



TECHNISCHE UNIVERSITÄT MÜNCHEN

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt
Lehrstuhl für Mikrobielle Ökologie

Experimental characterization of overlapping genes
in enterohemorrhagic *E. coli*: Overexpression phenotypes and
high-throughput NGS analysis of transcription start sites

Barbara Katrin Zehentner

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. Hanno Schaefer
Prüfer der Dissertation: 1. Prof. Dr. Siegfried Scherer
2. Prof. Dr. Wolfgang Liebl
3. Senior Lecturer Dr. Lindsay Hall

Die Dissertation wurde am 05.11.2019 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 14.04.2020 angenommen.

Contents

Abstract	1
Zusammenfassung	3
List of publications	5
Abbreviations	6
List of figures	8
List of tables	10
1 Introduction	13
1.1 Definition and emergence of overlapping genes	13
1.1.1 Biology of overlapping genes	13
1.1.2 Origin of overlapping genes	16
1.1.3 Properties of overlapping genes	17
1.1.4 Overlapping genes in the bacterial world and beyond	19
1.2 Model organism <i>E. coli</i> and the serovar O157:H7	22
1.2.1 Commensal and pathogenic <i>E. coli</i>	22
1.2.2 EHEC O157:H7 pathogenicity and infections	23
1.3 Identification and characterization of genes	25
1.3.1 Bacterial gene structure	25
1.3.2 Bioinformatic identification of genes	27
1.3.3 Experimental approaches to identify and characterize new genes	29
1.4 Perspectives of this study	31
2 Material and methods	33
2.1 Material	33
2.1.1 Bacterial strains and plasmids	33
2.1.2 Chemicals and enzymes	35
2.1.3 Primer	36

2.1.4	Media and buffer	38
2.1.5	Antibiotics and media supplements	41
2.1.6	Length marker and commercial kits	42
2.2	Cultivation and storage of bacteria	43
2.3	Isolation of nucleic acids	43
2.3.1	Isolation of plasmid DNA	43
2.3.2	Isolation of genomic DNA	43
2.3.3	Isolation of RNA	44
2.4	<i>In vitro</i> processing of nucleic acids	45
2.4.1	RNase digest	45
2.4.2	DNase digest	45
2.4.3	cDNA synthesis	45
2.4.4	Polymerase chain reaction (PCR)	46
2.4.5	Quantitative polymerase chain reaction (qPCR)	47
2.4.6	Restriction digest	48
2.4.7	Ligation	49
2.4.8	QuikChange mutagenesis	49
2.4.9	Separation of nucleic acids with agarose gel electrophoresis	50
2.4.10	Separation and Quality control of nucleic acids with capillary gel electrophoresis	50
2.4.11	Determination of nucleic acid concentration and purity	50
2.4.12	Sanger sequencing	50
2.5	Transformation	51
2.5.1	Preparation of electrocompetent cells and electrotransformation	51
2.5.2	Preparation of chemocompetent cells and chemotransformation	51
2.6	Genetic modification of <i>E. coli</i> strains	52
2.6.1	Construction of promoter test strains	52
2.6.2	Construction of protein-tag expression strains	52
2.6.3	Construction of translationally arrested plasmid knock-out transformants	53

2.6.4	Construction of translationally arrested genomic knock-out mutants	53
2.7	Promoter activity analysis	55
2.8	Phenotypic analysis	55
2.8.1	High-throughput analysis	55
2.8.2	Low-throughput analysis	56
2.8.3	Competitive growth with genomic knock-out mutants	58
2.9	Transcriptional start site determination	58
2.9.1	Determination of bacterial growth phases	59
2.9.2	Cappable-seq sample preparation	59
2.9.3	Processing of Cappable-seq sequencing reads	61
2.9.4	Bioinformatic determination of transcriptional start sites	63
2.9.5	Cappable-seq cutoff evaluation for antisense ORFs	63
2.9.6	Determination of gene associated TSS	64
2.9.7	5' UTR evaluation	64
2.9.8	Operon Structures	64
2.9.9	Analysis of TSS strength	65
2.9.10	Promoter motif identification with sequence logos	65
2.9.11	TSS for sense overlapping ORFs	65
2.10	Protein chemical techniques	66
2.10.1	Preparation of whole cell lysates	66
2.10.2	Tris-tricine SDS-PAGE	67
2.10.3	Western blot	67
2.11	Bioinformatic applications	68
2.11.1	Promoter determination	68
2.11.2	Terminator identification	69
2.11.3	Ribosome binding site determination	69
2.11.4	Gene prediction	69
3	Results	71
3.1	Overview of overlapping genes analyzed	71

3.2	Expression of overlapping genes and immunostaining of proteins	73
3.2.1	Evaluation of a suitable overexpression vector	73
3.2.2	Western blots of overlapping genes	74
3.2.3	Protein mass analysis for proteins of overlapping genes	75
3.3	High-throughput overexpression phenotypic analysis	78
3.3.1	Sequencing evaluation	78
3.3.2	Selection of candidates with overexpression phenotypes on the basis of z-scores	81
3.4	Low-throughput phenotypic analysis	85
3.4.1	Analysis of candidates in competitive growth assays	85
3.4.2	Overlapping genes with significant overexpression phenotypes	86
3.4.3	Stress specific phenotypes for overlapping gene candidates	88
3.5	Determination of transcriptional start sites	91
3.5.1	Cappable-seq sequencing output	91
3.5.2	Reproducibility of Cappable-seq	94
3.5.3	Cutoff analysis for genome wide TSS	95
3.5.4	Cutoff analysis for antisense TSS	98
3.5.5	Gene associated transcriptional start sites	100
3.5.6	Detection of putative operon structures of overlapping genes	104
3.5.7	Growth phase and condition dependent TSS strength	105
3.5.8	Bioinformatic and experimental analysis of promoters	113
3.6	Identification of sense overlapping ORFs	127
3.6.1	Differentiation of Cappable-TSS from gene background	127
3.6.2	Analysis of sense ORF associated TSS within annotated genes	130
3.7	Functional characterization of the overlapping gene <i>pop</i> encoded antisense to <i>ompA</i>	136
3.7.1	Genomic localization of <i>pop</i>	136
3.7.2	Effect of overexpression and a knock-out mutant of <i>pop</i> in competitive growth	140
3.7.3	Transcriptional unit of <i>pop</i>	141

3.7.4	pH-dependent detection of Pop in Western blots	145
3.7.5	Bioinformatic analyses of the functionality of <i>pop</i>	146
4	Discussion	149
4.1	Detection of gene products of overlapping gene candidates	149
4.1.1	Assessment of two experimental methods to detect small and low abundance proteins	149
4.1.2	Overlapping gene candidates produce stable proteins	151
4.2	High- and low-throughput overexpression assays detect growth phenotypes	153
4.2.1	Combination of overexpression assays revealed many OGCs producing growth phenotypes	154
4.2.2	Are phenotypes due to an overexpression burden or a specific protein activity?	157
4.3	Transcriptional start site determination using Cappable-seq	159
4.3.1	From dRNA-seq to Cappable-seq	159
4.3.2	Performance of Cappable-seq	160
4.3.3	Identification of hundreds of transcriptional start sites antisense to annotated genes upstream of overlapping genes	162
4.3.4	TSS functionality is supported by reproducible and differential signals as well as promoter activity	164
4.3.5	Cappable-seq gives initial access to sense overlapping genes	167
4.4	Analysis of the pH-regulated overlapping gene <i>pop</i>	169
4.4.1	<i>pop</i> was probably overlooked due to its prominent mother gene <i>ompA</i>	169
4.4.2	Typical gene structure of <i>pop</i>	170
4.4.3	pH-dependent, protein-coding functionality of <i>pop</i>	171
4.4.4	Pathogenesis related function of <i>pop</i> ?	172
4.5	Cumulative evidence for functionality of overlapping gene candidates	173
4.6	Concluding remarks	174
4.7	Outlook	175
5	References	177

6 Supplement	207
6.1 Supplementary files	207
6.2 Supplementary tables	209
6.3 Supplementary figures	284
Acknowledgment	288
Curriculum Vitae	289
Eidesstattliche Erklärung	290

Abstract

Escherichia coli O157:H7 strain EDL933 (EHEC) is a human pathogenic bacterium causing mainly foodborne gastrointestinal infections. The genome of this strain harbors 5498 annotated genes, but open reading frames (ORFs) in intergenic regions also show evidence of translation, likely coding for unknown proteins. In addition, very few protein-coding ORFs overlapping partially or completely with annotated genes in different antisense reading frames (overlapping genes) have been reported in EHEC. However, overlapping genes are mainly ignored in bacteria although they are a well-known feature of viral genomes. Therefore, large-scale experimental approaches to detect and characterize overlapping genes have not been conducted in prokaryotes so far and the structural and functional features of such constructs are largely unknown. The present study was performed to address this task focusing on, but not limited to, a set of 216 previously selected antisense Overlapping Gene Candidates (OGCs).

The expressability of overlapping ORFs was analyzed by Western blots and was verified for 202 OGCs using immunological detection of a fused protein tag. The impact of overexpression of these OGCs on the growth of EHEC was examined under 19 different environmental conditions applying a high-throughput Next Generation Sequencing based phenotyping approach, which revealed 53 OGCs significantly altering bacterial growth upon changing environmental conditions. In subsequent low-throughput competitive growth experiments the overexpression effect was directly compared to the overexpression effect of translationally arrested mutants of the respective overlapping ORFs to clarify the protein-coding potential of the OGC. For 15 candidates stress specific overexpression phenotypes were reproduced indicating functionality of the OGCs at the protein level. Next, transcription start sites (TSS) were determined under eight environmental conditions at a genomic scale using Cappable-seq. TSS were identified for 112 OGCs and another 7064 overlapping ORFs embedded in antisense. More than 40% of these transcription start sites had reproducible, increased TSS signal strengths depending on growth conditions or growth phases, which is evidence for regulation of antisense transcription. For OGCs a conserved promoter structure equivalent to annotated genes was detected. Besides antisense overlapping ORFs,

transcription start sites for 44 putative sense overlapping ORFs were also found.

One example of an overlapping gene candidate from the OGC set is *pop*, an unusually long 603 bp open reading frame. It overlaps completely in antisense with the coding region of the conserved outer membrane protein *ompA*. Measurements of mRNA levels unveiled differential expression of *pop* depending on the pH value of the growth medium. The activity of the promoter region upstream of the transcription start site was verified using a GFP assay and bioinformatic analyses identified a ribosome binding site upstream of the presumed start codon. A translationally arrested mutant of *pop* showed a pH dependent overexpression phenotype and evidence for native translation of *pop* into a protein was found in ribosome profiling data for different *E. coli* strains. These results indicate functionality of *pop* at the protein level although its molecular mechanism of action needs to be investigated in more detail.

Although EHEC is a well studied bacterium, the coding potential of its genome may be significantly underestimated as overlapping genes are still systematically excluded in genome annotation. This study provides experimental evidence for a protein-coding functionality of many overlapping open reading frames in this strain.

Zusammenfassung

Escherichia coli O157:H7 EDL933 (EHEC) ist ein humanpathogenes Bakterium, das oftmals nach Verzehr von kontaminierten Lebensmitteln Erkrankungen des Magen-Darm-Trakts auslöst. Im Genom von EHEC sind 5498 Gene annotiert, allerdings sind auch in intergenischen Bereichen vermeintlich Protein-kodierende offene Leserahmen (ORFs) identifiziert worden, die Translationssignale aufweisen, ebenso vereinzelt Protein-kodierende ORFs, die partiell oder vollständig mit annotierten Genen in unterschiedlichen antisense Leserahmen überlappen (überlappende Gene). Obwohl überlappende Gene in viralen Genomen ein weit verbreitetes Merkmal sind, bleiben jene bei prokaryotischen Genomannotationen größtenteils unbeachtet. Aus diesem Grund wurden bisher kaum umfassende Experimente durchgeführt, um deren funktionale und strukturelle Merkmale zu identifizieren und zu charakterisieren. In vorliegender Arbeit wurde das Ziel verfolgt, Hinweise zu finden, um die Funktionalität von unter anderem 216 zuvor ausgewählten antisense überlappenden Genkandidaten (OGCs) zu verifizieren.

Mit Hilfe von *Western blots* konnte immunologisch gezeigt werden, dass 202 OGCs exprimiert werden können und das kodierte Protein ein stabiles Produkt bildet. Der Einfluss auf das Wachstum von EHEC wurde nach Überexpression der Kandidaten in zwei Ansätzen untersucht. Einerseits zeigte ein Phänotypisierungsansatz in 19 unterschiedlichen Wachstumsbedingungen mittels Hochdurchsatz-Sequenzierung (NGS) die Aktivität von 53 Kandidaten, welche sich in signifikant verändertem bakteriellen Wachstum abhängig der verwendeten Kulturbedingung äußerte. Andererseits konnten in nachfolgenden kompetitiven Wachstumsexperimenten das Protein-kodierende Potential von 15 Kandidaten belegt werden, indem der Überexpressionseffekt intakter Kandidaten mit dem von translational arretierten Mutanten verglichen wurde. Somit konnte stressspezifische Phänotypen basierend auf einer Protein-kodierenden Funktionalität der Kandidaten reproduziert werden. Darüber hinaus wurden die Transkriptionsstarts (TSS) im Genom von EHEC unter acht verschiedenen Kulturbedingungen bestimmt. Signale konnten für 112 OGCs sowie für 7064 vollständig überlappende antisense ORFs gefunden werden. Über 40 % der jeweiligen Startpositionen zeichneten sich durch reproduzierbare, erhöhte Expressionswerte abhängig von den untersuchten Wachs-

tumsbedingungen aus. Für OGCs wurde zusätzlich eine konservierte Promoterstruktur vergleichbar zu der von annotierten Genen gefunden. Neben antisense überlappenden ORFs wurden außerdem für 44 vermutete *sense* überlappende Sequenzen Transkriptionsstarts gefunden.

Ein Beispiel eines überlappenden Kandidatengens aus der Gruppe der OGCs ist *pop*, ein mit 603 Basenpaaren außergewöhnlich langer offener Leserahmen, der vollständig mit der Sequenz des hochkonservierten Membranproteins *ompA* überlappt. Die Regulation der Transkription von *pop* wurde durch quantitative Messungen der mRNA untersucht. Dabei erwies sich der pH-Wert des Wachstumsmediums als kritischer Faktor für die differentielle Expression von *pop*. Die Aktivität des Promoters wurde in einem GFP-Assay bestätigt und eine bioinformatische Analyse zeigte eine ribosomale Bindestelle vor dem angenommenen Startcodon auf. Desweiteren wurden Hinweise für eine natürliche Translation von *pop* in ein Protein in *ribosome profiling* Daten verschiedener *E. coli* Stämme gefunden. Funktionalität basierend auf Wachstumsexperimenten mittels Überexpression zeigte einen pH-abhängigen Phänotyp für *pop*. Die Ergebnisse deuten auf eine Funktionalität von *pop* als Protein-kodierendes Gen hin, wobei die molekulare Funktionsweise noch genauer untersucht werden muss.

Obwohl EHEC bereits gut erforscht ist, scheint dessen Potential zur Kodierung von Proteinen bei weitem unterschätzt zu sein, da überlappende Gene bei Genomannotierungen noch immer systematisch ausgeschlossen werden. Die vorliegende Arbeit bietet allerdings experimentelle Hinweise für eine Protein-kodierende Funktionalität zahlreicher überlappenden offener Leserahmen.

List of publications

Publications

Zehentner, B., Ardern, Z., Kreitmeier, M., Scherer, S., and Neuhaus, K. (2020). A novel pH-regulated, unusual 603 bp overlapping protein coding gene *pop* is encoded antisense to *ompA* in *Escherichia coli* O157:H7 (EHEC). *Frontiers in microbiology* 11, 377

Zehentner, B., Neuhaus, K., Hücker, S. M., Kreitmeier, M., Vanderhaeghen, S., Ardern, Z., and Scherer, S. (2019). ‘Massive overlapping coding in the *E. coli* genome in which hundreds of overlapping genes form a hidden coding reserve’. submitted, currently in revision

Vanderhaeghen, S., Zehentner, B., Scherer, S., Neuhaus, K., and Ardern, Z. (2018). The novel EHEC gene *asa* overlaps the TEGT transporter gene in antisense and is regulated by NaCl and growth phase. *Scientific reports* 8(1), 17875

Conference posters

Zehentner, B., Landstorfer, R., Scherer, S., and Neuhaus, K. (2016). ‘Overexpression of overlapping ORFs in *Escherichia coli* O157:H7 reveals growth phenotypes’. *Society for Molecular Biology and Evolution (SMBE)*

Zehentner, B., Landstorfer, R., Scherer, S., and Neuhaus, K. (2017). ‘Combination of high-throughput and single growth assays detects overexpression phenotypes of translationally arrested overlapping ORFs in *E. coli*’. *European Molecular Biology Laboratory (EMBL)*

Zehentner, B., Scherer, S., and Neuhaus, K. (2018). ‘Genome wide TSS-identification revealed transcriptional start sites for open reading frames antisense to annotated genes’. *Gordon Research Seminar and Conference, Microbial Stress Response (GRS, GRC)*

Abbreviations

5' RACE	rapid amplification of 5' cDNA ends
AG	annotated gene
asRNA	antisense RNA
asTSS	antisense TSS
cDNA	complementary DNA
CDS	coding DNA sequence
DNA	deoxyribonucleic acid
dRNA-seq	differential RNA sequencing
embORF	antisense embedded overlapping ORFs
Gb3	globotriaosylceramide
HT	high-throughput
LB	lysogeny broth
LDF	linear discriminant function
LEE	locus of enterocyte effacement
LT	low-throughput
MM	M9 minimal medium
mRNA	messenger RNA
MS	mass spectrometry
NGS	Next Generation Sequencing
OD _x	optical density at the indicated wavelength x in nanometer
OGC	overlapping gene candidate
OLG	overlapping gene
ON	overnight
ORF	open reading frame
PCR	polymerase chain reaction
RBS	ribosome binding site

RNA	ribonucleic acid
RPKM	reads per kilobase per million mapped reads
rRNA	ribosomal RNA
RRS	relative read score
RT-PCR	reverse transcription PCR
S/N	signal-to-noise ratio
SD	Shine-Dalgarno
Stx	Shiga toxin
TIS	translation initiation site
tRNA	transfer RNA
TSS	transcription start site
UTR	untranslated region
Vol	volume

List of figures

1.1	Illustration of DNA with coding ORFs	14
1.2	Non-trivially overlapping gene types	15
1.3	<i>De novo</i> evolution model described by Carvunis <i>et al.</i> (2012)	16
1.4	Infection route of <i>E. coli</i> O157:H7 (Lim <i>et al.</i> , 2010)	24
1.5	Gene structure elements in the context of gene expression	26
1.6	General workflow of ribosome profiling	29
2.1	Cappable-seq	60
3.1	Western blot of control proteins with SPA- or His-tag	74
3.2	Examples of Western blots	76
3.3	Mass determination of proteins of overlapping genes	77
3.4	High-throughput overexpression phenotyping	78
3.5	Overlapping genes with phenotypes in single competitive growth assays	87
3.6	Overexpression phenotypes of OGC 231	87
3.7	Single competitive growth phenotypes in non-stress environments	90
3.8	Growth curves of EHEC for Cappable-seq	93
3.9	Cutoff value evaluation for genome wide transcriptional start sites	97
3.10	Evaluation of RRS for antisense and intergenic parts of the genome	99
3.11	Gene structure and 5' UTR characteristics	101
3.12	Inter-gene distance of 4146 <i>E. coli</i> K12 operon genes	104
3.13	Differential RRS signals of OGC associated TSS	109
3.14	Differential expression strength for TSS of antisense ORFs.	111
3.15	Analysis of embedded antisense ORFs with transcription start site	112
3.16	Sequence logos of TSS upstream regions and published <i>E. coli</i> promoters	114
3.17	GFP assay of test promoters	117
3.18	TSS and putative promoters for OGC 135 and OGC 85	121
3.19	Differentially expressed TSSs and promoters of OGC 96	123
3.20	TSSs and promoters of OGC 226	124
3.21	TSSs and promoters of OGC 136	125

3.22	TSS and promoter of OGC 207	126
3.23	Examples of TSS within annotated genes	128
3.24	Conservation of promoter regions upstream of 44 sense overlapping ORF associated TSS	132
3.25	Genomic organization of <i>pop</i>	138
3.26	<i>pop</i> translation in ribosome profiling	139
3.27	Effect of <i>pop</i> expression and knock-out in medium of various pH ranges . . .	142
3.28	Analysis of the transcriptional unit of <i>pop</i>	144
3.29	Western blots of Pop	145
4.1	Sources of contamination during Cappable-seq	161
4.2	Summary of structural and functional characterization of 216 OGCs with translation signal in ribosome profiling	173
S1	Relative change of TSS frequencies	284
S2	Length distribution for genes and ORFs of gene sets analyzed	285
S3	Length distribution for overlapping ORFs with TSS	286
S4	Genomic sequence of the overlapping gene <i>pop</i>	287

List of tables

2.1	Bacterial strains	33
2.2	Plasmids	35
2.3	List of primers	36
2.4	Culture media	38
2.5	Buffers	39
2.6	Media supplements	41
2.7	Antibiotics	42
2.8	DNA and protein length markers	42
2.9	Kits	42
2.10	Optical densities, culture volumes and Trizol volumes for RNA isolation . . .	44
2.11	Composition of reaction mix for Taq or Q5 PCR	47
2.12	PCR temperature programs for Taq and Q5 PCRs	48
2.13	Restriction digest of nucleic acids	49
2.14	Culture conditions used in HT phenotyping	57
2.15	Trimming file for Cappable-seq data processing	61
2.16	Composition of resolving and stacking gels for tricine SDS-PAGE	67
3.1	Categories of protein signals in Western blot	75
3.2	Correlation of HT sequencing experiments	79
3.3	Number of mapped reads ($\times 10^4$) for replicated sequencing experiments for HT phenotyping	80
3.4	Correlation of RPKM profiles of individual OGCs in HT phenotyping	81
3.5	Overlapping gene candidates with HT overexpression phenotype	83
3.6	Overlapping gene candidates with phenotype in LT phenotyping	88
3.7	Sequencing reads in Cappable-seq	91
3.8	Mapped reads in Cappable-seq	92
3.9	Correlation of Cappable-seq data sets	95
3.10	Short description of gene sets for TSS identification	102

3.11	Gene associated transcriptional start sites for annotated genes and antisense ORFs	103
3.12	Overlapping gene candidates possibly localized in operons	106
3.13	Gene associated TSS with significant differences of RRSs	107
3.14	Detailed expression patterns of TSS of embedded ORFs	110
3.15	Overview of candidates selected for promoter analysis	116
3.16	Characteristics and activity of promoters analyzed in GFP assay	118
3.17	Sense overlapping ORFs with TSS	129
3.18	Categorization of putative sense overlapping TSS after visual analysis	130
3.19	Characteristics of sense overlapping ORFs	133
3.20	<i>pop</i> prediction with Prodigal	147
4.1	Comparison of HT and LT phenotyping	155
S1	Additional plasmids used and constructed	209
S2	Reverse primer used for cloning pBAD/SPA+OGC x variants	211
S3	Mutation primer used to construct pBAD/ Δ OGC x variants	215
S4	Primer used to construct pProbe-NT+promoter variants	217
S5	Supplier information of chemicals	218
S6	List of analyzed overlapping gene candidates	220
S7	Blastp analysis of overlapping gene candidates	232
S8	Proteins of overexpressed overlapping genes in western blots	233
S9	Graphical summary of HT and LT phenotyping	242
S10	Genome wide transcription start sites determined with Cappable-seq	261
S11	Gene associated TSS for OGCs	262
S12	TSS for sense overlapping ORFs	275
S13	Evaluation of ribosome profiling and RNAseq of <i>pop</i> genomic region	283

1 Introduction

1.1 Definition and emergence of overlapping genes

1.1.1 Biology of overlapping genes

The genetic information of an organism is stored in its genome by deoxyribonucleic acid (DNA) or, in some viruses, ribonucleic acid (RNA) consisting of the nucleotides adenine (A), guanine (G) and cytosine (C) as well as thymine (T) or uracil (U) for DNA and RNA, respectively. To obtain information stored in the DNA, the RNA polymerase transcribes DNA into ribosomal RNA (rRNA), transfer RNA (tRNA), or complementary RNA molecules which either act as non-coding regulatory RNAs (Storz *et al.*, 2011) or serve as templates for the translation into proteins by ribosomes, i. e., messenger RNA (mRNA).

According to the genetic code, translation of the nucleotide sequence into amino acids occurs triplet-wise (i. e., per codon). This process is typically initiated at the start codon AUG, coding for methionine (more accurately, N-formylmethionine), whereas further NUG and AUN codons enable translational start at a reduced efficiency depending on the codon used (N represents any of the four nucleotides; Sussman *et al.*, 1996; Hecht *et al.*, 2017). Termination is induced if a release factor recognizes one of the three stop codons UAA, UGA, or UAG causing the release of the peptide chain from the ribosome (Poole and Tate, 2000). Start and stop codons form the boundaries of open reading frames (ORF), the coding units of genes (Section 1.3.1).

As the DNA is structured as a double helix and translation of DNA via RNA into proteins requires base triplets, six different reading frames exist at a single genomic locus (Figure 1.1). In theory, each of the reading frames can contain protein-coding ORFs. Consequently, two protein-coding genes are able to be located at the same DNA locus in different reading frames and to overlap each other. However, genome annotation algorithms programmed to find coding ORFs generally exclude long overlaps in prokaryotes (Hyatt *et al.*, 2010). Thus, only one of either ORFs is annotated and assumed genuine. In this work and elsewhere, the position of the second ORF of the gene pair, the overlapping gene (OLG), is specified regarding the annotated gene (AG, Figure 1.2). OLGs can be distinguished between sense or antisense orientation, as well as partial or embedded overlap in relation to annotated

genes. The overlap length classifies the overlapping gene pair as either trivial or non-trivial. The former share just a few base pairs, the latter at least 90 bp. While trivial overlaps were subject to several large-scale analyses (Saha *et al.*, 2015; Fonseca *et al.*, 2014; Johnson and Chisholm, 2004), non-trivially overlapping genes have not yet been characterized to any great extent in prokaryotes, but constitute a fascinating phenomenon.

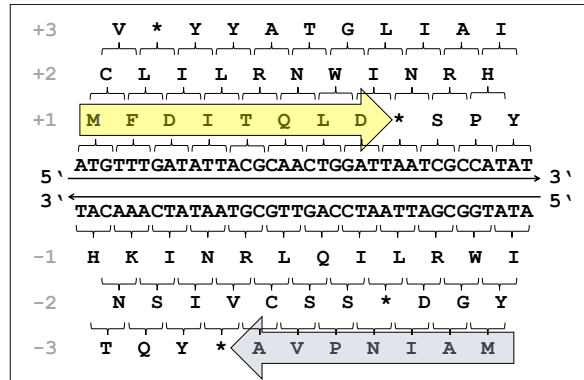


Figure 1.1: Illustration of DNA with coding ORFs. Translation of the DNA can occur in 3 readings frames on the sense and on the antisense strand, respectively, resulting in six possible reading frames. yellow: annotated gene in frame +1; blue: overlapping gene in frame -3.

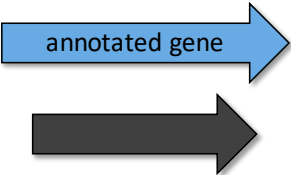
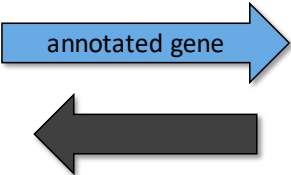
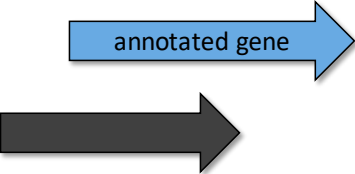
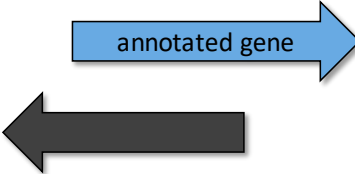
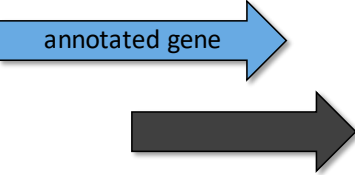
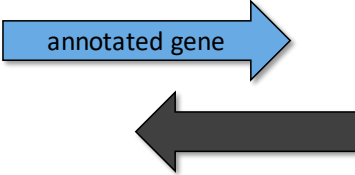
overlap	sense/parallel	antisense/antiparallel
embedded		
5' partial		
3' partial		

Figure 1.2: Non-trivially overlapping gene types. Orientation of overlapping genes (gray) is shown regarding the annotated gene (blue). Sense/parallel and antisense/antiparallel OLGs are categorized as embedded, 5' partial or 3' partial overlaps if the ORF is located within the annotated gene or overlaps the annotated gene at its 5' or 3' end, respectively.

1.1.2 Origin of overlapping genes

Different mechanisms for gene emergence have been described with gene duplication-divergence being the longest-discussed principle contributing to gene evolution (Bergthorsson *et al.*, 2007; Dittmar and Liberles, 2011). Besides this, gene fusion/fission and, in particular, horizontal gene transfer are important sources for bacterial diversity (Ochman *et al.*, 2000). A more recently discovered process for gene evolution is *de novo* evolution, which was rejected for a long time (Tautz, 2014; Jacob, 1977). In this scenario, new genes do not descend from preexisting genes, but arise from naturally occurring but formerly non-coding sequences (Keese and Gibbs, 1992). These sequences acquire regulatory elements indispensable for expression of proteins, which, in turn, interact with other gene products (Andersson *et al.*, 2015). A recently developed model formalizes this process (Figure 1.3; Carvunis *et al.*, 2012). As non-genic sequences are transcribed (Thomason *et al.*, 2015; Dornenburg *et al.*,



Figure 1.3: *De novo* evolution model described by Carvunis *et al.* (2012). Non-genic sequences are occasionally translated and form proto-genes, which either return into a non-genic state or evolve into (new) genes. Once being a gene, sequences can be template for duplication events (circular arrow). Accumulation of deactivating mutations forms non-functional pseudo-genes, which are first highly similar to the original genes but become over time indistinguishable from non-genic sequences.

2010) and frequently associated with ribosomes (Hücker, Ardern, *et al.*, 2017; Baek *et al.*, 2017), it is presumed that they are even translated as gene precursors (proto-genes) and may have adaptive potential, developed through exposing previously hidden sequences to selection (Carvunis *et al.*, 2012; Wilson and Masel, 2011). Beneficial proto-genes might be retained in the genome and evolve into genes, while deleterious or neutral precursors lose the ability to be translated and revert into non-genic sequences.

Evidence for *de novo* gene evolution can be found in orphan/taxonomically restricted genes which lack significant sequence homology to known genes, thus rendering gene duplication for new gene emergence an inadequate model (Tautz, 2014; Khalturin *et al.*, 2009; Schlötterer,

2015). There is evidence throughout both eukaryotes and prokaryotes for a considerable proportion of genes being orphans (Dujon, 1996; Yang *et al.*, 2013; Satoshi and Nishikawa, 2004), and even sequencing more genomes across taxonomic groups did not reduce the number of these unusual genes (Wilson *et al.*, 2005). Consequently, orphan genes are presumed to be involved in the development of taxonomic-/species-specific morphological traits and environmental adaptation (Khalturin *et al.*, 2009), but they have also been shown to be important for eukaryotic biology and evolution in general (Johnson, 2018; Verster *et al.*, 2017).

One mechanism of how *de novo* gene evolution can take place was first discussed in 1973 by Grassé (English translation 1977) describing a scenario termed ‘overprinting’. It comprises the utilization of an alternative reading frame in a preexisting gene, which results in the formation of an overlapping gene pair (Keese and Gibbs, 1992). As pointed out by Tautz (2014), it is clear that overprinted genes are not created by gene duplication and must originate *de novo*. Single overlapping gene examples were described starting mid 1970s in bacteriophage Φ X174, thus providing evidence for the process of overprinting quite early (Barrell *et al.*, 1976), though *de novo* evolution in general was not widely considered a realistic explanation of gene emergence (Stephens, 1951).

In addition to overprinting and *de novo* origination, overlapping genes might also arise by mutational events which create new upstream start codons or delete stop codons. This causes elongation of preexisting genes, which now overlap other genes (Fonseca *et al.*, 2014; Fukuda *et al.*, 1999; Rogozin *et al.*, 2002). However, such overlaps are more likely to be trivial and probably did not originate *de novo*.

1.1.3 Properties of overlapping genes

The predominant overlap type in bacterial genomes constitute trivial overlapping genes in sense orientation (Saha *et al.*, 2015; Fonseca *et al.*, 2014; Johnson and Chisholm, 2004). It has been known for years that these overlaps are regulators in gene expression creating translational coupling (Scherbakov and Garber, 2000; Normark *et al.*, 1983; Oppenheim and Yanofsky, 1980). Thus, these gene constructs can increase the growth rate of bacteria by reduction of generation time (Saha *et al.*, 2016). In contrast, non-trivially overlapping genes

seem to be less abundant in bacteria with only few examples analyzed (e. g., Delaye *et al.*, 2008; Tunca *et al.*, 2009, Section 1.1.4). However, statistical analyses revealed a significantly different length distribution of ORFs with long overlaps in alternative reading frames in comparison to expectations based on codon usage alone (Mir *et al.*, 2012). This indicates that from a statistical view bacterial genomes harbor a greater number of long overlapping ORFs than expected.

However, the existence of two protein-coding genes at a single genomic locus is accompanied with certain characteristics. Constraints for sequence adaptations are most obvious, as mutations within the overlapping gene pair affects both sequences and thus, their evolution is limited relative to non-overlapping genes (Krakauer, 2000). An analysis in viruses showed that in general proteins encoded by overlapping genes consist of more high-degeneracy amino acids than non-overlapping genes in order to tolerate more mutations during selection (Pavesi *et al.*, 2018). The selection pressure, however, differs for different reading frames and only OLGs in reading frame -2 (according to Figure 1.1) are automatically protected (to some extent) if purifying selection acts on the annotated gene in reading frame +1 (Mir and Schober, 2014; Rogozin *et al.*, 2002). Additionally, overlapping gene pairs are highly constrained in frames +1/-2 and in a lesser extent in +1/-3, +1/+2 and +1/+3 if amino acid sequence composition is examined in the form of di- and n -peptides (Lèbre and Gascuel, 2017). Related to this, overlapping open reading frames created by overprinting are characterized by a different codon usage than non-overlapping annotated genes (Pavesi *et al.*, 2013). This implies that annotated genes and OLGs have different gene ages, thus supporting the *de novo* gene evolution process where overlapping genes are rather young (Rogozin *et al.*, 2002). In contrast to trivially overlapping genes, which show a high degree of sequence conservation across bacterial genomes (Johnson and Chisholm, 2004), non-trivially overlapping genes are considerably less conserved than the corresponding annotated genes (e. g., Fellner *et al.*, 2015; Delaye *et al.*, 2008) and thus, classified as orphan genes or developing proto-genes. As such, they seem to have non-essential and probably poorly adapted functions (Moshensky and Alexeevski, 2019; Chen *et al.*, 2012; Jordan *et al.*, 2002; Rogozin *et al.*, 2002), which might explain why low transcription and translation rates have been observed (Landstorfer, 2014). It must be noted, however, that such analyses have not yet been conducted for a

high-confidence set of prokaryotic overlapping genes. Most of the aforementioned analyses were conducted with data from viruses, and mostly for same-strand overlaps. As such, substantially more work remains to be done in this area.

Although evolution and sequence composition of overlapping genes is restricted, a computational approach allowed construction of plenty of double and triple overlapping genes having intact functional domains of known proteins (Opou et al., 2017). The authors suggest as these gene pairs can be designed easily that they may have occurred frequently in genome evolution. Furthermore, overlapping coding might be of biotechnological interest to reduce genetic drift in expression strains.

Some detailed studies about the sequence characteristics of overlapping genes have been conducted (mostly in viruses), but large scale experimental analyses showing any phenotypic impact of overlapping genes in organisms are missing but necessary to verify functionality of these genes in the genomic context of bacterial cells.

1.1.4 Overlapping genes in the bacterial world and beyond

Overlapping genes were first identified in bacteriophages (Barrell et al., 1976) and extensive work has let them become an integral part of viral genomes (Keese and Gibbs, 1992; Pavese et al., 2018). Although overlapping genes are arguably the best examples of *de novo* gene evolution, as the sequence context of the overprinted gene is fixed, thus, easier to determine, it was long thought that the size constraint of the viral capsid, and hence the viral genome, is the driving force for gene overlaps (Chirico et al., 2010). However, gene novelty and evolutionary exploration were recently shown to explain the potential of OLGs better as viral genomes are often too small to completely fill the available capsid volume (Brandes and Linial, 2016). Thus, the capsid does not seem to be the limiting factor for the viral genome size or, consequently, also for the origin of overlapping genes.

Although OLGs have been identified in diverse genomes (e.g., Makalowska et al., 2005; Michel et al., 2012; Mouilleron et al., 2015; Capt et al., 2016; Balabanov et al., 2012) and more recent insights in the development of overlaps has been gained, the widespread occurrence of overlapping genes is still not generally accepted in bacteria (e.g., Warren et al., 2010). Moreover, genes in alternative reading frames have been rejected always or nearly al-

ways due to misannotations (Pallejà *et al.*, 2008). Nevertheless, more and more non-trivially overlapping genes are being detected and characterized in prokaryotes, although sometimes just by serendipity (e. g., Haycocks and Grainger, 2016). Systematic searches for transcribed and translated genomic regions have increased the opportunities to identify at least anti-sense overlapping ORFs (Landstorfer, 2014; Hücker, 2018). Based on these experiments, the experimental description of four antisense overlapping genes in the humanpathogenic bacterium *E. coli* O157:H7 was possible so far:

- *nog1*: novel overprinted protein-coding OLG with upregulated transcription in cow dung (Fellner *et al.*, 2015)
- *ano*: bicistronically expressed protein effective under anaerobic conditions (Hücker *et al.*, 2018a)
- *laoB*: recently evolved small protein responsive under arginine stress (Hücker *et al.*, 2018b)
- *asa*: disordered small protein with overexpression phenotype in high salt supplemented medium (Vanderhaeghen *et al.*, 2018).

Furthermore, evidence for the controversial gene pair *htgA/yaaW* was provided (Fellner, 2014; Delaye *et al.*, 2008; Missiakas *et al.*, 1993; Nonaka *et al.*, 2006).

While the list of characterized overlapping genes is growing for *E. coli* (McVeigh *et al.*, 2000; Behrens *et al.*, 2002; Balabanov *et al.*, 2012; Kurata *et al.*, 2013; Haycocks and Grainger, 2016), other prokaryotes also harbor overlapping genes. For instance, Kim *et al.* (2009) detected altogether 10 OLGs in *Pseudomonas fluorescens* using mass spectrometry. They identified one protein for sense and nine proteins for antisense overlapping genes. In the actinomycete *Streptomyces coelicolor*, the protein-coding gene *adm* overlaps in antisense the important iron regulator *dmdR1* (Tunca *et al.*, 2009). For the first time, strand-specific knock-outs were created and phenotypes were analyzed for both genes separately. It could be shown that a tightly regulated interplay between expression of the genes controls siderophore and antibiotic biosynthesis. In the thermophilic bacterium *Thermus thermophilus*, a rare case of non-trivially overlapping sense genes was described (*rpmH/rnpA*, Feltens *et al.*, 2003).

Expression of both genes is directed from the same ribosome binding site (RBS), thus linkage of ribosome and RNase biosynthesis could be assumed.

Bacterial genomes have a huge number of non-trivially overlapping open reading frames in sense as well as in antisense direction. In spite of a few examples, which have been characterized, functionality of the large majority has not been proven.

1.2 Model organism *E. coli* and the serovar O157:H7

1.2.1 Commensal and pathogenic *E. coli*

Escherichia coli is a Gram-negative, $1\ \mu\text{m} \times 3\ \mu\text{m}$ rod shaped bacterium in the family Enterobacteriaceae and class γ -proteobacteria. Since its description in 1886 by Theodor Escherich, *E. coli* has emerged as an important microorganism in various fields including basic research to understand biological mechanisms and processes (e.g., Inoue *et al.*, 2007; Lange and Hengge-Aronis, 1991), production of biopharmaceuticals (e.g., Baeshen *et al.*, 2015; Mamat *et al.*, 2015) or metabolic engineering for white biotechnology applications (e.g., Shomar *et al.*, 2018; Yim *et al.*, 2011).

E. coli was identified as a commensal bacterium which inhabits the gastrointestinal tract of humans and colonizes the gut of infants as early as a few hours after birth but may cause diseases merely in weakened e.g., immuno-compromised hosts (Taur and Pamer, 2013). In contrast to apathogenic strains, pathogenic *E. coli* are responsible for a broad range of diseases ranging from gastroenteritis to extraintestinal infections in for instance the urinary tract or the central nervous system (Kaper *et al.*, 2004; Johnson and Russo, 2002). The similarities of commensals and pathogens based on the genome sequence are limited and genome differences of more than 1 Mbp have been reported (Croxen *et al.*, 2013; Perna *et al.*, 2001). The core genome, however, is present in both groups and includes 1000 to 3000 genes exerting diverse functions including transport or energy metabolism (Kaas *et al.*, 2012; Lukjancenko *et al.*, 2010). Divergence from apathogenic into pathogenic strains occurred by the acquisition of virulence factors as well as gain and loss of genes involved in adaptation to different environments (Croxen *et al.*, 2013; Blount, 2015). Evolutionary analyses showed that different pathogenic *E. coli* strains appear to have evolved by way of successive transmission events of prophage regions and mobile elements (Wick *et al.*, 2005), with these additions quite often even occurring in parallel across different lineages (Reid *et al.*, 2000).

According to Kaper *et al.* (2004), pathogenic *E. coli* can be classified into seven groups: enteropathogenic *E. coli* (EPEC), enterohemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAEC), enteroinvasive *E. coli* (EIEC), diffusely

adherent *E. coli* (DAEC) and *E. coli* associated with extraintestinal infection (ExPEC). A further group of pathogenic strains comprises all Shiga-toxin producing *E. coli* (STEC) which includes EHEC (Sperandio and Nguyen, 2012). Each pathotype has a characteristic combination of virulence factors leading to a characteristic pathogenic potential. Within pathotypes clonal groups can be clustered into serotypes with typical polysaccharide (O) and flagellar surface antigen (H) combinations (Croxen *et al.*, 2013). The number of identified O-antigens is considerable (>180, DebRoy *et al.*, 2016). Therefore, it is not remarkable that for STEC alone more than 380 different serotypes (based on O and H) have been isolated (Sperandio and Nguyen, 2012). However, only a small subset is associated with human diseases (Blanco *et al.*, 2003). One serotype linked to several outbreaks is O157:H7 within the EHEC pathotype (e.g., Riley *et al.*, 1983; Michino *et al.*, 1999; Braeye *et al.*, 2014; Furukawa *et al.*, 2018). The strain EDL933 of this serotype is the subject of this thesis.

1.2.2 EHEC O157:H7 pathogenicity and infections

The natural reservoir of *Escherichia coli* O157:H7 is the gut of cattle and livestock (Fairbrother and Nadeau, 2006), but EHEC can also survive in water, soil, and on diverse plants like lettuce and sprouts (e.g., Carey *et al.*, 2009; Semenov *et al.*, 2010). EHEC is ingested by the susceptible human host mainly via contaminated food (meat, dairy products or plants, Figure 1.4, Lim *et al.*, 2010). Only a couple of hundred cells are required to induce an infection (Tuttle *et al.*, 1999) exhibiting symptoms like watery diarrhea and hemorrhagic colitis as well as in severe cases renal failure due to the hemolytic uremic syndrome (HUS, Croxen *et al.*, 2013).

After passing the acid barrier of the stomach (Hong *et al.*, 2012), EHEC attaches to endothelial cells of the large intestine (Croxen and Finlay, 2010). This mechanical cue of the initial attachment (Alsharif *et al.*, 2015), as well as chemical environmental signals (Abe *et al.*, 2002), induce the expression of the locus of enterocyte effacement (LEE) pathogenicity island. Different effector molecules including a type III secretion system are responsible for LEE-specific attaching and effacing lesions, one major characteristic of EHEC infections (Stevens and Frankel, 2014). One of the effectors is the protein receptor Tir, which is translocated by the type III secretion system into the host cell membrane (DeVinney *et al.*,

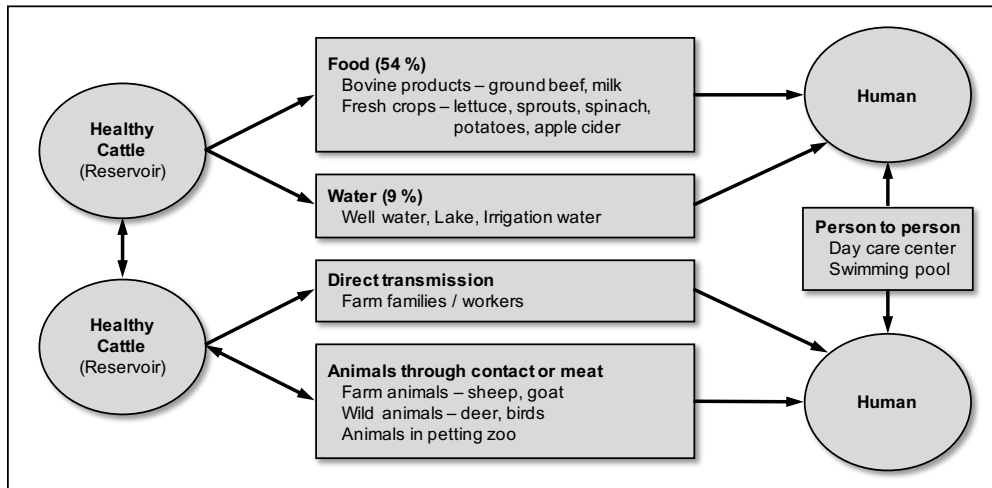


Figure 1.4: Infection route of *E. coli* O157:H7 (Lim *et al.*, 2010). EHEC naturally inhabits the gut of asymptomatic cattle. Humans are contracted with EHEC via contaminated food, water or direct contact with infected animals and humans.

1999). It interacts with Intimin, a cell surface protein of pathogenic *E. coli*, and forms an intimate connection of the bacteria to the host (Stevens and Frankel, 2014). This tight binding induces the remodeling of the host actin cytoskeleton: actin polymers initially used to build up microvilli accumulate at bacterial attachment sites and the intestinal membrane protrusions are destroyed (Knutton *et al.*, 1989; In *et al.*, 2016). This contributes to diarrhea associated with EHEC infections.

In addition to the LEE pathogenicity island, the expression of phage-encoded Shiga toxins (Stx) contributes to the severity of an EHEC infection. Genome analysis of *E. coli* O157:H7 strain EDL933 revealed the presence of the two known toxin variants Stx1 and Stx2 (Perna *et al.*, 2001) with Stx2 associated with a more severe course of the disease (Boerlin *et al.*, 1999). Upon bacterial stress, phage mediated expression of the toxins and cell lysis associated with toxin release is induced (Wagner *et al.*, 2002). Shiga toxins are AB₅-toxins. Five B subunits form a pentamer capable of interacting with the receptor Gb3 (globotriaosylceramide) on human target cells in the intestinal and vascular epithelium as well as in kidneys (Schüller, 2011; Bauwens *et al.*, 2013; Obrig, 2010). In contrast, cattle lack Gb3 receptors in the gut which explains resistance of newborn calves as well as adult cattle to STEC infections (Pruimboom-Brees *et al.*, 2000). After binding, receptor and toxin are internalized and the non-covalent associated, enzymatically active A subunit exhibits its N-glycosidase activity at

the 28S rRNA of ribosomes which results in abortion of protein synthesis and, consequently, in cell death (Endo *et al.*, 1988; Melton-Celsa, 2014). If Shiga toxins enter the blood stream and reach renal cells, the infection becomes systemic and the likelihood of developing HUS is increased.

The combination of Shiga toxins, LEE pathogenicity island and further virulence factors like the virulence plasmid pO157 encoded enterohemolysin *ehxA* leads to the clinical picture of EHEC infections (Lim *et al.*, 2010). Treatment options are often solely symptomatic as the application of antibiotics against *E. coli* is avoided to prevent increased release of Shiga toxins (Goldwater and Bettelheim, 2012). Consequently, if an infection is treated with antibiotics, the prevalence to establish HUS is increased and the situation can become life-threatening (Wong *et al.*, 2012). Alternative approaches might use recently developed single domain antibodies against the B subunits of Stx2, which were shown to efficiently neutralize the toxin (Mejías *et al.*, 2016). Additionally, as the microbial composition of the gut of infected individuals significantly influences the susceptibility to and severity of an EHEC infection (Koyanagi *et al.*, 2019), use of probiotic bacteria constitutes a promising approach to prevent and treat diseases (Saito *et al.*, 2019; Dini *et al.*, 2016).

1.3 Identification and characterization of genes

1.3.1 Bacterial gene structure

A standard prokaryotic protein-coding gene consists of several functional and regulatory elements (Figure 1.5). The part carrying sequence information for protein synthesis is the coding DNA sequence (CDS) in the form of an open reading frame. The initial step during protein expression is mRNA synthesis/transcription followed by protein synthesis/translation. Each of the steps has specific regulatory elements.

The element necessary to initiate mRNA synthesis is the promoter region. This sequence is up to several hundred bp long and located upstream of the CDS. It is the primary attachment site of the RNA polymerase to the DNA. σ factors, which are multi-domain proteins, bind to the initially inactive polymerase protein complex in order to activate it. Additionally, specific σ -subunits recognize the -10 and -35 elements of the promoter, thus guiding the

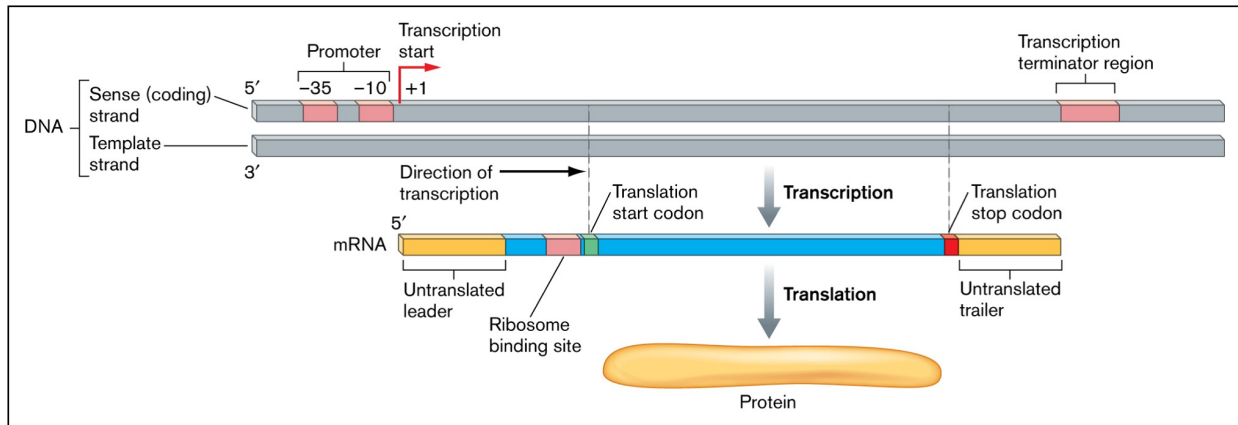


Figure 1.5: Gene structure elements in the context of gene expression. Slonczewski and Foster (2013)

polymerase holoenzyme to the accurate transcription initiation site (Davis *et al.*, 2016). Depending on the nature of the σ factors, different promoters are identified (Browning and Busby, 2004). For instance, the highly conserved -10 Pribnow-box (5' TATAAT 3') and the associated -35 region (5' TTGACA 3') are recognized by the housekeeping sigma factor σ^{70} responsible for driving elemental gene expression. Presence of the -35 element, however, is not mandatory for promoter functionality as long as one specific subunit of σ^{70} interacts with activator proteins and induces the enzymatically active form of the RNA polymerase (Paget and Helmann, 2003). Besides the two highly conserved regions, (a) the UP element, about -40 to -60 bp upstream of the -35 region (Estrem *et al.*, 1998), (b) the extended -10 region (Burr *et al.*, 2000), and (c) the spacer between -10 and -35 sequence (Singh *et al.*, 2011) were proven to be important for some promoters. Further upstream-located areas may contain binding sites for transcription factors which can affect and regulate transcription. mRNA is synthesized from position +1, the transcription start site (TSS), to the terminator sequence at the 3' end of the mRNA which induced the stopping of transcription either intrinsically by hairpin formation or in a factor-dependent manner for instance by the (ATP)-dependent RNA-DNA helicase Rho (Ray-Soni *et al.*, 2016).

Apart from the CDS enclosed by start and stop codon, the mRNA typically contains untranslated regions (UTRs) at the 5' as well as 3' end. The 3' UTR was long thought to be solely responsible for regulating transcription termination, though the function of regulatory processes of this region has been studied extensively in eukaryotes and is well

understood there (Mignone *et al.*, 2002). Nevertheless, recent findings showed that also in prokaryotes post-transcriptional regulation can take place in the 3' UTR by small RNAs or controlled mRNA decay (e.g., López-Garrido *et al.*, 2014; Ren *et al.*, 2017). In contrast, post-transcriptional and translational regulation by the 5' UTR is a well-known process and structures like RNA thermometers (Hücker, Simon, *et al.*, 2017; Kortmann and Narberhaus, 2012), riboswitches (Mellin and Cossart, 2015; Abduljalil, 2018), and small RNAs (Waters and Storz, 2009) are of high interest to understand for instance rapid adaptation of bacterial gene expression to environmental stresses. Finally, the ribosome binding site (also Shine-Dalgarno (SD) sequence) on the mRNA is the initial attachment site for the ribosome prior to translation initiation. The consensus sequence 5' GGAGG 3' was identified for *E. coli* with an optimal distance to the start codon of 5 bp. Base-pairing of the 3' end of the 16S rRNA in the bacterial ribosome (anti-SD sequence, 3' AUUCCUCCACUA 5') directs the ribosome to the start codon (Ma *et al.*, 2002). Although a SD sequence seemed to be mandatory for translation, protein synthesis of leaderless mRNAs without 5' UTR and solely a 5' ATG start codon was shown and indicates that the RBS is non-essential (Brock *et al.*, 2008; Lomsadze *et al.*, 2018). Additionally, even in transcripts with a leader sequence, extensive sequence deviations of the consensus still allow translation initiation, though at reduced efficiencies (Evfratov *et al.*, 2016).

A further striking feature in the prokaryotic genome is the arrangement of genes in operons allowing transcription of functionally related genes as single, polycistronic mRNA that codes for several proteins (e.g., Aksoy *et al.*, 1984).

1.3.2 Bioinformatic identification of genes

The initial bioinformatic analysis of newly sequenced bacterial genomes is based on the structural and functional annotation of genes and non-coding features such as structural RNAs or small RNAs (Tatusova *et al.*, 2016). The gene structure of prokaryotic genes is well-known and less complex than many eukaryotic genes. As bacteria lack a nucleus, transcription and translation proceeds rather simultaneously. Consequently, gene regulation on the post-transcriptional and translational level is important and differences in gene structure elements contribute to the exceptionally fine-tuned gene expression (Browning and Busby,

2004).

Despite sequences of regulatory elements of some genes deviating from consensus sequences or being completely absent (e. g., ribosome binding site), gene-finding algorithms are optimized to comprehensively identify genes based on the structural features. Several different methods have been developed for this task, for instance GeneMarkS (Besemer *et al.*, 2001), Glimmer (Delcher *et al.*, 2007), Prodigal (Hyatt *et al.*, 2010), which is included in the PROKKA genome annotation pipeline (Seemann, 2014), or a CDS prediction algorithm included into the FGeneSB genome annotation pipeline (Solovyev and Salamov, 2011). Start codons, ribosome binding sites, gene length, coding potential based on GC content and genomic localization are assessed and determine the potential of a sequence to be protein-coding (Hyatt *et al.*, 2010). Furthermore, supporting features like promoter binding sites and terminators are integrated to improve the structural annotations (Solovyev and Salamov, 2011). However, all known methods share in common avoidance of labeling extensively overlapping open reading frames as genes, as well as small open reading frames to minimize false positive discovery rates. Algorithms behind all these methods differ and may include Markov models or log-likelihood functions to find, score and annotate the ORFs most likely to be protein-coding genes.

The second step required for gene annotation is the assignment of a function to identified ORFs. By means of different protein databases (e. g., RefSeq, COG, KEGG) the proteins' function or participation in a specific pathway is deduced (Solovyev and Salamov, 2011). If no significant similarity with any entry on protein or domain level is found, the gene product is annotated with the label 'hypothetical' or similar terms (Seemann, 2014; Haft *et al.*, 2017). For example, the genome of *E. coli* O157:H7 str. EDL933 was annotated using the NCBI Prokaryotic Genome Annotation Pipeline (RefSeq annotation NZ_CP008957, 02/23/2017, Latif *et al.*, 2014). In total, 5498 protein-coding genes were defined. Of these, 4525 have a functional annotation as significant homologies to functionally characterized proteins were found. For 973 sequences, a hypothetical annotated function was assigned. Considering the phenomenon of orphan genes described in Section 1.1.2 and the fact that many true proteins have simply not yet been characterized, absence of a protein match in databases does not necessarily indicate missing function of the proteins or even further a false positive result in

gene annotation.

Exact bioinformatic prediction of functional protein-coding genes can successfully be applied for many sequences in a bacterial genome, if the reference databases to annotate functions are of high quality, but methods struggle with the annotation of unusual, new or underrepresented sequences. Therefore, experimental characterization is unavoidably required to uncover functionality of such putative genes.

1.3.3 Experimental approaches to identify and characterize new genes

Gene identification relies mainly on bioinformatic tools, but search algorithms exclude atypical sequences from annotation, for instance small and overlapping genes encoding new proteins. Ribosomal profiling is an emerging technology first developed in 2009 by Ingolia *et al.* which supplements theoretical gene assignments with experimental data about the translation status of the genome. A snapshot of the translated part of an organism's genome at a certain time point is created by analyzing mRNA molecules associated with ribosomes (Figure 1.6). By means of this method, translated genes in the intergenic regions of anno-

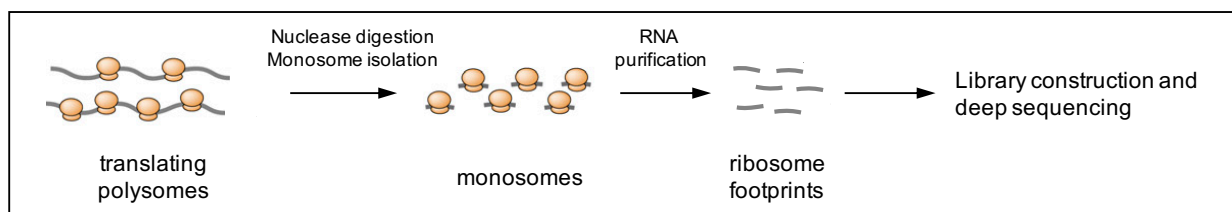


Figure 1.6: General workflow of ribosome profiling. Cell lysates with intact polyribosomes are prepared. Single ribosomes are separated by treating samples with endonucleases and mRNA overhangs are digested by exonucleases. Ribosome protected mRNA fragments (footprints) are isolated and analyzed by next generation sequencing. Genomic regions covered by a substantial amount of reads can be considered as translated. Figure adapted from Hsu *et al.* (2016).

tated genes, which were overlooked so far, were identified (Hücker, Ardern, *et al.*, 2017). Furthermore, several different studies have proven the translation of small genes into tiny proteins using ribosomal profiling applications in prokaryotes (e.g., Weaver *et al.*, 2019; Sberro *et al.*, 2019). More important, this method sheds light on the long neglected phenomenon of overlapping genes. Combined with findings of extensive antisense and sense overlapping transcription found with differential RNA-seq (dRNA-seq) or modified 5' RACE

(rapid amplification of 5' cDNA ends, e. g., Sharma and Vogel, 2014; Thomason *et al.*, 2015; Mendoza-Vargas *et al.*, 2009) ribosome profiling provides evidence for the targeted expression of out-of-frame ORFs in bacteria resulting in the ‘alternative proteome’ (Meydan *et al.*, 2019). Nevertheless, some critics deny the biological relevance of such signals, favoring an explanation in terms of uncontrolled or spurious transcription and translation (Ingolia *et al.*, 2014; Smith *et al.*, 2019; Lloréns-Rico *et al.*, 2016). Functional characterization of such candidate genes could force the argumentation towards the existence and functionality of overlapping genes.

Resolving the functions of genes has long been of interest to understand cellular processes, but experimental methods were revolutionized with large-scale functional genomic approaches and next generation sequencing (NGS) applications (Gray *et al.*, 2015; Brochado and Typas, 2013). Besides transposon-based strategies (Hensel *et al.*, 1995; Van Opijnen *et al.*, 2009), targeted single gene deletion libraries were constructed to assay the functions of genes in a genome wide manner (e. g., Baba *et al.*, 2006). Screening these libraries with several dozens of chemicals or small molecule allowed assigning specific phenotypes to almost all targeted genes in key model organism genomes (e. g., Nichols *et al.*, 2011; Hillenmeyer *et al.*, 2008; Deutschbauer *et al.*, 2014). Methods to functionally characterize genes often measure the growth fitness of bacteria, but also growth independent phenotypes for instance caused by genes influencing cell morphology (Sycuro *et al.*, 2013; Peters *et al.*, 2016) or envelope biogenesis (Paradis-Bleau *et al.*, 2014) were examined. Finally, it is not only essential to analyze functions of single genes, but also the role of the proteins in the cellular context to resolve the complexity of a bacterial cell. Thus, genetic interactions were evaluated by analyzing genomic double mutants (Typas *et al.*, 2008; Butland *et al.*, 2008) and, even further, protein interactions and the bacterial pathway architecture were explored (e. g., Zeghouf *et al.*, 2004; Rajagopala *et al.*, 2014; Typas and Sourjik, 2015).

Although most methods target loss-of-function phenotypic effects by gene disruption, gain-of-function approaches by gene overexpression are a valid method to complement the holistic view of phenotypes (Boyer *et al.*, 2004; Mutalik *et al.*, 2019). While single gene deletions cannot target essential genes and only expression modulatory techniques like CRISPRi are effective to screen such genes (Peters *et al.*, 2016), overexpression is a useful supplementary

method to characterize both essential and non-essential genes (Prelich, 2012). Similar to loss-of-function screens, growth benefits or disadvantages are analyzed either in a pool (e. g., Neme *et al.*, 2017; Soo *et al.*, 2011) or in a arrayed format where each single candidate is considered on its own (e. g., Boyer *et al.*, 2004). The mechanism of underlying overexpression phenotypes can include stoichiometric imbalance, promiscuous interaction or pathway modulation (Moriya, 2015).

Depending on the application, knock-out or overexpression or a combination of both strategies can be used to elucidate the function of a gene. Large-scale screens are important for an overview, but in depth analysis of functionality including interaction partners and metabolic or enzymatic pathways combined with details about the gene structure are necessary to describe new genes.

1.4 Perspectives of this study

The study aims to characterize overlapping genes in the human pathogenic bacterium *Escherichia coli* O157:H7 str. EDL933 at different levels of functionality:

- 1) protein-coding potential
- 2) overexpression phenotype
- 3) presence of transcription start sites.

One set of putative overlapping gene candidates (OGCs) analyzed in this thesis was suggested by Landstorfer, 2014. Landstorfer used ribosomal profiling and traditional RNA-seq to describe the translome as well as the transcriptome of EHEC. The results allowed distinguishing between coding and non-coding regions in the genome. Besides annotated genes, several overlapping ORFs were identified as likely being translated (compare Section 3.1). These candidates were analyzed in this project for their ability to form proteins detectable in Western blot. Furthermore, overexpression phenotypes were examined in two different approaches in high-throughput (HT) and low-throughput (LT). Both approaches examined the influence of overlapping proteins on the growth of EHEC either in a pooled competitive or dual competitive experiment, analyzing all candidates in parallel or one candidate in comparison to a control, respectively. For a more precise description of the gene structure of the

candidates, the new NGS technique Cappable-seq (Ettwiller *et al.*, 2016) was applied to experimentally determine the transcriptional start sites of these candidates. As Cappable-seq is not restricted to a special set of genes, but analyzes +1 sites across the whole genome, an extended set of overlapping ORFs not selected by Landstorfer, 2014, became available and was investigated concerning this structural feature. Additionally, transcriptional start sites within annotated genes were screened for putative sense overlapping genes.

Finally, the overlapping gene *pop* was functionally characterized in greater depth.

2 Material and methods

2.1 Material

2.1.1 Bacterial strains and plasmids

Bacterial strains used in this thesis are listed in Table 2.1, plasmids are listed in Table 2.2 and Supplementary table S1.

Table 2.1: Bacterial strains.

Bacterial Strain	Genotype/Key Characteristics	Source
<i>E. coli</i> O157:H7 EDL933 (EHEC)	outbreak strain isolated from raw hamburger meat (Wells <i>et al.</i> , 1983)	Collection de l’Institut Pasteur (collection number CIP 106327), Nov 2003
<i>E. coli</i> O157:H7 EDL933 Δpop	translational arrest in <i>pop</i> coding sequence	this work
<i>E. coli</i> Top10	F- <i>mcrA</i> $\Delta(mrr-hsdRMS-mcrBC)$ $\Phi 80lacZ\Delta M15 \Delta lacX74 recA1 araD139 \Delta(araA-leu)7697 galU galK rpsL endA1 nupG$	Invitrogen
<i>E. coli</i> DH5 α	F- $\Phi 80lacZ\Delta M15 \Delta(lacZYA-argF)$ U169 <i>recA1 endA1 hsdR17</i> (rK-, mK+) <i>phoA supE44</i> λ - <i>thi-1 gyrA96 relA1</i>	Invitrogen
<i>E. coli</i> CC118	<i>araD139 \Delta(ara, leu)7697 \Delta lacX74 phoA\Delta 20 galE galK thi rpsE rpoB argE_{Am} recA1</i> (λ pir)	Manoil and Beckwith, 1985

Table 2.1: Continued from previous page

Bacterial Strain	Genotype/Key Characteristics	Source
<i>E. coli</i> SM10 λ pir	<i>thi thr leu tonA lacY supE</i> <i>recA::RP4-2-Tc::Mu Km^R (λpir)</i>	Simon <i>et al.</i> , 1983
<i>S. aureus</i>	milk isolate	Weihenstephan strain collection (WS 7273)
<i>S. chromogenes</i>	milk isolate	Weihenstephan strain collection (WS 7309)

Table 2.2: Plasmids.

Plasmid	Key Characteristics	Source
pBAD/Myc-HisC	pBR322 derivative, medium copy origin of replication, <i>araBAD</i> promoter, Amp ^R	Invitrogen
pBAD/SPA	pBAD based plasmid with in-frame C-terminal SPA-tag (Zeghouf <i>et al.</i> , 2004)	this work
pBAD/His	pBAD based plasmid with in-frame 6xHis-tag	this work
pSLTS	pKDTS derivative, P _{<i>araB</i>} promoter for λ -Red recombinase, Amp ^R	Kim <i>et al.</i> , 2014
pMRS101	pKNG101 derivative, oriE1, Strep ^R , oriR6K, Amp ^R	Sarker and Cornelis, 1997
pProbe-NT	pBBR1 derivative, Kan ^R , promoterless GFP reporter	Miller <i>et al.</i> , 2000

2.1.2 Chemicals and enzymes

Chemicals used in this thesis and supplier information are listed in Supplementary table S5. Polymerases and PCR reagents are purchased from New England Biolabs, Ipswich, MA, USA. Restriction endonucleases are purchased from Thermo Fisher, Waltham, MA, USA. Supplier information of further enzymes used are given in the text.

2.1.3 Primer

Primer (custom DNA oligos) are purchased from Eurofins Genomics, Ebersberg (dissolved in H₂O at 50 pmol μL⁻¹) and listed in Table 2.3 and Supplementary tables S2, S3, and S4.

Table 2.3: List of primers. Universal primers, primers for pBAD/SPA and pBAD/His construction and *pop* primer. Restriction enzyme cut sites are underlined.

Plasmid/DNA Primer	Sequence (5' → 3')
pBAD-C+165F	CAGAAAAGTCCACATTGATT
pBAD-C+494R	TGATTTAATCTGTATCAGGC
pMRS101+8708F	GACACTGAATACGGGGCAAC
pMRS101+458R	CTTATCGATGATAAGCTGTC
pProbe-NT+3944F	AAACTGCCAGGAATTGGGGAT
pProbe-NT+4164R	CGTATGTTGCATCACCTTCA
rrsh-F (16S rDNA)	AATGTTGGGTAAAGTCCCGC
rrsh-R (16S rDNA)	GGAGGTGATCCAACCGCAGG
pBAD/SPA Plasmids	Sequence (5' → 3')
SPA-tag-F- <i>Hind</i> III (Annealing)	ATCA <u>AGCTT</u> ACAAGAGAAGATGGAAAAAGAATT TCATAGCCGTCTCAGCAGCCAACCGCTTTAAGAA AATCTCATCCTCCGGGGCACTTGATTATGATATT CCA <u>ACTACTGCTAGCGAGA</u>
SPA-tag-R- <i>Sal</i> I (Annealing)	TTC <u>GTCGACCTACTT</u> GTCATCGTCATCCTTGTAG TCGATGTCATGATCTTTATAATCACCGTCATGGT CTTTGTAGTCGAGCTCACCCCTGAAAATACAAAT TCTCGCTAGCAGTAGTTGG
SPA-1F- <i>Hind</i> III	ATCA <u>AGCTT</u> ACAAGAGAAGA
SPA-198R- <i>Sal</i> I	TTC <u>GTCGACCTACTT</u> GTCAT
pBAD-His-linker-F	AGCTTACAATAGCGCCG
pBAD-His-linker-R	TCGACGGCGCTATTGTA

Table 2.3: Continued from previous page

Control Proteins	Sequence (5' → 3')
rpmH+3F- <i>Nco</i> I	GATCCATGGGGAAACGCACTTTTCAACCGT
rpmH+119R- <i>Hind</i> III	TAGAAGCTTCCTTAGAAACGGTCAGACGAG
gst+3F- <i>Nco</i> I	GATCCATGGGGAAATTGTTCTACAAACC
gst+585R- <i>Hind</i> III	TAGAAGCTTCCTTTAAGCCTTCCGCTGA
Analysis of <i>pop</i>	Sequence (5' → 3')
<u>creation of Δpop plasmid and genomic knock-out</u>	
cloning primer	
Z1307+1F- <i>Apa</i> I	ATAGGGCCCTTGGATGATAACGAGGCGCA
Z1307+1221R- <i>Spe</i> I	GGACTAGTGATTGTTTCAGCTGATTGAAG
mutation primer	
Z1307+788FmutS- <i>Mae</i> I	TAGTTGTTCTAGGTTACACCG
Z1307+788RmutS- <i>Mae</i> I	CGGTGTAACCTAGAACAATA
sequencing primer	
Z1307+578F	TGGGTGTTTCCTACCGTTTC
Z1307+1004R	TTCGATCTCTACGCGACGAT
<u>RT-PCR of Δpop</u>	
RTPCR- <i>ycbG</i> -F	AGTGTCGACCGAAAGTCAGTTCAATTTAC
RTPCR- <i>pop</i> -F	TTCGATCTCTACGCGACGAT
RTPCR- <i>pop</i> -R	TGGGTGTTTCCTACCGTTTC
RTPCRterm- <i>pop</i> -F	CGGTGTAACCTAGAACAATA
RTPCRterm-dORF-R	ATAGGGCCCTTGGATGATAACGAGGCGCA
RTPCRterm-stemloop-R	TACGTTGTAGACTTTACATC
<u>RT-PCR of Δpop</u>	
<i>pop</i> -77F- <i>Sal</i> I	AGTGTCGACCGAAAGTCAGTTCAATTTAC
<i>pop</i> +25R- <i>Eco</i> RI	AGCGTGAATTTCGATCGAAGTTAAAGGTATC

2.1.4 Media and buffer

Media and buffers (Tables 2.4 and 2.5) are prepared with ultrapure water. Media are sterilized by autoclaving at 121 °C for 15 min. Heat labile (HL) components are filter sterilized (0.22 µm pore diameter) and added after cooling. Buffers are not sterilized unless otherwise stated.

Table 2.4: Culture media.

Medium	Ingredient	Concentration
LB	tryptone	10 g L ⁻¹
	yeast extract	5 g L ⁻¹
	NaCl	5 g L ⁻¹
	optional: agar	16 g L ⁻¹
SOC medium	tryptone	20 g L ⁻¹
	yeast extract	5 g L ⁻¹
	NaCl	0.5 g L ⁻¹
	KCl	0.186 g L ⁻¹
	1 M MgCl ₂ (HL)	10 mL L ⁻¹
	2 M Glucose (HL)	10 mL L ⁻¹
M9 minimal medium	Na ₂ HPO ₄ (anhydrous)	6 g L ⁻¹
	KH ₂ PO ₄	3 g L ⁻¹
	NH ₄ Cl	1 g L ⁻¹
	NaCl	0.5 g L ⁻¹
	CaCl ₂	3 mg L ⁻¹
	1 M MgSO ₄ (HL)	1 mL L ⁻¹
	25 % glucose (HL)	8 mL L ⁻¹
	20 % casamino acid (HL)	5 mL L ⁻¹
0.5 % thiamine (HL)	0.1 mL L ⁻¹	

Table 2.5: Buffers.

Buffer	Ingredient	Concentration
10x PBS	NaCl	80 g L ⁻¹
	KCl	2 g L ⁻¹
	Na ₂ HPO ₄ (anhydrous)	14.2 g L ⁻¹
	KH ₂ PO ₄	2.4 g L ⁻¹
	pH 7.4, adjusted sterilization by autoclaving	
50x TAE buffer	Tris	242.2 g L ⁻¹
	acetic acid	57.1 mL L ⁻¹
	Na ₂ EDTA	18.61 g L ⁻¹
	sterilization by autoclaving	
3x Gel buffer	Tris	363.42 g L ⁻¹
	HCl (25 %)	130.4 mL L ⁻¹
	SDS	3 g L ⁻¹
	pH 8.45, adjusted	
10x Anode buffer	Tris	121.14 g L ⁻¹
	HCl	29.3 mL L ⁻¹
	pH 8.9, adjusted	
10x Cathode buffer	Tris	121.14 g L ⁻¹
	tricine	179.18 g L ⁻¹
	SDS	10 g L ⁻¹
	pH 8.25, not adjusted	

Table 2.5: Continued from previous page

Buffer	Ingredient	Concentration
SDS sample buffer	SDS	2 %
	Coomassie Blue G250	0.04 %
	glycerin	40 %
	Tris (pH 6.8)	200 mM
	prior to use: <i>β</i> -mercaptoethanol	2 %
10x Blotting buffer	glycin	143 g L ⁻¹
	Tris	30 g L ⁻¹
	SDS	10 g L ⁻¹
1x Blotting buffer	10x Blotting buffer	100 mL L ⁻¹
	methanol	200 mL L ⁻¹
10x TBS	Tris	12.11 g L ⁻¹
	NaCl	87.66 g L ⁻¹
	pH 8, adjusted	
1x TBS-T	10x TBS	100 mL L ⁻¹
	Tween 20	0.5 mL L ⁻¹
AP buffer	Tris	12.11 g L ⁻¹
	MgCl ₂	0.38 g L ⁻¹
	pH 9.5, adjusted	
NBT	NBT	50 mg / 1 mL 70 % DMF
BCIP	BCIP	20 mg / 1 mL DMF

2.1.5 Antibiotics and media supplements

All media supplements and antibiotics listed in Tables 2.6 and 2.7 are dissolved in MQ-H₂O and sterilized by sterile filtration (0.22 μ m pore diameter).

Table 2.6: Media supplements.

Supplement	Working Concentration
arabinose	0.002 %
L-malic acid	4 mM
L-arginine	20 mM
CsCl	20 mM
malonic acid	4 mM
acetic acid	4 mM
1-methylimidazole	20 mM
NaCl	500 mM
NaOH	4 mM
Na ₃ VO ₄	4 mM
sodium salicylate	0.16 mM
HClO ₄	32 μ M
phytic acid	32 μ M
1,2-propanediol	100 mM
1-propanol	20 mM
pyridoxine HCl	20 mM
glucose	1 %
ZnCl ₂	0.8 mM
<i>Staphylococcus</i>	25 % (<i>S. aureus</i> (75 %) and <i>S. chromogenes</i> (25 %) mixture)
bicine (pH 8.5)	100 mM
MES (pH 5.8)	100 mM
MOPS (pH 7.4)	100 mM

Table 2.7: Antibiotics.

Antibiotic	Working Concentration
ampicillin	100–120 $\mu\text{g mL}^{-1}$
streptomycin	30 $\mu\text{g mL}^{-1}$
kanamycin	30 $\mu\text{g mL}^{-1}$

2.1.6 Length marker and commercial kits

Tables 2.8 and 2.9 list DNA and protein length standards as well as utilized reaction kits.

Table 2.8: DNA and protein length markers.

Type	Marker	Source
DNA	100 bp DNA ladder	New England Biolabs, Ipswich, MA, USA
	1 kb DNA ladder	New England Biolabs, Ipswich, MA, USA
	1 kb Plus DNA ladder	New England Biolabs, Ipswich, MA, USA
Protein	Spectra TM Multicolor Low Range	Thermo Fisher, Waltham, MA, USA

Table 2.9: Kits.

Kit	Source
GenElute Plasmid Miniprep Kit	Sigma-Aldrich, St. Louis, MO, USA
GenElute Gel Extraction Kit	Sigma-Aldrich, St. Louis, MO, USA
GenElute PCR Clean-up Kit	Sigma-Aldrich, St. Louis, MO, USA
TruSeq DNA PCR-Free Library Prep	Illumina, San Diego, CA, USA
High Sensitivity DNA Kit	Agilent, Santa Clara, CA, USA
RNA 6000 Nano Kit	Agilent, Santa Clara, CA, USA
Qubit dsDNA HS Assay Kit	Thermo Fisher, Waltham, MA, USA

2.2 Cultivation and storage of bacteria

Escherichia coli is cultivated aerobically at 37 °C unless stated otherwise. Growth in liquid cultures is carried out in lysogeny broth (LB) or M9 minimal medium (MM) with shaking at 150 rpm. Solid cultures are performed on LB agar plates. The culture medium is supplemented with the appropriate additives at the given working concentration according to Tables 2.6 and 2.7. For an overnight (ON) culture, LB medium is inoculated with a single colony and cultivated at least 12 h to stationary phase. Bacteria plates are stored at 4 °C for a maximum of four weeks. Glycerol stocks (1:1 mixture of 80 % glycerol and an ON culture) are used for long-term storage of bacteria at –80 °C.

2.3 Isolation of nucleic acids

2.3.1 Isolation of plasmid DNA

Plasmid DNA is isolated using the GenElute Miniprep kit according to the manufacturer’s instructions. A volume of 4–8 mL liquid culture is used for purification depending on cells and plasmids. Purified DNA is eluted in 50 µL nuclease free H₂O.

2.3.2 Isolation of genomic DNA

Isolation of genomic DNA of EHEC is carried out on ice. Cells of 5 mL of an ON culture are pelleted by centrifugation (5 min, 14 000 ×g, 4 °C). The cell pellet is dissolved in 700 µL Tris/EDTA solution (10 mM Tris, 1 mM EDTA, pH 8) and cells are disrupted mechanically (FastPrep-24, 100 µL 0.1 mm zirconia beads, three times 6.5 m s⁻¹ for 45 s). The cell debris is collected by centrifugation (14 000 ×g, 5 min, 4 °C). Subsequent centrifugation steps are performed with identical settings. DNA in the supernatant is extracted with Roti-Phenol/chloroform/isoamyl alcohol (25:24:1 ready-to-use mixture). One volume (Vol) of extraction solution is mixed thoroughly with the supernatant and centrifuged for 5 min. This step is performed a second time with the supernatant. The top layer is recovered and DNA is precipitated with 2 Vol 100 % ethanol and 0.1 Vol 5 M NaOAc (pH 5) at –20 °C for at least 30 min. DNA is collected by centrifugation for 10 min and washed two times with 1 mL 70 % ethanol (5 min incubation at room temperature, 5 min centrifugation). The

nucleic acid pellet is dried for 10–15 min at 30 °C and dissolved in 30 μ L nuclease free H₂O. Subsequent RNA digestion (Section 2.4.1) is followed by another Phenol/chloroform/isoamyl alcohol extraction.

2.3.3 Isolation of RNA

Cell material of a 50 mL liquid culture is harvested at defined time points based on optical density values at 600 nm (OD₆₀₀) by centrifugation (9000 \times g, 5 min, 4 °C, Table 2.10). Cell pellets are frozen in liquid nitrogen and stored at –80 °C.

Table 2.10: Optical densities, culture volumes and Trizol volumes for RNA isolation.

	Exponential Phase			Early Stationary Phase		
	OD ^a	harvest ^b	Trizol ^c	OD ^a	harvest ^b	Trizol ^c
LB	0.3	25 mL	2.4 mL	3.5	1.8 mL	6 mL
M9 minimal medium	0.2–0.3	25 mL	2.4 mL	1.6–1.8	1.8 mL	6 mL
LB + malic acid	0.3	25 mL	2.4 mL	3.9–4	1.8 mL	6 mL
LB + NaCl	0.2–0.3	25 mL	2.4 mL	1	1.8 mL	2.4 mL

^a, OD of the culture at 600 nm in the indicated growth phase

^b, amount of cells harvested for RNA isolation

^c, amount of Trizol used for RNA isolation

All steps for RNA isolation with Trizol are conducted on ice unless otherwise stated. Cell pellets are resuspended in an appropriate amount of cooled Trizol (Table 2.10) and disrupted mechanically. For this purpose, resuspended cells are split into four aliquots with 600 μ L or six aliquots with 1000 μ L and 400 μ L 0.1 mm Zirconia-beads are added for bead-beating (FastPrep-24, three times with 6.5 m s⁻¹ for 45 s, 5 min incubation on ice between the runs). Once cells are disrupted, they are incubated for 5 min at room temperature. Then, 0.2 Vol of cooled chloroform per initial amount of Trizol is added, samples are mixed vigorously and incubated for 5 min at room temperature. Phase separation is carried out by centrifugation (12 000 \times g, 15 min, 4 °C). The aqueous upper phase is recovered and RNA contained in this layer is precipitated with 0.1 Vol of aqueous phase of 3 M NaOAc, 1 μ L glycogen and

1 Vol (of mixture) of 2-propanol (RT) for 1 h at -20°C . RNA is collected by centrifugation ($12\,000 \times g$, 10 min, 4°C) and the pellet is washed twice with 1 mL 80 % Ethanol ($12\,000 \times g$, 5 min, 4°C). The remaining alcohol is collected (20–30 s centrifugation) and removed. The RNA pellet is dried at room temperature for about 15 min and subsequently dissolved in 40 μL RNase free H_2O .

2.4 *In vitro* processing of nucleic acids

2.4.1 RNase digest

RNA contamination in DNA samples is digested with 0.1 Vol RNaseA (10 mg mL^{-1} , Thermo Fisher). The mixture is incubated for 30 min at 30°C . DNA is extracted again with phenol, chloroform and isoamyl alcohol as described in Section 2.3.2.

2.4.2 DNase digest

DNA contamination in RNA samples is digested with AmbionTM Turbo DNase according to the manufacturers instructions (Invitrogen, Thermo Fisher). Briefly, a maximum amount of 10 μg total RNA is incubated in 1x Turbo DNase buffer and 2 U Turbo DNase in a total volume of 50 μL for at least 45 min at 37°C . DNase is inactivated subsequently at 75°C for 10 min. EDTA is added at a final concentration of 15 mM prior this heating step to prevent chemical scission of RNA during inactivation. The remaining RNA is precipitated with Ethanol (100 %)/NaOAc (3 M, pH 5.2)/glycogen (in a ratio of 690 μL /27.6 μL /1 μL) at -20°C ON. Precipitated RNA is collected by centrifugation ($12\,000 \times g$, 20 min, 4°C) and washed once with 80 % Ethanol ($12\,000 \times g$, 20 min, 4°C). The pellet is dried at room temperature and resuspended in 10 μL RNase free H_2O . Finally, DNA contamination was analyzed with a standard Taq-PCR (Section 2.4.4) using primers rrsh-F and rrsh-R for the housekeeping gene *rrsh* (16S rDNA gene).

2.4.3 cDNA synthesis

SuperScriptTM III Reverse Transcriptase (Invitrogen, Thermo Fisher) is used for first strand cDNA (complementary DNA) synthesis according to the manufacturer. Briefly, 500 ng total

DNA-depleted RNA is incubated in a volume of 13 μL with 1 μL dNTP mix (10 mM) and 1 μL of a gene specific reverse primer (10 μM) or a random nonamer primer (Sigma-Aldrich, 50 μM), as indicated, for 5 min at 65 °C. After this, 5x First-Strand buffer (4 μL), 1 M DTT (1 μL), 20 U/ μL SUPERase In RNase Inhibitor (1 μL) and 200 U/ μL SuperScript III Reverse Transcriptase (1 μL) are added and mixed by pipetting. The reaction mixture is incubated at 25 °C for 5 min, at 50 °C for 60 min, and finally at 70 °C for 15 min. The cDNA is stored at -20 °C. To verify specificity of subsequent analysis of the cDNA, ‘no reverse transcription’ controls are prepared. cDNA synthesis is performed as described apart from the reverse transcriptase, which is replaced by an equivalent amount of H₂O.

2.4.4 Polymerase chain reaction (PCR)

In vitro amplification of DNA is performed in a polymerase chain reaction (PCR) using a thermo stable polymerase (*Taq* or Q5[®] DNA polymerase, NEB) according to Table 2.11. The thermal cycling is conducted in a thermocycler (Flexcycler, Analytik Jena; Primus 96 advanced, Peqlab) as listed in Table 2.12. The annealing temperature T_a is calculated based on the melting temperature T_m of the primers used. The elongation time depends on the polymerase and the length of the amplified fragment, as indicated. The success of the PCR is verified with agarose gel electrophoresis (Section 2.4.9). If necessary, PCR products are purified with the GenElute PCR Clean-up Kit or GenElute Gel extraction Kit according to the manufacturer.

Standard PCR Standard PCRs are performed for cloning applications with Q5 polymerase. Purified genomic DNA (100 ng) or plasmid DNA (10 ng) is used as DNA template. The thermal cycling steps denaturation, annealing, and elongation are repeated 30 times.

Colony PCR Colony PCR is performed to confirm insert integration after cloning and transformation. *Taq* DNA Polymerase is used for this application. Parts of the colonies to be tested are added as DNA template. The initial denaturation step is increased to 15 min to break up the cells and expose the DNA for amplification. The thermal cycling steps denaturation, annealing, and elongation are repeated 30 times.

RT-PCR Reverse transcription (RT-)PCR is performed with *Taq* DNA Polymerase and cDNA (1 μ L of 500 ng reverse transcribed total RNA). The thermal cycling steps denaturation, annealing, and elongation are repeated 20 times.

Table 2.11: Composition of reaction mix for Taq or Q5 PCR (25 μ L).

Taq PCR	Working Concentration
10x ThermoPol buffer	1x
10 mM dNTPs	0.2 mM
10 μ M primer 1	0.5 μ M
10 μ M primer 2	0.5 μ M
DNA template	variable
5 U/ μ L <i>Taq</i> Polymerase	0.625 U
Q5 PCR	Working Concentration
5x Q5 reaction buffer	1x
5x Q5 high GC enhancer (optional)	(1x)
10 mM dNTPs	0.2 mM
10 μ M primer 1	0.5 μ M
10 μ M primer 2	0.5 μ M
DNA template	variable
2 U/ μ L Q5 Polymerase	0.5 U

2.4.5 Quantitative polymerase chain reaction (qPCR)

Quantitative polymerase chain reaction (qPCR) is conducted to quantify mRNA levels of *pop* relative to the 16S rRNA gene. One microliter of reverse transcribed cDNA (primer RTPCR-*pop*-R for *pop*, random primer for 16S rRNA gene) is mixed with 12.5 μ L SYBR Select Master Mix (Applied Biosystems) and 0.5 μ L of forward and reverse primers (50 μ L) in a total volume of 25 μ L. Amplification of *pop* is performed with primers RTPCR-*pop*-F and RTPCR-*pop*-R; the 16S rRNA gene was amplified with primers rrsh-F and rrsh-R. The qPCR is carried out with a CFX96 Touch Real-Time PCR Detection System (BioRad) with the

Table 2.12: PCR temperature programs for Taq and Q5 PCRs.

Cycles	Temperature Step	Taq	Q5
20-30x	Initial denaturation	95 °C, 2 min	98 °C, 2 min
	Denaturation	95 °C, 30 s	98 °C, 10 s
	Annealing	$T_a = T_m^* - 5 \text{ °C}$, 30 s	$T_a = T_m^* + 3 \text{ °C}$, 30 s
	Elongation	68 °C, 60 s/kbp	72 °C, 30 s/kbp
	Final elongation	68 °C, 5 min	72 °C, 2 min

T_m^* : lower melting temperature of the used primers.

following cycling parameters: 95 °C (5 min, initial denaturation), 40 cycles of denaturation, annealing and elongation at 95 °C (15 s), 61 °C (30 s) and 72 °C (30 s). A melting curve is recorded for quality control (61 °C to 95 °C in 0.5 °C steps for 5 s). Samples of three biological replicates are analyzed with three technical replicates for each sample. Specificity of cDNA amplification (e. g., exclude DNA contamination) is verified with ‘no reverse transcription’ controls for every RNA sample. The $\Delta\Delta C_q$ method is used to quantify fold-change of *pop* expression in relation to the 16S rRNA as reference in different growth conditions (Pfaffl, 2001). Statistical significance of differences in fold change values is calculated using a one-tailed Welch two sample t-test at significance level $\alpha = 0.05$.

2.4.6 Restriction digest

PCR products and plasmids are processed with restriction endonucleases (Thermo Fisher) following the protocol in Table 2.13. The reaction buffer is selected for the restriction enzymes used. The reaction mixture is incubated at 37 °C for at least 1 h. The enzymes are subsequently inactivated enzyme-specifically at 65 °C or 80 °C for 20 min.

For cloning applications, digested plasmids and PCR products are purified with the GenElute Gel extraction Kit and GenElute PCR clean-up Kit, respectively.

Table 2.13: Restriction digest of nucleic acids.

Component	PCR Product	Plasmid DNA
DNA	10 μ L (unpurified)	1 μ g
Reaction buffer		1x
Restriction enzyme 1		10 U
Restriction enzyme 2 (optional)		(10 U)
Reaction volume	32 μ L	20 μ L

2.4.7 Ligation

To ligate a PCR fragment into a vector, 20–100 ng linearized plasmid is used in a 20 μ L reaction. The digested PCR product is added at a molar ratio of 1:3 to 1:5 (vector:insert) as well as 1 U T4 DNA Ligase (Thermo Fisher) and 1x T4 DNA Ligase Buffer. For self-circularization of plasmid DNA, 10–50 ng plasmid, 5 Units T4 DNA Ligase and 1x T4 DNA Ligase Buffer are used in a 50 μ L reaction. The ligation reaction is performed either at 22 °C for 1 h or at 4 °C ON and is stopped by heating (10 min, 65 °C).

2.4.8 QuikChange mutagenesis

QuikChange mutagenesis is performed to introduce position-specific mutations in plasmids based on a protocol for a site directed mutagenesis published by Laible and Boonrod, 2009. In brief, one or in some cases two complementary mutagenesis primers are used in a Q5[®] PCR reaction to amplify the whole plasmid. Methylated template plasmid DNA is degraded with 10 U *DpnI* (Thermo Fisher) for 22 h. Undigested plasmids are subsequently precipitated with 10 Vol ice cooled 1- or 2-butanol for 1–5 min at room temperature and collected by centrifugation (30 min, 12 000 \times g, 4 °C). The pellet is washed once with 70% ethanol and centrifuged with the same settings. The pellet is air-dried and resuspended afterwards in 10 μ L nuclease free H₂O. Five microliter are used to transform competent *E. coli* Top10 cells (Section 2.5).

2.4.9 Separation of nucleic acids with agarose gel electrophoresis

PCR products, genomic DNA or total RNA are separated on an agarose gel to verify success of the PCR reaction or the isolation processes. Agarose is dissolved in 1x TAE buffer under heating. The concentration of agarose (1–2%) is adjusted to the nucleic acids to be separated. RedSafe Nucleid acid staining solution (1:20000 diluted) is dispensed in the slightly cooled agarose solution. Samples are mixed with 1x loading dye and loaded on the solidified agarose gel, which is covered with 1x TAE. Additionally, a suitable DNA marker as reference is loaded. Nucleid acids in the gel are separated (13 V cm^{-1}) for 30–45 min and visualized under UV light.

2.4.10 Separation and Quality control of nucleic acids with capillary gel electrophoresis

High resolution separation and quality control of nucleic acid for NGS applications is performed with capillary gel electrophoresis on a Bioanalyzer 2100 according to the manufacturer's instructions. DNA samples ($5\text{--}500 \text{ pg } \mu\text{L}^{-1}$) and RNA samples ($5\text{--}500 \text{ ng } \mu\text{L}^{-1}$) are analyzed with the High Sensitivity DNA Kit and RNA 6000 Nano Kit, respectively. The quality of RNA samples is measured via the RNA integrity number (RIN) and samples with $\text{RIN} < 9$ are discarded and not used for further sample preparations.

2.4.11 Determination of nucleic acid concentration and purity

Concentration of nucleic acids is measured for standard applications with the Nanodrop 1000 UV/Vis Spectrophotometer and require 260/280 values of 1.8 for DNA and 2.0 for RNA. Precise concentration measurements are performed with the Qubit dsDNA HS Assay Kit on a Qubit 2.0 Fluorometer according to the manufacturer's instructions.

2.4.12 Sanger sequencing

Sanger sequencing of plasmid DNA or purified PCR products is performed to verify insert sequences or to evaluate competitive growth (Section 2.8.2)

2.5 Transformation

2.5.1 Preparation of electrocompetent cells and electrotransformation

LB medium (100 mL) is inoculated with 2 mL of an ON culture and cultivated until an optical density of $OD_{600} = 0.4 - 0.6$. Cells are cooled on ice for 10 min and harvested afterwards by centrifugation ($3750 \times g$, 10 min, 4°C). Subsequent centrifugation steps are conducted with the same settings. The pellet is resuspended successively in 50 mL and 25 mL cooled H_2O followed by centrifugation. The last washing step of the pellet is performed in 20 mL 10% cooled glycerol. After centrifugation, the pellet is resuspended in 1 mL 10% cooled glycerol, quick-frozen in 40 μL aliquots in liquid nitrogen and stored at -80°C .

For electroporation, 10 ng purified plasmid or 10 μL desalted ligation mixture are mixed with cooled electrocompetent cells. The cells are transferred in a precooled 0.2 cm electroporation cuvette and an electric pulse (2.5 kV) is applied. Prewarmed SOC medium (960 μL , 37°C) is added immediately and cells are incubated for 1 h at 37°C and 150 rpm. Depending on the transformed sample, 100 μL for purified plasmids or the whole transformation mixture for ligation products is plated on LB-agar plates supplemented with appropriate antibiotics. To check successful ligation of insert and vector, a colony *Taq*-PCR is performed from colonies grown after transformation of ligation products (Section 2.4.4).

2.5.2 Preparation of chemocompetent cells and chemotransformation

LB medium (300 mL) is inoculated with 300 μL of an ON culture and cultivated until an optical density of $OD_{600} = 0.4 - 0.6$. The cells are cooled on ice for 15 min and harvested by centrifugation ($1800 \times g$, 10 min, 4°C). Subsequent centrifugation steps are conducted with the same settings. The pellet is resuspended in 60 mL cooled 1 M CaCl_2 and incubated on ice for 30 min. After centrifugation, the pellet is resuspended in 12 mL 0.1 M 20% glycerol and aliquots of 200 μL are quick-frozen in liquid nitrogen and stored afterwards at -80°C .

For chemical transformation, 10 ng purified plasmid or 10 μL ligation mixture are mixed with 200 μL chemocompetent cells and incubated on ice for 30 min. Cells are first heated at 42°C for 90 s, then incubated on ice for 2 min. Prewarmed SOC medium (800 μL , 37°C) is added and cells are incubated at 37°C and 150 rpm for 1 h. Depending on the trans-

formed sample, 100 μ L for purified plasmids or the whole transformation mixture for ligation products is plated on LB-agar plates supplemented with appropriate antibiotics. To check successful ligation of insert and vector, a colony *Taq*-PCR is performed from colonies grown after transformation of ligation products (section 2.4.4).

2.6 Genetic modification of *E. coli* strains

2.6.1 Construction of promoter test strains

Activity of promoter sequences is analyzed exogenously on the promoterless GFP vector pProbe-NT. Putative promoter sequences between 50 bp and 162 bp length are cloned in pProbe-NT using primers listed in Supplementary table S4. Genomic DNA or the plasmid pMRS101 Δ OGC 15 serve as DNA templates in PCRs. Promoter plasmids are verified by Sanger sequencing and transformed into *E. coli* Top10 for promoter activity testing (Section 2.7).

2.6.2 Construction of protein-tag expression strains

Novel pBAD based vectors are created for heterologous protein expression and immunological detection of proteins via Western blot (Section 2.10), pBAD/SPA and pBAD/His. For this, pBAD/myc-HisC is modified between the cut sites of restriction endonucleases *Hind*III and *Sal*I. In case of pBAD/His a seven amino acid linker sequence is used to replace the myc-cassette of pBAD/myc-HisC. The in-frame 6xHis tag remains unchanged. For pBAD/SPA, the sequence of SPA consisting of a calmodulin binding peptide, a TEV cleavage site, and a 3xFLAG epitope tag is designed after Zeghouf *et al.*, 2004, and replaces the myc-cassette of the original vector.

To construct these plasmids, 22.5 μ M or 4 μ M of each of the primers SPA-tag-F-*Hind*III and SPA-tag-R-*Sal*I or pBAD-His-linker-F and pBAD-His-linker-R are annealed in a 25 μ L reaction containing 10 mM Tris-HCl, pH 7.5, 100 mM NaCl and 1 mM EDTA. The reaction mixtures are heated at 90 $^{\circ}$ C or 70 $^{\circ}$ C, respectively, and cooled down slowly at room temperature. The His-linker annealing results in sticky end fragments suitable for cloning. The SPA-construct is used in a Q5-PCR, where the primers SPA-1F-*Hind*III and SPA-198R-*Sal*I

are added after five cycles to amplify the fragment to appropriate DNA amounts. Afterwards, the PCR fragment is digested with *HindIII* and *SalI*. Both fragments are cloned within the indicated cut sites into the original pBAD/myc-HisC vector using previously described techniques (Sections 2.4.6, 2.4.7 and 2.5).

PCR fragments of control genes (*rpmH* and *gst*) are cloned into each of the novel expression vectors by means of primers listed in Table 2.3. Overlapping gene candidates are cloned into pBAD/SPA using forward primers from Zehentner, 2015, and reverse primers listed in Supplementary table S2. Plasmids are transformed into *E. coli* Top10 for expression analysis if not stated otherwise.

2.6.3 Construction of translationally arrested plasmid knock-out transformants

Translationally arrested plasmids of selected overlapping gene candidates are created by cloning a mutation cassette constructed with an overlap extension PCR. Mutation fragments are amplified from genomic DNA in a Q5-PCR using one cloning primer and the corresponding forward or reverse mutation primer listed in Supplementary table S3. Both fragments are used as DNA template in the subsequent overlap extension PCR with the cloning primers. The purified mutation cassette (Δ OGC) is cloned in the plasmid pBAD/myc-HisC (selection with ampicillin). Alternatively, site directed QuikChange mutagenesis described in Section 2.4.8 is applied to create mutant plasmids. Template plasmids constructed in Zehentner, 2015, are sequenced with pBAD plasmid primers. If inserts have been verified, plasmids are used for mutagenesis. In any case, mutated plasmids were isolated from *E. coli* Top10 after cloning, sequenced and transformed into *E. coli* O157:H7 EDL933 for competitive growth.

2.6.4 Construction of translationally arrested genomic knock-out mutants

The translationally arrested genomic knock-out mutant *E. coli* O157:H7 EDL933 Δ *pop* is created with the plasmid pMRS101 from Sarker and Cornelis, 1997. The single base mutation introduced into the genome leads to a stop codon in the open reading frame of *pop* while the mutation is silent in the annotated mother gene.

Mutation fragments are amplified from genomic DNA in a Q5-PCR using one cloning primer and the corresponding forward or reverse mutation primer listed in Table 2.3. Both

fragments are used as DNA templates in a subsequent overlap extension PCR with the cloning primers. The purified mutation cassette Δpop is cloned in the plasmid pMRS101 using *ApaI* and *SpeI* (selection with ampicillin). The plasmid pMRS101+ Δpop is isolated and Sanger sequenced with both pMRS101 plasmid primers. A restriction digest with *NotI* is performed to remove the high copy ori of pMRS101. The linearized plasmid is self-circularized to the π -protein dependent low copy plasmid pKNG101+ Δpop . In general, maintenance of pKNG101-plasmids rely either on cells expressing the *pir* gene, which enables replication, or on homologous recombination and integration of the plasmid into the genome if the cell do not express the π -protein (Kaniga *et al.*, 1991). pKNG101+ Δpop is propagated in *E. coli* CC118 λ pir (selection with streptomycin) and isolated according to Section 2.3.1. The conjugation strain *E. coli* SM10 λ pir is transformed with pKNG101+ ΔOGC and used in a subsequent conjugation, where 500 μ L of an ON culture of these cells are mixed with the same amount of an ON culture of *E. coli* O157:H7 EDL933+pSLTS (selection marker ampicillin, temperature sensitive ori) and cultivated on LB plates (24 h, 30 °C). Successful conjugation and integration of pKNG into the genome via homologous recombination is verified in a colony PCR of ampicillin/streptomycin double resistant cells using primers pMRS101+8708F and the forward sequencing primer of *pop*. A positive clone is used for loop-out of the mutation plasmid. For this, EHEC is cultivated in LB at 30 °C at 150 rpm until an optical density of $OD_{600} = 0.5$ and counter-selected on saccharose agar (modified LB without NaCl containing saccharose) supplemented with 0.02% arabinose to induce the λ red recombination system on pSLTS. While cells with integrated pKNG101+ Δpop express the enzyme levansucrase which converts saccharose into toxic levans accumulating in the periplasm of Gram-negative bacteria (Reyrat *et al.*, 1998), cells which performed a second homologous recombination event can survive saccharose stress. Potential mutants are screened for streptomycin sensitivity and ampicillin resistance and a colony PCR using *pop* sequencing primers is performed. The PCR product is purified and Sanger sequenced to verify introduction of the desired mutation into the chromosome. *E. coli* O157:H7 EDL933 Δpop is cultivated at 37 °C to clear the cells from the plasmid pSLTS. Glycerin stocks are prepared for storage of the mutants.

2.7 Promoter activity analysis

The activity of putative promoter sequences is tested with a *gfp* promoter activity test. ON cultures of *E. coli* Top10, *E. coli* Top10 containing pProbe-NT and *E. coli* Top10 containing pProbe-NT+promoter-OGC variants are prepared. Ten milliliter LB medium, M9 medium or LB medium supplemented with indicated stressors are inoculated 1:100 with the prepared pre-cultures. Kanamycin is added to cultivate plasmid carrying cells. Equal volumes of cells are harvested (2 min, 6600 ×g, 4 °C) at an optical density of $OD_{600} = 0.5-0.6$. The cell pellet is washed once in 1–2 mL 1x PBS and finally resuspended in 1 mL 1x PBS. Cells are diluted 1:10 and the optical density is measured at 600 nm. Fluorescence of 4x 200 µL diluted cells is measured using a Wallac Victor3 multilabel reader (Perkin Elmer, excitation: 485 nm, emission: 535 nm, measuring time: 1 s). Mean fluorescence is normalized to $OD_{600} = 1$ and self-fluorescence of *E. coli* is subtracted. The experiment is conducted in biological triplicates. Mean values and standard deviations are calculated and statistical significances between different promoter constructs or stress conditions are evaluated with two tailed Welch two sample t-tests (significance level $\alpha = 0.05$). Significance of the difference between fluorescence of one promoter construct in two different growth phases is calculated with a two tailed paired t-test (significance level $\alpha = 0.05$).

2.8 Phenotypic analysis

2.8.1 High-throughput analysis

High-throughput (HT) overexpression phenotyping (Figure 3.4) is carried out according to Zehentner, 2015, and briefly described in the following paragraph.

Independent bacteria pools with *E. coli* O157:H7 EDL933 containing pBAD+OGC x variants are created. For this, cells are cultivated in 200 µL LB supplemented with ampicillin ($100 \mu\text{g mL}^{-1}$) in microtiter plates. Optical densities are measured (Wallac Victor3 multilabel reader, Perkin Elmer) and cultures are diluted to $OD_{600} = 0.5$. Equal amounts of the cells are combined and the cell count of the mixture is determined with plate counting. Bacteria pools are stored as glycerol stocks. LB medium (50 mL in 200 mL conical flasks) supplemented with $100 \mu\text{g mL}^{-1}$ ampicillin and different stressors (Table 2.14, Section 2.1.5) is inoculated

with approximately 1×10^5 cells representing about 500 cells per OGC variant. Cultures are incubated for 22 h and expression of the OGC-proteins is induced at time points t_{0h} and $t_{6.5h}$ with L-arabinose. Plasmids are isolated after cultivation. One microgram plasmid is linearized with *NcoI* and further fragmented by ultrasonication to 350 bp (Covaris settings: Peak Incident Power 140 W, Duty Factor 10 %, Cycles Per Burst 200, treatment time 80 s). Plasmid fragments are prepared for sequencing using the TruSeq DNA PCR-Free Library Prep Kit according to the manufacturer’s protocol. Samples of 23 culture conditions as well as the timepoint zero input sample are pooled and quantified with the Perfecta NGS Library Quantification Kit for Illumina (Quanta Bioscience). The library is sequenced paired end (25 bp) on an Illumina MiSeq using the MiSeq reagent Kit v2 (50 cycles) according to the instruction manual. Sequencing reads are processed on the Galaxy platform with Fastq Groomer, mapped to an artificial sequence containing all OGCs with Bowtie2 using standard settings and forward and reverse reads are merged with MergeSamFiles. RPKM (reads per kilobase per million mapped reads) values for each candidate gene and condition is determined using the Artemis browser (Rutherford *et al.*, 2000) and normalized to z -scores with equation 2.1.

$$z_{i,k} = \frac{x_{i,k} - x_i}{\sigma_i} \quad (2.1)$$

with $x_{i,k}$ the RPKM value of candidate i in condition k , x_i the mean RPKM and σ_i the standard deviation of the candidate in all conditions.

Pearson’s product-moment correlation coefficients are calculated between RPKM values of biological and technical replicates. Furthermore, Spearman’s rank correlation coefficients are determined between all RPKM values of one candidate between different biological replicates.

2.8.2 Low-throughput analysis

Low-throughput phenotypic analysis is conducted by means of single competitive growth assays. For this, ON cultures of two competing strains, EHEC pBAD+OGC and EHEC pBAD+ Δ OGC, are diluted to an optical density of $OD_{600} = 1$ and mixed in equal amounts. An appropriate amount of the mixture is centrifuged and the cell pellet is stored as input reference sample at -20°C for subsequent plasmid isolation. LB medium ($100 \mu\text{g mL}^{-1}$

Table 2.14: Culture conditions used in HT phenotyping.

1 LB (without stress)	8 1-methylimidazole	14 phytic acid
2 glucose	9 NaCl	15 1,2-propanediol
3 L-malic acid	10 NaOH	16 1-propanol
4 L-arginine	11 Na ₃ VO ₄	17 pyridoxine HCl
5 CsCl	12 sodium salicylate	18 <i>Staphylococcus</i>
6 acetic acid	13 HClO ₄	19 ZnCl ₂
7 malonic acid		

ampicillin) with and without selected stressors (concentration of stressors, Section 2.1.5) is inoculated with 100 μ L of a 1:300 dilution of the cell mixture. The cultures are incubated for 22 h and protein expression is induced with L-arabinose at two time points (t_{0h} and $t_{6.5h}$). Plasmids are isolated from cultivated cells as well as from the time point zero sample and Sanger sequenced with the primer pBAD-C+165F. Bacteria proportions are determined by calculating the amount of wild type and mutated plasmid. For this purpose, fluorescence signals at the mutated position(s) of the plasmids are measured. Percentage of wild type or mutant plasmids are calculated according to equation 2.2.

$$\%_{Wt} = \frac{Wt}{Wt + Mt} \quad \text{and} \quad \%_{Mt} = \frac{Mt}{Wt + Mt} \quad (2.2)$$

with Wt and Mt representing peak heights of mutated positions in wild type and mutant plasmids in sequencing electropherograms; values of more than one mutated position are averaged. The experiment is conducted at least in biological triplicates and solely replicates with adequate and consistent input ratios are used for further analysis. Peak height ratios are normalized to an 1:1 input ratio of the reference sample for visualization purposes. Mean values and standard deviations are calculated. Statistical significance is tested with two-tailed paired t-test between wild type ratios of input and cultured samples at a significance level $\alpha = 0.05$. Significance of wild type and mutant ratio differences within a sample are calculated with a Welch two sample t-test (significance level $\alpha = 0.05$). It is assumed for statistical calculation that samples follow a normal distribution.

2.8.3 Competitive growth with genomic knock-out mutants

ON cultures of EHEC wild type and EHEC Δpop are used for competitive growth. Adjusted cell numbers ($OD_{600} = 1$) are mixed in equal amounts in a total volume of 1 mL. An aliquot of 500 μ L is pelleted and stored at -20°C as t_0 reference sample. Ten milliliter LB medium supplemented with cultivation stressors (concentration of stressors, Section 2.1.5) are inoculated with 100 μ L of an 1:300 dilution of the cell mixture. Cultures are incubated for 18 h. After cultivation, 500 μ L cells are harvested and resuspended in sterile H_2O . A colony PCR is performed using 5 μ L of resuspended cells as template and *pop* specific sequencing primer (Section 2.4.4, Table 2.3). PCR products are purified and sequenced with the primer Z1307+578F. Bacteria proportions are determined as described in Section 2.8.2. Statistical significance was determined with Welch two sample t-test between mutant and wild type cell ratios in the tested conditions and with paired t-test between wild type ratios of input and cultured samples at significance level $\alpha = 0.05$. Normal distribution of samples is assumed.

2.9 Transcriptional start site determination

Transcriptional start sites are determined with the recently published approach Cappable-seq (Ettwiller *et al.*, 2016). The method is based on the separation of RNA species according to their phosphorylation status at the 5' end. While the most prevalent RNA species (processed rRNAs and tRNAs) are monophosphorylated, less abundant unprocessed mRNAs have a 5' triphosphate. This property is used to mark mRNAs, extract them from total RNA samples and sequence them to determine the transcriptional start site at single base resolution. TagRNA-seq is used during sample preparation to increase accuracy of TSS identification (Innocenti *et al.*, 2015). In this method, monophosphorylated RNA fragments are labeled with a different sequence tag to distinguish contaminating processed RNAs from Cappable-enriched mRNAs.

Data evaluation of Cappable-seq is conducted with individually built JAVA programs, R and bash scripts. Code details are shown, where necessary.

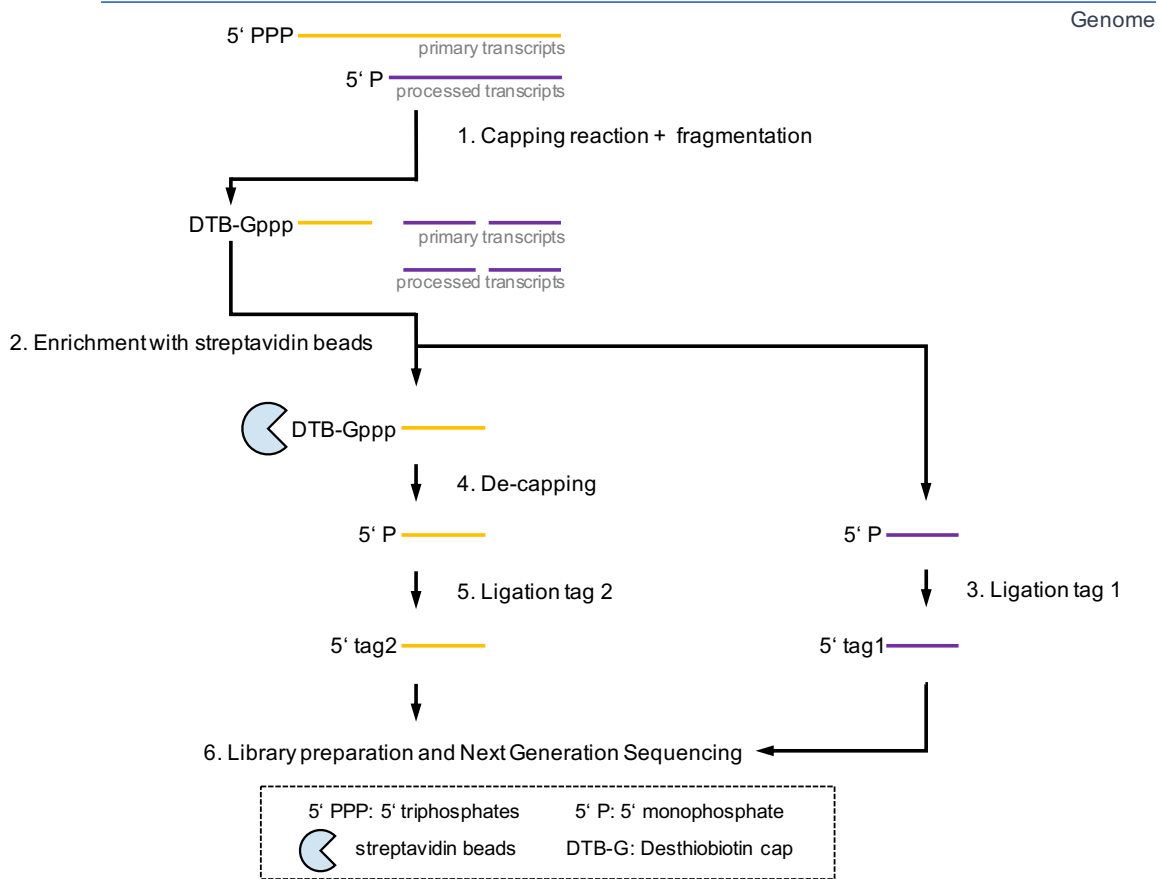
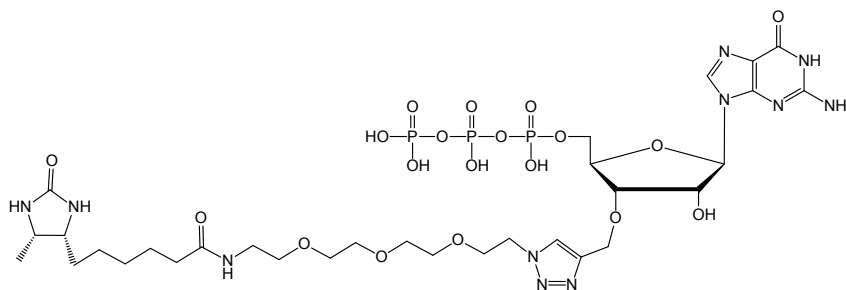
2.9.1 Determination of bacterial growth phases

Transcriptional start sites are analyzed in four growth conditions and two growth phases therein. To define points for cell harvest, growth curves are recorded. For this purpose 50 mL LB medium (pure, supplemented with 4 mM L-malic acid or supplemented with 500 mM NaCl) or M9 medium are inoculated with an EHEC ON culture to an optical density of $OD_{600} = 0.03$ in case of LB or with 500 μ L overnight culture diluted to an optical density of $OD_{600} = 3$ in case of M9 to maintain the amount of LB transferred to minimal medium in all experiments constant. The cultures are cultivated aerobically in 500 ml baffled flasks and optical densities at 600 nm are measured in appropriate time intervals until growth reaches stationary phase.

2.9.2 Cappable-seq sample preparation

Total DNase depleted RNA is applied to Cappable-seq. In the following paragraph, sample preparation and sequencing carried out by vertis Biotechnologie AG, Freising, is described briefly (Figure 2.1a).

In a first step, 5' triphosphorylated RNA is reversible capped with DTB-GTP (3' desthiobiotin-TEG-guanosine 5' triphosphate, Figure 2.1b) by the vaccinia capping enzyme. All transcripts are fragmented and size selected (> 70 nt). 5' labeled RNA fragments are bound to streptavidin beads and separated from uncapped RNA. Using a poly(A) polymerase, 3' ends are poly(A) tailed. Prior to the enzymatic removal of the desthiobiotin cap with Cap-Clip Acid Pyrophosphatase, 5' monophosphorylated contaminants are ligated to 5' Illumina TruSeq sequencing adapters, which carry special sequence tags 1 (ATTACTCG and TCCG-GAGA, equal proportions, PSS-set). Newly exposed 5' triphosphates of previous primary transcripts are ligated to 5' Illumina TruSeq sequencing adapters carrying the sequence tags 2 (CGCTCATT and GAGATTCC, equal proportions, TSS-set). Synthesis of first-strand cDNA is performed using oligo(dT)-adapter primer and M-MLV reverse transcriptase. The cDNA is amplified in a PCR (14–16 cycles) with primers binding at the 3' end of the first-strand cDNA exhibiting a biotinylation for a subsequent size selection step. For this, samples are enzymatically fragmented and 5' cDNA fragments are selected using streptavidin beads

(a) Cappable-seq following Ettwiller *et al.*, 2016

(b) DTB-TEG-GTP

Figure 2.1: Cappable-seq. (a) Overview of the Cappable-seq workflow of Ettwiller *et al.*, 2016, adapted by vertis Biotechnologie AG. 1, Primary transcripts (5' triphosphorylated RNA) are labeled using vaccinia capping enzyme attaching a desthiobiotin cap (DTB-TEG-GTP) to their 5' ends. 2, Biotinylated 5' fragments are captured with streptavidin beads. 3, Contaminating monophosphorylated fragments are marked with sequence tag 1 (PSS-set) prior to 4, de-capping and 5, tagging of the enriched primary transcripts with sequence tag 2 (TSS-set). 6, The sample is sequenced using next generation methods. (b) Structural formula of capping molecule DTB-TEG-GTP (3' desthiobiotin-tetraethylene glycol-guanosine 5' triphosphate).

(size range: 100–300 bp). Illumina sequencing adapters (3′) are ligated and the cDNA is finally amplified in a PCR reaction (4–8 cycles). PCR libraries are pooled, size fractionated (200–500 bp) and sequenced on an Illumina NextSeq 500 system (single end, 75 bp).

2.9.3 Processing of Cappable-seq sequencing reads

Demultiplexed raw sequencing reads are provided by vertis Biotechnologie AG, Freising. PSS- and TSS-tag containing reads are separated and subsequently quality trimmed with the programs cutadapt (Martin, 2011) and Trimmomatic (Bolger *et al.*, 2014), where the latter one removes low quality reads as well as reads with Poly-A-80-, Poly-T-80, Poly-G-80 and Poly-AAGGG-tail as defined in the file ‘adapter.fasta’ (Table 2.15). The remaining reads are mapped to the genome of *Escherichia coli* O157:H7 EDL 933 (NCBI accession no. NZ_CP008957) using bowtie2 (Langmead and Salzberg, 2012). Settings for the programs cutadapt, Trimmomatic and bowtie2 are given in Script 1.

Table 2.15: Trimming file for Cappable-seq data processing. Sequences included in the file ‘adapter.fasta’ for quality trimming of Cappable-seq sequencing reads.

name	sequence
TruSeq 3′fwd	AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
TruSeq 3′rev	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
TruSeq 5′fwd	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA
TruSeq 5′rev	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
PolyA80	(A) ₈₀
PolyT80	(T) ₈₀
PolyG80	(G) ₈₀
PolyAAGGG	AAAAAAAAAAGGGAGGGGGGGGGGGGGGG

Script 1: Processing of Cappable-seq sequencing reads

```
1 # Trimming adapter sequences
2 # INPUT: demultiplexed raw sequencing reads file
3 # PSS-tag data
4 cutadapt -u 2 -g ^ATTACTCG -g ^TCCGGAGA -e 0.14 -O 8 --no-indels -o
   temp_PSS.fastq --untrimmed-output untrimmed_file.fastq INPUT.fastq
5 cutadapt -u 3 -o OUTPUT_PSS.fastq temp_PSS.fastq
6 # TSS-tag data
7 cutadapt -u 2 -g ^CGCTCATT -g ^GAGATTCC -e 0.14 -O 8 --no-indels -o
   temp_TSS.fastq --untrimmed-output untrimmed_file.fastq INPUT.fastq
8 cutadapt -u 3 -o OUTPUT_TSS.fastq temp_TSS.fastq
9
10 # Trimming of low quality regions and reads
11 # INPUT: PSS-tag/TSS-tag separated fastq-files
12 java -jar trimmomatic-0.36.jar SE INPUT.fastq OUTPUT_trimmed.fastq -phred33
   ILLUMINACLIP:adapter.fasta:2:30:10 SLIDINGWINDOW:4:15 LEADING:3
   TRAILING:3 MINLEN:30
13
14 # Mapping with bowtie2 and conversion into bam-file
15 # INPUT: quality trimmed reads in fastq-files
16 bowtie2 --local -q -p 16 -x genome -U INPUT.fastq -S mapped.sam
17 samtools view -Sb mapped.sam > mapped.bam
18
19 # Processing of reads for visualization
20 samtools sort mapped.bam -o sorted.bam
21 samtools index sorted.bam sorted.bam.bai
```

2.9.4 Bioinformatic determination of transcriptional start sites

Transcriptional start sites are determined with two programs provided by Ettwiller *et al.*, 2016, where the first trims all mapped sequencing reads to 1 bp long reads beginning at the 3' end leaving the outermost 5' base of the read. A relative read score (RRS) is calculated with Equation 2.3.

$$RRS_{io} = \frac{n_{io}}{N} * 10^6 \quad (2.3)$$

In this formula, n_{io} represents the number of reads at position i and orientation o , whereas N is the total number of all mapped reads. All reads above the predefined threshold and input parameter $minRRS$ are maintained. The second program clusters putative TSS dynamically if they have a spacing below the threshold $dist$. The values for $minRRS$ are used as indicated. The clustering distance is set in all analyses to $dist = 5$. Execution details for both programs are shown in Script 2.

Script 2: Standard workflow for TSS determination according to Ettwiller *et al.* (2016)

```

1 # Selection of genomic positions with a minimum number of reads
2 # INPUT: sorted bam file
3 bam2firstbasegtf.pl --bam INPUTsorted.bam --cutoff minRRS --lib_type F --out
  TSS_enriched.gtf
4
5 # Clustering of positions within a defined distance
6 # INPUT: selected positions in gtf-format
7 cluster_tss.pl --tss selected.gtf --cutoff dist --out clustered.gtf

```

2.9.5 Cappable-seq cutoff evaluation for antisense ORFs

Two sets of genome positions are used to evaluate an optimal cutoff for antisense ORFs:

- 1) all genome positions having at least one read (i. e., $RRS > 0$)
- 2) all genome positions without annotated gene regions having at least one read (i. e., $RRS > 0$). Annotated gene regions include the coding DNA sequences of all annotated

genes (based on the annotation of the genome of *E. coli* O157:H7 EDL933, RefSeq annotation NZ_CP008957, 02/23/2017, Latif2014) as well as 100 bp upstream thereof.

Histograms are built to determine the amount of genome positions with a certain RRS for both sets (bin width 0.1). The relative change for each bin is calculated with Equation 2.4.

$$\text{relative change} = \frac{\text{amount of genome positions without annotated gene regions}}{\text{amount of all genome positions}} \quad (2.4)$$

Relative change values are plotted against the mean of the bin range (e. g., mean for bin range 1.45 to 1.55: 1.5). A cubic square smooth function is placed on the data to illustrate the curve shape.

2.9.6 Determination of gene associated TSS

All ORFs in given gene sets are investigated for upstream-located transcription start sites. TSS lists created in Section 2.9.4 with a given threshold $minRRS$ are screened for all possible TSS which are located 250 bp upstream of the start codon of the investigated ORFs. Screenings are performed for the eight analyzed conditions independently. A TSS is maintained if it is present simultaneously in the three biological replicates.

2.9.7 5' UTR evaluation

The distance between gene associated transcription start sites of annotated genes and the corresponding start codon is measured. Transcription start sites are determined as described in Section 2.9.6 except from the fact that a 500 bp upstream region is scanned for transcription start sites.

2.9.8 Operon Structures

Operon information of *E. coli* MG1655 are downloaded from the Database of prokaryotic Operons (DOOR). The inter-gene distance of polycistronic genes is calculated (ranging from 1 bp to 147 bp with one exception of 919 bp). The genomic locations of ORFs are compared and candidates with an inter-gene distance of less than 150 bp are categorized as putative polycistronic ORFs.

2.9.9 Analysis of TSS strength

RRSs for gene associated TSS in all conditions and replicates are extracted and visualized. Additionally, statistical differences between RRS are calculated with t-test. A paired t-test is applied when exponentially and stationary samples are compared, whereas a Welch two sample t-test is used when samples of stress conditions (minimal medium, LB supplemented with either L-malic acid or NaCl) are compared to the non-stress (LB) in the same growth phase. To categorize transcription start sites according to the signal strength, the mean RRS of the TSS position across all analyzed conditions is calculated and divides TSS into weak ($RRS \leq 1.5$), medium ($1.5 < RRS < 5$) and strong ($RRS \geq 5$) positions.

2.9.10 Promoter motif identification with sequence logos

Upstream regions of gene associated transcription start sites are analyzed for conservation patterns. For this, sequence information of 100 bp upstream of all transcription start sites (duplicates removed) in one gene set are extracted. The sequences are applied to sequence logo construction with WebLogo 3 (Crooks *et al.*, 2004) according to Script 3.

Script 3: Sequence logo creation with WebLogo 3

```

1 # Create sequence logo
2 # INPUT: fasta file with sequences to be analyzed for conservation
3 weblogo -l -50 -u 0 -n 51 --first-index -100 --yaxis 1 --errorbars No --format
  pdf --color-scheme classic --number-interval 10 < INPUT.fasta > logo_output.
  pdf

```

2.9.11 TSS for sense overlapping ORFs

All longest possible ORFs in the genome of EHEC are detected with the program getORF (Script 4). Out of frame sense overlapping ORFs, either completely embedded or 3' partially overlapping (according to Figure 1.2), having at least 93 bp in common with the annotated genes, are identified.

Script 4: getORF settings to find the longest ORFs for each stop in the genome of EHEC

```
1 # Find all ORFs in the genome
2 # INPUT: genome fasta file
3 getorf -filter -sequence INPUT.fna -table 11 -minsize 93 -find 1 -circular Y
```

Transcription start sites of sense overlapping ORFs are determined with the method described in Section 2.9.6 with $minRRS = 1.5$. The RRSs of the putative TSS in the single biological replicates and conditions are compared to the maximum background signal of the annotated gene, which is the position with the highest RRS within the respective annotated gene in the corresponding replicates and conditions. Only positions are considered as highest background, if they are not classified as transcription start sites for any gene or ORF or 5 bp upstream and downstream of transcription start sites due to the clustering algorithm (Section 2.9.4). A signal-to-noise ratio is calculated with Equation 2.5.

$$S/N = \frac{RRS_{TSS}}{RRS_{noise}} \quad (2.5)$$

Sense transcription start sites are treated as non-background for $S/N > 1.5$ in all biological replicates. All maintained TSS are visually inspected to verify the association to a sense overlapping ORF.

2.10 Protein chemical techniques

2.10.1 Preparation of whole cell lysates

Cells harboring expression vectors pBAD/SPA or pBAD/His with desired insert sequences (control proteins or OGCs) are cultivated until an optical density of $OD_{600} = 0.3$. Protein production is induced with 0.002% L-arabinose. After a maximum induction time of 4 h, 1 mL of cells is harvested. Cell pellets are dissolved in 50 μ L SDS sample buffer and heated at 95 °C for 10 min to lyse the cells. Whole cell lysates are centrifuged at 16 000 \times g for 10 min at RT to collect the cell debris. Samples are either used directly for SDS-PAGE or stored at -20 °C.

2.10.2 Tris-tricine SDS-PAGE

Tris-tricine SDS-PAGE according to Schägger, 2006, is used to separate small proteins. Changing the buffer system from glycine to tricine shifts the stacking limit in the stacking gel to a low-molecular-mass range and enables efficient separation of small proteins (Schägger and Von Jagow, 1987).

Whole cell lysates of test samples (10 μ L) and the positive control sample *gst* (2.5 μ L) are separated on a tricine gel consisting of a 4% stacking gel and a 16% resolving gel with each 3.3% crosslinking (Table 2.16). Diluted cathode buffer and anode buffer (1x each) is filled in the provided chambers of cathode and anode, respectively, and 35 mA current per gel is applied. Visualization of separated proteins is performed with Western blot (Section 2.10.3).

Table 2.16: Composition of resolving and stacking gels for tricine SDS-PAGE.

	resolving gel (16 %)	stacking gel (4 %)
40 % acrylamide:bisacrylamide (29:1)	3.2 mL	0.4 mL
3x gel buffer	2.7 mL	1 mL
Glycerin	0.8 mL	
H ₂ O	1.3 mL	2.6 mL
TEMED	2.67 μ L	3 μ L
APS (10 %)	26.67 μ L	30 μ L

2.10.3 Western blot

Immunological detection of separated proteins on SDS gels is performed with Western blots. In this process, proteins are transferred to a membrane capable to bind proteins and visualized with a specific antibody that binds to the coexpressed protein tag. An enzyme coupled to the antibody is used to stain the target proteins.

All following incubation steps are carried out with gentle shaking. Immobilon-P^{SQ} (0.2 μ M) membrane, a PVDF membrane optimized for binding small proteins, is activated in 100% methanol for 15 s and subsequently washed in ultrapure water (5 min). The membrane as well as the polyacrylamide gel from Tris-tricine SDS-PAGE are equilibrated in 1x blotting buffer

for 10 min. Three filter papers soaked with 1x blotting buffer, gel, membrane, and once again filter papers are stacked in this order precisely in a semi-dry electroblotting device. Proteins are transferred from the gel onto the membrane (12 V, 20 min). After the blotting step, proteins are fixed on the membrane with 3% trichloroacetic acid (5 min) and washed with ultrapure water (5 min). Unspecific binding of proteins is blocked with 5% nonfat dried milk powder in 1x TBS-T at 4°C ON. Afterwards, the membrane is washed three times with TBS-T for 10 min. The primary antibody (monoclonal mouse anti-FLAG M2-Alkaline Phosphatase (AP) antibody, Sigma Aldrich, or monoclonal mouse anti-6xHis antibody) is applied as 1:1000 dilution in TBS-T (10 mL) for 1 h at room temperature. The membrane is subsequently washed six times with TBS-T for 5 min, respectively. When using anti-His antibodies, a secondary antibody (Goat anti-mouse AP conjugated antibody, Dianova) is applied 1:10 000 diluted in TBS-T as described before.

For colorimetric detection with BCIP/NBT, AP buffer is added for 5 min and replaced by 10 mL reaction buffer supplemented with 100 μ L NBT-solution and 125 μ L fresh BCIP-solution. The chromogenic substrates are removed as soon as bands appear and the staining reaction is stopped with 3% trichloroacetic acid.

Chemiluminescent detection is carried out with an IVIS system. The membrane is treated with 500 μ L CDP Star chemiluminescent substrate. Bands are visualized after a maximum incubation time of 1 min at room temperature with an exposure time of 10 s.

2.11 Bioinformatic applications

2.11.1 Promoter determination

Two programs are used to identify putative promoter sequences, BPROM (Solovyev and Salamov, 2011) and bTSSfinder (Shahmuradov *et al.*, 2017). If not stated otherwise, input sequences for BPROM and bTSSfinder have a length of 100 bp and 300 bp, respectively, begin upstream of the TSS and end at the TSS. BPROM calculates a linear discriminant function (LDF) to rate the promoter strength. A promoter with LDF = 0.2 has 80% accuracy and specificity. BPROM identifies solely σ^{70} promoters. bTSSfinder can predict promoters of the classes σ^{70} , σ^{38} , σ^{32} , σ^{28} , and σ^{24} . Scoring threshold are 0.06, 0.00, 1.01, 1.24, and 0.31,

respectively.

2.11.2 Terminator identification

FindTerm (Solovyev and Salamov, 2011) is applied to identify a rho-independent terminator for *pop* (threshold -3). A 900 bp long testsequence downstream of the *ompA* coding region is analyzed. The identified terminator is split into 30 bp segments. Sequences are folded with the tool QuickFold of Mfold (<http://www.bioinfo.rpi.edu/applications/mfold>; Zuker, 2003) to determine the stem loop structure of the identified terminator.

2.11.3 Ribosome binding site determination

Shine-Dalgarno sequences are determined according to Ma *et al.*, 2002 in the region 30 bp upstream of the start codon. Sequences with a minimum free energy of $\Delta G^\circ = -2.9$ kcal/mol are predicted to be ribosome binding sites.

2.11.4 Gene prediction

Genome data of *Escherichia coli* O157:H7 str. EDL933 (Accession number CP008957), *Shigella dysenteriae* str. ATCC 13313 (Accession number CP026774.1), *Klebsiella pneumoniae* subsp. pneumoniae str. ATCC 13883 (BioProject PRJNA261239) and *Enterobacter cloacae* subsp. cloacae str. ATCC 13047 (Accession number CP001918) are downloaded from NCBI. Prodigal v2.60 (Hyatt *et al.*, 2010) is used with default settings for gene prediction.

3 Results

3.1 Overview of overlapping genes analyzed

In the course of this study, different sets of overlapping genes were analyzed and described briefly in the following.

1) 216 overlapping gene candidates (OGCs)

The translome and transcriptome of *E. coli* O157:H7 str. EDL933 was analyzed using ribosomal profiling (Figure 1.6) and strand-specific RNA-seq by Landstorfer (2014). He selected a set of 242 ORFs outside of prophage regions, which were covered with ribosome profiling reads and calculated the translational efficiency by defining the ribosomal coverage value ($RCV = \frac{RPKM_{translatome}}{RPKM_{transcriptome}}$; RPKM, reads per kilobase per million mapped reads). These candidates were reanalyzed by Zehentner (2015) in the updated genome of EHEC published by Latif *et al.* (2014) and a set of 216 unique candidates remained (134 embedded, 82 partial with 76 OGCs having a non-trivial overlap of ≥ 90 bp, Supplementary table S6). A total of 156 candidates (72%) had a strong indication for translation since the RCV exceeds the threshold of 0.355 established by Neuhaus *et al.* (2017) for translated mRNAs. As few as 21 candidates fall below the threshold 0.197 probably describing non-coding transcripts, whereas the translation status of the remaining 39 overlapping gene candidates is uncertain (thresholds according to Neuhaus *et al.*, 2017). A blastp analysis against proteins deposited in the RefSeq database yielded homologous proteins for 19 OGCs (Supplementary table S7, RefSeq database as of Sept 2019, analysis conducted by Dr. Zachary Ardern). All 216 candidates were analyzed regarding their potential to form stable proteins (Section 3.2), to confer an overexpression phenotype (Sections 3.3 and 3.4) and to exhibit a transcription start site (Section 3.5).

2) 30 870 antisense embedded overlapping ORFs (embORFs)

The genome of EHEC (RefSeq annotation NZ_CP008957, 02/23/2017, Latif2014) was screened bioinformatically for the longest possible open reading frames embedded in antisense to annotated genes. The analysis revealed 30 870 ORFs between 93 bp and

2529 bp (Supplementary Figure S2). The majority of ORFs are less than 200 bp long (71%). The predominant start codons are the rare codons ATC (5631 ORFs) and CTG (5503 ORFs, Meydan *et al.*, 2019), whereas GTG (3440 ORFs) and ATG (3520 ORFs) are found least frequently. A set of 218 ORFs was found to have a blastp hit; thus, indicating significant similarity with a protein deposited in the RefSeq protein database (e-value cutoff 10^{-10} , RefSeq database as of Feb 2017). Embedded OGCs from 1 (see above) are a subset of the embedded ORFs. The entire set was analyzed for transcriptional start sites (Section 3.5).

3) 17 601 sense overlapping ORFs

In addition to antisense ORFs, EHEC harbors 16 556 embedded as well as 1045 3' partial and 1773 5' partial sense overlapping ORFs with a length of at least 93 bp. Although transcriptome and translome analyses can be conducted strand specifically, sense overlapping genes cannot be assessed using these methods. In contrast to eukaryotic ribosome profiling (Ingolia *et al.*, 2009), bacterial profiling data typically lack a three nucleotide periodicity on single gene level which is used to determine the reading frame of translated genes (Hwang and Buskirk, 2016). Therefore, ribosome profiling reads mapping to a genomic region with annotated gene and sense overlapping ORF cannot be assigned unambiguously. Genome wide transcriptional start site data were investigated to find first hints of independently transcribed embedded or 3' partial sense overlapping ORFs (Section 3.6). The analysis was restricted to these two kinds of overlaps since transcription start sites associated with 5' partial ORFs are often false positive signals due to their location in near proximity to the respective annotated genes and can just as well be the start site of the latter ones.

In addition to overlapping genes, the set of annotated genes of *E. coli* O157:H7 EDL933 (RefSeq annotation NZ_CP008957, 02/23/2017, Latif *et al.*, 2014) were screened for transcriptional start sites (Section 3.5). A distinction was made between functional (4525) and hypothetical (973) annotated genes. While genes in the former set have a known protein function, functions of the proteins in the latter set are hypothetical.

3.2 Expression of overlapping genes and immunostaining of proteins

Protein products of overlapping gene candidates were analyzed by Western blots. After evaluating a reasonable expression vector, protein signals of overexpressed OGCs were evaluated and OGC-protein masses were determined using a semi-logarithmic approach. Reproducibility of the method was examined by measuring a subset of candidates twice.

3.2.1 Evaluation of a suitable overexpression vector

The vectors pBAD/SPA and pBAD/His were evaluated for their capability to produce tagged proteins in a wide mass range suitable for Western blots. Both, pBAD/SPA and pBAD/His, consist of the pBAD/myc-HisC backbone. The sequential peptide affinity tag (SPA tag) was designed after Zeghouf *et al.* (2004) and replaces the myc epitope of pBAD/myc-HisC. For pBAD/His, myc was removed without replacement and the original 6xHis tag was kept in frame.

Two proteins were selected for evaluation purposes. *rpmH* (50S ribosomal protein L34) and *gst* (glutathione S-transferase) are highly expressed genes in EHEC (verified in ribosomal profiling data of Landstorfer, 2014). Using the test proteins characterized by substantially different protein masses (5.4 kDa and 22.9 kDa), the Western blot protocol was adjusted to detect both proteins despite their wide range of weights. The average weights range from 3.3 kDa to 54.78 kDa for putative proteins encoded by OGCs.

His- as well as SPA-tagged proteins RpmH and Gst could be detected successfully (Figure 3.1). However, rpmH+His runs at 13 kDa instead of the expected 7.1 kDa. In contrast, rpmH+SPA (16 kDa) is detectable close to the expected size (14 kDa). The second control protein has calculated protein sizes of 31 kDa and 25 kDa for SPA- and His-tag, respectively, and was detected accordingly (30 kDa and 24 kDa). As expected, lysates of cells without any plasmid (i. e., empty cells) had no significant signals above background, whereas tag-expressing cells (i. e., empty vectors) produced a weak signal for the SPA-tag, but not the small His-tag (3 kDa). The band of SPA is slightly higher (13.5 kDa instead of 10 kDa), but in line with previous observations of Baek *et al.* (2017) (10 kDa instead of 8 kDa, Figure 3.1c).

In general, the background signal is higher in SPA blots probably due to application of an

alkaline phosphatase conjugated primary SPA antibody compared to a two-antibody based detection system for His-tagged proteins. Nevertheless, the detection was more reproducible and less error-prone for RpmH in SPA blots compared to His blots indicating a potential stabilizing function of SPA for small proteins. This might be advantageous for Western blot detection of small overlapping genes and, therefore, pBAD/SPA was used for further protein analysis. The chemiluminescent detection of alkaline phosphatase activity using the IVIS detection system leads to broad and indistinct band patterns. Consequently, protein detection in further experiments was carried out with colorimetric substrates BCIP and NBT, resulting in clearer protein bands on the blot (Section 3.2.2).

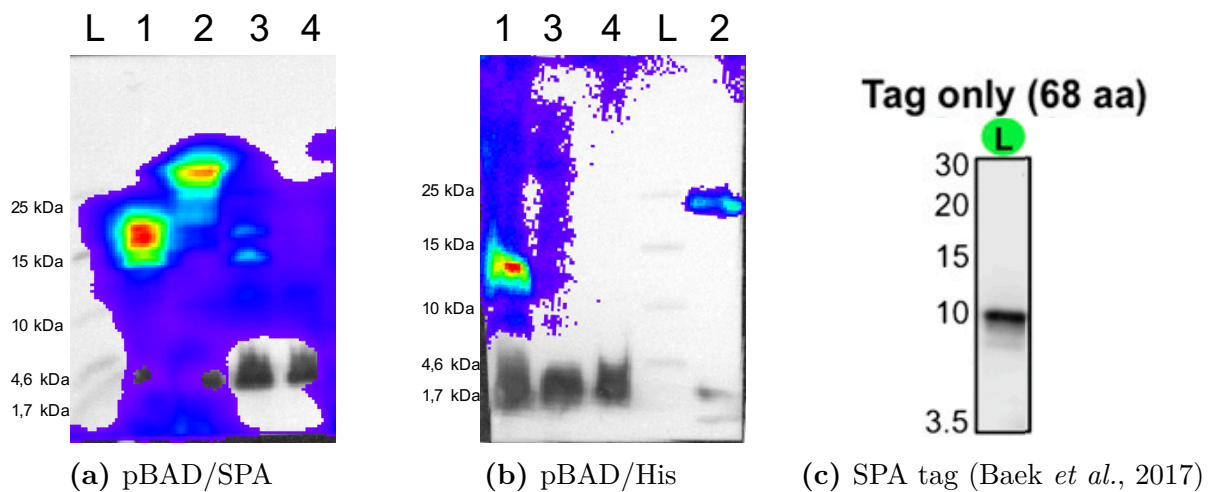


Figure 3.1: Western blot of control proteins with SPA- or His-tag. Chemiluminescent protein detection of SPA-tagged (a) and His-tagged (b) proteins. L, protein ladder; 1, RpmH; 2, Gst; 3, tag only; 4, cells without plasmid; the amount of Gst+His loaded was reduced to one fifth compared to the other samples. (c) Baek *et al.* (2017): Expression of SPA tag in LB medium (L). The weight of SPA (68 amino acids) is calculated to 8 kDa and SPA is detected at 10 kDa.

3.2.2 Western blots of overlapping genes

The vector pBAD/SPA was used to clone 210 out of 216 overlapping gene candidates for protein analysis with Western blots. Proteins for overall 202 candidates were detected (Supplementary table S8 and Supplementary file 1). Detailed analyses of OGC 57 and OGC 59 are presented in Section 3.7 and the publication of Vanderhaeghen *et al.* (2018), respectively. Protein signals of the remaining 200 candidates were categorized after visual inspection into

four groups (Table 3.1, examples in Figures 3.2a – 3.2d).

Table 3.1: Categories of protein signals in Western blot. Category and number of candidates with respective protein pattern detected in analysis.

Category	No. of Candidates
single protein band	107
high background	65
by-products visible	21
signal with smear	7

A single unambiguous protein band at the expected size was visible for 107 candidates indicating the overlapping protein (Figure 3.2a). Background signals, consisting of slight secondary bands or smear, were found in blot of 65 candidates (Figure 3.2b); however, one major protein band representing the desired product was still detectable for these. Secondary products were detected for 21 overlapping gene candidates (Figure 3.2c). For 85 % of these candidates, the signal of the assumed correct protein band is strongest or at least equal to the secondary band(s). In some cases, by-products have a substantially higher molecular mass than the assumed protein resulting from the overlapping gene. Seven candidates had smeared bands, possibly due to sub-optimal gel loading, blotting or staining, like OGC 121 (Figure 3.2d). To resolve this problem, the plasmid of this candidate was transformed into EHEC and the experiment was repeated to improve the protein signal. Indeed, expression of the overlapping gene in the native environment led to an increased stability and a single distinct protein band (Figure 3.2e). However, in several cases, expression in EHEC was found to be disadvantageous as no proteins were detected in Western blots, although Western blot of cell lysates prepared from *E. coli* Top10 resulted in clear signals (Supplementary file 1).

3.2.3 Protein mass analysis for proteins of overlapping genes

The molecular weights of proteins can be determined using standard proteins with known molecular masses (Shapiro *et al.*, 1967; Dunker and Rueckert, 1969). The logarithmic molecular weight of a protein has a negative linear correlation to the relative migration distance

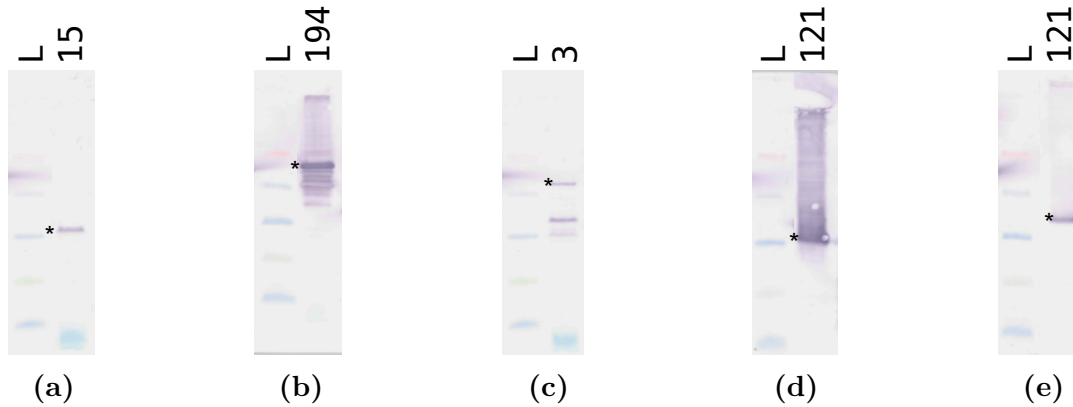


Figure 3.2: Examples of Western blots. Blots for candidates with (a) a single protein band, (b) a single protein band with high background signal, (c) several protein forms, and (d) a smeared protein band. Protein expression was performed in *E. coli* Top10 for (a) – (d). (e) Protein expression of candidate 121 shown in (d) performed in the native organism *E. coli* O157:H7 EDL933. Numbers indicate the overlapping gene candidate, asterisks the most probable protein band. L, Protein ladder. Irrelevant gel lanes were excised from the blot image. Original blots are included in Supplementary file 1.

(R_f) of the protein in an acrylamid gel. Using the linear regression line of standard proteins, the molecular mass of a target protein can be calculated using the R_f of this polypeptide.

Accordingly, the protein masses of 202 proteins were determined (Supplementary table S8). Comparison of experimental protein masses and theoretical ones calculated based on sequence composition revealed a stable linear relationship (coefficient of determination $R^2 = 0.89$, Figure 3.3a). However, the molecular weights calculated using the relative migration distance in the SDS gel have higher values than expected (solid line above dashed line in Figure 3.3a). This fact is in line with the observation that masses of the control protein RpmH and the SPA tag appeared to be larger when compared to their theoretical mass (Section 3.2.1). The mass difference between the expected and the observed sizes slightly decreases for larger proteins (i. e., reduced spacing between the linear regression and reference line at higher molecular weights). Again, this is in accordance with molecular weight measurements of the larger control protein Gst, which was detected at almost the correct size.

The reproducibility of protein mass determination with Western blot was examined. Therefore, expression in *E. coli* Top10 and Western blot were repeated for 50 candidates

showing a phenotype in the HT experiment (Section 3.3.2), as well as six further randomly picked candidates without HT phenotype. The molecular masses of the two replicates were plotted (Figure 3.3b). The overall consistency of the data was confirmed by a linear regression line with $R^2 = 0.925$ indicating a reproducible detection of proteins expressed from overlapping gene candidates with Western blot. Although just this subset of all in all 57 OGCs was tested twice, it may be presumed that the detected proteins of the remaining candidates are not artifacts, but reproducible signals, too.

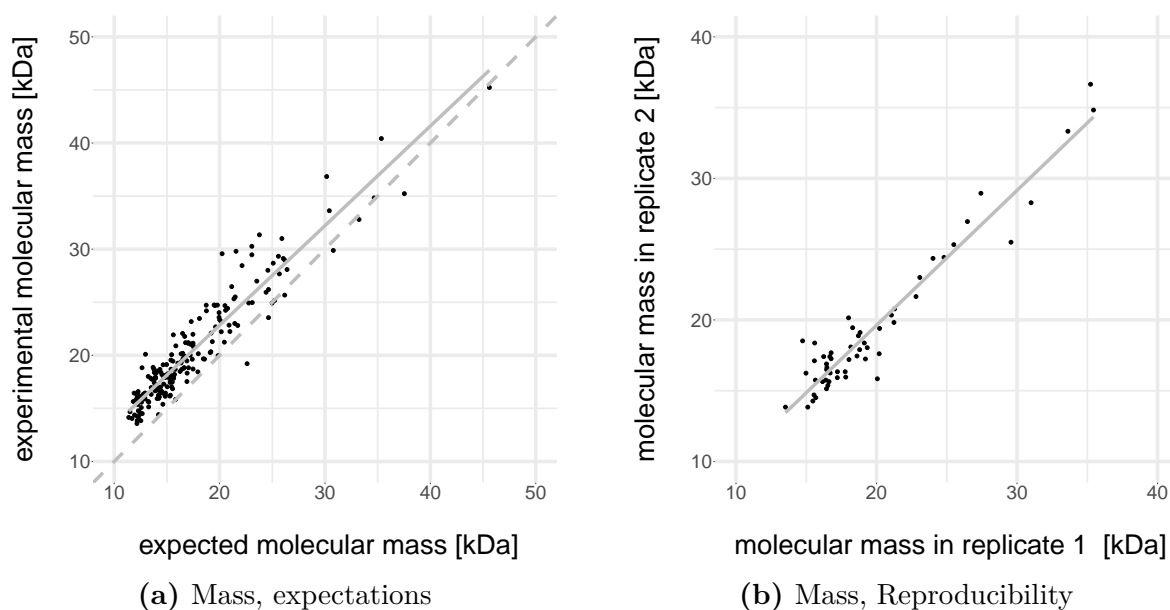


Figure 3.3: Mass determination of proteins of overlapping genes. **(a)** The experimentally determined molecular mass is plotted against the expected calculated molecular mass [kDa]; $n = 202$; solid line, linear regression line of data points with $R^2 = 0.891$; dashed line, theoretical perfect molecular weight match. **(b)** The experimentally determined molecular masses [kDa] for the candidates in two replicates are plotted; $n = 56$; solid line, linear regression line of data points with $R^2 = 0.925$.

3.3 High-throughput overexpression phenotypic analysis

A high-throughput phenotyping was performed to investigate the effect of proteins encoded by overlapping genes on the growth of EHEC (Figure 3.4). The experimental procedure was conducted in two independent biological and one technical replicate. The data set created in Zehentner (2015), was added for data evaluation (replicate I). For biological replicates, independent bacteria pools were used for cultivation, whereas the same pool was used to inoculate medium for the technical replicate. NGS data were evaluated and relative enrichment or depletion of OGC plasmids across different conditions was determined representing overexpression phenotypes.

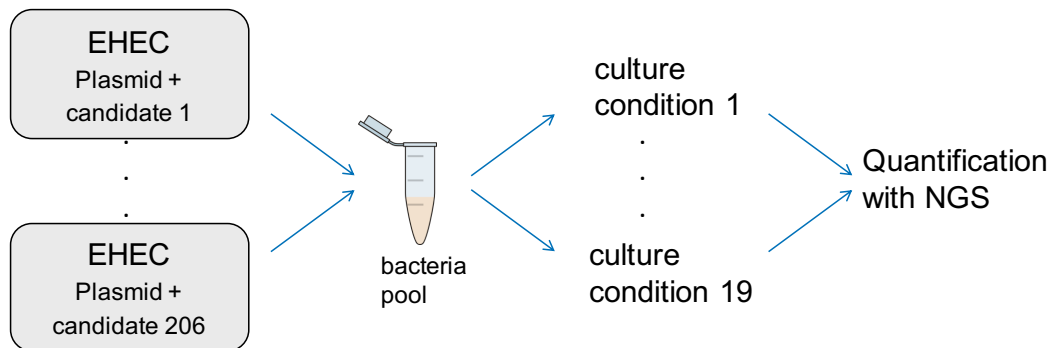


Figure 3.4: High-throughput overexpression phenotyping. Candidate genes were previously cloned and transformed into *E. coli* O157:H7 EDL933 (Zehentner, 2015). Bacteria liquid cultures are pooled in equal amounts. Different stress conditions are inoculated with the bacteria mixture. After competitive growth, plasmids are isolated and quantified by NGS.

3.3.1 Sequencing evaluation

Sequencing reads were mapped to an artificial genome-like sequence including all sequences of OGCs which were present in the bacterial pool. RPKM values of overlapping gene candidates were determined (Supplementary file 2) and linear Pearson correlations r of the replicates were calculated as summarized in Table 3.2.

In general, r can have values between -1 and $+1$. Values near these boundaries indicate strong negative or positive linear correlation, respectively, whereas a moderate linear relationship is assumed for $0.5 < |r| \leq 0.8$ (Peck *et al.*, 2015). The correlation coefficient of

independent biological replicates of the HT phenotyping range from $r = 0.671$ to $r = 0.791$, which implies a considerable similarity of the datasets. A correlation coefficient of $r = 0.964$ for technical replicates indicates a high reproducibility of the experimental workflow. Taken together, it can be assumed that the outcome of the experiment mainly depends on the input pool.

Table 3.2: Correlation of HT sequencing experiments. Pairwise Pearson’s product-moment correlation coefficients r of RPKM values for biological (I, IIA, III) and technical (IIA, IIB) replicates.

Replicates Compared	Correlation Coefficient r
I - IIA	0.707
I - III	0.791
IIA - III	0.671
IIA - IIB	0.964

A number of 10^4 to 10^5 sequencing reads in each condition map to the target sequences (Table 3.3). The condition Na_3VO_4 in replicate III was sequenced less efficiently with read numbers below 1×10^4 . Therefore, this condition was excluded from replicate III for further pairwise correlation calculations and the phenotype evaluation.

To determine the similarity of the HT phenotyping effects on gene level across biological replicates, pairwise correlations of biologically replicated RPKM values of each gene in all conditions were calculated. Spearman’s rank correlation function was chosen for evaluation as it uses ranks of the RPKM values (Härdle *et al.*, 2015) instead of absolute values required to calculate Pearson’s r . Thus, the performance of the Spearman’s function is more robust against strongly varying values. Similar to r , Spearman’s $|\rho|$ can have values between 0 and 1 and $0.5 < |\rho| \leq 0.8$ or $|\rho| > 0.8$ indicate moderate or strong correlation, respectively. Altogether, 23 candidates have at least moderately conserved RPKM patterns (Table 3.4), representing 11% of tested candidates. Nevertheless, RPKM tendencies for at least some candidates are well reproducible across completely independent biological replicates although the experimental setup of the phenotyping procedure is highly complex.

Table 3.3: Number of mapped reads ($\times 10^4$) for replicated sequencing experiments for HT phenotyping. The condition with notable less sequencing reads in replicate III (highlighted) was excluded from analysis.

Condition	Replicate I	Replicate IIA	Replicate IIB	Replicate III
t_0	4.2	2.2	1.5	5.4
LB	13.4	3.5	6.1	7.0
Glucose	4.8	3.2	2.8	3.3
L-malic acid	10.0	4.6	5.8	7.8
L-arginine	10.1	8.3	5.6	8.8
CsCl	5.8	5.8	5.6	6.1
Acetic acid	4.6	8.5	4.2	3.8
Malonic acid	7.1	2.9	2.5	4.1
1-Methylimidazole	4.0	5.5	5.6	8.3
NaCl	1.6	5.5	8.1	6.8
NaOH	7.3	5.6	11.1	5.9
Na_3VO_4	6.3	5.0	7.0	0.4
Sodium salicylate	7.0	7.5	7.9	9.0
Perchloric acid	5.6	6.5	5.4	5.0
Phytic acid	11.4	9.5	8.8	7.0
1,2-Propanediol	9.6	6.7	9.5	5.6
1-Propanol	10.7	4.5	5.8	6.6
Pyridoxin HCl	7.0	6.9	9.9	9.6
Staphylococcus	1.5	2.5	2.6	3.2
ZnCl_2	6.6	3.7	10.1	12.0

Table 3.4: Correlation of RPKM profiles of individual OGCs in HT phenotyping. Spearman’s rank correlation coefficient ρ is given for candidates with strong ($\rho > 0.8$) or moderate ($0.5 < \rho \leq 0.8$) positive correlations in three (left) or two (right) pairwise comparisons of replicates I, IIA, and III.

Candidate	Correlation Coefficient ρ	Candidate	Correlation Coefficient ρ
OGC 23	$0.53 < \rho < 0.85$	OGC 58	$\rho = 0.63$ and $\rho = 0.65$
OGC 57	$0.70 < \rho < 0.82$	OGC 78	$\rho = 0.57$ and $\rho = 0.58$
OGC 59	$0.63 < \rho < 0.71$	OGC 113	$\rho = 0.54$ and $\rho = 0.62$
OGC 81	$0.56 < \rho < 0.82$	OGC 117	$\rho = 0.67$ and $\rho = 0.75$
OGC 105	$0.53 < \rho < 0.65$	OGC 119	$\rho = 0.54$ and $\rho = 0.55$
OGC 121	$0.57 < \rho < 0.79$	OGC 153	$\rho = 0.59$ and $\rho = 0.68$
OGC 125	$0.51 < \rho < 0.68$	OGC 158	$\rho = 0.52$ and $\rho = 0.79$
OGC 141	$0.55 < \rho < 0.75$	OGC 171	$\rho = 0.52$
OGC 194	$0.52 < \rho < 0.65$	OGC 174	$\rho = 0.54$ and $\rho = 0.56$
OGC 231	$0.79 < \rho < 0.87$	OGC 189	$\rho = 0.60$ and $\rho = 0.74$
		OGC 201	$\rho = 0.54$ and $\rho = 0.60$
		OGC 226	$\rho = 0.59$ and $\rho = 0.61$
		OGC 241	$\rho = 0.63$ and $\rho = 0.77$

3.3.2 Selection of candidates with overexpression phenotypes on the basis of z -scores

In order to select candidates where overexpression led to a better or worse growth of bacteria in certain conditions, respectively, RPKM values of each overlapping gene candidate in each investigated condition were normalized to z -scores according to Equation 2.1 ($z_{i,k} = \frac{x_{i,k} - x_i}{\sigma_i}$) for biological replicates I, IIA, and III (Supplementary file 3 and 4). The z -scores are evaluated for each gene separately and specify for a candidate the relative growth alterations after overexpression in one condition compared to the average bacterial growth after overexpression of this candidate in all conditions. Additionally, the z -scores indicate the likelihood of a random event. For example, values following a normal distribution have $|z| < 1$ for 68%

of data points and $|z| \geq 1$ in 32%. Normal distribution of RPKM values was verified by visual inspection of quantile-quantile plots to apply a z -score evaluation. Using a selection criterion for potential phenotypes of $|z| \geq 2$, the probability that a z -score fulfills this criterion by chance is 4.6%. Applying the rule that significant z -scores have to be present in two or three independent biological replicates reduces the chance for erroneous assignment of a phenotype to 0.211% or 9.7×10^{-3} %, respectively, and makes random selection of putative active OGCs highly unlikely.

In summary, 59 overlapping gene candidates with high-throughput overexpression phenotypes were initially selected based on z -score evaluation. The overexpression plasmids were sequenced to exclude false positives due to unwanted sequence errors. Six plasmids revealed single base mutations within the open reading frame or wrongly cloned candidate ORFs. Thus, 53 OGCs with reliable phenotypes remained (Table 3.5 and Supplementary table S9). Nine of these have significantly conserved profiles across biological replicates, whereby six showed at least moderate correlations in all three experiment repetitions (compare with Table 3.4, an example is shown in Figure 3.6). It has to be noted that correlations were calculated based on RPKM values. Nevertheless, they are transferable to correlations of gene specific z -score profiles, as z -scores are normalized RPKM values and the method of normalization does not affect the relationship of data sets. It is striking that most phenotypes can be observed in salt stress conditions, especially in sodium chloride, which altered bacterial growth of 32% of the selected candidates significantly. Furthermore, despite the well-known negative effect of overproduction of unneeded proteins on the growth rate of *E. coli* (e.g. Shachrai *et al.*, 2010), not only disadvantageous, but also advantageous growth effects were detected. Nevertheless, as mentioned before, growth differences are relative effects concerning the average growth of the bacteria and, therefore, the absolute impact of one candidate cannot be deduced here.

Table 3.5: Overlapping gene candidates with HT overexpression phenotype. Relative growth effects in the indicated conditions are categorized in + and – representing better or worse growth upon overexpression than average growth of bacteria. Candidates with moderate correlation of phenotypic profiles in three (light gray) or two (dark gray) biological replicates are highlighted.

Candidate	Condition	Effect	Candidate	Condition	Effect	Candidate	Condition	Effect
salt conditions			acidic conditions			further conditions		
OGC 15	NaCl	-	OGC 116	malonic acid	+	OGC 6	<i>Staphylococcus</i>	+
OGC 18	NaCl	+	OGC 137	malonic acid	-	OGC 50	<i>Staphylococcus</i>	+
OGC 30	NaCl	+	OGC 140	malonic acid	-	OGC 145	<i>Staphylococcus</i>	+
OGC 31	NaCl	-	OGC 198	malonic acid	-	OGC 186	<i>Staphylococcus</i>	+
OGC 59	NaCl	-	OGC 218	malonic acid	-	OGC 195	<i>Staphylococcus</i>	+
OGC 68	NaCl	+	OGC 57	L-malic acid	+	OGC 232	<i>Staphylococcus</i>	+
OGC 75	NaCl	-	OGC 121	L-malic acid	+	OGC 3	HClO ₄	-
OGC 106	NaCl	-	OGC 146	L-malic acid	-	OGC 147	HClO ₄	-
OGC 107	NaCl	+	OGC 226	L-malic acid	+	OGC 191	HClO ₄	-
OGC 172	NaCl	+	OGC 153	acetic acid	-	OGC 241	HClO ₄	-
OGC 174	NaCl	+				OGC 71	sodium salicylate	-
OGC 178	NaCl	-				OGC 117	sodium salicylate	-
OGC 194	NaCl	-				OGC 164	1-propanol	-
OGC 213	NaCl	+				OGC 177	1-propanol	-
OGC 217	NaCl	+				OGC 25	glucose	+

Table 3.5: Continued from previous page

Candidate	Condition	Effect	Candidate	Condition	Effect	Candidate	Condition	Effect
salt conditions			acidic conditions			further conditions		
OGC 231	NaCl	-				OGC 119	LB	-
OGC 23	Na ₃ VO ₄	-						
OGC 26	Na ₃ VO ₄	+						
OGC 44	Na ₃ VO ₄	+						
OGC 51	Na ₃ VO ₄	-						
OGC 96	Na ₃ VO ₄	+						
OGC 183	Na ₃ VO ₄	+						
OGC 205	Na ₃ VO ₄	+						
OGC 24	CsCl	+						
OGC 85	CsCl	+						
OGC 139	CsCl	+						
OGC 167	ZnCl ₂	-						

3.4 Low-throughput phenotypic analysis

Candidates with high-throughput overexpression phenotypes were analyzed in follow-up experiments to investigate the absolute influence of overexpression on bacterial growth. In general, the effect of overlapping genes is assumed to be weak. As demonstrated by Fellner *et al.* (2015), standard growth curves lack sufficient sensitivity to measure slight growth differences between modified and unmodified bacteria. Therefore, direct competition experiments were conducted, where a mixed population of a wild type and a mutant transformant is grown and variations in bacterial numbers before and after growth are examined. Using this highly sensitive method, small alterations in bacterial growth can be assessed (Deutschbauer *et al.*, 2014).

3.4.1 Analysis of candidates in competitive growth assays

The high-throughput screening revealed 53 candidates with putative phenotypes (Table 3.5). Fifty-one candidates were investigated for overexpression phenotypes in the course of the low-throughput phenotypic analysis. Functional characterization of overlapping gene candidate 59, then designated *asa*, was previously published based on the results of the presented HT analysis (Vanderhaeghen *et al.*, 2018). OGC 57, then designated *pop*, was analyzed in-depth as part of this thesis and detailed results are presented in Section 3.7.

Plasmids of the remaining candidates were mutated. A stop codon was introduced at the beginning of the coding sequence by replacement of up to three bases. To conduct competitive overexpression growth, a mixture of cells containing plasmids with either the intact or the truncated candidate open reading frame was cultivated in the identified stress condition as well as without stress. In some cases additional growth conditions were tested, though they did not meet the HT phenotype selection criterion. Instead, highly correlated behavior across biological replicates and somewhat higher and consistent z -scores were found for these conditions.

3.4.2 Overlapping genes with significant overexpression phenotypes

Single competitive overexpression assays were conducted in at least biological triplicates for 51 overlapping gene candidates (Supplementary table S9). An overexpression phenotype is considered for stress conditions that significantly changed proportions of competing strains after growth. As described previously, the difference between the strains is minimal, the maximum are three base substitutions resulting in a stop codon. It is assumed that these small alterations do not change the activity and function of expressed RNA (if a function is present), but of proteins. Further, if significant growth differences are detected, it is implied that overexpression leads to substantial changes within the bacteria, which can be traced back to expression and presence or absence of a protein for cells carrying the intact or translationally arrested sequence, respectively.

Nine candidates showed a statistically valid growth alteration upon stress exposure (Table 3.6, candidates with p -value < 0.05). For the remaining analyzed candidates no statistically significant growth effects were detected. However, visual inspection of the graphical results of the experiment revealed clear growth differences for four further candidates and an overexpression phenotype is proposed even for these (Table 3.6 and Supplementary table S9). Altogether, 13 candidates with growth phenotype in at least one stress condition were detected representing 25 % of tested overlapping gene candidates (Figure 3.5).

One example is shown in detail in Figure 3.6. The high-throughput screening resulted in a phenotype for OGC 231, a relative growth disadvantage in sodium chloride, which was verified in the LT approach as wild type OGC 231 expressing cells grew significantly worse. Additionally, the organic acid L-malic acid was tested, though the phenotype criterion was just not met for this condition, but z -scores are consistent and high in all biological replicates. As can be seen, cells expressing the full-length sequence of OGC 231 grew significantly better than cells expressing the truncated sequence in the acidified medium. These results show that stress conditions of HT phenotyping retested in LT assays have to be selected carefully. Despite missing significance in the former approach, single competitive assays can lead to clear growth phenotypes.

Moreover, it can be seen from Table 3.6 that in some cases tendencies of the high-throughput approach are not reproducible in single assays. Nevertheless, the detected phe-

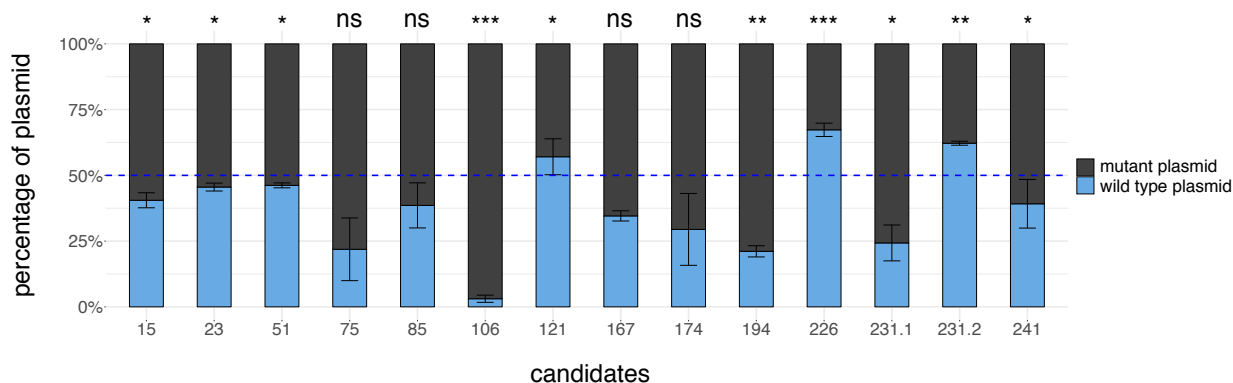


Figure 3.5: Overlapping genes with phenotypes in single competitive growth assays. Mean percentage of peak heights (fluorescence intensities) measured in sequencing electropherograms at wild type or mutated positions in the plasmids are shown for the candidates. Values for wild type and mutant are indicated in blue or grey, respectively. Only the phenotype causing stress condition is displayed according to Table 3.6 with 231.1 showing the primary and 231.2 the secondary stress for OGC 231. Values are adapted to 50% input ratio (blue dashed line) in relation to the corresponding t_0 condition. Error bars indicate standard deviations. Significance was tested with a two-tailed paired t-test ($\alpha = 0.05$; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$; ns, not significant).

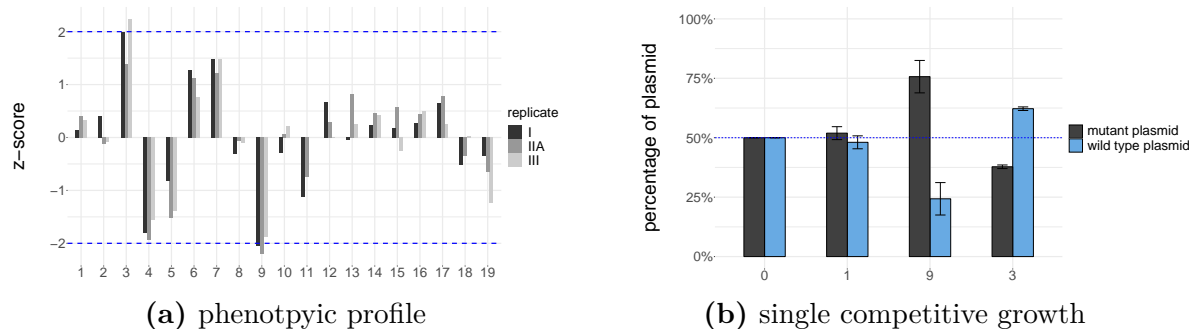


Figure 3.6: Overexpression phenotypes of OGC 231. **(a)** Phenotypic profile of HT overexpression screening. z -scores for three biological replicates and 19 culture conditions are shown. Significant relative growth disadvantage is proposed for condition 9 (NaCl as indicated in Table 2.14), with $z < -2$ in at least two biological replicates. **(b)** Graphical presentation of single competitive growth assay similar to Figure 3.5. Plasmid ratios before ($0 = t_0$) and after cultivation in non-stress ($1 = \text{LB}$) and stress conditions ($9 = \text{NaCl}$ and $3 = \text{L-malic acid}$) are given.

notypes are inferred to be genuine, since absolute growth effects between the two competing strains are measured, which are comparable only in part with the HT relative growth of the wild type transformant. Furthermore, the observation made for the high-throughput screen-

Table 3.6: Overlapping gene candidates with phenotype in LT phenotyping. Tested stress conditions for each OGC are listed, whereby secondary stresses without significant HT phenotype are indicated in square brackets. Relative growth tendencies in HT and detected growth in LT assays are symbolized with + and -. In case of LT, + indicates a growth advantage for cells overexpressing the wild type OGC sequence and - indicates a growth disadvantage for these cells. Statistically shifted growth after stress exposure was tested with a two-tailed paired t-test ($\alpha = 0.05$). ^a phenotype verification by visual inspection (Supplementary table S9), ^b no phenotype.

Candidate	Stress	HT	LT	p-value
OGC 15	NaCl	-	-	0.028
OGC 23	Na ₃ VO ₄	-	-	0.038
OGC 51	Na ₃ VO ₄	-	-	0.020
OGC 75	NaCl ^a	-	-	0.056
OGC 85	CsCl ^a	+	-	0.146
OGC 106	NaCl	-	-	2.16×10^{-7}
OGC 121	L-malic acid [malonic acid ^b]	+ [+]	+ [+]	0.015 [0.176]
OGC 167	ZnCl ₂ ^a	-	-	0.052
OGC 174	NaCl ^a	+	-	0.11
OGC 194	NaCl	-	-	0.005
OGC 226	L-malic acid [CsCl ^b]	+ [-]	+ [-]	0.001 [0.069]
OGC 231	NaCl [L-malic acid]	- [+]	- [+]	0.027 [0.002]
OGC 241	HClO ₄ ^b [CsCl]	+ [+]	[-]	0.512 [0.033]

ing that especially salt conditions lead to phenotypes was also made in the low-throughput approach. Summing up, NaCl, Na₃VO₄, CsCl, and ZnCl₂ cause competitive phenotypes for 11 of 13 tested candidates.

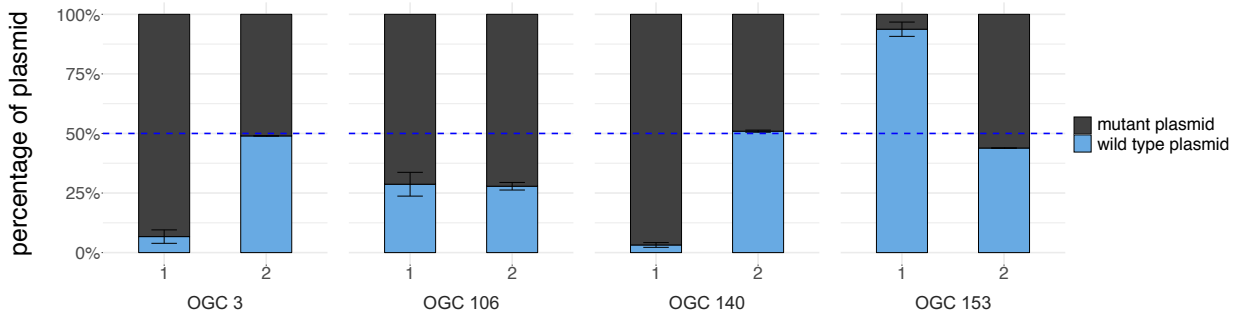
3.4.3 Stress specific phenotypes for overlapping gene candidates

Single competitive growth assays were conducted following the hypothesis that specific stress conditions drive the activity of overlapping genes. As shown in the previous section, especially salt but also other culture medium additives induce growth differences when over-

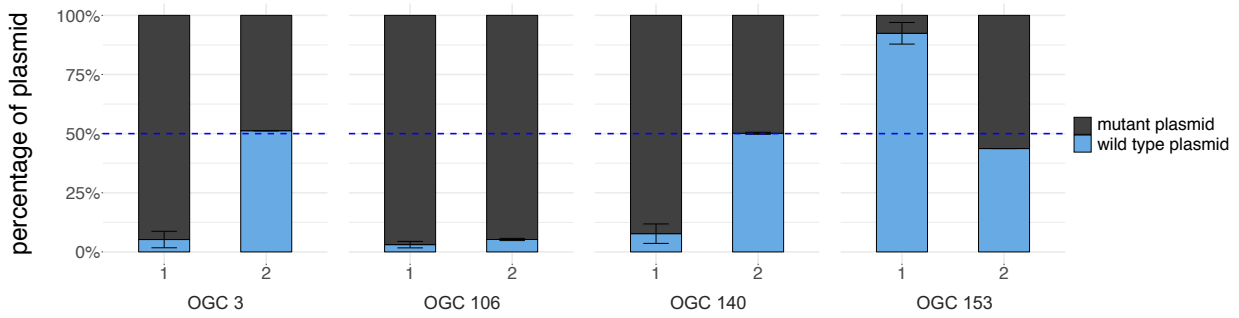
expressing intact or translationally arrested overlapping gene candidates from a plasmid. However, LB as a non-stress condition was tested for each candidate to verify the specificity of phenotypes for environmental stresses. As can be seen in Supplementary table S9, significant growth alterations were not seen in LB in most cases.

Contrary to predominant stress-specific phenotypes, four candidates showed visible significance in stress as well as non-stress conditions (p-values for LB: OGC 3: 2.77×10^{-4} , OGC 106: 1.81×10^{-4} , OGC 140: 2.31×10^{-5} , OGC 153: 2.67×10^{-3} , Figure 3.7a). To exclude genomic secondary effects which might influence the growth and be responsible for phenotypes, plasmids of these candidates were transformed in new EHEC cells and competitive growth was repeated to reproduce growth tendencies. As shown in Figure 3.7b, growth behaviors strongly differ for the repetitions of OGC 3, OGC 140 and OGC 153. Additionally, phenotypes found for these candidates (Supplementary table S9) disappeared and it can be assumed that spontaneous mutations in the genome of EHEC rather than overexpression of the OGCs are responsible for growth differences. Thus, LT phenotypes of these three candidates constitute most likely artifacts.

However, growth trends in LB as well as in stress conditions remained unchanged for OGC 106. This indicates possibly a more general function of the overlapping gene, which might not be restricted to stressful environments. Hence, it would be worthwhile to analyze this candidate in future projects.



(a) competitive growth in LB



(b) competitive growth in stress conditions

Figure 3.7: Single competitive growth phenotypes in non-stress environments. Mean percentage of wild type or mutated plasmids are shown for the original assay (1) and the repetition of the experiment in one replicate with new EHEC transformants (2) in (a) LB and (b) the stress conditions HClO_4 , NaCl, malonic acid, and acetic acid for OGC 3, OGC 106, OGC 140, and OGC 153. Values are adapted to 50% input ratio (blue dashed line) of the related t_0 condition. Error bars indicate standard deviations.

3.5 Determination of transcriptional start sites

Cappable-seq was applied to determine transcriptional start sites in the genome of EHEC in various stress and non-stress conditions. Using the sequencing output of Cappable-seq, the reproducibility of the sequencing experiment was investigated. Furthermore, optimal evaluation settings were established to detect TSS for overlapping genes. Using these criteria, genome wide as well as gene associated TSS were analyzed and their signal strengths across different conditions were examined. To strengthen evidence for functionality of TSS signals for overlapping gene candidates, bioinformatic and experimental analyses of promoter regions upstream of the TSS were conducted.

3.5.1 Cappable-seq sequencing output

Cappable-seq is a method published by Ettwiller *et al.* (2016), for transcriptional start site determination offered commercially by the vertis Biotechnologie AG, Freising. Total RNA isolated from EHEC sampled from four growth conditions and two growth phases (according to Table 2.10 and Figure 3.8) was used for Cappable-seq. The RNA was sequenced using Illumina NextSeq 500 for TSS determination in biological triplicates and technical replicates with up to 11.8 million reads in the single experiment (Table 3.7).

Table 3.7: Sequencing reads in Cappable-seq. Sequencing reads ($\times 10^6$) for three biological (I-III) and one technical replicate (IIIA, IIIB) for two growth phases and four growth conditions are given. Biological replicate III comprises of experiments IIIA and IIIB resulting from sequencing Cappable library III twice (see also Section 3.5.2).

Replicate	Exponential Phase				Early Stationary Phase			
	LB	MM	Acid	Salt	LB	MM	Acid	Salt
I	7.4	9.3	9.0	11.8	10.6	10.5	10.8	9.7
II	10.1	8.4	11.4	10.0	10.0	10.4	9.6	9.6
III	17.1	18.1	19.3	18.1	15.7	15.7	16.6	17.2
IIIA	7.9	8.2	10.4	9.1	6.5	6.7	7.4	7.9
IIIB	9.3	9.9	8.9	9.0	9.2	9.0	9.2	9.3

Table 3.8: Mapped reads in Cappable-seq. Numbers of mapped reads (in million) for three biological replicates (I-III) are given. **(a)** Total mapped read (added up). **(b)**, **(c)** Mapped reads for the extracted tagRNA-seq data sets (TSS-set, PSS-set). Values for two growth phases and four growth conditions are given in each case. rtRNA: Proportion of reads mapping to rRNA and tRNA sequences averaged over biological replicates.

(a) Total mapped reads

Replicate	Exponential Phase				Early Stationary Phase			
	LB	MM	Acid	Salt	LB	MM	Acid	Salt
I-III	30.9	31.6	35.5	34.9	31.8	31.3	32.2	31.7
rtRNA	19%	25%	19%	22%	23%	27%	30%	28%

(b) Mapped reads in TSS-set

Replicate	Exponential Phase				Early Stationary Phase			
	LB	MM	Acid	Salt	LB	MM	Acid	Salt
I	5.1	6.2	6.5	7.8	6.7	6.2	6.2	6.0
II	7.1	5.4	8.0	6.3	5.3	5.6	4.6	6.4
III	12.7	11.2	14.5	13.1	9.3	8.4	9.5	10.6
rtRNA	11%	14%	11%	13%	11%	15%	14%	14%

(c) Mapped reads in PSS-set

Replicate	Exponential Phase				Early Stationary Phase			
	LB	MM	Acid	Salt	LB	MM	Acid	Salt
I	0.7	1.7	1.2	2.2	2.0	2.6	2.3	1.8
II	1.9	1.9	2.0	2.4	3.3	3.4	3.0	1.8
III	3.4	5.2	3.4	3.1	5.1	5.7	5.9	5.1
rtRNA	51%	51%	50%	54%	50%	44%	66%	63%

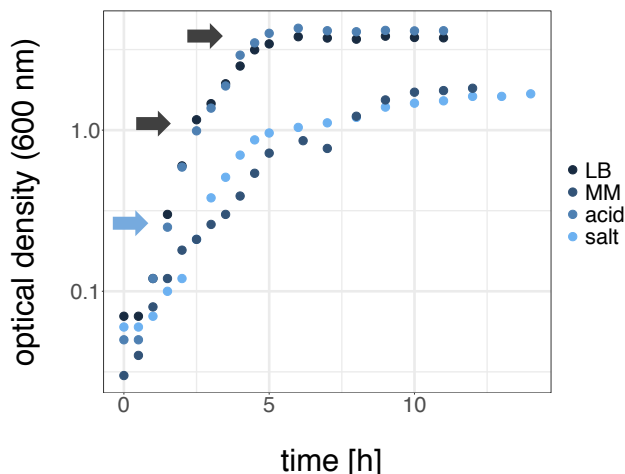


Figure 3.8: Growth curves of EHEC for Cappable-seq. Growth was recorded in LB, minimal medium M9 (MM), LB supplemented with 4 mM L-malic acid (acid), and LB supplemented with 500 mM NaCl (salt). Optical densities for cell harvest in exponential phase (blue arrow, $OD_{600} = 0.2\text{--}0.3$ in all conditions) and early stationary phase (black arrows, $OD_{600} = 3.5\text{--}4$ in LB and LB+acid, $OD_{600} = 1\text{--}1.8$ for LB+salt and MM) are indicated.

Cappable-seq includes an enrichment procedure of 5' mRNA fragments from total RNA by means of tagging RNA 5' triphosphates and, consequently, indirect depletion of processed 5' monophosphorylated RNAs like rRNAs and tRNAs. Therefore, the success of mRNA extraction can be examined by analyzing the proportion of reads mapping to rRNA and tRNA regions, which account for approximately 80 % and 14 % of total RNA, respectively (Westermann *et al.*, 2012). For the analyzed samples, more than 30 million reads (biological replicates added up) mapped to genomic regions. 19 %–30 % thereof mapped to rRNA and tRNA genes though no rRNA removal step was performed (Table 3.8a). This result indicates efficient accumulation of non-processed primary RNA (mRNA).

An approach called tagRNA-seq was additionally applied during sample preparation (Innocenti *et al.*, 2015). Differential ligation of varying sequence tags to Cappable-enriched RNA allowed labeling of primary and contaminating processed RNAs according to their 5' phosphorylation status (Figure 2.1). Consequently, two data sets were extracted from the sequencing reads, TSS-set and PSS-set, whereby the former constantly contains a higher number of mapped reads (Table 3.8b and 3.8c). Investigating the mapped reads regarding the RNA species present revealed that the percentage of rRNA and tRNA mapping reads is

decreased in the TSS data set despite a higher number of total mapped reads compared to the PSS-set; in particular, rRNA and tRNA represent on average 11%–15% in the TSS-set and 44%–66% in the PSS-set. Therefore, tagRNA-seq is an appropriate method within the Cappable-seq workflow to further reduce the proportion of rRNA and tRNA. Although only 53% (on average) of PSS-set reads can be explained by rRNA and tRNA, further processed RNAs and in parts degradation products are included in the PSS-set. Thus, tagRNA-seq lowers interfering background signals in general. However, the TSS-set is not completely without background, as shown in Section 3.5.4.

3.5.2 Reproducibility of Cappable-seq

For reproducibility calculations, mapped sequencing reads were applied to the first part of the Cappable-seq specific sample processing workflow (Script 2). In this analysis, reads are trimmed to one nucleotide long reads representing the 5' base of the original read. A relative read score ($RRS_{io} = \frac{n_{io}}{N} * 10^6$, equation 2.3) at each genome position on each genome strand is calculated (details in Section 3.5.3). The reproducibility of Cappable-seq was determined by calculating pairwise Pearson's product-moment correlation coefficients r between biological and technical replicates (Table 3.9). All genome positions showing a minimum of one trimmed read and therefore $RRS > 0$ in at least one of two compared replicates were included in the calculation.

Pearson's r has values > 0.6 in all comparisons of biological replicates which indicates a linear relationship of moderate strength (as described in Section 3.3.1). Further, more than 80% of correlations are strong ($r > 0.8$). The mean correlation in early stationary growth phase is slightly increased compared to exponential growth phase ($r = 0.90$, $r = 0.83$), which might indicate a less fluctuating RNA composition in stationary phase.

Sequencing of Cappable library III was performed twice and resulted in data sets IIIA and IIIB. Correlation analysis showed a very strong linear relationship between the sequencing experiments ($r > 0.999$) independent of the analyzed condition (Table 3.9). Therefore, these technical replicates were combined to dataset III.

Summing up, the reproducibility is excellent in all conditions considering that sample collection of biological replicates was completely independent regarding bacterial growth,

Table 3.9: Correlation of Cappable-seq data sets. Pairwise Pearson’s product moment correlation coefficients r were calculated for genome positions with $RRS > 0$ in at least one of two correlated replicates. Replicate III consists of the technical replicates IIIA and IIIB.

Replicate	Exponential Phase				Early Stationary Phase			
	LB	MM	Acid	Salt	LB	MM	Acid	Salt
I + II	0.85	0.81	0.66	0.77	0.91	0.91	0.91	0.82
I + III	0.95	0.85	0.91	0.84	0.97	0.89	0.98	0.86
II + III	0.88	0.92	0.67	0.84	0.94	0.91	0.97	0.74
mean r	Exponential phase: 0.83				Stationary phase: 0.90			
III:A+B	0.99997	0.99993	0.99997	0.99992	0.99997	0.99987	0.99998	0.99988

RNA isolation and Cappable-seq sample preparation. Variability in the data sets can be traced back to these experimental steps alone since sequencing with Illumina NextSeq is highly reliable as shown.

3.5.3 Cutoff analysis for genome wide TSS

In general, Cappable-seq is used to determine transcriptional start sites at a genome-wide scale. Ettwiller *et al.* (2016) provided a suite of two programs for this purpose, both taking input parameters to optimize data evaluation (Script 2). An analysis was conducted to evaluate the optimal parameter for the first program to detect genome wide TSS. For the second program, Ettwiller *et al.* (2016) worked out the optimal value for the input parameter, which was adopted (see next paragraph). The evaluation is presented for the data set of exponential growth in LB medium (replicate I), but trends observed are similar for all other conditions and replicates.

The first program, as described briefly in Section 3.5.2, trims all sequencing reads starting at the 3’ end to 1-bp long reads representing the first 5’ base. The number of reads at each genome position are counted and the RRS is calculated according to $RRS_{io} = \frac{n_{io}}{N} * 10^6$. Only genome positions with a minimum number of reads are maintained as putative transcription

start sites. The cutoff value $minRRS$ is the variable input parameter for the first program and specifies the threshold to retain a genomic position as a potential TSS.

As the DNA polymerase exhibits an uncertainty in transcription initiation (Ettwiller *et al.*, 2016; Hawley and McClure, 1983), nearby transcription start sites have to be clustered to the position with the highest RRS (Figure 3.9a). This is performed with the second program which takes the optimal input parameter $dist = 5$, specifying the distance of TSS which are clustered. Clustering transcriptional start sites yields a reduced set of TSS (Figure 3.9b), but it is a reasonable step get precise TSS information.

To define the best threshold for the input parameter $minRRS$, the number of transcription start sites in the genome are determined before and after clustering for $minRRS$ ranging from 0 to 20 (Figure 3.9b). As can be seen, the number of TSS decreases with increasing cutoff value in each case. The extraordinarily high number of genomic TSS positions for low values of $minRRS$ can be explained with the calculation of the RRS. Rearranging Equation 2.3 ($RRS_{io} = \frac{n_{io}}{N} * 10^6$) for the number of reads n at a genomic position i on the strand o (n_{io}) shows that for example as little as three or five reads are necessary to reach $minRRS = 0.5$ or $minRRS = 1$, respectively, for a total number of reads of $N = 5 \times 10^6$. Consequently, all positions with such low RRSs are classified as TSS, if a low value for $minRRS$ is applied. As said before, the number of TSS decreases for higher thresholds. However, the slope flattens over the data curve. While the number of clustered TSS is reduced by a factor of 4.2, if $minRRS$ is increased from 0 to 0.5, the fold change is lowered to 1.7 or 1.2, if $minRRS$ is increased from 0.5 to 1.0 or 1.0 to 1.5, respectively. As a further increase of $minRRS$ does not lead to a substantial reduction, $minRRS = 1.5$ is adequate for evaluation, which is in accordance with Ettwiller *et al.* (2016).

The TSS workflow detected 13 689 clustered genome wide transcription start sites for $minRRS = 1.5$ (Supplementary table S10), a value comparable to the unprecedented number of TSS published by Ettwiller *et al.* (2016), for *E. coli* MG1655 (16 359, $minRRS = 1.5$). However, the number of false positives might be reduced if an increased cutoff value is used, e.g., $minRRS = 5$ (Supplementary table S10, Figure 3.9c). Visual inspection might be a necessary step to reliably determine TSS, regardless of the RRS threshold.

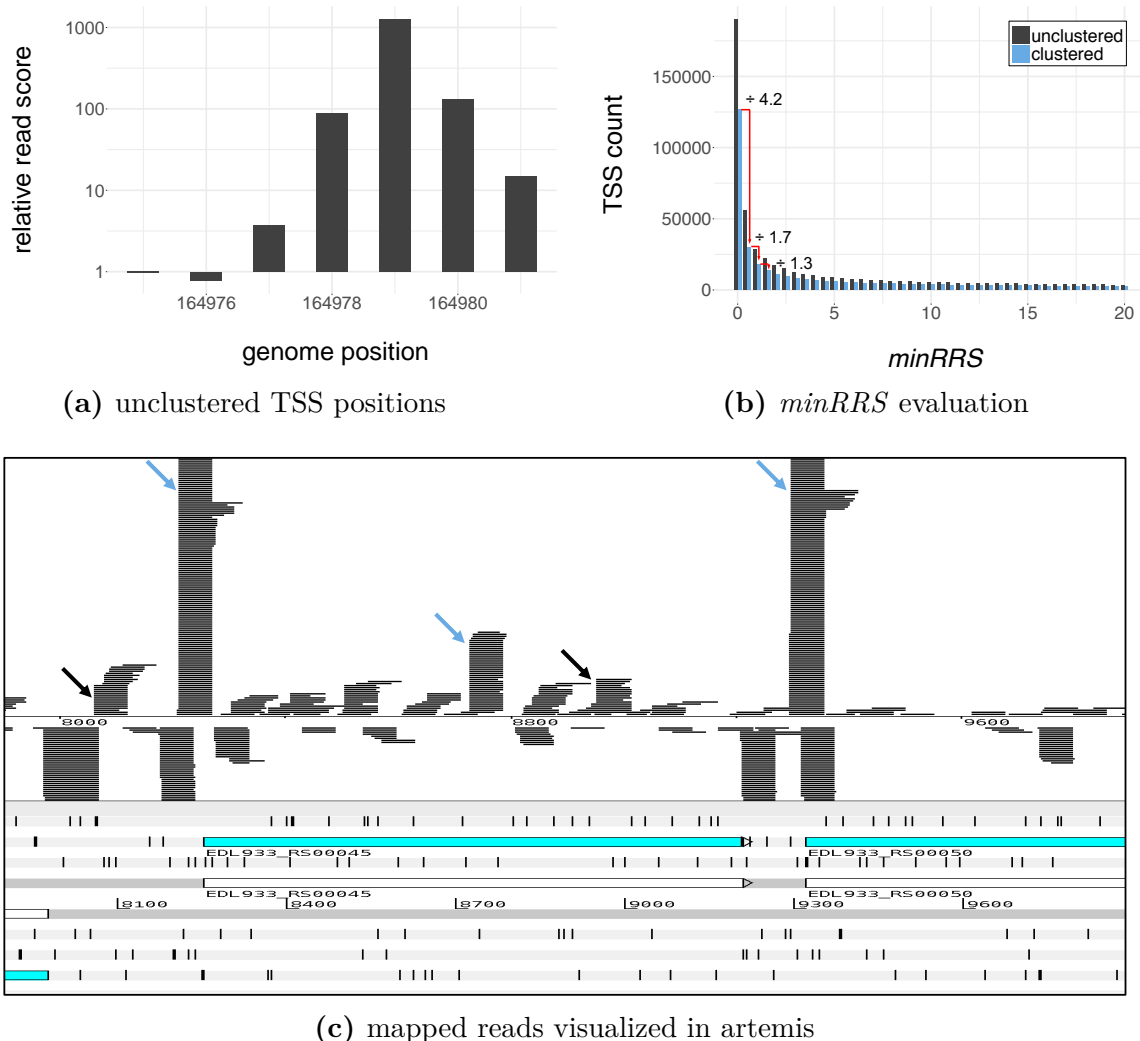


Figure 3.9: Cutoff value evaluation for genome wide transcriptional start sites. (a) Relative read score (RRS) for genomic positions around the clustered TSS are shown (position 164979 with highest RRS). (b) Number of TSS identified depending on the cutoff $minRRS$ before (gray) and after (blue) clustering nearby TSS positions within 5 bp. The decrease of clustered TSS is highest for low cutoff values ($\frac{n(TSS_{minRRS1})}{n(TSS_{minRRS2})}$ for $minRRS1 = 0$ and $minRRS2 = 0.5, 4.2$; for $minRRS1 = 0.5$ and $minRRS2 = 1, 1.7$; for $minRRS1 = 1$ and $minRRS2 = 1.5, 1.3$; for $minRRS1 = 1.5$ and $minRRS2 = 2, 1.3$) (c) Exemplary extract of the genome of EHEC with mapped reads from replicate I in LB, exponential phase. black arrows, TSS at $minRRS = 1.5$; blue arrows, TSS at $minRRS = 5.0$.

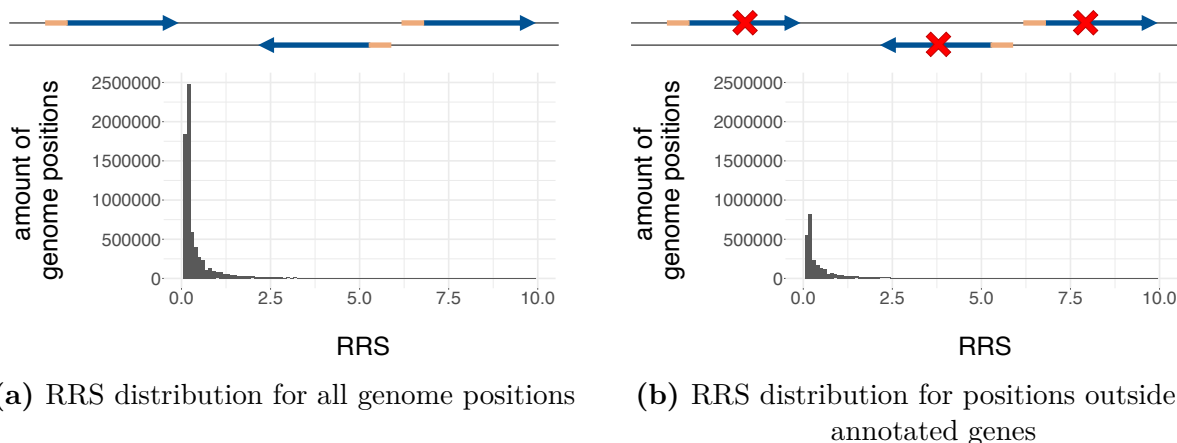
3.5.4 Cutoff analysis for antisense TSS

The main task of Cappable-seq in this project was the identification of transcription start sites for antisense overlapping genes. Therefore, a particular cutoff value analysis including all 24 sequenced samples was conducted to find reasonable parameters for this task.

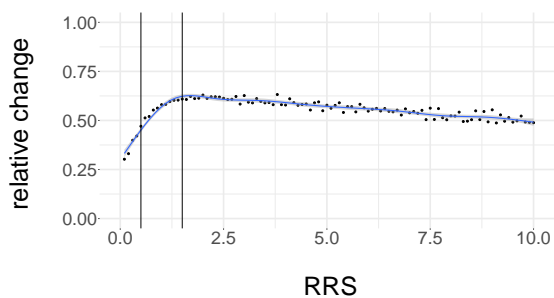
It is generally expected that transcription start sites are located closely upstream of the start codon of annotated genes (AGs), but many more signals are detected, either within or outside AGs, which might originate from degradation of RNA or represent individual start sites (Figure 3.9c). To distinguish real TSS signals outside annotated gene regions from background signals, the background noise arising from annotated genes is estimated and transferred to ORFs located in intergenic regions or antisense to annotated genes (methodological details are described in Section 2.9.5). For this, counts of genome positions with certain RRSs were analyzed. All positions in the genome as well as genomic regions without annotated genes regions (i. e., coding DNA sequence and upstream area where the TSS is expected are excluded) were investigated (Figures 3.10a and 3.10b). Similar to Figure 3.9b, the number of genome positions with a certain RRS decreases with increasing RRS for both genomic areas analyzed. Based on these histograms in Figure 3.10, the relative change of the number of genome positions at a certain RRS between both sets was calculated according to Equation 2.4 (Figure 3.10c). A low value for the relative change indicates a high change in the composition of the compared data sets at a specific RRS, whereas a high value indicates a low change. The highest change by far was seen for very low RRSs (< 0.5), which can be explained by an extensive removal of uninformative genomic positions representing background noise. The relative change increases rapidly until the maximum of approx. 0.62 is reached at $RRS = 1.5$, where most signals are maintained in both datasets indicating reliable TSS. Further on, the relative change declines slowly, which shows the deletion of reliable AG associated TSS resulting in smaller values for the relative change. In contrast, for extremely high values of the RRS the relative change reaches 1 (Supplementary figure S1). This indicates that highly expressed TSS are even located in genomic regions, where no TSS would be expected.

The analysis shows that genome positions representing noise of annotated genes probably have $RRS < 0.5$, but positions with $RRS \geq 1.5$ most likely reflect genuine transcription start

sites. Therefore, also genome positions in unexpected genomic regions between or antisense to annotated genes with a RRS greater or equal to 1.5 probably do not represent noise. Thus, using $\text{minRRS} = 1.5$ for the first program of the TSS workflow is an appropriate threshold to reliably assign transcription start sites for antisense or intergenic ORFs. Additionally, the shape of the curve indicates that TSS with $0.5 \leq \text{RRS} < 1.5$ might be true TSS as well, but these TSS have to be verified with independent methods, e. g., promoter activity analysis or transcriptional start site determination with 5' RACE.



(a) RRS distribution for all genome positions (b) RRS distribution for positions outside annotated genes



(c) relative change of distributions

Figure 3.10: Evaluation of RRS for antisense and intergenic parts of the genome. (a), (b) Frequencies of genomic positions with indicated RRS for (a) all genome positions and (b) genome positions outside annotated genes (AG and 100 bp upstream excluded). Upper panels illustrate analyzed genomic regions. (c) Relative change of frequencies (Equation 2.4, $\text{relative change} = \frac{(b)}{(a)}$) at the indicated RRS. A cubic square smooth function is placed on the data (blue line). Key RRSs at 0.5 and 1.5 are visualized (black vertical lines).

3.5.5 Gene associated transcriptional start sites

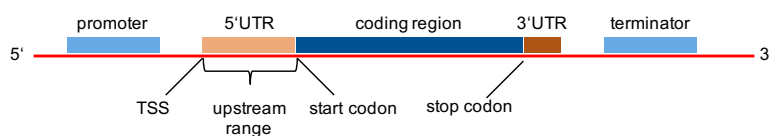
Based on the overall structure of a transcriptional unit (Figure 3.11a), previously identified genome wide TSS were examined for gene association, i. e., start sites in the proximity of the start codons of genes. The analysis focused only on reproducible TSS, i. e., start sites present in all three biological replicates of one analyzed condition (according to Section 2.9.6).

In general, the transcriptional start site is the first base of the 5' UTR, whose length varies between different transcripts. Therefore, a 5' UTR analysis was performed to estimate an appropriate range upstream of the start codon where TSS might be located. The distance between start codons and the gene associated TSS with a maximum 5' UTR of 500 bp was calculated for 4525 functionally annotated genes (according to Section 3.1, Figure 3.11b). The mean length of the 4265 analyzed UTRs is 149 bp, whereby half of the UTRs are 84 bp or shorter. The most common distance between TSS and start codon is 23 bp–27 bp, but a notable number of genes have a longer distance of up to 247 bp (75th percentile). The same analysis was conducted also with the smaller set of 973 hypothetical annotated genes and revealed a similar distribution of 5' UTR lengths, which is slightly skewed to larger distances (75th percentile: 317 bp). To include an adequate but suitably conservative analysis region, further work of gene associated TSS was carried out within a 250 bp upstream range.

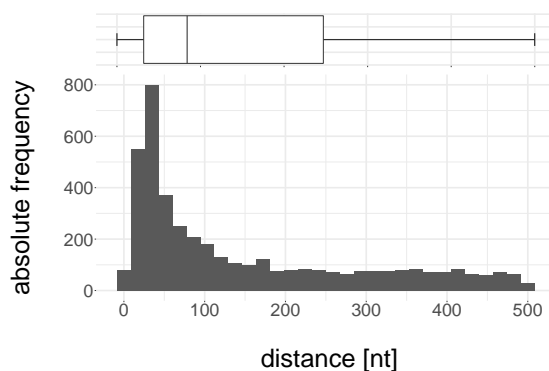
Using this analysis strategy, several gene clusters were investigated for the presence of upstream transcription start sites (Table 3.10, details described in Section 3.1). Annotated genes were divided into functional and hypothetical annotated genes. Additionally, all possible antisense embedded open reading frames, as well as a subset with blastp hits against the RefSeq database were investigated. Furthermore, overlapping gene candidates previously analyzed using Western blots (Section 3.2) and phenotyping (Sections 3.3 and 3.4) were examined regarding upstream TSS.

The numbers of genes with any upstream TSS for each of the gene sets are listed in Table 3.11. The analysis was conducted with two values of *minRRS* (*minRRS* = 1.5 and *minRRS* = 5, according to Sections 3.5.3 and 3.5.4).

For *minRRS* = 1.5, TSS were identified for 2571 functionally annotated genes representing 57% of all genes. A similar percentage was detected for translated OGCs, whereas substantially fewer hAGs as well as embedded ORFs are associated with a transcriptional



(a) DNA coding strand



(b) 5' UTR length distribution

Figure 3.11: Gene structure and 5' UTR characteristics. (a) Simplified structure of the coding strand of a bacterial gene including the coding region defined by start and stop codon, 5'/3' untranslated regions (5'/3' UTR), and regulatory elements controlling the transcription such as promoter and terminator. (b) 5' UTR length distribution of TSS upstream of functional annotated genes. The range 500 bp upstream of 4265 genes is screened for TSS with $RRS \geq 5$. Distance between TSS and start codons is specified in nucleotides (nt). Boxplot displays minimum (0 bp) and maximum (500 bp), 25th percentile (32 bp), median (84 bp), and 75th percentile (247 bp) of the 5' UTR lengths.

Table 3.10: Short description of gene sets for TSS identification. Details are described in Section 3.1.

Gene Set	Number of Genes	Characteristics
fAG	4525	functional annotated genes, i. e., gene product is not hypothetical
hAG	973	annotated genes, but hypothetical, i. e., gene product is hypothetical
embORF	30 870	antisense embedded ORFs, i. e., completely overlapping an annotated gene
<i>blastp</i>	218	subset of embORFs with blastp hit against RefSeq
OGCs	216	embedded and partial overlapping gene candidates with ribosomal profiling signal in EHEC

start sites. For an increased value of *minRRS*, the number of genes with upstream TSS decreases, as expected. Nevertheless, 46 % of functional annotated genes had a TSS signal at this threshold, whereas the number of overlapping ORFs with TSS at the increased cutoff is reduced to 9 % to 26 % (embORF, blastp, and OGCs in Table 3.11). This indicates weaker TSS signal strengths for overlapping genes than for annotated genes. For most gene sets, the absolute counts of unique TSS in a specific gene set (specified in Table 3.13) exceed the number of genes in the respective set. This finding implies transcription initiation of some genes at different start sites. All embedded ORFs show the opposite trend, possibly indicating that a single transcription start site is used to initiate transcription of several genes even as cistronic transcript. This observation can be explained by a huge number of genes in the gene set of embORFs ($\geq 30\,000$), resulting in a high ORF density which allows fewer TSS for the ORFs. This, in combination with a shorter average gene length of those genes compared to any other gene sets (Supplementary figure S2), might be responsible for operon-like structures detected with the TSS finding algorithms, where these ORF(s) are located in the 5' UTR of a downstream gene. Looking at the length distribution of those embORFs with a TSS shows that the ORFs are indeed short compared to OGCs with TSS

(see below) supporting the above stated assumption (Supplementary figure S3).

Table 3.11: Gene associated transcriptional start sites for annotated genes and antisense ORFs. Number of genes/ORFs with reliable TSS (i. e., present in three replicates of Cappable-seq) located at most 250 bp upstream of the start codon are listed. The percentage of genes/ORFs with TSS are indicated in brackets. Values are given for two *minRRS* values.

Gene Set	<i>minRRS</i> = 1.5	<i>minRRS</i> = 5
fAG	2571 (57 %)	2097 (46 %)
hAG	450 (46 %)	286 (29 %)
embORF	7064 (23 %)	2702 (9 %)
<i>blastp</i>	44 (20 %)	21 (10 %)
OGCs	112 (52 %)	56 (26 %)

TSS in proximity to OGC start codons are listed in Supplementary table S11. The algorithm identified transcription start sites for 127 overlapping gene candidates with a maximum distance of 250 bp between TSS and start codon at *minRRS* = 1.5. Visual inspection of the sequencing reads showed that TSS signals for 15 candidates likely belong to annotated genes (Supplementary file 5), thus reducing the number of OGCs with TSS to 112 as listed in Table 3.11. However, the arrangement of these overlapping genes and the corresponding annotated genes could indicate co-transcription (further investigated in Section 3.5.6). Transcriptional start sites of 23 candidates are stable in all eight investigated conditions indicating highly reliable TSS. In contrast, some TSS seem to be specific for growth phases or growth conditions. A detailed analysis of TSS signal strengths is shown in Section 3.5.7. Furthermore, TSS for 66 % of candidates with HT-phenotype (35 out of 53, Table 3.5) and 62 % of candidates with LT-phenotype (8 out of 13, Table 3.6) were identified. In general, the mean length of OGCs with TSS is approximately 235 bp, thus, slightly increased in comparison to the ORF length of emORFs with TSS. However, 89 OGCs do not have any TSS. The option for bi- or polycistronic expression of these candidates along with other genes was tested (Section 3.5.6).

3.5.6 Detection of putative operon structures of overlapping genes

As mentioned before, 89 overlapping gene candidates do not have a TSS within 250 bp upstream. Additionally, as also said above, the TSS of 15 candidates belong most likely to nearby annotated genes and hence, OGCs might be co-transcribed as bi- or polycistronic RNA. Therefore, overlapping genes were analyzed regarding their genomic localization and the possibility to be part of an operon.

A total number of 2379 operons of *E. coli* K12 (downloaded from Database of prokaryotic Operons, DOOR) were analyzed concerning their inter-gene distance (Figure 3.12). The space between genes within operons is < 150 bp, apart from one exception (919 bp). Thus, OGCs with a distance above 150 bp to an upstream gene were not considered for possible operon structures.

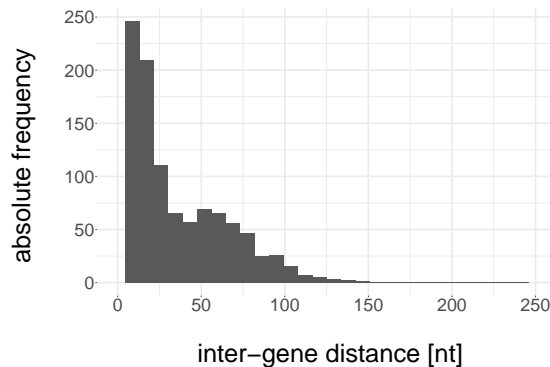


Figure 3.12: Inter-gene distance of 4146 *E. coli* K12 operon genes. The histogram is limited to 250 bp distance between genes within operons for better visualization.

The operon analysis revealed that 3 and 16 of those OGCs without TSS based on standard selection criteria according to Section 3.5.5 might be co-transcribed with upstream OGCs or annotated genes, respectively (Table 3.12a). Six annotated genes thereof do not have a TSS (standard selection criteria), but might be co-transcribed itself with the upstream AG.

Furthermore, eight candidates, which have a TSS, are putatively parts of operons together with annotated genes (Table 3.12b, left two columns). However, it cannot be deduced from the data whether the OGCs are transcribed only from the associated TSS or also as cistronic transcript along with the annotated gene. Further experiments are necessary to determine the transcriptional unit, for instance analyze whether a rho-independent terminator is located

between the annotated gene and overlapping gene candidate. Finally, seven small OGC pairs exhibited the same positions as transcription start site (Table 3.12b, right column). Thus, co-transcription of these candidate pairs could be assumed, too.

Putative operon structures of the previously excluded 15 OGCs (Section 3.5.5) could be detected for four candidates with an upstream AG and for a further seven candidates co-transcription with the downstream AG might be possible (Table 3.12c).

Although the analysis showed that some OGCs might be part of operons, absence of a TSS is not necessarily an indication of absence of transcription. For instance, 63 OGCs are neither directly associated with a TSS nor part of polycistronic transcriptional units. Nevertheless, this analysis allows for updating the number of overlapping gene candidates with either their own TSS or possibly situated within an operon from 112 to 142, which increases the overall percentage to 66 %.

3.5.7 Growth phase and condition dependent TSS strength

Transcriptional start sites associated with either annotated genes or overlapping gene candidates were analyzed regarding their signal strength in different growth conditions. RRSs of the three biological replicates were used to determine TSS with significantly varying values between either growth phases or growth conditions. To determine significance, paired or Welch two-sample t-tests were used, respectively. Although other statistical tests like limma and VarMixt were shown to perform slightly better when differential expression data were evaluated elsewhere, standard t-tests are still appropriate and easily applicable for this task (Jeanmougin *et al.*, 2010).

The analysis of TSS for significantly different RRS values in various conditions was an explorative approach to get a set of transcription start sites potentially regulated in any of the analyzed conditions. Methods to adjust p-values are often used to correct for multiple testing if many pair-wise comparisons are calculated (e.g., Bonferroni adjustment or false discovery rate calculations according to Benjamini and Hochberg, 1995), as the probability to get a significant result by chance increases with an increasing number of tests performed. Since the approach was not used to unveil the transcriptional regulation of TSS sets for searching candidates, p-value adjustments were not applied. This enables to retain more

Table 3.12: Overlapping gene candidates possibly localized in operons. OGCs **(a)** without TSS and **(b)** with TSS fulfilling the criterion for operon arrangement with either annotated genes or OGCs. **(c)** OGCs with TSS previously excluded from evaluation due to possible co-transcription with annotated genes analyzed for operon structures. ^a annotated genes without TSS, possibly cotranscribed with another upstream gene.

(a) OGC without TSS co-transcribed with annotated genes or OGCs

Annotated Genes		OGCs
OGC 4 / RS_00630	OGC 173 / RS_21005	OGC 6 / OGC 5
OGC 68 / RS_07890	OGC 179 / RS_21490 ^a	OGC 9 / OGC 7/8
OGC 80 / RS_09340 ^a	OGC 180 / RS_21520 ^a	OGC 151 /
OGC 90 / RS_11020	OGC 181 / RS_21605	OGC 152/153
OGC 107 / RS_13205	OGC 186 / RS_22470	
OGC 149 / RS_19035 ^a	OGC 224 / RS_27235	
OGC 162 / RS_20120 ^a	OGC 227 / RS_27420	
OGC 167 / RS_20435 ^a	OGC 231 / RS_27620	

(b) OGC with individual TSS putatively co-transcribed with annotated genes or OGCs

Annotated Genes		OGCs
OGC 1 / RS_00035	OGC 200 / RS_23910	OGC 8 / OGC 7
OGC 24 / RS_02740	OGC 203 / RS_24685	OGC 48 / OGC 47
OGC 44 / RS_04670	OGC 230 / RS_27580	OGC 70 / OGC 71
OGC 140 / RS_17790	OGC 241 / RS_28465	OGC 74 / OGC 73
		OGC 145 / OGC 146
		OGC 159 / OGC 158
		OGC 165 / OGC 164

(c) initially excluded OGCs with or without co-transcription along with annotated genes

Upstream AGs	Downstream AGs	No Operon Structure
OGC 58 / RS_06010	OGC 12 / RS_01095	OGC 91
OGC 103 / RS_12865	OGC 20 / RS_02655	OGC 171
OGC 128 / RS_16085	OGC 23 / RS_29480	OGC 172
OGC 185 / RS_22405	OGC 41 / RS_04125	OGC 220
	OGC 70/71 / RS_08055	
	OGC 225 / RS_27320	

candidates for subsequent detailed analyses. Follow-up experiments, however, are obligatory in order to make conclusions concerning the regulation status of a TSS and the corresponding ORF or gene.

Table 3.13: Gene associated TSS with significant differences of RRSs. Transcriptional start sites identified for more than one gene within a gene set were maintained in the set of unique TSS once. RRS of three biological replicates were applied to t-test calculations (significance level: $\alpha = 0.05$). Differences in growth phases within one stress condition were assessed with a paired, differences of LB to stress conditions in the same growth phase with an unpaired t-test. TSS with at least one significant expression difference are listed. Threshold for TSS identification, $minRRS = 1.5$; maximal 5' upstream range, 250 bp.

Gene Set	Unique TSS	Significant Differences of RRS Between	
		Growth Phases	Growth Conditions
fAG	5027	3169 (63 %)	2765 (55 %)
hAG	773	453 (59 %)	390 (50 %)
embORF	4844	2662 (55 %)	2351 (49 %)
<i>blastp</i>	52	30 (58 %)	27 (52 %)
OGCs	148	83 (56 %)	61 (41 %)

A total number of 83 TSS of OGCs show significantly different RRS values when comparing exponential or early stationary growth phases, whereas expression strength of 61 TSS is significantly different in stress conditions in comparison to LB medium; 43 TSS are candidates for regulation in both growth phase and growth condition (Table 3.13, Figure 3.13, Supplementary file 6). For example, the transcription start site for OGC 171 has higher RRSs in exponential phase (Figure 3.13a). Although statistical significance was achieved for only one condition (paired t-test, p-value 0.016 in LB + acid), similar tendencies for increased RRSs are found in all remaining growth conditions. In contrast, the transcription start site of OGC 189 has an enhanced RRS when cells are grown to early stationary phase in minimal medium (Figure 3.13b). It can be seen that the RRS is significantly higher compared to exponential phase solely in minimal medium (paired t-test, p-value 6.9×10^{-3}).

Additionally, the RRS in low nutrient medium is substantially increased in comparison to RRSs in LB as well as LB based stress media in stationary phase (Welch t-test, p-value ≤ 0.05 for all conditions tested against minimal medium). Therefore, it is assumed that expression of OGC 189 increases in minimal medium specifically in stationary phase.

The percentage of TSS associated with OGCs which are differentially expressed is lower than the percentage of TSS associated with annotated genes (41 % to 56 % compared to 50 % to 63 %). However, reproducible differences in the signal for overlapping gene candidates strengthen the assumption that RNA for these putative genes is expressed in a controlled and targeted manner.

As listed in Table 3.13, 4844 TSS are localized upstream of embedded antisense overlapping ORFs. Statistical calculations revealed that more than 70 % of the TSS have significantly different RRS values in conditions analyzed (examples are shown in Figure 3.14). A more detailed analysis revealed that most TSS have differential expression patterns in growth condition and growth phase simultaneously (1560; 45 %), fewer solely in growth phase (1102; 32 %, Table 3.14), and the least in growth condition (791; 20 %) which probably can be explained by the limited number of growth conditions analyzed (i. e., LB, minimal medium, LB + L-malic acid, LB + NaCl). TSS were categorized according to their mean RRS, averaged over all conditions and replicates of one candidate, for strongly (mean RRS > 5), moderately ($1.5 \leq \text{mean RRS} \leq 5$) and weakly (mean RRS < 1.5) expressed positions. Although the selection criterion for TSS determination was set to $\text{minRRS} = 1.5$ according to Section 3.5.4, the mean RRS can drop below 1.5 since even data below this threshold were included in calculations. Therefore, the expression level reflects an overall rather than a specific expression of the TSS. Regardless, where differential expression was detected, the proportion of TSS with either weak and strong expression is low (16 %–35 % and 21 %–27 %, respectively), as most TSS have a moderate RRS between 1.5 and 5. The antisense embedded ORFs associated with a TSS were analyzed regarding their length and the start codon of the longest open reading frame in order to find patterns for specifically expressed TSS (Figure 3.15). The mean length of the ORFs in each group is almost equal, thus, there is no tendency for an increased length of genes expressed at higher values (Figure 3.15a). Furthermore, there seems to be no general trend for differences in start codon choice of antisense embedded

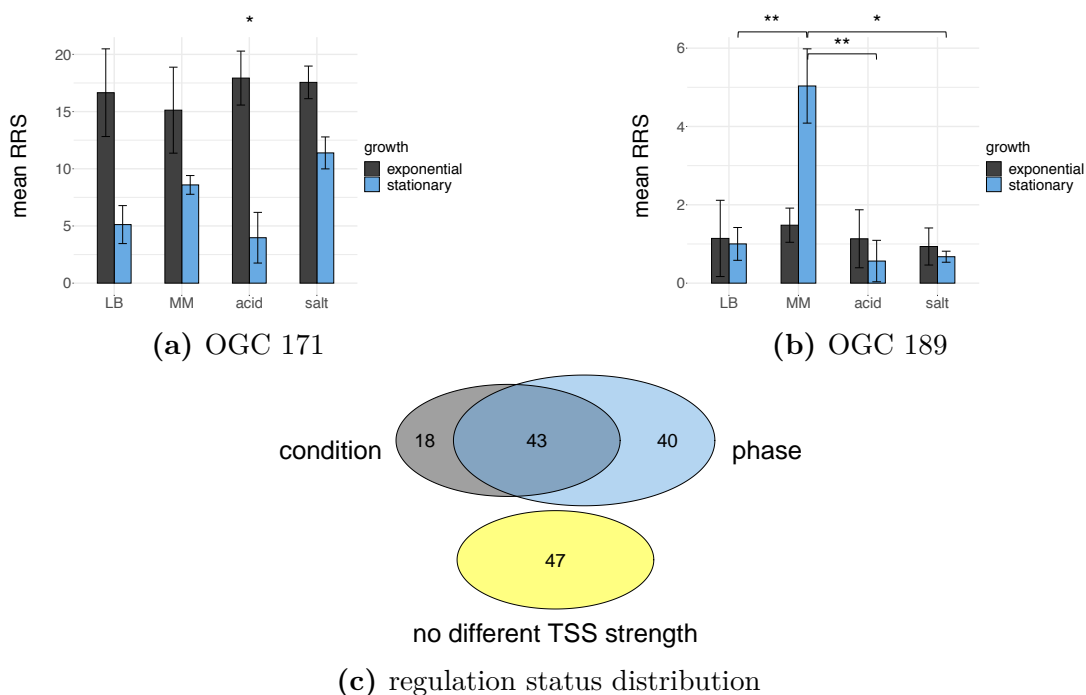


Figure 3.13: Differential RRS signals of OGC associated TSS. Examples for **(a)** growth phase dependent RRS for TSS, and **(b)** growth condition dependent RRS for TSS. Mean RRS of TSS expressed in cells grown in LB medium, M9 minimal medium (MM), or LB supplemented with L-malic acid (acid) or NaCl (salt) to exponential (grey) or early stationary phase (blue) are shown. Error bars indicate the standard deviation. Statistical significance tested with a paired (growth phase dependent) or a Welch (growth conditions dependent) t-test ($\alpha = 0.05$; * $p \leq 0.05$; ** $p \leq 0.01$). **(c)** Distribution of TSS with significant different expression strength visualized in a Venn diagram. TSS were categorized by RRS differences in growth condition compared to LB (gray), in growth phase (blue), in growth phase and growth condition (darker blue) or no significant RRS differences (yellow).

ORFs (Figure 3.15b). Nevertheless, ATC and CTG start codons are most abundant for ORFs with a differentially expressed TSS at any level of expression. The start codons ATG and GTG have lowest frequencies in almost all categories. In any case, ATG as well as further NTG start codons, apart from CTG, do not seem to be the most preferential start codons for the longest possible ORFs, but it cannot be ruled out that shorter ORFs, starting with ATG/GTG/TTG, exist in the respective reading frame. These results show that there is a huge number of transcription start sites antisense to annotated genes having significantly different RRS values in growth phases or growth conditions. Furthermore, these antisense TSS are localized upstream of overlapping open reading frames, but the TSS analyses do not allow drawing any conclusions on whether an antisense RNA (asRNA, i. e., non-coding RNA) or a specific ORF-bearing RNA (i. e., mRNA) is transcribed and translated in the latter case. Further experiments are necessary to address this issue.

Table 3.14: Detailed expression patterns of TSS of embedded ORFs. Absolute and relative numbers of differentially expressed gene associated TSS of embedded ORFs are given. TSS are categorized according to the overall expression level of the TSS (mean RRS, averaged over all conditions and replicates; weak, mean RRS < 1.5; moderate, $1.5 \leq \text{mean RRS} \leq 5$; strong: mean RRS > 5) and the expression status (de: differentially expressed; nd: not differentially expressed; p: significant RRS differences in growth phase; c: significant RRS differences in growth condition; pc: significant RRS differences in growth phase and condition). Threshold for TSS identification, $\text{minRRS} = 1.5$; maximal 5' upstream range, 250 bp.

	Differentially Expressed in				
	de	nd	p	c	pc
no. of TSS	3453	1391	1102	791	1560
weak	991 (29 %)	224 (16 %)	243 (22 %)	207 (26 %)	541 (35 %)
moderate	1631 (47 %)	826 (59 %)	560 (51 %)	416 (53 %)	655 (42 %)
strong	831 (24 %)	341 (25 %)	299 (27 %)	168 (21 %)	364 (23 %)

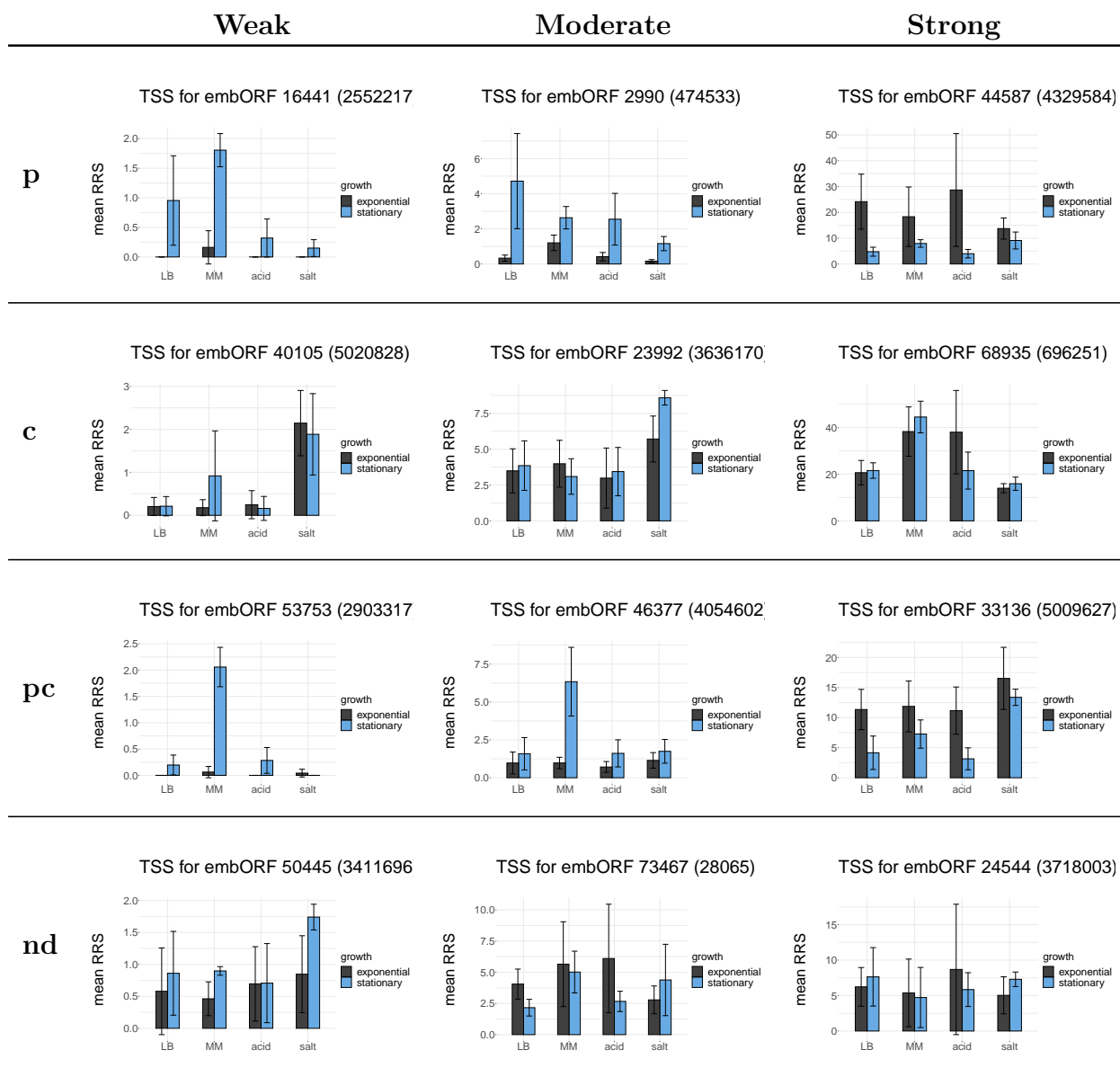


Figure 3.14: Differential expression strength for TSS of antisense ORFs. Examples for three different expression strengths (weak: mean RRS < 1.5 ; moderate: $1.5 \leq \text{mean RRS} \leq 5$; strong: mean RRS > 5) and four differential expression states (p: growth phase; c: growth condition; pc: growth phase and condition; nd: not differentially expressed) are displayed.

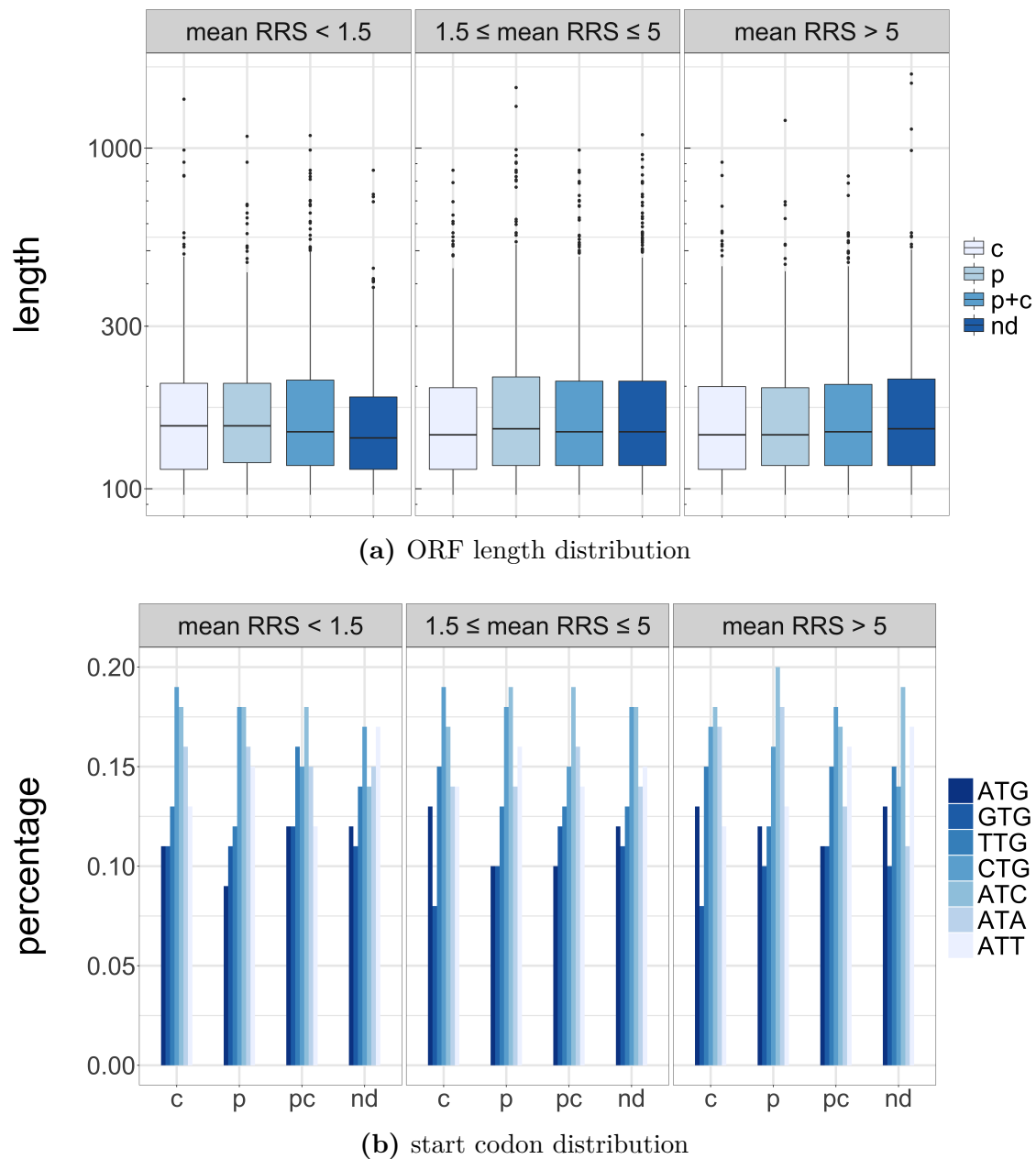


Figure 3.15: Analysis of embedded antisense ORFs with transcription start site. ORFs with the most upstream-localized start codon in the respective reading frame were analyzed. (a) Logarithmic representation of length of ORFs downstream of regulated and unregulated TSS at three different overall expression strengths of the corresponding TSS. (b) Percentage of ORFs with start codons as indicated. Signals are distributed according overall TSS strengths and expression types according to Figure 3.14.

3.5.8 Bioinformatic and experimental analysis of promoters

Promoters are a regulatory element upstream of transcription start sites. Extensive studies on bacterial promoters revealed that their structure is well conserved. A bacterial promoter consists of a highly conserved -10 region (Pribnow-box) and a less conserved -35 region (Section 1.3.1).

Conservation patterns of upstream regions of TSS identified for functionally annotated genes and OGCs were investigated to rate the activity of their start sites. Thus, genomic sequences covering 100 bp upstream of TSS in either gene set were used to construct a sequence logo. Random genome positions and randomly chosen transcription start sites regardless of location and association with genes or ORFs were used as negative and positive controls, respectively. As expected, a highly conserved -10 region was found in the upstream region of TSS associated with functionally annotated genes, but also with OGCs. The positive control (i. e., randomly chosen set of reliable transcription start sites) showed the same conserved regions (Figure 3.16). In accordance with published studies (Figure 3.16e, Singh *et al.*, 2011), the degree of sequence conservation of the Pribnow-box is in all cases higher in comparison to the -35 region with a slightly conserved thymine residue (position -36). Furthermore, the TSS shows a slight tendency for having more thymine in the DNA sequence, whereas the -1 position has a higher preference for purine bases. In contrast to TSS specific promoter patterns, random genome regions revealed no conserved pattern in this analysis.

The tools bTSSfinder and BPROM were used to identify specific promoter sequences of selected candidates bioinformatically (Shahmuradov *et al.*, 2017; Solovyev and Salamov, 2011). OGCs were picked from different categories: TSS cutoff value, differential TSS pattern, number of gene associated TSS (Table 3.15). Promoter test sequences of ranges between 50 bp and 162 bp (according to Table 3.16) were introduced into a promoterless GFP-plasmid (pProbe-NT). Promoter activity was assessed by measuring GFP fluorescence conferred by the test sequences and compared to background fluorescence caused by the empty plasmid.

The ability of the applied GFP assay to detect promoter activity was tested in advance, although the assay was used successfully in previous studies (e. g. Fellner *et al.*, 2015). For this, promoters of the annotated gene *heliD*, coding for a DNA helicase, as well as the

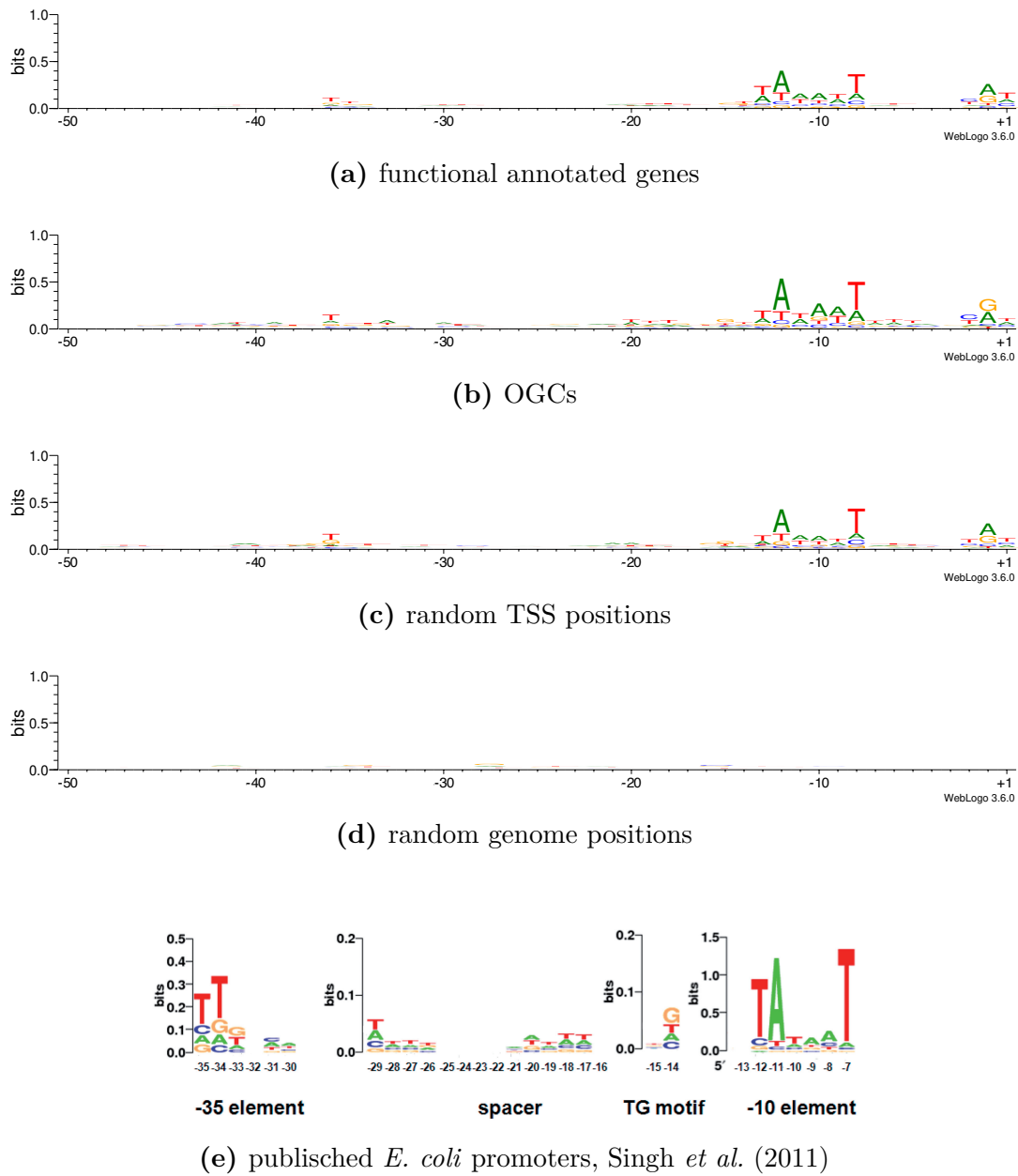


Figure 3.16: Sequence logos of TSS upstream regions and published *E. coli* promoters. Sequence conservation of aligned upstream regions (100 bp) of gene associated TSS for (a) functional annotated genes ($n = 5027$), (b) OGCs ($n = 148$), (c) random TSS positions ($n = 148$), and (d) random genome positions ($n = 148$) is shown. TSS, +1 position. Sequence logos were created with WebLogo 3 (Crooks *et al.*, 2004). (e) Sequence logo of 554 *E. coli* promoters with mapped TSS and determined -10 region (Singh *et al.*, 2011; Mitchell *et al.*, 2003). Promoter elements are indicated.

promoter of OGC 15 were examined. In both cases, significantly increased fluorescence was detected for cells carrying the plasmid with GFP under the control of the promoter test sequences (Figures 3.17a and 3.17b, Table 3.16). Consequently, assay performance as well as promoter activity were verified.

In a second experiment, OGC 15 promoter variants were analyzed to assess the impact of three core promoter regions (-35 region, spacer region, and -10 Pribnow-box, compare with Section 1.3.1) on the promoter activity and to examine the relationship between activity and sequence conservation (Figure 3.17c). Four or five point mutations were introduced independently in each of these promoter sections and GFP fluorescence mediated by each of the newly created sequences was measured. Mutations in the -35 region hardly alter the activity of the promoter. Selected changes in the separating sequence led to a reduced fluorescence by roughly ten-fold. However, fluorescence is still significantly higher than the background signal (Table 3.16). An even higher reduction was seen for the modified -10 region. Values comparable to the promoter-less GFP construct were measured.

In summary, the highly conserved -10 region seems to be essential for promoter activity and, to a lesser extent, the spacer region despite lacking sequence conservation in any overall comparison. The less conserved -35 region affects the promoter activity only slightly - as can be seen from mean fluorescence values obtained by intact and mutated constructs having a similar order of magnitude. Thus, this region is not crucial for the basic promoter activity tested here.

Table 3.15: Overview of candidates selected for promoter analysis. Gene name, TSS detection criterion (*minRRS*), presence of differential TSS expression, and number of gene associated TSS (thereof analyzed in GFP assay) are listed.

Candidate	<i>minRRS</i>	Differential Regulation	TSS (Analyzed)
<i>helD</i>	5	no	1
OGC 15	5	no	1
OGC 85	5	↑ in minimal medium	1
OGC 96	1.5	↑ in salt	2 (2)
OGC 135	1.5	↑ in stationary phase	3 (1)
OGC 136	5	no	3 (2)
OGC 207	≫5	no	1
OGC 226	0.5	no	2 (2)

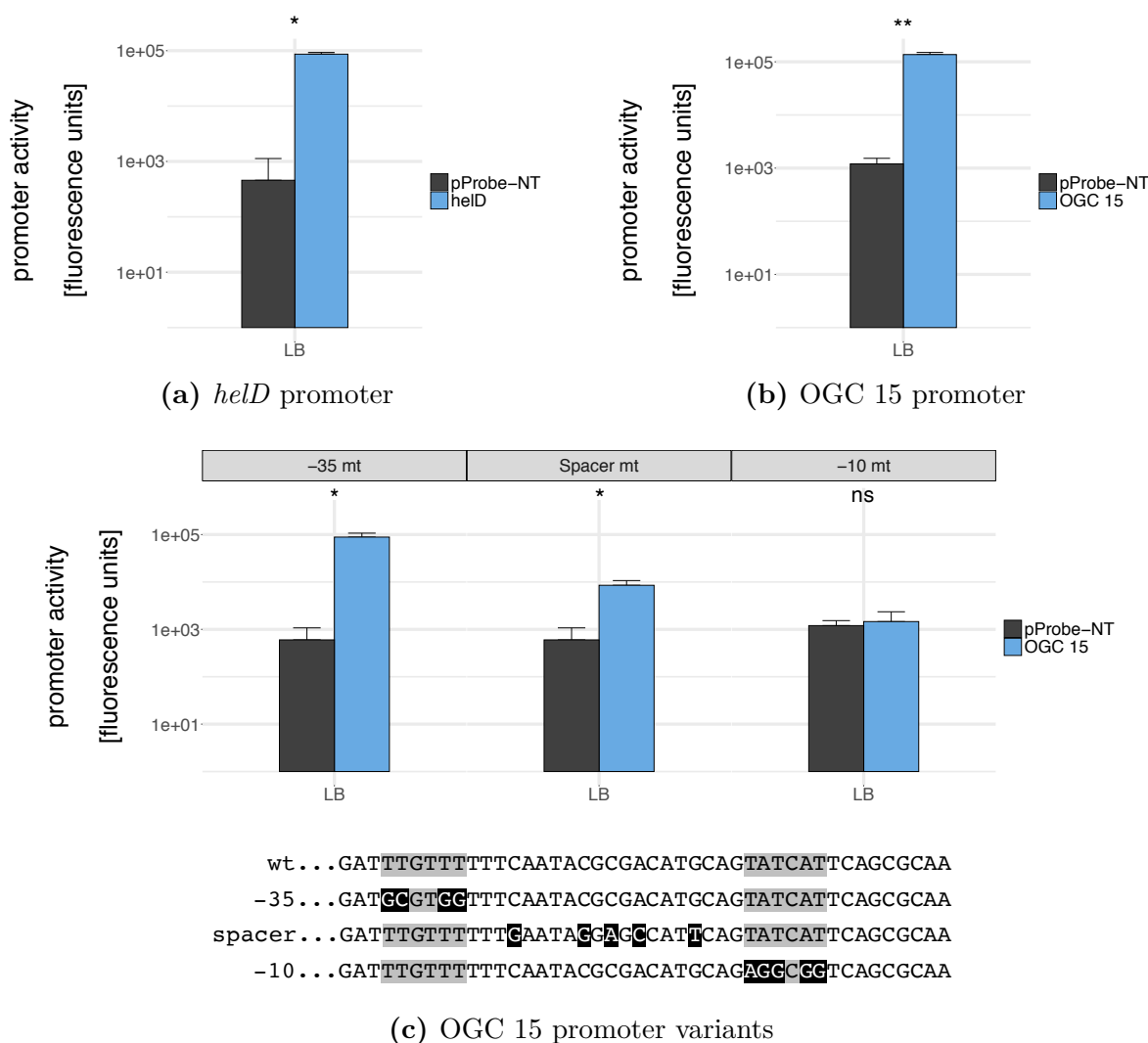
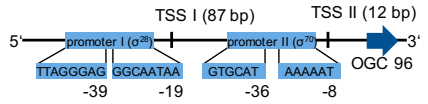
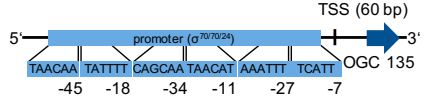
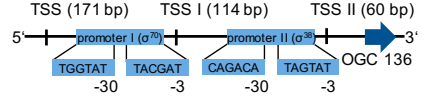
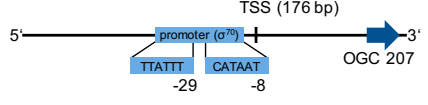
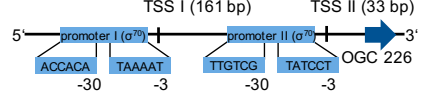


Figure 3.17: GFP assay of test promoters. Promoter activity of (a) *helD* promoter, (b) OGC 15 promoter, and (c) OGC 15 promoter variants (created by point mutations (highlighted black) in the respective promoter regions). All experiments were conducted in biological triplicates in *E. coli* Top10 cultivated in LB medium. As control, fluorescence of promoter-less vector pProbe-NT was measured. Mean values are shown. Error bars indicate the standard deviation. Statistical significance between fluorescence levels of tested plasmids was calculated with a Welch t-test ($\alpha = 0.05$). * $p \leq 0.05$; ** $p \leq 0.01$; ns, not significant

Table 3.16: Characteristics and activity of promoters analyzed in GFP assay. Promoter name, position of the gene associated TSS, length of the promoter test sequence with position of sequence ends indicated in brackets, identified promoter activity based on p-value (two-tailed Welch t-test, significant increased fluorescence of promoter construct compared to empty vector, $\alpha = 5 \times 10^{-2}$) and schematic overview of promoter localization are listed. Sequence differences of OGC 15 promoter variants are shown in Figure 3.17c. In the promoter representations, the position of the TSS (black vertical bar) with respect to the start codon of the gene is indicated in base pairs in brackets. Numbers below promoter boxes represent the position of the 3' base of the respective box regarding the TSS.

Promoter	TSS	Test Sequence (End)	Activity	p-value	Schematic Promoter Region
<i>helD</i>	1241246	100 bp (TSS)	yes	LB: 2×10^{-3}	
OGC 15	300491	63 bp (TSS)			
wt			yes	LB: 3×10^{-3}	
-10 mt			no	LB: 7×10^{-1}	
spacer mt			yes	LB: 2×10^{-2}	
-35 mt			yes	LB: 1×10^{-2}	
OGC 85	1985980	100 bp (TSS)	yes	LB: 1×10^{-2} MM: 1×10^{-2} acid: 3×10^{-3} salt: 2×10^{-2}	

Table 3.16: Continued from previous page

Promoter	TSS	Test Sequence (End)	Activity	p-value	Schematic Promoter Region
OGC 96 I	2285573	60 bp (TSS)	yes	LB: 2×10^{-2} salt: 7×10^{-3}	
OGC 96 II	2285498	60 bp (TSS)	yes	LB: 3×10^{-2} salt: 7×10^{-3}	
OGC 135	3218689	70 bp (TSS)	no	exp: 9×10^{-1} stat: 5×10^{-1}	
OGC 136 I	3226911	50 bp (TSS)	yes	LB: 8×10^{-3} salt: 1×10^{-2}	
OGC 136 II	3226857	50 bp (TSS)	yes	LB: 4×10^{-2} salt: 2×10^{-2}	
OGC 207	4867699	100 bp (TSS)	yes	LB: 6×10^{-2} MM: 8×10^{-3}	
OGC 226 I	5307090	162 bp (TSS+22 bp)	yes	LB: 6×10^{-3} acid: 2×10^{-3}	
OGC 226 II	5306962	72 bp (TSS+2 bp)	yes	LB: 1×10^{-1} acid: 8×10^{-2}	

Further on, remaining promoters (listed in Table 3.15) were investigated for their activity using the GFP assay. Three promoter sequences were tested, in which the TSS showed different expression patterns between growth phases or growth conditions: OGC 135, OGC 85 and OGC 96. Two further candidates, OGC 226 and OGC 136, exhibited several TSS upstream of the assumed start codon. For their transcriptional start sites, different promoter regions have been analyzed.

For OGC 135, Cappable-seq revealed three gene associated TSS in the proximity of the start codon, of which one displayed notably higher RRSs in stationary phase compared exponential phase (mean RRS: 0.5 (exponential) and 3.1 (stationary), Figure 3.18a). bTSSfinder and BPROM detected three different promoters (one σ^{24} and two σ^{70} promoters, Table 3.16) within 50 bp upstream of this TSS. The promoter test fragment (70 bp) included all three possible promoter regions and the promoter activity was determined in exponential as well as in stationary growth phase of the cells. Although the strengths of the predicted promoters were rated relatively high by the programs compared to other promoters analyzed here (BPROM, LDF σ^{70} , -6.53 ; bTSSfinder, σ^{70} -score, 1.96 ; bTSSfinder, σ^{24} -score, 1.93), no increased fluorescence above background was detected in any of the tested conditions (Figure 3.18b, Table 3.16). Therefore, promoter activity is absent and no statement can be made about differential expression of the TSS in different growth phases based on the activity of these associated promoters.

The transcriptional start sites of OGC 85 showed high RRSs and thus, seems to be a reliably determined position (Figure 3.18c). It is striking that especially cultivation to stationary phase in minimal medium resulted in a significantly increased TSS signal when compared to complex LB medium (disregarding any supplements). Promoter activity of the upstream region of this TSS was verified in all tested culture conditions (Table 3.16). Additionally, a slightly increased activity in minimal medium compared to the activity in LB was found for the promoter region, which is in line with the expression pattern of the associated TSS. Nevertheless, this trend was just outside a classification as significant (two-tailed Welch t-test, p-value 0.057). In contrast, the promoter activity in M9 minimal medium did not differ from the activity recorded in supplemented LB (two-tailed Welch t-test, p-value 0.50 for LB+L-malic acid and p-value 0.32 for LB+NaCl). In summary, activity of the promoter region of

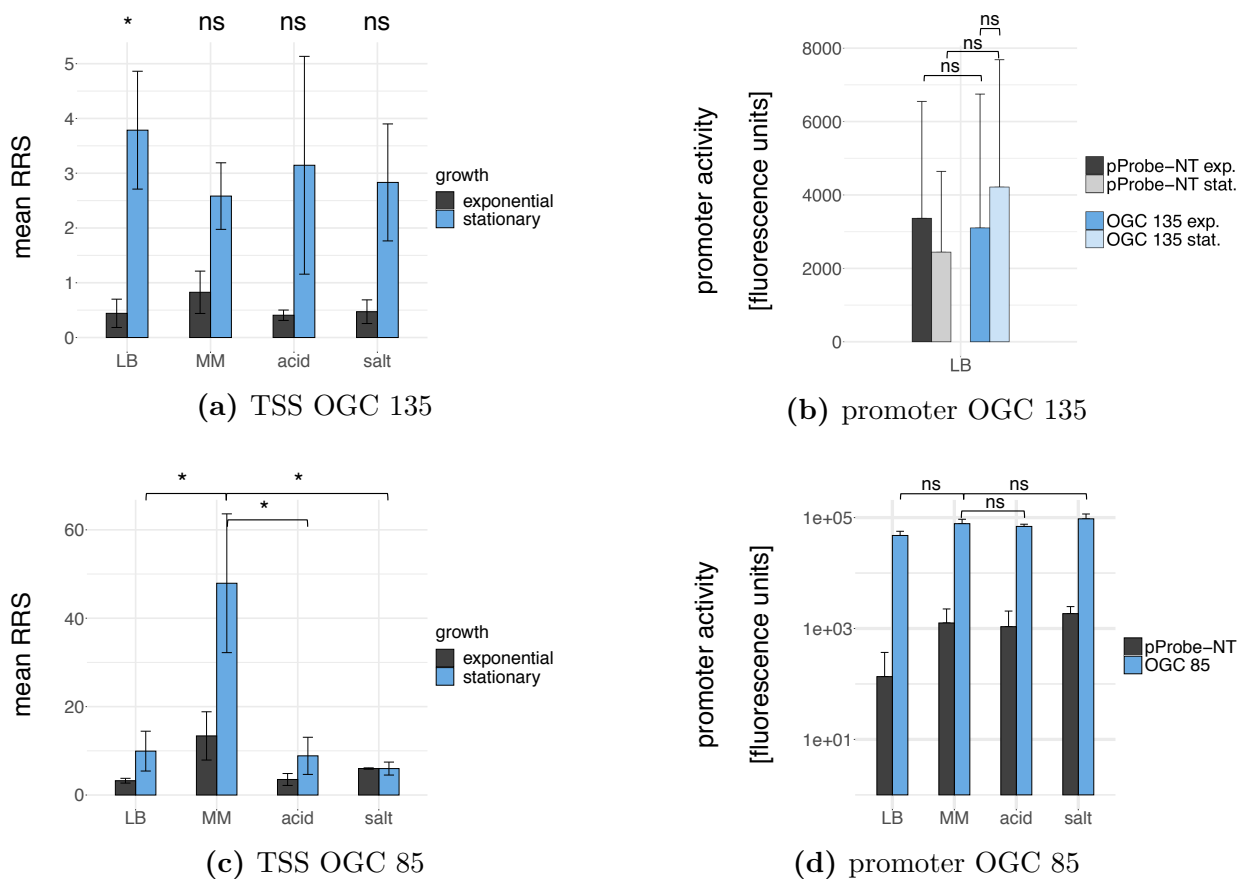


Figure 3.18: TSS and putative promoters for OGC 135 and OGC 85. **(a)**, **(b)** OGC 135. **(c)**, **(d)** OGC 85. Experiments (Cappable-seq, GFP assay) were conducted in biological triplicates. Mean RRSs of gene associated transcriptional start sites (**(a)**, **(c)**) and mean fluorescence of the promoter in GFP assay representing promoter activity (**(b)**, **(d)**) are shown. Error bars indicate standard deviations. Statistical significance of RRS values was calculated as described in Figure 3.13. Statistical significance of relevant comparisons was tested with a two-tailed Welch and a paired t-test ($\alpha = 0.05$, Section 2.7). * $p \leq 0.05$; ns, not significant.

OGC 85 was proven. Slightly increased GFP fluorescence in minimal medium compared to plain LB might confirm the observed differential expression.

OGC 96 was analyzed for TSS specifically expressed in salt supplemented LB. A special feature of this OLG is the presence of two transcriptional start sites upstream of the start codon. Both TSS have tendencies for or significantly increased RRSs when cells are grown to stationary phase in LB supplemented with NaCl (two-tailed Welch t-test; TSS I, p-value 0.082; TSS II, p-value 0.037; Figure 3.19). Promoter activities of predicted σ^{70} promoters were examined in the standard growth medium LB, as well as in the aforementioned cultivation stress (NaCl), in which the increased TSS expression was found. The analysis revealed that, on the one side, both tested promoters result in significantly increased GFP fluorescence indicating active promoters (Table 3.16) and on the other side it could be shown that salt stress leads to increased promoter activity of both promoters.

OGC 226 exhibited two transcriptional start sites ($\text{minRRS} = 0.5$) as with OGC 96. The strengths of the bioinformatically predicted σ^{70} promoters, rated according to the LDF score by BPROM, varied (promoter I, 1.58; promoter II, 0.5). The decreased LDF score for promoter II indicates reduced accuracy and specificity compared to promoter I. This tendency found bioinformatically is supported by a significantly decreased fluorescence in the GFP assay. Consequently, the activity of promoter II was lower compared to promoter I (Figure 3.20). Nevertheless, even promoter II tends to exhibit some promoter activity, although far lower than promoter I. This finding is independent of the growth conditions tested, which were selected based on overexpression effects described earlier (Tables 3.16 and 3.6). Besides this, the activity of promoter I of this gene is lower compared to any of the previously described regulatory sequences found to be active, which is in agreement with a low RRS of the associated transcriptional start sites.

A further candidate with several independent transcriptional start sites is OGC 136. Two TSS and the upstream sequences have been tested in promoter experiments. The mean RRS of TSS I of this gene in stationary phase was substantially lower than the mean RRS of TSS II ($6.1 < 54.9$, Figure 3.21). Although both independently tested promoter sequences caused significantly increased fluorescence by GFP expression, the activity of promoter I is considerably lower than of promoter II, which is in agreement with the strength of the

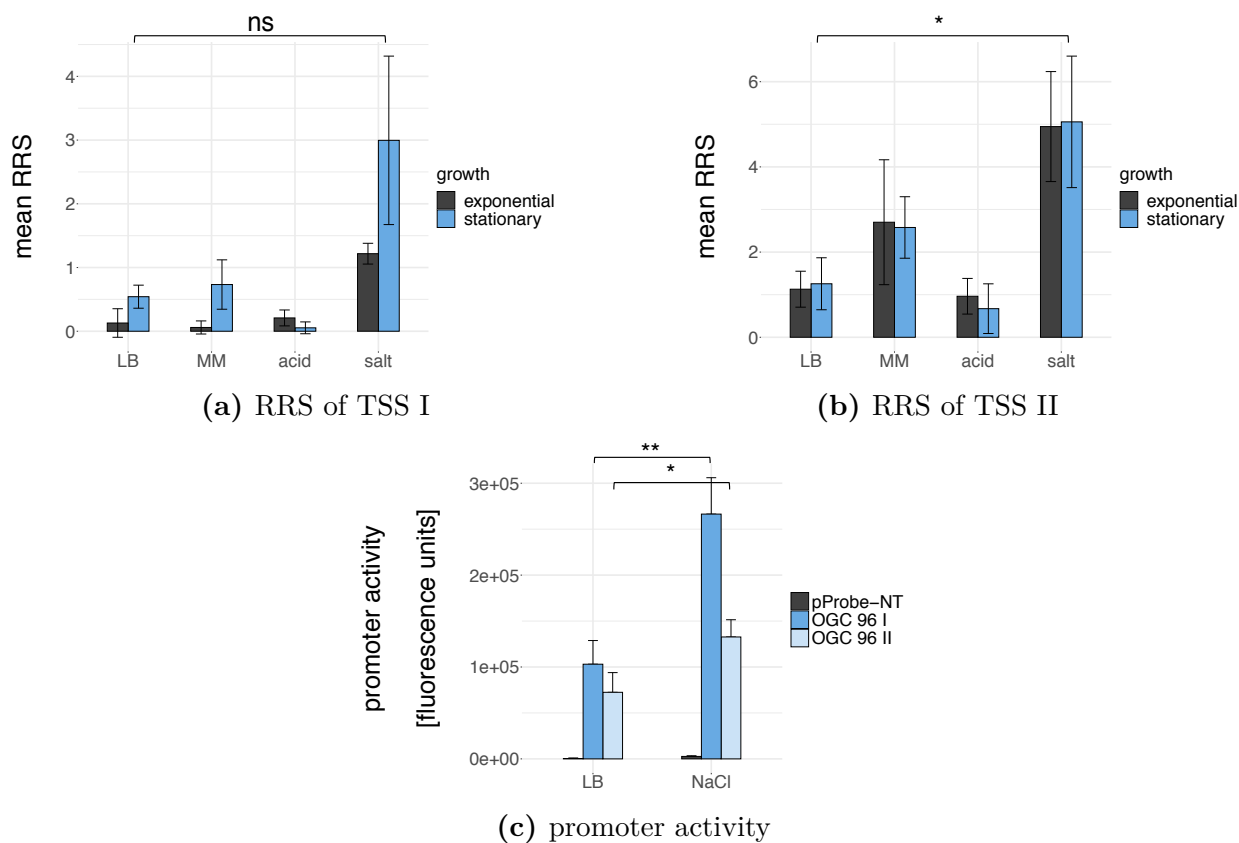


Figure 3.19: Differentially expressed TSSs and promoters of OGC 96. **(a)**, **(b)** Mean RRSs of gene associated transcriptional start sites. **(c)** Mean fluorescence representing activity of the promoter in GFP assay. Experiments (Cappable-seq, GFP assay) were conducted in biological triplicates. Error bars indicate standard deviations. Statistical significance of relevant comparisons was tested with a two-tailed Welch t-test ($\alpha = 0.05$). * $p \leq 0.05$; ** $p \leq 0.01$; ns, not significant.

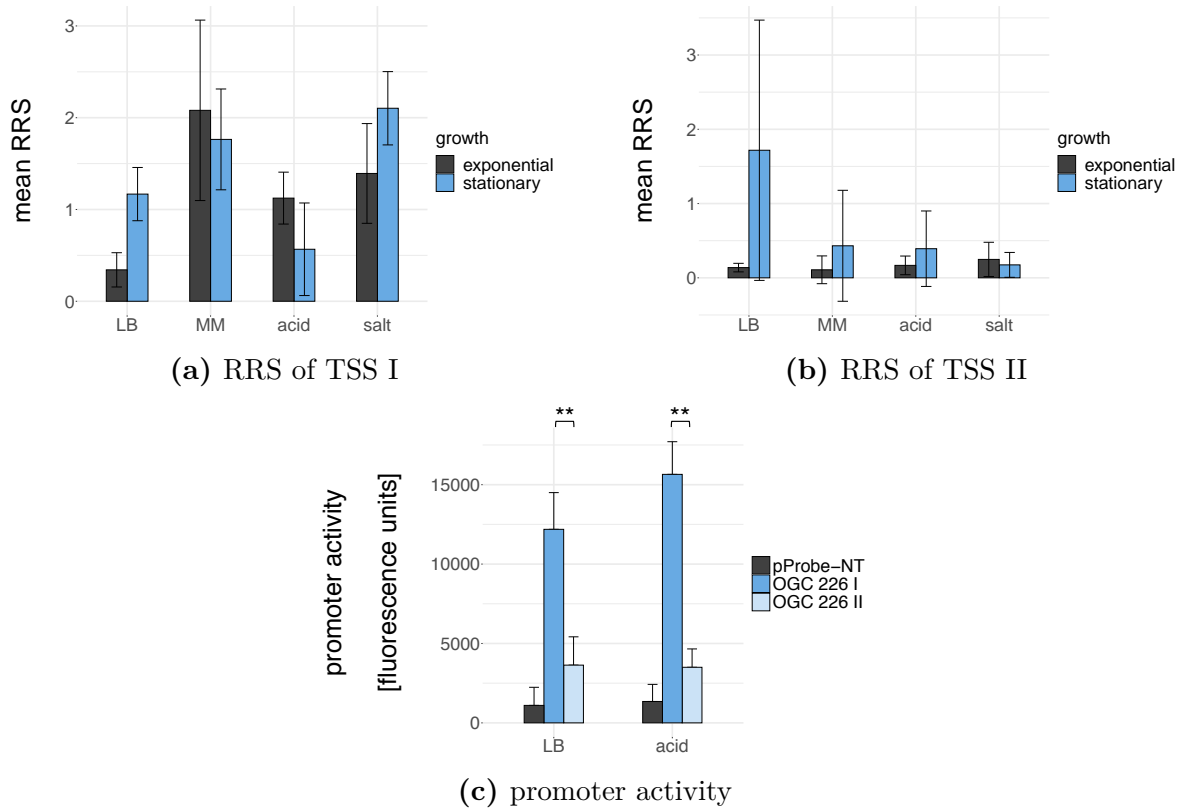


Figure 3.20: TSSs and promoters of OGC 226. (a), (b) Mean RRSs of gene associated transcriptional start sites. (c) Mean fluorescence representing activity of the promoter in GFP assay. Experiments (Cappable-seq, GFP assay) were conducted in biological triplicates. Error bars indicate standard deviations. Statistical significance of relevant comparisons was tested with a two-tailed Welch t-test ($\alpha = 0.05$). * $p \leq 0.05$; ** $p \leq 0.01$.

identified TSS (i. e., their RRS).

Finally, a promoter was tested which is located upstream of a transcriptional start site with an exceptionally high RRS throughout all experiments (mean RRS in exponential phase, 1957; in stationary phase, 3158; Figure 3.22a). Therefore, it was presumed that also the promoter activity will be higher compared to any other promoter analyzed. First, a fluorescence signal was detected, indicating the tested sequence to be active. Furthermore, promoter activity is ten-fold higher than any of the other promoters tested (Figure 3.22b). The overlapping ORF downstream of this TSS is OGC 207, which produces a detectable protein visible in Western blots, but was unremarkable in phenotypic analysis so far. How-

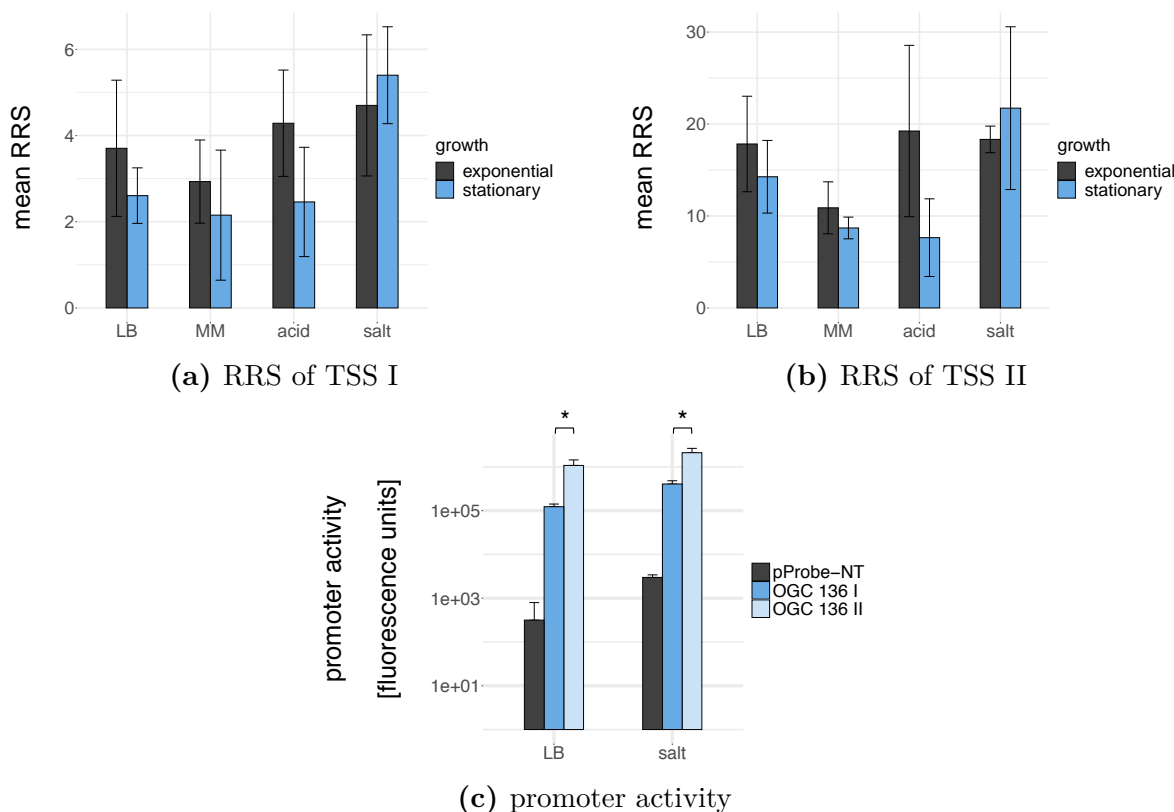


Figure 3.21: TSSs and promoters of OGC 136. (a), (b) Mean RRSs of gene associated transcriptional start sites. (c) Mean fluorescence representing activity of the promoter in GFP assay. Experiments (Cappable-seq, GFP assay) were conducted in biological triplicates. Error bars indicate standard deviations. Statistical significance of relevant comparisons was tested with a two-tailed Welch t-test ($\alpha = 0.05$). * $p \leq 0.05$; ** $p \leq 0.01$.

ever, due to the strong TSS and the stable protein product, this ORF might be a promising overlapping gene candidate for further experimental analysis.

In summary, it could be shown that unambiguously predicted promoters exhibit detectable activity when tested in GFP assays. Furthermore, the activity recorded is connected to the strength of the TSS (i. e., RRS). However, for a valid description of any promoter and the TSS connected to it, experimental validation of the native promoter is essential.

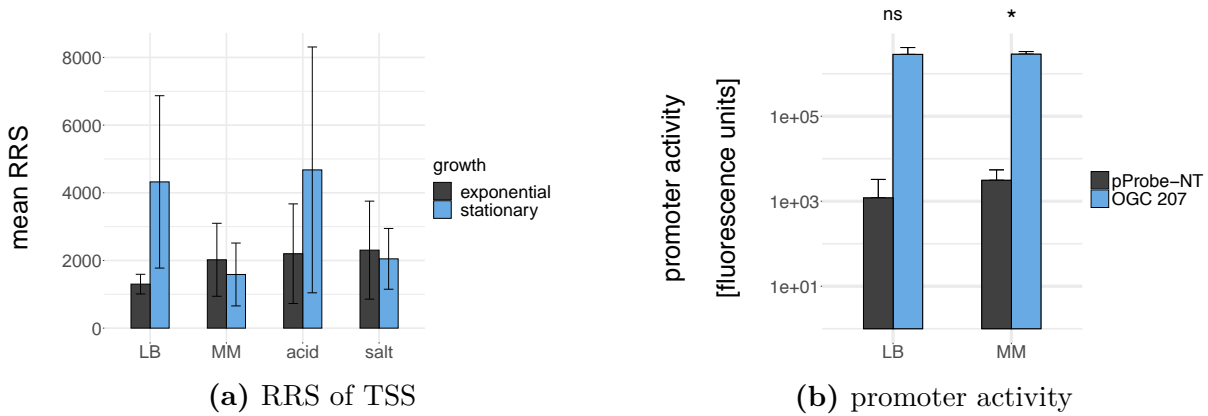


Figure 3.22: TSS and promoter of OGC 207. **(a)** Mean RRSs of gene associated transcriptional start site. **(b)** Mean fluorescence representing activity of the promoter in GFP assay. Experiments (Cappable-seq, GFP assay) were conducted in biological triplicates. Error bars indicate standard deviations. Statistical significance of relevant comparisons was tested with a two-tailed Welch t-test ($\alpha = 0.05$). * $p \leq 0.05$; ns, not significant.

3.6 Identification of sense overlapping ORFs

Cappable-seq data were screened for transcriptional start sites within annotated genes, which might be a hint for the transcription of sense overlapping genes. After distinguishing putative TSS from gene background (specified in Section 3.6.1), start sites were visually examined and the possibility of these TSS to be associated with sense overlapping ORFs were evaluated.

3.6.1 Differentiation of Cappable-TSS from gene background

Transcriptional start sites identified by Cappable-seq are distributed all over the genome but the number of genome wide TSS exceeds by far the number of annotated genes (> 10 000 TSS compared to 5498 annotated genes). Although some TSS are associated with antisense overlapping genes and ORFs (Section 3.5.5), a considerable number of TSS have no clear connection to either annotated genes or antisense OLGs based on chosen selection criteria. Visual inspection of the Cappable-seq data in Artemis revealed that many start sites are found even within annotated genes. However, 5' ends are typically not expected at these positions (Figure 3.23). There are two possibilities of the origin of these positions:

- 1) gene background: degradation sites of mRNA labeled during sample preparation ('contaminants', discussed in Section 4.3.2)
- 2) individual transcriptional start sites: TSS for perhaps alternative short products of annotated genes or sense overlapping ORFs.

Cappable-seq data were evaluated regarding potential TSS for 16 556 embedded and 1045 3' partial sense overlapping ORFs, which are reproducible across biological replicates. In total, transcriptional start sites were automatically detected with methods described in Section 2.9.6 for between 114 and 171 partial and between 1530 and 2348 fully embedded ORFs, respectively, depending on different growth media and growth phases (Table 3.17a). These putative sense ORF associated TSS had to be differentiated from the background signals originating from mRNA degradation of annotated genes (i. e., gene background). For this, the signal-to-noise ratios (S/N) of RRSs of the TSS and the highest background signal position were calculated ($\frac{RRS_{TSS}}{RRS_{noise}}$, illustrated in Figure 3.23a, details of the method are also

described in Section 2.9.11). Start sites with an increased ratio of at least 1.5 above background in all biological replicates were considered as true TSS signals (Table 3.17b). With this restriction, the number of genes exhibiting an upstream TSS is lesser than before, but up to 63% of the sense overlapping ORFs, depending on the condition, maintain their start site above the annotated genes' background signals.

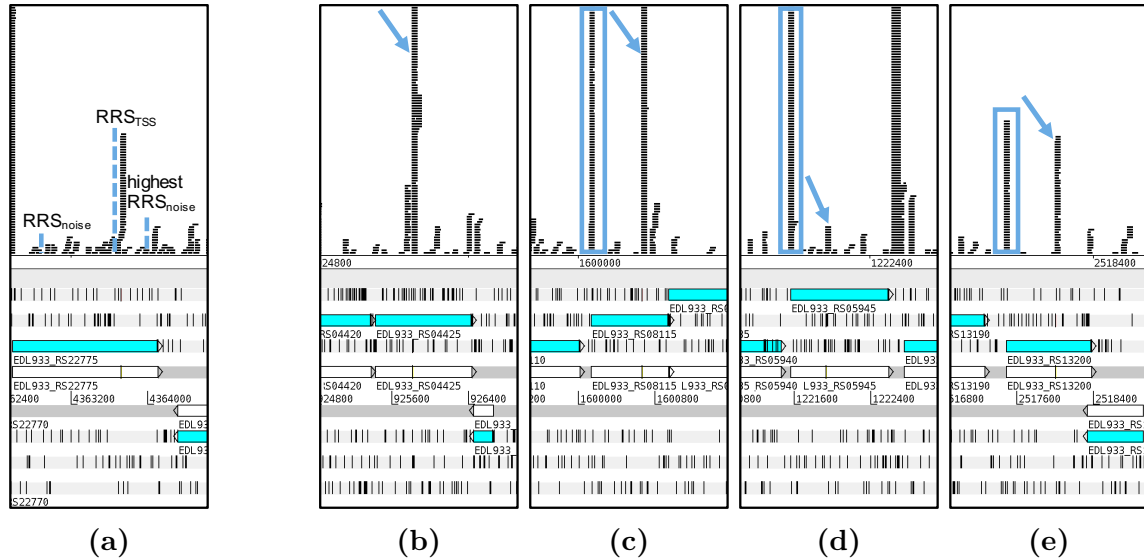


Figure 3.23: Examples of TSS within annotated genes. Mapped Cappable-seq reads are visualized. For clarity, reads of replicate 1 (condition: exponential LB) are shown, but tendencies are similar in other replicates and conditions. (a) Illustration of internal TSS selection method. The ratio of the RRS of a putative sense ORF TSS, RRS_{TSS} , and the highest RRS of positions within the annotated gene, for which association to any ORF can be excluded, RRS_{noise} , is calculated ($\frac{RRS_{TSS}}{RRS_{noise}}$). (b)-(e) Examples for automatically detected TSS with increased RRS of at least 1.5 background RRS (blue arrows) in three replicates and expected TSS signals upstream of annotated genes (blue boxes) are shown.

Table 3.17: Sense overlapping ORFs with TSS. Number of automatically detected sense overlapping ORFs with TSS within annotated genes for eight analyzed Cappable-seq conditions are given. **(a)** All sense ORFs with TSS and **(b)** sense ORFs with TSS with a RRS above annotated gene background are listed (> 1.5 -fold increase). Sense ORFs are restricted to 3' partial (5' end of ORF overlaps with 3' end of AG) and embedded ORFs.

(a) Sense overlapping ORFs with associated TSS

Condition	3' Partial		Embedded	
	Exponential	Stationary	Exponential	Stationary
LB	132	158	1621	1936
MM	148	165	1742	2034
acid	114	123	1573	1530
salt	123	171	1711	2348

(b) Sense overlapping ORFs with associated TSS, 1.5-fold increase of RRS above background

Condition	3' Partial		Embedded	
	Exponential	Stationary	Exponential	Stationary
LB	83	66	993	639
MM	77	63	825	714
acid	63	62	818	602
salt	68	73	73	837

3.6.2 Analysis of sense ORF associated TSS within annotated genes

To get a deeper insight into the start sites within annotated genes, transcriptional start sites associated with sense overlapping ORFs were visually examined for one condition (i. e., LB medium in exponential phase; Table 3.18)

The genomic region of the TSS was investigated to unveil any ambiguous TSS assignments due to

- a misannotated start codon for the annotated gene: The TSS belongs most likely to the annotated gene, which has no upstream TSS. The TSS was only erroneously associated to a sense ORF (Figure 3.23b).
- an in-frame start codon of the annotated gene: The TSS is the starting point for an alternative mRNA producing a shortened protein (i. e., protein isoform).
- a second downstream annotated gene: The TSS for a downstream genes is located within the first annotated gene and probably does not constitute a TSS for an assumed sense overlapping ORF within the first AG (Figure 3.23c).

For 33 and 3 TSS of embedded and partial sense ORFs, respectively, for which a connection to sense embedded ORFs was assumed, visual inspection showed that they very likely belong

Table 3.18: Categorization of putative sense overlapping TSS after visual analysis. TSS either associated only with OLG or with AG. Mostly, TSS association status is uncertain; TSS belongs either to OLG or to AG (downstream AG, alternative in-frame ATG).

TSS Category	3' Partial	Embedded
TSS for AG	3	33
TSS for OLG	6	38
TSS for OLG or AG (leaderless transcript)	1	34
TSS for OLG or downstream AG or alternative in-frame ATG	72	885
TSS as degradation product	1	3
total	83	993

to upstream annotated genes. Nearby downstream start codons suggest a misannotation of the start codon of the corresponding annotated genes.

The TSS of a further 35 sense overlapping ORFs could just as well belong to annotated genes producing an mRNA coding for a protein isoform of the annotated gene. The TSS is located directly within the internal start codon (ATG, GTG) of the annotated gene. Ettwiller *et al.* (2016) reported a codon position preference of the identified intragenic sense TSS at the first nucleotide of the in-frame start codon of the annotated gene, which may point to leaderless transcription of the ORFs (Brock *et al.*, 2008). Although the TSS is preferentially located at the second nucleotide in the presented data, leaderless transcription could explain the TSS position within the start codon.

For the vast majority of TSS within annotated genes (> 80%), no explicit statement about their association can be made. Thus, they belong either to sense overlapping ORFs or to an annotated downstream gene with 5' UTRs of more than 250 bp. Additionally, an association of the TSS to the annotated gene in which it was found might be possible due to in-frame start codons.

Nevertheless, clear signals for TSS of partial or embedded sense overlapping ORFs were found for 6 and 38 candidates, respectively (e. g., Figures 3.23d and 3.23e, Supplementary figure S12). For these ORFs, in-frame start codons of the annotated genes were not detected, therefore transcription of an alternative mRNA of the annotated gene can likely be excluded. Furthermore, the respective annotated genes have its own TSS, thus, misannotation of the start codon can be excluded. Finally, downstream annotated genes, if present at all, do have separate transcription start sites. Therefore, a relation of the start site to a downstream gene can also be excluded. Characteristics of the 44 putative sense ORFs are shown in Table 3.19. The overall length of the ORFs is rather short ranging from 96 bp to 453 bp with a mean length of 166 bp (Supplementary figure S3). The most frequent start codon of the ORF is TTG, followed by ATG and CTG. The median distance of start codon and transcription start site is 102 bp and thus, slightly increased compared to annotated genes (Figure 3.11). The signal of upstream sequences has a quite good conserved Pribnow-box, but comparably high noise in the remaining parts of the test sequence. However, this might be caused by the small set of only 44 sequences aligned (Figure 3.24). For 13 ORFs, a ribosome binding

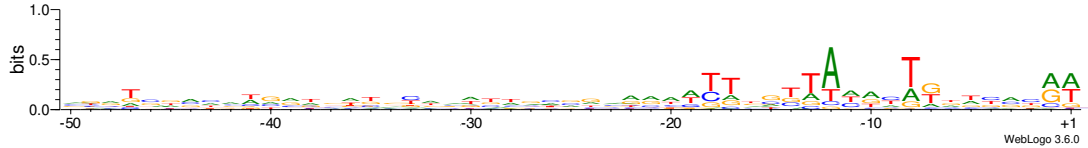


Figure 3.24: Conservation of promoter regions upstream of 44 sense overlapping ORF associated TSS. Sequence logos were created with WebLogo 3 (Crooks *et al.*, 2004). TSS, +1 position.

site upstream of the assumed start codon was detected according to Ma *et al.* (2002).

The number of TSS detected with an increased RRS_{TSS} to RRS_{noise} ratio of at least 1.5 is quite high (1076 ORFs) but is reduced by almost two thirds for a minimum ratio of 5 (376 ORFs). Although this more stringent selection results in a smaller set easier to analyze manually, only 8 out of 44 candidate ORFs remain with the increased S/N ratio.

To summarize, many TSS upstream of sense overlapping ORFs, which are putative sense overlapping genes, are found within annotated genes, but in most cases several possibilities for TSS association exist (either to common annotated genes or to sense overlapping ORFs). For less than 1% of all investigated sense embedded and 3' partial overlapping ORFs ($\frac{44}{16556+1045}$) an unambiguous TSS with a signal-to-noise ratio of $S/N > 1.5$ was identified. Nevertheless, for these 44 sense overlapping genes, the presence of a TSS is an indication of their transcription, but further studies are necessary to verify protein expression and functionality of these sense overlapping ORFs as protein-coding genes.

Table 3.19: Characteristics of sense overlapping ORFs. For each sense overlapping ORF with TSS, identifier (sID), genome position (strand, start, and stop), start codon, length of the ORF in nucleotides and overlap type are given as well as the genome position of the TSS and the distance between the TSS and start codon. The last column shows the signal-to-noise ratios (S/N, $\frac{RRS_{TSS}}{RRS_{noise}}$) for three biological replicates.

sID	Strand	Start	Stop	Start Codon	Length [nt]	Overlap	TSS	Distance	S/N
sID_72617	-	154635	154763	CTG	129	embedded	154830	67	26.9 / 60.1 / 23.7
sID_4053	+	655670	655882	CTG	213	embedded	655612	58	2.4 / 2.8 / 1.7
sID_5860	+	925839	925976	ATG	138	embedded	925812	27	6.5 / 11.8 / 12.7
sID_5929	+	936579	936794	TTG	216	embedded	936579	0	2.5 / 3.4 / 3.5
sID_65425	-	1202741	1202836	TTG	96	embedded	1203049	213	4.3 / 1.6 / 2.0
sID_65424	-	1202884	1202991	ATT	108	embedded	1203049	58	4.3 / 1.6 / 2.0
sID_7723	+	1222058	1222153	ATG	96	embedded	1221932	126	1.7 / 1.8 / 1.6
sID_63872	-	1416720	1416893	ATG	174	embedded	1416976	83	12.0 / 3.6 / 14.0
sID_63851	-	1421103	1421237	ATG	135	embedded	1421303	66	5.3 / 3.6 / 3.8
sID_60970	-	1824800	1824931	CTG	132	embedded	1825087	156	3.4 / 3.6 / 6.0
sID_60810	-	1849037	1849210	ATC	174	partial	1849385	175	13.5 / 6.7 / 4.3
sID_60435	-	1896092	1896241	GTG	150	embedded	1896427	186	2.0 / 2.5 / 2.0
sID_13183	+	2087954	2088067	ATA	114	embedded	2087770	184	2.9 / 1.9 / 1.9
sID_58937	-	2110811	2110960	ATA	150	embedded	2111077	117	13.7 / 7.5 / 12.1
sID_58689	-	2153696	2153887	ATG	192	embedded	2153913	26	2.6 / 2.0 / 2.6
sID_57486	-	2333149	2333346	ATG	198	embedded	2333372	26	1.5 / 2.6 / 2.9
sID_16193	+	2518136	2518234	GTG	99	embedded	2518001	135	4.7 / 5.0 / 3.1

Table 3.19: Continued from previous page

sID	Strand	Start	Stop	Start Codon	Length [nt]	Overlap	TSS	Distance	S/N
sID_16557	+	2572159	2572470	ATA	312	embedded	2572060	99	7.0 / 3.0 / 3.4
sID_54453	-	2790437	2790628	ATG	192	embedded	2790654	26	2.7 / 1.9 / 2.2
sID_18263	+	2809638	2809763	TTG	126	embedded	2809403	235	9.0 / 4.1 / 8.2
sID_19269	+	2964892	2964996	CTG	105	embedded	2964804	88	2.2 / 2.1 / 1.5
sID_52549	-	3079318	3079458	ATT	141	partial	3079477	19	2.2 / 3.2 / 3.3
sID_51233	-	3289302	3289484	ATT	183	embedded	3289663	179	3.0 / 3.7 / 2.9
sID_51232	-	3289465	3289602	ATT	138	embedded	3289663	61	3.0 / 3.7 / 2.9
sID_51231	-	3289527	3289640	GTG	114	embedded	3289663	23	3.0 / 3.7 / 2.9
sID_22528	+	3428605	3428751	TTG	147	embedded	3428417	188	1.9 / 4.6 / 2.9
sID_23476	+	3559655	3559879	ATT	225	embedded	3559509	146	4.0 / 5.7 / 3.6
sID_24441	+	3701092	3701349	TTG	258	embedded	3701074	18	5.8 / 8.0 / 4.7
sID_24442	+	3701273	3701569	CTG	297	embedded	3701074	199	5.8 / 8.0 / 4.7
sID_47204	-	3930483	3930647	ATG	165	embedded	3930658	11	17.5 / 32.2 / 27.1
sID_27733	+	4191596	4191805	ATA	210	embedded	4191454	142	2.7 / 1.8 / 3.3
sID_44986	-	4267427	4267639	ATA	213	embedded	4267858	219	3.7 / 2.1 / 1.6
sID_44985	-	4267655	4267753	TTG	99	embedded	4267858	105	3.7 / 2.1 / 1.6
sID_43122	-	4556226	4556339	TTG	114	embedded	4556343	4	3.2 / 2.5 / 1.8
sID_43081	-	4563596	4563700	TTG	105	embedded	4563758	58	1.8 / 2.4 / 2.9
sID_30910	+	4671974	4672078	CTG	105	embedded	4671922	52	4.3 / 3.8 / 2.6
sID_31442	+	4749831	4750022	ATA	192	partial	4749619	212	342.8 / 26.5 / 40.6

Table 3.19: Continued from previous page

sID	Strand	Start	Stop	Start Codon	Length [nt]	Overlap	TSS	Distance	S/N
sID_31442	+	4749831	4750022	ATA	192	partial	4749647	184	10.6 / 6.1 / 14.0
sID_41605	-	4808895	4809347	ATC	453	partial	4809585	238	2.1 / 5.1 / 3.5
sID_41173	-	4866091	4866222	TTG	132	embedded	4866259	37	6.0 / 6.4 / 8.3
sID_35348	+	5342691	5342816	TTG	126	embedded	5342553	138	2.0 / 1.7 / 2.5
sID_35348	+	5342691	5342816	TTG	126	embedded	5342570	121	6.1 / 10.4 / 5.5
sID_37787	-	5369599	5369793	CTG	195	embedded	5370001	208	2.0 / 1.6 / 2.3
sID_36884	-	5512087	5512212	CTG	126	partial	5512303	91	1.7 / 6.1 / 2.4

3.7 Functional characterization of the overlapping gene *pop* encoded antisense to *ompA*

When conducting high-throughput phenotyping, the overlapping gene candidate OGC 57, later designated *pop* (**p**H-regulated **o**verlapping **p**rotein-coding gene), had highly correlated *z*-scores in biological replicates as well as a positive overexpression effect in LB medium supplemented with L-malic acid (Tables 3.4 and 3.5, Supplementary table S9). Further, *pop* was unusually long for an overlapping gene and, thus, attracted attention to characterize this candidate in more detail.

3.7.1 Genomic localization of *pop*

The overlapping gene *pop* probably starts at genome position 1 236 020 and ends at position 1 236 622 (coordinates following the genome annotation of Latif *et al.* (2014), GenBank accession CP008957). It has a length of 603 bp and is located in reading frame -1 with respect to the highly conserved outer membrane protein gene *ompA* (1065 bp, Figure 3.25 and Supplementary figure S4). The coding sequence of *pop* is completely embedded in antisense to the coding sequence of *ompA*.

Ribosome profiling of EHEC EDL933 produced evidence for translation of this OLG reproducible across biological replicates in LB medium (Figure 3.26a, Supplementary table S13, data provided by Landstorfer (2014), evaluated and prepared for visualization by Dr. Zachary Ardern). The expression of *ompA*, compared to *pop*, is higher by a factor of 150. This observation is not surprising, as OmpA is one of the most highly expressed proteins in *E. coli* (Ortiz-Suarez *et al.*, 2016). Upstream of *pop* is another annotated gene, *ycbG* (453 bp), which encodes a macrodomain ter protein. RPKM values are on average three times higher for *ycbG* than for *pop* (Figure 3.26b, Supplementary table S13). However, the average translational efficiency expressed by the ribosome coverage value ($RCV = \frac{RPKM_{\text{translatome}}}{RPKM_{\text{transcriptome}}}$, Hücker, Ardern, *et al.*, 2017) is better for the overlapping gene *pop* compared to *ycbG* (Supplementary table S13) and the RCVs for *pop* (> 1 in all instances) exceed translation efficiency of *ycbG* throughout all biological experiments. Furthermore, the threshold of translated transcripts specified by Neuhaus *et al.* (2017) is $RCV = 0.355$ which is clearly lower

than the RCVs obtained for *pop* (Figure 3.26c). Both findings strengthen the hypothesis of a meaningful translation signal of this coding region and provide evidence for translation of *pop*. In addition, both, a transcription start site and a σ^{70} promoter are located in the region between *ycbG* and *pop* (Figure 3.25, panel **B**, details Section 3.7.3), indicating *pop*'s independent translation. Downstream of *pop* two open reading frames are found in frames -1 and -2 (regarding to *ompA*). Each of the ORFs is a little over 200 bp in length and still largely overlaps with *ompA* (Figure 3.25). Although a rho-independent terminator is located downstream of these (Figure 3.25, panel **D**, details in Section 3.7.3), none of the ORFs appears to be either transcribed or translated (Supplementary table S13). Furthermore, a possible, but rare start codon CTG at genome positions 1236020 to 1236022 (Supplementary figure S4) was detected at the 5' end of the *pop*-ORF as well as a Shine-Dalgarno sequence upstream thereof ($\Delta G^\circ = -3.6$ kcal/mol, Figure 3.25, panel **C**).

In summary, *pop* was identified as a translated open reading frame based on ribosome profiling experiments. The translation signals are reproducible across biological replicates, thus there is strong evidence for specific expression of *pop* rather than the signal being merely due to stochastic background expression.

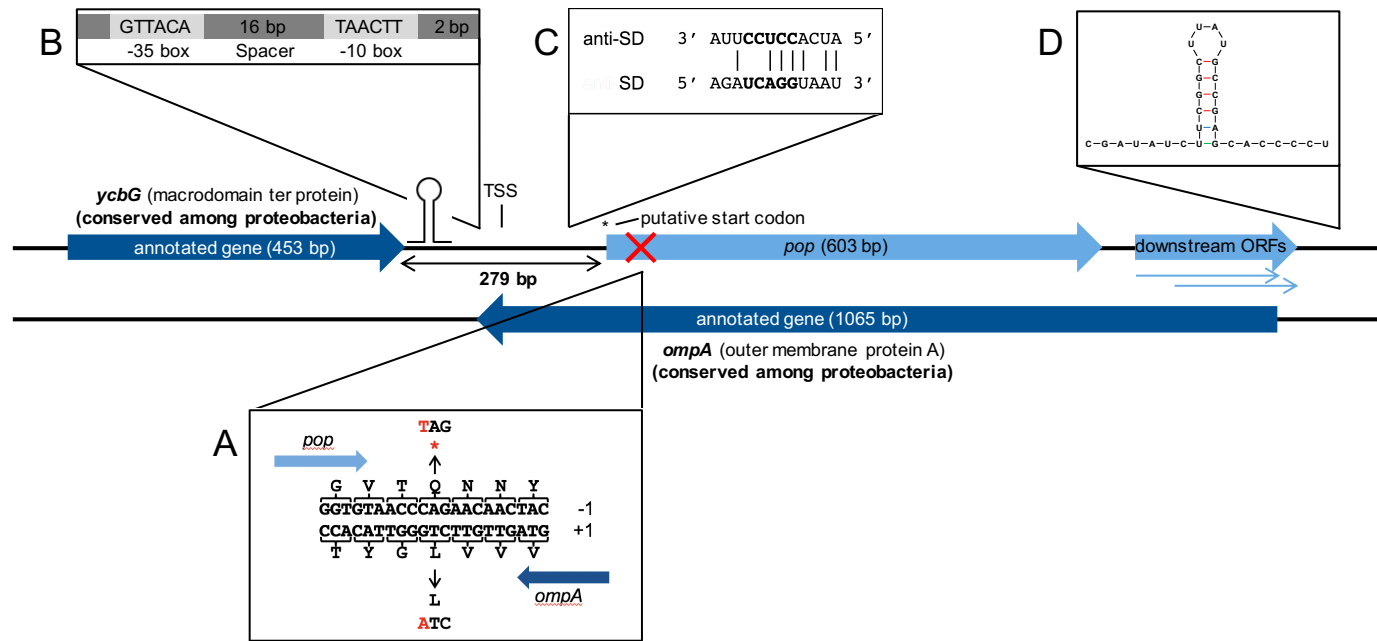
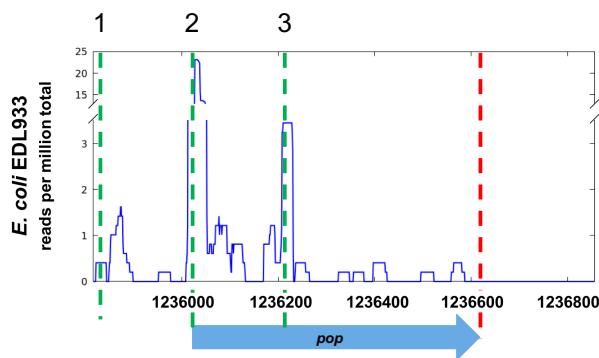
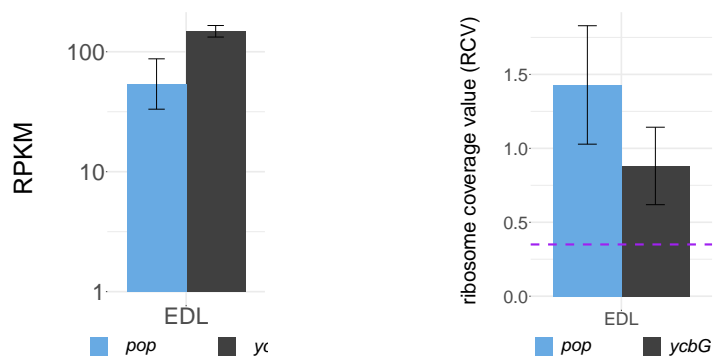


Figure 3.25: Genomic organization of *pop*. *pop* (603 bp) is located downstream of the annotated gene *ycbG* and completely embedded antisense in the sequence of *ompA*. Downstream of *pop* are two smaller overlapping open reading frames localized, which overlap *ompA* in parts and are referred to as downstream ORFs. The transcription start site was identified with Cappable-seq in the intergenic region of *pop* and *ycbG*. The full genomic sequence of *pop* is shown in Supplementary figure S4. **(A)** Translationally arrested mutant of *pop*. *pop* is located in reading frame -1 with respect to *ompA*. A single base substitution C → T at genome position 1236083 (red cross) was introduced in mutants for phenotypic characterization as indicated, resulting in a stop codon (*) in *pop* and a synonymous amino acid change in *ompA*. **(B)** Promoter sequence. Sequences of -10 box and -35 box predicted by BRPOM and bTSSfinder as well as the length of the spacer between the conserved boxes and the distance to the TSS are shown. **(C)** Alignment of SD sequence of *pop* and anti-SD sequence of the 16S rRNA in the 30S ribosomal subunit. The SD sequence ($\Delta G^\circ = -3.6$ kcal/mol) was predicted 10 bp upstream of the putative start codon according to Ma *et al.* (2002). The core of the ribosome binding site is displayed in bold letters. **(D)** Secondary stem loop structure of the first 40 bp of the predicted terminator with a final energy of $\Delta G = -8.6$ kcal/mol.



(a) distribution of ribosome profiling reads across the sequence of *pop*



(b) RPKM in ribosome profiling

(c) translatability (RCV)

Figure 3.26: *pop* translation in ribosome profiling. (a) Alignment of sequence and ribosome profiling reads of *pop* in *E. coli* strain O157:H7 EDL933. Graph shows normalized sequencing reads (RPKM) of ribosome profiling experiments in LB medium; the sum signal of two biological replicates is visualized. Three putative start codons are indicated with green dashed lines in region 1 (TTG), 2 (CTG) and 3 (GTG). The stop codon is indicated by a red dashed line. (b) Averaged RPKM values of translation for overlapping gene *pop* and the upstream annotated gene *ycbG*. (c) Averaged ribosomal coverage values (RCV) of *pop* and *ycbG*. Purple dashed line shows the threshold for translated ORFs ($RCV = 0.35$). Error bars display upper and lower RPKM value or RCV used for calculation in each case.

3.7.2 Effect of overexpression and a knock-out mutant of *pop* in competitive growth

Competitive growth experiments, similar to those in Section 3.4, were performed to analyze the influence of *pop* overexpression on the growth of EHEC. The intact and a translationally arrested mutant ORF were cloned under the control of an arabinose-inducible promoter in the overexpression plasmid pBAD/myc-HisC (pBAD+*pop* and pBAD+ Δ *pop*). The mutated sequence differs in one base pair from the wild type gene. This single base substitution leads to a stop codon in *pop* (Figure 3.25, panel **A**), whereas the mother frame of *ompA* remains unchanged. Thus, any growth variations after extensive expression of both *pop*-ORFs can be explained by the presence or absence of a protein encoded by this OLG. The competition experiment was conducted in previously identified acid stress conditions as well as in pH-adjusted LB media (Figure 3.27a). Altered growth of cells overexpressing mutant or wild type sequences was not detected in plain LB medium, whereas LB-based media supplemented with different stressors had a significant influence on the relative growth of mutant and wild type transformants. Addition of the organic acids L-malic acid and malonic acid for example led to enhanced growth of cells containing the plasmid expressing the intact sequence compared to cells expressing the mutated sequence. Thus, the presence of *pop* is beneficial in these conditions. The addition of those acidic substances resulted in a pH shift from 7.4 to 5.8 at the beginning of cultivation. Reversed proportions were detected for LB buffered to alkaline conditions (pH 8.7) with bicine. In contrast, LB adjusted with the biologic buffers MES and MOPS to acidic (pH 5.8) or near neutral (pH 7.4) environments, respectively, did not result in significant growth differences. Nevertheless, wild type cells grew slightly better in MES-buffering, which is in accordance with observations of increased wild type cell proportions in other acidic conditions.

Quantitative PCR was conducted to determine native mRNA levels of *pop* compared to the mRNA of the 16S rRNA gene. In line with the growth advantage of the wild type during overexpression in malic acid (Figure 3.27a), a tendency for upregulated mRNA levels of *pop* was detected in the presence of L-malic acid (fold change 2.4, one-tailed Welch two sample t-test, p-value = 0.17). In contrast, significant downregulation of its mRNA in bicine buffered LB medium was observed (fold change 0.35, one-tailed Welch two sample t-test,

p-value = 0.03). The overall fold change expression between malic acid and bicine based on these qPCR is 6.9, therefore a pH-dependent differential regulation of *pop* is presumed.

In a next step, the genome of *E. coli* O157:H7 EDL933 was modified to create a knock-out for *pop*. As for the overexpression variant, the single base mutation causes a stop codon in *pop* while the amino acid sequence of *ompA* remains unchanged (Figure 3.25, panel **A**). Several pH-relevant stress conditions were tested in competitive growth of the EHEC mutant Δpop and EHEC wild type, but no significant growth difference in any condition was detected (Figure 3.27b).

In summary, opposite overexpression phenotypes in alkaline buffered and acidified media indicate a protein-coding function for *pop* as differences of mRNAs produced by the intact and translationally arrested sequence variants are very small. Therefore, a pH-dependent function of *pop* is proposed which is supported by differential regulation of *pop* in high and low pH media.

3.7.3 Transcriptional unit of *pop*

A reproducible transcriptional start site of *pop* was determined with Cappable-seq (detection criteria, $minRRS = 0.5$, 250 bp upstream range) at genome position 1 235 862 in the intergenic region between *ycbG* and *pop* (Figure 3.25, Supplementary figure S4). As for TSS in Section 3.5.8, the bioinformatic tools BPRM and bTSSfinder were used to identify a promoter. In contrast to these analyses, the input sequences for the programs started 65 bp and 197 bp upstream of the TSS, respectively. The σ^{70} promoter, shown in Figure 3.25, panel **B**, and Supplementary figure S4, was predicted with both programs (BPRM, LDF score 0.59; bTSSfinder, score 1.86). Despite a suboptimal distance of only 2 bp between the -10 box of the promoter and the TSS, promoter activity could be verified in the GFP-assay (Figure 3.28a). Significantly enhanced fluorescence for cells containing pProbe-NT+promoter-*pop* compared to cells with the promoter-less control plasmid pProbe-NT was measured in LB and bicin-buffered medium. Although the fluorescence signal produced by the promoter was strikingly increased in basic milieu (pH 8.7), it might result from GFP accumulation in the cells during longer incubation times necessary in this growth condition (Miller *et al.*, 2000).

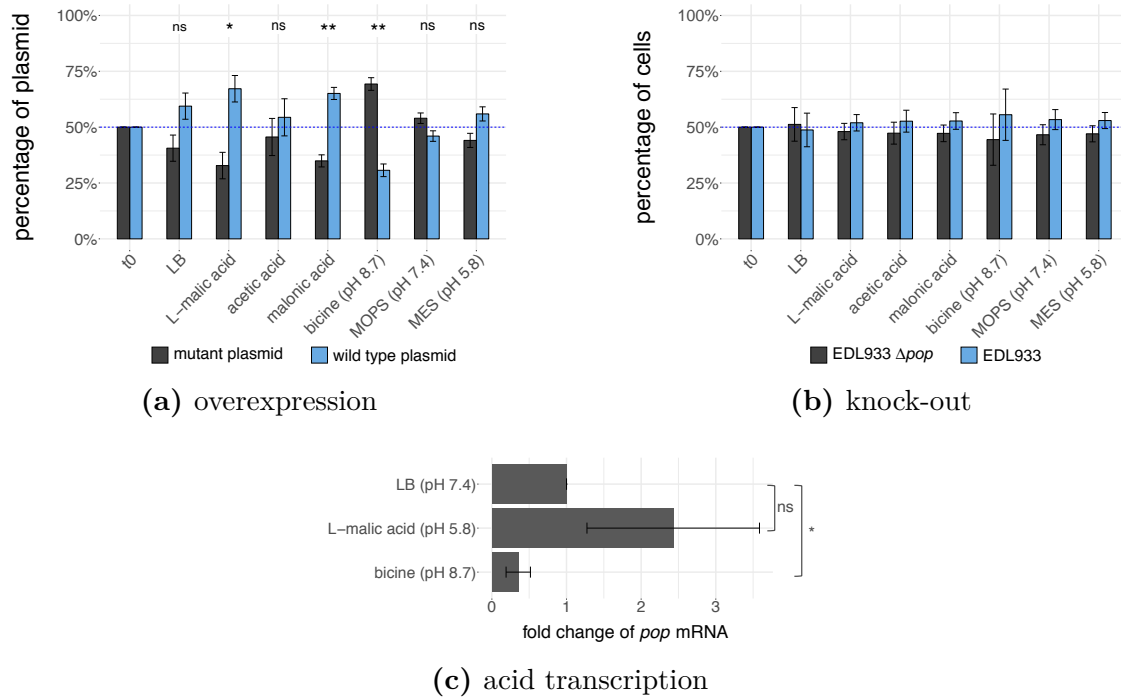


Figure 3.27: Effect of *pop* expression and knock-out in medium of various pH ranges. **(a)** Overexpression and **(b)** knock-out competitive growth of *pop*. Competitive growth of EHEC while overexpressing either intact (pBAD+*pop*) or translationally arrested *pop* (pBAD+ Δpop) or EHEC wild type and EHEC Δpop was conducted in conditions as indicated, i.e., LB medium supplemented with an organic acid or a biological buffer. Mean percentages of peak heights (fluorescence intensities) for wild type (blue bars) or mutated position (grey bars) measured in sequencing electropherograms are shown before (t_0) and after 22 h growth for overexpression and 18 h growth for knock-out experiments. Values are normalized to 50% input ratio (blue dashed line). Error bars indicate standard deviations of three biological replicates. Statistical significance between wild type plasmids before and after growth was tested with a paired t-test. **(c)** Relative quantification of *pop* mRNA by qPCR. Fold change of *pop* mRNA with respect to 16S rRNA of EHEC grown to early exponential phase ($OD_{600} = 0.3$) in neutral (LB), acidic (LB + L-malic acid) and alkaline (bicine buffered LB) growth medium. Mean values and standard deviations of three biological replicates are presented. Significance was tested with a one-tailed Welch two sample t-test at significance level $\alpha = 0.05$. Statistical significance is indicated in each of the panels as follows: * $p \leq 0.05$; ** $p \leq 0.01$; ns, not significant.

As the promoter activity for *pop* is considerably lower compared to other annotated and overlapping gene promoters, co-transcription along with the upstream promoter for *ycbG* was examined with RT-PCR (Figure 3.28b). No mRNA molecule spanning both genes was detected, thus, monocistronic transcription of *pop* from the tested promoter is proposed.

FindTerm predicted a rho-independent terminator with a length of 120 bp, 295 bp downstream of the stop codon of *pop*. Secondary structures of 30 bp segments of the supposed terminator region were constructed using the tool Quickfold of Mfold and a stable stem loop structure was detected ($\Delta G = -8.6$ kcal/mol, bases 35 to 78 of the terminator, Figure 3.25, panel **D**). RT-PCRs were used to validate the 3' end of the mRNA (Figure 3.28b). *pop* and the downstream ORFs seem to be co-transcribed and transcription is terminated downstream of the stem loop structure.

Based on these results, a transcriptional unit with a length of 1120 bp can be proposed. It covers almost the entire open reading frame of *ompA* in antisense. However, the gene *ycbG* upstream of *pop* is not part of the transcript, while two downstream ORFs located upstream of the stem loop structure of a rho-independent terminator are included in the mRNA.

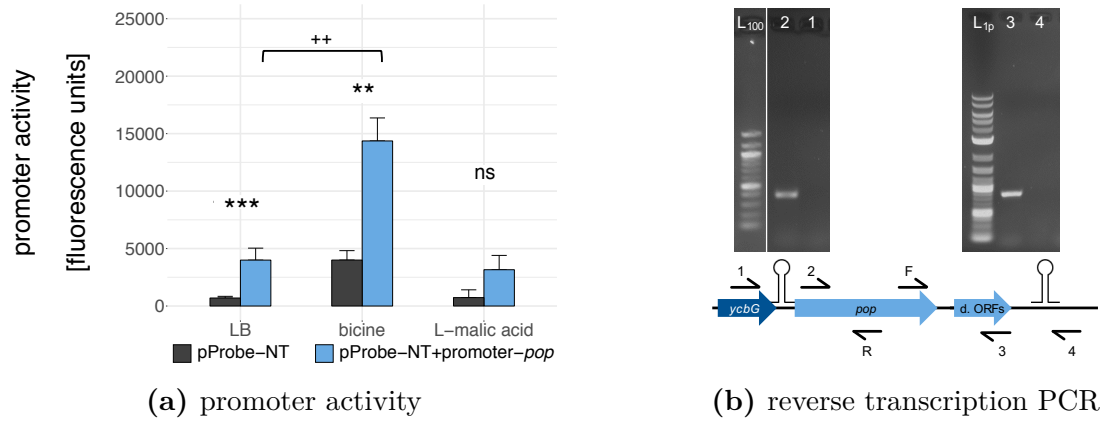


Figure 3.28: Analysis of the transcriptional unit of *pop*. **(a)** Promoter activity assay for the promoter of *pop*. Mean fluorescence units of *E. coli* Top10 cells with either promoterless GFP-plasmid or plasmid under the control of *pop*-promoter in culture conditions as indicated are given. Error bars show standard deviations. Significant differences between fluorescence of empty vector (grey bars) and the promoter construct (blue bars) and between growth conditions was tested with a Welch two sample t-test (**/++ $p \leq 0.01$; *** $p \leq 0.001$; ns, not significant). **(b)** Test for mono- or polycistronic mRNA and verification of the stem loop secondary structure in RT-PCR. Agarose gels of RT-PCRs are shown. Two different forward primers, binding within *ycbG* (1: RTPCR-*ycbG*-F) or within *pop* (2: RTPCR-*pop*-F), were combined with a *pop* reverse primer (R: RTPCR-*pop*-R) to verify independent transcription of the two ORFs. cDNA synthesis was performed with primer R. Two different reverse primers, binding upstream (3: RTPCRterm-dORF-R) or downstream (4: RTPCRterm-stemloop-R) of the stem loop structure, were combined with a *pop* forward primer (F: RTPCRterm-*pop*-F) to prove 3' end of mRNA. cDNA synthesis was performed with primer 3 and 4, respectively. 1, PCR with primers 1+R; 2, PCR with primers 2+R. 3, PCR with primers F+3; 4, PCR with primers F+4; L₁₀₀: 100 bp DNA Ladder (NEB); L_{1p}: 1 kbPlus DNA Ladder (NEB); d. ORFs, downstream ORFs.

3.7.4 pH-dependent detection of Pop in Western blots

According to Section 3.2, *pop* was cloned in-frame into pBAD/SPA and overexpressed to detect Pop in Western blots (Figure 3.29). The experiment was conducted in LB (pH 7.4) and bicin-buffered LB (pH 8.7). The full-length protein (theoretically 30 kDa, detected approx. 34 kDa) as well as shorter products (approx. 20 kDa and 24 kDa) were detected in immunoblots. After protein induction in LB, the amount of the full-length Pop-protein increases for the first 1.5 h, but decreases afterwards (Figure 3.29a). This suggests an unstable protein. When buffering the medium with bicine, protein expression is somewhat higher shortly after induction and the protein signal of Pop does not decrease over time (Figure 3.29b). A more detailed analysis of the smaller products showed that band intensities increase constantly over time in LB medium, whereas the signal is constant in bicine.

Differences in *pop* expression at varying pH values of the growth medium were detected. However, as expression of the protein from the plasmid is highly artificial, it probably does not reflect natural occurrence of the protein. Nevertheless, Pop itself and a different stability according to conditions was detected and this supports the protein-coding potential of *pop*.

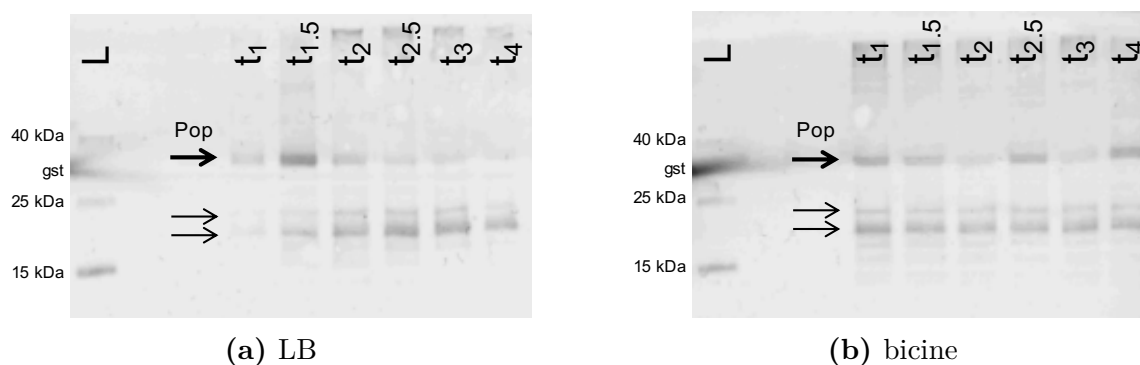


Figure 3.29: Western blots of Pop. Expression of *pop* in frame with a C-terminal SPA-tag was performed in (a) LB at pH 7.4 and (b) bicine-buffered LB at pH 8.7. Cells were harvested after induction with arabinose (1 to 4 h, t_x) and normalized cell numbers were separated on 16% Tris-tricine gels. The arrows point at the band of the putative full-length protein Pop (bold arrows) and the two shorter proteins (thin arrows). L, Spectra Multicolor Low Range Protein Ladder (Thermo Scientific) with additional internal Western blot control glutathione-S-transferase (31 kDa, remaining band sizes as indicated); t_x , whole cell extract from samples harvested after $x = 1, 1.5, 2, 2.5, 3,$ and 4 h after induction.

3.7.5 Bioinformatic analyses of the functionality of *pop*

A protein database search was performed to find hints of a specific function of Pop. No significant Pop homologs were found in PDB (Protein Data Bank), UniProtKB/Swiss-Prot and the RefSeq protein database using the blastp algorithm, whereas significant similarities with annotated proteins were detected in NCBI's non-redundant protein sequence (nr) database. The top hit is an uncharacterized protein in *Shigella sonnei* with 67% coverage and 99% identity at an e-value of 4×10^{-91} . The genomic sequence of the target organism revealed ambiguous bases at the 5' end of the *ompA* homolog, which resulted in a missing start codon. Thus, *ompA* was not annotated because an obvious gene structure was absent. Consequently, the *ab initio* prediction of *pop* was possible. This result confirms the operating principles of annotation algorithms described in Section 1.3.2, which systematically prevent annotation of long (non-trivially) overlapping genes. To examine whether the gene structure of *pop* is accepted by gene finding algorithm in case of an absent, i. e., masked, *ompA*, gene prediction was applied to the genomes of *E. coli* O157:H7 EDL933, *Shigella dysenteriae*, *Klebsiella pneumoniae* and *Enterobacter cloacae*, four representatives in the family Enterobacteriaceae. N bases were introduced in each genome at potential start codons of *ompA* to mask the start codon and thus, 'hide' the annotated gene for the algorithm. In all four cases, *ompA* was not predicted, in contrast to *pop* which was predicted each time (Table 3.20). Protein-coding genes are rated with a prediction score which ranged from -0.5 to > 1000 in EHEC. Although the total score of *pop* (14.37) falls within the lowest 10% of all 5351 predicted EHEC coding sequences, conserved annotated genes, e. g., a fimbrial chaperon or the entericidin A protein, were scored worse. This analysis showed that *pop* clearly exhibits distinct gene structure elements, which enable its identification as protein-coding gene in case of 'absent' *ompA*.

Although no hints about the function of the protein Pop were found, *pop* seems to be a protein-coding gene which is normally overlooked as it is encoded in the shadow of its conserved overlapping gene partner *ompA*.

Table 3.20: *pop* prediction with Prodigal. For each species (EHEC EDL933, *Shigella dysenteriae*, *Klebsiella pneumoniae*, *Enterobacter cloacae*) chromosome coordinates (start, stop, strand) of predicted open reading frames are given. The best hit is shaded grey, respectively. Total score (total s.), coding potential (coding pot.), start score (start s.), start codon, RBS (ribosome binding site) motif, spacer, RBS score (RBS s.), upstream score (upstream s.), type score (type s.), and GC content of the ORF (GC) are listed.

Strain	Start	Stop	Strand	Total S.	Coding Pot.	Start S.	Start Codon	RBS Motif	Spacer	RBS S.	Upstream S.	Type S.	GC
EHEC EDL933	1235822	1236622	+	14.37	26.57	-12.20	TTG	GGA/GAG/AGG	5-10bp	2.57	-1.97	-12.79	0.552
	1236020	1236622	+	-8.00	12.90	-20.90	TTG	3Base/5BMM	13-15bp	-6.49	-1.62	-12.79	0.557
	1236215	1236622	+	-31.46	-17.59	-13.87	GTG	GGA/GAG/AGG	3-4bp	-8.08	1.11	-6.40	0.566
	1236341	1236622	+	-27.86	-20.06	-7.80	GTG	GGA/GAG/AGG	5-10bp	2.57	-3.46	-6.40	0.535
	1236449	1236622	+	-44.94	-30.97	-13.97	GTG	4Base/6BMM	13-15bp	-2.45	-1.66	-9.36	0.511
Shigella	996209	996364	-	-45.53	-32.96	-12.57	GTG	None	None	-8.18	0.49	-4.39	0.558
	996209	996397	-	-38.78	-32.36	-6.42	GTG	4Base/6BMM	13-15bp	-1.67	-0.64	-3.61	0.540
	996209	996505	-	-27.65	-23.66	-3.99	GTG	GGA/GAG/AGG	5-10bp	2.31	-3.11	-2.69	0.542
	996209	996631	-	-27.26	-18.30	-8.96	GTG	None	None	-5.00	-0.77	-2.69	0.567
	996209	996826	-	-3.54	15.83	-19.37	TTG	None	None	-5.00	-3.18	-11.18	0.560
	996209	997024	-	23.26	33.37	-10.11	TTG	GGA/GAG/AGG	5-10bp	2.31	-1.24	-11.18	0.556
Klebsiella	1612552	1613646	+	31.93	50.39	-18.47	TTG	None	None	-7.78	0.24	-10.93	0.567
	1612744	1613646	+	4.56	25.19	-20.62	TTG	3Base/5BMM	13-15bp	-7.64	-2.05	-10.93	0.571
	1612939	1613646	+	-12.53	-9.93	-2.60	GTG	GGA/GAG/AGG	5-10bp	2.36	2.08	-6.54	0.582
	1613173	1613646	+	-26.61	-17.70	-8.92	GTG	GGA/GAG/AGG	11-12bp	-1.35	-0.53	-6.54	0.561
	1613491	1613646	+	-76.84	-53.31	-23.53	GTG	None	None	-12.71	0.36	-10.69	0.538
Enterobacter	2728537	2728710	-	-60.66	-47.88	-12.78	GTG	GGA/GAG/AGG	11-12bp	-2.27	-1.15	-8.86	0.580
	2728537	2728947	-	-16.60	-6.23	-10.37	GTG	GGA/GAG/AGG	3-4bp	-5.19	1.38	-6.06	0.582
	2728537	2729142	-	12.89	29.43	-16.53	TTG	3Base/5BMM	13-15bp	-5.87	-0.68	-9.99	0.559
	2728537	2729334	-	43.76	61.06	-17.30	TTG	None	None	-7.43	0.12	-9.99	0.545

4 Discussion

4.1 Detection of gene products of overlapping gene candidates

4.1.1 Assessment of two experimental methods to detect small and low abundance proteins

Detection and characterization of proteins is an essential step in molecular and cell biology for detailed analysis of molecular processes in an organism. While highly abundant proteins are detected well in gel-based (e. g., SDS-PAGE, Western blot) and gel-free (e. g., mass spectrometry, MS) methods, scientists struggle to detect small and low abundance proteins (Garbis *et al.*, 2005). For analysis of overlapping genes and the corresponding proteins, these limitations have to be overcome, as it was reported that functional protein-coding overlapping genes are weakly expressed (e. g. Fellner *et al.*, 2015), which might be traced back to their status as evolutionary young genes (Carvunis *et al.*, 2012; Donoghue *et al.*, 2011) and smaller size compared to annotated genes (Supplementary figure S2). In the following, some technical aspects of SDS-PAGE and Western blot as well as mass spectrometry are reflected and the methods are assessed for suitability of the analysis of overlapping proteins.

MS is a technology used to analyze the proteomes and their post-translational modifications in various organisms (e. g., Impens *et al.*, 2017; Hebert *et al.*, 2014; Merrihew *et al.*, 2008; Bouwmeester *et al.*, 2004). The success in protein detection, however, relies on the accessibility of the proteins to the method (i. e., trypsin cut sites), as well as their abundance in the investigated culture condition. Especially the latter aspect restricts MS to the most frequent proteins found in a specific sample and makes it inappropriate to search for low abundance proteins, when using standard methods. In contrast, targeted MS searches for specific proteins/biomarkers and does not analyze the proteome globally (Shi *et al.*, 2016). Even though sensitivity of this method especially in complex samples is still a concern, approaches like antibody-based enrichment (Larsson *et al.*, 2000) are successfully applied to detect low abundance proteins. However, the limited throughput of targeted MS is a major drawback (Arsène-Ploetze *et al.*, 2015; Shi *et al.*, 2016). Furthermore, detection of small proteins is challenging, as the number of proteolytic peptides is lower for such molecules, thus identification with confidence is more difficult (Neuhaus *et al.*, 2016).

Standard SDS polyacrylamide gel electrophoresis (SDS-PAGE, Laemmli, 1970) was improved by Schägger and Von Jagow (1987) to increase the capability of the method to separate proteins with molecular masses < 30 kDa efficiently using a Tris-tricine buffer system. For instance, Hemm *et al.* (2008) were able to detect almost 40 small proteins in the mass range of 1.7 kDa to 5.5 kDa with this method when combined with Western blot. Insertion of the SPA-tag (7.7 kDa) to the C-terminal end of small target ORFs on the chromosome allowed labelling of natively expressed proteins, which could be detected via a tag-specific antibody. Like other commonly used epitope tags (e. g., maltose binding protein, 42 kDa, or glutathione-S-transferase, 26 kDa) SPA increases the mass of tagged proteins artificially and thus, improves protein detection due to stabilizing effects, as it was shown for the product of *rpmH*, where protein detection was highly reproducible in this study. However, the metabolic burden for cells expressing proteins attached to a larger tag is considerably higher (Waugh, 2005). A notably smaller protein tag commonly used for protein purification is 6xHis. It was shown for this modification that the protein structure is not influenced significantly (Carson *et al.*, 2007). Therefore, the nature of the target protein is maintained and, unlike larger tags, the likelihood for false positive detection of small proteins due to stabilizing effects is reduced. Hence, using SPA as protein tag is a compromise between a sufficient size to stabilize otherwise somewhat unstable proteins and a limited size to maintain the nature of the tagged sequences, thus, reducing artificial stabilizing effects.

Nevertheless, independent of the tag chosen for protein visualization, chromosomal tagging relies on proteins expressed from their natural genomic environment. As mentioned in Section 1.1.3, overlapping genes are often expressed weakly and even using a highly sensitive detection with tag specific antibodies in Western blot might not yield signals. Further, for some overlapping genes of which the functionality of the annotated gene is essential in EHEC, insertion of an epitope tag destroys the gene on the opposite strand causing cell death. A further drawback of this method is the limited scalability for high-throughput analysis, as every single gene has to be tagged separately. Additionally, creation of genomic mutants by homologous recombination is complex and laborious (Datsenko and Wanner, 2000; Miller and Mekalanos, 1988), despite improved techniques which increase the success rate (Sarker and Cornelis, 1997; Kim *et al.*, 2014; Herring *et al.*, 2003; Yu *et al.*, 2000).

4.1.2 Overlapping gene candidates produce stable proteins

To overcome several of the limitations described above, protein detection was carried out with high-percentage tricine SDS polyacrylamide gels. An increased throughput in SDS-PAGE and Western blot was achieved by tagging the sequences in frame with SPA (Baek *et al.*, 2017) using a modified overexpression plasmid pBAD-SPA created for this study. The cloned plasmids allowed for analyzing more than 200 overlapping gene candidates for their protein-coding potential. Nonetheless, it has to be kept in mind that expression of *E. coli* O157:H7 EDL933 overlapping gene candidates from an arabinose inducible plasmid in the cloning strain *E. coli* Top10 is of artificial nature.

Protein detection was successful for more than 95 % of candidates analyzed. On average, the detected protein weights were higher than expected, but reproducible (Figure 3.3 on page 77). Mass differences have been observed before for protein separation with SDS-PAGE (Dolnik and Gurske, 2011). The accuracy of the method for molecular weight calculation applied is $\pm 10\%$ (corresponds to approx. 1 kDa–2 kDa for small proteins up to 20 kDa, Guttman and Nolan, 1994). Deviations of expected and calculated masses are larger than 2 kDa for 75 % and 3 kDa for 53 % (Supplementary table S8). These larger variations cannot be explained solely by measurement inaccuracies due to errors in mass calculation or distorted migration quantification on the basis of e.g., broad bands. But, as hypothesized for the protein Gir2 of *S. cerevisiae* and verified for mammalian α -crystallin A chain, the amino acid compositions of proteins impact their run behavior in SDS-PAGE (Alves *et al.*, 2004; Jong *et al.*, 1978). Especially a high content of acidic amino acids seems to reduce binding of SDS to Gir2. As a consequence, the protein is denatured incompletely, which probably results in slow migration in the gel matrix. Nevertheless, folding/unfolding has a minor effect on mobility of a protein, as it was shown for SDS-containing PAGE compared to non SDS-PAGE (Dunker and Rueckert, 1969). Only a substantially higher amount of SDS molecules bound to the unfolded protein in comparison to the folded protein lead to an increased negative charge of the linearized protein which might surpass the enhanced frictional resistance of the unfolded protein. Although protein denaturation was conducted, i.e., addition of protein denaturing agents and heating, the predominantly increased masses detected in the experiments are in accordance with SPA-tagged protein detection conducted by Weaver *et al.* (2019).

More than half of the proteins in Western blot appeared unambiguously, whereas the signals of the remaining proteins showed more uncertain patterns (Table 3.1). In most cases, background signals are increased though one particular band could be assigned to the desired protein. For seven candidates, smeared signals were discovered. In any case, optimization of loaded sample volume, blotting and staining probably improves protein signals. As reported by Weaver *et al.* (2019), different behaviors of individual samples complicate simultaneous detection of unknown proteins and optimized method adjustments are difficult to conduct for a high-throughput approach without having knowledge about the protein behavior in advance.

In contrast, defined by-products in addition to the main protein signals were detected on the blots for 21 candidates. Most often, the additional stained bands had a lower molecular mass. As hypothesized for the manganese regulated small protein MntS, signals could be caused by posttranslational modifications or different protein conformations (Waters *et al.*, 2011). Moreover, additional bands could result from protein degradation. As protein half-lives can range from several minutes to many hours (Goldberg, 2003) and protein lysates were analyzed four hours after protein induction, degradation is a reasonable cause of additional bands. Using a test series analyzing proteins at different time points after induction might allow distinguishing between stable protein conformations and unstable/degraded products, as it was shown for the overlapping gene *asa* (Vanderhaeghen *et al.*, 2018). Furthermore, there is increasing evidence that protein isoforms are translated in bacteria from alternative translation initiation sites (e.g. Meydan *et al.*, 2019). This assumption is reinforced by different studies detecting internal transcription start sites within annotated genes (e.g. Sharma *et al.*, 2010; Thomason *et al.*, 2015). These additional sites are reported to be at least in part responsible for the transcription of alternative mRNAs which can be translated into protein isoforms (Ten-Caten *et al.*, 2018). Analysis of present Cappable-seq TSS data indeed revealed that six OGCs with by-products have an internal transcription start site. Five of these (OGC 3, OGC 75, OGC 76, OGC 204, OGC 215) have a downstream in-frame start codon which could produce a protein of the desired, smaller size. Follow up analysis including an MS approach measuring N-terminal peptides (N-terminomics, Impens *et al.*, 2017) or chromosomal tagging could on the one hand verify different translation start sites

and on the other hand examine the natural expression of protein isoforms. Preliminary tests to confirm promoter activity or native expression of the candidates could increase the success of the methods as both rely on endogenous protein presence at suitable amounts. In addition to ORFs producing smaller by-products, seven candidates with additional stained proteins at substantially higher molecular weights were detected. For these OGCs (No. 33, 39, 73, 76, 144, 172, and 177) it is unknown whether the most probable band was wrongly assigned or technical/biological reasons exist for the abnormal band patterns. Possible explanations include inappropriate antibody concentrations and overloaded SDS polyacrylamide gels which might result in unspecific binding to and detection of other proteins. Furthermore, incompletely denatured proteins due to insufficient SDS or heating might not lose their quaternary structure and high molecular weight protein complexes are visible in Western blots. In all cases of stained by-products, it is inconclusive with this first high-throughput Western blot analysis, whether possible isoforms or conformational variants are real. Additional experiments would be necessary to improve Western blot protocols and reduce artificial signals due to technical errors.

In general, the protein analysis showed that the overlapping gene candidates analyzed mostly form stable proteins under the examined condition, even though functionality cannot be inferred from protein presence. Even random DNA sequences are reported to be soluble and detectable in SDS-PAGE and Western blots (Priambada *et al.*, 1996; Doi *et al.*, 2005). However, there are hints that random sequences can sometimes exhibit bioactivity (Neme *et al.*, 2017) and even unfolded and unstructured proteins can be functional (Dyson and Wright, 2005).

4.2 High- and low-throughput overexpression assays detect growth phenotypes

The need for methods to functionally characterize unexplored gene products becomes obvious when the enormous number of 1431 uncharacterized proteins in *E. coli*, one of the best studied organism, is considered (Hu *et al.*, 2009). Availability of NGS methods enabled researchers to develop large-scale approaches for effective and cost efficient functional genome studies (Gray *et al.*, 2015). Although most studies target loss-of-function phenotypes (e. g. Baba *et al.*, 2006; Giaever *et al.*, 2002), gain-of-function screenings are powerful tools to

complement missing phenotypic effects (Prelich, 2012).

4.2.1 Combination of overexpression assays revealed many OGCs producing growth phenotypes

The present study used two different overexpression assays to measure the effect of previously identified overlapping gene candidates on the growth of EHEC when overexpressed (Table 4.1).

Sets of 206 and 51 overlapping gene candidates were analyzed in high- and low-throughput phenotyping approaches, respectively, and 25 % of each exhibited altered growth upon overexpression. As a preselected set of OGCs with overexpression phenotype in the HT approach was applied in the single competitive assays, it was presumed that the success rate in follow up analyses is increased (> 25 %). Although this hypothesis was not met and 75 % of candidates tested did not have changed growth in each of the applied experimental settings, it is hypothesized that at least some phenotypes escaped detection due to the following reasons:

- 1) Experimental design: Substantial differences in the experimental settings of applied phenotyping approaches led to a limited comparability of the screenings (Table 4.1). For instance, the composition of the competitive pools with transformants overexpressing wild type sequences only in the HT approach or a mixture of bacteria expressing wild type and mutant sequences in the LT approach. Therefore, different growth effects are measured, in particular relative effects in the HT approach and absolute growth effects between wild type- and mutant-sequence expressing strains. Thus, a missing phenotype in the LT approach could be explained by the limited comparability of the two assays, but also a false positive phenotype in the HT analysis might be expected if one strain is growing much better and suppresses all others in a condition. Furthermore, a non-coding functionality of the overlapping ORF in general (Housman and Ulitsky, 2016) could also explain the disagreement of the two assays as only the LT approaches distinguishes between protein-coding and non-coding functionality of the sequences. Competing strains in single competitive growth assays expressed a wild type and a mutated form of the ORF producing significantly different putative proteins, a full length or a truncated version. In contrast, transcripts are nearly identical

Table 4.1: Comparison of HT and LT phenotyping.

	HT approach	LT approach
hypothesis	most conditions do not alter growth of bacteria; specific stresses lead to growth (dis-)advantage upon overexpression of overlapping gene candidates	
candidates selected	overlapping ORFs with ribosomal footprints	overlapping ORFs with phenotype in HT approach
assayed stress conditions	19	LB and condition(s) with HT growth phenotype
assayed bacteria per stress	206	2
plasmid of transformants in pool	includes intact OGC	includes intact or translationally arrested OGC
phenotype induction	overexpression from L-arabinose inducible plasmid	
analysis of bacteria / plasmids	NGS on Illumina MiSeq of isolated plasmids	Sanger sequencing of isolated plasmids
evaluation	transformation of sequencing reads to RPKM values and calculation of z-scores across all conditions	calculation of percentage of Wt and Mt plasmid represented by fluorescence intensities at the mutated position(s) per condition
phenotype criterion	$ z \geq 2$	significant ($p < 0.5$) altered plasmid proportions after growth, visual inspection
interpretation of phenotype	overexpression of an intact sequence leads to a relative growth dis-/advantage only in selected conditions and indicates functionality	overexpression of an intact sequence leads to a growth dis-/advantage compared to translationally arrested sequence and indicates a protein-coding potential

and vary in a maximum of three bases. Although it was shown for the *trans*-encoded small non-coding RNA SgrS that changing six nucleotides is sufficient to alter binding to the mRNA of the target gene *ptsG* as well as regulating its expression (Kawamoto *et al.*, 2006), it is assumed that nucleotide changes for up to three positions do not affect the activity of the non-coding RNA particularly for *cis*-encoded antisense RNAs which bind to the target gene over their entire length (Bobrovskyy and Vanderpool, 2013).

- 2) Experimental implementation: A comparatively low number of stress conditions was tested (19 conditions). Published chemical genetic screens are characterized by a broad coverage of diverse stress categories with several hundreds of different components (e. g. Nichols *et al.*, 2011, 324 conditions). Even a recently published overexpression phenotyping highly similar to the present HT approach investigated more than 50 experimental conditions including different carbon sources as well as inhibitory compounds like salts, metals, and antibiotics (Mutalik *et al.*, 2019). Despite an increased number of assayed conditions, the percentage of genes with high-confidence growth effect was comparable (813 genes with phenotype out of 4151 analyzed genes, 20%) but still low likely due to the increased number of assayed genes. However, a survey of yeast deletion mutants in more than 400 environmental treatments or stresses yielded growth phenotypes for almost all genes, i. e., 97% (Hillenmeyer *et al.*, 2008). Therefore, it can reasonably be assumed that phenotypes are awaiting to be discovered for the remainder of the set of 206 overlapping gene candidates.
- 3) Phenotype definition: Phenotypes in present and previously published studies are defined as altered growth behavior of bacteria (e.g. Boyer *et al.*, 2004; Deutschbauer *et al.*, 2014). However, all genes without effect on growth of the bacteria remain unnoticed. Further phenotypic effects are conceivable for instance genes influencing mobility of the bacterium (Bogomolnaya *et al.*, 2014) or cell morphology (Sycuro *et al.*, 2013). Although assays have been developed to identify genes impacting a specific function, e. g., envelope biogenesis of bacteria by means of a high-throughput colorimetric *lacZ* based assay (Paradis-Bleau *et al.*, 2014), screenings of overlapping genes for specific

phenotypes are challenging since they often lack homologs in other bacteria, thus, selecting the right assay is difficult.

Major differences of the two applied phenotypic assays make it difficult to assign phenotypes unambiguously. Nevertheless, both methods generate sets of candidates with putative growth effects, which are reproducible on their own and, thus, are reasonably no artifacts. Therefore, it is proposed that both experiments are appropriate to acquire growth phenotypes and either of the methods is applicable.

4.2.2 Are phenotypes due to an overexpression burden or a specific protein activity?

The growth of EHEC was influenced upon overexpression of overlapping gene candidates positively as well as negatively (Tables 3.5 and 3.6). However, statistically valid negative phenotypic effects dominated the single competitive growth analysis (85%), whereas phenotype directions were more balanced in the HT-approach. As relative growth effects were measured in the HT-approach, the phenotype directions (i. e., growth advantage or disadvantage) have only a little informative value and the absolute effect of an overlapping gene cannot be derived.

Overall, inhibitory effects on the growth of *E. coli* upon overexpression of genes (Mutalik *et al.*, 2019) and even random sequences (Neme *et al.*, 2017) have been described, but the cause of the negative phenotype remains unclear. As summarized by Bolognesi and Lehner (2018), toxicity of protein overexpression can for example occur due to overloaded protein transport systems taking the overexpressed protein to the target compartment of the cell, excessive catalytic activity of the protein unbalancing the cell metabolism, or forming protein aggregates which are harmful for the cell. Furthermore, protein overproduction can be harmful for cells if the cost of the translation process from the nucleic acid sequence into proteins as well as associated steps are exceptionally high (Stoebel *et al.*, 2008; Kafri *et al.*, 2016). In further consequence, other and more important proteins are not translated and the fitness of the cell decreases. Moriya and coworkers figured out that a metabolic burden is only reached if proteins are expressed at levels of 15% of the total protein content in yeast for GFP (Kintaka *et al.*, 2016) and several glycolytic proteins (Eguchi *et al.*, 2018). Such a

threshold of massive and burdening protein expression was probably not reached here, since any detection of overlapping gene proteins even under optimal conditions in *E. coli* Top10 was challenging and protein signals remained weak even if overexpressed (Section 3.2; Fellner *et al.*, 2014). Consequently, growth effects detected in competitive growth assays are presumed to result from any functionality of the gene product rather than being the sole effect of cell stress due to overexpression. Furthermore, competing strains are induced at the same level since they grow together, and the overall burden - if any - should be similar. To overcome the limitation of metabolic burden upon overexpression, one approach could be similar to Mutalik *et al.* (2019), who relied on endogenous promoter sequences upstream of the genes of interest cloned on a promoterless plasmid. As overlapping genes are thought to be young genes (Fellner *et al.*, 2015; Tautz, 2014) and might have still evolving gene structure elements like the promoter (Carvunis *et al.*, 2012), any native expression is probably insufficient to induce a detectable phenotype.

Although mainly negative fitness effects were detected in the LT approach, which is in agreement with other overexpression studies (e.g. Neme *et al.*, 2017), positive growth phenotypes detected for OGC 121, OGC 226, OGC 231, and *pop* are unambiguous effects induced by the respective overlapping gene. The positive influence on the growth can exclude toxicity and a certain function of the particular protein can be assumed. The low amount of positive growth phenotypes might also be explained by the identical genomic wild type background of the cells. The wild type genomic OLG copy in mutant transformants might be induced upon stress exposure and probably compensates phenotypes mediated by the plasmid encoded OGC. In any case, also negative growth effects hint towards a function beyond a purely burdening overproduction (Prelich, 2012). The OGCs tested are most likely no enzymes but rather regulating proteins of some sort. Overexpression would cause metabolic interference and this, in turn, would cause a negative effect on growth.

In summary, high- and low-throughput overexpression assays revealed positive as well as negative growth phenotypes and a reasonable inference of functionality of the underlying proteins and the corresponding genes can be made for both. However, the mechanism is unknown in all cases.

4.3 Transcriptional start site determination using Cappable-seq

4.3.1 From dRNA-seq to Cappable-seq

Genome-wide precise transcriptional start site detection started in 2010, when Sharma *et al.* (2010) developed the method dRNA-seq to discriminate RNA molecules of different phosphorylation states. The enzyme TEX (terminator exonuclease) enables a selective depletion of processed monophosphorylated transcripts, the predominant RNA species consisting of rRNAs and tRNAs as well as RNA degradation products. In contrast, primary triphosphorylated molecules (mRNAs) remain unchanged and provide information about transcription start sites. To analyze the relative enrichment patterns between primary and processed transcripts, a control library without TEX treatment is obligatory (Thomason *et al.*, 2015). The activity of TEX was reported to be specific for degradation of 5' monophosphorylated RNAs (Sharma *et al.*, 2010), but it is inhibited by secondary structures of the RNA which protect RNA against degradation (Sharma *et al.*, 2010; Zhelyazkova *et al.*, 2012; Jäger *et al.*, 2014). Hence, TEX treated libraries still contain processed RNAs (25% rRNA and 27% tRNA reported by Sharma *et al.*, 2010).

To overcome such limitations associated with an indirect enrichment of triphosphorylated RNA species, Ettwiller *et al.* (2016) established Cappable-seq where triphosphorylated transcripts are enzymatically biotinylated and captured with streptavidin beads. As suggested by the inventors of Cappable-seq, an untreated/non enriched control library does not need to be sequenced as the false positive rate only decreases from 3.7% to 1.4%, which is not pivotal. Furthermore, the fraction of usable and informative reads not mapping to rRNA and tRNA regions is increased by this direct enrichment strategy (3% rRNA mapped reads reported by Ettwiller *et al.*, 2016, 19% to 30% rRNA + tRNA mapped reads reported in this study, Table 3.8a). In combination with a high information depth facilitated by NGS, the amount of weakly expressed transcription start sites increases and the probability to detect further weak TSS likely increases with sequencing depth. Therefore, Cappable-seq was used here to find TSS for overlapping ORFs.

4.3.2 Performance of Cappable-seq

In Cappable-seq combined with tagRNA-seq (Innocenti *et al.*, 2015) applied here, mono- and triphosphorylated transcripts were differentially labeled and after deep sequencing each sequencing read was assigned to either the set of processing sites (PSS-set) or transcriptional start sites (TSS-set). Although this method constitutes an elegant method to separate different transcripts, the proportion of reads with a sequence tag identifier reported by Innocenti *et al.* (2015) was low (2.1% to 4.7% representing 1.4 to 3.3 million reads with either sequence tag). However, if the approach is combined with the Cappable-seq enrichment procedure, the number of reads with discriminative sequence tags is significantly increased (> 94% of reads contain sequence tag in each single experiment).

Despite enrichment for 5' triphosphorylated transcripts, a considerable number of sequencing reads were marked with a PSS-tag (up to 35%, cumulative proportion across biological replicates). Unspecific binding of the hydrophilic backbone of RNA molecules without biotin cap to hydrophilic magnetic beads is a known source of contamination in bead based applications and may justify the PSS-set in part (Figure 4.1, path 1). From a statistical perspective, rRNAs and tRNAs should represent the main RNA species in the PSS-set as they dominate total RNA samples (Karpinets *et al.*, 2006) and indeed, 53% of reads on average map to these regions. However, 47% of reads are located in intergenic, coding and antisense parts of the genome. They might be explained by spontaneous hydrolysis of the biotin cap from labeled transcripts during the first ligation step, making transcripts accessible for ligation of the PSS sequence tag (Figure 4.1, path 2, Innocenti *et al.*, 2015). Furthermore, hydrolysis of 5' triphosphates into 5' monophosphates by the enzyme RppH, the first step of mRNA decay, produces additional fragments available for different Cappable-seq contamination paths (Figure 4.1, path 3, Deana *et al.*, 2008; Celesnik *et al.*, 2007).

Besides RNAs contaminating the Cappable-library in form of the PSS-set, a small fraction of rRNAs and tRNAs, as well as degraded primary transcripts, are found in the TSS-set (Sections 3.5.1 and 3.5.4, Table 3.8b). Because the specificity and efficiency of the enzymes used is limited, it is assumed that, on the one hand, the capping enzyme adds biotin caps even to processed monophosphorylated transcripts (Figure 4.1, path 4, Fritz Thümmel, vertis Biotechnologie AG, personal communication) and, on the other hand, monophosphorylated

fragments which escaped ligation of the sequence tag in the first ligation step are substrates for the second ligation step (Figure 4.1, path 5, Innocenti *et al.*, 2015; Raabe *et al.*, 2013). In both cases processed transcripts are incorrectly marked with the TSS-set specific sequence tag. However, the number of processed transcripts as measured by the fraction of reads mapping to rRNA and tRNA regions in the genome, which is maximal 15 % of total mapped reads in each condition and, therefore, low enough in order to be able to confidently specify transcription start sites.

Although the PSS-data set was originally used by Innocenti *et al.* (2015) to discriminate between true start sites and processing sites according to the fraction of reads mapping in

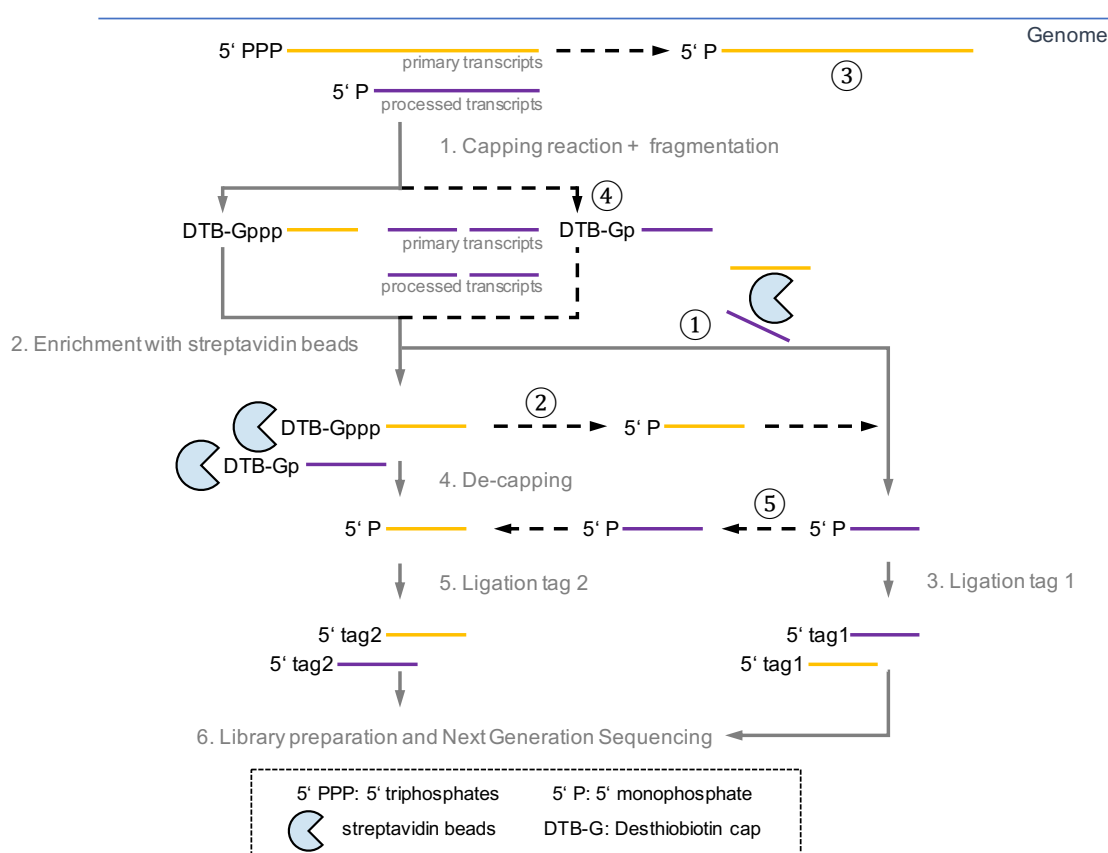


Figure 4.1: Sources of contamination during Cappable-seq. The workflow of Cappable-seq of Ettwiller *et al.* (2016), adapted by vertis Biotechnologie AG, is shown (gray lines). Contamination pathways are shown in black dashed lines. 1, unspecific binding of RNA molecules to beads. 2, spontaneous hydrolysis of capped fragments. 3, degradation products of mRNA decay. 4, limited capping enzyme specificity. 5, failed adapter ligation during ligation of tag 1.

each set at a certain position, they were used in the present study solely to increase the purity of the TSS-set (i. e., reduction of rRNA/tRNA mapping reads from 30 % to 11 %). Probably, preselection with streptavidin beads prior to adapter ligation as well as previously described flows of individual fragments skews the proportion unpredictably, which might complicate or even rule out differentiation of primary and processing transcripts.

Sources of ‘contaminating’ reads are diverse as discussed previously, but the low amount of reads mapping to rRNAs and tRNAs especially in the TSS-set indicate trustworthy data for determining transcription start sites reliably. Ettwiller *et al.* (2016) compared TSS from Cappable-seq to dRNA-seq derived TSS and other deposited TSS (RegulonDB v 8.6, Salgado *et al.*, 2012) to reinforce the reliability of Cappable-seq. Indeed, a total number of 9600 out of 16 359 TSS are present in at least one further data set. Since no further high-throughput TSS analysis in the pathogenic *E. coli* O157:H7 EDL933 has been conducted so far, TSS could not directly be compared to existing data. Nevertheless, TSS of 3685 homologous genes of *E. coli* O157:H7 EDL933 and *E. coli* MG1655 were compared (gene data was provided by Dr. Zachary Ardern). A total of 1685 genes have reported TSS in both *E. coli* strains, pathogenic (this study) and apathogenic (RegulonDB v 10.5, Santos-Zavaleta *et al.*, 2018). The distance between TSSs and start codons for 56 % of genes differed for at most 2 bp and for 75 % for no more than 10 bp between the corresponding genes. Although data sets are only partially comparable due to genome differences of analyzed bacterial strains, it can be presumed that tagRNA-seq coupled Cappable-seq is a precise method for TSS identification as a notable number of TSS appear to have been predicted correctly. Furthermore, experimentally identified TSS of the annotated genes *gadA* and *gadB* from *E. coli* O157:H7 EDL933 (Bhagwat and Bhagwat, 2004) match perfectly to TSS positions determined in the present study. This finding supports the precision of TSS determination with Cappable-seq as well as the data evaluation applied.

4.3.3 Identification of hundreds of transcriptional start sites antisense to annotated genes upstream of overlapping genes

Investigation of the transcriptional landscapes of bacterial cells was revolutionized with the introduction of dRNA-seq (Sharma *et al.*, 2010). But even before, microarrays and modified

5' RACE protocols began to uncover the complexity of prokaryotic transcriptomes regarding for example transcriptional control (Tjaden *et al.*, 2002; Mendoza-Vargas *et al.*, 2009). Combination of high-resolution methods and precise transcription start site mapping allowed a more defined definition of weakly transcribed transcripts compared to stand alone RNAseq, which can be insufficient for this application (Landstorfer *et al.*, 2014). Therefore, TSS determination is required for precise analysis of the transcriptome of bacteria.

The presence of a TSS is evidence for transcription of a certain genomic part into RNA, but it cannot be deduced whether these RNAs are translated. Although different studies identified larger numbers of antisense TSS in diverse prokaryotes (Jäger *et al.*, 2014; Thomason *et al.*, 2015) and the importance of small non-coding RNAs encoded either in *cis* or in *trans* for gene regulation has been experimentally verified (Thomason and Storz, 2010; Storz *et al.*, 2011; Bobrovskyy and Vanderpool, 2014; Papenfort *et al.*, 2015), the specificity of genome wide transcription of antisense regions is debated (Raghavan *et al.*, 2012; Dornenburg *et al.*, 2010). In addition, despite clear indications for start codon signals driving translation of some antisense transcripts (Weaver *et al.*, 2019; Meydan *et al.*, 2019), hypotheses of functional expression of overlapping genes are rejected in favor of 'translational noise' (Smith *et al.*, 2019). Nevertheless, finding TSS for overlapping gene candidates definitely strengthens arguments for the presence and functionality of overlapping genes.

Cappable-seq was specifically used for this application, namely to identify TSS in antisense regions upstream of overlapping genes and ORFs. In general, the TSS identification algorithms of Ettwiller *et al.* (2016) were adopted, but an analysis was performed to find optimal criteria to reliably identify antisense TSS (Section 3.5.4, Figure 3.10c). It could be shown that genome positions with a RRS of at least 1.5 in antisense regions are trustworthy TSS and most probably do not represent transcriptional noise. In addition, genome positions had to have reproducible signals in order to be identified as TSS with 'reproducible' defined here as signals present in all three biological replicates of one condition analyzed. Thus, TSS identified are highly reproducible. With these criteria TSS for 112 out of 216 OGCs were detected as well as another 4000 antisense TSS for a further set of embedded antisense overlapping ORFs. Nevertheless, the entirety of start sites antisense to annotated genes was not analyzed

TSSpredator is a program developed by Dugar *et al.* (2013) for automated TSS annotation in dRNA-seq data, which uses similar principles as applied in the presented work for reproducible TSS signals within biological replicates. The number of antisense TSS (asTSS) predicted with this method in different *Campylobacter jejuni* strains is clearly lower (< 1000) but still remarkable high considering the genome length of < 1.8 Mb. In contrast, Thomason *et al.* (2015) adapted settings of TSSpredator and reported almost 5500 asTSS in the substantially larger genome of *E. coli* MG1655 (4.6 Mb). However, they used relaxed reproducibility criteria to define positions as TSS, e. g., signal in 2/3 or 3/5 replicates including technical replicates analyzed on different sequencing machines, or for TSS with a sufficient high signal strength a weaker signals in further replicates is acceptable. Although it was observed that higher variability is rather achieved for RNA sequenced on various platforms than for RNA sampled in different conditions (Wade, 2015), reliability of some TSS might be questioned. Nevertheless, these numbers indicate that many TSS antisense of annotated genes can be detected, but signals are more confident if more stringent selection criteria are used as was done for TSS determination of OGCs with Cappable-seq.

In any case, automated TSS annotation is difficult and manual revision is mandatory as suggested by Kröger *et al.* (2013) and conducted for the set of TSS of OGCs. For some OGCs it was hypothesized that they are parts of operons, but TSS determination on its own is insufficient to unambiguously assign the transcriptional units. Combining start sites with high-throughput data of transcriptional terminators (e. g., by Term-seq, Dar *et al.*, 2016) and deep analysis of RNA coverage of the genome by RNA-seq can give insights into the operon architecture of bacterial genes (Conway *et al.*, 2014).

4.3.4 TSS functionality is supported by reproducible and differential signals as well as promoter activity

Independent of the method used for transcription start site determination, either dRNA-seq, Cappable-seq or genome wide 5' RACE-based protocols, the number of positions enriched in the primary transcriptome is surprising (Sharma *et al.*, 2010; Ettwiller *et al.*, 2016; Mendoza-Vargas *et al.*, 2009).

As mentioned, some researchers doubt the functionality of antisense transcription start

sites (Raghavan *et al.*, 2012; Dornenburg *et al.*, 2010). However, increasing evidence is reported which supports their functional nature:

- 1) Conservation and reproducibility
- 2) Differential expression
- 3) Enrichment of promoter elements.

The specificity of antisense transcription is reinforced by reproducible and conserved expression patterns. As signals in TSS-studies are indeed reproducible and detectable in various experiments across several biological replicates it can generally be excluded that antisense TSS are artifacts of library preparation, and must instead be real TSS (Sharma *et al.*, 2010). Several large scale analyses showed that asTSS are conserved between four strains of the species *C. jejuni* (Dugar *et al.*, 2013), two species of the genus *Listeria* (Wurtzel *et al.*, 2012) or eight species of the genus *Shewanella* (Shao *et al.*, 2014). Although they are less conserved than primary TSS of annotated genes, at least some asTSS appear to be conserved, thus, targeted and specific expression can be presumed. For antisense RNAs, it was thought that missing conservation is an indication for missing functionality (Raghavan *et al.*, 2012). However, Wade and Grainger (2014) mentioned, that this is not necessarily true, because if antisense and otherwise pervasive transcription has no function, it is questionable why selection has not eliminated this unproductive transcription. It seems that functionality may be the best response (Wade and Grainger, 2014).

In contrast to conserved expression, transcription start sites can be highly specific for distinct growth conditions. Kröger *et al.* (2013) showed that 86 % of all genes in *Salmonella* have a TSS and are expressed in at least one of 22 infection relevant growth and stress conditions. Each individual condition has less genes with active TSS, indicating condition specific differences and hence regulation. Similarly, antisense transcripts are differentially expressed depending on the induced stress (Kröger *et al.*, 2013). Though only presumed for the group of more conserved internal TSS, transcriptional start sites which evolved with a regulated expression are likely functional (Shao *et al.*, 2014). Despite lower conservation and expression levels of antisense TSS, regulation and differential expression might indicate functionality of these and any other transcription start site (Thomason *et al.*, 2015).

Another argument for the functionality of (antisense) TSS are enriched upstream promoter elements. Especially the A/T-rich Pribnow-box was found to be conserved in different studies, even for antisense transcripts (Thomason *et al.*, 2015; Raghavan *et al.*, 2012; Shao *et al.*, 2014). Nevertheless, Mendoza-Vargas *et al.* (2009) pointed out that the information content of a promoter is low and spurious transcription from promoters distributed all over the genome can occur easily, especially as promoter sequences can easily emerge due to mutation bias in prokaryotic genomes towards adenine and thymine (Hershberg and Petrov, 2010). Therefore, presence of a promoter upstream of the TSS only indicates genuine expression of a gene product from this start site, but a final conclusion about functionality of the transcription output cannot be drawn.

Each of the points just mentioned were raised to support antisense transcription for (small) antisense encoded RNAs, but most arguments are still disputed as remarked above. However, additional strong evidence is provided here for transcription and function of antisense encoded overlapping genes and ORFs, since their functionalities and/or activities has been verified by phenotypes (see above) and in these ways:

- 1) Reproducibility: TSS are reproducibly detected in independent biological replicates. Apart from a first analysis of annotated gene associated TSS of *E. coli* MG1655 and EHEC, which showed a consistent TSS/start codon distance for 75 % of genes present in both *E. coli* strains, conservation analyses especially for antisense regions across species are still missing due to the lack of published data, but might be subject of evolution oriented prospective projects.
- 2) Differential expression: More than a hundred TSS upstream of overlapping gene candidates were identified in at least one of eight samples analyzed, either from cells grown to exponential or stationary phase in rich, minimal, acid stress or salt stress medium. Certainly, more different conditions would increase the number of TSS associated with overlapping genes and their differential activation. Here, specific differential expression was observed for more than 80 TSS for OGCs (Figure 3.13c, Table 3.13), which strengthens their targeted and regulated expression. Additionally, 4844 TSS for embedded ORFs in antisense regions were identified and at least 71 % have a significantly

different TSS expression strength.

- 3) Promoter elements: Conserved promoter elements were observed for TSS upstream of OGCs as well as of antisense embedded ORFs. Enrichment for A/T rich Pribnow-boxes was detected for these sequences. In contrast, no sequence conservation was found for random genome positions analyzed as negative set. Thus, most TSS expression occurs due to active promoters. For 7 OGCs, 10 putative promoter sequences were tested and 9 could be verified. Concerning these seven OGCs, four and three exhibited high- and low-throughput overexpression phenotypes, respectively. Additionally, it was possible to distinguish between the main and minor TSSs and promoters based on the RRSs of TSS and GFP fluorescence mediated by the promoters (Figures 3.20 and 3.21). For transcription start sites of OGC 96, differential regulation was verified based on promoter-reporter fusion experiments.

4.3.5 Cappable-seq gives initial access to sense overlapping genes

In the past, analysis and characterization of overlapping genes was limited to antisense overlapping genes. Consequently, the set of 216 translated OGCs was preselected by means of ribosome profiling (Section 3.1, Landstorfer, 2014). Although many insights into the translational status of bacterial cells were gained in different bacteria by this method (e.g. Jeong *et al.*, 2016; Fisunov *et al.*, 2015; Woolstenhulme *et al.*, 2015), the standard application for prokaryotes suffers from a restricted precision to unambiguously uncover the codon-periodicity of translating ribosomes at single gene level (Landstorfer, 2014; Hücker, Ardern, *et al.*, 2017; Dingwall *et al.*, 1981). Therefore, it is currently not possible to map strand-specific ribosome profiling reads to recognize a certain reading frame, neither antisense nor sense.

However, Cappable-seq data is able to detect sense TSS within annotated genes. Although for most sense overlapping genes a clear assignment of the TSS to a certain ORF was not possible (992/1076, > 90%), at least 44 TSS seem to be specific for new transcripts overlapping annotated genes on the sense strand. Further on, their upstream sequences show a slightly conserved Pribnow box; thus, it can reasonably be assumed that ORFs are transcribed. Additionally, 13 candidates were found to have a ribosome binding site.

However, further experiments must verify sense overlapping gene transcription.

Intragenic transcriptional start sites are, like antisense TSS, more often identified in primary transcriptome studies (e. g. Papenfort *et al.*, 2015; Sharma *et al.*, 2010). For instance, the transcription of *micL* initiates from a σ^E dependent promoter within the copper homeostasis protein *cutC* in *E. coli* (Guo *et al.*, 2014). Thus, *micL* was identified as a sense encoded small RNA which regulates the expression of the membrane lipoprotein Lpp. As observed by Thomason *et al.* (2015), an intragenic TSS located at the 3' end of an annotated gene most likely belongs to the neighboring downstream annotated gene. Nevertheless, they identified even plenty internal sense TSS not either close to the 5' or 3' end of annotated genes indicating rather independent transcripts. Unfortunately, the translation status of sense overlapping regions of such TSS remains unclear, but recent methodological improvements allow better insight on the bacterial transcriptome.

For instance, Hwang and Buskirk (2016) performed ribosome profiling using the nuclease RelE. Usually, microcococcus nuclease or a mixture of different endo- and exonucleases are added to cell lysates to cut and digest mRNA not protected by translating ribosomes to obtain the ribosomal footprints (Oh *et al.*, 2011; Hücker, Arden, *et al.*, 2017). In contrast, RelE cuts mRNA specifically within the ribosomal A-site at the second nucleotide position of a codon with low or nearly no sequence specificity (Pedersen *et al.*, 2003; Hurley *et al.*, 2011). This property of RelE allows the determination of the reading frame of the translating ribosomes at single gene level. This has been shown exemplarily for *prfB* (Hwang and Buskirk, 2016). This improvement will allow examining regions in which different sense reading frames are translated.

Another novel strategy involves to stall initiating ribosomes at the start codon by using the antimicrobial peptide onc112 or the antibiotics retapamulin or tetracycline (Weaver *et al.*, 2019; Meydan *et al.*, 2019; Nakahigashi *et al.*, 2016). Here, translation initiation sites (TIS), i. e., start codons, are mapped on a genomic scale. All currently available studies identified several unusual TIS within annotated genes or antisense thereof. These developing techniques will give even deeper insights into the field of sense overlapping ORFs in future.

Further verification of sense overlapping genes proposed here are made by recent proteomic techniques. In N-terminal proteomics, the TIS is identified based on sequence of proteins

present in the cell. For *Listeria monocytogenes*, 19 internal translation initiation sites were described, creating shorter protein isoforms compared to the currently known full-length annotated gene (Impens *et al.*, 2017). This method probably could be used to examine out-of-frame products, i. e., proteins of sense overlapping genes.

Although most studies ignore sense overlapping genes and only provide indirect hints towards these genes, Feltens *et al.* (2003) actually characterized the RNaseP protein gene (*rnpA*) overlapping completely the gene for the ribosomal protein L34 *rpmH* in the sense direction. This demonstration of alternative proteins in alternative sense reading frames, even though it is currently only a rare finding, already suggests that further layer of complexity not only to the transcriptome, but also to the translome of a bacterial cell has been overlooked so far. The data and studies mentioned above give ample evidence for (sense) overlapping genes.

4.4 Analysis of the pH-regulated overlapping gene *pop*

Single overlapping gene characterization studies are important but unfortunately not of highest priority in the era of genome wide analyses, despite such data provide tremendous starting points for such experiments. In recent years, some non-trivially overlapping genes have been characterized in detail, but most are typically short (Fellner *et al.*, 2014; Haycocks and Grainger, 2016). Thus, characterization of *pop* and its protein Pop with a length of 200 amino acids is of particular interest, since such long overlapping genes should not originate from random translation events of genomic loci (Smith *et al.*, 2019). The gene was named according to exhibited features, regulated by pH, overlapping, and protein-coding. However, it should not be mistaken for *hemC/F/H/L*, previously referred *popA/B/C/E*.

4.4.1 *pop* was probably overlooked due to its prominent mother gene *ompA*

Algorithms like Glimmer or Prodigal are used for automated genome annotations. However, the number of coding sequences in bacterial genomes predicted is often underestimated, especially as small genes as well as overlapping ORFs with extensive overlaps are neglected by these programs (Burge and Karlin, 1998; Delcher *et al.*, 2007; Hücker, Ardern, *et al.*, 2017). Nevertheless, it has been shown that *pop* is translated in the pathogen *E. coli* O157:H7

EDL933, as well as two further pathogenic *E. coli* strains (*E. coli* O157:H7 Sakai and *E. coli* LF82, Dr. Zachary Ardern, personal communication). As even short proteins expressed in *E. coli* can be bioenergetically cost intensive (Lynch and Marinov, 2015), it is presumed that selection removes such non-functional ORFs quickly. According to molecular clock estimations, these pathogenic *E. coli* strains diverged more than 4 million years ago (≈ 1 billion generations, Reid *et al.*, 2000). Therefore, any non-functional sequence shared with the ancestor of these strains should have been lost. This strongly assumes that *pop* was simply overlooked so far.

4.4.2 Typical gene structure of *pop*

The transcriptional unit of *pop* was studied and a gene structure was identified including the following elements:

- 1) transcriptional start site
- 2) active σ^{70} promoter
- 3) rho-independent terminator
- 4) coding open reading frame *pop* with the putative start codon CTG.

While experimental evidences for TSS, promoter, and terminator have a high precision, ribosome profiling data is less clear, as three regions with peaks in read coverage were found (regions 1-3 in Figure 3.26a). Each of these contains a putative start codon (Supplementary figure S4) and could, therefore, represent a translation initiation site (Oh *et al.*, 2011; Woolstenhulme *et al.*, 2015). Ribosome profiling reads cover region 2 best, thus, it is proposed that translation of *pop* starts at this point. Indeed, a Shine-Dalgarno motif for ribosome binding as well as a nearby CTG start codon was identified. Although CTG represents a rare start codon, it is used in prokaryotes (Yamamoto *et al.*, 2018; Hecht *et al.*, 2017; Sussman *et al.*, 1996; Meydan *et al.*, 2019).

In region 1, a TTG start codon was identified, which would serve as starting point for the longest potential open reading frame for *pop*. Additionally, this start codon was predicted by Prodigal as the most probable one. However, no ribosome binding site was identified

upstream thereof, though this is not a prerequisite for gene expression (Gualerzi and Pon, 2015; Moll *et al.*, 2002), as well as no TSS initiating transcription of *pop*. As experiments conducted did not show bicistronic expression of *pop* along with *ycbG*, possibly due to a predicted terminator downstream of the annotated gene ($\Delta G = -12$ kcal/mol, indicated in Figures 3.25 and 3.28b), TTG is unlikely the correct start codon.

Peak region 3 contains a GTG, which is located 45 amino acids downstream of the mutation introduced in *pop* for competitive growth analysis. In general, this position cannot be excluded as potential translation initiation site. However, the growth phenotypes found are not caused by the protein translated from this start site, because they depend on the artificially introduced stop codon further upstream. Though this is a strong argument against the GTG as the correct start codon, translation of a smaller protein isoform not carrying these phenotypes cannot be ruled out.

4.4.3 pH-dependent, protein-coding functionality of *pop*

Besides gene structure, protein-coding potential and functionality were investigated for the gene *pop* and the protein Pop. Western blots revealed that cells grown in different culture media express Pop at various stabilities. For instance, Pop seems to be unstable in LB, but more stable in bicin-buffered medium. Moreover, Western blot profiles indicated stable protein isoforms in growth medium with elevated pH levels, a phenomenon recently reported for other proteinaceous gene products in some bacteria (Meydan *et al.*, 2019; Vanderhaeghen *et al.*, 2018; Nakahigashi *et al.*, 2016; Waters *et al.*, 2011). VirF in *Shigella* and AerR in *Rhodobacter capsulatus* are two such examples (Di Martino *et al.*, 2016; Yamamoto *et al.*, 2018). While the former one, an AraC like activator, possesses two TSS and consequently two mRNAs are transcribed which are templates for the two protein forms, the cobalamin-binding photoreceptor AerR is translated from a single RNA either into a long or a short isoform. It is noteworthy that translation of the short AerR protein initiates at a CTG codon (Yamamoto *et al.*, 2018) which is also proposed for *pop*. However, putative Pop isoforms are highly unlikely since no internal TSS was identified in Cappable-seq. In any case, data strongly indicate that *pop* codes for a protein.

Maybe more important than the protein on Western blots, overexpression phenotypes in

competitive growth assays provide clear evidence for a proteinaceous nature of the *pop* gene product. Overexpression of *pop* was not only connected to decreased growth rates, which could just indicate the burden of the overexpression (Dong *et al.*, 1995; Shachrai *et al.*, 2010), but strains overexpressing *pop* grew better in medium acidified with L-malic acid. Thus, a growth defect solely caused by stressed cells due to protein overexpression is excluded. A genomic knock-out was analyzed, but results suggest that absence of the protein is not critical for EHEC under conditions investigated. However, it was shown that knock-out and overexpression phenotypes do not need to be complementary to be meaningful for either one (Prelich, 2012). For instance, separate overexpression of *CLN1* and *CLN2* in *Saccharomyces cerevisiae* can both compensate a cell cycle kinase mutation (Hadwiger *et al.*, 1989). In contrast to single gene deletions of *CLN1* or *CLN2*, respectively, where knock-out effects can be compensated by each of the intact genes, only a double mutant shows a phenotype. Similarly, only simultaneous deletion of four different genes coding for cold shock proteins in *E. coli* leads to a detectable low temperature sensitivity (Xia *et al.*, 2001). Therefore, it could be assumed that *pop* function is redundant and a knock-out could be compensated by the cell.

4.4.4 Pathogenesis related function of *pop*?

The exact cellular action of Pop was not examined, but it could be speculated that the positive overexpression growth effect in acidic medium is associated with an acid tolerance of EHEC which is necessary to survive the acid barrier in the stomach after ingestion (Nguyen and Sperandio, 2012). In further consequence, a pathogenicity or host-environment related function could be assigned to *pop* in pathogenic *E. coli* which is solely activated upon specific stresses. In line with a pathogenesis-related function is the finding for absent ribosome profiling reads in a pathogenic *E. coli* (Dr. Zachary Arden, personal communication).

4.5 Cumulative evidence for functionality of overlapping gene candidates

The work presented provides evidence at several levels for specific expression and, thus, function of overlapping genes and their corresponding proteins (Figure 4.2). A total of 216 OGCs were analyzed and 207 showed signals in at least one of the following experiments: Western blots, indicating a protein-coding potential, high-throughput overexpression phenotyping, indicating functionality of the ORF, and transcriptional start site determination, indicating specific and in some cases regulated transcription of an RNA molecule.

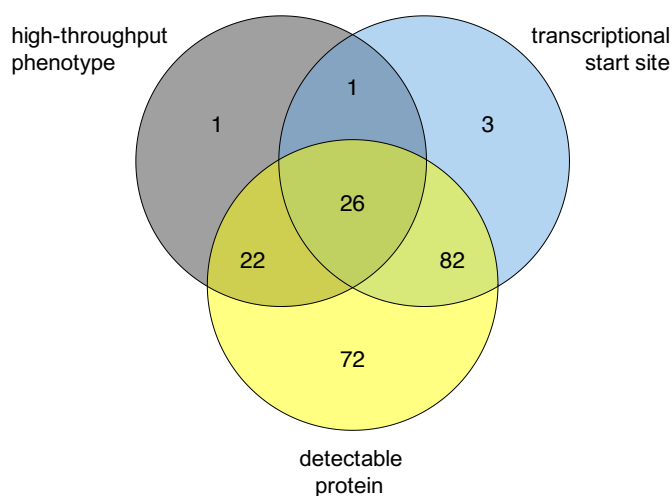


Figure 4.2: Summary of structural and functional characterization of 216 OGCs with translation signal in ribosome profiling. Venn diagram shows overlapping gene candidates supported by different types of experimental evidence.

In general, signals and effects for some candidates are weak compared to other genes, especially annotated ones. This correlates with a supposed recent evolutionary origin of emerging overlapping genes resulting from previously non-coding regions (Rogozin *et al.*, 2002) as is, for instance, seen for OGC 135. The sequence of this candidate can be translated into a stable protein, however, the effect of overexpression was too small to be seen in phenotypic analyses. Furthermore, the TSS associated with OGC 135 attracted attention due to differential signal strengths in different growth phases, but bioinformatic promoter identification was inconclusive and corroborated by the missing promoter activity.

It was shown that if a candidate has either an overexpression phenotype or a TSS, it is very likely that a stable protein can be detected in Western blots (130 candidates). Only

one candidate with TSS and phenotype has no protein. This overlapping gene candidate, OGC 31, revealed no positive signal in single competitive assay as well. As this method also investigates protein-coding potential of the candidates, OGC 31 possibly acts on the RNA level.

In contrast, 72 candidates had a detectable protein, but no TSS or phenotype. As discussed in Sections 4.2.1 and 4.3.4, increasing the number of analyzed conditions probably increases the number of candidates mediating a growth alteration upon overexpression or expressing a native transcription start site.

Overall, 26 candidates (highlighted in Supplementary table S6) have in each of the experiments positive signals providing best evidence for the functionality of overlapping genes. Furthermore, seven candidates of these (OGC 15, OGC 51, OGC 75, OGC 85, OGC 121, OGC 174, OGC 241) have also an overexpression phenotype in single competitive growth assays. As a general toxic effect of overexpression products in LT phenotyping experiments can probably be excluded as the majority of growth effects are stress dependent (Section 3.4.3), functionality of these seven candidates as proteins in EHEC is expected.

4.6 Concluding remarks

Browning and Busby (2004) stated "Although all genes are equal, some are more equal than others!". Even though most overlapping genes typically have weaker functional evidence and differ in their structure from textbook prokaryotic genes, they should be considered as true genes despite being 'less equal'.

This study showed accumulated data pointing towards functionality of overlapping genes. Nevertheless, their existence and functionality is controversially discussed among bacteriologists, questioned or even rejected (e. g., Raghavan *et al.*, 2012; Wade and Grainger, 2014). For instance, Almeida *et al.* (2019) investigated the primary antisense transcriptome of halophilic archaea and mapped hundreds of asTSS to *Halobacterium salinarum* annotated genes. Integration of ribosome profiling reads unveiled that one tenth of antisense transcripts are ribosome associated, but translation of these is not accepted and rather association of the ncRNAs with ribosomes for translational control is assumed (rancRNAs, ribosome associated non-protein-coding RNAs, Pircher *et al.*, 2014). However, especially long ORFs antisense to

annotated genes like *pop* or long non-coding asRNAs may form a hitherto greatly underestimated source of functional units exhibiting protein-coding potential (Pircher *et al.*, 2014) and await discovery.

Raghavan *et al.* (2012) suggested for the functional characterization of asRNAs that expression data should be linked with results of genetic and biochemical experiments as well as evolutionary analysis to get a comprehensive view of asRNAs. Following this proposal, large scale experimental studies were performed successfully in this dissertation to start characterizing not asRNAs, but antisense overlapping, protein-coding genes. The experiments yielded dozens of promising gene candidates. These could be worthwhile to be analyzed in more detail as Raghavan *et al.* (2012) stated that "there are undoubtedly some individual asRNAs that serve some biological function". As this holds true for at least some asRNAs, such biological functionality has been shown for selected overlapping genes.

4.7 Outlook

Detailed characterization of overlapping gene candidates with low-throughput overexpression phenotype according to the analysis of *pop* can add further examples to the slowly growing list of functional, protein-coding OLGs. Experiments could include loss-of-function analysis of a genomic knock-out and native mRNA expression as well as experiments describing the transcriptional unit including promoter and terminator.

OGC 60, OGC 95, and OGC 223 could be of particular interest. None of the candidates could be cloned so far either for phenotyping or for protein detection. This could be due to a toxic phenotype. Detailed experimental analysis could shed light on the function of these overlapping gene candidates.

Besides the wealth of information hidden in the data set provided by Cappable-seq, further experimental approaches should complement the assignment of transcriptional units. Dar *et al.* (2016) developed Term-seq, a high-throughput method to detect intrinsic terminator sequences. Combining knowledge about TSS and terminator should provide insights into the transcription status of overlapping genes, including operon structures. Furthermore, results of varying strength of transcription start sites in different culture conditions might be examined in follow-up experiments to indicate a potential differential regulation of the

underlying ORFs. Further data of TIS sequencing will allow to pinpoint each translational start site located on each transcriptional unit, either in antisense or sense to known genes.

The discovery of sense transcription start sites associated with sense OLGs was briefly mentioned. However, the identification of translation initiation sites within annotated genes (Weaver *et al.*, 2019; Meydan *et al.*, 2019) indicates that sense overlapping genes should be given a high priority. Combining TSS data with unusual translation initiation sites found by several researchers could provide initial promising candidates for future experimental analysis of probably hidden overlapping genes.

Even though a wealth of information is gained by whole genome methods like ribosome profiling, single gene characterizations are equally important to strengthen the argument for functional overlapping genes.

5 References

- Abduljalil, J. M. (2018). Bacterial riboswitches and RNA thermometers: Nature and contributions to pathogenesis. *Non-coding RNA research* **3**(2), 54–63.
- Abe, H., Tatsuno, I., Tobe, T., Okutani, A., Sasakawa, C. (2002). Bicarbonate ion stimulates the expression of locus of enterocyte effacement-encoded genes in enterohemorrhagic *Escherichia coli* O157:H7. *Infection and immunity* **70**(7), 3500–3509.
- Aksoy, S., Squires, C. L., Squires, C. (1984). Translational coupling of the *trpB* and *trpA* genes in the *Escherichia coli* tryptophan operon. *Journal of bacteriology* **157**(2), 363–367.
- Almeida, J. P. P. de, Vêncio, R. Z., Lorenzetti, A. P., Caten, F. t., Gomes-Filho, J. V., Koide, T. (2019). The Primary Antisense Transcriptome of *Halobacterium salinarum* NRC-1. *Genes* **10**(4), 280.
- Alsharif, G., Ahmad, S., Islam, M. S., Shah, R., Busby, S. J., Krachler, A. M. (2015). Host attachment and fluid shear are integrated into a mechanical signal regulating virulence in *Escherichia coli* O157: H7. *Proceedings of the National Academy of Sciences* **112**(17), 5503–5508.
- Alves, V. S., Pimenta, D. C., Sattlegger, E., Castilho, B. A. (2004). Biophysical characterization of Gir2, a highly acidic protein of *Saccharomyces cerevisiae* with anomalous electrophoretic behavior. *Biochemical and biophysical research communications* **314**(1), 229–234.
- Andersson, D. I., Jerlström-Hultqvist, J., Näsvall, J. (2015). Evolution of new functions de novo and from preexisting genes. *Cold Spring Harbor perspectives in biology* **7**(6), a017996.
- Arsène-Ploetze, F., Bertin, P. N., Carapito, C. (2015). Proteomic tools to decipher microbial community structure and functioning. *Environmental Science and Pollution Research* **22**(18), 13599–13612.
- 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. *Molecular systems biology* **2**(1).
- Baek, J., Lee, J., Yoon, K., Lee, H. (2017). Identification of Unannotated Small Genes in *Salmonella*. *G3: Genes, Genomes, Genetics* **7**(3), 983–989.

- Baeshen, M. N., Al-Hejin, A. M., Bora, R. S., Ahmed, M., Ramadan, H., Saini, K. S., Baeshen, N. A., Redwan, E. M. (2015). Production of biopharmaceuticals in *E. coli*: current scenario and future perspectives. *Journal of Microbiology and Biotechnology* **25**(7), 953–962.
- Balabanov, V. P., Kotova, V. Y., Kholodii, G. Y., Mindlin, S. Z., Zavlilgelsky, 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 microbiology letters* **337**(1), 55–60.
- Barrell, B. G., Air, G., Hutchison III, C. (1976). Overlapping genes in bacteriophage ϕ X174. *Nature* **264**(5581), 34.
- Bauwens, A., Betz, J., Meisen, I., Kemper, B., Karch, H., Müthing, J. (2013). Facing glycosphingolipid–Shiga toxin interaction: dire straits for endothelial cells of the human vasculature. *Cellular and Molecular Life Sciences* **70**(3), 425–457.
- Behrens, M., Sheikh, J., Nataro, J. P. (2002). Regulation of the overlapping *pic/set* locus in *Shigella flexneri* and enteroaggregative *Escherichia coli*. *Infection and immunity* **70**(6), 2915–2925.
- Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**(1), 289–300.
- Bergthorsson, U., Andersson, D. I., Roth, J. R. (2007). Ohno’s dilemma: evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences* **104**(43), 17004–17009.
- Besemer, J., Lomsadze, A., Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic acids research* **29**(12), 2607–2618.
- Bhagwat, A. A., Bhagwat, M. (2004). Comparative analysis of transcriptional regulatory elements of glutamate-dependent acid-resistance systems of *Shigella flexneri* and *Escherichia coli* O157: H7. *FEMS microbiology letters* **234**(1), 139–147.

- Blanco, M., Blanco, J., Mora, A., Rey, J., Alonso, J., Hermoso, M., Hermoso, J., Alonso, M., Dahbi, G., González, E. (2003). Serotypes, virulence genes, and intimin types of Shiga toxin (verotoxin)-producing *Escherichia coli* isolates from healthy sheep in Spain. *Journal of clinical microbiology* **41**(4), 1351–1356.
- Blount, Z. D. (2015). The natural history of model organisms: The unexhausted potential of *E. coli*. *eLife* **4**, e05826.
- Bobrovskyy, M., Vanderpool, C. K. (2013). Regulation of bacterial metabolism by small RNAs using diverse mechanisms. *Annual review of genetics* **47**, 209–232.
- Bobrovskyy, M., Vanderpool, C. K. (2014). The small RNA SgrS: roles in metabolism and pathogenesis of enteric bacteria. *Frontiers in cellular and infection microbiology* **4**, 61.
- Boerlin, P., McEwen, S. A., Boerlin-Petzold, F., Wilson, J. B., Johnson, R. P., Gyles, C. L. (1999). Associations between virulence factors of Shiga toxin-producing *Escherichia coli* and disease in humans. *Journal of clinical microbiology* **37**(3), 497–503.
- Bogomolnaya, L. M., Aldrich, L., Ragoza, Y., Talamantes, M., Andrews, K. D., McClelland, M., Andrews-Polymenis, H. L. (2014). Identification of novel factors involved in modulating motility of *Salmonella enterica* serotype typhimurium. *PLoS ONE* **9**(11), e111513.
- Bolger, A. M., Lohse, M., Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15), 2114–2120.
- Bolognesi, B., Lehner, B. (2018). Protein Overexpression: Reaching the limit. *eLife* **7**, e39804.
- Bouwmeester, T., Bauch, A., Ruffner, H., Angrand, P.-O., Bergamini, G., Croughton, K., Cruciat, C., Eberhard, D., Gagneur, J., Ghidelli, S. (2004). A physical and functional map of the human TNF- α /NF- κ B signal transduction pathway. *Nature cell biology* **6**(2), 97.
- 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 biology* **5**(9), R72.
- Braeye, T., Denayer, S., De Rauw, K., Forier, A., Verluyten, J., Fourie, L., Dierick, K., Botteldoorn, N., Quoilin, S., Cosse, P. (2014). Lessons learned from a textbook outbreak: EHEC-O157:H7 infections associated with the consumption of raw meat products, June 2012, Limburg, Belgium. *Archives of public health* **72**(1), 44.

- Brandes, N., Linial, M. (2016). Gene overlapping and size constraints in the viral world. *Biology direct* **11**(1), 26.
- Brochado, A. R., Typas, A. (2013). High-throughput approaches to understanding gene function and mapping network architecture in bacteria. *Current opinion in microbiology* **16**(2), 199–206.
- Brock, J. E., Pourshahian, S., Giliberti, J., Limbach, P. A., Janssen, G. R. (2008). Ribosomes bind leaderless mRNA in *Escherichia coli* through recognition of their 5'-terminal AUG. *Rna* **14**(10), 2159–2169.
- Browning, D. F., Busby, S. J. (2004). The regulation of bacterial transcription initiation. *Nature Reviews Microbiology* **2**(1), 57.
- Burge, C. B., Karlin, S. (1998). Finding the genes in genomic DNA. *Current opinion in structural biology* **8**(3), 346–354.
- Burr, T., Mitchell, J., Kolb, A., Minchin, S., Busby, S. (2000). DNA sequence elements located immediately upstream of the –10 hexamer in *Escherichia coli* promoters: a systematic study. *Nucleic acids research* **28**(9), 1864–1870.
- Butland, G., Babu, M., Díaz-Mejía, J. J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A. G., Pogoutse, O. (2008). eSGA: *E. coli* synthetic genetic array analysis. *Nature methods* **5**(9), 789.
- Capt, C., Passamonti, M., Breton, S. (2016). The human mitochondrial genome may code for more than 13 proteins. *Mitochondrial DNA Part A* **27**(5), 3098–3101.
- Carey, C. M., Kostrzynska, M., Thompson, S. (2009). *Escherichia coli* O157:H7 stress and virulence gene expression on Romaine lettuce using comparative real-time PCR. *Journal of microbiological methods* **77**(2), 235–242.
- Carson, M., Johnson, D. H., McDonald, H., Brouillette, C., DeLucas, L. J. (2007). His-tag impact on structure. *Acta Crystallographica Section D: Biological Crystallography* **63**(3), 295–301.
- Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charlotteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B. (2012). Proto-genes and *de novo* gene birth. *Nature* **487**(7407), 370–374.

- Celesnik, H., Deana, A., Belasco, J. G. (2007). Initiation of RNA decay in *Escherichia coli* by 5' pyrophosphate removal. *Molecular cell* **27**(1), 79–90.
- Chen, W.-H., Trachana, K., Lercher, M. J., Bork, P. (2012). Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Molecular biology and evolution* **29**(7), 1703–1706.
- Chirico, N., Vianelli, A., Belshaw, R. (2010). Why genes overlap in viruses. *Proceedings of the Royal Society B: Biological Sciences* **277**(1701), 3809–3817.
- Conway, T., Creecy, J. P., Maddox, S. M., Grissom, J. E., Conkle, T. L., Shadid, T. M., Teramoto, J., San Miguel, P., Shimada, T., Ishihama, A. (2014). Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. *MBio* **5**(4), e01442–14.
- Crooks, G. E., Hon, G., Chandonia, J.-M., Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome research* **14**(6), 1188–1190.
- Croxen, M. A., Finlay, B. B. (2010). Molecular mechanisms of *Escherichia coli* pathogenicity. *Nature Reviews Microbiology* **8**(1), 26.
- 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*. *Clinical microbiology reviews* **26**(4), 822–880.
- Dar, D., Shamir, M., Mellin, J., Koutero, M., Stern-Ginossar, N., Cossart, P., Sorek, R. (2016). Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science* **352**(6282), aad9822.
- Datsenko, K. A., Wanner, B. L. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proceedings of the National Academy of Sciences* **97**(12), 6640–6645.
- Davis, M. C., Kesthely, C. A., Franklin, E. A., MacLellan, S. R. (2016). The essential activities of the bacterial sigma factor. *Canadian journal of microbiology* **63**(2), 89–99.
- Deana, A., Celesnik, H., Belasco, J. G. (2008). The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. *Nature* **451**(7176), 355.

- DebRoy, C., Fratamico, P. M., Yan, X., Baranzoni, G., Liu, Y., Needleman, D. S., Tebbs, R., O'Connell, C. D., Allred, A., Swimley, M. (2016). Comparison of O-antigen gene clusters of all O-serogroups of *Escherichia coli* and proposal for adopting a new nomenclature for O-typing. *PLoS ONE* **11**(1), e0147434.
- Delaye, L., DeLuna, A., Lazcano, A., Becerra, A. (2008). The origin of a novel gene through overprinting in *Escherichia coli*. *BMC Evolutionary Biology* **8**(1), 31.
- Delcher, A. L., Bratke, K. A., Powers, E. C., Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**(6), 673–679.
- Deutschbauer, A., Price, M. N., Wetmore, K. M., Tarjan, D. R., Xu, Z., Shao, W., Leon, D., Arkin, A. P., Skerker, J. M. (2014). Towards an informative mutant phenotype for every bacterial gene. *Journal of bacteriology* **196**(20), 3643–3655.
- DeVinney, R., Stein, M., Reinscheid, D., Abe, A., Ruschkowski, S., Finlay, B. B. (1999). Enterohemorrhagic *Escherichia coli* O157:H7 produces Tir, which is translocated to the host cell membrane but is not tyrosine phosphorylated. *Infection and immunity* **67**(5), 2389–2398.
- Di Martino, M. L., Romilly, C., Wagner, E. G. H., Colonna, B., Prosseda, G. (2016). One gene and two proteins: a leaderless mRNA supports the translation of a shorter form of the *Shigella* VirF regulator. *MBio* **7**(6), e01860–16.
- Dingwall, C., Lomonosoff, G. P., Laskey, R. A. (1981). High sequence specificity of micrococcal nuclease. *Nucleic acids research* **9**(12), 2659–2674.
- Dini, C., Bolla, P. A., Urraza, P. J. de (2016). Treatment of in vitro enterohemorrhagic *Escherichia coli* infection using phage and probiotics. *Journal of applied microbiology* **121**(1), 78–88.
- Dittmar, K., Liberles, D. (2011). *Evolution after gene duplication*. John Wiley & Sons.
- Doi, N., Kakukawa, K., Oishi, Y., Yanagawa, H. (2005). High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. *Protein Engineering Design and Selection* **18**(6), 279–284.
- Dolnik, V., Gurske, W. A. (2011). Chemical modification of proteins to improve the accuracy of their relative molecular mass determination by electrophoresis. *Electrophoresis* **32**(20), 2893–2897.

- Dong, H., Nilsson, L., Kurland, C. G. (1995). Gratuitous overexpression of genes in *Escherichia coli* leads to growth inhibition and ribosome destruction. *Journal of bacteriology* **177**(6), 1497–1504.
- Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H., Spillane, C. (2011). Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evolutionary Biology* **11**(1), 47.
- Dornenburg, J. E., DeVita, A. M., Palumbo, M. J., Wade, J. T. (2010). Widespread antisense transcription in *Escherichia coli*. *MBio* **1**(1), e00024–10.
- Dugar, G., Herbig, A., Förstner, K. U., Heidrich, N., Reinhardt, R., Nieselt, K., Sharma, C. M. (2013). High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. *PLoS genetics* **9**(5), e1003495.
- Dujon, B. (1996). The yeast genome project: what did we learn? *Trends in Genetics* **12**(7), 263–270.
- Dunker, A., Rueckert, R. R. (1969). Observations on molecular weight determinations on polyacrylamide gel. *Journal of Biological Chemistry* **244**(18), 5074–5080.
- Dyson, H. J., Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nature reviews Molecular cell biology* **6**(3), 197.
- Eguchi, Y., Makanae, K., Hasunuma, T., Ishibashi, Y., Kito, K., Moriya, H. (2018). Estimating the protein burden limit of yeast cells by measuring the expression limits of glycolytic proteins. *eLife* **7**, e34595.
- Endo, Y., Tsurugi, K., Yutsudo, T., Takeda, Y., Ogasawara, T., Igarashi, K. (1988). Site of action of a Vero toxin (VT2) from *Escherichia coli* O157:H7 and of Shiga toxin on eukaryotic ribosomes: RNA N-glycosidase activity of the toxins. *European Journal of Biochemistry* **171**(1-2), 45–50.
- Escherich, T. (1886). *Die darmbakterien des Säuglings und ihre beziehungen zur physiologie der Verdauung*. F. Enke.
- Estrem, S. T., Gaal, T., Ross, W., Gourse, R. L. (1998). Identification of an UP element consensus sequence for bacterial promoters. *Proceedings of the National Academy of Sciences* **95**(17), 9761–9766.

- Ettwiller, L., Buswell, J., Yigit, E., Schildkraut, I. (2016). A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. *BMC Genomics* **17**(1), 199.
- Evfratov, S. A., Osterman, I. A., Komarova, E. S., Pogorelskaya, A. M., Rubtsova, M. P., Zatsepin, T. S., Semashko, T. A., Kostryukova, E. S., Mironov, A. A., Burnaev, E. (2016). Application of sorting and next generation sequencing to study 5'-UTR influence on translation efficiency in *Escherichia coli*. *Nucleic acids research* **45**(6), 3487–3502.
- Fairbrother, J., Nadeau, E. (2006). *Escherichia coli*: on-farm contamination of animals. *Rev Sci Tech* **25**(2), 555–69.
- Fellner, L. (2014). 'Functional characterization of overlapping genes in the food-borne pathogen *Escherichia coli* O157:H7'. Thesis.
- 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 microbiology letters* **350**(1), 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 Evolutionary Biology* **15**(1), 1.
- Feltens, R., Göbringer, M., Willkomm, D. K., Urlaub, H., Hartmann, R. K. (2003). An unusual mechanism of bacterial gene expression revealed for the RNase P protein of *Thermus* strains. *Proceedings of the National Academy of Sciences* **100**(10), 5724–5729.
- Fisunov, G., Evsyutina, D., Arzamasov, A., Butenko, I., Govorun, V. (2015). Profiling of *Mycoplasma gallisepticum* ribosomes. *Acta Naturae* **7**(4 (27)).
- Fonseca, M. M., Harris, D. J., Posada, D. (2014). Origin and length distribution of unidirectional prokaryotic overlapping genes. *G3: Genes, Genomes, Genetics* **4**(1), 19–27.
- Fukuda, Y., Tomita, M., Washio, T. (1999). Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. *Nucleic acids research* **27**(8), 1847–1853.

- Furukawa, I., Suzuki, M., Masaoka, T., Nakajima, N., Mitani, E., Tasaka, M., Teranishi, H., Matsumoto, Y., Koizumi, M., Ogawa, A. (2018). An outbreak of enterohemorrhagic *Escherichia coli* O157:H7 infection associated with minced meat cutlets in Kanagawa, Japan. *Japanese Journal of Infectious Diseases*, JJID. 2017.495.
- Garbis, S., Lubec, G., Fountoulakis, M. (2005). Limitations of current proteomics technologies. *Journal of Chromatography A* **1077**(1), 1–18.
- Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., Arkin, A. P., Astromoff, A., El Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.-D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Güldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kötter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C.-y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W., Johnston, M. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**(6896), 387–91.
- Goldberg, A. L. (2003). Protein degradation and protection against misfolded or damaged proteins. *Nature* **426**(6968), 895.
- Goldwater, P. N., Bettelheim, K. A. (2012). Treatment of enterohemorrhagic *Escherichia coli* (EHEC) infection and hemolytic uremic syndrome (HUS). *BMC medicine* **10**(1), 12.
- Grassé, P.-P. (1973). L'évolution du vivant, matériaux pour une nouvelle théorie transformiste, Pierre Grassé. *Sciences*.
- Gray, A. N., Koo, B.-M., Shiver, A. L., Peters, J. M., Osadnik, H., Gross, C. A. (2015). High-throughput bacterial functional genomics in the sequencing era. *Current opinion in microbiology* **27**, 86–95.
- Gualerzi, C. O., Pon, C. L. (2015). Initiation of mRNA translation in bacteria: structural and dynamic aspects. *Cellular and Molecular Life Sciences* **72**(22), 4341–4367.

- Guo, M. S., Updegrove, T. B., Gogol, E. B., Shabalina, S. A., Gross, C. A., Storz, G. (2014). MicL, a new σ^E -dependent sRNA, combats envelope stress by repressing synthesis of Lpp, the major outer membrane lipoprotein. *Genes & development* **28**(14), 1620–1634.
- Guttman, A., Nolan, J. (1994). Comparison of the separation of proteins by sodium dodecyl sulfate-slab gel electrophoresis and capillary sodium dodecyl sulfate-gel electrophoresis. *Analytical biochemistry* **221**(2), 285–289.
- Hadwiger, J. A., Wittenberg, C., Richardson, H. E., Barros Lopes, M. de, Reed, S. I. (1989). A family of cyclin homologs that control the G1 phase in yeast. *Proceedings of the National Academy of Sciences* **86**(16), 6255–6259.
- Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R. (2017). RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic acids research* **46**(D1), D851–D860.
- Härdle, W. K., Klinke, S., Rönz, B. (2015). *Introduction to statistics: using interactive MM* Stat elements*. Springer.
- Hawley, D. K., McClure, W. R. (1983). Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic acids research* **11**(8), 2237–2255.
- Haycocks, J. R., Grainger, D. C. (2016). Unusually Situated Binding Sites for Bacterial Transcription Factors Can Have Hidden Functionality. *PLoS ONE* **11**(6), e0157016.
- Hebert, A. S., Richards, A. L., Bailey, D. J., Ulbrich, A., Coughlin, E. E., Westphall, M. S., Coon, J. J. (2014). The one hour yeast proteome. *Molecular & Cellular Proteomics* **13**(1), 339–347.
- Hecht, A., Glasgow, J., Jaschke, P. R., Bawazer, L. A., Munson, M. S., Cochran, J. R., Endy, D., Salit, M. (2017). Measurements of translation initiation from all 64 codons in *E. coli*. *Nucleic acids research* **45**(7), 3615–3626.
- 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. *Molecular microbiology* **70**(6), 1487–1501.
- Hensel, M., Shea, J. E., Gleeson, C., Jones, M. D., Dalton, E., Holden, D. W. (1995). Simultaneous identification of bacterial virulence genes by negative selection. *Science* **269**(5222), 400–403.

- Herring, C. D., Glasner, J. D., Blattner, F. R. (2003). Gene replacement without selection: regulated suppression of amber mutations in *Escherichia coli*. *Gene* **311**, 153–163.
- Hershberg, R., Petrov, D. A. (2010). Evidence that mutation is universally biased towards AT in bacteria. *PLoS genetics* **6(9)**, e1001115.
- Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W., Proctor, M., St. Onge, R. P., Tyers, M., Koller, D., Altman, R. B., Davis, R. W., Nislow, C., Giaever, G. (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320(5874)**, 362–365.
- Hong, W., Wu, Y. E., Fu, X., Chang, Z. (2012). Chaperone-dependent mechanisms for acid resistance in enteric bacteria. *Trends in microbiology* **20(7)**, 328–335.
- Housman, G., Ulitsky, I. (2016). Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1859(1)**, 31–40.
- Hsu, P. Y., Calviello, L., Wu, H.-Y. 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* **113(45)**, E7126–E7135.
- Hu, P., Janga, S. C., Babu, M., Díaz-Mejía, J. J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P. (2009). Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS biology* **7(4)**, e1000096.
- Hücker, S. M. (2018). ‘RIBOseq-based discovery of non-annotated genes in *Escherichia coli* O157:H7 Sakai and their functional characterization’. Thesis.
- Hücker, S. M., Ardern, Z., Goldberg, T., Schafferhans, A., Bernhofer, M., Vestergaard, G., Nelson, C. W., Schlöter, M., Rost, B., Scherer, S. (2017). Discovery of numerous novel small genes in the intergenic regions of the *Escherichia coli* O157: H7 Sakai genome. *PLoS ONE* **12(9)**, e0184119.
- Hücker, S. M., Vanderhaeghen, S., Abellan-Schneyder, I., Scherer, S., Neuhaus, K. (2018a). The novel anaerobiosis-responsive overlapping gene *ano* is overlapping antisense to the annotated gene ECs2385 of *Escherichia coli* O157: H7 Sakai. *Frontiers in microbiology* **9**.

- Hücker, S. M., Vanderhaeghen, S., Abellan-Schneyder, I., Wecko, R., Simon, S., Scherer, S., Neuhaus, K. (2018b). A novel short L-arginine responsive protein-coding gene (*laoB*) antiparallel overlapping to a CadC-like transcriptional regulator in *Escherichia coli* O157: H7 Sakai originated by overprinting. *BMC Evolutionary Biology* **18**(1), 21.
- Hücker, S. M., Simon, S., Scherer, S., Neuhaus, K. (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 microbiology letters* **364**(2).
- Hurley, J. M., Cruz, J. W., Ouyang, M., Woychik, N. A. (2011). Bacterial toxin RelE mediates frequent codon-independent mRNA cleavage from the 5' end of coding regions in vivo. *Journal of Biological Chemistry* **286**(17), 14770–14778.
- Hwang, J.-Y., Buskirk, A. R. (2016). A ribosome profiling study of mRNA cleavage by the endonuclease RelE. *Nucleic acids research*, gkw944.
- Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* **11**(1), 119.
- Impens, F., Rolhion, N., Radoshevich, L., Bécavin, C., Duval, M., Mellin, J., Del Portillo, F. G., Pucciarelli, M. G., Williams, A. H., Cossart, P. (2017). N-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in *Listeria monocytogenes*. *Nature microbiology* **2**(5), 17005.
- In, J., Foulke-Abel, J., Zachos, N. C., Hansen, A.-M., Kaper, J. B., Bernstein, H. D., Halushka, M., Blutt, S., Estes, M. K., Donowitz, M. (2016). Enterohemorrhagic *Escherichia coli* reduces mucus and intermicrovillar bridges in human stem cell-derived colonoids. *Cellular and molecular gastroenterology and hepatology* **2**(1), 48–62. e3.
- 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 reports* **8**(5), 1365–1379.
- Ingolia, N. T., Ghaemmighami, S., Newman, J. R., Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**(5924), 218–223.

- Innocenti, N., Golumbeanu, M., d'Hérœuel, A. F., Lacoux, C., Bonnin, R. A., Kennedy, S. P., Wessner, F., Serror, P., Bouloc, P., Repoila, F. (2015). Whole-genome mapping of 5' RNA ends in bacteria by tagged sequencing: a comprehensive view in *Enterococcus faecalis*. *Rna* **21(5)**, 1018–1030.
- Inoue, T., Shingaki, R., Hirose, S., Waki, K., Mori, H., Fukui, K. (2007). Genome-wide screening of genes required for swarming motility in *Escherichia coli* K-12. *Journal of bacteriology* **189(3)**, 950–957.
- Jacob, F. (1977). Evolution and tinkering. *Science* **196(4295)**, 1161–1166.
- Jäger, D., Förstner, K. U., Sharma, C. M., Santangelo, T. J., Reeve, J. N. (2014). Primary transcriptome map of the hyperthermophilic archaeon *Thermococcus kodakarensis*. *BMC Genomics* **15(1)**, 684.
- Jeanmougin, M., De Reynies, A., Marisa, L., Paccard, C., Nuel, G., Guedj, M. (2010). Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PLoS ONE* **5(9)**, e12336.
- 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. (2016). The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3 (2). *Nature communications* **7**, 11605.
- Johnson, B. R. (2018). Taxonomically restricted genes are fundamental to biology and evolution. *Frontiers in genetics* **9**, 407.
- Johnson, J. R., Russo, T. A. (2002). Extraintestinal pathogenic *Escherichia coli*: “the other bad *E coli*”. *Journal of Laboratory and Clinical Medicine* **139(3)**, 155–162.
- Johnson, Z. I., Chisholm, S. W. (2004). Properties of overlapping genes are conserved across microbial genomes. *Genome research* **14(11)**, 2268–72.
- Jong, W. W. de, Zweers, A., Cohen, L. H. (1978). Influence of single amino acid substitutions on electrophoretic mobility of sodium dodecyl sulfate-protein complexes. *Biochemical and biophysical research communications* **82(2)**, 532–539.
- Jordan, I. K., Rogozin, I. B., Wolf, Y. I., Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome research* **12(6)**, 962–968.

- Kaas, R. S., Friis, C., Ussery, D. W., Aarestrup, F. M. (2012). Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics* **13**(1), 577.
- Kafri, M., Metzl-Raz, E., Jona, G., Barkai, N. (2016). The cost of protein production. *Cell reports* **14**(1), 22–31.
- Kaniga, K., Delor, I., Cornelis, G. R. (1991). A wide-host-range suicide vector for improving reverse genetics in gram-negative bacteria: inactivation of the *blaA* gene of *Yersinia enterocolitica*. *Gene* **109**(1), 137–141.
- Kaper, J. B., Nataro, J. P., Mobley, H. L. (2004). Pathogenic *Escherichia coli*. *Nature Reviews microbiology* **2**(2), 123.
- Karpinets, T. V., Greenwood, D. J., Sams, C. E., Ammons, J. T. (2006). RNA: protein ratio of the unicellular organism as a characteristic of phosphorous and nitrogen stoichiometry and of the cellular requirement of ribosomes for protein synthesis. *BMC biology* **4**(1), 30.
- Kawamoto, H., Koide, Y., Morita, T., Aiba, H. (2006). Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. *Molecular microbiology* **61**(4), 1013–1022.
- Keese, P. K., Gibbs, A. (1992). Origins of genes: "big bang" or continuous creation? *Proceedings of the National Academy of Sciences* **89**(20), 9489–93.
- Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., Bosch, T. C. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics* **25**(9), 404–413.
- 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 biotechnology* **14**(1), 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**(12), e8455.
- Kintaka, R., Makanae, K., Moriya, H. (2016). Cellular growth defects triggered by an overload of protein localization processes. *Scientific reports* **6**, 31774.

- Knutton, S., Baldwin, T., Williams, P., McNeish, A. (1989). Actin accumulation at sites of bacterial adhesion to tissue culture cells: basis of a new diagnostic test for enteropathogenic and enterohemorrhagic *Escherichia coli*. *Infection and immunity* **57**(4), 1290–1298.
- Kortmann, J., Narberhaus, F. (2012). Bacterial RNA thermometers: molecular zippers and switches. *Nature Reviews Microbiology* **10**(4), 255.
- Koyanagi, Y., Suzuki, R., Ihara, K., Miyagi, H., Isogai, H., Yoneyama, H., Isogai, E. (2019). Intestinal *Clostridium* species lower host susceptibility to enterohemorrhagic *Escherichia coli* O157:H7 infection. *Pathogens and disease* **77**(4), ftz036.
- Krakauer, D. C. (2000). Stability and evolution of overlapping genes. *Evolution* **54**(3), 731–739.
- Kröger, C., Colgan, A., Srikumar, S., Händler, K., Sivasankaran, S. K., Hammarlöf, D. L., Canals, R., Grissom, J. E., Conway, T., Hokamp, K. (2013). An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. *Cell host & microbe* **14**(6), 683–695.
- Kurata, T., Katayama, A., Hiramatsu, M., Kiguchi, Y., Takeuchi, M., Watanabe, T., Ogasawara, H., Ishihama, A., Yamamoto, K. (2013). Identification of the set of genes, including nonannotated *morA*, under the direct control of ModE in *Escherichia coli*. *Journal of bacteriology* **195**(19), 4496–505.
- Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *nature* **227**(5259), 680.
- Laible, M., Boonrod, K. (2009). Homemade site directed mutagenesis of whole plasmids. *Journal of visualized experiments: JoVE*(**27**).
- Landstorfer, R. (2014). ‘Comparative transcriptomics and translatomics to identify novel overlapping genes, active hypothetical genes, and ncRNAs in *Escherichia coli* O157:H7 EDL933’. Thesis.
- 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.

- Lange, R., Hengge-Aronis, R. (1991). Identification of a central regulator of stationary-phase gene expression in *Escherichia coli*. *Molecular microbiology* **5**(1), 49–59.
- Langmead, B., Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**(4), 357.
- Larsson, T., Bergström, J., Nilsson, C., Karlsson, K.-A. (2000). Use of an affinity proteomics approach for the identification of low-abundant bacterial adhesins as applied on the Lewisb-binding adhesin of *Helicobacter pylori*. *FEBS letters* **469**(2-3), 155–158.
- Latif, H., Li, H. J., Charusanti, P., Palsson, B. Ø., Aziz, R. K. (2014). A gapless, unambiguous genome sequence of the enterohemorrhagic *Escherichia coli* O157: H7 strain EDL933. *Genome Announcements* **2**(4), e00821–14.
- Lèbre, S., Gascuel, O. (2017). The combinatorics of overlapping genes. *Journal of theoretical biology* **415**, 90–101.
- Lim, J. Y., Yoon, J. W., Hovde, C. J. (2010). A brief overview of *Escherichia coli* O157: H7 and its plasmid O157. *Journal of microbiology and biotechnology* **20**(1), 5.
- Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H., Bork, P., Glass, J. I., Serrano, L., Lluch-Senar, M. (2016). Bacterial antisense RNAs are mainly the product of transcriptional noise. *Science advances* **2**(3), e1501363.
- Lomsadze, A., Gemayel, K., Tang, S., Borodovsky, M. (2018). Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome research* **28**(7), 1079–1089.
- López-Garrido, J., Puerta-Fernández, E., Casadesús, J. (2014). A eukaryotic-like 3' untranslated region in *Salmonella enterica hilD* mRNA. *Nucleic acids research* **42**(9), 5894–5906.
- Lukjancenko, O., Wassenaar, T. M., Ussery, D. W. (2010). Comparison of 61 sequenced *Escherichia coli* genomes. *Microbial ecology* **60**(4), 708–720.
- Lynch, M., Marinov, G. K. (2015). The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences* **112**(51), 15690–15695.
- Ma, J., Campbell, A., Karlin, S. (2002). Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *Journal of bacteriology* **184**(20), 5733–5745.

- Makalowska, I., Lin, C.-F., Makalowski, W. (2005). Overlapping genes in vertebrate genomes. *Computational biology and chemistry* **29**(1), 1–12.
- Mamat, U., Wilke, K., Bramhill, D., Schromm, A. B., Lindner, B., Kohl, T. A., Corchero, J. L., Villaverde, A., Schaffer, L., Head, S. R. (2015). Detoxifying *Escherichia coli* for endotoxin-free production of recombinant proteins. *Microbial cell factories* **14**(1), 57.
- Manoil, C., Beckwith, J. (1985). Tn*phoA*: a transposon probe for protein export signals. *Proceedings of the National Academy of Sciences* **82**(23), 8129–33.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**(1), 10–12.
- McVeigh, A., Fasano, A., Scott, D. A., Jelacic, S., Moseley, S. L., Robertson, D. C., Savarino, S. J. (2000). IS1414, an *Escherichia coli* insertion sequence with a heat-stable enterotoxin gene embedded in a transposase-like gene. *Infection and immunity* **68**(10), 5710–5715.
- Mejías, M. P., Hiriart, Y., Lauché, C., Fernández-Brando, R. J., Pardo, R., Bruballa, A., Ramos, M. V., Goldbaum, F. A., Palermo, M. S., Zylberman, V. (2016). Development of camelid single chain antibodies against Shiga toxin type 2 (Stx2) with therapeutic potential against Hemolytic Uremic Syndrome (HUS). *Scientific reports* **6**, 24913.
- Mellin, J., Cossart, P. (2015). Unexpected versatility in bacterial riboswitches. *Trends in Genetics* **31**(3), 150–156.
- Melton-Celsa, A. R. (2014). Shiga toxin (Stx) classification, structure, and function. *Microbiology spectrum* **2**(2).
- Mendoza-Vargas, A., Olvera, L., Olvera, M., Grande, R., Vega-Alvarado, L., Taboada, B., Jimenez-Jacinto, V., Salgado, H., Juárez, K., Contreras-Moreira, B., Huerta, A. M., Collado-Vides, J., Morett, E. (2009). Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. *PLoS ONE* **4**(10), e7526.
- Merrihew, G. E., Davis, C., Ewing, B., Williams, G., Käll, L., Frewen, B. E., Noble, W. S., Green, P., Thomas, J. H., MacCoss, M. J. (2008). Use of shotgun proteomics for the identification, confirmation, and correction of *C. elegans* gene annotations. *Genome research* **18**(10), 1660–1669.

- Meydan, S., Marks, J., Klepacki, D., Sharma, V., Baranov, P. V., Firth, A. E., Margus, T., Kefi, A., Vazquez-Laslop, N., Mankin, A. S. (2019). Retapamulin-assisted ribosome profiling reveals the alternative bacterial proteome. *Molecular cell*.
- Michel, A. M., Choudhury, K. R., Firth, A. E., Ingolia, N. T., Atkins, J. F., Baranov, P. V. (2012). Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome research* **22**(11), 2219–2229.
- Michino, H., Araki, K., Minami, S., Takaya, S., Sakai, N., Miyazaki, M., Ono, A., Yanagawa, H. (1999). Massive outbreak of *Escherichia coli* O157:H7 infection in schoolchildren in Sakai City, Japan, associated with consumption of white radish sprouts. *American journal of epidemiology* **150**(8), 787–796.
- Mignone, F., Gissi, C., Liuni, S., Pesole, G. (2002). Untranslated regions of mRNAs. *Genome biology* **3**(3), reviews0004. 1.
- Miller, V. L., Mekalanos, J. J. (1988). A novel suicide vector and its use in construction of insertion mutations: osmoregulation of outer membrane proteins and virulence determinants in *Vibrio cholerae* requires *toxR*. *Journal of bacteriology* **170**(6), 2575–2583.
- Miller, W. G., Leveau, J. H., Lindow, S. E. (2000). Improved *gfp* and *inaZ* broad-host-range promoter-probe vectors. *Molecular Plant-Microbe Interactions* **13**(11), 1243–1250.
- Mir, K., Neuhaus, K., Scherer, S., Bossert, M., Schober, S. (2012). Predicting statistical properties of open reading frames in bacterial genomes. *PLoS ONE* **7**(9), e45103.
- Mir, K., Schober, S. (2014). Selection pressure in alternative reading frames. *PLoS ONE* **9**(10), e108768.
- Missiakas, D., Georgopoulos, C., Raina, S. (1993). The *Escherichia coli* heat shock gene *htpY*: mutational analysis, cloning, sequencing, and transcriptional regulation. *Journal of bacteriology* **175**(9), 2613–2624.
- Mitchell, J. E., Zheng, D., Busby, S. J., Minchin, S. D. (2003). Identification and analysis of ‘extended-10’ promoters in *Escherichia coli*. *Nucleic acids research* **31**(16), 4689–4695.
- Moll, I., Grill, S., Gualerzi, C. O., Bläsi, U. (2002). Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Molecular microbiology* **43**(1), 239–246.

- Moriya, H. (2015). Quantitative nature of overexpression experiments. *Molecular biology of the cell* **26**(22), 3932–3939.
- Moshensky, D., Alexeevski, A. (2019). Long antiparallel open reading frames are unlikely to be encoding essential proteins in prokaryotic genomes. *BioRxiv*, 724807.
- Mouilleron, H., Delcourt, V., Roucou, X. (2015). Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic acids research* **44**(1), 14–23.
- Mutalik, V. K., Novichkov, P. S., Price, M. N., Owens, T. K., Callaghan, M., Carim, S., Deutschbauer, A. M., Arkin, A. P. (2019). Dual-barcoded shotgun expression library sequencing for high-throughput characterization of functional traits in bacteria. *Nature communications* **10**(1), 308.
- 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 Research* **23**(3), 193–201.
- Neme, R., Amador, C., Yildirim, B., McConnell, E., Tautz, D. (2017). Random sequences are an abundant source of bioactive RNAs or peptides. *Nature ecology & evolution* **1**(6), 0127.
- 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* **17**, 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**(1), 216.
- Nguyen, Y., Sperandio, V. (2012). Enterohemorrhagic *E. coli* (EHEC) pathogenesis. *Frontiers in Cellular and Infection Microbiology* **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**(1), 143–56.

- 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 & development* **20**(13), 1776–1789.
- Normark, S., Bergström, S., Edlund, T., Grundström, T., Jaurin, B., Lindberg, F. P., Olsson, O. (1983). Overlapping genes. *Annual review of genetics* **17**(1), 499–525.
- Obrig, T. G. (2010). *Escherichia coli* Shiga toxin mechanisms of action in renal disease. *Toxins* **2**(12), 2769–2794.
- Ochman, H., Lawrence, J. G., Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**(6784), 299.
- Oh, E., Becker, A. H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R. J., Typas, A., Gross, C. A., Kramer, G. (2011). Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* **147**(6), 1295–1308.
- Oppenheim, D. S., Yanofsky, C. (1980). Translational coupling during expression of the tryptophan operon of *Escherichia coli*. *Genetics* **95**(4), 785–795.
- Opuu, V., Silvert, M., Simonson, T. (2017). Computational design of fully overlapping coding schemes for protein pairs and triplets. *Scientific reports* **7**(1), 15873.
- Ortiz-Suarez, M. L., Samsudin, F., Piggot, T. J., Bond, P. J., Khalid, S. (2016). Full-length OmpA: structure, function, and membrane interactions predicted by molecular dynamics simulations. *Biophysical journal* **111**(8), 1692–1702.
- Paget, M. S., Helmann, J. D. (2003). The σ^{70} family of sigma factors. *Genome biology* **4**(1), 203.
- Pallejà, A., Harrington, E. D., Bork, P. (2008). Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics* **9**(1), 335.
- Papenfort, K., Förstner, K. U., Cong, J.-P., Sharma, C. M., Bassler, B. L. (2015). Differential RNA-seq of *Vibrio cholerae* identifies the VqmR small RNA as a regulator of biofilm formation. *Proceedings of the National Academy of Sciences* **112**(7), E766–E775.
- Paradis-Bleau, C., Kritikos, G., Orlova, K., Typas, A., Bernhardt, T. G. (2014). A genome-wide screen for bacterial envelope biogenesis mutants identifies a novel factor involved in cell wall precursor metabolism. *PLoS genetics* **10**(1), e1004056.

- 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 computational biology* **9**(8), e1003162.
- Pavesi, A., Vianelli, A., Chirico, N., Bao, Y., Blinkova, O., Belshaw, R., Firth, A., Karlin, D. (2018). Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLoS ONE* **13**(10), e0202513.
- Peck, R., Olsen, C., Devore, J. L. (2015). *Introduction to statistics and data analysis*. Cengage Learning Custom Publishing.
- Pedersen, K., Zavialov, A. V., Pavlov, M. Y., Elf, J., Gerdes, K., Ehrenberg, M. (2003). The bacterial toxin RelE displays codon-specific cleavage of mRNAs in the ribosomal A site. *Cell* **112**(1), 131–140.
- Perna, N. T., Plunkett III, G., Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A. (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**(6819), 529.
- Peters, J. M., Colavin, A., Shi, H., Czarny, T. L., Larson, M. H., Wong, S., Hawkins, J. S., Lu, C. H., Koo, B.-M., Marta, E. (2016). A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. *Cell* **165**(6), 1493–1506.
- Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic acids research* **29**(9), e45–e45.
- Pircher, A., Gebetsberger, J., Polacek, N. (2014). Ribosome-associated ncRNAs: An emerging class of translation regulators. *RNA biology* **11**(11), 1335–1339.
- Poole, E., Tate, W. (2000). Release factors and their role as decoding proteins: specificity and fidelity for termination of protein synthesis. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression* **1493**(1-2), 1–11.
- Prelich, G. (2012). Gene overexpression: uses, mechanisms, and interpretation. *Genetics* **190**(3), 841–854.
- Prijambada, I. D., Yomo, T., Tanaka, F., Kawama, T., Yamamoto, K., Hasegawa, A., Shima, Y., Negoro, S., Urabe, I. (1996). Solubility of artificial proteins with random sequences. *FEBS letters* **382**(1-2), 21–25.

- Pruimboom-Brees, I. M., Morgan, T. W., Ackermann, M. R., Nystrom, E. D., Samuel, J. E., Cornick, N. A., Moon, H. W. (2000). Cattle lack vascular receptors for *Escherichia coli* O157: H7 Shiga toxins. *Proceedings of the National Academy of Sciences* **97**(19), 10325–10329.
- Raabe, C. A., Tang, T.-H., Brosius, J., Rozhdestvensky, T. S. (2013). Biases in small RNA deep sequencing data. *Nucleic acids research* **42**(3), 1414–1426.
- Raghavan, R., Sloan, D. B., Ochman, H. (2012). Antisense transcription is pervasive but rarely conserved in enteric bacteria. *MBio* **3**(4), e00156–12.
- Rajagopala, S. V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., Franca-Koh, J., Pakala, S. B., Phanse, S., Ceol, A. (2014). The binary protein-protein interaction landscape of *Escherichia coli*. *Nature biotechnology* **32**(3), 285.
- Ray-Soni, A., Bellecourt, M. J., Landick, R. (2016). Mechanisms of bacterial transcription termination: all good things must end. *Annual review of biochemistry* **85**, 319–347.
- Reid, S. D., Herbelin, C. J., Bumbaugh, A. C., Selander, R. K., Whittam, T. S. (2000). Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406**(6791), 64–7.
- Ren, G.-X., Guo, X.-P., Sun, Y.-C. (2017). Regulatory 3' untranslated regions of bacterial mRNAs. *Frontiers in microbiology* **8**, 1276.
- Reyrat, J.-M., Pelicic, V., Gicquel, B., Rappuoli, R. (1998). Counters selectable markers: untapped tools for bacterial genetics and pathogenesis. *Infection and immunity* **66**(9), 4011–4017.
- Riley, L. W., Remis, R. S., Helgerson, S. D., McGee, H. B., Wells, J. G., Davis, B. R., Hebert, R. J., Olcott, E. S., Johnson, L. M., Hargrett, N. T. (1983). Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. *New England Journal of Medicine* **308**(12), 681–685.
- Rogozin, I. B., Spiridonov, A. N., Sorokin, A. V., Wolf, Y. I., Jordan, I. K., Tatusov, R. L., Koonin, E. V. (2002). Purifying and directional selection in overlapping prokaryotic genes. *Trends in Genetics* **18**(5), 228–232.
- Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A., Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* **16**(10), 944–945.

- Saha, D., Panda, A., Podder, S., Ghosh, T. C. (2015). Overlapping genes: a new strategy of thermophilic stress tolerance in prokaryotes. *Extremophiles* **19**(2), 345–353.
- Saha, D., Podder, S., Panda, A., Ghosh, T. C. (2016). Overlapping genes: a significant genomic correlate of prokaryotic growth rates. *Gene* **582**(2), 143–147.
- Saito, K., Suzuki, R., Koyanagi, Y., Isogai, H., Yoneyama, H., Isogai, E. (2019). Inhibition of enterohemorrhagic *Escherichia coli* O157:H7 infection in a gnotobiotic mouse model with pre-colonization by *Bacteroides* strains. *Biomedical reports* **10**(3), 175–182.
- Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muñiz-Rascado, L., García-Sotelo, J. S., Weiss, V., Solano-Lira, H., Martínez-Flores, I., Medina-Rivera, A. (2012). RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic acids research* **41**(D1), D203–D213.
- Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., García-Sotelo, J. S., Alquicira-Hernández, K., Muñiz-Rascado, L. J., Peña-Loredo, P. (2018). RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic acids research* **47**(D1), D212–D220.
- Sarker, M. R., Cornelis, G. R. (1997). An improved version of suicide vector pKNG101 for gene replacement in Gram-negative bacteria. *Molecular microbiology* **23**(2), 410–411.
- Satoshi, F., Nishikawa, K. (2004). Estimation of the number of authentic orphan genes in bacterial genomes. *DNA Research* **11**(4), 219–231.
- Sberro, H., Fremin, B. J., Zlitni, S., Edfors, F., Greenfield, N., Snyder, M. P., Pavlopoulos, G. A., Kyrpides, N. C., Bhatt, A. S. (2019). Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell* **178**(5), 1245–1259. e14.
- Schägger, H. (2006). Tricine-sds-page. *Nature protocols* **1**(1), 16.
- Schägger, H., Von Jagow, G. (1987). Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Analytical biochemistry* **166**(2), 368–379.
- Scherbakov, D. V., Garber, M. B. (2000). Overlapping genes in bacterial and phage genomes. *Molecular Biology* **34**(4), 485–495.

- Schlötterer, C. (2015). Genes from scratch—the evolutionary fate of *de novo* genes. *Trends in Genetics* **31**(4), 215–219.
- Schüller, S. (2011). Shiga toxin interaction with human intestinal epithelium. *Toxins* **3**(6), 626–639.
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**(14), 2068–2069.
- Semenov, A. M., Kuprianov, A. A., Van Bruggen, A. H. (2010). Transfer of enteric pathogens to successive habitats as part of microbial cycles. *Microbial ecology* **60**(1), 239–249.
- Shachrai, I., Zaslaver, A., Alon, U., Dekel, E. (2010). Cost of unneeded proteins in *E. coli* is reduced after several generations in exponential growth. *Molecular cell* **38**(5), 758–767.
- Shahmuradov, I. A., Mohamad Razali, R., Bougouffa, S., Radovanovic, A., Bajic, V. B. (2017). bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and *Escherichia coli*. *Bioinformatics* **33**(3), 334–340.
- Shao, W., Price, M. N., Deutschbauer, A. M., Romine, M. F., Arkin, A. P. (2014). Conservation of transcription start sites within genes across a bacterial genus. *MBio* **5**(4), e01398–14.
- Shapiro, A. L., Viñuela, E., Maizel Jr, J. V. (1967). Molecular weight estimation of polypeptide chains by electrophoresis in SDS-polyacrylamide gels. *Biochemical and biophysical research communications* **28**(5), 815–820.
- Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiß, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R. (2010). The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**(7286), 250.
- Sharma, C. M., Vogel, J. (2014). Differential RNA-seq: the approach behind and the biological insight gained. *Current opinion in microbiology* **19**, 97–105.
- Shi, T., Song, E., Nie, S., Rodland, K. D., Liu, T., Qian, W., Smith, R. D. (2016). Advances in targeted proteomics and applications to biomedical research. *Proteomics* **16**(15-16), 2160–2182.
- Shomar, H., Gontier, S., Broek, N. J. van den, Mora, H. T., Noga, M. J., Hagedoorn, P.-L., Bokinsky, G. (2018). Metabolic engineering of a carbapenem antibiotic synthesis pathway in *Escherichia coli*. *Nature chemical biology*, 1.

- Simon, R., Prierer, U., Pühler, A. (1983). A broad host range mobilization system for in vivo genetic engineering: transposon mutagenesis in gram negative bacteria. *Bio/technology* **1(9)**, 784.
- Singh, S. S., Typas, A., Hengge, R., Grainger, D. C. (2011). *Escherichia coli* σ^{70} senses sequence and conformation of the promoter spacer region. *Nucleic acids research* **39(12)**, 5109–5118.
- Slonczewski, J. L., Foster, J. W. (2013). *Microbiology: An Evolving Science: Third International Student Edition*. WW Norton & Company.
- Smith, C., Canestrari, J., Wang, J., Derbyshire, K., Gray, T., Wade, J. (2019). Pervasive Translation in *Mycobacterium tuberculosis*. *bioRxiv*, 665208.
- Solovyev, V., Salamov, A. (2011). Automatic annotation of microbial genomes and metagenomic sequences. *Metagenomics and its applications in agriculture*. Nova Science Publishers, Hauppauge, 61–78.
- Soo, V. W., Hanson-Manful, P., Patrick, W. M. (2011). Artificial gene amplification reveals an abundance of promiscuous resistance determinants in *Escherichia coli*. *Proceedings of the National Academy of Sciences* **108(4)**, 1484–1489.
- Sperandio, V., Nguyen, Y. (2012). Enterohemorrhagic *E. coli* (EHEC) pathogenesis. *Frontiers in Cellular and Infection Microbiology* **2**, 90.
- Stephens, S. (1951). ‘Possible significance of duplication in evolution’. *Advances in genetics*. **4**. Elsevier, 247–265.
- Stevens, M. P., Frankel, G. M. (2014). The Locus of Enterocyte Effacement and Associated Virulence Factors of Enterohemorrhagic *Escherichia coli*. *Microbiology spectrum* **2(4)**, 131–155.
- Stoebel, D. M., Dean, A. M., Dykhuizen, D. E. (2008). The cost of expression of *Escherichia coli lac operon* proteins is in the process, not in the products. *Genetics* **178(3)**, 1653–1660.
- Storz, G., Vogel, J., Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Molecular cell* **43(6)**, 880–891.
- Sussman, J. K., Simons, E. L., Simons, R. W. (1996). *Escherichia coli* translation initiation factor 3 discriminates the initiation codon in vivo. *Molecular microbiology* **21(2)**, 347–360.

- Sycuro, L. K., Rule, C. S., Petersen, T. W., Wyckoff, T. J., Sessler, T., Nagarkar, D. B., Khalid, F., Pincus, Z., Biboy, J., Vollmer, W. (2013). Flow cytometry-based enrichment for cell shape mutants identifies multiple genes that influence *Helicobacter pylori* morphology. *Molecular microbiology* **90**(4), 869–883.
- Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic acids research* **44**(14), 6614–6624.
- Taur, Y., Pamer, E. G. (2013). The intestinal microbiota and susceptibility to infection in immunocompromised patients. *Current opinion in infectious diseases* **26**(4), 332.
- Tautz, D. (2014). The discovery of de novo gene evolution. *Perspectives in biology and medicine* **57**(1), 149–161.
- Ten-Caten, F., Vêncio, R. Z., Lorenzetti, A. P. R., Zaramela, L. S., Santana, A. C., Koide, T. (2018). Internal RNAs overlapping coding sequences can drive the production of alternative proteins in archaea. *RNA biology* **15**(8), 1119–1132.
- Thomason, M. K., Bischler, T., Eisenbart, S. K., Förstner, K. U., Zhang, A., Herbig, A., Nieselt, K., Sharma, C. M., Storz, G. (2015). Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli*. *Journal of bacteriology* **197**(1), 18–28.
- Thomason, M. K., Storz, G. (2010). Bacterial antisense RNAs: how many are there, and what are they doing? *Annual review of genetics* **44**, 167–188.
- Tjaden, B., Saxena, R. M., Stolyar, S., Haynor, D. R., Kolker, E., Rosenow, C. (2002). Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic acids research* **30**(17), 3732–3738.
- Tunca, S., Barreiro, C., Coque, J. R., Martín, J. F. (2009). Two overlapping antiparallel genes encoding the iron regulator DmdR1 and the Adm proteins control sidephore and antibiotic biosynthesis in *Streptomyces coelicolor* A3 (2). *The FEBS journal* **276**(17), 4814–4827.
- Tuttle, J., Gomez, T., Doyle, M., Wells, J., Zhao, T., Tauxe, R., Griffin, P. (1999). Lessons from a large outbreak of *Escherichia coli* O157:H7 infections: insights into the infectious dose and method of widespread contamination of hamburger patties. *Epidemiology & Infection* **122**(2), 185–192.

- Typas, A., Nichols, R. J., Siegele, D. A., Shales, M., Collins, S. R., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Wanner, B. L. (2008). High-throughput, quantitative analyses of genetic interactions in *E. coli*. *Nature methods* **5**(9), 781.
- Typas, A., Sourjik, V. (2015). Bacterial protein networks: properties and functions. *Nature Reviews Microbiology* **13**(9), 559.
- Van Opijnen, T., Bodi, K. L., Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature methods* **6**(10), 767.
- Vanderhaeghen, S., Zehentner, B., Scherer, S., Neuhaus, K., Ardern, Z. (2018). The novel EHEC gene *asa* overlaps the TEGT transporter gene in antisense and is regulated by NaCl and growth phase. *Scientific reports* **8**(1), 17875.
- Verster, A. J., Styles, E. B., Mateo, A., Derry, W. B., Andrews, B. J., Fraser, A. G. (2017). Taxonomically restricted genes with essential functions frequently play roles in chromosome segregation in *Caenorhabditis elegans* and *Saccharomyces cerevisiae*. *G3: Genes, Genomes, Genetics* **7**(10), 3337–3347.
- Wade, J. T. (2015). Where to begin? Mapping transcription start sites genome-wide in *Escherichia coli*. *Journal of bacteriology* **197**(1), 4–6.
- Wade, J. T., Grainger, D. C. (2014). Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nature Reviews Microbiology* **12**(9), 647.
- Wagner, P. L., Livny, J., Neely, M. N., Acheson, D. W., Friedman, D. I., Waldor, M. K. (2002). Bacteriophage control of Shiga toxin 1 production and release by *Escherichia coli*. *Molecular microbiology* **44**(4), 957–970.
- Warren, A. S., Archuleta, J., Feng, W.-c., Setubal, J. C. (2010). Missing genes in the annotation of prokaryotic genomes. *BMC bioinformatics* **11**(1), 131.
- Waters, L. S., Sandoval, M., Storz, G. (2011). The *Escherichia coli* MntR mini regulon includes genes encoding a small protein and an efflux pump required for manganese homeostasis. *Journal of bacteriology*, 5887–5897.
- Waters, L. S., Storz, G. (2009). Regulatory RNAs in bacteria. *Cell* **136**(4), 615–628.
- Waugh, D. S. (2005). Making the most of affinity tags. *Trends in biotechnology* **23**(6), 316–320.

- Weaver, J., Mohammad, F., Buskirk, A. R., Storz, G. (2019). Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes. *MBio* **10**(2), e02819–18.
- Wells, J., Davis, B., Wachsmuth, I., Riley, L., Remis, R. S., Sokolow, R., Morris, G. (1983). Laboratory investigation of hemorrhagic colitis outbreaks associated with a rare *Escherichia coli* serotype. *Journal of clinical microbiology* **18**(3), 512–520.
- Westermann, A. J., Gorski, S. A., Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology* **10**(9), 618.
- Wick, L. M., Qi, W., Lacher, D. W., Whittam, T. S. (2005). Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7. *Journal of Bacteriology* **187**(5), 1783–1791.
- Wilson, B. A., Masel, J. (2011). Putatively noncoding transcripts show extensive association with ribosomes. *Genome biology and evolution* **3**, 1245–52.
- Wilson, G., Bertrand, N., Patel, Y., Hughes, J., Feil, E., Field, D. (2005). Orphans as taxonomically restricted and ecologically important genes. *Microbiology* **151**(8), 2499–2501.
- Wong, C. S., Mooney, J. C., Brandt, J. R., Staples, A. O., Jelacic, S., Boster, D. R., Watkins, S. L., Tarr, P. I. (2012). Risk factors for the hemolytic uremic syndrome in children infected with *Escherichia coli* O157:H7: a multivariable analysis. *Clinical Infectious Diseases* **55**(1), 33–41.
- 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 reports* **11**(1), 13–21.
- Wurtzel, O., Sesto, N., Mellin, J. R., Karunker, I., Edelheit, S., Bécavin, C., Archambaud, C., Cossart, P., Sorek, R. (2012). Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species. *Molecular systems biology* **8**(1).
- Xia, B., Ke, H., Inouye, M. (2001). Acquisition of cold sensitivity by quadruple deletion of the *cspA* family and its suppression by PNPase S1 domain in *Escherichia coli*. *Molecular microbiology* **40**(1), 179–188.
- Yamamoto, H., Fang, M., Dragnea, V., Bauer, C. E. (2018). Differing isoforms of the cobalamin binding photoreceptor AerR oppositely regulate photosystem expression. *eLife* **7**, e39028.

- Yang, L., Zou, M., Fu, B., He, S. (2013). Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. *BMC Genomics* **14**(1), 65.
- Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J. D., Osterhout, R. E., Stephen, R. (2011). Metabolic engineering of *Escherichia coli* for direct production of 1, 4-butanediol. *Nature chemical biology* **7**(7), 445.
- Yu, D., Ellis, H. M., Lee, E.-C., Jenkins, N. A., Copeland, N. G. (2000). An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proceedings of the National Academy of Sciences* **97**(11), 5978–5983.
- Zeghouf, M., Li, J., Butland, G., Borkowska, A., Canadien, V., Richards, D., Beattie, B., Emili, A., Greenblatt, J. F. (2004). Sequential Peptide Affinity (SPA) system for the identification of mammalian and bacterial protein complexes. *Journal of proteome research* **3**(3), 463–468.
- Zehentner, B. (2015). ‘Expression und Funktion von überlappenden ORFs in EHEC’. Thesis.
- Zehentner, B., Ardern, Z., Kreitmeier, M., Scherer, S., Neuhaus, K. (2020). A novel pH-regulated, unusual 603 bp overlapping protein coding gene *pop* is encoded antisense to *ompA* in *Escherichia coli* O157:H7 (EHEC). *Frontiers in microbiology* **11**, 377.
- Zehentner, B., Landstorfer, R., Scherer, S., Neuhaus, K. (2016). Overexpression of overlapping ORFs in *Escherichia coli* O157:H7 reveals growth phenotypes. *Society for Molecular Biology and Evolution (SMBE)*.
- Zehentner, B., Landstorfer, R., Scherer, S., Neuhaus, K. (2017). Combination of high-throughput and single growth assays detects overexpression phenotypes of translationally arrested overlapping ORFs in *E. coli*. *European Molecular Biology Laboratory (EMBL)*.
- Zehentner, B., Neuhaus, K., Hücker, S. M., Kreitmeier, M., Vanderhaeghen, S., Ardern, Z., Scherer, S. (2019). ‘Massive overlapping coding in the *E. coli* genome in which hundreds of overlapping genes form a hidden coding reserve’. submitted, currently in revision.
- Zehentner, B., Scherer, S., Neuhaus, K. (2018). Genome wide TSS-identification revealed transcriptional start sites for open reading frames antisense to annotated genes. *Gordon Research Seminar and Conference, Microbial Stress Response (GRS, GRC)*.

- Zhelyazkova, P., Sharma, C. M., Förstner, K. U., Liere, K., Vogel, J., Börner, T. (2012). The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. *The Plant Cell* **24**(1), 123–136.
- Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research* **31**(13), 3406–3415.

6 Supplement

6.1 Supplementary files

Supplementary files S1 to S6 are found on the enclosed CD-ROM.

Supplementary file S1: Western blots of overexpressed OGCs.

All original Western blot membranes without image correction are shown. Numbers indicate the overlapping gene candidate (OGC). All blots contain the protein length marker (L) and most contain the control protein glutathione S-transferase (G). Blot 1 shows additionally cells expressing the small control protein RpmH (R). Blot 2 contains separated proteins of non-vector carrying *Escherichia coli* cells (C, empty cells) and tag-expressing *Escherichia coli* cells (T, empty vector). Protein expression was performed in *E. coli* Top10 except for blots 41 and 42, where proteins were overexpressed in *E. coli* O157:H7 EDL933. The protein band of **OGC 137** is overlaid by the signal of the adjacent candidate and therefore invisible. No proteins were detected for candidates 12, 36, 151, and 214, as well as for 161, 211, and 240 where band patterns probably result from sample transfer of neighboring samples.

Supplementary file S2: RPKM values of overlapping gene candidates in high-throughput phenotyping.

RPKM values for 206 overlapping gene candidates (OGCs) analyzed in high-throughput phenotyping in 20 different culture conditions are listed.

Supplementary file S3: z -scores of overlapping gene candidates in HT phenotyping.

z -scores of 206 overlapping gene candidates (OGCs) are listed. z -scores are calculated with $z_{i,k} = \frac{x_{i,k} - x_i}{\sigma_i}$ for $x_{i,k}$ the RPKM value of candidate i in condition k , x_i the mean RPKM value of candidate i in all analyzed high-throughput conditions and σ_i the standard deviation of RPKM values of candidate i in all analyzed conditions.

Supplementary file S4: Phenotypic profiles of overlapping gene candidates.

Phenotypic profiles show z -scores of overlapping gene candidates (OGCs) for three biological

replicates of high-throughput phenotyping in conditions analyzed: 0 t_0 , 1 LB, 2 glucose, 3 L-malic acid, 4 L-arginine, 5 CsCl, 6 acetic acid, 7 malonic acid, 8 1-methylimidazole, 9 NaCl, 10 NaOH, 11 Na_3VO_4 , 12 sodium salicylate, 13 HClO_4 , 14 phytic acid, 15 1,2-propanediol, 16 1-propanol, 17 pyridoxine HCl, 18 *Staphylococcus*, 19 ZnCl_2

Table 1 shows profiles for candidates with high-throughput phenotype, Table 2 shows profiles for those without phenotype.

Supplementary file S5: Sequencing read visualization of Cappable-seq.

Visualization of 214 overlapping gene candidates in Artemis in the lower panel. Annotated genes are marked in blue, OGCs are marked in red/yellow. Sequencing reads of Cappable-seq mapped to the genome of *E. coli* O157:H7 EDL933 are shown in the upper panel. For reasons of clarity, sequencing reads of biological replicate I are shown only. TSS are highlighted with arrows. If the TSS signals are present in different growth conditions, one representative condition is shown, as indicated.

Supplementary file S6: Relative read score visualization of TSS.

Relative read scores (RRS) of transcriptional start sites (TSS) from Cappable-seq associated with overlapping gene candidates in eight analyzed conditions are shown. Genome positions of TSS are indicated in brackets. TSS are grouped according their significant different RRS: growth condition (Table 1), growth phase (Table 2), growth phase and growth condition (Table 3), no difference (Table 4). Categorization was performed according to p-values in t-tests (two-tailed Welch two-sample t-test or two-tailed two-sample t-test for analysis of growth conditions or growth phases, respectively; significance level $\alpha = 0.05$).

6.2 Supplementary tables

Supplementary table S1: Additional plasmids used and constructed.

Plasmid	Description	Source
pBAD+OGC x	pBAD/Myc-HisC expressing one of 206 overlapping genes candidates included in the HT-phenotyping, Myc- and his-tag expression prevented by natural stop codons of OGCs	Zehentner, 2015
pBAD+ΔOGC x	pBAD/Myc-HisC expressing one of 51 translationally arrested overlapping genes candidates included in the LT-phenotyping, Myc- and his-tag expression prevented by natural stop codons of OGCs	this work
pBAD+Δ <i>pop</i>	pBAD/Myc-HisC expressing translationally arrested <i>pop</i> , Myc- and his-tag expression prevented by natural stop codons of <i>pop</i>	this work
pBAD/SPA+OGC x	pBAD/SPA expressing one of 210 SPA-tagged overlapping genes candidates for western blots	this work
pBAD/SPA+ <i>gst</i>	expression vector for SPA-tagged glutathion S-transferase gene (<i>gst</i>) for western blots	this work
pBAD/SPA+ <i>rpmH</i>	expression vector for SPA-tagged ribosomal protein L34 gene (<i>rpmH</i>) for western blots	this work
pBAD/His+ <i>gst</i>	expression vector for His-tagged glutathion S-transferase gene (<i>gst</i>) for western blots	this work
pBAD/His+ <i>rpmH</i>	expression vector for His-tagged ribosomal protein L34 gene (<i>rpmH</i>) for western blots	this work
pMRS101+Δ <i>pop</i>	vector construct to produce the genomic knock-out EHEC Δ <i>pop</i>	this work
pKNG101+Δ <i>pop</i>	vector construct to produce the genomic knock-out EHEC Δ <i>pop</i>	this work
pMRS101+ΔOGC 15	vector construct to produce the genomic knock-out EHEC ΔOGC 15	this work
pProbe-NT+promoter- <i>helD</i>	vector construct including putative promoter sequence of the helicase D gene (<i>helD</i>) for promoter activity determination in a GFP-assay	this work
pProbe-NT+promoter-OGC 15 wt	vector construct including putative promoter sequence of OGC 15 for promoter activity determination in a GFP-assay	this work
pProbe-NT+promoter-OGC 15 -10 mt	vector construct including putative promoter sequence of OGC 15 for promoter activity determination in a GFP-assay, mutation within the predicted -10 box	this work
pProbe-NT+promoter-OGC 15 spacer mt	vector construct including putative promoter sequence of OGC 15 for promoter activity determination in a GFP-assay, mutation within the spacer region between -10 box and -35 box	this work
pProbe-NT+promoter-OGC 15 -35 mt	vector construct including putative promoter sequence of OGC 15 for promoter activity determination in a GFP-assay, mutation within the predicted -35 box	this work

Supplementary table S1: Continued from previous page

Plasmid	Description	Source
pProbe-NT+promoter-OGC 85	vector construct including putative promoter sequence of OGC 85 for promoter activity determination in a GFP-assay	this work
pProbe-NT+promoter-OGC 96 I	vector construct including first putative promoter sequence of OGC 96 for promoter activity determination in a GFP-assay	this work
pProbe-NT+promoter-OGC 96 II	vector construct including second putative promoter sequence of OGC 96 for promoter activity determination in a GFP-assay	this work
pProbe-NT+promoter-OGC 135	vector construct including putative promoter sequence of OGC 135 for promoter activity determination in a GFP-assay	this work
pProbe-NT+promoter-OGC 136 I	vector construct including first putative promoter sequence of OGC 136 for promoter activity determination in a GFP-assay	this work
pProbe-NT+promoter-OGC 136 II	vector construct including second putative promoter sequence of OGC 136 for promoter activity determination in a GFP-assay	this work
pProbe-NT+promoter-OGC 174	vector construct including putative promoter sequence of OGC 174 for promoter activity determination in a GFP-assay	this work
pProbe-NT+promoter-OGC 207	vector construct including putative promoter sequence of OGC 207 for promoter activity determination in a GFP-assay	this work
pProbe-NT+promoter-OGC 226 I	vector construct including first putative promoter sequence of OGC 226 for promoter activity determination in a GFP-assay	this work
pProbe-NT+promoter-OGC 226 II	vector construct including second putative promoter sequence of OGC 226 for promoter activity determination in a GFP-assay	this work
pProbe-NT+promoter- <i>pop</i>	vector construct including putative promoter sequence of <i>pop</i> for promoter activity determination in a GFP-assay	this work

Supplementary table S2: Reverse primer used for cloning pBAD/SPA+OGC **x** variants. Cut sites for restriction endonucleases indicated in the name of the primers are underlined.

name of primer	sequence (5' → 3')	name of primer	sequence (5' → 3')
OGC1+438R- <i>KpnI</i>	TAGGTACCCGTTAGTCCCGTCAGTAAA	OGC3+459R- <i>HindIII</i>	ATCAAGCTTCTTGATGGGGCTGGGAGCGT
OGC4+288R- <i>HindIII</i>	ATCAAGCTTCTGTACCTTCCGCGCCGATA	OGC5+156R- <i>HindIII</i>	ATCAAGCTTTCGCGTGGAAACGCGGGCAGA
OGC6+141R- <i>HindIII</i>	ATCAAGCTTCCGGTTGCGGCGGCGTACAA	OGC7+84R- <i>HindIII</i>	ATCAAGCTTCTTGCCGGGCAAGAACTGAA
OGC8+90R- <i>HindIII</i>	ATCAAGCTTCAGTTCGATTGCCGACAAAT	OGC9+567R- <i>HindIII</i>	ATCAAGCTTCATTATCGAAACCCCTGCCAC
OGC10+207R- <i>HindIII</i>	ATCAAGCTTCACCGTAAACAGTAACCTGA	OGC11+195R- <i>HindIII</i>	ATCAAGCTTCTATGGTGGGCGATAAATTA
OGC12+105R- <i>HindIII</i>	ATCAAGCTTCCCTCTCAGTGTGAAACGGA	OGC13+186R- <i>HindIII</i>	ATCAAGCTTCGGCACACTGCCTGACGGCA
OGC14+165R- <i>HindIII</i>	ATCAAGCTTCTAATAATGGTTTATTAAGT	OGC15+114R- <i>HindIII</i>	ATCAAGCTTCACCAAAGAATAAACGGCTC
OGC16+444R- <i>HindIII</i>	ATCAAGCTTCATCAAGGCATTGACCATCA	OGC17+315R- <i>HindIII</i>	ATCAAGCTTCCCCTGTTTGTAAAACCGG
OGC18+126R- <i>HindIII</i>	ATCAAGCTTCTCCAGAACGTGTTAAACGG	OGC19+423R- <i>HindIII</i>	ATCAAGCTTCGTCAACGCGGAAATCCTCA
OGC20+84R- <i>HindIII</i>	ATCAAGCTTCCGAACATATGGCACGAAAA	OGC21+258R- <i>HindIII</i>	ATCAAGCTTCCGTTTCTGGCAGAAACATC
OGC22+201R- <i>HindIII</i>	ATCAAGCTTCTCGCCGATCACACCAGCAT	OGC23+183R- <i>HindIII</i>	ATCAAGCTTCGGCGGCGCATAACAGGTATG
OGC24+132R- <i>HindIII</i>	ATCAAGCTTCTTTATGTTTCCGGCGGCAA	OGC25+81R- <i>HindIII</i>	ATCAAGCTTCCCTCTTTTCTACCCAACCGC
OGC26+144R- <i>HindIII</i>	ATCAAGCTTCGCATGATATTTACAAAGG	OGC27+144R- <i>HindIII</i>	ATCAAGCTTCTTGCTGATTATTGCCGGTG
OGC28+105R- <i>HindIII</i>	ATCAAGCTTCCAGTGCCGACGTCAAACGT	OGC29+207R- <i>HindIII</i>	ATCAAGCTTCCCTCAGGACCCGATAGGGCT
OGC30+336R- <i>EcoRI</i>	TAGACGAATTCCTTTTGTGAGGAAGGGTAA	OGC31+132R- <i>HindIII</i>	ATCAAGCTTCCGTTAACTCTGAGGTCTGG
OGC32+285R- <i>HindIII</i>	ATCAAGCTTCTAGTAGGGCACTTTTTTTTA	OGC33+303R- <i>HindIII</i>	ATCAAGCTTCGCGCCAGGTGTAAGGAAAG
OGC34+126R- <i>HindIII</i>	ATCAAGCTTCACCTGAAACGCCAGTCTGC	OGC35+141R- <i>HindIII</i>	ATCAAGCTTCCCTTATTGAGGTGAATAATG
OGC36+105R- <i>HindIII</i>	ATCAAGCTTCCGCCAATCCTCGGTGGCTT	OGC39+228R- <i>HindIII</i>	ATCAAGCTTCCCTGTTGAATAATGGACAA
OGC40+195R- <i>HindIII</i>	ATCAAGCTTCCGACCCGGCCTGCTTTGCT	OGC41+207R- <i>HindIII</i>	ATCAAGCTTCACCGCTAAGCACAGAAAAG
OGC42+159R- <i>HindIII</i>	ATCAAGCTTCTAACGATATTAATCCTGG	OGC43+534R- <i>HindIII</i>	ATCAAGCTTCGACCCGGGATACTGCGCGCG
OGC44+171R- <i>HindIII</i>	ATCAAGCTTCGCAAATTTCCAGGTGCCT	OGC45+135R- <i>HindIII</i>	ATCAAGCTTCTAACAGCGATAAATTCCCC
OGC46+372R- <i>HindIII</i>	ATCAAGCTTCTGACGTGACACCCGGTTCT	OGC47+99R- <i>HindIII</i>	ATCAAGCTTCTCTGCCCTGAAGGCGGCGG
OGC48+237R- <i>HindIII</i>	ATCAAGCTTCTACCTGCCCCCTGTCCOCT	OGC50+96R- <i>HindIII</i>	ATCAAGCTTCCCTGCTGTGCGGGCTGGGTGG
OGC51+177R- <i>HindIII</i>	ATCAAGCTTCGCCTGGTGGTCTGGTTTTG	OGC55+387R- <i>HindIII</i>	ATCAAGCTTCCAAACCCGCCGACCACAAAAG
OGC56+129R- <i>HindIII</i>	ATCAAGCTTCACCCGGCTTTTTATTTCATC	OGC57+780R- <i>HindIII</i>	ATCAAGCTTCCGTCGTATGCCGTACAAAG
OGC58+132R- <i>HindIII</i>	ATCAAGCTTCCGCACGGCAATTACAGTG	OGC59+243R- <i>HindIII</i>	ATCAAGCTTCTCTGTCTGCCGGAATGGGT
OGC60+375R- <i>HindIII</i>	ATCAAGCTTCACAACTTTTCGCGATGCG	OGC68+87R- <i>HindIII</i>	ATCAAGCTTCATTGTGGTGAGCATCATGG
OGC69+186R- <i>HindIII</i>	ATCAAGCTTCTTCGACCCGGACGAAAAAG	OGC70+84R- <i>HindIII</i>	ATCAAGCTTTCGCTCAAAGAGGCGCAGAGT

Supplementary table S2: Continued from previous page

name of primer	sequence (5' → 3')	name of primer	sequence (5' → 3')
OGC71+99R- <i>Hind</i> III	ATCAAGCTTCACTGCCTGAAAGATCAATA	OGC72+201R- <i>Hind</i> III	ATCAAGCTTCGCAAAAGGGCAAAACGGTG
OGC73+81R- <i>Hind</i> III	ATCAAGCTTCATGAAGGCGCTGATACTTA	OGC74+123R- <i>Hind</i> III	ATCAAGCTTCACCTCCACGAGCTTTGCTG
OGC75+318R- <i>Hind</i> III	ATCAAGCTTCAGATGCCAGGTTTTGGCAT	OGC76+417R- <i>Hind</i> III	ATCAAGCTTCTCCATTTCCGATAACGTCT
OGC77+360R- <i>Hind</i> III	ATCAAGCTTCTTCTCGAAGCCAGCGCCAA	OGC78+144R- <i>Hind</i> III	ATCAAGCTTCTATAAGAGACAGCGTAATC
OGC79+456R- <i>Hind</i> III	ATCAAGCTTCTGCCGCCAGCCGCATCAAC	OGC80+162R- <i>Hind</i> III	ATCAAGCTTCCAACAAGGCGGCTATATGA
OGC81+180R- <i>Hind</i> III	ATCAAGCTTCTTTGATGCAACAAGATTTG	OGC82+132R- <i>Hind</i> III	ATCAAGCTTCCCTTCCTGTGGCGATGTGGT
OGC83+399R- <i>Hind</i> III	ATCAAGCTTCTTTTAAAGCAAGAGTAAAT	OGC84+180R- <i>Eco</i> RI	TAGACGAATTCCTCACGGAGAGAATAAAAA
OGC85+75R- <i>Hind</i> III	ATCAAGCTTCTCGGATTCGCTTAATTTTA	OGC86+144R- <i>Hind</i> III	ATCAAGCTTCTCGCAGGGGTGACGCGGCA
OGC88+141R- <i>Hind</i> III	ATCAAGCTTCCGCATGATGCCGCGTAAAC	OGC89+639R- <i>Hind</i> III	ATCAAGCTTCGTTCTTAAATCCAGCATCC
OGC90+255R- <i>Hind</i> III	ATCAAGCTTCCAGCATCATCGCTTTGTGC	OGC91+186R- <i>Hind</i> III	ATCAAGCTTCAGCCAGATAGTGCGCCGTA
OGC92+81R- <i>Hind</i> III	ATCAAGCTTCTTTCCAGGGCAACCCGGTT	OGC93+180R- <i>Hind</i> III	ATCAAGCTTCGTTTTCGGGTAACGCAAA
OGC94+135R- <i>Hind</i> III	ATCAAGCTTCGGCATTGAGTCTGTATGCA	OGC95+417R- <i>Hind</i> III	ATCAAGCTTCCGTATCCGTGCCCCGCCTA
OGC96+87R- <i>Hind</i> III	ATCAAGCTTCTATGACCACAATGCACTCA	OGC98+240R- <i>Hind</i> III	ATCAAGCTTCCGTATCTGGTTTGTTTATA
OGC100+228R- <i>Eco</i> RI	TAGACGAATTCGCAATGGACTATGGCTTCA	OGC101+195R- <i>Hind</i> III	ATCAAGCTTTCGTTGCCAGCAGCTGGATCG
OGC102+318R- <i>Hind</i> III	ATCAAGCTTCCCGAAACTGCCGAGCTGCC	OGC103+165R- <i>Hind</i> III	ATCAAGCTTTCGAAGTTAGTCGATAAAGCG
OGC104+168R- <i>Hind</i> III	ATCAAGCTTCCCTGGAGATTATTAATGAT	OGC105+93R- <i>Hind</i> III	ATCAAGCTTCCCGTTCCAGGGTGGTGTGG
OGC106+258R- <i>Hind</i> III	ATCAAGCTTCGGTTGCCCGGAACACCTTT	OGC107+90R- <i>Hind</i> III	ATCAAGCTTCCGCTTGCGCCAGTCTCTGG
OGC108+162R- <i>Hind</i> III	ATCAAGCTTCGCAACCGAAGAGTGCGCCA	OGC109+105R- <i>Hind</i> III	ATCAAGCTTCCGGCGGATTATGGGAGTTT
OGC110+156R- <i>Hind</i> III	ATCAAGCTTCAATAAGTTGAGATGACACT	OGC111+168R- <i>Hind</i> III	ATCAAGCTTCCACAGGCCGATCTGAGCCAA
OGC112+72R- <i>Hind</i> III	ATCAAGCTTCGATTATCGCCCGCATCTCG	OGC113+177R- <i>Hind</i> III	ATCAAGCTTCAAAAGAGGCACTGGTTGAA
OGC114+144R- <i>Hind</i> III	ATCAAGCTTCCCTACCCATGAACAGCAGC	OGC115+135R- <i>Hind</i> III	ATCAAGCTTCTCCGTATAGCCGCTTTGAT
OGC116+303R- <i>Hind</i> III	ATCAAGCTTCTTTACCGTCATGGATTTCT	OGC117+147R- <i>Hind</i> III	ATCAAGCTTTCGAGCGCAAAAGTTGCCGAGG
OGC118+456R- <i>Hind</i> III	ATCAAGCTTCCCGTTGCGCGACCGTTTTG	OGC119+183R- <i>Hind</i> III	ATCAAGCTTCCCGTAAAACGTGAGCTGTA
OGC121+171R- <i>Hind</i> III	ATCAAGCTTCCAACCTAATACCGCCAAAA	OGC123+105R- <i>Hind</i> III	ATCAAGCTTTCGACAGGTTTAAAGAGGAAT
OGC124+156R- <i>Hind</i> III	ATCAAGCTTCATCGAGAACTCGCCAGCTT	OGC125+132R- <i>Hind</i> III	ATCAAGCTTCAGGATGGAGTAATGAGAAA
OGC126+291R- <i>Hind</i> III	ATCAAGCTTCACCTCCGATACTTATTCGC	OGC128+189R- <i>Hind</i> III	ATCAAGCTTTCGCGACTGCGTAAGGTCGAG
OGC129+960R- <i>Hind</i> III	ATCAAGCTTCGCCACCATTGCGGTGGTTG	OGC130+114R- <i>Hind</i> III	ATCAAGCTTCCCTTCGTAATTTTAAAGGC
OGC131+315R- <i>Hind</i> III	ATCAAGCTTCTTTTGCCGACCTGAAATCC	OGC132+171R- <i>Hind</i> III	ATCAAGCTTCCCGCGTATCTGGGCGATAC
OGC133+213R- <i>Hind</i> III	ATCAAGCTTCCGCTTGAATAGCCAGCCTG	OGC134+216R- <i>Hind</i> III	ATCAAGCTTCCATCGCGCTTACTTCGGTA
OGC135+432R- <i>Hind</i> III	ATCAAGCTTCATAAAGCAGATATTCCTG	OGC136+195R- <i>Hind</i> III	ATCAAGCTTCTGCGCTACGCTCTATGGCT

Supplementary table S2: Continued from previous page

name of primer	sequence (5' → 3')	name of primer	sequence (5' → 3')
OGC137+282R- <i>Hind</i> III	ATCAAGCTTCTTCTGGCAGGTAGGCGGAC	OGC138+291R- <i>Hind</i> III	ATCAAGCTTCTCGTTTTCAGCACCAATTGC
OGC139+378R- <i>Hind</i> III	ATCAAGCTTCTGCCGGACCAGACCCCGCC	OGC140+189R- <i>Hind</i> III	ATCAAGCTTTCGTGCCGGTGGTGACGTGAC
OGC141+462R- <i>Hind</i> III	ATCAAGCTTCGGAACGGTATGCTGAATTC	OGC142+171R- <i>Hind</i> III	ATCAAGCTTCTTTTGGCAACGAGTCACCG
OGC143+231R- <i>Hind</i> III	ATCAAGCTTCCCTCCGTCACTGCTTGGCGTG	OGC144+96R- <i>Hind</i> III	ATCAAGCTTTCGCACCAGACCCTGACTGCG
OGC145+270R- <i>Hind</i> III	ATCAAGCTTCACAGCGCCTCAGAGTATGA	OGC146+135R- <i>Hind</i> III	ATCAAGCTTCCACGGCTATCTGGCGCGAG
OGC147+132R- <i>Hind</i> III	ATCAAGCTTCACTACAGCGATGGTGTAAT	OGC148+276R- <i>Hind</i> III	ATCAAGCTTCCAGGAAATTATCATCCTTA
OGC149+282R- <i>Hind</i> III	ATCAAGCTTCCCTGCAACACTACAGTTTTC	OGC150+444R- <i>Hind</i> III	ATCAAGCTTTCGGTGAACACCGGTAAAGGC
OGC151+72R- <i>Hind</i> III	ATCAAGCTTCCATCTCTCTTACACCGCCG	OGC152+213R- <i>Hind</i> III	ATCAAGCTTCCCTGCCTGAAAACGTTGAGT
OGC153+96R- <i>Hind</i> III	ATCAAGCTTCCCTTCTTCTGCCAGCATATT	OGC154+99R- <i>Hind</i> III	ATCAAGCTTCATAAAAAAGAAGGCCAGAT
OGC156+213R- <i>Hind</i> III	ATCAAGCTTCACCAAACATACTGATGTGA	OGC157+213R- <i>Hind</i> III	ATCAAGCTTCAAGTCGTACGCCGGTTAAG
OGC158+177R- <i>Hind</i> III	ATCAAGCTTCCAATTGCGCACCCGCGCAT	OGC159+150R- <i>Hind</i> III	ATCAAGCTTTCGGTGTTATCAATATTGGCG
OGC160+192R- <i>Hind</i> III	ATCAAGCTTCACGGCGGCAAAGCCCTGAC	OGC161+75R- <i>Hind</i> III	ATCAAGCTTCCCTACCCAGACGCATCTGA
OGC162+384R- <i>Hind</i> III	ATCAAGCTTCATGGAGCAGTACGATGTGC	OGC163+171R- <i>Hind</i> III	ATCAAGCTTCAACGGCACGTTGGAACAGA
OGC164+147R- <i>Hind</i> III	ATCAAGCTTCCACCGTGCTGGTGTCTAAT	OGC165+405R- <i>Hind</i> III	ATCAAGCTTCCCTGCACTTCCACCTCGGTT
OGC167+450R- <i>Hind</i> III	ATCAAGCTTCGCAGAAACGAGATTTGAT	OGC168+483R- <i>Hind</i> III	ATCAAGCTTCCAGTCGTGAATTTAAAATC
OGC169+177R- <i>Hind</i> III	ATCAAGCTTCGAAAATTTCACTTAGTGAT	OGC171+348R- <i>Hind</i> III	ATCAAGCTTTCGCGCAGCCAGCTGCGCCGC
OGC172+90R- <i>Hind</i> III	ATCAAGCTTCCCACAAGGAATGCAAATGA	OGC173+345R- <i>Hind</i> III	ATCAAGCTTTCGGTGGGATGCTGATGGGGG
OGC174+156R- <i>Hind</i> III	ATCAAGCTTCCGCGTCAACCAGAGTGATA	OGC175+210R- <i>Hind</i> III	ATCAAGCTTTCGTTTTCTGCGTAATGCCCCG
OGC176+225R- <i>Hind</i> III	ATCAAGCTTCGCGCCTAAAATCGACCTCC	OGC177+342R- <i>Hind</i> III	ATCAAGCTTCCCTCCAGGCTCGCCAACCTCA
OGC178+168R- <i>Hind</i> III	ATCAAGCTTCACATTTTCTCGTTTTGAAAG	OGC179+420R- <i>Hind</i> III	ATCAAGCTTCCATCCTTGCAGTACTGGTG
OGC180+168R- <i>Hind</i> III	ATCAAGCTTCGCAGCACCAAAGCGGCAAA	OGC181+258R- <i>Hind</i> III	ATCAAGCTTCAAAGATCGCCGCGCCTCGG
OGC182+192R- <i>Hind</i> III	ATCAAGCTTCCATTGAGCACCTGCGTGAC	OGC183+267R- <i>Eco</i> RI	TAGACGAATTCCGGTGATGCCCTTGCCGAA
OGC184+102R- <i>Hind</i> III	ATCAAGCTTCACATGAAGCGGCTCGGTCA	OGC185+228R- <i>Hind</i> III	ATCAAGCTTTCACGTTGAAGTTGTGGCGAT
OGC186+117R- <i>Hind</i> III	ATCAAGCTTCTTGGTCATCTGAACACCAT	OGC187+204R- <i>Hind</i> III	ATCAAGCTTTCGCTGGTGACGAACGTGAGC
OGC188+105R- <i>Hind</i> III	ATCAAGCTTCCGACCCGGTGAAAATGGCGG	OGC189+228R- <i>Hind</i> III	ATCAAGCTTCTCCGTTATTTCTCGGCTTT
OGC190+339R- <i>Hind</i> III	ATCAAGCTTCTAACTCAAATTCCTGATA	OGC191+93R- <i>Hind</i> III	ATCAAGCTTTCACGGCATTGACGAAGTGCG
OGC192+165R- <i>Hind</i> III	ATCAAGCTTCGCTTTCTGGCGGTGAACAA	OGC193+129R- <i>Hind</i> III	ATCAAGCTTTCACACGCTGTTTGAAAAATC
OGC194+570R- <i>Hind</i> III	ATCAAGCTTCGGCCATAAATTCGGTCTGG	OGC195+333R- <i>Hind</i> III	ATCAAGCTTCTCCGCAGCGTCGGGAGCTT
OGC196+144R- <i>Hind</i> III	ATCAAGCTTCTTTCTCATTTTTGTGATG	OGC197+210R- <i>Hind</i> III	ATCAAGCTTCTGTGCCGGAAGATATCACT
OGC198+105R- <i>Hind</i> III	ATCAAGCTTCCAGCYTGCATCGCTTGCGC	OGC199+237R- <i>Hind</i> III	ATCAAGCTTCCGCAATGAGGAAGATTGCC

Supplementary table S2: Continued from previous page

name of primer	sequence (5' → 3')	name of primer	sequence (5' → 3')
OGC200+132R- <i>Hind</i> III	ATCAAGCTTCCCGACGACGAGATCCTTGG	OGC201+219R- <i>Hind</i> III	ATCAAGCTTCCCGTGGATATGCCGACACC
OGC202+306R- <i>Hind</i> III	ATCAAGCTTCTGGAAAGACCTTGAGTGGA	OGC203+99R- <i>Hind</i> III	ATCAAGCTTTCGTCTGCATTTAACTGGCAT
OGC204+219R- <i>Hind</i> III	ATCAAGCTTCCAGCCGCGCCAACAATCCT	OGC205+873R- <i>Eco</i> RI	TAGACGAATTCCGGTTGATTTCAGGAGTGCG
OGC206+198R- <i>Hind</i> III	ATCAAGCTTCCCTGGGCGTTCATTCTTGTC	OGC207+165R- <i>Hind</i> III	ATCAAGCTTCCGCATGGCTTGCCGACGCG
OGC208+120R- <i>Hind</i> III	ATCAAGCTTCGGCAATGTGATTTGTTGCA	OGC209+231R- <i>Hind</i> III	ATCAAGCTTCCCTGGCGCTTGACCGCCAGC
OGC210+165R- <i>Hind</i> III	ATCAAGCTTCTCGTTGAATCGCGACAGAA	OGC211+105R- <i>Hind</i> III	ATCAAGCTTACCCTTACAACAACGGCGC
OGC212+747R- <i>Hind</i> III	ATCAAGCTTCCACCGCCGTGGCTTTGCGCC	OGC213+207R- <i>Hind</i> III	ATCAAGCTTCCCCTTCATTCCACAATACTG
OGC214+273R- <i>Hind</i> III	ATCAAGCTTCCACCGCCTGCAAGGGATCGA	OGC215+111R- <i>Hind</i> III	ATCAAGCTTCCATGATATGTTGAATCCTA
OGC217+105R- <i>Hind</i> III	ATCAAGCTTTCGTGAGCGATGCAGCTGAAC	OGC218+243R- <i>Hind</i> III	ATCAAGCTTTCGCAGATACCATTGATGTGG
OGC219+135R- <i>Hind</i> III	ATCAAGCTTTCAGTTATCTGCGGCATCTGC	OGC220+84R- <i>Hind</i> III	ATCAAGCTTCTGAGTTTTCAACCGACGAG
OGC221+315R- <i>Hind</i> III	ATCAAGCTTCTATCGATACGACTGAATGC	OGC222+105R- <i>Hind</i> III	ATCAAGCTTCTAAATTAATGGTGCCGGTT
OGC223+1476R- <i>Hind</i> III	ATCAAGCTTCAACAACATTCGTATCGAAG	OGC224+72R- <i>Hind</i> III	ATCAAGCTTCCCCTGTGGATACTCTCCCGC
OGC225+150R- <i>Hind</i> III	ATCAAGCTTCGCGCAAAAAATTAACAGT	OGC226+300R- <i>Hind</i> III	ATCAAGCTTCCAAGGTCAGGAAGAAGCGG
OGC227+216R- <i>Hind</i> III	ATCAAGCTTCGATCTCAGTTAGCAATATT	OGC228+312R- <i>Hind</i> III	ATCAAGCTTCTTTTCATGCCACAAGGCAAA
OGC229+231R- <i>Hind</i> III	ATCAAGCTTCTACCCTATCATTAAATGAAT	OGC230+138R- <i>Hind</i> III	ATCAAGCTTCCC GCATCGACCAGCTGCTG
OGC231+648R- <i>Hind</i> III	ATCAAGCTTCCCTCAACGTCCCTGCGGGTA	OGC232+324R- <i>Hind</i> III	ATCAAGCTTCCGAGTGTGGCTTTACCGGT
OGC235+180R- <i>Hind</i> III	ATCAAGCTTCGATGTCCGGCGAGTTCCCC	OGC236+129R- <i>Hind</i> III	ATCAAGCTTCCGTTACTGACTGGCTGGTC
OGC237+240R- <i>Hind</i> III	ATCAAGCTTCTGCGCTGGAACAGGCGGGC	OGC238+255R- <i>Hind</i> III	ATCAAGCTTCCCAGGATGAAGGTAAAGTT
OGC239+141R- <i>Hind</i> III	ATCAAGCTTCCCGGTAGGGGCCAGCGGCC	OGC240+135R- <i>Hind</i> III	ATCAAGCTTCATTTTCTGAGTAATGCTGA
OGC241+153R- <i>Hind</i> III	ATCAAGCTTCGCCCAGGATCAGGCAGATG	OGC242+366R- <i>Hind</i> III	ATCAAGCTTCCGCCTGACTGCGCGTCCGA

Supplementary table S3: Mutation primer used to construct pBAD/ Δ OGC x variants. Restriction endonucleases indicated in some primer names refer to cut site deleted or created by the mutations introduced.

name of primer	sequence (5' → 3')	name of primer	sequence (5' → 3')
OGC3+45F-mutS	TCAACCGCCATAGAGTCTCGAT	OGC145+42F-mutS	ATAAGCGATGACTTCTCGCGCCA
OGC6+17F-mutS	ACAAGGTATCCTGACGGCGCTCT	OGC146+31F-mutS	TTCGCGTGACAGCAACATCGA
OGC15+11F-mutS	TCGAGCGTCTTTAGCCTGTTGCCA	OGC147+30F-mutS	AGACCGAAAGTAGTGGTGTGGT
OGC18+25F-mutS	GATAAACTGCCTGAAGCGGGTTTC	OGC153+34F-mutS	AGGAACGCGTAGATACCGTTATC
OGC18+25R-mutS	GAAACCCGCTTCAGGCAGTTTTATC	OGC164+29F-mutS	GGCGAGTTCCCTAATCGTGCATCTC
OGC23+25F-mutS	CTGATAGCACCTTAAATACCAAACG	OGC164+29R-mutS	GAGATGCACGATTAGGGAACCTCGCC
OGC24+31F-mutS	GCATCCGGCACATAACTTAATCGCGG	OGC167+37F-mutS	GTAGGTCTGATATGACGCGCAAG
OGC24+31R-mutS	CCGCGATTAAGTTATGTGCCGGATGC	OGC172+26F-mutS	TCGGCCTGGCTTAACGAACTGA
OGC25+12F-mutS	TACCGTATTCTGACCGCAGCCCCA	OGC174+32F-mutS	GTCAAGCGTAGCATCCGGCA
OGC26+42F-mutS	GATTGCCGGGTAGAGAGATAT	OGC177+33F-mutS	GATCTTTGCGAATTAGCGCCTTGAGG
OGC30+28F-mutS	CTACTTCTTCGGTAAGTAAGCGAGAAC	OGC178+35F-mutS	TGCATCTGCTGTAGATAAAAAGGCA
OGC30+28R-mutS	GTTCTCGCTTACTTACCGAAGAAGTAG	OGC183+30F-mutS	GATGGTAAAATAATTGCGGCACTG
OGC31+29F-mutS	GTTTATATTGCTAACATGCGAAGAAG	OGC183+30R-mutS	CAGTGCCGCAATTATTTTACCATC
OGC31+29R-mutS	CTTCTTCGCATGTTAGCAATATAAAC	OGC186+28F-mutS	TGCATTACGCCTTAGAATCCCAGCAACTT
OGC44+45F-mutS	GTATTTGATGTAAATTTTCCTTC	OGC191+33F-mutS	GAAACGCGCCTGACCAGCCCGGTCAA
OGC50+14F-mutS	TCATCGGCATGTAGTTTGCCGT	OGC191+33R-mutS	TTGACCGGGCTGGTCAGGCGCGTTTC
OGC51+45F-mutS	ATGCTTGAGCGCTAGCGGAAAAA	OGC194+54F-mutS	GGTGATCGCTTAAATATTTTCAGG
OGC68+34F-mutS	CTGCCCCGGTAGAGTGCGGCTAA	OGC194+54R-mutS	CCTGAAATATTTAAGCGATCACC
OGC71+14F-mutS	CGTGTCCATCCTAGTTAAAACAAGA	OGC195+24F-mutS	ACCCAGCCGATAGTTAAAGCGTT
OGC75+45F-mutS	AGCAAAGCTCGTGTAGGTTAGCAGT	OGC198+25F-mutS	TGGCTCATTTAGCTTTCCGGGCCA
OGC85+16F-mutS	TGGACTTGAGCTACCGGTCGTT	OGC205+31F-mutS	GAGGTGAGTTTGTGAGGCAATATTTTC
OGC96+28F-mutS	TTCCGGAACCTAAGGGTTCAC	OGC213+52F-mutS	AAGCTGCGTTGAGCACTTAATTG
OGC106+84F-mutS	GACGGCGTTTTAGTTTCGGTTTTCA	OGC217+27F-mutS	CATTAATATATAGATAGCCACGA
OGC107+32F-mutS	TGCTGTTCTTTTAGCGCCAC	OGC218+52F-mutS	ACCAAACCTGTAGCGCGTCAA
OGC116+21F-mutS	TCAGCGGTTTTGATGCCAGCAAC	OGC226+34F-mutS	TTAACGCGTTGAGGAAGTCGGCGT
OGC117+44F-mutS	ATGCTGCGACATAAACGATTACAC	OGC226+34R-mutS	ACGCCACTTCCTCAACGCGTTAA
OGC119+64F-mutS	GCAATGGCGTAGGTAACGGCGCT	OGC231+30F-mutS	GAGGCGTTATTAAGCTTGCAGGCG
OGC121+30F-mutS	GAAATTCAACTAAGATCTCAGGG	OGC231+30R-mutS	CGCCTGCAAGCTTAATAACGCCTC

Supplementary table S3: Continued from previous page

name of primer	sequence (5' → 3')	name of primer	sequence (5' → 3')
OGC121+30R-mutS	CCCTGAGATCTTAGTTGAATTC	OGC232+31F-mutS	CTACAGCCGTAGCCAAATGTA
OGC137+39F-mutS	AGTTCTTGACTGACGGATACATAG	OGC241+40F-mutS	CGTTCCGTATAGCGTCAGGATAAAC
OGC139+15F-mutS	ATGGGGATGGTAAAATAATCCGGT	OGC241+40R-mutS	GTTTATCCTGACGCTATACGGAACG
OGC140+26F-mutS- <i>BspHI</i>	CGGCTATTATTCATGATTTCAATCAC		
OGC140+26R-mutS- <i>BspHI</i>	GTGATTGAAATCATGAAATAATAGCCG		

Supplementary table S4: Primer used to construct pProbe-NT+promoter variants. Cut sites for restriction endonucleases indicated in the name of the primers are underlined.

name of primer	sequence (5' → 3')	name of primer	sequence (5' → 3')
OGC85-165F+ <i>SalI</i>	ATGGT <u>TCGAC</u> AGTTTCACCGGACGCATATT	OGC174-110F+ <i>SalI</i>	AGT <u>GTCGAC</u> CCAGCCGACTGGAAGACGCG
OGC85-83R+ <i>EcoRI</i>	AGAGT <u>GAAATTC</u> ACGCAGAAAATAATACTCT	OGC174-29R+ <i>EcoRI</i>	AGAGT <u>GAAATTC</u> TGATAAGATGCGCCAGCAT
OGC96-146F+ <i>SalI</i> I	AGT <u>GTCGAC</u> ATTGTATCAAGAATTAGGGA	OGC207-276F+ <i>SalI</i>	AGT <u>GTCGAC</u> AGGAAGTTATTACTCAGGAA
OGC96-105R+ <i>EcoRI</i> I	AGAGT <u>GAAATTC</u> GCAAAATTAATTTTACTAG	OGC207-194R+ <i>EcoRI</i>	AGAGT <u>GAAATTC</u> ACTCGTTTATTATGCCACA
OGC96-71F+ <i>SalI</i> II	AGT <u>GTCGAC</u> CCATACAGTCAAATTTCTAGT	OGC226-301F- <i>SalI</i> I	AGT <u>GTCGAC</u> AGAGTTACAGCACTTTTTTGC
OGC96-30R+ <i>EcoRI</i> II	AGAT <u>CGAATTC</u> AGAACAATATTTTTGCATA	OGC226-157R- <i>EcoRI</i> I	AGCGT <u>GAAATTC</u> CATCTCTGCGATGACCAATT
OGC135-129F+ <i>SalI</i>	ATGGT <u>TCGAC</u> ACACGCTTTTGTCAATCCAT	OGC226-103F- <i>SalI</i> II	ACT <u>GTCGAC</u> TAAACCTTCCAGTACCAAAAC
OGC135-78R+ <i>EcoRI</i>	AGAT <u>CGAATTC</u> CTGCTACAATGATGTTAAA	OGC226-51R- <i>SacI</i> II	AGT <u>GAGCTCT</u> TTTCTATAAGGATAATGAATG
OGC136-164F+ <i>SalI</i> I	ATGGT <u>TCGAC</u> GATCACTACGGAGCTGGTAT	helD-127F+ <i>SalI</i>	AGT <u>GTCGAC</u> CCGATAAAAACCTCGCTTTAC
OGC136-132R+ <i>EcoRI</i> I	AGAGT <u>GAAATTC</u> GTCACCGTATCGTACACGG	helD-46R+ <i>EcoRI</i>	AGAGT <u>GAAATTC</u> TCTTATCAGTGTAACCGTC
OGC136-113F+ <i>SalI</i> II	AGT <u>GTCGAC</u> AGCAGTCAGCTGCGCAGACA		
OGC136-78R+ <i>EcoRI</i> II	AGTAC <u>GAAATTC</u> TTAATACTATTCTGGCACA		
OGC15-146F- <i>SalI</i>	AGT <u>GTCGAC</u> CCCCAACGGTGATGATTACCG		
OGC15-141R- <i>EcoRI</i> -35mt	AGAGT <u>GAAATTC</u> TTGCGCTGAATGATACTGCATGTCGCGTATTGAAACCACGCATCGGTAATCATCACCG		
OGC15-130R- <i>EcoRI</i> -Spacer-mt	AGAGT <u>GAAATTC</u> TTGCGCTGAATGATACTGAATGGCTCCTATTCAAAAAACAAATCGGT		
OGC15-114R- <i>EcoRI</i> -10mt	AGAGT <u>GAAATTC</u> TTGCGCTGACCGCCTCTGCATGTCGCGTATT		
OGC15-103R- <i>EcoRI</i>	AGAGT <u>GAAATTC</u> TTGCGCTGAATGATACTGCA		

Supplementary table S5: Supplier information of chemicals.

Supplier	Chemicals
Applichem	nitro blue tetrazolium chloride (NBT)
Baker	ethanol
Fluka	caesium chloride (CsCl), casamino acid, D-(+)-glucose monohydrate, perchoric acid (HClO ₄), sodium salicylate
iNtRON Biotechnology	RedSafe
Invitrogen	Trizol
Merck	calcium chloride (CaCl ₂), sodium hydroxide (NaOH), vitamin B ₁ (thiamine hydrochloride), trichloroacetic acid, 1-butanol, 1-methylimidazole
Oxoid	bacteriological agar, tryptone, yeast extract
Promega	β -mercaptoethanol
Riedle-de-haen	formic acid
Roth	acetic acid, ammonium chloride (NH ₄ Cl), ampicillin sodium salt, arabinose, chloroform, disodium ethylenediaminetetraacetate (Na ₂ EDTA), disodium phosphate (Na ₂ HPO ₄), glycerine, glycine, hydrogen chloride (HCl), kanamycin sulfate, magnesium chloride (MgCl ₂), magnesium sulfate (MgSO ₄), MES, methanol, MOPS, phytic acid, potassium chloride (KCl), potassium dihydrogenphosphate (KH ₂ PO ₄), pyridoxine hydrochloride, Roti [®] -Phenol/Chloroform/Isoamyl alcohol, sodium acetate (NaOAc), sodium chloride (NaCl), sodium dodecyl sulfate (SDS), tricine, tris(hydroxymethyl)aminomethane (Tris), Tween20, 1-propanol, 2-butanol, 2-propanol, 5-Bromo-4-chloro-3-indolyl phosphate (BCIP)

Supplementary table S5: Continued from previous page

Supplier	Chemicals
Sigma-Aldrich	1,2-Propanediol, bicine, Coomassie brilliant blue G 250, dithiothreitol (DTT), L-arginine, L-malic acid, malonic acid, nalidixic acid, sodium orthovanadate (Na_3VO_4), streptomycin, zinc chloride (ZnCl_2)
ThermoFischer	glycogen (RNA grade)

Supplementary table S6: List of analyzed overlapping gene candidates.

Candidate number, genomic start and genomic stop positions of putative translated overlapping genes candidates are listed, as well as the length of the ORF in nucleotides [nt] and amino acids [aa], the translational efficiency specified by the ribosomal coverage value ($RCV = \frac{RPKM_{translatome}}{RPKM_{transcriptome}}$) and the overlap type. OGC numbers of candidates with positive signals in all three experimental analyses conducted, i. e., protein analysis in Western blots, high-throughput phenotyping, and transcriptional start site determination are shaded in light gray. The RCVs of candidates having $RCV \leq 0.197$ are shaded in dark gray. For partial overlaps, orientation of the overlapping gene pair is given according to Table 1.2 and the overlapping region is indicated in brackets. If one partial overlapping gene overlaps substantially (> 30 bp) with two annotated gene, the longer overlap is depicted. All overlapping genes are located antisense in respect to the annotated gene, apart from OGC 86.

^a: candidates not cloned for protein detection

^b: candidates not cloned for overexpression phenotyping

^c: a second overlap of > 30 bp is present but not listed

candidate	start	stop	length [nt]	length [aa]	RCV	overlap
1	5785	5327	459	152	0.489	3' partial (221 bp)
3	91432	90953	480	159	0.709	embedded
4	139266	139574	309	102	0.524	embedded
5	146372	146548	177	58	0.628	3' partial (168 bp)
6	146589	146750	162	53	0.319	embedded
7	152604	152708	105	34	0.554	embedded
8	152656	152766	111	36	0.504	embedded
9	152797	153384	588	195	0.227	embedded
10	156342	156569	228	75	0.309	embedded
11	226771	226556	216	71	0.358	3' partial (184 bp)

Supplementary table S6: Continued from previous page

candidate	start	stop	length [nt]	length [aa]	RCV	overlap
12	238462	238337	126	41	0.472	5' partial (96 bp)
13	255599	255805	207	68	1.569	embedded
14	284781	284596	186	61	1.293	5' partial (104 bp)
15	300575	300709	135	44	0.094	embedded
16	397692	397228	465	154	0.313	3' partial (351 bp) ^c
17	459830	460165	336	111	0.072	embedded
18	487981	488127	147	48	1.360	embedded
19	536838	536395	444	147	0.784	3' partial (312 bp) ^c
20	551920	551816	105	34	0.229	5' partial (95 bp)
21 ^b	557232	556954	279	92	0.989	3' partial (201 bp) ^c
22	569151	569372	222	73	0.497	3' partial (107 bp)
23	570371	570574	204	67	1.479	5' partial (116 bp)
24	572347	572499	153	50	0.901	3' partial (107 bp)
25	573442	573543	102	33	0.199	embedded
26	573736	573900	165	54	0.140	5' partial (135 bp)
27	578933	578769	165	54	1.658	embedded
28	605196	605071	126	41	4.763	embedded
29	620654	620427	228	75	0.729	embedded
30	690295	689939	357	118	0.460	annotated gene embedded

Supplementary table S6: Continued from previous page

candidate	start	stop	length [nt]	length [aa]	RCV	overlap
31 ^a	690646	690494	153	50	0.493	5' partial (28 bp)
32	744843	745148	306	101	0.135	5' partial (255 bp)
33	745267	745590	324	107	0.251	embedded
34	751822	751968	147	48	2.173	embedded
35	783269	783430	162	53	0.370	5' partial (143 bp)
36	818676	818801	126	41	2.283	embedded
39	834715	834467	249	82	0.948	embedded
40	847645	847860	216	71	0.249	embedded
41	878158	878385	228	75	0.417	5' partial (131 bp)
42	887023	886844	180	59	1.666	3' partial (121 bp)
43	960070	959516	555	184	0.325	embedded
44	975155	975346	192	63	0.885	embedded
45	1004840	1004685	156	51	1.218	embedded
46	1014103	1014495	393	130	0.337	embedded
47	1053268	1053387	120	39	0.349	embedded
48	1053332	1053589	258	85	2.310	5' partial (105 bp)
50	1068989	1069105	117	38	1.020	embedded
51	1110879	1110682	198	65	0.558	embedded
55	1170587	1170994	408	135	0.758	embedded

Supplementary table S6: Continued from previous page

candidate	start	stop	length [nt]	length [aa]	RCV	overlap
56	1224245	1224394	150	49	1.055	embedded
57	1235822	1236622	801	266	0.925	embedded
58	1237252	1237404	153	50	1.186	embedded
59	1247671	1247934	264	87	0.928	embedded
60 ^{a,b}	1397192	1397587	396	131	0.440	embedded
68	1560568	1560461	108	35	2.997	embedded
69	1572884	1572678	207	68	1.674	embedded
70	1591427	1591323	105	34	0.177	5' partial (85 bp)
71	1591577	1591458	120	39	0.190	embedded
72	1595079	1594858	222	73	0.403	embedded
73	1753846	1753947	102	33	0.216	embedded
74	1753958	1754101	144	47	0.307	embedded
75	1754037	1754375	339	112	0.737	5' partial (110 bp), elongates annotated gene
76	1759756	1760193	438	145	0.115	3' partial (345 bp)
77	1766909	1767289	381	126	1.167	3' partial (242 bp) ^c
78	1768657	1768821	165	54	0.573	5' partial (128 bp)
79	1785186	1785662	477	158	0.902	3' partial (361 bp), elongates annotated gene
80	1800178	1800360	183	60	0.430	annotated gene embedded ()
81	1820227	1820027	201	66	1.299	embedded

Supplementary table S6: Continued from previous page

candidate	start	stop	length [nt]	length [aa]	RCV	overlap
82	1951148	1951300	153	50	0.939	3' partial (134 bp)
83	1957028	1956609	420	139	1.936	5' partial (357 bp)
84	1985693	1985493	201	66	0.889	3' partial (194 bp)
85	1985915	1985820	96	31	0.579	embedded
86	2070776	2070612	165	54	0.501	sense, 5' partial (125 bp)
88	2086712	2086873	162	53	0.016	3' partial (30 bp)
89	2097766	2097107	660	219	0.404	3' partial (594 bp)
90	2116902	2116627	276	91	0.358	embedded
91	2194737	2194943	207	68	1.371	embedded
92	2199885	2199784	102	33	0.744	embedded
93	2200092	2199892	201	66	2.304	embedded
94	2239591	2239746	156	51	0.719	embedded
95 ^{a,b}	2283712	2283275	438	145	0.285	embedded (in pseudogene)
96	2285486	2285379	108	35	0.528	embedded
98	2348952	2348701	252	83	0.514	embedded (in pseudogene)
100	2351010	2350762	249	82	0.324	3' partial (246 bp)
101	2365208	2364993	216	71	0.281	embedded
102 ^b	2427881	2428219	339	112	3.098	3' partial (279 bp)
103	2451861	2452046	186	61	0.239	3' partial (154 bp)

Supplementary table S6: Continued from previous page

candidate	start	stop	length [nt]	length [aa]	RCV	overlap
104 ^b	2459634	2459822	189	62	1.682	5' partial (172 bp)
105	2508653	2508540	114	37	1.276	embedded
106	2517304	2517026	279	92	1.177	3' partial (234 bp)
107	2518208	2518098	111	36	0.576	embedded
108	2524985	2525167	183	60	0.410	embedded
109	2535669	2535544	126	41	0.301	embedded
110	2559999	2559850	150	49	0.633	5' partial (86 bp)
111	2570050	2569862	189	62	1.952	embedded
112	2573882	2573790	93	30	0.175	embedded
113	2574440	2574243	198	65	2.585	embedded
114	2590840	2591004	165	54	1.960	3' partial (162 bp)
115	2591370	2591215	156	51	0.294	embedded
116	2597331	2597654	324	107	1.189	embedded
117	2633160	2632993	168	55	0.358	embedded
118	2640768	2641244	477	158	0.472	embedded
119	2724603	2724806	204	67	1.056	3' partial (146 bp)
121	2758320	2758129	192	63	2.685	embedded
123	2850662	2850537	126	41	0.455	5' partial (98 bp)
124 ^a	2879456	2879632	177	58	2.788	5' partial (129 bp) ^c

Supplementary table S6: Continued from previous page

candidate	start	stop	length [nt]	length [aa]	RCV	overlap
125	2903381	2903229	153	50	0.589	5' partial (139 bp)
126	2922867	2922556	312	103	1.603	embedded
128	3032991	3033200	210	69	1.045	embedded
129	3037014	3037994	981	326	0.126	embedded
130	3057632	3057498	135	44	0.529	embedded
131	3094477	3094142	336	111	1.631	embedded
132 ^b	3126316	3126125	192	63	0.329	3' partial (140 bp)
133	3186694	3186461	234	77	1.600	embedded
134	3204597	3204833	237	78	5.229	3' partial (213 bp)
135	3218749	3219201	453	150	1.468	3' partial (323 bp)
136	3226797	3226582	216	71	0.207	5' partial (152 bp)
137	3234399	3234097	303	100	1.003	embedded
138	3324998	3324687	312	103	0.266	5' partial (230 bp)
139	3325403	3325005	399	132	0.299	embedded
140	3396024	3396233	210	69	2.466	3' partial (176 bp)
141	3544592	3544110	483	160	0.465	annotated gene embedded
142	3549358	3549167	192	63	1.280	embedded
143	3555142	3554891	252	83	3.913	embedded
144	3606140	3606024	117	38	0.719	embedded

Supplementary table S6: Continued from previous page

candidate	start	stop	length [nt]	length [aa]	RCV	overlap
145	3614589	3614299	291	96	2.659	embedded
146	3614693	3614538	156	51	0.311	embedded
147	3620967	3620815	153	50	0.093	embedded
148	3625381	3625085	297	98	0.643	embedded
149	3642357	3642659	303	100	0.686	embedded
150	3663248	3662784	465	154	0.610	3' partial (400 bp) ^c
151	3663610	3663518	93	30	0.171	embedded
152	3663844	3663611	234	77	0.180	embedded
153	3664068	3663952	117	38	0.214	embedded
154	3690649	3690768	120	39	0.396	embedded
156	3711252	3711019	234	77	0.005	3' partial (99 bp)
157	3724602	3724835	234	77	0.318	embedded
158	3793574	3793771	198	65	1.788	embedded
159	3793750	3793920	171	56	0.725	embedded
160	3854004	3853792	213	70	1.188	embedded
161	3854833	3854928	96	31	51.613	embedded
162	3866157	3865753	405	134	0.169	3' partial (399 bp)
163	3899578	3899387	192	63	1.446	embedded
164	3913753	3913920	168	55	0.181	embedded

Supplementary table S6: Continued from previous page

candidate	start	stop	length [nt]	length [aa]	RCV	overlap
165	3913797	3914222	426	141	0.306	embedded
167	3927557	3928027	471	156	0.476	3' partial (250 bp)
168	3942911	3942408	504	167	0.386	3' partial (382 bp) ^c
169	3955707	3955510	198	65	0.748	embedded
171	3962979	3963347	369	122	5.183	3' partial (313 bp)
172	4011928	4011818	111	36	0.208	5' partial (93 bp)
173	4030384	4030749	366	121	1.566	3' partial (285 bp)
174	4044641	4044465	177	58	0.207	3' partial (96 bp)
175	4063615	4063385	231	76	0.830	embedded
176	4063997	4063752	246	81	1.131	5' partial (200 bp)
177	4074561	4074199	363	120	0.321	embedded
178	4123722	4123534	189	62	0.368	embedded
179	4127122	4127562	441	146	0.734	3' partial (340 bp)
180	4130362	4130174	189	62	0.687	embedded
181	4150230	4150508	279	92	0.369	embedded
182 ^b	4174510	4174298	213	70	1.716	embedded
183	4234579	4234292	288	95	0.711	embedded
184	4243757	4243635	123	40	1.768	embedded
185	4285388	4285636	249	82	1.276	3' partial (197 bp)

Supplementary table S6: Continued from previous page

candidate	start	stop	length [nt]	length [aa]	RCV	overlap
186	4296457	4296320	138	45	1.439	3' partial (107 bp)
187	4336256	4336480	225	74	0.234	5' partial (150 bp)
188	4350011	4350136	126	41	0.991	embedded
189	4364636	4364884	249	82	3.692	embedded
190	4378141	4378500	360	119	0.316	5' partial (165 bp)
191	4408252	4408365	114	37	0.380	embedded
192	4409026	4409211	186	61	1.556	embedded
193	4491862	4491713	150	49	9.173	embedded
194	4495934	4496524	591	196	0.855	3' partial (585 bp)
195	4503343	4502990	354	117	0.552	5' partial (221 bp) ^c
196	4503508	4503344	165	54	0.115	embedded
197	4527492	4527262	231	76	0.677	embedded
198	4528288	4528163	126	41	1.160	embedded
199	4529845	4530102	258	85	0.159	3' partial (81 bp), elongates annotated gene
200	4604545	4604393	153	50	0.784	embedded
201	4614994	4614755	240	79	0.610	embedded
202	4615317	4614991	327	108	0.641	embedded
203	4746767	4746886	120	39	1.508	embedded
204	4771121	4770882	240	79	0.980	5' partial (137 bp) ^c

Supplementary table S6: Continued from previous page

candidate	start	stop	length [nt]	length [aa]	RCV	overlap
205 ^a	4781041	4781934	894	297	2.480	annotated genes embedded
206	4802350	4802568	219	72	0.597	embedded
207	4867875	4868060	186	61	1.065	3' partial (109 bp)
208 ^b	4900039	4900179	141	46	0.318	5' partial (66 bp)
209	4902482	4902231	252	83	0.331	embedded
210	4993551	4993366	186	61	0.472	5' partial (146 bp)
211	4999632	4999757	126	41	0.543	embedded
212	5013888	5014655	768	255	1.001	embedded
213	5058151	5057924	228	75	2.835	5' partial (137 bp)
214 ^b	5059030	5059323	294	97	0.103	3' partial (263 bp)
215	5154269	5154138	132	43	2.268	embedded
217	5177791	5177666	126	41	0.456	embedded
218	5205265	5205528	264	87	0.840	embedded
219	5215945	5215790	156	51	0.150	embedded
220	5252719	5252823	105	34	0.747	embedded
221	5260228	5260563	336	111	1.863	embedded
222	5264836	5264961	126	41	1.636	embedded
223 ^{a,b}	5270145	5271641	1497	498	0.378	3' partial (1434 bp)
224	5280370	5280462	93	30	0.237	embedded

Supplementary table S6: Continued from previous page

candidate	start	stop	length [nt]	length [aa]	RCV	overlap
225	5296461	5296631	171	56	0.544	embedded
226	5306929	5306609	321	106	1.259	3' partial (288 bp)
227	5318533	5318769	237	78	1.003	3' partial (160 bp)
228	5328673	5328341	333	110	0.332	annotated gene embedded
229	5340788	5341039	252	83	0.353	5' partial (107 bp)
230	5344162	5344320	159	52	0.161	3' partial (103 bp)
231	5353324	5353992	669	222	0.684	3' partial (525 bp) ^c
232	5364789	5365133	345	114	1.038	3' partial (208 bp)
235	5423818	5423618	201	66	0.586	embedded
236	5500863	5501012	150	49	0.613	embedded
237	5516294	5516034	261	86	2.335	3' partial (208 bp)
238	5534380	5534655	276	91	0.334	3' partial (130 bp)
239	5538104	5537943	162	53	1.193	embedded
240	5539285	5539440	156	51	0.305	5' partial (132 bp)
241	5540205	5540378	174	57	0.761	embedded
242	5540400	5540786	387	128	0.408	embedded

Supplementary table S7: Blastp analysis of overlapping gene candidates.


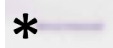


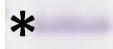



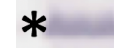





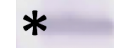



ID of overlapping gene candidates, homologous proteins deposited in the RefSeq database (Sept 2019) and genome where protein was found are listed.

candidate	homologous protein	genome
1	hypothetical protein	<i>Escherichia coli</i> TA206 supercont1.31
14	hypothetical protein	<i>Escherichia coli</i> strain 2d11B
30	hypothetical protein	<i>Escherichia coli</i> strain HT2012ST04 Scaffold25
32	hypothetical protein	<i>Escherichia coli</i> XH140A Contig54
51	hypothetical protein	<i>Escherichia coli</i> strain upec-34 upec-34_ctg_11666
75	hypothetical protein	<i>Escherichia coli</i> strain YDC774 scf7180000000021
79	PTS-dependent dihydroxyacetone kinase operon transcriptional regulator DhaR	<i>Shigella flexneri</i> 2a str. 301
80	hypothetical protein, partial	<i>Escherichia coli</i> strain HT2012173 Scaffold_cov_30.00_scaffold76
100	hypothetical protein	<i>Shigella dysenteriae</i> Sd197
101	methylated-DNA-[protein]-cysteine S-methyltransferase	<i>Lelliottia aquatilis</i> strain 9827-07 NODE_2_length_416864_cov_12.999
119	alpha-amylase	<i>Escherichia coli</i> strain TZ43_S scaffold_11
156	hypothetical protein	Multispecies: <i>Shigella</i>
171	IS3 family transposase	<i>Escherichia coli</i> UMN026
192	hypothetical protein	<i>Croceicoccus mobilis</i> strain Ery22 Ery22_C25
194	hypothetical protein, partial	<i>Shigella sonnei</i> strain 201312273_1
199	hypothetical protein	<i>Escherichia coli</i> strain MOD1-EC6540 MOD1-EC6540_81_length_16051_cov_32.7192
213	HTH-type transcriptional repressor FabR	<i>Escherichia coli</i> strain FWSEC0057 FWSEC0057_contig00269
224	conserved hypothetical protein	<i>Clavispora lusitaniae</i> ATCC 42720 scaffold_4 genomic scaffold
231	hypothetical protein	<i>Escherichia</i> sp. E4742 chromosome

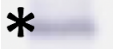









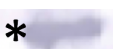


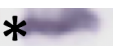
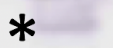



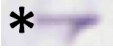

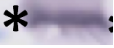

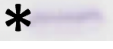

Supplementary table S8: Proteins of overexpressed overlapping genes in western blots.

Blot snippets show the most probable protein band for the indicated overlapping gene candidate. For candidates where more than one western blot exists, the blot with the highest quality was used for visualization. Original blots are shown in Supplementary file 1. The evaluation categories are specified as ‘single’ (single protein band), ‘bkg.’ (high background signal), ‘smear’ (smeared protein bands), and ‘by-pr.’ (clear by-products). Masses indicate the *theoretical* and experimentally determined molecular weight of the protein in kDa.











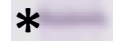













^a: image correction performed

candidate	protein	cat.	mass	candidate	protein	cat.	mass	candidate	protein	cat.	mass
OGC 1		bkg.	<i>25.23</i> 25.17	OGC 3		by-pr.	<i>26.20</i> 28.95	OGC 4		single	<i>19.86</i> 19.97
OGC 5		smear	<i>14.20</i> 14.41	OGC 6		single	<i>13.80</i> 17.41	OGC 7 ^a		single	<i>12.33</i> 13.93
OGC 8		single	<i>12.16</i> 13.57	OGC 9		bkg.	<i>30.79</i> 29.88	OGC 10		single	<i>16.64</i> 18.84
OGC 11		smear	<i>15.72</i> 18.20	OGC 13		bkg.	<i>15.54</i> 16.85	OGC 14		bkg.	<i>14.84</i> 16.14
OGC 15		single	<i>12.93</i> 16.45	OGC 16		bkg.	<i>26.18</i> 25.67	OGC 17		bkg.	<i>20.90</i> 22.84
OGC 18 ^a		single	<i>13.28</i> 15.60	OGC 19 ^a		single	<i>24.64</i> 23.55	OGC 20 ^a		single	<i>12.08</i> 15.85

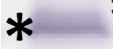

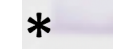











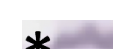





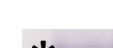



Supplementary table S8: Continued from previous page

candidate	protein	cat.	mass	candidate	protein	cat.	mass	candidate	protein	cat.	mass
OGC 21		single	19.48 24.74	OGC 22		bkg.	16.59 18.78	OGC 23		single	15.33 16.24
OGC 24		bkg.	14.07 18.52	OGC 25		single	12.16 16.24	OGC 26		bkg.	14.14 16.77
OGC 27		single	14.44 18.39	OGC 28		single	12.46 15.99	OGC 29		single	16.68 19.89
OGC 30		single	22.61 19.21	OGC 32		bkg.	19.20 22.09	OGC 33		by-pr.	20.17 23.16
OGC 34		single	12.89 16.28	OGC 35		single	13.81 18.38	OGC 39		by-pr.	17.21 20.01
OGC 40		single	15.42 18.04	OGC 41		single	16.69 21.78	OGC 42		bkg.	15.05 18.24
OGC 43		single	30.16 36.84	OGC 44		bkg.	15.31 17.45	OGC 45		bkg.	13.80 16.84
OGC 46		bkg.	23.06 30.27	OGC 47		single	12.89 16.40	OGC 48		single	17.55 20.77













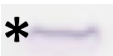









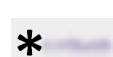

Supplementary table S8: Continued from previous page

candidate	protein	cat.	mass	candidate	protein	cat.	mass	candidate	protein	cat.	mass
OGC 50		single	<i>12.55</i> 15.75	OGC 51		single	<i>15.82</i> 15.84	OGC 55 ^a		single	<i>23.78</i> 31.35
OGC 56		single	<i>13.69</i> 18.87	OGC 57			<i>37.53</i> 35.23	OGC 58 ^a		bkg.	<i>13.92</i> 17.83
OGC 59			<i>17.48</i> 18.71	OGC 68 ^a		single	<i>11.89</i> 16.44	OGC 69		bkg.	<i>15.64</i> 17.83
OGC 70		bkg.	<i>11.93</i> 14.38	OGC 71		single	<i>12.59</i> 15.58	OGC 72		single	<i>15.79</i> 18.68
OGC 73		by-pr.	<i>12.29</i> 15.87	OGC 74		single	<i>13.22</i> 15.66	OGC 75		by-pr.	<i>20.45</i> 21.24
OGC 76		by-pr.	<i>24.99</i> 30.96	OGC 77		bkg.	<i>22.76</i> 24.93	OGC 78		by-pr.	<i>14.34</i> 18.68
OGC 79		bkg.	<i>25.67</i> 27.67	OGC 80 ^a		single	<i>14.74</i> 18.10	OGC 81		bkg.	<i>15.61</i> 21.93
OGC 82 ^a		single	<i>13.91</i> 17.95	OGC 83		bkg.	<i>24.64</i> 26.20	OGC 84		by-pr.	<i>16.25</i> 18.48










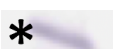


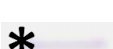


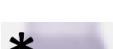








Supplementary table S8: Continued from previous page

candidate	protein	cat.	mass	candidate	protein	cat.	mass	candidate	protein	cat.	mass
OGC 85		single	<i>11.81</i> 15.65	OGC 86		single	<i>14.40</i> 17.01	OGC 88		single	<i>14.07</i> 15.92
OGC 89		bkg.	<i>33.23</i> 32.79	OGC 90 ^a		bkg.	<i>18.57</i> 19.64	OGC 91		bkg.	<i>16.04</i> 19.26
OGC 92		single	<i>11.69</i> 14.05	OGC 93		single	<i>15.45</i> 17.03	OGC 94 ^a		single	<i>13.87</i> 16.66
OGC 96		single	<i>12.18</i> 13.84	OGC 98		bkg.	<i>18.46</i> 19.67	OGC 100		bkg.	<i>17.94</i> 18.44
OGC 102		single	<i>20.98</i> 22.24	OGC 103		bkg.	<i>14.62</i> 15.39	OGC 104		single	<i>15.51</i> 17.47
OGC 105		single	<i>12.88</i> 15.89	OGC 106 ^a		bkg.	<i>19.17</i> 20.34	OGC 107		single	<i>12.53</i> 13.84
OGC 108		bkg.	<i>14.74</i> 16.36	OGC 109 ^a		single	<i>12.37</i> 14.70	OGC 110 ^a		by-pr.	<i>13.76</i> 18.71
OGC 111		single	<i>14.64</i> 19.11	OGC 112		bkg.	<i>11.50</i> 14.68	OGC 113		bkg.	<i>15.39</i> 18.69


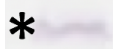





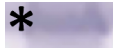
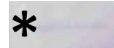

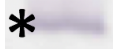





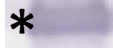




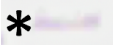
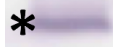

Supplementary table S8: Continued from previous page

candidate	protein	cat.	mass	candidate	protein	cat.	mass	candidate	protein	cat.	mass
OGC 114		bkg.	<i>14.22</i> 15.94	OGC 115		bkg.	<i>13.51</i> 16.96	OGC 116		single	<i>20.76</i> 24.42
OGC 117		bkg.	<i>14.50</i> 16.62	OGC 118		single	<i>26.40</i> 28.08	OGC 119		single	<i>15.57</i> 19.45
OGC 121		smear	<i>15.28</i> 16.16	OGC 123		bkg.	<i>12.97</i> 20.10	OGC 125		bkg.	<i>14.00</i> 16.70
OGC 126		bkg.	<i>19.62</i> 22.68	OGC 128		single	<i>15.25</i> 17.81	OGC 129 ^a		single	<i>45.60</i> 45.23
OGC 130 ^a		single	<i>12.92</i> 15.96	OGC 131		bkg.	<i>20.01</i> 23.34	OGC 132		single	<i>15.45</i> 20.06
OGC 133		bkg.	<i>16.64</i> 19.92	OGC 134 ^a		bkg.	<i>17.31</i> 21.07	OGC 135		single	<i>25.09</i> 28.68
OGC 136		single	<i>16.34</i> 19.28	OGC 137		single	<i>19.11</i> 20.26	OGC 138		smear	<i>19.57</i> 24.69
OGC 139		bkg.	<i>21.44</i> 23.00	OGC 140		single	<i>16.36</i> 19.40	OGC 141		smear	<i>25.58</i> 29.33

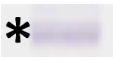









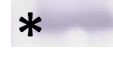
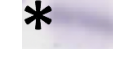






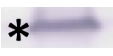





Supplementary table S8: Continued from previous page

candidate	protein	cat.	mass	candidate	protein	cat.	mass	candidate	protein	cat.	mass
OGC 142		single	15.23 18.32	OGC 143		bkg.	17.47 21.98	OGC 144		by-pr.	12.27 16.59
OGC 145 ^a		single	20.24 29.57	OGC 146		single	13.79 17.77	OGC 147		single	13.90 17.20
OGC 148		single	19.79 24.73	OGC 149		by-pr.	19.95 23.58	OGC 150		single	24.41 25.94
OGC 152		single	16.92 18.32	OGC 153		single	12.24 15.37	OGC 154		single	12.45 15.66
OGC 156		single	17.05 18.83	OGC 157		single	17.03 21.19	OGC 158		smear	15.03 17.04
OGC 159		bkg.	14.20 16.24	OGC 160		bkg.	16.49 22.06	OGC 162		bkg.	23.53 26.99
OGC 163		bkg.	15.38 19.55	OGC 164		single	13.95 17.26	OGC 165 ^a		bkg.	24.58 28.01
OGC 167 ^a		by-pr.	25.91 31.00	OGC 168		single	26.05 29.12	OGC 169		bkg.	15.65 18.08







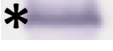
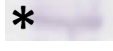


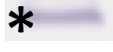

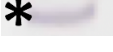
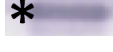

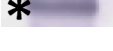
Supplementary table S8: Continued from previous page

candidate	protein	cat.	mass	candidate	protein	cat.	mass	candidate	protein	cat.	mass
OGC 171		by-pr.	<i>21.48</i> 25.50	OGC 172		by-pr.	<i>12.33</i> 14.27	OGC 173		single	<i>22.13</i> 28.45
OGC 174		single	<i>14.84</i> 17.90	OGC 175		single	<i>16.41</i> 18.68	OGC 176		single	<i>17.31</i> 23.18
OGC 177		by-pr.	<i>21.72</i> 22.81	OGC 178		bkg.	<i>14.97</i> 19.13	OGC 179		single	<i>23.06</i> 29.47
OGC 180 ^a		single	<i>14.85</i> 16.25	OGC 181		single	<i>18.75</i> 24.19	OGC 182		bkg.	<i>15.85</i> 20.89
OGC 183		single	<i>19.31</i> 21.29	OGC 184		single	<i>12.57</i> 15.08	OGC 185		single	<i>18.09</i> 23.47
OGC 186		single	<i>12.77</i> 16.34	OGC 187 ^a		bkg.	<i>16.73</i> 19.67	OGC 188		single	<i>12.62</i> 15.91
OGC 189		single	<i>16.90</i> 17.53	OGC 190		bkg.	<i>21.56</i> 29.80	OGC 191		single	<i>12.21</i> 16.47
OGC 192		single	<i>14.54</i> 18.17	OGC 193 ^a		bkg.	<i>13.47</i> 17.84	OGC 194		bkg.	<i>30.41</i> 33.62

Supplementary table S8: Continued from previous page

candidate	protein	cat.	mass	candidate	protein	cat.	mass	candidate	protein	cat.	mass
OGC 195		bkg.	<i>21.36</i> 25.32	OGC 196		single	<i>14.00</i> 17.07	OGC 197		by-pr.	<i>16.53</i> 19.68
OGC 198		single	<i>12.64</i> 18.83	OGC 199 ^a		bkg.	<i>17.41</i> 19.63	OGC 200		single	<i>13.59</i> 17.74
OGC 201		single	<i>16.77</i> 21.19	OGC 202		single	<i>20.51</i> 24.68	OGC 203		single	<i>12.43</i> 16.16
OGC 204 ^a		by-pr.	<i>16.90</i> 19.47	OGC 206		smear	<i>16.35</i> 20.14	OGC 207		bkg.	<i>14.70</i> 17.29
OGC 208		bkg.	<i>13.98</i> 18.33	OGC 209 ^a		bkg.	<i>17.48</i> 20.92	OGC 210		bkg.	<i>14.68</i> 18.01
OGC 212		single	<i>35.35</i> 40.42	OGC 213		single	<i>16.50</i> 20.19	OGC 215		by-pr.	<i>13.23</i> 18.01
OGC 217		single	<i>12.66</i> 14.51	OGC 218		single	<i>17.98</i> 20.15	OGC 219		single	<i>14.09</i> 16.93
OGC 220		single	<i>12.16</i> 14.27	OGC 221		bkg.	<i>20.57</i> 24.24	OGC 222		single	<i>12.29</i> 14.77

Supplementary table S8: Continued from previous page

candidate	protein	cat.	mass	candidate	protein	cat.	mass	candidate	protein	cat.	mass
OGC 224		single	<i>11.38</i> 14.15	OGC 225		single	<i>14.28</i> 16.08	OGC 226		by-pr.	<i>19.95</i> 24.02
OGC 227		bkg.	<i>17.00</i> 21.14	OGC 228		bkg.	<i>20.16</i> 22.22	OGC 229		single	<i>17.48</i> 21.14
OGC 230		bkg.	<i>13.51</i> 16.32	OGC 231		by-pr.	<i>34.67</i> 34.83	OGC 232		single	<i>21.15</i> 26.47
OGC 235		single	<i>15.50</i> 18.65	OGC 236		single	<i>13.27</i> 15.14	OGC 237		bkg.	<i>17.24</i> 21.05
OGC 238		single	<i>18.73</i> 24.72	OGC 239		single	<i>14.38</i> 16.44	OGC 241		single	<i>14.86</i> 16.44
OGC 242		by-pr.	<i>23.10</i> 24.97								

Supplementary table S9: Graphical summary of HT and LT phenotyping.

Overlapping gene candidates with phenotypes within the high-throughput approach are listed. Primary stress conditions fulfill the phenotype criterion $|z| \geq 2$ in at least two biological replicates, secondary conditions (listed in square brackets) were additionally tested in the LT approach although the phenotype criterion is not met. Phenotypic profiles visualize advantageous ($z \geq 2$) or disadvantageous ($z \leq -2$) relative growth effects for different stress conditions. Graphs of LT phenotyping show competitive growth of cells overexpressing wild type (blue bars) or translationally arrested (gray bars) copy of overlapping ORFs. Values are normalized for visualization to 50:50 input ratio (blue dotted line). Phenotypes of OGCs in bold are selected in HT and reproduced in LT phenotyping.

^a growth conditions: 0 t₀, 1 LB, 2 glucose, 3 L-malic acid, 4 L-arginine, 5 CsCl, 6 acetic acid, 7 malonic acid, 8 1-methylimidazole, 9 NaCl, 10 NaOH, 11 Na₃VO₄, 12 sodium salicylate, 13 HClO₄, 14 phytic acid, 15 1,2-propanediol, 16 1-propanol, 17 pyridoxine HCl, 18 *Staphylococcus*, 19 ZnCl₂

^b HT phenotype reproduced in single competitive growth assays, verified by statistical analysis

^c HT phenotype reproduced in single competitive growth assays, verified by visual inspection

^d phenotype in secondary stress condition detected

^e phenotype artifact, details in Section 3.4.3

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
3 ^e	HClO ₄		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
6	<i>Staphylococcus</i>		
15 ^b	NaCl		
18	NaCl		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
23 ^b	Na ₃ VO ₄		
24	CsCl		
25	glucose		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
26	Na ₃ VO ₄		
30	NaCl		
31	NaCl		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
44	Na ₃ VO ₄		
50	<i>Staphylococcus</i>		
51 ^b	Na ₃ VO ₄		

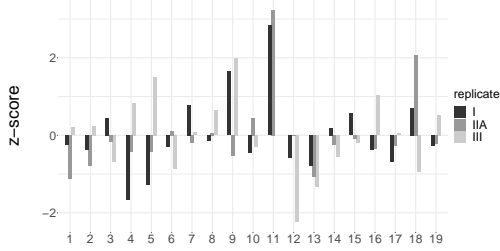
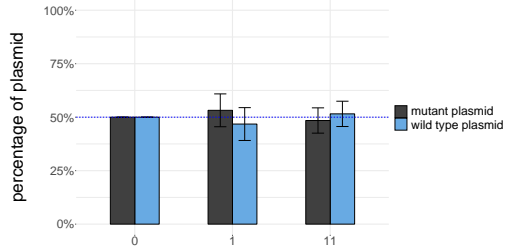
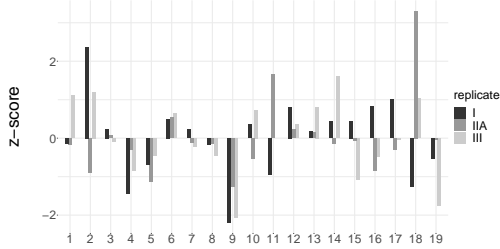
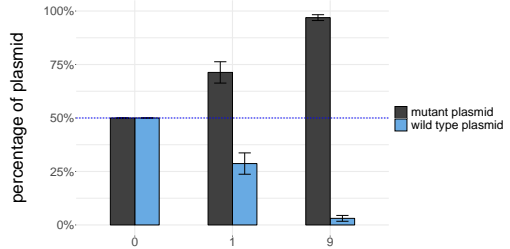
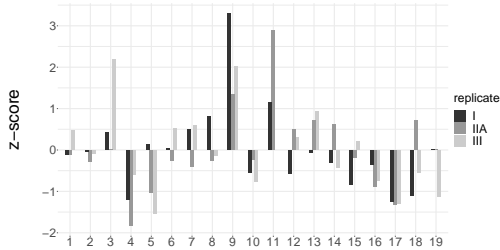
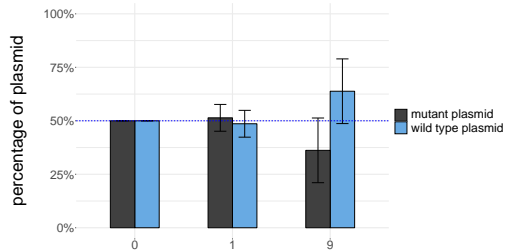
Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
57	L-malic acid [acetic acid]		detailed characterization in Section 3.7
59	NaCl		detailed characterization published by Vanderhaeghen <i>et al.</i> , 2018
68	NaCl		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
71	sodium salicylate		
75 ^c	NaCl		
85 ^c	CsCl		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
96	Na ₃ VO ₄		
106 ^b	NaCl		
107	NaCl		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
116	malonic acid		
117	sodium salicylate		
119	LB [ZnCl ₂]		

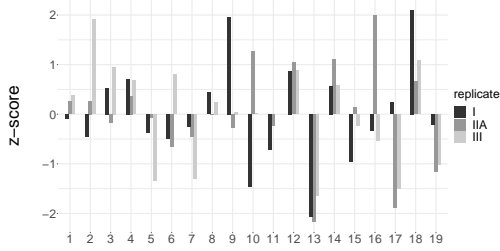
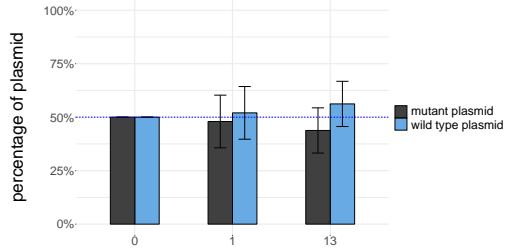
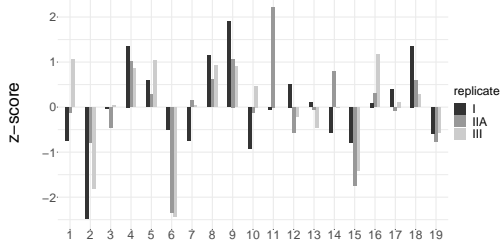
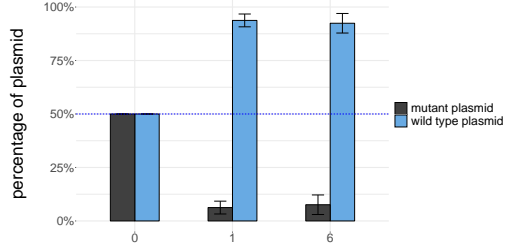
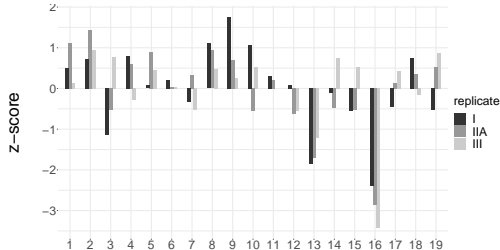
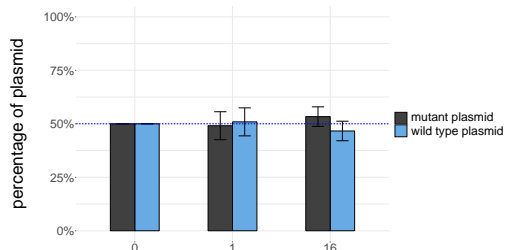
Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
121 ^b	L-malic acid [malonic acid]		
137	malonic acid		
139	CsCl		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
140 ^e	malonic acid		
145	<i>Staphylococcus</i>		
146	L-malic acid		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
147	HClO ₄		
153 ^e	acetic acid		
164	1-propanol		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
167 ^c	ZnCl ₂		
172	NaCl		
174 ^c	NaCl		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
177	1-propanol		
178	NaCl		
183	Na ₃ VO ₄		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
186	<i>Staphylococcus</i>		
191	HClO ₄		
194 ^b	NaCl		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
195	<i>Staphylococcus</i>		
198	malonic acid		
205	Na ₃ VO ₄		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
213	NaCl		
217	NaCl		
218	malonic acid		

Supplementary table S9: Continued from previous page

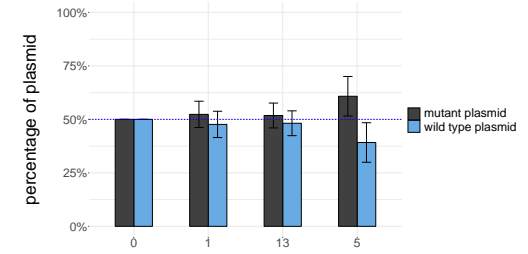
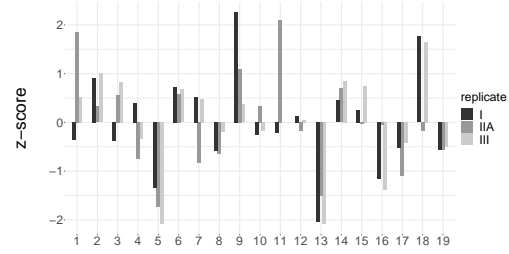
candidate	stress	HT phenotyping ^a	LT phenotyping ^a
226 ^b	L-malic acid [CsCl]		
231 ^b	NaCl [L-malic acid]		
232	<i>Staphylococcus</i>		

Supplementary table S9: Continued from previous page

candidate	stress	HT phenotyping ^a	LT phenotyping ^a
-----------	--------	-----------------------------	-----------------------------

241^{b,d}

HClO₄
[CsCl]



Supplementary table S10: Genome wide transcription start sites determined with Cappable-seq.

$minRRS = 1.5$

exponential	I	II	III	stationary	I	II	III
LB	13533	13689	11115	LB	17238	14367	16374
MM	16228	13036	13887	MM	19831	19547	18540
acid	17033	11142	11625	acid	15484	11125	16240
salt	13569	12734	14751	salt	18574	15318	17593

$minRRS = 5$

exponential	I	II	III	stationary	I	II	III
LB	5606	5830	4919	LB	7660	6270	7087
MM	7066	5848	5780	MM	8369	7590	7432
acid	7212	4918	4904	acid	6957	4924	7183
salt	5744	5515	6213	salt	7820	6764	7824

Supplementary table S11: Gene associated TSS for OGCs.

Genome positions of TSS are given for each candidate in analyzed growth conditions and growth phases. TSS were determined with the selection criterion $minRRS = 1.5$. The maximum distance of TSS and start codon of the associated OGC is 250 bp.

^a growth conditions: LB medium (LB), M9 minimal medium (MM), LB medium supplemented with 4 mM L-malic acid (acid), LB medium supplemented with 500 mM NaCl (salt).

^b TSS for indicated candidates were misannotated and belong most likely to annotated genes.

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
1		5858						
5		146134				146134		146134
7		152541						
8		152541						
10	156284	156284	156284	156284		156284		156284
11					226796	226796	226796	226796
12 ^b	238497		238497	238497	238497	238497, 238506		238497
14						284813, 284820		

Supplementary table S11: Continued from previous page

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
15	300491	300491	300491	300491	300491	300491	300491	300491
17					459613			
18					487949		487949	
19	536847	536847	536847	536847	536847	536847	536847	536847
20 ^b	552044, 552150	552044, 552150	552044, 552150	552044, 552150	552044, 552090, 552150	552044, 552150	552044, 552150	552044, 552150
21	557333	557333	557333	557333	557324, 557333	557324, 557333	557333	557333, 557395
22	569126, 569147	569126, 569147	569126	569126		569126, 569147		569126
23 ^b					570334			570334
24					572144		572144	
28					605352	605352	605352, 605412	

Supplementary table S11: Continued from previous page

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
29				620886				620886
30		690488		690488		690488	690488	690488
31						690725		
33					745155		745155	
35	783061, 783125	783061, 783125	783061, 783125	783061, 783125	783125	783061, 783125	783125	783061, 783125
40	847582							
41 ^b		878124		878124		878124	878124	878124
42						887079		
44						974914		
45					1005016	1005016	1005016	
47		1053207			1053207		1053207	
48		1053207			1053207		1053207	
50	1068854	1068854	1068854	1068854	1068854	1068854	1068854	1068854

Supplementary table S11: Continued from previous page

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
51		1111018						
56	1224125		1224125	1224125	1224125	1224125		1224125
58 ^b	1237008, 1237105	1237008, 1237105	1237105	1237008, 1237105	1237077, 1237105	1237077, 1237105	1237077, 1237105	1237008, 1237105
70 ^b	1591568, 1591602	1591520, 1591568, 1591602	1591520, 1591568, 1591602	1591520, 1591568, 1591602	1591520, 1591568, 1591602	1591520, 1591568, 1591602	1591520, 1591568, 1591602	1591568, 1591602
71 ^b	1591602, 1591718	1591602, 1591718	1591602, 1591718	1591602, 1591718	1591602, 1591718	1591602, 1591718	1591602, 1591718	1591602, 1591718
73	1753780		1753780		1753636	1753636		1753791
74	1753780		1753780					1753791
75	1754009	1754009		1754009	1754009	1754009	1754009	1753791, 1754009
76	1759709	1759709	1759709	1759709	1759709		1759709	1759709

Supplementary table S11: Continued from previous page

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
79	1785119, 1785151	1785119	1785119	1785119, 1785152	1785119	1785119	1785119	1785119, 1785152
81	1820368, 1820395	1820368, 1820395	1820368, 1820395	1820368, 1820395	1820368, 1820395	1820368, 1820395	1820368, 1820395	1820368, 1820395
82	1950922	1950921	1950922		1950990			
84		1985698		1985698	1985698	1985698		1985698, 1985829
85	1985980	1985980	1985980	1985980	1985980	1985980	1985980	1985980
86					2070801	2070801		
88		2086676		2086676	2086676	2086676	2086676	2086676
89	2097839	2097839			2097818, 2097839	2097818, 2097839	2097818, 2097839	2097818
91 ^b	2194678	2194505, 2194678	2194678	2194678	2194505, 2194678	2194505, 2194678	2194505, 2194678	2194505, 2194678
92		2200066						

Supplementary table S11: Continued from previous page

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
93					2200262		2200262	
96	2285637	2285498, 2285637	2285637	2285498, 2285637		2285498, 2285637		2285498, 2285573
98	2349088	2349088	2349088	2349088	2349088	2349088	2349088	2349088
100	2351146	2351146	2351146	2351146	2351129, 2351146	2351146	2351146	2351129, 2351146
101	2365331, 2365394	2365394	2365394	2365394	2365331, 2365394	2365331, 2365394	2365331, 2365385, 2365394	2365331, 2365394
102								2427678
103 ^b					2451618, 2451681	2451635	2451635	
111	2570242	2570242	2570242	2570242	2570242	2570242	2570242	2570242
113	2574619	2574619	2574619	2574619		2574619		2574619
114	2590681	2590681	2590681	2590681	2590681	2590681	2590681	2590681

Supplementary table S11: Continued from previous page

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
115				2591539	2591539	2591539	2591539	2591539
117	2633176	2633176	2633176	2633176	2633176	2633176		2633176
119	2724503	2724503	2724503	2724503	2724503	2724503	2724503	
121								2758419
123					2850767	2850767	2850767	
124	2879448	2879448			2879448			2879448
126	2922960	2922960	2922960	2922960				2922960
128 ^b	3032829	3032829	3032829	3032829	3032829	3032829	3032829, 3032844	3032829, 3032844
129						3036985		
131					3094487		3094487	
132	3126382	3126382	3126382	3126382		3126382		3126382
133	3186891	3186795, 3186891	3186795, 3186891	3186891	3186891	3186795, 3186891	3186891	3186891

Supplementary table S11: Continued from previous page

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
135	3218525	3218525, 3218616, 3218644	3218525	3218525	3218644, 3218689	3218689	3218644	3218644, 3218689
136	3226857, 3226911, 3226968	3226857, 3226911, 3226968	3226857, 3226911, 3226968	3226857, 3226911, 3226968	3226857, 3226911, 3226968	3226857, 3226911, 3226968	3226857, 3226911, 3226968	3226857, 3226911, 3226968
137					3234464	3234464		
138	3325070, 3325206	3325070, 3325206	3325206	3325070, 3325206	3325070, 3325206	3325070, 3325206	3325070, 3325206	3325070, 3325206
140							3395817	
141					3544805	3544805	3544805	
144	3606221	3606221	3606221	3606221	3606221	3606174, 3606221		3606221
145					3614707	3614707	3614707	3614707, 3614785

Supplementary table S11: Continued from previous page

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
146					3614707	3614707	3614707	3614707, 3614785
147			3620991					
152	3664050	3664050	3664050		3664071	3664050	3664050, 3664071	
153					3664071		3664071	
156	3711337	3711337	3711337	3711337	3711337	3711337	3711337	3711337
157	3724490, 3724543	3724490, 3724543	3724543		3724543, 3724553			3724490, 3724543, 3724553
158	3793535	3793535		3793535	3793535	3793535	3793535	3793535
159	3793535	3793535		3793535	3793535	3793535	3793535	3793535
160								3854027
164			3913738			3913738		

Supplementary table S11: Continued from previous page

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
165			3913738			3913738		
171 ^b	3962829	3962829	3962829	3962829	3962829	3962829	3962829	3962829
172 ^b	4012032, 4012074		4012032		4012074	4012074	4012074	4012074
174						4044652		
177	4074735		4074735	4074735				
182					4174585			
183	4234712	4234712		4234712		4234712		4234712
185 ^b	4285149	4285149	4285149	4285149	4285149	4285149	4285149	4285149
187					4336246		4336246	4336246
189			4364424			4364424, 4364614		
190		4377914, 4378087	4377914	4377914, 4378088		4378087		4378087

Supplementary table S11: Continued from previous page

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
191	4408102	4408102	4408102	4408102	4408239	4408102, 4408239	4408099, 4408239	4408102
195			4503429		4503429	4503429	4503429	
200	4604638	4604638	4604638	4604638				
201	4615094	4615094	4615094	4615094	4615094	4615094	4615094	4615094
203						4746709		
206	4802304				4802304	4802304	4802304	4802304
207	4867699	4867699	4867699	4867699	4867699	4867699	4867699	4867699
208	4899935		4899935	4899935				4899935
210					4993618	4993618		4993618
212	5013828		5013828					5013797, 5013828
214	5058977	5058977	5058977	5058977	5058977	5058977	5058977	5058977
215	5154301	5154301	5154301	5154301		5154301		5154301

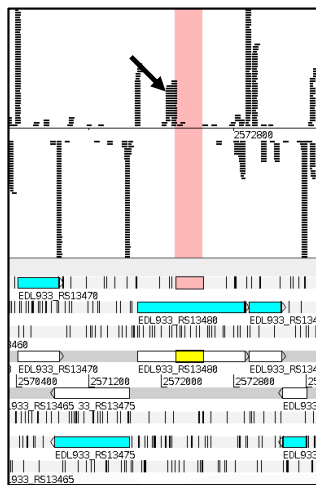
Supplementary table S11: Continued from previous page

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
217			5177839			5177839		
218								5205263
219	5215958	5215958	5215958	5215958	5215958	5215958	5215958	5215958
220 ^b								5252536
221	5260097	5260040	5260040, 5260097	5260097				5260097
222				5264685, 5264750				5264750
225 ^b	5296268, 5296368	5296251, 5296268, 5296368	5296268, 5296368	5296268	5296268	5296251, 5296268	5296268	5296268, 5296368
228					5328686	5328686	5328686	
230	5344039	5344039	5344039	5344039	5344039	5344039	5343974, 5344039	5344039
235	5423923	5423923	5423923	5423923	5423923	5423923	5423923	5423923

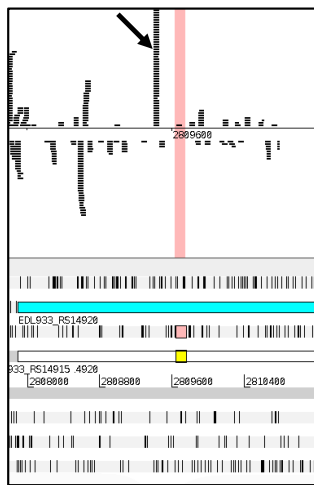
Supplementary table S11: Continued from previous page

OGC	exponential phase ^a				early stationary phase ^a			
	LB	MM	acid	salt	LB	MM	acid	salt
237	5516318, 5516399		5516399	5516399		5516399		5516399
238	5534257	5534257	5534257	5534257	5534257	5534257		
239					5538351	5538345		
241	5539957		5539957	5539957		5539957		5539957

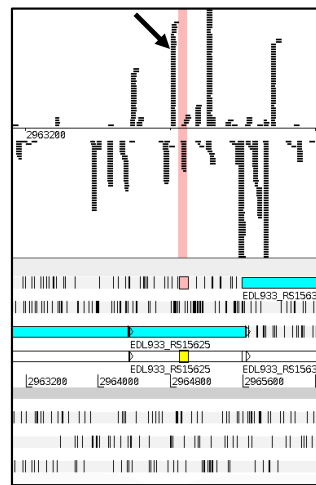
Supplementary table S12: Continued from previous page



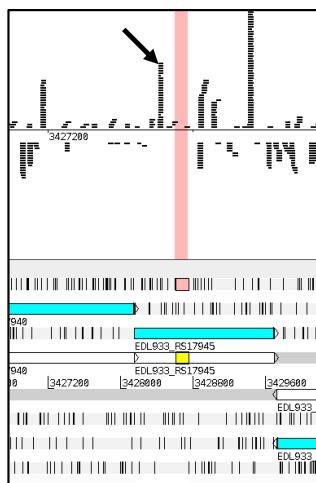
sID_16557



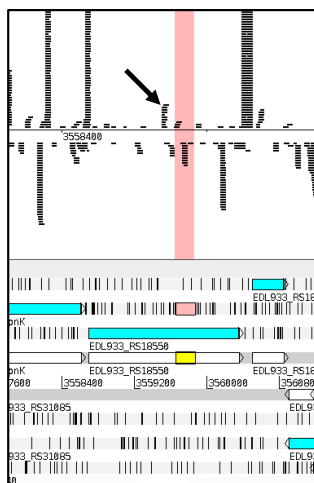
sID_18263



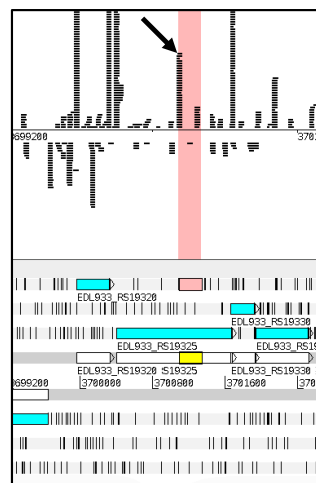
sID_19269



sID_22528

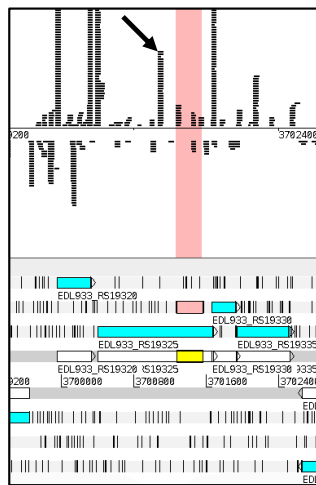


sID_23476

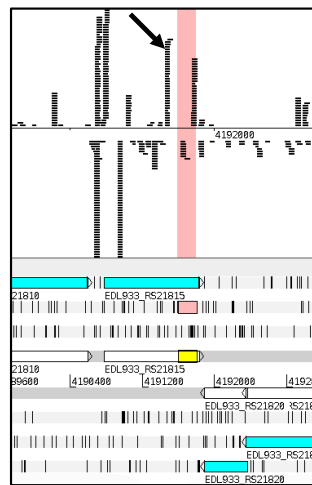


sID_24441

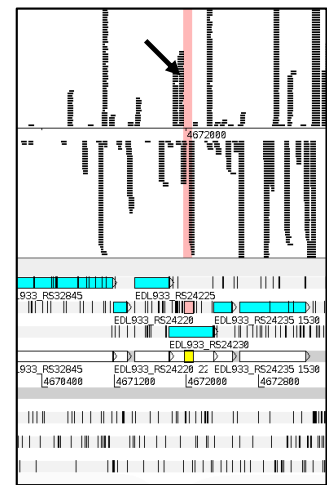
Supplementary table S12: Continued from previous page



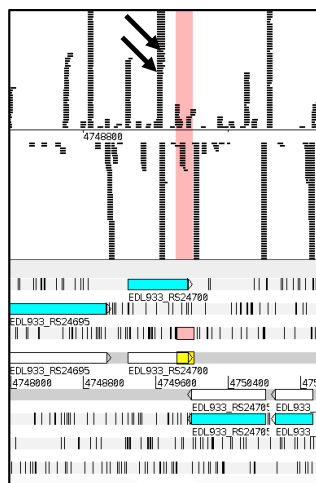
sID_24442



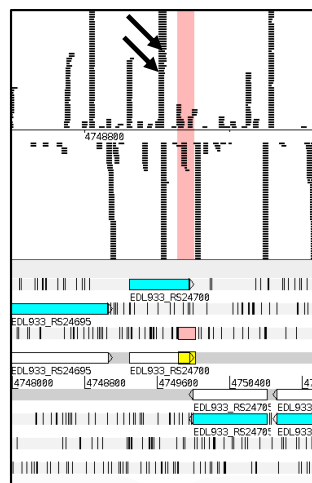
sID_27733



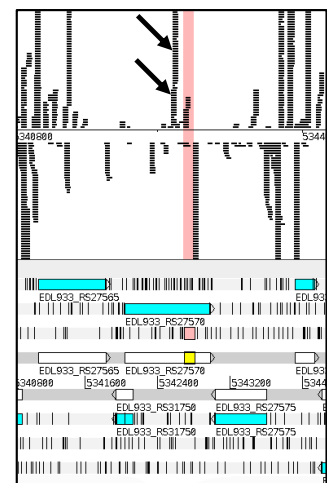
sID_30910



sID_31442

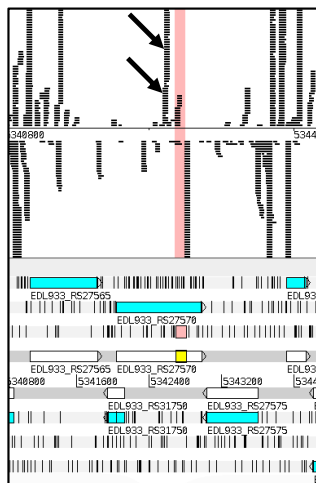


sID_31442

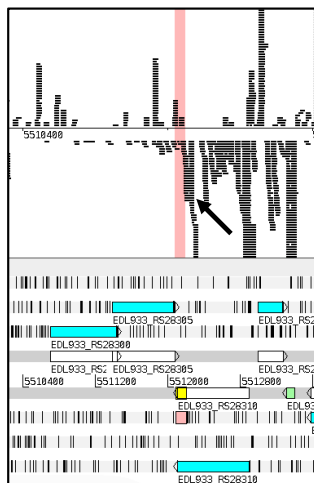


sID_35348

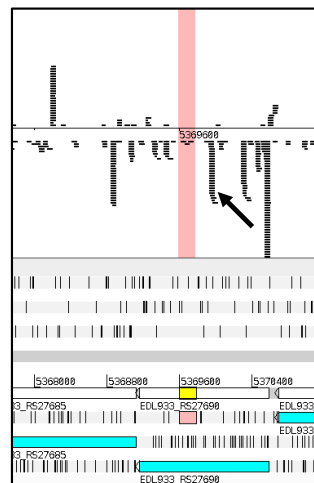
Supplementary table S12: Continued from previous page



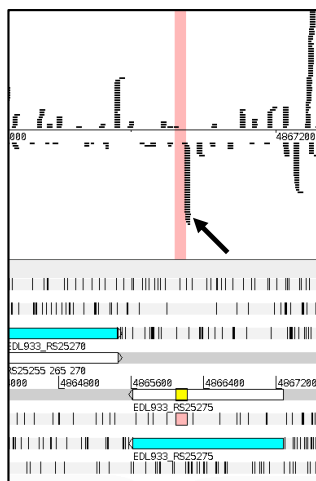
sID_35348



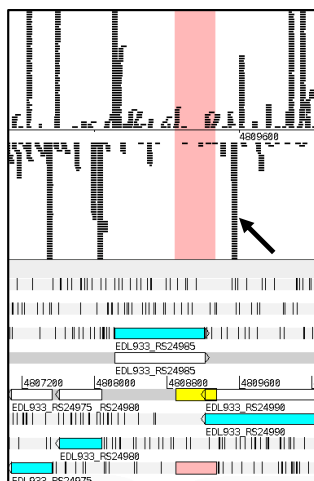
sID_36884



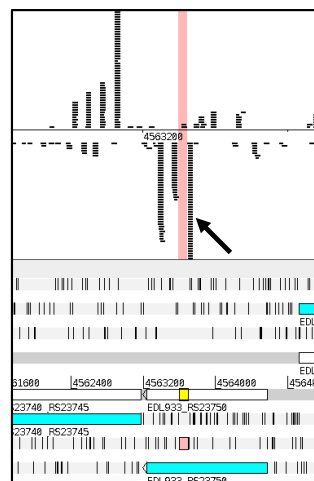
sID_37787



sID_41173

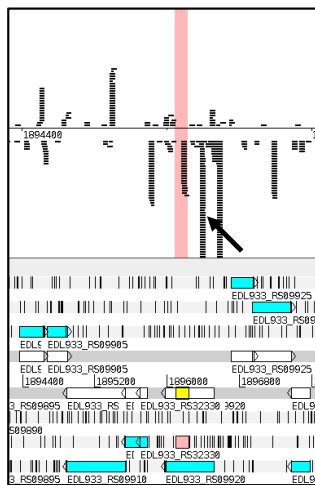


sID_41605

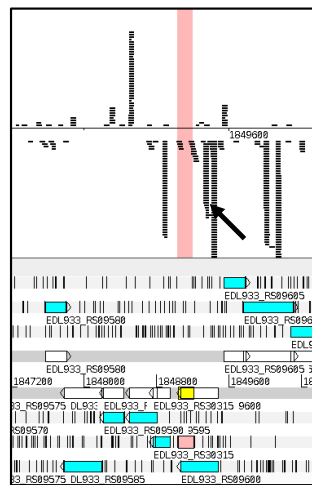


sID_43081

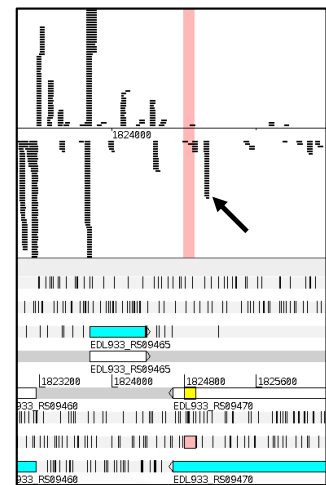
Supplementary table S12: Continued from previous page



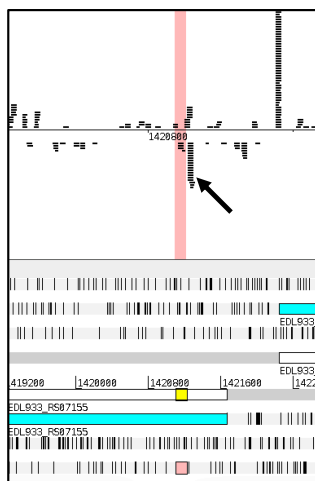
sID_60435



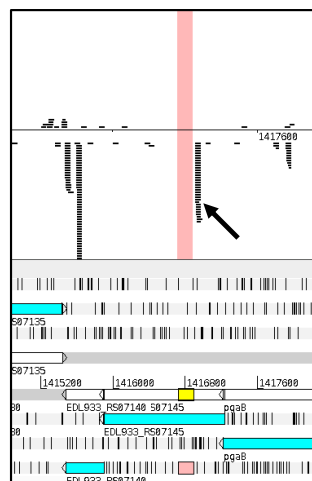
sID_60810



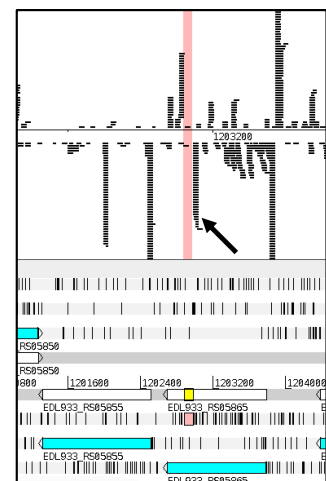
sID_60970



sID_63851

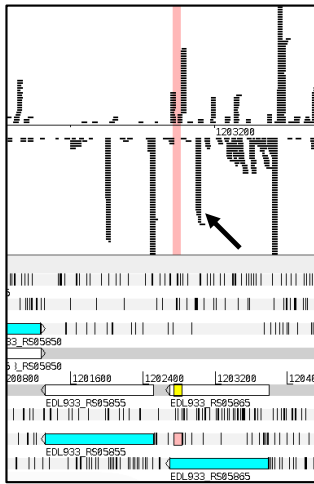


sID_63872

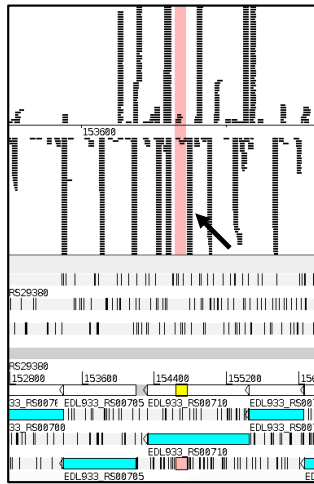


sID_65424

Supplementary table S12: Continued from previous page



sID_65425



sID_72617

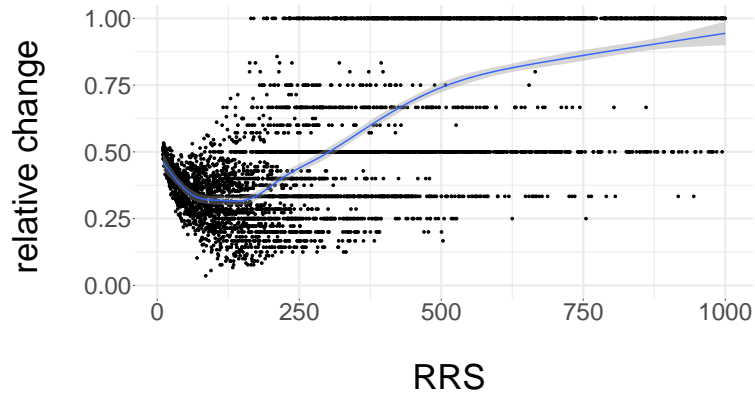
Supplementary table S13: Evaluation of ribosome profiling and RNAseq of *pop* genomic region.

Experiments were conducted by Landstorfer, 2014, and data were published by Neuhaus *et al.*, 2017. Chromosome coordinates of the open reading frames in the genomic region of *pop* are given (orientation in brackets). Read count and RPKM values of two replicates of ribosome profiling (light gray, sample numbers in sequence read archive: SRR5266618 (upper part), SRR5266620 (lower part)) and RNA-seq (dark gray, sample numbers in sequence read archive: SRR5266617 (upper part), SRR5266619 (lower part)) were determined and RCV values were calculated ($\frac{\text{RPKM ribosome profiling}}{\text{RPKM RNA-seq}}$).

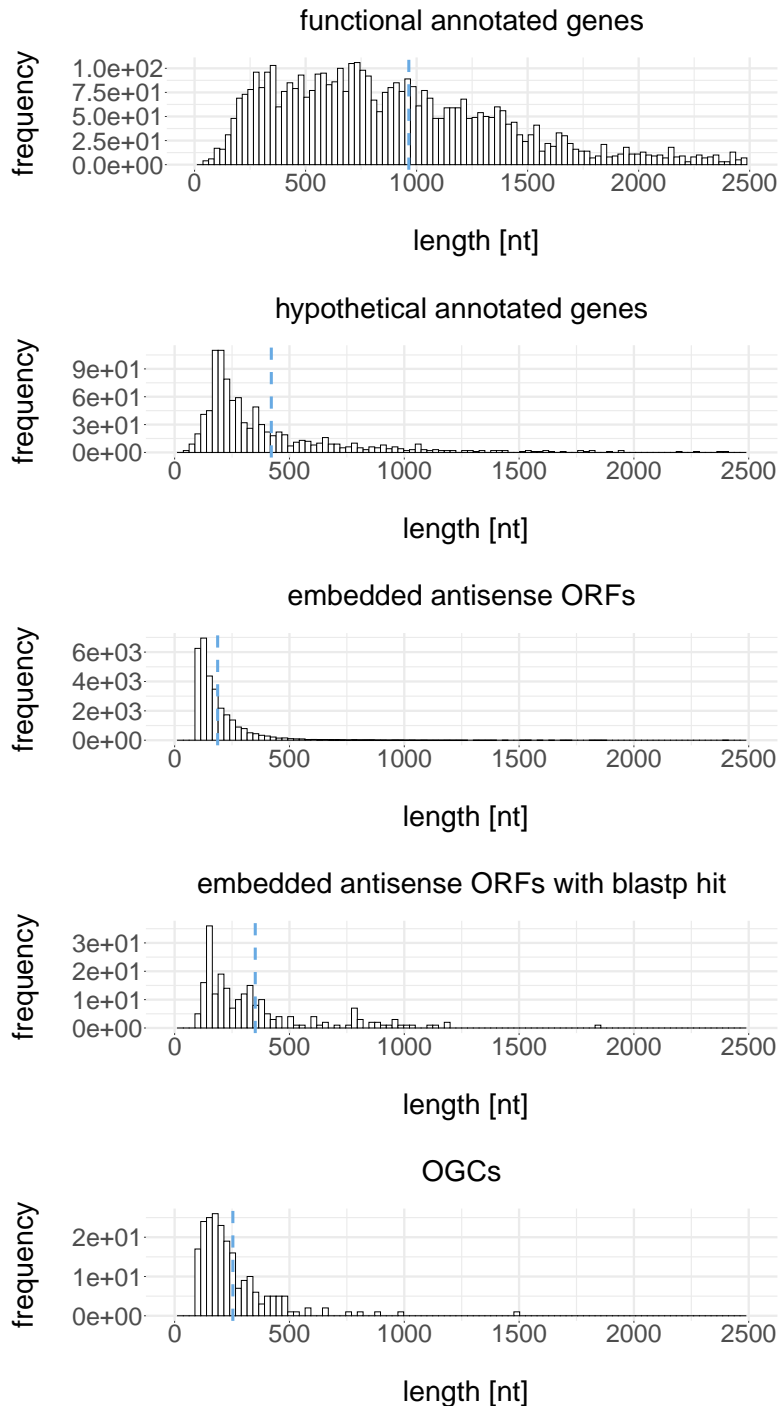
	Feature	Start	Stop	Read Count	RPKM	Read Count	RPKM	RCV
Replicate 1	ycbG (+)	1235288	1235740	141	165.55	212	144.86	1.14
	ompA (-)	1235816	1236856	24725	12632.70	9758	2901.43	4.35
	pop (+)	1236020	1236622	99	87.32	93	47.74	1.83
	d. ORF 1 (+)	1236662	1236892	0	0	3	4.02	NA
	d. ORF 2 (+)	1236714	1236950	0	0	2	2.61	NA
Replicate 2	ycbG (+)	1235288	1235740	183	132.63	304	214.27	0.62
	ompA (-)	1235816	1236856	15864	5003.04	5269	1616.06	3.10
	pop (+)	1236020	1236622	61	33.21	61	32.30	1.03
	d. ORF 1 (+)	1236662	1236892	0	0	1	1.38	NA
	d. ORF 2 (+)	1236714	1236950	0	0	1	1.34	NA

6.3 Supplementary figures

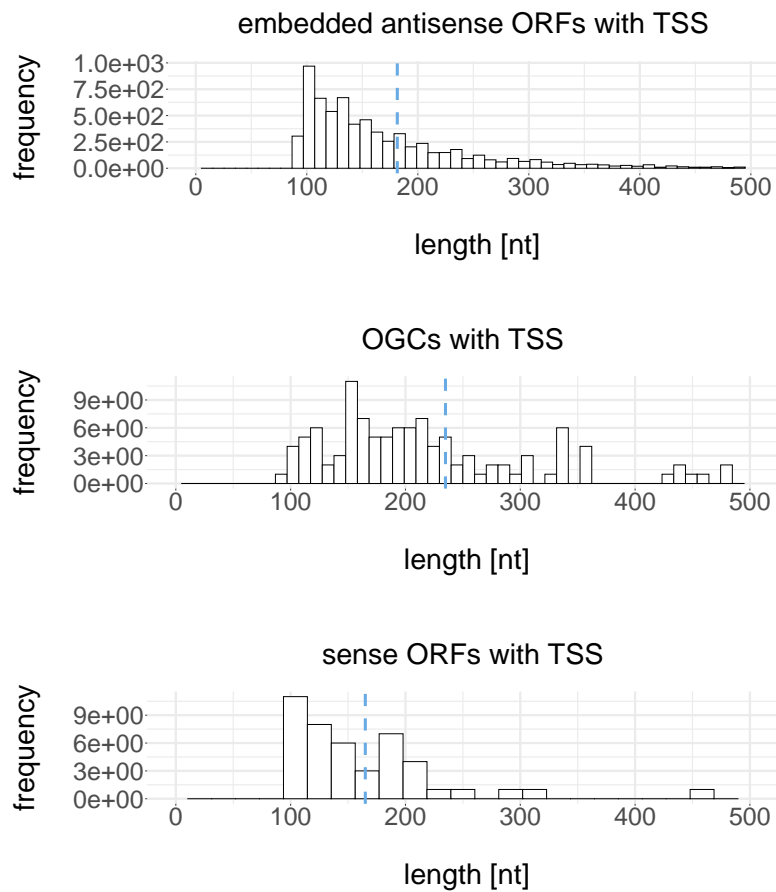
Supplementary figure S1: Relative change of TSS frequencies. Relative change for two gene set described in Figures 3.10a and 3.10b at the indicated RRS ranging from 10 to 1000. Cubic square smooth function is placed on the data (blue line).



Supplementary figure S2: Length distribution for genes and ORFs of gene sets analyzed. Lengths of functional and hypothetical annotated genes, all embedded antisense ORFs and those with blastp hit as well as overlapping gene candidates are shown. The mean length is indicated with the blue dashed line.



Supplementary figure S3: Length distribution for overlapping ORFs with TSS. Lengths of embedded antisense ORFs, OGCs, and sense ORFs with TSS are shown. The mean length is indicated with a blue dashed line.



Supplementary figure S4: Genomic sequence of the overlapping gene *pop*. Nucleic acid and proposed amino acid sequence are shown. black, TSS; blue, possible NTG start codons (1, 2, 3); red, stop codon (*); green, mutated position (C → T) resulting in a premature stop codon in *pop*. Black boxes and dashed line indicate the predicted promoter with -35 and -10 box and spacer region, respectively. The Shine-Dalgarno sequence upstream of start codon 2 is underlined.

CTGACGAAAGTCAGTTCAATTTACTAAAGGCCAAAAAAACCCCGCAGCAGCGGGGTTTTTCTAC
 CAGACGAGAACTTAAGC^①TTGCGGCTGAGTTACAA^②CGTCTTTGATACCTTTAACTT^③CGATCTCTA
 CGCGACGATCCGGAGCCAGGCAGTCGATCAGTGCAGCACGCTGTTTCACGTTGTCACAGGTGTT
 GCCAGTAACCGGGTTGGATTCGCCCATACCACGTGCGGAGATCTTGTCTGCCGGGATACCTTTG
 GAGATCAGGTAATCAACAACAGAC^②CTGAGCACGGCGCTCGGACAGACCCTGGTTGTAAGCGTCAG
 M S T A L G Q T L V V S V R
 AACCGATGCGGTTCGGTGTAAACC^③CAGAACAACACTACGGAACCGTCTTTCGGATCCAGGTTGCTCAG
 T D A V G V T Q N N Y G T V F R I Q V A Q
 CTGGCTGTACAGCTGATCCAGAGCAGCCTGACCTTCCGGTTTCAGGGTTGCTTTGTTGAAGTTG
 L A V Q L I Q S S^③L T F R F Q G C F V E V
 AACAGAACGTCAGACTTCAGAGTGAA^③GTGCTTGGTCTGTACTTCCGGTGCCGGAGCTGGAGCCG
 E Q N V R L Q S E V L G L Y F R C R S W S R
 GAGCAACTACTGGAGCTGCTTCGCCCTGACCGAAACGGTAGGAAACACCCAGGCTCAGCATGCC
 S N Y W S C F A L T E T V G N T Q A Q H A
 GTTGTCCGGACGAGTGCCGATGGTGTGTGCGTCACCGATGTTGTTGGTCCACTGGTATTCCAGA
 V V R T S A D G V C V T D V V G P L V F Q
 CGGGTAGCGATTTTCAGGAGTGATCGCGTACTCAACACCGCCAGCGAAGACCGGAGAAACGCCCG
 T G S D F R S D R V L N T A S E D R R N A G
 TGTCGTGGTTTTTACCATAAACGTTGGATTTAGTGTCTGCACGCCATAACCATAACCACCCAGACG
 V V V F T I N V G F S V C T P Y H T T Q T
 AGTGTAGATGTCCAGGTCGTCAGTGATTGGGTAACCCAGTTTAGCGGTCAGTTGAACGCCCTGA
 S V D V Q V V S D W V T Q F S G Q L N A L
 GCTTTGTATGCACCGTTTTCAACGCTGCCTTTGTACGGCATAACGACC^{TAA}
 S F V C T V F N A A F V R H T T *

Acknowledgment

Mein besonderer Dank gilt meinem Doktorvater Herrn Prof. Dr. Siegfried Scherer. Vielen Dank für die Möglichkeit zur Promotion an Ihrem Lehrstuhl sowie für Ihre Unterstützung, Ihr entgegengebrachtes Vertrauen bei der Umsetzung meines Projekts und Ihr allzeit offenes Ohr bei kleineren und größeren Problemen.

Für die schnelle und unkomplizierte Erstellung von Zweit- und Drittgutachten möchte ich mich bei Herrn Prof. Dr. Wolfgang Liebl und Frau Dr. Lindsay Hall bedanken, sowie bei Herrn Prof. Dr. Hanno Schäfer für die Übernahme des Prüfungsvorsitzes.

Meinem Mentor PD Dr. Klaus Neuhaus möchte ich herzlich danken für die experimentelle Betreuung. Danke auch für deine Geduld bei all meinen vielen Anliegen und Fragen.

Besonders möchte ich mich bei Romy Wecko bedanken, die mir zuverlässig und kompetent im Labor geholfen hat. Ohne dich, Romy, wäre vor allem das Western Blot Projekt nicht schaffbar gewesen! Auch meinen Studenten Jacqueline, Letyfee, Tobias, Caroline, Katharina, Katja, Veronika und Natascha dank ich für die Hilfe bei den Experimenten.

Außerdem möchte ich mich bei meinen tollen Bürokollegen Michaela Kreitmeier, Alina Glaub, Anika Wahl und zuvor Stefan Wichmann, sowie bei allen weiteren und ehemaligen Kollegen der AG OLG, Dr. Christopher Huptas, Dr. Sarah Hücker und Dr. Sonja Vanderhaeghen, für die harmonische und kollegiale Arbeitsatmosphäre bedanken. Desweiteren danke ich Dr. Zachary Ardern für die bereitwillige Unterstützung bei diversen bioinformatischen Analysen und Fragestellungen.

Vielen Dank auch an alle Mitarbeiter des Lehrstuhls Mikrobielle Ökologie für die tolle Zeit.

Der letzte Dank gilt meiner Familie. Ein großes Dankeschön geht an meinen Bruder Wolfgang und meine Eltern Marianne und Georg dafür, dass sie mich auf diesem Weg unterstützt und den nötigen Rückhalt gegeben haben. Ihr wart immer da, wenn ich euch brauchte. Ebenso danke ich dir, Christopher, dass du mich während der letzten Jahre unterstützt und meine Arbeitsweise geduldet hast, aber auch stets die richtigen Worte gefunden hast, um mich aufzubauen und zu motivieren.

Curriculum Vitae**Barbara Katrin Zehentner**

Date of birth December 19, 1990

Place of birth Landshut

Nationality German

Education

- 01.2016 - 11.2019 **PhD candidate**
Technical University of Munich, Chair of Microbial Ecology
(Prof. Dr. Scherer):
Experimental characterization of overlapping genes in entero-
hemorrhagic *E. coli*: Overexpression phenotypes and high-
throughput NGS analysis of transcription start sites
- 10.2013 - 11.2015 **Master of Science**
Molecular Biotechnology, Technical University of Munich
Master thesis, Chair of Microbial Ecology (Prof. Dr. Scherer):
Expression und Funktion von überlappenden ORFs in EHEC
- 10.2010 - 09.2013 **Bachelor of Science**
Molecular Biotechnology, Technical University of Munich
Bachelor thesis, Chair of Biotechnology of Natural Products
(Prof. Dr. Schwab): Klonierung und Charakterisierung der
mutmaßlichen Erdbeerallergene aus *Fragaria ananassa*
- 09.2001 - 07.2010 **Abitur**
Ruperti-Gymnasium Mühldorf am Inn

Eidesstattliche Erklärung

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

**Experimental characterization of overlapping genes in enterohemorrhagic
E. coli: Overexpression phenotypes and high-throughput NGS analysis of
transcription start sites**

am Lehrstuhl für Mikrobielle Ökologie, ZIEL - Institute for Food & Health, unter der Anleitung und Betreuung durch Herrn Prof. Dr. Siegfried Scherer ohne sonstige Hilfe erstellt und bei der Abfassung nur die gemäß § 6 Ab. 6 und 7 Satz 2 angebotenen Hilfsmittel benutzt habe.

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

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

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

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

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

Ort, Datum, Unterschrift