome control[b]	RPKM translatome stress[b]	Reference(s) about gene function and regulation
3. Regulation by cold or osmotic stress															
ECs1739 (*hns*)	3401	2925	0.9	0.637	0.895	5848	5253	7147	2597	0.4*	0.002	0.01	4836	366	Gualerzi, Giuliodori and Pon (2003); Barria *et al.* (2013);
ECs2533 (*cspC*)	232	223	1.0	0.914	1	782	789	305	144	0.5*	0.023	0.075	407	40	Duffitt *et al.* (2011); Barria *et al.* (2013);
ECs3460 (*raiA*)	582	686	1.2	0.612	0.878	1205	1491	418	1987	4.8*	3.25E-07	4.25E-06	339	351	Di Pietro *et al.* (2013)
ECs3595 (*rpoS*)	715	2787	3.9*	3.91E-05	0.001	511	2087	239	1317	5.5*	4.77E-07	7.50E-06	50	76	Sledjeski, Gupta and Gottesman (1996); Duffitt *et al.* (2011)
ECs5155 (*rnr*)	666	1035	1.6	0.17	0.439	194	315	646	1926	3.0*	0.001	0.005	67	46	Barria *et al.* (2013)
ECs0360 (*betT*)	69	404	5.9*	3.76E-07	1.32E-05	24	148	23	409	17.8*	5.00E-14	2.24E-12	3	12	Wood (2007)
ECs0725 (*kdpB*)	8	12	1.5	0.496	0.787	3	5	9	24	2.7*	0.028	0.088	1	1	Wood (2007)
ECs2137 (*marR*)	240	94	0.4*	0.005	0.04	393	161	100	27	0.3*	2.49E-04	0.002	47	4	Cohen (2014)
ECs2138 (*marA*)	341	159	0.5	0.02	0.11	630	307	167	32	0.2*	3.40E-06	4.38E-05	106	5	Cohen (2014)
ECs2438 (*katE*)	11	38	3.5*	0.002	0.017	3	13	20	138	6.9*	1.71E-07	2.93E-06	1	4	Weber, Kogl and Jung (2006)
ECs2705 (*hchA*)	70	12	0.2*	7.69E-06	1.85E-04	59	10	130	8	0.1*	1.85E-11	5.86E-10	48	1	Weber, Kogl and Jung (2006)
ECs3542 (*proX*)	97	731	7.5*	7.05E-09	4.05E-07	70	548	53	595	11.2*	1.92E-11	6.05E-10	14	35	Weber, Kogl and Jung (2006)
ECs4399 (*treF*)	124	80	0.6	0.196	0.476	54	36	25	84	3.4*	7.93E-04	0.005	4	3	Weber, Kogl and Jung (2006)
ECs5044 (*soxS*)	49	400	8.2*	3.96E-09	2.42E-07	108	843	66	769	11.7*	9.26E-12	3.07E-10	48	135	Duval and Lister (2013)

Notes: [a]Mean count of two biological replicates normalized to the smallest library.
[b]Mean RPKM value of two biological replicates.

ROSE-like RNA thermometers require a temperature above 37°C for melting of the RNA secondary structure. As expected, neither transcription nor translation of the heat shock chaperone *ibpA* (ECs4627) is significantly changed, possessing this specific RNA thermometer (Kortmann and Narberhaus 2012; Krajewski, Nagel and Narberhaus 2013).

All in all, only CspE and DeaD show a similar regulation at the investigated COS adaptation compared to cold shock (Phadtare and Severinov 2010). In contrast, CspA and CspG are downregulated after adaptation. The heme transporter ChuA is increased, even though a decrease was expected because of the FourU-RNA thermometer. After adaptation, gene regulation by RNA thermometers seems to play a minor role at COS. RNA thermometers respond rapidly to temperature changes, but after adaptation other regulation mechanisms prevail. Further, it is unclear whether the osmotic stress is a greater challenge; therefore, osmotic adaptation induced responses may dominate the EHEC transcriptome and translatome.

Genes containing riboswitches are transcriptionally and translationally regulated

Annotated riboswitches for EHEC Sakai were downloaded from *Rfam* (Nawrocki *et al*. 2015). As expected, the majority is neither transcriptionally nor translationally regulated in COS (Table 1), because the concentration of the ligands inducing a conformational change is not altered here. However, few genes with riboswitches showed differences as described below.

Expression of *hisG* (ECs2820), which is part of histidine operon, is controlled by the histidine leader. This is a riboswitch causing premature attenuation of mRNA transcription after histidine binding (Chan and Landick 1993). Transcription of the coding region of *hisG* is 5-fold upregulated at COS, whereas transcription of the 5′UTR containing the riboswitch is unchanged (Fig. 2A). Also, translation of *hisG* is 6-fold increased at COS (Fig. 2B). Corroborating, Kocharunchitt *et al*. (2012) observed an upregulation of HisG in the transcriptome and proteome after adaptation.

The riboswitch upstream of *mgtA* (ECs5219) causes transcriptional attenuation and mRNA instability after Mg^{2+} binding (Spinelli *et al*. 2008). At COS, a 2-fold reduction of *mgtA* transcription and 2.5-fold downregulation at the translational level occur (Table 1). The RCV is even 28-fold decreased. Comparable to the histidine leader, transcription of the riboswitch is unchanged. Probably, osmotic stress downregulates this Mg transporter, because bacteria decrease membrane permeability for cations under osmotic stress (Cohen 2014).

ECs5007 has a lysine riboswitch in its 5′UTR. It is unlikely that the slight translational upregulation is caused by regulation through the riboswitch, since it functions by transcriptional attenuation (Sudarsan *et al*. 2003; Serganov, Huang and Patel 2008). Transcription of the riboswitch and the coding region is not altered.

The cobalamin riboswitch upstream of *btuB* (ECs4897) inhibits translation after ligand binding by blocking the ribosomal binding site (Vitreschak *et al*. 2003). Here, translation of *btuB* is 2.3-fold increased at COS. The same was observed by Kocharunchitt *et al*. (2014) for the protein under osmotic stress only. This might be explained by the fact that, in addition to cobalamin, magnesium also influences the cobalamin riboswitch: a Mg^{2+} concentration above 0.5 mM is necessary for the conformational change after cobalamin binding (Choudhary and Sigel 2014). As mentioned above, expression of the Mg-transporter MgtA is decreased in COS. Low magnesium levels render the ribosomal binding site of *btuB* accessible regardless of any cobalamin.

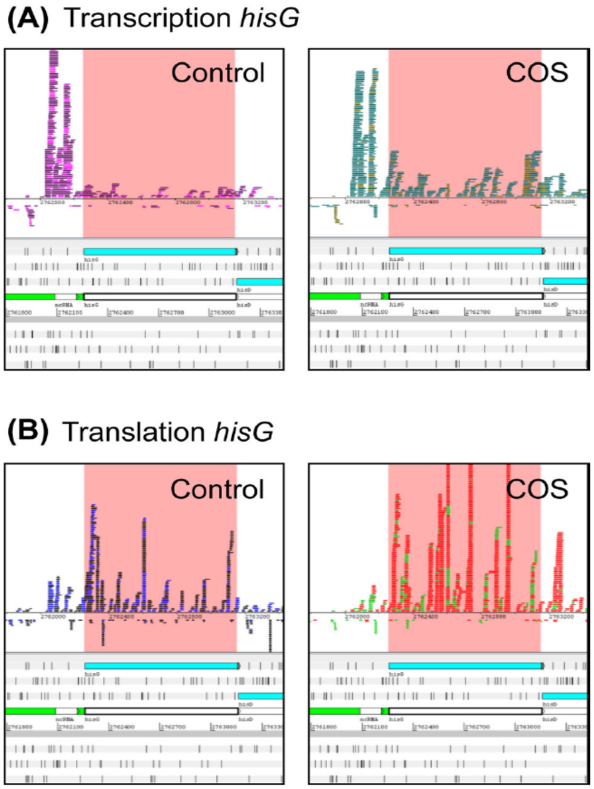


Figure 2. ECs2820 (*hisG*)—an example of transcriptional and translational regulation by a riboswitch. RNAseq and RIBOseq reads mapped to *hisG* visualized in Artemis. Please note, that some reads are not visible due to scaling artifacts, but the exact count number is recorded in Table 1. The coding region of *hisG* is highlighted in rose. RNAseq reads of the control condition replicate I are colored in pink, and of replicate II in purple; reads of the COS condition replicate I are colored in turquoise, and of replicate II in olive. RIBOseq reads of the control condition replicate I are colored in black, and of replicate II in blue; reads of the COS condition replicate I are colored in red, and of replicate II in green. **(A)** Transcription of *hisG* at the control and the COS condition. There is a 5.2-fold increase of transcription at stress. **(B)** Translation of *hisG* at the control and the stress condition. The mean read number normalized to the smallest library mapping to the coding region of *hisG* increases from 50 reads at the control to 298 reads at COS. This represents a 6-fold upregulation.

The sRNA DsrA and the alternative σ-factor RpoS are induced at COS adaptation

The sRNA *dsrA* is driven by a cold-inducible promotor (Repoila and Gottesman 2003). Accordingly, its transcription is 8.8-fold induced in COS (Fig. 3A). DsrA has several target genes and regulates at both transcriptional and translational levels. At the transcriptional level, DsrA overcomes the silencing of H-NS targeted genes; at the translational level, it activates RpoS (ECs3595) at low temperature (Sledjeski, Gupta and Gottesman 1996). Without DsrA, the 5′UTR of *rpoS* mRNA forms a secondary structure masking its ribosomal binding site. Mediated by Hfq, *dsrA* forms base pairs over 20 nts with the 5′UTR of *rpoS*, resolving the secondary structure allowing translation

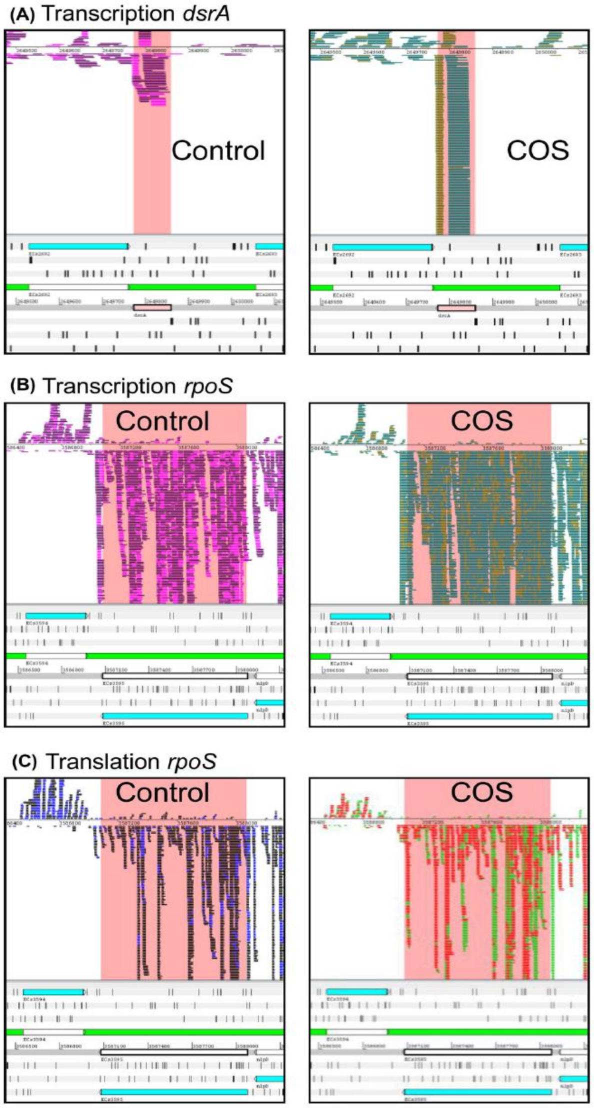


Figure 3. RNAseq and RIBOseq reads mapped to the sRNA *dsrA* and to ECs3595 (*rpoS*). Data are visualized using Artemis. The exact numbers of mapped reads can be found in Table 1. The coding region of *rpoS* and the sRNA *dsrA* are highlighted in rose. RNAseq reads of the control condition replicate I are colored in pink, and of replicate II in purple; reads of the COS condition replicate I are colored in turquoise, and of replicate II in olive. RIBOseq reads of the control condition replicate I are colored in black and of replicate II in blue; reads of the COS condition replicate I are colored in red, and of replicate II in green. **(A)** Transcription of *dsrA* at the control and stress condition. At COS the number of mapped reads is 8.8-fold higher. **(B)** Transcription of *rpoS* at the control and the COS condition. There is a 3.9-fold increase of transcription at stress. **(C)** Translation of *rpoS* at the control and the stress condition. At COS, translation is 5.5-fold upregulated.

(Majdalani *et al.* 1998). Here, not only translational, but also transcriptional upregulation of *rpoS* occurred at stress of about 5.5-fold and 3.9-fold, respectively (Fig. 3B and C). Corroborating, *rpoS* is transcriptionally increased in EHEC adapted to growth in soil (Duffitt *et al.* 2011) and expression of proteins of the RpoS regulon is enriched 13.8-fold (Kocharunchitt *et al.* 2012). Bartholomäus *et al.* (2016) reported an upregulation of *rpoS* at the transcriptional and translational level in *Escherichia coli* upon osmotic shock. Of course, an important regulator like RpoS is not only controlled by DsrA. Transcriptional regulation is also mediated by alternative promotors, mRNA stability and the transcription factors ArcA, CRP and Fur (reviewed in Landini *et al.* 2014). Expression of *arcA* and *crp* are not altered in COS; however, *fur* is translationally increased in COS (data not shown). Fur also activates iron uptake genes that fit well to an increase of *chuA* expression as described above. Furthermore, two additional sRNAs enhance *rpoS* expression: ArcZ and RprA (Landini *et al.* 2014). Transcription of *arcZ* is not altered, but abundance of *rprA* is increased 10.5-fold under stress.

Transcriptional and translational regulation of other cold/osmotic stress-inducible genes

Many other genes reported to be induced after cold or osmotic shock show changed expression in COS (Table 1), but their regulatory mechanism(s) are yet unknown. For instance, CspC (ECs2533), pY (ECs3460), H-NS (ECs1739) and RNase R (ECs5155) increase after cold shock (Barria, Malecki and Arraiano 2013). In agreement to this, *raiA* (encoding pY) is 4.8-fold translationally upregulated and *rnr* (encoding RNase R) is 3-fold increased (Table 1). pY binds to the 30S ribosomal subunit to modify it and thus, reducing translation of specific mRNAs and RNase R digests RNA secondary structures (Di Pietro *et al.* 2013). Obviously, these functions are also required for cold adaptation and not only after cold shock. Surprisingly, for several genes, the direction of regulation at combined stress adaptation is opposite compared to cold shock alone. *CspC* is translationally 2-fold downregulated and is also transcriptionally decreased in EHEC adapted to cold stream water (Duffitt *et al.* 2011). Additionally, the translatability of *cspC* mRNA is 8.5-fold reduced in COS. The exact role of H-NS in cold shock is unknown, but it is required for extended growth in the cold (Gualerzi, Giuliodori and Pon 2003). Anyway, in COS translation but not transcription of *hns* is 3-fold reduced caused by an 11.8-fold decrease of the RCV.

Genes reported to be regulated by osmotic stress show larger changes in expression than genes regulated by cold stress under our conditions (Table 1). For instance, the transcription factors *marR* (ECs2137) and *marA* (ECs2138) are significantly decreased, whereas *soxS* (ECs5044) is 8.2-fold transcriptionally and 11.7-fold translationally upregulated. In contrast, all transcription factors are transcriptionally upregulated in EHEC adapted to growth in soil (Duffitt *et al.* 2011). The choline transporter *betT* (ECs0360) is 5.9-fold upregulated transcriptionally and even 17.8-fold translationally. Additionally, the potassium transporting ATPase *kdpB* (ECs0725) is 2.7-fold increased translationally. Weber, Kogl and Jung (2006) investigated changes of the proteome in *E. coli* after osmotic stress induced with sorbitol or NaCl. Three proteins, induced in their study, were also upregulated at the stress condition used here. The betaine transporter *proX* (ECs3542) is increased 7.5-fold transcriptionally and 11.2-fold translationally. Interestingly, the cytoplasmic trehalase *treF* (ECs4399) shows differential regulation: its transcription in slightly downregulated, whereas the translation increases 3.4-fold at COS. The hydroxyperoxidase KatE (ECs2438) is part of the RpoS regulon. As explained above, *rpoS* is increased at the stress condition. *katE* is 3.5-fold transcriptionally and 6.9-fold translationally upregulated. Transcriptional increase is confirmed by Kocharunchitt *et al.* (2012) for the same COS condition. The chaperon HchA (ECs2705) was upregulated in osmotic stress (Weber, Kogl and Jung 2006). Since HchA is a heat shock protein, it is not surprising

Loh E, Dussurget O, Gripenland J *et al.* A trans-acting riboswitch controls expression of the virulence regulator PrfA in *Listeria monocytogenes*. *Cell* 2009;**139**:770–9.

Majdalani N, Cunning C, Sledjeski D *et al.* DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *P Natl Acad Sci USA* 1998;**95**:12462–7.

Mortazavi A, Williams BA, McCue K *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.

Narberhaus F, Waldminghaus T, Chowdhury S. RNA thermometers. *FEMS Microbiol Rev* 2006;**30**:3–16.

Nawrocki EP, Burge SW, Bateman A *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* 2015;**43**:D130–7.

Neuhaus K, Landstorfer R, Fellner L *et al.* Translatomics combined with transcriptomics and proteomics reveals novel functional, recently evolved orphan genes in *Escherichia coli* O157:H7 (EHEC). *BMC Genomics* 2016;**7**:133.

Neuhaus K, Rapposch S, Francis KP *et al.* Restart of exponential growth of cold-shocked *Yersinia enterocolitica* occurs after down-regulation of *cspA1/A2* mRNA. *J Bacteriol* 2000;**182**:3285–8.

Phadtare S, Severinov K. RNA remodeling and gene regulation by cold shock proteins. *RNA Biol* 2010;**7**:788–95.

Raghavan R, Groisman EA, Ochman H. Genome-wide detection of novel regulatory RNAs in *E. coli*. *Genome Res* 2011;**21**:1487–97.

Regulski EE, Moy RH, Weinberg Z *et al.* A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism. *Mol Microbiol* 2008;**68**:918–32.

Repoila F, Gottesman S. Temperature sensing by the *dsrA* promoter. *J Bacteriol* 2003;**185**:6609–14.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2009;**26**:139–40.

Semenov AM, Kuprianov AA, van Bruggen AH. Transfer of enteric pathogens to successive habitats as part of microbial cycles. *Microb Ecol* 2010;**60**:239–49.

Serganov A, Huang L, Patel DJ. Structural insights into amino acid binding and gene control by a lysine riboswitch. *Nature* 2008;**455**:1263–7.

Sledjeski DD, Gupta A, Gottesman S. The small RNA, DsrA, is essential for the low temperature expression of RpoS during exponential growth in *Escherichia coli*. *EMBO J* 1996;**15**:3993–4000.

Spinelli SV, Pontel LB, Garcia Vescovi E *et al.* Regulation of magnesium homeostasis in *Salmonella*: Mg(2+) targets the *mgtA* transcript for degradation by RNase E. *FEMS Microbiol Lett* 2008;**280**:226–34.

Storz G, Vogel J, Wassarman KM. Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell* 2011;**43**:880–91.

Sudarsan N, Wickiser JK, Nakamura S *et al.* An mRNA structure in bacteria that controls gene expression by binding lysine. *Gene Dev* 2003;**17**:2688–97.

Uppal S, Akkipeddi VS, Jawali N. Posttranscriptional regulation of *cspE* in *Escherichia coli*: involvement of the short 5'-untranslated region. *FEMS Microbiol Lett* 2008;**279**:83–91.

Uppal S, Jawali N. Cyclic AMP receptor protein (CRP) regulates the expression of *cspA, cspB, cspG* and *cspI*, members of *cspA* family, in *Escherichia coli*. *Arch Microbiol* 2015;**197**:497–501.

Vitreschak AG, Rodionov DA, Mironov AA *et al.* Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA* 2003;**9**:1084–97.

Weber A, Kogl SA, Jung K. Time-dependent proteome alterations under osmotic stress during aerobic and anaerobic growth in *Escherichia coli*. *J Bacteriol* 2006;**188**:7165–75.

Wood JM. Bacterial osmosensing transporters. *Methods Enzymol* 2007;**428**:77–107.

Microbiology 2013;**159**:2437–43.

Bartholomäus A, Fedyunin I, Feist P *et al.* Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philos T Roy Soc A* 2016;**374**:20150069.

Bastet L, Dube A, Masse E *et al.* New insights into riboswitch regulation mechanisms. *Mol Microbiol* 2011;**80**:1148–54.

Brandi A, Spurio R, Gualerzi CO *et al.* Massive presence of the Escherichia coli 'major cold-shock protein' CspA under non-stress conditions. *EMBO J* 1999;**18**:1653–9.

Chan CL, Landick R. Dissection of the his leader pause site by base substitution reveals a multipartite signal that includes a pause RNA hairpin. *J Mol Biol* 1993;**233**:25–42.

Choudhary PK, Sigel RK. Mg(2+)-induced conformational changes in the *btuB* riboswitch from *E. coli*. *RNA* 2014;**20**:36–45.

Cohen BE. Functional linkage between genes that regulate osmotic stress responses and multidrug resistance transporters: challenges and opportunities for antibiotic discovery. *Antimicrob Agents Ch* 2014;**58**:640–6.

Di Pietro F, Brandi A, Dzeladini N *et al.* Role of the ribosome-associated protein PY in the cold-shock response of *Escherichia coli*. *Microbiologyopen* 2013;**2**:293–307.

Duffitt AD, Reber RT, Whipple A *et al.* Gene expression during survival of *Escherichia coli* O157:H7 in soil and water. *Int J Microbiol* 2011;**2011**:340506.

molecular zippers and switches. *Nat Rev Microbiol* 2012;**10**:255–65.

Kouse AB, Righetti F, Kortmann J *et al.* RNA-mediated thermoregulation of iron-acquisition genes in *Shigella dysenteriae* and pathogenic *Escherichia coli*. *PLoS One* 2013;**8**:e63781.

Krajewski SS, Nagel M, Narberhaus F. Short ROSE-like RNA thermometers control IbpA synthesis in *Pseudomonas* species. *PLoS One* 2013;**8**:e65168.

Landini P, Egli T, Wolf J *et al.* sigmaS, a major player in the response to environmental stresses in *Escherichia coli*: role, regulation and mechanisms of promoter recognition. *Environ Microbiol Rep* 2014;**6**:1–13.

Landstorfer R, Simon S, Schober S *et al.* 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* 2014;**15**:353.

Lange R, Hengge-Aronis R. The cellular concentration of the σ^S subunit of RNA polymerase in *Escherichia coli* is controlled at the levels of transcription, translation, and protein stability. *Gene Dev* 1994;**8**:1600–12.

Larsson O, Tian B, Sonenberg N. Toward a genome-wide landscape of translational control. *Cold Spring Harb Perspect Biol* 2013;**5**:a012302.

2.5 Differential expression of annotated genes in LB compared to BHI

In Hücker et al. (2017b), only the two BHI conditions were compared, the condition LB at 37°C is not included in this publication. Similar to the publication, differential expression of transcriptional and/or translational levels of all annotated EHEC genes were identified using the Bioconductor package *edgeR* (Robinson et al., 2009). BHI at 37°C was used as the reference condition to which gene expression at LB at 37°C was compared. A gene is considered differentially expressed, when the p-value is ≤ 0.05 and the false discovery rate is ≤ 0.1.

Both growth conditions, LB and BHI at 37°C, reflect optimal temperature with high nutrient availability. A major difference between both conditions was the time point of harvest: in BHI, the sample was taken in early exponential growth phase, whereas in LB, gene expression at mid-exponential growth was investigated. Therefore, changes in gene expression are likely to be caused by those differences. Overall, 30% of annotated genes show differential regulation (listed in Supplementary Table S1). At transcriptional level, more genes are regulated than at translational level (Figure 10). In addition, more genes are downregulated in LB than upregulated (40% transcriptionally downregulated, 10% translationally downregulated, and 8.6% downregulated on both levels, respectively). The overlap between significant regulation on transcriptional and translational level is only moderate. Accordingly, comparisons of *E. coli* K-12 after heat shock and after osmotic shock show only a moderate overlap between transcriptionally and translationally regulated genes (Bartholomaus et al., 2016). However, in most cases the change in expression of either transcription or translation goes in the same direction, indicating that RNAseq and RIBOseq results generally confirm each other (Vogel and Marcotte, 2012). Despite, 21 genes were found downregulated on transcriptional level, but upregulated on translational level and for two other genes, the regulation is vice versa. This number is relatively small, since King et al. (2014) report up to 6% of EHEC genes with significant regulation into different directions on RNA compared to protein level after cold shock.

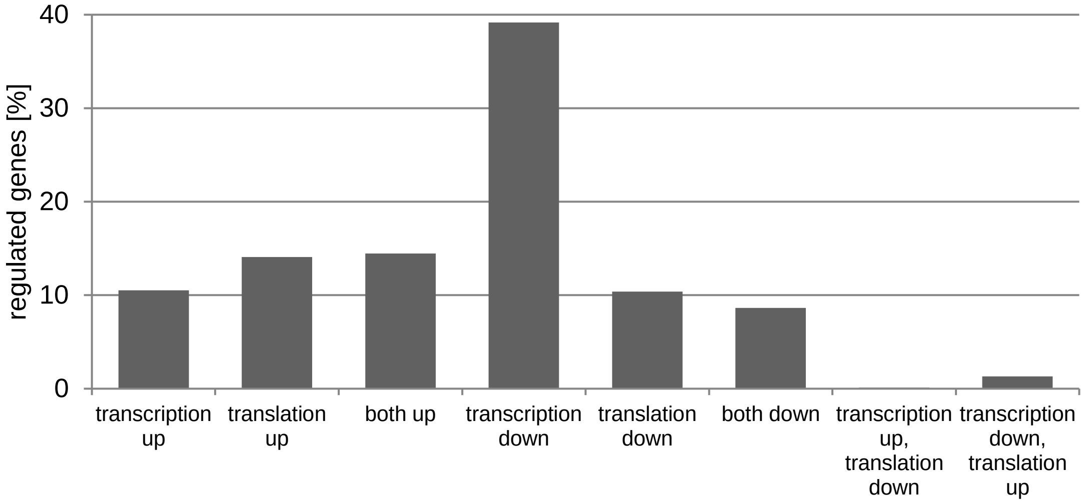


Figure 10: Differentially regulated genes in LB at 37°C compared to BHI at 37°C. Differential regulation was calculated using *edgeR*. The diagram shows the abundance of upregulation/downregulation and of transcriptional/translational regulation in percent.

Part II: High-throughput discovery of novel genes

The focus of this work is the discovery of putative novel genes in the EHEC genome. It is possible that intergenic regions harbor small protein-coding ORFs, which were overlooked at initial genome annotation. Therefore, every intergenic ORF of at least 93 bp (corresponding to a 30 AA protein) was investigated regarding its coverage with RNAseq and RIBOseq reads. The ORF was assumed to be translated, if covered ≥ 50% with RIBOseq reads, and showing a translatability (i.e., number of ribosomes per mRNA) comparable to annotated genes.

2.6 Publication 2: Discovery of numerous novel small genes in the intergenic regions of the *Escherichia coli* O157:H7 Sakai genome

Discovery of numerous novel small genes in the intergenic regions of the *Escherichia coli* O157:H7 Sakai genome

Sarah M. Hücker[1,5], Zachary Ardern[1,5], Tatyana Goldberg[2], Andrea Schafferhans[2], Michael Bernhofer[2], Gisle Vestergaard[3], Chase W. Nelson[4], Michael Schloter[3], Burkhard Rost[2], Siegfried Scherer[1,5], and Klaus Neuhaus[1,6]*

[1]Chair for Microbial Ecology, Technische Universität München, Freising, Germany [2]Department of Informatics - Bioinformatics & TUM-IAS, Technische Universität München, Garching, Germany

[3]Research Unit Environmental Genomics, Helmholtz Zentrum München, Neuherberg, Germany

[4]Sackler Institute for Comparative Genomics, American Museum of Natural History New York, New York 10024, USA

[5]ZIEL - Institute for Food & Health, Technische Universität München, Freising, Germany

[6]Core Facility Microbiome/NGS, ZIEL - Institute for Food & Health, Technische Universität München, Freising, Germany

*Corresponding author: Klaus Neuhaus, neuhaus@tum.de

Abstract

In the past, short protein-coding genes were often disregarded by genome annotation pipelines. Transcriptome sequencing (RNAseq) signals outside of annotated genes have usually been interpreted to indicate either ncRNA or pervasive transcription. Therefore, in addition to the transcriptome, the translatome (RIBOseq) of the enteric pathogen *Escherichia coli* O157:H7 strain Sakai was determined at two optimal growth conditions and a severe stress condition combining low temperature and high osmotic pressure. All intergenic open reading frames potentially encoding a protein of ≥ 30 amino acids were investigated with regard to coverage by transcription and translation signals and their translatability expressed by the ribosomal coverage value. This led to discovery of 465 unique, putative novel genes not yet annotated in this *E. coli* strain, which are evenly distributed over both DNA strands of the genome. For 255 of the novel genes, annotated homologs in other bacteria were found, and a machine-learning algorithm, trained on small protein-coding *E. coli* genes, predicted that 89% of these translated open reading frames represent bona fide genes. The remaining 210 putative novel genes without annotated homologs were compared to the 255 novel genes with homologs and to 250

short annotated genes of this *E. coli* strain. All three groups turned out to be similar with respect to their translatability distribution, fractions of differentially regulated genes, secondary structure composition, and the distribution of evolutionary constraint, suggesting that both novel groups represent legitimate genes. However, the machine-learning algorithm only recognized a small fraction of the 210 genes without annotated homologs. It is possible that these genes represent a novel group of genes, which have unusual features dissimilar to the genes of the machine-learning algorithm training set.

Introduction

The pathogenic *E. coli* strain O157:H7 Sakai (EHEC) was first isolated in 1996 from an outbreak in Japan [1]. When contaminated food is consumed, EHEC can cause bloody diarrhea and the disease may progress to the life-threatening hemolytic uremic syndrome [2]. In addition to humans [3] and contaminated food, EHEC persists in many environments, such as soil [4], plants [5], invertebrates [6], and cattle [7]. These environments represent various challenges requiring expression of a different set of bacterial genes [8]. Since there is no vaccination or targeted therapy available [9], it is important to better understand the biology of this enteric pathogen in order to prevent infections.

In contrast to eukaryotic genomes, bacterial genomes are densely covered with annotated protein-coding genes, e.g., 88.1% of the EHEC Sakai genome consists of protein-coding genes according to the most recent genome annotation [1]. Nevertheless, it is still possible that intergenic regions harbor overlooked short genes [10, 11]. After sequencing a bacterial genome, bioinformatics tools, such as GLIMMER [12] or RAST [13] are used for gene prediction and annotation. Especially for short genes, these tools are biased in that open reading frames (ORFs) shorter than 150 bp are often rejected [14] and in some cases are not even permitted for database entry [15]. Thus, the sensitivity of automated annotation processes in predicting short genes is quite low [16]. Additionally, the experimental detection of small proteins in proteome studies is difficult: Many small proteins are lost during proteome purification and many more are not detectable by classic mass spectrometry, because they do not produce enough tryptic peptides of the proper size [17]. Therefore, small proteins have been largely ignored in the past and our knowledge of their structures and functions is very limited [15]. Although small proteins have recently come more into focus [18, 19], the majority of them still belong to the ‘dark proteome’ lacking known folds or domains, thus rendering putative functional assignments using bioinformatics tools impossible [20, 21].

The rise of next-generation sequencing technologies allows high-throughput investigation of the expression status of genomes without any restriction to gene length. RNAseq strand-specifically determines the global transcriptome and widespread transcription outside of annotated genes has become increasingly obvious [22-25]. In the past, these transcription signals were generally interpreted as ncRNAs [26, 27] or just pervasive transcription without any biological significance [28-30]. However, ribosomal footprinting (RIBOseq) can be used to determine the coverage of RNA with ribosomes, indicating translation into a peptide of the associated RNA, thus, facilitating the global investigation of the translatome [31, 32]. Even more, RIBOseq reads usually show a triplet periodicity reflecting the codon-wise movement of the ribosome during the translation process [31, 33]. Combining ribosomal footprinting with RNAseq allows estimation of the translatability of an ORF, expressed by the ribosomal coverage value (RCV), which is the ratio of the reads per kilobase (of gene) per million sequenced reads (RPKM) value for the translatome over the RPKM value for the transcriptome. The RCV can be used to distinguish ncRNA from translated mRNA, and RIBOseq allows the discovery of many non-annotated short translated ORFs [33-39]. In bacteria, RIBOseq is less frequently applied. However, Baek et al. [40] recently reported 130 novel short genes in *Salmonella*, the smallest gene encoding a peptide of only 7 amino acids (AA). The translatome of EHEC strain EDL933 under a single growth condition yielded 72 novel genes encoded in intergenic regions, 95% of them encoding proteins smaller than 100 AA [11].

In this study, RIBOseq and RNAseq analysis of *E. coli* O157:H7 Sakai was compared at three different growth conditions to identify translated ORFs in the intergenic regions. The resulting candidates for novel genes were further characterized using bioinformatics analysis.

Results

Translatome signals of putative novel genes

The transcriptome and the translatome of EHEC Sakai were determined at three different growth conditions. Two standard lab conditions (lysogeny broth (LB) at 37°C; Brain-heart-infusion (BHI) at 37°C) and combined cold and osmotic stress (COS; BHI supplemented with 4% NaCl at 14°C) in two biological replicates each. Details about total read number and amount of rRNA, tRNA, and mRNA are listed in S1 Table. All intergenic ORFs of at least 30 AA length were considered as potentially encoding a protein if significant RIBOseq signals were found. A RIBOseq signal was assumed significant at a threshold of at least 1 RPKM, at least 50% ORF coverage, and an RCV of at least 0.25. This analysis resulted in 1271 potentially translated intergenic ORFs, which were manually examined for the following additional criteria before consideration as candidate genes. First, ORFs with identical sequences to others were removed. Next, every ORF with its mapped RIBOseq reads was visualized in the Artemis viewer [41]. False positives were assumed if the signal could have been caused by neighboring annotated genes and not by the putative ORF of interest and, as such were excluded. In the case of same-strand overlapping ORFs in different reading frames, the ORF with the better fit to the RIBOseq signal was selected. After individual inspection in which 806 candidates were excluded, we arrived at a conservative estimate of 465 intergenic ORFs, which were considered to show convincing evidence of translation in the RIBOseq experiments. The novel putative genes were consecutively numbered in the order they appear in the EHEC genome (XECs001-XECs465). The novel genes were approximately uniformly distributed within the whole genome, occurring on both strands of the chromosome (Fig 1). Details about position on the genome, length, RPKM value, coverage, and RCV of all novel genes are found in S2 Table.

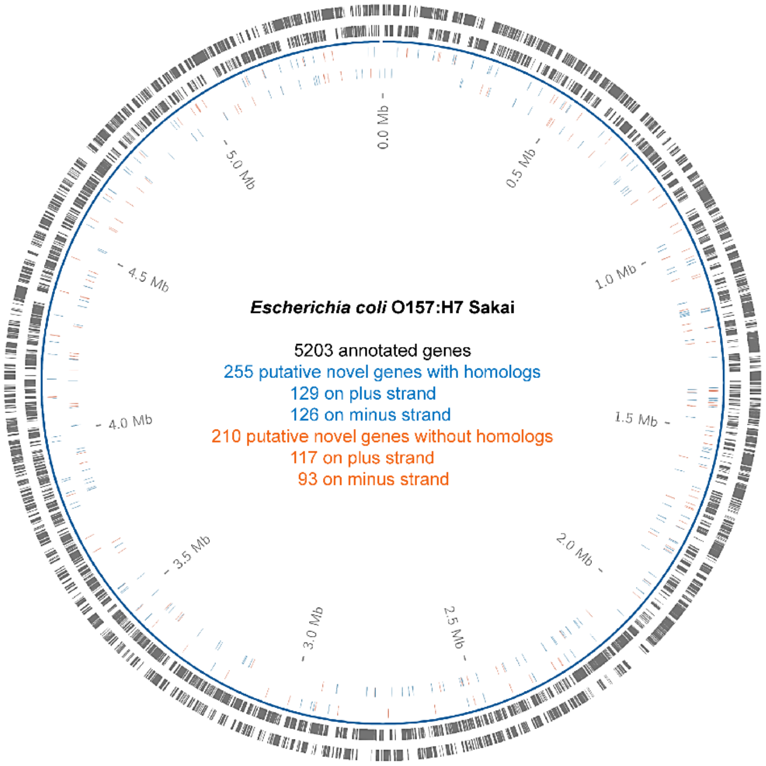


Fig 1. Distribution of 465 small novel genes within the EHEC genome. The circles from outside to inside show: annotated genes on the plus strand, annotated genes on the minus strand, novel genes on the plus strand and novel genes on minus strand. Novel genes with annotated homologs are colored in blue and novel genes without annotated homologs are colored in orange.

Two-hundred-eleven (211) novel genes show translation at both optimal growth conditions (LB and BHI at 37°C), 210 novel genes are detected in LB only, and four are detected in BHI control only. RIBOseq signals of 32 novel genes are shared under all three conditions but no gene fulfills the criteria for candidate gene inclusion in BHI COS only (Fig 2 and S2 Table). One example of a translated intergenic ORF for each growth condition is visualized in Fig 3. The three novel gene candidates depicted are clearly covered by RIBOseq reads over their entire length and it is considered highly unlikely that the translation signals are caused by neighboring annotated genes. Additionally, the novel genes show sufficient RCVs of 0.51 (XECs135), 0.58 (XECs029) and 0.29 (XECs459), confirming translation.

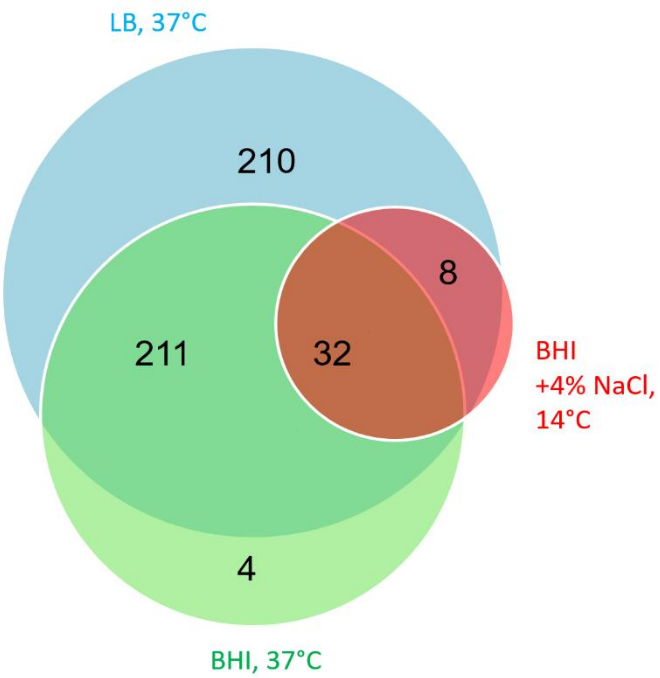
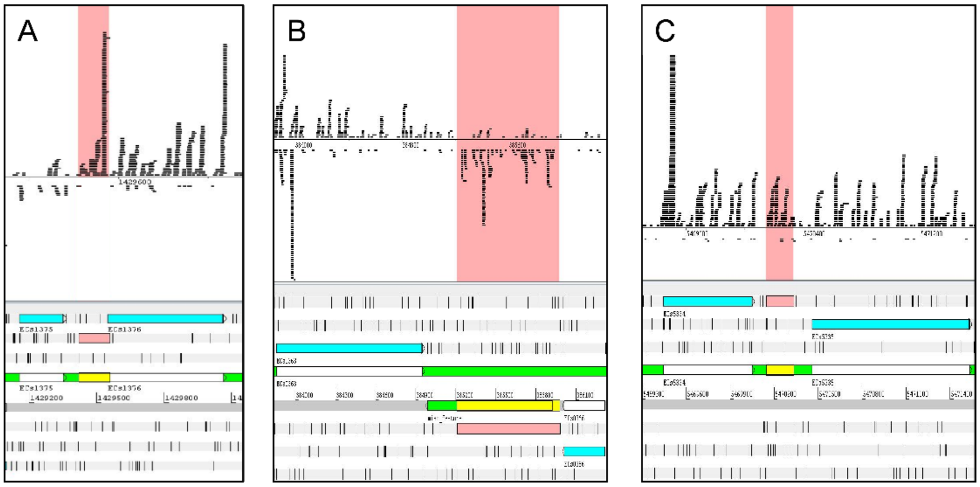


Fig 2. Growth conditions where the novel genes reach or exceed translation thresholds. The Venn-Diagram shows how many ORFs are translated under the three growth conditions investigated. The majority of novel genes are translated at optimal growth conditions leading to a large overlap between LB and BHI control. Blue: LB at 37°C, green: BHI at 37°C, red: BHI + 4% NaCl at 14°C.



XECs135, 138 bp
BHI control

XECs029, 768 bp
LB

XECs459, 189 bp
BHI stress

Fig 3. Three novel genes with RIBOseq signals as examples. In the lower part, the corresponding section of the genome is shown with the novel gene highlighted in pink. In the upper part, the strand-specifically mapped RIBOseq reads are displayed, whereby each black line represents a sequenced read.

Annotated homologs of novel genes

The amino acid sequences of the novel genes were used as a query to find annotated homologous proteins in other bacteria with blastp using default parameters against the RefSeq database. With an e-value threshold of ≤ 10^{-3}, 55% of the putative proteins encoded in the novel genes match an annotated homolog (Table 1). When a more stringent e-value threshold of ≤ 10^{-10} was applied, 42% of novel genes still possess annotated homologs. The hits with the lowest e-value for each novel gene are listed in S3 Table. Interestingly, 34 of the novel genes are annotated in other *E. coli* O157:H7 strains, of which twelve were found in the EHEC strain EDL933 [42], which is the closest relative to strain Sakai used in this study. Additionally, eleven of the novel genes detected in the intergenic regions of EHEC EDL933 in a previous study [11] were confirmed for EHEC Sakai, as well.

Table 1. Summary of the properties of the short annotated genes, novel genes with annotated homologs and novel genes without annotated homologs.

	(i) Short annotated genes (control group)	**(ii) Translated ORFs with annotated homologs**	**(iii) Translated ORFs without annotated homologs**
Number of ORFs analyzed	250	255	210
Mean length [bp]	192	172	127
Mean RCV LB	1.55	2.04	1.74
Mean RCV BHI control	0.55	0.44	0.5
Mean RCV BHI COS	0.12	0.11	0.1
Regulated genes (BHI control versus LB)	82 (32.8%)	103 (40.4%)	76 (36.2%)
Regulated genes (BHI control versus BHI COS)	90 (36%)	210 (82.4%)	170 (81%)
Promoter predicted	242 (96.8%)	242 (94.9%)	210 (100%)
Mean promoter localization (bp upstream start)	187	137	128
Mean promoter strength (LDF score)	3.43	3.44	3.86
Terminator predicted	55 (22%)	53 (20.8%)	32 (15.2%)
Mean terminator localization (bp downstream stop)	68	107	127
Mean terminator score	-16.86	-16.72	-15.87
Shine-Dalgarno (SD) motif predicted	200 (80%)	114 (44.7%)	74 (35.2%)
Mean ΔG° of the SD motifs	-5.17	-4.61	-4.47
Mean SD localization (bp upstream start)	7	8	11
Machine-learning algorithm prediction “real“	248 (99.2%)	226 (88.6%)	5 (2.4%)
Machine-learning algorithm prediction “pseudo”	2 (0.8%)	29 (11.4%)	205 (97.6%)
$K_A>K_S$	7 (2.8%)	0	0
$K_A<K_S$	5 (2%)	12 (4.7%)	5 (2.4%)

Based on the blastp analysis with an e-value threshold of ≤ 10^{-3}, the 465 novel genes were divided into two groups: one group of 255 ORFs, which have annotated homologs in other bacteria (‘with annotated homolog’), and a second group of 210 ORFs for which no annotated homologs were found in the database (‘without annotated homolog’). Furthermore, the 250 shortest annotated genes of EHEC Sakai with an RCV of at least 0.25 in LB (S2 Table, S4 Table) were compared to the two groups of novel genes (see also below; S3 Table). Even though the shortest annotated genes were used, they are on average longer (mean 192 bp) than the novel genes (mean 172 bp). The novel genes without annotated homologs being the shortest, with a mean length of 127 bp (Table 1). More than 50% of the latter group would encode a protein of just 30-39 AA (Fig 4A). However, the largest novel gene would encode a protein of 425 AA. For the three groups, the RCV distribution is shown for LB in Figure 4B. All groups show a comparable pattern: the majority of genes have a moderate translatability and a subset of genes is translated with high efficiency. Growth in BHI control and in BHI COS also yield RCV distributions which are similar among the three gene groups (S1 Fig). Overall, translatability is somewhat decreased under BHI control, but there is a massive decline of translatability under BHI COS condition (Table 1). However, the decline is in a similar range for all three groups and attributable to the stress condition.

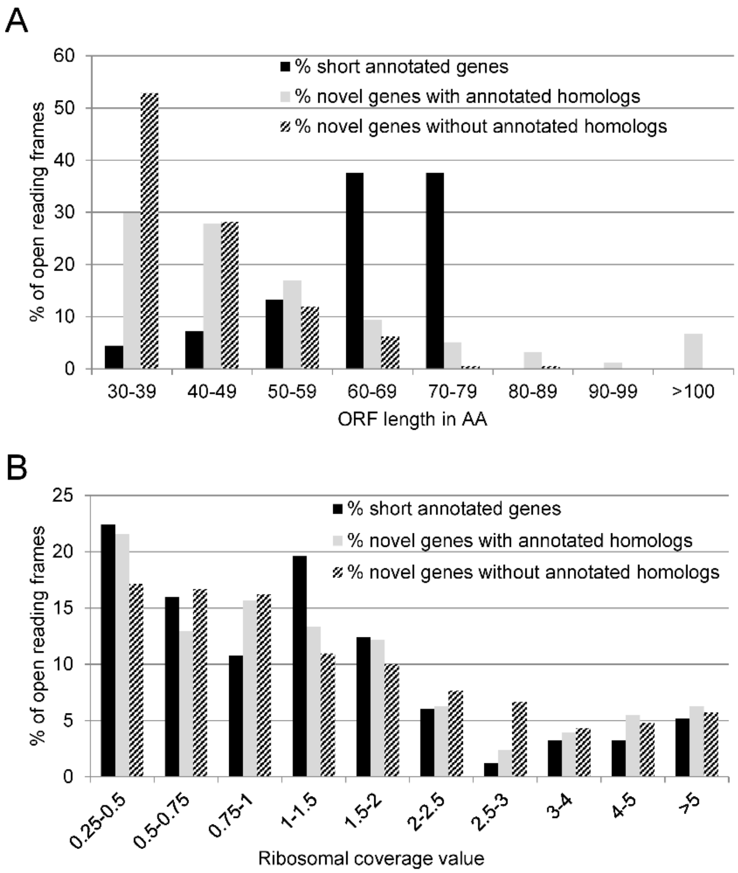


Fig 4. Length and RCV distribution of short annotated genes, novel genes with annotated homologs, and novel genes without annotated homologs. (A) The ORF length in AA was binned into eight categories and the number of ORFs for each gene group belonging to every category was determined. On average, the annotated genes are longer than the novel genes. The novel genes without annotated homologs have the shortest length. (B) The translatability expressed by the ribosomal coverage value (RCV) when growing in LB. The RCV was binned into ten groups. All three gene categories show a similar RCV distribution.

Sequence conservation

A tblastn search for non-annotated homologs of the novel genes in other organisms, using the RefSeq genomic database, shows high conservation levels within the *Escherichia* genus and often more widely (Fig 5). Six novel genes with annotated homologs (blastp) and three putative novel genes without annotated homologs did not have tblastn hits. Thus, 249 and 207 genes with unique sequences are shown in Fig 5A and 5B, respectively. The novel genes with annotated homologs (blastp) show more unannotated homologs (tblastn) with greater average evolutionary distance and AA similarity compared to those novel genes without annotated homologs (blastp). A two-tailed t-test comparing the maximum distance of intact homologs (tblastn) for the novel genes with and without annotated homologs (blastp) gives a p-value of p=0.002. Thus, the maximum evolutionary distance of the homologs found using tblastn is significantly different for both groups (i.e., genes with and without annotated homologs using blastp).

There is some evidence for horizontal gene transfer of some ORFs, with highly similar sequences found in distant bacterial genera, and even eukaryotes, for instance multiple matches between XECs029 and *Drosophila* genomes. The sequences in the RefSeq database might be misidentified. However, the phenomenon of transfer of bacterial genome regions to arthropods has been described [43].

Intergenic sequences upstream and downstream of the novel genes were analyzed as above. As expected, sequence similarity is less preserved in the upstream and downstream regions when compared to the ORF-sequence of the novel genes (S2 Fig). For intact homologs (i.e., no stop codon) of the novel genes, the average sequence similarity for intact tblastn hits outside of the *Escherichia/Shigella* genera is 69% (S5 Table). Average sequence similarity for all homologs of the sequences upstream and downstream of the novel genes is lower, at 47% (S2 Fig).

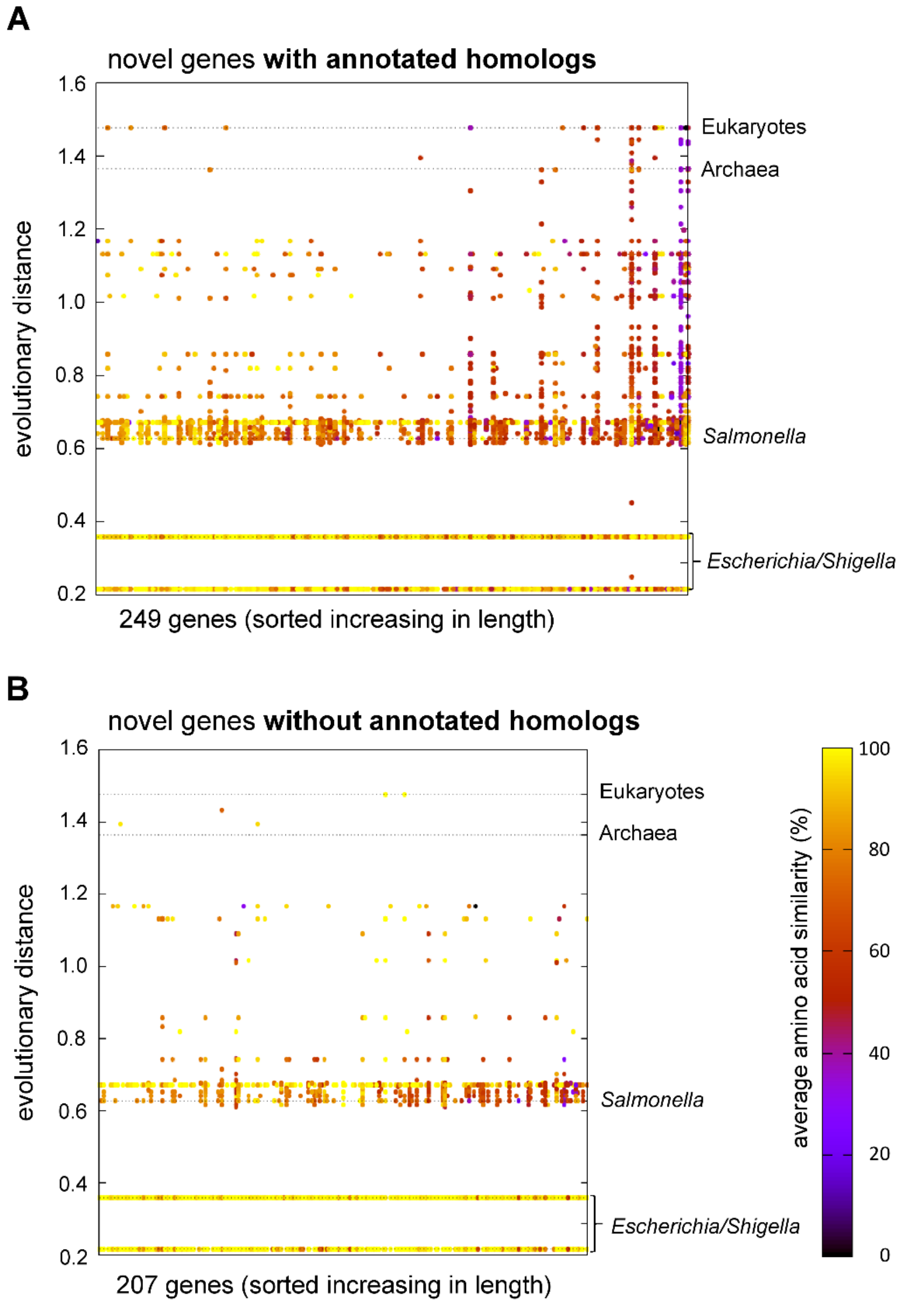
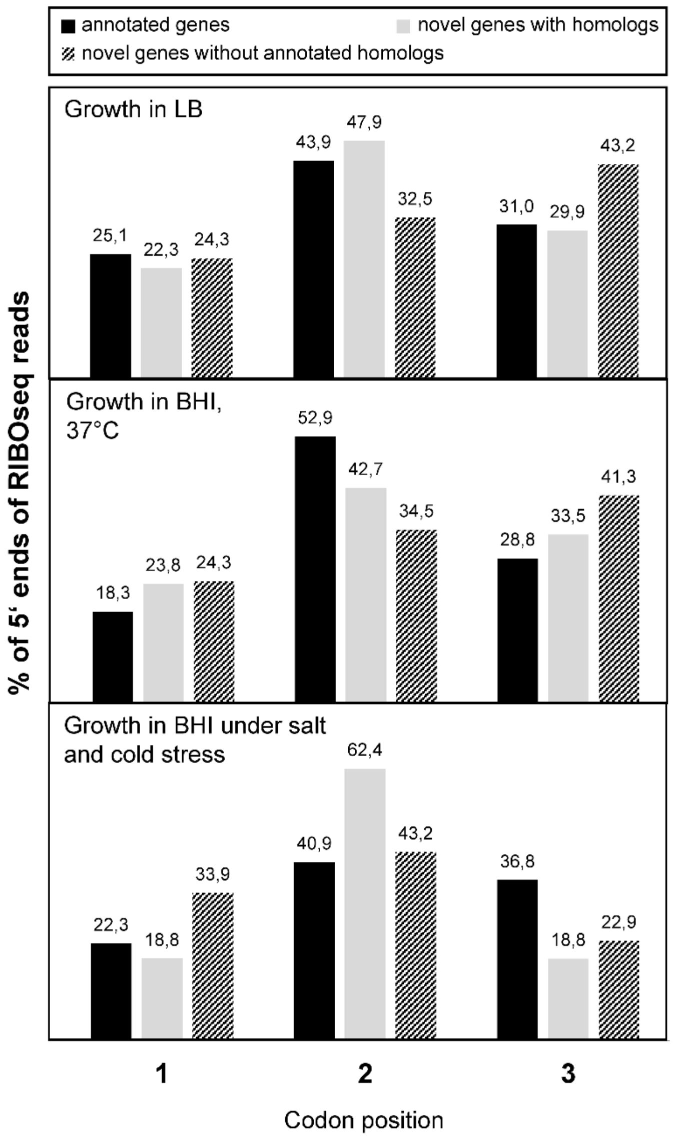


Fig 5. Conservation of novel genes with and without annotated homologs. Average AA sequence similarity (according to the color scale) for all target sequences from a tblastn search of the RefSeq genomic database, for each ORF is shown. Each dot represents a hit in the database for a given novel gene, with points combined and similarity averaged by genus. Novel genes are spread across the X-axis ordered by their length; the Y-axis shows the taxonomic distance of each genus, using the SILVA database 16S rRNA alignment guide tree. (A) Novel genes with at least one annotated homologous protein sequence. (B) Novel genes without annotated homologs. Those with annotated homologs tend to be found across more genera. Note that the number of homologs found in each genus is not indicated, with the vast majority being in *Escherichia* and *Shigella*. Data overview is provided in S5 Table.

Triplet periodicity of the RIBOseq signal

A characteristic of RIBOseq data, at least from eukaryotes, is that the reads show a triplet periodicity reflecting the codon-wise translation by the ribosome [31]. Thus, the codon positions of 5' ends of all RIBOseq reads with read length 20 bp were determined in the sum signal of all annotated genes and of the novel genes with and without annotated homologs. Indeed, the annotated genes and the novel genes with annotated



homologs show a reading frame signal at codon position two for all investigated growth conditions (Fig 6). However, the signal is weak and the novel genes without annotated homologs only show a reading frame at codon position two when grown in BHI COS.

Fig 6. Reading frame in the sum signal of annotated genes, novel genes with annotated homologs, and without annotated homologs. The 5' ends of RIBOseq reads of length 20 bp were investigated with regard to their codon position. The bar diagrams show the percentage of 5' ends on every codon position for the three investigated growth conditions. Annotated genes and novel genes with annotated homologs have the majority of 5' ends at position two for every condition. The novel genes without annotated homologs only show a reading frame at codon position two at the condition BHI + 4% NaCl at 14°C.

Differential regulation of the novel genes

Differential expression at transcriptional and translational levels between growth conditions indicates regulation of gene expression, which implies functionality. Therefore, we investigated the novel genes for significantly changed transcription and translation using BHI control as the reference condition in comparison to LB and BHI COS. In addition, the 250 shortest annotated genes were analyzed as a control group. Comparing growth in BHI and LB medium at 37°C showed that about one third of the genes in each group is differentially expressed (Table 1). XECs170 is an example of a transcriptionally and translationally upregulated novel gene (Fig 7A): the transcription in

LB is 2.7-fold increased and the translation is even 9.8-fold higher. For all groups, downregulation in LB is more frequent than upregulation. Downregulation occurs more often at the transcriptional level, whereas for upregulation translational changes are more frequent (Fig 7B). Fold changes, p-values and false discovery rates determined with edgeR [44] for all significantly regulated genes are listed in S6 Table.

When the two BHI conditions are compared, even more genes show differential regulation. For example, the novel gene XECs197 is clearly expressed at the control condition, but transcription and translation are almost switched off at BHI COS (Fig 7C). For the short annotated genes, 40% are regulated, but for the novel genes without annotated homologs and the novel genes with annotated homologs 81% and 82.4% are differentially expressed, respectively (Table 1, S7 Table). Although the absolute number of regulated genes is higher for the novel genes, all three gene groups show the same trend (Fig 7D): the majority are downregulated at BHI COS, where translational regulation clearly dominates.

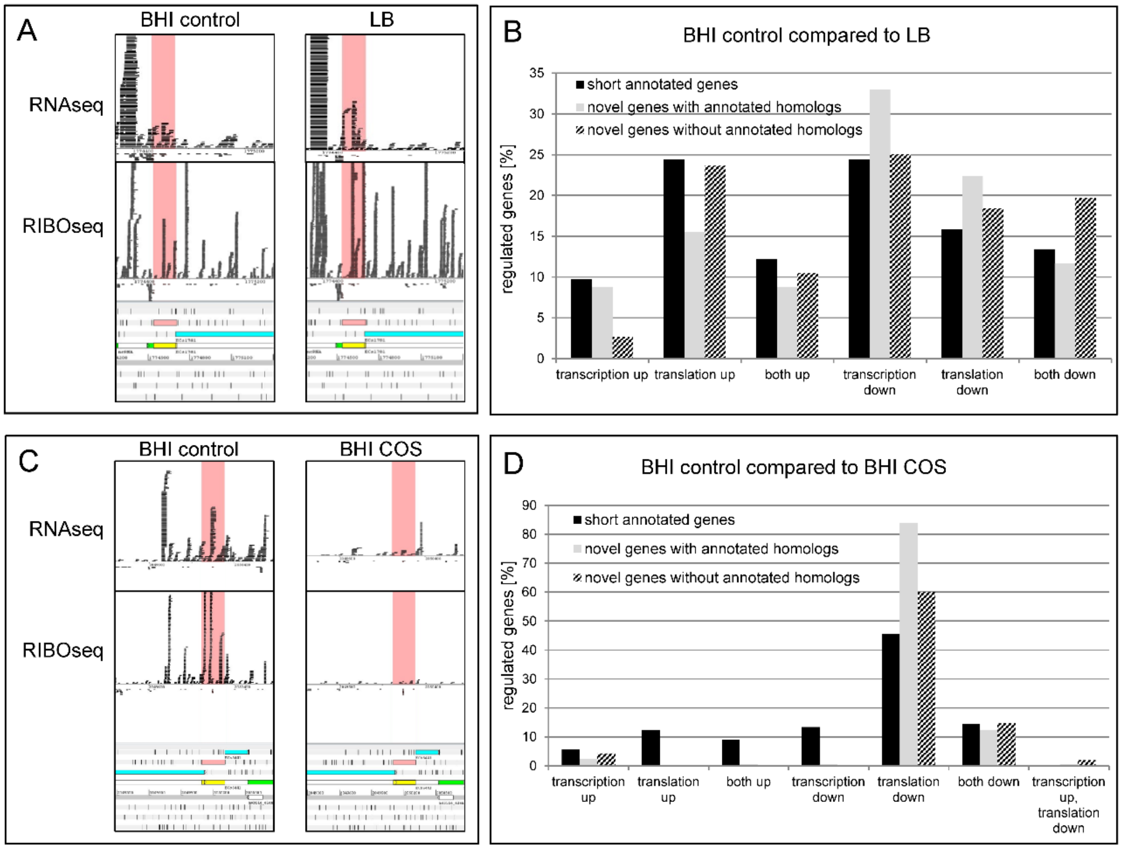


Fig 7. Differentially regulated genes under different growth conditions. (A) Example of a transcriptionally and translationally upregulated gene in LB compared to BHI control. The novel gene XECs170 is highlighted in pink. The transcription of XECs170 is increased 2.7-fold and translation 9.8-fold. (B) Summary of differentially regulated genes in LB compared to BHI control. For all three gene categories, downregulation dominates. (C) Example of a transcriptionally and translationally downregulated gene in BHI COS compared to BHI control. Transcription of XECs197 is 5.5-fold and translation is 129-fold reduced at the stress condition. (D) Summary of differentially regulated genes in BHI COS compared to BHI control. Downregulation at the translational level clearly dominates for all gene categories.

Bioinformatics analyses

Predicted protein characteristics. The software PredictProtein [45, 46] predicts many parameters of an amino acid sequence including composition, secondary structure, protein localization, disordered regions, as well as the number of DNA/RNA binding sites, disulfide bonds and transmembrane helices. Prediction of secondary structures is very similar for the three groups (Fig 8A). The proteins mainly fold into α-helices and loops, β-sheet-like structures are less common. Concerning disordered regions, the three groups contain a similar average portion of disorder of about 20% regarding the UCON prediction [47] (S8 Table, S9 Table). Forty-four (9.5%) novel genes show evidence of transmembrane helices (Fig 8B). The proportion of short annotated genes with predicted transmembrane helices is higher (18%). Novel genes with annotated homologs also more often contain a transmembrane helix than do novel genes without annotated homologs (12.9% compared to 5.2%, respectively). For the number of predicted disulfide bonds an opposite picture was obtained. The novel genes without annotated homologs more often have one or more disulfide bonds predicted, followed by the novel genes with annotated homologs, but 90% of the short annotated genes seem not to contain any disulfide bond (Fig 8C). The localization of the putative proteins was also predicted: 34 putative novel proteins should localize in the inner or outer membrane, while surprisingly, 85% are predicted to be secreted (Fig 8D). Whereas the localization prediction of the novel genes with and without annotated homologs is similar, the result for the short annotated genes is slightly different: Many of them should still be secreted (45%), but the number predicted to be cytoplasmic and inner membrane proteins is higher. Further details and additional properties of the novel genes and the short annotated genes are listed in S8 Table and S9 Table.

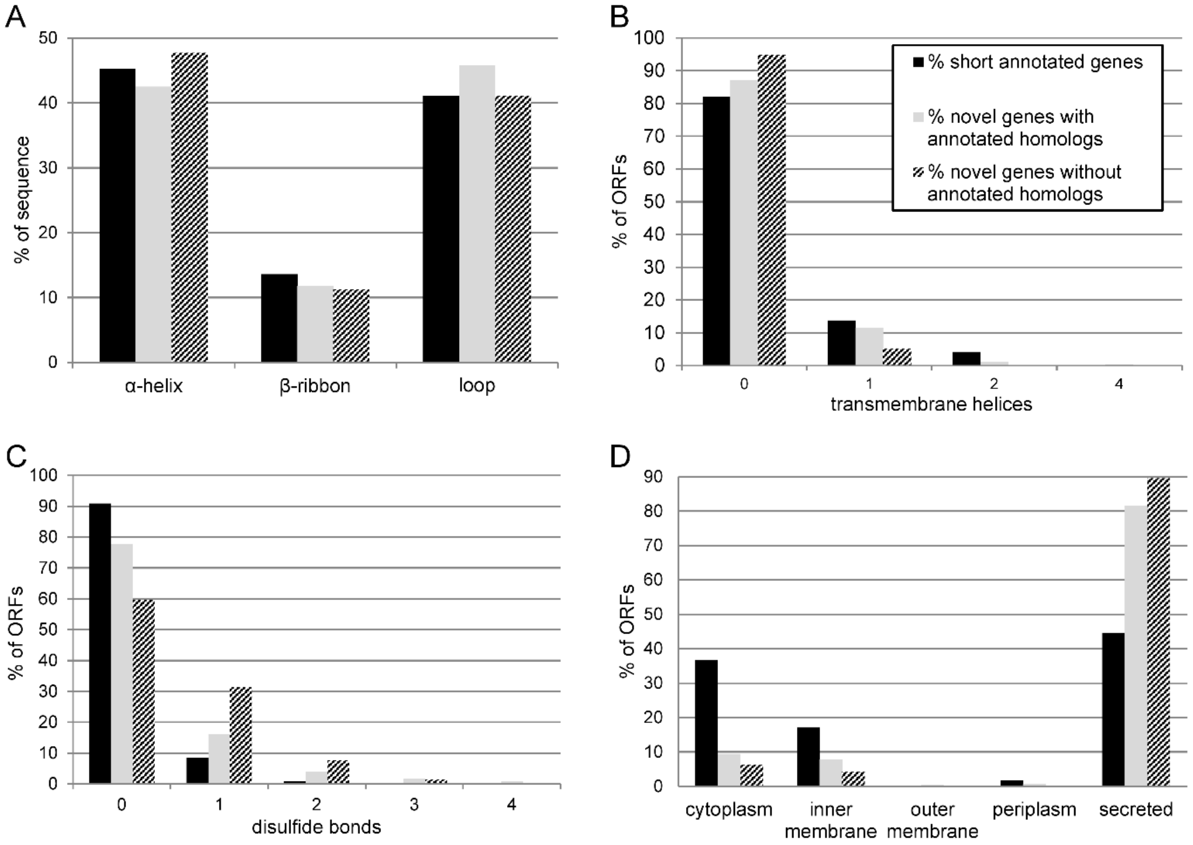


Fig 8. Selected results of PredictProtein for the short annotated genes, and the novel genes with and without annotated homologs. (A) Average secondary structure composition. (B) Number of predicted transmembrane helices. (C) Number of predicted disulfide bonds. (D) Predicted localization of the proteins within the *E. coli* cell. Percentage values for every gene separately can be found in S8 Table and S9 Table.

Machine learning trained on known EHEC proteins confirms blastp hits. The above-mentioned parameters were also predicted for a number of short annotated proteins of *Escherichia coli* O157:H7 EDL933 to obtain a positive control set. As a negative control set, these natural proteins were scrambled (for each positive control sequence, 100 randomly scrambled sequences were used) and submitted for PredictProtein analysis. A machine-learning algorithm was trained on the positive and negative control sets to distinguish between ‘real’ protein sequences and scrambled ones (‘pseudo’) [11]. This algorithm was used to investigate the 465 translated ORFs found in this study (S3 Table) and the 250 short annotated genes of EHEC Sakai (S4 Table). Again, every amino acid sequence was scrambled 10-times as a negative control. As expected, the algorithm recognized 99.4% of the scrambled proteins as ‘pseudo’ and 99.2% of the short annotated genes as ‘real’ based on predicted parameters of those sequences. Overall, 50% of the novel genes were recognized as ‘real’. However, the presence of an annotated homolog (found via blastp) correlates well with being predicted as ‘real’ by the machine-learning algorithm and vice versa (Table 1, S10 Table). Only five novel genes without annotated homologs were recognized by machine-learning algorithm as ‘real’ proteins. Conversely, 29 novel genes with annotated homologs were predicted as ‘pseudo’ proteins (Table 1 and S3 Table).

Promoter and terminator prediction. A promoter is required to initiate transcription of an ORF and is recognized by the σ-factor of the RNA polymerase holoenzyme. The housekeeping σ-factor in *E. coli* is σ^{70} (reviewed in [48]). Therefore, σ^{70} promoter sequences were searched in the regions 300 bp upstream of putative start codons of the novel genes using BPROM. Interestingly, all novel genes without annotated homologs have a predicted promoter in their upstream region and in the upstream regions for the novel genes with annotated homologs a promoter sequence appears to be present in 95% of the cases (Table 1 and S3 Table). On average, the predicted promoter sequence localizes 187 bp upstream of the start codon for the annotated genes. In the case of the novel genes, the distance to the start codon is slightly shorter. The LDF score is a measure of the promoter strength and a promoter is considered active with an LDF score of at least 0.2. The average LDF score of the predicted promoters for the three gene groups is similar: 3.43 for the short annotated genes, 3.44 for the novel genes with annotated homologs and 3.86 for the novel genes without annotated homologs, respectively (Table 1).

Transcription termination mediated by ρ-independent terminators [49] in the region 300 bp downstream of the stop codon was investigated using FindTerm. For 20.8% of the novel genes with annotated homologs a terminator was predicted. For those without annotated homologs, the fraction was slightly lower (Table 1).

Shine-Dalgarno sequence and start codons. The presence of a Shine-Dalgarno (SD) sequence upstream of the start codon promotes efficient translation initiation [50]. The consensus SD motif for *E. coli* is uaAGGAGGu and base pairing of this sequence with the anti-SD of the 16S rRNA results in a free energy of ΔG° -9.6 [51]. Within the region 30 bp upstream of the start codons 41% of the novel genes with annotated homologs and 35.2% without annotated homologs have a SD sequence (Table 1). A high proportion of the annotated genes have a SD sequence (80%). Additionally, the average free energy of the SD is lower for the annotated genes (-5.17 compared to -4.61 and - 4.47, respectively). The upstream regions of XECs059 (novel gene with annotated homolog) and XECs428 (novel gene without annotated homolog) contain a perfect SD sequence (S3 Table).

ATG is the most common start codon, but also GTG, TTG, and the rare start codons CTG, ATT, ATA, and ATC can initiate translation in *E. coli* [52]. Genome annotation algorithms only search for the three most common start codons (ATG, GTG, and TTG, respectively) [12] and in accordance with this, the group of the annotated genes shows for 90% of genes an ATG start codon, for 7.2% a GTG start codon, and for 2.8% a TTG start codon, whereas rare start codons are not present at all. In case of the novel genes, the real start codon is unknown. Because of that the potential start codon farthest upstream of the coding region, but within the transcriptome signal, was chosen no matter whether it was a frequent or rare start codon. Therefore, only 42% of the novel genes with annotated homologs and 32.8% of the novel genes without annotated homologs start with either ATG, GTG, or TTG. All other genes, putatively, have rare start codons. However, it cannot be excluded that some of these genes possess an ATG, GTG, or TTG start codon further downstream of the open reading frame.

Evolutionary sequence analysis of novel genes. The rates of non-synonymous (amino acid changing) and synonymous (not amino acid changing) substitutions per site, kA and kS respectively, reflect the evolutionary processes underlying the divergence of related genes. In the absence of selection, it is expected that kA ≈ kS, indicating neutrality. On the other hand, when purifying selection acts to eliminate disadvantageous mutations, the fact that most fitness-altering mutations are nonsynonymous implies that selection will disproportionately slow the rate of divergence at non-synonymous sites, leading to kA < kS. On the other hand, when positive selection acts to promote advantageous mutations, this will disproportionately increase the rate of divergence at non-synonymous sites, leading to kA > kS. Although intergenic junk sequences are expected to evolve neutrally, functional genes can also exhibit kA ≈ kS because of near-neutrality or a balance between positive and purifying selective forces. We reasoned that only functional protein-coding sequences would show significant signs of positive or negative selection and, based on the hypothesis that our novel genes are functional, we predicted that the proportion of genes exhibiting significant signatures of selection should be similar between novel candidate genes and annotated genes.

To test this hypothesis, the most distant homologous sequences matching the genes, with 100% coverage and no gaps, were identified using tblastn. Due to the short size of most of the genes, many sequences were too similar for a kA/kS comparison, leaving 175 of 250 annotated genes, 153 of 255 novel genes with annotated homologs, and 116 of 210 novel genes without annotated homologs available for analysis (S3 Table, S4 Table). Of these remaining genes, 12 (4.8%), 12 (4.7%), and 5 (2.4%) genes showed significant selection in the three respective classes using a Holm-Bonferroni multiple comparisons procedure, which was not a significant difference between classes (p=0.335, Fisher's Exact Test). However, only annotated genes exhibited any genes under significant positive selection (5 genes), which was a significant difference among classes (p=0.001, Fisher's Exact Test; Table 1).

Discussion

RIBOseq is a powerful tool to detect translated mRNA

Ribosomal footprinting has been used to detect translation of non-annotated ORFs previously. In eukaryotes, hundreds of non-annotated ORFs show evidence of translation, e.g., in yeast [53], in *Drosophila* [54], in zebrafish [34], in *Arabidopsis* [37], and even in humans [55]. Additionally, the translation of previously annotated ncRNAs was reported frequently [36, 39, 56]. In bacteria, 130 novel genes were detected in *Salmonella* [40] and 72 novel genes were detected in EHEC strain EDL933 [11]. For the latter strain, translation is also reported for a number of RNAs that were previously classified as ncRNA. For instance, the ncRNA *ryhB* encodes a nonamer peptide RyhP

[39]. Although it was not the focus of their study, Jeong et al. [57] report translation signals for 31 annotated ncRNAs in *Streptomyces coelicolor*. Even the well-studied λ-phage with a very small genome of 48.5 kB shows translation of 50 non-annotated ORFs [58].

RIBOseq experiments with eukaryotes allow reading frame determination for individual genes [33, 37, 38]. The reading frame resolution of prokaryotic RIBOseq data is lower such that we cannot determine a reading frame in the RIBOseq signal of single ORFs. This may be caused by bacterial ribosomes being more flexible and incorporating changing numbers of mRNA nucleotides [59]. In addition, the RIBOseq method, formerly developed for eukaryotes, has been adapted for bacteria and footprints of more variable read length are obtained [60]. Furthermore, the composition of ribosomal proteins and rRNAs can be heterogeneous dependent on the growth condition; especially at stress conditions, specialized ribosomes are responsible for the translation of a subset of mRNAs [61, 62]. Putatively, the specialized ribosomes protect an mRNA stretch of deviating length. Recent findings indicate that the usage of a translational inhibitor influences ribosome conformation, which weakens the reading frame signal [63]. For instance chloramphenicol, as used in this study, preferentially arrests translation at positions encoding alanine, serine, or threonine [64] which dilutes the triplet signal. Also, the choice of the ribonuclease used for digestion of mRNA not protected by ribosomes influences RIBOseq results [65]. To minimize the influence of any sequence specificity for a single RNase, we applied a mixture of five RNases (RNase I, MNase, XRN-1, RNase R, and RNase T). Here, we show a reading frame in the sum signal for all genes for the first time in bacteria using conventional RIBOseq. Very recently, the addition of the endonuclease RelE to the ribosome preparation has been reported to improve reading frame determination. The RelE toxin cuts the mRNA within the ribosome very precisely at a specific position in the codon [66]. However, as shown in Fig. 6, under our three conditions a reading frame in the sum signal can be extracted from the data, at least for the group of novel genes that have annotated homologs in other bacterial strains or species.

RIBOseq based evidence for translation of 465 intergenic ORFs

In this study, 465 intergenic ORFs have been detected, which show a clear RIBOseq signal (S2 Table). The average size of the novel-gene encoded proteins is only 50 AA. Standard genome annotation algorithms do usually not predict such very short genes or proteins [14, 16]. In this study, an arbitrary size minimum of 30 AA was applied to restrict the number of ORFs to be investigated and to reduce the possibility of false positives, but even smaller peptides can be functional [39, 40]. Knowledge about the functions of small proteins in bacteria is limited, but small proteins have recently achieved attention (reviewed in [15, 18]). For instance, Baumgartner et al. [67] confirmed five small proteins in *Synechocystis* by Western blot. Neuhaus et al. [11] detected 72 novel small genes in the intergenic regions of the *E. coli* strain EDL933 by evaluating RNAseq and RIBOseq data of a single growth condition (LB, 37°C). Compared to their work, this study on a different EHEC strain achieves a higher sequencing depth and two additional growth conditions including severe stress were investigated. Moreover, translated ORFs were not only selected by an RPKM value threshold, but further conservative thresholds for coverage and RCV were applied. Translation of eleven novel small genes found in EHEC EDL933 by Neuhaus et al. [11] is present in EHEC Sakai and twelve translated ORFs of EHEC Sakai are annotated proteins in EDL933. Vice versa, 28 of the 72 novel EDL933 genes are annotated proteins in strain Sakai.

The 255 translated ORFs with annotated homologs most likely represent protein-coding genes

Blastp analysis revealed that a group of 255 out of the 465 novel ORFs with a clear RIBOseq signal found in this work, have annotated homologs in other bacteria. In addition, many of these 255 genes display predicted protein structures (Fig 8), as well as σ^{70} promoters, and in some cases ρ-independent terminators and SD sites, like annotated short proteins. Even ORFs without these predicted extra features can encode proteins, because those genes could be part of an operon, the promoter could be recognized by an alternative σ-factor [68], termination could be ρ-dependent [69], and translation of leaderless mRNAs occurs [70]. Overall, these novel genes behave similarly in all parameters investigated when compared to 250 short annotated genes of

EHEC Sakai. Both gene groups are transcribed and translated at the same magnitude and the RCV distributions of all growth conditions are comparable. A similar fraction of genes is differentially transcribed and/or translated, when BHI control is compared to BHI COS or LB. Even the directions of up/down regulation compare well (Fig 7). Additionally, active translation is supported by the presence of a reading frame on codon position two for every growth condition in the sum signal caused by codon-wise progression of the ribosome. Furthermore, a machine-learning algorithm trained with short annotated proteins of EHEC EDL933 predicted 88.6% of these genes with annotated homologs as being ‘real’ proteins. Finally, there is no significant difference between the number of genes under selection in this class as compared to either annotated genes or novel genes without annotated homologs. However, unlike annotated genes, for which the majority of selected genes showed evidence of positive selection, all selected genes in this class were under purifying selection. This is not unexpected under the hypothesis of functionality, because purifying selection is the most common form of selection in nature [52], and because this result was obtained despite choosing the most distant homolog. However, it is also likely that ascertainment bias plays a role in this result, as it is probable that more emphasis has historically been placed on the annotation of genes which are shared by more distantly related organisms. This would especially be true if many of the novel genes we identified are orphan genes, since such genes lack distantly related homologs by definition. Therefore, we conclude that these 255 translated intergenic ORFs indeed represent novel small protein-coding genes of EHEC strain Sakai.

Unusual features of the 210 novel genes without annotated homologs

A second group, 210 out of 465 novel genes, had no annotated homologs when using blastp. However, homologs in other bacteria may be present but were missed during annotation of these genomes due to their unusual features. Indeed, a tblastn search confirmed that many non-annotated homologs in the *Escherichia* genus and, in some cases, in farther related species as well, exist (Fig. 5B). The majority of these ORFs were not classified by the machine-learning algorithm to encode ‘real’ proteins. This appears to be more significant and raises the question whether these ORFs indeed code for proteins. The following analysis is based on a comparison between three groups: (i) 250 annotated small genes, (ii) 255 novel small genes with annotated homologs and (iii) the group of 210 ORFs without annotated homologs, which may or may not code for proteins (Table 1). Several arguments support the hypothesis that these ORFs are functional and not residues due to pervasive transcription [29]: first, their expression obviously does not lead to a fitness disadvantage, as in misfolded proteins, which are cytotoxic [71]. Second, a promoter is present upstream of all 210 ORFs, and thirdly, the same fraction of these ORFs is differentially transcribed, compared to both control groups (i) and (ii) (Fig 7). However, these data would fit the hypothesis either that these ORFs represent ncRNA or that they are protein-coding genes. The following observations are in favor of the hypothesis that these novel ORFs are protein-coding genes and not ncRNAs: most significantly, RIBOseq signals, and hence significant RCVs, are in the same order of magnitude as those of short annotated genes, many ORFs without homologs are differentially regulated at the translational level, SD sequences are present upstream of one third of the ORFs, and the number of predicted protein structures is very similar to that of annotated protein-coding genes. Finally, a similar proportion of genes appear to be under selection as among the annotated genes and novel genes with annotated homologs, with the caveat that ascertainment bias has likely favored the detection of genes under purifying selection.

Why, then, does the machine-learning algorithm not recognize these ORFs as protein-coding genes? A first explanation is that the algorithm will only predict sequences as ‘real’, which are within the known parameter space of the training set. Proteins of unknown structure and folds may reside outside the parameter space of ‘established’ proteins and, thus, will fail to be classified as ‘real’ and inevitability binned as ‘pseudo’. The majority of all established proteins belong to a protein family with known secondary structure or which contains characterized domains. But 25% of all protein sequences do not match to any family and, therefore, belong to the ‘dark proteome’ [72]. In pro-karyotes, 13% of all proteins are ‘dark’ [20]. Their properties are different when compared to known proteins: They are shorter, they are often secreted, contain more disulfide bonds, have a lower evolutionary reuse [20], are more disordered, have a different hydrophobic amino acid topology, and have a higher energy [21]. Many of these properties fit well with the PredictProtein data of the proteins encoded by the novel genes without annotated homologs: accordingly, the majority of putative proteins without annotated homologs are very short, are predicted to be secreted, and more often contain disulfide bonds. Thus, these properties render it unlikely that the machine-learning algorithm will predict these unusual proteins correctly.

A second possibility is that the novel genes without annotated homologs may represent very young taxonomically restricted or ‘orphan’ genes. Yomtovian et al. [73] reported that orphan genes of EHEC show an amino acid composition more comparable to random sequences than to annotated genes, since they may not yet have a fully adapted function, which makes it difficult for any annotation program, including our machine-learning algorithm, to distinguish them from scrambled proteins. Also, young genes without annotated homologs are shorter [74], which is true for our data set. Additionally, evolutionary young genes often use uncommon start codons [75], which is also true for our data set. This hypothesis is further supported by the evolutionary distances of the non-annotated homologs detected using tblastn, when comparing the novel genes without annotated homologs to the novel genes with annotated homologs (Fig. 5). The genes with annotated homologs show intact tblastn hits (i.e., ORFs without stop codons) with a significantly greater evolutionary distance compared to the genes without annotated homologs.

In summary, we believe that our data provide evidence supporting the hypothesis that most of these 210 ORFs are evolutionarily young genes coding for proteins with unusual features. The data set may contain some false positives, since in a few cases, ribosome binding of the RNA may exert a regulatory function, comparable to a translation regulating riboswitch instead of translation into protein [76, 77]; however, this will not invalidate our general findings.

Conclusion

This study supports the fact, that, in contrast to earlier beliefs, bacterial genomes are probably under-annotated due to small genes having been overlooked. In *E. coli*

O157:H7 Sakai, at least 465 non-annotated short ORFs are covered with significant RIBOseq reads indicating active translation and the majority of these ORFs show features of protein-coding genes. Since the EHEC Sakai genome harbors about 5200 annotated protein-coding genes, these additional genes would significantly increase the number of protein-coding genes in this bacterium. Obviously, much further work is required for functional characterization of the novel genes. It would not be surprising if other bacterial genomes also harbor many overlooked short genes in their intergenic regions, which could be investigated by combined RNAseq and RIBOseq. In addition, the high-throughput discovery of small proteins in proteome analysis requires modified or improved methods since these proteins likely escape attention with most currently available methods [17, 78, 79]. Our study supports the notion that it is advisable to improve genome annotation algorithms in order to reduce bias against annotation of short genes [16, 75].

Material and Methods

Transcriptome and translatome sequencing

Strand-specific RNAseq and RIBOseq of *Escherichia coli* O157:H7 Sakai (GenBank accession number BA000007.2 and RefSeq accession NC_002695.1, version from February 2014) [1] were performed at three different growth conditions in two biological replicates each. An overnight culture of EHEC was inoculated 1:100 in lysogeny broth (LB medium) and incubated at 37°C and 150 rpm until an OD_{600} of 0.4 was reached. Additionally, two conditions using brain-heart infusion broth (BHI; Merck KGaA) were investigated. For the BHI control condition, an overnight culture of EHEC was inoculated 1:100 and incubated at 37°C and 150 rpm until an OD_{600} of 0.1 was reached. For the stress condition of combined cold and osmotic stress (COS), 4% NaCl were added to the BHI medium and incubation was performed at 14°C until an OD_{600} of 0.1 was reached.

RNAseq was performed as described by Landstorfer et al. [8] for the Illumina system. For ribosomal footprinting, the method published by Ingolia et al. [31] was adapted to bacteria as described [11] with the following further modifications: mRNA not protected by ribosomes was digested with a mixture of five RNases to exclude sequence specificity. Buffer NEB 4 plus 1 mM $CaCl_2$ was added to 1 ml cell extract and the solution was incubated for 1 h at RT with 250 U MNase (Roche), 5 U XRN-1 (NEB), 250 U RNase I (Thermo Fisher Scientific), 50 U RNase R (Biozym) and 12 U RNase T (NEB). The monosome fraction was harvested by sucrose density gradient centrifugation and unprotected mRNA digestion was repeated once. For the LB condition, rRNA was depleted using the MICROBExpress kit (Thermo Fisher Scientific) and for the BHI conditions rRNA depletion was performed using the RiboZero kit for Gram-negative bacteria (Illumina). All libraries were prepared using the TruSeq Small RNA Sample Preparation Kit (Illumina) and sequenced on a HiSeq 2500 machine according to the manufacturer. The sequencing raw data is available at the Sequence Read Archive (SRA, NCBI) under the accession SRP113660.

Read mapping and RCV calculation

For processing and mapping of the sequencing raw data, the Galaxy platform was used [80] as described [11]. The data were visualized using BamView [81] implemented in Artemis 16.0 [41]. The RPKM values for all intergenic non-annotated ORFs in EHEC which would encode a peptide of ≥ 30 AA (~12,000 ORFs) were calculated in R, whereas reads mapping to rRNA or tRNA were excluded [82]. Besides the canonical DTG start codons, the rare start codons CTG, ATT, ATA and ATC were allowed according to genetic code table 11 (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi). The ratio of RPKM translatome over RPKM transcriptome gives the ribosomal coverage value (RCV), which is a measure for the translatability of a certain ORF [39]. Novel gene candidates had to fulfill the following criteria for at least one growth condition in both biological replicates to be considered translated: RPKM translatome at least 1 read per million mapped reads, coverage translatome ≥ 0.5 and RCV ≥ 0.25. To exclude false positives, all novel gene candidates were manually inspected in Artemis.

Reading frame determination

Adapter removal and quality trimming were performed using AdapterRemoval v2.1.7 [83] and non-rRNA reads longer than 18 bp were extracted using sortMeRNA v2.0 [84].

Extracted reads were mapped to previously annotated genes, novel genes with annotated homologs and novel genes without annotated homologs, in *Escherichia coli* O157:H7 Sakai using Vsearch v2.1.2 [85]. The reading frame of the 5' end of each mapped read of length 20 bp (maximum of read length distribution) was determined using a custom script (S1 File), which counts the number of 5' ends for the three codon positions and sums the values for the three gene groups (annotated genes, novel genes with annotated homologs, and novel genes without annotated homologs).

Differential gene expression

The condition 'BHI at 37°C' was used as the reference data set and for the LB and BHI COS conditions significant changes on transcriptional and translational level were determined. Read counts were normalized to the smallest library and differential expression was analyzed by an exact test implemented in the Bioconductor package edgeR (version 3.2.4) [44]. A p-value ≤ 0.05 and a false discovery rate (FDR) ≤ 0.1 were used to delineate significant expression changes.

Prediction of σ^{70} promoters

The region 300 bp upstream of the start codon was searched for the presence and strength of a σ^{70} promoter with the program BPROM (Softberry [86]). It searches for the -35 and -10 consensus motif and recognition sequences for transcription factors. With this data, an LDF score (linear discriminant function) is calculated, whereupon increasing values indicate growing promoter strength. An LDF score of 0.2 gives the threshold for promoter prediction with 80% accuracy and specificity.

Prediction of ρ-independent terminators

The region 300 bp downstream of the stop codon was searched for the presence and strength of a ρ-independent terminator using FindTerm (Softberry [86]). This program searches thymidine-rich regions, and calculates the energy of possible terminator structures. Low energy values indicate strong terminators.

Prediction of Shine-Dalgarno sequence

The region 30 bp upstream of the start codon was examined for the presence of a Shine-Dalgarno sequence (optimum uaAGGAGGu). ΔG° was calculated according to Ma et al. [51] with a threshold of Δ -2.9 kcal/mol.

Calculation of kA/kS

The most distantly related homologs of the short annotated genes and the novel genes were determined with tblastn by selecting the hit with the highest e-value which still has 100% coverage and no gaps. In case the sequence pairs were too similar, meaningful kA/kS calculation was not possible. The ratio of synonymous to non-synonymous substitutions between those gene pairs was computed using the KaKs_Calculator 2.0 [87]. The "bacterial and plant plastid code" was selected and the method model selection (MS) was used. The ORF is assumed to be under positive selection when kA/kS is significantly greater than 1 and under purifying selection when kA/kS is significantly less than 1. Significance was determined using a Holm-Bonferroni multiple comparisons procedure with respect to the family, an error rate of 0.05. A Fisher's Exact Test was performed in R version 3.3.2. Unless otherwise noted, all p-values refer to two-sided tests.

Detection of annotated homologs

Novel gene sequences were translated into the corresponding proteins sequences, which were used to query the GenBank database using blastp with default parameters [88]. An e-value cutoff of 10^{-3} was applied.

Sequence Conservation

Sequences of the novel genes were aligned against the full RefSeq genomic database downloaded on 5 April 2017, using a tblastn search in the local BLAST utilities 2.6.0+ from the NCBI [89] with a maximum e-value of 0.001. The putative homologues were extracted from the database and those without stop codons were retained as 'intact'. The amino acid similarity of each intact subject sequence with the query ORF was calculated using the Needle-Wunsch algorithm "Needleall" from EMBOSS [90]. The *Achromobacter* sp. ATCC35328 sequences with names beginning NZ_CYUC010 were removed from the analysis, due to abnormally high similarity with *E. coli* for a very large number of genes. Thus, we assumed this species to be mislabeled. To map the results gained using NCBI databases to the SILVA taxonomy, hits were conflated to genus level, which allowed inclusion of over 90% of genera with hits in each case. To obtain approximate relative evolutionary distances, the average distance from EHEC Sakai to the last common ancestor with each genus was calculated from the 16S rRNA SILVA reference NR99 guide tree [91], release 128, using Newick Utilities [92]. A custom shell script for these tasks, ORFage, was used (S2 File). A similar pipeline was used to check the conservation of intergenic sequences upstream and downstream of the novel genes. For the upstream regions, the sequences between the stop codon of the nearest annotated gene upstream of the start codon of the novel gene was taken. Similarly, for downstream regions, the sequence between the stop codon of the novel gene and the start codon of the next annotated gene downstream was taken. Some of the regions were too short to obtain (meaningful) tblastn hits and were excluded. Further regions were excluded, when containing another of the novel genes before an annotated gene was reached. One downstream sequence was abnormally long and could not be processed (tblastn search > 1 day), hence, this region was also excluded. Within the upstream and downstream sequences, stop codons were allowed. The shell script used for preparation of the intergenic sequences including the use of ENTREZ DIRECT [93] is included in S3 File.

Predicted protein characteristics

The amino acid sequences encoded in the 250 short annotated genes and the 465 novel genes were submitted to PredictProtein [46] using default parameters. This software predicts structural and functional features of the putative proteins. The results of PROFphd (secondary structure) [94], TMSEG (transmembrane helices) [95], DISULFIND (disulfide bonds) [96], UCON (disordered regions) [47] and LocTree3 (subcellular localization) [97] were analyzed in further detail.

Machine learning based protein recognition

A machine-learning algorithm, as described by Neuhaus et al. [11], was used to classify the novel proteins based on predicted protein parameters. Briefly, about 279 short annotated proteins were picked from EHEC EDL933 and these sequences shuffled 100-times. All sequences, natural and shuffled, were submitted to a PredictProtein analysis [45, 46]. The machine-learning algorithm was trained using the predicted parameters for the annotated proteins (positive control) and their shuffled counterparts (negative control). Both strains, EDL933 and Sakai are very closely related to each other [98] and, thus, the trained algorithm was used here, as well. We not only examined the protein sequences of the novel genes in Sakai, but also shuffled those 10-times to detect false positives.

Localization of novel genes

Visualization of the gene's localization was created using Circos [99].

Acknowledgement

We thank Svenja Simon for providing her R scripts for RPKM value and coverage calculation.

References

1. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, et al. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res. 2001;8(1):11-22. PubMed PMID: 11258796.
2. Lim JY, Yoon J, Hovde CJ. A brief overview of *Escherichia coli* O157:H7 and its plasmid O157. J Microbiol Biotechnol. 2010;20(1):5-14. PubMed PMID: 20134227; PubMed Central PMCID: PMC3645889.
3. Lewis SB, Cook V, Tighe R, Schuller S. Enterohemorrhagic *Escherichia coli* colonization of human colonic epithelium *in vitro* and *ex vivo*. Infect Immun. 2015;83(3):942-9. doi: 10.1128/IAI.02928-14. PubMed PMID: 25534942; PubMed Central PMCID: PMC4333473.
4. Ma J, Ibekwe AM, Yi X, Wang H, Yamazaki A, Crowley DE, et al. Persistence of *Escherichia coli* O157:H7 and its mutants in soils. PLoS One. 2011;6(8):e23191. doi: 10.1371/journal.pone.0023191. PubMed PMID: 21826238; PubMed Central PMCID: PMC3149627.
5. Hou Z, Fink RC, Sugawara M, Diez-Gonzalez F, Sadowsky MJ. Transcriptional and functional responses of *Escherichia coli* O157:H7 growing in the lettuce rhizoplane. Food microbiology. 2013;35(2):136-42. doi: 10.1016/j.fm.2013.03.002. PubMed PMID: 23664265.
6. Castro BG, Souza MM, Regua-Mangia AH, Bittencourt AJ. Occurrence of Shiga-toxigenic *Escherichia coli* in *Stomoxys calcitrans* (Diptera: Muscidae). Rev Bras Parasitol Vet. 2013;22(2):318-21. doi: 10.1590/S1984-29612013000200052. PubMed PMID: 23856725.
7. Naylor SW, Low JC, Besser TE, Mahajan A, Gunn GJ, Pearce MC, et al. Lymphoid follicle-dense mucosa at the terminal rectum is the principal site of colonization of enterohemorrhagic *Escherichia coli* O157:H7 in the bovine host. Infect Immun. 2003;71(3):1505-12. PubMed PMID: 12595469; PubMed Central PMCID: PMC148874.

8. Landstorfer R, Simon S, Schober S, Keim D, Scherer S, Neuhaus K. 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. 2014;15:353. doi: 10.1186/1471-2164-15-353. PubMed PMID: 24885796; PubMed Central PMCID: PMC4048457.

9. Trachtman H, Austin C, Lewinski M, Stahl RA. Renal and neurological involvement in typical Shiga toxin-associated HUS. Nat Rev Nephrol. 2012;8(11):658-69. doi: 10.1038/nrneph.2012.196. PubMed PMID: 22986362.

10.Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. Small membrane proteins found by comparative genomics and ribosome binding site models. Mol Microbiol. 2008;70(6):1487-501. PubMed PMID: 19121005.

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

12.Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007;23(6):673-9. PubMed PMID: 17237039.

13.Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75. doi: 10.1186/1471-2164-9-75. PubMed PMID: 18261238; PubMed Central PMCID: PMC2265698.

14.Boekhorst J, Wilson G, Siezen RJ. Searching in microbial genomes for encoded small proteins. Microb Biotechnol. 2011;4(3):308-13. doi: 10.1111/j.1751-7915.2011.00261.x. PubMed PMID: 21518296.

15.Storz G, Wolf YI, Ramamurthi KS. Small proteins can no longer be ignored. Annu Rev Biochem. 2014;83:753-77. doi: 10.1146/annurev-biochem-070611-102400. PubMed PMID: 24606146.

16.Warren AS, Archuleta J, Feng WC, Setubal JC. Missing genes in the annotation of prokaryotic genomes. BMC Bioinformatics. 2010;11:131. PubMed PMID: 20230630.

17.Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol. 2013;9(1):59-64. doi: 10.1038/nchembio.1120. PubMed PMID: 23160002; PubMed Central PMCID: PMC3625679.

18.Kemp G, Cymer F. Small membrane proteins–elucidating the function of the needle in the haystack. Biol Chem. 2014;395(12):1365-77.

19.Brylinski M. Exploring the "dark matter" of a mammalian proteome by protein structure and function modeling. Proteome Sci. 2013;11(1):47. doi: 10.1186/1477-5956-11-47. PubMed PMID: 24321360; PubMed Central PMCID: PMC3866606.

20.Perdigão N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, et al. Unexpected features of the dark proteome. Proc Natl Acad Sci U S A. 2015;112(52):15898-903.

21.Bitard-Feildel T, Callebaut I. Exploring the dark foldable proteome by considering hydrophobic amino acids topology. Scientific reports. 2017;7:41425. doi: 10.1038/srep41425. PubMed PMID: 28134276; PubMed Central PMCID: PMC5278394.

22.Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, et al. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. PLoS Genet. 2009;5(7):e1000569. PubMed PMID: 19609351.

23.Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature. 2010;464(7286):250-5. doi: 10.1038/nature08756. PubMed PMID: 20164839.

24.Flaherty BL, Van Nieuwerburgh F, Head SR, Golden JW. Directional RNA deep sequencing sheds new light on the transcriptional response of *Anabaena* sp. strain PCC 7120 to combined-nitrogen deprivation. BMC Genomics. 2011;12:332. PubMed PMID: 21711558.

25.Neme R, Tautz D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence. eLife. 2016;5:e09977. doi: 10.7554/eLife.09977. PubMed PMID: 26836309; PubMed Central PMCID: PMC4829534.

26.Lasa I, Toledo-Arana A, Dobin A, Villanueva M, de los Mozos IR, Vergara-Irigaray M, et al. Genome-wide antisense transcription drives mRNA processing in bacteria. Proc Natl Acad Sci U S A. 2011;108(50):20172-7. doi: 10.1073/pnas.1113521108. PubMed PMID: 22123973; PubMed Central PMCID: PMC3250193.

27.Dornenburg JE, Devita AM, Palumbo MJ, Wade JT. Widespread Antisense Transcription in *Escherichia coli*. mBio. 2010;1(1). PubMed PMID: 20689751.

28.Lin YF, A DR, Guan S, Mamanova L, McDowall KJ. A combination of improved differential and global RNA-seq reveals pervasive transcription initiation and events in all stages of the life-cycle of functional RNAs in *Propionibacterium acnes*, a major contributor to wide-spread human disease. BMC Genomics. 2013;14:620. doi: 10.1186/1471-2164-14-620. PubMed PMID: 24034785; PubMed Central PMCID: PMC3848588.

29.Wade JT, Grainger DC. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol. 2014;12(9):647-53. doi: 10.1038/nrmicro3316. PubMed PMID: 25069631.

30.Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013;5(3):578-90.

31.Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324(5924):218-23. PubMed PMID: 19213877.

32.Aeschimann F, Xiong J, Arnold A, Dieterich C, Grosshans H. Transcriptome-wide measurement of ribosomal occupancy by ribosome profiling. Methods. 2015. doi: 10.1016/j.ymeth.2015.06.013. PubMed PMID: 26102273.

33.Smith JE, Alvarez-Dominguez JR, Kline N, Huynh NJ, Geisler S, Hu W, et al. Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. Cell Rep. 2014;7(6):1858-66. doi: 10.1016/j.celrep.2014.05.023. PubMed PMID: 24931603; PubMed Central PMCID: PMC4105149.

34.Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. 2014;33(9):981-93. doi: 10.1002/embj.201488411. PubMed PMID: 24705786.

35.Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. 2014;8(5):1365-79. doi: 10.1016/j.celrep.2014.07.045. PubMed PMID: 25159147.

36.Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. eLife. 2014;3:e03523. doi: 10.7554/eLife.03523. PubMed PMID: 25233276; PubMed Central PMCID: PMC4359382.

37.Hsu PY, Calviello L, Wu HL, Li FW, Rothfels CJ, Ohler U, et al. Super-resolution ribosome profiling reveals unannotated translation events in *Arabidopsis*. Proc Natl Acad Sci U S A. 2016. doi: 10.1073/pnas.1614788113. PubMed PMID: 27791167.

38.Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, et al. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods. 2016;13(2):165-70. doi: 10.1038/nmeth.3688. PubMed PMID: 26657557.

39.Neuhaus K, Landstorfer R, Simon S, Schober S, Wright PR, Smith C, et al. 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. 2017;18(1):216. doi: 10.1186/s12864-017-3586-9. PubMed PMID: 28245801; PubMed Central PMCID: PMC5331693.

40.Baek J, Lee J, Yoon K, Lee H. Identification of Unannotated Small Genes in *Salmonella*. G3. 2017;7(3):983-9. doi: 10.1534/g3.116.036939. PubMed PMID: 28122954; PubMed Central PMCID: PMC5345727.

41.Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics. 2000;16(10):944-5. PubMed PMID: 11120685.

42.Latif H, Li HJ, Charusanti P, Palsson BØ, Aziz RK. A gapless, unambiguous genome sequence of the enterohemorrhagic *Escherichia coli* O157: H7 strain EDL933. Genome Announc. 2014;2(4):e00821-14.

43.Dunning Hotopp JC, Clark ME, Oliveira DC, Foster JM, Fischer P, Munoz Torres MC, et al. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science. 2007;317(5845):1753-6. Epub 2007/09/01. doi: 10.1126/science.1142490. PubMed PMID: 17761848.

44.Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26(1):139-40. PubMed PMID: 19910308.

45.Rost B, Yachdav G, Liu J. The predictprotein server. Nucleic Acids Res. 2004;32(suppl 2):W321-W6.

46.Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, et al. PredictProtein--an open resource for online prediction of protein structural and functional features. Nucleic Acids Res. 2014;42(Web Server issue):W337-43. doi: 10.1093/nar/gku366. PubMed PMID: 24799431; PubMed Central PMCID: PMC4086098.

47.Schlessinger A, Punta M, Rost B. Natively unstructured regions in proteins identified from contact predictions. Bioinformatics. 2007;23(18):2376-84.

48.Browning DF, Busby SJ. The regulation of bacterial transcription initiation. Nat Rev Microbiol. 2004;2(1):57-65. doi: 10.1038/nrmicro787. PubMed PMID: 15035009.

49.Wilson KS, von Hippel PH. Transcription termination at intrinsic terminators: the role of the RNA hairpin. Proc Natl Acad Sci U S A. 1995;92(19):8793-7. PubMed PMID: 7568019; PubMed Central PMCID: PMC41053.

50.Vimberg V, Tats A, Remm M, Tenson T. Translation initiation region sequence preferences in *Escherichia coli*. BMC Mol Biol. 2007;8:100. doi: 10.1186/1471-2199-8-100. PubMed PMID: 17973990; PubMed Central PMCID: PMC2176067.

51.Ma J, Campbell A, Karlin S. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J Bacteriol. 2002;184(20):5733-45. PubMed PMID: 12270832.

52.Hughes AL. Adaptive Evolution of Genes and Genomes. Oxford University Press, New York. 1999.

53.Wilson BA, Masel J. Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol Evol. 2011;3:1245-52. doi: 10.1093/gbe/evr099. PubMed PMID: 21948395; PubMed Central PMCID: PMC3209793.

54.Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MA, Brocard M, et al. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. eLife. 2014;3:e03528. doi: 10.7554/eLife.03528. PubMed PMID: 25144939; PubMed Central PMCID: PMC4359375.

55.Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. eLife. 2015;4:e08890. doi: 10.7554/eLife.08890. PubMed PMID: 26687005; PubMed Central PMCID: PMC4739776.

56.Carlevaro-Fita J, Rahim A, Guigo R, Vardy LA, Johnson R. Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. RNA. 2016;22(6):867-82. doi: 10.1261/rna.053561.115. PubMed PMID: 27090285; PubMed Central PMCID: PMC4878613.

57.Jeong Y, Kim JN, Kim MW, Bucca G, Cho S, Yoon YJ, et al. The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3(2). Nat Commun. 2016;7:11605. doi: 10.1038/ncomms11605. PubMed PMID: 27251447; PubMed Central PMCID: PMC4895711.

58.Liu X, Jiang H, Gu Z, Roberts JW. High-resolution view of bacteriophage lambda gene expression by ribosome profiling. Proc Natl Acad Sci U S A. 2013;110(29):11928-33. doi: 10.1073/pnas.1309739110. PubMed PMID: 23812753; PubMed Central PMCID: PMC3718152.

59.O'Connor PB, Li GW, Weissman JS, Atkins JF, Baranov PV. rRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome profiling experiments. Bioinformatics. 2013;29(12):1488-91. doi: 10.1093/bioinformatics/btt184. PubMed PMID: 23603333; PubMed Central PMCID: PMC3673220.

60.Mohammad F, Woolstenhulme CJ, Green R, Buskirk AR. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. Cell Rep. 2016;14(4):686-94. doi: 10.1016/j.celrep.2015.12.073. PubMed PMID: 26776510; PubMed Central PMCID: PMC4835026.

61.Xue S, Barna M. Specialized ribosomes: a new frontier in gene regulation and organismal biology. Nat Rev Mol Cell Biol. 2012;13(6):355-69. doi: 10.1038/nrm3359. PubMed PMID: 22617470; PubMed Central PMCID: PMC4039366.

62.Byrgazov K, Vesper O, Moll I. Ribosome heterogeneity: another level of complexity in bacterial translation regulation. Curr Opin Microbiol. 2013;16(2):133-9. doi: 10.1016/j.mib.2013.01.009. PubMed PMID: 23415603; PubMed Central PMCID: PMC3653068.

63.Gerashchenko MV, Gladyshev VN. Translation inhibitors cause abnormalities in ribosome profiling experiments. Nucleic Acids Res. 2014;42(17):e134. doi: 10.1093/nar/gku671. PubMed PMID: 25056308; PubMed Central PMCID: PMC4176156.

64.Marks J, Kannan K, Roncase EJ, Klepacki D, Kefi A, Orelle C, et al. Context-specific inhibition of translation by ribosomal antibiotics targeting the peptidyl transferase center. Proc Natl Acad Sci U S A. 2016;113(43):12150-5. doi: 10.1073/pnas.1613055113. PubMed PMID: 27791002; PubMed Central PMCID: PMC5086994.
65.Gerashchenko MV, Gladyshev VN. Ribonuclease selection for ribosome profiling. Nucleic Acids Res. 2017;45(2):e6. doi: 10.1093/nar/gkw822. PubMed PMID: 27638886.
66.Hwang JY, Buskirk AR. A ribosome profiling study of mRNA cleavage by the endonuclease RelE. Nucleic Acids Res. 2017;45(1):327-36. doi: 10.1093/nar/gkw944. PubMed PMID: 27924019; PubMed Central PMCID: PMC5224514.
67.Baumgartner D, Kopf M, Klahn S, Steglich C, Hess WR. Small proteins in cyanobacteria provide a paradigm for the functional analysis of the bacterial micro-proteome. BMC Microbiol. 2016;16(1):285. doi: 10.1186/s12866-016-0896-z. PubMed PMID: 27894276; PubMed Central PMCID: PMC5126843.
68.Cho BK, Kim D, Knight EM, Zengler K, Palsson BO. Genome-scale reconstruction of the sigma factor network in *Escherichia coli*: topology and functional states. BMC Biol. 2014;12:4. doi: 10.1186/1741-7007-12-4. PubMed PMID: 24461193; PubMed Central PMCID: PMC3923258.
69.Banerjee S, Chalissery J, Bandey I, Sen R. Rho-dependent transcription termination: more questions than answers. J Microbiol. 2006;44(1):11-22. Epub 2006/03/24. doi: 2342 [pii]. PubMed PMID: 16554712.
70.Zheng X, Hu G-Q, She Z-S, Zhu H. Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. BMC Genomics. 2011;12(1):361.
71.Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 2008;134(2):341-52. doi: 10.1016/j.cell.2008.05.042. PubMed PMID: 18662548; PubMed Central PMCID: PMC2696314.
72.Levitt M. Nature of the protein universe. Proc Natl Acad Sci U S A. 2009;106(27):11079-84. doi: 10.1073/pnas.0905029106. PubMed PMID: 19541617; PubMed Central PMCID: PMC2698892.
73.Yomtovian I, Teerakulkittipong N, Lee B, Moult J, Unger R. Composition bias and the origin of ORFan genes. Bioinformatics. 2010;26(8):996-9. doi: 10.1093/bioinformatics/btq093. PubMed PMID: 20231229; PubMed Central PMCID: PMC2853687.
74.Tatarinova TV, Lysnyansky I, Nikolsky YV, Bolshoy A. The mysterious orphans of *Mycoplasmataceae*. Biol Direct. 2016;11(1):1.
75.Oheigeartaigh SS, Armisen D, Byrne KP, Wolfe KH. SearchDOGS bacteria, software that provides automated identification of potentially missed genes in annotated bacterial genomes. J Bacteriol. 2014;196(11):2030-42. doi: 10.1128/JB.01368-13. PubMed PMID: 24659774; PubMed Central PMCID: PMC4010983.
76.Hücker SM, Simon S, Scherer S, Neuhaus K. 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. 2017;364(2). doi: 10.1093/femsle/fnw262. PubMed PMID: 27856567.
77.Dar D, Shamir M, Mellin JR, Koutero M, Stern-Ginossar N, Cossart P, et al. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. Science. 2016;352(6282):aad9822. doi: 10.1126/science.aad9822. PubMed PMID: 27120414.
78.Zur H, Aviner R, Tuller T. Complementary Post Transcriptional Regulatory Information is Detected by PUNCH-P and Ribosome Profiling. Sci Rep. 2016;6.
79.Cassidy L, Prasse D, Linke D, Schmitz RA, Tholey A. Combination of Bottom-up 2D-LC-MS and Semi-top-down GelFree-LC-MS Enhances Coverage of Proteome and Low Molecular Weight Short Open Reading Frame Encoded Peptides of the Archaeon *Methanosarcina mazei*. J Proteome Res. 2016;15(10):3773-83. doi: 10.1021/acs.jproteome.6b00569. PubMed PMID: 27557128.
80.Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016;44(W1):W3-W10. doi: 10.1093/nar/gkw343. PubMed PMID: 27137889; PubMed Central PMCID: PMC4987906.
81.Carver T, Bohme U, Otto TD, Parkhill J, Berriman M. BamView: viewing mapped read alignment data in the context of the reference sequence. Bioinformatics. 2010;26(5):676-7. PubMed PMID: 20071372.

82.Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621-8. doi: 10.1038/nmeth.1226. PubMed PMID: 18516045.
83.Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016;9:88. doi: 10.1186/s13104-016-1900-2. PubMed PMID: 26868221; PubMed Central PMCID: PMCPMC4751634.
84.Kopylova E, Noe L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28(24):3211-7. doi: 10.1093/bioinformatics/bts611. PubMed PMID: 23071270.
85.Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584. doi: 10.7717/peerj.2584. PubMed PMID: 27781170; PubMed Central PMCID: PMCPMC5075697.
86.Solovyev VV, Tatarinova TV. Towards the integration of genomics, epidemiological and clinical data. Genome Med. 2011;3(7):48. doi: 10.1186/gm264. PubMed PMID: 21867574; PubMed Central PMCID: PMC3221549.
87.Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics. 2010;8(1):77-80.
88.Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.
89.Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. Epub 2009/12/17. doi: 10.1186/1471-2105-10-421. PubMed PMID: 20003500; PubMed Central PMCID: PMCPMC2803857.
90.Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16(6):276-7. Epub 2000/05/29. PubMed PMID: 10827456.
91.Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(D1):D590-D6.
92.Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 2010;26(13):1669-70. Epub 2010/05/18. doi: 10.1093/bioinformatics/btq243. PubMed PMID: 20472542; PubMed Central PMCID: PMCPMC2887050.
93.Kans J. Entrez Direct: E-utilities on the UNIX Command Line: Bethesda (MD): National Center for Biotechnology Information; 2013.
94.Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins. 1994;19(1):55-72. PubMed PMID: 8066087.
95.Bernhofer M, Kloppmann E, Reeb J, Rost B. TMSEG: Novel prediction of transmembrane helices. Proteins. 2016;84(11):1706-16. doi: 10.1002/prot.25155. PubMed PMID: 27566436; PubMed Central PMCID: PMC5073023.
96.Ceroni A, Passerini A, Vullo A, Frasconi P. DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. Nucleic Acids Res. 2006;34(suppl 2):W177-W81.
97.Goldberg T, Hecht M, Hamp T, Karl T, Yachdav G, Ahmed N, et al. LocTree3 prediction of localization. Nucleic Acids Res. 2014;42(Web Server issue):W350-5. doi: 10.1093/nar/gku396. PubMed PMID: 24848019.
98.Zhang W, Qi W, Albert TJ, Motiwala AS, Alland D, Hyytia-Trees EK, et al. Probing genomic diversity and evolution of *Escherichia coli* O157 by single nucleotide polymorphisms. Genome Res. 2006;16(6):757-67. doi: 10.1101/gr.4759706. PubMed PMID: 16606700; PubMed Central PMCID: PMC1473186.
99.Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639-45. doi: 10.1101/gr.092759.109. PubMed PMID: 19541911; PubMed Central PMCID: PMC2752132.

2.7 Detection of 380 transcribed and translated antiparallel overlapping novel genes

Novel genes cannot only be located in intergenic regions, but they can overlap to annotated protein-coding genes in different reading frames as well. In this study, only antiparallel overlapping genes were investigated, because for same-strand overlapping genes it is most of the time not possible to discriminate from which ORF the transcription or translation signal is caused, especially for embedded short ORFs within annotated longer ORFs. According to Hücker et al. (2017a) every ORF encoding at least 30 AAs and overlapping antisense at least 93 bp to annotated genes was determined and investigated regarding its transcription or translation at three tested growth conditions mentioned above.

2.7.1 Three-hundred-and-eighty antiparallel overlapping ORFs show evidence of translation

At three investigated growth conditions combined, 380 antisense ORFs are considered to be translated. All these thresholds have to be fulfilled in at least one biological replicate: i) the RPKM value for the translatome is at least one read per million sequenced reads, ii) the RCV value is at least 0.25, iii) the ORF coverage by RIBOseq reads is ≥ 50%, and iv) individual visual inspection in the Artemis genome browser (Rutherford et al., 2000) confirmed that the translation signal indeed was caused by the specific ORF and not by neighboring annotated genes. Overall, the EHEC Sakai genome contains 38,128 antisense ORFs of at least 93 bp. Of those, 754 fulfilled the upper three thresholds mentioned, and after visual inspection, 380 ORFs remained representing putative novel OLGs. The putative antisense genes were consecutively numbered after their appearance in the genome, and named OLGECs###. The genome position, ORF length, and expression level at every growth condition for the 380 OLGs are listed in Supplementary Table S2. This data set does not contain ORFs from genomic regions of prophage origin, because OLGs are an already accepted feature for viral genomes. The majority of the 380 OLGs is translated at optimal growth conditions: 39% of OLGs are translated only in LB medium and 27% are translated in both LB and BHI at 37°C (Figure 11). Only five OLGs are translated at all growth conditions. A similar distribution was observed for the putative intergenic novel genes (Hücker et al., 2017a). In case the thresholds for RPKM value translatome, RCV, and coverage should be met in both biological replicates of at least one condition, then 142 OLGs remain. All of them are translated at the optimal conditions, and only seven OLGs are translated at BHI stress as well.

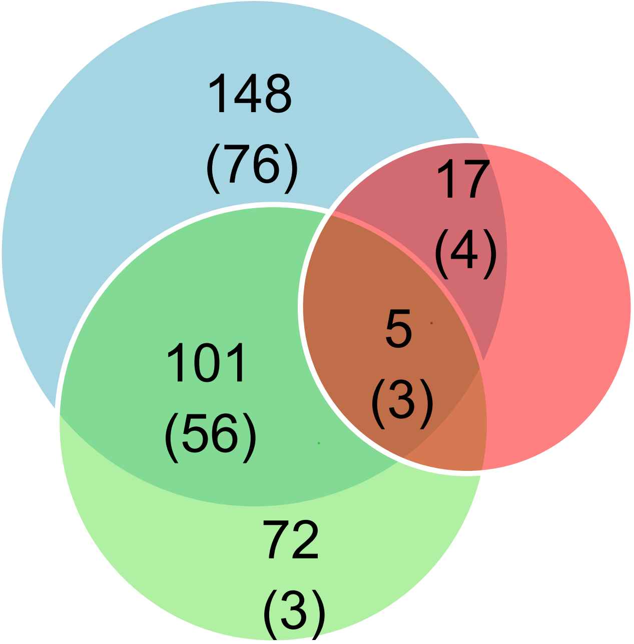


Figure 11: Venn-Diagram showing the number of translated OLGs for the three growth conditions. The blue circle represents the condition LB at 37°C, the green circle BHI at 37°C, and the red circle BHI + 4% NaCl at 14°C. The numbers of translated OLGs for every condition are given. How many OLGs fulfill all criteria in both biological repli-cates is indicated in brackets.

Previous research showed that also rare start codons are used in case of novel translated ORFs (Neuhaus et al., 2017). Since the correct start codon of the OLGs is undetermined, the farthest upstream start codon was used, including the rare start codons CTG, ATT, ATA, and ATC according to the genetic code table 11 (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi). Thus, only 15% of the translated OLGs described here have an ATG start codon, 24% have either a GTG or TTG start codon, and 61% have a rare start codon. In some cases, the true start codon might be located downstream of the rare start codon. However, in other cases, the translated OLGs do not possess a canonical start codon further downstream, indicating that translation initiation at rare codons is more frequent than presumed. In addition, this is confirmed by mammalian RIBOseq data: the software ORF-RATER predicts 575 novel genes using rare start codons, usually encoding N-terminally extended proteins (Fields et al., 2015). Up to 30% of short upstream ORFs seem to use rare start codons (Ji et al., 2015).

2.7.2 Reproducibility of RNAseq and RIBOseq signals of overlapping ORFs

Similar as for the annotated genes (see 2.1.1), the reproducibility of RNAseq and RIBOseq data for the overlapping ORFs of the biological replicates was investigated. RPKM values of the 380 putative novel OLGs were plotted against each other for every growth condition. In LB at 37°C, the reproducibility is excellent with a correlation of R=0.98 (Figure 12), which is as good as the correlation of biological replicates for the annotated genes (Figure 6). However, in BHI at 37°C the annotated genes show a better correlation than the 380 OLGs: RNAseq shows a correlation of R=0.83 and RIBOseq of 0.65, respectively (data not shown). The lower correlation might be explained by the lower expression level of OLGs, thus, increased noise. Moreover, due to the composition of LB and BHI medium, the latter is expected to have larger batch-to-batch variability. At BHI stress, only the RNAseq experiment shows acceptable correlation of R=0.73. For the RIBOseq data, correlation almost disappears (R=0.1). This occurs, because the majority of the 380 OLGs is not translated at the stress condition (Figure 11 + Supplementary Table S2). Similarly, RPKM values of many (annotated) genes are below background levels, and small changes without biological meaning greatly influence the correlation.

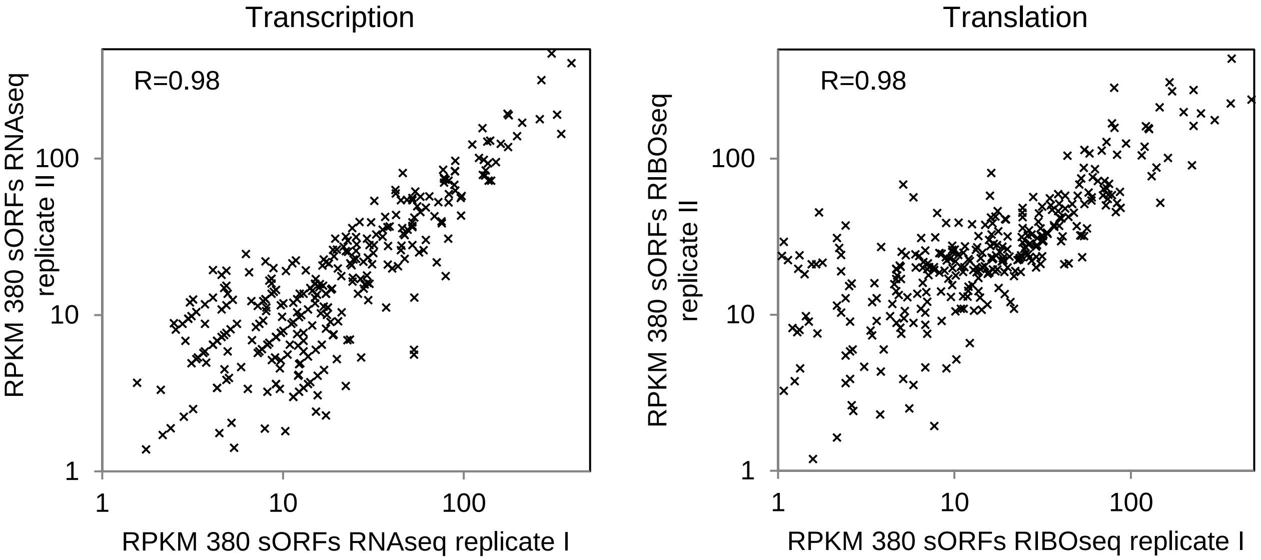


Figure 12: Correlation of the two RNAseq and RIBOseq biological replicates for the 380 putative OLGs at the condition LB at 37°C. RPKM values of the 380 OLGs were plotted against each other in logarithmic scale and Pearson correlation was calculated.

2.7.3 Correlation of transcription and translation for overlapping ORFs

The correlation of RNAseq to RIBOseq data for the 380 translated OLGs was investigated (see 2.2). Data from growth in LB at 37°C leads to the best correlation of R=0.64 (Figure 13), which is similar to the correlation of the annotated genes (Figure 8). However, for the BHI conditions, the annotated genes show a better correlation than the translated OLGs. The low correlation in BHI COS is explained by the missing correlation of the two RIBOseq replicates due to very low expression levels (see 2.7.2), which leads to high deviations of the RPKM mean value. Many OLGs are translated only to a very low extent at this condition, whereas they are still transcribed.

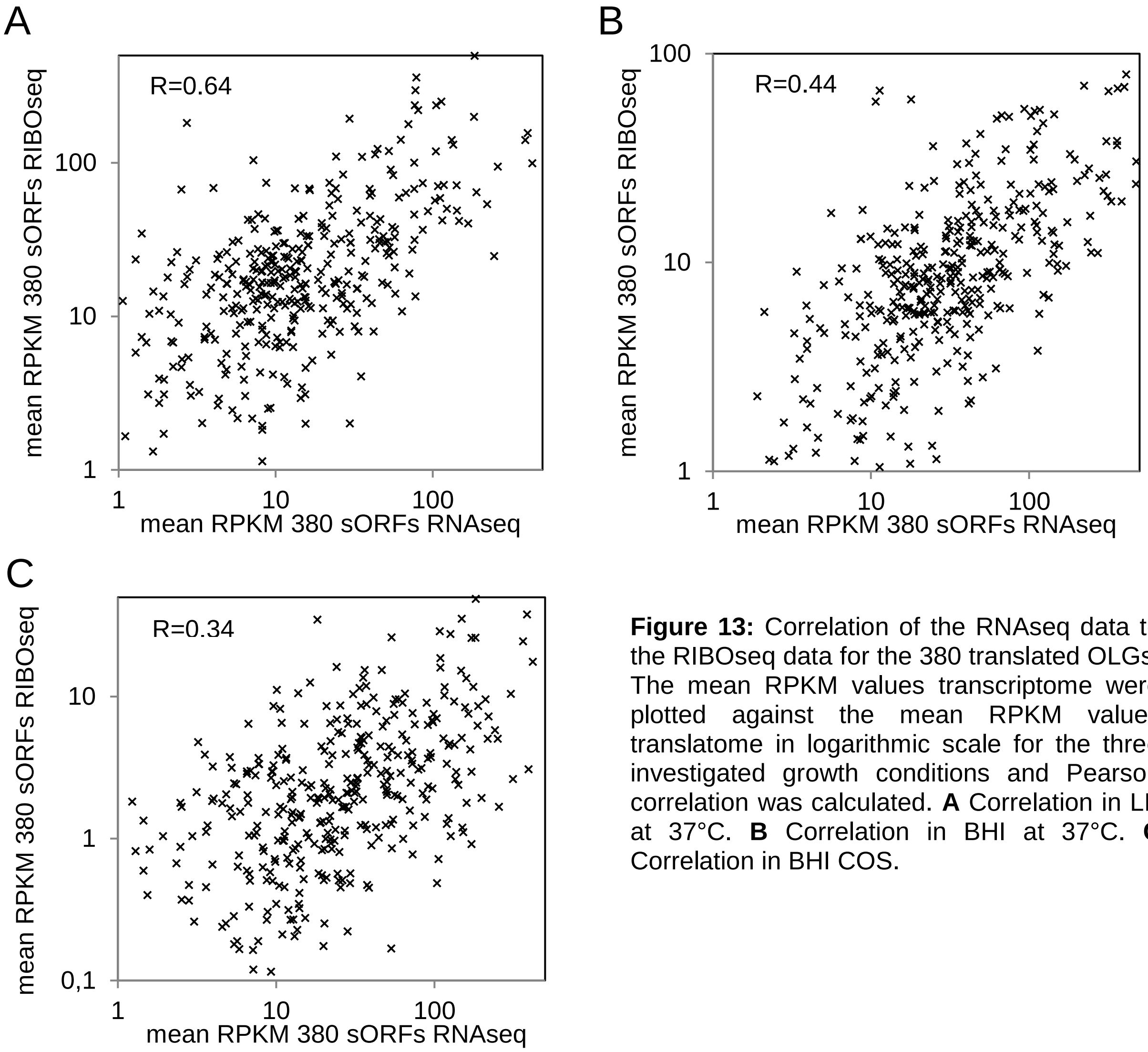


Figure 13: Correlation of the RNAseq data to the RIBOseq data for the 380 translated OLGs. The mean RPKM values transcriptome were plotted against the mean RPKM values translatome in logarithmic scale for the three investigated growth conditions and Pearson correlation was calculated. **A** Correlation in LB at 37°C. **B** Correlation in BHI at 37°C. **C** Correlation in BHI COS.

2.7.4 Size and RCV distribution of the potential novel genes

The translated OLGs have an average length of 168 bp, which corresponds to a protein of 56 AAs. The size distribution of the translated OLGs was compared to the size distribution of the 250 shortest annotated genes of EHEC Sakai. For the comparison, the same set of short annotated EHEC genes was used as in Hücker et al. (2017a). One third of the putative antisense genes encode proteins smaller than 40 AAs, whereas the majority of short annotated genes has a size of 60-79 AAs (Figure 14). However, translation of some longer antiparallel ORFs was detected as well: 6% are larger than 100 AA, and the longest ORF OLGECs066 encodes a protein of 328 AA (Supplementary Table S2).

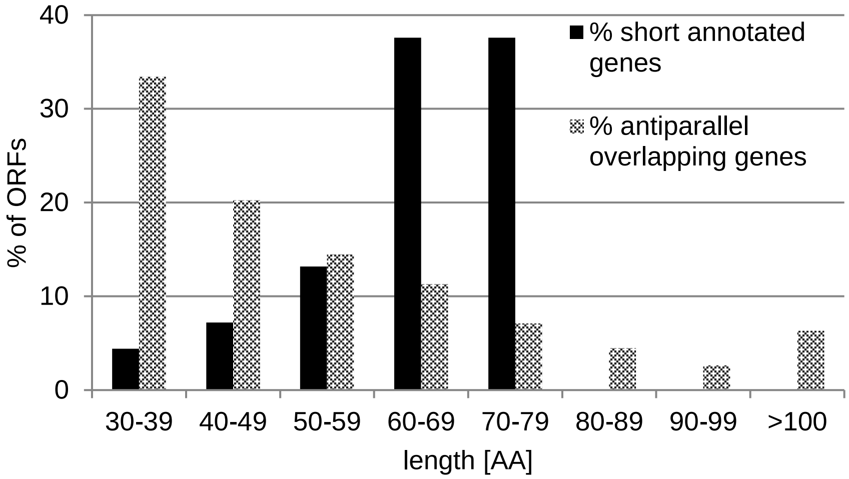


Figure 14: Size distribution of the 380 translated OLGs compared to the 250 shortest annotated genes of EHEC Sakai. The size in AA was grouped in steps to 10 AAs, and the number of ORFs in percent was determined for every group.

It was investigated, whether the putative novel antisense genes are translated to the same extent as the 250 short annotated genes. A measure of translation efficiency is the RCV calculated by the RPKM value translatome over the RPKM value transcriptome. The RCVs of the translated OLGs were binned in distinct groups, and the number of ORFs in each group was determined for every growth condition (Figure 15). The overall RCV distribution of translated antisense ORFs and short annotated genes was similar at every condition. In LB, most ORFs have a moderate translatability, but some ORFs are also translated with high efficiency. The mean RCV of the OLGs is 1.92 and the mean RCV of the annotated genes is with 1.55 even slightly lower (Hücker et al., 2017a). At the BHI conditions, the global translatability decreases, but putative OLGs and short annotated genes still show a comparable RCV distribution. The mean RCV of the OLGs at BHI control is 0.22, and at BHI stress it is

0.12 (compared to 0.55, and 0.12 of the short annotated genes, respectively). In conclusion, the putative OLGs are translated with the same efficiency as short annotated genes supporting their protein-coding character.

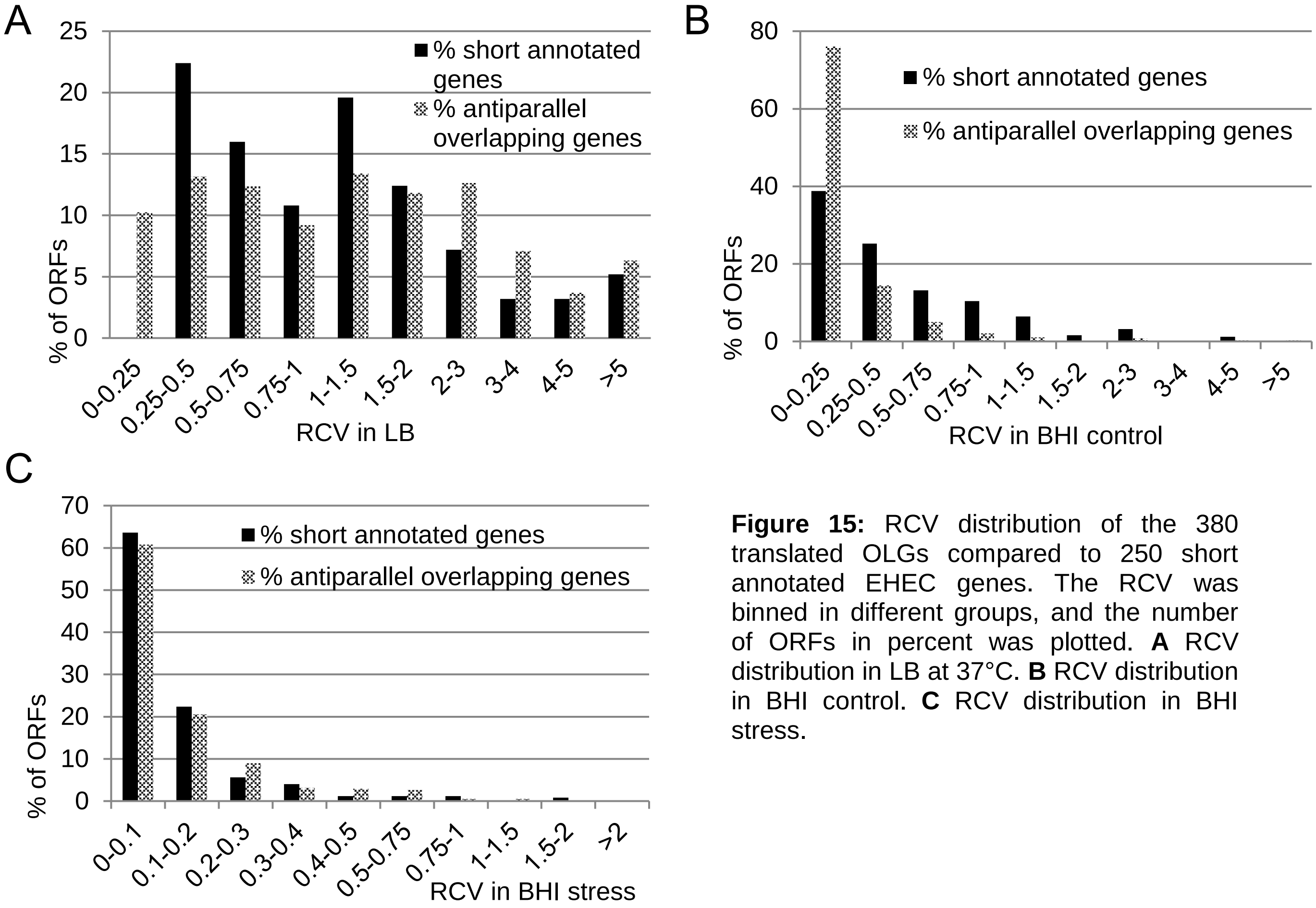


Figure 15: RCV distribution of the 380 translated OLGs compared to 250 short annotated EHEC genes. The RCV was binned in different groups, and the number of ORFs in percent was plotted. **A** RCV distribution in LB at 37°C. **B** RCV distribution in BHI control. **C** RCV distribution in BHI stress.

2.7.5 Translated OLGs with annotated homologs

Using the AA sequences of the 380 translated OLGs, a BLASTP search against the RefSeq database was performed in order to detect annotated homologs in other bacteria. For an e-value threshold ≤ 10^{-10}, homologs for 44 OLGs (11.6%) were found. Using a less conservative e-value threshold of ≤ 10^{-3}, the number of OLGs with annotated homologs increased to 66 (17.4%). For each OLG, the BLASTP hit with the lowest e-value is listed in Supplementary Table S3. Usually, the detected homologs are in closely related strains/species. The presence of annotated homologs supports the claim that these 66 OLGs are indeed protein-coding. In contrast to OLGs, more than

50% of the putative novel intergenic genes had annotated homologs (Hücker et al., 2017a). Annotation algorithms usually discard overlapping ORFs, which is most likely the reason for the omission of annotation in case of overlapping ORFs, even though the OLG is present in other organisms (Delcher et al., 2007). Thus, only the often better conserved mother gene becomes annotated. In addition, many translated OLGs would encode short proteins, which are also often discarded by annotation algorithms. Therefore, it is highly unlikely for a combination of a short ORF in antisense to a conserved gene to become annotated. Accordingly, short intergenic ORFs have less frequently an annotated homolog, but their sequence is conserved in closely related species, as detected by a TBLASTN analysis (Hücker et al., 2017a).

2.7.6 Translated OLGs are differentially regulated under three growth conditions

Differential regulation on translational and/or transcriptional level of the 380 translated OLGs was investigated using *edgeR* (compare to 2.5). First, BHI stress was compared to BHI control. Overall, 18% of OLGs are differentially expressed. Compared to the short annotated genes, of which 33% were differentially regulated, this value is smaller. However, the direction of regulation shows the same trend: the majority of ORFs is downregulated on translational level (Figure 16A). Figure 16B shows the translational downregulation of OLGECs297 as an example gene. The transcription does not change significantly between the BHI conditions for this gene, whereas the translation is decreased 14-fold at stress. All differentially regulated OLGs at the BHI conditions, including the mean number of reads normalized to library size, fold change, p-value, and FDR are listed in Supplementary Table S4. Next, differential regulation between LB and BHI, both at 37°C, was investigated. Again, more short annotated genes (36%) are differentially regulated comparing to 16% of translated OLGs (significantly regulated OLGs are listed in Supplementary Table S5). No clear trend for a preferred level of regulation is observed (Figure 16C). Downregulation on transcriptional level dominates for OLGs. Further, more annotated short genes than OLGs are upregulated in LB. An example of a translationally downregulated OLG, OLGECs052, is shown in Figure 16D. This gene was decreased 1.5-fold on transcriptional level (which was not significant), but translation was 8.5-fold decreased significantly.

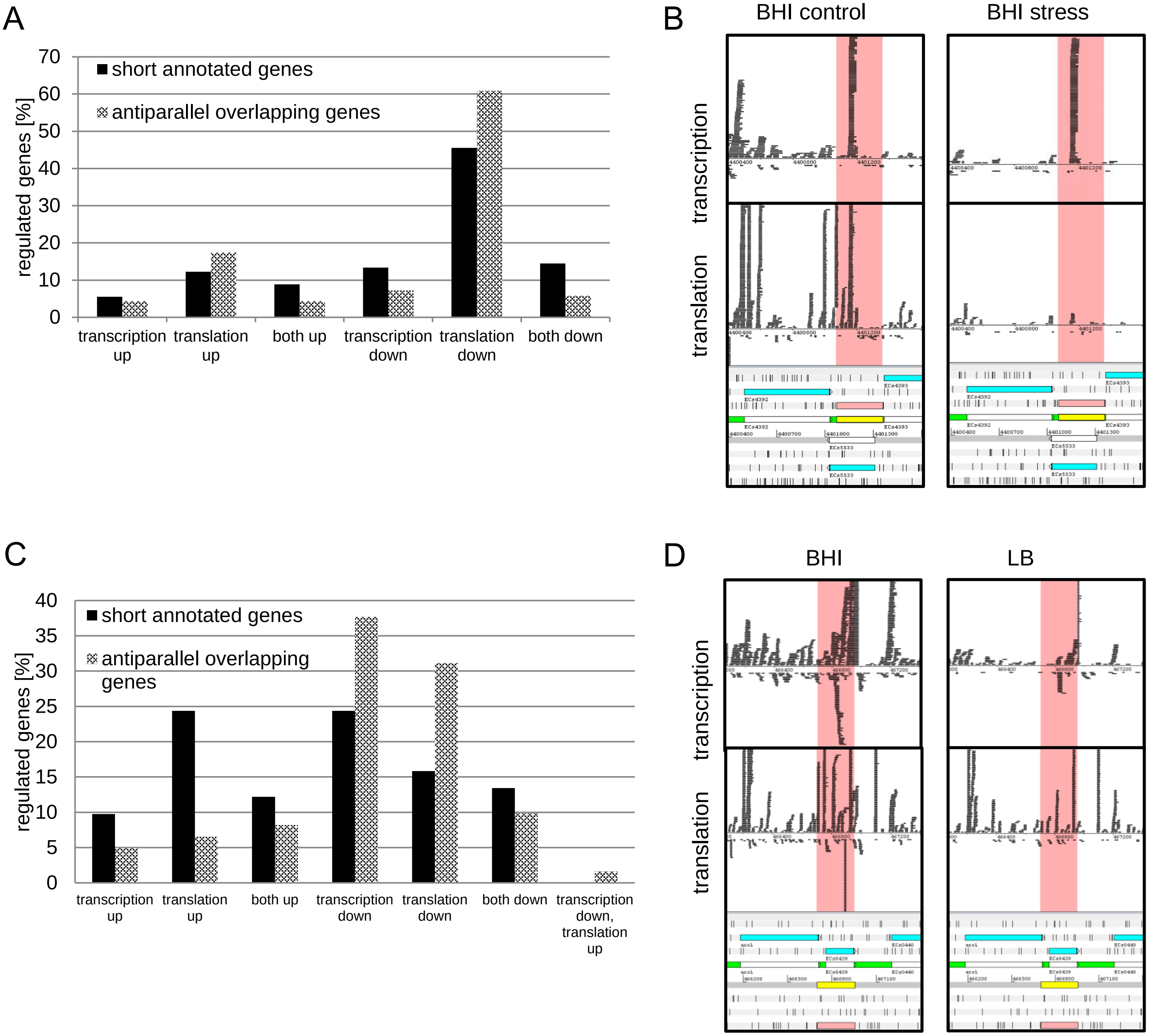


Figure 16: Differentially regulated OLGs. **A** Differential regulation in BHI stress. For the 69 OLGs showing differential expression on transcriptional and/or on translational level, the mechanism of regulation was determined, plotted in percent, and compared to the regulation of 250 short annotated EHEC genes. **B** Translational downregulation of OLGECs297 visualized in the Artemis genome browser. OLGECs297 is highlighted in pink and annotated genes are colored in blue. The upper panel shows the strand-specifically mapped RNAseq and RIBOseq reads. At the stress condition, translation is almost turned off. **C** Differential regulation in LB. Sixty-one OLGs are differentially transcribed and/or translated. The type of regulation was determined, and compared to short annotated EHEC genes. **D** Translational down-regulation of OLGECs052 visualized in the Artemis genome browser. The transcription of OLGECs052 (highlighted in pink) is decreased 1.5-fold and the translation 8.5-fold.

Significant differences in expression between growth conditions hint towards a biological meaning of these translated OLGs. Regulation depending on growth conditions would not be expected for non-functional sequences. Comparison of ORF expression in RNAseq and RIBOseq data of different growth conditions provides a first indication for the functionality of the gene (Fellner et al., 2015; Landstorfer et al., 2014).

2.7.7 Prediction of σ^{70} promoters

For all translated OLGs, the sequences 300 bp upstream of their putative start codon were submitted to BPROM to predict σ^{70} promoters (Supplementary Table S3). Except for three OLGs, all OLGs are predicted to contain a σ^{70} promoter in their upstream region (Table 1). This number of ORFs is even higher than in case of the short annotated genes. In contrast, the mean promoter strength predicted, expressed by the LDF-score, is higher for the short annotated genes. However, the distance of the predicted promoter to the start codon is shorter for the OLGs.

Table 1: Prediction of σ^{70} promoters for the translated OLGs and for the short annotated genes.

	σ^{70} promoter predicted	**mean LDF-score**	**mean distance to start codon**
translated OLGs	99%	2.75	135
short annotated genes	97%	3.43	187

Even if no σ^{70} promoter was predicted upstream of the start codon, transcription of the ORF is still possible. Besides the housekeeping σ-factor σ^{70}, *E. coli* uses six alternative σ-factors (Burrows et al., 2003; Nonaka et al., 2006; Yu et al., 2006), which are not predicted by the program used. Furthermore, a gene might be part of an operon and in this case, transcription would initiate at a promoter of a gene upstream.

2.7.8 Prediction of ρ-independent terminators

The software FindTerm was used to search in the region 300 bp downstream of the stop codon for a ρ-independent terminator. For 44 OLGs, such a terminator is predicted (Table 2, Supplementary Table S3). The percentage of short annotated genes with a predicted ρ-independent terminator is slightly higher, but also for this group, four-fifths do not have a predicted terminator. The mean free energy is similar for both gene groups, whereas the distance of the stop codon to the terminator is shorter for the annotated genes.

Table 2: Prediction of ρ-independent terminators for the translated OLGs and for the short annotated genes.

	ρ-independent terminator predicted	**mean free energy**	**mean distance to stop codon**
translated OLGs	12%	-15.4	131
short annotated genes	22%	-16.9	68

Comparable to the presence of a σ^{70} promoter, the prediction of a ρ-independent terminator is evidence for a protein-coding gene, but conversely, absence does not indicate a non-coding ORF. The translated OLG could be part of an operon and, thus, the terminator is located downstream of further genes. In addition, *E. coli* possesses another termination mechanism, which is even more frequent than ρ-independent termination, using the protein Rho (Banerjee et al., 2006). Such ρ-dependent terminators cannot be predicted by bioinformatics.

2.7.9 Prediction of Shine-Dalgarno sequences

The 16S rRNA of the ribosome contains the so-called anti-Shine-Dalgarno sequence, which is complementary to the Shine-Dalgarno (SD) sequence located upstream of the start codon on the mRNA. Base pairing between SD and anti-SD leads to ribosome binding followed by translation initiation. The optimal SD sequence (taAGGAGGt) has a ΔG° of -9.6, and is ideally located 8-10 bp upstream of the start codon. However, a location of 4-16 bp upstream of the start codon is still in the optimal range (Ma et al., 2002). Presence of a SD sequence is also an important criterion for annotation algorithms (Delcher et al., 2007). Therefore, the region 30 bp upstream of the translated OLGs was searched for a SD sequence. Indeed, 46% of translated OLGs have a SD sequence predicted (Table 3, Supplementary Table S3). For the short annotated genes, 80% have a SD sequence predicted, the mean ΔG° is lower, and the mean distance to the start codon is smaller.

Table 3: Prediction of Shine-Dalgarno sequences for the translated OLGs and for the short annotated genes.

	Shine-Dalgarno sequence predicted	mean ∆G°	mean distance to start codon
translated OLGs	46%	-4.5	11
short annotated genes	80%	-5.2	7

However, OLGs without a predicted SD sequence can still be translated, because translation of leaderless mRNAs occurs, too (Zheng et al., 2011). Especially under stress, leaderless mRNAs are preferentially translated by modified ribosomes, where the anti-SD sequence was cleaved (Vesper et al., 2011). The higher average SD strength of the short annotated genes is not mirrored in the translatability distribution (compare 2.7.4), because OLGs and short annotated genes were found to have similar RCVs. However, the mean RPKM values of the transcriptome and translatome are higher for the short annotated genes; thus, global expression is higher for those genes.

2.7.10 Factors influencing RIBOseq signals in experiments

Due to different protocols used, it is difficult to compare the RIBOseq data of this study to other published bacterial RIBOseq data. The choice of RNase(s) for digestion of unprotected mRNA seems to impact footprint size and distribution. In many bacterial RIBOseq protocols *Micrococcus* nuclease (MNase) is used (Li et al., 2012; Oh et al., 2011). The disadvantage of this nuclease is its high sequence specificity leading to footprints with an average size of 28 bp similar to eukaryotic footprints. In eukaryotic protocols, normally RNase I is used (Ingolia et al., 2009), which shows much less sequence specificity. Some publications claim RNase I to be inactive in bacteria, because it binds to the ribosome (Gerashchenko and Gladyshev, 2017), but when high concentrations are used, also bacterial footprints with an average size of 21-23 bp can be produced using this nuclease (Landstorfer, 2014; Neuhaus et al., 2016). Ribosomes of different species seem to show different sensitivity towards RNases (Miettinen and Bjorklund, 2015). Partial digestion of ribosomal RNA must be avoided, and has to be tested beforehand. The mixture of five different RNases used here is an important protocol improvement to get rid of sequence specificity.

Also, RIBOseq results are influenced by the usage of translational inhibitors (Gerashchenko and Gladyshev, 2014). The inhibitor might enrich one ribosomal conformation artificially (Lareau et al., 2014). In this study, chloramphenicol was used to stall translation. Recently, it turned out that chloramphenicol is not a universal inhibitor of peptide bond formation, but translation is preferentially stalled at alanine, serine, and threonine (Marks et al., 2016). In contrast, specific inhibitors, like tetracycline, can be used to stall ribosomes preferentially at the start codon, which allows a different type of read-out for RIBOseq data (Nakahigashi et al., 2016). Thus, the choice of inhibitor depends on the experiments, but for a general overview, they should be avoided in RIBOseq experiments. Without translational inhibitors, it is mandatory to harvest the cells rapidly. Otherwise the ribosomes could run off the mRNA.

Different buffers reported for cell lysis and sucrose gradient centrifugation influence RIBOseq results. Hsu et al. (2016) improved the resolution of *Arabidopsis* footprints greatly by changing ion strengths of buffer components. In EHEC, the buffer used influences the resolution on sub-codon level as well, and the distribution of reads mapping to rRNA, tRNA, or mRNA (Abellan-Schneyder, 2017).

2.7.11 Comparison to published RIBOseq data

All-in-all the analyses conducted, support the claim that the majority of the 380 translated OLGs represents protein-coding genes. Coverage of antisense ORFs with RIBOseq reads is already strong evidence for translation into a protein. Usually, the overlapping annotated gene is also showing translation signals, and its ORF is much longer. Therefore, it is unlikely that the mother gene was just falsely annotated. Theoretically, RIBOseq reads might also have been caused by other co-purified RNA binding proteins or randomly bound ribosomes, not actively translating the mRNA, but these options are unlikely (Liu and Qian, 2016). Furthermore, an RCV in the same range as for annotated protein-coding genes gives additional evidence that the translated OLGs do not represent novel antisense RNAs (Neuhaus et al., 2017). Detection of annotated homologs, differential regulation between growth conditions, and structural genetic elements (σ^{70} promoter, ρ-independent terminator, and Shine-Dalgarno sequence) are further indicators for the protein-coding nature of these antisense OLGs.

Landstorfer (2014) discovered 472 transcribed antiparallel overlapping ORFs in EHEC strain EDL933 at nine different growth conditions (Landstorfer et al., 2014), and translation of 242 antisense ORFs at the condition LB at 37°C, OD_{600} 0.4. Because EDL933 is very closely related to EHEC Sakai, the sequences of all transcribed and/or translated ORFs are present in both strains, and the overlap between the data sets was investigated. Only 39 (10.3%) OLGs are considered translated in both strains. Additional 25 OLGs (6.6%) of the 380 translated OLGs are also found transcribed in at least one of the tested conditions. Reasons for this quite low overlap could be that different growth conditions were used: Landstorfer (2014) worked with 0.5xLB, whereas in this study, full LB was used. In addition, a different RIBOseq protocol was used here. The mRNA not protected by ribosomes was digested with a mixture of five RNases, and rRNA was depleted using another kit. Moreover, in this study, further parameters for deciding if an ORF is translated, were applied: Landstorfer (2014) applied RPKM value thresholds and visual inspection in the Artemis genome browser only, but no RCV and coverage thresholds were used. This increases the probability of false positives in this data set. For instance, the signal was assigned to the wrong ORF.

Former RIBOseq studies in bacteria were focused on the translation process, like decoding rate determination (Dana and Tuller, 2014; Subramaniam et al., 2014), translational pausing (Li et al., 2012; Mohammad et al., 2016), ribosome drop-off (Sin et al., 2016), and termination (Baggett et al., 2017). The translation of non-annotated ORFs was only investigated in a few studies, and until today, there is no publication that deals with translation of ORFs antisense to annotated genes. However, the existing publications prove that the combination of RNAseq and RIBOseq is a valuable method to discover novel genes: Jeong et al. (2016) report translation of 31 ORFs of *Streptomyces coelicolor*, which were previously annotated as ncRNAs. Baek et al. (2017) discovered translation of 130 non-annotated short ORFs in *Salmonella,* and Neuhaus et al. (2016) found translation of 72 intergenic ORFs of EHEC strain EDL933. In contrast, in eukaryotes, massive translation outside of annotated genes is reported frequently for many organisms from yeast to human. Many translation signals occur upstream of annotated genes, and previously annotated lncRNAs often appear to be translated. In addition, ORFs antisense to annotated genes are covered by RIBOseq reads (Aspden et al., 2014; Bazzini et al., 2014; Calviello et al., 2016; Ingolia et al., 2014; Ji et al., 2015; Ruiz-Orera et al., 2014; Smith et al., 2014). Meanwhile, software tools are available to distinguish protein-coding genes from background (Calviello et al., 2016; Fields et al., 2015; Malone et al., 2017), and a data base for non-annotated translated ORFs exists (Olexiouk et al., 2016).

All those observations indicate that many more parts of a genome are translated, and probably that the number of protein-coding genes in EHEC and other organisms is much higher than presumed in the past. In this study, translation of 465 intergenic ORFs and 380 ORFs overlapping antiparallel to annotated genes in *Escherichia coli* O157:H7 Sakai was uncovered. Further analyses support that these translated ORFs indeed represent protein-coding genes. If all 845 translated ORFs really encode proteins, the number of genes in EHEC Sakai would increase of 16%. Functions of the novel proteins are unknown. Putatively, several proteins might be associated to the cell membrane, which was reported for small proteins described earlier (Hemm et al., 2008; Kemp and Cymer, 2014; Storz et al., 2014). Some researchers doubt that RIBOseq signals are only caused by translation, and point towards protein-coding genes. For instance, Ingolia et al. (2014) assume simple pervasive translation analogous to pervasive transcription (Wade and Grainger, 2014). Nevertheless, several reports confirmed that RIBOseq signals of an ORF indicate a protein-coding gene (Chu et al., 2015; Landry et al., 2015; Liu and Qian, 2016).

The discovery of 380 translated ORFs, antiparallel overlapping to annotated genes, is also interesting from an evolutionary point of view. These novel OLGs likely originated *de novo* by overprinting (Grassé, 1977; Keese and Gibbs, 1992). This mechanism seems to be more frequent for gene birth than previously presumed. Non-trivial OLGs appear not to be a rare phenomenon in bacteria, but seem to occur frequently, as in case of viruses.

Part III: Functional characterization of selected antiparallel overlapping genes

Candidates, which are highly likely to be protein-coding OLGs, were selected from RNAseq and RIBOseq data for functional characterization. The selected candidates are transcribed and translated in at least one condition, have an RCV ≥ 0.25, have annotated homologs, and should have some other structural features predicted (σ^{70} promoter, ρ-independent terminator, Shine-Dalgarno sequence). Additionally, the mother gene should be transcribed and translated as well. Functional characterization was undertaken for those novel genes by cloning strand-specific knock-out mutants and search for a phenotype in competitive growth experiments against the wild type. Furthermore, transcriptional start and termination site were determined and activity of the putative promoter region was investigated. Feasibility of protein expression was tested by expressing an EGFP-fusion protein.

2.8 Publication 3: 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

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

Sarah M. Hücker[1], Sonja Vanderhaeghen[1], Isabel Abellan-Schneyder[1], Romy Wecko[1], Svenja Simon[2], Siegfried Scherer[1,3] and Klaus Neuhaus[1,4*]

[1]Chair for Microbial Ecology, Technische Universität München, Weihenstephaner Berg 3, 85354 Freising, Germany; [2]Chair for Data Analysis and Visualization, Department of Computer and Information Science, University of Konstanz, Box 78, 78457 Konstanz, Germany; [3]ZIEL – Institute for Food & Health, Technische Universität München, Weihenstephaner Berg 3, 85354 Freising, Germany; [4]Core Facility Microbiome/NGS, ZIEL – Institute for Food & Health, Technische Universität München, Weihenstephaner Berg 3, 85354 Freising, Germany.

*Correspondence: neuhaus@tum.de

ABSTRACT

Background: Due to the DNA triplet code, it is possible that the sequences of two or more protein-coding genes overlap to a large degree. However, such non-trivial overlaps are usually excluded by genome annotation pipelines and, thus, only a few overlapping gene pairs have been described in bacteria. In contrast, transcriptome and translatome sequencing reveals many signals originated from the antisense strand of annotated genes, of which we analyzed an example gene pair in more detail.

Results: A small open reading frame of *Escherichia coli* O157:H7 Sakai, designated *laoB* (L-arginine responsive overlapping gene), is embedded in reading frame -2 in the antisense strand of ECs5115, encoding a CadC-like transcriptional regulator. This overlapping gene shows evidence of transcription and translation in LB and BHI medium based on RNAseq and RIBOseq. The transcriptional start site is 289 bp upstream of the start codon and transcription termination is 155 bp downstream of the stop codon. Overexpression of LaoB fused to an EGFP reporter was possible. The sequence upstream of the transcriptional start site displayed strong promoter activity under different conditions, whereas promoter activity was significantly decreased in the presence of L-arginine. A strand-specific translationally arrested mutant of *laoB* provided a significant growth advantage in competitive growth experiments in the presence of L-arginine compared to the wild type, which returned to wild type level after complementation of *laoB in trans*. A phylostratigraphic analysis indicated that the novel gene is restricted to the *Escherichia/Shigella* clade and might have originated recently by overprinting leading to the expression of part of the antisense strand of ECs5115.

Conclusions: Here, we present evidence of a novel small protein-coding gene *laoB* encoded in the antisense frame -2 of the annotated gene ECs5115. Clearly, *LaoB* is evolutionarily young, and it originated in the *Escherichia/Shigella* clade by overprinting, a process which may be more important for the *de novo* evolution of bacterial genes than previously assumed.

Keywords: overlapping gene, overprinting, small protein, *de novo* gene, EHEC

INTRODUCTION

The DNA triplet code is constructed such that the majority of amino acids (AAs) can be encoded by more than one codon, leading to the so-called degeneration of the genetic code. Codon position three shows the highest degeneration (wobble position), whereas position one is only slightly degenerated, and position two is not degenerated [1]. Thus, a DNA double strand contains six possible reading frames, each of which has the capacity to encode a protein and it is feasible that the sequences of two or more protein-coding genes overlap. Most generally, overlapping genes (OLGs) share at least one nucleotide between the coding regions of two genes. When the reading frame of the evolutionary older established gene (mother gene) is defined as frame +1, a same-strand overlap in reading frames +2 or +3 relative to the annotated gene is possible. Same-strand OLGs originate from programmed ribosomal frameshift [2, 3] or programmed transcriptional realignment [4]. Additionally, a second gene can overlap the mother gene antisense in frames -1, -2 or -3. It is under debate, which antisense frame is preferred for the occurrence of OLGs. In *E. coli,* most long antisense open reading frames (ORFs) are detected in frame -1 [5]. However, this finding might be caused by codon bias of the mother gene [6, 7]. Whereas Krakauer [1] predicted highest constraints on frame -2, in *E. coli* more long ORF are found in frame -2 than in -3 [6, 8]. Lèbre and Gascuel [9] investigated the constraints of OLGs at the AA level and detected the highest constraints on frame -3 due to a high number of “forbidden dipeptides” within the protein encoded, which would cause a stop codon in the established gene.

In prokaryotes, many genes are organized in operons, which are transcribed as a polycistronic mRNA. In these cases, trivial same-strand overlaps of only a few base pairs are very common and facilitate translational coupling [10]. In contrast, almost no long OLGs (overlap ≥ 90 bp) have been described in bacteria [11-14], while longer OLGs are well known in viral genomes, probably leading to genome size reduction, since in viruses, 38% of all AAs are encoded overlapping, and in many cases the OLGs encode accessory proteins with unusual sequence composition like many disordered regions [15-17].

OLGs may originate by overprinting [18]: By chance, an overlapping reading frame is expressed in a bacterial population. However, encoding two functional genes at one locus leads to severe constraints of sequence evolution, since many mutations will influence the AA sequence of two genes carrying completely different functions [1, 8, 9]. This may be one reason, why the overprinting hypothesis has been neglected as being rather unlikely [7, 19]. Instead, the gene duplication followed by subfunctionalization or neofuctionalization hypothesis [20] has been favored for the origin of novel genes.

Here, we present an initial functional characterization of the novel OLG *laoB* of *Escherichia coli* O157:H7 strain Sakai, the expression of which was seen in transcriptome data and ribosomal profiling [21]. *LaoB* overlaps antiparallel to the annotated gene ECs5115, and this overlapping gene pair is a novel example of this seemingly rare form of bacterial gene organization. We propose that *laoB* originated very recently by overprinting.

MATERIAL AND METHODS

Bacterial strains and plasmids used in this study are listed in Additional file S1. Oligonucleotides are listed in Additional file S2.

Determination of transcriptional start site by 5’ RACE

An overnight culture of *Escherichia coli* O157:H7 Sakai (Genbank accession number NC_002695) [22] was inoculated 1:100 in 0.5 LB with 400 mM NaCl, and incubated at

37°C and 150 rpm until an OD_{600} of 0.8 was reached. Total RNA of 500 µl EHEC culture was isolated with Trizol, and the remaining DNA was digested using 2 U TURBO^{TM} DNase (Thermo Fisher Scientific). The 5'RACE System for Rapid Amplification of cDNA Ends, Version 2.0 (Invitrogen) was used according to the manual. After the second PCR, the dominant product was excised from the agarose gel, and purified with the GenElute^{TM} Gel Extraction Kit (Sigma-Aldrich). The PCR product was Sanger sequenced by Eurofins with oligonucleotide *laoB*+25R.

Determination of transcriptional stop site by 3'RACE

Total RNA of 500 µl EHEC Sakai overnight culture in LB medium was isolated using Trizol and the remaining DNA was digested using 2 U TURBO^{TM} DNase (Thermo Fisher Scientific). The 5'/3' RACE Kit, 2nd Generation (Roche Applied Science) was applied according to the manual, but instead of an oligo dT primer for cDNA synthesis the gene specific primer *laoB*-12F was used. A nested PCR was performed for product amplification. The dominant product was excised from the agarose gel, purified with the GenElute^{TM} Gel Extraction Kit (Sigma-Aldrich), and Sanger sequenced (Eurofins) with oligonucleotide *laoB*+31F.

Cloning of pProbe-NT plasmids and determination of promoter activity

The genomic region 300 bp upstream of the transcriptional start site was amplified by PCR, and restriction enzyme cut sites for *Sal*I and *EcoR*I were introduced. The PCR products were cloned into the plasmid pProbe-NT [23], and transformed into *Escherichia coli* Top10. The plasmid sequence was verified by Sanger sequencing (Eurofins). Overnight cultures of *E. coli* Top10 + pProbe-NT (negative control) and pProbe-NT-PromoterTSS were used for 1:100 inoculation of 10 ml 0.5 LB medium with 30 µg/ml kanamycin. The following conditions were investigated for promoter activity in 0.5xLB medium each: plain LB, at pH 5, at pH 8.2, plus 400 mM NaCl, plus 0.5 mM $CuCl_2$, plus 2 mM formic acid, plus 2.5 mM malonic acid, or plus 10 mM L-arginine. Cultures were incubated at 37°C and 150 rpm until an OD_{600} of 0.5 was reached. Then, the cells were pelleted, washed once with PBS, and resuspended in 1 ml PBS. The OD_{600} was adjusted to 0.3 and 0.6. Four-times each 200 µl of both OD-adjusted suspensions were pipetted in a black microtiter plate and the fluorescence was measured (Wallac Victor^3, Perkin Elmer Life Science, excitation 485 nm, emission 535 nm, measuring time 1 s). The fluorescence of *E. coli* Top10 without vector was subtracted as background. To measure promoter activity after L-arginine supplementation, the experiment was repeated in modified MOD medium [24] without L-glutamic acid, L-arginine, and L-aspartic acid, since these AAs are easily convertible within the cell. Depleted MOD medium and MOD medium supplemented with 10 mM L-arginine were tested. The experiments were performed in triplicate. Significance of changes was calculated by the Student's t-test.

Cloning of a C-terminal LaoB-EGFP fusion protein and overexpression of LaoB-EGFP protein

The *laoB* sequence without the stop codon was amplified by PCR, and restriction enzyme cut sites for *Pst*I and *Nco*I were introduced. The PCR product was cloned into the plasmid pEGFP, and transformed into *Escherichia coli* Top10. The plasmid sequence was verified by Sanger sequencing (Eurofins). For overexpression of the fusion protein, overnight cultures of *E. coli* Top10 + pEGFP and *E. coli* Top10 + pEGFP-*laoB* were inoculated 1:100 in 10 ml 0.5 LB medium with 120 µg/ml ampicillin in duplicates. Cultures were incubated at 37°C and 150 rpm until an OD_{600} of 0.3 was reached. For one culture each, protein expression was induced using 10 mM IPTG. Incubation of induced and uninduced cultures was continued for 1 h. Cells were pelleted, washed once with PBS, and the pellet was resuspended in 1 ml PBS. The OD_{600} was adjusted to 0.3 and 0.6. Four times each 200 µl of the OD-adjusted bacterial suspensions were pipetted in a black microtiter plate and the fluorescence was measured as before. The experiment was performed in triplicate. Significance of changes was calculated by the Student's t-test.

Cloning of a translationally arrested *laoB* mutant

For cloning of the genomic knock-out mutant Δ*laoB,* the method described by Kim *et al* [25] was adapted. The mutations introduced do not change the AA sequence of the overlapping gene ECs5115. The pHA_{1887} fragment and the selection cassette were amplified by PCR from the plasmid pTS2Cb. Three consecutive point mutations, leading to a premature stop codon (5th codon) and a restriction enzyme cut site deletion (see below), were introduced into the *laoB* sequence by PCR using the oligonucleotides HA3*laoB*-139F and SM5*laoB*mut+42R (3' mutation fragment), and SM3*laoB*mut-16F and HA5*laoB*+183R (5' mutation fragment). Because the plasmid pTS2Cb-Δ*laoB* was obtained by Gibson Assembly, the four PCR fragments contain overlapping sequences. In a total reaction volume of 20 µl, 200 fmol of each PCR fragment and the NEBuilder® HiFi DNA Assembly Master Mix (NEB) were incubated at 50°C for 4 h. Two µl of the reaction were transformed into *E. coli* Top10 and plated on LB agar with 120 µg/ml ampicillin and 20 µg/ml chloramphenicol. Next, the mutation cassette was amplified by PCR using pTS2Cb-Δ*laoB* as template, and the PCR product of correct size was purified from an agarose gel (GenElute™ Gel Extraction Kit; Sigma-Aldrich). *E. coli* O157:H7 Sakai [22] was transformed with the plasmid pSLTS and, subsequently, transformed with 75 ng of the mutation cassette. After incubation for 3 h at 30°C and 150 rpm in SOC medium, the cells were plated on LB-agar plates with 120 µg/ml ampicillin and 20 µg/ml chloramphenicol, and incubated at 30°C. One colony per plate was suspended in PBS. One-hundred µl of a 1:10 dilution in PBS were plated on LB agar with 120 µg/ml ampicillin and 100 ng/ml anhydrotetracycline for I-SceI induction, and incubated at 30°C over night. Several colonies were streaked on LB agar with 20 µg/ml chloramphenicol and plain LB agar, and incubated at 37°C over night. Colonies that were only able to grow on LB were selected, and the genomic area surrounding the point mutations introduced was amplified by PCR. Additional to the premature stop codon, the restriction enzyme cut site for *Mnl*I was deleted, which was screened for by restriction digest of PCR products with this enzyme. Correct introduction of the three point mutations was assumed for *Mnl*I-digestion negative PCR products, and confirmed by Sanger sequencing (Eurofins).

Competitive growth assays

Overnight cultures in LB medium of EHEC Sakai wild type and EHEC Sakai Δ*laoB* were adjusted to an OD_{600} of 1.0 and then mixed in equal quantities (500 µl wild type + 500 µl mutant). Five-hundred µl of the mixture were pelleted, and the cells were snap frozen in liquid nitrogen (control, t=0). Ten ml 0.5 LB medium were inoculated 1:3000 with the mixed EHEC culture. The following conditions were investigated in 0.5xLB: plain LB, at pH 5, at pH 8.2, plus 400 mM NaCl, plus 0.5 mM $CuCl_2$, plus 2 mM formic acid, plus 2.5 mM malonic acid, plus 4 mM malic acid, plus 400 µM $ZnCl_2$, or plus 20 mM L-arginine. Cultures were incubated for 18 h at 37°C and 150 rpm. Then, 500 µl of the culture were pelleted, 100 µl water were added to the pellet, and the sample was heated to 95°C for 10 min. Using this crude DNA preparation, a PCR was performed with the primer pair *laoB*-38F and *laoB*+140R. The PCR product was Sanger sequenced (Eurofins), and the ratio between wild type and mutant Δ*laoB* was determined by comparing peak heights. The absolute numbers were transformed into percentage values for each condition, and the values were normalized to a t=0 ratio for 50:50 wild type over mutant. Thus, the competitive index was calculated using the following formula:

$$CI = \frac{mutant_{end}[\%]\ \times\ wild\ type_{start}[\%]}{mutant_{start}[\%]\ \times\ wild\ type_{end}[\%]}$$

The experiment was performed in biological triplicates. Significance was calculated by the Student's t-test.

Complementation of EHEC Δ*laoB*

To compensate the *laoB* genomic knock-out mutation, the intact *laoB* ORF was supplemented *in trans* on a plasmid. First, the sequence of *laoB* was amplified by PCR, and restriction enzyme cut sites for *Nco*I and *Hind*III were introduced. The PCR product was cloned into the plasmid pBAD/*Myc-His*-C, and the plasmid was transformed into *E. coli* O157:H7 Sakai Δ*laoB*. As a negative control, the plasmid containing the mutated *laoB* gene (Δ*laoB*) was cloned. Next, competitive growth experiments were performed as described above using *E. coli* O157:H7 Sakai Δ*laoB* + pBAD-*laoB* (complementation) and *E. coli* O157:H7 Sakai Δ*laoB* + pBAD-Δ*laoB* (control). Both overnight cultures were supplemented with 120 µg/ml ampicillin, and the cultures were mixed in equal ratio. Ten ml of either 0.5 LB or 0.5 LB + 20 mM L-arginine were inoculated 1:3000 in quadruplicates. Induction of the *laoB* frame (present either as wild type or as Δ*laoB*) was performed with 0.002% arabinose. After incubation at 37°C and 150 rpm for 18 h, plasmids were isolated using the GenElute™Plasmid Miniprep Kit (Sigma-Aldrich). Using 20 ng isolated plasmid, PCR was performed with the oligonucleotides pBAD+208F and pBAD+502R. The PCR products were Sanger sequenced (Eurofins), and the ratio of intact *laoB* over translationally arrested Δ*laoB* was determined in percent. The experiment was performed in biological triplicates. Significant changes were calculated by the Student’s t-test.

Transcriptome and translatome sequencing

RNAseq and RIBOseq data sets of Hücker *et al* [26] were investigated with respect to translated ORFs located in antisense to annotated genes. Briefly, the bacteria had been grown under the following growth conditions: LB medium at 37°C, harvested at OD_{600} 0.4, BHI medium at 37°C, harvested at OD_{600} 0.1, and BHI medium supplemented with 4% NaCl at 14°C, harvested at OD_{600} 0.1. An ORF was considered translated, when (i) it was covered with at least one read per million mapped sequenced reads normalized to 1 kbp, (ii) ≥ 50% of the ORF is covered with RIBOseq reads, and (iii) the ribosomal coverage value (RCV) is at least 0.25 in both biological replicates. Promising candidates were verified by visual inspection using the Artemis genome browser [27].

Bioinformatics methods

Prediction of σ^{70} promoters

The region 550 bp upstream of the start codon of *laoB* was searched for the presence of a σ^{70} promoter with the program BPROM (Softberry) [28]. The LDF score given is a measure of promoter strength, whereupon an LDF score of 0.2 indicates presence of a σ^{70} promoter with 80% accuracy and specificity.

Prediction of alternative σ-factors

The search for alternative σ-factors was performed manually. The sequence 50 bp upstream of the detected TSS was compared to the consensus motifs of σ^{28} [29], σ^{32} [30], and σ^{54} [31].

Prediction of ρ-independent terminators

The region 300 bp downstream of the stop codon of *laoB* was searched for the presence and folding energy of a ρ-independent terminator using the program FindTerm (Softberry) [28].

Prediction of Shine-Dalgarno sequence

The free energy ∆G° of the region 30 bp upstream of the start codon of *laoB* was calculated according to Ma *et al* [32]. The perfect Shine-Dalgarno (SD) sequence taAGGAGGt has a ∆G° of -9.6. A ∆G° of -2.9 is considered the threshold for the presence of an SD sequence [32].

Detection of annotated homologs

The AA sequence LaoB, corresponding to *laoB*, was used to query the data base GeneBank with blastp using default parameters [33].

PredictProtein

LaoB was submitted to the software PredictProtein [34]. The methods PROFphd (secondary structure) [35], TMSEG (transmembrane helices) [36], DISULFIND (disulfide bonds) [37], and LocTree2 (subcellular localization) [38] were used.

Phylogenetic tree construction

For evolutionary analysis of *laoB* and ECs5115, tblastn was used with an e-value cutoff of 0.001 and at least 50% identity, which allows a search of nucleotide sequences homologous to a protein sequence query in all genomic sequences of the database independent of their annotation status. Blast hits with an e-value cutoff of 0.001 for ECs5115 were considered [39, 40]. For LaoB, tblastn was not sensitive enough to detect all existing genomic sequences, hence exemplary sequences within a broad range of sequence identities were downloaded from the database, and used for phylogenetic analysis. Multiple sequence alignments were conducted using MUSCLE implemented in MEGA6 [41]. The automated alignments were manually checked and adapted, where necessary. Homologous genes, in which the homolog of ECs5115 was intact, but *laoB* had no tblastn hit, were individually checked by pairwise alignments of the nucleotide sequences [EMBOSS Needle, 42]. The area in which *laoB* aligned with the (often) disintegrated *laoB* homologous sequences was translated to its AA sequence, and aligned by multiple sequence alignment as before.

Reference phylogenetic trees of the strains and species examined were constructed according to Fellner *et al* [14]. Briefly, a concatenated sequence of the housekeeping genes 16S rDNA*, atpD, adk, gyrB, purA,* and *recA* was used. The sequences were aligned using ClustalW in MEGA6. Columns with gaps or ambiguities were removed. The final dataset contains 7484 positions. The best nucleotide substitution model was searched for using MEGA6. The final Maximum-Likelihood tree was calculated using Neighbor Joining, and bootstrapped 1000-times. The best nucleotide substitution model for tree construction was identified to be the General Time reversible model (GTR with a lowest Bayesion Information Criterion of 123336.358). The non-uniformity of evolutionary rates among substitution sites was modeled using a discrete Gamma distribution with five rate categories (+G, parameter = 0.5494). The log likelihood value of the final tree was -61963.20.

RESULTS

Detection of a transcribed and translated antiparallel overlapping ORF

Combined RNAseq and RIBOseq data of *E. coli* O157:H7 Sakai grown at three different growth conditions were searched for ORFs, which are antiparallel overlapping to annotated genes, and show signals for transcription and translation. Further, the translatability of the ORFs was calculated using the ribosomal coverage value (RCV), which is defined as the quotient of translatome RPKM over transcriptome RPKM [43]. The annotated ECs5115 (1,539 bp) encodes a transcriptional regulator of the CadC family, and shows signals for transcription (RNAseq) and translation (RIBOseq) on the sense and antisense strand (Figure 1). The latter reads correspond to a small ORF completely embedded within ECs5115 in the reading frame -2 relative to ECs5115, encoding a short hypothetical protein of 41 AAs. A blastp search for annotated homologs resulted in a single hit to a hypothetical protein of *Escherichia albertii* TW07627 (gene bank accession number EDS93387.1) with an e-value of 5×10^{-13} and 78% identity. The software PredictProtein could not detect transmembrane helices or disulfide bonds, and the hypothetical protein is predicted to be secreted. The strongest transcription of the ORF was found in BHI medium at 37°C (Table 1A), whereas translation (RPKM) and translatability (RCV) are highest in LB at 37°C. ECs5115 is only weakly transcribed, and read numbers decrease over the length of the gene (Figure 1 + Table 1B). Translation of both reading frames, ECs5115 and *laoB* is almost completely switched off at combined cold and osmotic stress.

Table 1: Transcription and translation of *laoB* (part 1A) and its mother gene ECs5115 (part 1B) at the three different growth conditions indicated. The RPKM values of the transcriptome (RNAseq) and the translatome (RIBOseq) data for the overlapping novel gene and annotated mother gene are listed, including the RCV, indicating their translatability. ORF coverage is the fraction of a gene sequence, which is covered by RIBOseq reads. In addition, the corresponding data for the putative overlapping gene *laoA* (compare Fig. 2A) are shown (part 1C).

1A *laoB*				
condition	**RPKM transcriptome***	**RPKM translatome***	**ribosomal coverage value***	**ORF coverage***
LB, 37°C	29.5	194	6.83	0.7
BHI, 37°C	49.4	23.6	0.48	0.6
BHI + 4% NaCl, 14°C	28.3	0.2	0.01	0.07

1B ECs5115				
condition	**RPKM transcriptome***	**RPKM translatome***	**ribosomal coverage value***	**ORF coverage***
LB, 37°C	19	12.3	0.65	0.35
BHI, 37°C	26.9	7.1	0.27	0.45
BHI + 4% NaCl, 14°C	10.2	0.6	0.07	0.16

1C *laoA*				
condition	**RPKM transcriptome***	**RPKM translatome***	**ribosomal coverage value***	**ORF coverage***
LB, 37°C	38.3	9.6	0.27	0.51
BHI, 37°C	36.3	1.9	0.05	0.37
BHI + 4% NaCl, 14°C	12.7	0.4	0.02	0.16

*Mean values of the two biological replicates are shown.

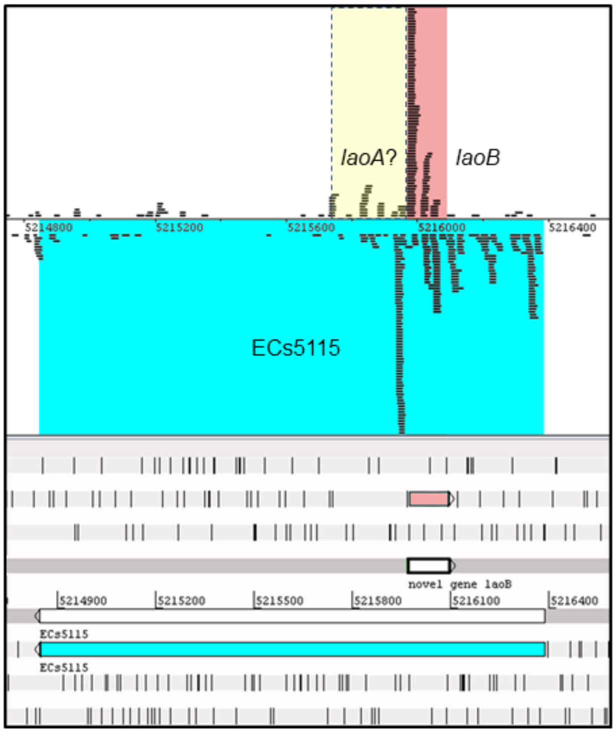
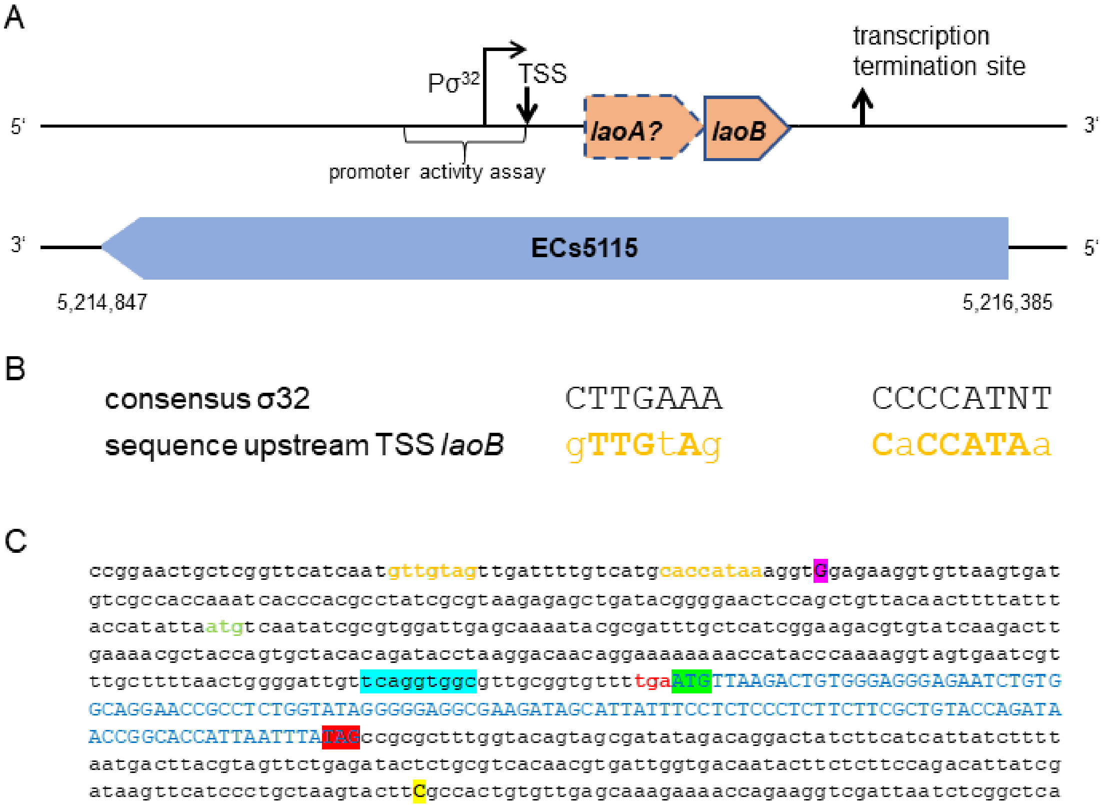


Figure 1: Translation of *laoB* in LB medium. RIBOseq reads mapped strand-specifically to the overlapping gene pair *laoB*/ECs5115 are visualized in Artemis. The annotated gene ECs5115 is highlighted in blue. The novel gene *laoB* is highlighted in pink. A potential, non-characterized overlapping gene *laoA* is high-lighted in yellow.

Characterization of *laoB* promoter region

A predicted SD sequence (ΔG° of -6.8) is present 15 bp upstream of the putative start codon (Figure 2C). A single transcriptional start site was identified 289 bp upstream of the start codon by 5’RACE. This would imply a very long 5’UTR. Therefore, the region was searched for additional ORFs. Indeed, another ORF (*laoA*), which would encode a protein of 61 AA, is located directly upstream of the OLG *laoB* (Figure 2A). However, this ORF is only weakly translated (Figure 1 + Table 1C). Furthermore, it does not have annotated homologs, and in its upstream region, no SD sequence was detected. Thus, this ORF was not characterized further. While FindTerm does not predict a ρ-independent terminator in the region 300 bp downstream of the stop codon, a transcriptional stop site was determined 155 bp downstream of the stop codon by 3’RACE (Figure 2C).



```
ccggaactgctcggttcatcaatgttgtagttgattttgtcatgcaccataaaggtGgagaaggtgttaagtgat
gtcgccaccaaatcacccacgcctatcgcgtaagagagctgatacggggaactccagctgttacaacttttattt
accatattaatgtcaatatcgcgtggattgagcaaaatacgcgatttgctcatcggaagacgtgtatcaagactt
gaaaacgctaccagtgctacacagatacctaaggacaacaggaaaaaaaaccatacccaaaaggtagtgaatcgt
ttgcttttaactggggattgttcaggtggcgttgcggtgttttgaATGTTAAGACTGTGGGAGGGAGAATCTGTG
GCAGGAACCGCCTCTGGTATAGGGGGAGGCGAAGATAGCATTATTTCCTCTCCCTCTTCTTCGCTGTACCAGATA
ACCGGCACCATTAATTTATAGccgcgctttggtacagtagcgatatagacaggactatcttcatcattatctttt
aatgacttacgtagttctgagatactctgcgtcacaacgtgattggtgacaatacttctcttccagacattatcg
ataagttcatccctgctaagtacttCgccactgtgttgagcaaagaaaaccagaaggtcgattaatctcggctca
```

Figure 2: The overlapping gene pair *laoB*/ECs5115 and the regions upstream and downstream. **A** Schematic view of the overlapping gene pair *laoAB*/ECs5115. Positions of the promoter, TSS, and transcription termination are indicated (not to scale). **B** Alignment of the proposed σ^{32} promoter to the consensus motif. **C** DNA sequence of the novel gene *laoB* and the upstream and downstream regions. The sequence of *laoB* is colored in blue and written in capital letters. The start codon is highlighted in green and the stop codon in red. Also the start and stop codon of the potential upstream gene *laoA* are marked (lower case, green and red, respectively). The TSS detected by 5'RACE is highlighted in pink, and the transcriptional stop determined by 3'RACE is highlighted in yellow. The putative σ^{32} promoter is colored in orange. The Shine-Dalgarno sequence upstream *laoB* is highlighted in light blue.

The software BPROM did not predict a σ^{70} promoter in the upstream region of *laoB* in a suitable distance to the TSS. Therefore, the region upstream of the TSS was manually investigated for the presence of alternative σ-factor consensus motifs. Interestingly, a sequence with high similarity to the σ^{32} consensus motif was detected in proper distance to the TSS (Figure 2B). The sequence 300 bp upstream of the TSS, containing the potential σ^{32} promoter, was cloned into pProbe-NT for investigation of promoter activity at different growth conditions. Significant promoter activity was detectable at all conditions tested (Figure 3A). LB supplemented with 400 mM NaCl lead to the highest fluorescence intensity with a 2.9-fold increase compared to LB. Additionally, the conditions LB + 2.5 mM malonic acid and LB at pH 5 showed a significantly increased promoter activity. Promoter activity was reduced in LB supplemented with 10 mM L-arginine of about 1.3-fold. However, LB medium already contains L-arginine. Therefore, the experiment was repeated in depleted MOD medium without L-arginine and the convertible AAs, which leads to a more pronounced decline of promoter activity (Figure 3B).

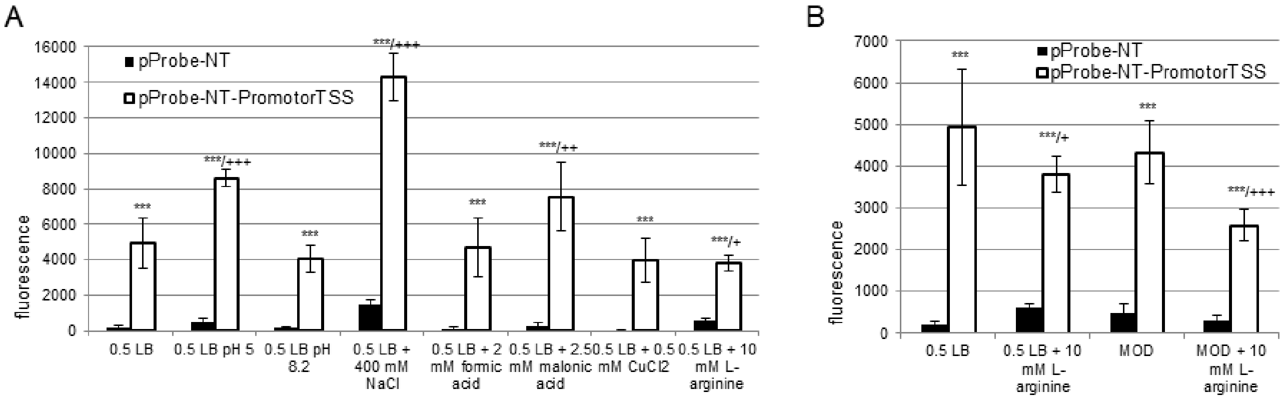


Figure 3: Promoter activity of the region 300 bp upstream of the *laoB* TSS. Significant changes between vector control and pProbe-NT-PromoterTSS are marked with asterisks (*** p<0.001). Significant differences between 0.5 LB and investigated stress conditions are marked with pluses (+ p<0.05, ++ p<0.01, +++ p<0.001). **A** Promoter activity of the region 300 bp upstream of the determined TSS in LB medium with different supplementations. **B** Promoter activity of the region 300 bp upstream of the TSS in modified MOD medium and in MOD medium supplemented with 10 mM L-arginine.

Expression of a LaoB-EGFP fusion protein

Next, it was investigated whether the LaoB protein can be expressed in *E. coli.* For this purpose, the *laoB* sequence was cloned in-frame and upstream of EGFP, and transformed into *E. coli* Top10. After induction with IPTG, a fluorescent LaoB-EGFP fusion protein was produced. The induced culture shows an 11.7-fold increased fluorescence intensity compared to the uninduced culture demonstrating expression of the fusion protein (Figure 4).

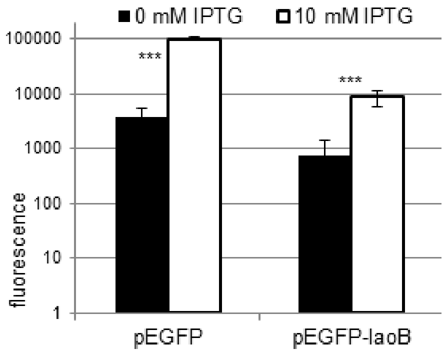


Figure 4: Overexpression of LaoB C-terminally fused to EGFP. *E. coli* Top10 was transformed with the empty pEGFP vector as positive control (left) and with pEGFP-LaoB (right). Fluorescence values in logarithmic scale without (black) and with (white) induction are depicted. Expression of the fusion protein was induced with 10 mM IPTG after adjusting the OD_{600} to 0.6. The experiment was performed in triplicate. *** p<0.001.

The translationally arrested mutant Δ*laoB* shows a growth advantage in arginine-containing media

For functional characterization of *laoB* the knock-out mutant Δ*laoB* was created using genome editing [25]. A premature stop codon at the fifth codon of *laoB* was generated by a point mutation (Figure 5A). Two additional point mutations were introduced in adjacent nucleotides to delete an *MnI*I restriction enzyme cut site (required for easier selection). The AA sequence of ECs5115 is not changed by the point mutations, because the affected codon of the mother gene still encodes serine.

To find a potential phenotype, competitive growth experiments with EHEC wild type and Δ*laoB* were performed. The equal-ratio mixture of wild type and mutant was incubated under different conditions, and a potential growth advantage was determined by the ratio of the wild type and mutant genes at the endpoint. When LB medium was supplemented with 20 mM L-arginine, a phenotype was detected: EHEC Δ*laoB* displayed a significant growth advantage indicated by a ratio of wild type to mutant of 15:85 (Figure 5B). Thus, the competitive index is 13.6. No phenotype was found for any other conditions tested (Additional file S3).

Intact *laoB*, cloned into pBAD-*myc*-*His*-C, should restore the phenotype of EHEC wild type. Therefore, competitive growth experiments using EHEC Δ*laoB* carrying pBAD-*laoB* against EHEC Δ*laoB* + pBAD-Δ*laoB* (mutant control) were performed. Expression of *laoB* was induced with arabinose. As expected, in LB, the ratio between the two strains tested did not change independent whether the plasmid borne *laoB* was induced or not (Figure 5C). In contrast, in LB supplemented with 20 mM L-arginine, the strain carrying the functional *laoB*-copy shows a significant growth disadvantage if induced with arabinose. This reflects the competitive growth phenotype of the wild type strain compared to the mutant strain (Figure 5B). Thus, translation arrested *laoB* can be complemented *in trans.*

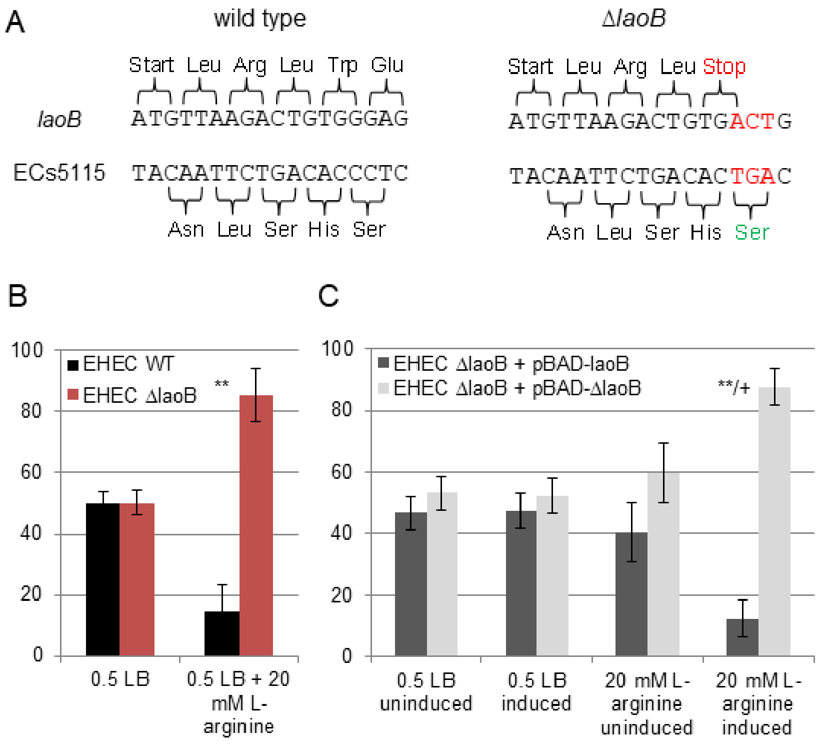


Figure 5: Nucleotide sequence and phenotype of EHEC Δ*laoB*. **A** Construction of a translationally arrested Δ*laoB* mutant. Introduction of a point mutation in the DNA sequence of *laoB* changed the fifth codon encoding glutamine to a premature stop codon. Because of two adjacent mutations, a cut site for the restriction enzyme *Mnl*I is deleted at this position. The three point mutations do not influence the AA sequence of the antiparallel overlapping annotated gene ECs5115. **B** Ratio in percent of EHEC wild type to EHEC Δ*laoB* after competitive growth. Wild type and mutant were mixed in equal ratios and after 18 h incubation at different growth conditions, their ratio was determined. In 0.5 LB, no change compared to the inoculation ratio occurred, but when the medium was supplemented with 20 mM L-arginine, EHEC Δ*laoB* shows a significant growth advantage. The experiment was performed in triplicate. ** $p<0.01$. **C** Complementation of EHEC Δ*laoB* using a plasmid-borne *laoB*. The diagram shows the ratios in percent of EHEC Δ*laoB* + pBAD-*laoB* and EHEC Δ*laoB* + pBAD-Δ*laoB* after competitive growth. The experiment was performed in triplicate. Significant changes between uninduced and induced conditions are marked with a plus (+ $p<0.05$). Significant changes between 0.5 LB and 0.5 LB + 20 mM L-arginine are marked with asterisks (** $p<0.01$).

Phylostratigraphic analysis of *laoB*

Two tblastn searches with ECs5115 and *laoB* as queries were performed to determine the taxonomic distribution of both genes. *LaoB* was only detected in a few *Escherichia* and *Shigella* strains (Figure 6 + Additional file S5), whereas the antiparallel overlapping, annotated ECs5115 has homologs in multitude bacterial phyla (Additional file S4). However, in those sequences the embedded *laoB* ORF is disintegrated due to several point mutations and indels leading to frame shifts and stop codons. When a *laoB* homolog is present, its sequence is always highly conserved, showing only a few AA substitutions, but no premature stop codons or frameshift mutations. The highest sequence variability occurs in *E. fergusonii* (Figure 6).

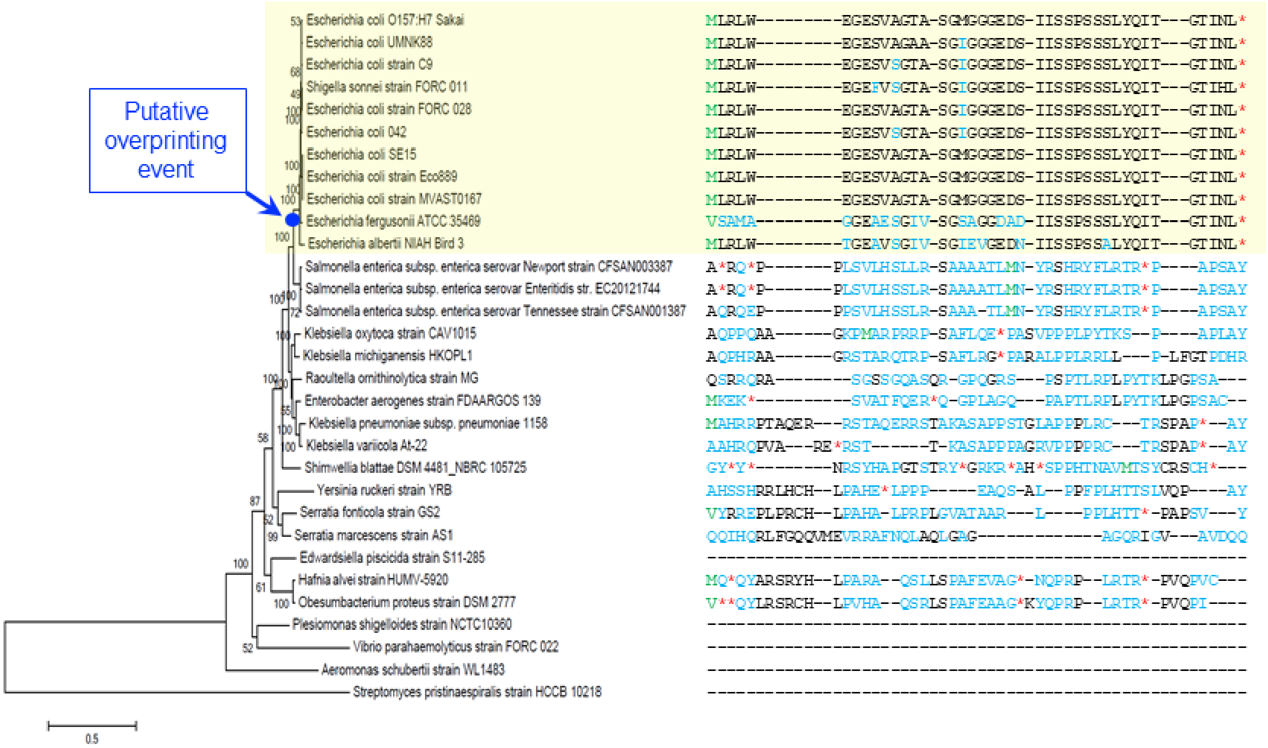


Figure 6: Phylogenetic tree and alignment of *laoB*. The phylogenetic tree to the left was constructed based on a concatemer of 16S RNA, *atpD*, *adk*, *gyrB*, *purA*, and *recA.* To the right, the different amino acid sequences of LaoB are aligned. Start codons are colored in green, AA changes in blue, and stop codons (*) in red.

DISCUSSION

This study provides evidence for a novel overlapping gene pair, *laoB*/ECs5115, in *E. coli* O157:H7 Sakai. Transcription and translation of a short ORF, embedded in the antisense reading frame -2 to a CadC-like transcriptional regulator, was detected by RNAseq and RIBOseq at optimal growth conditions. Translational knock-out of the ORF by a premature stop codon resulted in a significant growth advantage of the mutant strain in LB medium supplemented with L-arginine over the wild type strain in competitive growth. Consistently, the activity of the putative σ^{32} promoter is repressed by L-arginine. Whether *laoB* is part of an overlapping operon together with *laoA*, located upstream of *laoB*, is unknown. *LaoA* was not examined, since transcription and translation of *laoA* appeared to be very weak under the conditions tested.

Is *laoB* a protein-coding gene?

LaoB might function as a novel ncRNA instead of a novel protein-coding gene. However, due to the following reasons this appears to be unlikely: first, the same ORF has been annotated in *E. albertii* as a protein-coding gene, which demonstrates that genome annotation programs recognize this reading frame as potentially protein coding. Second, 15 bp upstream of the start codon a SD sequence is present (Figure 2C). The distance of the SD to the start codon is within the natural ranges observed and the detected sequence is close to the SD consensus motif, resulting in strong ribosomal binding [32]. Third, experimental data confirm the protein-coding character of *laoB*, since the ORF is covered by RIBOseq reads (Figure 1). RIBOseq signals clearly indicate active translation of an RNA molecule [44, 45]. In LB medium, the ORF has a very high RCV (Table 1A), which is much higher than the mean RCV of 1.55, which we found for short annotated EHEC genes [26, 43]. Also, stable translation into a protein was further confirmed by the expression of a LaoB-EGFP fusion protein (Figure 4). Fourth, a translationally arrested mutant lead to a clear phenotype, which could be complemented by the wild type sequence *in trans* by using just the *laoB* ORF without any adjacent sequence attached (Figure 5). If *laoB* would function as an antisense RNA, it would regulate its targets by base pairing with complementary mRNAs [46]. It appears to be unlikely that a translationally arrested mutant, which changes only ~0.5% of the nucleotides compared to the complete transcript of the *laoB* sequence, would exert such a dramatic phenotype.

Putative function of LaoB

The results presented in this work provide first hints towards a potential LaoB function. The region 300 bp upstream of the TSS determined shows significant promoter activity at all investigated conditions (Figure 3). The *laoB* promoter is probably recognized by the alternative σ-factor σ^{32}, since a sequence very similar to the σ^{32} consensus motif is present in the proper distance to the TSS (Figure 2B) [30]. The first T of the -35 box and the A of the -10 box are completely conserved in σ^{32} promoters, and both nucleotides are present in the σ^{32} promoter region of *laoB*. Additionally, the spacer between the -35 and -10 box has the optimal distance of 14 bp, and σ^{32} promoters with this spacer distance tolerate a substitution of the second C of the tetra-C motif of the -10 box without losing promoter strength [47], which is also the case here. In addition, the distance between the -10 box and the TSS is in the optimal range of 6 bp [30]. Transcription of heat shock genes is induced by σ^{32}. Accordingly, transcription of *laoB* is almost switched off at cold stress (Table 1A). The σ^{32} stress regulon includes chaperons, transcription factors, DNA/RNA surveillance proteins, and many membrane-associated proteins [30]. In this study, the promoter has the highest activity in LB supplemented with NaCl and at acidic conditions (Figure 3). Interestingly, σ^{32} is also the master regulator of the transcription factor PhoPQ, which is also induced at acid stress [30].

In our hands, EHEC Δ*laoB* only showed a clear phenotype after supplementing the medium with L-arginine (Figure 5B). As a proteinogenic AA, L-arginine is involved in many central metabolic pathways. Bacteria synthesize L-arginine from glutamate [48], or take it up from the environment by three different transporters [49]. Arginine can be utilized as sole carbon and nitrogen source, and is the substrate for the synthesis of polyamines [48]. Here, high L-arginine concentrations resulted in a significantly reduced activity of the *laoB* promoter, and the EHEC wild type has a clear growth disadvantage in competitive growth. These observations would agree with the speculation that LaoB might be involved in enhancing L-arginine uptake. In many EHEC reservoirs, nutrient concentrations, including L-arginine concentrations, are low and efficient uptake represents an advantage. The high arginine concentrations used in this study are unlikely to occur naturally. Therefore, under environmental conditions, which are low in arginine, intact LaoB may confer a growth advantage. The hypothesis that LaoB somehow interacts with arginine transport is supported by the facts that a high proportion of small proteins – LaoB has a size of only 41 AAs - associates with the cell membrane, in which transporters are located [50, 51], and that the σ^{32} regulon includes many membrane proteins [30]. However, testing this speculation and further functional characterization of LaoB must await future studies.

Origin of *laoB* by overprinting

The time of origin of an OLG can be estimated by phylostratigraphic analysis, comparing the phylogenetic distribution of the mother gene and the overlapping gene [18, 52]. The intact *laoB* ORF is only present in *Escherichia* and *Shigella* strains (Figure 6), while the annotated gene ECs5115 has a much broader taxonomical distribution (i.e., higher conservation level), and is present in both Gram-negative and Gram-positive bacteria (Additional file S4). It is concluded that *laoB* originated recently and might be an interesting example of *de novo* gene birth by overprinting [18, 52, 53]. This would mean that a number of point mutations in the ECs5115 sequence would have created the *laoB* ORF including its regulatory sequences after the *Escherichia/Shigella* clade separated from *Salmonella*. One may postulate that a weak σ^{32} promoter sequence was already present at the proper location by chance and, later, may have been further optimized by additional point mutations leading to an increased transcription of the novel ORF. The resulting (m)RNA must have been used as template for translation, perhaps based on a weak ribosomal binding site, which happened to be present upstream of the start codon. Now, one must assume that the AA chain, at this point, was functional *ab initio* by chance, conferring a fitness advantage to the cell. At this early evolutionary stage, a novel gene is volatile and the process is reversible, such that the novel ORF can get lost again [54]. A fitness gain related to the L-arginine metabolism may have led to fixation of the functional allele in the population by Darwinian evolution. Because EHEC colonizes many hosts and environments [55], which requires expression of different sets of genes [56, 57], LaoB might improve its fitness in one of those species specific niches. Alternatively, the novel ORF could have been fixed by neutral evolution together with the mutated mother gene [58]. Later on, extension at the 3’ end by the loss of a stop codon may occur, leading to an elongation of the novel protein, which would be more likely than 5’ elongation due to regulatory elements in the 5’UTR [59]. This speculative order of events has some similarities to the proto-gene hypothesis of Carvunis *et al* [39], which deals with the potential *de novo* origin of short genes in intergenic regions of the yeast *S. cerevisiae*.

In EHEC, only two other antiparallel overlapping gene pairs, in which a young gene also may have originated recently *de novo* by overprinting, have been characterized functionally [13, 14]. Nevertheless, *de novo* birth of genes in antisense to annotated genes may be more frequent than presumed in the past, as has also been suggested by

Haycocks and Grainger [60] based on the frequent binding of transcriptional regulator in intragenic locations. In contrast to duplication followed by neofunctionalization or sub-functionalization, which is the established theory for the origin of new genes [20], but produces just variants of existing sequences, overprinting would allow for the rapid creation of true novelty [61].

CONCLUSION

Strand-specific RNAseq and RIBOseq are well suited to identify translated ORFs located in antisense to annotated genes. Frequent antisense transcription is observed in all RNAseq experiments, but almost all signals have been interpreted as ncRNA [62]. However, RIBOseq already confirmed translation of many antisense RNAs in eukaryotes [63-65], and this method identified numerous overlooked small genes in the intergenic regions of different bacteria [66-68]. Therefore, improved genome annotation algorithms are required which do not systematically dismiss small and/or overlapping genes [8, 69, 70]. Integration of transcriptomic, translatomic, and other experimental data into annotation pipelines would increase specificity and sensitivity for the prediction of novel small genes [71-73]. Additionally, improved proteomic methods are necessary, which do not miss small non-annotated proteins [74, 75]. In any case, functional characterization of novel short genes overlooked to date presents a major future challenge to experimental microbiology. In this paper, we provide evidence for a small protein encoded in antisense to an annotated protein-coding gene as well as its initial functional characterization. We assume that the number of small protein-coding genes located in antisense to annotated genes in EHEC may be significant and that their origin by overprinting may be a frequent mechanism of *de novo* gene birth, possibly in bacteria in general.

REFERENCES

1. Krakauer DC: **Stability and evolution of overlapping genes**. *Evolution* 2000, **54**(3):731-739.
2. Caliskan N, Katunin VI, Belardinelli R, Peske F, Rodnina MV: **Programmed–1 Frameshifting by Kinetic Partitioning during Impeded Translocation**. *Cell* 2014, **157**(7):1619-1631.
3. Meydan S, Klepacki D, Karthikeyan S, Margus T, Thomas P, Jones JE, Khan Y, Briggs J, Dinman JD, Vazquez-Laslop N *et al*: **Programmed Ribosomal Frameshifting Generates a Copper Transporter and a Copper Chaperone from the Same Gene**. *Mol Cell* 2017, **65**(2):207-219.
4. Sharma V, Firth AE, Antonov I, Fayet O, Atkins JF, Borodovsky M, Baranov PV: **A pilot study of bacterial genes with disrupted ORFs reveals a surprising profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment**. *Mol Biol Evol* 2011:msr155.

5. Merino E, Balbas P, Puente JL, Bolivar F: **Antisense overlapping open reading frames in genes from bacteria to humans**. *Nucleic Acids Res* 1994, **22**(10):1903-1908.

6. Mir K, Neuhaus K, Scherer S, Bossert M, Schober S: **Predicting statistical properties of open reading frames in bacterial genomes**. *PloS one* 2012, **7**(9):e45103.

7. Veloso F, Riadi G, Aliaga D, Lieph R, Holmes DS: **Large-scale, multi-genome analysis of alternate open reading frames in bacteria and archaea**. *Omics* 2005, **9**(1):91-105.

8. Firth AE, Brown CM: **Detecting overlapping coding sequences with pairwise alignments**. *Bioinformatics* 2005, **21**(3):282-292.

9. Lèbre S, Gascuel O: **The combinatorics of overlapping genes**. *J Theor Biol* 2017, **415**:90-101.

10. Johnson ZI, Chisholm SW: **Properties of overlapping genes are conserved across microbial genomes**. *Genome Res* 2004, **14**(11):2268-2272.

11. Tunca S, Barreiro C, Coque JJ, Martin JF: **Two overlapping antiparallel genes encoding the iron regulator DmdR1 and the Adm proteins control siderophore and antibiotic biosynthesis in *Streptomyces coelicolor* A3(2)**. *FEBS J* 2009, **276**(17):4814-4827.

12. Silby MW, Levy SB: **Overlapping protein-encoding genes in *Pseudomonas fluorescens* Pf0-1**. *PLoS Genet* 2008, **4**(6):e1000094.

13. Fellner L, Bechtel N, Witting MA, Simon S, Schmitt-Kopplin P, Keim D, Scherer S, Neuhaus K: **Phenotype of *htgA* (*mbiA*), a recently evolved orphan gene of *Escherichia coli* and *Shigella*, completely overlapping in antisense to *yaaW***. *FEMS Microbiol Lett* 2014, **350**(1):57-64.

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

15. Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D: **Overlapping genes produce proteins with unusual sequence properties and offer insight into *de novo* protein creation**. *J Virol* 2009, **83**(20):10719-10736.

16. Loessner MJ, Gaeng S, Scherer S: **Evidence for a holin-like protein gene fully embedded out of frame in the endolysin gene of *Staphylococcus aureus* bacteriophage 187**. *J Bacteriol* 1999, **181**(15):4452-4460.

17. Chirico N, Vianelli A, Belshaw R: **Why genes overlap in viruses**. *Proc Royal Soc B: Biol Sci* 2010, **277**(1701):3809-3817.

18. Keese PK, Gibbs A: **Origins of genes: "big bang" or continuous creation?** *Proceedings of the National Academy of Sciences of the United States of America* 1992, **89**(20):9489-9493.

19. Boldogköirid Z: **Coding in the noncoding DNA strand: A novel mechanism of gene evolution?** *J Mol Evol* 2000, **51**(6):600-606.

20. Kondrashov FA: **Gene duplication as a mechanism of genomic adaptation to a changing environment**. *Proceedings Biological sciences / The Royal Society* 2012, **279**(1749):5048-5057.

21. Hücker SM, Simon S, Scherer S, Neuhaus K: **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* 2017, **364**(2).

22. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T *et al*: **Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12**. *DNA research : an international journal for rapid publication of reports on genes and genomes* 2001, **8**(1):11-22.

23. Miller WG, Leveau JH, Lindow SE: **Improved *gfp* and *inaZ* broad-host-range promoter-probe vectors.** *Mol Plant Microbe Interact* 2000, **13**(11):1243-1250.

24. Rosenfeld E, Duport C, Zigha A, Schmitt P: **Characterization of aerobic and anaerobic vegetative growth of the food-borne pathogen Bacillus cereus F4430/73 strain**. *Canadian journal of microbiology* 2005, **51**(2):149-158.

25. Kim J, Webb AM, Kershner JP, Blaskowski S, Copley SD: **A versatile and highly efficient method for scarless genome editing in Escherichia coli and Salmonella enterica**. *BMC Biotechnol* 2014, **14**:84.

26. Hücker SM, Goldberg T, Schafferhans A, Bernhofer M, Vestergaard G, Nelson CW, Ardern Z, Rost B, Scherer S, Neuhaus K: **Discovery of numerous novel small genes in the intergenic regions of the Escherichia coli O157:H7 Sakai genome**. *PloS one* 2017.

27. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation**. *Bioinformatics* 2000, **16**(10):944-945.

28. Solovyev VV, Tatarinova TV: **Towards the integration of genomics, epidemiological and clinical data**. *Genome medicine* 2011, **3**(7):48.

29. Yu HH, Di Russo EG, Rounds MA, Tan M: **Mutational analysis of the promoter recognized by Chlamydia and Escherichia coli sigma(28) RNA polymerase**. *J Bacteriol* 2006, **188**(15):5524-5531.

30. Nonaka G, Blankschien M, Herman C, Gross CA, Rhodius VA: **Regulon and promoter analysis of the *E. coli* heat-shock factor, σ32, reveals a multifaceted cellular response to heat stress**. *Genes Dev* 2006, **20**(13):1776-1789.

31. Burrows PC, Severinov K, Ishihama A, Buck M, Wigneshweraraj SR: **Mapping sigma 54-RNA polymerase interactions at the -24 consensus promoter element**. *J Biol Chem* 2003, **278**(32):29728-29743.

32. Ma J, Campbell A, Karlin S: **Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures**. *J Bacteriol* 2002, **184**(20):5733-5745.

33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

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

35. Rost B, Sander C: **Combining evolutionary information and neural networks to predict protein secondary structure**. *Proteins: Structure, Function, and Bioinformatics* 1994, **19**(1):55-72.

36. Bernhofer M, Kloppmann E, Reeb J, Rost B: **TMSEG: Novel prediction of transmembrane helices**. *Proteins* 2016, **84**(11):1706-1716.

37. Ceroni A, Passerini A, Vullo A, Frasconi P: **DISULFIND: a disulfide bonding state and cysteine connectivity prediction server**. *Nucleic Acids Res* 2006, **34**(suppl 2):W177-W181.

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

39. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B *et al*: **Proto-genes and *de novo* gene birth**. *Nature* 2012, **487**(7407):370–374.

40. Domazet-Lošo T, Tautz D: **An ancient evolutionary origin of genes associated with human genetic diseases**. *Mol Biol Evol* 2008, **25**(12):2699-2707.

41. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: Molecular Evolutionary Genetics Analysis version 6.0**. *Mol Biol Evol* 2013, **30**(12):2725-2729.

42. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park YM, Buso N, Lopez R: **The EMBL-EBI bioinformatics web and programmatic tools framework**. *Nucleic Acids Res* 2015, **43**(W1):W580-W584.

43. Neuhaus K, Landstorfer R, Simon S, Schober S, Wright PR, Smith C, Backofen R, Wecko R, Keim DA, Scherer S: **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* 2017, **18**(1):216.

44. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS: **Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling**. *Science* 2009, **324**(5924):218-223.

45. Liu B, Qian SB: **Characterizing inactive ribosomes in translational profiling**. *Translation* 2016, **4**(1):e1138018.

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

47. Koo BM, Rhodius VA, Campbell EA, Gross CA: **Dissection of recognition determinants of Escherichia coli sigma32 suggests a composite -10 region with an 'extended -10' motif and a core -10 element**. *Mol Microbiol* 2009, **72**(4):815-829.

48. Cunin R, Glansdorff N, Pierard A, Stalon V: **Biosynthesis and metabolism of arginine in bacteria**. *Microbiological reviews* 1986, **50**(3):314-352.

49. Wissenbach U, Six S, Bongaerts J, Ternes D, Steinwachs S, Unden G: **A third periplasmic transport system for L-arginine in Escherichia coli: molecular characterization of the artPIQMJ genes, arginine binding and transport**. *Mol Microbiol* 1995, **17**(4):675-686.

50. Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE: **Small membrane proteins found by comparative genomics and ribosome binding site models**. *Mol Microbiol* 2008, **70**(6):1487-1501.

51. Kemp G, Cymer F: **Small membrane proteins–elucidating the function of the needle in the haystack**. *Biol Chem* 2014, **395**(12):1365-1377.

52. Pavesi A, Magiorkinis G, Karlin DG: **Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of Deltaretroviruses**. *PLoS Comput Biol* 2013, **9**(8):e1003162.

53. Delaye L, Deluna A, Lazcano A, Becerra A: **The origin of a novel gene through overprinting in *Escherichia coli***. *BMC Evol Biol* 2008, **8**(1):31.

54. Huvet M, Stumpf MP: **Overlapping genes: a window on gene evolvability**. *BMC Genomics* 2014, **15**(1):721.

55. Lim JY, Yoon J, Hovde CJ: **A brief overview of *Escherichia coli* O157:H7 and its plasmid O157**. *J Microbiol Biotechnol* 2010, **20**(1):5-14.

56. Duffitt AD, Reber RT, Whipple A, Chauret C: **Gene expression during survival of *Escherichia coli* O157:H7 in soil and water**. *Int J Microbiol* 2011, **2011**:340506.

57. Landstorfer R, Simon S, Schober S, Keim D, Scherer S, Neuhaus K: **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* 2014, **15**:353.

58. Lillo F, Krakauer DC: **A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes**. *Biology direct* 2007, **2**(1):22.

59. Fonseca MM, Harris DJ, Posada D: **Origin and Length Distribution of Unidirectional Prokaryotic Overlapping Genes**. *G3: Genes| Genomes| Genetics* 2014, **4**(1):19-27.

60. Haycocks JR, Grainger DC: **Unusually Situated Binding Sites for Bacterial Transcription Factors Can Have Hidden Functionality**. *PloS one* 2016, **11**(6):e0157016.

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

62. Dornenburg JE, Devita AM, Palumbo MJ, Wade JT: **Widespread Antisense Transcription in *Escherichia coli***. *mBio* 2010, **1**(1).

63. Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MA, Brocard M, Couso JP: **Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq**. *eLife* 2014, **3**:e03528.

64. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC *et al*: **Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation**. *EMBO J* 2014, **33**(9):981-993.

65. Carlevaro-Fita J, Rahim A, Guigo R, Vardy LA, Johnson R: **Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells**. *RNA* 2016, **22**(6):867-882.

66. Jeong Y, Kim JN, Kim MW, Bucca G, Cho S, Yoon YJ, Kim BG, Roe JH, Kim SC, Smith CP *et al*: **The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3(2)**. *Nature communications* 2016, **7**:11605.

67. Baek J, Lee J, Yoon K, Lee H: **Identification of Unannotated Small Genes in *Salmonella***. *G3* 2017, **7**(3):983-989.

68. Neuhaus K, Landstorfer R, Fellner L, Simon S, Marx H, Ozoline O, Schafferhans A, Goldberg T, Rost B, Küster B *et al*: **Translatomics combined with transcriptomics and proteomics reveals novel functional, recently evolved orphan genes in *Escherichia coli* O157:H7 (EHEC)**. *BMC Genomics* 2016, **7**:133.

69. Warren AS, Archuleta J, Feng WC, Setubal JC: **Missing genes in the annotation of prokaryotic genomes**. *BMC Bioinformatics* 2010, **11**:131.

70. Oheigeartaigh SS, Armisen D, Byrne KP, Wolfe KH: **SearchDOGS bacteria, software that provides automated identification of potentially missed genes in annotated bacterial genomes**. *J Bacteriol* 2014, **196**(11):2030-2042.

71. Karp PD, Keseler IM, Shearer A, Latendresse M, Krummenacker M, Paley SM, Paulsen I, Collado-Vides J, Gama-Castro S, Peralta-Gil M *et al*: **Multidimensional annotation of the Escherichia coli K-12 genome**. *Nucleic Acids Res* 2007, **35**(22):7577-7590.

72. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A *et al*: **RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more**. *Nucleic Acids Res* 2013, **41**(Database issue):D203-213.

73. Olexiouk V, Crappe J, Verbruggen S, Verhegen K, Martens L, Menschaert G: **sORFs.org: a repository of small ORFs identified by ribosome profiling**. *Nucleic Acids Res* 2016, **44**(D1):D324-329.

74. Zur H, Aviner R, Tuller T: **Complementary Post Transcriptional Regulatory Information is Detected by PUNCH-P and Ribosome Profiling**. *Scientific reports* 2016, **6**.

75. Willems P, Ndah E, Jonckheere V, Stael S, Sticker A, Martens L, Van Breusegem F, Gevaert K, Van Damme P: **N-terminal proteomics assisted profiling of the unexplored translation initiation landscape in Arabidopsis thaliana**. *Mol Cell Proteomics* 2017.

2.9 Publication 4: The novel, antiparallel overlapping gene pair *anoG*/ECs2385 in *Escherichia coli* O157:H7 Sakai

The novel, antiparallel overlapping gene pair *anoG*/ECs2385 in *Escherichia coli* O157:H7 Sakai

Sarah M. Hücker[1], Sonja Vanderhaeghen[1], Isabel Abellan-Schneyder[1], Romy Wecko[1], Siegfried Scherer[1,2] and Klaus Neuhaus[1,3*]

[1]Chair for Microbial Ecology, Technische Universität München, Freising, Germany; [2]ZIEL – Institute for Food & Health, Technische Universität München, Freising, Germany; [3]Core Facility Microbiome/NGS, ZIEL – Institute for Food & Health, Technische Universität München, Freising, Germany.

*Correspondence: neuhaus@tum.de

Abstract

Standard genome annotation presumes that only one protein is encoded at a given bacterial dsDNA locus. In contrast to this assumption, transcription and translation of an overlapping open reading frame of 186 bp length were discovered by RNAseq and RIBOseq experiments. This open reading frame is completely embedded in the annotated gene ECs2385 in *Escherichia coli* O157:H7 Sakai in the antiparallel reading frame -3. The open reading frame is transcribed as part of a polycistronic mRNA, which includes the annotated upstream gene ECs2384, encoding a murein lipoprotein. The transcriptional start site of the operon resides 38 bp upstream of the ECs2384 start codon, driven by a predicted σ^{70} promoter, which is constitutively active at different growth conditions. The polycistronic operon contains a ρ-independent terminator just upstream of the novel gene, significantly decreasing its transcription. The novel gene can be stably expressed as an EGFP-fusion protein and a translationally arrested mutant shows a growth advantage under anaerobiosis in competitive growth compared to the wild type. Therefore, the novel antiparallel overlapping gene is named *anoG* – anaerobiosis responsive overlapping gene. A phylostratigraphic analysis indicates that

anoG originated recently *de novo* by overprinting after the *Escherichia/Shigella* clade separated from other enterobacteria.

1. Introduction

Escherichia coli strains are classified as EHEC when they possess Shiga-toxin genes and the locus of enterocyte effacement [1]. The EHEC strain O157:H7 Sakai was isolated from an outbreak in Japan in 1996. It has a genome of 5.5 Mb [2], which is 20% larger than the genome of *E. coli* K12, probably due to DNA acquired by horizontal gene transfer and integration of 24 prophages [1]. In humans, EHEC causes hemorrhagic colitis, and the disease can progress to the life-threatening hemolytic uremic syndrome [3]. To date, neither targeted therapy nor vaccination is available, and antibiotics even promote a fatal outcome by Shiga-toxin induction [4]. The serotype O157:H7 is the most frequent clinical isolate causing 100,000 reported infections per year in USA [5]. Transmission mainly occurs via consumption of contaminated food, e.g., undercooked beef or fresh produce, but also person-to-person and animal-to-person spread is possible [3]. Additionally, EHEC thrives in many environmental niches: while the major reservoir are cattle, EHEC also colonizes other mammals, birds, fish, insects [6], and the protozoan *Acantamoeba polyphaga* [7]. Another important reservoir are green leaf plants, where EHEC colonizes the stomata [8] and roots [9], and is even internalized in seedlings [10]. In addition, EHEC persists for several weeks in sterilized soil and water at cold temperature [11]. Cycling between those different hosts and habitats occurs frequently, and insects can serve as transmission vectors [12; 13]. These different life styles display variable challenges, and require expression of changing sets of genes.

Next Generation Sequencing is a valuable tool to investigate global gene expression at different levels. Strand-specific RNAseq allows the quantification of transcription [14]. For example, the transcriptome of EHEC strain EDL933 was determined under eleven different growth conditions (i.e., radish sprouts, cow dung, antibiotic treatment), and shows differential expression of many genes [15]. Besides signals mapping to annotated genes, RNAseq experiments resulted in many reads mapping to intergenic regions or antisense to annotated genes [16]. In the past, those signals were interpreted to represent ncRNA [17; 18] or just pervasive transcription [19]. Today, RIBOseq allows investigation of the global translatome [20] by sequencing only mRNA, which is protected by ribosomes. When RIBOseq and RNAseq are combined, the translatability of a certain open reading frame (ORF) can be determined, and ncRNA can be distinguished from protein coding mRNA [21]. Indeed, many RNAseq signals outside of annotated genes also show RIBOseq signals, leading to the discovery of hundreds of translated ORFs in eukaryotes [22; 23; 24; 25; 26; 27]. Combined RNAseq and RIBOseq also detected 130 novel genes in *Salmonella enterica* Typhimurium [28], 72 novel genes in the intergenic regions of EHEC EDL933 [29], and 465 novel genes in EHEC Sakai [30]. However, functional characterization of all those translated ORFs is largely lacking.

In this study, the enterohemorrhagic *E. coli* strain Sakai [2] is used as a model organism. In EHEC, only two antiparallel overlapping gene pairs have been characterized: *htgA/yaaW* [31; 32] and *nog1/citC* [33]. Five additional OLG pairs are known in different *E. coli* strains: *yghW/morA* [34], *pic/setB* [35], *ardD/tniA* [36], *aatS/aatC* [37], and *tnpA/astA* [38]. Here, we report experimental evidence for the third OLG pair in EHEC: The novel gene *anoG* overlaps antisense to the annotated gene ECs2385, and encodes a functional protein.

2. Material and methods

Bacterial strains and plasmids used in this study are listed in Supplementary Table S1. Oligonucleotides are listed in Supplementary Table S2.

2.1 Determination of transcriptional start site by 5'RACE

The total RNA of an overnight culture in LB medium of *Escherichia coli* O157:H7 Sakai (Genbank accession number NC_002695) [2] was isolated with Trizol. The kit 5'RACE System for Rapid Amplification of cDNA Ends, Version 2.0 (Invitrogen) was used according to the manual. After the second PCR, the dominant product was excised from the agarose gel and purified with the GenElute^{TM} Gel Extraction Kit (Sigma-Aldrich). The PCR product was Sanger sequenced by Eurofins with oligonucleotide *anoG*+56R.

2.2 RT-PCR

EHEC total RNA of 500 µl overnight culture in LB medium was isolated with Trizol. Remaining DNA was digested using 2 U TURBO^{TM} DNase (Thermo Fisher Scientific) for 1 h at 37°C. After RNA purification by ethanol precipitation, reverse transcription with 500 ng RNA as template was performed using 200 U SuperScript^{TM} III Reverse Transcriptase (Thermo Fisher Scientific) according to the manual. The obtained cDNA was used as template for a PCR with a primer pair spanning ECs2384 and *anoG*.

2.3 qRT-PCR

Relative quantification of ECs2384 and *anoG* mRNA was performed at the following conditions: 0.5xLB at 37°C aerobically, OD_{600}=0.5 and 0.5xLB at 37°C anaerobically, OD_{600}=0.5. RNA of 2 ml EHEC culture was isolated using the RNeasy Mini Kit (Qiagen). Cell lysis was performed using 200 µl 15 mg/ml lysozyme in TE buffer at pH 8. Then, 15 µl 20 mg/ml proteinase K were added, and the sample was incubated for 15 min at room temperature. The following steps were performed according to the manual except the on-column DNase digestion was skipped, and instead 10 µg RNA were incubated with 2 U TURBO^{TM} DNase (Thermo Fisher Scientific) for 1 h at 37°C. After RNA purification by ethanol precipitation, reverse transcription with 2 µg RNA as template was performed using 200 U SuperScript^{TM} III Reverse Transcriptase (Thermo Fisher Scientific) and a Random Nonamer primer (GE Healthcare) according to the manual. One µl cDNA was used as template for the qRT-PCR with the SYBR® Select Master Mix (Applied Biosystems) on a CFX96^{TM} Real-Time machine (Bio-Rad). The $\Delta\Delta$Ct method was used for quantification [39], and 16S rRNA was used as the reference gene.

2.4 Cloning of pProbe-NT EGFP reporter plasmid and determination of promoter activity

The genomic region 300 bp upstream of the determined transcriptional start site (TSS) of *anoG* was amplified by PCR, and restriction enzyme cut sites for *Sal*I and *EcoR*I were introduced. The PCR product was cloned into the plasmid pProbe-NT [40], and transformed into *Escherichia coli* Top10. The plasmid sequence was verified by Sanger sequencing (Eurofins). Overnight cultures of *E. coli* Top10 + pProbe-NT and *E. coli* Top10 + pProbe-NT-PromoterTSS were used for 1:100 inoculation of 10 ml 0.5 LB medium with 30 µg/ml kanamycin. Growth in 0.5xLB was investigated for promoter activity using the following conditions: plain medium, at pH 5, at pH 8.2, plus 400 mM NaCl, plus 0.5 mM $CuCl_2$, plus 2 mM formic acid, or plus 2.5 mM malonic acid. Cultures were incubated at 37°C and 150 rpm until an OD_{600} of 0.5 was reached (3-8 h, dependent on growth condition). Next, the cells were pelleted, washed once with PBS, and resuspended in 1 ml PBS. The OD_{600} was adjusted to 0.3 and 0.6. Four-times each 200 µl bacterial suspension were pipetted in a black microtiter plate and the fluorescence was measured (Wallac Victor^3, Perkin Elmer Life Science, excitation 485 nm, emission 535 nm, measuring time 1 s). The fluorescence of *E. coli* Top10 without plasmid was subtracted as background. Promoter activity at anaerobic conditions was determined with the following changes of the protocol: 15 ml 0.5xLB with 30 µg/ml kanamycin (investigated conditions see above) inoculated 1:100 with overnight cultures of *E. coli* Top10 + pProbe-NT or *E. coli* Top10 + pProbe-NT-PromoterTSS in tightly closed 15 ml falcon tubes were incubated at 37°C and 150 rpm until an OD_{600} of 0.5 was reached (6-12 h dependent on growth condition). Subsequently, the cultures were transferred into Schott flasks and incubated for 15 min aerobically at 37°C and 150 rpm to allow the EGFP to mature. Cell harvest and measurement of fluorescence intensity was performed as described above. The experiment was performed in triplicate. Significance of changes was calculated by Student’s t-test.

2.5 Cloning of C-terminal AnoG-EGFP fusion proteins and overexpression of AnoG proteins

The *anoG* sequence without the stop codon was amplified by PCR, and restriction enzyme cut sites for *Pst*I and *Nco*I were introduced. The PCR product was cloned into the plasmid pEGFP, and transformed into *Escherichia coli* Top10. Because the correct start codon of *anoG* is unknown, pEGFP plasmids for the possible start codons 1 (CTG), 2 (ATG), 3 (GTG) and 5 (CTG) were constructed (Figure 2). Cloning of an EGFP fusion protein for the possible start codon 4 (GTG) failed. As negative controls, also C-terminal

EGFP fusion plasmids with translationally arrested Δ*anoG* sequences for every possible start codon were cloned (see below). The plasmid sequences were verified by Sanger sequencing (Eurofins). For fusion protein overexpression, overnight cultures of *E. coli* Top10 + pEGFP, *E. coli* Top10 + pEGFP-*anoG*_start1-5 and *E. coli* Top10 + pEGFP-Δ*anoG*_start1-5 were inoculated 1:100 in 10 ml 0.5xLB medium with 120 µg/ml ampicillin in duplicates. Cultures were incubated at 37°C and 150 rpm until an OD_{600} of 0.3 was reached. For one culture each, protein expression was induced by 10 mM IPTG. Incubation of induced and uninduced cultures was continued for 1 h, and then cells were pelleted. The cells were washed once with PBS, and the pellet was resuspended in 1 ml PBS. The OD_{600} was adjusted to 0.3 and 0.6. Four-times each 200 µl diluted culture were pipetted in a black microtiter plate and the fluorescence was measured (Wallac Victor^3, Perkin Elmer Life Science, excitation 485 nm, emission 535 nm, measuring time 1 s). The fluorescence of *E. coli* Top10 without plasmid was subtracted as background. The experiment was performed in triplicate. Significance of changes was calculated by Student's t-test.

2.6 Cloning of a translationally arrested *anoG* mutant

For cloning of the genomic knock-out mutant Δ*anoG* the genome editing method of Kim *et al.* [41] was adapted. The pHA_{1887} fragment and the selection cassette were amplified by PCR from the plasmid pTS2Cb. A point mutation leading to a premature stop codon was introduced into the *anoG* sequence by PCR with the oligonucleotides HA3*anoG*-115F and SM5*anoG*mut+19R (3' mutation fragment), and SM3*anoG*mut-5F and HA5*anoG*+174R (5' mutation fragment). Because the plasmid pTS2Cb-Δ*anoG* was constructed by Gibson Assembly, the four PCR fragments contain overlapping sequences. In a total reaction volume of 20 µl, 200 fmol of each PCR fragment and the NEBuilder® HiFi DNA Assembly Master Mix (NEB) were incubated at 50°C for 4 h. Two µl of the reaction were transformed into *E. coli* Top10, and plated on LB agar with 120 µg/ml ampicillin and 20 µg/ml chloramphenicol. Next, the mutation cassette was amplified by PCR using pTS2Cb-Δ*anoG* as template, and the PCR product of correct size is purified from an agarose gel (GenElute™ Gel Extraction Kit; Sigma-Aldrich).

E. coli O157:H7 Sakai [2] transformed with the plasmid pSLTS were subsequently transformed with 75 ng of the mutation cassette. After incubation for 3 h at 30°C and 150 rpm in SOC medium, cultures were plated on LB agar plates with 120 µg/ml ampicillin and 20 µg/ml chloramphenicol, and incubated at 30°C. One colony per plate was suspended in PBS. One-hundred µl of a 1:10 dilution with PBS were plated on LB agar with 100 µg/ml ampicillin and 100 ng/ml anhydrotetracycline for I-SceI induction, and incubated at 30°C over night. Several colonies were steaked on LB agar with 20 µg/ml chloramphenicol and LB agar only, and incubated at 37°C over night. Colonies, which were only able to grow on LB, were selected, and the genomic area surrounding the mutation was amplified by PCR. Additional to the premature stop codon, the restriction enzyme cut site for *Hga*I was deleted. Introduction of the point mutation was confirmed by Sanger sequencing (Eurofins) for *Hga*I digestion negative PCR products.

2.7 Competitive growth assays

Overnight cultures of EHEC Sakai wild type and EHEC Sakai Δ*anoG* were adjusted to an OD_{600} of 1.0, and then mixed in equal quantities (500 µl wild type + 500 µl mutant). Five-hundred µl of the mixture were pelleted, and the cells were snap frozen in liquid nitrogen (control t=0). Ten ml 0.5xLB medium were inoculated 1:3000 with the mixed EHEC culture. The following conditions were investigated under aerobic conditions in 0.5xLB: plain medium, at pH 5.2, at pH 8, plus 400 mM NaCl, plus 0.5 mM $CuCl_2$, plus 2 mM formic acid, or plus 2.5 mM malonic acid. Additionally, the experiment was carried out anaerobically with the same supplementations as described above using 15 ml 0.5xLB medium in tightly closed 15 ml falcon tubes. Cultures were incubated for 18 h at 37°C and 150 rpm. Next, 500 µl of culture were pelleted, 100 µl ddH_2O were added, and the sample was heated to 95°C for 10 min. The crude DNA-preparation was used as template for a PCR with the primer pair *anoG*-78F and *anoG*+124R. The PCR product was Sanger sequenced (Eurofins), and the ratio between wild type and mutant was determined by comparing peak heights. The absolute numbers were transformed into percentage values of each condition, and the values were normalized to a t=0 ratio of 50:50 wild type to mutant. The experiment was performed in biological triplicates. Significance of changes was calculated by Student’s t-test.

2.8 Complementation of EHEC Δ*anoG*

To compensate the *anoG* genomic knock-out, the intact *anoG* ORF (start codon 2, ATG) was complemented on the plasmid pBAD/*Myc-His*-C *in trans*. In addition, plasmids of truncated *anoG* sequences using the alternative start codons 4 (GTG) and 5 (CTG) were cloned. First, the sequence of *anoG* was amplified by PCR, and restriction enzyme cut sites for *Nco*I and *Hind*III were introduced. The PCR product was cloned into the plasmid pBAD/*Myc-His*-C, and the plasmid was transformed into *E. coli* O157:H7 Sakai Δ*anoG*. As a negative control, similar plasmids containing the mutated *anoG* ORF (Δ*anoG*) were cloned. Next, competitive growth experiments for the three possible start codons were performed as described above using *E. coli* O157:H7 Sakai Δ*anoG* + pBAD-*anoG*-Start2/4/5 (complementation) and *E. coli* O157:H7 Sakai Δ*anoG* + pBAD-Δ*anoG*-Start2/4/5*.* Overnight cultures were supplemented with 120 µg/ml ampicillin, adjusted to OD_{600} 1, mixed in equal ratio, and inoculated into 15 ml 0.5xLB in tightly closed 15 ml falcon tubes in duplicates. One culture each was induced with 0.002% arabinose, the other left uninduced. After incubation at 37°C and 150 rpm for 18 h, plasmids were isolated using GenElute™Plasmid Miniprep Kit (Sigma-Aldrich). With 20 ng isolated plasmids, a PCR was performed using oligonucleotides pBAD+208F and *anoG*+124R. The PCR products were Sanger sequenced (Eurofins), and the ratio of intact *anoG* over translationally arrested *anoG* was determined in percent. The experiment was performed in biological triplicates. Significant changes were calculated by Student's t-test.

2.9 Transcriptome and translatome sequencing

RNAseq and RIBOseq data [30] were investigated regarding translated ORFs in antisense to annotated genes. Briefly, the bacteria had been grown at the following growth conditions: LB medium at 37°C, harvested at $OD_{600} = 0.4$, BHI medium at 37°C, harvested at $OD_{600} = 0.1$, and BHI medium supplemented with 4% NaCl at 14°C, harvested at $OD_{600} = 0.1$. An ORF is considered translated, when it is covered with at least one read per million mapped sequenced reads normalized to 1 kbp, ≥ 50% of the ORF is covered with RIBOseq reads, and the ribosomal coverage value (RCV) is at least 0.25 in both biological replicates. Promising candidates were verified by visual inspection using the Artemis genome browser [42].

2.10 Bioinformatics methods

2.10.1 Prediction of σ^{70} promoters

The region 300 bp upstream of the TSS of *anoG* was searched for the presence of a σ^{70} promoter with the program BPROM (Softberry) [43]. The given LDF score is a measure of promoter strength, whereupon an LDF score of 0.2 indicates presence of a σ^{70} promoter with 80% accuracy and specificity.

2.10.2 Prediction of ρ-independent terminators

The regions 300 bp downstream of the stop codons of ECs2384 and *anoG* were searched for the presence and folding energy of a ρ-independent terminator with the program FindTerm (Softberry) [43].

2.10.3 Prediction of the terminator secondary structure

The RNA sequence of the ρ-independent terminator predicted with FindTerm was submitted to the Mfold web server RNA Folding Form using default parameters to determine the secondary structure [44].

2.10.4 Detection of annotated homologs

The AA sequence of the putative protein AnoG was used as a query for a blastp search against the data base refseq using default parameters [45].

2.10.5 PredictProtein

The AA sequence of AnoG was submitted to the software PredictProtein [46]. The results of PROFphd (prediction of secondary structures) [47], TMSEG (number of transmembrane helices) [48], DISULFIND (number of disulfide bonds) [49], and LocTree2 (prediction of subcellular localization) [50] were examined in further detail.

2.10.6 Phylogenetic tree construction

The novel gene *anoG* and the annotated genes ECs2384 and ECs2385 were phylostratigraphically analyzed to trace back the sequence evolution during species evolution. Tblastn (NCBI, e-value cutoff 10^{-10}, identity cutoff 50%) was used to search for homologous nucleotide sequences in all genomic sequences of the nr database independent of their annotation status. Exemplary sequences within a broad range of sequence identities were downloaded. Multiple sequence alignments were conducted using MUSCLE implemented in MEGA6 [51]. The automated alignments were manually checked and adapted, where necessary. Homologous gene pairs, in which ECs2385 was intact, but *anoG* had no tblastn hit, were individually checked by pairwise alignments of the nucleotide sequences [EMBOSS Needle, 52]. The area, in which *anoG* aligned with the (often) disintegrated *anoG* homologous sequences, was translated to the AA sequence, and aligned by multiple sequence alignment as before.

Phylogenetic trees of the strains and species examined were constructed according to Fellner *et al* [33]. Briefly, a concatenated sequence of the housekeeping genes 16S rDNA*, atpD, adk, gyrB, purA,* and *recA* was used. The sequences were aligned using ClustalW in MEGA6. Columns with gaps or ambiguities were removed, and the final dataset contains 8025 positions. The best nucleotide substitution model was searched using MEGA6. The final Maximum-Likelihood tree was calculated using Neighbor Joining, and bootstrapped 1000-times. The best nucleotide substitution model for tree construction was identified to be the General Time Reversible model (GTR), assuming that a certain fraction of sites is evolutionarily invariable (+I, 20.3394% sites). The non-uniformity of evolutionary rates among substitution sites was modeled using a discrete Gamma distribution with five rate categories (+G, parameter = 0.3102). The log likelyhood value of the final tree was -61620.9271.

3. Results

3.1 Detection of an overlapping ORF covered with RNAseq and RIBOseq reads

RNAseq and RIBOseq data sets of the enteric pathogen *Escherichia coli* O157:H7 Sakai at three different growth conditions were analyzed with regard to transcription and translation of ORFs antiparallel overlapping to annotated genes (aerobic growth). Thereby, the ORF, termed *anoG*, overlapping in reading frame -3 to the annotated gene

ECs2385 was discovered. ECs2385 encodes a conserved hypothetical protein containing a transpeptidase domain. *AnoG* and ECs2385 are covered with RNASeq and RIBOSeq reads at all growth conditions, but to different extents (Figure 1A + Table 1). *AnoG* shows highest translation in LB medium. Furthermore, the translatability in LB is extremely high indicated by an RCV of 18.18. In BHI medium at 37°C, the transcription of *anoG* is 2-fold increased, whereas the translation is 3.5-fold reduced compared to LB. Even though the translatability is reduced 6.6-fold, the RCV is still clearly above the threshold of 0.25. At combined cold and osmotic stress, translation and translatability clearly decrease (60-fold and 55-fold reduction compared to LB, respectively) (Table 1A). The annotated gene ECs2384 upstream of *anoG*, which encodes a murein lipoprotein, is very highly transcribed and translated at all conditions investigated (Figure 1A + Table 1B). A qRT-PCR analysis confirms that ECs2384 is transcribed to a much higher extend than *anoG*, the transcription of *anoG* is 303-fold lower under aerobic growth and 234-fold under anaerobic growth, respectively (Figure 1B). The overlapping annotated gene ECs2385 is moderately expressed, showing the highest transcription at BHI stress and the highest translation in LB (Table 1C).

Table 1: Transcription and translation of **A** *anoG,* **B** ECs2384, and **C** ECs2385. The RPKM values of the transcriptome and translatome data for the novel gene, the overlapping annotated gene, and the upstream annotated gene are listed. The ribosomal coverage value, a measure of the translatability, was calculated by the ratio of RPKM translatome to RPKM transcriptome. ORF coverage gives the percentage of gene sequence, which is covered by RIBOseq reads.

A *anoG*

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	50	923	18.18	87%
BHI, 37°C	95	261	2.72	90%
BHI + 4% NaCl, 14°C	46	16	0.33	58%

B ECs2384

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	39792	68550	1.83	100%
BHI, 37°C	16006	16979	1.06	100%
BHI + 4% NaCl, 14°C	10197	2892	0.32	100%

C ECs2385

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	16	21	1.42	58%
BHI, 37°C	13	8	0.73	56%
BHI + 4% NaCl, 14°C	49	7	0.15	61%

*mean values of the two biological replicates are shown.

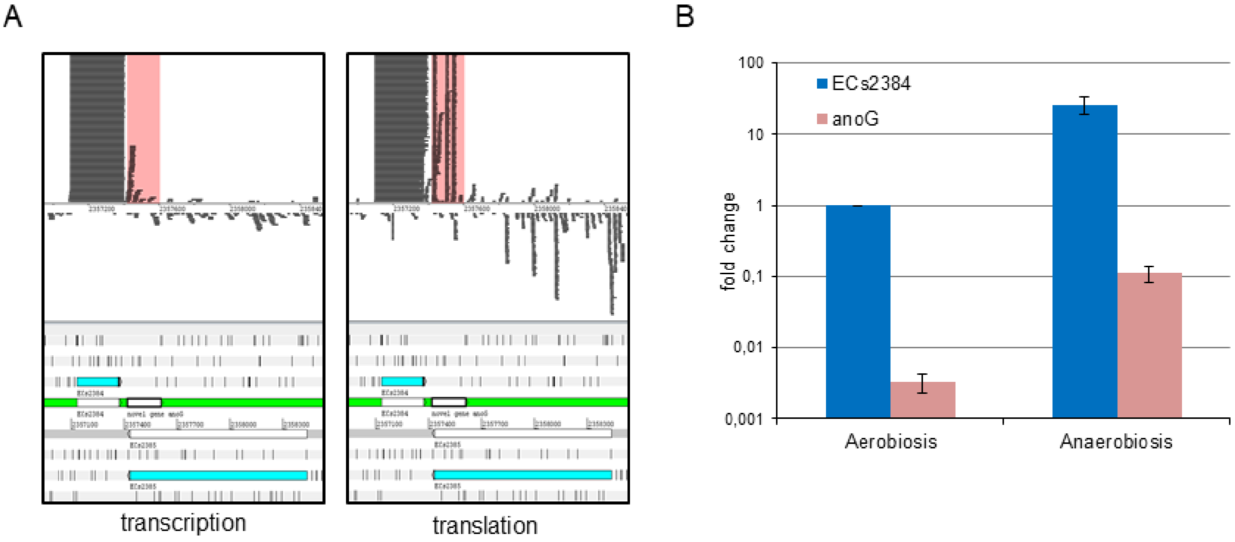


Figure 1: Transcription and translation of ECs2384, ECs2385, and *anoG.* **A** Visualization of RNAseq and RIBOseq reads of the anoG/ECs2385 region with the upstream ECs2384. Using the genome browser Artemis, strand-specifically mapping reads of the growth condition LB at 37°C are shown. Both annotated genes, ECs2385 and ECs2384, are shown in blue. The reads mapping to the novel gene *anoG* are highlighted in pink. All three depicted genes show transcription and translation signals in LB to a different extent. **B** qRT-PCR of ECs2384 and *anoG*. Transcription level of ECs2384 and *anoG* was investigated at the following conditions: 0.5×LB aerobically, OD_{600}=0.5, and 0.5×LB anaerobically, OD_{600}=0.5. The fold change compared to the transcription level of ECs2384 under aerobiosis is depicted in logarithmic scale. The experiment was performed in duplicate.

3.2 Determination of the transcriptional start site and promoter activity

Transcriptional start site (TSS) determination using 5’RACE resulted in a single signal 324 bp upstream of the proposed *anoG* start codon (Figure 2A). The TSS detected is also 38 bp upstream of the annotated gene ECs2384, and fits well to a predicted σ^{70} promoter 9 bp upstream of the start site using the software BPROM (Figure 2B). An LDF-score of 5.56 indicates high promoter strength, which was confirmed by assaying the region upstream of the TSS using an EGFP-reporter plasmid. High fluorescence intensity was measured at all investigated anaerobic growth conditions (Figure 3). Compared to the vector control, the fluorescence intensity is about 1000-fold increased. In addition, differential promoter activity between the tested stress conditions occurs: the promoter activity at pH 5 and in LB medium supplemented with 400 mM NaCl or 2.5 mM malonic acid is 2-fold increased compared to plain LB. The only condition with significantly decreased promoter activity is LB at pH 8.2. The region upstream of the TSS shows promoter activity at aerobic growth, as well (data not shown).

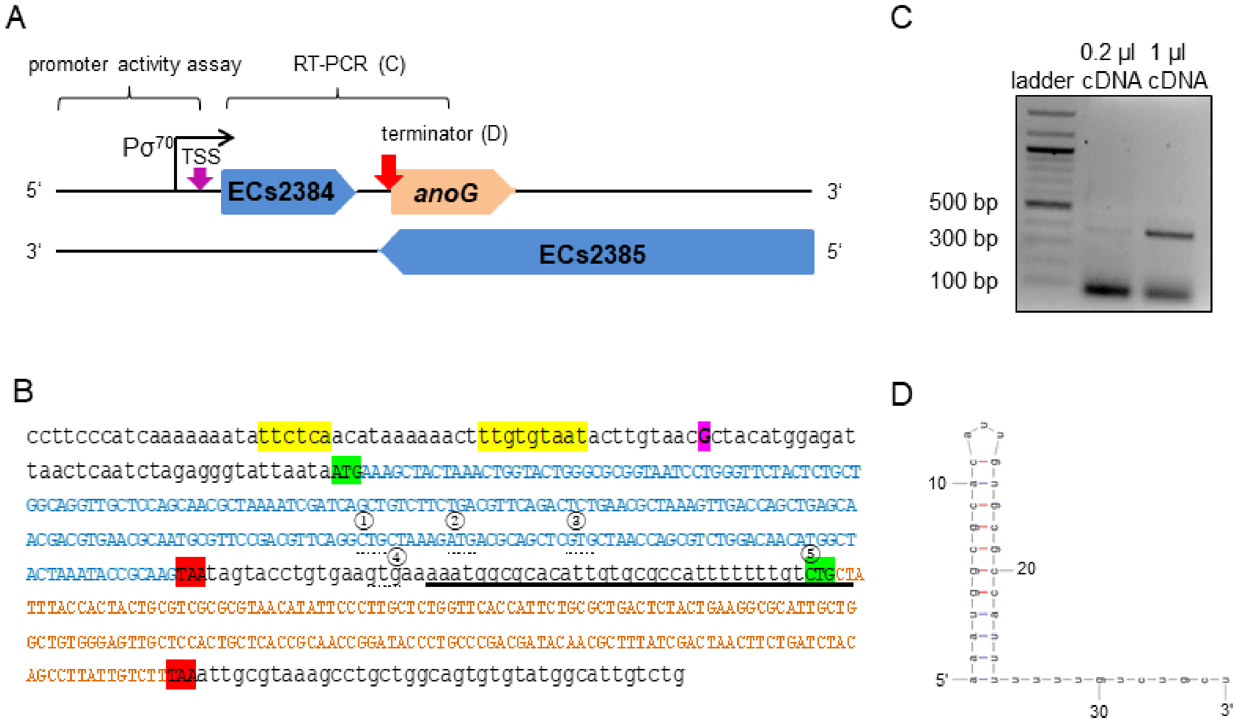

A
promoter activity assay
RT-PCR (C)
Pσ70
TSS
terminator (D)
5'
ECs2384
anoG
3'
3'
ECs2385
5'
C
0.2 µl 1 µl
ladder cDNA cDNA
500 bp
300 bp
100 bp
B
ccttcccatcaaaaaaatattctcaacataaaaaacttgtgtaatacttgtaacGctacatggagat
taactcaatctagagggtattaataATGAAAGCTACTAAACTGGTACTGGGCGCGGTAATCCTGGGTTCTACTCTGCT
GGCAGGTTGCTCCAGCAACGCTAAAATCGATCAGCTGTCTTCTGACGTTCAGACTCTGAACGCTAAAGTTGACCAGCTGAGCA
ACGACGTGAACGCAATGCGTTCCGACGTTCAGGCTGCTAAAGATGACGCAGCTCGTGCTAACCAGCGTCTGGACAACATGGCT
ACTAAATACCGCAAGTAAtagtacctgtgaagtgaaaaatggcgcacattgtgcgccatttttttgtCTGCTA
TTTACCACTACTGCGTCGCGCGTAACATATTCCCTTGCTCTGGTTCACCATTCTGCGCTGACTCTACTGAAGGCGCATTGCTG
GCTGTGGGAGTTGCTCCACTGCTCACCGCAACCGGATACCCTGCCCGACGATACAACGCTTTATCGACTAACTTCTGATCTAC
AGCCTTATTGTCTTTAAattgcgtaaagcctgctggcagtgtgtatggcattgtctg
1
2
3
4
5
D
10
20
5'
30
3'

Figure 2: Genomic organization of ECs2384 and *anoG*/ECs2385. **A** Overview of ECs2384 and *anoG*/ECs2385. The annotated genes are depicted in blue, and the novel OLG *anoG* is depicted in pink. The predicted σ^{70} promoter, the experimentally determined TSS (purple arrow), and the predicted ρ-independent terminator between ECs2384 and *anoG* (red arrow) are sketched. The DNA stretches, which

were used for the promoter activity assay and RT-PCR are indicated. **B** DNA sequence of ECs2384 and *anoG*. The sequence of the annotated gene ECs2384 is written in blue capital letters. The sequence of *anoG* is written in orange capital letters. The start codons are highlighted in green and the stop codons in red. The four alternative upstream start codons of *anoG* are marked by a dashed line and numbered consecutively. The fifth start codon we propose as the correct one (see text). The TSS detected by 5'RACE is highlighted in purple. The predicted σ^{70} promoter is colored in yellow, and the predicted ρ-independent terminator is underlined. **C** Agarose gel picture of the RT-PCR product ECs2384-*anoG*. A 100 bp DNA ladder (NEB) was used as a size standard. A primer pair was used for PCR with EHEC cDNA as template spanning the sequences of ECs2384 and *anoG*. The gel shows a product of the anticipated size of 380 bp, indicating that ECs2384 and *anoG* are transcribed as a polycistronic mRNA. **D** Secondary structure of the ρ-independent terminator predicted with Mfold.

RT-PCR using a primer pair spanning both genes, ECs2384 and *anoG*, resulted in a PCR product of 350 bp, which indicates that the two genes are co-transcribed as a bicistronic mRNA (Figure 2C). The regions downstream of the ECs2384 and *anoG* stop codons were investigated for the presence of a ρ-independent terminator using the software FindTerm. Downstream of *anoG,* no terminator is predicted. In contrast, the intergenic sequence between ECs2384 and *anoG* contains a terminator with a binding energy of -20.8 (Figure 2B and D). This terminator explains, why ECs2384 is transcribed to a much higher extend than *anoG* (Table 1), even though the two genes are organized in an operon. Most transcription events probably stop at this terminator and only a monocistronic ECs2384 mRNA is produced. However, termination is not 100% efficient and, thus, some transcription events produce a polycistronic mRNA comprising both ECs2384 and *anoG*.

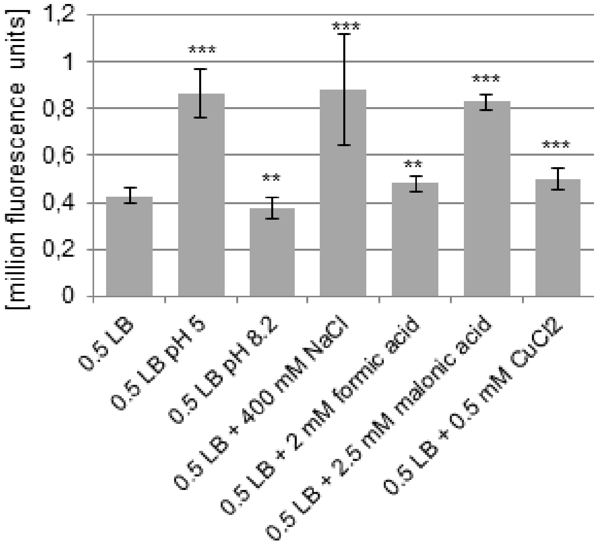


Figure 3: Promoter activity of the region upstream of the TSS at anaerobic conditions. At an OD_{600} of 0.5, the fluorescence caused by the pProbe-NT-PromoterTSS plasmid was measured. As negative control, the fluorescence of *E. coli* transformed with pProbe-NT was also determined, which was 289 ± 110 (not shown in the diagram). The fluorescence value of pProbe-NT-PromoterTSS was at all conditions significantly higher than the vector control (p<0.001). Significant changes between 0.5xLB and investigated stress conditions were calculated by Student's t-test and marked with asterisks (** p<0.01, *** p<0.001).

3.3 Properties of the hypothetical protein AnoG

Five potential start codons are present in the ORF under discussion (Figure 2B, circles 1-5): The first start codon CTG would lead to a 101 AA protein, the second start codon ATG to a 98 AA protein, the third start codon GTG to a 94 AA protein, a fourth start codon GTG to a 74 AA protein, and the fifth rare start codon CTG producing a 62 AA protein. The *anoG* sequence was cloned upstream of EGFP into the plasmid pEGFP. Plasmids for the possible start codons 1, 2, 3 and 5 were constructed and transformed into *E. coli* Top10. After induction with 10 mM IPTG, AnoG-EGFP fusion proteins were expected to be expressed, indicated by an increase in fluorescence intensity. Plasmids with the translationally arrested Δ*anoG* ORF (see below) were used as negative controls. As expected, the empty pEGFP plasmid (positive control) leads to a high increase of fluorescence intensity after induction (data not shown). The fluorescence of constructs using the putative start codons 1, 2 and 3 was zero (Table 2). Only *anoG* start codon 5 caused a 3.7-fold increase of fluorescence intensity compared to the uninduced culture and had the highest fluorescence value for all *anoG* start codons tested, indicating translation of the fusion protein. In contrast, induction of Δ*anoG* did not change the observed fluorescence intensity at all. Since ECs2385 is annotated only as a hypothetical protein, expression of an ECs2385-EGFP fusion protein was tested as well, and induction leads to a clear increase of fluorescence intensity (Table 2).

Table 2: Expression of a C-terminally AnoG-EGFP fusion protein. *E. coli* Top10 was transformed with the different pEGFP-*anoG* plasmids, and expression of the fusion protein was induced with 10 mM IPTG. OD_{600} was adjusted to 0.6 and fluorescence was measured. Fluorescence values of empty *E. coli* Top10 were subtracted (causing zero values for some readings). The experiment was performed in triplicate. The first three potential start codons had zero fluorescence, when not induced, therefore, calculation of significance was meaningless (N/A). Significant changes between induced and uninduced cultures were calculated by Student's t-test (*** p<0.001).

Sample	Fluorescence 0 mM IPTG	Fluorescence 10 mM IPTG	Significance
E. coli Top10 + pEGFP-*anoG*-Start1	0	0	N/A
E. coli Top10 + pEGFP-∆*anoG*-Start1	0	1124±152	N/A
E. coli Top10 + pEGFP-*anoG*-Start2	0	0	N/A
E. coli Top10 + pEGFP-∆*anoG*-Start2	0	0	N/A
E. coli Top10 + pEGFP-*anoG*-Start3	0	0	N/A
E. coli Top10 + pEGFP-∆*anoG*-Start3	0	291±56	NA
E. coli Top10 + pEGFP-*anoG*-Start5	883±703	3305±236	***
E. coli Top10 + pEGFP-∆*anoG*-Start5	466±312	397±169	-/-
E. coli Top10 + pEGFP-ECs2385	527±430	8662±2444	***

Therefore, the experimental data rather supports the fifth start codon, which is a rare CTG. The derived AA sequence of AnoG was analyzed using PredictProtein. The secondary structure consists mainly of loops and hydrophilic α-helices, but no membrane helices were predicted. One disulfide bond was predicted, and the protein might be secreted. A blastp search for annotated homologs in other bacteria did not obtain any hit.

3.4 Phenotype of *anoG* under anaerobic conditions

In order to search for a phenotype of *anoG,* a strand-specific translationally arrested mutant EHEC ∆*anoG* was cloned by changing a single nucleotide of the seventh *anoG* codon leading to a premature stop codon. The point mutation introduced localizes downstream of the ECs2384 stop codon leaving its sequence is unaffected. The AA sequence of the overlapping ECs2385 is not changed, because the mutation is synonymous in this frame (Figure 4A).

Competitive growth experiments were performed with equal inoculation ratios of EHEC wild type and EHEC Δ*anoG* to search for a phenotype. When the cultures were incubated aerobically, the ratio between wild type and mutant did not change significantly (Figure 4B). In contrast, anaerobic incubation resulted in a small but significant and consistent growth advantage of EHEC Δ*anoG*. The anaerobic competitive growth experiment was also performed at several stress conditions: the observed phenotype was similar to plain LB, i.e., the mutant strain showed a small growth advantage compared to the wild type (Supplementary Figure S1). Interestingly, transcription of *anoG* and also of its upstream gene ECs2384 are strongly increased at anaerobiosis compared to aerobic incubation 26-fold and 33-fold, respectively (Figure 1B).

Although the translational arrest of *anoG* leads to a weak phenotype only, a complementation *in trans* was performed, transforming EHEC Δ*anoG* with the plasmid pBAD-*myc/His*-C carrying an intact *anoG* ORF under the control of an arabinose inducible promoter. As a negative control, the same mutant was also transformed with a plasmid containing the translationally arrested Δ*anoG* ORF. Competitive growth experiments were performed as before. Furthermore, complementation plasmids for different putative *anoG* start codons (Figure 2B) were tested. Plasmids using the putative start codons 2 (ATG) and 4 (GTG) did not show significant changes of the ratio of *anoG* over Δ*anoG* (Figure 4C). In contrast, when putative start codon 5 (CTG) was used, induction resulted in a small growth disadvantage of the complemented strain after competitive growth compared to translationally arrested Δ*anoG.* However, the observed difference between wild type and mutant is larger (Figure 4B), therefore, only a partial complementation was possible.

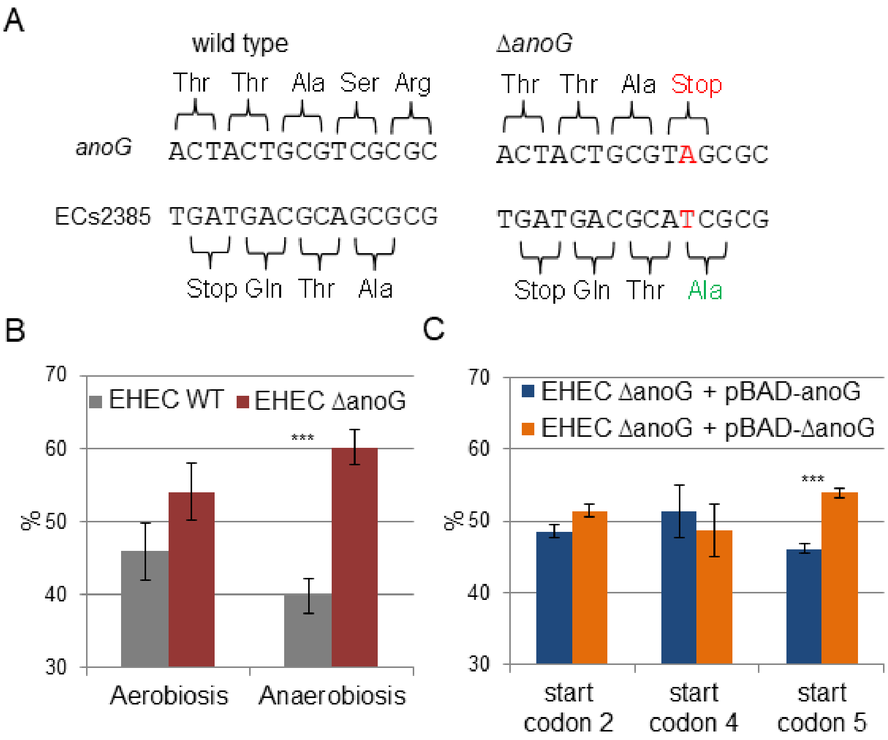


Figure 4: Creation and phenotype of EHEC Δ*anoG*. **A** Mutation strategy to obtain a translationally arrested Δ*anoG* mutant. Intro-duction of a point mutation in the DNA sequence resulted in a premature stop codon for *anoG*. The point mutation does not influence the AA sequence of the anti-parallel over-lapping gene ECs2385. In addition, a cut site for the restriction enzyme *Hga*I happened to be deleted by this mutation. **B** Phenotype of *anoG*. The ratios in percent of EHEC wild type (WT) over EHEC Δ*anoG* are shown after com-petitive growth in 0.5xLB medium aerobically and anaerobically. Wild type and mutant were mixed in equal ratio, and after 18 h incubation their ratio was determined by Sanger sequencing. The mutant has a significant growth advantage under anaerobiosis. The experiment was performed in triplicate. Significant changes were calculated by Student's t-test (*** p<0.001). **C** Complementation of EHEC Δ*anoG in trans.* EHEC Δ*anoG* was transformed with the plasmid pBAD carrying the intact *anoG* ORF, and competitive growth was performed against EHEC Δ*anoG* + pBAD-Δ*anoG*. The plasmid was induced using 0.002% arabinose. Only use of start codon 5 restores the phenotype of the wild type partly. The experiment was performed in triplicate. Significant changes were calculated by Student's t-test (*** p<0.001).

3.5 Phylogeny of ECs2384 and the OLG pair *anoG*/ECs2385

The phylogenetic distribution of the annotated genes ECs2384, ECs2385, and of the novel gene *anoG* was investigated to estimate the relative age of these genes. Homologous sequences were searched using tblastn applying an e-value cutoff of 10^{-10}. Annotated homologs of the mother gene ECs2385 are present in many bacterial species, but 99% of the hits were found in *Enterobacteriacea* (Supplementary Figure S2), with very few exceptions. The upstream gene ECs2384 is annotated in all *Enterobacteriacea* investigated, and the AA sequence is highly conserved showing only a very few AA substitutions (Supplementary Figure S3). In conclusion, ECs2384 and ECs2385 are highly conserved. In contrast, AnoG homologs are not annotated elsewhere. Non-annotated, intact conserved homologs of *anoG* are found only in *Escherichia coli* and *Shigella* strains (Figure 5). Homologs with low similarity were found in *E. fergusonii* and *E. albertii.* The sequence in both *E. albertii* strains is intact and extended at the 3' end. The sequence in *E. fergusonii* has an internal variable region containing stop codons, and is probably dysfunctional.

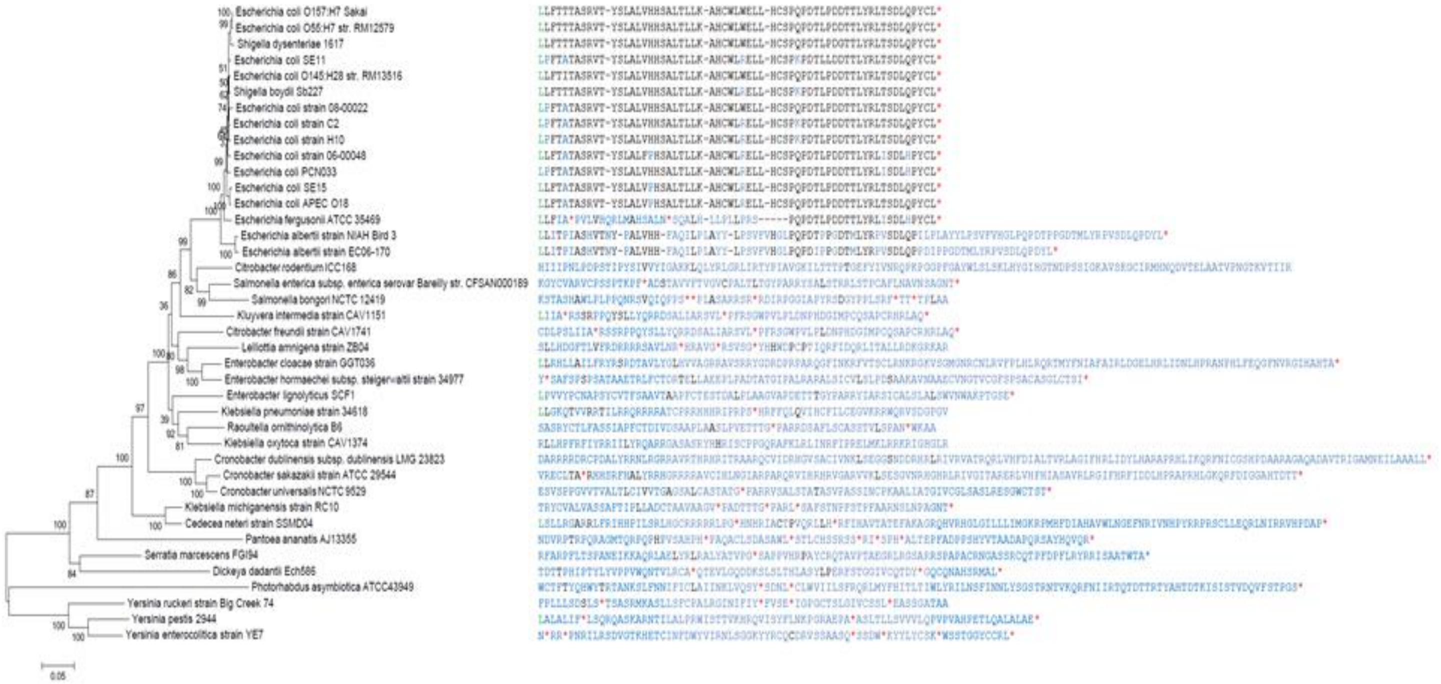

Escherichia coli O157:H7 Sakai
Escherichia coli O55:H7 str. RM12579
Shigella dysenteriae 1617
Escherichia coli SE11
Escherichia coli O145:H28 str. RM13516
Shigella boydii Sb227
Escherichia coli strain 08-00022
Escherichia coli strain C2
Escherichia coli strain H10
Escherichia coli strain 06-00048
Escherichia coli PCN033
Escherichia coli SE15
Escherichia coli APEC O18
Escherichia fergusonii ATCC 35469
Escherichia albertii strain NIAH Bird 3
Escherichia albertii strain EC06-170
Citrobacter rodentium ICC168
Salmonella enterica subsp. enterica serovar Bareilly str. CFSAN000189
Salmonella bongori NCTC 12419
Kluyvera intermedia strain CAV1151
Citrobacter freundii strain CAV1741
Lelliottia amnigena strain ZB04
Enterobacter cloacae strain GGT036
Enterobacter hormaechei subsp. steigerwaltii strain 34977
Enterobacter lignolyticus SCF1
Klebsiella pneumoniae strain 34618
Raoultella ornithinolytica B6
Klebsiella oxytoca strain CAV1374
Cronobacter dublinensis subsp. dublinensis LMG 23823
Cronobacter sakazakii strain ATCC 29544
Cronobacter universalis NCTC 9529
Klebsiella michiganensis strain RC10
Cedecea neteri strain SSMD04
Pantoea ananatis AJ13355
Serratia marcescens FGI94
Dickeya dadantii Ech586
Photorhabdus asymbiotica ATCC43949
Yersinia ruckeri strain Big Creek 74
Yersinia pestis 2944
Yersinia enterocolitica strain YE7
0.05

Figure 5: Phylogenetic tree of *anoG.* The phylogenetic tree on the left was constructed from representative species possessing homologs of *anoG* or ECs2385, antiparallel overlapping to each other. The tree is based on a concatemer of 16S RNA, *atpD*, *adk*, *gyrB*, *purA*, and *recA.* On the right, the different amino acid sequences of AnoG (if present) are aligned. Possible start codons are colored in green and stop codons in red (*). Variable regions with no detectable amino acid homology to AnoG are colored in blue.

4. Discussion

4.1 Is *anoG* a protein-coding gene?

In bacteria, regulatory RNAs are frequently encoded antisense to annotated protein-coding genes [16; 53], whereas only a few examples of non-trivial protein-coding OLGs are known [32; 33; 54]. Thus, instead of a novel protein-coding gene, *anoG* might encode a novel ncRNA. Coverage of the ORF with RNAseq reads (Table 1A + Figure 1) and detectable promoter activity (Figure 3) would support both, a ncRNA and a protein encoding gene. However, several observations contradict the hypothesis that *anoG* is a solely ncRNA. First, the ORF is clearly covered by RIBOseq reads indicating active translation. RIBOseq has been used successfully in the past to detect translation of non-annotated genes in eukaryotes [23; 55; 56] and prokaryotes [28; 29]. It is highly unlikely that such a high RIBOseq signal is caused by contaminating RNA binding proteins (catching the RNA and causing the carry-over), or *anoG* sequence randomly bound to ribosomes [57]. Furthermore, start and stop of the RIBOseq signal fit very well to the *anoG* ORF (Figure 1A). Second, the translatability of *anoG* in LB medium is exceptionally high (Table 1A). Short annotated EHEC genes only have a mean RCV of 1.55 at this condition [30]. Neuhaus *et al.* [58] report that the mean RCV of tRNAs, which are also not translated such as ncRNAs, is 0.06 and that an ORF can be considered a protein-coding gene with an RCV of at least 0.3. Third, it is uncommon that ncRNAs are transcribed as a polycistronic RNA together with a protein-coding gene (Figure 2C). Fourth, expression of an AnoG-EGFP fusion protein was possible (Table 2). Fifth, it is not expected that the change of a single nucleotide in the sequence of a ncRNA will lead to a phenotype, because ncRNAs regulate the expression of target mRNAs by base pairing over a stretch of several nucleotides, and a single base substitution will hardly abolish pairing. We consider this combined evidence to make it very likely that *anoG* encodes AnoG as a protein-coding gene.

4.2 *AnoG* uses the rare start codon CTG

According to the genetic code table 11 (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi), ATG is the most frequent start codon in bacteria. In addition, GTG and TTG can be used as start codons as well. Only rare cases of CTG and ATT start codons have been reported. Five putative in-frame start codons are present (CTG, ATG, GTG, GTG, and CTG, Figure 2B) leading to potential *anoG* variants, which are all translationally arrested by introduction of a stop codon (Figure 4A). Induction of expression of an AnoG-EGFP fusion protein was only possible when using the fifth CTG start codon (Table 2). Furthermore, partial complementation of the Δ*anoG* phenotype was achieved using the downstream CTG start codon, which was not possible using any of the other start codons (Figure 4C). These results indicate that the second CTG start codon (overall fifth) and not the canonical ATG (overall second) probably is the correct start codon of *anoG*. In *E. coli,* only one other gene, the plasmid borne *repA,* is confirmed to start with CTG [59]. RIBOseq data of *E. coli* using the antibiotic tetracycline to stall translation at the translational start site (i.e., the start codon) indicates three additional genes with CTG start codons as alternative start sites [60]. Moreover, the S12 ribosomal protein *rpsL* of *Deinococcus desertii* starts with CTG [61]. New methods, like sequencing of translation initiation (QTIseq) [62] and N-terminal proteomics [56; 63], may additionally confirm initiation by rare start codons. However, only one point mutation is required to change the rare CTG start codon of *anoG* into an optimal ATG codon.

4.3 Evolution of the novel OLG pair *anoG*/ECs2385

Phylostratigraphic analyses indicate that the upstream gene ECs2384 as well as the opposite strand gene ECs2385 are conserved proteins, which probably originated before the *Enterobacteriaceae* diversified (Supplementary Figures 2 and 3). Very few homologs of the mother gene ECs2385 are found in distantly related organisms, and may indicate horizontal gene transfer. Those do not harbor an *anoG* homolog. *AnoG* in contrast, is a strongly taxonomically restricted gene, and probably evolved after the separation of the *Escherichia/Shigella* clade from other *Enterobacteriaceae* (Figure 5). Evolutionary young genes may be volatile, and can get lost, when they do not encode a protein with a beneficial function under an environment experienced currently by a bacterium [64]. *AnoG* overlaps antiparallel to ECs2385, certainly causing some constraints in the evolution of both genes [65; 66]. However, *anoG* is encoded in frame -3 relative to ECs2385, and this combination provides the highest freedom for variation in two evolutionary coupled overlapping genes [67]. The annotated gene ECs2384 is transcribed to a high extent (Table 1B), and transcription is usually terminated at a downstream ρ-independent terminator (Figure 2BD). Nevertheless, the RNA polymerase may occasionally read through a ρ-independent terminator, and a longer mRNA will be produced (Figure 2C). This extended mRNA now contains the small ORF *anoG,* which may have originated by several point mutations. However, it is unknown which features of an overlapping open reading frame are required to enable the appearance of a functional protein by overprinting in the first place. Finally, the ORF must be translated into a protein. In case the novel protein will lead to some fitness advantage at a certain condition, the novel ORF will eventually become fixed, and will evolve further under positive selection [68]. In *E. coli* O157:H7 Sakai, *anoG* exhibits a small, but detectable and significant phenotype at anaerobiosis (Figure 4B), at least demonstrating a cellular impact of AnoG under certain conditions. However, it is impossible to infer potential functions of this protein based on the data available, and a functional characterization of AnoG certainly requires additional studies.

Acknowledgments

We thank Svenja Simon (University of Konstanz, Chair for Data Analysis and Visualization) to determine positions in the *anoG*/ECs2385 sequence, where mutation(s) introduce a premature stop codon into the sequence of *anoG*, are synonymous regarding the AA sequence of ECs2385, and additionally create or destroy a restriction enzyme cut site.

References

[1] S.M. Sadiq, T.H. Hazen, D.A. Rasko, and M. Eppinger, EHEC Genomics: Past, Present, and Future. Microbiology spectrum 2 (2014) EHEC-0020-2013.

[2] T. Hayashi, K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C.G. Han, E. Ohtsubo, K. Nakayama, T. Murata, M. Tanaka, T. Tobe, T. Iida, H. Takami, T. Honda, C. Sasakawa, N. Ogasawara, T. Yasunaga, S. Kuhara, T. Shiba, M. Hattori, and H. Shinagawa, Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA research: an international journal for rapid publication of reports on genes and genomes 8 (2001) 11-22.

[3] J.Y. Lim, J. Yoon, and C.J. Hovde, A brief overview of *Escherichia coli* O157:H7 and its plasmid O157. J Microbiol Biotechnol 20 (2010) 5-14.

[4] C.S. Wong, S. Jelacic, R.L. Habeeb, S.L. Watkins, and P.I. Tarr, The risk of the hemolytic-uremic syndrome after antibiotic treatment of *Escherichia coli* O157:H7 infections. N Engl J Med 342 (2000) 1930-6.

[5] M. Eppinger, and T.A. Cebula, Future perspectives, applications and challenges of genomic epidemiology studies for food-borne pathogens: A case study of Enterohemorrhagic *Escherichia coli* (EHEC) of the O157:H7 serotype. Gut microbes 6 (2015) 194-201.

[6] A.K. Persad, and J.T. LeJeune, Animal Reservoirs of Shiga Toxin-Producing *Escherichia coli*. Microbiology spectrum 2 (2014) EHEC-0027-2014.

[7] J. Barker, T.J. Humphrey, and M.W. Brown, Survival of *Escherichia coli* O157 in a soil protozoan: implications for disease. FEMS Microbiol Lett 173 (1999) 291-5.

[8] Z. Saldana, E. Sanchez, J. Xicohtencatl-Cortes, J.L. Puente, and J.A. Giron, Surface structures involved in plant stomata and leaf colonization by shiga-toxigenic *Escherichia coli* O157:H7. Front Microbiol 2 (2011) 119.

[9] Z. Hou, R.C. Fink, M. Sugawara, F. Diez-Gonzalez, and M.J. Sadowsky, Transcriptional and functional responses of *Escherichia coli* O157:H7 growing in the lettuce rhizoplane. Food Microbiol 35 (2013) 136-42.

[10] D. Jayaraman, O. Valdes-Lopez, C.W. Kaspar, and J.M. Ane, Response of Medicago truncatula seedlings to colonization by *Salmonella enterica* and *Escherichia coli* O157:H7. PloS one 9 (2014) e87970.

[11] A.D. Duffitt, R.T. Reber, A. Whipple, and C. Chauret, Gene expression during survival of *Escherichia coli* O157:H7 in soil and water. Int J Microbiol 2011 (2011) 340506.

[12] A.M. Semenov, A.A. Kuprianov, and A.H. van Bruggen, Transfer of enteric pathogens to successive habitats as part of microbial cycles. Microb Ecol 60 (2010) 239-49.

[13] L. Wasala, J.L. Talley, U. Desilva, J. Fletcher, and A. Wayadande, Transfer of *Escherichia coli* O157:H7 to spinach by house flies, *Musca domestica* (Diptera: Muscidae). Phytopathology 103 (2013) 373-80.

[14] B.L. Flaherty, F. Van Nieuwerburgh, S.R. Head, and J.W. Golden, Directional RNA deep sequencing sheds new light on the transcriptional response of *Anabaena* sp. strain PCC 7120 to combined-nitrogen deprivation. BMC Genomics 12 (2011) 332.

[15] R. Landstorfer, S. Simon, S. Schober, D. Keim, S. Scherer, and K. Neuhaus, 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 (2014) 353.

[16] J.E. Dornenburg, A.M. Devita, M.J. Palumbo, and J.T. Wade, Widespread Antisense Transcription in *Escherichia coli*. mBio 1 (2010).

[17] R. Raghavan, E.A. Groisman, and H. Ochman, Genome-wide detection of novel regulatory RNAs in *E. coli*. Genome Res 21 (2011) 1487-97.

[18] I. Lasa, A. Toledo-Arana, A. Dobin, M. Villanueva, I.R. de los Mozos, M. Vergara-Irigaray, V. Segura, D. Fagegaltier, J.R. Penades, J. Valle, C. Solano, and T.R. Gingeras, Genome-wide antisense transcription drives mRNA processing in bacteria. Proceedings of the National Academy of Sciences of the United States of America 108 (2011) 20172-7.

[19] J.T. Wade, and D.C. Grainger, Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol 12 (2014) 647-53.

[20] N.T. Ingolia, S. Ghaemmaghami, J.R. Newman, and J.S. Weissman, Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. Science 324 (2009) 218-23.

[21] K. Neuhaus, R. Landstorfer, S. Simon, S. Schober, P.R. Wright, C. Smith, R. Backofen, R. Wecko, D.A. Keim, and S. Scherer, 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 (2017).

[22] J.L. Aspden, Y.C. Eyre-Walker, R.J. Phillips, U. Amin, M.A. Mumtaz, M. Brocard, and J.P. Couso, Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. eLife 3 (2014) e03528.

[23] A.A. Bazzini, T.G. Johnstone, R. Christiano, S.D. Mackowiak, B. Obermayer, E.S. Fleming, C.E. Vejnar, M.T. Lee, N. Rajewsky, T.C. Walther, and A.J. Giraldez, Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J 33 (2014) 981-93.

[24] N.T. Ingolia, G.A. Brar, N. Stern-Ginossar, M.S. Harris, G.J. Talhouarne, S.E. Jackson, M.R. Wills, and J.S. Weissman, Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep 8 (2014) 1365-79.

[25] Z. Ji, R. Song, A. Regev, and K. Struhl, Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. eLife 4 (2015) e08890.

[26] J. Ruiz-Orera, X. Messeguer, J.A. Subirana, and M.M. Alba, Long non-coding RNAs as a source of new peptides. eLife 3 (2014) e03523.

[27] L. Calviello, N. Mukherjee, E. Wyler, H. Zauber, A. Hirsekorn, M. Selbach, M. Landthaler, B. Obermayer, and U. Ohler, Detecting actively translated open reading frames in ribosome profiling data. Nat Methods 13 (2016) 165-70.

[28] J. Baek, J. Lee, K. Yoon, and H. Lee, Identification of Unannotated Small Genes in *Salmonella*. G3 7 (2017) 983-989.

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

[30] S.M. Hücker, T. Goldberg, A. Schafferhans, M. Bernhofer, G. Vestergaard, C.W. Nelson, Z. Ardern, B. Rost, S. Scherer, and K. Neuhaus, Discovery of numerous novel small genes in the intergenic regions of the *Escherichia coli* O157:H7 Sakai genome. PloS one (2017).

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

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

[33] L. Fellner, S. Simon, C. Scherling, M. Witting, S. Schober, C. Polte, P. Schmitt-Kopplin, D.A. Keim, S. Scherer, and K. Neuhaus, Evidence for the recent origin of a bacterial protein-coding, overlapping orphan gene by evolutionary overprinting. BMC Evol. Biol. 15 (2015) 1.

[34] T. Kurata, A. Katayama, M. Hiramatsu, Y. Kiguchi, M. Takeuchi, T. Watanabe, H. Ogasawara, A. Ishihama, and K. Yamamoto, Identification of the set of genes, including nonannotated *morA*, under the direct control of ModE in *Escherichia coli*. J Bacteriol 195 (2013) 4496-505.

[35] M. Behrens, J. Sheikh, and J.P. Nataro, Regulation of the overlapping *pic*/*set* locus in *Shigella flexneri* and enteroaggregative *Escherichia coli*. Infect. Immun. 70 (2002) 2915-2925.

[36] V.P. Balabanov, V.Y. Kotova, G.Y. Kholodii, S.Z. Mindlin, and G.B. Zavilgelsky, 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 (2012) 55-60.

[37] J.R. Haycocks, and D.C. Grainger, Unusually Situated Binding Sites for Bacterial Transcription Factors Can Have Hidden Functionality. PloS one 11 (2016) e0157016.

[38] A. McVeigh, A. Fasano, D.A. Scott, S. Jelacic, S.L. Moseley, D.C. Robertson, and S.J. Savarino, IS1414, an *Escherichia coli* insertion sequence with a heat-stable enterotoxin gene embedded in a transposase-like gene. Infect Immun 68 (2000) 5710-5715.

[39] M.W. Pfaffl, A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29 (2001) e45.

[40] W.G. Miller, J.H. Leveau, and S.E. Lindow, Improved *gfp* and *inaZ* broad-host-range promoter-probe vectors. Mol Plant Microbe Interact. 13 (2000) 1243-50.

[41] J. Kim, A.M. Webb, J.P. Kershner, S. Blaskowski, and S.D. Copley, A versatile and highly efficient method for scarless genome editing in *Escherichia coli* and *Salmonella enterica*. BMC Biotechnol 14 (2014) 84.

[42] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.A. Rajandream, and B. Barrell, Artemis: sequence visualization and annotation. Bioinformatics 16 (2000) 944-5.

[43] V.V. Solovyev, and T.V. Tatarinova, Towards the integration of genomics, epidemiological and clinical data. Genome medicine 3 (2011) 48.

[44] M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31 (2003) 3406-15.

[45] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, Basic local alignment search tool. J. Mol. Biol. 215 (1990) 403-410.

[46] G. Yachdav, E. Kloppmann, L. Kajan, M. Hecht, T. Goldberg, T. Hamp, P. Honigschmid, A. Schafferhans, M. Roos, M. Bernhofer, L. Richter, H. Ashkenazy, M. Punta, A. Schlessinger, Y. Bromberg, R. Schneider, G. Vriend, C. Sander, N. Ben-Tal, and B. Rost, PredictProtein--an open resource for online prediction of protein structural and functional features. Nucleic Acids Res 42 (2014) W337-43.

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

[48] M. Bernhofer, E. Kloppmann, J. Reeb, and B. Rost, TMSEG: Novel prediction of transmembrane helices. Proteins 84 (2016) 1706-1716.

[49] A. Ceroni, A. Passerini, A. Vullo, and P. Frasconi, DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. Nucleic Acids Res. 34 (2006) W177-W181.

[50] T. Goldberg, M. Hecht, T. Hamp, T. Karl, G. Yachdav, N. Ahmed, U. Altermann, P. Angerer, S. Ansorge, K. Balasz, M. Bernhofer, A. Betz, L. Cizmadija, K.T. Do, J. Gerke, R. Greil, V. Joerdens, M. Hastreiter, K. Hembach, M. Herzog, M. Kalemanov, M. Kluge, A. Meier, H. Nasir, U. Neumaier, V. Prade, J. Reeb, A. Sorokoumov, I. Troshani, S. Vorberg, S. Waldraff, J. Zierer, H. Nielsen, and B. Rost, LocTree3 prediction of localization. Nucleic Acids Res 42 (2014) W350-5.

[51] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 30 (2013) 2725-9.

[52] W. Li, A. Cowley, M. Uludag, T. Gur, H. McWilliam, S. Squizzato, Y.M. Park, N. Buso, and R. Lopez, The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res. 43 (2015) W580-W584.

[53] J. Georg, and W.R. Hess, cis-antisense RNA, another level of gene regulation in bacteria. Microbiology and molecular biology reviews : MMBR 75 (2011) 286-300.

[54] S. Tunca, C. Barreiro, J.J. Coque, and J.F. Martin, Two overlapping antiparallel genes encoding the iron regulator DmdR1 and the Adm proteins control siderophore and antibiotic biosynthesis in *Streptomyces coelicolor* A3(2). FEBS J 276 (2009) 4814-27.

[55] C. Fritsch, A. Herrmann, M. Nothnagel, K. Szafranski, K. Huse, F. Schumann, S. Schreiber, M. Platzer, M. Krawczak, J. Hampe, and M. Brosch, Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res 22 (2012) 2208-18.

[56] P. Van Damme, D. Gawron, W. Van Criekinge, and G. Menschaert, N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. Mol Cell Proteomics 13 (2014) 1245-61.

[57] B. Liu, and S.B. Qian, Characterizing inactive ribosomes in translational profiling. Translation 4 (2016) e1138018.

[58] K. Neuhaus, R. Landstorfer, S. Simon, S. Schober, P.R. Wright, C. Smith, R. Backofen, R. Wecko, D.A. Keim, and S. Scherer, 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 18 (2017) 216.

[59] A.J. Spiers, and P.L. Bergquist, Expression and regulation of the RepA protein of the RepFIB replicon from plasmid P307. J Bacteriol 174 (1992) 7533-41.

[60] K. Nakahigashi, Y. Takai, M. Kimura, N. Abe, T. Nakayashiki, Y. Shiwa, H. Yoshikawa, B.L. Wanner, Y. Ishihama, and H. Mori, Comprehensive identification of translation start sites by tetracycline-inhibited ribosome profiling. DNA Res 23 (2016) 193-201.

[61] M. Baudet, P. Ortet, J.C. Gaillard, B. Fernandez, P. Guerin, C. Enjalbal, G. Subra, A. de Groot, M. Barakat, A. Dedieu, and J. Armengaud, Proteomics-based refinement of *Deinococcus deserti* genome annotation reveals an unwonted use of non-canonical translation initiation codons. Mol Cell Proteomics 9 (2010) 415-26.

[62] X. Gao, J. Wan, B. Liu, M. Ma, B. Shen, and S.B. Qian, Quantitative profiling of initiating ribosomes *in vivo*. Nat Methods 12 (2015) 147-53.

[63] P. Willems, E. Ndah, V. Jonckheere, S. Stael, A. Sticker, L. Martens, F. Van Breusegem, K. Gevaert, and P. Van Damme, N-terminal Proteomics Assisted Profiling of the Unexplored Translation Initiation Landscape in *Arabidopsis thaliana*. Mol Cell Proteomics 16 (2017) 1064-1080.

[64] M. Huvet, and M.P. Stumpf, Overlapping genes: a window on gene evolvability. BMC Genomics 15 (2014) 721.

[65] F. Lillo, and D.C. Krakauer, A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes. Biol Direct 2 (2007) 22.

[66] Z.I. Johnson, and S.W. Chisholm, Properties of overlapping genes are conserved across microbial genomes. Genome Res 14 (2004) 2268-72.

[67] K. Mir, and S. Schober, Selection pressure in alternative reading frames. PLoS One 9 (2014) e108768.

[68] A.R. Carvunis, T. Rolland, I. Wapinski, M.A. Calderwood, M.A. Yildirim, N. Simonis, B. Charloteaux, C.A. Hidalgo, J. Barbette, B. Santhanam, G.A. Brar, J.S. Weissman, A. Regev, N. Thierry-Mieg, M.E. Cusick, and M. Vidal, Proto-genes and *de novo* gene birth. Nature 487 (2012) 370–374.

2.10 <u>Publication 5</u>: Discovery of the novel gene *slyC* antiparallel overlapping the transcriptional regulator *slyA* in *Escherichia coli* O157:H7 Sakai, and characterization of the influence of L-arginine on its gene expression

Discovery of the novel gene *slyC* antiparallel overlapping the transcriptional regulator *slyA* in *Escherichia coli* O157:H7 Sakai, and characterization of the influence of L-arginine on its gene expression

Sarah M. Hücker[1], Sonja Vanderhaeghen[1], Lena Dübbel[1], Romy Wecko[1], Siegfried Scherer[1,2] and Klaus Neuhaus[3*]

[1]Chair for Microbial Ecology, Technische Universität München, Weihenstephaner Berg 3, 85354 Freising, Germany; [2]ZIEL – Institute for Food & Health, Technische Universität München, Weihenstephaner Berg 3, 85354 Freising, Germany; [3]Core Facility Microbiome/NGS, ZIEL – Institute for Food & Health, Technische Universität München, Weihenstephaner Berg 3, 85354 Freising, Germany.

*Correspondence: neuhaus@tum.de

ABSTRACT

Transcription and translation of an open reading frame, named *slyC*, which overlaps antiparallel in reading frame -2 to the transcriptional regulator *slyA,* was detected in RNAseq and RIBOseq data of the enteric pathogen *Escherichia coli* O157:H7 Sakai. S*lyC* is annotated as an outer membrane lipoprotein in other *E. coli* strains, and the open reading frame is present in many *Enterobacteriales*, where it also overlaps to *slyA.* The transcriptional start site is located upstream of the annotated gene *slyB,* and RT-PCR confirmed polycistronic transcription of the operon *slyBC*. The sequence upstream of the transcriptional start contains the predicted consensus motif of two ARG boxes overlapping with the promoter, probably binding to the L-arginine dependent transcription factor ArgR. Promoter activity was decreased after L-arginine supplementation, and the strand-specific translationally arrested mutant Δ*slyC* shows a growth disadvantage in LB medium containing L-arginine compared to the wild type in competitive growth experiments. A SlyC-EGFP fusion protein could be expressed. Therefore, *slyC* represents an arginine regulated, novel antiparallel overlapping gene.

Keywords: overlapping gene, EHEC, arginine, ArgR, SlyA

INTRODUCTION

Since an amino acid (AA) is encoded by three nucleotides, the DNA double strand contains six possible reading frames. Therefore, it is feasible that the sequences of two or more protein-coding genes overlap. Overlapping genes (OLGs) are a common feature of viral genomes, because of limited space for a genome inside the capsid (RANCUREL *et al.* 2009; CHIRICO *et al.* 2010). In contrast, in bacteria, only trivial overlaps of a few nucleotides are a well-known phenomenon, especially for genes encoded in an operon facilitating translational coupling (LILLO AND KRAKAUER 2007). However, 5.3% of all annotated *E. coli* K12 genes contain a completely embedded (but non-annotated) overlapping open reading frame (ORF) larger than 300 bp (MERINO *et al.* 1994). Interestingly, alternative reading frames in relation to the annotated mother gene show significantly less stop codons than expected statistically (MIR *et al.* 2012). In the past, the existence of non-trivial OLGs was neglected in bacteria due to increased evolutionary constraints on both sequences (KRAKAUER 2000; LÈBRE AND GASCUEL 2017), or the long ORFs antisense to annotated genes are supposed to be present only because of the codon bias of the mother gene, and should not have biological meaning (VELOSO *et al.* 2005). Algorithms used for genome annotation still reject OLGs, and only the ORF with the better score becomes annotated (DELCHER *et al.* 2007).

Therefore, only a handful of longer prokaryotic OLGs have been described in literature, e.g., (WANG *et al.* 1999; MCVEIGH *et al.* 2000; BEHRENS *et al.* 2002; SILBY AND LEVY 2008). In some cases, the OLG was predicted bioinformatically (JENSEN *et al.* 2006), or peptides not matching annotated genes were detected by mass spectrometry (KIM *et al.* 2009; ZHAO *et al.* 2011). Phenotypic characterization was performed for even less OLGs. TUNCA *et al.* (2009) described the overlapping gene pair *dmdR1/adm* in *Streptomyces coelicolor*: Both genes encode regulators and strand-specific knock-out mutants showed a phenotype.

In this study, the enteric pathogen *Escherichia coli* O157:H7 Sakai is used as a model organism (HAYASHI *et al.* 2001). EHEC colonizes many habitats (i.e., the intestine of mammals, insects, plants, foodstuff, and soil) (LIM *et al.* 2010; SALDANA *et al.* 2011), and all these environments represent different challenges regarding nutrient availability, host defense mechanisms, and competition with other bacteria. Consequently, expression of distinct sets of genes is required (LANDSTORFER *et al.* 2014; HÜCKER *et al.* 2017b). The few known OLGs in EHEC are evolutionary young (KRAKAUER 2000; FELLNER *et al.* 2014; FELLNER *et al.* 2015), and might contribute to the species specific adaption for colonization of a new niche (HUVET AND STUMPF 2014). EHEC strain EDL933 (LATIF *et al.* 2014) is closely related to strain Sakai used here, and in EDL933 already two OLGs are discovered and characterized (FELLNER *et al.* 2014; FELLNER *et al.* 2015): the ORF *htgA* is embedded in antisense into the sequence of the mother gene *yaaW,* and *htgA* might have originated *de novo* by overprinting (DELAYE *et al.* 2008; FELLNER *et al.* 2014). Overprinting indicates that a novel gene was created step-wise from previously non-coding sequence (KEESE AND GIBBS 1992). Strand-specific knock-out mutants of *htgA* and *yaaW* showed altered biofilm formation and changes in the metabolome (FELLNER *et al.* 2014). Also, for the second OLG pair *nog1/citC* it is likely that the novel gene originated by overprinting antisense to the citrate lyase ligase *citC*. The following observations support the protein-coding character of *nog1*: the ORF is weakly transcribed in cow dung (LANDSTORFER *et al.* 2014), the region upstream the transcriptional start site (TSS) shows promoter activity, the Nog1 protein can be expressed, and the translationally arrested mutant has a growth disadvantage in competitive growth experiments compared to the wild type (FELLNER *et al.* 2015).

Next generation sequencing is a powerful method for the discovery of transcription and translation signals antisense to annotated genes. Transcription of an ORF can be detected by strand-specific RNAseq (FLAHERTY *et al.* 2011), and ribosomal footprinting (INGOLIA *et al.* 2009) demonstrates mRNA translation. The combination of both methods allows the discrimination between ncRNAs and novel protein-coding genes by calculating the translatability (NEUHAUS *et al.* 2017). In this study, the novel OLG pair *slyC/slyA* in EHEC was discovered: the ORF *slyC*, antiparallel overlapping to the transcriptional regulator *slyA,* shows evidence of transcription and translation, furthermore, phylostratigraphic analysis, promoter activity, and a phenotype confirm its protein-coding character. The expression of the novel OLG *slyC* seems to be regulated by intracellular L-arginine concentration.

MATERIAL AND METHODS

Bacterial strains and plasmids used in this study are listed in Supplementary Table S1. Oligonucleotides are listed in Supplementary Table S2.

Determination of transcriptional start site by 5' RACE

The total RNA of an overnight culture in LB medium of *Escherichia coli* O157:H7 Sakai (Genbank accession number NC_002695) (Hayashi *et al.* 2001) was isolated with Trizol. The 5'RACE System for Rapid Amplification of cDNA Ends, Version 2.0 (Invitrogen) was used according to the manual. After the second PCR, the dominant product(s) were excised from the agarose gel, and purified with the GenElute^{TM} Gel Extraction Kit (Sigma-Aldrich). The PCR product(s) were Sanger sequenced (Eurofins) using oligonucleotide *slyC*+152R.

Determination of transcriptional termination site by 3'RACE

Total RNA of 500 µl EHEC Sakai overnight culture in LB medium was isolated using Trizol, and the remaining DNA was digested using 2 U TURBO^{TM} DNase (Thermo Fisher Scientific). The 5'/3' RACE Kit, 2nd Generation (Roche Applied Science) was applied according to the manual, but instead of an oligo dT primer for cDNA synthesis the gene specific primer *slyC*-3F was used. A nested PCR was performed for product amplification. The dominant product was excised from the agarose gel, purified with the GenElute^{TM} Gel Extraction Kit (Sigma-Aldrich), and Sanger sequenced (Eurofins) with oligonucleotide *slyC*+49F.

RT-PCR

EHEC total RNA of 500 µl overnight culture in LB medium was isolated using Trizol. Remaining DNA was digested using 2 U TURBO^{TM} DNase (Thermo Fisher Scientific) for 1 h at 37°C. Reverse transcription with 500 ng RNA as template was performed with

200 U SuperScript^{TM} III Reverse Transcriptase (Thermo Fisher Scientific) according to the manual. The obtained cDNA was used as template for a PCR with a primer pair spanning *slyB* and *slyC*.

Cloning of pProbe-NT plasmids and determination of promoter activity

The genomic regions 300 bp upstream of the determined transcriptional start sites were amplified by PCR, and restriction enzyme cut sites for *Sal*I and *EcoR*I were introduced. The PCR products were cloned into the plasmid pProbe-NT (MILLER *et al.* 2000), and transformed into *Escherichia coli* Top10. The plasmid sequence was verified by Sanger sequencing (Eurofins). Overnight cultures of *E. coli* Top10 + pProbe-NT and *E. coli* Top10 + pProbe-NT-Promoter_*slyC,* or pProbe-NT-PromoterTSS were used for 1:100 inoculation of 10 ml 0.5 LB medium with 30 µg/ml kanamycin. The following conditions were investigated for promoter activity in 0.5xLB each: plain LB, at pH 5, at pH 8.2, plus 400 mM NaCl, plus 0.5 mM $CuCl_2$, plus 2 mM formic acid, plus 2.5 mM malonic acid, or plus 10 mM L-arginine. Cultures were incubated at 37°C and 150 rpm until an OD_{600} of 0.5 was reached. Then the cells were pelleted, washed once with PBS, and resuspended in 1 ml PBS. The OD_{600} was adjusted to 0.3 and 0.6. Four times 200 µl were pipetted in a black microtiter plate, and the fluorescence was measured (Wallac Victor^3, Perkin Elmer Life Science, excitation 485 nm, emission 535 nm, measuring time 1 s). The fluorescence of *E. coli* Top10 without plasmid was subtracted as background. For *E. coli* Top10 + pProbe-NT-PromoterTSS the experiment was repeated in modified MOD medium (ROSENFELD *et al.* 2005) without the AAs L-arginine, L-aspartic acid, and L-glutamic acid. The conditions MOD and MOD plus 10 mM L-arginine were investigated as described above. The experiment was performed in triplicate. Significant changes were calculated by the Student's t-test.

Cloning of a C-terminal SlyC-EGFP fusion protein and overexpression of SlyC fusion protein

The *slyC* sequence without the stop codon was amplified by PCR, and restriction enzyme cut sides for *Pst*I and *Nco*I were introduced. The PCR product was cloned into the plasmid pEGFP, and transformed into *Escherichia coli* Top10. The plasmid sequence was verified by Sanger sequencing (Eurofins). For fusion protein overexpression, overnight cultures of *E. coli* Top10 + pEGFP and *E. coli* Top10 + pEGFP-*slyC* were inoculated 1:100 in 10 ml 0.5 LB medium with 120 µg/ml ampicillin in duplicates. Cultures were incubated at 37°C and 150 rpm until an OD_{600} of 0.3 was reached. For one culture each, protein expression was induced by 10 mM IPTG. Incubation of induced and uninduced cultures was continued for 1 h, and then cells were pelleted. The cells were washed once with PBS, and the pellet was resuspended in 1 ml PBS. The OD_{600} was adjusted to 0.3 and 0.6. Four times 200 µl were pipetted in a black microtiter plate, and the fluorescence was measured (Wallac Victor^3, Perkin Elmer Life Science, excitation 485 nm, emission 535 nm, measuring time 1 s). The fluorescence of *E. coli* Top10 without plasmid was subtracted as background. The experiment was performed in triplicate. Significant differences were calculated by the Student's t-test.

Cloning of a translationally arrested *slyC* mutant

For cloning of the genomic knock-out mutant Δ*slyC,* the genome editing method described by KIM *et al.* (2014) was adapted. The pHA_{1887} fragment and the selection cassette were amplified by PCR from the plasmid pTS2Cb. Two point mutations leading to a premature stop codon at the *slyC* sequence, but do not change the AA sequence of *slyA*, were introduced by PCR with the oligonucleotides HA3*slyC*-145F and SM5*slyC*mut+36R (3' mutation fragment), and SM3*slyC*mut-22F and HA5*slyC*+156R (5' mutation fragment). Because the four PCR fragments contain overlapping sequences, the plasmid pTS2Cb-Δ*slyC* can be obtained by Gibson Assembly. In a total reaction volume of 20 µl, 200 fmol of each PCR fragment and 10 µl of the NEBuilder® HiFi DNA Assembly Master Mix (NEB) were incubated at 50°C for 4 h. Two µl of the reaction were transformed into *E. coli* Top10, and plated on LB agar with 120 µg/ml ampicillin and 20 µg/ml chloramphenicol. Next, the mutation cassette was be amplified by PCR using pTS2Cb-Δ*slyC* as template, and the PCR product of correct size was purified from an agarose gel (GenElute^{TM} Gel Extraction Kit; Sigma-Aldrich). *E. coli* O157:H7 Sakai was transformed with the plasmid pSLTS. These EHEC were transformed with 75 ng mutation cassette, and incubated 3 h at 30°C and 150 rpm in SOC medium. Then, the cells were plated on LB agar plates with 120 µg/ml ampicillin and 20 µg/ml chloramphenicol, and incubated at 30°C. One colony per plate was suspended in PBS. One-hundred µl of a 1:10 PBS-dilution were plated on LB agar with 100 µg/ml ampicillin and 100 ng/ml anhydrotetracycline for I-SceI induction, and incubated at 30°C over night. Several colonies were plated and incubated at 37°C over night on LB agar with 20 µg/ml chloramphenicol and plain LB agar. Colonies, which were able to grow only on LB, were selected, and the genomic area surrounding the mutation was amplified by PCR. Additional to the premature stop codon, a restriction enzyme cut site for *Alu*I was created. Introduction of the two point mutations was confirmed by Sanger sequencing (Eurofins) for *Alu*I digestion positive PCR products.

Competitive growth assays

Overnight cultures of EHEC Sakai wild type and EHEC Sakai Δ*slyC* were adjusted to an OD_{600} of 1.0, and then mixed in equal quantities (500 µl wild type + 500 µl mutant). Five-hundred µl of the mixture were pelleted, and the cells were snap frozen in liquid nitrogen (control t=0). Ten ml 0.5 LB medium were inoculated 1:3000 with the mixed EHEC culture. The following conditions were investigated in 0.5xLB each: plain LB, at pH 5, at pH 8.2, plus 400 mM NaCl, plus 0.5 mM $CuCl_2$, plus 2 mM formic acid, plus 2.5 mM malonic acid, plus 4 mM sodium-orthovanadate, plus 4 mM acidic acid, plus 4 mM malic acid, plus 250 mM NaH_2PO_4, plus 5 mM LiOH, plus 160 µM formaldehyde, or plus 20 mM L-arginine. Cultures were incubated for 18 h at 37°C and 150 rpm. Then, 500 µl of culture were pelleted, 100 µl water were added to the pellet, and the sample was heated to 95°C for 10 min. Using this crude DNA extraction, a PCR was performed with the primer pair *slyC*-95F and *slyC*+109R. The PCR product was Sanger sequenced (Eurofins), and the ratio between wild type and mutant was determined by comparing peak heights. The absolute numbers were transformed into percentage values of each condition, and the values were calculated for a t=0 ratio of 1:1 wild type to mutant ratio, respectively. Thus, the competitive index was calculated using the following formula:

$$CI = \frac{mutant_{end}[\%] \times wild\ type_{start}[\%]}{mutant_{start}[\%] \times wild\ type_{end}[\%]}$$

The experiment was performed in biological triplicates. Significance was calculated by the Student’s t-test.

Complementation

To compensate the *slyC* genomic knock-out mutation, the intact *slyC* ORF was supplemented *in trans* on a plasmid. First, the sequence of *slyC* was amplified by PCR, and restriction enzyme cut sites for *Nco*I and *Hind*III were introduced. The PCR product was cloned into the plasmid pBAD/*Myc-His*-C, and the plasmid was transformed into *E. coli* O157:H7 Sakai Δ*slyC*. As a negative control, the plasmid containing the mutated *slyC* gene (Δ*slyC*) was cloned. Next, competitive growth experiments were performed as described above using *E. coli* O157:H7 Sakai Δ*slyC* + pBAD-*slyC* (complementation) and *E. coli* O157:H7 Sakai Δ*slyC* + pBAD-Δ*slyC* (control). Both overnight cultures were supplemented with 120 µg/ml ampicillin, and the cultures were mixed in equal ratio. Ten ml of either 0.5 LB or 0.5 LB + 20 mM L-arginine were inoculated 1:3000 in quadruplicates. Induction of the *slyC* frame (either present as wild type or as Δ*slyC*) was performed with 0.002% arabinose. After incubation at 37°C and 150 rpm for 18 h, plasmids were isolated using the GenElute™Plasmid Miniprep Kit (Sigma-Aldrich). Using 20 ng isolated plasmid, PCR was performed with the oligonucleotides pBAD+208F and pBAD+502R. The PCR products were Sanger sequenced (Eurofins), and the ratio of intact *slyC* over translationally arrested Δ*slyC* was determined in percent. The experiment was performed in biological triplicates.

Transcriptome and translatome sequencing

RNAseq and RIBOseq data sets of HÜCKER *et al.* (2017a) were investigated with respect to translated ORFs located in antisense to annotated genes. Briefly, the bacteria had been grown under the following growth conditions: LB medium at 37°C, harvested at OD_{600} 0.4, BHI medium at 37°C, harvested at OD_{600} 0.1, and BHI medium supplemented with 4% NaCl at 14°C, harvested at OD_{600} 0.1. An ORF was considered translated, when (i) it was covered with at least one read per million mapped sequenced reads normalized to 1 kbp, (ii) ≥ 50% of the ORF is covered with RIBOseq reads, and (iii) the ribosomal coverage value (RCV) is at least 0.25 in both biological replicates. Promising candidates were verified by visual inspection using the Artemis genome browser (RUTHERFORD *et al.* 2000).

Bioinformatics methods

Prediction of σ^{70} promoters

The region 300 bp upstream of the TSS was searched for the presence of a σ^{70} promoter with the program BPROM (Softberry) (SOLOVYEV AND TATARINOVA 2011). The LDF score is a measure of promoter strength, whereupon an LDF score of 0.2 indicates presence of a σ^{70} promoter with 80% accuracy and specificity. Furthermore, the program predicts transcription factor consensus motifs.

Prediction of ρ-independent terminators

The region 300 bp downstream of the stop codon of *slyC* was searched for the presence and folding energy of a ρ-independent terminator with the program FindTerm (Softberry) (SOLOVYEV AND TATARINOVA 2011). The secondary structure of the detected terminator was determined using mfold (ZUKER 2003).

Prediction of Shine-Dalgarno sequence

Presence of a Shine-Dalgarno sequence in the region 30 bp upstream of the *slyB* and the *slyC* start codons was investigated. The free energy $\Delta G°$ was calculated according to MA *et al.* (2002). The Shine-Dalgarno consensus motif taAGGAGGt has a $\Delta G°$ of -9.6, and a minimum $\Delta G°$ of -2.9 is required for the presence of a Shine-Dalgarno sequence.

Detection of annotated homologs

SlyC was translated into the corresponding protein sequence, which was used to query the data base GeneBank with blastp for annotated homologous proteins using default parameters (ALTSCHUL *et al.* 1990).

PredictProtein

The AA sequence of *slyC* was submitted to the software PredictProtein (YACHDAV *et al.* 2014). The methods PROFphd (secondary structure) (ROST AND SANDER 1994), TMSEG (transmembrane helices) (BERNHOFER *et al.* 2016), DISULFIND (disulfide bonds) (CERONI *et al.* 2006), and LocTree2 (subcellular location) (GOLDBERG *et al.* 2014) were used.

Prediction of signal peptides

The AA sequence of *slyC* was submitted to the signal peptide prediction programs SignalP 4.1 (PETERSEN *et al.* 2011) and Phobius (KALL *et al.* 2004). Default parameters were used.

Phylogenetic tree construction

For phylostratigraphic analysis of *slyC* and *slyA,* tblastn was used with an e-value cutoff of 0.001 and at least 50% identity, which allows the search of homologous nucleotide sequences to an AA query in all genomic sequences of the database independent from its annotation status. Sequences within a broad range of sequence identities were downloaded from the database and used for phylogenetic analysis. Multiple sequence alignments were conducted with MUSCLE, MEGA6 and manually adapted (TAMURA *et al.* 2013). Those homologous gene pairs in which *slyA* was intact, but *slyC* had no tblastn hit were individually checked by pairwise alignments of the nucleotide sequences (EMBOSS Needle, LI *et al.* 2015). The sequence, in which *slyC* aligned with the disintergrated *slyC* homolog, was translated to the AA sequence and aligned as multiple sequence alignment.

All phylogenetic trees were constructed after (FELLNER *et al.* 2015) from a concatenated sequence of the housekeeping genes 16S rDNA*, atpD, adk, gyrB, purA,* and *recA*. The sequences were aligned with ClustalW with default parameters in MEGA6. Columns with gaps or ambiguities were manually removed. The best nucleotide substitution model was computed in MEGA6, and the model with the lowest Bayesian Information Criterion (BIC) and the lowest log likelihood was used. The final Maximum Likelihood tree was calculated with Neighbor Joining and bootstrapped 1000 times.

The final dataset contains 7,240 positions. The best nucleotide substitution model for phylogenetic tree construction was identified to be the General Time Reversible model (GTR) by assuming that a certain fraction of sites are evolutionarily invariable (+I, 34.950% sites). The non-uniformity of evolutionary rates among substitution sites was modeled using a discrete Gamma distribution with five rate categories (+G, parameter = 0.2379). In this case, the BIC was with a value of 107061.259 the lowest, hence, the best one. The log likelihood value of the final tree was -52951.6792.

RESULTS

Discovery of a transcribed and translated non-annotated ORF

The sequence of a 192 bp ORF, named *slyC,* is completely embedded in antisense into the sequence of the annotated transcriptional regulator *slyA* (ECs2351) (Figure 1A). Transcription and translation of *slyC* were discovered by analyzing RNAseq and RIBOseq data of *Escherichia coli* O157:H7 Sakai grown at three different conditions (HÜCKER *et al.* 2017a). Figure 2 visualizes the RNAseq and RIBOseq reads in BHI medium at 37°C mapped to the annotated genes *slyB* (ECs2350), *slyA* and to the putative novel gene *slyC*. All genes clearly show transcription and translation signals. Further transcription and translation signals occur downstream of *slyC* indicating putative additional novel antiparallel overlapping genes. However, the ORFs downstream of *slyC* were not further characterized. *SlyA, slyB,* and *slyC* are also expressed to different extents at the other two conditions investigated (Table 1). *SlyC* shows highest transcription in BHI at 37°C and highest translation in LB medium at 37°C (Table 1A). Accordingly, the translatability, expressed by the ribosomal coverage value, is highest in LB. The annotated *slyA* shows the highest transcription in BHI + 4% NaCl at 14°C and the highest translation in LB, as well (Table 1B). In EHEC, the gene upstream of the expressed ORF *slyC* is annotated as the outer membrane protein *slyB*. This gene shows the highest absolute expression compared to *slyA* and *slyC,* and, again, highest translatability in LB (Table 1C).

Table 1: Transcription and translation of *slyA, slyB,* and *slyC*. The RPKM values for the novel gene, the overlapping annotated gene, and the upstream annotated gene are listed. The ribosomal coverage value, a measure of the translatability, was calculated by the ratio of RPKM translatome over RPKM transcriptome. ORF coverage gives the percentage of gene sequence, which is covered by RIBOseq reads.

A *slyC*

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	113.5	250.7	2.28	0.84
BHI, 37°C	444.5	148.6	0.34	0.82
BHI + 4% NaCl, 14°C	181.9	26	0.15	0.62

B ECs2351 (*slyA*)

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	291.5	219.9	0.76	0.87
BHI, 37°C	329.1	72.9	0.26	0.93
BHI + 4% NaCl, 14°C	1327.4	21.8	0.03	0.9

C ECs2350 (*slyB*)

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	4544.6	4786.8	1.06	1
BHI, 37°C	2408.5	1895	0.76	1
BHI + 4% NaCl, 14°C	1250.6	123.7	0.1	0.98

*mean values of the two biological replicates are shown.

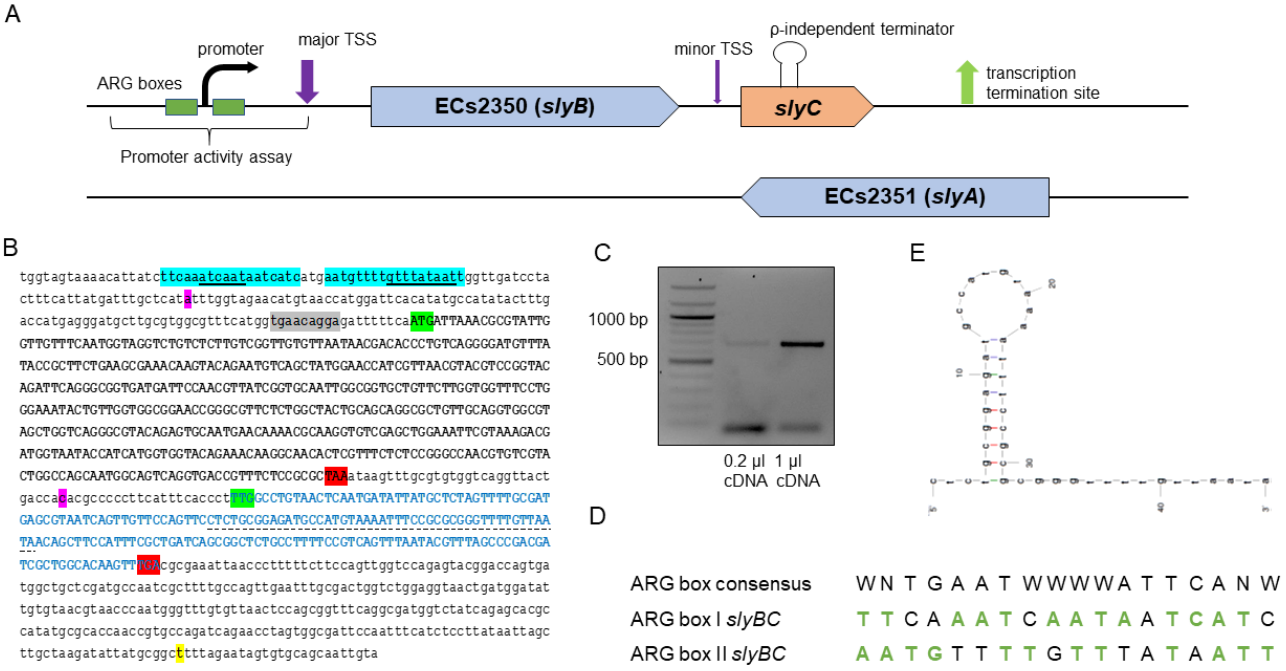

A
ARG boxes
promoter
major TSS
ECs2350 (slyB)
minor TSS
ρ-independent terminator
slyC
transcription
termination site
Promoter activity assay
ECs2351 (slyA)
B
tggtagtaaaacattatcttcaaatcaataatcatcatgaatgttttgtttataattggttgatccta
ctttcattatgatttgctcatatttggtagaacatgtaaccatggattcacatatgccatatactttg
accatgagggatgcttgcgtggcgtttcatggtgaacaggagatttttcaATGATTAAACGCGTATTG
GTTGTTTCAATGGTAGGTCTGTCTCTTGTCGGTTGTGTTAATAACGACACCCTGTCAGGGGATGTTTA
TACCGCTTCTGAAGCGAAACAAGTACAGAATGTCAGCTATGGAACCATCGTTAACGTACGTCCGGTAC
AGATTCAGGGCGGTGATGATTCCAACGTTATCGGTGCAATTGGCGGTGCTGTTCTTGGTGGTTTCCTG
GGAAATACTGTTGGTGGCGGAACCGGGCGTTCTCTGGCTACTGCAGCAGGCGCTGTTGCAGGTGGCGT
AGCTGGTCAGGGCGTACAGAGTGCAATGAACAAAACGCAAGGTGTCGAGCTGGAAATTCGTAAAGACG
ATGGTAATACCATCATGGTGGTACAGAAACAAGGCAACACTCGTTTCTCTCCGGGCCAACGTGTCGTA
CTGGCCAGCAATGGCAGTCAGGTGACCGTTTCTCCGCGCTAAataagtttgcgtgtggtcaggttact
gaccacacgccccttcatttcaccctTTGGCCTGTAACTCAATGATATTATGCTCTAGTTTTGCGAT
GAGCGTAATCAGTTGTTCCAGTTCCTCTGCGGAGATGCCATGTAAAATTTCCGCGCGGTTTTGTTAA
TAACAGCTTCCATTTCGCTGATCAGCGGCTCTGCCTTTTCCGTCAGTTTAATACGTTTAGCCCGACGA
TCGCTGGCACAAGTTTGAcgcgaaattaacccttttttcttccagttggtccagagtacggaccagtga
tggctgctcgatgccaatcgcttttgccagttgaatttgcgactggtctggaggtaactgatggatat
tgtgtaacgtaacccaatgggtttgtgttaactccagcggtttcaggcgatggtctatcagagcacgc
catatgcgcaccaaccgtgccagatcagaacctagtggcgattccaatttcatctccttataattagc
ttgctaagatattatgcggctttagaatagtgtgcagcaattgta
C
1000 bp
500 bp
0.2 µl cDNA
1 µl cDNA
E
D
ARG box consensus   W N T G A A T W W W W A T T C A N W
ARG box I slyBC   T T C A A A T C A A T A A T C A T C
ARG box II slyBC   A A T G T T T T G T T T A T A A T T

Figure 1: Structure of the genomic region in EHEC containing the annotated genes *slyA* and *slyB* and the novel gene *slyC*. **A** Schematic view of the annotated genes *slyA, slyB,* and the novel gene *slyC*. The determined transcriptional start sites and the transcription termination site are delineated. Also, the predicted σ^{70} promoter overlapping with the sequences of the two ARG boxes and the ρ-independent terminator are shown. **B** DNA sequence of the novel gene *slyC* and its upstream and downstream sequence. The sequence of *slyC* is colored in blue and written in capital letters. The sequence of *slyB* is also written in capital letters. Start codons are highlighted in green, and stop codons in red. The two transcriptional start sites detected by 5'RACE are highlighted in pink, and the transcriptional termination site detected by 3'RACE is highlighted in yellow. The predicted σ^{70} promoter is underlined, and the two overlapping ARG boxes are highlighted in blue. The Shine-Dalgarno sequence is highlighted in light gray. The predicted ρ-independent terminator is indicated by a dashed line. **C** Agarose gel picture of the RT-PCR product of *slyBC* cDNA. The 100 bp DNA ladder (NEB) was used as a size standard. A primer pair was used spanning the sequence of *slyB* and *slyC*. The gel shows one product of 670 bp indicating that

slyB and *slyC* are transcribed as a polycistronic mRNA. **D** Comparison of the detected ARG boxes of *slyBC* to the consensus motif. The ARG box consensus motif published by CHARLIER *et al.* (1992) was applied. The first ARG box matches in 12 positions the consensus motif, and the second ARG box in nine positions. **E** Secondary structure of the predicted ρ-independent terminator using mfold.

A blastp search using the AA sequence SlyC, detected three annotated homologs in other *E. coli* strains with e-values between 5×10^{-27} and 2×10^{-32}. In these *E. coli* strains, this ORF is annotated as an outer membrane lipoprotein, and the AA sequences show between 87-100% query coverage and 98-100% identity to SlyC of EHEC Sakai. Outer membrane proteins require a signal peptide, which is recognized by the tat secretion system. Indeed, the programs Phobius and SignalP 4.1 both predict a signal peptide and a cleavage site after AA 24, which would lead to a 40 AA mature SlyC protein. In contrast, the subcellular location prediction indicates that *slyC* is secreted. In agreement with an outer membrane localization of SlyC, the software PredictProtein predicts two transmembrane helices. Furthermore, *slyC* secondary structure should consist to 50% of loops, 40% of α-helices, 10% of β-sheets, and the protein should contain two disulfide bonds.

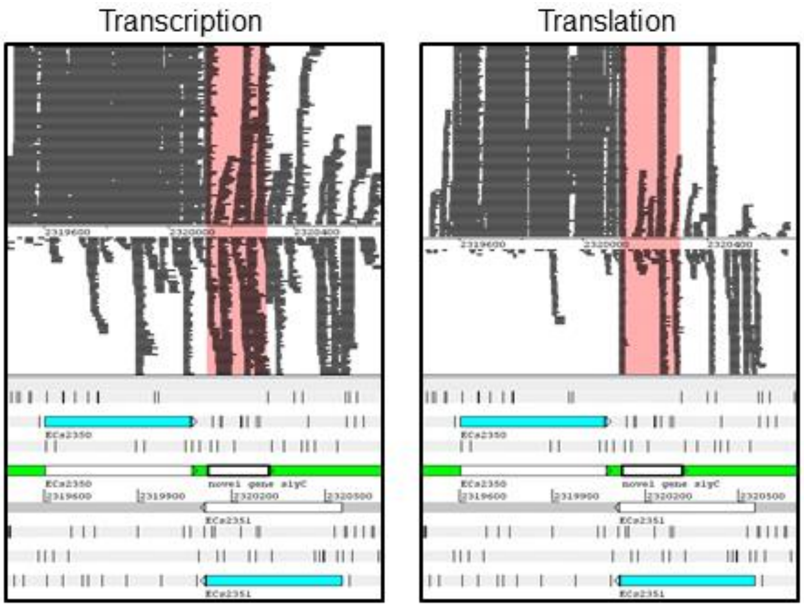


Figure 2: Visualization of RNAseq and RIBOseq reads mapped strand-specifically to the annotated genes ECs2350 (*slyB*), ECs2351 (*slyA*), and to the novel OLG *slyC* in Artemis. The annotated genes are highlighted in blue. The novel gene *slyC* is highlighted in pink.

Phylogeny of *slyC*/*slyA*

A tblastn search allows not only the detection of annotated homologs but also identifies non-annotated homologous nucleotide sequences in other bacteria. To determine the putative evolutionary age of *slyA* and *slyC,* a tblastn search with an e-value threshold of

0.001 was performed. The search identified homologous nucleotide sequences of *slyC* in 1,032 bacterial genomes and homologs of *slyA* in 2,774 organisms. Additionally, a tblastn search with the AA sequence of *slyB* resulted in 1,542 organisms with homologs. While all organisms, in which *slyC* was found, belong to the order *Enterobacteriales*, *slyA* has homologous sequences throughout all bacterial species, even in *Mycobacteria*, where it is also annotated as a transcriptional regulator**.** For the evolutionary tree, 40 homologous sequences were chosen. In all of them a *slyA* homolog is present and annotated as “transcriptional regulator SlyA”. The *slyC* sequences are always embedded in antisense into the sequence of *slyA*. Figure 3 shows that the *slyC* sequence is intact in all species, except in *Salmonella*, where the sequence has an internal stop codon in the middle of the sequence, which is followed by the start codon AA methionine. The *slyC* sequence is highly conserved in *Escherichia/Shigella.* In distantly related species, several AA substitutions have occurred, and the sequences are elongated. Enlargement occurs more frequently at the 5’ end of *slyC* than at the 3’ end. The sequences of *slyA* and *slyB* are highly conserved, only some distantly related *slyA* homologs contain variable regions (Supplementary Figures S1 and S2).

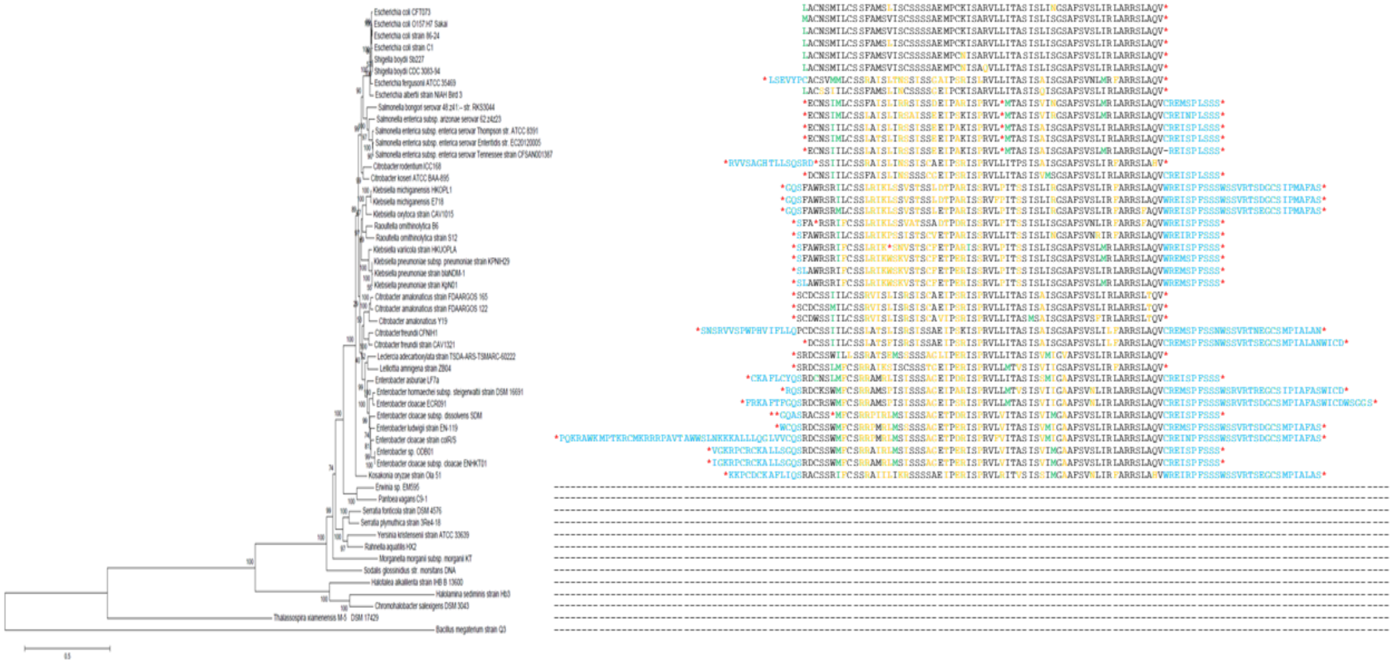

Escherichia coli CFT073
Escherichia coli O157:H7 Sakai
Escherichia coli strain 86-24
Escherichia coli strain C1
Shigella boydii Sb227
Shigella boydii CDC 3083-94
Escherichia fergusonii ATCC 35469
Escherichia albertii strain NIAH Bird 3
Salmonella bongori serovar 48:z41:-- str. RKS3044
Salmonella enterica subsp. arizonae serovar 62:z4z23
Salmonella enterica subsp. enterica serovar Thompson str. ATCC 8391
Salmonella enterica subsp. enterica serovar Enteritidis str. EC20120005
Salmonella enterica subsp. enterica serovar Tennessee strain CFSAN001387
Citrobacter rodentium ICC168
Citrobacter koseri ATCC BAA-895
Klebsiella michiganensis HKOPL1
Klebsiella michiganensis E718
Klebsiella oxytoca strain CAV1015
Raoultella ornithinolytica B6
Raoultella ornithinolytica strain S12
Klebsiella variicola strain HKUOPLA
Klebsiella pneumoniae subsp. pneumoniae strain KPNIH29
Klebsiella pneumoniae strain blaNDM-1
Klebsiella pneumoniae strain KpN01
Citrobacter amalonaticus strain FDAARGOS 165
Citrobacter amalonaticus strain FDAARGOS 122
Citrobacter amalonaticus Y19
Citrobacter freundi CFNIH1
Citrobacter freundi strain CAV1321
Leclercia adecarboxylata strain TSDA-ARS-TSMARC-60222
Lelliottia amnigena strain Z804
Enterobacter asburiae LF7a
Enterobacter hormaechei subsp. steigerwaltii strain DSM 16691
Enterobacter cloacae ECR091
Enterobacter cloacae subsp. dissolvens SDM
Enterobacter ludwigii strain EN-119
Enterobacter cloacae strain colR/S
Enterobacter sp. ODB01
Enterobacter cloacae subsp. cloacae ENHKU01
Kosakonia oryzae strain Ola 51
Erwinia sp. EM595
Pantoea vagans C9-1
Serratia fonticola strain DSM 4576
Serratia plymuthica strain 3Re4-18
Yersinia kristensenii strain ATCC 33639
Rahnella aquatilis HX2
Morganella morganii subsp. morganii KT
Sodalis glossinidius str. morsitans DNA
Halotalea alkalilenta strain IHB B 13600
Halomonas sediminis strain Hb3
Chromohalobacter salexigens DSM 3043
Thalassospira xiamenensis M-5 DSM 17429
Bacillus megaterium strain Q3
0.5

Figure 3: Phylostratigraphic distribution of *slyC.* The phylogenetic tree on the left was constructed from representative species possessing the *slyC* sequence embedded into the sequence of *slyA* based on a

concatemer of 16S RNA, *atpD*, *adk*, *gyrB*, *purA*, and *recA.* On the right, the homologous AA sequences of SlyC are aligned. Start codons are colored in green, stop codons (*) in red, AA changes in orange, and enlargement is highlighted in blue.

Characterization of the promoter region of *slyBC*

The transcriptional start site(s) of *slyC* were determined by 5'RACE. This resulted in a minor TSS 22 bp upstream of the TTG start codon of *slyC* and a major TSS 618 bp upstream of the start codon (Figure 1AB). The major TSS is also 97 bp upstream of the start codon of the annotated gene *slyB*. This indicates that *slyB* and *slyC* can be transcribed as a polycistronic mRNA, and build the operon *slyBC*. Polycistronic transcription was confirmed by RT-PCR (Figure 1C). However, transcription of *slyB* is 7- to 40-fold higher than transcription of *slyC* (Table 1) indicating that *slyB* can be transcribed alone as a single mRNA, as well. This is confirmed by prediction of a ρ-independent terminator using FindTerm, located 119 bp downstream of the *slyB* stop codon, which overlaps with the *slyC* sequence (Figure 1ABE). Inefficient termination will than lead to the transcription of the polycistronic *slyBC* mRNA. No further ρ-independent terminator was predicted downstream of the *slyC* stop codon. Anyway, the transcription termination site could be determined by 3'RACE 275 bp downstream of the *slyC* stop codon. This unusually long 3'UTR together with the downstream transcription and translation signals (Figure 2) support the hypothesis of additional translated ORFs downstream of *slyC.* A Shine-Dalgarno sequence with ∆G° of -4.6 is present 9 bp upstream of the start codon of *slyB*, but upstream of *slyC* no Shine-Dalgarno sequence is detectable (Figure 1B).

The software BPROM predicts a strong σ^{70} promoter 35 bp upstream of the major TSS with an LDF score of 6.66 (Figure 1B). Interestingly, two ARG boxes are predicted, which overlap with the predicted promoter. ARG boxes are binding sites for the transcription factor ArgR, which regulates L-arginine biosynthesis and uptake genes (Cunin *et al.* 1986). At high intracellular L-arginine concentrations, an ArgR hexamer binds to two ARG boxes, and downregulates transcription of its target genes by blocking promoter recognition of the RNA polymerase (Caldara *et al.* 2006). Indeed, the predicted ARG boxes upstream of *slyB* show high similarity to the consensus motif (Figure 1D), and the distance of three nucleotides between the boxes corresponds to the optimal value (Charlier *et al.* 1992; Tian *et al.* 1992).

The sequence upstream of *slyBC* including the potential promoter and the ARG boxes was cloned into the plasmid pProbe-NT. An active promoter leads to EGFP expression, which can be measured by an increase in fluorescence intensity. Indeed, the sequence 300 bp upstream of the major TSS shows significant promoter activity (285- to 365-fold increased fluorescence compared to the empty plasmid) with different fluorescence intensities dependent on growth condition (Figure 4A). In LB supplemented with 400 mM NaCl, the highest promoter activity is detected leading to a 2.9-fold increase of fluorescence intensity compared to plain LB. Additionally, the conditions LB at pH 5 and supplementation with 2.5 mM malonic acid lead to significantly increased promoter activity (2.5-fold and 1.4-fold, respectively). As expected, based on the presence of the ARG boxes, supplementation with 10 mM L-arginine leads to a 1.2-fold reduced promoter activity. However, LB medium already contains L-arginine, therefore, the experiment was repeated in modified MOD medium lacking L-arginine. Interestingly, the promoter activity in MOD medium is 2.3-fold higher than in LB medium (Figure 4B). Supplementation with 10 mM L-arginine causes a 1.9-fold reduction of fluorescence intensity.

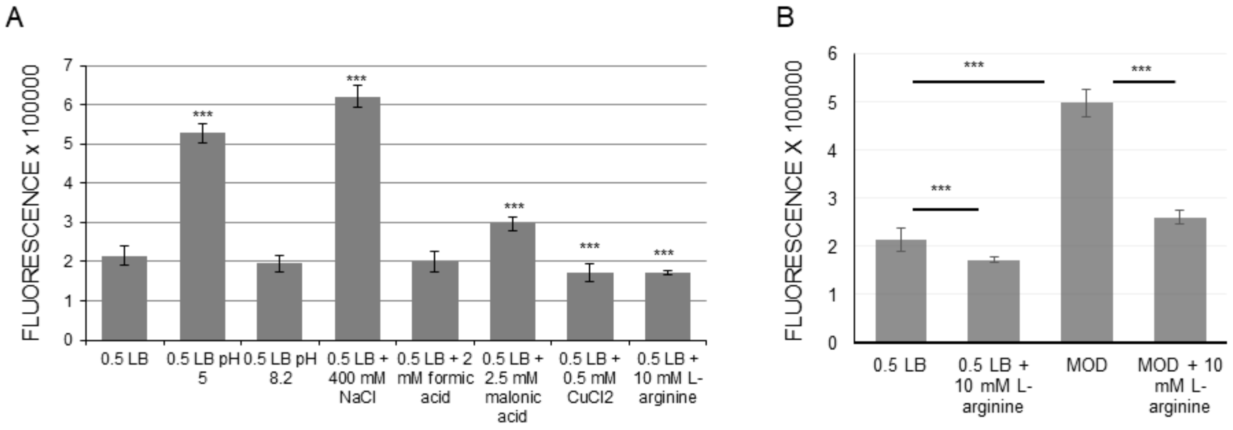


Figure 4: Promoter activity of the region upstream of the major TSS. *E. coli* Top10 transformed with pProbe-NT-majorTSS was incubated until an OD_{600} of 0.5 was reached, and fluorescence was measured. The measured fluorescence of the empty vector was 393 ± 410, which is significantly lower than the fluorescence of pProbe-NT-majorTSS at every investigated condition ($p<0.001$). Significant changes between 0.5 LB and investigated stress conditions were calculated by Student's t-test and marked with asterisks (*** $p<0.001$). **A** Promoter activity in 0.5 LB medium at different conditions **B** Promoter activity in MOD medium with and without L-arginine.

The region 300 bp upstream of the minor TSS (Figure 1A) was also investigated for promoter activity. The sequence shows significant promoter activity at all tested conditions indicated by a 4- to 14-fold increase in fluorescence intensity compared to the vector control (Supplementary Figure S3).

Initial functional characterization of SlyC

To investigate, whether a SlyC protein can be expressed, the plasmid pEGFP-SlyC, containing SlyC C-terminally fused to EGFP, was transformed into *E. coli* Top10, and fusion protein expression was induced by IPTG. The fluorescence intensity was measured 1 h after induction. Compared to the uninduced culture, a 5.7-fold increased fluorescence intensity was observed after induction, indicating that a SlyC-EGFP fusion protein is stably expressed in *E. coli* (Figure 5A).A strand-specific translationally arrested mutant of *slyC* was cloned using the genome editing method (Kim *et al.* 2014). The third codon of the *slyC* sequence was changed into a premature stop codon by the introduction of two point mutations (Figure 5B). The AA sequence of the annotated gene *slyA* is not affected by these point mutations, since they lead to a CTT instead of an TTA codon still encoding leucine.

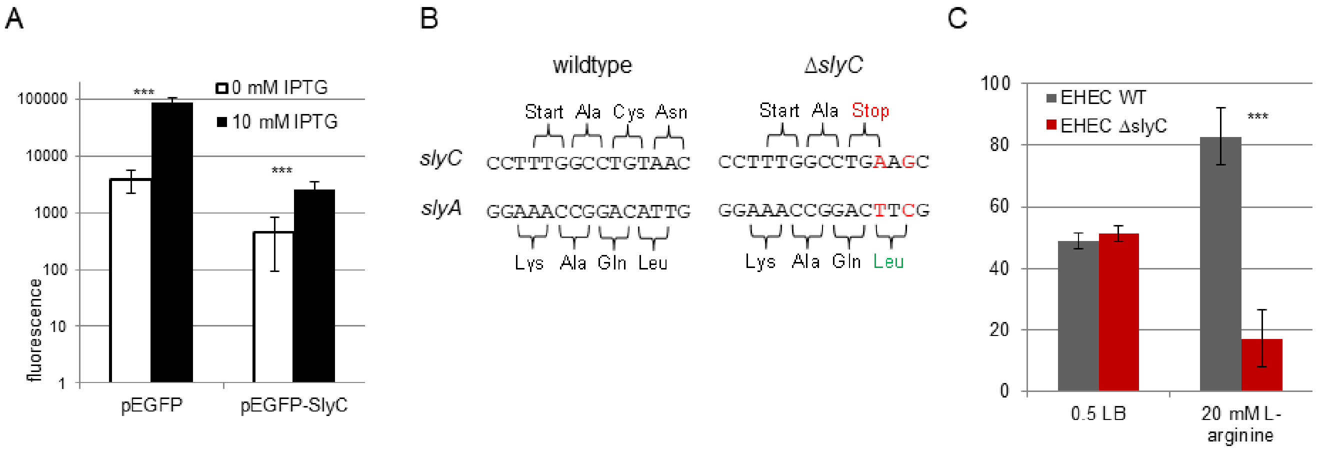


Figure 5: Phenotype of SlyC. **A** Expression of SlyC C-terminally fused with EGFP. *E. coli* Top10 was transformed with pEGFP-SlyC, and expression of the fusion protein was induced with 10 mM IPTG. Fluorescence values in logarithmic scale with and without induction are depicted. The empty pEGFP vector was used as a positive control. The experiment was performed in triplicate. *** p<0.001. **B** Cloning of a translationally arrested Δ*slyC* mutant. Introduction of two point mutations into the DNA sequence of *slyC* changed the third codon encoding cysteine into a premature stop codon. Additionally, a cut site for the restriction enzyme *Alu*I is created. The two point mutations do not influence the AA sequence of the antiparallel overlapping gene *slyA.* **C** Ratio in percent of EHEC wild type over EHEC Δ*slyC* after competitive growth. Wild type and mutant were mixed in equal ratio, and after 18 h incubation the ratio was determined by Sanger sequencing. In 0.5 LB, no change compared to the inoculation ratio occurred, but when the medium was supplemented with 20 mM L-arginine the wild type shows a significant growth advantage. The experiment was performed in triplicate. *** p<0.001.

Competitive growth experiments were performed to search for a phenotype of ∆*slyC*. EHEC wild type and the translationally arrested mutant Δ*slyC* were mixed in equal ratio and grown in competition under different conditions. After 18 h, the ratio of wild type to ∆*slyC* was determined. In 0.5 LB, neither the wild type nor the mutant has an advantage (Figure 5C). However, when the medium is supplemented with 20 mM L-arginine, the ratio shifted significantly: the wild type possesses a clear growth advantage resulting in a 4:1 ratio after competition expressed through a competitive index of 0.24. All further tested growth conditions do not show a phenotype (Supplementary Figure S4). It was tried to restore the phenotype of the wild type by transforming EHEC ∆*slyC* with an arabinose inducible plasmid carrying the intact *slyC* ORF. However, complementation *in trans* was not successful, the ratio of EHEC ∆*slyC* + pBAD-*slyC* over EHEC ∆*slyC* + pBAD-Δ*slyC* was 1:1 after competitive growth (data not shown).

DISCUSSION

Combined RNAseq and RIBOseq are high-throughput next generation sequencing methods, which led to the discovery of numerous non-annotated ORFs in eukaryotes (Aspden *et al.* 2014; Bazzini *et al.* 2014; Ingolia *et al.* 2014; Smith *et al.* 2014), of which some are also detected antisense to annotated genes (Fields *et al.* 2015). In bacteria, hundreds of non-annotated short intergenic ORFs appeared to be translated and have been overlooked by genome annotation algorithms (Neuhaus *et al.* 2016; Baek *et al.* 2017; Hücker *et al.* 2017a). Also, the OLG pair *nog1/citC* (Fellner *et al.* 2015) was initially discovered by transcription of *nog1* in RNAseq data (Landstorfer *et al.* 2014). Therefore, careful examination of RNAseq and RIBOseq data might discover additional OLG pairs.

Transcriptome and translatome sequencing of EHEC Sakai resulted in the discovery of the novel protein-coding gene *slyC,* which is completely embedded in the antisense reading frame -2 in the gene of the transcriptional regulator *slyA* (Figure 2). *SlyC* forms an operon together with the outer membrane protein *slyB* (Figure 1C), and it is even possible that further ORFs downstream of *slyC*, also in antisense to *slyA*, are part of this operon. Translation of *slyC* starts at a TTG start codon, which is used for initiation of only 3% of all *E. coli* genes (BLATTNER *et al.* 1997). However, it is highly unlikely that a downstream ATG is the correct start codon, because changing the third *slyC* codon into a premature stop codon, resulted in a phenotype in competitive growth experiments (Figure 5C), which is not expected if this sequence would be located upstream of the real start codon. Furthermore, it can be excluded that *slyC* is a non-coding RNA instead of a protein-coding gene due to the following reasons: (i) the ORF is annotated as a protein-coding gene in other *E. coli* strains, (ii) *slyC* shows RIBOseq signals at all investigated growth conditions (Table 1) with translatability (ribosomal coverage value) of its mRNA being in the range of annotated EHEC genes (NEUHAUS *et al.* 2017), (iii) a SlyC-EGFP fusion protein could be expressed (Figure 5A), and (iv) a phenotype was detected after changing only two nucleotides, which is not expected for a ncRNA, because those regulate expression of target genes by base pairing of longer sequences (GOTTESMAN *et al.* 2006), which should not be impeded after such a small change.

The phylostratigraphic analysis shows that *slyA* is evolutionary old, since homologs are present in all bacterial phyla (Supplementary Figure S1). In contrast, the evolutionary age of *slyC* is much younger, because homologs are only present in *Enterobacteriales,* always overlapping in antisense to *slyA*, whereas in distantly related species the *slyC* ORF is not detectable any more (Figure 3). Also, *slyB* has a broader taxonomic distribution than *slyC* (Supplementary Figure S2). This indicates that *slyC* has originated *de novo* by overprinting (KEESE AND GIBBS 1992; PAVESI *et al.* 2013) before the separation of the *Enterobacteriales*. We hypothesize that the *slyC* ORF was created by some point mutations. Occasionally, the ORF was co-transcribed together with the upstream gene *slyB*, because the ρ-independent terminator downstream of *slyB* failed to terminate transcription in some transcription events. A promoter and other regulatory elements are not necessary for *slyC* transcription in this scenario. Then, the elongated mRNA was used as a template of translation and an ancestral SlyC protein may have been produced. If the novel protein would have been detrimental, i.e., misfolded proteins are cytotoxic (DRUMMOND AND WILKE 2008), the *slyC* ORF would have been lost again (HUVET AND STUMPF 2014). Therefore, SlyC should have provided some initial, unknown beneficial function. Further point mutations may have optimized SlyC function.

SlyC is now the third gene discovered in EHEC, which likely originated by overprinting in antisense to an established gene (Delaye *et al.* 2008; Fellner *et al.* 2015). For other known OLGs the evolutionary origin was not investigated (Silby and Levy 2008; Tunca *et al.* 2009). Maybe overprinting is a more common mechanism for the creation of novel genes than presumed in the past. Indeed, deep RNAseq discovered that every area of the genome can be transcribed (Neme and Tautz 2016), and potentially novel genes may arise from such transcripts.

This study also gives some hints of the putative SlyC function. Supposably, SlyC is integrated into the cell membrane, because in other *E. coli* strains the ORF is annotated as an outer membrane lipoprotein, two transmembrane helices were predicted, and a signal sequence is present. The proteinogenic AA L-arginine clearly plays a role in SlyC function. The σ^{70} promoter of *slyBC* overlaps with the sequences of two ARG boxes (Figure 1), which is also the case for the L-arginine biosynthesis genes (Charlier *et al.* 1992). At high L-arginine concentrations, the transcription factor ArgR forms a hexamer together with six L-arginine molecules, which stabilize the structure and function as corepressors (Van Duyne *et al.* 1996). Binding of an ArgR hexamer to two ARG boxes separated by a 3 bp spacer nucleotides inhibits target gene transcription (Cunin *et al.* 1986). Later, it was discovered that ArgR can repress transcription without overlapping with the promoter and also activates transcription of other target genes, e.g., the ast-pathway, which leads to L-arginine catabolism, is induced by ArgR binding (Caldara *et al.* 2007). Cho *et al.* (2015) investigated the ArgR regulon in *E. coli* using ChIP-exo-seq, and they discovered 62 ArgR binding regions, of which all contain two ARG boxes. However, they did not detect ArgR binding upstream of *slyBC*. Experimental data confirms that activity of the *slyBC* promoter is somewhat repressed in the presence of L-arginine with highest promoter activity in a medium without any L-arginine (Figure 4), indicating that the ARG boxes upstream of *slyBC* are real ArgR binding sites and not false positives only identified by a bioinformatic prediction. Furthermore, perfect agreement to the ARG box consensus motif is not required especially for the second ARG box (Figure 1D), because ArgR still binds even if one half of the second box is mutated (Charlier *et al.* 1992). The phenotype detected in the competitive growth experimentsconfirms the importance of L-arginine for SlyC function: the EHEC Δ*slyC* mutant has a significant growth disadvantage in the presence of 20 mM L-arginine compared to the wild type (Figure 5C). This phenotype is L-arginine specific and does not occur at other stress conditions (Supplementary Figure S4). Maybe a complementation *in trans* failed, because the plasmid only contained the protein-coding sequence of SlyC but not its promoter region. Thus, regulation by the ARG boxes was absent.

The promoter of *slyBC* also shows significantly altered activity at other stress conditions: under salt and acid stress the activity is increased (Figure 4A). In the environment, EHEC faces acid stress, when it passes through the stomach of mammalian hosts. EHEC possesses three different acid resistance systems, of which one system uses the arginine decarboxylase AdiA and, thus, this system is L-arginine dependent (CASTANIE-CORNET *et al.* 1999; LIM *et al.* 2010). Indeed, the arginine dependent system can protect EHEC challenged at pH 2, and it is effective against weak organic acids (LIN *et al.* 1996). Hypothetically, SlyC could be involved in this L-arginine dependent acid resistance system. Alternatively, SlyC might interact with one of the three L-arginine uptake systems in the cell membrane (CUNIN *et al.* 1986).

The gene *slyA,* antisense to *slyC,* was first annotated as a hemolysin in *Salmonella* (LIBBY *et al.* 1994), but then it turned out that *slyA* is a transcriptional regulator inducing the expression of a hemolysin (LUDWIG *et al.* 1995; LUDWIG *et al.* 1999). Proteome analysis of SlyA knock-out and overexpression indicate that also chaperons, histidine biosynthesis genes, and acid resistance genes belong to the SlyA regulon (SPORY *et al.* 2002). Whether SlyA also regulates the expression of *slyBC* is not known, however, the SlyA consensus motif (MCVICKER *et al.* 2011) was not detected upstream of the *slyBC* transcriptional start site. Further functional characterization of SlyC must await future studies.

REFERENCES

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local alignment search tool. Journal of molecular biology 215**:** 403-410.

Aspden, J. L., Y. C. Eyre-Walker, R. J. Phillips, U. Amin, M. A. Mumtaz *et al.*, 2014 Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. Elife 3**:** e03528.

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

Bazzini, A. A., T. G. Johnstone, R. Christiano, S. D. Mackowiak, B. Obermayer *et al.*, 2014 Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J 33**:** 981-993.

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

Bernhofer, M., E. Kloppmann, J. Reeb and B. Rost, 2016 TMSEG: Novel prediction of transmembrane helices. Proteins 84**:** 1706-1716.

Blattner, F. R., G. Plunkett, 3rd, C. A. Bloch, N. T. Perna, V. Burland *et al.*, 1997 The complete genome sequence of *Escherichia coli* K-12. Science 277**:** 1453-1462.

Caldara, M., D. Charlier and R. Cunin, 2006 The arginine regulon of *Escherichia coli*: whole-system transcriptome analysis discovers new genes and provides an integrated view of arginine regulation. Microbiology 152**:** 3343-3354.

Caldara, M., P. N. Minh, S. Bostoen, J. Massant and D. Charlier, 2007 ArgR-dependent repression of arginine and histidine transport genes in *Escherichia coli* K-12. J Mol Biol 373**:** 251-267.

Castanie-Cornet, M. P., T. A. Penfound, D. Smith, J. F. Elliott and J. W. Foster, 1999 Control of acid resistance in *Escherichia coli*. J Bacteriol 181**:** 3525-3535.

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

Charlier, D., M. Roovers, F. Van Vliet, A. Boyen, R. Cunin *et al.*, 1992 Arginine regulon of *Escherichia coli* K-12. A study of repressor-operator interactions and of in vitro binding affinities versus in vivo repression. J Mol Biol 226**:** 367-386.

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

Cho, S., Y. B. Cho, T. J. Kang, S. C. Kim, B. Palsson *et al.*, 2015 The architecture of ArgR-DNA complexes at the genome-scale in *Escherichia coli*. Nucleic Acids Res 43**:** 3079-3088.

Cunin, R., N. Glansdorff, A. Pierard and V. Stalon, 1986 Biosynthesis and metabolism of arginine in bacteria. Microbiol Rev 50**:** 314-352.

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

Delcher, A. L., K. A. Bratke, E. C. Powers and S. L. Salzberg, 2007 Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23**:** 673-679.

Drummond, D. A., and C. O. Wilke, 2008 Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134**:** 341-352.

Fellner, L., N. Bechtel, M. A. Witting, S. Simon, P. Schmitt-Kopplin *et al.*, 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., S. Simon, C. Scherling, M. Witting, S. Schober *et al.*, 2015 Evidence for the recent origin of a bacterial protein-coding, overlapping orphan gene by evolutionary overprinting. BMC evolutionary biology 15**:** 1.

Fields, A. P., E. H. Rodriguez, M. Jovanovic, N. Stern-Ginossar, B. J. Haas *et al.*, 2015 A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Mol Cell 60**:** 816-827.

Flaherty, B. L., F. Van Nieuwerburgh, S. R. Head and J. W. Golden, 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.

Goldberg, T., M. Hecht, T. Hamp, T. Karl, G. Yachdav *et al.*, 2014 LocTree3 prediction of localization. Nucleic Acids Res 42**:** W350-355.

Gottesman, S., C. A. McCullen, M. Guillier, C. K. Vanderpool, N. Majdalani *et al.*, 2006 Small RNA regulators and the bacterial response to stress. Cold Spring Harb Symp Quant Biol 71**:** 1-11.

Hayashi, T., K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii *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.

Hücker, S. M., T. Goldberg, A. Schafferhans, M. Bernhofer, G. Vestergaard *et al.*, 2017a Discovery of numerous novel small genes in the intergenic regions of the *Escherichia coli* O157:H7 Sakai genome. PloS One.

Hücker, S. M., S. Simon, S. Scherer and K. Neuhaus, 2017b 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 364.

Huvet, M., and M. P. Stumpf, 2014 Overlapping genes: a window on gene evolvability. BMC genomics 15**:** 721.

Ingolia, N. T., G. A. Brar, N. Stern-Ginossar, M. S. Harris, G. J. Talhouarne *et al.*, 2014 Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep 8**:** 1365-1379.

Ingolia, N. T., S. Ghaemmaghami, J. R. Newman and J. S. Weissman, 2009 Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. Science 324**:** 218-223.

Jensen, K. T., L. Petersen, S. Falk, P. Iversen, P. Andersen *et al.*, 2006 Novel overlapping coding sequences in *Chlamydia trachomatis*. FEMS Microbiol Lett 265**:** 106-117.

Kall, L., A. Krogh and E. L. Sonnhammer, 2004 A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338**:** 1027-1036.

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

Kim, J., A. M. Webb, J. P. Kershner, S. Blaskowski and S. D. Copley, 2014 A versatile and highly efficient method for scarless genome editing in *Escherichia coli* and *Salmonella enterica*. BMC Biotechnol 14**:** 84.

Kim, W., M. W. Silby, S. O. Purvine, J. S. Nicoll, K. K. Hixson *et al.*, 2009 Proteomic detection of non-annotated protein-coding genes in *Pseudomonas fluorescens* Pf0-1. PLoS One 4**:** e8455.

Krakauer, D. C., 2000 Stability and evolution of overlapping genes. Evolution 54**:** 731-739.

Landstorfer, R., S. Simon, S. Schober, D. Keim, S. Scherer *et al.*, 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.

Latif, H., H. J. Li, P. Charusanti, B. Ø. Palsson and R. K. Aziz, 2014 A gapless, unambiguous genome sequence of the enterohemorrhagic *Escherichia coli* O157: H7 strain EDL933. Genome announcements 2**:** e00821-00814.

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

Li, W., A. Cowley, M. Uludag, T. Gur, H. McWilliam *et al.*, 2015 The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic acids research 43**:** W580-W584.

Libby, S. J., W. Goebel, A. Ludwig, N. Buchmeier, F. Bowe *et al.*, 1994 A cytolysin encoded by *Salmonella* is required for survival within macrophages. Proc Natl Acad Sci U S A 91**:** 489-493.

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

Lim, J. Y., J. Yoon and C. J. Hovde, 2010 A brief overview of *Escherichia coli* O157:H7 and its plasmid O157. J Microbiol Biotechnol 20**:** 5-14.

Lin, J., M. P. Smith, K. C. Chapin, H. S. Baik, G. N. Bennett *et al.*, 1996 Mechanisms of acid resistance in enterohemorrhagic *Escherichia coli*. Appl Environ Microbiol 62**:** 3094-3100.

Ludwig, A., S. Bauer, R. Benz, B. Bergmann and W. Goebel, 1999 Analysis of the SlyA-controlled expression, subcellular localization and pore-forming activity of a 34 kDa haemolysin (ClyA) from *Escherichia coli* K-12. Mol Microbiol 31**:** 557-567.

Ludwig, A., C. Tengel, S. Bauer, A. Bubert, R. Benz *et al.*, 1995 SlyA, a regulatory protein from *Salmonella typhimurium*, induces a haemolytic and pore-forming protein in *Escherichia coli*. Mol Gen Genet 249**:** 474-486.

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

McVeigh, A., A. Fasano, D. A. Scott, S. Jelacic, S. L. Moseley *et al.*, 2000 IS1414, an *Escherichia coli* insertion sequence with a heat-stable enterotoxin gene embedded in a transposase-like gene. Infect Immun 68**:** 5710-5715.

McVicker, G., L. Sun, B. K. Sohanpal, K. Gashi, R. A. Williamson *et al.*, 2011 SlyA protein activates *fimB* gene expression and type 1 fimbriation in *Escherichia coli* K-12. J Biol Chem 286**:** 32026-32035.

Merino, E., P. Balbas, J. L. Puente and F. Bolivar, 1994 Antisense overlapping open reading frames in genes from bacteria to humans. Nucleic Acids Res 22**:** 1903-1908.

Miller, W. G., J. H. Leveau and S. E. Lindow, 2000 Improved *gfp* and *inaZ* broad-host-range promoter-probe vectors. Mol Plant Microbe Interact. 13**:** 1243-1250.

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

Neme, R., and D. Tautz, 2016 Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence. Elife 5**:** e09977.

Neuhaus, K., R. Landstorfer, L. Fellner, S. Simon, H. Marx *et al.*, 2016 Translatomics combined with transcriptomics and proteomics reveals novel functional, recently evolved orphan genes in *Escherichia coli* O157:H7 (EHEC). BMC Genomics 7**:** 133.

Neuhaus, K., R. Landstorfer, S. Simon, S. Schober, P. R. Wright *et al.*, 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 18**:** 216.

Pavesi, A., G. Magiorkinis and D. G. Karlin, 2013 Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of Deltaretroviruses. PLoS Comput Biol 9**:** e1003162.

Petersen, T. N., S. Brunak, G. von Heijne and H. Nielsen, 2011 SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature methods 8**:** 785-786.

Rancurel, C., M. Khosravi, A. K. Dunker, P. R. Romero and D. Karlin, 2009 Overlapping genes produce proteins with unusual sequence properties and offer insight into *de novo* protein creation. J Virol 83**:** 10719-10736.

Rosenfeld, E., C. Duport, A. Zigha and P. Schmitt, 2005 Characterization of aerobic and anaerobic vegetative growth of the food-borne pathogen *Bacillus cereus* F4430/73 strain. Can J Microbiol 51**:** 149-158.

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

Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice *et al.*, 2000 Artemis: sequence visualization and annotation. Bioinformatics 16**:** 944-945.

Saldana, Z., E. Sanchez, J. Xicohtencatl-Cortes, J. L. Puente and J. A. Giron, 2011 Surface structures involved in plant stomata and leaf colonization by shiga-toxigenic *Escherichia coli* O157:H7. Front Microbiol 2**:** 119.

Silby, M. W., and S. B. Levy, 2008 Overlapping protein-encoding genes in *Pseudomonas fluorescens* Pf0-1. PLoS Genet 4**:** e1000094.

Smith, J. E., J. R. Alvarez-Dominguez, N. Kline, N. J. Huynh, S. Geisler *et al.*, 2014 Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. Cell Reports 7**:** 1858-1866.

Solovyev, V. V., and T. V. Tatarinova, 2011 Towards the integration of genomics, epidemiological and clinical data. Genome Med 3**:** 48.

Spory, A., A. Bosserhoff, C. von Rhein, W. Goebel and A. Ludwig, 2002 Differential regulation of multiple proteins of *Escherichia coli* and *Salmonella enterica* serovar Typhimurium by the transcriptional regulator SlyA. J Bacteriol 184**:** 3549-3559.

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

Tian, G., D. Lim, J. Carey and W. K. Maas, 1992 Binding of the arginine repressor of *Escherichia coli* K12 to its operator sites. J Mol Biol 226**:** 387-397.

Tunca, S., C. Barreiro, J. J. Coque and J. F. Martin, 2009 Two overlapping antiparallel genes encoding the iron regulator DmdR1 and the Adm proteins control siderophore and antibiotic biosynthesis in *Streptomyces coelicolor* A3(2). FEBS J 276**:** 4814-4827.

Van Duyne, G. D., G. Ghosh, W. K. Maas and P. B. Sigler, 1996 Structure of the oligomerization and L-arginine binding domain of the arginine repressor of *Escherichia coli*. J Mol Biol 256**:** 377-391.

Veloso, F., G. Riadi, D. Aliaga, R. Lieph and D. S. Holmes, 2005 Large-scale, multi-genome analysis of alternate open reading frames in bacteria and archaea. Omics 9**:** 91-105.

Wang, L. F., S. S. Park and R. H. Doi, 1999 A novel *Bacillus subtilis* gene, *antE*, temporally regulated and convergent to and overlapping *dnaE*. J Bacteriol 181**:** 353-356.

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

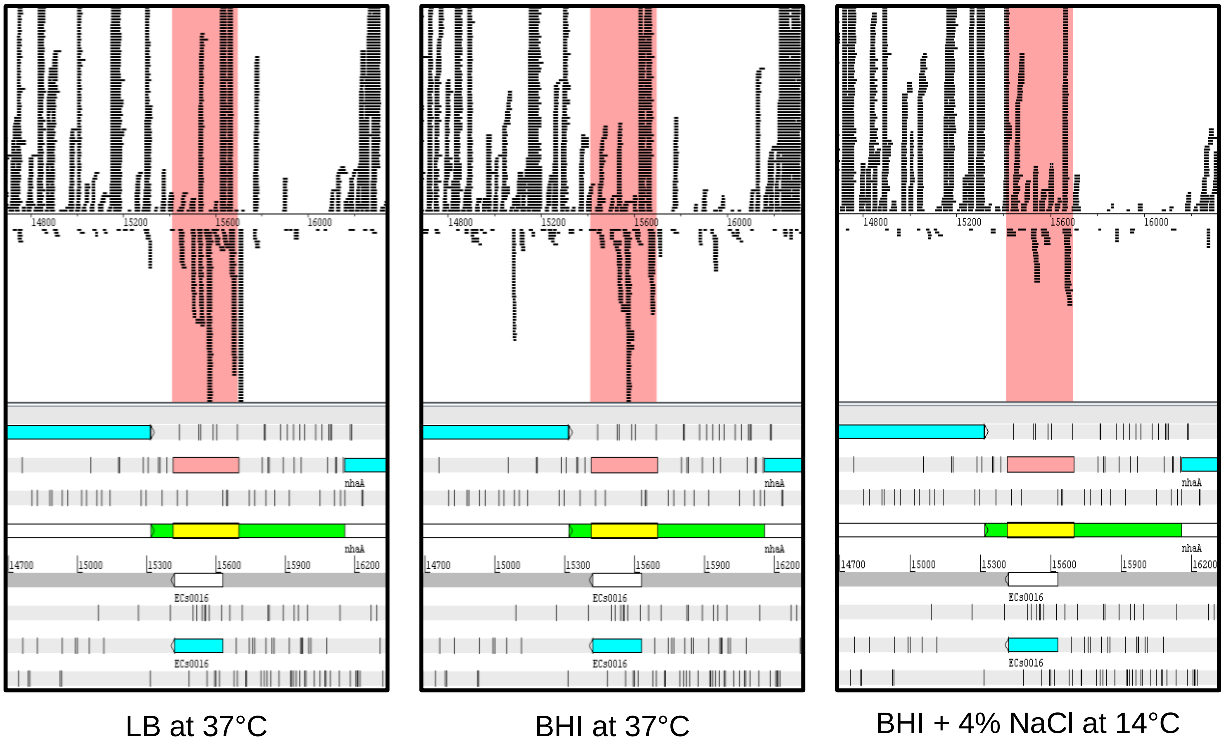
Zhao, L., L. Liu, W. Leng, C. Wei and Q. Jin, 2011 A Proteogenomic analysis of *Shigella flexneri* using 2D LC-MALDI TOF/TOF. BMC Genomics 12**:** 528.

Zuker, M., 2003 Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31**:** 3406-3415.

2.11 Overlapping gene OLGECs0016

2.11.1 Transcription and translation

A 285 bp ORF overlaps antiparallel in reading frame -1 to the annotated gene ECs0016, which encodes a GEF-like protein. Expression of this ORF, designated OLGECs0016, was discovered in the LB condition (Table 4A, Figure 17). Transcription of OLGECs0016 is even higher in the BHI conditions, but under these conditions the ORF is only weakly translated. The RCV in BHI is below the threshold at which OLGECs0016 would be considered expressed. The annotated gene ECs0016 is transcribed in all conditions with the highest transcription in LB (Table 4B), whereas it is only weakly translated at all conditions (Figure 17), showing low RCVs.



LB at 37°C

BHI at 37°C

BHI + 4% NaCl at 14°C

Figure 17: Translation of ECs0016 and OLGECs0016 visualized in the Artemis genome browser. The RIBOseq reads of every tested growth condition were mapped strand specifically to the EHEC Sakai genome. The annotated genes are colored in blue and OLGECs0016 is highlighted in pink.

Table 4: Expression of OLGECs0016 (A) and the annotated gene ECs0016 (B).

A

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	878.4	222.3	0.25	0.76
BHI, 37°C	2722.9	60.8	0.03	0.77
BHI + 4% NaCl, 14°C	2105.7	19.5	0.01	0.32

B

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	423.8	67.8	0.16	0.78
BHI, 37°C	142.7	15.2	0.12	0.82
BHI + 4% NaCl, 14°C	93.3	2.0	0.02	0.62

*Mean values of the two biological replicates are shown.

The start codon of OLGECs0016 is unknown. The most upstream positioned start codon is a TTG, which would lead to a 94 AA protein. Downstream of the TTG, a CTG start codon is present and then the protein would consist of 80 AAs. The third putative start codon is a GTG leading to a 75 AA protein (Figure 18). The first and the second start codon are supported by the prediction of a putative Shine-Dalgarno sequence upstream (Ma et al., 2002). There are SD sequences predicted 15 bp upstream of the first start codon with ΔG° of -4.2, and 17 bp upstream of start codon 2 with ΔG° of -4.1. The sequence of OLGECs0016 with 94 AAs was used as a query for a BLASTP search against the Refseq database to find annotated homologs. Indeed, several homologs were detected in other *E. coli* strains, in which the ORF is annotated as 'hypothetical'. The best hit has an e-value of $4×10^{-61}$, a query coverage of 100%, and an identity of 98%. The software PredictProtein (Yachdav et al., 2014) predicts that OLGECs0016 is secreted, however, no signal peptide was predicted by SignalP 4.1 (Petersen et al., 2011). PredictProtein did not find any disulfide bonds or transmembrane helices.

```
tgtgaagaagttttttgacgacctgacccgctaacctccccaaaagcctgcccgtgggcaggcctgggt
aaaaAtagggtgcgttgaagatatgcgagcacctgtaaagtggcggggatcactccccgccgTTGCTC
TTACTCGGATTCGTAAGCCGTGAAAACAGCAACCTCCGTCTGGCCAGTTCGGATGTGAACCTCACAGA
GGTCTTTTCTCGTTACCAGCGCCGCCACTACGGCGGTGATACAGATGACGATCAGGGCGACAATCATC
ACCTTATGCTGCTTCATTGCTCTCTTCTCCTTGACCTTACGGTCAGTAAGAGGCACTCTACATGTGTT
CAGCATATAGGGGGCCTCGGGTTGATGGTAAAATATCACTCGGGGCTTTTCTCTATCTGCCGTTCAGC
TAATGCCTGAgacagacagcctcaagcacccgccgctatta
```

Figure 18: DNA sequence of OLGECs0016 including its upstream and downstream region. The sequence of OLGECs0016 is written in capital letters and in blue print. The three putative start codons are highlighted in green, and the stop codon is highlighted in red. The predicted SD sequences are highlighted in light blue. The predicted σ^{70} promoter is underlined. The transcriptional start site determined using 5' RACE is highlighted in pink.

2.11.2 Characterization of the promoter region

The software BPROM predicts a σ^{70} promoter 95 bp upstream of the first start codon with an LDF-score of 1.69 (Figure 18). To confirm an active promoter, 300 bp of the sequence upstream of the first start codon were cloned into the plasmid pProbe-NT (Table 5). An active promoter causes EGFP production, measurable by an increase in fluorescence intensity. However, no promoter activity was detectable at any of the investigated conditions (LB medium with different supplementations): the fluorescence intensity was similar to the plasmid control (data not shown).

Table 5: Oligonucleotides used for the characterization of the OLGECs0016 promoter region.

name	sequence	purpose
OLGECs0016+265R	attagctgaacggcagat	5' RACE, reverse transcription
OLGECs0016+179R	tgtagagtgcctcttactgaccgtaa	5' RACE, 1st PCR
OLGECs0016+137R	gcaatgaagcagcataaggtgatgat	5' RACE, 2nd PCR
OLGECs0016-300F-*Sal*I	tacgGTCGACtcaagtctgtccgcggtg	cloning pProbe-OLGECs0016
OLGECs0016-18R-*Eco*RI	attaGAATTCcggcggggagtgatcccc	cloning pProbe-OLGECs0016

The transcriptional start site (TSS) was determined using 5' RACE (Table 5). A single TSS was identified 58 bp upstream of the first start codon (Figure 18), despite no promoter activity was found beforehand. It is possible that OLGECs0016 does not possess its own promoter, but instead it is transcribed in an operon together with the annotated gene ECs0015 upstream. In this case, the TSS is expected upstream of ECs0015, which is not the case. Maybe the detected TSS is a degradation product of the putative polycistronic mRNA ECs0015-OLGECs0016. However, RNAseq shows that

OLGECs0016 is transcribed at LB. Probably, the promoter activity is dependent on regulatory elements, which bind further upstream of the 300 bp tested.

2.11.3 EGFP-fusion protein expression

In order to investigate if an OLGECs0016 protein can be expressed, a C-terminal EGFP-fusion protein was cloned (Table 6). The plasmid pEGFP-OLGECs0016 was transformed into *E. coli* Top10 and protein expression was induced with 10 mM IPTG. As a positive control, the empty plasmid was used. The protein expression was tested for every putative start codon. The induced cultures for all putative start codons show a significant increase in fluorescence intensity compared to the uninduced cultures (Figure 19). However, using the first start codon, the fluorescence intensity is only 4.4-fold higher, whereas OLGECs0016 with start codons 2 or 3 lead to a 9-fold increased fluorescence intensity. This experiment does not provide a distinct answer, which of the three possible start codons is the correct one.

Table 6: Oligonucleotides used for cloning of an OLGECs0016-EGFP fusion protein.

name	sequence	purpose
OLGECs0016+1F-*Pst*I	caggCTGCAGgttgctcttactcggattcgt	cloning pEGPF-OLGECs0016 S1
OLGECs0016_2+46F-*Pst*I	attcCTGCAGgctggccagttcggatgtgaa	cloning pEGPF-OLGECs0016 S2
OLGECs0016_3+61F-*Pst*I	ctacCTGCAGggtgaacctcacagaggtctt	cloning pEGPF-OLGECs0016 S3
OLGECs0016+266R-*Nco*I	taatCCATGGcggcattagctgaacggcaga	cloning pEGPF-OLGECs0016

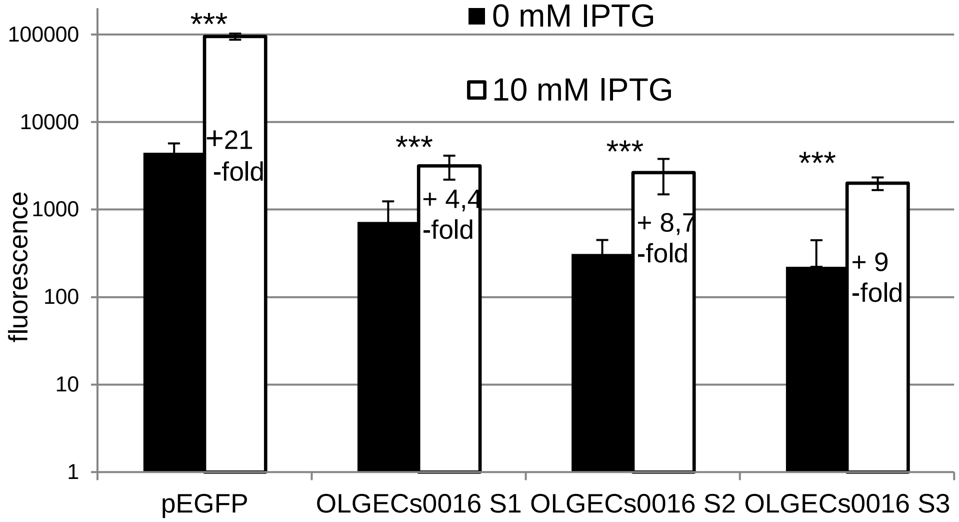


Figure 19: Expression of OLGECs0016-EGFP fusion proteins. *E. coli* Top10 carrying pEGFP plasmids was incubated in 0.5×LB until an OD_{600} of 0.5 was reached. Then, one culture each was induced with 10 mM IPTG and incubation was continued for 1 h. The cells were harvested, washed with PBS and adjusted to an OD_{600} of 0.6. The fluorescence intensity was measured in quadruplicates in black 96-well microtiter plates (Wallac Victor^3, excitation 485 nm, emission 535 nm, measuring time 1 s). The fluorescence of *E. coli* Top10 was subtracted as background. The experiment was performed in triplicate. Significant changes between induced and uninduced cultures were calculated using the Student's t-test (*** $p<0.001$).

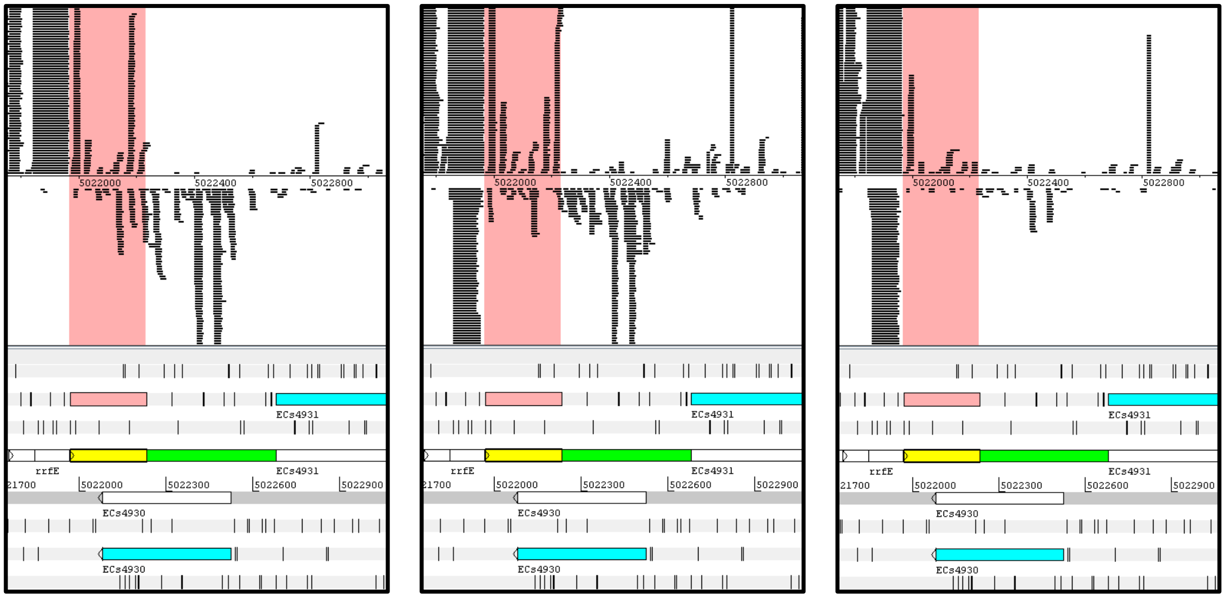
The RIBOseq data and the expression of an OLGECs0016-EGFP fusion protein suggest that OLGECS0016 is translated into a protein. The function of this protein is unknown, because in this study, no strand-specific knock-out mutant could be cloned and the BLASTP search only discovered homologs encoding hypothetical proteins.

In conclusion, RNAseq supports that OLGECs0016 is transcribed (Table 4A). Further, translation of the OLGECs0016 mRNA was confirmed by RIBOseq (Figure 17). Evidence of an OLGECs0016 protein was obtained by expression of an OLGECs0016-EGFP fusion protein (Figure 19). Therefore, OLGECs0016/ECs0016 likely represent another novel OLG pair of EHEC Sakai. However, determination of the correct start codon and the search for a phenotype must await future studies. The predicted σ^{70} promoter (Figure 18) is probably not the correct promoter, because it is too far upstream of the detected TSS and the DNA stretch containing this promoter did not show activity. The location and the type of the OLGECs0016 promoter remain unclear.

2.12 Overlapping gene OLGECs4930

2.12.1 Transcription and translation

RNAseq and RIBOseq data showed transcription and translation at optimal growth conditions of a 267 bp ORF located downstream of the 5S rRNA gene *rrfE* (Figure 20). Additionally, the ORF overlaps tail-to-tail in reading frame -1 to the hypothetical protein ECs4930, thus, it was designated OLGECs4930. *RrfE* appears to be translated, because the mRNA footprints are co-purified together with the ribosome including the ribosomal RNA, and the rRNA depletion step performed later is not able to remove all rRNA. Interestingly, OLGECs4930 shows the highest transcription in BHI stress (Table 7A), but translation is almost turned off. The highest translation occurs in LB at 37°C, even though transcription was lowest here compared to the other two conditions. Therefore, the translatability in LB is very high, resulting in a higher RCV than that of many of the short annotated genes (Figure 15A). The annotated mother gene also shows a high translatability in LB at 37°C (Table 7B). At BHI stress, ECs4930 is almost not expressed (Figure 20).



LB at 37°C BHI at 37°C BHI + 4% NaCl at 14°C

Figure 20: Translation of ECs4930 and OLGECs4930 visualized in the Artemis genome browser. The RIBOseq reads of every tested growth condition were mapped strand specifically to the EHEC Sakai genome. The annotated genes are colored in blue and OLGECs4930 is highlighted in pink.

Table 7: Expression of OLGECs4930 (A) and the annotated mother gene ECs4930 (B).

A

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	24.2	110.0	6.24	0.73
BHI, 37°C	93.1	54.5	0.59	0.79
BHI + 4% NaCl, 14°C	168.3	4.2	0.03	0.43

B

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	26.0	136.6	5.92	0.86
BHI, 37°C	23.4	25.6	1.29	0.85
BHI + 4% NaCl, 14°C	9.0	2.4	0.34	0.60

*Mean values of the two biological replicates are shown.

The farthest upstream possible start codon of OLGECs4930 is a very rare ATA codon, which would lead to an 88 AA protein. Another rare CTG start codon is present downstream, which would lead to an 82 AA protein. In addition, a canonical TTG start codon directly followed by a GTG would be possible, and then the protein would consist of 74 or 73 AAs, respectively (Figure 21). Upstream of all putative start codons a SD sequence is predicted. The SD sequence 16 bp upstream of the first putative start codon has a ΔG° of -5,5, another SD is only 1 bp upstream of the second putative start codon with a ΔG° of -3.8 and the third SD sequence is 14 bp upstream with a ΔG° of -4.4. Annotated homologs were searched using the AA sequence of the 88 AA protein using BLASTP. The closest homolog was detected in *E. coli* strain 90.0091 with an e-value of 1×10^{-56}, 100% query coverage, and 100% identity. The protein is annotated as disulfide interchange domain protein DsbA. Additional homologs to further hypothetical proteins with higher e-value, lower coverage and lower identity values are present in many other *Escherichia* and *Shigella* strains. The software PredictProtein does predict neither transmembrane helices, nor a signal peptide, but OLGECs4930 should be located in the inner membrane. Furthermore, a disulfide bond is predicted.

cagattaaatcagaacgcagaagcggtctgataaaacagaatttgcctggcggccttagcgcggtggt

cccacctgaccccatgccgaactcagaagtgaaacgccgtagcgccgatggtagtgtggggtctcccc

atgcg**agagtaggg**aactgccaggcatcaa**ATAAAGCGAAAGGCCATCCTGACGGATGGCCTTTTTGCGT**

TGGTGCAAAAAAATGCCGGATGCGACGCTGGCGCGTTTTATCCAGCTTACGCAGGCACGATAGGGGGCAGC

TTATTCCCCCACATACGCCAGATCCAGCAACGGATACGGTTTCCCCAAATCGTCCACCTCAGAGCGTCCCGTA

ACCTTAAAACCCACCTTCTTATAGAACCCAACCGCCTGCTCATTTTGCTCATTAACATTGGTTGTCAGTTCCGG

GGCCATTGAgagcgcatgcttcacca

Figure 21: DNA sequence of OLGECs4930 including its upstream and downstream region. The sequence of OLGECs4930 is written in capital letters and colored in blue. The three putative start codons are highlighted in green, and the stop codon is highlighted in red. The predicted SD sequences are highlighted in light blue. The predicted σ^{70} promoter is underlined.

2.12.2 Promoter activity

A σ^{70} promoter with an LDF-score of 2.52 is predicted 135 bp upstream of the first putative start codon using BPROM (Figure 21). The sequence 300 bp upstream of the first start codon was cloned into pProbe-NT (Table 8) and plasmids were transformed in *E. coli* Top10. Promoter activity of this sequence was detected under all conditions investigated (Figure 22). The highest fluorescence intensity was measured in 0.5×LB supplemented with 400 mM NaCl. Additionally, the conditions pH 5 and supplementation with malonic acid show a significantly increased fluorescence intensity. In contrast, at pH 8.2 and after supplementation with $CuCl_2$, the promoter activity is significantly decreased. Because determination of the TSS failed, it is not sure if the detected promoter activity was caused by the predicted σ^{70} promoter, or if the sequence contains other promoters.

Table 8: Oligonucleotides used for cloning of pProbe-OLGECs4930.

name	sequence	purpose
OLGECs4930-300F-*Sal*I	ttcgGTCGACaagacgacgacgttgata	cloning pProbe-OLGECs4930
OLGECs4930-18R-*EcoR*I	attgGAATTCttgatgcctggcagttcc	cloning pProbe-OLGECs4930

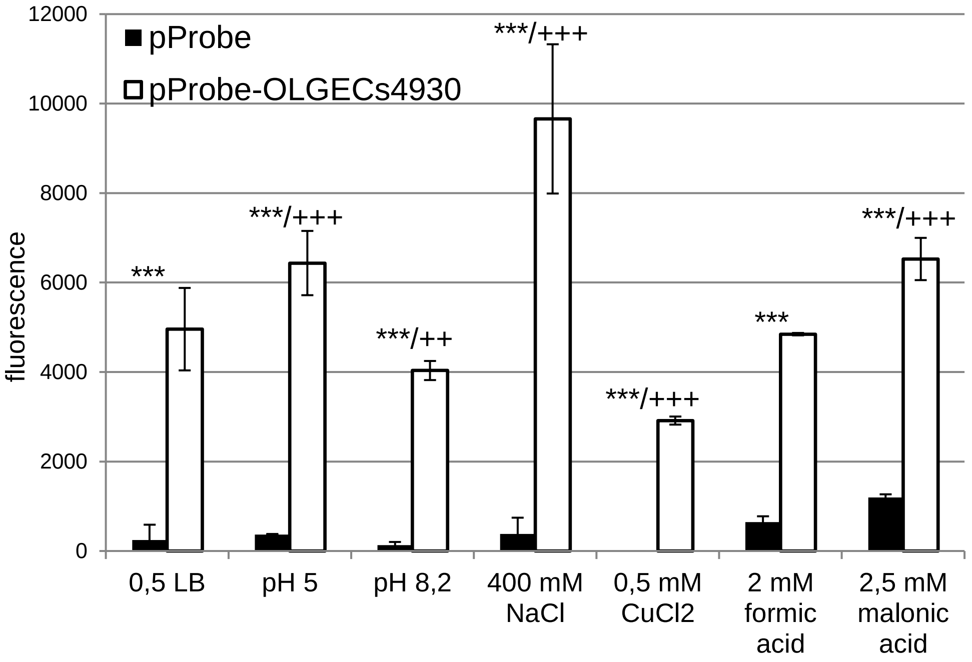


Figure 22: Promoter activity of the sequence upstream of the first putative start codon of OLGECs4930. *E. coli* Top10 carrying pProbe plasmids was incubated in 0.5×LB with different supplementations until an OD_{600} of 0.5 was reached. Then, the cells were harvested, washed with PBS and adjusted to an OD_{600} of 0.6. The fluorescence intensity was measured in quadruplicates in black 96-well microtiter plates (Wallac Victor^3, excitation 485 nm, emission 535 nm, measuring time 1 s). The fluorescence of *E. coli* Top10 was subtracted as background. The experiment was performed in triplicate. Significant changes between the empty plasmid and pProbe-OLGECs4930 plasmids were calculated using the Student's t-test and marked with asterisks (*** $p<0.001$). Significant changes between 0.5×LB and different stress conditions were calculated by the Student's t-test and marked with pluses (++ $p<0.01$; +++ $p<0.001$).

2.12.3 Cloning of a strand-specific knock-out mutant

To obtain evidence for a possible function of OLGECs4930, the strand-specific knock-out mutant ΔOLGECs4930 was cloned using the genome editing method of Kim et al.

(2014) (Table 9). The 44th codon of OLGECs4930 (counted from the first putative start codon) was mutated into a premature stop codon by introducing two point mutations. These point mutations do not influence the AA sequence of the annotated gene ECs4930. Competitive growth experiments with equal ratios of EHEC wild type and ΔOLGECs4930 mutant were performed to search for a phenotype. Overall, 17 different conditions were tested aerobically, and seven of them were tested anaerobically as well. However, for none of them the ratio of wild type to mutant changed (data not shown). Therefore, the function of OLGECs4930, if there is any, remains unknown. An option for future studies would be to determine the metabolome of wild type and mutant, maybe at different conditions (Fellner et al., 2014; Fellner et al., 2015).

Table 9: Oligonucleotides used for cloning of the strand-specific knock-out mutant ΔOLGECs4930.

name	sequence	purpose
HA3OLGECs4930-22F	aggcgtatcacgaggccctttagggaactgccaggcat	amplification 3’ mutation fragment
SM5OLGECs4930mut+159R	accgctgccactcttgagatttggggaaaccgtatccgttgctggatTtAgcg	amplification 3’ mutation fragment
SM3OLGECs4930mut+103F	gcaaggaggtgcataagggggcagcttattccccacatacgcTaAatc	amplification 5’ mutation fragment
HA5OLGECs4930+282R	ctcacatgttctttcctgcggtgaagcatgcgctctca	amplification 5’ mutation fragment
OLGECs4930+30F	cctttttgcgttggtgca	sequencing
OLGECs4930+232R	gcaaaatgagcaggcggt	sequencing

C-terminal OLGECs4930-EGFP fusion proteins were cloned for every putative start codon (compare to 2.11.3). However, none of the start codons used led to significantly higher fluorescence intensity compared to the uninduced cultures (data not shown). But the RIBOseq data clearly shows translation of OLGECs4930 (Table 7A). An explanation for the failure to detect a fusion protein might be that OLGECs4930 is predicted to localize to the inner membrane: the C-terminus of the putative protein might face the periplasm, in which EGFP does not fold. A different method showing protein expression would be to fuse a *myc/hisC* tag to OLGECs4930 and to perform a Western blot directed against the tag. Localization to the periplasm could be shown by an alkaline phosphatase PhoA assay (Manoil, 1991), in which a C-terminal OLGECs4930-PhoA fusion protein is cloned. PhoA is only active in the acidic environment of the periplasm, dephosphorylating target molecules. In the assay, p-nitrophenyl phosphate is added as

PhoA target. Dephosphorylation of the target leads to a measurable change in adsorption.

Concluding, OLGECs4930 is clearly transcribed (Table 7A). Moreover, the region upstream of the first putative start codon contains an active promoter (Figure 22). Maybe OLGECs4930 encodes a novel ncRNA instead of a novel protein, because the expression of an OLGECs4930-EGFP fusion protein failed for every putative start codon, and the mutant ΔOLGECs4930 did not show a phenotype in competitive growth experiments, which is expected in case of OLGECs4930 as a ncRNA, since functionality of a ncRNA will likely not be abolished by only two point mutations. However, the presence of annotated homologous proteins in other bacteria and the coverage of OLGECs4930 with RIBOseq reads (Figure 20) argues against a ncRNA. Future work is necessary to confirm that OLGECs4930 indeed encodes a protein and to unravel the function of this putative protein.

2.13 A novel operon antiparallel overlapping to ECs0535

2.13.1 Transcription and translation of a tri-cistronic operon

Initially, RNAseq and RIBOseq data in LB at 37°C pointed towards the expression of a 165 bp ORF antiparallel overlapping in reading frame -3 to the annotated ligase ECs0535. Visual inspection in the Artemis genome browser revealed convincing additional translation signals upstream of this ORF (Figure 23). Indeed, the upstream part contains two additional small ORFs also overlapping in reading frame -3 to ECs0535: the first ORF of the putative operon has a length of 63 bp and the second ORF of 162 bp. Most likely, these three ORFs represent the first described antiparallel overlapping operon (OLGECs0535 I to III). All ORFs are transcribed and translated at optimal conditions, whereas under BHI stress, translation is almost turned off (Table 10ABC). The annotated gene ECs0535 is only weakly translated at all conditions, also showing the lowest translation at BHI stress (Figure 23, Table 10D).

Table 10: Expression of OLGECs0535 I (A), OLGECs0535 II (B), OLGECs0535 III (C), and the annotated gene ECs0535 (D).

A

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	264.5	421	1.59	1
BHI, 37°C	594	102	0.17	1
BHI + 4% NaCl, 14°C	541.5	13.5	0.02	0.88

B

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	141.6	71.4	0.51	0.88
BHI, 37°C	314.7	20.9	0.07	0.75
BHI + 4% NaCl, 14°C	250.8	5.1	0.02	0.33

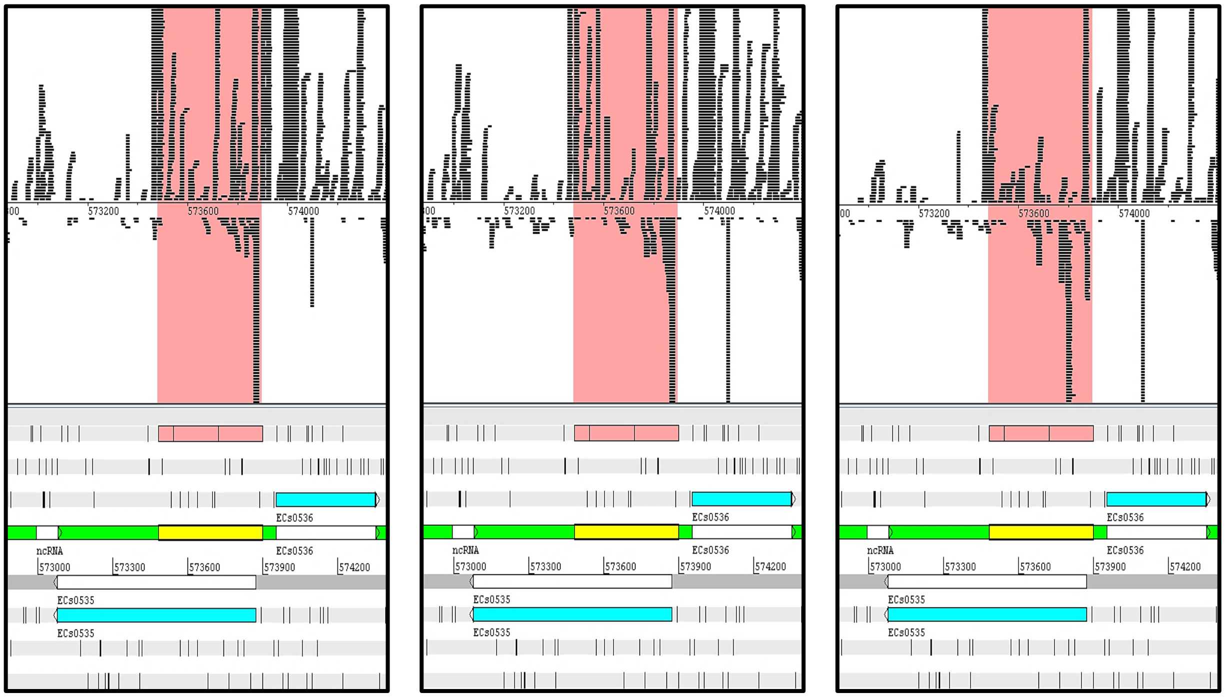
C

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	155.6	304.5	1.98	0.89
BHI, 37°C	359.1	66.5	0.20	0.74
BHI + 4% NaCl, 14°C	260.6	17.3	0.07	0.29

D

condition	RPKM transcriptome*	RPKM translatome*	ribosomal coverage value*	ORF coverage*
LB, 37°C	24.7	33.1	1.32	0.36
BHI, 37°C	53.9	19.1	0.38	0.49
BHI + 4% NaCl, 14°C	74.3	4.4	0.06	0.54

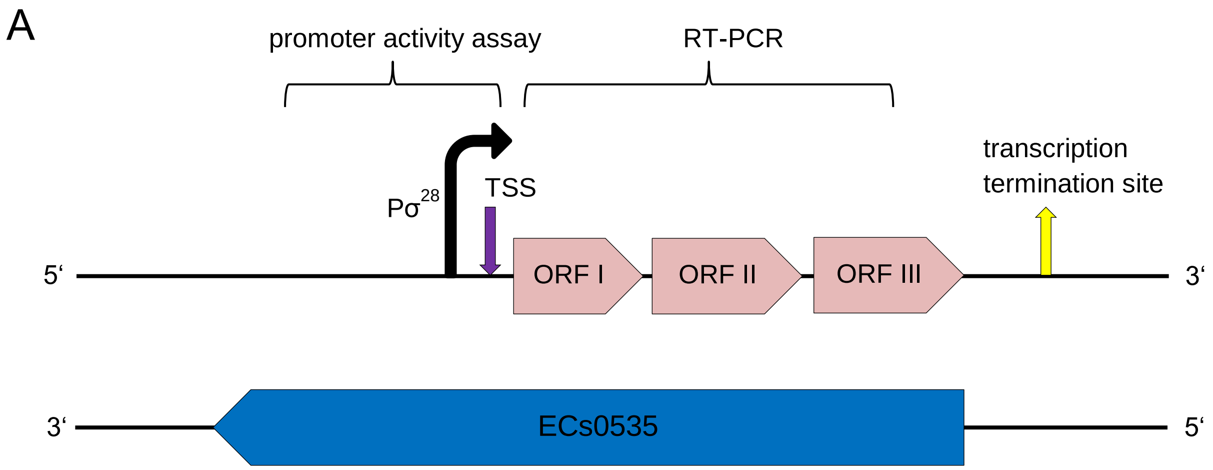
*Mean values of the two biological replicates are shown.



LB at 37°C BHI at 37°C BHI + 4% NaCl at 14°C

Figure 23: Translation of ECs0535 and OLGECs0535 I-III visualized in the Artemis genome browser. The RIBOseq reads of every tested growth condition were mapped strand-specifically to the EHEC Sakai genome. The annotated genes are colored in blue and the putative operon OLGECs0535 I-III is highlighted in pink.

A BLASTP search using the AA sequences of the putative proteins as a query was performed against the Refseq database to detect annotated homologs. The very short first ORF, OLGECs0535 I, obtained one hit to an uncharacterized protein of *Shigella sonnei* with an e-value of 1×10^{-11}, query coverage of 100%, and 95% identity. OLGECs0535 II has five annotated homologs in *Shigella* and *Escherichia.* The best hit is again an uncharacterized protein of *Shigella sonnei* with an e-value of 3×10^{-25}, 100% query coverage, and 92% identity. This 53 AA protein would use a rare CTG start codon (Figure 24B), but also a downstream GTG or TTG start codon would be possible. However, only the CTG start codon shows a predicted SD sequence 12 bp upstream with a ΔG° of -6.9. In case of the GTG start codon, this SD sequence would be 21 bp upstream, which is outside the optimal range (Ma et al., 2002). For OLGECs0535 III, the BLASTP search detected several annotated homologs in *Escherichia* and *Shigella.* The best hit is a hypothetical protein of *E. coli* strain 5.0588 with an e-value of 3×10^{-31}, 100% query coverage, and 93% identity. The 54 AA protein would start with a GTG start codon, but an ATG is also present downstream, which would encode a 50 AA protein (Figure 24). A SD sequence is predicted 6 bp upstream of the GTG start codon with a ΔG° of -5.0, this SD sequence would be 18 bp upstream of the ATG start codon. The software PredictProtein indicates that all three putative proteins are secreted. Furthermore, it predicts a disulfide bridge for OLGECs0535 III.



B

gtatcgtccagcagcgccagtcctttgtcagggagctggagcaacatggcaatctggttttccgccccttccagttcaatcacgggtttatgttgttg
cttcgccgcctgcaatagctgataatcgataccgtattctgcccgcagccccagtttTTGTGCCTGCGTCGCCTGAAGAACCATCGC
GATTTGCCACAGCGGTTGGGTAGAAAAGAGTGAgggagaaatgcccatctcCTGGCTAATGTGCTGCAGTTTTTGG
AGTTGCTCCTCGCTAATGCGCTCTTCCAGCGCCTCGCAGGCAGGCAAATTAGCAAAAGGCGTATCGCTGGTG
GAAACATCCGCCTCGACGATCAGCGCATCGGCGTTTTTGAGCTTTTTGAGCAAACGGGTGGGTAGgggagccat
ATCGTGGCTACCCATATGAATACTGCCAATCAGATGAAAATGGCGATTGCCGGGAAGAGAGATATCAATGG
CAGGCCAGGTGTAATGATTACCGCGCAGCGCGGCCCAAAGCGTTTTTACCCGGTACAACAGATCCATACGAC
CTCCCTTTGTGAAATATCATGCTAGcgcgcggggagggatggcgcaaatggagttggggacttgtttaatgtgtttgtatgattcagta
tgttcttgtatcgctatttacagggaagcaacagtt

Figure 24: Organization of the tri-cistronic operon OLGECs0535 I-III. **A** Schematic representation of the genomic region containing the mother gene ECs0535 and the overlapping operon OLGECs0535 I-III. The determined transcriptional start and stop site and the putative σ^{28} promoter are indicated. **B** DNA sequence of the putative overlapping operon OLGECs0535 I-III including its upstream and downstream region. The sequences of OLGECs0535 I-III are written in capital letters and colored in blue. The putative start codons are highlighted in green, and the stop codons are highlighted in red. The predicted SD sequences are wavelike underlined. The TSS determined by 5' RACE is highlighted in pink, and the transcriptional termination site determined by 3' RACE is highlighted in yellow. The predicted σ^{70} promoter is underlined. The detected σ^{28} -10 consensus motif is written in orange print.

2.13.2 Characterization of the polycistronic mRNA and of the promoter region

To investigate if the three ORFs are indeed transcribed as a polycistronic mRNA, RT-PCR was used. By means of oligonucleotides spanning all three genes, OLGECs0535 I-III, an RT-PCR product of the expected size was amplified (Abellan-Schneyder, 2017). Thus, polycistronic mRNA was confirmed. Next, the precise start and stop of transcription were determined (Table 11). The TSS is located 24 bp upstream of the OLGECs0535 I start codon, and transcription terminates 91 bp downstream from the stop codon of OLGECs0535 III (Figure 24). Even though BPROM predicts a σ^{70} promoter 101 bp upstream of OLGECs0535 I, this is probably not the correct promotor, because it is located too far upstream of the TSS. The sequence upstream of the TSS was investigated manually for the consensus motifs of alternative σ-factors. A sequence with high similarity to the -10 element of a σ^{28} promoter (Yu et al., 2006) was detected 9 bp upstream of the TSS (Figure 24). Only one nucleotide deviates from the consensus motif GCCGATAA. Nevertheless, a sequence similar to the -35 element consensus motif is not present. Anyway, Abellan-Schneyder (2017) tested the promoter activity of the sequence 300 bp upstream of the OLGECs0535 I start codon. This sequence contains an active promoter, causing highest fluorescence intensity, when measured in 0.5×LB supplemented with NaCl.

Table 11: Oligonucleotides used for the characterization of the OLGECs0535 I-III promoter region.

name	sequence	purpose
OLGECs0535-10F	cgtcgcctgaagaaccat	3' RACE, reverse transcription
OLGECs0535+4F	gctacccatatgaatactgccaatca	3' RACE, 1st PCR
OLGECs0535+37F	atggcgattgccgggaagag	3' RACE, 2nd PCR
OLGECs0535+147R	agcatgatatttcacaaagg	5' RACE, reverse transcription
OLGECs0535+117R	cgtatggatctgttgtaccgggtaa	5' RACE, 1st PCR
OLGECs0535+66R	atcattacacctggcctgccattg	5' RACE, 2nd PCR
OLGECs0535-387F-*Sal*I	ctatGTCGACggcgcgtaatttatcccgcc	cloning pProbe-OLGECs0535
OLGECs0535-107R-*Eco*RI	ccagGAATTCggtatcgattatcagctatt	cloning pProbe-OLGECs0535

2.13.3 EGFP-fusion protein expression

For OLGECs0535 II and OLGECs0535 III C-terminal EGFP-fusion proteins were cloned (Table 12). OLGECs0535 I was not investigated because of its short length, making false positive results very likely. Plasmids for all putative start codons of the two genes were tested. Expression of an EGFP-fusion protein could be determined for both OLGs and irrespective of the putative start codon (Figure 25). The putative CTG and GTG start codons of OLGECs0535 II caused a lesser increase of fluorescence intensity after induction than the putative TTG start codon. For the two putative start codons of OLGECs0535 III, the increases of fluorescence intensity are similar. Concluding, this experiment showed that OLGECs0535 II and III fusion proteins can be expressed, but the correct start codons are still unknown.

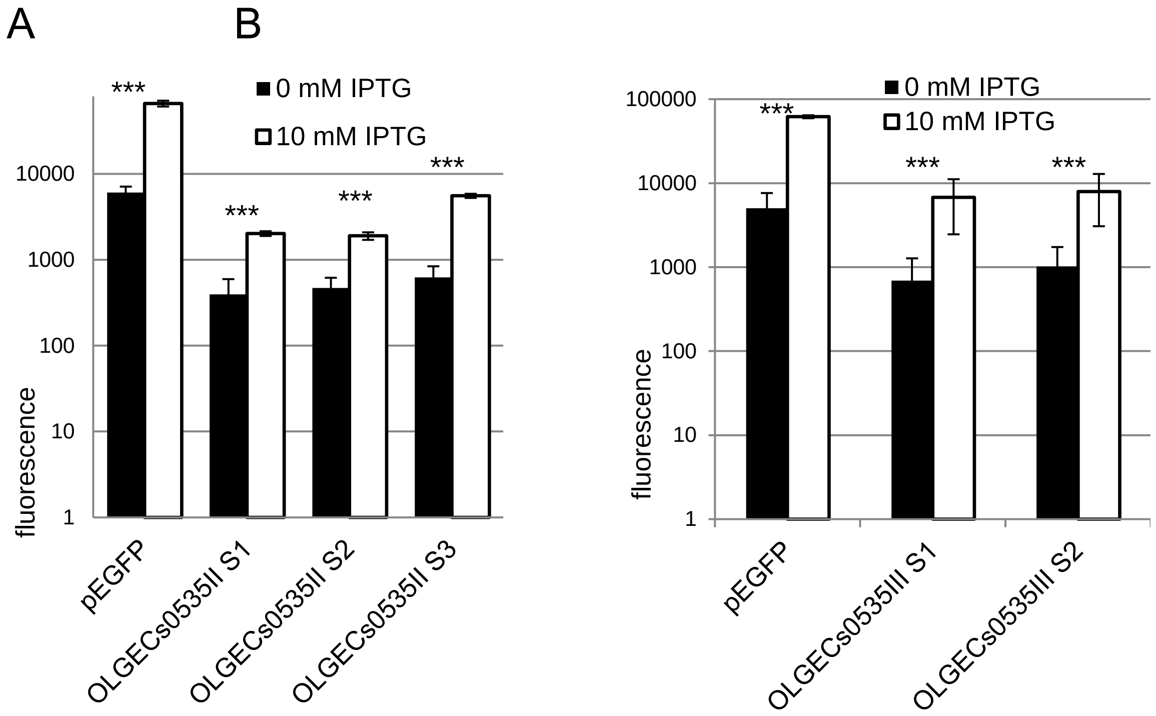


Figure 25: Expression of EGFP-OLGECs0535 II and EGFP-OLGECs0535 III fusion proteins. *E. coli* Top10 carrying pEGFP plasmids was incubated in 0.5×LB until an OD_{600} of 0.5 was reached. Then, one culture each was induced with 10 mM IPTG and incubation was continued for 1 h. The cells were harvested, washed with PBS and adjusted to an OD_{600} of 0.6. The fluorescence intensity was measured in quadruplicates in black 96-well microtiter plates (Wallac Victor^3, excitation 485 nm, emission 535 nm, measuring time 1 s). The fluorescence of *E. coli* Top10 was subtracted as background. The experiment was performed in triplicate. Significant changes between induced and uninduced cultures were calculated using the Student's t-test (*** p<0.001). **A** Expression of EGFP-OLGECs0535 II. **B** Expression of EGFP-OLGECs0535 III.

Table 12: Oligonucleotides used for cloning of pEGFP-OLGECs0535 II and pEGFP-OLGECs0535 III.

name	sequence	purpose
OLGECs0535II+1F-*BamH*I	aatcggatcccctggctaatgtgctgcagtt	cloning pEGFP-OLGECs0535II S1
OLGECs0535II+10F-*BamH*I	aatcggatcccgtgctgcagtttttggagtt	cloning pEGFP-OLGECs0535II S2
OLGECs0535II+28F-*BamH*I	aatcggatcccttgctcctcgctaatgcgct	cloning pEGFP-OLGECs0535II S3
OLGECs0535II+140R-*Nco*I	gtagccatggccccacccgtttgctcaaaaa	cloning pEGFP-OLGECs0535II
OLGECs0535III+1F-*Pst*I	tgatCTGCAGgatcgtggctacccatatgaa	cloning pEGFP-OLGECs0535III S1
OLGECs0535III+16F-*Pst*I	ggtcCTGCAGgatgaatactgccaatcagat	cloning pEGFP-OLGECs0535III S2
OLGECs0535III+146R-*Nco*I	ctagCCATGGcgcatgatatttcacaaaggg	cloning pEGFP-OLGECs0535III

To analyze possible gene functions, strand-specific knock-out mutants for each ORF of the overlapping operon were undertaken. However, despite applying two different cloning strategies (Fellner, 2015; Kim et al., 2014), no mutant could be created (Abellan-Schneyder, 2017). Therefore, the functions of this overlapping operon are still unknown.

In summary, the ORFs OLGECs0535 I-III represent protein-coding genes and build the first operon, which is completely embedded antiparallel into an annotated gene. RIBOseq clearly shows that all three genes are translated (Figure 23) and expression of a protein was confirmed for OLGECs0535 II and III (Figure 25). Additionally, all novel OLGs have annotated homologs. Transcription of the overlapping operon is putatively driven by a σ^{28} promoter, and the TSS is located 24 bp upstream of the start codon of OLGECs0535 I (Figure 24). Follow-up studies might determine the functions of this overlapping operon. In addition, phylostratigraphic analyses of the mother gene ECs0535 and the three OLGs could be performed to determine gene ages, and to investigate if the whole operon originated at once, or if the novel genes appeared one after the other.

2.14 Overlapping gene OLGZ5561 of EHEC EDL933

2.14.1 Summary of previous results concerning OLGZ5561

Fellner (2015) worked on the functional characterization of OLG candidates in EHEC strain EDL933. One candidate, representing the longest non-annotated overlapping ORF of EHEC EDL933, has a size of 2463 bp (corresponding to an 820 AA protein). It overlaps antiparallel in reading frame -2 to the annotated gene Z5561, which encodes the β-subunit of the DNA-directed RNA polymerase. However, OLGZ5561 shows only very weak transcription signals in the RNAseq data of Landstorfer et al. (2014), whereas the mother gene is strongly transcribed (Figure 26). In LB medium, only 10 reads map to OLGZ5561 (corresponding to an RPKM of 0.2). Transcription at the other conditions is even lower (0 to 6 reads). A non-annotated homolog of OLGZ5561 is also present in EHEC Sakai, where it overlaps antiparallel to ECs4911. Accordingly, in EHEC Sakai the ORF is weakly transcribed and translated with RPKM values below 1 for every growth condition tested.

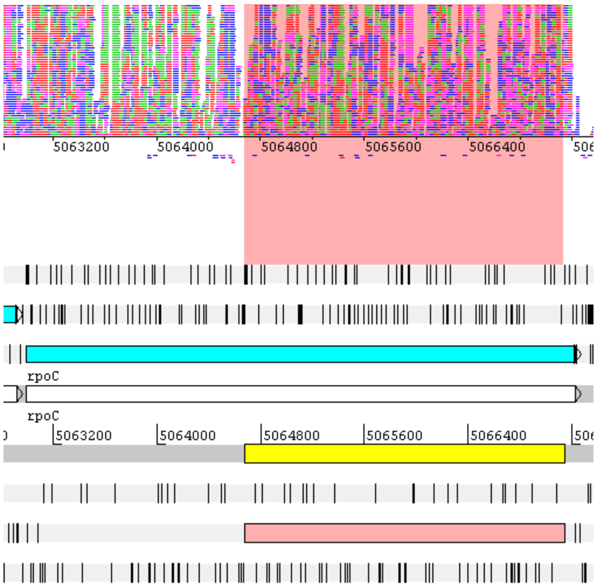


Figure 26: Transcription of OLGZ5561 and *rpoC* visualized in the Artemis genome browser. The RNAseq reads of Landstorfer et al. (2014) were strand-specifically mapped to the EHEC EDL933 genome. The annotated gene *rpoC* (Z5561) is colored in blue. OLGZ5561 is highlighted in pink.

Fellner (2015) cloned the translationally arrested mutant ΔOLGZ5561 by changing the 60th codon of OLGZ5561 into a premature stop codon (Figure 27). Then, a competitive growth experiment was performed: EHEC ΔOLGZ5561 was mixed in equal ratio with translationally arrested mutants of other OLGs and incubated at different conditions. The proportion of the mutants after competitive growth was determined using NGS. The proportion of EHEC ΔOLGZ5561 strongly increased in 0.5xLB supplemented with 78 µM erythromycin, and smaller increases were observed after supplementation with 0.2 M $MgCl_2$, 16 mM sodium orthovanadate, and 0.32 mM menadione. On the other hand, the proportion of EHEC ΔOLGZ5561 clearly decreased in medium supplemented with 5 mM malonic acid. Furthermore, the metabolome of EHEC EDL933 wild type and ΔOLGZ5561 was determined after growth in LB. Phenylalanine and tyrosine appeared to be reduced in the mutant, whereas citric and heptadecanonic acid appeared to be increased. However, these changes were not significant. In addition to the above experiments, Fellner (2015) investigated the promoter activity of the sequence 300 bp upstream of the start codon. No significant differences between the empty pProbe-NT plasmid and the plasmid with the cloned sequence were detected.

```
CGCAAAGATTGGAAAGATAAAGCGGGATTACTCGTTATCAGAACCGCCCAGACCTGCGTTCAGCAGTTCTGCCAGGCTGGCAG
ATGCGTCTTCTGCAGTCACCTGCGGTGCAGCCGGAGCTTCACCCGCAGCACGGCGACGCATACGATCCTGGTGGTACGCGTAA
CCGGTACCTGCCGGGATCAGACGACCCACGATAACGTTCTCTTTCAGGCCGCGCAGTTCGTCGCGTTTGCCCGCAACGGCTGC
TTCGGTCAGCACGCGAGTGGTCTCCTGGAACGATGCCGCGGAGATGAAGGACTCGGTTGCCAGAGACGCTTTGGTGATACCCA
GCAGATCGCGGGAGTAAGTTGCACCCACTTTGCCGTTCGCTTCCAGTTCGCGGTTTGCGATCTTGACGCGAGAGTATTCAACC
TGTTCGCCTTCCAGGAAGTCGGAGCTGCCCGCGTTAACGATGGTAGCTTTACGCAGCATCTGACGAACGATAACTTCGATGTG
TTTATCGTTAATCTTAACGCCCTGCAGACGGTATACGTCCTGTACTTCGTTAACGATGTAACGAGTCACAGCATGAACACCAC
GCAGACGCAGAATGTCGTGCGGCGCTTCCGGACCGTCGGAAATTACGTCACCACGTTCTACACGTTCACCTTCGAACACGTTG
AGCTGACGCCATTTCGGAATCATCTCTTCGTACGGATCGCTACCGTCTACCGGGGTGATAACCAGACGACGTTTACCTTTGGT
TTCTTTACCGAAGGAAACGATACCGCTGATTTCAGCCAGGATTGCCGGCTCTTTCGGACGACGTGCTTCGAACAGGTCCGCAA
CGCGTGGCAGACCACCGGTGATGTCCTTGGTACCGCCGGATTCCTGCGGAATACGCGCCAGGGTGTCACCAGAGCTGATCTGT
ACGCCATCTTCCAGCTGAACAATCGCTTTACCCGGCAGGAAGTACTGCGCAGGCATATCGGTACCTGGGATCAGAACGTCGTT
ACCCTGAGCATCAACGATTTTCAGTGCCGGACGCAGATCTTTACCACCTGCGGTACGTTCTGCGGAATCCAGAACCACCAGCG
AAGACAGACCGGTCAGTTCGTCGGTCTGACGAGTAATGGTCTGGCCGTCGATCATGTCAGTAAAGCGTACAAAACCGCTTACT
TCGGTGATAACCGGCATGGTGTGCGGGTCCCAGTTTGCAACGGTTTCGCCGCCAGCAACCTGTTCGCCATCGCCTTTTGCCAG
TACCGCACCGTAAGGTACTTTGTAGCTTTCTTTGGTACGACCGAATTCGTCGATCAGTTTCAGTTCGGTGTTACGGGAAGTGA
TAACCAGTTTACCGCTGGAGTTCACAACCGACTTCACGTTGCTGAGCTTGATGCTACCTTTGTTTTTCACCTGGATGCTGGAT
TCAGCAGCCGCACGAGATGCCGCACCACCGATGTGGAACGTACGCATGGTCAGCTGTGTACCCGGTTCACCGATGGACTGTGC
CGCGATAACACCGATTGCTTCACCCTTGTTGATGATGTGGCCACGCGCCAGGTCACGACCGTAGCAGTGCGCACATACACCAA
AGTCGGTGTCACAAGATACAACAGAACGTACTTTAACCGCGTCGACAGAGTTCTCTTCCAGCAGGTCACACCACTGTTCGTGC
AGCAGCGTGTTGCGCGGAACGAGGATATCAGCAGTACCCGGCTTCAGAACGTCTTCAGCAGTTACACGACCCAATACGCGATC
GCGCAGCGGCTCTTTAACGTCACCACCCTCGATAACCGGAGTCATCATGATACCTTCATGGGTACCACAATCGTCTTCGGTAA
CCACCAGGTCCTGCGCCACGTCAACCAGACGACGAGTCAGGTAACCGGAGTTCGCAGTTTTCAGTGCGGTATCCGCCAGACCT
TTACGAGCACCGTGGGTGGAGATGAAGTACTGGAGTACGTTCAGACCTTCACGGAAGTTCGCGGTGATTGGCGTTTCGATGAT
GGAGCCATCCGGCTTCGCCATCAGACCACGCATACCAGCAAGCTGACGAATCTGTGCCGCAGAACCACGCGCACCGGAGTCGG
CCATCATGTAGATGCTGTTGAAGGAAACCTGCTTCTCTTCCTGACCGTCACGGTTAATAACGGTTTCAGTTTGCAGGTTATCC
ATCATCGCTTTGGATACACGATCGTTCGCCGCAGCCCAGATATCGATAACTTTGTTGTAGCGTTCGCCCGCAGTTACCAGACC
AGACTGGAACTGCTCCTGAATTTCAGCAACTTCTGCTTCTGCCTCGGAGATGATTTCGTGTTTCTTCTCCGGGATGACCATGT
CATCGATACCAACAGATGCACCAGAACGCGCTGCATATGCAAAGCCGGTGTACATGATCTGGTCCGCAAAAATAACGGTCGGT
TTCAGACCGAGAATGCGGTAGCAGGTGTTCAGCATTTTGGAGATTGCTTTTTTACCCAGCGCCTGGTTGACGATGGAGTAAGG
CAGACCTTTCGGTACAATCATCCACAGAATGGCACGGCCAACAGTCGTGTCTTTCAGGCTGGTTTTCGCTACTAATTCACCGT
TAGCATCTTTTTCATACTCGGTGATACGCACTTTAAC
```

Figure 27: DNA sequence of OLGZ5561 including its upstream and downstream region. The sequence of OLGZ5561 is written in blue. The putative start codons are highlighted in green and the stop codon in red. The nucleotide, which was mutated (C → T) to create a premature stop codon, is underlined, and colored in red. The TSS determined using 5' RACE is highlighted in pink.

2.14.2 Predicted properties of OLGZ5561

The search for annotated homologs using BLASTP resulted in many hits. The best hit is an uncharacterized protein of *Shigella sonnei* with an e-value of 0, query coverage and identity of 97% and 99%, respectively. Hits with higher e-value, lower coverage, and lower identity were obtained in diverse bacterial phyla including Gram positive bacteria and even in amoeba. Mostly, these homologous proteins are annotated as 'hypothetical', only in the amoeba *Acytostelium leptosomum* the homolog is annotated as 'RNA polymerase I', and in *Salmonella* strain Enteritidis EC20120916 the homolog is annotated as 'lipase'. The putative protein does not contain conserved domains.

PredictProtein predicts that secondary structures of OLGZ5561 mainly consist of loops (45%), followed by β-barrels (40%), and 15% α-helices. The software did not find a signal peptide, disulfide bonds, or transmembrane helices, and OLGZ5561 should be located in the cytoplasm.

Further investigation of OLGZ5561 DNA sequence revealed two additional putative start codons upstream of the introduced point mutation (Figure 27). Besides the CTG start codon of the longest ORF, a GTG start codon is present leading to an 803 AA protein. An additional CTG start codon downstream would encode a 763 AA protein. None of the putative start codons is supported by a predicted upstream Shine-Dalgarno sequence. The TSS was determined using 5’ RACE 77 bp upstream of the first putative CTG start codon (Table 13).

Table 13: Oligonucleotides used for 5’ RACE.

name	sequence	purpose
OLGZ5561+378R	ACATCGAAGTTATCGTTC	5’ RACE, RT-PCR
OLGZ5561+329R	ATCGTTAACGCGGGCAGCTCCGACTT	5’ RACE, 1st PCR
OLOZ5561+283R	TACTCTCGCGTCAAGATCGCAAAC	5’ RACE, 2nd PCR

2.14.3 Competitive growth experiments

In order to confirm the results of the global competitive growth experiment using multiple translationally arrested EHEC strains at once (Fellner, 2015), single competitive growth assays were conducted with EHEC EDL933 wild type against ΔOLGZ5561. Overnight cultures of the two strains were mixed in equal ratio and incubated for 18 h under a number of different conditions. After competitive growth, the ratio between wild type and mutant changed significantly (Figure 28A). Interestingly, there are conditions, in which the mutant shows a significant growth advantage and others, in which it shows a significant growth disadvantage. In plain LB, the wild type shows an advantage, which is more pronounced after supplementation with malonic acid. On the other hand, the mutant has a very strong advantage in LB supplemented with erythromycin; the wild type almost disappeared in these cultures. A lesser advantage of the mutant was detected in menadione and $MgCl_2$ supplemented medium. These results are in good agreement with the competitive growth experiments conducted by Fellner (2015). Next, it was investigated if the phenotype still occurs, when the inoculation ratio is changed from 1:1 to 5:95. In case of an inoculation ratio of 95% mutant to 5% wild type, the wild type was not able to outgrow the mutant in medium containing malonic acid anymore, but at an inoculation ratio of 95% wild type and 5% mutant, the erythromycin phenotype still occured comparable to the inoculation ratio of 1:1 (Figure 28B).

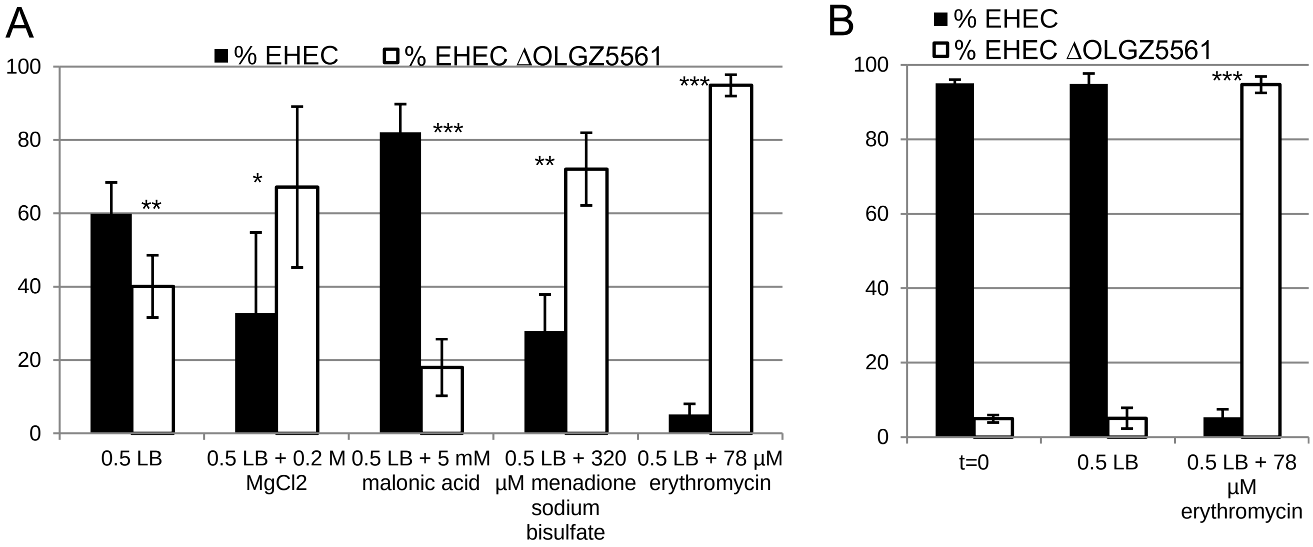


Figure 28: Competitive growth of EHEC wild type against ΔOLGZ5561. Overnight cultures of EHEC wild type and ΔOLGZ5561 were mixed and incubated at different conditions. After 18 h, DNA was isolated and the genomic region containing the introduced point mutation was amplified by PCR. The ratio of wild type to mutant was determined by comparing peak heights of Sanger sequencing. **A** Competitive growth with an inoculation ratio of 50% wild type to 50% ΔOLGZ5561. The experiment was performed in seven biological replicates. Significant changes between wild type and mutant were calculated using the Student's t-test and marked with asterisks (* p<0.05, ** p<0.01, *** p<0.001). **B** Competitive growth with an inoculation ratio of 95% wild type to 5% ΔOLGZ5561. The experiment was performed in triplicate. Significant changes between inoculation (t=0) and after competitive growth were calculated using the Student's t-test (*** p<0.001).

2.14.4 Complementation of the malonic acid and erythromycin phenotype

The intact OLGZ5561 ORF was cloned into an arabinose inducible plasmid (Table 14). Plasmids for all putative start codons were cloned and transformed into EHEC EDL933 ΔOLGZ5561. As negative control, plasmids with the mutated sequence ΔOLGZ5561 were produced as well. Competitive growth experiments with EHEC ΔOLGZ5561 + pBAD-OLGZ5561 (complementation) against EHEC ΔOLGZ5561 + pBAD-ΔOLGZ5561 (mutant control) were performed to test if the malonic acid and erythromycin phenotype can be complemented *in trans*. A complementation would prove that the detected phenotype was indeed caused by translational arrest of OLGZ5561 and not by a secondary mutation, which occurred somewhere else in the EHEC genome. Expression of OLGZ5561 was induced with increasing arabinose concentrations. Using the first two putative start codons, complementation failed, the ratio between wild type and mutant ORF stayed at the inoculation ratio (data not shown). Only for start codon 3, a partial complementation of the malonic acid phenotype was possible (Figure 29). Surprisingly, the complementation shows a strong growth advantage in LB even without induction, which indicates leakiness of the plasmid promoter. At an induction with 0.02% arabinose, the advantage decreases and at induction with 0.002% arabinose the mutant has a growth advantage. Supplementation with erythromycin leads to a similar distribution compared to LB medium only. As expected from the competitive growth (Figure 28A), increased induction of OLGZ5561 results in a growth advantage of the mutant, but the difference to plain LB is not significant. In contrast, supplementation with malonic acid showed a significant growth advantage of the complementation compared to LB at induction with 0.02% and 0.2% arabinose, somehow reflecting the phenotype of the wild type. One reason, why only a partial complementation was possible, might be that the OLGZ5561 concentrations obtained by induction of the plasmid do not correspond to the physiological concentrations. Induction with 0.2% arabinose causes overexpression of OLGZ5561, whereas the NGS results indicate that OLGZ5561 is only expressed at a very low level. Alternatively, regulatory sequences encoded upstream of OLGZ5561 are missing in this experimental setup. The complementation indicates that the third putative start codon might be the correct one, because for the other two start codons complementation failed.

Table 14: Oligonucleotides used for cloning of the pBAD complementation plasmids.

name	sequence	purpose
OLGZ5561+2456R-*Hind*III	tgccaagcttctaacggtgaattagtagcg	complementation S1
OLGZ5561+1F-*Nco*I	tgtgccatggcactgcggtgcagccggagctt	complementation S1
OLGZ5561+52F-*Nco*I	aattccatgggagtggtacgcgtaaccggtac	complementation S2
OLGZ5561+175F-*Nco*I	aattccatgaaagaacgatgccgcggagatga	complementation S3 wild type
ΔOLGZ5561+175F-*Nco*I	aattccatgaaagaatgatgccgcggagatga	complementation S3 mutant
OLGZ5561+2444R-*Hind*III	gggaagcttagttagtagcgaaaaccagcct	complementation S2/S3

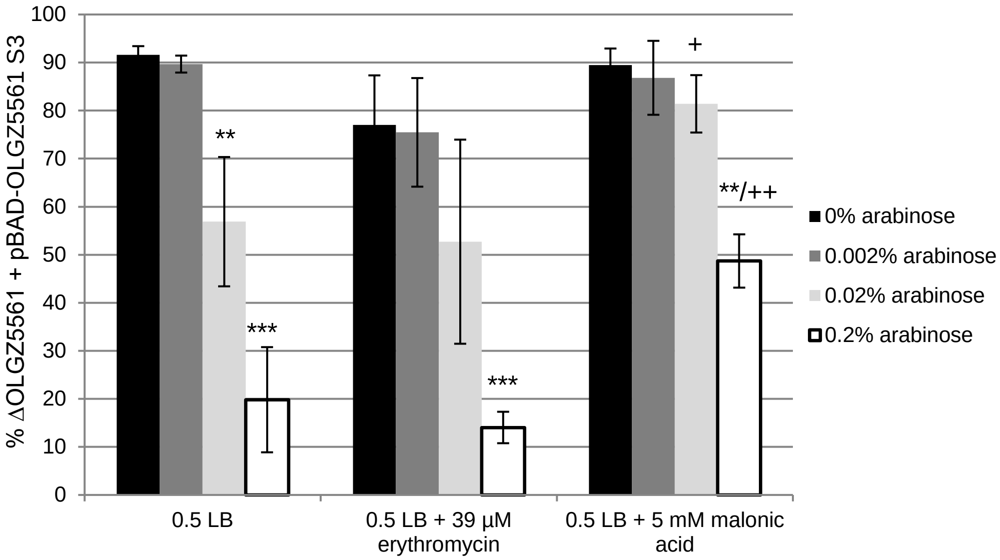


Figure 29: Complementation of EHEC ∆OLGZ5561 *in trans*. Competitive growth of EHEC ∆OLGZ5561 + pBAD-OLGZ5561 against EHEC ∆OLGZ5561 + pBAD-∆OLGZ5561 at different conditions was performed. One culture each was induced with 0.002%, 0.02%, or 0.2% arabinose (one culture was not induced). After 18 h, the plasmids were isolated and the region containing the introduced point mutation was amplified by PCR. The ratio of intact OLGZ5561 to ∆OLGZ5561 was determined. Depicted is the proportion of EHEC ∆OLGZ5561 + pBAD-OLGZ5561 in percent. The experiment was performed in triplicate. Significant changes between induced and uninduced cultures were calculated using the Student's t-test and marked with asterisks (** p<0.01, *** p<0.001). Significant changes between 0.5xLB and the stress conditions were calculated using the Student's t-test and marked with pluses (+ p<0.05, ++ p<0.01).

2.14.5 qRT-PCR

Because RNAseq is not convincingly showing that OLGZ5561 can be transcribed (Figure 26), qRT-PCR at different conditions was performed. Total RNA of EHEC EDL933 wild type and ∆OLGZ5561 was isolated and reverse transcribed into cDNA. The cDNA was used as template in a qPCR with SYBR green. Relative quantification was performed with the ∆∆Ct method (Pfaffl, 2001) using 16S rRNA for normalization (Table 15). Transcription of ∆OLGZ5561 is increased 3-fold in LB medium compared to the wild type ORF (Figure 30). If the culture is supplemented with erythromycin, transcription of OLGZ5561 is increased 7.5-fold compared to plain LB and the mutant shows a similar transcription level as in plain LB. Supplementation with malonic acid results in a 2-fold reduced transcription of the mutant. The qRT-PCR proved that OLGZ5561 is transcribed, and that erythromycin induces transcription.

Table 15: Oligonucleotides used for qRT-PCR of OLGZ5561.

name	sequence	purpose
rrsHR	GGAGGTGATCCAACCGCAGG	qRT-PCR 16S rRNA
rrsHF	AATGTTGGGTTAAGTCCCGC	qRT-PCR 16S rRNA
Z5561+3923F	CGTCTCTGGCAACCGAGTCC	qRT-PCR OLGZ5561
Z5561+4073R	CGCGTAACCGGTACCTGCCG	qRT-PCR OLGZ5561

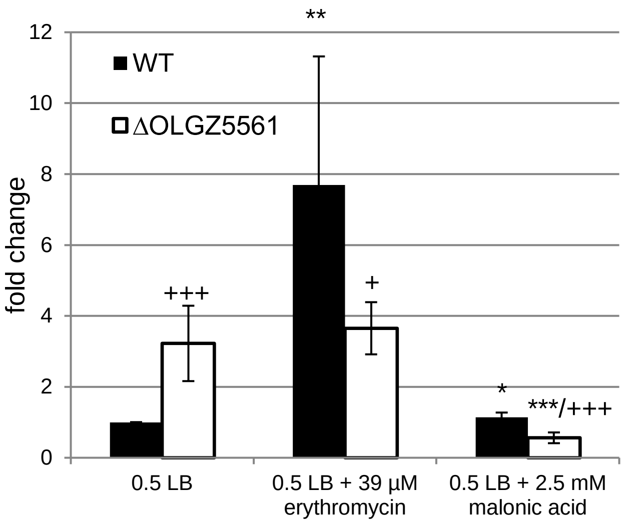


Figure 30: Transcription of OLGZ5561 and ΔOLGZ5561 determined by qRT-PCR. Fold changes relative to transcription of OLGZ5561 in 0.5xLB are depicted. The experiment was performed four times. Significant changes between the LB condition and the stress conditions were calculated with the Student's t-test and marked with asterisks (* p<0.05, ** p<0.01, *** p<0.001). Significant changes between wild type and mutant were calculated with the Student's t-test and marked with pluses (+ p<0.05, +++ p<0.001).

2.14.5 Failure to show protein expression

Translation of OLGZ5561 is not supported by the RIBOseq data either. Therefore, C-terminally EGFP-fusion proteins for every putative start codon were cloned. However, fusion protein expression was not detectable for any of them, because no difference in fluorescence intensity between induced and uninduced cultures could be measured (data not shown). Both, *E. coli* Top10 and EHEC EDL933 background were tested. The OLGZ5561-EGFP protein would have a size of above 1,000 AAs, and maybe the protein failed to fold properly. To avoid this, fusion proteins with a truncated OLGZ5561 sequence (220 AAs) were cloned, but still no protein expression was detectable. Transport to the periplasm could be one reason, since EGFP will not fold there, even though no signal sequence was predicted. Therefore, plasmids with a C-terminally *myc*-tagged OLGZ5561 were cloned, transformed into *E. coli* BL21 and expression of the protein was induced with arabinose. There was no detectable signal in a Western blot, using an antibody against the *myc*-tag for the OLGZ5561 protein using any of the putative start codons (data not shown). Thus, none of the experiments confirmed that an

OLGZ5561 protein exists. However, both EGFP-fusion protein expression and Western blot are plasmid-based and require OLGZ5561 overexpression. Speculating, OLGZ5561 was expressed in concentrations detrimental for the cell and rapid protein degradation occurred. Due to its overlapping nature, it is not possible to clone a C-terminal tag downstream of OLGZ5561 into the EHEC genome and search for the protein by Western blot (Baek et al., 2017), because the tag will destroy the annotated gene Z5561, which is an essential gene. Another possibility might be that OLGZ5561 represents a ncRNA instead of a protein-coding gene, but an ORF length of 2463 bp renders this option unlikely.

Although protein detection was not possible, OLGZ5561 is a very interesting OLG candidate due to its strong phenotype (Figure 28). Translational arrest of OLGZ5561 leads to a strong growth advantage in the presence of erythromycin. The macrolide antibiotic erythromycin targets the ribosome, and interrupts translation by context-specific inhibition of peptide bond formation (Kannan et al., 2014). But erythromycin is not able to inhibit the translation of every protein, because the protein synthesis is still at 6% compared to untreated cultures even at erythromycin concentrations 100-fold above the minimal inhibitory concentration (Kannan et al., 2012). The EHEC ΔOLGZ5561 ribosomes appear less sensitive against inhibition of translation. However, it is unlikely that a putative OLGZ5561 protein is part of the ribosome itself, because ribosomal proteins are very abundant and OLGZ5561 is only lowly transcribed and translated (Figure 26). A putative function of OLGZ5561 might be that it modifies a ribosomal protein. Interestingly, under acid stress, presence of intact OLGZ5561 represents an advantage for the cell (Figure 28). RT-PCR confirmed that both, erythromycin and malonic acid, influence the transcription of OLGZ5561 (Figure 30). Probably, repeating the metabolome experiments of Fellner (2015) in the presence of erythromycin and malonic acid, could shed light on the function of the putative OLGZ5561 protein. In rare cases, also synonymous mutations like the nucleotide change of Z5561 can influence bacterial fitness (Knoppel et al., 2016).

3. Conclusion and Outlook

In this work, I could demonstrate the great potential of combined high-throughput transcriptome and translatome sequencing. Especially, if conducted at different growth conditions, differentially regulated genes were determined, and transcriptional regulation distinguished from post-transcriptional regulation. This comparison made the importance of post-transcriptional regulation obvious. In the future, it would be interesting to investigate further growth conditions mimicking relevant habitats of EHEC, like the intestine of ruminants (anaerobiosis, 37°C, competition for nutrients), or plant surfaces (biofilm, cold temperature, limited nutrient availability). The genome of EHEC Sakai still contains many genes without functional annotation. RNAseq and RIBOseq reads mapping to hypothetical proteins would prove, on the one hand, that these genes are expressed, and do not represent annotation artefacts and, on the other hand, the expression condition(s) might indicate potential gene functions (Baric et al., 2016; Landstorfer et al., 2014).

3.1 Improvements required for accurate detection of novel genes using RIBOseq

In addition to already annotated and well-known genes, this study verified that combined RNAseq and RIBOseq convincingly detect abundant translation of non-annotated intergenic and antiparallel overlapping ORFs. These novel protein-coding genes were confirmed by bioinformatics characterization. Additional support for the protein-coding nature of these translated ORFs would be the discovery of a reading frame for single ORFs reflecting the codon-wise progression of the ribosome (Ingolia et al., 2009). However, a reading frame was only detectable in the sum signal (Hücker et al., 2017a), but at the single gene level, the resolution was too poor. In contrast, eukaryotic RIBOseq data allow reading frame detection of single genes (Smith et al., 2014). Presence of a reading frame was useful as a criterion to detect translated non-annotated genes (Calviello et al., 2016; Legendre et al., 2015; Malone et al., 2017). In the past, the insufficient resolution of prokaryotic RIBOseq data was either explained by a sequence specificity of the nuclease used (Gerashchenko and Gladyshev, 2017), by the application of chloramphenicol as a translational inhibitor (Marks et al., 2016), or by a greater conformational flexibility of the bacterial ribosome compared to the eukaryotic ribosome (O'Connor et al., 2013). This study confirms the third hypothesis, because the reading frame resolution did not improve using a mixture of five RNases to get rid of any sequence specificity a single enzyme might have, and RIBOseq data of EHEC Sakai without using chloramphenicol show no significant improvement in reading frame resolution (Abellan-Schneyder, 2017). Instead of trying to improve the composition of RIBOseq buffers or other protocol elements (Hsu et al., 2016), the most promising approach for future RIBOseq experiments is to use an additional enzyme, as in the protocol of Hwang and Buskirk (2017). Addition of the ribonuclease RelE during digestion of unprotected mRNA resulted in a reading frame with high resolution allowing frame determination of single genes, since RelE cuts with high precision after the second nucleotide for most codons. Detection of a reading frame for non-annotated ORFs covered by RIBOseq reads would be further evidence of their translation into proteins. Furthermore, same-strand overlapping ORFs could be studied, because when both, the annotated gene and the overlapping ORF, are expressed, a mixed signal for the frame is expected.

In the future, it would be very helpful to extend bioinformatics tools developed to detect non-annotated translated ORFs in eukaryotic RIBOseq data (Baranov and Michel, 2016; Fields et al., 2015; Legendre et al., 2015; Malone et al., 2017) towards prokaryotic RIBOseq data. In addition, already published RIBOseq data sets could be reinvestigated regarding non-annotated translated ORFs. Improved genome annotation algorithms are required, which do not categorical exclude ORFs below a certain size threshold, and ORFs overlapping to annotated genes (Boekhorst et al., 2011; Firth and Brown, 2005; Oheigeartaigh et al., 2014; Storz et al., 2014; Warren et al., 2010). Maybe, including experimental data like RNAseq, RIBOseq, and proteomics into the annotation pipeline will improve accuracy and precision for small ORF prediction. Finally, non-ATG start codons should be considered as well (Chu et al., 2015; Iwasaki and Ingolia, 2017; Nakahigashi et al., 2016).

3.2 High-throughput detection of proteins corresponding to the novel genes

It is desirable to prove with a complementary method that the translated ORFs indeed represent protein-coding genes. Direct detection of the encoded proteins would be definitive evidence of their existence. However, high-throughput determination of the proteome using tandem mass spectrometry is biased against small proteins. First, many small proteins get lost during protein purification. Second, small proteins are reported to frequently associate to the membrane (Kemp and Cymer, 2014), and the discovery of membrane proteins requires special protocols. Third, a protein is only considered to be present, when at least two unique tryptic peptides are identified (a criterion which is hard to meet for short proteins). Fourth, the tryptic peptides must be in a certain size range. Finally, even if two tryptic peptides of proper size could be obtained, the protein might be missed, when it is present in low abundance only (Slavoff et al., 2013), which is presumed for many translated non-annotated ORFs. Therefore, most small proteins cannot be identified by mass spectrometry. Neuhaus et al. (2016) verified only seven proteins of 72 translated intergenic ORFs using mass spectrometry. Additionally, classical proteomics measures the steady state levels of all proteins, whereas RIBOseq determines the translation at a certain time point. Because many proteins have a half-life of several hours (Maier et al., 2011), the two methods give different results. However, specialized mass spectrometry protocols were recently developed detecting newly synthesized proteins only: BONCAT uses azide-tagged AAs to distinguish already present from newly synthesized proteins (Dieterich et al., 2006). Further, pSILAC uses pulse labeling with stable isotopes (Liu et al., 2017), QuaNCAT combines tagging with non-canonical AAs and stable isotope labeling (Howden et al., 2013), and PUNCH-P applies labeling of nascent peptides with biotinylated puromycin (Zur et al., 2016). Finally, in N-terminal proteomics, only the N-terminal peptide needs to be detected, and the presence of two tryptic peptides is not necessary (Slavoff et al., 2013; Willems et al., 2017).

3.3 Functional characterization of novel genes

The functional characterization of novel proteins represents a major bottleneck and high-throughput phenotyping methods are required (Baric et al., 2016). In this work, three novel OLGs could be functionally characterized, and some results are available for four additional OLGs, but the functions of the other 374 translated antiparallel overlapping ORFs and 465 translated intergenic ORFs remain completely unknown. Usually, growth phenotypes are determined by measuring the growth of knock-out mutants at different conditions. Competitive growth experiments are more sensitive than comparing two growth curves of single strains. The Keio collection contains knock-outs of every annotated gene of *E. coli* K-12 (Baba et al., 2006). Using this collection, Nichols et al. (2011) found growth phenotypes for 49% of genes, testing 300 conditions on agar plates. However, for the study of non-annotated genes the knock-out mutants have to be cloned one by one, and especially the cloning of strand-specific translationally arrested mutants for OLGs is very time consuming, and sometimes not successful with the applied genome editing method (Kim et al., 2014). Maybe the cloning efficiency can be enhanced using CRISPR-Cas9. CRISPR stands for ‘clustered regularly interspaced short palindromic repeats’, and together with the endonuclease Cas9 it is part of the adaptive bacterial immunity against viral infections. This system was engineered for site-specific, scarless genome editing. A guide RNA is required, which base pairs to the target sequence in close proximity to a PAM motive (NGG), and recruits Cas9, which will introduce a double-strand break (Doudna and Charpentier, 2014). Garst et al. (2017) report a CRISPR-Cas9 method, which allows the mutation of 10,000 loci in parallel and using barcodes to track a phenotype back to the genotype. An alternative to the study of knock-out mutants is the investigation of overexpression phenotypes. Analogous to the Keio collection, the ASKA collection contains every annotated gene of *E. coli* K-12 cloned into an IPTG inducible plasmid (Kitagawa et al., 2005). Cloning of overexpression plasmids would be much easier than mutating the genome of EHEC. Additionally, random genome fragments could be overexpressed: Boyer et al. (2004) analyzed the toxicity of overexpression of random DNA fragments of yeast. Interestingly, 23% of the fragments showing toxicity matched intergenic regions or overlap to annotated genes. If knock-out and overexpression are not showing a phenotype, probably due to functional redundancy or wrong conditions tested, *in vitro* biochemical characterization (e.g., activity based metabolic profiling, determination of protein-protein interactions by mass spectrometry) of the protein would represent another option (Baric et al., 2016). In the future, the transcriptional start and stop sites of putative novel genes could be determined in high-throughput using the NGS methods TSSseq (Filiatrault et al., 2011) and Termseq (Dar et al., 2016), which would make single 5'/3' RACE experiments obsolete. Furthermore, the correct start codon is detectable in RIBOseq experiments after the addition of tetracycline (Nakahigashi et al., 2016).

3.4 Characterization of the novel genes at a single cell level

RNAseq and RIBOseq show transcription and translation level averaged over the bacterial population, but even isogenic populations show a great cell-to-cell heterogeneity (Wang and Roy, 2017). It would be interesting to study the expression of novel genes on single cell level as well. Single-cell RNAseq protocols for eukaryotes already exist (Bacher and Kendziorski, 2016; Zhu et al., 2017), but they are not useable in bacteria, because they require polyadenylated mRNA. Additionally, a single prokaryotic cell contains only 3-9 fg total RNA. Despite these difficulties, recent protocols facilitated single cell RNAseq of prokaryotes (Fu et al., 2016; Wang and Roy, 2017). However, transcription on single cell level can also be investigated by fluorescence microscopy. Buskila et al. (2014) analyzed the mRNA localization in bacteria, and found different patterns with mRNA preferentially localizing to the genome, to the cytoplasm, to the membrane, or to the cell poles. Taniguchi et al. (2010) detected the number of RNA molecules per bacterial cell of a certain mRNAs, showing high variations over time, because transcription occurs in bursts. Especially for the OLGs, it would be fascinating to view *in vivo* at which frequencies the mother gene and the OLG are transcribed and translated. New methods allow observing *in vivo* real-time translation of a single mRNA into a protein in eukaryotic cells (Wu et al., 2016). Even the translation initiation rate (every 30 s), elongation rate (10 amino acids/s), and the space between translating ribosomes could be determined (200-900 bp) (Morisaki et al., 2016).

3.5 Origin, evolution and fate of *de novo* genes

Origin and evolution of the novel genes with special emphasis on overlapping genes is another promising direction for future research. With the three overlapping gene pairs discovered and characterized in this study, the EHEC genome now contains five confirmed OLG pairs, and my NGS data indicate that there are many more. This contradicts the theory that non-trivial OLGs are an exception in bacterial genomes due to high constraints on sequence evolution (Veloso et al., 2005). Putatively, the novel genes have arisen *de novo* from non-coding DNA. If there is indeed a continuum from non-coding sequence to protein-coding as it was hypothesized by Carvunis et al. (2012), or if the genes originated at a certain time point in the history of life in a 'big bang' (Keese and Gibbs, 1992), is not sufficiently answered today. Presence of genes restricted to certain lineages or genes without any homologs argue against a big bang creation (Light et al., 2014). Phylostratigraphic analysis can be used to determine the gene age (Pavesi et al., 2013). However, recent publications criticize phylostratigraphy, because it underestimates the gene age (Moyers and Zhang, 2016), or the results should be confirmed by investigating synteny as well (Lu et al., 2017). Carvunis et al. (2012) postulate, based on their analysis of non-overlapping genes of different age in yeast, that *de novo* gene birth is as frequent as duplication for the creation of new genes. Abrusan (2013) further analyzed their data and detected that novel genes are quickly integrated into cellular networks, whereas there is assumedly a bottleneck from a structural point of view: conserved genes have lower aggregation propensity and contain less β-barrels, meaning that only a small fraction of the proto-genes becomes fixed, and the rest is lost again. All-in-all, the study supports the results of Carvunis et al. (2012). In contrast, Moyers and Zhang (2016) disagree with widespread *de novo* gene birth: they performed an *in silico* simulation of gene evolution with exclusion of *de novo* gene birth, and still found the same trends as Carvunis et al. (2012). Therefore, they concluded that the genes could have arisen by other mechanisms as well. Lu et al. (2017) also investigated *de novo* gene birth in yeast including overlapping ORFs. They could not confirm the majority of *de novo* genes of Carvunis et al. (2012) due to too low transcript levels or false age assignment, but they detected 8,871 *de novo* genes in *Saccharomyces sensu stricto*, of which 65% arose from transcript isoforms of ancient genes and 79% are overlapping to annotated genes. Anyway, in case of the OLGs, creation by duplication is not possible and they have to have originated *de novo* likely by overprinting (Delaye et al., 2008; Keese and Gibbs, 1992). It would be interesting to find pairs of OLGs, in which the two genes become decoupled and evolve independently later. Overlapping genes might be important for species-specific adaption to colonize a new niche.

4. References

Abellan-Schneyder, I.S., 2017. Analyse überlappender Gene in *E. coli* O157:H7 Sakai mittels RIBOseq. Master's Thesis, TU München, Freising.

Abrusan, G., 2013. Integration of new genes into cellular networks, and their structural maturation. Genetics 195, 1407-1417.

Aspden, J.L., Eyre-Walker, Y.C., Phillips, R.J., Amin, U., Mumtaz, M.A., Brocard, M., Couso, J.P., 2014. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. eLife 3, e03528.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., Mori, H., 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2, 2006 0008.

Bacher, R., Kendziorski, C., 2016. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol 17, 63.

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

Baggett, N.E., Zhang, Y., Gross, C.A., 2017. Global analysis of translation termination in *E. coli*. PLoS Genet 13, e1006676.

Balabanov, V.P., Kotova, V.Y., Kholodii, G.Y., Mindlin, S.Z., Zavilgelsky, G.B., 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.

Banerjee, S., Chalissery, J., Bandey, I., Sen, R., 2006. Rho-dependent transcription termination: more questions than answers. Journal of microbiology 44, 11-22.

Baranov, P.V., Michel, A.M., 2016. Illuminating translation with ribosome profiling spectra. Nat Methods 13, 123-124.

Baric, R.S., Crosson, S., Damania, B., Miller, S.I., Rubin, E.J., 2016. Next-Generation High-Throughput Functional Annotation of Microbial Genomes. mBio 7.

Barker, J., Humphrey, T.J., Brown, M.W., 1999. Survival of *Escherichia coli* O157 in a soil protozoan: implications for disease. FEMS Microbiol Lett 173, 291-295.

Bartholomaus, A., Fedyunin, I., Feist, P., Sin, C., Zhang, G., Valleriani, A., Ignatova, Z., 2016. Bacteria differently regulate mRNA abundance to specifically respond to various stresses. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences 374.

Battle, S.E., Brady, M.J., Vanaja, S.K., Leong, J.M., Hecht, G.A., 2014. Actin pedestal formation by enterohemorrhagic *Escherichia coli* enhances bacterial host cell attachment and concomitant type III translocation. Infect Immun 82, 3713-3722.

Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C., Giraldez, A.J., 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J 33, 981-993.

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

Betran, E., 2015. The "life histories" of genes. J Mol Evol 80, 186-188.

Binns, N., Masters, M., 2002. Expression of the *Escherichia coli pcnB* gene is translationally limited using an inefficient start codon: a second chromosomal example of translation initiated at AUU. Mol Microbiol 44, 1287-1298.

Bitard-Feildel, T., Callebaut, I., 2017. Exploring the dark foldable proteome by considering hydrophobic amino acids topology. Scientific reports 7, 41425.

Boekhorst, J., Wilson, G., Siezen, R.J., 2011. Searching in microbial genomes for encoded small proteins. Microbial biotechnology 4, 308-313.

Boldogköirid, Z., 2000. Coding in the noncoding DNA strand: A novel mechanism of gene evolution? J. Mol. Evol. 51, 600-606.

Boyer, J., Badis, G., Fairhead, C., Talla, E., Hantraye, F., Fabre, E., Fischer, G., Hennequin, C., Koszul, R., Lafontaine, I., 2004. Large-scale exploration of growth inhibition caused by overexpression of genomic fragments in *Saccharomyces cerevisiae*. Genome Biol. 5, R72.

Brylinski, M., 2013. Exploring the "dark matter" of a mammalian proteome by protein structure and function modeling. Proteome science 11, 47.

Burland, V., Shao, Y., Perna, N.T., Plunkett, G., Sofia, H.J., Blattner, F.R., 1998. The complete DNA sequence and analysis of the large virulence plasmid of *Escherichia coli* O157:H7. Nucleic Acids Res 26, 4196-4204.

Burrows, P.C., Severinov, K., Ishihama, A., Buck, M., Wigneshweraraj, S.R., 2003. Mapping sigma 54-RNA polymerase interactions at the -24 consensus promoter element. J Biol Chem 278, 29728-29743.

Buschmann, D., Haberberger, A., Kirchner, B., Spornraft, M., Riedmaier, I., Schelling, G., Pfaffl, M.W., 2016. Toward reliable biomarker signatures in the age of liquid biopsies - how to standardize the small RNA-Seq workflow. Nucleic Acids Res 44, 5995-6018.

Buskila, A.A., Kannaiah, S., Amster-Choder, O., 2014. RNA localization in bacteria. RNA Biol 11, 1051-1060.

Caliskan, N., Katunin, V.I., Belardinelli, R., Peske, F., Rodnina, M.V., 2014. Programmed -1 frameshifting by kinetic partitioning during impeded translocation. Cell 157, 1619-1631.

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., Ohler, U., 2016. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods 13, 165-170.

Carlevaro-Fita, J., Rahim, A., Guigo, R., Vardy, L.A., Johnson, R., 2016. Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. RNA 22, 867-882.

Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B., Brar, G.A., Weissman, J.S., Regev, A., Thierry-Mieg, N., Cusick, M.E., Vidal, M., 2012. Proto-genes and *de novo* gene birth. Nature 487, 370–374.

Chang, C.P., Chen, S.J., Lin, C.H., Wang, T.L., Wang, C.C., 2010. A single sequence context cannot satisfy all non-AUG initiator codons in yeast. BMC Microbiol 10, 188.

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

Chu, Q., Ma, J., Saghatelian, A., 2015. Identification and characterization of sORF-encoded polypeptides. Crit Rev Biochem Mol Biol 50, 134-141.

Croxen, M.A., Law, R.J., Scholz, R., Keeney, K.M., Wlodarska, M., Finlay, B.B., 2013. Recent advances in understanding enteric pathogenic *Escherichia coli*. Clin Microbiol Rev 26, 822-880.

Dana, A., Tuller, T., 2014. Mean of the typical decoding rates: a new translation efficiency index based on the analysis of ribosome profiling data. G3 5, 73-80.

Dar, D., Shamir, M., Mellin, J.R., Koutero, M., Stern-Ginossar, N., Cossart, P., Sorek, R., 2016. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. Science 352, aad9822.

de Groot, A., Roche, D., Fernandez, B., Ludanyi, M., Cruveiller, S., Pignol, D., Vallenet, D., Armengaud, J., Blanchard, L., 2014. RNA Sequencing and Proteogenomics Reveal the Importance of Leaderless mRNAs in the Radiation-tolerant Bacterium *Deinococcus deserti*. Genome biology and evolution.

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, A.L., Bratke, K.A., Powers, E.C., Salzberg, S.L., 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23, 673-679.

Diament, A., Tuller, T., 2016. Estimation of ribosome profiling performance and reproducibility at various levels of resolution. Biology direct 11, 24.

Dieterich, D.C., Link, A.J., Graumann, J., Tirrell, D.A., Schuman, E.M., 2006. Selective identification of newly synthesized proteins in mammalian cells using bioorthogonal noncanonical amino acid tagging (BONCAT). Proc Natl Acad Sci U S A 103, 9482-9487.

Dornenburg, J.E., Devita, A.M., Palumbo, M.J., Wade, J.T., 2010. Widespread Antisense Transcription in *Escherichia coli*. mBio 1.

Doudna, J.A., Charpentier, E., 2014. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science 346, 1258096.

Duffitt, A.D., Reber, R.T., Whipple, A., Chauret, C., 2011. Gene expression during survival of *Escherichia coli* O157:H7 in soil and water. Int J Microbiol 2011, 340506.

Eppinger, M., Cebula, T.A., 2015. Future perspectives, applications and challenges of genomic epidemiology studies for food-borne pathogens: A case study of Enterohemorrhagic *Escherichia coli* (EHEC) of the O157:H7 serotype. Gut microbes 6, 194-201.

Etienne-Mesmin, L., Chassaing, B., Sauvanet, P., Denizot, J., Blanquet-Diot, S., Darfeuille-Michaud, A., Pradel, N., Livrelli, V., 2011. Interactions with M cells and macrophages as key steps in the pathogenesis of enterohemorrhagic *Escherichia coli* infections. PloS one 6, e23594.

Fellner, L., 2015. Fuctional characterization of overlapping genes in the food-borne pathogen *Escherichia coli* O157:H7, Lehrstuhl für Mikrobielle Ökologie. PhD Thesis, TU München, Freising.

Fellner, L., Bechtel, N., Witting, M.A., 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 Microbiol. Lett. 350, 57-64.

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

Fields, A.P., Rodriguez, E.H., Jovanovic, M., Stern-Ginossar, N., Haas, B.J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S.A., Ingolia, N.T., Regev, A., Weissman, J.S., 2015. A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Mol Cell 60, 816-827.

Filiatrault, M.J., Stodghill, P.V., Myers, C.R., Bronstein, P.A., Butcher, B.G., Lam, H., Grills, G., Schweitzer, P., Wang, W., Schneider, D.J., Cartinhour, S.W., 2011. Genome-wide identification of transcriptional start sites in the plant pathogen *Pseudomonas syringae* pv. tomato str. DC3000. PLoS One 6, e29335.

Firth, A.E., Brown, C.M., 2005. Detecting overlapping coding sequences with pairwise alignments. Bioinformatics 21, 282-292.

Flaherty, B.L., Van Nieuwerburgh, F., Head, S.R., Golden, J.W., 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.

Fonseca, M.M., Harris, D.J., Posada, D., 2014. Origin and Length Distribution of Unidirectional Prokaryotic Overlapping Genes. G3: Genes| Genomes| Genetics 4, 19-27.

Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., Schumann, F., Schreiber, S., Platzer, M., Krawczak, M., Hampe, J., Brosch, M., 2012. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res 22, 2208-2218.

Fu, Y., Chen, H., Liu, L., Huang, Y., 2016. Single Cell Total RNA Sequencing through Isothermal Amplification in Picoliter-Droplet Emulsion. Analytical chemistry 88, 10795-10799.

Gao, X., Wan, J., Liu, B., Ma, M., Shen, B., Qian, S.B., 2015. Quantitative profiling of initiating ribosomes *in vivo*. Nat Methods 12, 147-153.

Garst, A.D., Bassalo, M.C., Pines, G., Lynch, S.A., Halweg-Edwards, A.L., Liu, R., Liang, L., Wang, Z., Zeitoun, R., Alexander, W.G., Gill, R.T., 2017. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. Nat Biotechnol 35, 48-55.

Gerashchenko, M.V., Gladyshev, V.N., 2014. Translation inhibitors cause abnormalities in ribosome profiling experiments. Nucleic Acids Res 42, e134.

Gerashchenko, M.V., Gladyshev, V.N., 2017. Ribonuclease selection for ribosome profiling. Nucleic Acids Res 45, e6.

Gilsdorf, A., 2016. Infektionsepidemiologisches Jahrbuch meldepflichtiger Krankheiten für 2015, in: Koch-Institut, R. (Ed.). Robert Koch-Institut, Berlin, p. 236.

Giuliodori, A.M., Di Pietro, F., Marzi, S., Masquida, B., Wagner, R., Romby, P., Gualerzi, C.O., Pon, C.L., 2010. The *cspA* mRNA is a thermosensor that modulates translation of the cold-shock protein CspA. Mol Cell 37, 21-33.

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

Guimaraes, J.C., Rocha, M., Arkin, A.P., 2014. Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. Nucleic Acids Res 42, 4791-4799.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.G., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M., Shinagawa, H., 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA research : an international journal for rapid publication of reports on genes and genomes 8, 11-22.

Haycocks, J.R., Grainger, D.C., 2016. Unusually Situated Binding Sites for Bacterial Transcription Factors Can Have Hidden Functionality. PloS one 11, e0157016.

He, S., Wurtzel, O., Singh, K., Froula, J.L., Yilmaz, S., Tringe, S.G., Wang, Z., Chen, F., Lindquist, E.A., Sorek, R., Hugenholtz, P., 2010. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nat Methods 7, 807-812.

Hemm, M.R., Paul, B.J., Schneider, T.D., Storz, G., Rudd, K.E., 2008. Small membrane proteins found by comparative genomics and ribosome binding site models. Mol Microbiol 70, 1487-1501.

Hou, Z., Fink, R.C., Radtke, C., Sadowsky, M.J., Diez-Gonzalez, F., 2013. Incidence of naturally internalized bacteria in lettuce leaves. Int J Food Microbiol 162, 260-265.

Howden, A.J., Geoghegan, V., Katsch, K., Efstathiou, G., Bhushan, B., Boutureira, O., Thomas, B., Trudgian, D.C., Kessler, B.M., Dieterich, D.C., Davis, B.G., Acuto, O., 2013. QuaNCAT: quantitating proteome dynamics in primary cells. Nat Methods 10, 343-346.

Hsu, P.Y., Calviello, L., Wu, H.L., Li, F.W., Rothfels, C.J., Ohler, U., Benfey, P.N., 2016. Super-resolution ribosome profiling reveals unannotated translation events in *Arabidopsis*. Proceedings of the National Academy of Sciences of the United States of America.

Hücker, S.M., Ardern, Z., Goldberg, T., Schafferhans, A., Bernhofer, M., Vestergaard, G., Nelson, C.W., Schloter, M., Rost, B., Scherer, S., Neuhaus, K., 2017a. Discovery of numerous novel small genes in the intergenic regions of the *Escherichia coli* O157:H7 Sakai genome. PLoS One, accepted.

Hücker, S.M., Simon, S., Scherer, S., Neuhaus, K., 2017b. 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 364.

Huvet, M., Stumpf, M.P., 2014. Overlapping genes: a window on gene evolvability. BMC Genomics 15, 721.

Hwang, J.Y., Buskirk, A.R., 2017. A ribosome profiling study of mRNA cleavage by the endonuclease RelE. Nucleic Acids Res 45, 327-336.

Iguchi, A., Iyoda, S., Terajima, J., Watanabe, H., Osawa, R., 2006. Spontaneous recombination between homologous prophage regions causes large-scale inversions within the *Escherichia coli* O157:H7 chromosome. Gene 372, 199-207.

Ingolia, N.T., 2014. Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet 15, 205-213.

Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J., Jackson, S.E., Wills, M.R., Weissman, J.S., 2014. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep 8, 1365-1379.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., Weissman, J.S., 2009. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. Science 324, 218-223.

Iwasaki, S., Ingolia, N.T., 2017. The Growing Toolbox for Protein Synthesis Studies. Trends Biochem Sci.

Jayapal, K.P., Philp, R.J., Kok, Y.J., Yap, M.G., Sherman, D.H., Griffin, T.J., Hu, W.S., 2008. Uncovering genes with divergent mRNA-protein dynamics in Streptomyces coelicolor. PloS one 3, e2097.

Jayaraman, D., Valdes-Lopez, O., Kaspar, C.W., Ane, J.M., 2014. Response of Medicago truncatula seedlings to colonization by *Salmonella enterica* and *Escherichia coli* O157:H7. PloS one 9, e87970.

Jensen, K.T., Petersen, L., Falk, S., Iversen, P., Andersen, P., Theisen, M., Krogh, A., 2006. Novel overlapping coding sequences in *Chlamydia trachomatis*. FEMS Microbiol Lett 265, 106-117.

Jeong, Y., Kim, J.N., Kim, M.W., Bucca, G., Cho, S., Yoon, Y.J., Kim, B.G., Roe, J.H., Kim, S.C., Smith, C.P., Cho, B.K., 2016. The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3(2). Nature communications 7, 11605.

Ji, Z., Song, R., Regev, A., Struhl, K., 2015. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. eLife 4, e08890.

Johnson, Z.I., Chisholm, S.W., 2004. Properties of overlapping genes are conserved across microbial genomes. Genome Res 14, 2268-2272.

Kannan, K., Kanabar, P., Schryer, D., Florin, T., Oh, E., Bahroos, N., Tenson, T., Weissman, J.S., Mankin, A.S., 2014. The general mode of translation inhibition by macrolide antibiotics. Proceedings of the National Academy of Sciences of the United States of America 111, 15958-15963.

Kannan, K., Vazquez-Laslop, N., Mankin, A.S., 2012. Selective protein synthesis by ribosomes with a drug-obstructed exit tunnel. Cell 151, 508-520.

Karch, H., Denamur, E., Dobrindt, U., Finlay, B.B., Hengge, R., Johannes, L., Ron, E.Z., Tonjum, T., Sansonetti, P.J., Vicente, M., 2012. The enemy within us: lessons from the 2011 European *Escherichia coli* O104:H4 outbreak. EMBO Mol. Med. 4, 841-848.

Keese, P.K., Gibbs, A., 1992. Origins of genes: "big bang" or continuous creation? Proceedings of the National Academy of Sciences of the United States of America 89, 9489-9493.

Kemp, G., Cymer, F., 2014. Small membrane proteins–elucidating the function of the needle in the haystack. Biol Chem 395, 1365-1377.

Kim, J., Webb, A.M., Kershner, J.P., Blaskowski, S., Copley, S.D., 2014. A versatile and highly efficient method for scarless genome editing in *Escherichia coli* and *Salmonella enterica*. BMC Biotechnol 14, 84.

Kim, W., Silby, M.W., Purvine, S.O., Nicoll, J.S., Hixson, K.K., Monroe, M., Nicora, C.D., Lipton, M.S., Levy, S.B., 2009. Proteomic detection of non-annotated protein-coding genes in *Pseudomonas fluorescens* Pf0-1. PloS one 4, e8455.

King, T., Kocharunchitt, C., Gobius, K., Bowman, J.P., Ross, T., 2014. Global genome response of *Escherichia coli* O157:H7 Sakai during dynamic changes in growth kinetics induced by an abrupt temperature downshift. PloS one 9, e99627.

Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H., Mori, H., 2005. Complete set of ORF clones of *Escherichia coli* ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. DNA Res 12, 291-299.

Knoppel, A., Nasvall, J., Andersson, D.I., 2016. Compensating the Fitness Costs of Synonymous Mutations. Mol Biol Evol 33, 1461-1477.

Kocharunchitt, C., King, T., Gobius, K., Bowman, J.P., Ross, T., 2012. Integrated transcriptomic and proteomic analysis of the physiological response of *Escherichia coli* O157:H7 Sakai to steady-state conditions of cold and water activity stress. Mol Cell Proteomics 11, M111 009019.

Kouse, A.B., Righetti, F., Kortmann, J., Narberhaus, F., Murphy, E.R., 2013. RNA-mediated thermoregulation of iron-acquisition genes in *Shigella dysenteriae* and pathogenic *Escherichia coli*. PloS one 8, e63781.

Kozak, M., 1983. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol Rev 47, 1-45.

Krakauer, D.C., 2000. Stability and evolution of overlapping genes. Evolution 54, 731-739.

Kröger, C., Dillon, S.C., Cameron, A.D., Papenfort, K., Sivasankaran, S.K., Hokamp, K., Chao, Y., Sittka, A., Hebrard, M., Handler, K., Colgan, A., Leekitcharoenphon, P., Langridge, G.C., Lohan, A.J., Loftus, B., Lucchini, S., Ussery, D.W., Dorman, C.J., Thomson, N.R., Vogel, J., Hinton, J.C., 2012. The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. Proceedings of the National Academy of Sciences of the United States of America 109, E1277-1286.

Kuersten, S., Radek, A., Vogel, C., Penalva, L.O., 2013. Translation regulation gets its 'omics' moment. Wiley Interdisciplinary Reviews: RNA 4, 617-630.

Landry, C.R., Zhong, X., Nielly-Thibault, L., Roucou, X., 2015. Found in translation: functions and evolution of a recently discovered alternative proteome. Curr. Opin. Struct. Biol. 32, 74-80.

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, R.B., 2014. Comparative transcriptomics and translatomics to identify novel overlapping genes, active hypothetical genes, and ncRNAs in *Escherichia coli* O157:H7 EDL933. PhD Thesis, Technische Universität München, München.

Lareau, L.F., Hite, D.H., Hogan, G.J., Brown, P.O., 2014. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. eLife 3, e01257.

Larsson, O., Tian, B., Sonenberg, N., 2013. Toward a genome-wide landscape of translational control. Cold Spring Harbor perspectives in biology 5, a012302.

Lasa, I., Toledo-Arana, A., Dobin, A., Villanueva, M., de los Mozos, I.R., Vergara-Irigaray, M., Segura, V., Fagegaltier, D., Penades, J.R., Valle, J., Solano, C., Gingeras, T.R., 2011. Genome-wide antisense transcription drives mRNA processing in bacteria. Proceedings of the National Academy of Sciences of the United States of America 108, 20172-20177.

Lèbre, S., Gascuel, O., 2017. The combinatorics of overlapping genes. J. Theor. Biol. 415, 90-101.

Legendre, R., Baudin-Baillieu, A., Hatin, I., Namy, O., 2015. RiboTools: a Galaxy toolbox for qualitative ribosome profiling analysis. Bioinformatics 31, 2586-2588.

Levitt, M., 2009. Nature of the protein universe. Proceedings of the National Academy of Sciences of the United States of America 106, 11079-11084.

Lewis, S.B., Cook, V., Tighe, R., Schuller, S., 2015. Enterohemorrhagic *Escherichia coli* colonization of human colonic epithelium in vitro and ex vivo. Infect Immun 83, 942-949.

Li, G.W., Burkhardt, D., Gross, C., Weissman, J.S., 2014. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell 157, 624-635.

Li, G.W., Oh, E., Weissman, J.S., 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484, 538-541.

Light, S., Basile, W., Elofsson, A., 2014. Orphans and new gene origination, a structural and evolutionary perspective. Curr. Opin. Struct. Biol. 26, 73-83.

Lillo, F., Krakauer, D.C., 2007. A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes. Biology direct 2, 22.

Lim, J.Y., Hong, J.B., Sheng, H., Shringi, S., Kaul, R., Besser, T.E., Hovde, C.J., 2010a. Phenotypic diversity of *Escherichia coli* O157:H7 strains associated with the plasmid O157. Journal of microbiology 48, 347-357.

Lim, J.Y., Yoon, J., Hovde, C.J., 2010b. A brief overview of *Escherichia coli* O157:H7 and its plasmid O157. J Microbiol Biotechnol 20, 5-14.

Lin, Y.F., A, D.R., Guan, S., Mamanova, L., McDowall, K.J., 2013. A combination of improved differential and global RNA-seq reveals pervasive transcription initiation and events in all stages of the life-cycle of functional RNAs in *Propionibacterium acnes*, a major contributor to wide-spread human disease. BMC Genomics 14, 620.

Liu, B., Qian, S.B., 2016. Characterizing inactive ribosomes in translational profiling. Translation 4, e1138018.

Liu, T.Y., Huang, H.H., Wheeler, D., Xu, Y., Wells, J.A., Song, Y.S., Wiita, A.P., 2017. Time-Resolved Proteomics Extends Ribosome Profiling-Based Measurements of Protein Synthesis Dynamics. Cell Syst 4, 636-644 e639.

Liveris, D., Schwartz, J.J., Geertman, R., Schwartz, I., 1993. Molecular cloning and sequencing of infC, the gene encoding translation initiation factor IF3, from four enterobacterial species. FEMS Microbiol Lett 112, 211-216.

Lu, T.-C., Leu, J.-Y., Lin, W.-c., 2017. A comprehensive analysis of transcript-supported *de novo* genes in *Saccharomyces sensu stricto* yeasts. Mol Biol Evol.

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.

Ma, J., Ibekwe, A.M., Yi, X., Wang, H., Yamazaki, A., Crowley, D.E., Yang, C.H., 2011. Persistence of *Escherichia coli* O157:H7 and its mutants in soils. PloS one 6, e23191.

Maier, T., Schmidt, A., Guell, M., Kuhner, S., Gavin, A.C., Aebersold, R., Serrano, L., 2011. Quantification of mRNA and protein and integration with protein turnover in a bacterium. Mol Syst Biol 7, 511.

Malone, B., Atanassov, I., Aeschimann, F., Li, X., Grosshans, H., Dieterich, C., 2017. Bayesian prediction of RNA translation from ribosome profiling. Nucleic Acids Res 45, 2960-2972.

Manoil, C., 1991. Analysis of Membrane Protein Topology Using Alkaline Phosphatase and b-Galactosidase Gene Fusions, Methods in Cell Biology. Academic Press Inc., pp. 61-75.

Marks, J., Kannan, K., Roncase, E.J., Klepacki, D., Kefi, A., Orelle, C., Vazquez-Laslop, N., Mankin, A.S., 2016. Context-specific inhibition of translation by ribosomal antibiotics targeting the peptidyl transferase center. Proceedings of the National Academy of Sciences of the United States of America 113, 12150-12155.

Martens, A.T., Taylor, J., Hilser, V.J., 2015. Ribosome A and P sites revealed by length analysis of ribosome profiling data. Nucleic Acids Res 43, 3680-3687.

Merino, E., Balbas, P., Puente, J.L., Bolivar, F., 1994. Antisense overlapping open reading frames in genes from bacteria to humans. Nucleic Acids Res 22, 1903-1908.

Miettinen, T.P., Bjorklund, M., 2015. Modified ribosome profiling reveals high abundance of ribosome protected mRNA fragments derived from 3' untranslated regions. Nucleic Acids Res 43, 1019-1034.

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.

Mir, K., Schober, S., 2014. Selection pressure in alternative reading frames. PloS one 9, e108768.

Mohammad, F., Woolstenhulme, C.J., Green, R., Buskirk, A.R., 2016. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. Cell Rep 14, 686-694.

Mohawk, K.L., O'Brien, A.D., 2011. Mouse models of *Escherichia coli* O157:H7 infection and shiga toxin injection. Journal of biomedicine & biotechnology 2011, 258185.

Monk, J.M., Charusanti, P., Aziz, R.K., Lerman, J.A., Premyodhin, N., Orth, J.D., Feist, A.M., Palsson, B.O., 2013. Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. Proceedings of the National Academy of Sciences of the United States of America 110, 20338-20343.

Morisaki, T., Lyon, K., DeLuca, K.F., DeLuca, J.G., English, B.P., Zhang, Z., Lavis, L.D., Grimm, J.B., Viswanathan, S., Looger, L.L., Lionnet, T., Stasevich, T.J., 2016. Real-time quantification of single RNA translation dynamics in living cells. Science 352, 1425-1429.

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

Muniesa, M., Hammerl, J.A., Hertwig, S., Appel, B., Brussow, H., 2012. Shiga toxin-producing *Escherichia coli* O104:H4: a new challenge for microbiology. Appl Environ Microbiol 78, 4065-4073.

Nakahigashi, K., Takai, Y., Kimura, M., Abe, N., Nakayashiki, T., Shiwa, Y., Yoshikawa, H., Wanner, B.L., Ishihama, Y., Mori, H., 2016. Comprehensive identification of translation start sites by tetracycline-inhibited ribosome profiling. DNA Res 23, 193-201.

Nakahigashi, K., Takai, Y., Shiwa, Y., Wada, M., Honma, M., Yoshikawa, H., Tomita, M., Kanai, A., Mori, H., 2014. Effect of codon adaptation on codon-level and gene-level translation efficiency in vivo. BMC Genomics 15, 1115.

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

Neuhaus, K., Landstorfer, R., Fellner, L., Simon, S., Marx, H., Ozoline, O., Schafferhans, A., Goldberg, T., Rost, B., Küster, B., Keim, D.A., Scherer, S., 2016. Translatomics combined with transcriptomics and proteomics reveals novel functional, recently evolved orphan genes in *Escherichia coli* O157:H7 (EHEC). BMC Genomics 7, 133.

Neuhaus, K., Landstorfer, R., Simon, S., Schober, S., Wright, P.R., Smith, C., Backofen, R., Wecko, R., Keim, D.A., 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 18, 216.

Nguyen, Y., Sperandio, V., 2012. Enterohemorrhagic *E. coli* (EHEC) pathogenesis. Front Cell Infect Microbiol 2, 90.

Nichols, R.J., Sen, S., Choo, Y.J., Beltrao, P., Zietek, M., Chaba, R., Lee, S., Kazmierczak, K.M., Lee, K.J., Wong, A., Shales, M., Lovett, S., Winkler, M.E., Krogan, N.J., Typas, A., Gross, C.A., 2011. Phenotypic landscape of a bacterial cell. Cell 144, 143-156.

Nonaka, G., Blankschien, M., Herman, C., Gross, C.A., Rhodius, V.A., 2006. Regulon and promoter analysis of the *E. coli* heat-shock factor, σ^{32}, reveals a multifaceted cellular response to heat stress. Genes Dev 20, 1776-1789.

Normark, S., Bergström, S., Edlund, T., Grundström, T., Jaurin, B., Lindberg, F.P., Olsson, O., 1983. Overlapping genes. Annu Rev Genet 17, 499-525.

O'Connor, P.B., Li, G.W., Weissman, J.S., Atkins, J.F., Baranov, P.V., 2013. rRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome profiling experiments. Bioinformatics 29, 1488-1491.

Oh, E., Becker, A.H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R.J., Typas, A., Gross, C.A., Kramer, G., Weissman, J.S., Bukau, B., 2011. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell 147, 1295-1308.

Oheigeartaigh, S.S., Armisen, D., Byrne, K.P., Wolfe, K.H., 2014. SearchDOGS bacteria, software that provides automated identification of potentially missed genes in annotated bacterial genomes. J Bacteriol 196, 2030-2042.

Olexiouk, V., Crappe, J., Verbruggen, S., Verhegen, K., Martens, L., Menschaert, G., 2016. sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res 44, D324-329.

Pavesi, A., Magiorkinis, G., Karlin, D.G., 2013. Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of Deltaretroviruses. PLoS Comput Biol 9, e1003162.

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K.S., Buckley, M.J., Tabor, B., Signal, B., Gloss, B.S., Hammang, C.J., Rost, B., 2015. Unexpected features of the dark proteome. Proc. Natl. Acad. Sci. USA 112, 15898-15903.

Perkins, T.T., Kingsley, R.A., Fookes, M.C., Gardner, P.P., James, K.D., Yu, L., Assefa, S.A., He, M., Croucher, N.J., Pickard, D.J., Maskell, D.J., Parkhill, J., Choudhary, J., Thomson, N.R., Dougan, G., 2009. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. PLoS Genet 5, e1000569.

Persad, A.K., LeJeune, J.T., 2014. Animal Reservoirs of Shiga Toxin-Producing *Escherichia coli.* Microbiology spectrum 2, EHEC-0027-2014.

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

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

Raghavan, R., Groisman, E.A., Ochman, H., 2011. Genome-wide detection of novel regulatory RNAs in *E. coli*. Genome Res 21, 1487-1497.

Rancurel, C., Khosravi, M., Dunker, A.K., Romero, P.R., Karlin, D., 2009. Overlapping genes produce proteins with unusual sequence properties and offer insight into *de novo* protein creation. J Virol 83, 10719-10736.

Reiland, H.A., Omolo, M.A., Johnson, T.J., Baumler, D.J., 2014. A Survey of *Escherichia coli* O157:H7 Virulence Factors: The First 25 Years and 13 Genomes. Advances in Microbiology 4, 390-423.

Rex, G., Surin, B., Besse, G., Schneppe, B., McCarthy, J.E., 1994. The mechanism of translational coupling in *Escherichia coli*. Higher order structure in the *atpHA* mRNA acts as a conformational switch regulating the access of de novo initiating ribosomes. J Biol Chem 269, 18118-18127.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2009. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140.

Ruiz-Orera, J., Messeguer, X., Subirana, J.A., Alba, M.M., 2014. Long non-coding RNAs as a source of new peptides. eLife 3, e03523.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., Barrell, B., 2000. Artemis: sequence visualization and annotation. Bioinformatics 16, 944-945.

Sabath, N., Ferrada, E., Barve, A., Wagner, A., 2013. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. Genome biology and evolution 5, 966-977.

Sadiq, S.M., Hazen, T.H., Rasko, D.A., Eppinger, M., 2014. EHEC Genomics: Past, Present, and Future. Microbiology spectrum 2, EHEC-0020-2013.

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

Saldana, Z., Sanchez, E., Xicohtencatl-Cortes, J., Puente, J.L., Giron, J.A., 2011. Surface structures involved in plant stomata and leaf colonization by shiga-toxigenic *Escherichia coli* O157:H7. Front Microbiol 2, 119.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., Selbach, M., 2011. Global quantification of mammalian gene expression control. Nature 473, 337-342.

Semenov, A.M., Kuprianov, A.A., van Bruggen, A.H., 2010. Transfer of enteric pathogens to successive habitats as part of microbial cycles. Microb Ecol 60, 239-249.

Sharma, V., Firth, A.E., Antonov, I., Fayet, O., Atkins, J.F., Borodovsky, M., Baranov, P.V., 2011. A pilot study of bacterial genes with disrupted ORFs reveals a surprising profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment. Mol. Biol. Evol., msr155.

Silby, M.W., Levy, S.B., 2004. Use of in vivo expression technology to identify genes important in growth and survival of *Pseudomonas fluorescens* Pf0-1 in soil: discovery of expressed sequences with novel genetic organization. J Bacteriol 186, 7411-7419.

Silby, M.W., Rainey, P.B., Levy, S.B., 2004. IVET experiments in *Pseudomonas fluorescens* reveal cryptic promoters at loci associated with recognizable overlapping genes. Microbiology 150, 518-520.

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.

Sin, C., Chiarugi, D., Valleriani, A., 2016. Quantitative assessment of ribosome drop-off in *E. coli*. Nucleic Acids Res 44, 2528-2537.

Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., Saghatelian, A., 2013. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 9, 59-64.

Smith, J.E., Alvarez-Dominguez, J.R., Kline, N., Huynh, N.J., Geisler, S., Hu, W., Coller, J., Baker, K.E., 2014. Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. Cell Rep 7, 1858-1866.

Sousa, C.P., 2003. East1 toxin and its presence in a changing microbial world. J Venomous Anim Toxins Trop Dis 9, 4-52.

Spiers, A.J., Bergquist, P.L., 1992. Expression and regulation of the RepA protein of the RepFIB replicon from plasmid P307. J Bacteriol 174, 7533-7541.

Stein, R.A., Katz, D.E., 2017. *Escherichia coli*, cattle and the propagation of disease. FEMS Microbiol Lett 364.

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V.T., Hein, M.Y., Huang, S.X., Ma, M., Shen, B., Qian, S.B., Hengel, H., Mann, M., Ingolia, N.T., Weissman, J.S., 2012. Decoding human cytomegalovirus. Science 338, 1088-1093.

Storz, G., Wolf, Y.I., Ramamurthi, K.S., 2014. Small proteins can no longer be ignored. Annu Rev Biochem 83, 753-777.

Subramaniam, A.R., Zid, B.M., O'Shea, E.K., 2014. An integrated approach reveals regulatory controls on bacterial translation elongation. Cell 159, 1200-1211.

Taniguchi, Y., Choi, P.J., Li, G.W., Chen, H., Babu, M., Hearn, J., Emili, A., Xie, X.S., 2010. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science 329, 533-538.

Tunca, S., Barreiro, C., Coque, J.J., Martin, J.F., 2009. Two overlapping antiparallel genes encoding the iron regulator DmdR1 and the Adm proteins control siderophore and antibiotic biosynthesis in *Streptomyces coelicolor* A3(2). FEBS J 276, 4814-4827.

Van Damme, P., Gawron, D., Van Criekinge, W., Menschaert, G., 2014. N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. Mol Cell Proteomics 13, 1245-1261.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., Thermes, C., 2014. Ten years of next-generation sequencing technology. Trends Genet 30, 418-426.

Veloso, F., Riadi, G., Aliaga, D., Lieph, R., Holmes, D.S., 2005. Large-scale, multi-genome analysis of alternate open reading frames in bacteria and archaea. Omics 9, 91-105.

Vesper, O., Amitai, S., Belitsky, M., Byrgazov, K., Kaberdina, A.C., Engelberg-Kulka, H., Moll, I., 2011. Selective translation of leaderless mRNAs by specialized ribosomes generated by MazF in Escherichia coli. Cell 147, 147-157.

Vogel, C., Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature Reviews Genetics 13, 227-232.

Wade, J.T., Grainger, D.C., 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol 12, 647-653.

Wang, J., Roy, B., 2017. Single-cell RNA-seq reveals lincRNA expression differences in Hela-S3 cells. Biotechnol Lett 39, 359-366.

Wang, Z., Sun, X., Zhao, Y., Guo, X., Jiang, H., Li, H., Gu, Z., 2015. Evolution of gene regulation during transcription and translation. Genome biology and evolution 7, 1155-1167.

Warren, A.S., Archuleta, J., Feng, W.C., Setubal, J.C., 2010. Missing genes in the annotation of prokaryotic genomes. BMC Bioinformatics 11, 131.

Wasala, L., Talley, J.L., Desilva, U., Fletcher, J., Wayadande, A., 2013. Transfer of *Escherichia coli* O157:H7 to spinach by house flies, *Musca domestica* (Diptera: Muscidae). Phytopathology 103, 373-380.

Willems, P., Ndah, E., Jonckheere, V., Stael, S., Sticker, A., Martens, L., Van Breusegem, F., Gevaert, K., Van Damme, P., 2017. N-terminal Proteomics Assisted Profiling of the Unexplored Translation Initiation Landscape in *Arabidopsis thaliana*. Mol Cell Proteomics 16, 1064-1080.

Wong, C.S., Jelacic, S., Habeeb, R.L., Watkins, S.L., Tarr, P.I., 2000. The risk of the hemolytic-uremic syndrome after antibiotic treatment of *Escherichia coli* O157:H7 infections. N Engl J Med 342, 1930-1936.

Woolstenhulme, C.J., Guydosh, N.R., Green, R., Buskirk, A.R., 2015. High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. Cell Rep 11, 13-21.

Wu, B., Eliscovich, C., Yoon, Y.J., Singer, R.H., 2016. Translation dynamics of single mRNAs in live cells and neurons. Science 352, 1430-1435.

Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., Honigschmid, P., Schafferhans, A., Roos, M., Bernhofer, M., Richter, L., Ashkenazy, H., Punta, M., Schlessinger, A., Bromberg, Y., Schneider, R., Vriend, G., Sander, C., Ben-Tal, N., Rost, B., 2014. PredictProtein--an open resource for online prediction of protein structural and functional features. Nucleic Acids Res 42, W337-343.

Yomtovian, I., Teerakulkittipong, N., Lee, B., Moult, J., Unger, R., 2010. Composition bias and the origin of ORFan genes. Bioinformatics 26, 996-999.

Yu, H.H., Di Russo, E.G., Rounds, M.A., Tan, M., 2006. Mutational analysis of the promoter recognized by *Chlamydia* and *Escherichia coli* sigma(28) RNA polymerase. J Bacteriol 188, 5524-5531.

Zhao, L., Liu, L., Leng, W., Wei, C., Jin, Q., 2011. A Proteogenomic analysis of *Shigella flexneri* using 2D LC-MALDI TOF/TOF. BMC Genomics 12, 528.

Zheng, X., Hu, G.-Q., She, Z.-S., Zhu, H., 2011. Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. BMC Genomics 12, 361.

Zhu, S., Qing, T., Zheng, Y., Jin, L., Shi, L., 2017. Advances in single-cell RNA sequencing and its applications in cancer research. Oncotarget.

Zupanic, A., Meplan, C., Grellscheid, S.N., Mathers, J.C., Kirkwood, T.B., Hesketh, J.E., Shanley, D.P., 2014. Detecting translational regulation by change point analysis of ribosome profiling data sets. RNA 20, 1507-1518.

Zur, H., Aviner, R., Tuller, T., 2016. Complementary Post Transcriptional Regulatory Information is Detected by PUNCH-P and Ribosome Profiling. Scientific reports 6.

5. Supplement

Supplementary Table S2: Genome location, expression level, translatability and ORF coverage of the novel OLGs.

description				LB. 37°C				BHI control				BHI stress			
gene name	start	stop	length [bp]	RPKM transcriptome*	RPKM translatome*	RCV*	coverage*	RPKM transcriptome*	RPKM translatome*	RCV*	coverage*	RPKM transcriptome*	RPKM translatome*	RCV*	coverage*
OLGECs001	6877	6975	99	13.2	9.4	1.26	0.38	42.9	2.2	0.05	0.63	19.5	0.6	0.02	0.20
OLGECs002	7097	7192	96	4.9	4.5	0.26	0.35	4.8	4.8	0.39	0.54	9.3	0.1	0.04	0.26
OLGECs003	8256	8354	99	13.2	12.2	1.18	0.29	21.7	5.1	0.46	0.65	5.2	1.4	0.26	0.39
OLGECs004	9768	10313	546	39.7	45.3	1.21	0.78	86.8	17.9	0.23	0.80	31.4	2.1	0.07	0.13
OLGECs005	15413	15697	285	878.4	222.3	0.25	0.76	2722.9	60.8	0.03	0.77	2105.7	19.5	0.01	0.32
OLGECs006	18419	18541	123	7.2	17.3	4.31	0.46	18.5	4.7	0.27	0.44	20.5	2.3	0.11	0.24
OLGECs007	24472	24621	150	15.4	3.1	0.13	0.06	13.5	8.8	0.71	0.39	94.2	4.0	0.02	0.24
OLGECs008	55330	55656	327	32.8	49.0	1.49	0.55	67.9	9.0	0.13	0.67	32.6	2.7	0.08	0.62
OLGECs009	57569	57670	102	8.2	1.8	0.25	0.14	6.3	8.1	0.11	0.43	1.4	0.6	0.11	0.07
OLGECs010	74157	74267	111	12.5	8.1	0.62	0.40	3.3	2.8	0.19	0.41	0.0	1.7	0.00	0.30
OLGECs011	82148	82252	105	3.5	7.1	1.01	0.14	33.3	5.9	0.19	0.54	12.7	1.5	0.13	0.02
OLGECs012	102752	102847	96	10.6	6.8	0.59	0.21	10.1	2.3	0.11	0.22	182.0	48.9	0.13	0.25
OLGECs013	116069	116329	261	3.4	2.0	1.30	0.16	9.1	2.1	0.25	0.52	15.0	6.4	0.42	0.22
OLGECs014	116532	116720	189	9.3	16.5	1.76	0.24	8.6	13.0	1.01	0.36	10.6	8.2	0.49	0.29
OLGECs015	121015	121119	105	2.2	6.8	0.53	0.46	12.7	14.5	0.79	0.28	2.8	0.5	0.04	0.07
OLGECs016	133796	134005	210	9.5	13.4	1.37	0.48	10.6	3.1	0.29	0.45	34.3	8.6	0.34	0.74
OLGECs017	146371	146544	174	21.9	52.9	2.53	0.93	138.9	24.2	0.23	0.86	86.8	1.4	0.02	0.28
OLGECs018	146653	146886	234	52.5	24.9	0.50	0.73	152.0	9.1	0.06	0.72	123.0	4.5	0.04	0.38
OLGECs019	152528	152704	177	76.1	46.0	0.60	0.78	239.8	28.2	0.14	0.87	90.8	3.7	0.04	0.67
OLGECs020	152574	152762	189	104.5	119.1	1.32	0.90	360.2	36.5	0.11	0.91	113.7	5.1	0.04	0.74
OLGECs021	152659	152811	153	104.9	237.3	2.74	0.88	400.8	69.5	0.18	0.92	124.9	4.7	0.04	0.69
OLGECs022	152744	152854	111	70.0	179.0	3.21	0.78	318.0	66.3	0.21	0.90	88.8	1.9	0.02	0.71
OLGECs023	152858	152971	114	23.6	16.8	0.72	0.54	81.8	13.4	0.17	0.59	29.3	0.5	0.02	0.42
OLGECs024	155495	155617	123	7.2	104.1	18.61	0.47	24.7	36.1	1.24	0.54	3.6	1.1	0.06	0.45
OLGECs025	156164	156334	171	29.1	19.7	0.69	0.56	76.8	23.7	0.28	0.61	62.9	9.0	0.13	0.26
OLGECs026	156531	156689	159	15.2	13.4	1.22	0.56	23.0	9.4	0.42	0.67	12.6	2.0	0.15	0.20
OLGECs027	159905	160027	123	1.5	6.8	0.76	0.65	6.9	4.5	0.74	0.46	5.4	0.2	0.04	0.00
OLGECs028	177396	177566	171	0.0	5.6	0.00	0.37	9.6	7.0	0.73	0.50	3.0	0.3	0.15	0.24
OLGECs029	178675	178917	243	7.1	2.2	0.45	0.15	16.9	6.0	0.17	0.30	34.3	5.2	0.11	0.15
OLGECs030	181180	181326	147	13.3	19.7	1.75	0.63	37.5	10.4	0.26	0.60	53.3	1.3	0.04	0.06
OLGECs031	216670	216774	105	8.8	16.5	0.71	0.57	15.3	4.3	0.33	0.40	2.8	0.4	0.00	0.22

OLGECs032	237138	237236	99	8.9	2.5	0.55	0.10	25.2	6.2	0.26	0.25	28.5	6.4	0.15	0.13
OLGECs033	238339	238515	177	8.9	21.6	2.72	0.53	17.1	6.1	0.39	0.56	25.9	5.5	0.27	0.47
OLGECs034	245019	245207	189	5.4	10.5	1.81	0.38	12.3	12.4	1.00	0.47	13.8	10.6	0.38	0.17
OLGECs035	245032	245190	159	5.3	11.3	2.40	0.37	14.1	13.9	1.00	0.43	16.4	12.6	0.38	0.25
OLGECs036	255598	255801	204	13.7	22.9	2.09	0.62	41.1	2.7	0.07	0.46	24.4	0.6	0.03	0.29
OLGECs037	269980	270252	273	6.6	9.2	1.32	0.40	8.9	1.5	0.18	0.37	19.3	4.5	1.09	0.35
OLGECs038	271101	271304	204	46.3	43.1	1.12	0.58	122.2	17.3	0.15	0.85	68.0	3.9	0.07	0.16
OLGECs039	326539	326649	111	51.0	25.9	0.50	0.81	121.0	12.7	0.12	0.83	70.7	2.5	0.04	0.26
OLGECs040	326799	326924	126	8.1	14.0	1.68	0.46	26.5	5.2	0.20	0.50	11.2	1.0	0.06	0.00
OLGECs041	340060	340233	174	9.3	9.8	0.94	0.41	16.2	2.0	0.18	0.50	11.9	0.3	0.02	0.16
OLGECs042	346865	347047	183	5.3	30.5	8.02	0.62	42.4	5.7	0.14	0.72	27.5	3.9	0.14	0.62
OLGECs043	346933	347046	114	8.5	43.3	7.04	0.69	49.1	6.8	0.14	0.65	38.3	5.3	0.14	0.54
OLGECs044	346938	347072	135	19.6	38.2	1.98	0.67	75.5	6.0	0.09	0.63	66.4	4.9	0.07	0.15
OLGECs045	347556	347663	108	3.9	15.4	4.10	0.31	26.4	5.2	0.27	0.69	13.1	0.2	0.04	0.11
OLGECs046	379017	379136	120	37.9	13.1	0.35	0.53	59.3	8.5	0.26	0.80	171.6	0.9	0.03	0.08
OLGECs047	383798	384025	228	11.4	30.1	3.51	0.52	39.2	7.3	0.19	0.57	12.7	2.1	0.21	0.13
OLGECs048	423441	423587	147	2.8	3.6	1.39	0.38	2.3	1.1	0.59	0.56	0.0	1.6	0.00	0.34
OLGECs049	435745	435852	108	0.0	10.6	0.00	0.48	3.9	1.6	0.43	0.60	9.6	0.0	0.00	0.03
OLGECs050	442225	442353	129	11.5	18.0	2.10	0.28	12.0	4.1	0.36	0.24	11.5	3.6	0.16	0.14
OLGECs051	459830	460162	333	19.4	19.5	1.07	0.45	55.0	20.1	0.31	0.67	72.7	7.7	0.08	0.17
OLGECs052	466701	466958	258	22.5	5.6	0.26	0.36	43.5	18.9	0.42	0.52	26.2	0.5	0.02	0.11
OLGECs053	475453	475566	114	0.0	1.9	0.00	0.09	3.7	2.2	0.62	0.50	0.0	1.4	0.00	0.28
OLGECs054	487585	487689	105	57.1	20.9	0.42	0.71	143.3	11.0	0.08	0.69	126.5	1.0	0.01	0.14
OLGECs055	495547	495654	108	56.5	26.4	0.78	0.21	62.4	48.9	0.46	0.34	41.3	58.2	0.70	0.53
OLGECs056	543104	543298	195	9.8	16.5	1.98	0.54	142.6	13.9	0.10	0.69	71.8	3.5	0.06	0.33
OLGECs057	543728	543847	120	1.9	3.1	0.00	0.28	32.2	9.4	0.29	0.60	13.6	0.2	0.02	0.13
OLGECs058	551819	551968	150	111.2	59.1	0.54	0.71	277.9	25.4	0.09	0.71	313.4	2.6	0.01	0.18
OLGECs059	566926	567030	105	12.9	9.7	2.35	0.44	31.8	7.9	0.25	0.50	37.5	0.5	0.02	0.32
OLGECs060	569359	569523	165	2.5	5.3	2.28	0.45	11.4	1.0	0.14	0.36	3.1	2.1	0.52	0.31
OLGECs061	573300	573530	231	80.8	221.1	2.71	0.61	181.7	33.1	0.18	0.69	159.0	13.5	0.09	0.24
OLGECs062	573442	573540	99	184.2	499.8	2.69	0.88	411.4	79.8	0.19	0.89	362.9	24.5	0.06	0.41
OLGECs063	573506	573748	243	141.6	71.4	0.51	0.88	314.7	20.9	0.07	0.75	250.8	5.1	0.02	0.33
OLGECs064	573562	573720	159	117.3	72.0	0.62	0.85	296.2	22.0	0.08	0.78	241.7	5.8	0.02	0.21
OLGECs065	574555	575199	645	6.9	14.8	2.49	0.53	41.3	5.9	0.15	0.63	20.2	1.9	0.10	0.19
OLGECs066	619829	620812	984	4.6	13.3	2.92	0.29	30.3	5.2	0.18	0.39	6.5	2.9	0.40	0.25
OLGECs067	620743	620847	105	26.1	31.8	1.37	0.83	139.1	14.2	0.11	0.95	28.2	7.0	0.26	0.25
OLGECs068	651760	651891	132	1.4	34.7	8.07	0.58	49.1	41.3	0.70	0.76	14.0	0.4	0.02	0.15

OLGECs069	657241	657408	168	24.9	58.3	2.37	0.74	116.9	54.0	0.56	0.82	28.2	2.6	0.11	0.14
OLGECs070	657705	657803	99	6.1	10.9	1.66	0.71	55.1	5.6	0.12	0.69	20.2	4.1	0.18	0.28
OLGECs071	678261	678368	108	2.2	6.9	0.00	0.19	12.7	5.3	0.44	0.56	4.8	0.3	0.04	0.10
OLGECs072	678556	678687	132	29.9	30.2	1.04	0.56	35.6	14.1	0.65	0.64	20.8	8.6	0.43	0.32
OLGECs073	690223	690459	237	114.7	42.3	0.37	0.39	235.1	12.6	0.07	0.70	119.7	1.3	0.01	0.15
OLGECs074	717267	717458	192	20.8	14.2	0.76	0.51	33.6	11.1	0.29	0.39	116.2	11.6	0.05	0.14
OLGECs075	717349	717495	147	32.9	15.2	0.52	0.60	42.2	12.1	0.19	0.32	147.7	15.2	0.05	0.19
OLGECs076	727989	728081	93	30.0	12.0	0.60	0.40	48.2	4.8	0.11	0.64	119.7	3.4	0.03	0.14
OLGECs077	740595	740792	198	34.9	40.9	1.19	0.78	54.0	8.4	0.16	0.79	50.2	2.1	0.04	0.17
OLGECs078	752045	752188	144	23.6	29.9	1.48	0.58	0.0	2.0	0.00	0.39	0.0	0.0	0.00	0.00
OLGECs079	797490	797597	108	2.2	22.6	2.63	0.48	18.9	14.3	1.14	0.39	24.7	1.9	0.07	0.10
OLGECs080	868432	868578	147	7.9	15.6	2.52	0.52	4.1	2.1	0.62	0.28	12.6	1.3	0.09	0.16
OLGECs081	882190	882318	129	55.8	83.2	1.68	0.74	108.4	15.6	0.15	0.84	90.8	6.3	0.06	0.59
OLGECs082	947589	947687	99	14.5	23.1	1.58	0.51	23.9	5.8	0.24	0.26	27.8	2.1	0.18	0.07
OLGECs083	963145	963258	114	2.0	17.9	1.19	0.55	13.3	1.5	0.11	0.36	9.8	0.7	0.04	0.18
OLGECs084	963185	963298	114	4.1	18.7	0.59	0.50	8.6	1.4	0.19	0.39	4.6	0.2	0.04	0.51
OLGECs085	975384	975572	189	44.5	123.2	3.02	0.64	114.9	23.7	0.20	0.58	125.9	27.6	0.21	0.26
OLGECs086	975707	975880	174	4.2	23.8	1.39	0.49	29.9	11.3	0.40	0.78	24.7	5.6	0.50	0.19
OLGECs087	1015683	1015907	225	53.0	28.1	0.53	0.40	50.7	8.6	0.20	0.53	12.8	0.3	0.04	0.51
OLGECs088	1053555	1053815	261	86.4	74.0	0.85	0.72	111.7	15.1	0.14	0.61	171.2	3.0	0.02	0.71
OLGECs089	1102257	1102388	132	7.7	46.3	6.05	0.61	11.0	3.6	0.33	0.55	3.9	1.8	0.60	0.64
OLGECs090	1107647	1107748	102	3.6	0.9	0.13	0.07	3.3	4.6	0.97	0.27	10.9	4.3	0.00	0.31
OLGECs091	1113673	1113801	129	3.6	13.9	0.53	0.47	14.3	2.4	0.17	0.34	5.7	0.6	0.00	0.12
OLGECs092	1131730	1131822	93	0.0	14.7	0.00	0.55	6.2	1.9	0.65	0.39	0.0	0.9	0.00	0.20
OLGECs093	1148618	1148848	231	23.9	16.4	0.70	0.71	63.0	6.2	0.15	0.60	49.2	1.2	0.05	0.08
OLGECs094	1148903	1149052	150	21.4	22.0	1.16	0.63	116.2	5.7	0.11	0.61	22.8	1.9	0.09	0.22
OLGECs095	1149920	1150069	150	4.0	68.7	18.27	0.57	16.4	14.7	0.96	0.48	14.3	0.7	0.05	0.15
OLGECs096	1159597	1159764	168	62.5	141.6	2.37	0.40	122.8	46.5	0.37	0.63	171.5	25.8	0.14	0.54
OLGECs097	1218187	1218282	96	6.3	3.9	0.50	0.27	10.1	5.7	0.28	0.44	10.8	6.5	0.42	0.32
OLGECs098	1228603	1228914	312	6.2	17.1	3.21	0.38	54.5	9.0	0.34	0.65	22.1	0.9	0.04	0.17
OLGECs099	1228622	1228996	375	5.7	14.0	3.05	0.30	43.5	7.4	0.31	0.54	17.4	0.9	0.05	0.23
OLGECs100	1238178	1238282	105	33.6	8.0	0.25	0.19	13.8	5.2	0.19	0.29	49.6	6.8	0.07	0.24
OLGECs101	1241752	1241862	111	26.4	14.2	0.58	0.19	40.0	37.2	1.04	0.37	110.5	63.1	0.29	0.11
OLGECs102	1321903	1322082	180	10.0	28.6	3.02	0.40	18.6	11.4	0.61	0.71	7.4	2.8	0.14	0.26
OLGECs103	1459972	1460094	123	0.0	9.6	0.00	0.31	13.9	2.3	0.33	0.37	18.0	1.8	0.01	0.70
OLGECs104	1518936	1519049	114	0.0	0.8	0.00	0.07	7.5	1.8	0.22	0.24	7.8	3.3	0.26	0.39
OLGECs105	1668571	1668672	102	7.7	27.5	4.52	0.58	36.4	12.7	0.42	0.67	14.5	2.5	0.17	0.33

OLGECs106	1674745	1674861	117	12.3	20.6	1.68	0.74	32.7	9.7	0.32	0.80	37.4	1.2	0.03	0.21
OLGECs107	1685213	1685305	93	57.5	36.9	0.84	0.52	79.8	19.4	0.26	0.71	175.8	11.7	0.03	0.14
OLGECs108	1718444	1718587	144	13.2	68.4	5.59	0.65	6.5	9.4	1.46	0.44	7.7	0.2	0.04	0.30
OLGECs109	1731925	1732128	204	168.1	40.3	0.26	0.76	244.3	16.8	0.09	0.90	159.5	1.8	0.02	0.44
OLGECs110	1734879	1734986	108	19.8	7.7	0.39	0.25	27.0	4.3	0.16	0.38	12.4	2.7	0.21	0.47
OLGECs111	1737226	1737369	144	10.3	19.7	2.14	0.59	89.1	14.7	0.16	0.63	23.6	1.2	0.03	0.08
OLGECs112	1751359	1751475	117	10.3	12.5	1.01	0.36	11.1	3.9	0.55	0.40	14.0	1.4	0.06	0.21
OLGECs113	1836539	1836631	93	0.0	3.0	0.00	0.24	4.6	2.5	0.48	0.25	0.0	2.1	0.00	0.27
OLGECs114	1836917	1837021	105	42.0	8.0	0.23	0.08	20.7	11.9	0.29	0.38	56.7	9.6	0.08	0.08
OLGECs115	1851636	1851743	108	12.9	6.3	0.43	0.40	11.5	5.9	0.33	0.29	10.3	3.9	0.18	0.11
OLGECs116	1896510	1896608	99	22.5	25.3	1.16	0.73	151.0	9.9	0.06	0.79	135.6	2.6	0.02	0.43
OLGECs117	1898865	1898963	99	13.2	15.0	1.56	0.52	16.2	3.8	0.27	0.54	19.5	0.8	0.03	0.09
OLGECs118	1905789	1905950	162	9.8	11.3	4.81	0.34	30.4	5.8	0.19	0.51	28.4	2.1	0.08	0.13
OLGECs119	1909185	1909325	141	23.0	9.4	0.43	0.42	34.2	5.9	0.27	0.59	19.0	1.3	0.13	0.07
OLGECs120	1978720	1978929	210	135.1	131.6	0.97	0.72	42.6	22.2	0.62	0.83	35.3	1.9	0.05	0.49
OLGECs121	2009519	2009614	96	387.8	141.0	0.37	0.87	68.8	11.0	0.17	0.50	38.5	3.3	0.19	0.32
OLGECs122	2013187	2013279	93	5.0	20.6	0.86	0.69	3.9	3.9	0.34	0.44	1.6	0.8	0.26	0.00
OLGECs123	2025009	2025170	162	51.6	30.9	0.60	0.78	24.4	9.5	0.39	0.70	31.1	1.8	0.07	0.34
OLGECs124	2033641	2033796	156	42.9	27.7	0.77	0.64	58.0	12.2	0.20	0.51	58.9	3.9	0.06	0.45
OLGECs125	2089671	2089778	108	31.9	8.6	0.42	0.34	35.6	15.9	0.29	0.38	24.1	16.2	0.34	0.17
OLGECs126	2109870	2109962	93	122.9	50.2	0.45	0.54	133.9	21.9	0.17	0.62	73.4	1.2	0.02	0.18
OLGECs127	2114091	2114258	168	1.1	1.7	0.76	0.08	9.9	13.3	1.05	0.49	6.6	1.8	0.13	0.18
OLGECs128	2257366	2257470	105	2.2	4.7	0.66	0.42	2.3	0.9	0.06	0.31	0.0	1.3	0.00	0.25
OLGECs129	2265962	2266255	294	13.8	12.5	0.94	0.61	16.6	7.8	0.42	0.49	3.5	0.1	0.02	0.07
OLGECs130	2267997	2268089	93	2.5	4.7	0.13	0.21	7.2	6.8	0.78	0.45	5.6	2.4	0.37	0.08
OLGECs131	2270403	2270606	204	131.8	141.1	1.07	0.66	59.9	17.8	0.44	0.65	20.0	0.5	0.03	0.14
OLGECs132	2270536	2270748	213	183.2	199.2	1.09	0.85	67.0	30.7	0.58	0.85	14.9	0.5	0.04	0.27
OLGECs133	2282356	2282457	102	6.4	3.0	0.64	0.29	7.4	2.6	0.27	0.39	9.4	2.9	0.40	0.35
OLGECs134	2288754	2288879	126	10.0	6.4	0.72	0.42	13.7	12.2	0.53	0.33	14.8	3.0	0.10	0.08
OLGECs135	2314750	2314929	180	9.8	17.2	2.77	0.42	25.0	24.6	0.93	0.60	10.3	1.0	0.22	0.61
OLGECs136	2320123	2320314	192	113.5	250.7	2.28	0.83	444.5	148.5	0.33	0.82	181.9	26.0	0.15	0.73
OLGECs137	2332660	2332809	150	16.7	11.4	0.76	0.38	14.3	2.7	0.18	0.38	32.2	2.1	0.03	0.26
OLGECs138	2357400	2357582	183	1179.6	951.0	0.80	0.91	848.5	267.2	0.31	0.92	735.7	16.4	0.02	0.29
OLGECs139	2357972	2358166	195	2.1	10.3	4.88	0.63	11.2	10.4	1.11	0.67	8.4	0.8	0.10	0.04
OLGECs140	2358040	2358135	96	1.9	13.5	1.51	0.78	8.9	17.9	2.03	0.82	10.0	0.3	0.03	0.16
OLGECs141	2430505	2430702	198	39.9	31.5	0.87	0.66	37.8	10.0	0.26	0.62	18.3	2.0	0.11	0.21
OLGECs142	2430665	2430772	108	11.2	30.0	2.78	0.56	18.8	8.9	0.53	0.78	16.5	1.9	0.21	0.48

OLGECs143	2438350	2438442	93	7.9	4.3	0.19	0.32	0.0	6.9	0.00	0.38	21.6	3.3	0.12	0.17
OLGECs144	2441237	2441365	129	24.1	68.5	2.85	0.81	0.7	2.4	0.97	0.36	20.2	0.8	0.02	0.20
OLGECs145	2451164	2451304	141	1.7	1.3	0.00	0.12	3.0	0.8	0.27	0.31	0.0	1.4	0.00	0.33
OLGECs146	2452940	2453044	105	26.9	12.1	0.41	0.46	26.2	8.6	1.29	0.63	155.8	8.4	0.03	0.18
OLGECs147	2455003	2455128	126	74.1	27.2	0.37	0.65	171.7	9.7	0.06	0.69	220.1	7.2	0.03	0.27
OLGECs148	2478647	2478886	240	107.2	584.4	5.95	0.61	56.8	119.8	2.10	0.45	109.1	18.6	0.35	0.13
OLGECs149	2496874	2496972	99	53.9	90.6	1.67	0.67	54.6	11.3	0.23	0.58	36.0	3.9	0.18	0.11
OLGECs150	2508106	2508237	132	29.3	34.6	3.81	0.63	386.0	19.6	0.05	0.67	63.3	1.0	0.03	0.40
OLGECs151	2508279	2508398	120	9.3	22.7	5.10	0.58	59.5	15.1	0.27	0.74	11.8	0.7	0.08	0.19
OLGECs152	2508453	2508704	252	2.8	5.4	0.31	0.31	20.6	6.5	0.33	0.59	8.8	0.3	0.03	0.06
OLGECs153	2524143	2524268	126	6.2	17.4	2.75	0.66	4.1	5.4	1.27	0.57	3.0	1.0	0.15	0.12
OLGECs154	2524147	2524296	150	6.5	15.7	2.58	0.68	5.1	4.6	1.10	0.49	2.5	0.9	0.15	0.19
OLGECs155	2526936	2527133	198	108.0	70.7	0.73	0.83	789.2	76.7	0.09	0.94	187.4	6.3	0.03	0.74
OLGECs156	2533544	2533675	132	40.9	12.2	0.43	0.57	65.6	6.0	0.09	0.63	53.4	0.2	0.01	0.15
OLGECs157	2533768	2533923	156	35.5	109.5	3.45	0.50	36.2	23.5	0.79	0.51	34.7	5.1	0.14	0.73
OLGECs158	2539184	2539291	108	9.9	21.8	2.63	0.72	27.9	8.4	0.31	0.59	58.3	2.0	0.04	0.31
OLGECs159	2540686	2540895	210	3.1	23.2	9.19	0.50	12.4	8.9	0.91	0.35	54.7	1.4	0.03	0.47
OLGECs160	2591936	2592067	132	8.1	25.9	3.34	0.46	31.5	4.8	0.16	0.45	84.1	2.1	0.03	0.46
OLGECs161	2624878	2624997	120	86.3	36.7	0.44	0.23	102.1	34.7	0.28	0.30	164.2	7.6	0.03	0.17
OLGECs162	2643486	2643605	120	1.9	3.9	0.40	0.22	4.3	0.9	0.23	0.27	1.2	1.8	0.00	0.27
OLGECs163	2665421	2665525	105	7.0	42.4	1.64	0.53	42.5	12.8	0.32	0.51	21.2	1.1	0.05	0.20
OLGECs164	2720782	2720883	102	11.3	24.5	3.72	0.75	33.9	4.5	0.15	0.60	34.2	1.2	0.04	0.22
OLGECs165	2721046	2721273	228	7.3	22.6	3.23	0.53	5.6	17.3	3.38	0.50	11.1	1.6	0.16	0.18
OLGECs166	2726900	2727052	153	2.4	9.1	1.39	0.42	112.7	42.6	0.35	0.66	14.0	0.3	0.04	0.12
OLGECs167	2737012	2737194	183	6.9	25.7	3.81	0.74	43.7	6.4	0.15	0.73	22.7	3.9	0.18	0.59
OLGECs168	2737080	2737193	114	7.3	37.1	5.14	0.85	56.5	9.0	0.16	0.89	22.1	4.8	0.20	0.61
OLGECs169	2737085	2737219	135	20.3	33.4	1.70	0.77	73.5	8.6	0.12	0.87	53.7	4.2	0.07	0.27
OLGECs170	2756807	2756908	102	1.8	2.7	0.76	0.15	9.2	4.9	0.54	0.66	6.5	0.6	0.09	0.37
OLGECs171	2788274	2788399	126	4.8	4.2	0.76	0.15	14.7	8.3	0.47	0.56	33.6	11.5	0.18	0.26
OLGECs172	2791148	2791321	174	1.1	12.6	4.79	0.60	2.4	1.1	0.45	0.32	5.1	1.6	0.22	0.11
OLGECs173	2817480	2817695	216	12.6	12.6	1.16	0.61	15.6	6.8	0.49	0.77	23.4	1.0	0.05	0.52
OLGECs174	2824843	2825094	252	9.4	24.7	2.94	0.34	42.4	13.4	0.32	0.43	98.9	7.5	0.08	0.65
OLGECs175	2840030	2840143	114	5.3	2.4	0.38	0.15	2.1	5.8	1.22	0.51	1.3	0.8	0.04	0.17
OLGECs176	2842514	2842771	258	1695.6	4526.4	2.72	0.73	81.9	168.8	2.29	0.57	46.5	15.3	0.37	0.60
OLGECs177	2860955	2861077	123	26.4	13.6	0.52	0.49	75.8	16.8	0.23	0.68	13.3	0.9	0.06	0.20
OLGECs178	2881708	2881857	150	0.0	2.5	0.00	0.17	3.2	1.3	0.20	0.16	2.5	1.8	0.33	0.25
OLGECs179	2882138	2882233	96	4.4	2.9	0.76	0.19	0.0	3.1	0.00	0.31	7.7	3.7	0.24	0.27

OLGECs180	2944672	2944878	207	16.3	33.3	2.07	0.60	46.2	17.7	0.38	0.63	98.1	6.6	0.07	0.64
OLGECs181	2967894	2968037	144	7.4	13.1	2.00	0.59	23.2	8.3	0.43	0.56	12.4	1.9	0.09	0.16
OLGECs182	2979022	2979171	150	0.0	6.0	0.00	0.18	3.2	0.3	0.04	0.17	5.4	2.4	0.48	0.48
OLGECs183	2991104	2991259	156	57.7	14.1	0.24	0.58	20.7	6.6	0.27	0.59	20.0	0.2	0.01	0.18
OLGECs184	3005755	3005862	108	8.2	1.1	0.09	0.04	3.9	4.2	0.94	0.22	9.6	3.3	0.24	0.25
OLGECs185	3012123	3012236	114	7.8	6.8	1.85	0.39	14.1	3.4	0.28	0.36	30.6	10.4	0.54	0.52
OLGECs186	3012181	3012282	102	8.7	7.6	1.85	0.36	19.0	3.9	0.25	0.38	37.1	11.9	0.42	0.49
OLGECs187	3012706	3012801	96	12.0	23.9	1.93	0.49	41.0	3.6	0.11	0.66	72.5	3.4	0.05	0.34
OLGECs188	3019079	3019276	198	45.5	31.8	0.71	0.58	113.4	3.8	0.03	0.70	106.1	0.7	0.01	0.17
OLGECs189	3037472	3037594	123	10.6	6.3	0.79	0.33	9.8	6.2	0.20	0.51	9.6	8.6	0.07	0.50
OLGECs190	3083966	3084079	114	5.7	2.2	0.57	0.22	12.7	3.7	0.15	0.30	6.5	3.0	0.22	0.33
OLGECs191	3087665	3087796	132	14.4	11.2	0.87	0.67	22.5	7.1	0.33	0.78	49.4	2.2	0.05	0.36
OLGECs192	3087754	3087846	93	20.0	11.3	0.57	0.78	22.9	8.1	0.37	0.67	41.4	3.3	0.09	0.63
OLGECs193	3087771	3087863	93	15.5	13.3	0.85	0.72	20.3	8.0	0.40	0.59	50.2	3.0	0.07	0.52
OLGECs194	3103514	3103606	93	13.5	19.6	1.53	0.52	47.5	12.6	0.23	0.75	19.1	1.3	0.07	0.14
OLGECs195	3148595	3148798	204	9.6	23.0	3.20	0.55	15.8	6.1	0.40	0.49	8.7	0.5	0.07	0.17
OLGECs196	3166540	3166632	93	23.0	45.2	2.21	0.61	107.3	31.2	0.35	0.75	133.2	9.3	0.07	0.19
OLGECs197	3206562	3206690	129	14.4	16.1	1.12	0.51	51.0	2.8	0.08	0.38	22.4	0.8	0.03	0.19
OLGECs198	3211976	3212077	102	0.0	17.3	0.00	0.32	17.2	1.3	0.11	0.17	8.7	0.3	0.02	0.18
OLGECs199	3213777	3213878	102	5.9	8.8	1.55	0.58	35.4	9.5	0.32	0.85	12.3	0.3	0.04	0.08
OLGECs200	3247471	3247563	93	15.5	12.7	0.80	0.52	20.3	5.6	0.28	0.67	36.6	1.2	0.03	0.25
OLGECs201	3258789	3258911	123	11.7	6.8	0.57	0.11	8.6	13.0	0.82	0.26	25.4	8.6	0.18	0.07
OLGECs202	3260734	3260907	174	30.1	25.9	0.84	0.66	34.2	7.2	0.21	0.80	50.7	3.7	0.07	0.20
OLGECs203	3260966	3261061	96	30.0	30.2	1.01	0.68	36.7	14.7	0.39	0.60	30.0	2.2	0.17	0.42
OLGECs204	3282816	3282917	102	1.8	3.9	0.76	0.37	17.5	23.3	1.07	0.48	18.2	34.8	0.95	0.27
OLGECs205	3308007	3308147	141	0.0	42.3	0.00	0.43	18.7	8.6	0.47	0.61	3.7	1.2	0.26	0.26
OLGECs206	3309133	3309339	207	7.2	12.6	2.08	0.49	19.3	10.8	0.57	0.62	8.6	1.5	0.19	0.35
OLGECs207	3309273	3309401	129	6.5	5.5	0.95	0.27	27.6	7.3	0.25	0.60	4.0	1.9	0.44	0.45
OLGECs208	3319622	3319762	141	32.9	10.6	0.27	0.32	31.5	14.2	0.23	0.17	50.1	2.5	0.03	0.12
OLGECs209	3330622	3330888	267	8.2	19.8	3.45	0.60	16.6	5.6	0.35	0.52	5.8	0.2	0.05	0.06
OLGECs210	3330638	3330907	270	8.1	19.6	3.45	0.58	18.9	5.7	0.32	0.53	7.2	0.2	0.05	0.20
OLGECs211	3333506	3333604	99	93.3	48.5	0.51	0.23	108.7	53.3	0.39	0.49	537.0	43.4	0.04	0.22
OLGECs212	3333763	3334131	369	14.4	2.9	0.19	0.28	43.5	12.0	0.24	0.50	24.9	0.5	0.02	0.19
OLGECs213	3333975	3334070	96	9.6	4.2	0.15	0.33	37.6	6.6	0.24	0.68	13.9	0.3	0.03	0.12
OLGECs214	3394401	3394616	216	1.9	1.7	1.01	0.11	2.8	1.7	0.77	0.20	6.9	3.0	0.21	0.38
OLGECs215	3436056	3436250	195	78.4	360.2	4.47	0.67	126.1	23.1	0.20	0.68	7.6	1.2	0.16	0.56
OLGECs216	3436966	3437169	204	0.0	19.9	0.00	0.45	17.7	1.1	0.06	0.40	11.3	0.5	0.04	0.08

OLGECs217	3437120	3437215	96	0.0	15.5	0.00	0.63	26.0	1.1	0.05	0.33	29.4	0.6	0.01	0.10
OLGECs218	3461522	3461635	114	47.7	16.6	0.39	0.27	45.9	33.2	0.46	0.44	53.5	26.1	0.25	0.19
OLGECs219	3463926	3464018	93	4.5	5.0	1.15	0.39	26.8	1.9	0.07	0.46	11.2	2.5	0.20	0.42
OLGECs220	3469721	3470092	372	14.0	43.1	3.07	0.42	16.6	8.8	0.50	0.47	32.1	2.5	0.07	0.35
OLGECs221	3469887	3470096	210	16.3	68.2	4.43	0.40	18.8	14.7	0.77	0.55	50.1	2.0	0.04	0.26
OLGECs222	3504735	3504887	153	7.6	11.1	2.30	0.43	23.9	7.6	0.30	0.62	1.9	1.0	0.13	0.23
OLGECs223	3518406	3518519	114	5.3	18.7	3.18	0.52	38.0	3.2	0.09	0.40	20.8	0.5	0.03	0.43
OLGECs224	3532643	3532753	111	1.7	14.5	4.03	0.50	6.8	5.1	0.72	0.33	22.1	1.0	0.03	0.40
OLGECs225	3538978	3539241	264	401.4	156.3	0.39	0.58	364.1	68.2	0.18	0.83	137.1	2.9	0.03	0.41
OLGECs226	3597447	3597542	96	0.0	4.8	0.00	0.36	8.5	5.5	2.57	0.60	1.5	0.4	0.04	0.11
OLGECs227	3636976	3637206	231	9.6	25.0	2.69	0.36	34.3	8.6	0.27	0.49	10.6	1.7	0.15	0.05
OLGECs228	3662343	3662579	237	21.0	37.4	2.31	0.64	49.4	11.2	0.23	0.76	37.8	3.6	0.14	0.63
OLGECs229	3665285	3665575	291	142.4	48.8	0.35	0.72	477.6	30.5	0.07	0.78	255.6	1.7	0.01	0.33
OLGECs230	3667403	3667597	195	14.1	27.7	2.10	0.59	70.2	8.8	0.13	0.60	44.5	1.0	0.02	0.28
OLGECs231	3670104	3670352	249	10.7	15.8	2.60	0.61	224.6	26.2	0.21	0.73	52.7	2.8	0.05	0.08
OLGECs232	3670482	3670649	168	5.8	31.2	4.27	0.60	97.3	8.9	0.10	0.74	28.7	1.6	0.07	0.26
OLGECs233	3699402	3699509	108	5.6	11.8	1.77	0.37	13.4	5.8	0.17	0.48	103.2	7.0	0.15	0.51
OLGECs234	3702174	3702332	159	45.9	30.2	0.84	0.50	112.7	14.0	0.13	0.74	97.0	2.2	0.02	0.17
OLGECs235	3706115	3706291	177	15.8	33.5	2.40	0.77	51.6	15.6	0.32	0.87	20.5	1.0	0.06	0.10
OLGECs236	3753535	3753768	234	10.3	36.4	3.83	0.50	35.5	9.0	0.33	0.67	10.5	0.5	0.05	0.12
OLGECs237	3778152	3778379	228	8.6	21.3	2.90	0.55	68.1	14.7	0.21	0.65	61.2	2.9	0.05	0.42
OLGECs238	3778357	3778473	117	42.8	36.7	0.92	0.63	134.6	10.0	0.08	0.82	89.3	9.1	0.11	0.48
OLGECs239	3780148	3780267	120	1.5	3.1	1.01	0.30	3.5	3.5	1.07	0.60	33.4	4.7	0.16	0.53
OLGECs240	3780351	3780668	318	55.2	38.2	0.67	0.59	43.2	14.8	0.33	0.51	82.7	3.1	0.04	0.29
OLGECs241	3792783	3792929	147	15.5	4.6	0.30	0.27	14.6	10.3	0.48	0.36	10.1	11.1	0.55	0.14
OLGECs242	3794979	3795248	270	2.7	18.1	2.39	0.34	13.2	9.8	0.75	0.54	8.2	0.6	0.07	0.31
OLGECs243	3795055	3795354	300	2.6	16.2	6.14	0.34	8.1	9.4	1.15	0.59	4.0	0.7	0.15	0.23
OLGECs244	3814409	3814570	162	12.3	18.5	1.62	0.70	47.5	7.4	0.21	0.70	21.5	1.3	0.09	0.15
OLGECs245	3823114	3823275	162	44.2	41.4	0.99	0.67	174.9	15.6	0.09	0.73	217.2	5.0	0.03	0.67
OLGECs246	3827634	3827777	144	293.5	503.8	1.74	1.00	268.5	168.0	0.66	1.00	148.8	35.4	0.24	0.74
OLGECs247	3897967	3898101	135	12.7	27.5	2.83	0.54	12.4	2.1	0.20	0.50	6.6	2.0	0.18	0.27
OLGECs248	3929126	3929344	219	4.9	26.2	6.01	0.75	32.0	8.1	0.26	0.63	15.2	0.3	0.02	0.27
OLGECs249	3931359	3931460	102	8.2	16.7	2.11	0.47	25.1	5.8	0.26	0.50	26.1	1.7	0.10	0.36
OLGECs250	3934494	3934688	195	259.6	94.6	0.40	0.87	309.0	38.1	0.13	0.78	148.8	5.1	0.04	0.59
OLGECs251	3934511	3934654	144	76.1	100.3	1.33	0.85	194.5	31.2	0.16	0.71	95.1	3.7	0.04	0.75
OLGECs252	3942489	3942668	180	3.6	8.6	3.18	0.42	20.0	5.7	0.28	0.55	16.0	2.3	0.24	0.41
OLGECs253	3948722	3948928	207	22.7	63.4	3.22	0.80	43.2	15.6	0.41	0.76	32.6	2.6	0.07	0.34

OLGECs254	3948730	3948894	165	22.0	74.2	3.84	0.81	48.2	16.6	0.38	0.82	37.8	2.9	0.07	0.48
OLGECs255	3973089	3973310	222	2.7	182.2	61.82	0.64	17.9	60.4	3.44	0.59	6.7	6.4	0.96	0.40
OLGECs256	4007663	4007791	129	35.0	4.1	0.14	0.46	74.4	17.9	0.24	0.53	53.9	0.9	0.02	0.15
OLGECs257	4018707	4018868	162	2.9	3.0	0.33	0.27	7.7	1.8	0.24	0.21	3.2	4.8	1.06	0.28
OLGECs258	4029254	4029361	108	15.5	16.3	1.09	0.57	57.4	7.5	0.13	0.62	51.5	4.4	0.09	0.31
OLGECs259	4035910	4036095	186	52.4	119.4	2.28	0.74	102.8	50.4	0.48	0.70	74.9	6.4	0.08	0.28
OLGECs260	4036868	4036999	132	63.5	10.8	0.15	0.51	102.1	21.4	0.22	0.64	72.6	4.1	0.07	0.14
OLGECs261	4039495	4039932	438	6.5	16.9	3.51	0.51	20.2	4.2	0.22	0.57	7.3	1.0	0.19	0.19
OLGECs262	4042550	4042753	204	77.4	297.3	3.80	0.86	61.9	16.6	0.30	0.67	35.6	4.9	0.14	0.75
OLGECs263	4062372	4062479	108	3.9	7.7	2.05	0.44	3.4	9.1	0.92	0.64	6.8	0.5	0.00	0.17
OLGECs264	4070264	4070404	141	76.4	31.5	0.41	0.66	669.5	49.7	0.08	0.95	393.9	3.1	0.01	0.53
OLGECs265	4072250	4072525	276	4.4	18.0	4.26	0.41	37.0	8.0	0.22	0.63	22.0	2.2	0.10	0.26
OLGECs266	4072681	4072776	96	8.7	74.1	9.04	0.85	86.9	21.4	0.26	0.85	47.9	2.4	0.06	0.61
OLGECs267	4072706	4072831	126	6.6	42.7	6.66	0.75	86.7	12.9	0.16	0.75	45.9	2.9	0.06	0.37
OLGECs268	4077302	4077397	96	10.1	36.4	4.08	0.71	38.7	24.2	0.67	0.73	62.5	5.4	0.09	0.46
OLGECs269	4087759	4087887	129	4.7	10.8	1.89	0.34	17.8	3.5	0.12	0.43	27.1	1.1	0.02	0.27
OLGECs270	4091628	4091960	333	1.4	7.3	0.92	0.35	29.4	11.3	0.43	0.61	11.3	1.1	0.25	0.40
OLGECs271	4098342	4098443	102	6.4	6.4	1.46	0.31	8.0	4.4	0.82	0.52	5.1	3.7	0.51	0.30
OLGECs272	4098999	4099094	96	8.2	1.9	0.17	0.21	5.0	7.8	0.74	0.48	27.1	1.1	0.02	0.28
OLGECs273	4119394	4119507	114	33.1	15.2	1.14	0.43	24.4	1.3	0.03	0.22	141.5	2.4	0.01	0.28
OLGECs274	4144786	4144926	141	12.6	7.9	1.14	0.45	39.4	6.2	0.19	0.70	11.1	1.0	0.06	0.20
OLGECs275	4149206	4149328	123	10.2	7.3	0.73	0.46	14.8	9.2	0.46	0.55	12.1	0.7	0.03	0.28
OLGECs276	4157946	4158038	93	15.5	2.0	0.13	0.31	7.9	1.1	0.02	0.37	32.8	2.7	0.26	0.26
OLGECs277	4173482	4173589	108	15.5	16.3	1.14	0.66	123.4	7.0	0.09	0.83	418.8	17.6	0.14	0.53
OLGECs278	4197561	4197677	117	15.0	45.2	3.43	0.70	31.2	14.9	0.49	0.65	35.5	13.6	0.40	0.46
OLGECs279	4198232	4198375	144	3.2	3.2	0.20	0.39	25.5	4.8	0.19	0.58	5.7	0.2	0.04	0.16
OLGECs280	4208686	4208787	102	24.1	16.3	0.69	0.79	14.1	5.3	0.36	0.60	21.8	2.0	0.11	0.49
OLGECs281	4227554	4227715	162	14.6	3.4	0.38	0.12	21.7	5.8	0.21	0.48	33.0	4.2	0.08	0.19
OLGECs282	4237499	4237606	108	24.5	13.2	0.49	0.56	15.4	4.4	0.44	0.49	32.3	6.4	0.22	0.55
OLGECs283	4237905	4238003	99	9.8	35.7	2.39	0.29	0.9	1.5	0.19	0.37	12.8	0.8	0.04	0.43
OLGECs284	4248614	4248850	237	71.0	19.0	0.28	0.70	154.6	12.0	0.08	0.62	45.6	4.4	0.11	0.20
OLGECs285	4255986	4256171	186	36.5	18.1	0.57	0.49	35.2	29.6	0.65	0.43	107.9	28.9	0.18	0.73
OLGECs286	4263989	4264081	93	2.5	67.1	6.98	0.40	20.3	16.9	0.84	0.70	1.6	0.0	0.00	0.00
OLGECs287	4267952	4268107	156	19.6	40.6	2.08	0.44	46.2	26.1	0.57	0.69	42.9	7.9	0.10	0.18
OLGECs288	4279212	4279340	129	28.4	11.3	0.39	0.34	20.9	6.7	0.38	0.65	10.9	0.2	0.01	0.07
OLGECs289	4284874	4285020	147	3.2	0.0	0.00	0.00	8.8	1.7	0.28	0.33	2.5	1.7	0.33	0.33
OLGECs290	4309564	4309659	96	0.0	4.5	0.00	0.14	1.9	2.3	0.00	0.34	0.0	4.0	0.00	0.35

OLGECs291	4320625	4320735	111	39.8	61.2	1.53	0.72	36.4	21.4	0.58	0.81	16.7	2.4	0.14	0.26
OLGECs292	4327793	4327987	195	19.7	15.1	0.71	0.55	42.4	4.4	0.13	0.60	32.6	4.4	0.16	0.60
OLGECs293	4329768	4329935	168	16.0	29.1	1.85	0.67	35.1	3.8	0.11	0.53	56.1	2.0	0.04	0.70
OLGECs294	4344194	4344355	162	4.9	5.7	1.52	0.30	17.5	6.1	0.55	0.41	13.8	0.9	0.03	0.26
OLGECs295	4347243	4347395	153	8.2	8.7	1.11	0.24	18.6	7.5	0.51	0.67	4.8	2.0	0.08	0.45
OLGECs296	4348718	4348822	105	11.5	11.8	1.10	0.50	44.8	12.7	0.29	0.55	26.7	1.7	0.00	0.51
OLGECs297	4401072	4401362	291	1247.1	405.0	0.34	0.91	71.2	34.9	0.67	0.62	72.8	0.8	0.01	0.21
OLGECs298	4401131	4401268	138	2576.0	704.3	0.28	0.97	144.5	51.2	0.50	0.66	149.8	1.2	0.01	0.23
OLGECs299	4404092	4404256	165	2.8	20.2	1.19	0.53	11.2	2.5	0.26	0.60	6.7	0.3	0.07	0.13
OLGECs300	4408310	4408897	588	4.9	16.3	5.07	0.43	21.3	10.9	0.49	0.57	14.1	0.6	0.04	0.15
OLGECs301	4465224	4465367	144	1.3	5.8	1.76	0.32	8.2	1.4	0.14	0.16	7.7	3.3	0.20	0.26
OLGECs302	4465651	4465752	102	1.8	10.9	1.51	0.54	22.1	6.2	0.27	0.62	16.0	1.0	0.10	0.18
OLGECs303	4483573	4483701	129	25.5	7.9	0.31	0.23	9.9	2.2	0.20	0.40	10.3	1.5	0.13	0.48
OLGECs304	4485351	4485518	168	7.2	20.5	2.33	0.25	26.0	3.0	0.11	0.33	4.0	3.2	0.73	0.30
OLGECs305	4486780	4487052	273	14.5	23.3	1.62	0.45	29.9	13.1	0.43	0.57	9.5	0.5	0.06	0.21
OLGECs306	4528334	4528450	117	15.9	20.6	1.30	0.68	30.4	3.3	0.11	0.36	133.8	4.6	0.04	0.57
OLGECs307	4543439	4543594	156	0.0	1.4	0.00	0.23	4.5	1.2	0.22	0.31	5.2	3.1	0.58	0.51
OLGECs308	4550443	4550574	132	14.4	34.8	2.46	0.59	65.8	9.3	0.15	0.54	20.2	0.3	0.01	0.00
OLGECs309	4563317	4563493	177	49.5	30.5	0.61	0.73	35.0	12.3	0.35	0.68	25.1	0.8	0.03	0.23
OLGECs310	4563366	4563704	339	18.6	17.3	0.90	0.43	18.7	5.9	0.31	0.36	9.2	0.6	0.12	0.31
OLGECs311	4624979	4625080	102	12.3	13.0	1.14	0.63	14.6	6.1	0.44	0.78	1.4	1.3	0.00	0.18
OLGECs312	4625396	4625515	120	8.5	13.7	1.55	0.51	9.4	3.0	0.32	0.43	9.3	0.0	0.00	0.00
OLGECs313	4643516	4643653	138	221.6	53.9	0.26	0.86	359.9	38.2	0.11	0.84	189.2	8.6	0.10	0.44
OLGECs314	4645707	4645994	288	42.9	113.8	2.66	0.55	107.2	36.7	0.48	0.65	55.7	7.4	0.23	0.17
OLGECs315	4645711	4645842	132	76.7	237.2	3.08	0.68	223.1	70.4	0.45	0.70	109.1	16.0	0.24	0.23
OLGECs316	4658938	4659159	222	37.2	23.0	0.62	0.78	38.1	5.8	0.15	0.67	40.0	0.9	0.02	0.40
OLGECs317	4716572	4716712	141	6.9	9.2	1.15	0.27	17.9	7.7	0.50	0.48	59.6	9.6	0.09	0.75
OLGECs318	4716582	4716737	156	6.2	11.3	1.74	0.34	16.2	7.7	0.54	0.49	55.3	8.9	0.11	0.32
OLGECs319	4721524	4721892	369	16.1	29.1	1.80	0.56	41.8	30.0	1.10	0.76	9.9	0.7	0.09	0.04
OLGECs320	4740766	4740870	105	245.5	24.8	0.15	0.61	247.0	11.2	0.05	0.82	210.1	9.5	0.05	0.41
OLGECs321	4743195	4743302	108	5.6	7.7	1.28	0.44	17.8	9.5	0.09	0.70	21.9	6.5	0.41	0.41
OLGECs322	4750106	4750336	231	9.3	2.5	0.30	0.26	21.3	8.3	0.34	0.41	27.6	1.7	0.05	0.21
OLGECs323	4759164	4759553	390	11.2	17.7	1.60	0.50	44.3	8.5	0.26	0.57	38.6	0.5	0.01	0.22
OLGECs324	4780249	4780431	183	103.1	56.9	0.58	0.68	475.0	23.8	0.06	0.89	304.0	10.5	0.03	0.51
OLGECs325	4894219	4894344	126	7.7	12.8	1.54	0.43	14.0	2.4	0.18	0.47	8.2	0.9	0.07	0.12
OLGECs326	4905743	4905925	183	51.5	16.0	0.31	0.75	41.8	6.7	0.16	0.67	25.5	0.5	0.02	0.07
OLGECs327	4908629	4908793	165	39.6	67.7	1.74	0.54	74.6	50.0	0.65	0.62	42.7	1.2	0.03	0.37

OLGECs328	4919092	4919286	195	22.4	8.9	0.42	0.33	39.9	16.8	0.29	0.38	36.2	15.4	0.25	0.19
OLGECs329	4920564	4920710	147	0.0	3.8	0.00	0.19	3.9	6.2	1.10	0.33	2.5	0.4	0.07	0.20
OLGECs330	4934190	4934288	99	5.6	22.8	1.43	0.45	14.7	12.3	0.83	0.70	6.7	1.1	0.15	0.12
OLGECs331	4935553	4935648	96	8.7	26.7	3.14	0.56	8.6	3.4	0.13	0.54	5.4	0.3	0.04	0.03
OLGECs332	4964727	4964870	144	1.3	23.5	8.82	0.53	4.6	1.4	0.34	0.42	3.6	0.5	0.11	0.14
OLGECs333	4965735	4965875	141	4.3	2.6	0.65	0.27	3.0	1.2	0.39	0.36	11.6	3.7	0.30	0.56
OLGECs334	4971395	4971694	300	190.0	64.5	0.34	0.76	272.9	11.1	0.04	0.91	198.1	1.9	0.01	0.70
OLGECs335	4972411	4972599	189	13.0	10.2	0.76	0.34	31.1	9.4	0.29	0.52	115.6	10.2	0.06	0.23
OLGECs336	5004272	5004469	198	17.1	5.2	0.23	0.10	11.9	9.7	0.53	0.32	41.3	9.8	0.11	0.20
OLGECs337	5021969	5022232	264	24.2	110.0	6.24	0.73	93.1	54.5	0.59	0.79	168.3	4.2	0.03	0.43
OLGECs338	5035514	5035921	408	6.1	4.7	2.33	0.21	11.2	6.0	0.47	0.38	10.0	2.4	0.11	0.05
OLGECs339	5035776	5035913	138	16.2	11.7	2.05	0.40	30.7	16.0	0.42	0.61	24.2	6.9	0.14	0.14
OLGECs340	5102903	5103073	171	8.7	13.9	1.72	0.71	18.6	7.7	0.55	0.81	11.3	1.0	0.16	0.34
OLGECs341	5112395	5112619	225	8.7	6.6	1.04	0.16	26.3	6.7	0.28	0.62	12.5	1.5	0.17	0.22
OLGECs342	5116284	5116508	225	430.1	99.3	0.22	0.77	599.5	14.1	0.03	0.65	561.4	6.2	0.01	0.34
OLGECs343	5118499	5118720	222	11.3	4.0	0.35	0.42	40.4	5.1	0.23	0.56	2.3	0.7	0.34	0.11
OLGECs344	5128803	5128925	123	29.6	2.0	0.26	0.24	28.7	9.8	0.33	0.45	10.9	0.1	0.01	0.00
OLGECs345	5149271	5149423	153	8.8	12.0	4.82	0.22	21.6	8.9	0.36	0.41	34.5	4.7	0.11	0.30
OLGECs346	5166369	5166740	372	4.4	25.3	5.50	0.51	21.6	11.5	0.74	0.60	6.8	0.6	0.08	0.31
OLGECs347	5166927	5167100	174	60.8	59.7	0.99	0.77	331.9	19.6	0.06	0.79	69.8	1.6	0.03	0.18
OLGECs348	5186632	5186736	105	35.4	18.5	0.53	0.70	65.1	10.0	0.15	0.65	14.1	1.4	0.04	0.47
OLGECs349	5186636	5186755	120	25.1	17.3	0.70	0.59	65.1	9.3	0.14	0.63	18.5	1.9	0.09	0.06
OLGECs350	5192012	5192113	102	11.8	3.6	0.25	0.20	14.0	7.9	0.48	0.42	37.2	8.8	0.13	0.26
OLGECs351	5203944	5204132	189	9.8	17.2	1.75	0.66	23.6	7.2	0.30	0.71	22.0	1.4	0.05	0.11
OLGECs352	5203973	5204065	93	11.0	20.3	1.85	0.75	27.4	7.9	0.30	0.82	31.9	2.3	0.05	0.10
OLGECs353	5215818	5215913	96	77.5	13.5	0.21	0.51	61.6	3.1	0.05	0.46	18.6	0.6	0.02	0.27
OLGECs354	5215970	5216092	123	29.5	193.9	6.83	0.70	49.4	23.6	0.48	0.60	28.3	0.2	0.01	0.07
OLGECs355	5220629	5220733	105	3.5	7.4	0.63	0.70	13.9	6.5	0.47	0.62	3.5	3.9	0.55	0.20
OLGECs356	5227712	5227963	252	57.3	32.9	0.57	0.63	133.4	6.8	0.06	0.65	151.8	1.1	0.01	0.24
OLGECs357	5231437	5231682	246	40.8	63.9	1.57	0.64	142.5	22.5	0.16	0.74	384.5	37.9	0.11	0.70
OLGECs358	5254034	5254300	267	14.8	15.2	1.90	0.32	15.2	7.3	0.43	0.37	7.5	1.1	0.15	0.12
OLGECs359	5271232	5271360	129	9.3	19.4	1.99	0.74	22.0	5.7	0.28	0.74	20.1	2.5	0.14	0.30
OLGECs360	5280370	5280483	114	11.8	24.4	2.82	0.75	306.5	26.4	0.09	0.91	122.9	1.4	0.01	0.46
OLGECs361	5280487	5280621	135	14.8	26.3	2.02	0.79	201.5	24.6	0.12	0.74	103.8	0.5	0.00	0.28
OLGECs362	5304431	5305123	693	6.8	16.1	2.34	0.38	31.8	8.6	0.31	0.63	15.6	1.1	0.07	0.43
OLGECs363	5312328	5312639	312	48.5	33.3	1.20	0.69	92.2	17.7	0.19	0.81	65.2	10.5	0.17	0.37
OLGECs364	5312751	5313092	342	21.4	9.4	0.42	0.20	29.8	13.5	0.50	0.48	47.2	6.1	0.09	0.40

OLGECs365	5329164	5329259	96	4.9	16.4	0.99	0.45	8.5	6.2	2.62	0.58	9.3	2.7	0.20	0.37
OLGECs366	5334211	5334417	207	28.3	16.1	0.61	0.50	48.9	6.3	0.13	0.69	14.3	1.5	0.10	0.26
OLGECs367	5339346	5339528	183	26.9	84.2	3.08	0.60	67.3	50.7	0.73	0.63	19.8	1.3	0.08	0.22
OLGECs368	5418482	5418592	111	7.5	20.0	2.70	0.57	11.8	10.2	0.79	0.60	7.3	0.1	0.01	0.00
OLGECs369	5418486	5418578	93	9.0	23.9	2.70	0.64	11.1	12.2	1.06	0.60	7.2	0.1	0.02	0.10
OLGECs370	5427029	5427295	267	1.6	10.4	7.02	0.35	11.4	66.5	7.36	0.55	5.8	0.8	0.21	0.11
OLGECs371	5427198	5427497	300	0.8	9.9	1.98	0.32	10.7	59.0	6.43	0.48	5.9	1.5	0.30	0.18
OLGECs372	5451976	5452137	162	147.2	41.9	0.28	0.56	144.8	12.2	0.10	0.67	62.7	1.9	0.03	0.09
OLGECs373	5452273	5452386	114	11.0	12.5	1.16	0.50	16.8	9.9	0.58	0.59	14.3	1.5	0.11	0.25
OLGECs374	5464147	5464245	99	2.4	26.3	1.58	0.47	21.8	22.8	0.99	0.60	21.0	0.0	0.00	0.16
OLGECs375	5485907	5486020	114	4.1	7.1	0.26	0.60	11.7	3.6	0.42	0.52	4.6	1.8	0.29	0.52
OLGECs376	5489042	5489143	102	8.2	8.2	1.08	0.57	18.8	2.7	0.14	0.50	9.4	3.0	0.27	0.36
OLGECs377	5490342	5490506	165	76.1	67.9	0.89	0.45	111.7	18.7	0.19	0.35	91.1	2.3	0.03	0.24
OLGECs378	5490361	5490564	204	67.4	63.8	0.93	0.45	94.9	18.1	0.23	0.47	79.1	3.1	0.05	0.25
OLGECs379	5491275	5491502	228	16.5	66.6	4.04	0.89	60.6	11.5	0.19	0.86	96.8	6.8	0.07	0.69
OLGECs380	5491693	5491800	108	13.3	18.0	1.35	0.57	41.5	2.1	0.04	0.34	29.5	2.5	0.07	0.54

*The mean value of the two biological replicates is shown.

Supplementary Table S3: Properties of the novel OLGs. The table shows from left to right the annotated homolog with the lowest e-value, position and strength of the predicted σ^{70} promoter, position and ΔG° of the predicted Shine-Dalgarno sequence, Position and binding energy of the predicted ρ-independent terminators and comparison to the RNAseq and RIBOseq results of Landstorfer (2014).

	BLASTP		promoter		Shine-Dalgarno Sequence		terminator		
gene name	**e-value**	**organism**	**bps upstream start codon**	**LDF score**	**bps upstream start codon**	**ΔG°**	**bps downstream stop codon**	**score**	**expression in EDL933**
OLGECs001	-	-	140	3.8	-	-	-	-	-
OLGECs002	-	-	28	2.43	-	-	-	-	-
OLGECs003	-	-	208	1.64	1	-7.6	-	-	-
OLGECs004	3E-20	*Ceratitis capitata*	78	0.74	7	-6.5	-	-	-
OLGECs005	2E-60	*Escherichia coli* 8.0586	95	1.69	15	-4.2	-	-	RIBOseq
OLGECs006	6E-11	*Escherichia coli* P0298942.8	216	1.66	2	-2.9	-	-	-
OLGECs007	5E-23	*Escherichia coli* DEC8C	162	5.67	10	-4.2	-	-	-
OLGECs008	-	-	216	2.06	4	-5.7	-	-	-
OLGECs009	-	-	175	1.1	-	-	-	-	-
OLGECs010	-	-	-	-	9	-5.4	-	-	-
OLGECs011	-	-	182	1.7	5	-3.6	-	-	-
OLGECs012	-	-	99	0.71	2	-3.5	-	-	-
OLGECs013	1E-39	*Escherichia coli* 83972	201	0.39	15	-5.9	-	-	RNAseq
OLGECs014	-	-	206	2.35	-	-	-	-	-
OLGECs015	-	-	81	0.93	-	-	-	-	-
OLGECs016	5E-19	*Escherichia coli* O157:H7 str. G5101	54	2.71	-	-	-	-	-
OLGECs017	6E-12	*Escherichia coli* E110019	142	2.56	13	-2.9	-	-	RIBOseq
OLGECs018	4E-14	*Cedecea davisae* DSM 4568	96	3.14	4	-3.1	-	-	RNAseq
OLGECs019	-	-	197	5.23	-	-	-	-	RIBOseq
OLGECs020	-	-	243	5.23	4	-3.6	-	-	RIBOseq
OLGECs021	-	-	139	4.58	-	-	-	-	-
OLGECs022	-	-	224	4.58	-	-	-	-	-
OLGECs023	-	-	193	1.46	-	-	-	-	RNAseq
OLGECs024	-	-	44	3.53	-	-	-	-	-
OLGECs025	-	-	118	5.62	-	-	-	-	-
OLGECs026	-	-	95	2.45	-	-	-	-	RNAseq
OLGECs027	-	-	38	3.79	-	-	49	-12.4	-
OLGECs028	-	-	166	3.86	-	-	-	-	-
OLGECs029	3E-04	*Pseudomonas fuscovaginae*	194	1.94	-	-	-	-	-
OLGECs030	-	-	31	1.12	-	-	-	-	-
OLGECs031	-	-	194	1.09	-	-	-	-	-
OLGECs032	-	-	108	3.3	15	-4.6	-	-	-
OLGECs033	-	-	206	1.76	-	-	-	-	-
OLGECs034	-	-	122	2.39	0	-5.1	-	-	-
OLGECs035	-	-	135	2.39	13	-5.1	-	-	-
OLGECs036	-	-	149	2.84	-	-	-	-	RIBOseq
OLGECs037	9E-31	*Escherichia coli* TA007	-	-	21	-5.4	-	-	-
OLGECs038	-	-	222	3.57	-	-	-	-	-

OLGECs039	-	-	228	2.61	-	-	-	-	-
OLGECs040	-	-	30	3.03	-	-	-	-	-
OLGECs041	-	-	242	2.96	-	-	-	-	-
OLGECs042	-	-	73	2.22	14	-3.4	-	-	-
OLGECs043	-	-	74	2.22	15	-3.4	-	-	RNAseq
OLGECs044	-	-	48	2.22	13	-5.2	-	-	RNAseq
OLGECs045	-	-	134	1.95	3	-3.7	-	-	-
OLGECs046	-	-	177	9.53	8	-3.6	-	-	RNAseq
OLGECs047	-	-	123	5.56	-	-	10	-20.4	-
OLGECs048	-	-	130	1.96	2	-5	-	-	-
OLGECs049	-	-	181	3.22	-	-	-	-	-
OLGECs050	-	-	174	1.86	8	-4.1	-	-	-
OLGECs051	-	-	224	3.67	-	-	157	-15.2	RIBOseq
OLGECs052	1E-50	*Escherichia coli* 908675	89	5.5	-	-	-	-	-
OLGECs053	-	-	227	0.64	20	-3.7	-	-	-
OLGECs054	-	-	244	2.54	-	-	-	-	-
OLGECs055	-	-	163	2.93	-	-	-	-	-
OLGECs056	-	-	215	0.48	11	-3.6	137	-14.2	-
OLGECs057	-	-	195	2.93	-	-	-	-	-
OLGECs058	1E-04	*Limnobacter* sp.	191	2.43	-	-	-	-	-
OLGECs059	-	-	84	1.7	-	-	75	-14.2	-
OLGECs060	3E-05	*Shigella sonnei* 53G	228	3.2	-	-	-	-	-
OLGECs061	-	-	201	5.29	-	-	-	-	-
OLGECs062	-	-	60	1.13	-	-	-	-	RIBOseq
OLGECs063	-	-	57	2.74	-	-	-	-	RNAseq
OLGECs064	2E-21	*Escherichia coli* 900105	113	2.74	12	-6.9	-	-	RNAseq
OLGECs065	2E-08	*Oceanicola batsensis* HTCC2597	109	0.89	8	-2.9	-	-	RNAseq
OLGECs066	2E-170	*Escherichia coli* chi7122	76	5.98	12	-4	-	-	-
OLGECs067	-	-	41	5.98	-	-	-	-	-
OLGECs068	-	-	36	4.75	-	-	96	-21.9	-
OLGECs069	-	-	91	6.66	14	-4	-	-	-
OLGECs070	-	-	196	7.15	-	-	-	-	-
OLGECs071	-	-	106	2.66	12	-4.8	-	-	-
OLGECs072	-	-	147	2.87	-	-	-	-	-
OLGECs073	3E-09	*Escherichia coli* 907710	155	1.74	-	-	-	-	-
OLGECs074	-	-	87	3.42	-	-	-	-	-
OLGECs075	-	-	169	3.42	20	-3.6	-	-	-
OLGECs076	-	-	38	3.06	-	-	12	-12	-
OLGECs077	-	-	30	1.44	-	-	-	-	-
OLGECs078	-	-	120	2.55	9	-6.2	-	-	RIBOseq
OLGECs079	-	-	48	1.81	-	-	-	-	-
OLGECs080	-	-	114	1.76	-	-	238	-18.5	-
OLGECs081	-	-	106	3.47	-	-	-	-	-
OLGECs082	-	-	80	2.4	2	-3.3	-	-	-

OLGECs083	-	-	200	4.9	-	-	-	-	-
OLGECs084	-	-	240	4.9	6	-4.2	-	-	-
OLGECs085	-	-	44	2.92	-	-	-	-	RIBOseq
OLGECs086	-	-	250	1.69	-	-	-	-	-
OLGECs087	-	-	71	2.99	-	-	-	-	-
OLGECs088	4E-17	*Enterobacter cloacae* EcWSU1	146	2.72	-	-	-	-	RIBOseq
OLGECs089	-	-	-	-	16	-3.1	-	-	-
OLGECs090	-	-	-	-	-	-	-	-	-
OLGECs091	-	-	145	2.4	-	-	-	-	-
OLGECs092	-	-	76	3.84	-	-	-	-	-
OLGECs093	-	-	128	1.42	1	-4.6	252	-13.2	RNAseq
OLGECs094	-	-	31	0.74	8	-3	55	-13.2	-
OLGECs095	-	-	178	4.12	-	-	-	-	RIBOseq
OLGECs096	-	-	53	5.48	10	-8.2	-	-	-
OLGECs097	-	-	44	1.67	21	-4.4	-	-	-
OLGECs098	-	-	101	3.19	0	-5.6	-	-	-
OLGECs099	3E-35	*Escherichia coli* MS 115-1	120	3.19	19	-5.6	-	-	-
OLGECs100	-	-	83	1.43	-	-	-	-	-
OLGECs101	-	-	55	2.11	-	-	-	-	-
OLGECs102	-	-	207	2.72	1	-3.5	-	-	-
OLGECs103	-	-	230	0.79	-	-	-	-	-
OLGECs104	-	-	136	2.53	1	-3.7	-	-	-
OLGECs105	-	-	101	3.2	-	-	-	-	RNAseq
OLGECs106	-	-	229	3.94	6	-3.3	-	-	-
OLGECs107	-	-	234	2.73	-	-	-	-	-
OLGECs108	-	-	56	1.6	-	-	-	-	RNAseq
OLGECs109	7E-35	*Escherichia coli* chi7122	109	2.14	18	-5.3	127	-13.8	-
OLGECs110	-	-	237	2.84	10	-3.6	-	-	-
OLGECs111	-	-	94	5.76	2	-3.4	-	-	-
OLGECs112	2E-12	*Salmonella enterica* subsp. enterica serovar Minnesota str. A4-603	153	3.34	-	-	-	-	-
OLGECs113	-	-	126	1.72	10	-3.2	-	-	-
OLGECs114	-	-	61	1.7	-	-	-	-	-
OLGECs115	-	-	202	2.6	-	-	122	-13.3	-
OLGECs116	-	-	84	3.49	-	-	-	-	RIBOseq
OLGECs117	-	-	167	2.77	3	-4.6	-	-	-
OLGECs118	-	-	61	0.86	-	-	-	-	-
OLGECs119	9E-14	*Escherichia coli* 96.0497	127	1.33	-	-	-	-	-
OLGECs120	-	-	203	3.09	7	-4.2	121	-21.5	-
OLGECs121	-	-	150	2.42	-	-	28	-20.3	-
OLGECs122	-	-	229	1.98	-	-	-	-	-
OLGECs123	7E-29	*Escherichia coli* UMEA 3065-1	122	4.48	8	-6.3	-	-	RIBOseq

OLGECs124	3E-29	*Escherichia coli*	160	0.59	8	-7.2	-	-	-
OLGECs125	-	-	102	1.21	-	-	-	-	-
OLGECs126	-	-	74	4.92	-	-	-	-	RIBOseq
OLGECs127	-	-	233	2.87	-	-	-	-	-
OLGECs128	-	-	48	3.18	-	-	-	-	-
OLGECs129	3E-12	*Salmonella enterica* subsp. enterica serovar Choleraesuis str. ATCC 10708	107	0.44	17	-3.8	-	-	-
OLGECs130	-	-	116	2.61	15	-4.9	-	-	-
OLGECs131	2E-07	*Erwinia tracheiphila* PSU-1	216	2.51	-	-	245	-13.8	-
OLGECs132	-	-	194	3.75	-	-	-	-	RIBOseq
OLGECs133	-	-	189	2.18	-	-	-	-	-
OLGECs134	-	-	98	4.02	4	-4.2	122	-16.3	-
OLGECs135	-	-	189	1.98	9	-3.3	-	-	RNAseq
OLGECs136	2E-25	*Escherichia coli* 8.0586	154	0.68	-	-	-	-	RNAseq
OLGECs137	-	-	85	2.06	-	-	-	-	-
OLGECs138	-	-	188	3.28	-	-	-	-	RIBOseq
OLGECs139	-	-	66	1.13	-	-	-	-	-
OLGECs140	-	-	134	1.13	-	-	-	-	-
OLGECs141	4E-04	*Citrobacter koseri* ATCC BAA-895	-	-	-	-	-	-	RIBOseq
OLGECs142	-	-	131	1.98	7	-6.2	-	-	-
OLGECs143	-	-	77	2.2	-	-	-	-	-
OLGECs144	-	-	111	2.79	-	-	-	-	RNAseq
OLGECs145	-	-	79	2.46	-	-	78	-15.5	-
OLGECs146	-	-	81	5.27	-	-	-	-	-
OLGECs147	3E-06	*Escherichia coli* 908658	55	1.14	-	-	-	-	RNAseq
OLGECs148	-	-	244	1.15	-	-	-	-	-
OLGECs149	-	-	41	1.56	4	-5	95	-13.5	-
OLGECs150	-	-	130	2.2	-	-	-	-	-
OLGECs151	-	-	170	1.5	-	-	-	-	-
OLGECs152	-	-	188	2.88	9	-3.8	273	-13.1	-
OLGECs153	-	-	235	5.6	12	-5.4	-	-	-
OLGECs154	1E-25	*Escherichia coli* 2749250	239	5.6	16	-5.2	-	-	RNAseq
OLGECs155	-	-	117	3.04	8	-4.8	-	-	-
OLGECs156	-	-	73	0.92	-	-	18	-18.6	-
OLGECs157	-	-	26	1.2	-	-	242	-18.6	-
OLGECs158	-	-	108	4.12	-	-	-	-	-
OLGECs159	-	-	75	3.57	-	-	-	-	-
OLGECs160	1E-13	*Escherichia coli* 97.1742	64	1.18	-	-	-	-	-
OLGECs161	-	-	138	3.93	-	-	-	-	-
OLGECs162	5E-15	*Escherichia coli* 576-1	159	2.99	9	-3.7	-	-	-
OLGECs163	-	-	241	5.23	-	-	-	-	-
OLGECs164	-	-	188	3	-	-	-	-	-

OLGECs165	-	-	61	3.33	1	-3.3	-	-	-
OLGECs166	-	-	60	8.08	-	-	-	-	-
OLGECs167	-	-	73	2.22	14	-3.4	-	-	-
OLGECs168	-	-	74	2.22	15	-3.4	-	-	-
OLGECs169	-	-	48	2.22	13	-5.2	-	-	-
OLGECs170	-	-	116	2.73	8	-3.9	145	-16.8	RNAseq
OLGECs171	2E-05	*Klebsiella pneumoniae* subsp. pneumoniae DSM 30104	127	3.35	10	-5.1	-	-	-
OLGECs172	-	-	112	1.67	21	-6.1	-	-	RIBOseq
OLGECs173	-	-	104	2.49	2	-3.1	-	-	-
OLGECs174	-	-	225	0.57	5	-4.6	-	-	-
OLGECs175	-	-	237	1.31	-	-	-	-	-
OLGECs176	3E-05	*Escherichia coli* PA10	145	3.29	-	-	-	-	-
OLGECs177	-	-	57	4.55	0	-3.1	-	-	-
OLGECs178	-	-	108	1.32	-	-	264	-14.6	-
OLGECs179	-	-	61	1.77	-	-	-	-	-
OLGECs180	-	-	170	3.93	5	-6.3	-	-	-
OLGECs181	-	-	28	2.04	-	-	-	-	-
OLGECs182	-	-	39	4.28	-	-	-	-	-
OLGECs183	-	-	238	5.46	-	-	-	-	-
OLGECs184	-	-	193	3.79	3	-4.8	194	-15.6	-
OLGECs185	-	-	44	0.82	-	-	-	-	-
OLGECs186	-	-	186	0.32	7	-4.2	-	-	-
OLGECs187	-	-	31	4.99	-	-	-	-	-
OLGECs188	-	-	30	3.2	-	-	-	-	RNAseq
OLGECs189	-	-	33	1.57	2	-4.3	-	-	-
OLGECs190	-	-	144	1.7	11	-4.3	-	-	-
OLGECs191	-	-	78	1.71	-	-	-	-	-
OLGECs192	-	-	167	1.71	-	-	-	-	-
OLGECs193	-	-	184	1.71	8	-3.6	-	-	-
OLGECs194	-	-	197	1.2	20	-3.4	140	-15.7	-
OLGECs195	-	-	70	0.32	2	-3.1	166	-15.5	-
OLGECs196	-	-	-	-	-	-	-	-	-
OLGECs197	-	-	180	2.95	-	-	-	-	-
OLGECs198	-	-	134	3.13	-	-	-	-	-
OLGECs199	-	-	237	3.29	-	-	-	-	-
OLGECs200	-	-	153	3.81	-	-	-	-	-
OLGECs201	-	-	234	0.95	-	-	214	-11	-
OLGECs202	-	-	163	0.55	0	-2.9	-	-	-
OLGECs203	-	-	163	5.84	13	-3.4	-	-	-
OLGECs204	-	-	213	1.66	-	-	-	-	-
OLGECs205	-	-	161	1.53	-	-	-	-	-
OLGECs206	-	-	215	0.76	-	-	-	-	RIBOseq
OLGECs207	-	-	42	2.44	5	-4.4	-	-	-

OLGECs208	-	-	246	0.45	13	-5.9	-	-	-
OLGECs209	-	-	233	1.77	-	-	-	-	-
OLGECs210	2E-18	*Escherichia coli* P0304816.10	214	1.77	9	-4	-	-	-
OLGECs211	-	-	80	2.92	10	-2.9	-	-	-
OLGECs212	-	-	71	1.12	10	-3.1	-	-	-
OLGECs213	-	-	132	1.12	14	-3.5	-	-	-
OLGECs214	-	-	220	2.08	13	-4.6	-	-	-
OLGECs215	-	-	167	1.48	6	-5.6	-	-	RNAseq
OLGECs216	-	-	71	1.14	18	-3.5	173	-11.5	-
OLGECs217	-	-	240	0.95	-	-	-	-	-
OLGECs218	-	-	209	8.27	-	-	-	-	-
OLGECs219	-	-	215	3.45	-	-	-	-	-
OLGECs220	-	-	104	1.79	-	-	-	-	-
OLGECs221	2E-12	*Salmonella enterica* subsp. enterica serovar Johannesburg str. S5-703	101	1.79	-	-	-	-	-
OLGECs222	-	-	117	5.7	-	-	-	-	-
OLGECs223	5E-06	*Klebsiella pneumoniae* RYC492	95	3.01	2	-3.1	136	-15.7	RIBOseq
OLGECs224	-	-	252	0.89	20	-4	-	-	-
OLGECs225	-	-	57	4.71	-	-	-	-	-
OLGECs226	-	-	237	3.24	-	-	-	-	-
OLGECs227	-	-	66	1.8	-	-	-	-	RIBOseq
OLGECs228	9E-32	*Escherichia coli* 2845350	58	2.48	20	-4.5	-	-	-
OLGECs229	3E-59	*Escherichia coli* UMN026	245	5.15	-	-	-	-	-
OLGECs230	5E-04	*Plautia stali* symbiont	232	1.71	21	-5.7	-	-	-
OLGECs231	-	-	233	2.92	-	-	-	-	-
OLGECs232	-	-	130	1.54	-	-	-	-	-
OLGECs233	-	-	200	2.51	16	-3.3	-	-	-
OLGECs234	1E-09	*Salmonella enterica*	235	2.32	-	-	-	-	-
OLGECs235	-	-	86	2.62	14	-4.1	156	-13.4	RIBOseq
OLGECs236	3E-45	*Escherichia coli* E24377A	27	4.83	8	-9	-	-	-
OLGECs237	-	-	69	1.25	4	-5.1	-	-	-
OLGECs238	-	-	176	1.84	17	-5.1	-	-	-
OLGECs239	-	-	242	4.08	-	-	-	-	-
OLGECs240	-	-	-	-	20	-3.6	-	-	-
OLGECs241	-	-	150	0.66	-	-	-	-	-
OLGECs242	-	-	209	1.07	-	-	-	-	-
OLGECs243	8E-16	*Aggregatibacter actinomycetemcomitans* serotype a str. A160	27	2.71	-	-	-	-	-
OLGECs244	2E-22	*Shigella flexneri* K-315	32	1.35	20	-4.7	-	-	-
OLGECs245	-	-	112	2.9	5	-3.4	-	-	-
OLGECs246	2E-25	*Escherichia coli* EPECa14	126	6.12	17	-3.6	162	-15.2	RIBOseq

OLGECs247	-	-	134	2.66	-	-	107	-14.2	-
OLGECs248	-	-	133	3.23	-	-	-	-	-
OLGECs249	4E-09	*Citrobacter koseri* ATCC BAA-895	56	1.5	16	-3.7	-	-	-
OLGECs250	8E-04	*Escherichia coli* KTE222	87	2.62	-	-	12	-17.3	-
OLGECs251	2E-16	*Escherichia coli* 908555	30	3.42	-	-	29	-17.3	-
OLGECs252	-	-	224	1.55	18	-5.9	-	-	-
OLGECs253	-	-	-	-	-	-	-	-	-
OLGECs254	-	-	-	-	-	-	-	-	-
OLGECs255	-	-	146	1.38	-	-	-	-	-
OLGECs256	-	-	209	6.58	20	-4.3	-	-	-
OLGECs257	-	-	65	1.81	12	-8.3	-	-	-
OLGECs258	-	-	50	3.25	10	-5.4	-	-	-
OLGECs259	-	-	175	3.2	-	-	-	-	RIBOseq
OLGECs260	-	-	130	1.76	5	-3	-	-	-
OLGECs261	3E-04	*Shigella sonnei* str. Moseley	159	0.47	-	-	-	-	RIBOseq
OLGECs262	2E-07	*Edwardsiella tarda* ATCC 23685	152	0.56	-	-	-	-	RIBOseq
OLGECs263	-	-	199	2.33	7	-6.7	-	-	-
OLGECs264	2E-04	*Klebsiella oxytoca* E718	182	4.98	-	-	-	-	-
OLGECs265	-	-	42	0.41	19	-3.4	-	-	-
OLGECs266	-	-	90	4.83	-	-	-	-	-
OLGECs267	-	-	35	4.83	-	-	-	-	-
OLGECs268	-	-	91	2.68	-	-	-	-	-
OLGECs269	-	-	131	1.93	-	-	-	-	-
OLGECs270	-	-	28	1.6	13	-3	-	-	-
OLGECs271	-	-	96	2.66	-	-	220	-14.9	-
OLGECs272	-	-	151	1.75	5	-3.7	-	-	-
OLGECs273	-	-	163	2.56	1	-3.2	-	-	-
OLGECs274	-	-	177	5.85	-	-	-	-	-
OLGECs275	-	-	156	3.48	-	-	-	-	-
OLGECs276	-	-	64	2.19	9	-3.8	-	-	-
OLGECs277	-	-	187	2.09	8	-3	-	-	-
OLGECs278	-	-	166	2.33	-	-	-	-	-
OLGECs279	-	-	198	1.09	-	-	-	-	-
OLGECs280	-	-	246	0.77	-	-	-	-	-
OLGECs281	-	-	191	4.48	11	-4.2	246	-16.3	-
OLGECs282	-	-	116	2.92	-	-	197	-15.5	-
OLGECs283	-	-	131	3.98	-	-	-	-	-
OLGECs284	4E-18	*Cronobacter malonaticus* 681	142	0.79	8	-3.4	-	-	RIBOseq
OLGECs285	-	-	63	1.4	-	-	113	-14.6	-
OLGECs286	-	-	177	1.25	20	-3.9	-	-	-
OLGECs287	-	-	96	1.92	20	-3.7	-	-	-

OLGECs288	-	-	189	3.04	10	-3.3	-	-	-
OLGECs289	-	-	157	2.18	-	-	-	-	-
OLGECs290	-	-	178	1.03	11	-3.8	-	-	-
OLGECs291	-	-	159	4.14	3	-4.6	-	-	RIBOseq
OLGECs292	-	-	35	2.05	13	-7.2	-	-	-
OLGECs293	-	-	121	1.55	-	-	258	-13.3	-
OLGECs294	-	-	207	0.89	-	-	2	-14.8	-
OLGECs295	-	-	56	1.53	20	-3.5	-	-	-
OLGECs296	-	-	228	2	-	-	-	-	-
OLGECs297	2E-23	*Escherichia coli* DEC5B	204	7.56	0	-4.6	-	-	-
OLGECs298	4E-10	*Escherichia coli* 908658	241	6.5	14	-5.1	-	-	-
OLGECs299	-	-	79	4.64	5	-4.1	-	-	RIBOseq
OLGECs300	4E-74	*Escherichia coli* EC1737	213	4.97	21	-3.8	-	-	RIBOseq
OLGECs301	-	-	52	2.61	-	-	-	-	-
OLGECs302	-	-	237	1.6	-	-	-	-	-
OLGECs303	-	-	32	1.48	-	-	44	-14.1	-
OLGECs304	-	-	216	2.6	-	-	-	-	-
OLGECs305	-	-	26	4.53	13	-5.2	-	-	-
OLGECs306	-	-	64	0.87	-	-	-	-	RNAseq
OLGECs307	-	-	175	5.04	-	-	-	-	-
OLGECs308	-	-	32	3.06	5	-4.1	-	-	-
OLGECs309	-	-	175	1.75	0	-4.7	-	-	-
OLGECs310	-	-	224	1.75	6	-4.8	-	-	-
OLGECs311	-	-	153	3.86	17	-6.2	-	-	-
OLGECs312	-	-	118	4.35	11	-4.9	-	-	-
OLGECs313	4E-11	*Escherichia coli* SE11	148	3.98	-	-	174	-18	-
OLGECs314	-	-	217	0.8	-	-	-	-	-
OLGECs315	-	-	221	0.8	-	-	-	-	-
OLGECs316	-	-	195	2.24	-	-	77	-14.8	-
OLGECs317	-	-	86	8.28	18	-6.7	-	-	-
OLGECs318	-	-	96	8.28	-	-	-	-	-
OLGECs319	2E-30	*Salmonella enterica* subsp. enterica serovar Montevideo str. OH_2009072675	216	0.73	-	-	-	-	-
OLGECs320	-	-	210	5.84	-	-	-	-	-
OLGECs321	-	-	65	3.16	-	-	-	-	-
OLGECs322	-	-	116	1.82	17	-5.5	-	-	-
OLGECs323	2E-15	*Enterobacter cloacae*	27	2.96	-	-	-	-	-
OLGECs324	-	-	184	4.07	-	-	-	-	RIBOseq
OLGECs325	-	-	-	-	-	-	-	-	-
OLGECs326	-	-	75	2.78	-	-	-	-	RIBOseq
OLGECs327	2E-13	*Escherichia coli* DEC8A	56	5.91	15	-5.9	-	-	-
OLGECs328	-	-	206	2.8	16	-4.3	-	-	-
OLGECs329	-	-	217	2.85	-	-	-	-	-

OLGECs330	3E-14	*Escherichia coli* O104:H4 str. 01-09591	-	-	-	-	-	-	-
OLGECs331	-	-	123	1.29	-	-	-	-	-
OLGECs332	-	-	212	5.08	14	-3	-	-	-
OLGECs333	-	-	172	0.76	-	-	-	-	-
OLGECs334	-	-	142	3.11	12	-6.8	-	-	RIBOseq
OLGECs335	-	-	101	1.41	-	-	-	-	-
OLGECs336	-	-	44	1.79	-	-	-	-	-
OLGECs337	3E-04	*Escherichia coli* O26:H11 str. CVM10026	135	2.52	16	-5.5	-	-	-
OLGECs338	-	-	93	3.73	7	-4.8	-	-	-
OLGECs339	-	-	101	3.73	15	-4.8	-	-	-
OLGECs340	-	-	130	1.93	16	-4.8	-	-	-
OLGECs341	2E-04	*Mesorhizobium* sp. LSHC424B00	162	2.12	14	-7.6	-	-	-
OLGECs342	-	-	169	1.78	14	-5.3	-	-	-
OLGECs343	-	-	169	7.96	-	-	-	-	RNAseq
OLGECs344	-	-	43	3.42	10	-2.9	-	-	RIBOseq
OLGECs345	-	-	204	3.63	-	-	-	-	-
OLGECs346	-	-	176	1.47	-	-	-	-	-
OLGECs347	-	-	235	0.77	-	-	-	-	RIBOseq
OLGECs348	-	-	96	1.62	14	-4.1	-	-	-
OLGECs349	-	-	100	1.62	18	-4.1	-	-	RNAseq
OLGECs350	-	-	67	3.92	9	-4.5	-	-	-
OLGECs351	-	-	207	3.25	-	-	-	-	-
OLGECs352	-	-	236	3.25	-	-	-	-	-
OLGECs353	-	-	52	2.67	13	-3	-	-	-
OLGECs354	1E-12	*Escherichia albertii* TW07627	204	2.67	15	-6.8	-	-	RIBOseq
OLGECs355	-	-	128	1.34	20	-2.9	-	-	-
OLGECs356	-	-	76	4.3	3	-3.1	-	-	RNAseq
OLGECs357	-	-	178	0.95	-	-	-	-	-
OLGECs358	6E-20	*Klebsiella pneumoniae* subsp. pneumoniae DSM 30104	147	3.26	-	-	-	-	-
OLGECs359	-	-	141	2.19	16	-4.7	-	-	-
OLGECs360	-	-	86	0.91	14	-3.9	-	-	-
OLGECs361	-	-	130	1.97	-	-	-	-	-
OLGECs362	5E-12	*Streptomyces cattleya* NRRL 8057 = DSM 46488	55	2.65	11	-4.8	-	-	RIBOseq
OLGECs363	-	-	65	3.63	0	-4.7	-	-	-
OLGECs364	-	-	195	0.86	3	-4.2	-	-	-
OLGECs365	-	-	149	0.76	10	-4.2	-	-	-
OLGECs366	-	-	99	2.6	-	-	-	-	-
OLGECs367	-	-	179	2.07	15	-4.8	-	-	-

OLGECs368	-	-	233	6.55	-	-	-	-	-
OLGECs369	-	-	237	6.55	-	-	-	-	-
OLGECs370	-	-	176	0.7	2	-4.9	-	-	-
OLGECs371	-	-	168	5.65	-	-	-	-	-
OLGECs372	-	-	250	1.77	16	-4.8	-	-	RIBOseq
OLGECs373	-	-	233	2.39	3	-2.9	-	-	-
OLGECs374	-	-	129	3.63	-	-	-	-	-
OLGECs375	-	-	163	2.56	18	-4.9	-	-	-
OLGECs376	-	-	45	2.12	12	-6.7	-	-	-
OLGECs377	-	-	175	1.48	-	-	-	-	-
OLGECs378	1E-18	*Escherichia coli* DEC1D	194	1.48	1	-5	-	-	RIBOseq
OLGECs379	-	-	199	1.35	17	-3.4	-	-	-
OLGECs380	-	-	147	1.6	16	-5.4	-	-	-

Supplementary Table S4: Differentially regulated OLGs in BHI stress compared to BHI control. Significant changes are highlighted in gray.

gene name	counts transcriptome control*	counts transcriptome stress*	log fold change	p-value	FDR	counts translatome control*	counts translatome stress*	log fold change	p-value	FDR
OLGECs007	3	19	2.623	4.65E-04	0.068	8	6	-0.383	0.649	1.000
OLGECs012	2	24	3.869	2.44E-06	0.002	1	41	5.279	2.17E-10	3.01E-07
OLGECs013	5	8	0.812	0.347	1.000	4	60	4.076	1.71E-09	1.93E-06
OLGECs016	4	11	1.432	0.068	1.000	4	46	3.513	1.55E-07	7.15E-05
OLGECs017	46	30	-0.629	0.238	1.000	29	6	-2.195	0.001	0.063
OLGECs021	107	36	-1.570	0.002	0.196	71	17	-2.041	1.70E-04	0.025
OLGECs022	63	19	-1.764	0.001	0.128	48	6	-3.051	1.27E-06	4.12E-04
OLGECs024	6	1	-2.325	0.046	1.000	35	3	-3.390	1.51E-06	4.73E-04
OLGECs037	5	8	0.721	0.421	1.000	3	45	4.125	1.10E-08	8.23E-06
OLGECs052	20	12	-0.794	0.207	1.000	29	2	-3.723	9.60E-07	3.21E-04
OLGECs057	7	3	-1.191	0.199	1.000	8	0	-5.983	2.73E-04	0.037
OLGECs066	52	13	-1.976	4.72E-04	0.068	35	106	1.636	0.001	0.123
OLGECs068	12	3	-1.962	0.015	0.637	41	1	-5.959	9.64E-11	1.50E-07
OLGECs069	34	10	-1.827	0.002	0.211	69	11	-2.553	7.02E-06	0.002
OLGECs087	20	6	-1.717	0.010	0.515	14	2	-3.057	0.001	0.061
OLGECs094	29	6	-2.336	3.20E-04	0.051	6	8	0.360	0.768	1.000
OLGECs108	2	2	-0.001	1.000	1.000	10	0	-6.300	7.46E-05	0.012
OLGECs129	9	2	-2.493	0.009	0.485	18	0	-7.091	3.21E-07	1.30E-04
OLGECs131	21	8	-1.467	0.023	0.806	27	2	-3.986	7.70E-07	2.62E-04
OLGECs132	25	6	-2.007	0.002	0.199	49	3	-3.897	1.63E-08	1.07E-05
OLGECs135	9	4	-1.071	0.207	1.000	33	2	-3.860	2.88E-07	1.20E-04
OLGECs138	272	226	-0.264	0.576	1.000	318	42	-2.859	2.71E-08	1.75E-05
OLGECs146	4	22	2.434	0.001	0.077	6	8	0.597	0.470	1.000
OLGECs150	93	17	-2.452	6.60E-06	0.004	18	2	-2.989	2.06E-04	0.029
OLGECs155	286	70	-2.036	3.89E-05	0.013	110	35	-1.593	0.002	0.144
OLGECs166	32	5	-2.780	3.19E-05	0.011	49	1	-5.359	5.97E-11	9.82E-08
OLGECs176	37	23	-0.685	0.216	1.000	315	97	-1.646	0.001	0.073
OLGECs183	6	6	-0.127	1.000	1.000	8	0	-5.888	4.29E-04	0.053
OLGECs185	3	6	0.848	0.418	1.000	3	41	3.732	1.18E-07	5.82E-05
OLGECs186	4	6	0.758	0.452	1.000	3	41	3.751	1.01E-07	5.16E-05
OLGECs211	17	72	2.083	1.40E-04	0.028	30	38	0.401	0.450	1.000
OLGECs212	28	17	-0.714	0.218	1.000	27	5	-2.522	1.57E-04	0.023
OLGECs215	43	3	-4.039	1.75E-08	4.37E-05	33	3	-3.308	2.61E-06	0.001
OLGECs225	174	71	-1.296	0.008	0.460	126	16	-2.945	7.02E-08	3.92E-05
OLGECs229	261	151	-0.815	0.087	1.000	62	9	-2.694	3.68E-06	0.001
OLGECs231	93	25	-1.912	2.74E-04	0.045	48	27	-0.765	0.153	1.000
OLGECs233	3	17	2.670	0.001	0.092	4	21	2.540	0.001	0.061
OLGECs236	15	5	-1.659	0.023	0.806	15	1	-3.676	7.93E-05	0.013
OLGECs238	28	19	-0.568	0.329	1.000	8	40	2.338	8.79E-05	0.014

OLGECs239	1	7	2.662	0.014	0.628	3	24	2.951	6.27E-05	0.011
OLGECs242	7	4	-0.872	0.362	1.000	19	3	-2.766	3.22E-04	0.043
OLGECs243	5	2	-1.125	0.336	1.000	20	3	-2.585	4.65E-04	0.056
OLGECs255	7	3	-1.444	0.126	1.000	91	18	-2.251	3.01E-05	0.006
OLGECs264	167	98	-0.778	0.108	1.000	52	13	-1.969	0.001	0.061
OLGECs277	26	100	1.942	2.31E-04	0.041	6	66	3.583	1.53E-08	1.07E-05
OLGECs278	7	8	0.094	1.000	1.000	12	56	2.300	5.32E-05	0.010
OLGECs286	4	1	-2.537	0.098	1.000	11	0	-6.438	3.29E-05	0.007
OLGECs292	14	13	-0.105	0.930	1.000	6	28	2.208	0.001	0.081
OLGECs297	35	40	0.177	0.764	1.000	71	5	-3.733	3.25E-09	3.20E-06
OLGECs298	34	39	0.182	0.759	1.000	50	4	-3.722	3.00E-08	1.90E-05
OLGECs300	23	14	-0.716	0.241	1.000	47	8	-2.545	2.41E-05	0.005
OLGECs305	15	5	-1.516	0.035	0.998	25	3	-2.921	5.07E-05	0.009
OLGECs307	2	2	-0.001	1.000	1.000	2	17	3.433	6.79E-05	0.012
OLGECs323	30	24	-0.320	0.578	1.000	23	4	-2.425	4.48E-04	0.055
OLGECs327	22	13	-0.758	0.219	1.000	60	5	-3.472	4.26E-08	2.53E-05
OLGECs333	1	3	1.473	0.334	1.000	1	22	4.289	1.52E-06	4.73E-04
OLGECs343	15	1	-3.744	5.83E-05	0.017	8	4	-1.103	0.229	1.000
OLGECs344	6	3	-1.224	0.221	1.000	10	1	-3.878	0.001	0.083
OLGECs346	14	5	-1.607	0.028	0.870	31	6	-2.403	1.94E-04	0.028
OLGECs347	104	24	-2.109	6.09E-05	0.017	25	5	-2.340	0.001	0.065
OLGECs354	11	6	-0.863	0.255	1.000	19	0	-7.230	1.26E-07	6.14E-05
OLGECs357	64	180	1.480	0.003	0.235	38	227	2.617	3.67E-07	1.45E-04
OLGECs360	64	26	-1.319	0.013	0.608	21	2	-3.215	4.46E-05	0.009
OLGECs361	49	25	-0.986	0.067	1.000	24	1	-4.352	5.55E-07	1.98E-04
OLGECs367	22	7	-1.639	0.012	0.565	68	7	-3.181	1.39E-07	6.68E-05
OLGECs370	5	3	-0.713	0.529	1.000	131	4	-4.908	3.80E-14	1.42E-10
OLGECs371	6	4	-0.634	0.560	1.000	130	6	-4.322	1.72E-12	4.15E-09
OLGECs374	4	4	-0.188	1.000	1.000	17	0	-7.009	6.12E-07	2.14E-04

*Mean values of the two biological replicates are listed.

Supplementary Table S5: Differentially regulated OLGs in LB compared to BHI control. Significant changes are highlighted in gray.

gene name	counts transcriptome BHI*	counts transcriptome LB*	log fold change	p-value	FDR	counts translatome BHI*	counts translatome LB*	log fold change	p-value	FDR
OLGECs009	2	2	0.386	1.000	1.000	9	1	-3.868	0.001	0.083
OLGECs017	48	9	-2.405	5.03E-05	0.008	53	42	-0.326	0.545	1.000
OLGECs021	112	37	-1.614	0.002	0.095	128	159	0.313	0.516	1.000
OLGECs022	66	18	-1.898	4.92E-04	0.040	87	88	0.015	0.986	1.000
OLGECs052	21	14	-0.632	0.307	1.000	51	6	-3.059	9.01E-07	4.40E-04
OLGECs055	12	14	0.234	0.779	1.000	54	12	-2.213	1.05E-04	0.019
OLGECs056	53	5	-3.510	6.42E-08	5.20E-05	34	14	-1.325	0.022	0.674
OLGECs057	7	1	-3.511	0.004	0.187	15	2	-3.220	2.11E-04	0.031
OLGECs065	51	10	-2.339	6.10E-05	0.009	46	41	-0.166	0.769	1.000
OLGECs066	55	11	-2.294	6.52E-05	0.010	63	56	-0.157	0.768	1.000
OLGECs067	27	7	-2.033	0.001	0.094	19	15	-0.349	0.598	1.000
OLGECs068	13	1	-4.338	6.11E-05	0.009	75	20	-1.894	4.20E-04	0.052
OLGECs078	0	8	6.022	2.73E-04	0.027	4	19	2.361	0.001	0.111
OLGECs089	3	3	0.000	1.000	1.000	6	28	2.169	0.001	0.081
OLGECs098	29	5	-2.500	1.48E-04	0.018	36	24	-0.615	0.269	1.000
OLGECs099	29	6	-2.341	3.20E-04	0.031	36	23	-0.626	0.262	1.000
OLGECs101	7	7	0.006	1.000	1.000	43	7	-2.578	2.69E-05	0.007
OLGECs111	25	4	-2.766	9.53E-05	0.013	27	13	-1.107	0.063	1.000
OLGECs116	28	6	-2.322	3.79E-04	0.034	13	12	-0.123	0.923	1.000
OLGECs120	16	67	2.060	1.78E-04	0.020	61	130	1.076	0.029	0.764
OLGECs121	13	91	2.852	3.96E-07	1.96E-04	13	61	2.261	6.15E-05	0.012
OLGECs127	3	1	-2.320	0.157	1.000	24	1	-4.391	5.55E-07	2.99E-04
OLGECs132	26	93	1.864	3.76E-04	0.034	89	193	1.109	0.023	0.691
OLGECs135	9	4	-1.149	0.166	1.000	59	14	-2.058	2.36E-04	0.034
OLGECs136	159	51	-1.634	0.001	0.073	347	217	-0.682	0.149	1.000
OLGECs144	0	8	5.931	4.29E-04	0.036	4	40	3.279	9.17E-07	4.41E-04
OLGECs150	97	8	-3.584	1.64E-09	2.43E-06	32	20	-0.678	0.233	1.000
OLGECs155	299	50	-2.593	3.28E-07	1.66E-04	199	62	-1.692	0.001	0.068
OLGECs166	33	1	-4.881	8.33E-09	9.14E-06	87	6	-3.833	5.24E-10	6.22E-07
OLGECs176	39	993	4.665	2.24E-17	1.90E-13	567	5518	3.272	7.01E-11	8.91E-08
OLGECs204	3	1	-2.318	0.157	1.000	25	2	-3.921	1.53E-06	0.001
OLGECs212	29	14	-1.092	0.061	0.852	48	5	-3.230	4.66E-07	2.68E-04
OLGECs215	45	36	-0.340	0.528	1.000	59	335	2.503	6.32E-07	3.26E-04
OLGECs216	7	0	-5.833	0.001	0.052	3	17	2.668	0.001	0.074
OLGECs218	9	13	0.473	0.527	1.000	39	8	-2.335	1.31E-04	0.022
OLGECs231	98	6	-3.990	9.88E-11	2.24E-07	86	18	-2.283	2.62E-05	0.006
OLGECs232	30	3	-3.494	2.01E-06	0.001	18	21	0.226	0.743	1.000
OLGECs237	30	5	-2.681	6.39E-05	0.010	44	23	-0.970	0.076	1.000

OLGECs256	18	10	-0.841	0.194	1.000	31	3	-3.573	1.12E-06	0.001
OLGECs260	25	21	-0.253	0.684	1.000	35	6	-2.520	7.20E-05	0.014
OLGECs262	23	37	0.717	0.196	1.000	41	285	2.786	5.84E-08	4.25E-05
OLGECs264	175	26	-2.768	1.53E-07	1.02E-04	93	22	-2.115	7.39E-05	0.014
OLGECs265	19	3	-2.574	0.001	0.046	27	23	-0.265	0.668	1.000
OLGECs266	16	2	-2.877	4.62E-04	0.038	25	31	0.315	0.590	1.000
OLGECs267	20	2	-3.206	5.72E-05	0.009	20	24	0.259	0.678	1.000
OLGECs270	20	1	-4.129	4.37E-06	0.001	47	11	-2.136	2.43E-04	0.035
OLGECs277	27	4	-2.696	9.33E-05	0.013	10	8	-0.336	0.707	1.000
OLGECs283	0	3	4.393	0.078	0.909	2	14	3.126	3.80E-04	0.048
OLGECs285	12	16	0.430	0.536	1.000	58	15	-1.976	3.79E-04	0.048
OLGECs297	37	844	4.510	1.83E-16	1.04E-12	128	531	2.046	2.56E-05	0.006
OLGECs298	36	826	4.539	1.38E-16	9.39E-13	91	440	2.275	3.91E-06	0.002
OLGECs322	9	5	-0.831	0.310	1.000	20	3	-2.939	1.22E-04	0.021
OLGECs324	174	44	-2.005	7.05E-05	0.010	57	48	-0.241	0.659	1.000
OLGECs328	13	11	-0.299	0.694	1.000	34	8	-2.137	0.001	0.062
OLGECs344	7	8	0.096	1.000	1.000	17	1	-3.890	2.37E-05	0.006
OLGECs347	109	25	-2.113	5.63E-05	0.009	45	47	0.059	0.925	1.000
OLGECs360	66	4	-4.190	3.72E-10	6.89E-07	38	13	-1.577	0.006	0.327
OLGECs361	51	5	-3.319	2.12E-07	1.27E-04	43	16	-1.425	0.011	0.452
OLGECs364	18	18	0.006	1.000	1.000	51	14	-1.891	0.001	0.081
OLGECs370	6	1	-2.321	0.046	0.725	236	13	-4.225	1.60E-13	3.16E-10
OLGECs371	6	1	-3.169	0.016	0.404	235	13	-4.166	2.72E-13	5.10E-10

*Mean values of the two biological replicates are listed.