TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Mikrobielle Ökologie

Identification of overlapping genes in the human pathogens *Escherichia coli* LF82 and *Pseudomonas aeruginosa* PAO1 using transcriptomics, translatomics and proteomics

MICHAELA VERONIKA KREITMEIER

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:		Prof. Dr. Hanno Schäfer
Prüfer der Dissertation:	1.	Prof. Dr. Siegfried Scherer
	2.	Prof. Dr. Wolfgang Liebl

Die Dissertation wurde am 09.06.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 08.09.2021 angenommen.

Für meine Eltern

Table of Contents

Abstract ... 1

Zusammenfassung ... 2

List of Publications ... 4

Abbreviations ... 5

List of Figures ... 7

List of Equations ... 9

List of Tables ... 10

1. Introduction ... 11

  1.1 Overlapping genes in bacterial genomes ... 11

    1.1.1 The genetic code and gene expression ... 11

    1.1.2 Definition and occurrence of overlapping genes ... 13

    1.1.3 Origin and evolution of overlapping genes ... 15

  1.2 The model organisms ... 17

    1.2.1 *Escherichia coli* ... 17

      1.2.1.1 Discovery and properties ... 17

      1.2.1.2 Genome and classification ... 18

      1.2.1.3 AIEC pathotype and strain LF82 ... 18

    1.2.2 *Pseudomonas aeruginosa* ... 19

      1.2.2.1 Discovery and properties ... 19

      1.2.2.2 Genome ... 20

      1.2.2.3 Antibiotic resistance mechanisms ... 20

      1.2.2.4 Pathogenicity and clinical relevance ... 21

  1.3 Identification of protein-coding genes ... 22

    1.3.1 *In silico* gene prediction ... 22

    1.3.2 Experimental gene identification by high-throughput methods ... 23

      1.3.2.1 RNA-seq & Cappable-seq ... 23

      1.3.2.2 Conventional Ribo-seq ... 24

      1.3.2.3 Modified Ribo-seq variants ... 25

      1.3.2.4 Mass spectrometry ... 27

  1.4 Perspectives of the study ... 28

2. Materials & Methods ... 29

  2.1 Materials ... 29

    2.1.1 Antibiotics ... 29

    2.1.2 Bacterial strains and plasmids ... 29

    2.1.3 Commercial enzymes, ladders and kits ... 30

2.1.4 Growth media and buffers 31
2.1.5 Oligonucleotides 31
2.1.6 Peptides 33
2.1.7 Reagents and chemicals 33
2.2 Methods 34
2.2.1 Microbiological methods 34
2.2.1.1 Cultivation and storage of bacteria 34
2.2.1.2 Bacterial counting 34
2.2.1.3 Determination of minimum inhibitory concentration (MIC) 34
2.2.2 Molecular biological methods 35
2.2.2.1 Isolation and purification of nucleic acids 35
2.2.2.1.1 Genomic DNA extraction 35
2.2.2.1.2 RNase digestion 35
2.2.2.1.3 RNA extraction 35
2.2.2.1.4 DNase digestion 36
2.2.2.1.5 Plasmid isolation 36
2.2.2.1.6 Concentration of nucleic acids 36
2.2.2.2 Nucleic acid quantitation and quality control 36
2.2.2.2.1 Agarose gel electrophoresis 36
2.2.2.2.2 Capillary electrophoresis 37
2.2.2.2.3 Measurement of nucleic acid concentration 37
2.2.2.3 Modification and detection of nucleic acids 37
2.2.2.3.1 Polymerase chain reaction (PCR) 37
2.2.2.3.2 Reverse transcription 39
2.2.2.3.3 Quantitative PCR 39
2.2.2.3.4 Restriction digest 39
2.2.2.3.5 Ligation 40
2.2.2.3.6 Sanger sequencing 40
2.2.2.4 Construction of genetically modified *E. coli* 40
2.2.2.4.1 Preparation of electrocompetent cells 40
2.2.2.4.2 Transformation 40
2.2.2.4.3 Conjugation 40
2.2.2.4.4 Gene inactivation according to Datsenko & Wanner (2000) 41
2.2.2.4.5 Gene inactivation via conjugation of a suicide plasmid 41
2.2.3 Next generation sequencing-based methods 42
2.2.3.1 Transcriptome sequencing (RNA-seq) 42
2.2.3.1.1 Inhibition of translation and cell lysis 42

2.2.3.1.2 Ribosomal RNA depletion 44
2.2.3.1.3 Ultrasonic fragmentation 44
2.2.3.1.4 RNA (de-)phosphorylation 44
2.2.3.1.5 Library preparation and normalization 44
2.2.3.1.6 Sequencing 45
2.2.3.2 Ribosome profiling (Ribo-seq) 45
2.2.3.2.1 Nuclease digestion 45
2.2.3.2.2 Sucrose gradient and monosome fractionation 45
2.2.3.2.3 Size selection and recovery of RNA 45
2.2.3.3 Cappable-seq for transcription start site (TSS) identification 46
2.2.4 Protein chemical methods 46
2.2.4.1 Overexpression of *relE* 46
2.2.4.1.1 Cultivation and induction of RelE synthesis 47
2.2.4.1.2 Cell lysis 47
2.2.4.2 Purification of RelE 47
2.2.4.2.1 Kit-based affinity chromatography 47
2.2.4.2.2 FPLC-based affinity chromatography 47
2.2.4.2.3 Protein refolding by diffusion dialysis 48
2.2.4.2.4 Protein concentration via centrifugation 48
2.2.4.2.5 Protein storage 48
2.2.4.3 Protein detection and quantitation 49
2.2.4.3.1 Electrophoretic separation via SDS-PAGE 49
2.2.4.3.2 Western blot and immunodetection 49
2.2.4.3.3 Determination of protein concentration 50
2.2.4.4 Analysis of proteins by mass spectrometry 50
2.2.4.4.1 Detection of RelE 50
2.2.4.4.2 Data dependent acquisition (DDA) for novel gene identification 50
2.2.4.4.2.1 Sample collection and preparation 50
2.2.4.4.2.2 Sample purification and offline HPLC-fractionation 51
2.2.4.4.2.3 LC-MS/MS measurement 51
2.2.4.4.3 Parallel reaction monitoring (PRM) for novel gene verification and quantitation 51
2.2.4.4.3.1 Sample collection and preparation 51
2.2.4.4.3.2 Sample purification and fractionation using StageTips 51
2.2.4.4.3.3 PRM assay development 52
2.2.5 Data processing and bioinformatic analyses 52
2.2.5.1 Evaluation of next generation sequencing (NGS) data 52

2.2.5.1.1 Transcriptome sequencing 52

2.2.5.1.2 Conventional Ribo-seq 53

2.2.5.1.3 RelE-supported Ribo-seq 54

2.2.5.1.4 Retapamulin-assisted Ribo-seq 54

2.2.5.1.5 Cappable-seq 55

2.2.5.2 Evaluation of mass spectroscopical data 55

2.2.5.2.1 DDA 55

2.2.5.2.2 PRM 56

2.2.5.3 *In silico* analyses of identified overlapping gene candidates 56

2.2.5.3.1 Genomic distribution 56

2.2.5.3.2 Promoter determination 56

2.2.5.3.3 Terminator identification 56

2.2.5.3.4 Ribosome binding site identification 56

2.2.5.3.5 Detection of homologues 56

2.2.5.3.6 Gene prediction 57

2.2.5.3.7 Evolutionary analyses 57

3. Results 58

3.1 Identification of protein-coding genes in *E. coli* LF82 58

3.1.1 Conventional RNA-seq & Ribo-seq 59

3.1.1.1 Sequencing output and data reproducibility 59

3.1.1.2 Analysis of anORF expression 61

3.1.2 Ribo-RET for translation initiation site (TIS) detection 62

3.1.2.1 Construction of a Δ*tolC* deletion mutant 62

3.1.2.2 Determination of the minimum inhibitory concentration of retapamulin 64

3.1.2.3 Cut-off specification for genome wide TISs analysis 65

3.1.3 RelE-enhanced Ribo-seq to visualize triplet periodicity 67

3.1.3.1 Expression and downstream processing of RelE 67

3.1.3.2 Evaluation of reading frame information 70

3.1.4 Detection and analyses of novel intergenic and overlapping genes 72

3.1.4.1 Application of prediction algorithms for individual and collective datasets 72

3.1.4.2 Bioinformatic characterization of the selected ORFs 75

3.2 Identification of protein-coding genes in *P. aeruginosa* PAO1 78

3.2.1 Conventional RNA-seq & Ribo-seq 79

3.2.1.1 Adaptation and optimization of the Ribo-seq protocol 79

3.2.1.2 Sequencing output and data reproducibility 80

3.2.2 Proteomic verification of anORF translation by mass spectrometry 81

3.2.3 Ribo-RET for translation initiation site detection 81

3.2.3.1 Determination of the minimum inhibitory concentration of retapamulin 82

3.2.3.2 Cut-off specification for genome wide TISs analysis 82

3.2.4 RelE-enhanced Ribo-seq to visualize triplet periodicity 83

3.2.4.1 Evaluation of RelE functionality in PAO1 84

3.2.5 Cappable-seq for transcription start site (TSS) identification 85

3.2.5.1 Sequencing output and reproducibility of biological replicates 85

3.2.5.2 Cut-off analysis for global detection of TSS 86

3.2.6 Detection and analyses of novel intergenic and overlapping genes 87

3.2.6.1 Application of prediction algorithms for individual and collective datasets 87

3.2.6.2 Bioinformatic characterization of the selected ORFs 90

3.2.6.3 Characteristics of two exceptionally long OLGs 94

3.2.6.3.1 Genomic localization 94

3.2.6.3.2 Gene-like structural features of *olg1* and *olg2* 95

3.2.6.3.3 Transcriptome and translatome analysis 96

3.2.6.3.4 Detection of translated peptides by mass spectrometry 99

3.2.6.3.5 Evolutionary analyses of *olg1* and *olg2* 100

4. Discussion 103

4.1 High-throughput discovery of novel genes – strengths and limitations 103

4.1.1 Transcriptome sequencing & Cappable-seq 103

4.1.2 Classical ribosome profiling 106

4.1.3 Modified variants of ribosome profiling 108

4.1.3.1 RelE-assisted Ribo-seq for reading frame analysis 108

4.1.3.2 Retapamulin-supported Ribo-seq for TIS detection 111

4.1.4 Mass spectrometry 112

4.2 Cumulative evidence for protein-coding capacity of identified gene candidates 113

4.3 *olg1* and *olg2* as prime examples for overlooked OLGs in *Pseudomonas* 117

4.4 Concluding remarks and outlook 120

5. References 122

6. Supplement 146

6.1 Supplementary Files 146

6.2 Supplementary Figures 147

6.3 Supplementary Tables 162

Danksagung 168

Curriculum Vitae 169

Eidesstaatliche Erklärung 170

Abstract

A double stranded DNA locus allows six possible reading frames due to the triplet character of the genetic code. All reading frames have the potential to encode proteins, which, theoretically, allows for an overlapping encoding of two or more protein-coding genes at the same locus. In prokaryotes, sense overlaps of a few base pairs are common and are often associated with joint translational regulation. In contrast, the existence of large gene overlaps (≥30 nt) has mainly been accepted for viruses. However, studies addressing these nested gene constructs in prokaryotes are rare. In this study, two human pathogenic γ-Proteobacteria, *Escherichia coli* LF82 and *Pseudomonas aeruginosa* PAO1, were analysed for the presence and characteristics of overlapping genes. For both strains, multiple high-throughput experiments including transcriptome sequencing (RNA-seq) and ribosome profiling (Ribo-seq) were conducted under different conditions. In addition, application of the toxin RelE and the antibiotic retapamulin in modified Ribo-seq experiments was implemented to improve resolution and power of conventional Ribo-seq. Concerning RelE-supported Ribo-seq, the *E. coli* RelE toxin was overexpressed, purified and refolded prior to application in *E. coli* LF82 and *P. aeruginosa* PAO1 experiments. Retapamulin-assisted Ribo-seq required the generation of efflux deletion mutants and determination of the minimum inhibitory concentration for efficient stalling of ribosomes at start codons. Both methods were successfully adapted for the target organisms as indicated by a RelE induced periodicity signal as well as a retapamulin induced redistribution of ribosomes, both observed for annotated, protein-coding genes of *E. coli* LF82 (n = 4,586) and *P. aeruginosa* (n = 5,572), respectively. Different Ribo-seq prediction tools were applied to each of the Ribo-seq datasets generated, and the results were combined for reliable overlapping gene identification. In total, predicted hits resulting from 19 and 16 different prediction combinations based on 7 and 5 different Ribo-seq datasets were evaluated for *E. coli* LF82 and *P. aeruginosa* PAO1, respectively. Candidates being detected in more than half of all prediction combinations were visually inspected and curated, yielding 104 (*E. coli* LF82) and 63 (*P. aeruginosa* PAO1) promising overlapping gene candidates. Furthermore, 12 and 61 novel open reading frames in intergenic regions were identified. All candidates were bioinformatically characterized with respect to their expressability. The results obtained were comparable to those of annotated genes implying the genuine protein-coding capability of the novel overlapping gene candidates. For *P. aeruginosa* PAO1, data-dependent acquisition-based mass spectrometry confirmed the protein-coding potential of 47 gene candidates, and Cappable-seq aided in the determination of their transcription start sites. Two of the overlapping gene candidates showing large antisense overlaps of 957 and 1,536 nucleotides with annotated genes in *P. aeruginosa* PAO1 were characterized in more detail. Bioinformatic analyses confirmed the gene-like structure, transcription and translation of both overlapping gene candidates. Multiple, high-confident peptides covering a wide range of both open reading frames were detected via mass spectrometry, and a subset of the peptides discovered were unequivocally validated by parallel reaction monitoring. Quantification of these peptides revealed a growth-phase dependent expression of both overlapping gene candidates, emphasising the potential functionality of the encoded protein products. Further support for functionality was obtained by evolutionary analyses indicating purifying selection. The novel gene candidates identified in this study expand the known coding capacity of *E. coli* LF82 and *P. aeruginosa* PAO1, thereby facilitating a deeper understanding of their genome complexity.

Zusammenfassung

Durch den Triplett-Charakter des genetischen Codes ermöglicht ein doppelsträngiger DNA-Lokus die Existenz von insgesamt sechs verschiedenen Leserastern. Alle sechs Leseraster können potenziell für Proteine kodieren, was zu einer Überlappung der kodierenden Sequenz zweier oder mehrerer Gene am selben Lokus führen kann. Solche Gene werden als überlappende Gene bezeichnet. In Prokaryoten sind *sense* Überlappungen von einigen wenigen Basenpaaren nicht unüblich, da diese häufig für die gekoppelte Regulation mehrerer Gene genutzt werden. Längere Überlappungen von mehr als 30 Nucleotiden hingegen sind vorwiegend in viralen Genomen akzeptiert; Studien zu entsprechenden Genen in Prokaryoten liegen jedoch kaum vor. In dieser Arbeit wurden die beiden humanpathogenen γ-Proteobakterien *Escherichia coli* LF82 und *Pseudomonas aeruginosa* PAO1 hinsichtlich der Existenz und der Eigenschaften von überlappenden Genen untersucht. Für beide Stämme wurden verschiedene Hochdurchsatz-Experimente wie beispielsweise Transkriptom-Sequenzierung (RNA-seq) und *Ribosome profiling* (Ribo-seq) unter mehreren Bedingungen durchgeführt. Zusätzlich sollten modifizierte Ribo-seq Protokolle unter Anwendung des Toxins RelE und des Antibiotikums Retapamulin durchgeführt werden, um die Aussagekraft von konventionellen Ribo-seq Experimenten zu erhöhen. Für die Durchführung von Ribo-seq Experimenten mit RelE musste das endogen in *E. coli* vorkommende Toxin zuerst überexprimiert, aufgereinigt und dann rückgefaltet werden. Retapamulin-gestütztes Ribo-seq erforderte die Erstellung von Efflux-Deletionsmutanten sowie die Bestimmung der minimalen Hemmkonzentration, um eine effiziente Inhibierung von initiierenden Ribosomen und somit eine optimale Funktionalität von Retapamulin zu gewährleisten. Wie anhand des RelE induzierten Periodizitätssignals sowie der durch Retapamulin bedingten Umverteilung der Ribosomen an annotierten, protein-kodierenden Genen (n = 4586 und n = 5572) ersichtlich, wurden beide Methoden erfolgreich in den Ziel-Stämmen implementiert. Die generierten Ribo-seq Datensätze wurden unter Anwendung verschiedenster Vorhersagealgorithmen ausgewertet und die resultierenden Ergebnisse wurden vereinigt, um eine verlässliche Identifizierung überlappender Gene zu gewährleisten. Insgesamt wurden die Ergebnisse von 19 beziehungsweise 16 verschiedenen Vorhersagekombinationen, die auf 7 und 5 Ribo-seq Datensätzen basierten, für *E. coli* LF82 und *P. aeruginosa* PAO1 ausgewertet. Kandidaten, die in mehr als der Hälfte aller Vorhersage-Kombinationen detektiert wurden, wurden visuell inspiziert und gegebenenfalls manuell kuriert. Mit dieser Methode wurden 104 (*E. coli* LF82) und 63 (*P. aeruginosa* PAO1) vielversprechende überlappende Gen-Kandidaten sowie 12 beziehungsweise 61 neue, translatierte Leserahmen in intergenischen Regionen identifiziert. Alle ermittelten Kandidaten wurden bioinformatisch hinsichtlich des Expressionspotentials charakterisiert. Die Eigenschaften der neuen Gen-Kandidaten waren mit denen aller annotierten Gene vergleichbar, was die Protein-kodierenden Fähigkeiten dieser neuen Gen-Kandidaten bestätigt. Das Protein-kodierende Potential von 47 in *P. aeruginosa* PAO1 identifizierten Gen-Kandidaten wurde mittels datenabhängiger Massenspektrometrie verifiziert und die Durchführung von Cappable-seq ermöglichte die reproduzierbare Ermittlung der Transkriptionsstarts. Zwei Gen-Kandidaten, die außergewöhnlich lange *antisense* Überlappungen von 957 und 1536 Nucleotiden mit annotierten Genen in *P. aeruginosa* PAO1 zeigten, wurden näher untersucht. Bioinformatische Analysen bestätigten das Vorhandensein Gen-spezifischer Elemente, und die Transkription sowie Translation beider Gen-Kandidaten wurde nachgewiesen. Mehrere Peptide, die einen großen Bereich der kodierten Proteine abdeckten, wurden mittels Massenspektrometrie detektiert und eine Auswahl hiervon im Anschluss durch parallele Reaktionsüberwachung validiert. Die Quantifizierung der validierten Peptide ergab Hinweise auf eine Wachstumsphasen-abhängige Expression beider Gen-Kandidaten, welche auf eine potenzielle Funktionalität der kodierten Proteinprodukte hinwies. Weitere Hinweise auf Funktionalität wurden durch evolutionäre Analysen erhalten, deren Ergebnisse auf negative Selektion hindeuteten.

Die in dieser Arbeit identifizierten, neuen Gen-Kandidaten erweitern das bislang bekannte Kodierungspotential von *E. coli* LF82 und *P. aeruginosa* PAO1 und ermöglichen dadurch ein vertieftes Verständnis für die Genomkomplexität dieser Organismen.

List of Publications

Publications

Kreitmeier, M., Ardern, Z., Abele, M., Ludwig, C., Scherer, S., and Neuhaus, K. (2021). Shadow ORFs illuminated: long overlapping genes in *Pseudomonas aeruginosa* are translated and under purifying selection. *Submitted, in revision.*

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.

Conference contributions (posters)

Kreitmeier, M., Ardern, Z., Neuhaus, K., and Scherer, S. (2019). "Ribosome profiling reveals new weakly translated ORFs overlapping annotated genes in *Escherichia coli* LF82". *24th Annual Meeting, RNA Society (USA), Krakow, Poland.*

Huptas, C., **Kreitmeier, M.**, Wenning, M., and Scherer, S. (2017). "Delineation of *Pseudomonas* species based on genomic sequences". *16th International Conference on Pseudomonas, Microbiology Society (UK), Liverpool, United Kingdom.*

Abbreviations

%C	crosslinker percentage
%T	acrylamide monomer concentration
A	adenine
AA	amino acid
ACN	acetonitrile
AGC	automatic gain control
AIEC	adherent-invasive *E. coli*
anORF	annotated ORF
aSD	anti-Shine-Dalgarno
blast	Basic Local Alignment Search Tool
bp	base pair
BSA	bovine serum albumin
C	cytosine
CD	Crohn´s disease
cDNA	complementary DNA
CFU	colony-forming units
CV	column volume
DAEC	diffusely adherent *E. coli*
DDA	data-dependent acquisition
DE	differential gene expression
DEPC	diethyl pyrocarbonate
DIA	data-independent acquisition
DMSO	dimethyl sulfoxide
dN/dS	ratio of non-synonymous and synonymous variants
DNA	deoxyribonucleic acid
DTB-Gppp	3'-desthiobiotin-TEG-guanosine 5' triphosphate
EAEC	enteroaggregative *E. coli*
EHEC	enterohemorrhagic *E. coli*
EIEC	enteroinvasive *E. coli*
EPEC	enteropathogenic *E. coli*
ETEC	enterotoxigenic *E. coli*
ExpX	experiment no. X
FA	formic acid
FDR	false discovery rate
FPLC	fast protein liquid chromatography
G	guanine
HCD	higher energy collision induced dissociation
His	histidine
iBAQ	Intensity Based Absolute Quantification
IMAC	immobilized metal ion affinity chromatography
iORF	ORF in intergenic region
kDa	kilodalton
LB	lysogeny broth
LC	liquid chromatography
LDF	linear discriminant function
m/z	mass-to-charge ratio
maxIT	maximum injection time
MFP	membrane fusion protein
MIC	minimal inhibitory concentration
MNase	micrococcal nuclease
mRNA	messenger RNA
MS	mass spectrometry
MS/MS	tandem mass spectrometry
N	any nucleotide
NC	negative control
NCE	normalized collision energy
ND	no drug
NGS	next generation sequencing
nt	nucleotide
NTA	nitrilotriacetic acid
OD_{600nm}	optical density at 600 nm
OLG	overlapping gene
OLG_EA	antisense embedded OLG
OLG_ES	sense embedded OLG
OLG_PA3	partial antisense OLG with overlap at the 3´ end
OLG_PA5	partial antisense OLG with overlap at the 5´ end
OLG_PS3	partial sense OLG with overlap at the 3´ end
OLG_PS5	partial sense OLG with overlap at the 5´ end
OLG_TL	trivial OLG
OMF	outer membrane factor
ON	overnight
ORF	open reading frame
ori	origin of replication

PAA	polyacrylamide
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PRM	parallel reaction monitoring
PSM	peptide spectrum match
qPCR	quantitative PCR
RCV	ribosome coverage value
RepX	replicate no. X
RET	retapamulin
RF	reading frame
RFP	ribosome footprint
Ribo-RET	retapamulin-assisted Ribo-seq
Ribo-seq	ribosome profiling
RIN	RNA integrity number
RNA	ribonucleic acid
RNAP	RNA polymerase
RNase I	ribonuclease I
RNase T	Exonuclease T
RNA-seq	total RNA sequencing
RND	resistance-nodulation-cell division
RPKM	reads per kilobase per million mapped reads
RPM	reads per million mapped reads
rRNA	ribosomal RNA
RRS	relative read score
RT	reverse transcription
SD	Shine-Dalgarno
SDS-PAGE	sodium dodecyl sulphate–polyacrylamide gel electrophoresis
SOC	super optimal broth with catabolite repression
T	thymine
T1SS	type I secretion system
T2SS	type II secretion system
T3SS	type III secretion system
T6SS	type VI secretion system
TDA	targeted data acquisition
TEX	terminator 5′-phosphate-dependent exonuclease
TIS	translation initiation site
tRNA	transfer RNA
TSS	transcription start site
UTR	untranslated region
UV	ultraviolet
v	version
ΔG_{SD}	minimum SD free energy

List of Figures

Figure 1. Schematic illustration of a double-stranded DNA locus with all possible reading frames 11

Figure 2. Schematic representation of a bacterial gene including structural elements.. 12

Figure 3. Classification of overlapping genes... 13

Figure 4. *De novo* evolution of new genes via proto-genes 16

Figure 5. Schematic illustration of an RND pump 21

Figure 6. Scheme of the Cappable-seq pipeline for transcription start site identification..................................... 24

Figure 7. Overview of the experimental procedure of Ribo-seq and RNA-seq.. 25

Figure 8. Regulation and organisation of the RelE-encoding operon.. 27

Figure 9. Growth curve of *E. coli* LF82 for RNA-seq and Ribo-seq 59

Figure 10. Reproducibility of biological Ribo-seq and RNA-seq experiments 60

Figure 11. Scheme of the genomic integration and subsequent removal of the *tolC* deletion fragment in *E. coli* LF82 63

Figure 12. Agarose gel analysis of putative *E. coli* LF82Δ*tolC* deletion mutants 64

Figure 13. Results of minimum inhibitory concentration testing of retapamulin for *E. coli* LF82 and *E. coli* LF82Δ*tolC*.. 64

Figure 14. Translation signals of highly expressed proteins of the S10 ribosomal protein operon 66

Figure 15. Metagene analysis of highly expressed genes in *E. coli* LF82. 67

Figure 16. Results of SDS-PAGE and Western blot analysis of *relE* and *relB* overexpression samples 68

Figure 17. Results of SDS-PAGE and mass spectrometry analysis of RelE and RellB purification samples........... 69

Figure 18. Chromatogram and SDS-PAGE analysis of the final ÄKTA purification process................................. 69

Figure 19. Read length distribution of RelE-supported Ribo-seq in *E. coli* LF82 70

Figure 20. Reading frame analysis of gene *ompA* in *E. coli* LF82.. 71

Figure 21. Reading frame sum signal of all annotated genes for Ribo-seq and RNA-seq datasets in *E. coli* LF82.. 71

Figure 22. Reproducibility of Ribo-seq prediction results in *E. coli* LF82 73

Figure 23. Ribo-seq signals of three novel gene examples identified in *E. coli* LF82.. 74

Figure 24. Distribution and length of novel gene candidates in the *E. coli* LF82 genome 75

Figure 25. Expression metrics of the novel gene candidates in comparison to protein-coding, annotated genes in *E. coli* LF82 76

Figure 26. Translation initiation site analysis of novel gene candidates in *E. coli* LF82.. 77

Figure 27. Reading frame sum signal of all novel gene candidates in *E. coli* LF82 78

Figure 28. Read length distributions of all mRNA reads of conventional Ribo-seq in *P. aeruginosa* PAO1........... 80

Figure 29. Comparison of translatome and proteome data in *P. aeruginosa* PAO1... 81

Figure 30. Results of minimum inhibitory concentration testing of retapamulin for *P. aeruginosa* PAO1 and *P. aeruginosa* PAO397.. 82

Figure 31. Metagene and threshold analysis of all anORFs in *P. aeruginosa* PAO397 83

Figure 32. Results of RelE sequence bias and periodicity analysis in *P. aeruginosa* PAO1 84

Figure 33. Results of the cut-off and 5`UTR length analysis of transcription start sites in *P. aeruginosa* PAO1.. 86

Figure 34. Ribo-seq signals of three novel gene examples identified in *P. aeruginosa* PAO1 89

Figure 35. Distribution and length of novel gene candidates in the *P. aeruginosa* PAO1 genome 90
Figure 36. Sequence logos of promoter regions and putative Shine-Dalgarno sequences in *P. aeruginosa* PAO1 91
Figure 37. Expression metrics of the novel gene candidates in comparison to protein-coding, annotated genes in *P. aeruginosa* PAO1 92
Figure 38. Reading frame sum signal of all novel gene candidates in *P. aeruginosa* PAO1 93
Figure 39. Genomic location of *olg1* and *olg2* 95
Figure 40. Results of RNA-seq, Ribo-seq and proteome analysis at the *olg1/tle3* and *olg2*/PA1383 locus 97
Figure 41. Normalized protein intensities measured by DDA-MS 99
Figure 42. Intensities of all peptides measured for Tle3, Olg1, PA1383 and Olg2 using targeted proteomics 100
Figure 43. Stop codon depletion as indicator for purifying selection acting on *olg1* and *olg2* 101
Figure 44. Sequence constraint analysis of *olg1* and *olg2* using FRESCo and OLGenie 102
Figure 45. Overview of all novel gene candidates identified in *E. coli* LF82 and *P. aeruginosa* PAO1 117

List of Equations

Equation 1. Calculation of bacterial cell counts 34
Equation 2. Optimal annealing temperature for Q5 High-Fidelity DNA Polymerase PCR 37
Equation 3. Optimal annealing temperature for *Taq* DNA Polymerase PCR 38
Equation 4. Normalization equation for NGS-libraries 45
Equation 5. Normalization of sequencing reads to sequencing depth 52
Equation 6. Normalization of sequencing reads to sequencing depth and gene length 53
Equation 7. Equation for calculation of the ribosome coverage value 53
Equation 8. Calculation of the relative read score 55

List of Tables

Table 1. Antibiotics ... 29

Table 2. Bacterial strains ... 29

Table 3. Plasmids ... 30

Table 4. Enzymes, kits and ladders ... 30

Table 5. Growth media with their composition ... 31

Table 6. Primers used for PCR and cloning applications ... 31

Table 7. Isotopically labelled peptides used for targeted proteomics ... 33

Table 8. Pipetting scheme for PCR with Q5 High-Fidelity or *Taq* DNA Polymerase ... 38

Table 9. Thermal cycler settings for PCR with Q5 High-Fidelity or *Taq* DNA Polymerase ... 38

Table 10. Detailed set-up of all Ribo-seq and RNA-seq experiments ... 43

Table 11. Number and classification of non-trivial overlapping genes predicted for the *E. coli* LF82 genome ... 58

Table 12. Overview of Ribo-seq and RNA-seq reads obtained for Exp1 in *E. coli* LF82 after sequencing ... 60

Table 13. Overview of expression metrics obtained for all annotated genes of *E. coli* LF82 ... 61

Table 14. Translation metrics of highly expressed proteins of the S10 ribosomal protein operon in *E. coli* LF82 ... 65

Table 15. Number of novel gene candidates identified in *E. coli* LF82 ... 75

Table 16. Number and classification of non-trivial overlapping genes predicted for the *P. aeruginosa* PAO1 genome ... 78

Table 17. Overview of Ribo-seq and RNA-seq reads obtained for Exp1 in *P. aeruginosa* PAO1 after sequencing ... 80

Table 18. Overview of Cappable-seq reads after sequencing ... 85

Table 19. Calculation of Pearson correlation coefficients r for Cappable-seq experiments in *P. aeruginosa* PAO1 ... 85

Table 20. Reproducibility of Ribo-seq prediction results for *P. aeruginosa* PAO1 ... 87

1. Introduction

1.1 Overlapping genes in bacterial genomes

1.1.1 The genetic code and gene expression

The genetic information of all living organisms is stored in nucleic acids, which are molecule chains composed of phosphate, sugar and nucleobases. Depending on their components and their structure, two different types can be distinguished: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA is the most common carrier of genetic information and consists of the purin nucleobases adenine (A) and guanine (G) and the pyrimidine nucleobases cytosine (C) and thymine (T), which are linked via a phosphate-deoxyribose backbone. The sequence of the nucleobases determines the genetic code, whereby always three nucleobases encode for one of the 20 proteinogenic amino acids (AA). Due to hydrogenic interactions between A and T or C and G of two antiparallel DNA strands, DNA naturally forms a double helix structure (Watson & Crick, 1953). Consequently, one DNA locus enables six possible reading frames (RF); all of them might be protein encoding, in theory (**Figure 1**).

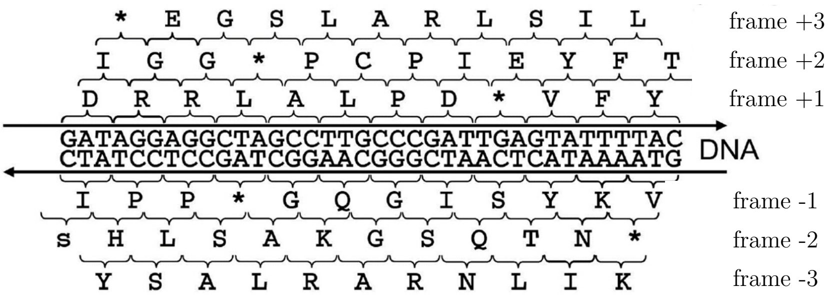


Figure 1. Schematic illustration of a double-stranded DNA locus with all possible reading frames (frame +1, +2 & +3 on the sense strand; frame -1, -2 & -3 on the antisense strand). Figure from Scherer *et al.* (2018).

A prerequisite for translation of nucleic acids into proteins are open reading frames (ORFs), which are defined as DNA segments between start and stop codons. Together with regulatory elements, the DNA sequence of an ORF forms the structural basis of a protein-coding gene. To exert their function, those genes are transcribed into messenger RNA (mRNA) and translated into the final polypeptide (**Figure 2**).

The first step in gene expression is the recruitment of the RNA polymerase (RNAP) guided by a promoter upstream of the ORF. This process is mediated by σ factors that recognize specific promoter elements and activate the RNAP core complex by forming the RNAP holoenzyme (Browning & Busby, 2004). The type of σ factor used for transcription initiation thereby depends on conserved sequence patterns and may vary in response to extracellular and intracellular signals (Paget & Helmann, 2003). The principal σ factor of *Escherichia coli*, σ^{70}, for instance, directs general transcription of housekeeping genes by recognizing conserved -35 (5'TTGACA'3) and -10 elements (5'TATAAT'3) upstream of the transcription start site (TSS). However, deviations from the consensus sequence as well as the complete absence of the -35 element are frequently reported (Shultzaberger *et al.*, 2007, Barne *et al.*, 1997, Keilty & Rosenberg, 1987). Other promoter-associated features influencing transcriptional activity and efficiency include an extended -10 sequence (Keilty & Rosenberg, 1987), presence of UP elements -60 to -40 nt upstream of the TSS (Estrem *et al.*, 1998), the spacer region between the -35 and -10 element (Aoyama *et al.*, 1983) as well as the distance between the -10 element and the TSS (Lewis & Adhya, 2004). Additional upstream sequences, e.g., *cis*-elements, can further modulate and regulate the transcriptional process by interacting with *trans*-elements like transcription factors.

In bacteria, ORFs are commonly organized in operons, which are clusters of co-regulated genes sharing the same promoter (Ermolaeva *et al.*, 2001). Once recruited, the RNAP starts to synthesize mRNA at the TSS (position +1) and elongates the nascent mRNA until it reaches a specific area at the end of a single gene or a whole operon, where the RNAP-DNA-mRNA complex dissociates, and transcription terminates. Termination can either be initiated by a GC-rich symmetric region followed by a stretch of Ts forming a RNA hairpin terminator (intrinsic termination; Gusarov & Nudler, 1999) or by adenosine triphosphate-dependent hydrolysis of the RNA induced by factor Rho (Rho-dependent termination; Peters *et al.*, 2009).

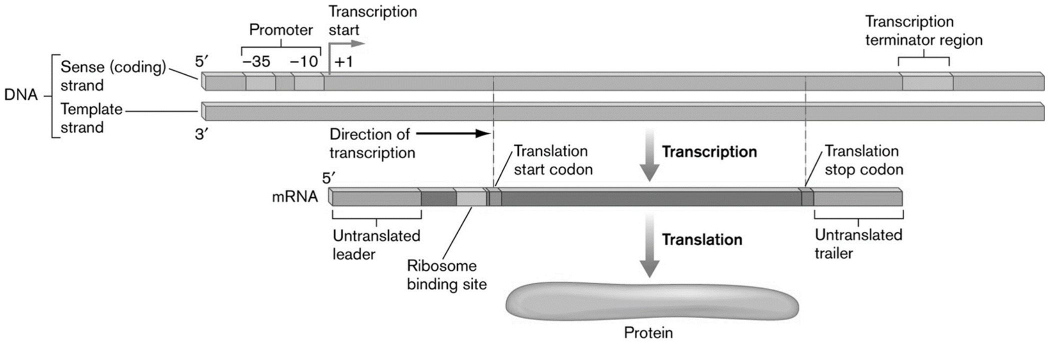


Figure 2. Schematic representation of a bacterial gene including structural elements necessary for its transcription and translation. Figure by Slonczewski & Foster (2009).

After transcription, the synthesized mRNA is translated into a polypeptide in a three-stage process. Firstly, translation initiation is mediated by start codon recognition and the subsequent binding of a N-formylmethionyl-transfer RNA (tRNA) within the P site of the ribosome. The canonical codon AUG represents the most frequent start codon in *E. coli* (~83% of all genes), followed by GUG (14%) and UUG (3%). Other non-canonical start codon variants including NUG and AUN codons (N represents all possible nucleobases) may also initiate translation in rare cases, but at lower efficiency (Hecht *et al.*, 2017, Chengguang *et al.*, 2017). In a second step, the succeeding codon is recognised by the complementary aminoacyl-tRNA in the A site of the ribosome and a peptide bond with the previous AA located in the P site is formed. Finally, the ribosome translocates along the mRNA in a three-nucleotide (nt) wise manner, allowing the next codon to be decoded by the respective aminoacyl-tRNA. Several rounds of decoding, peptide bond formation and translocation lead to elongation of the nascent polypeptide chain. Translation ceases when termination factors like RF1 or RF2 recognize one of the stop codons UAG, UAA or UGA. After hydrolysis of the peptidyl-tRNA ester bond, those factors dissociate and ribosome recycling takes place (for review see Rodnina, 2018).

The protein-coding sequence of an mRNA is flanked by untranslated regions (UTRs) at the 5´ and 3´ end. Those regions are involved in regulation of gene expression both at the transcriptional and at the translational level. The upstream 5´ UTR, for instance, harbours a Shine-Dalgarno (SD) sequence, which is complementary to a region at the 3´ end of the 16S ribosomal RNA (rRNA), known as the anti-Shine-Dalgarno (aSD). Interaction between both mediates the correct positioning of the ribosome for translational initiation (Shine & Dalgarno, 1974). In ɣ-Proteobacteria, the aSD core sequence 'CCUCC' hybridizes with the SD sequence in an optimal aligned spacing of 7-9 nt relative to the start codon (Ma *et al.*, 2002). However, deviations from the SD core sequence and the optimal spacing distance may also enable translation initiation, although with reduced efficiency (Komarova *et al.*, 2020, Evfratov *et al.*, 2017, Ma *et al.*, 2002, Ringquist *et al.*, 1992). Even in the complete absence of a SD sequence, translation can take place at high levels, indicating that SD presence is not mandatory for initiation (Skorski *et al.*, 2006). Despite the SD sequence, other 5´UTR structures like riboswitches (Breaker, 2018), RNA thermometers (Loh *et al.*, 2018), small upstream open reading frames (Orr *et al.*, 2019) or small RNAs (Storz *et al.*, 2011) exert a regulatory activity on gene expression.

Further regulation of gene expression is mediated by the 3´UTR. Despite being involved in transcriptional regulation by enabling intrinsic or Rho-dependent termination, the 3´UTR was also shown to be involved in post-transcriptional regulation, e.g., by interacting with small RNAs or by affecting mRNA stability (Ren *et al.*, 2017).

1.1.2 Definition and occurrence of overlapping genes

The presence of six possible reading frames facilitates the existence of overlapping genes (OLGs). Those are defined as genes sharing at least one nucleotide of their protein-coding region with the protein-coding region of another gene. The type of overlapping gene is determined by its location and orientation relative to the annotated ORF (anORF), also called mother gene. By definition, the mother gene is always encoded in frame +1 and the frame of the OLG is specified referring to the mother gene´s frame (**Figure 3A**). Both can be located either on the same strand (sense) or on different strands (antisense), whereby the OLG can either overlap completely (embedded) or partially with the mother gene (**Figure 3B**). Another criterion for OLG classification represents the length of the overlap. Small overlaps of a few base pairs (bp) are called non-trivial, whereas large overlaps of more than 30 bp are referred to as non-trivial overlaps.

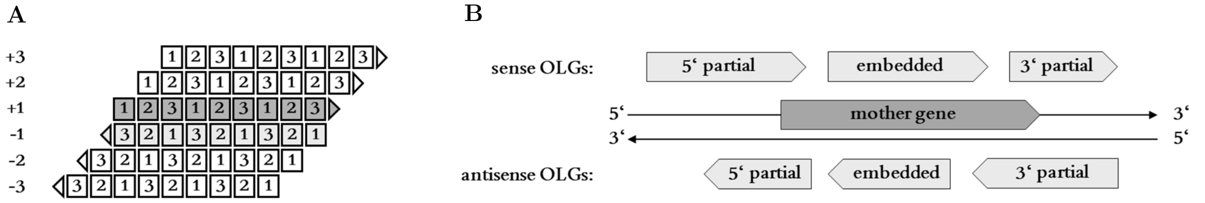


Figure 3. Classification of overlapping genes (OLGs). (**A**) The frame of an OLG is designated in relation to the position of its mother gene encoded in frame +1. (**B**) OLGs can overlap on the same strand (sense) or on different strands (antisense) with their annotated mother gene, while overlapping either completely (embedded) or partially at the 5' or 3' end.

OLGs exist in all domains of life, including viruses, eukaryotes as well as prokaryotes. Probably the best studied genomes harbouring characterized OLGs are those of viruses. The first overlapping gene pair was indeed identified in a virus, the bacteriophage ΦX174, as early as 1976 (Barrell *et al.*). Since then, many trivial as well as non-trivial gene overlaps have been detected (e.g., Neuhaus *et al.*, 2010). Schlub & Holmes (2020) recently analysed 5,976 genomes of RNA and DNA viruses with different genome structures and found 53% of them to have at least one gene overlap of more than 50 bp. This high prevalence of viral OLGs was for a long time believed to be an exclusive result of spatial limitations of the viral capsid, thus favouring small genome sizes (Chirico *et al.*, 2010). In this context, OLGs were hypothesized to expand the coding capacity without affecting genome size (Belshaw *et al.*, 2007). For these reasons, OLGs have been generally presumed to be a commonplace feature of viral genomes. In contrast, the presence of OLGs is rarely considered in other living organisms due to their larger genome sizes. Nevertheless, several publications demonstrated the occurrence of OLGs in eukaryotes, including mammalians like humans and mice (Sanna *et al.*, 2008, Veeramachaneni *et al.*, 2004). However, since eukaryotic OLGs are often located within the intragenic region of annotated genes and removed after splicing, these genes cannot be considered as ‘real’ OLGs according to our definition. In prokaryotes, gene overlaps are very common and account for approximately 30% of all microbial genes (Johnson & Chisholm, 2004). The vast majority among them shows same strand overlaps of a few base pairs, which often facilitate the joint transcriptional and translational regulation of both genes (Johnson & Chisholm, 2004, Scherbakov & Garber, 2000). In contrast, non-trivial OLGs are assumed to be rare due to a severe evolutionary constraint of the genetic information.

Mutational changes in the coding sequence of one gene may also affect the other overlapping gene, restricting its functionality and adaptability (Krakauer, 2000, Miyata & Yasunaga, 1978). Therefore, the existence of non-trivial OLGs in prokaryotes is generally called into question, and already annotated OLGs have been designated as mis-annotations (Pallejà *et al.*, 2008). In concordance with this assumption, detection of non-trivial OLGs is hampered by prokaryotic prediction algorithms like Glimmer (Delcher *et al.*, 2007) or Prodigal (Hyatt *et al.*, 2010) due to their systematic exclusion of overlapping ORFs (Warren *et al.*, 2010). Hence, only the ORF with the highest score is chosen for annotation. Consequently, few non-trivial OLGs have been described in literature so far, often merely discovered by serendipity (e.g., Haycocks & Grainger, 2016, Balabanov *et al.*, 2012, Jensen *et al.*, 2006). Even in the well-studied model organism *E. coli* the number of known, non-trivial OLGs is severely limited, and a detailed experimental characterization was implemented for even fewer. So far, the best studied prokaryotic OLGs in this organism are (in chronological order):

a) *astA/tnpA*: This overlapping gene pair was found to be encoded on a virulence plasmid in a human-derived *E. coli* strain. The 117 bp long gene *astA* is fully embedded in the transposon-like gene *tnpA* and exhibits enterotoxin activity after expression (McVeigh *et al.*, 2000).

b) *setAB/pic*: The protein products of *setAB* constitute different subunits of the oligomeric ShET1 enterotoxin. Both genes overlap with the *pic* gene encoding for a 109-kilodalton (kDa) secreted mucinase. *SetA* and *setB* seemed to be transcribed from the same transcript and showed a regulated expression, which was divergent from their mother gene (Behrens *et al.*, 2002).

c) *ardD/tniA*: *ArdD* is located within the *tniA* gene, which is a non-conjugative transposon gene mediating mercury resistance. *ArdD* confers anti-restriction activity against EcoKI, while showing only modest similarities to other known Ard proteins (Balabanov *et al.*, 2012).

d) *morA/yghX*: The *morA* gene encodes for a small peptide with an unusal AA composition and overlaps completely with the putative pseudogene *yghX*. Transcription and translation of *morA* were experimentally confirmed, and the molybdenum-dependent transcription factor ModE was shown to repress promoter activity (Kurata *et al.*, 2013).

e) *htgA/yaaW*: Transcription of this antisense overlapping gene pair was confirmed by promoter-fusions and 5'RACE, and its functionality was indicated by differential phenotypes after introduction of stand-specific knockouts (Fellner *et al.*, 2014). *HtgA* encodes for a positive regulator of the σ^{32} heat shock promoter (Delaye *et al.*, 2008, Missiakas *et al.*, 1993); the function of *yaaW* is unknown.

f) *nog1/citC*: *Nog1* represents a novel OLG, which presumably arose by genetic overprinting. It is completely embedded in the citrate lyase ligase-encoding *citC* gene and its expression is specifically upregulated in cow dung (Fellner *et al.*, 2015).

g) *aatS/aatC*: Haycocks & Grainger (2016) showed that transcription of *aatS*, a small gene located within the type I secretion system gene *aatC*, is regulated by the cAMP receptor protein CRP. *AatS* shows structural features of a protein-coding gene, but a potential functionality remains to be elucidated.

h) *laoB/*ECs5115: *LaoB* encodes for a hypothetical protein of 41 AAs with unknown function and is completely embedded within the annotated gene, ECs5115. The latter encodes for a transcriptional regulator of the CadC family. Both genes showed signs of expression in transcriptome sequencing and ribosome profiling data and *laoB* functionality was analysed in competitive growth experiments (Hücker *et al.*, 2018b).

i) *ano/*ECs2385: *Ano* is encoded in frame -3 and overlaps almost completely with the transpeptidase gene, ECs2385. It has been shown to be part of a bicistronic operon and has a strong phenotype under anaerobic conditions (Hücker *et al.*, 2018a).

j) *asa/yccA*: Vanderhaeghen *et al.* (2018) identified the novel gene *asa*, which is strongly regulated by growth phase and NaCl. It is predicted to encode for a disordered, secreted protein and overlaps completely in frame -2 with the putative TEGT transporter gene *yccA*.

k) *pop/ompA*: *Pop* and *ompA* represent the most recent example of an overlapping gene pair. With a length of 603 nt, *pop* is one of the longest, hitherto known OLGs, which is completely embedded into the mother gene. It seems to be regulated by pH and its final protein could be detected after overexpression (Zehentner *et al.*, 2020b).

Additional studies in *E. coli* broadened the general knowledge of OLGs and their abundance in this species. Especially the application of novel high-throughput methods like ribosome profiling and modified variants of it made a substantial contribution to the detection of several OLGs (e.g., Meydan *et al.*, 2019, Weaver *et al.*, 2019). In addition to growing experimental evidence, statistical analyses of bacterial genomes revealed that ORFs in alternative reading frames are more frequent than expected. Mir *et al.* (2012), for instance, showed that the occurrence of long OLGs in frame -1 can be explained partly by the codon usage of the annotated gene. However, frame -2 and -3 also showed a significant depletion of stop codons, favouring the emergence of long OLGs. These findings provided initial evidence for a protective selection pressure on stop codon fixation caused by the coding capacity of an alternative, overlapping ORF. The average length of the ORF thereby depended on the overall GC-content of the respective genome. In general, GC-rich organisms could harbour longer protein-coding genes because they contain less AT-rich stop codons (Mir *et al.*, 2012, Oliver & Marín, 1996). As a consequence, those organisms are also expected to have a higher number of non-trivial OLGs (Merino *et al.*, 1994).

In fact, several OLGs have been successfully identified in the GC-rich genus *Pseudomonas* so far. By performing *in vivo* expression technology, Silby & Levy (2004) detected ten soil-induced genes, which overlapped antisense with annotated genes in *Pseudomonas fluorescens* Pf0-1. Four years later, one of the detected overlapping gene pairs, named *cosA*/Pfl_0939, was further characterized. Translation of both genes was confirmed by reverse transcription (RT)-polymerase chain reaction (PCR) and mutational as well as complementation studies revealed a potential role of *cosA* in soil colonization (Silby & Levy, 2008). The role of the overlapping gene pairs *iiv19/leuA2 and iiv8/ppk* in soil environments was examined in follow-up studies (Silby *et al.*, 2009, Kim & Levy, 2008). In an additional mass spectrometry study, peptide mapping confirmed the protein-coding nature of ten further OLGs, nine of them overlapping antisense and one of them overlapping sense with their mother genes (Kim *et al.*, 2009). Despite *P. fluorescens,* strand-specific transcriptome sequencing revealed antisense transcription of several genes in the plant pathogen, *Pseudomonas syringae* pathovar tomato DC3000 (Filiatrault *et al.*, 2010). A similar study was also performed for the human pathogenic strain, *Pseudomonas aeruginosa* PA14, indicating the potential presence of OLGs in this species (Eckweiler & Häussler, 2018). However, the existence of OLGs in the following genera, *Streptomyces* (Tunca *et al.*, 2009), *Mycobacterium* (Smith *et al.*, 2019), *Bacillus* (Wang *et al.*, 1999a), *Deinococcus* (Willems *et al.*, 2020), *Synechocystis* (Cheregi *et al.*, 2012) and others, indicates that non-trivial OLGs are a ubiquitous feature of prokaryotic organisms belonging to diverse phyla.

1.1.3 Origin and evolution of overlapping genes

Although initially assumed, compression of the genome size seems unlikely to be the key driver of OLG emergence in prokaryotes (Johnson & Chisholm, 2004). In concordance with this assumption, low volume utilization of the viral capsid necessitates another explanation for gene overlapping, e.g., evolution exploration or gene novelty (Brandes & Linial, 2016).

Several possibilities exist of how novel genes can arise. One of the first discovered mechanisms was gene duplication (Ohno, 1970). For a long time, this mechanism was believed to be the only way how new genes originate. Consequently, the almost exclusive prevailing dogma was that genetic novelty can only emerge from existing genes (Tautz, 2014). During the process of gene duplication, a redundant copy of a subsisting gene is generated, which then has four possible destinies: It can either be maintained to conserve the original function, become a pseudogene (pseudogenization), or develop a new (neofunctionalization) or subdivided function (subfunctionalization) of the former ancestral gene (Zhang, 2003). The latter can contribute to the emergence of genetic novelty due to an evolutionary loss of similarity to the parental gene (Domazet-Loso & Tautz, 2003, Schmid & Aquadro, 2001). Other sources of genetic novelty based on present structures include domestication of transposable elements (Jangam *et al.*, 2017), gene fusion and fission (Pasek *et al.*, 2006) as well as horizontal transfer of genes from unrelated genomes (Boto, 2010). All mentioned processes have the advantage of providing existent structures like promoters or SD sequences necessary for gene expression. However, the widespread detection of orphan genes without any gene homology in other lineages suggested an alternative way of gene emergence. Orphan genes were detected as early as 1992 when the first entire yeast chromosome III was completely sequenced (Oliver *et al.*, 1992). At this time, approximately a third of all detected genes showed no sequence similarity to other known genes, including those of the same species (Dujon, 1996, Casari *et al.*, 1996). This observation was initially thought to be the result of a lack of sequenced genomes, and thus, of known homologues at that time. Notwithstanding, this hypothesis has not been confirmed with increasing completion of genome projects (Khalturin *et al.*, 2009, Wilson *et al.*, 2005). Quite the contrary, orphans have not only been discovered in a multitude of eukaryotic genomes in the subsequent years, but also with varying percentages in bacteria and archaea (Fukuchi & Nishikawa, 2004). The mere number and the ubiquitous appearance of orphans suggests the possibility of another type of gene evolution despite duplication and rapid-divergence, namely *de novo* gene birth (for review see Van Oss & Carvunis, 2019). This process describes the evolution of genes from formerly non-coding sequences and had been considered highly unlikely for a long time (Jacob, 1977). The underlying mechanisms of *de novo* gene birth are still obscure; possible mechanisms include continuum evolution or preadaptation. In the first case, a formerly non-coding sequence becomes coding by random mutational events. For example, the emergence of a start codon or the deletion of a stop codon may lead to the development of transitory proto-genes (Carvunis *et al.*, 2012). These proto-genes are transcribed and translated pervasively at low levels and constitute a pool of evolutionary innovation. If the novel translated peptide is beneficial for organisms, the proto-gene can gradually maturate into a functional gene fixed by positive selection (**Figure 4**). Alternatively, proto-genes without adaptive potential can lose their protein-coding ability over time.

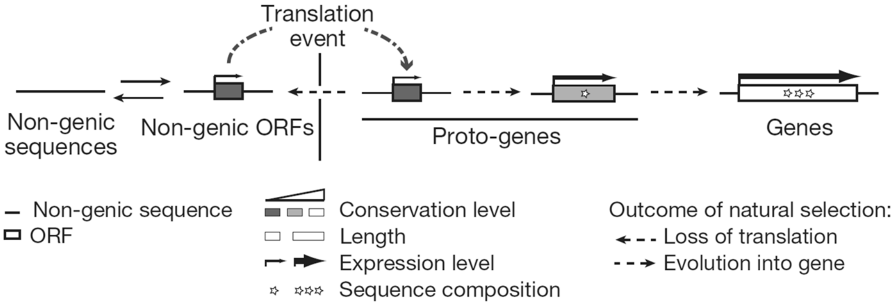


Figure 4. *De novo* evolution of new genes via proto-genes. Non-coding sequences become coding by random mutational events. Upon translation, the risen proto-genes could successively evolve into genes, while other proto-genes without beneficial properties might get lost over time. Figure by Carvunis *et al.* (2012).

However, the continuum evolution hypothesis does not unequivocally explain how genes can emerge without forming deleterious aggregations (Monsellier & Chiti, 2007). As an alternative, preadaptation was suggested to be the underlying mechanism of *de novo* gene emergence. In contrast to continuum evolution, preadaptation describes an all or nothing process, during which only a preadapted sequence with gene-like structure becomes fixed (Wilson & Masel, 2011). One special type of preadaptation is overprinting (Grassé, 1977). In this case, a formerly non-coding sequence, overlapping an existing gene located in a different reading frame, becomes coding. It has been suggested that OLGs might be the result of overprinting in viruses (Sabath *et al.*, 2012, Keese & Gibbs, 1992), eukaryotes (Neme & Tautz, 2013, Makalowska *et al.*, 2005) and prokaryotes (Hücker *et al.*, 2018b, Fellner *et al.*, 2015). Indeed, OLGs show several indications of originating from *de novo* gene birth. They are typically characterized by a divergent codon composition compared to non-overlapping genes and are usually less conserved than their annotated counterparts (Fellner *et al.*, 2015, Delaye *et al.*, 2008). These findings indicate different evolutionary ages and support the assumption that OLGs evolved rather recently (Rogozin *et al.*, 2002a). As a consequent, OLGs often encode for structural disordered proteins (Willis & Masel, 2018, Rancurel *et al.*, 2009) with accessory, non-essential functions (Moshensky & Alexeevski, 2019, Chen *et al.*, 2012), offering an possible explanation for their rather low expression (Landstorfer *et al.*, 2014). Their large percentage of high-degeneracy amino acids, in addition, may favour their birth and enhances the toleration towards point mutations (Pavesi *et al.*, 2018). Nonetheless, two genes encoded at the same DNA locus are subjected to a severe evolutionary constraint (Krakauer, 2000), which differs depending on the frame. On the one hand, a sequence constraint provided by the mother gene could favour the emergence of OLGs. An OLG encoded in frame -2, for instance, is protected from non-synonymous substitutions if the annotated gene in frame +1 is under purifying selection (Mir *et al.*, 2012, Rogozin *et al.*, 2002a). On the other hand, the sequence constraint restricts the evolvability, and thus, the functionality of the overlapping gene pair. In this context, a *de novo* gene overlapping in frame -3 with the annotated gene in +1 is evolutionary favoured due to a higher degree of sequence freedom (Krakauer, 2000). However, the vast majority of the aforementioned studies have been conducted using eukaryotes or viruses, whereas detailed analyses of the evolution and the properties of OLGs in prokaryotes are still lacking.

1.2 The model organisms

1.2.1 *Escherichia coli*

1.2.1.1 Discovery and properties

Escherichia coli was first described in 1885 by the German-Austrian paediatrician, Theodor Escherich. He noticed a hitherto unknown bacterial population in the meconium and faeces of neonates, which he termed *Bacterium coli commune* (Escherich & Bettelheim, 1988). In 1919, Castellani and Chalmers introduced the name *Escherichia coli* out of respect for its initial discoverer.

E. coli is a rod-shaped, flagellated Gram-negative bacterium belonging to the class of γ-Proteobacteria. This species is characterized by a high phenotypic diversity, flexibility in energy and carbon utilization and the capability to colonize a wide range of environmental niches (Brennan *et al.*, 2013, Schaechter, 2009, Ishii & Sadowsky, 2008, Díaz *et al.*, 2001). As a commensal bacterium, *E. coli* is frequently found in the intestine of warm-blooded animals including humans, where it constitutes the most abundant facultative anaerobic species (Tenaillon *et al.*, 2010, Savageau, 1983). In the laboratory context, *E. coli* is considered as the “working horse” for biologists and biotechnologists of several fields. This organism is not merely utilized to unravel fundamental biological and evolutional processes, but also to produce compounds for medicinal and industrial applications by metabolic engineering (for review see Blount, 2015, or Chen *et al.*, 2013) for several reasons. Those include, amongst others, a short generation time, availability of manipulation techniques and detailed knowledge about the genetic and genomic features. Based on whole-genome data as well as multi-locus sequence typing, eight different phylogenetic groups (A, B1, B2, C, D, E, F and *Escherichia* clade I) can be distinguished (Clermont *et al.*, 2013); all of them showing different phenotypic and genotypic properties (Méric *et al.*, 2013).

1.2.1.2 Genome and classification

The first sequenced genome of *E. coli* was reported in 1997 for the laboratory strain K-12 MG1655. With a size of 4.6 Mbp and a nearly balanced GC content of 50.8%, this genome harbours 4,288 protein-coding genes, of which up to one-third are of unknown function (Hu *et al.*, 2009, Blattner *et al.*, 1997). To date, the genomes of many other strains have been sequenced. All of them share a core genome consisting of 1,000 to 3,000 genes (Lukjancenko *et al.*, 2010, Rasko *et al.*, 2008), which can be distributed among the chromosome and one or multiple plasmids. Mobile genetic elements like transposable elements, plasmids, bacteriophages, and genomic islands additionally expand the core genome, and thus, contribute to the genomic plasticity of *E. coli.* Those elements often carry virulence and antimicrobial resistance-associated genes and can be disseminated via horizontal gene transfer (Rankin *et al.*, 2011). Acquisition or loss of mobile genetic elements can convert commensal into virulent strains and *vice versa* (Touchon *et al.*, 2009), resulting in genome size differences of up to 1 Mb within the species. In contrast to apathogenic *E. coli*, pathogenic strains can cause various diseases including enteric illnesses like gastroenteritis or extra-intestinal diseases like urinary tract infections or meningitis (Kaper *et al.*, 2004). According to Nataro & Kaper (1998), six enteric pathotypes can be classified, namely, which are: diffusely adherent *E. coli* (DAEC), enteroaggregative *E. coli* (EAEC), enterohemorrhagic *E. coli* (EHEC), enteroinvasive *E. coli* (EIEC), enteropathogenic *E. coli* (EPEC) and enterotoxigenic *E. coli* (ETEC). Each pathotype is characterized by the presence of specific virulence attributes resulting in different pathogenicity profiles. Additionally, Boudeau *et al.* (1999) described a new pathotype of *E. coli*, recommending it to be named adherent-invasive *E. coli* (AIEC).

1.2.1.3 AIEC pathotype and strain LF82

The AIEC pathotype has been proposed to be involved in the etiology of inflammatory bowel diseases like Crohn´s disease (CD) or ulcerative colitis (Eaves-Pyles *et al.*, 2008, Darfeuille-Michaud & Colombel, 2008). These illnesses are characterized by a chronic inflammation of the gastrointestinal tract and are often associated with a local dysbiosis in the microbial composition (Tamboli *et al.*, 2004). Several mouse as well as human studies suggest AIEC strains play an important role in the induction of intestinal inflammation, and thus, may contribute to the development of inflammatory bowel diseases. Darfeuille-Michaud *et al.* (2004), for instance, showed that the prevalence of AIEC strains in the ileal mucosa is significantly increased in CD patients compared to healthy control individuals. Other independent research groups (e.g., Dogan *et al.*, 2013, Martinez-Medina *et al.*, 2009, Baumgart *et al.*, 2007) made similar observations. In addition, AIECs are able to induce typical histopathological hallmarks of CD like the formation of granulomas (Meconi *et al.*, 2007). However, there is still no unequivocal proof of whether AIECs are the initial trigger or just an enhancing factor in disease development and maintenance. The presence in healthy humans indicates the potential pathosymbiontic nature of AIECs, taking advantage of inflammation-related perturbations in the gut of susceptible individuals (Palmela *et al.*, 2018, Dogan *et al.*, 2014). In concordance with this hypothesis are the findings of Craven *et al.* (2012), who showed that a moderate to severe ileitis accompanied by inflammation favours the invasion and colonization of AIECs in a mouse model. In terms of virulence genes, AIECs are very diverse and clearly separate from the diarrheagenic pathotypes. Typical characteristics for this pathotype include 1) the ability to adhere to intestinal cells (Darfeuille-Michaud *et al.*, 1998), 2) the potential to invade epithelia cells via interacting with microtubule and actin microfilaments of the host cell (Boudeau *et al.*, 1999), and 3) the capacity to survive and replicate intracellularly in macrophages without inducing cell death (Glasser *et al.*, 2001). Despite sharing these common traits, AIEC strains show a high genetic and clonally diversity, and belong to different phylogenetic groups and serotypes (Camprubí-Font *et al.*, 2020, Céspedes *et al.*, 2017).

The strain LF82 (serotype O83:H1; phylogroup B2) represents one of the current best characterized AIECs and was also a subject of this study. It was originally isolated from an ileal lesion of a patient with Crohn's disease (Boudeau *et al.*, 1999), and shares several virulence determinants with extraintestinal pathogenic bacteria, indicating its potential to colonize other sites despite the gut (Conte *et al.*, 2016). Important virulence factors include the type I pili. They mediate the adhesion of the bacterium to the intestinal epithelium by binding to the antigenic cell adhesion molecule 6 on the apical surface of ileal enterocytes (Barnich *et al.*, 2007, Boudeau *et al.*, 2001). The subsequent invasion is promoted by the outer membrane protein OmpA, which recognizes the stress response factor Gp96 on the ileal epithelium, leading to their fusion with outer membrane vesicles (Rolhion *et al.*, 2010). After internalization, LF82 upregulates further virulence factors like the stress protein HtrA (Bringer *et al.*, 2005) or enzymes like the oxidoreductase DsbA *(Bringer et al.*, 2007) to survive and to replicate within macrophages, defying harsh acidic and oxidative conditions. Further knowledge of the genetic determinants for *E. coli* LF82s' remarkable adaptability to variable environments is necessary to elucidate the role of AIECs in inflammatory bowel diseases and to elaborate prevention and treatment strategies.

1.2.2 *Pseudomonas aeruginosa*

1.2.2.1 Discovery and properties

Pseudomonas aeruginosa, originally named *Bacillus pyocyaneus,* was first isolated in 1882 by the pharmacist, Carle Gessard. Gessard noticed that the isolated organism produces blue-green pus in cutaneous wounds of patients. This blue-green purulence, most likely caused by the phenazine metabolite pyocyanin, formed the basis for its later renaming. After Walter Migula (1894) proposed the generic term, "*Pseudomonas*", for the whole genus, *Bacillus pyocyaneus* was renamed as *Pseudomonas aeruginosa.* The etymology of the species name thereby originates from the Latin word *aerūgō* (*"*verdigris"*),* thus referring to the ability of this organism to produce blue-green pigments (Palleroni, 2010).

Like the genus *Escherichia, P. aeruginosa* belongs to the class of γ-Proteobacteria and is a Gram-negative, rod-shaped and motile bacterium. As a facultative anaerobic organism, it is able to acquire energy from both aerobic and anaerobic respiration; for the latter, nitrite or nitrate is used as a terminal electron acceptor instead of oxygen (Davies *et al.*, 1989). In the entire absence of electron acceptors, *P. aeruginosa* can also ferment arginine (Vander Wauven *et al.*, 1984) or pyruvate (Eschbach *et al.*, 2004) for energy production. More than 100 different substances including aromatic hydrocarbons (Diggle & Whiteley, 2020) can be utilized as carbon and energy sources by different strains, confirming the exceptional metabolic versality (Frimmersdorf *et al.*, 2010, Ramos, 2004) of this organism. Simultaneously, *P. aeruginosa* has only modest nutritional requirements (Palleroni *et al.*, 1984) and can tolerate a broad range of temperatures from 4 to 42 °C (LaBauve & Wargo, 2012). These properties enable the colonization of a variety of different habitats with varying nutritional and physical conditions, including terrestrial ecosystems like soil (Szoboszlay *et al.*, 2003, Green *et al.*, 1974) as well as aquatic environments like sewage (Wheater *et al.*, 1980) or sea water (Kimata *et al.*, 2004).

1.2.2.2 Genome

The comparatively large and complex genome of *P. aeruginosa* mirrors its impressive adaptability. In 2000, Stover *et al.* sequenced the first genome of *P. aeruginosa,* which was those of the strain, PAO1. The researchers observed a genome with a size of 6.3 Mbp, an average GC content of 66.6% and a high genetic complexity. The genome sequenced harboured 5,570 predicted genes, many of which belong to paralogous gene families and encode functionally diverse products. In concordance with the exceptional versatility of *P. aeruginosa*, a high percentage of genes encoded for proteins involved in nutrient uptake and metabolism. In addition, 8.4% of all PAO1 genes were shown to have a regulatory function. This is one of the highest proportions of regulatory genes detected for bacterial genomes and enables *P. aeruginosa* to respond to fluctuating environments. After initial sequencing of strain PAO1, many other strains were also subjected to genome sequencing. Nowadays, fully assembled genome sequences of 224 different *P. aeruginosa* strains are available on NCBI with sizes ranging from 6.12 to 7.76 Mbp (https://www.ncbi.nlm.nih.gov/; state: 04/16/20). All strains share a gene density of approximately 0.9 genes per 1 kb (Rogozin *et al.*, 2002a) and have a highly conserved core genome (Wolfgang *et al.*, 2003). In contrast, the accessory genome, constituting up to 18% of the total genome (Ozer *et al.*, 2014), can vary widely. In both, the core and the accessory genome, several genes involved in antibiotic resistance and virulence can be found (Valot *et al.*, 2015).

1.2.2.3 Antibiotic resistance mechanisms

P. aeruginosa shows a low susceptibility to many available antimicrobial agents, including β-lactam antibiotics, aminoglycosides, polymyxins and fluoroquinolone-based substances (for review see Poole, 2011). Moreover, three different modes of antibiotic resistance can be distinguished: intrinsic, acquired & adaptive resistance.

Intrinsic resistance refers to an organism's inherent ability to circumvent the adverse effects of antibiotics due to innate structural or functional features (Blair *et al.*, 2015). For *P. aeruginosa*, intrinsic resistance is primarily conferred by its limited outer membrane permeability (Angus *et al.*, 1982, Yoshimura & Nikaido, 1982) and the capacity to synthesise antibiotic-inactivating enzymes (Lambert, 2002). An additional factor, which contributes to the intrinsic resistance of *P. aeruginosa,* is the presence of resistance-nodulation-cell division (RND) multidrug efflux systems. Those efflux systems restrict the effect of various substances like toxins, heavy metals, organic molecules, endogenous metabolites and antibiotics due to proton-driven excretion from the cells' cytoplasm (Blanco *et al.*, 2016). To exert this function, three subunits are necessary (**Figure 5**): the actual RND transporter located in the cytoplasmic membrane (Murakami *et al.*, 2002), an protein complex spanning the outer membrane (outer membrane factor, OMF) (Wong *et al.*, 2001, Koronakis *et al.*, 2000) and an periplasmic "adaptor" protein (membrane fusion protein, MFP), thereby stabilizing the interaction between the other two subunits (Takatsuka & Nikaido, 2009, Zgurskaya & Nikaido, 1999). All three components are necessary for the functionality of the efflux system (Ma *et al.*, 1995, Ma *et al.*, 1993). In *P. aeruginosa* PAO1, up to 10 different genes encoding for RND efflux systems are present (Stover *et al.*, 2000). Each system is regulated independently (Poole, 2008), and they partially complement each other with respect to their substrate specificity (Fernando & Kumar, 2013). Up to now, five RND pumps are known to be involved in the expulsion of antibiotics out of the cell, namely: MexA$^{\mathrm{MFP}}$-MexB$^{\mathrm{RND}}$-OprM$^{\mathrm{OMF}}$, MexX$^{\mathrm{MFP}}$-MexY$^{\mathrm{RND}}$-OprM$^{\mathrm{OMF}}$, MexC$^{\mathrm{MFP}}$-MexD$^{\mathrm{RND}}$-OprJ$^{\mathrm{OMF}}$, MexE$^{\mathrm{MFP}}$-MexF$^{\mathrm{RND}}$-OprN$^{\mathrm{OMF}}$ and MexJ$^{\mathrm{MFP}}$-MexK$^{\mathrm{RND}}$-OprM$^{\mathrm{OMF}}$ (Housseini B Issa *et al.*, 2018, Li *et al.*, 2015, Lister *et al.*, 2009). Probably the best studied efflux pump in *P. aeruginosa* is MexA$^{\mathrm{MFP}}$-MexB$^{\mathrm{RND}}$-OprM$^{\mathrm{OMF}}$, which is a homologue to the AcrA$^{\mathrm{MFP}}$-AcrB$^{\mathrm{RND}}$-TolC$^{\mathrm{OMF}}$ efflux system in *E. coli.* This is the only efflux system with constitutive expression and is able to eject a large number of structurally unrelated groups of antibiotics (Masuda *et al.*, 2000). A chromosomal deletion of one or more MexA$^{\mathrm{MFP}}$-MexB$^{\mathrm{RND}}$-OprM$^{\mathrm{OMF}}$ genes increases, *inter alia*, the susceptibility against quinolones, β-lactam antibiotics, tetracyclin, chlorampehicol and aminoglycoside antibiotics compared to the wild type strain (Lomovskaya *et al.*, 1999, Yoneyama *et al.*, 1997, Gotoh *et al.*, 1994, Poole *et al.*, 1993). When deleting multiple efflux pump genes, susceptibility is even more pronounced. The *P. aeruginosa* strain, PAO397 is a derivate of PAO1, lacking all five efflux pumps involved in antibiotic expulsion (Δ*mexAB-oprM,* Δ*mexCD-oprJ*, Δ*mexJKL*, Δ*mexXY* & Δ*mexEF-oprN*), and additionally, carrying a Δ*opmH*362 mutation (Chuanchuen *et al.*, 2005). This strain, for example, shows a reduced minimal inhibitory concentration (MIC) of triclosan by a factor ≥64, when compared to the PAO1 parent strain (Chuanchuen *et al.*, 2003). In addition, an elevated MIC was also observable for several other antibiotics including substances used for the treatment of human *P. aeruginosa* infections, e.g., aztreonam and carbenicillin (Iyer & Erwin, 2015, Giamarellou & Antoniadou, 2001).

Spontaneous mutations affecting efflux pump systems also play an important role in acquired resistance. In this case, overexpression of the efflux pumps MexA$^{\mathrm{MFP}}$-MexB$^{\mathrm{RND}}$-OprM$^{\mathrm{OMF}}$ (Ziha-Zarifi *et al.*, 1999), MexC$^{\mathrm{MFP}}$-MexD$^{\mathrm{RND}}$-OprJ$^{\mathrm{OMF}}$ (Jakics *et al.*, 1992) and MexE$^{\mathrm{MFP}}$-MexF$^{\mathrm{RND}}$-OprN$^{\mathrm{OMF}}$ (Fukuda *et al.*, 1995) is induced by mutational changes in their transcriptional regulators, thus, leading to a de-repression of the efflux pump expression (Stickland *et al.*, 2010, Srikumar *et al.*, 2000, Saito *et al.*, 1999). Besides mutational alterations, resistance can also be acquired by the uptake of foreign antimicrobial genes like extended-spectrum β-lactamases or metallo-β-lactamases via transformation, transduction or conjugation (Arber, 2014, Bush, 2010).

In contrast to intrinsic and acquired resistance, adaptive resistance is triggered by environmental stimuli leading to transient alterations in gene and protein expression (Fernandez & Hancock, 2012). A still enigmatic acquired resistance mechanism in *P. aeruginosa* is the formation of persister cells, which can enter a dormant state, when exposed to antibiotics or other environmental stresses (Balaban *et al.*, 2013). Especially in cystic fibrosis patients, persistence and the thereby conferred multidrug tolerance constitutes an emerging challenge (Bianconi *et al.*, 2019, Mulcahy *et al.*, 2010).

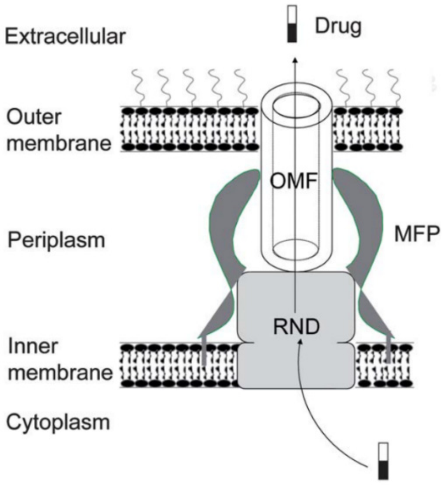


Figure 5. Schematic illustration of an RND pump by Wang *et al.* (2012). OMF: outer membrane factor. MFP: membrane fusion protein. RND: resistance-nodulation-cell division transporter.

1.2.2.4 Pathogenicity and clinical relevance

P. aeruginosa possesses a sophisticated virulence machinery leading to a high degree of pathogenicity. One of the first steps in pathogenesis include colonization and adhesion, which are facilitated by surface structures like type IV pili (Tang *et al.*, 1995) and flagella (Balloy *et al.*, 2007). Once attached to the host cells, *P. aeruginosa* can secret effector proteins via six different secretion systems (type I – type VI secretion systems; T1SS -T6SS) (Filloux, 2011); of which

T2SS and T3SS are of decisive importance during pathogenesis (Jyot *et al.*, 2011, Roy-Burman *et al.*, 2001). The T2SS mediates secretion of enzymes like alkaline phosphatase PhoA, elastase LasA and LasB, exotoxin A as well as hemolytic and nonhemolytic phospholipases C into the extracellular space (Cianciotto, 2005). In contrast, T3SS enables direct injection of toxins including exotoxin S, T, U and Y into eukaryotic cells, where they can promote immune evasion and induce tissue damage (Engel & Balachandran, 2009). Other secreted molecules involved in pathogenesis are alginates as well as siderophores. The former promote biofilm formation, and thus, enhance resistance towards antibiotics and the host immune system (May *et al.*, 1991); the latter ensure a sufficient availability of iron for bacterial growth and dissemination (Takase *et al.*, 2000). This extensive repertoire of virulence factors enables *P. aeruginosa* to infect various hosts, such as humans, animals, plants, insects, amoebas, and nematodes (Pukatzki *et al.*, 2002, Mahajan-Miklos *et al.*, 2000, Plotnikova *et al.*, 2000, Rahme *et al.*, 2000, Mahajan-Miklos *et al.*, 1999). In humans, a transient, asymptomatic colonization of *P. aeruginosa* is not unusual. Estepa *et al.* (2014), for instance, reported occurrence of *P. aeruginosa* in the faeces of 8.2% of all healthy individuals tested. Similar results were obtained by Morrison & Wenzel (1984), who reported an endogenous colonization rate in faecal samples ranging from 2.6% to 24%. However, colonization rates may be even higher during hospitalization (Valles *et al.*, 2004). This is in concordance with the opportunistic nature of *P. aeruginosa*, which predominately infects individuals with compromised immune systems like patients suffering from severe burns, cancer, AIDS or cystic fibrosis (Govan & Deretic, 1996). A point prevalence study conducted in 75 countries revealed that 51.5% of all patients in intensive care units suffered from a bacterial infection, nearly 20% of them caused by *P. aeruginosa* (Vincent *et al.*, 2009). The most common hospital-acquired infections induced by *P. aeruginosa* are ventilator-associated infections of the respiratory tract, followed by infections of the urinary tract, surgical sites and bloodstream infections; all of them associated with high morbidity and mortality rates (Weinstein *et al.*, 2005). Due to a global increase in the prevalence of multi-drug resistant strains, effective treatment options are limited (Bassetti *et al.*, 2018, Livermore, 2009, Peña *et al.*, 2009). Therefore, *P. aeruginosa* infections remain an ongoing challenge for today's health care systems.

1.3 Identification of protein-coding genes

1.3.1 *In silico* gene prediction

In the last few years, new sequencing technologies have enabled affordable, high-throughput generation of sequence data (Kahvejian *et al.*, 2008, Metzker, 2005). This has led to a massive increase in the number of available eukaryotic as well as prokaryotic genomes. After genome assembly, a crucial part for subsequent studies is an accurate and reliable prediction of genes and their structural and functional annotation. To cope with the tremendous amount of sequencing data, this step is usually carried out by automated, computational methods with only limited manual curation. In general, two different approaches are used for gene delineation: *ab initio* prediction and homology-based prediction (for review see Overbeek *et al.*, 2007, Zhang, 2002). In the first case, typical gene motifs like promoters, start and stop codons but also statistical properties, e.g., the GC-content or codon usage, were utilized for gene annotation by various programs like Glimmer (Delcher *et al.*, 2007, Delcher *et al.*, 1999, Salzberg *et al.*, 1998), GeneMark (Besemer & Borodovsky, 2005, Borodovsky & McIninch, 1993) or Prodigal (Hyatt *et al.*, 2010). In contrast, homology-based prediction, for example, using blast (Altschul *et al.*, 1990), exclusively relies on sequence similarity to genes with known functions. However, both approaches suffer from certain limitations, particularly regarding sensitivity. Especially novel genes, without any significant homology to the ones deposited in databases as well as genes not complying with the criteria used in *ab initio* predictions, were often missed. In addition, many prediction pipelines have an arbitrary minimal gene size threshold of 110-150 bp (Wood *et al.*, 2012, Boekhorst *et al.*, 2011, Basrai *et al.*, 1997) in order to minimize false-positive hits and avoid over-prediction at the expense of under-prediction. As a result, many prediction methods fail to detect small protein-encoding genes (e.g., Kremer *et al.*, 2016, Wood *et al.*, 2012, Warren *et al.*, 2010). Another restricting factor for prokaryotic prediction algorithms is that they eliminate overlapping ORFs by allowing only one protein-coding ORF prediction per DNA locus (Warren *et al.*, 2010, Delcher *et al.*, 2007).

1.3.2 Experimental gene identification by high-throughput methods

1.3.2.1 RNA-seq & Cappable-seq

To overcome the limitations of computational prediction, experimental data should be consulted in order to identify genomic regions with functionality (Richardson & Watson, 2013). However, experimental verification might be rather challenging, especially for small genes with weak expression. Molecular biological methods like sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS-PAGE) or Western blot often fail to detect small proteins due to inadequate resolving power, improper membrane transfer or poor retention. Next generation sequencing (NGS)-based methods are size-independent, and thus, also suitable for the verification of the coding capacity of short genes. Total RNA sequencing (RNA-seq), for example, enables cheap and rapid detection of RNA transcripts, and hence, is of increasing importance in transcriptome studies (Landstorfer *et al.*, 2014, Costa *et al.*, 2010). In contrast to methods like quantitative polymerase chain reaction (qPCR) or microarrays, RNA-seq enables whole genome expression analysis without prior specification of the transcripts to be identified (for review see Wang *et al.*, 2009). The typical RNA-seq workflow starts with cultivation of the target organism followed by total RNA purification. Since rRNAs constitute up to 95% of the total RNA within the bacterial cell (Giannoukos *et al.*, 2012), rRNA depletion is a crucial step during RNA-seq. After depletion, residual RNA is fragmented, ligated to adapters, and finally, converted to a complementary DNA (cDNA) library by RT and amplification. Several technologies are available for sequencing, whereby Illumina´s short-read sequencing represents the hitherto most commonly used technology. For this sequencing technique, the cDNA molecules are attached to the surface of a flow cell and copied via bridge amplification. During the sequencing process, fluorescent-labelled nucleotides are sequentially added, and incorporated nucleotides are visually detected after each cycle. The sequenced raw reads are bioinformatically processed, aligned against a reference genome, and then, evaluated according to the initial research question. Potential applications of RNA-seq include differential expression analysis (e.g., Landstorfer *et al.*, 2014), unveiling of transcriptional dynamics (e.g., Wilhelm *et al.*, 2008), identification of novel transcripts and revision of annotated gene and exon boundaries (e.g., Nagalakshmi *et al.*, 2008).

A more precise method to map boundaries of mRNA transcripts constitutes Cappable-seq (**Figure 6**), which was first described by Ettwiller *et al.* (2016). This method enables precise determination of TSS by selective 5´ end enrichment of primary transcripts. Those are characterized by a triphosphorylated nucleotide at their 5´ end, which is the first nucleotide that has been incorporated by RNAP. In contrast, the 5´ ends of processed transcripts including those of rRNAs and tRNAs are monophosphorylated. The different phosphorylation states are utilized to separate the two types of RNA: Only triphosphorylated fragments are capped with 3´-desthiobiotin-TEG-guanosine 5´triphosphate (DTB-Gppp), fragmented to 200 bp, and subsequently, captured by reversible interaction with streptavidin beads. In contrast, monophosphorylated transcripts do not interact with streptavidin beads, and thus, are eliminated from the sample. After elution of the captured primary transcripts, the desthiobiotinylated cap is removed, and the RNA fragments are subjected to library preparation and sequencing. This technique enables robust genome-wide identification of TSS at single base resolution, while reducing data complexity and archiving high rRNA depletion rates. By applying Cappable-seq to *E. coli*, Ettwiller *et al.* (2016) identified 16,539 TSS, whereby 41% have not been identified before, probably due to their weak expression. Therefore, this method seems to be rather sensitive, and thus, also suitable for TSS identification of transcribed OLGs.

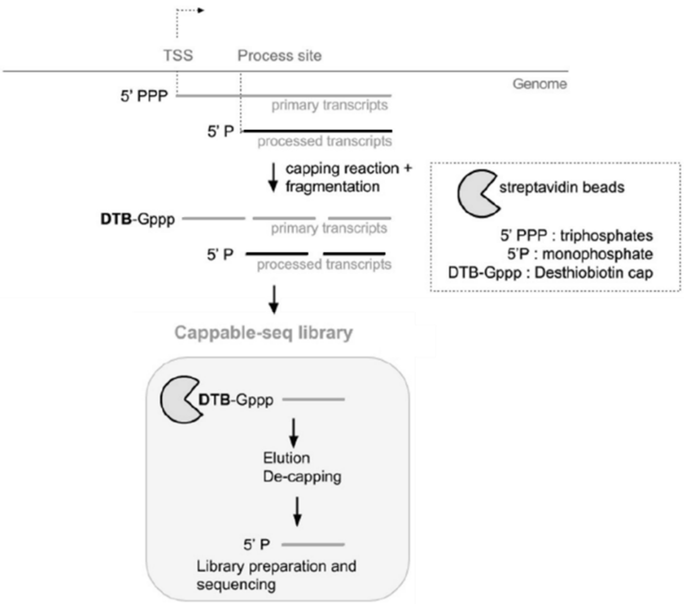


Figure 6. Scheme of the Cappable-seq pipeline for transcription start site (TSS) identification. After capping of 5´ triphosphorylated primary transcripts with 3´-desthiobiotin-TEG-guanosine 5´triphosphate (DTB-Gppp), RNA fragments are enriched by interaction with streptavidin beads. Processed transcripts with monophosphorylated 5´ ends are depleted. After elution and de-capping of the enriched primary transcripts, library preparation and sequencing are performed. Figure adapted from Ettwiller *et al.* (2016).

1.3.2.2 Conventional Ribo-seq

Another NGS-based method, which is often combined with RNA-seq, is ribosome profiling (Ribo-seq). This method allows the exploration of the global translational landscape by sequencing mRNA fragments, the so-called ribosome footprints (RFPs), covered by ribosomes (Ingolia *et al.*, 2009). Like RNA-seq, the Ribo-seq procedure starts with sample collection (**Figure 7**). During this process, stalling of the ribosomes is crucial to preserve their original *in vivo* position on the mRNA. There are several possibilities for immobilizing ribosomes. Most commonly translation-inhibiting drugs like harringtonine (e.g., Ingolia *et al.*, 2012), cycloheximide (e.g., Schneider-Poetsch *et al.*, 2010) or chloramphenicol (e.g., Oh *et al.*, 2011) are used to avoid ribosomal run-off. However, the use of drugs is often accompanied by ongoing translation, especially of resistant genes, altered translational speed or redistribution of ribosomes, leading to biased results (Sharma *et al.*, 2019, Hussmann *et al.*, 2015, Gerashchenko & Gladyshev, 2014). Thus, alternatively, cells can either be harvested via rapid filtration or by cooling with dry ice, followed by centrifugation and flash freezing using liquid nitrogen (Hücker *et al.*, 2017, Becker *et al.*, 2013). The next step in Ribo-seq comprises cryogenic cell homogenisation under conditions, which maintain the original ribosomal state. Cell lysis is followed by a nuclease footprinting assay. During this assay, RFPs, covered by ribosomes, and thus, physically protected from nuclease digestion, are generated. Several nucleases like ribonuclease I (RNase I, e.g., Neuhaus *et al.*, 2016), micrococcal nuclease (MNase, e.g., Mohammad *et al.*, 2019) or a mixture of different nucleases (Hücker *et al.*, 2017) have been used in bacteria in order to obtain precise RFPs. The choice of nuclease, therefore, plays a pivotal role for ensuring data quality. Enzyme concentration as well as incubation time have to be optimized to digest all exposed mRNA as accurately as possible without affecting ribosomal integrity (Gerashchenko & Gladyshev, 2017). A sucrose gradient density centrifugation is optionally used to purify the generated monosome complexes. Following extraction, RNA fragments are separated by denaturing polyacrylamide (PAA) gel electrophoresis, and RFPs are excised from the gel. The size to be excised depends on several factors, *inter alia*, on the nuclease used for footprinting as well as on the research issue to be addressed (Glaub *et al.*, 2020, Mohammad *et al.*, 2016). Once separated, RFPs are subjected to rRNA depletion, library preparation and sequencing as described for RNA-seq.

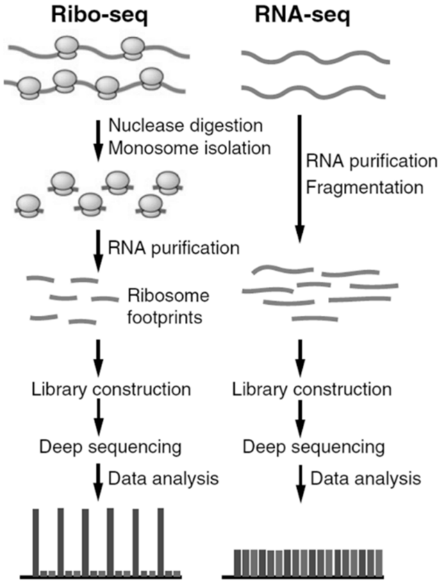


Figure 7. Overview of the experimental procedure of Ribo-seq and RNA-seq. Both methods are carried out on a split sample with stalled ribosomes. During the Ribo-seq workflow, monosomes are generated by nuclease footprinting, and then, isolated by gradient density centrifugation. Once purified, RNA is extracted and subjected to gel electrophoresis for size selection. Recovered RFPs were used for library construction and deep sequencing. After mapping, a three-nucleotide periodicity indicates genuine translation. For RNA-seq, total RNA is extracted, fragmented, and sequenced. As a result, obtained reads should be distributed evenly across all positions of transcribed regions. Figure adapted from Hsu *et al.* (2016).

Although initially applied to yeast (Ingolia *et al.*, 2009), Ribo-seq offers insights into translational processes of a variety of different organisms including eukaryotes (e.g., Ingolia *et al.*, 2011), prokaryotes (e.g., Miranda-CasoLuengo *et al.*, 2016) and viruses (e.g., Stern-Ginossar & Ingolia, 2015). In these organisms, Ribo-seq was used for quantification of mRNA abundance, recording of translational dynamics, deciphering of control mechanisms and compartmentalization of gene expression, and the identification of novel, translated genes (for excellent review see Brar & Weissman, 2015). By applying this method, several novel genes of small length were identified, for instance, in *Salmonella enterica* Typhimurium 14028s (Baek *et al.*, 2017) and *E. coli* O157:H7 (Hücker *et al.*, 2017, Neuhaus *et al.*, 2016).

1.3.2.3 Modified Ribo-seq variants

Recently, several modifications of the classical Ribo-seq protocol have been developed in order to optimize the resolution and power of such experiments, e.g., the implementation of nucleotide-precise footprints to determine the frame of translated ORFs. In eukaryotes, a three nt periodicity signal, indicating the translational frame, was obtained since the early beginnings employing Ribo-seq, mainly due to the use of the specific nuclease, RNase I (Ingolia *et al.*, 2009). Although RNase I was used successfully for bacterial Ribo-seq in the past (e.g., Neuhaus *et al.*, 2016), some researchers claimed RNase I activity to be impaired in *E. coli* (e.g., Kitahara & Miyazaki, 2011). Consequently, MNase was commonly used as an alternative for nuclease footprinting. A major drawback of MNase, though, is its partially sequence-specific cleavage pattern (Dingwall *et al.*, 1981), which leads to a sequence bias at both ends of the RFP, and thus, rules out a precise reading frame determination. In 2017, Hwang and Buskirk proposed the use of the endogenous type II toxin RelE to improve nucleotide resolution in bacteria. In an *in vitro* Ribo-seq experiment with *E. coli* MG1655, a combination of MNase and RelE enabled the detection of a highly resolved reading frame in the sum signal of all annotated genes as well as for single genes. This result was achieved by exploiting RelE´s unique feature of cleaving translating mRNA within the ribosomal A site codon (**Figure 8**, Pedersen *et al.*, 2003). Under physiological conditions, cleavage by RelE induces a translational arrest and inhibits cell growth, which, in turn, decreases energy consumption and enhances cell survival (Christensen *et al.*, 2001). Therefore, RelE is directly involved in several stress-response pathways, e.g., the stringent response (Christensen *et al.*, 2001). The toxic endonuclease RelE as well as its cognate antitoxin RelB are both encoded in the *relBE* operon (Bech *et al.*, 1985). This operon is autoregulated by RelB, which represses the transcription of both genes during steady-state cell growth (Overgaard *et al.*, 2009). Under non-stress conditions, the antitoxin RelB is present in ~10-fold excess over the toxin RelE, leading to an efficient inactivation of the latter by direct protein-protein interaction (Overgaard *et al.*, 2009, Galvani *et al.*, 2001). However, under stress conditions, e.g., nutrient starvation, the transcription of the operon is strongly activated, and the antitoxin RelB is subsequently cleaved by the activated protease, Lon (Gerdes *et al.*, 2005, Christensen *et al.*, 2001). As a result, the global translation is arrested reversibly by RelEs cleavage between the second and third nucleotide of the codon located in the A site of the ribosome (Pedersen *et al.*, 2003), yielding precise RFP 3´ ends. In addition, RelE only cleaves in a ribosomal context (Pedersen *et al.*, 2003) and shows a modest sequence preference (Hwang & Buskirk, 2017, Hurley *et al.*, 2011). These singular properties of RelE are essential for obtaining exact positional information and for visualising the three nt periodicity, which reflects the codon-wise movement of ribosomes during translation (Wen *et al.*, 2008). Consequently, translated regions can be unambiguously identified by differentiating RFPs generated by translational processes from RFPs caused by mere ribosomal occupancy.

Other modified variants of Ribo-seq also take advantage of special inhibitors to facilitate translation initiation site (TIS) detection in prokaryotes, aiding the discovery of hitherto unknown genes. The accuracy of TIS identification strongly depends on the inhibitor's mode of action. Ideally, assembly and positioning of the 70S ribosomal complex at the start codon of an mRNA should be guaranteed, whereas elongation after initiation must be effectively inhibited. Nakahigashi *et al.* (2016), for instance, used tetracycline for global mapping of TIS in a Ribo-seq experiment. The suitability of tetracycline in determining TSS, however, is limited due to its potential to act as a mere elongation inhibitor, able to stall also elongating and not only initiating ribosomes. Therefore, the distinction between RFPs generated by initiating ribosomes at TISs and RFP caused by pausing ribosomes is challenging. As a consequence, identification of unannotated TIS is hampered (Nakahigashi *et al.*, 2016).

Retapamulin-assisted ribosome profiling (Ribo-RET) represents an improved method for TIS detection. Meydan *et al.* (2019) showed that this method is capable of identifying not only TISs of annotated genes, but also a variety of alternative, yet functional TISs in the *E. coli* genome. Retapamulin (RET), the inhibitor used in this study, is a semi-synthetic derivate of the pleuromutilin family and specifically traps initiating ribosomes at the start codon of a translated ORF. RET targets the peptidyl transferase centre, and thus, interferes with the placement of AAs in the P and A site of the ribosome, hence impeding peptide bond formation (Davidovich *et al.*, 2007, Schlünzen *et al.*, 2004). By displacing the aminoacyl moiety of the first tRNA and impairing the correct positioning of an elongator tRNA in the A site, cessation of the early elongation phase is initiated (Meydan *et al.*, 2019). Even brief treatment with RET was sufficient for potent accumulation of ribosomes at start codons and for identification of primary as well as alternative TIS. However, deletion of the *tolC* gene, encoding for the outer part of the AcrA^{MFP}-AcrB^{RND}-TolC^{OMF} multi-drug efflux system, was a prerequisite for RET to exert its full effect within *E. coli.* Ribo-seq with the proline-rich antimicrobial peptide Onc112 seems to be a promising alternative to Ribo-RET, since application of Onc112 also resulted in reliable mapping of primary and alternative TIS by stalling ribosome initiation complexes. By combining both methods, Weaver *et al.* (2019) discovered novel ORFs in *E. coli,* including small and overlapping ones. One of the most recent modifications of the Ribo-seq protocol is the application of the antimicrobial peptide apidaecin. By sequestering release factors necessary for termination of translation, apidaecin arrests ribosomes at the stop codon of the translated ORFs (Mangano *et al.*, 2020). This method offers another interesting option for further genome exploration.

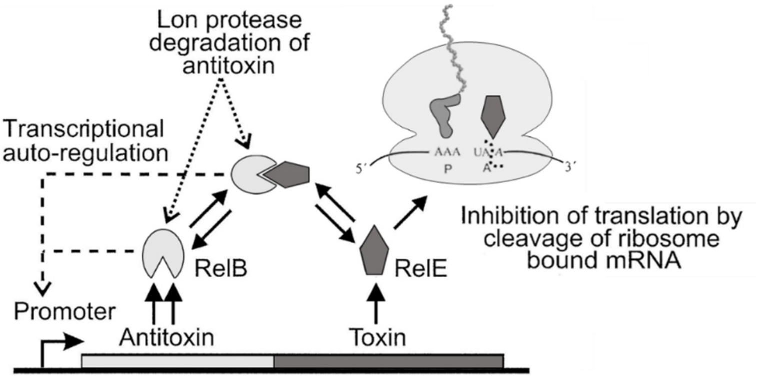

Lon protease degradation of antitoxin
Transcriptional auto-regulation
Promoter
RelB
RelE
Antitoxin
Toxin
Inhibition of translation by cleavage of ribosome bound mRNA
5′
3′
AAA
P
UA A
A

Figure 8. Regulation and organisation of the RelE-encoding operon. The toxin RelE and the respective antitoxin RelB are encoded by the *relBE* operon, which is autoregulated at the transcriptional level by the amount of free RelB as well as the presence of the non-toxic protein complex formed by RelB and RelE. RelB can be degraded by protease Lon, leading to an abrogation of neutralisation, and thus, to the activation of toxic RelE. Released RelE induces a translational arrest by cleaving mRNA between the second and the third codon which is located in the A site of the ribosome. Figure adapted from Pedersen *et al.* (2003).

1.3.2.4 Mass spectrometry

Mass spectrometry (MS), which is capable of detecting and quantifying proteins on a large scale, is considered as one of the most effective methods for proteome analysis. A typical MS experiment starts with the enzymatic digest of all proteins present in a sample by trypsin or any other suitable protease to generate a pool of peptides (Tsiatsiani & Heck, 2015). Optionally, additional pre-fractionation can be performed at this stage to increase sensitivity, and thus, also the detectability of low abundant proteins (Nissum *et al.*, 2007, Tang *et al.*, 2005). In the next step, complex peptide mixtures are resolved by liquid chromatography (LC), which separates peptides according to their intrinsic properties, e.g., charge, polarity, size or hydrophobicity (for review see Manadas *et al.*, 2010). Eluted peptides are converted into gas-phase ions using techniques like electrospray ionization (Fenn *et al.*, 1989) or matrix-assisted laser desorption/ionization (Karas & Hillenkamp, 1988). Ionized molecules are then separated according to their mass-to-charge ratio (m/z) by a mass analyser instrument. Most commonly, quadrupole, ion trap or time-of-flight mass analysers have been used, but also hybrid instruments, called tandem mass spectrometers, which combine multiple mass analysers in one system (for review see Pitt, 2009, Han *et al.*, 2008). For each precursor and product ion, a MS spectrum is generated, which is then used for peptide identification employing search engines such as Andromeda/MaxQuant (Cox *et al.*, 2011, Cox & Mann, 2008) or Mascot (Perkins *et al.*, 1999). For tandem mass spectrometry (MS/MS), multiple MS spectra are recorded, one for each mass analyser. In between, selected ions are additionally fragmented, e.g., by collision-induced dissociation (Wells & McLuckey, 2005).

Apart from the technical requirements, the type of data acquisition is also of central importance for a successful implementation of MS experiments. Three types of acquisition strategies can be distinguished: data-dependent acquisition (DDA), data-independent acquisition (DIA), and targeted data acquisition (TDA). In DDA, the most intense peptide precursors, obtained from the first mass spectrometer, are selected for further fragmentation and analysis by a second mass analyser. This process allows discovery-driven proteomics without the need for *a priori* knowledge about the proteins to be detected, but is accompanied by limited reproducibility and a bias against proteins with low abundance (Venable *et al.*, 2004). In contrast, DIA enables more accurate and reproducible quantification by sequential isolation and fragmentation of precursors in a flexible m/z window. Repetition of this process with shifted m/z windows leads to a full scanning over the whole m/z range, thereby achieving excellent protein coverage rates (Gillet *et al.*, 2012). The highest selectivity and sensitivity, however, is obtained by TDA, which is referred to as the gold standard for protein quantification (Peterson *et al.*, 2012). During the targeted approach of parallel reaction monitoring (PRM), all product ions are simultaneously monitored based on the m/z and retention time information prior obtained for the target peptides. In addition, PRM efficiency can further be enhanced by the use of heavy isotope-labelled standards derived from endogenous peptide sequences (for review see Rauniyar, 2015). Even though MS is under constant development, this method still reaches its limits when analysing low abundant proteins (Baldwin, 2004) and those lacking sufficient tryptic cleavage sites (Landry *et al.*, 2015, Slavoff *et al.*, 2013). Since most OLGs produce short proteins (e.g., Zehentner *et al.*, 2020a), the encoded proteins are difficult to predict and to detect directly using conventional mass spectrometry (Hemm *et al.*, 2020, Storz *et al.*, 2014). However, substantial progress has been made in recent years detecting small proteins in mass spectrometry (Petruschke *et al.*, 2020, Friedman *et al.*, 2017).

1.4 Perspectives of the study

The aim of this study was to identify and characterize novel overlapping gene candidates in the human pathogenic bacteria *E. coli* LF82 and *P. aeruginosa* PAO1 by multiple high-throughput approaches. In detail, the following methods, covering different levels of gene expression (transcriptome, translatome & proteome), were applied:

a)	transcriptome sequencing (RNA-seq)	transcriptome
b)	Cappable-seq	
c)	ribosome profiling (Ribo-seq)	translatome
d)	retapamulin-enhanced Ribo-seq (Ribo-RET)	
e)	RelE-supported Ribo-seq	
f)	mass spectrometry with data-dependent acquisition (DDA)	proteome
g)	mass spectrometry with targeted data acquisition (TDA)	

The results of the next generation sequencing-based methods a) to e) were combined for experimental delineation of translated ORFs. RNA-seq as well as Cappable-seq were applied for whole transcriptome description and for precise mapping of transcription start sites. Conventional Ribo-seq was used for the identification and quantification of translation events and for differentiation between coding and non-coding regions in both genomes. To expand the entire translatome analysis, the modified Ribo-seq variants Ribo-RET and RelE-supported Ribo-seq were performed. The use of the translational inhibitor, retapamulin enabled determination of genome-wide translation initiation sites. RelE-supported Ribo-seq was implemented to confirm genuine translation by visualizing the three nucleotide step-wise migration of the translating ribosome. All methods allowed global exploration of the genome, and thus, facilitated additional investigation of unannotated regions, e.g., novel intergenic genes. Promising novel gene candidates were further characterized bioinformatically. An organism-specific database covering all possible ORFs present in the genome of *P. aeruginosa* PAO1 was used for peptide discovery in a DDA tandem mass spectrometry experiment. Proteomics using TDA-based mass spectrometry validated and quantified some candidate peptides previously discovered by DDA. In addition, two exceptionally long OLGs in *Pseudomonas aeruginosa* PAO1 were analysed in more detail.

2. Materials & Methods

2.1 Materials

2.1.1 Antibiotics

All antibiotics used in this study are listed in **Table 1**.

Table 1. Antibiotics.

Antibiotic	Manufacturer	Working concentration [µg/mL]
ampicillin	Carl Roth	100-120
chloramphenicol	AppliChem	20
kanamycin	Carl Roth	10-30
retapamulin (RET)	Sigma-Aldrich	variable
spectinomycin	AppliChem	50
streptomycin	Sigma-Aldrich	30

2.1.2 Bacterial strains and plasmids

All bacterial strains and plasmids used in this thesis are listed in **Table 2** and **3**, respectively.

Table 2. Bacterial strains.

Bacterial strain	Relevant genotype/characteristics	Source
***E. coli* BL21**	F$^-$ *omp*T *gal dcm lon hsdS*$_B$($r_B^- m_B$) [*mal*B$^+$]$_{\text{K-12}}$(λ^S)	Studier (1991)
***E. coli* BL21(DE3) pLysS**	F$^-$ *omp*T *gal dcm lon hsd*S$_B$($r_B^- m_B$) λ(DE3 [*lac*I *lac*UV5-*T7p07 ind1 sam7 nin5*]) [*mal*B$^+$]$_{\text{K-12}}$(λ^S) pLysS[*T7p20 ori*$_{\text{p15A}}$] (Cm^R)	Studier (1991), Promega
***E. coli* CC118λpir**	A(*ara-leu*) *ara*D Δ*lac*X74 *gal*E *gal*K *pho*A20 thi1 *rps*E *rpo*B *arg*E(Am) *rec*Al, λpir	Manoil & Beckwith (1985)
***E. coli* LF82**	isolated from the ileal mucosa of a CD patient (Boudeau *et al.*, 1999)	Boudeau *et al.* (1999), obtained from Dr. Dirk Haller
E. coli* LF82Δ*tolC	TolC deletion strain of LF82	this work
***E. coli* SM10λpir**	*hi thr leu ton*A *lac*Y *sup*E *rec*A::RP4-2 Tc::Muλ*pirR6K*; Km^R	Simon *et al.* (1983)
***E. coli* Top10**	F- *mcrA* Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*araleu*)7697 *gal*U *gal*K *rps*L (Str^R) *end*A1 *nup*G	Invitrogen, Thermo Fisher Scientific
***P. aeruginosa* PAO1**	laboratory strain of the original Australian PAO strain isolated from a wound (Holloway, 1955)	German Collection of Microorganism and Cell Cultures (DSM No. 19880), 05/18.
***P. aeruginosa* PAO397**	Δ(*mexAB-oprM*) *nfxB* Δ(*mexCD-oprJ*) Δ(*mexJKL*) Δ(*mexXY*) Δ*opmH*362 Δ(*mexEF-oprN*)	Chuanchuen *et al.* (2005), obtained from Dr. Stephen Lory

Table 3. Plasmids.

Plasmid	Relevant characteristics	Source
pBAD/His C	*ori*pBR322, P$_{araB}$, Amp^{R}, 6×His:MCS	Invitrogen, Thermo Fisher Scientific
pET-22b(+)His$_6$:*relB*$_{\Delta9}$-*relE*$_{\mathrm{WT}}$	T7 promoter, *pelB*, *lacI*, 6×His:*relB*$_{\Delta9}$, *relE,* Amp^{R}	Dunican *et al.* (2015), Griffin *et al.* (2013), obtained by Dr. Scott Strobel
pKD4	*ori*R6Kγ, Amp^{R}, Neo^{R}/Kan^{R}, FRT	Datsenko & Wanner (2000), addgene
pKDsgRNA-p15	*ori*pSC101, P$_{araB}$, *exo*, *bet*, *gam,* Sm^{R}	Reisch & Prather (2015), addgene
pKNG101	*ori*R6K, *mob*RK2, *sac*BR, Sm^{R}	Kaniga *et al.* (1991)
pKNG101-TolC	pKNG101 with a *tolC*-deletion fragment in the multiple cloning site	this study
pMRS101	*ori*R6K, *ori*E1, *mob*RK2, *sac*BR, Sm^{R}, Amp^{R}	Sarker & Cornelis (1997)
pMRS101-TolC	pMRS101 with a *tolC*-deletion fragment in the multiple cloning site	this study

2.1.3 Commercial enzymes, ladders and kits

Table 4 lists all commercial enzymes, ladders and kits used in this study.

Table 4. Enzymes, kits and ladders.

Category/Product	Manufacturer
Enzymes	
Antarctic Phosphatase	New England Biolabs
Benzonase	Qiagen
Exonuclease T (RNase T)	New England Biolabs
Lysozym	Carl Roth
Micrococcal nuclease (MNase)	Thermo Fisher Scientific
Proteinase K	Analytik Jena
Q5 High-Fidelity DNA Polymerase	New England Biolabs
restriction enzymes (diverse)	Thermo Fisher Scientific
RNase A	Sigma-Aldrich
RNase R	Lucigen
SsoAdvanced Universal SYBR Green Supermix	Bio-Rad Laboratories
SUPERase · In RNase Inhibitor	Invitrogen, Thermo Fisher Scientific
SuperScript III Reverse Transcriptase	Invitrogen, Thermo Fisher Scientific
SuperScript II Reverse Transcriptase	Invitrogen, Thermo Fisher Scientific
T4 DNA Ligase	Thermo Fisher Scientific
T4 Polynucleotide Kinase	New England Biolabs
T4 RNA Ligase 2, truncated	New England Biolabs
Taq DNA Polymerase	New England Biolabs
TURBO DNase	Invitrogen, Thermo Fisher Scientific
XRN-1	New England Biolabs
DNA ladders	
100 bp DNA ladder	New England Biolabs
1 kb DNA ladder	New England Biolabs
1 kb Plus DNA ladder	New England Biolabs
RNA ladders	
random N$_{10}$	Biomers
random N$_{19}$	Biomers
random N$_{21}$	Biomers
random N$_{23}$	Biomers
random N$_{25}$	Biomers
random N$_{27}$	Biomers
random N$_{40}$	Biomers

Protein ladders	
Page Ruler Prestained Protein Ladder	Thermo Fisher Scientific
Spectra Multicolor Low Range Protein Ladder	Thermo Fisher Scientific
Kits	
Agilent High Sensitivity DNA Kit	Agilent Technologies
Agilent RNA 6000 Nano Kit	Agilent Technologies
GenElute Gel Extraction Kit	Sigma-Aldrich
GenElute PCR Clean-Up Kit	Sigma-Aldrich
GenElute Plasmid Miniprep Kit	Sigma-Aldrich
HiSeq Rapid SBS Kit v2 (50 cycles)	Illumina
miRNeasy Mini Kit	Qiagen
MiSeq Reagent Kit v3 (150-cycles)	Illumina
Mix2Seq Kit	Eurofins Genomics
QIAexpress Ni-NTA Fast Start Kit	Qiagen
Qubit dsDNA HS Assay Kit	Invitrogen, Thermo Fisher Scientific
Qubit RNA HS Assay Kit	Invitrogen, Thermo Fisher Scientific
riboPOOL Kit (*P. aeruginosa*)	siTOOLs Biotech
Ribo-Zero rRNA Depletion Kit (discontinued in 2018)	Illumina
TruSeq Small RNA Library Prep Kit	Illumina

2.1.4 Growth media and buffers

The composition of all growth media and buffers are listed in **Table 5** or in the subsequent chapters, respectively. All solutions were prepared with ultrapure water purified by a Milli-Q Elix Water Purification System (Merck). For RNA applications, ultrapure water was additionally treated with diethyl pyrocarbonate (DEPC) according to the supplier´s recommendation. If necessary, solutions were sterilized by filtration using a membrane filter with a pore size of 0.22 µm or by autoclaving at 121 °C for 15 min. For automated fast protein liquid chromatography (FPLC) applications, all solutions were degassed for 1 h at room temperature using a Sonorex Super RK 156 ultrasonic bath (Bandelin electronic).

Table 5. Growth media with their composition.

Medium	Component	Concentration	Remarks
Lysogeny broth (LB)	tryptone	10 g/L	adjusted to pH 7.4
	yeast extract	5 g/L	+ autoclaved
	NaCl	5 g/L	
	agar (optional)	16 g/L	
Schaedler broth	Schaedler bouillon	28.4 g/L	autoclaved
Super optimal broth with catabolite repression (SOC)	tryptone	20 g/L	autoclaved
	yeast extract	5 g/L	
	NaCl	0.5 g/L	
	KCl	0.186 g/L	
	$MgCl_2$	10 mM	
	$MgSO_4$	10 mM	
	glucose	20 mM	

2.1.5 Oligonucleotides

All primer used in this thesis are listed in **Table 6** and were purchased from Eurofins Genomics (Ebersberg) dissolved in H_2O in a final concentration of 50 µM.

Table 6. Primers used for PCR and cloning applications.

Intended use/name (organism)	Sequence (5’ → 3’)
16S *rRNA* gene primer (*E. coli*)	
rrsHF	AATGTTGGGTTAAGTCCCGC
rrsHR	GGAGGTGATCCAACCGCAGG

16S *rRNA* gene primer (*P. aeruginosa*)	
PA_16S_F	GATGTTGGGTTAAGTCCCGT
PA_16S_R	CCCCTACGGCTACCTTGTTA
vector primer	
pBAD-C+165F	CAGAAAAGTCCACATTGATT
pBAD-C+271F	TCTACTGTTTCTCCATACC
pBAD-C+494R	TGATTTAATCTGTATCAGGC
pET22b(+)+69F	GCTAGTTATTGCTCAGCGG
pET22b(+)+359R	TAATACGACTCACTATAGG
pKD4_KanR-389F	CTTGCCGCCAAGGATCTGAT
pKD4_KanR-1323R	GTGGAATCGAAATCTCGTGA
pKNG101+2358R	TCAGATCCTCTACGCCGGAC
pKNG101+976F	CTGGAGCGGATTTGCTCAAA
pMRS101+2136F	ACCTTTGTCTCGATCCTAGA
pMRS101+8640F	CGCAGGTATCGTATTAATTG
RelE expression (*E. coli*)	
relB+184R	ATTACGAAGCCGTTCTTTCAC
relE+3F-XhoI	AGTACTCGAGGGGCGTATTTTCTGGATTTT
relE+270R-HindI	ATTAAGCTTTCAGAGAATGCGTTTGACC
inactivation of *tolC* according to Datsenko & Wanner (*E. coli*)	
pKD4_KanR-31TolCF	AATTTTACAGTTTGATCGCGCTAAATACTGCTTCACCACAAGGAATGCAAGTGTAGGCTGGAGCTGCTTC
pKD4_KanR-389F	CTTGCCGCCAAGGATCTGAT
pKD4_KanR+935R	GTCCAGATCATCCTGATCGACAAGA
pKD4_KanR-1507TolCR	CAGACGGGGCCGAAGCCCCGTCGTCGTCATCAGTTACGGAAAGGGTTATGCATATGAATATCCTCCTTA
pKD4_KanR-1526TolCR	ATCTTTACGTTGCCTTACGTTCAGACGGGGCCGAAGCCCCGTCGTCGTCAATGGGAATTAGCCATGGTCC
pKDsgRNA-3182R	AATACCCAGCCTCGCTTTGT
pKDsgRNA_Redbeta+678F	GATATTTCGCCGCGACATTG
TolC-50F	AATTTTACAGTTTGATCGCGCTAAA
TolC-50R	ATCTTTACGTTGCCTTACGTTCAGA
inactivation of *tolC* using conjugation (*E. coli*)	
pMRS101+458R	CTTATCGATGATAAGCTGTC
pMRS101+8708F	GACACTGAATACGGGGCAAC
TolC-17R_US-EcoRI	TAACTGAATTCCATTTGCATTCCTTGTGGTG
TolC+37F	CTCTCTGGGTTCAGTTCGTT
TolC-133R	ACCAGTGGTAAATACCCATCAGAAT
TolC-285F	GCCGCGATAAAGTGTTTCTC
TolC-879R_DS-XbaI	AGTCTAGAGTGTGAAAATTGAAACCGTA
TolC-900F_US-BamHI	GCTTAGGATCCCAATTGTGAAAAGTTCATCT
TolC+1391R	GCTATCAGGCGCATAACCAT
TolC+1462_DS-EcoRI	AATCTGAATTCCATAACCCTTTCCGTAACTG
full-length mRNA primer for *olg1* & *olg2* (*P. aeruginosa)*	
OLG1+34F	CTCGTAGGGAGTTTCCGCGCG
OLG2+20F	TGACCAATACGCGCATCTCG
PA0260+800F	ACGGCCTGTTCGAACCCCTC
PA1383+278F	CCTACACCATCGATCCAGTG
qPCR primer for *gyrA*, *olg1* & *olg2* (*P. aeruginosa)*	
OLG1+524F	TCGCCATTGCGCTTGCGTAC
OLG1+640R	AGACGGTGGGACTTGCCAAC
PA_gyrA+346F	AACGCCGCAGCCATGCGATA
PA_gyrA+458R	CATGACCGCCGGGATCTGCT
RT primer for *olg1* & *olg2* (*P. aeruginosa)*	
OLG1+730R_RT	GCTGCCAGACGACCATCGAC
OLG2+1083R_RT	CGTCGGTGGCGATAACACCG
PA0260+575F	GCACCATGACCCATCCGCTGT
PA0260+423F	CGAATGGCTGGACCGCAACG
PA1383+172F	GTGGAAAATGGTGCCAACCT

2.1.6 Peptides

All peptides used for targeted proteomics are listed in **Table 7**. Peptides were synthesized in crude purity by JPT Peptide Technologies (Berlin) as freeze-dried, isotopically labelled peptides that terminate with a heavy arginine (U-13C6; U-15N4) or heavy lysine (U-13C6; U-15N2).

Table 7. Isotopically labelled peptides used for targeted proteomics.

Name	**Sequence (N-Terminus → C-Terminus)**
OLG1_1	AGDQNHAGAQFTSGK
OLG1_2	LAALATHPAGAAYR
OLG1_3	AVALAVAQR
OLG1_4	MASAADELEDTFER
Tle3_1	TIVSAQSITLPK
Tle3_2	FASGAGGAAVR
Tle3_3	TGLPQGFHQR
Tle3_4	AETPYEAR
Tle3_5	NSQLIDATVAYR
OLG2_1	ISPHLDTYX
OLG2_2	IPELGGX
OLG2_3	LAEVEEHHVAAGX
OLG2_4	LVEAQLB
PA1383_1	VSLQNDPLNELX
PA1383_2	VLHPLNLNNQDX
PA1383_3	FDVGDLB
PA1383_4	VDFENVX
PA1383_5	AHETGVYTVTEVAB

2.1.7 Reagents and chemicals

All reagents and chemicals used in this study are listed in **Supplementary Table 1**.

2.2 Methods

2.2.1 Microbiological methods

2.2.1.1 Cultivation and storage of bacteria

Unless otherwise stated, all *E. coli* and *P. aeruginosa* strains were cultivated aerobically in LB while shaking at 150 rpm and 37 °C. For Ribo-seq experiments, *E. coli* LF82 was routinely cultivated in glass flasks filled with Schaedler broth (1/5 of the initial flask volume) under aerobic as well as anaerobic conditions. *P. aeruginosa* was exclusively cultivated in chicane flasks ensuring an optimal oxygen supply. Main cultures were inoculated with overnight (ON) cultures in a ratio of 1:100 (*P. aeruginosa*) or 1:1,000 (*E. coli*) respectively. For ON cultures, 5 mL LB or Schaedler broth was inoculated with 50 µL of the respective glycerol stock and incubated for 24 h. An additional ON culture was prepared for *P. aeruginosa*, which was inoculated with 100 µL of the first ON culture (in 10 mL LB) and again incubated for 24 h. Cultivation on solid media was performed on LB agar plates supplemented with an antibiotic, if necessary. Agar plates with bacterial colonies were stored at 4 °C for up to six weeks. For long-term deposit, glycerol stocks were prepared by mixing the cell suspension with the same volume of 80% glycerol and stored at -80 °C.

2.2.1.2 Bacterial counting

Bacterial numbers were either determined by plate counting or by spectrophotometry. For plate counting, ten-fold serial dilutions were prepared by using sterile phosphate-buffered saline (PBS; 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, pH 7.4) or LB and plated on solid agar plates (100 µL, each) in duplicates. After incubation at 37 °C for at least two days, plates with bacterial counts between 3 and 300 were used for the final calculation of colony-forming units (CFU) per mL according to the following equation:

$$\mathrm{CFU/mL} = \frac{\sum \mathrm{n}}{1\mathrm{fa}+0.1\mathrm{fb}} \times \mathrm{DF} \times 10, \qquad \textbf{(Equation 1)}$$

where:
n = number of plates
fa = number of plates of the lowest dilution stage
fb = number of plates of the highest dilution stage
DF = dilution factor

Alternatively, total cell numbers were determined by measuring the optical density at a wavelength of 600 nm (OD_{600nm}) using a spectrophotometer (LAMBDA Bio, Perkin Elmer). The respective cultivation media was used as a reference and to dilute the initial cell suspension, if necessary.

2.2.1.3 Determination of minimum inhibitory concentration (MIC)

The MIC of retapamulin (RET) was determined for the wild type strains *P. aeruginosa* PAO1 and *E. coli* LF82 as well as their deletion mutants *P. aeruginosa* PAO397 and *E. coli* LF82Δ*tolC* via the broth microdilution method. For this purpose, RET was dissolved in dimethyl sulfoxide (DMSO) and diluted with the respective growth media to obtain RET stock concentrations of 64-0.25 µg/mL in two-fold steps. Bacteria were grown to a cell density of 1×10^8 CFU/mL as described previously and diluted to 1×10^6 CFU/mL. One hundred microliters of the cell culture were dispensed into each well of a 100-well plate and 100 µL of the different RET stock solutions were added to the wells yielding final inoculum sizes of approximately 5×10^5 CFU/mL and RET working concentrations of 32-0.125 µg/mL. Plain growth media without cell inoculum was used as a negative control (NC). Inoculated growth media (including 0.128% DMSO) without RET was prepared as a positive control. The plate was incubated for 24 h at 37 °C while shaking, and the optical density of each well was measured in an interval of 20 min using a Bioscreen C reader (Oy Growth Curves Ab Ltd). The MIC was subsequently determined as the lowest concentration of RET which inhibits visual growth of bacteria after the incubation period. This experiment was conducted in biological triplicates with three technical replicates each.

2.2.2 Molecular biological methods

All steps described in this chapter were carried out at room temperature unless otherwise stated. DNA and RNA samples were prepared with ultrapure or nuclease-free water and stored at -20 °C or -80 °C, respectively.

2.2.2.1 Isolation and purification of nucleic acids

2.2.2.1.1 Genomic DNA extraction

Five milliliters of an ON culture were centrifuged for 10 min at 5,000 ×g. Pelleted cells were dissolved in 567 µL Tris (10 mM)/EDTA(1 mM, pH 8) solution and mechanically lysed by two cycles of bead-beating at 6.5 m/s for 45 sec with 0.1 mm zirconia beads using a FastPrep-24 5G instrument (MP Biomedicals). To avoid degradation, samples were incubated on ice for 5 min after each round of bead-beating. Following centrifugation (1 min, 13,000 ×g), the supernatant was incubated with 30 µL SDS (10%) and 3 µL Proteinase K for 3 h at 37 °C. After incubation with 100 µL NaCl (5 M) and 80 µL CTAB/NaCl (10% CTAB in 700 mM NaCl) for 30 min at 65 °C, the same volume of RotiPhenol was added, and the sample was centrifuged for 5 min at 15,000 ×g. The aqueous phase was mixed with the same volume of RotiPhenol/chloroform/isoamylalcohol and centrifuged for 5 min at 15,000 ×g. This step was carried out twice, before DNA was precipitated with 0.6 volume fraction of cooled isopropyl at 4 °C for at least 20 min. Precipitated DNA was sedimented by centrifugation (5 min, 15,000 ×g, 4 °C), and washed twice by the addition of 1 mL EtOH (70%) and subsequent centrifugation (5 min, 15,000 ×g, 4 °C). After removal of EtOH, the pellet was air-dried for 20 min and dissolved in 100 µL H_2O ON at 4 °C. DNA concentration was measured by NanoDrop and RNase digestion was performed as described in the following chapter.

2.2.2.1.2 RNase digestion

RNase digestion with RNase A was performed to remove residual RNA from genomic DNA samples. For this purpose, 100 µg of DNA were incubated with 2 µL RNase A at 37 °C for 30 min. After adding 300 µL H_2O, the sample was mixed with the same volume of RotiPhenol/chloroform/isoamylalcohol. Following centrifugation (5 min, 15,000 ×g), the upper phase was supplemented with 1 mL cold EtOH (100%) and 0.1 volume fractions of sodium acetate and incubated at -20 °C for at least 30 min. Precipitated DNA was sedimented by centrifugation (5 min, 15,000 ×g, 4 °C), washed, dried and dissolved as described previously. Quality of the extracted DNA was assessed via gel electrophoresis.

2.2.2.1.3 RNA extraction

Total RNA was extracted either from homogenized cell lysate (RNA-seq, Ribo-seq & Cappable-seq) or from pelleted cells (any other application) using TRIzol Reagent. For NGS-based applications, total RNA was incubated in 1 mL cooled TRIzol for 5 min. For other applications (e.g., qPCR), cell pellets were resuspended in 1 mL cooled TRIzol and mechanically lysed by bead-beating as described for genomic DNA extraction using three cycles of beating instead of two. After 5 min of incubation with 200 µL ice-cold chloroform, phases were separated by centrifugation (15 min, 12,000 ×g, 4 °C) and the aqueous phase was incubated for 30 min with 500 µL isopropyl and 1 µL glycogen to precipitate the RNA. Subsequently, RNA was pelleted by centrifugation (10 min, 12,000 ×g, 4 °C) and washed twice with 1 mL EtOH (70%) as described for DNA extraction. After removal of the supernatant, RNA was air-dried and dissolved in an appropriate volume of nuclease-free water. Concentration of the RNA was determined by NanoDrop and RNA integrity was verified by agarose gel electrophoresis or Bioanalyzer measurement prior to DNase digestion.

2.2.2.1.4 DNase digestion

Up to 10 µg of RNA were subjected to DNA removal by incubation with 2 U TURBO DNase in 1× TURBO DNase Buffer supplemented with 50 U SUPERase · In RNase Inhibitor for 1 h at 60 °C. After inactivation of the reaction with 15 mM EDTA for 10 min at 65 °C, RNA was precipitated ON at -20 °C using ethanol (70%), 3 M sodium acetate and glycogen (690, 27.6, and 1 µL, respectively). Afterward, the RNA was pelleted, washed, dried, and dissolved in nuclease-free water as described for RNA extraction. After quantity and quality assessment, entire DNA removal was verified by PCR with 16S *rRNA* gene primers using *Taq* DNA Polymerase (section 2.2.2.3.1). DNase digestion was repeated multiple times, if necessary.

2.2.2.1.5 Plasmid isolation

Three to five milliliters of bacterial culture were used for plasmid isolation using the GenElute Plasmid MiniPrep Kit according to the manufacturer’s instructions. Extracted plasmid DNA was eluted twice with in total 50 µL H_2O and subjected to NanoDrop measurement and agarose gel electrophoresis as described in the subsequent chapters.

2.2.2.1.6 Concentration of nucleic acids

Concentration of nucleic acid was performed by centrifugation in a SpeedVac Vacuum Concentrator 5301 (Eppendorf) using standard settings (room temperature or 30 °C) for the evaporation of aqueous solutions. Alternatively, the sample volume was reduced by ethanol precipitation, followed by washing, drying and resuspension in a suitable volume of H_2O as described previously. The latter method was also used when an additional purification was required.

2.2.2.2 Nucleic acid quantitation and quality control

2.2.2.2.1 Agarose gel electrophoresis

RNA, DNA and PCR products were subjected to agarose gel electrophoresis to check the quality of the nucleic acids or to evaluate the size of PCR amplicons. Agarose was fully dissolved in TAE buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA) by microwaving. The ratio of agarose/TAE was adjusted to the type and size of the nucleic acid to be separated yielding a final agarose concentration of 1-2%. After heating, the evaporation loss was replenished with H_2O and 15 µL of diluted RedSafe Nucleic Acid Staining Solution (1:20,000) were added. For sample preparation, 5 µL of PCR products and 400-800 ng of DNA or RNA were mixed with either 6× DNA Loading Dye (Thermo Fisher Scientific) or 2× RNA Loading Dye (New England Biolabs). Samples were loaded on the polymerized agarose gel alongside with 7 µL of an appropriate marker. Nucleic acids were separated in TAE for 30-45 min at 110 V and visualized using an UVsolo TS imaging system (Analytik Jena). For electrophoretic separation of RNA samples, all solutions were prepared with DEPC-treated water.

2.2.2.2.2 Capillary electrophoresis

Additional quality and sizing controls were performed by capillary gel electrophoresis on a 2100 Bioanalyzer system (Agilent Technologies) following the manufacturer's instructions. The High Sensitivity DNA Kit and the RNA 6000 Nano Kit were used for separation of DNA and RNA samples, respectively. The RNA integrity number (RIN) allowed to evaluate the integrity of an RNA sample, ranging from 1 (totally degraded) to 10 (intact; Schroeder *et al.*, 2006). Only samples with a RIN ≥8 were used for further applications.

2.2.2.2.3 Measurement of nucleic acid concentration

Nucleic acid concentration was determined by a NanoDrop UV-Vis spectrophotometer (Thermo Fisher Scientific). One microliter of the sample was loaded on the pedestal and measured using the same solution as that of the unknown sample as blank. The purity of nucleic acids was determined by the absorbance ratio of 260 nm/280 nm, allowing to differentiate between pure DNA (260/280 = ~1.8) and pure RNA (260/280 = ~2) as well as the 260 nm/230 nm ratio, indicting the presence of contaminates when being lower than ~2.

As an alternative, fluorescence-based quantitation using a Qubit 4.0 Fluorometer (Thermo Fisher Scientific) was performed. The dsDNA HS Assay Kit and the RNA HS Assay Kit were used according to the manufacturer's instructions for measuring DNA and RNA samples, respectively.

2.2.2.3 Modification and detection of nucleic acids

2.2.2.3.1 Polymerase chain reaction (PCR)

Polymerase chain reaction (PCR) was carried out using two different polymerases (*Taq* and Q5 High-Fidelity DNA Polymerase). The choice of polymerase depended on the desired application and further processing of the PCR amplicons. Q5 High-Fidelity DNA Polymerase was used standardly for the amplification of DNA sequences required for subsequent cloning experiments due to its robust performance and low error rate. *Taq* DNA Polymerase, in contrast, was used for applications without need for a high-fidelity sequence amplification, e.g., in colony-PCR or RT-PCR. All PCR reactions were carried out in a thermal cycler (MWG Biotech Inc Primus 96 Thermal Cycler, MWG Biotech) with the settings listed below. Based on the specification of the polymerase used, elongation time during PCR was calculated for each fragment size individually. A negative control with H_2O as template as well as an appropriate positive control was included for each PCR reaction. Amplicons were subjected to agarose gel electrophoresis and, if necessary, to subsequent purification using the GenElute PCR Clean-Up Kit or the GenElute Gel Extraction Kit according to manufacturer's instructions.

Conventional PCR (Q5)

Conventional PCR was conducted as described in **Tables 8** and **9** using Q5 High-Fidelity DNA Polymerase, enabling the amplification of 1 kb DNA in 30 sec. The appropriate annealing temperature during PCR was determined based on this equation:

$$T_A\ [°C] = T_{M,\ lowest}[°C] + 3°C, \qquad \textbf{(Equation 2)}$$

where:
T_A = annealing temperature
$T_{M,\ lowest}$ = lowest melting temperature of the primers used

Colony-PCR

Colony-PCR was conducted for the verification of insert integration after cloning and for the selection of correct transformants. Therefore, colonies were picked, resuspended in 15 µL sterile H_2O and used as template for amplification by *Taq* DNA Polymerase (1 kb DNA in 60 sec) as described in **Tables 8** and **9**. The appropriate annealing temperature during PCR was determined based on this equation:

$$T_A\ [°C] = \frac{T_{M1} + T_{M2}}{2}\ [°C] - 5°C, \qquad (\textbf{Equation 3})$$

where:

T_A = annealing temperature

T_{M1} = melting temperature of the forward primer

T_{M2} = melting temperature of the reverse primer

Table 8. Pipetting scheme for PCR with Q5 High-Fidelity or *Taq* DNA Polymerase.

Type of PCR/ reagent	**Volume [µL]**	**final concentration**
PCR using Q5 High-Fidelity DNA Polymerase		
Template DNA	variable	<1 µg
Forward primer [10 µM]	2.5	0.5 µM
Reverse primer [10 µM]	2.5	0.5 µM
5× Q5 reaction buffer	10	1×
dNTPs [10 mM each]	1	200 µM
Q5 DNA Polymerase	0.5	0.02 U/µL
H_2O	ad 50	-
PCR using *Taq* DNA Polymerase		
Template DNA	2	-
Forward primer [10 µM]	0.5	0.2 µM
Reverse primer [10 µM]	0.5	0.2 µM
10× ThermoPol buffer	2.5	1×
dNTPs [10 mM each]	0.5	200 µM
Taq DNA Polymerase	0.125	0.025 U/µL
H_2O	ad 25	-

Table 9. Thermal cycler settings for PCR with Q5 High-Fidelity or *Taq* DNA Polymerase.

Step	**Temperature [°C]**	**Time [sec]**	**Cycles**
PCR using Q5 High-Fidelity DNA Polymerase			
Initial denaturation	98	30	1
Denaturation	98	10	30
Annealing	variable (**Equation 2**)	30	
Elongation	72	1 kb/30 sec	
Final elongation	72	300	1
Hold	4	∞	-
PCR using *Taq* DNA Polymerase			
Initial denaturation	95	300	1
Denaturation	95	30	30
Annealing	variable (**Equation 3**)	60	
Elongation	68	1 kb/60 sec	
Final elongation	68	300	1
Hold	4	∞	-

16S-PCR

16S-PCR was performed to check for the presence of DNA in RNA samples using 16S *rRNA* gene primers. This type of PCR was carried out as described for colony-PCR with 1 µL of template RNA. For determination of transcription termination sites, 1 µL of reverse transcribed cDNA was used for amplification following the colony-PCR protocol.

2.2.2.3.2 Reverse transcription

cDNA synthesis was performed using SuperScript III Reverse Transcriptase according to the manufacturer´s instructions with 500 ng of DNA-free RNA, 20 U SUPERase · In RNase Inhibitor and 50 pmol of random nonamers (Sigma-Aldrich) or 10 pmol of gene-specific primers. For each sample, a “no RT” control without reverse transcriptase was included for subsequent quantitative PCR analysis.

In addition, reverse transcription (RT) was performed to identify transcription termination sites downstream of selected novel gene candidates. For this purpose, single-stranded cDNA was generated by using appropriate primers binding downstream of the stop codon. After synthesis, 1 µL cDNA were amplified according to the colony-PCR protocol with a respective forward primer located within the ORFs coding region. Binding of all primers was verified in an analogous PCR with genomic DNA as template.

2.2.2.3.3 Quantitative PCR

Quantitative PCR (qPCR) was used for relative quantification of the mRNA level of the selected OLG *olg1*. The housekeeping gene *gyrA* encoding for DNA gyrase subunit A served as internal control. Each reaction contained 10 µL SsoAdvanced Universal SYBR Green Supermix (Bio-Rad Laboratories), 1 µL of cDNA and 0.5 µM of each forward and reverse primer in a total rection volume of 20 µL. All primer used in qPCR were listed in **Table 6** and were confirmed to have an amplification efficiency of 95 to 105%. Reactions with water as template were included for each run and served as negative controls. PCR was conducted using a CFX96 Touch Real-Time PCR Detection System (Bio-Rad Laboratories) with the subsequent settings: initial denaturation at 95 °C for 30 sec, and 40 cycles of denaturation (95 °C, 15 sec) and annealing (60 °C, 30 sec). A melt curve analysis with increments of 0.5 °C in a range of 65 to 95 °C ensured reaction specificity. All qPCR experiments were conducted in biological triplicates with three technical replicates each. Resulting data was evaluated using the ΔΔCt method according to Livak & Schmittgen (2001). A two-tailed Welch two sample t-test was used for statistical analysis at a significance level of 5% (p-value ≤ 0.05).

2.2.2.3.4 Restriction digest

Vector DNA or amplified PCR fragments were digested using different restriction enzymes and suitable buffers. For the digestion of vector DNA, 1 µg template was incubated with 10 U of each restriction enzyme, 2 µL of buffer and H_2O in a final reaction volume of 20 µL for 1.5 h at 37 °C. Digestion of 0.5 µg PCR amplicons in a volume of 10 µL was performed analogously in an increased reaction volume of 32 µL. If necessary, heat-inactivation of the restriction enzymes was carried out as described by the manufacturer. After digestion, plasmids and PCR products were purified using the GenElute Gel Extraction Kit or GenElute PCR Clean-Up Kit according to manufacturer’s instructions.

2.2.2.3.5 Ligation

For ligation, 50 ng of linearized vector was mixed with insert DNA in an molar ration of 1:1 to 1:2 (vector:insert). Alternatively, two inserts were ligated in equimolar amounts. The mixture was incubated with 1 U T4 DNA Ligase, 2 µL 10× Ligase Buffer and H_2O in a final reaction volume of 20 µL for 1 h at 22 °C. The reaction was stopped by enzyme inactivation at 65 °C for 10 min. Self-circularization of vector DNA was carried out as described with 50 ng of linearized plasmid, 5 U T4 DNA Ligase and 5 µL 10× Ligase Buffer filled up to a volume of 50 µL with H_2O.

2.2.2.3.6 Sanger sequencing

Sanger sequencing using the Mix2Seq Kit was performed to validate the correct sequence of plasmids, PCR products and cloning constructs. For this purpose, 15 µL of nucleic acids were mixed with 2 µL of an appropriate primer and sent to Eurofins Genomics (Ebersberg) for sequencing.

2.2.2.4 Construction of genetically modified *E. coli*

2.2.2.4.1 Preparation of electrocompetent cells

Electrocompetent cells were prepared by cultivating bacteria in LB to exponential phase (OD_{600nm} = ~0.5) and subsequent incubation on ice for 10 min while inverting several times. For cell harvest, 1 mL cell culture were centrifuged at 6,000 rpm for 10 min at 4 °C, and sedimented cells were resuspended in 1 mL cold H_2O. After centrifugation and removal of the supernatant, cells were mixed with 0.5 mL cold H_2O and once more sedimented by centrifugation, followed by another round of resuspension in 0.5 mL cold glycerol (10%) and centrifugation. Finally, pelleted cells were either dissolved in 50 µL H_2O for immediate use or in 50 µL cold glycerol (10%) followed by flash freezing in liquid nitrogen for long-term storage at -80 °C.

2.2.2.4.2 Transformation

For transformation, template DNA was mixed with 50 µL of electrocompetent cells in a cooled electroporation cuvette (0.2 cm). Up to 100 ng vector DNA or up to 0.5 µg of PCR product (desalted a 0.025 µm Millipore membrane floating on H_2O for 15 min) served as template; a negative control with H_2O as template was included. An electric pulse of 2.5 kV was applied using a MicroPulser Electroporator (Bio-Rad Laboratories) before cells were recovered in 450 µL pre-warmed SOC broth and incubated for 75 min at an appropriate temperature (28 °C or 37 °C, depending on the application) while shaking at 150 rpm. After incubation, 100 µL of the electroporated cells were plated on agar plates supplemented with a suitable antibiotic. Plates were incubated ON either at 28 °C or 37 °C, and grown colonies were subjected to colony-PCR with appropriate primer pairs to evaluate transformation success.

2.2.2.4.3 Conjugation

ON cultures of the donor and the recipient strain were mixed in equal volume fractions (500 µL each). After centrifugation (10,000 rpm, 5 min), the supernatant was decanted, and the retarded cells were resuspended in 100 µL LB and plated on LB agar. Conjugation plates were incubated ON at 37 °C and grown cells were flushed off the plate using 1.8 mL LB media. After preparing 10-fold dilutions (up to 10^{-3}) with LB, 100 µL of the undiluted as well as each of the diluted solutions were plated on agar plates supplemented with appropriate antibiotics. The agar plates were incubated at 37 °C for a least 4 h followed by additional incubation at room temperature for three days. Colony-PCR was performed for selected colonies to confirm plasmid transfer into the recipient strain.

2.2.2.4.4 Gene inactivation according to Datsenko & Wanner (2000)

One-step gene inactivation according to Datsenko & Wanner (2000) was implemented to generate a *tolC* deletion mutant strain of *E. coli* LF82. For this purpose, the plasmid pKD4 was propagated in *E. coli* BL21 and subsequently isolated as described in section 2.2.2.1.5. The kanamycin resistance cassette flaked by FLP recognition target sites of pKD4 was amplified in a conventional PCR with Q5 Polymerase. The primer pairs used for this PCR contained at least 50 bp of sequences homologous to regions flanking chromosomal *tolC*. The amplified resistance cassette was purified using the GenElute Gel Extraction Kit and digested with DpnI for vector DNA removal. Electrocompetent *E. coli* LF82 cells were prepared (section 2.2.2.4.1) and transformed with pKDsgRNA-15, carrying a spectinomycin resistance cassette, an arabinose-inducible λ Red recombinase gene and a temperature-sensitive origin of replication (ori). For this purpose, the cells were transformed as described in section 2.2.2.4.2 and incubated on spectinomycin-supplemented LB plates at 28 °C to maintain the plasmid. Successfully transformed cells were cultivated in LB_{Spec50} to $OD_{600nm} = 0.1$, before expression of λ Red recombinase was induced by the addition of arabinose in a final concentration of 10 mM. Induced cells were harvested at $OD_{600nm} = 0.5$ and electroporated for the immediate transformation with the amplified and desalted kanamycin cassette. After transformation, incubation in SOC broth was either performed at 37 °C to remove the temperature-sensitive pKDsgRNA-15 plasmid or at 28 °C (control samples without induction) to maintain the plasmid. Hundred microliters of the samples were plated on kanamycin-supplemented LB agar plates and incubated ON at 37 °C. Grown colonies were re-streaked on LB plates supplemented either with spectinomycin or kanamycin to confirm curing of pKDsgRNA-15. Colony-PCRs and Sanger sequencing were performed to check for the correct insertion of the kanamycin cassette and the absence of the *tolC* locus. Elimination of the kanamycin cassette via thermal induction of a FLP recognition target sites recombinase encoded by the heat-curable helper plasmid pCP20 was not performed since the correct integration of the kanamycin cassette into the *tolC* locus failed multiple times.

2.2.2.4.5 Gene inactivation via conjugation of a suicide plasmid

Alternatively, the deletion mutant *E. coli* LF82Δ*tolC* was generated using the plasmid pMRS101, which harboured two genes conferring resistance against the antibiotics streptomycin and ampicillin, a NotI-flanked high-copy number ori, a pir-protein dependent ori, a multiple cloning site as well as the *sacAB* genes imparting sucrose sensitivity. In a first step, 900 nt of the up- and downstream region of the *tolC* gene were amplified using a conventional Q5-PCR with genomic DNA as template. The primers used introduced BamHI and EcoRI as well as EcoRI and XbaI cleavage sites to the upstream and downstream amplicons, respectively. Both PCR products were purified, digested with EcoRI and subsequently ligated. The ligated deletion fragment was amplified via PCR using Q5 Polymerase. The deletion fragment as well as pMRS101 were digested with XbaI and BamHI and ligated, yielding the vector pMRS101-TolC. Propagation of pMRS101-TolC was carried out by transformation into electrocompetent *E. coli* Top10, followed by cultivation and plasmid isolation. Correct insertion of the deletion fragment into the vector was confirmed by Sanger sequencing. Subsequently, pMRS101-TolC was digested with NotI to remove the NotI-flanked high-copy ori and the ampicillin resistance gene. Re-ligation of the residual vector backbone resulted in the suicide vector pKNG101-TolC carrying the deletion fragment, a streptomycin resistance gene, and the low copy number pir-dependent ori. Amplification of this vector was performed by transformation in *E. coli* CC118λpir, cultivation and plasmid isolation. The propagated suicide vector was transformed in *E. coli* SM10λpir and the success of the transformation was verified by colony-PCR. After conjugation of pKNG101-TolC to *E. coli* LF82, cells were plated on LB agar supplemented with streptomycin and ampicillin to verify successful transfer of the plasmid while inhibiting growth of ampicillin-sensitive

E. coli SM10λpir cells. Chromosomal integration mediated by the homologous regions in the multiple cloning site of pKNG101-TolC was confirmed by colony-PCR. A loop out was performed in order to remove the chromosomally integrated plasmid including the *sacAB* genes and the intrinsic *tolC* gene. For this purpose, a liquid culture of *E. coli* LF82 with the chromosomally integrated pKNG101-TolC backbone was cultivated until OD_{600nm} = 0.7, and 25, 50 and 100 µL of this culture were plated on sucrose-supplemented LB agar. The plates were incubated ON at 37 °C to select for cells lacking the *sacB* gene which codes for a levansucrase conferring cell toxicity in the presence of sucrose. Cells grown on LB agar plates supplemented with sucrose were tested for the absence of *tolC* using colony-PCR. In addition, colonies were re-streaked on LB plates containing streptomycin to counterselect for the presence of the plasmid backbone. Finally, the whole genomic locus was amplified by PCR using Q5 Polymerase. The PCR product was purified and subjected to Sanger sequencing to validate the genomic *tolC* deletion.

2.2.3 Next generation sequencing-based methods

The processing of RNA samples described in this chapter was preferably performed in an RNase-free environment at 4 °C. RNA was routinely dissolved in nuclease-free water and stored at -80 °C. All buffers were prepared with nuclease-free or DEPC-treated water.

2.2.3.1 Transcriptome sequencing (RNA-seq)

2.2.3.1.1 Inhibition of translation and cell lysis

Cells were cultivated to a variable OD_{600nm} (**Table 10**) and translation was inhibited by continual addition of ~80 g dry ice per 200 mL of initial cell culture while stirring. For the Ribo-RET experiments, 100× MIC of RET was added and incubated for 5 min while stirring before rapid cooling with dry ice was performed. A no drug (ND) control without retapamulin was included. After the cultures reached a final temperature of 4 °C, cells were sedimented by centrifugation in pre-chilled beakers for 5 min at 8,000 ×g and 4 °C. Pelleted cells were resuspended in 200-650 µL cold, sterile Polysome-Lysis-Buffer (100 mM NH_4Cl, 20 mM Tris-HCl pH 8, 10 mM $MgCl_2$, 5 mM $CaCl_2$, 0.4% Triton X-100, 0.1% NP-40), optionally supplemented with 3 mM GMP-PNP and 0.32 U/µL SUPERase · In RNase Inhibitor. The resuspension was dropped into liquid nitrogen and stored at -80 °C. Cell lysis was conducted mechanically using a rechargeable driller (TE-CD 21 Li, Einhell Germany) with a pre-cooled pestle drill bit in the presence of liquid nitrogen. The lysate was cleared by several rounds of centrifugation (5 min, 8,000 ×g, 4 °C) until the cell debris was completely removed. The clarified lysate was stored at -80 °C until use for RNA-seq, Ribo-seq and Cappable-seq.

Table 10. Detailed set-up of all Ribo-seq and RNA-seq experiments.

Organism	Condition	OD	Translational stalling	Footprinting nucleases	Digestion buffer	Reaction inactivation	Size selection	rRNA depletion	RNA-seq	Replicates
E. coli LF82	aerobic	2	dry ice	750 U MNase 250 U RNase I 50 U RNase R 12 U RNase T 5 U XRN-1	Buffer 4 1 mM $CaCl_2$	150 U SUPERase · In (10 min)	19-27 nt	Ribo-Zero	yes	1×
E. coli LF82	anaerobic	2	dry ice	750 U MNase 250 U RNase I 50 U RNase R 12 U RNase T 5 U XRN-1	Buffer 4 1 mM $CaCl_2$	150 U SUPERase · In (10 min)	19-27 nt	Ribo-Zero	yes	1×
E. coli LF82Δ*tolC*	aerobic	1.4	dry ice + RET	450 U MNase	2 mM $CaCl_2$	6 mM EGTA (10 min)	27-40 nt	-	-	-
E. coli LF82Δ*tolC*	aerobic	1.4	dry ice	450 U MNase	2 mM $CaCl_2$	6 mM EGTA (10 min)	27-40 nt	-	-	-
E. coli LF82	aerobic	1.4	dry ice	450 U MNase 1 nmol RelE	6 mM $CaCl_2$	6 mM EDTA (10 min)	10-40 nt	Ribo-Zero	yes	-
P. aeruginosa PAO1	aerobic	1	dry ice	250 U MNase 250 U RNase I 50 U RNase R 12 U RNase T 5 U XRN-1	Buffer 4 1 mM $CaCl_2$	150 U SUPERase · In 6 mM EDTA (5 min each)	19-27 nt	Ribo-Zero	yes	-
P. aeruginosa PAO1	aerobic	6	dry ice	250 U MNase 250 U RNase I 50 U RNase R 12 U RNase T 5 U XRN-1	Buffer 4 1 mM $CaCl_2$	150 U SUPERase · In 6 mM EDTA (5 min each)	19-27 nt	Ribo-Zero	yes	-
P. aeruginosa PAO1	aerobic	1	dry ice	62.5 U MNase 18.75 U RNase R 4.375 U RNase T 1.875 U XRN-1	Buffer 4 1 mM $CaCl_2$	50 U SUPERase · In 6 mM EGTA (5 min each)	19-27 nt	riboPOOL	yes	1×
P. aeruginosa PAO397	aerobic	1	dry ice + RET	450 U MNase	2 mM $CaCl_2$	6 mM EGTA (10 min)	27-40 nt	-	-	-
P. aeruginosa PAO397	aerobic	1	dry ice	450 U MNase	2 mM $CaCl_2$	6 mM EGTA (10 min)	27-40 nt	-	-	-
P. aeruginosa PAO1	aerobic	1	dry ice	450 U MNase 2.5 nmol RelE	6 mM $CaCl_2$	6 mM EGTA (10 min)	10-40 nt	-	-	-
P. aeruginosa PAO1	aerobic	1	dry ice	450 U MNase	6 mM $CaCl_2$	6 mM EGTA (10 min)	10-40 nt	-	-	-

2.2.3.1.2 Ribosomal RNA depletion

After RNA extraction and DNase digestion, DNase-free RNA was optionally subjected to rRNA removal using either the Ribo-Zero rRNA Depletion Kit or the *P. aeruginosa*-specific riboPOOL Kit (protocol v1-5, siTOOLs Biotech) according to manufacturer's instructions. After depletion, samples were purified by ethanol precipitation. In case of using the *P. aeruginosa*-specific riboPOOL Kit for depletion, another DNase digestion step followed by precipitation was performed to remove residual DNA-based probes.

2.2.3.1.3 Ultrasonic fragmentation

One microgram of RNA was adjusted to a final volume of 50 µL with H_2O, transferred to a microTUBE AFA Fiber Pre-Slit Snap-Cap (Covaris) and fragmented using a S220 Focused-ultrasonicator (Covaris) with the following settings: 175 W, 10% duty cycle, 200 cycles for 180 sec. If necessary, the volume of the fragmented RNA was reduced by ethanol precipitation.

2.2.3.1.4 RNA (de-)phosphorylation

Samples were dephosphoylated using 10-20 U Antarctic phosphatase, 1× Antarctic Phosphatase Reaction Buffer and 0.3 U/µL SUPERase · In RNase Inhibitor in variable reaction volumes for 30 min at 37 °C. Purification of the samples was performed using the miRNeasy Mini Kit according to manufacturer's instructions with the following deviation: Samples were mixed with Buffer RWT in a final volume of 600 µL and transferred to a mini spin column. From this step on, purification was conducted as described by the manufacturer. Elution of the RNA was carried out using 30 µL H_2O and repeated once with the flow through of the first elution round. Eluted samples were phosphorylated with 20 U T4 Polynucleotide Kinase, 1× T4 Ligase Buffer and 0.3 U/µL SUPERase · In RNase Inhibitor in a final reaction volume of 30 µL for 60 min at 37 °C. Afterwards, samples were purified using the miRNeasy Mini Kit as described previously.

2.2.3.1.5 Library preparation and normalization

After volume reduction to 5 µL via vacuum concentration, samples were prepared for sequencing using the TruSeq Small RNA Library Prep Kit (Illumina) according to the manufacturer's instructions. In short, adapters were ligated to the 5` and 3` ends of the footprints, cDNA synthesis was performed, and libraries were amplified using 11 PCR cycles. For size selection, samples were mixed with 6× DNA Loading Dye (Thermo Fisher Scientific) and electrophoretically separated in 1× TBE buffer (10× TBE: 0.9 M Tris, 0.9 M boric acid, 20 mM EDTA) for 90 min at 145 V using a 10% polyacrylamide gel (4.2 mL H_2O, 2.47 mL Rotiphorese 30 NF 29:1, 0.75 mL 10× TBE, 75 µL 10% APS & 4.5 µL TEMED). Following electrophoresis, the gel was incubated in 50 mL 1× TBE and 15 µL SYBR Gold for 15 min while shaking before DNA fragments in a range of 140 to 160 nt were excised. If necessary, the excised region was expanded depending on the expected fragment length. Excised gel fragments were chopped by centrifugation (20,000 ×g, 2 min) and incubated ON in 300 µL H_2O at 22 °C and 700 rpm. The gel debris was transferred to a 2.2 µm cellulose-acetate filter and centrifuged at 6,000 ×g for 1 min before the flow through was precipitated ON at -20 °C using 100% EtOH, 3 M sodium acetate and glycogen (975, 30, and 2 µL, respectively). After fragment recovery, DNA pellets were dried at room temperature and resuspended in 15 µL 10 mM Tris-HCl (pH 8.5). Sample concentration was measured by Qubit and fragment size was determined using capillary electrophoresis. All libraries were diluted to a final concentration to 2 nM according to the following equation:

$$\text{nM}= \frac{1\ \text{ng/µL}}{660\frac{\text{g}}{\text{mol}} \times \emptyset\ \text{DNA length [bp]}} \times 10^6, \qquad \textbf{(Equation 4)}$$

where:

$\emptyset$ DNA length [bp] = mean fragment length in bp obtained after Bioanalyzer measurement

If necessary, multiple libraries were mixed for sequencing.

2.2.3.1.6 Sequencing

Libraries were sequenced single-end either on a MiSeq System (Illumina) using a MiSeq Reagent Kit v3 (150-cycles) or on a HiSeq2500 System (Illumina) using a HiSeq Rapid SBS Kit v2 (50 cycles). Alternatively, sequencing was commissioned from the Laboratory for Functional Genome Analysis (LMU, Munich).

2.2.3.2 Ribosome profiling (Ribo-seq)

Ribosome profiling was conducted as described for transcriptome sequencing. The following, additional steps were carried out in order to generate and enrich RFPs for sequencing.

2.2.3.2.1 Nuclease digestion

Variable amounts of the clarified lysate were digested for 1 h at 25 °C and 850 rpm using one or a combination of the following nucleases: MNase, RNase I, RNase R, RNase T, XRN-I, and RelE. To ensure optimal activity of the nucleases, the lysate was supplemented with Buffer 4 (New England Biolabs), $CaCl_2$ and/or SUPERase · In RNase Inhibitor, if necessary. The reaction was stopped by further incubation with either 6 mM EGTA, 6 mM EDTA and/or SUPERase · In RNase Inhibitor for 5-10 min. The detailed conditions and concentrations used for the single experiments are listed in **Table 10**.

2.2.3.2.2 Sucrose gradient and monosome fractionation

Monosome-mRNA-complexes were separated using sucrose gradient density centrifugation. For this purpose, cold, sterile Polysome-Gradient-Buffer (100 mM NH_4Cl, 20 mM Tris-HCl pH 8, 10 mM $MgCl_2$, 2 mM DTT) and sucrose solution (50% sucrose in Polysome-Gradient-Buffer) were used for the preparation of nine different gradient solutions with sucrose concentrations ranging from 10% to 50%. After the addition of an appropriate amount of SYBR Gold (1:10 diluted in H_2O), 1.5 mL of each sucrose gradient solution starting with the highest concentration were transferred to an ultracentrifugation tube. The digested cell extract was loaded onto the gradient and centrifuged at 28,000 rpm and 4 °C for 3 h in a L7-65 Ultracentrifuge (Beckman). Subsequently, the layer containing the intact ribosome-mRNA-complexes was visualized using black light lamps. For harvesting, the ultracentrifugation tube was pierced with a hot, sharp needle, which led to a slow release of gradient solution. The collected solution was aliquoted into 200 µL samples and subjected to RNA extraction and DNase digestion.

2.2.3.2.3 Size selection and recovery of RNA

For size selection, DNase-free RNA as well as random marker oligonucleotides of variable size (**Table 4**, RNA ladders) were mixed with 2× Novex TBE-Urea Sample Buffer (Thermo Fisher), incubated at 80 °C for 2 min and loaded to a 16% denaturing polyacrylamide gel (2.6 mL Rotiphorese Sequencing gel diluent, 6.4 mL Rotiphorese Sequencing gel concentrate, 1 mL Rotiphorese Sequencing gel buffer concentrate, 10 µL APS & 5 µL TEMED). To optimize separation of the RNA fragments according to their size, a defined amount of RNA was loaded, e.g., 50-100 µg per gel. After gel electrophoresis in 1× TBE for 110 min at 200 V, gels were stained with SYBR Gold and visualized using a FAS Nano Gel Documentation System (Nippon Genetics) before a region of variable size was excised from the gel (for detailed information see **Table 10**). The excised gel fragments were homogenized via centrifugation at 13,000 rpm for 2 min and 400 µL gel extraction buffer (300 mM NaOAc pH 5.5, 1 mM EDTA, 0.1 U/µL SUPERase · In RNase Inhibitor) were added to the gel debris. After ON incubation, the solution was transferred to a 0.2 µm cellulose-acetate filter and centrifuged for 2 min at 10,000 ×g. The RNA was precipitated ON at -20 °C using 1 µL glycogen and 690 µL 100% ethanol. Samples were centrifuged at 12,000 ×g at 4 °C for 20 min and the RNA was resuspended in 15 µL H_2O. After NanoDrop measurement, RNA samples were processed as described for transcriptome sequencing.

2.2.3.3 Cappable-seq for transcription start site (TSS) identification

Cappable-seq was carried out by vertis Biotechnologie AG (Freising). For this purpose, *P. aeruginosa* was cultivated until OD_{600nm} = 1 in biological triplicates and subjected to RNA extraction and DNase digestion. After quality control, 5' triphosphorylated RNA was reversibly capped with 3'-desthiobiotin-TEG-guanosine 5' triphosphate using the Vaccinia Capping Enzyme. Capped RNA was purified and heat-fragmented followed by two rounds of streptavidin beads enrichment to capture biotinylated primary RNA transcripts. Enriched transcripts were poly(A)-tailed using Poly(A) Polymerase and dephosphorylated with Antarctic Phosphatase prior to elution. The desthiobiotin cap of the eluted RNA species was removed by RNA 5´ Pyrophosphohydrolase resulting in 5´ monophosphorylated RNA. The obtained RNA was ligated to adapters and converted into a cDNA using M-MLV Reverse Transcriptase and oligo(dT)-adapter primer. The cDNA was amplified using 15-16 cycles of PCR while introducing sequencing barcodes. After fragmentation, cDNAs with a size of 200-600 bp were specifically isolated using magnetic beads. Bead-bound fragments were blunted and ligated to a 3` Illumina sequencing adapter. The final amplification of the size-selected cDNA was carried out using seven PCR cycles before samples were pooled and sequenced on an Illumina NextSeq 500 system using a read length of 75 bp.

2.2.4 Protein chemical methods

2.2.4.1 Overexpression of *relE*

Two different approaches were tested in order to synthesize RelE. In a first attempt, the sequence of a Q5-PCR-amplified *relE* sequence originating from *E. coli* Top10 was cloned in the expression vector pBAD/His C, which harboured an N-terminal histidine (His)-tag within the multiple cloning site. After transformation in *E. coli* Top10, *relE* expression was induced at OD_{600nm} = 0.5 by the addition of different amounts of arabinose (0.00002 – 0.2%) and the cells were incubated for additional 4 h. This experiment was discontinued due to the toxic behaviour and small yield of RelE. Since the sole expression of *relE* failed, a co-expression of the toxin RelE together with the respective antitoxin RelB as described by Griffin *et al.* (2013) was performed. The plasmid which was used for this approach, pET-22b(+)His_6:$relB_{\Delta 9}$-$relE_{WT}$, was received from the group of Dr. Scott Strobel and harboured a mutant His_6-tagged *relB* sequence as well as the *relE* gene with an overlapping stop/start codon as found in the native system under the control of T7 RNA polymerase. An internal deletion of nine AAs within *relB* should disrupt the antitoxin´s interaction with the toxin, thus, enabling the selective purification of RelE after expression. The detailed experimental procedure of this approach is described in the following sections. If necessary, volumes were downscaled for pilot expression and purification experiments. All steps were carried out at room temperature unless otherwise stated.

2.2.4.1.1 Cultivation and induction of RelE synthesis

pET-22b(+)His$_6$:*relB*$_{\Delta 9}$-*relE*$_{WT}$ was transformed into competent *E. coli* BL21 (DE3) pLysS cells. For protein expression, 600 mL LB supplemented with ampicillin were inoculated with 5 mL ON culture of the transformed *E. coli* BL21 (DE3) pLysS cells and incubated at 37 °C while shaking at 150 rpm. At $OD_{600nm} = 0.8$, protein expression was induced by the addition of 1 mM IPTG. After additional 3 h of incubation, cells were centrifuged at 10,000 rpm and 4 °C for 20 min. The supernatant was decanted, and the cell pellet was stored at -80 °C until further use. Samples of variable volume were taken throughout the entire cultivation process for SDS-PAGE analysis (section 2.2.4.3.1).

2.2.4.1.2 Cell lysis

Cell pellets were thawed on ice for 30 min and resuspended in 40 mL cold, sterile lysis buffer (300 mM NaCl, 50 mM NaH_2PO_4, 10 mM imidazole, 5 mM 2-mercaptoethanol, 0.1 mg/mL lysozyme, pH 8) supplemented with 4 complete Mini, EDTA-free Protease Inhibitor Cocktail tablets (Roche). After incubation for 30 min on ice, cell lysis was conducted by sonification at 4 °C (25% pulse intensity, 8× 20 sec with 20 sec pauses between each cycle) using a Sonopuls HD 2200 ultrasonic homogenizer (Bandelin electronic). The lysate was supplemented with 600 U Benzonase and incubated for 10 min at 4 °C before it was clarified by multiple rounds of centrifugation (9,000 ×g or 15,000 ×g at 4 °C for 30 min, each).

2.2.4.2 Purification of RelE

Purification of RelE was performed according to Griffin *et al.* (2013) and Dunican *et al.* (2015) using immobilized metal ion affinity chromatography (IMAC). Interaction between the His-tagged RelB-RelE complex and a nickel (Ni)-charged nitrilotriacetic acid (NTA) affinity resin enabled immobilization of the complex under native conditions. Selective elution of RelE was carried out under denaturing conditions ensuring elimination of the conformation-dependent interaction between the resin-bound RelB and its interaction partner RelE. In total, three different purification methods were tested in order to obtain pure RelE fractions: Purification using the QIAexpress Ni-NTA Fast Start Kit (section 2.2.4.2.1) or an FPLC system (section 2.2.4.2.2), either with a pre-packed or a manually packed chromatography column. Samples for SDS-PAGE and mass spectrometry (2.2.4.4.1) were taken throughout the entire process to evaluate the purification success.

2.2.4.2.1 Kit-based affinity chromatography

The clarified cell lysate was purified using the QIAexpress Ni-NTA Fast Start Kit with the buffers provided. In short, cell lysates were resuspended in native lysis buffer and applied to a Fast Start column according to the manufacturer's instructions. The column was washed twice with native wash buffer before RelE was eluted twice using 4 mL of denaturing wash buffer. Finally, RelB was eluted by adding 2 mL of denaturing elution buffer.

2.2.4.2.2 FPLC-based affinity chromatography

For the use of a pre-packed chromatography column, cell pellets were lysed as described (section 2.2.4.1.2) but with a lower volume of lysis buffer (5 mL instead of 40 mL). The clarified lysate was sterile filtered and injected into a 5 mL sample loop of an ÄKTApurifier 10 FPLC system (GE Healthcare) operated with a HisTrap HP 1 mL column (GE Healthcare). The column was washed with at least 10 column volumes (CV) of cold lysis buffer and 20 CV of cold wash buffer (300 mM NaCl, 50 mM NaH_2PO_4, 35 mM imidazole, 5 mM 2-mercaptoethanol, 0.1 mg/mL lysozyme, pH 8) under a constant flow rate of 1 mL/min. RelE was selectively eluted by adding 15 CV denaturation buffer (9.8 M urea, 100 mM NaH_2PO_4, 10 mM Tris-HCl, 1 mM 2-mercaptoethanol, pH 8). Optionally, the antitoxin RelB was eluted by washing with elution buffer (500 mM imidazole, 300 mM NaCl, 50 mM NaH_2PO_4, pH 8).

Alternatively, 40 mL clarified cell lysate were sterile filtered and mixed with 8 mL of a nickel-charged nitrilotriacetic acid agarose resin (Machery-Nagel), which was equilibrated prior to use for 15 min at 4 °C in 10 CV plain wash buffer. The sample was incubated for 1 h at 4 °C while agitating and applied to an unpacked XK 16/20 column (Amersham Bioscience). After draining by gravity, the resin was washed thrice with 10 CV wash buffer before the column was packed at 4 °C. The column was connected to an ÄKTApurifier FPLC system, which was previously flushed with H_2O and cold wash buffer under a constant flow of 2 mL/min. The connected column was additionally washed with wash buffer (~10 CV) until the ultraviolet (UV) signal at 280 nm reached the baseline. Finally, RelE was eluted by washing with 12 CV pre-warmed denaturation buffer and RelB was eluted by washing with 4 CV elution buffer. Fractions of 0.8 mL were collected throughout the entire purification process.

2.2.4.2.3 Protein refolding by diffusion dialysis

Fractions containing eluted RelE were pooled (in total 4.5 mL), subjected to concentration measurement via Bradford assay (section 2.2.4.3.3) and diluted to a final concentration of ≤40 µg/mL using denaturation buffer. Subsequently, the diluted sample was transferred to a pre-equilibrated dialysis cassette (Thermo Fisher) with a molecular weight cut-off of 7 kDa and dialyzed into 4 L of dialysis buffer (8 M urea, 50 mM bicine, pH 8.4) for 2 h at room temperature. The last step was repeated twice with fresh buffer, once ON and once for additional 2 h. Afterwards, the dialysis cassette was changed and the RelE protein was dialyzed thrice into 10 L of refolding buffer (300 mM KCl, 70 mM NH_4Cl, 50 mM Tris-HCl pH 7.5, 7 mM $MgCl_2$, 1 mM dithiothreitol), each round for at least 8 h. The temperature during dialysis against refolding buffer was gradually reduced from room temperature to 4 °C in order to avoid precipitation of urea.

2.2.4.2.4 Protein concentration via centrifugation

Refolded protein was centrifuged in a swinging bucket rotor centrifuge 5810 R (Eppendorf) at 5,000 ×g and 4 °C for 10 min using an Amicon Ultra-15 Centrifugal Filter Device (Merck) with a molecular weight cut-off of 10 kDa. Multiple rounds of centrifugation were necessary for the concentration of the entire protein solution. In between, the retentate was mixed by pipetting in order to avoid concentration gradients and clogging of the filter membrane. As a precaution, the flow through was additionally subjected to centrifugation using Amicon Ultra-15 Centrifugal Filter Device (Merck) with a molecular weight cut-off of 3 kDa. The retentates of all centrifugation steps were pooled and stored on ice. Filter membranes were flushed with refolding buffer, incubated for 30 min at 4 °C and then merged with the protein retentate.

2.2.4.2.5 Protein storage

The entire solution obtained after centrifugation (~3 mL) was centrifuged at 5,000 ×g for 15 sec to remove protein aggregates. Afterwards, the protein solution was dialysed into 1 L storage buffer (70 mM NH_4Cl, 50 mM Tris-HCl pH 7.5, 30 mM KCl, 7 mM $MgCl_2$, 1 mM dithiothreitol, 20% glycerol) at 4 °C twice (ON & 4 h). After buffer exchange, aliquots of 10 µL were flash frozen in liquid nitrogen and stored at -80 °C.

2.2.4.3 Protein detection and quantitation

2.2.4.3.1 Electrophoretic separation via SDS-PAGE

Samples were taken at different time points during the expression, purification, or refolding procedure in order to evaluate the success of the individual steps. For this purpose, whole cell pellets or liquid protein samples were mixed with an appropriate amount of sample buffer and heated to 95 °C for 10 min for cell lysis and protein denaturation. Whole cell lysates were additionally centrifuged at 13,200 rpm for 10 min to sediment cell debris. Samples were stored at -20 °C or directly used for SDS-PAGE analysis following the protocol of Laemmli (1970; Glycine–SDS-PAGE) or Schägger (2006; Tricine–SDS-PAGE).

For SDS-PAGE analysis according to Laemmli, samples were mixed with 2× Laemmli sample buffer (62.5 mM Tris-HCl pH 6.8, 25% glycerol, 5% 2-mercaptoethanol, 2% SDS, 0.01% bromphenolblue) and loaded on a discontinued SDS gel along with 5 µL of an appropriate size marker. The stacking gel had a total acrylamide monomer concentration (%T) of 6% and a crosslinker percentage (%C) of 2.7%; the resolving gel consisted of %T = 16% and %C = 2.7%. Electrophoresis was conducted in a chamber filled with running buffer (192 mM glycine, 25 mM Tris, 0.1% SDS) at an initial voltage of 60 V. After the migration front had reached the transition from stacking gel to resolving gel, the voltage was increased to 160 V. Electrophoretically-separated bands were visualized in staining solution (0.025% Coomassie Brilliant Blue G-250 in 10% acetic acid) for 2× 45 min and subsequently discoloured by washing twice with destain solution (10% acetic acid) before waving ON in H_2O.

To increase resolution of proteins smaller than 30 kDa, SDS-PAGE was alternatively performed according to Schägger using discontinuous gels with varying %T and %C percentages (stacking gel: %T = 4-5%, %C = 3,3%; resolving gel: %T = 16-18%, %C = 3.3-5%). Optionally, a spacer gel with %T = 10% and %C = 3.3% was casted between stacking and resolving gel. For sample preparation, varying volumes were mixed with 3× Schägger sample buffer (200 mM Tris-HCl pH 6.8, 40% glycerol, 2% SDS, 2% 2-mercaptoethanol, 0.04% Coomassie Brilliant Blue G-250) and heated before loading atop the gel. Gel electrophoresis was performed like described above with a few exceptions: Firstly, a constant electric current of 40 mA per gel was applied. Secondly, anode buffer and cathode buffer had slightly different compositions (anode buffer: 100 mM Tris, 22.5 mM HCl, pH 8.9; cathode buffer: 100 mM Tris, 100 mM tricine, 0.1% SDS, pH 8.25). Thirdly, the gels were incubated for at least 30 min in fixation solution (50% methanol, 10% acetic acid, 100 mM ammonium acetate) before they were stained as described.

2.2.4.3.2 Western blot and immunodetection

For Western blot analysis of expression as well as purification samples, SDS-PAGE was performed as described previously. A polyvinylidene difluoride membrane (Immobilon PSQ transfer membrane; Merck) and six filter papers were prepared. The membrane was incubated in 100% methanol for 10 sec and in H_2O for 5 min prior to blotting. Both the membrane as well as the gel were additionally equilibrated in blotting buffer (14.3 g/L glycine, 3 g/L Tris, 1 g/L SDS & 20% methanol) for 10 min before the transfer sandwich (3× filter paper, gel, membrane, 3× filter paper) was assembled. Semi dry blotting was carried out at 12 V for 45 min. After blotting, gels were waved 2× 10 min in fixation solution (0.2% glutaraldehyde, 0.1% Tween 20, in PBS), rinsed three times for 5 min in H_2O and incubated in quenching buffer (200 mM glycine, 0.1% Tween 20, in PBS) for 10 min. After washing (3× 5 min in H_2O), the membrane was incubated ON in TBS-T (150 mM NaCl, 20 mM Tris pH 8.0, 0.1% Tween 20) supplemented with 5% milk powder, rinsed 3× 10 min with TBS-T and covered with primary antibody solution (6× His-tag Monoclonal Antibody (Thermo Fisher), 1:1,000 diluted in TBS-T) for 1 h. After rinsing in TBS-T for 6× 5 min, the membrane was overlaid with alkaline phosphatase secondary antibody solution (goat anti-mouse (Dianova); 1:10,000 diluted in TBS-T) for 1 h. The membrane was rinsed in TBS-T six times for 5 min each prior to incubation with 500 µL of alkaline phosphate solution (100 mM Tris-HCl pH 9.5, 150 mM NaCl & 5 mM $MgCl_2$) for 2 min. For visualisation, 500 µl of the solution CDP-Star AP substrate were pipetted atop of the membrane. Emitted chemiluminescence was detected using an IVIS Lumina *in vivo* imaging system (PerkinElmer) with exposure times of 10 to 30 sec.

To evaluate successful transfer of the proteins, blotted membranes were stained using Ponceau S staining solution (0.5% Ponceau S in 10% acetic acid) for 5 min while agitating. Destaining was performed by washing with H_2O before the membrane was dried at room temperature.

2.2.4.3.3 Determination of protein concentration

The concentration of protein samples was determined either spectroscopically at a wavelength of 280 nm by NanoDrop or calorimetrically via Bradford assay. The latter was performed using Roti-Quant 5X-staining solution as described by the manufacturer for sample quantification in microtiter plates. Each sample was serially diluted and measured three times at a wavelength of 600 nm using a Wallac Victor3 multilabel reader (Perkin Elmer). Multiple dilutions of bovine serum albumin (BSA) were used for the generation of a calibration curve. The total protein concentration of the samples was calculated based on the absorption values measured for the calibration curve samples.

2.2.4.4 Analysis of proteins by mass spectrometry

2.2.4.4.1 Detection of RelE

Overexpression and purification samples were analysed regarding the presence and purity of RelE using mass spectrometry. Dr. Christina Ludwig (BayBioMS, TUM) or Dr. Per Haberkant (Proteomics Core Facility, EMBL Heidelberg) peformed data collection and analysis. Intensity Based Absolute Quantification (iBAQ; Schwanhäusser *et al.*, 2011) or top3 (Silva *et al.*, 2006) values were used for protein quantity estimations.

2.2.4.4.2 Data dependent acquisition (DDA) for novel gene identification

2.2.4.4.2.1 Sample collection and preparation

P. aeruginosa PAO1 was cultivated as described previously (section 2.2.1.1). Samples of 1 mL were taken at an OD_{600nm} = 1, sedimented by centrifugation at 12,000 ×g and 4 °C for 10 min and flash frozen in liquid nitrogen. All samples were stored at -80 °C prior to further preparation and analysis, which was carried out at the Bavarian Center for Biomolecular Mass Spectrometry (Freising).

Samples were prepared following the protocol by Doellinger *et al.* (2020). In short, cells were resuspended in 100 µL absolute trifluoroacetic acid and lysed for 5 min at 55 °C while shaking at 1,000 rpm. For sample neutralisation, 900 µL Tris (2 M) were added and protein concentration was determined using Bradford reagent according to manufacturer's instructions. Reduction and alkylation were performed by adding 10 mM Tris(2-carboxyethyl)phosphine and 55 mM chloroacetamide to 75 µg of each sample followed by incubation for 5 min at 95 °C. After dilution with the same volume of H_2O, proteins were digested with trypsin ON at 30 °C and 400 rpm at a protein/enzyme ratio of 50:1. The digestion reaction was stopped by adding 3% formic acid (FA).

2.2.4.4.2.2 Sample purification and offline HPLC-fractionation

StageTips (Rappsilber *et al.*, 2007) were prepared by placing 3 discs of Empore C18 material (3M) in a conventional 200 µL pipette tip. The conditioning of the StageTip was performed using 100% acetonitrile (ACN) followed by equilibration with 40% ACN/0.1% FA and 2% ACN/0.1% FA. Samples were loaded on the StageTip and desalted by washing with 2% ACN/0.1% FA before peptides were eluted with 40% ACN/0.1% FA. Offline peptide fractioning was performed using an XBridge BEH130 C18 3.5 µm 2.1 × 250 mm (Waters Corporation) reverse phase chromatography column operated by an 1100 series HPLC system (Agilent) at a constant flow rate of 200 µL/min. After sample loading, peptides were separated by applying a gradient from 4% to 32% buffer B (80% ACN) over 45 min followed by a gradient from 32% to 85% buffer B in 6 min. Fractions of 200 µL were collected every 30 sec throughout the entire elution process. If necessary, fractions were pooled to a maximum of 48 peptide fractions prior to MS measurement.

2.2.4.4.2.3 LC-MS/MS measurement

For MS analysis, 0.5 µg of peptides were loaded on a self-packed ReproSil-pur C18-AQ, 5 µm, 20 mm×75 µm (Dr. Maisch) trap column operated by an Ultimate 3000 RSLCnano system (Thermo Fisher Scientific) for 10 min using a flow rate of 5 µL/min and solvent A (0.1% FA and 5% DMSO in HPLC-grade H_2O). Peptides were separated with a self-packed ReproSil Gold C18-AQ, 3 µm, 450 mm×75 µm (Dr. Maisch) analytical column using a linear gradient from 4% to 32% of solvent B (0.1% FA and 5% DMSO in ACN) for 50 min at a flow rate of 300 nL/min. Eluted peptides were analysed using a downstream Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific) operated in DDA and positive ionization mode with the following parameters: For recording of MS1 spectra, the scan spectrum was 360–1300 m/z using a resolution of 60,000, an automatic gain control (AGC) target value of 4×10^5 and a maximum injection time (maxIT) of 50 msec. Fragmentation of up to 20 peptide precursors with charge states of 2+ to 6+ was performed using higher energy collision induced dissociation (HCD) with a normalized collision energy (NCE) of 30%. The dynamic exclusion duration was set to 20 sec and a precursor isolation window width of 1.3 m/z was selected. MS2 spectra were recorded at a resolution of 15,000 using an AGC target value of 5×10^4 and a maxIT of 22 msec.

2.2.4.4.3 Parallel reaction monitoring (PRM) for novel gene verification and quantitation

2.2.4.4.3.1 Sample collection and preparation

Cell harvest and further sample preparation were carried out as described for DDA with the following modifications: Samples of variable volumes (1 to 150 mL) were taken 1 h, 2 h, 4 h, 6 h, 8 h, and 24 h after inoculation as well as at OD_{600nm} = 1 (~160 min). If necessary, cell sedimentation by centrifugation was carried out multiple times with variable speed and time settings. Following cell lysis, 20 µg proteins per sample were reduced, alkylated and digested enzymatically using trypsin as described previously.

2.2.4.4.3.2 Sample purification and fractionation using StageTips

StageTips were packed, conditioned, and equilibrated as described for DDA-MS. After sample loading, 25 mM ammonium formate (pH 10) was added atop of the StageTips, and the flow through was collected (fraction 1). Further five-fold fractionation was achieved by washing the StageTips using 25 mM ammonium formate (pH 10) supplemented with variable ACN concentrations (5%, 10%, 15%, 25% and 50%).

Fraction 1 & 5, 2 & 6, and 3 & 4 were pooled, and the solvent was removed by centrifugation in a SpeedVac vacuum concentrator before dried peptides were dissolved in 2% ACN/0.1% FA.

2.2.4.4.3.3 PRM assay development

Peptides were separated as described previously using an Ultimate 3000 RSLCnano system (Thermo Fisher Scientific). PRM-MS analysis of eluted peptides was performed on a Q-Exactive HF-X mass spectrometer (Thermo Fisher Scientific) using the positive ionisation mode. MS1 spectra were measured at a resolution of 60,000 with an isolation window width of 360–1300 m/z, an AGC target value of 3×10^6, and a maxIT of 100 msec. Following HCD fragmentation with a NCE of 26%, MS2 spectra were acquired with a starting mass of 100 m/z and an isolation window size of 0.7 m/z using an orbitrap resolution of 60,000, an AGC target value of 1×10^6 and a maxIT of 118 ms. A selected set of 18 OLG and mother gene peptide precursors as well as 12 ProteomeTools Calibration Standard peptide precursors (Zolg *et al.*, 2017) purchased from JPT were targeted within one single run. The scheduled retention time window was set to 5 min, and a cycle time of ~2.1 sec was chosen in order to collect ~10 data points per peak for precise quantification.

Selection of the 18 target peptides was made based on the results obtained for the deep proteome DDA experiment (section 2.2.4.4.2, 48 fractions). The following criteria were used for the selection: peptide intensity, location within the coding region, charge state, modification type as well as the Andromeda score measured in the DDA experiment. For high confident peptide validation, selected peptides were synthesized as isotopically labelled internal reference peptides with either a heavy arginine (U-13C6; U-15N4) or heavy lysine (U-13C6; U-15N2) at the C-terminus. All peptides were pooled equally and spiked into the target sample for nano-flow PRM measurement yielding confident detection of all 18 peptides with MaxQuant scores >90. DDA experiment-derived spectral libraries as well as predicted spectral libraries obtained by the deep neural network Prosit (Gessulat *et al.*, 2019) were generated using the Skyline-daily (64-bit) software (v20.1.9.234 ; MacLean *et al.*, 2010).

2.2.5 Data processing and bioinformatic analyses

2.2.5.1 Evaluation of next generation sequencing (NGS) data

2.2.5.1.1 Transcriptome sequencing

Processing of FASTQ files was carried out using custom perl, bash and python scripts. In short, read quality was evaluated using FastQC (Andrews, 2010) with default settings, and identified adapter sequences were trimmed using fastp (Chen *et al.*, 2018). Clipped reads were mapped to the reference genome (*E. coli* LF82: GCF_000284495.1; *P. aeruginosa* PAO1: GCF_000006765.1_ASM676v1) using Bowtie2 v2.2.6 (Langmead & Salzberg, 2012) with a seed length of 17 nt in the --very-sensitive end-to-end mode. SAMTools (Li *et al.*, 2009) and BEDTools (Quinlan & Hall, 2010) were applied to remove reads mapping to rRNAs and tRNAs. Reads were visualized using the genome browser Artemis (Rutherford *et al.*, 2000) and normalized to sequencing depth by calculating reads per million mapped reads (RPM) according to the following equation:

$$\mathrm{RPM}= \frac{\mathrm{RC_{total}}}{10^6}, \qquad \textbf{(Equation 5)}$$

where:

$\mathrm{RC_{total}}$ = total number of mapping reads (tRNA & rRNA excluded)

To compare the expression of different genes, reads were also normalized to the gene length using the following equation for determining reads per kilobase per million mapped reads (RPKM):

$$\mathrm{RPKM}=\frac{\mathrm{RC_{gene}}}{\mathrm{RPM}\times \mathrm{L_{gene}}}, \qquad \textbf{(Equation 6)}$$

where:

$\mathrm{RC_{gene}}$ = number of reads mapping to a gene
RPM = reads per million mapped reads
$\mathrm{L_{gene}}$ = length of the gene in kb

Normalized reads were averaged over biological replicates and plotted using pyGenomeTracks (Ramírez *et al.*, 2018). Quantitative changes in expression levels were assessed by differential gene expression analysis using an exact test implemented in the Bioconductor package edgeR (Robinson *et al.*, 2010). Prior to statistical analysis, read counts were scaled to the smallest library size for normalization.

2.2.5.1.2 Conventional Ribo-seq

Translatome FASTQ files were processed and analysed as described for transcriptome sequencing data. Translated ORFs were predicted by using three different tools: REPARATION (Ndah *et al.*, 2017), DeepRibo (Clauwaert *et al.*, 2019) and the method according to Giess *et al.* (2017). The REPARATION algorithm is a *de novo* machine learning algorithm for delineation of bacterial ORFs based on experimental Ribo-seq data. REPARATION predicts ORFs as small as 10 AA based on i) start region RPKM; ii) stop region RPKM; iii) proportion of nucleotides within the ORF covered by RFPs; iv) proportion of nucleotides within the start region covered by RPFs; v) the ratio of the average RPF read count within the start region divided by the average RPF read count within the remaining ORF; and vi) ribosome binding site energy (Ndah *et al.*, 2017). DeepRibo relies on convolutional as well as recurrent neural networks to delineate protein-coding ORFs based on DNA sequence motifs and Ribo-seq information. The scripts by Giess *et al.* (2017) utilize specific RFP read length patterns around the TISs in combination with the sequence context for prediction. All programs were executed using default settings, and obtained results were combined using a custom perl script. All hits including redundant ORFs arising from multiple start codons were filtered for a minimum length of 93 nt and their annotation status was ascertained using the respective GFF file (*E. coli*: GCF_000284495.1_ASM28449v1_genomic.gff; *P. aeruginosa*: GCF_000006765.1_ASM676v1_genomic.gff). After removal of anORF predictions, the remaining ORFs were classified according to their type of overlap, and RPKM and coverage values were appraised. In addition, the translatability of the ORFs was determined by calculating the ribosome coverage value (RCV; Neuhaus *et al.*, 2017) according to the following equation:

$$\mathrm{RCV}=\frac{\mathrm{RPKM_{Ribo\text{-}seq}}}{\mathrm{RPKM_{RNA\text{-}seq}}}, \qquad \textbf{(Equation 7)}$$

where:

$\mathrm{RPKM_{Ribo\text{-}seq}}$ = reads per kilobase per million mapped reads of the translatome
$\mathrm{RPKM_{RNA\text{-}seq}}$ = reads per kilobase per million mapped reads of the transcriptome

2.2.5.1.3 RelE-supported Ribo-seq

Due to the shorter size of RelE-cleaved RFPs (Hwang & Buskirk, 2017), all RelE-supported Ribo-seq datasets as well as the respective control Ribo-seq and RNA-seq datasets were processed using adjusted parameters. Firstly, the respective adapter sequences were trimmed with fastp (Chen *et al.*, 2018) using a reduced minimal length requirement of -l 10 after clipping. Trimmed reads were aligned to the reference genome with Bowtie v1.1.2 (Langmead *et al.*, 2009) in the --best --strata mode reporting only unique mappers with a maximum of two seed mismatches (-m 1 --best --strata -n 2). All sequencing reads which mapped to rRNA or tRNA were removed using BEDTools (Quinlan & Hall, 2010) and SAMTools (Li *et al.*, 2009). The latter was also used for data quality and read length assessment.

The cleavage preference of RelE was determined based on the nucleobase composition before and after the cleavage site using custom python scripts. For this purpose, the type of nucleobase was determined in a range of -3 (within read) to +3 (outside read) and summarized for all RFPs. Reading frame analysis was performed as described by Hwang & Buskirk (2017). In short, the 3´ends of all RFPs mapping to a certain genome region were assigned to the first, second or third sub-codon position. If necessary, reads mapping to the first and the last 30 nt of each ORF were excluded. In addition, reads mapping to the third sub-codon position and ending with the nucleobase C were shifted from the third to the second sub-codon position in order to balance out RelE´s cleavage specificity.

2.2.5.1.4 Retapamulin-assisted Ribo-seq

Ribo-RET datasets as well as the respective no drug (ND) datasets were processed as described for conventional Ribo-seq experiments. For P site positioning, 15 nt or 17 nt were subtracted from the 3' end of the RFPs. Strand-specific RPM values were calculated according to **Equation 5** for every genome position. In addition, RPKM values were determined according to **Equation 6** for all ND datasets.

Genes with an RPM exceeding 100 in both the RET dataset as well as the ND dataset were used for metagene analysis according to Meydan *et al.* (2019). To avert inconclusive read assignment, same-strand encoded genes separated by fewer than 50 nt were excluded. For all genes complying with these criteria, normalized reads were calculated for each nucleotide position within the ORF as well as in a 30 nt region flanking the start and stop codon. For this purpose, the RPM value of each position was divided by the total PRM value obtained for the entire region. Afterwards, normalized values were averaged over an area ranging from 10 nt upstream to 50 nt downstream of the first start codon nucleotide.

The assignment of the TIS obtained in the RET dataset was performed using the python scripts provided by Meydan *et al.* (2019). The original algorithm quarries the entire genome for RET peaks with values >1 RPM for the identification of annotated genes or values >5 for the more stringent delineation of novel genes. Once a matching peak is found, the algorithm checks for the presence of one of the start codons AUG, GUG, CUG, UUG, AUU and AUC in a 3 nt wide region around the RET peak. To optimize translated ORF detection, the scripts by Meydan *et al.* (2019) were adapted to the generated datasets by adjusting the identification parameters, e.g., RPM values of the peaks to be detected, as described in the Results sections. All predicted ORFs exceeding the defined threshold value were merged with the predictions obtained by DeepRibo (Clauwaert *et al.*, 2019), REPARATION (Ndah *et al.*, 2017) and the scripts by Giess *et al.* (2017) and processed as described previously. In addition, the ND dataset was also analysed using the three prediction programs in order to add another level of gene identification regardless of the results obtained for the RET dataset.

2.2.5.1.5 Cappable-seq

Cappable-seq reads were processed as described for conventional RNA-seq and Ribo-seq. Resulting bam files were evaluated using the script provided by Ettwiller *et al.* (2016) for determination of TSS. In a first step, this script trims all reads to their leftmost 5´ends yielding 1-bp long reads. The trimmed reads were then used for the calculation of a strand-specific relative read score (RRS) for each position of the genome according to the following equation:

$$\mathrm{RRS} = \frac{\mathrm{n}}{\mathrm{N}} \times 10^6, \qquad \textbf{(Equation 8)}$$

where:
n = number of reads mapping to a certain genome position
N = number of reads mapping to the entire genome

Genome positions being covered with at least one read (RRS > 0) were used for reproducibility analysis by calculating pairwise Pearson´s product-moment correlation coefficients of three biological experiments. A suitable cut-off value for TSS prediction (RRS = 1.5) was determined by comparing fold changes of TSS counts obtained after application of different threshold values for each of the three experiments as described in section 3.2.5. The optimal search distance for TSS detection was ascertained by analysing 5`UTR lengths of annotated, protein-coding genes showing a TSS with an RRS ≥1.5 within a 500 bp upstream region in three biological replicates. A region of 200 bp upstream of the start codon of novel gene candidates was searched for the presence of a TSS with a RRS ≥1.5. Only TSS detected in at least two of the three replicates were considered to be of genuine origin. Sequence conservation of promoter regions was analysed in a 20 bp window upstream the identified TSS, and the results were visualized using the tool WebLogo (Crooks *et al.*, 2004).

2.2.5.2 Evaluation of mass spectroscopical data

2.2.5.2.1 DDA

Data analysis was performed using the search engine Andromeda (Cox *et al.*, 2011) integrated into the MaxQuant environment (v1.6.3.4; Tyanova *et al.*, 2016). MS2 spectra were either searched against all protein-coding genes listed in the RefSeq protein file (GCF_000006765.1_ASM676v1_protein, 5,572 reviewed entries, downloaded on 2020/12/07) supplemented with the AA sequences of the OLGs of interest, or against a six-frame translation of the genomic sequence of *P. aeruginosa* PAO1. The following settings were used for peptide identification and quantification: Common laboratory-originating contaminants were included into all databases, Trypsin/P was selected as digestion enzyme, and precursor and fragment ion tolerance were adjusted to 4.5 ppm and 20 ppm, respectively. Peptide spectrum match (PSM) and protein false discovery rates (FDR) were set to 1% using a target-decoy database comprised of reversed AA sequences. Seven AAs were selected as minimum peptide length, and the “match-between-run” function was turned off. Methionine oxidation as well as N-terminal acetylation were specified as variable modifications and carbamidomethylation of cysteine residues was defined as a fixed modification. The Skyline-daily (64-bit) software (v20.1.9.234; MacLean *et al.*, 2010) was used for the determination of dot products between experimental and predicted spectra. Finally, intensity and iBAQ (Schwanhäusser *et al.*, 2011) values were calculated for protein quantification.

2.2.5.2.2 PRM

The Skyline-daily (64-bit) software (v20.1.9.234; MacLean *et al.*, 2010) was used for PRM data analysis. Data quality was controlled by visual inspection and adjustment of peak integration, integration boundaries as well as transition interferences. Two to six transitions per peptide were chosen for reliable quantification of the target proteins, and peptides were filtered using a "Library Dot Product" (correclation between aquired product ion intensity and library reference spectrum intensity) value ≥0.8, a "DotProductLightToHeavy" (the ratio of the summarized light transition peak areas and the summarized heavy transition peak areas) value ≥0.9 and a "Average Mass Error PPM" of less than +/−20 ppm. For each sample, the intensities of all light peptides passing quality control filtering were added up. All peptides were checked for uniqueness against the RefSeq protein database.

2.2.5.3 *In silico* analyses of identified overlapping gene candidates

2.2.5.3.1 Genomic distribution

Phage regions within the genomes of *E. coli* LF82 and *P. aeruginosa* PAO1 were identified using PHASTER (Arndt *et al.*, 2016) and their location as well as the genomic location of the novel gene candidates was visualized using GView (Petkau *et al.*, 2010).

2.2.5.3.2 Promoter determination

The bacterial promoter prediction algorithm BPROM (Solovyev & Salamov, 2011) was used to identify putative σ^{70} promoters within a 300 nt region upstream of the respective start codon. A linear discriminant function (LDF) value ≥0.2 served as threshold for differentiation between promoter and non-promoter sequences based on functional motifs and oligonucleotide composition.

2.2.5.3.3 Terminator identification

Putative ρ-independent terminators within a 300 nt region downstream of the stop codon were predicted by FindTerm (Solovyev & Salamov, 2011) using a threshold ≤−3. To identify the hairpin structure necessary for termination, the predicted sequence was split into 30 nt fragments and subjected to secondary structure analysis using Mfold (Zuker, 2003)

2.2.5.3.4 Ribosome binding site identification

Detection of SD sequences was carried out within a 30 nt region upstream of the start codon according to Hyatt *et al.* (2010) or Ma *et al.* (2002). For the latter, a minimum free energy (ΔG_{SD}) threshold of ≤−2.9 kcal/mol was used. If necessary, a window of 20 bp around the start codon was used for SD sequence motif representation using WebLogo (Crooks *et al.*, 2004).

2.2.5.3.5 Detection of homologues

Homologous protein sequences were identified by the Basic Local Alignment Search Tool (blast; Altschul *et al.*, 1990) using default settings. AA sequence queries were compared to sequences included in the "Non-redundant protein sequences (nr)" and "RefSeq Select proteins (refseq_select)" databases, respectively. The e-value cut-off was set to $\leq 1\times10^{-3}$.

2.2.5.3.6 Gene prediction

Prodigal (Hyatt *et al.*, 2010) with standard settings was applied in order to identify protein-coding genes in the genome of *P. aeruginosa* PAO1. For prediction of novel OLGs, the annotated mother genes were masked by replacing every nucleotide of all possible start codons upstream and within the coding region by any nucleotide (N).

2.2.5.3.7 Evolutionary analyses

Phylostratigraphic, stop codon and sequence constraint analyses were conducted by Dr. Zachary Ardern as described in Kreitmeier *et al.* (2021). Briefly, gene homologues were identified using blastp, DIAMOND blastp (Buchfink *et al.*, 2015), and the Entrez Programming Utilities (Kans, 2013) accessing the Identical Protein Groups database. After downsampling, QuickProbs2 (Gudyś & Deorowicz, 2017) was used to align unique sequences followed by the conversion into the corresponding codon alignments using PAL2NAL (Suyama *et al.*, 2006). Aligned sequences were used for the reconstruction of phylogenetic trees by maximum likelihood using IQ-TREE (Nguyen *et al.*, 2015) with 1000 bootstrap iterations.

Testing of ORF lengths was performed using the codon permutation and synonymous mutation method by Schlub *et al.* (2018). Furthermore, evolution of the OLG loci was simulated along phylogenies according to Cassan *et al.* (2016) using Pyvolve (Spielman & Wilke, 2015). For *olg1*, an intact sequence from *P. prosekii* was chosen as outgroup, whereas various non-*P. aeruginosa* outgroup sequences without stop codons were available for *olg2*.

The tools FRESCo (Sealfon *et al.*, 2015) and OLGenie (Nelson *et al.*, 2020) were applied to determine the constraint in synonymous and non-synonymous sites in the mother genes *tle3* and PA1383 using a sliding window size of 50 codons.

3. Results

3.1 Identification of protein-coding genes in *E. coli* LF82

In this study, the *E. coli* LF82 genome was globally screened for protein-coding genes by using high-throughput methods. With a size of 4,773,108 bp and a GC content of 50.7%, the genome harbours 4,586 annotated, protein-coding genes (anORFs) according to RefSeq (Assembly: GCF_000284495.1). In addition, further 61,401 ORFs, constituting the longest possible ORFs (i.e., any codon as start codon allowed) between two stop codons with a minimum length of 93 bp, were detected and may also have the potential to encode for proteins. They are either located in intragenic regions (iORFs; 4,404) or overlap trivially (<30 nt; 1,170) or non-trivially (≥30 nt; 51,360) with annotated genes; the latter being divided into the categories listed in **Table 11**. For the remaining ORFs (4,467), a clear allocation was not possible because they were assigned to two or more categories (e.g., overlapping with multiple anORFs).

Table 11. Number and classification of non-trivial overlapping genes (OLGs) predicted for the *E. coli* LF82 genome. All ORFs exceeding 93 nt and overlapping more than 30 nt with an annotated gene were determined and classified according to the type of overlap and their strand location.

strand	overlap type	designation	number
sense	embedded	OLG_ES	15,771
	partially at 5`end	OLG_PS5	1,071
	partially at 3`end	OLG_PS3	905
antisense	embedded	OLG_EA	28,641
	partially at 5`end	OLG_PA5	2,664
	partially at 3`end	OLG_PA3	2,308

In order to discover new translation products encoded by unannotated ORFs in *E. coli* LF82, in total twelve NGS datasets were generated, which were:

- **Experiment 1 (Exp1) – eight datasets:** Ribo-seq & RNA-seq at two different cultivation conditions (aerobic & anaerobic) enable a general insight into the transcriptional and translational landscape as well as evaluation of differential gene expression as an indicator of gene functionality. This experiment was conducted in two biological replicates to capture random biological variation (section 3.1.1).
- **Experiment 2 (Exp2) – two datasets:** Ribo-RET facilitates the identification of translated genes by TIS mapping and allows correct assignment of their start codon position. An ND experiment without RET served as an internal expression control (section 3.1.2).
- **Experiment 3 (Exp3) – two datasets:** Ribo-seq using endoribonuclease RelE was conducted to increase RFP resolution and to detect a triplet periodicity allowing differentiation between background signals and genuine translation signals. An RNA-seq experiment was performed as a negative control for reading frame analysis (section 3.1.3).

In the following sections, all experiments are discussed and analysed individually with respect to annotated as well as unannotated ORFs. Finally, the results of the single experiment are combined and filtered in the final section to aid reliable OLG and iORF identification. The identified ORFs are further characterised by using different bioinformatic analyses.

3.1.1 Conventional RNA-seq & Ribo-seq

Transcriptome sequencing and ribosome profiling were applied to detect transcribed and translated regions in the *E. coli* LF82 genome under aerobic as well as anaerobic conditions. The sequencing output of two biological replicates were analysed regarding data quality, read length distribution and reproducibility. Afterwards, signals obtained for annotated ORFs were investigated in order to provide metrics for genuine expression and to evaluate suitability for differential gene expression analysis.

3.1.1.1 Sequencing output and data reproducibility

E. coli LF82 was cultivated in Schaedler broth with or without oxygen and harvested at the stationary phase (OD_{600nm} = 2, **Figure 9**), ensuring the availability of a sufficient amount of RNA for Ribo-seq. The cell lysate was processed for RNA-seq and Ribo-seq, and the extracted RNA was subjected to sequencing using an Illumina HiSeq 2500 system, yielding 60.6 to 98.6 million reads for the single experiments (**Table 12**).

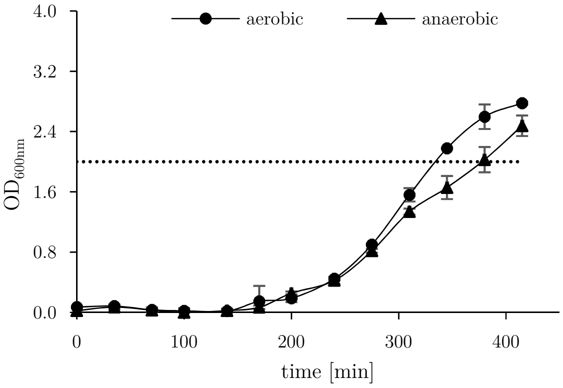


Figure 9. Growth curve of *E. coli* LF82 for RNA-seq and Ribo-seq. The optical density at 600 nm (OD_{600nm}) was measured in LB in biological triplicates. The dotted line indicates the sampling time for the experiment Exp1.

Quality trimming and genome alignment resulted in 15.4 to 51.6 million mappable reads; the remaining reads were either too short for mapping or did not align to the target genome (**Supplementary Table 2**). Since the majority of reads mapping to rRNA were located at the antisense strand (**Supplementary Figure 1**), a carryover of the rRNA-based hybridization probes used for rRNA depletion with the Ribo-Zero Kit (Illumina) seems most likely. Nevertheless, rRNA depletion is mandatory for RNA-seq experiments, since more than 80% of all cellular RNAs in bacteria are of ribosomal origin, and only 5% account for mRNAs (Westermann *et al.*, 2012). After mapping, 32.6 to 93.6% of all reads could be assigned to mRNA, thus confirming rRNA depletion efficiency. On average, the percentage of reads mapping to mRNA was higher in Ribo-seq experiments (Ø aerobic = 78.6%; Ø anaerobic = 62.6%) compared to the RNA-seq experiments (Ø aerobic = 40.8%; Ø anaerobic = 51.1%) due to the implementation of additional RFP enrichment steps, e.g., by PAA gel electrophoresis or ultracentrifugation. In total, 5.9 to 29.2 and 6.9 to 21.2 million mRNA reads were obtained for the RNA-seq and Ribo-seq experiments, respectively.

Table 12. Overview of Ribo-seq and RNA-seq reads obtained for Exp1 in *E. coli* LF82 after sequencing. Shown are the total number of sequenced reads as well as the number of reads mapping to the *E. coli* LF82 genome in millions for two biological replicates. Percentages in brackets indicate the proportion of reads mapping to rRNA and tRNA as well as mRNA regions of the *E. coli* LF82 genome.

			Aerobic cultivation				
Experiment	**Replicate**	**Total**	**Mapped**	**rRNA & tRNA**		**mRNA**	
Ribo-seq	I	65.1	15.4	5.6	(36.4%)	9.8	(63.6%)
	II	79.6	22.6	1.4	(6.4%)	21.2	(93.6%)
RNA-seq	I	64.5	18.2	12.3	(67.4%)	5.9	(32.6%)
	II	81.2	44.1	22.5	(51.0%)	21.6	(49.0%)
			Anaerobic cultivation				
Experiment	**Replicate**	**Total**	**Mapped**	**rRNA**		**mRNA**	
Ribo-seq	I	60.6	17.2	10.3	(59.9%)	6.9	(40.1%)
	II	82.8	24.7	3.7	(15.0%)	21.0	(85.0%)
RNA-seq	I	70.2	34.0	18.4	(54.3%)	15.5	(45.7%)
	II	98.6	51.6	22.4	(43.5%)	29.2	(56.5%)

Data reproducibility was determined by calculating Pearson correlation coefficients *r* between reads numbers obtained in biological replicates. For this purpose, reads mapping to anORFs were normalized to sequencing depth and gene length, and obtained RPKM values were used for the pairwise calculation of Pearson's *r*. Pearson's *r* coefficient ranged between 0.65 and 0.95 (**Figure 10**), indicating a moderate to very strong linear relationship between the biological replicates.

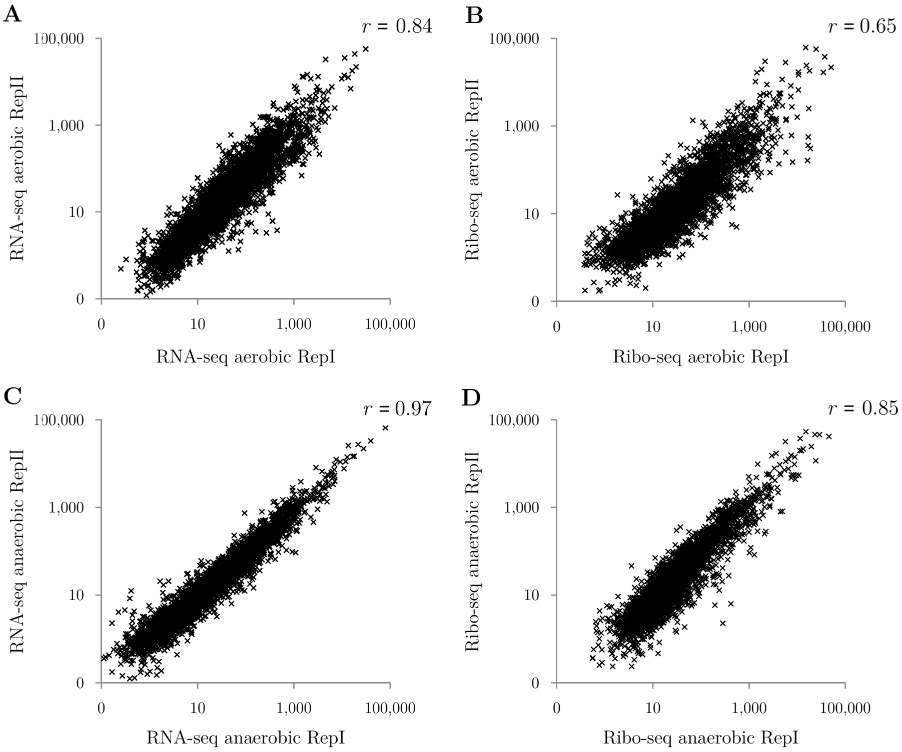


Figure 10. Reproducibility of biological Ribo-seq and RNA-seq experiments. Shown are the RPKM values obtained for all annotated genes (n = 4,586) of two biological replicates cultivated under (**A, B**) aerobic and (**C, D**) anaerobic conditions with their respective Pearson correlation coefficients (*r*).

On average, the aerobic experiments exhibited a lower correlation coefficient than the anaerobic and the linear relationship was stronger for the RNA-seq experiments compared to the Ribo-seq experiments. The latter observation might be explained by the increased number of experimental steps required for Ribo-seq offering a higher possibility for variation throughout the experimental procedure. Technical replicates were not included since reproducibility of sequencing using the Illumina technology was shown to be excellent (e.g., Marioni *et al.*, 2008). Therefore, variability as observed for the replicates can solely be traced back to biological and experimental variation.

3.1.1.2 Analysis of anORF expression

The RNA-seq and Ribo-seq data was investigated regarding transcription and translation of anORFs in the *E. coli* LF82 genome. In general, only slight differences between the aerobic and anaerobic datasets were observed when calculating RPKM and RCV values (**Table 13**). In both datasets, the median RPKM value of the translatome of all anORFs was around 15 and the respective values of the transcriptome marginally higher. To determine the number of active genes in transcriptome and translatome datasets, a RPKM threshold of 10 was applied. This threshold is commonly used to differentiate background signals from genuine signals (e.g., Fremin *et al.*, 2020, Neuhaus *et al.*, 2016) and is 20-fold higher than the RPKM value obtained for background transcription as suggested by Landstorfer *et al.* (2014). When applying this conservative threshold on the translatome data, 59.6% (aerobic) and 62.5% (anaerobic) of all anORFs had similar or higher expression levels under the conditions tested. This observation is in concordance with the results obtained from mass spectrometry detecting approximately 53% of all proteins encoded by anORFs after aerobic cultivation of *E. coli* LF82 (unpublished data).

Table 13. Overview of expression metrics obtained for all annotated genes (n = 4,586) of *E. coli* LF82. Reads per million mapped reads (RPKM) were calculated for RNA-seq and Ribo-seq of Exp1 and averaged over biological replicates. Ribosome coverage values (RCVs) indicating translatability were calculated by dividing the RPKM values of the translatome by the RPKM values of the transcriptome.

	RPKM RNA-seq		**RPKM Ribo-seq**		**RCV**	
	aerobic	**anaerobic**	**aerobic**	**anaerobic**	**aerobic**	**anaerobic**
Median	21.7	16.5	14.8	15.4	0.8	1.1
1st quartile	6.6	4.4	5.5	6.7	0.5	0.7
3rd quartile	76.0	59.8	54.3	51.3	1.3	2.1

In order to detect genes regulated in response to the presence or absence of oxygen, differential gene expression analysis was performed. In total, 210 genes with a logFC ≥|1| and a p-value ≤0.05 (**Supplementary Table 3**) were detected to be differentially regulated depending on data type (RNA-seq or Ribo-seq) and cultivation conditions (aerobic vs. anaerobic). At the translational level, 143 genes were regulated when comparing the anaerobic condition with the aerobic reference condition. Eighty thereof showed upregulation, and the remaining 63 were downregulated during cultivation in the absence of oxygen (**Supplementary Figure 2**). At the transcriptional level, a lower number of 93 genes were differentially expressed with a slightly higher proportion of downregulated genes under anaerobic conditions (**Supplementary Table 3**). Remarkably, the genes showing the highest downregulation in the anaerobic dataset were genes belonging to the *flg* and *fli* family, which are structural components of flagella or constitute regulatory factors of flagella-associated genes (Fitzgerald *et al.*, 2014, Macnab, 1992). Transcription of such genes was shown to be repressed by RpoS (Dong *et al.*, 2011, Dong & Schellhorn, 2009), which represents the general stress response σ factor in *E. coli.* Bayramoglu *et al.* (2017), who performed global transcriptome sequencing of planktonic *E. coli* MG1655 cultures with and without oxygen, observed a relation between anaerobic cultivation and downregulation of RpoS-regulated genes as well.

In addition, some of the genes reported to be upregulated by Bayramoglu *et al.* (2017) under anaerobic conditions were also detected in this study, including stress-associated genes like the cold shock inducible genes *cspB* and *cspI* (Wang *et al.*, 1999b). Of all differential expressed genes, only 26 were regulated both at the transcriptional as well at the translational level.

3.1.2 Ribo-RET for translation initiation site (TIS) detection

The aim of this experiment was to identify TISs in *E. coli* LF82 by applying the recently published method Ribo-RET. For this purpose, the *tolC* gene encoding for the TolC outer membrane efflux protein mediating resistance to RET had to be deleted in a first step to guarantee the proper functioning of the antibiotic. In a next step, RET´s efficiency in deletion mutant cells was examined by MIC testing in order to specify the amount of RET necessary for complete translational inhibition. Finally, the Ribo-RET experiment was conducted applying the 100× MIC.

3.1.2.1 Construction of a Δ*tolC* deletion mutant

The method according to Datsenko & Wanner (2000) was intended to be used for the creation of a *tolC* deletion mutant. An FRT-flanked kanamycin resistance cassette was amplified via PCR and ligated to 50 nt long regions which were homologous to the flanking sequences of the chromosomal *tolC* gene. Sequencing confirmed the construct. Transformation into competent LF82 cells expressing λ Red recombinase yielded several kanamycin-resistant colonies. However, colony-PCR using primers binding within *tolC* as well as sequencing confirmed that the *tolC* gene was still present in the genome of the putative transformants. Even after increasing the homologous sequences up to 900 nt, the correct genomic integration of the kanamycin cassette failed (results not shown). After multiple unsuccessful attempts, the one-step gene inactivation method was discontinued.

Alternatively, a suicide plasmid containing upstream and downstream flanking regions of *tolC* was used for gene inactivation. In a first step, a 925 bp region located upstream of *tolC* as well as a 938 bp region located downstream of *tolC* were amplified via PCR and ligated using previously introduced restriction enzyme cleavage sites. This deletion fragment was inserted into the multiple cloning site of the vector pMRS101, yielding vector pMRS101-TolC, before plasmid and deletion fragment integrity were confirmed by sequencing (results not shown). After plasmid propagation, the high copy origin of pMRS101-TolC along with a gene conferring ampicillin resistance were removed from the plasmid. Agarose gel electrophoresis confirmed the excision of the 1,766 bp large fragment resulting in the 8,551 bp large suicide vector pKNG101-TolC after self-ligation. Circularized pKNG101-TolC was propagated and transformed into *E. coli* SM10λpir cells. After verifying the uptake of the plasmid via colony-PCR, a positive colony was chosen for plate mating with the ampicillin-resistant recipient strain *E. coli* LF82. Incubation on agar supplemented with ampicillin and kanamycin allowed for the selection of transformed *E. coli* LF82. Genomic insertion of pKNG101-TolC via homologous recombination either took place up- or downstream of the intrinsic *tolC* gene as indicated in **Figure 11**.

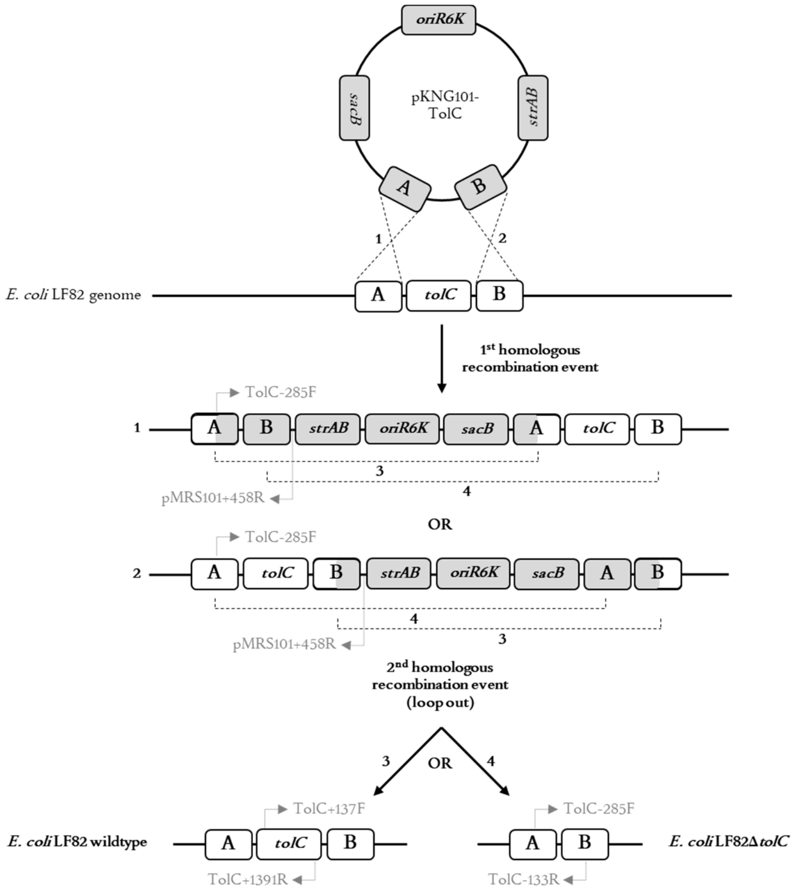


Figure 11. Scheme of the genomic integration and subsequent removal of the *tolC* deletion fragment in *E. coli* LF82. The suicide plasmid pKNG101-TolC with its origin of replication (*oriR6K*) and the genes *sacB* and *strAB* encoding for a levansucrase and enzymes conferring resistance to streptomycin were integrated into the genome of *E. coli* LF82 by homologous recombination between the upstream flanking regions of *tolC* named "A" (possibility 1) or the downstream flanking regions named "B" (possibility 2). Transformants with genomic integration of the plasmid were selected by PCR using the primer pairs TolC-285F and pMRS101+458R as indicated. A loop out of the integrated plasmid backbone facilitated by cultivation on sucrose agar resulted either in restoration of the wild type genotype (possibility 3) or successful deletion of the *tolC* gene (possibility 4), which was confirmed by PCR using the primer pairs TolC+137F & TolC+1391R and TolC-285F & TolC-133R. Figure adapted from Graf (2019).

Colony-PCR using a *tol*C-flanking primer (TolC-285F) and a vector-specific primer (pMRS101+458R) confirmed that half of the tested colonies integrated the vector according to possibility 1 (**Figure 11**), whereas the remaining colonies had a crossover event according to possibility 2 (**Figure 11**) resulting in PCR band of 1,415 bp and 2,897 bp, respectively (data not shown). To remove the intrinsic *tolC* gene as well as the vector backbone, one mutant was picked and cultivated on sucrose agar to promote a loop out by homologous recombination. The conversion of sucrose to the cell-toxic levan catalysed by the *sacB* encoded enzyme levansucrase (Li *et al.*, 2013, Gay *et al.*, 1983) enabled selection of mutants, which have eliminated the vector backbone. Depending on the location of the second homologous recombination, the selected cells had either the wild type gene structure including the *tolC* gene or they had successfully eliminated the gene of interest (**Figure 11**, possibility 3 or 4). Multiple mutants were screened with colony-PCR using primer pairs binding either in *tolC* flanking regions (primer TolC-285F & primer TolC-133R) or within *tolC* (primer TolC+137F & TolC+1391R) yielding PCR amplicons of 1,900 and 1,374 bp (**Figure 12**), respectively. Only clone number 7 exhibited the desired band pattern, showing a band of the same size (448 bp) as the pKNG101-TolC vector after PCR with *tolC* flanking primers (**Figure 12A**, lane 7), and the absence of a 1,374 bp band after PCR using primers binding within *tolC* (**Figure 12B**, lane 7). Sequencing of clone number 7 unequivocally confirmed the absence of *tolC*, and therefore, this clone was used for the subsequent experiments.

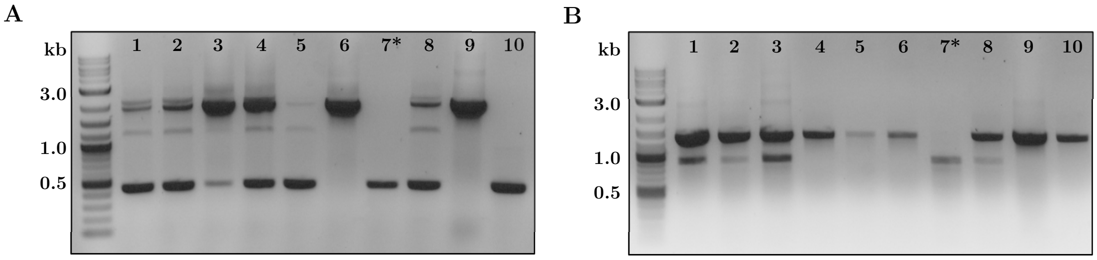


Figure 12. Agarose gel analysis of putative *E. coli* LF82Δ*tolC* deletion mutants. Shown are the band patterns for seven clones (lane 1-7) obtained after PCR using the primer pairs (**A**) TolC-285F & TolC-133R or (**B**) TolC+137F & TolC+1391R. Genomic DNA of *E. coli* LF82 (lane 9) and pure vector DNA of pKNG101-TolC (lane 10) were used as an internal control. Only clone 7 was confirmed to lack the *tolC* gene as indicated by the absence of 1,900 and 1,374 bp bands in A and B, respectively.

3.1.2.2 Determination of the minimum inhibitory concentration of retapamulin

Clone number 7 was cultivated and diluted to approximately 1×10^6 CFU/mL with Schaedler broth. The adjusted suspension was used for MIC testing, using the microdilution method in 96-well plates in three biological and technical replicates. For each experiment, 100 µL of the cell suspension were mixed with 100 µL of Schaedler broth supplemented with varying RET concentrations in order to obtain an optimal inoculum size of 1×10^5 CFU/mL. *E. coli* LF82 wild type cells were used to compare the effect of RET, depending on the presence or absence of the *tolC* gene.

Agar plating confirmed an initial cell concentration of 2.0×10^5 (±standard deviation 2.6×10^5) for the deletion mutant and 1.2×10^6 (±standard deviation 1.2×10^5) for the wild type strain yielding a final inoculum of 1.0×10^5 and 6.0×10^6 CFU/mL, respectively. After 24 h of cultivation, the MIC was defined as the lowest concentration of RET which inhibited bacterial growth as indicated by a lack of turbidity. The wild type strain exhibited a MIC value of ≥32 µg/mL, while *E. coli* LF82Δ*tolC* was more susceptible to RET demonstrated by a lower MIC value of 0.25 µg/mL (**Figure 13**). Those values were replicated by all three biological experiments confirming their reliability after 24 h of cultivation. Continuous measurement throughout the entire cultivation time revealed modest variations in growth for the deletion mutant compared to the wild type strain, probably caused by larger fluctuations in the initial cell inoculum. However, the overall growth trend was fairly consistent (**Supplementary Figure 3**).

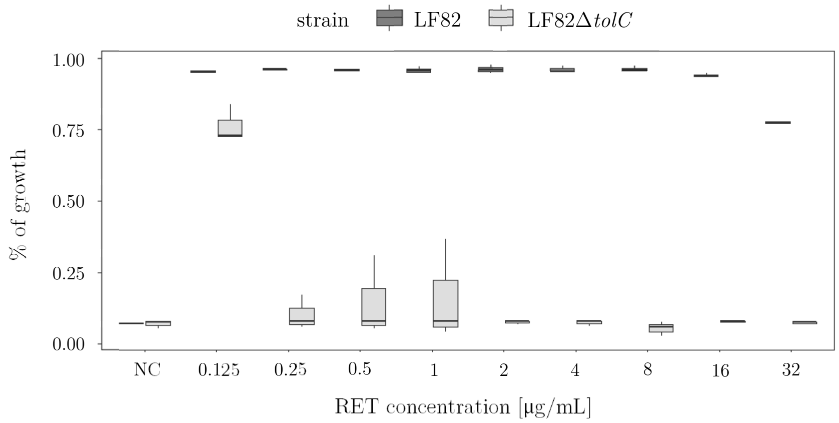


Figure 13. Results of minimum inhibitory concentration (MIC) testing of retapamulin (RET) for *E. coli* LF82 and *E. coli* LF82Δ*tolC*. The overall growth of both strains in the presence of different RET concentrations ranging from 0.125 to 32 µg/mL was calculated after 24 h by dividing the measured optical density by the respective value of bacterial cultures without RET (positive control). Plain broth without cell inoculum was used as a negative control (NC).

3.1.2.3 Cut-off specification for genome wide TISs analysis

The strain *E. coli* LF82Δ*tolC* was cultivated for approximately 300 min until reaching an OD_{600nm} of 1.4 (**Figure 9**) and incubated with the 100× MIC of RET to stall initiating ribosomes. After sequencing, reads were processed and aligned to the *E. coli* LF82 genome resulting in 23.8 and 65.8 mio mRNA reads for the Ribo-RET and the ND control Ribo-seq experiment. With values of 13.5% and 28.5%, the overall percentage of reads mapping to mRNA regions within the *E. coli* LF82 genome was substantially lower than those of the first experiments (see **Table 12**). One possible explanation for this observation may be the lack of rRNA depletion due to the discontinuation of the Ribo-Zero rRNA Depletion Kit by Illumina. In order to compensate for the missing rRNA depletion, the sequencing depth was increased (**Supplementary Table 2**). All reads mapping to mRNA regions were analysed regarding their read length to confirm proper size selection. Obtained read lengths were on average larger (30 – 40 nt, **Supplementary Figure 4E**) compared to those from the first experiment (20-26 nt, **Supplementary Figure 4AB**), which was consistent with the different selection ranges after PAA gel electrophoresis.

In order to evaluate the efficiency of RET in stalling ribosomes at the start codon of translated genes, the first nucleotide located at the P site of all reads was determined by subtracting 15 nucleotides from the 3' end as described by Meydan *et al.* (2019). All P site positions were normalized to sequencing depth resulting in RPM values, followed by visual inspection using a genome browser. Treatment with RET resulted in a pronounced redistribution of ribosomes as exemplary as shown for the genes *rpsJ*, *rplC*, *rplD*, *rplW* and *rplB* of the S10 ribosomal protein operon (Zurawski & Zurawski, 1985), known to encode for highly expressed proteins (Roymondal *et al.*, 2009, Karlin *et al.*, 2001). Those genes were among the 10% of the highest expressed anORFs in the ND dataset with RPKM values >400 (**Table 14**).

Table 14. Translation metrics of highly expressed proteins of the S10 ribosomal protein operon in *E. coli* LF82. RPM values at the start peak positions of the genes *rplB, rplW, rplD, rplC* and *rpsJ* were calculated in the retapamulin (RET) and no drug (ND) dataset and fold changes were determined. For the ND dataset, reads were additionally normalized to gene length yielding RPKM values.

gene	encoded protein	ND [RPKM]	RET peak at start position [RPM]	ND peak at start position [RPM]	Fold change peak
rplB	50S ribosomal protein L2	1,698.6	643.8	4.2	154.1
rplW	50S ribosomal protein L23	3,261.6	2,1682.0	660.6	32.8
rplD	50S ribosomal protein L4	1,836.8	1897.0	25.8	73.4
rplC	50S ribosomal protein L3	1,373.6	722.4	14.8	48.9
rpsJ	30S ribosomal protein S10	438.4	123.1	1.0	126.6

In the RET data, an accumulation of ribosomes at start codons was observable, whereas in the ND dataset ribosomes were distributed over the entire coding region (**Figure 14**). The factor by which the RET start peak was higher than the respective ND signals at the start position ranged from 32.8 for *rplW* to 154.1 for gene *rplB* (**Table 14**). Comparison of RET peak height with their respective expression strength as specifies by RPKM values in the ND dataset indicated a weak correlation between both metrics. This observation was confirmed by a low Pearson correlation coefficient *r* of 0.15, when analysing PRKM and RPM values of all anORFs. However, when determining the correlation according to Spearman, which ranks absolute values, and thus, is less sensitive towards outliers (de Winter *et al.*, 2016), a moderate correlation of 0.64 was obtained.

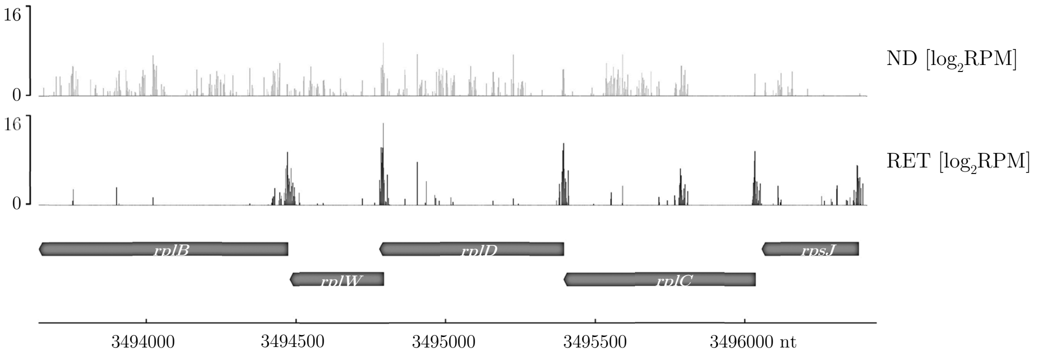


Figure 14. Translation signals of highly expressed proteins of the S10 ribosomal protein operon. Logarithmic RPM values are shown at each position of the genes *rplB, rplW, rplD, rplC* and *rpsJ* in the no drug (ND; top panel) and the retapamulin (RET; bottom panel) dataset.

Preciseness of the ribosome accumulation at start codons was evaluated by metagene analysis. For this analysis, all anORFs separated more than 50 nt from adjacent genes, and exhibiting RPM values >100 in the ND as well as in the RET dataset were chosen. RPM values at each position of the selected anORFs were normalized by dividing them by the entire RPM value of the respective gene including 30 nt flanking regions. As seen in **Figure 15A**, metagene analysis confirmed the inhibitory effect of RET on initiating ribosomes at start codons of 224 highly expressed genes. Compared to the ND control, the normalized read count was approximately six-fold increased at the start position +1 (i.e., 0 nt distance from start; see **Figure 15**) after RET treatment. The majority of reads was mapping in a distance window of -2 to +8 nt around the first nucleotide of the start codon resulting in a broader peak than observed for the data by Meydan *et al.* (2019) upon reanalysis (**Supplementary Figure 5**). In addition, the maximum RPM counts were obtained two nucleotides downstream of the first start codon nucleotide indicating suboptimal P site mapping. To compensate for P site imprecision, mapping was repeated with a corrected offset of 17 nt. This correction resulted in a sharp peak at the distance from start codon position 0 nt (**Figure 15B**).

The prediction algorithm provided by Meydan *et al.* (2019) was applied to the Ribo-RET dataset for the identification of translated genes. This algorithm globally identifies RET peaks exceeding a certain threshold and searches for a suitable start codon within a ±3 nt window around the identified peak. Increasing the offset to 17 nt optimized the location of the search area (**Figure 15B**) and resulted in slightly more predictions compared to the data mapped with a 15 nt offset (**Supplementary Figure 6**). However, the optimal threshold for the peak height had to be defined prior to subsequent analyses. Therefore, the algorithm was applied using thresholds ranging from 0.2 to 5, and the final number of predictions was determined. As expected, both variables correlated inversely; a higher threshold resulted in fewer predictions (**Supplementary Figure 6**). To determine the optimal threshold value, fold changes between the numbers of hits obtained for two consecutive values were calculated. Above a threshold of 1, a further increase of the value did not lead to a substantial reduction of predicted hits (**Supplementary Figure 6**) indicating that this threshold may be suitable for whole genome analysis. By applying this threshold, 88% of all anORFs with a RPKM value larger than 100 in the ND dataset were successfully predicted. This result is in concordance with those reported by Meydan *et al.* (2019), who also defined a threshold of 1 for ORF prediction in two *E. coli* strains, thereby obtaining an anORF prediction rate of 86%.

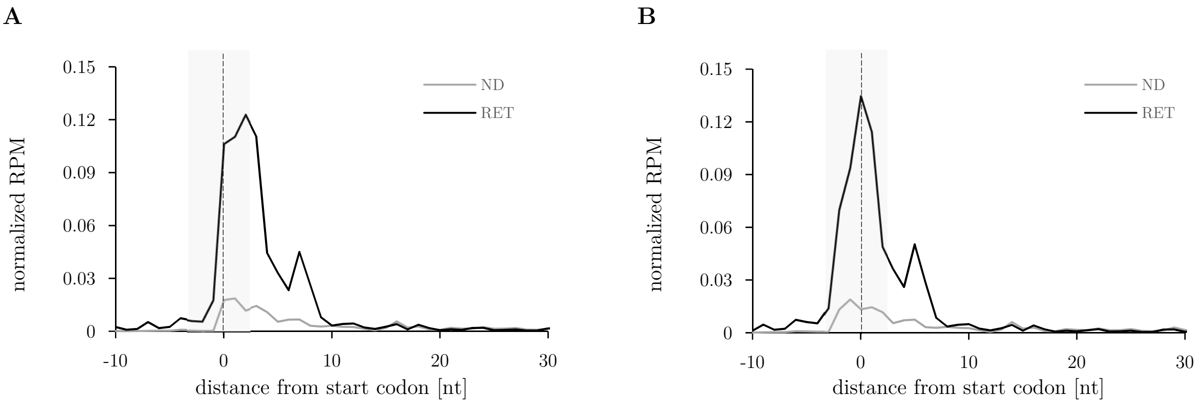


Figure 15. Metagene analysis of highly expressed genes in *E. coli* LF82. Shown are normalized RPM values of each position in a -10 to 30 nt window around the start codon (dashed line, distance 0 nt) after determining the P site position by subtracting (**A**) 15 nt or (**B**) 17 nt of all genes with RPM values >100 in the ND and RET datasets (n = 224 or n = 196, respectively).

3.1.3 RelE-enhanced Ribo-seq to visualize triplet periodicity

3.1.3.1 Expression and downstream processing of RelE

To utilize RelE´s unique cleavage preciseness in Ribo-seq experiments, the endogenous enzyme had to be expressed in sufficient amounts using a suitable expression system. For this purpose, the sequence of the PCR-amplified *relE* gene originating from *E. coli* MG1655 was fused to an N-terminal His$_6$-tag and cloned into the expression vector pBAD/HisC. After transformation in *E. coli* Top10, *relE* expression was induced at OD$_{600nm}$ = 0.5 by the addition of different amounts of arabinose. However, this experiment was aborted due to the cytotoxic effect of RelE, leading to a massive reduction of cell growth upon overexpression (data not shown). The lethal or inhibitory effect of overexpressed RelE is well known, and therefore, several groups tried to co-express RelE together with its respective antitoxin RelB with successful outcomes (e.g., Cherny *et al.*, 2007, Pedersen *et al.*, 2002, Gotfredsen & Gerdes, 1998). In a second experiment, the co-expression approach was implemented using the plasmid pET-22b(+)His$_6$:*relB*$_{\Delta 9}$-*relE*$_{WT}$ received from Dr. Scott Strobel (Dunican *et al.*, 2015, Griffin *et al.*, 2013). Sequencing verified that the obtained plasmid contained a mutant His$_6$-tagged *relB* sequence and a *relE* sequence with an overlapping stop/start codon as found in the native system under T7 RNA polymerase control. The plasmid pET-22b(+)His$_6$:*relB*$_{\Delta 9}$-*relE*$_{WT}$ was transformed into *E. coli* BL21 (DE3) pLysS cells, and protein expression was induced by the addition of IPTG. SDS-PAGE analysis of samples taken throughout the entire cultivation process confirmed the induction of a ~10 kDa protein (**Figure 16A**, highlighted by a box) compared to *E. coli* BL21 (DE3) pLysS without overexpression plasmid. A clear visual assignment of the observed band to either RelE or RelB was hampered by their similar molecular weight of 11.2 kDa and 9 kDa (Gotfredsen & Gerdes, 1998), respectively. The fusion to a 6× His-tag theoretically increased the molecular weight of RelB by 0.8 kDa (Young *et al.*, 2012), suggesting that the 10 kDa band might have originated from the His-tagged RelB. Western blot analysis using an anti-His-antibody for detection of RelB confirmed this assumption (**Figure 16B**). To validate the presence of the untagged RelE, semi-quantitative whole proteome analysis via MS was performed of a sample taken 3 h after induction. Both proteins were present in this sample (**Supplementary Figure 7A**). With an iBAQ value of 7.8×10^{10}, RelB represented the most abundant protein in the sample, followed by RelE (iBAQ = 4.1×10^{10}).

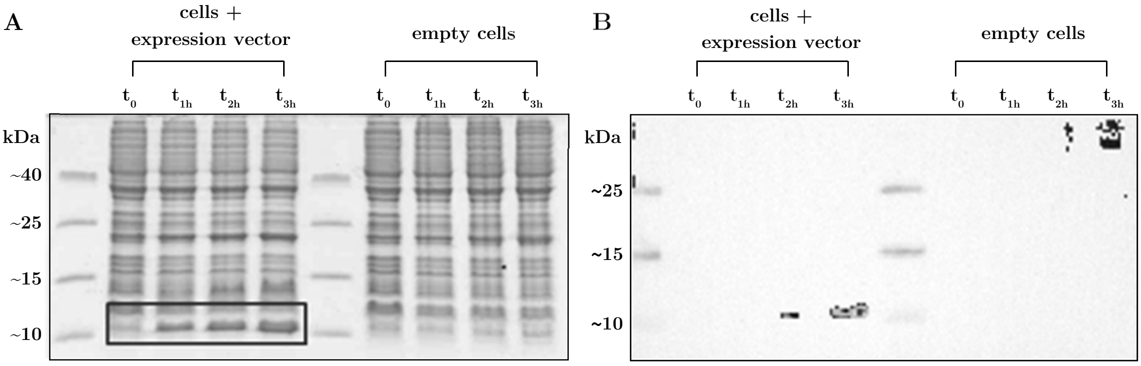

A
cells + expression vector
empty cells
t_0 t_{1h} t_{2h} t_{3h}
t_0 t_{1h} t_{2h} t_{3h}
kDa
~40
~25
~15
~10
B
cells + expression vector
empty cells
t_0 t_{1h} t_{2h} t_{3h}
t_0 t_{1h} t_{2h} t_{3h}
kDa
~25
~15
~10

Figure 16. Results of (**A**) SDS-PAGE and (**B**) Western blot analysis of *relE* and *relB* overexpression samples. *E. coli* BL21 (DE3) pLysS cells with and without expression plasmid pET-22b(+)His$_6$:r*elB*$_{\Delta 9}$-*relE*$_{\mathrm{WT}}$ were cultivated and samples were taken immediately (t_0), 1 h (t_{1h}), 2h (t_{2h}) and 3h (t_{3h}) after induction with IPTG.

After verifying successful overexpression, RelE should be selectively purified using IMAC. The underlying principle was to enrich the RelE/His-tagged RelB-complex by direct interaction between the His residues and matrix-immobilized Ni^{2+} ions. In the next step, the conformation-dependent interaction between both proteins should be abolished by denaturing conditions enabling selective elution of RelE. In a first attempt, RelE was purified using the Ni-NTA Fast Start Kit (Qiagen) under native as well as denaturing conditions. When applying denaturing buffer to the provided Ni-NTA columns and taking samples of the flow through, two bands were observed, a moderate band of 10 kDa representing RelB and a very diffuse band of approximately 13-14 kDa (**Figure 17A**, Lane 5 & 6), most likely caused by RelE. MS analysis again confirmed the presence of both proteins in the flow through after washing with denaturing buffer and confirmed a slight change in the ratio of RelE to RelB compared to initial overexpression sample (**Figure 17B**). However, SDS-PAGE analysis demonstrated several limitations of this purification method. Firstly, the large amount of unbound RelB present in the flow through after loading of the lysate indicated insufficient interaction with the resin (**Figure 17A**, Lane 2). Secondly, selective elution of RelE using denaturing conditions failed due to the co-elution of RelB (**Figure 17A**, Lane 5 & 6), which constituted the most abundant protein according to the MS analysis (**Figure 17B**). Thirdly, the denaturing strength was not sufficient to abolish the toxin´s strong interaction with the antitoxin because large amounts of RelE were still present after eluting RelB by adding a high-imidazole-containing buffer (**Figure 17A**, Lane 7 & 8). Consequently, the maximum yield of pure RelE was reduced.

The use of chromatography columns coupled to FPLC systems is a more sophisticated way of protein purification and allows the application of customized buffers as well as the optimization of flow speeds and interval lengths. By using a FPLC system with a pre-packed His-column, RelE was eluted more efficiently over RelB (**Supplementary Figure 7B**). However, a high protein background in the eluate indicated a high proportion of non-specifically bound proteins. To further optimize the purification process, a semi-batch procedure including sample binding at 4 °C was combined with an automated procedure including RelE and RelB elution at room temperature. As a first step, a Ni-NTA resin was incubated with the lysate for 1 h at 4 °C. Afterwards, the resin was washed several times with washing buffer at 4 °C before the column was packed and connected to the ÄKTA FPLC system. Finally, RelE was eluted with denaturing buffer at room temperature. In contrast to the completely automated procedure described above, the bed volume of the column and the length of the washing interval were increased in this approach. Those modifications resulted in two characteristic increases in the UV_{280nm} signals indicating successful protein elution (**Figure 18A**). Subsequent SDS-PAGE analysis confirmed the successful binding of RelB to the resin (presence of the 10 kDa protein in the lysate, **Figure 18B**, lane 1; and absence in the flow through, **Figure 18B**, lane 2) and the highly-selective separation of non-specifically bound proteins (**Figure 18**, peak 1 & lane 3) from eluted RelE (**Figure 18**, peak 2 & lane 4). A RelB band was observable after elution with 500 mM imidazole buffer (**Figure 18B**, lane 5). The fractions of the second peak were pooled, and RelE was subsequently refolded by dialysis, concentrated, again dialysed and stored at –80 °C.

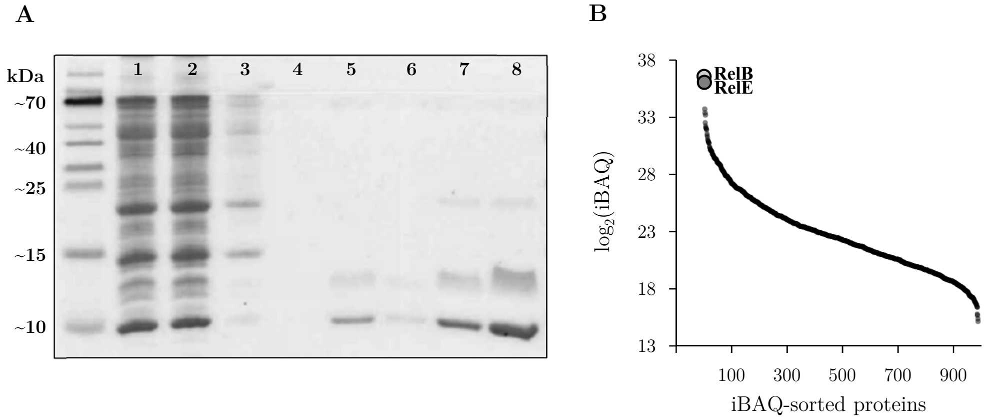


Figure 17. Results of (**A**) SDS-PAGE and (**B**) mass spectrometry analysis of RelE and RellB purification samples. The whole cell lysate before purification (**A**, lane 1), the flow through after applying the cell lysate on top of a Ni-NTA column (**A**, lane 2), the flow through after two rounds of washing with native wash buffer (**A**, lane 3 & 4), the flow through after two rounds of washing with denaturing wash buffer (**A**, lane 5 & 6), and the final flow through after elution with elution buffer (**A**, lane 7 & 8) are shown. The sample obtained after washing with denaturing wash buffer (**A**, lane 5) was subjected to (**B**) mass spectrometry analysis to validate the presence of the target proteins. The logarithmic iBAQ value of each detected protein (y-axis) was plotted against the detected proteins sorted by their iBAQ values (x-axis). Each dot represents one detected protein.

After downstream processing, a final concentration of 387 µg/mL in a total volume of 1.35 mL was measured via Bradford assay (**Supplementary Figure 8**). The SDS-PAGE analysis of the final sample showed an identical band pattern to that which was obtained directly after the purification process (see **Figure 18B**, lane 4). A total of 159 different *E. coli* proteins with a minimum of 2 unique peptide matches were identified in the final sample via MS. The protein with the highest abundance was RelE (top3 = 8.9), followed by the chaperone DnaK (top3 = 7.7) and the outer membrane protein OmpF (top3 = 7.2). According to the top3 values, the final protein solution contained 84.1% RelE, 15.6% other *E. coli* proteins and only 0.3% RelB, indicating the high purity of the RelE solution.

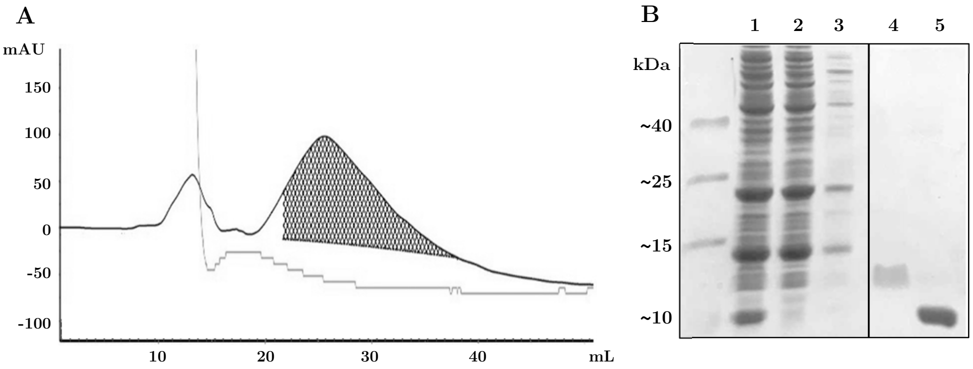


Figure 18. (**A**) Chromatogram and (**B**) SDS-PAGE analysis of the final ÄKTA purification process. The column was washed with 35 mM imidazole washing buffer (**A**, first peak); afterwards RelE was eluted by adding denaturing buffer (**A**, second peak). The dark line indicates the UV signal at 280 nm; the light line represents the conductivity, and thus, indicates the time point of buffer exchange. Samples were taken throughout the entire purification process and subjected to (**B**) SDS-PAGE analysis. Shown are the whole cell lysate before purification (lane 1), the flow through after resin incubation (lane 2), the flow through after washing with 35 mM imidazole buffer (lane 3), the elution of RelE using denaturing buffer (lane 4) and the final elution of RelB with 500 mM imidazole elution buffer (lane 5).

3.1.3.2 Evaluation of reading frame information

Purified RelE was added to the cell lysate during the Ribo-seq procedure in order to improve resolution of the experiment and to obtain RFPs with precise 3`ends. Nuclease digestion was performed in the presence of 450 U MNase for additional cleavage at the 5´end of the RFPs. Afterwards, RFPs with a size of 10-40 nt were subjected to library preparation and sequencing. Due to RelE´s cleavage within the A site of the ribosomes, expected RFPs were on average shorter than those of the other datasets, and thus, required different processing and mapping parameters (for details see section 2.2.5.1.3). After mapping, 23.4 mio and 0.4 mio mRNA reads were obtained for the Ribo-seq and RNA-seq experiment, respectively (**Supplementary Table 2**). Regarding read length, the predominant number of RFPs in the Ribo-seq experiment had a size of 26 to 31 nt (**Figure 19**). Additionally, a smaller proportion of shorter reads (14 to 18 nt) was observable. A similar result was reported by Hwang & Buskirk (2017), who noticed a high read proportion with sizes of ~25 nt as well as a substantial number of short RFPs (13-16 nt) after RelE cleavage. In contrast, nuclease footprinting with MNase solely resulted in a read distribution centred at 25 nt. Reanalysis of the data published by Hwang & Buskirk (2017) confirmed these results (**Figure 19**).

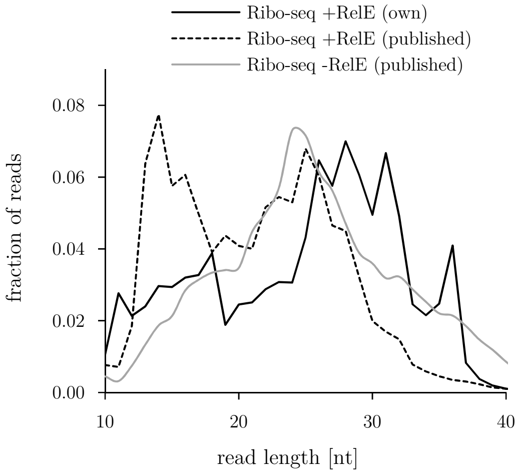


Figure 19. Read length distribution of RelE-supported Ribo-seq in *E. coli* LF82. The total proportion of reads mapping to mRNA regions of the Exp3_RelE dataset generated in this study ("Ribo-seq +RelE") as well as of the datasets with ("Ribo-seq +RelE (published)") and without RelE ("Ribo-seq -RelE (published)") published by Hwang & Buskirk (2017) are shown.

To obtain reading frame information, all 3´ends of RFPs were mapped to the genome and their sub-codon positions (1, 2 or 3) were ascertained within the coding region. Reads within the first and the last 30 nt of each coding region were excluded to avoid distortion by reads mapping to adjacent coding region. RelE cleaves specifically after the second nucleotide of the codon which is located in the A site of the ribosome. However, RelE has a strong sequence bias: If the codon in the A site is ending with a C, RelE preferentially cleaves after the third nucleotide instead of the second (Hwang & Buskirk, 2017). To compensate for this cleavage bias, 3´ends of RPFs mapping to sub-codon position 3 and ending with a C were shifted from position 3 to position 2. This analysis was firstly performed for the gene *ompA*, which was the highest-expressed gene in the Ribo-seq RelE dataset with an RPKM value of 120,369. Based on visual inspection of read number and coverage, the region between 600 nt and 632 nt was selected for reading frame analysis. As expected, most reads mapped to position 2, whereas reads mapping to the other positions were clearly underrepresented (**Figure 20**).

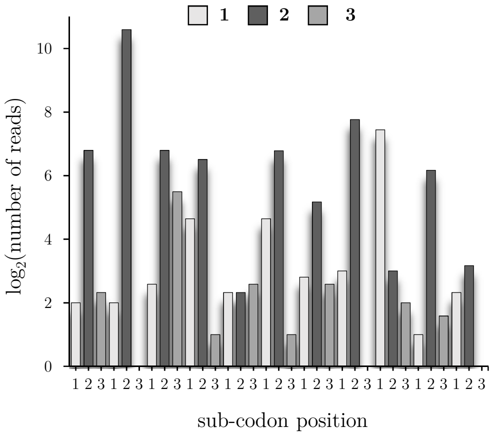


Figure 20. Reading frame analysis of gene *ompA* in *E. coli* LF82. Sub-codon positions of the 3´ ends of all reads obtained in the Ribo-seq experiment prepared with RelE mapping to a region between 600 and 632 nt of *ompA* were determined. A shift of reads ending in C was performed. Absolute number of reads are specified in a logarithmic scale.

The ribosomal reading frame was even more pronounced when calculating the sum signal of all anORFs, even without shifting for the NNC cleavage bias of RelE. 48% of all RFPs mapped to sub-codon position 2, 32% to sub-codon position 3 and 20% to sub-codon position 1 (**Figure 21A**). A NNC shift resulted in an improved reading frame signal of 57%, 24% and 20% for sub-codon positions 2, 3 and 1, respectively (**Figure 21B**).

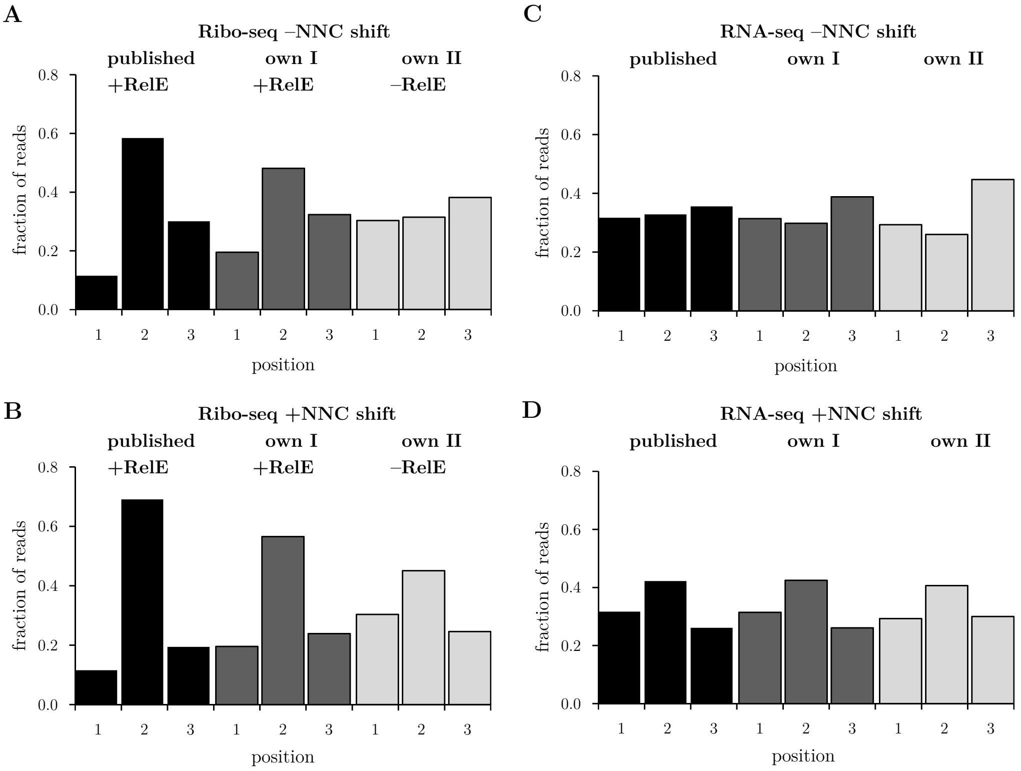


Figure 21. Reading frame sum signal of all annotated genes for (**A, C**) Ribo-seq and (**B, D**) RNA-seq datasets in *E. coli* LF82. Subtraction of 30 nt from start and stop positions was performed. Figure **A** and **C** display the results of the raw data; Figure **B** and **D** show the reading frame signal after shifting of reads arising from codons ending in C. In addition to the Ribo-seq and the corresponding RNA-seq dataset of this study (specified as "own I"), the data published by Hwang & Buskirk (2017) was also reanalysed (specified as "published"). The datasets of Exp1_aerobic_RepI (specified as "own II") served as an independent negative control for reading frame analysis.

A similar trend was observable for the published data yielding the highest read density at position 1 (69%), followed by position 3 (19%) and 1 (12%) after NNC shift (**Figure 21B**). Deviating proportions of reads mapping to sub-codon position 1 mainly caused the discrepancy in the sum signal resolution between the published data and the data of this study. To exclude that the observed signals are solely originating from the nucleotide bias in the coding regions, a previously generated Ribo-seq dataset (Exp1_aerobic_RepI, section 3.1.1), showing similar mRNA read numbers as well as a comparable experimental treatment, was reanalysed using the RelE-optimized processing and mapping parameters. Reading frame analysis of this dataset revealed the complete absence of a periodicity signal in the unshifted data (**Figure 21A**, "own II -RelE") and only a modest preference of sub-codon position 2 caused by NNC read shifting (**Figure 21B**, "own II -RelE"). Analysis of the corresponding RNA-seq datasets confirmed a reproducible lack of periodicity without read shift (**Figure 21C**) and a slight reading frame signal of ~41% for position 2 after NNC shift (**Figure 21D**), suggesting that the signal observed for Ribo-seq without RelE was of artificial nature. An identical result was obtained after reanalysis of a Ribo-seq dataset generated with MNase only by Hwang & Buskirk (2017) (**Supplementary Figure 9**).

3.1.4 Detection and analyses of novel intergenic and overlapping genes

For identification of novel, translated ORFs, DeepRibo (Clauwaert *et al.*, 2019), REPARATION (Ndah *et al.*, 2017) and the scripts by Giess *et al.* (2017) were applied to each of the Ribo-seq datasets. In addition, the Ribo-RET dataset was evaluated analogous to the method described by Meydan *et al.* (2019). After evaluation of the results of the single predictions, all results were combined to aid reliable OLG delineation. The identified ORFs were further characterized regarding expression strength, location, type of overlap, differential expression, reading frame periodicity, presence of homologues, length, and choice of start codon.

3.1.4.1 Application of prediction algorithms for individual and collective datasets

When applying the three prediction tools to each of the six Ribo-seq datasets (4× Exp1, 1× Exp2_ND & 1× Exp3_RelE), a marked discrepancy in the number of predictions was observed. DeepRibo predicted on average 2,863 translated ORFs, followed by REPARATION with 8,609 hits. The by far highest number of predictions was obtained for the scripts by Giess *et al.* (2017) with a mean value of 74,902. For the tools DeepRibo and REPARATION, the largest number of predictions was obtained for the Exp2_ND datasets and the lowest number of predictions for the Exp3_RelE dataset (**Supplementary Table 4**). Interestingly, the exact inverse effect was seen for the number of predictions obtained by the scripts published by Giess *et al.* (2017). For the Ribo-RET dataset, ORFs with a TIS peak larger than 1 RPM were predicted to be translated by the algorithm provided by Meydan *et al.* (2019). This analysis resulted in 18,587 predicted candidates.

For prediction performance evaluation, all hits obtained by one prediction tool were compared for all Ribo-seq datasets and the percentage of the overlapping fraction was calculated. As shown for the results by DeepRibo (**Figure 22A**), the reproducibility of the predicted hits in multiple datasets was limited and ranged from 14.0% (Exp2_ND vs. Exp3_RelE) to 60.3% (Exp1_aerobic_RepII vs. Exp1_anaerobic_RepII). When comparing the results of the two biological replicates, medium identity percentages of 42.7% (Exp1_aerobic) and 44.3% (Exp1_anaerobic) were observed. Comparable results were obtained for REPARATION predictions (**Supplementary Figure 10A**), whereas hits predicted by the scripts by Giess *et al.* (2017) showed on average higher identity percentages across different datasets (**Supplementary Figure 10B**), probably due to the higher number of total predictions.

Limited congruence values were also received when estimating the prediction overlap of the three prediction tools on a single Ribo-seq dataset. For Exp2_ND, which showed the highest number of predictions in terms of DeepRibo and REPARATION, for instance, only 2,522 out of in total 69,635 predictions were forecasted by all three tools (**Figure 22B**).

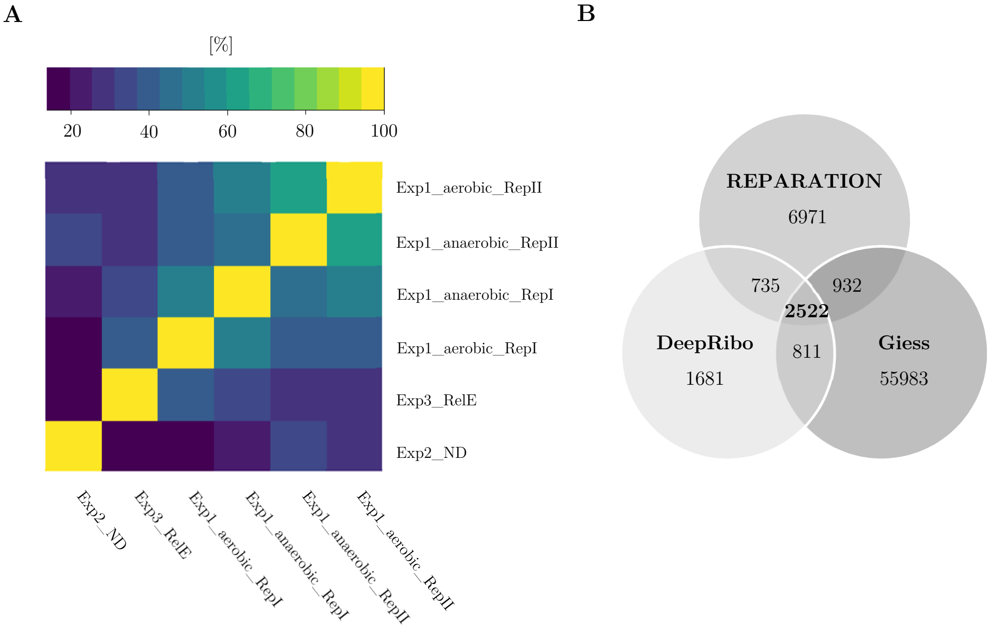


Figure 22. Reproducibility of Ribo-seq prediction results in *E. coli* LF82. Shown are (**A**) the percentages of identical predictions obtained by the tool DeepRibo for all possible dataset combinations as well as (**B**) the overlap and absolute number of predictions obtained by all three prediction tools (DeepRibo, REPARATION, scripts by Giess *et al.* (2017)) for the dataset Exp2_ND.

Since the reproducibility and comparability of the results obtained were limited, all hits predicted by of all tools were combined in order to facilitate reliable identification of novel ORFs. In total, the results of six Ribo-seq datasets (Exp1_aerobic_RepI+II, Exp1_anaerobic_RepI+II, Exp2_ND & Exp3_RelE) were separately analysed with DeepRibo, REPARATION, and the scripts by Giess *et al.* (2017) and afterwards merged with the results obtained for the Ribo-RET dataset (Exp2_RET), which was evaluated according to the method described by Meydan *et al.* (2019). Merging of the results of the 19 different prediction possibilities resulted in 64,877 translated ORF candidates. All predicted ORFs were scored according to how often they have been predicted, resulting in a scoring scale ranging from 1 (ORF predicted by one tool in one dataset) to 19 (ORF predicted by all tools in all datasets). Only hits found in more than a half of all prediction combinations (score > 9) were considered to be valid gene candidates. Hits were classified according to the RefSeq annotation, and ORFs matching to annotated genes were removed. The remaining hits were visually inspected in a genome browser and false positive hits, e.g., due to read cross talk caused by adjacent anORFs, were excluded. One hundred and sixteen high confident hits with scores ranging from 10 to 17 passed the individual inspection. Among them were also the three novel gene candidates whose signals are shown exemplarily in **Figure 23**. Finally, all novel gene candidates were further characterized as described in the following section.

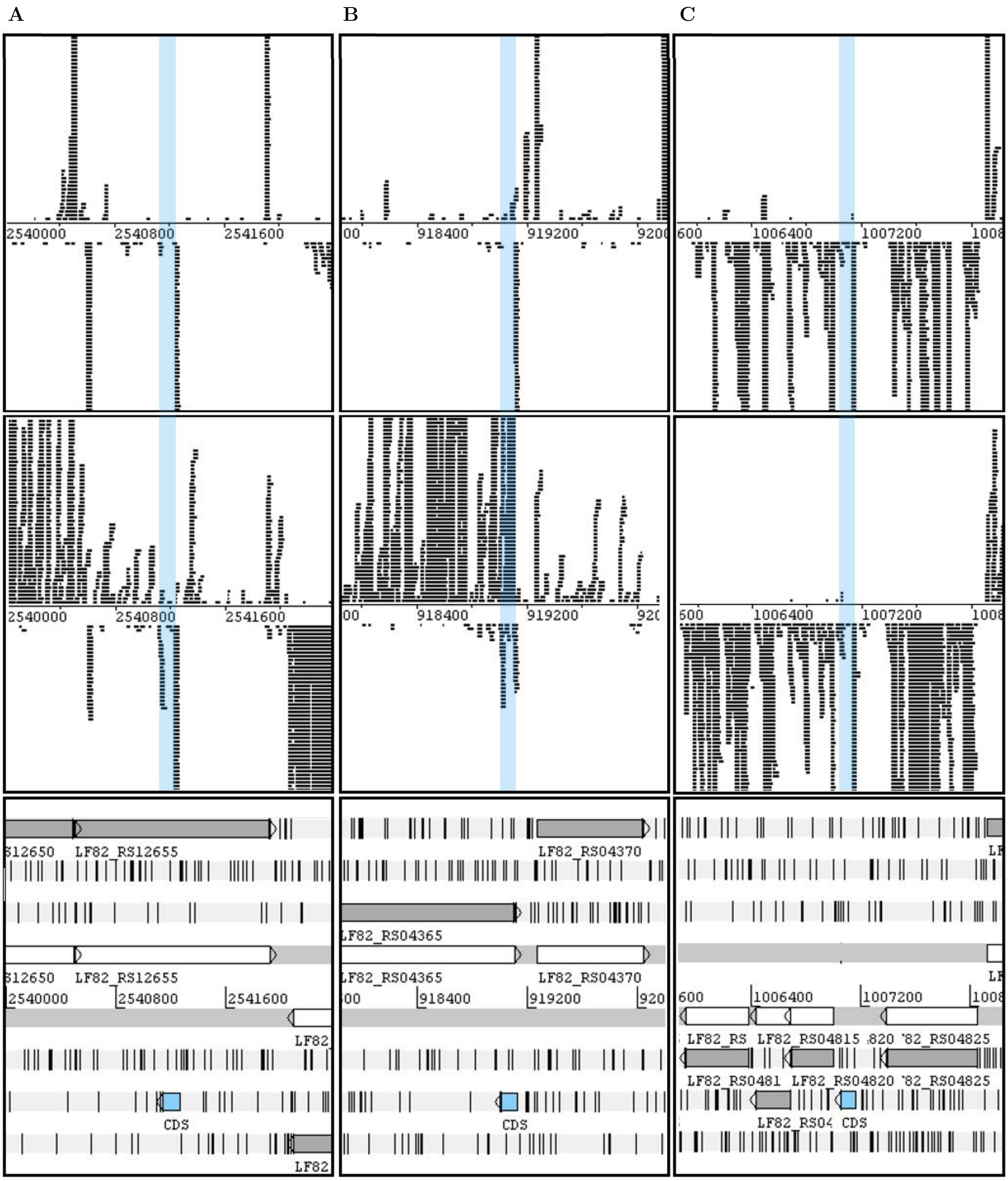


Figure 23. Ribo-seq signals of three novel gene examples identified in *E. coli* LF82. Strand-specific Ribo-seq signals of the RET dataset (Exp2_RET, top row) and the ND dataset (Exp2_ND, middle row) were visualized using the Artemis genome browser (Rutherford *et al.*, 2000). Translatome reads mapping to the forward strand are displayed above the centre line and reads mapping to the antisense strand are plotted below the centre line. The bottom row represent a six-frame translation of the genomic loci coding for the novel overlapping gene candidates (**A**) LF82_71 and (**B**) LF82_14 as well as for the novel intergenic gene candidate (**C**) LF82_18. All ORFs are displayed in their respective frame and anORFs are highlighted in grey. Black bars indicate stop codon positions. The genomic regions coding for the novel ORFs are shaded in blue.

3.1.4.2 Bioinformatic characterization of the selected ORFs

The novel ORFs were distributed across the entire genome of *E. coli* LF82 (**Figure 24A**). Interestingly, the ORFs encoded on the antisense strand (n = 106) were almost evenly spread, whereas ORF occurring on the sense strand (n = 10) were accumulated in a region from ~1 Mbp to 1.2 Mbp. However, this observation might be due to the smaller number of predicted ORFs on the sense strand. Four gene candidates were located within regions predicted to be of phage origin, one encoded on the sense strand, the others on the antisense strand (**Figure 24A**). The ORF length ranged from 93 bp to 669 bp (median value = 129 bp), and thus, the novel ORFs were on average shorter than the annotated, protein-coding genes with a median value of 807 bp (**Figure 24B**). 60.4% of all novel ORFs had an ATG start codon, 22.4% a TTG and 17.2% a GTG start codon. In contrast, the ORFs of annotated genes start predominately with ATG codons (87.7%), followed by GTG and TTG codons (8.5 and 3%, respectively). For 0.7% of all anORFs, a rare start codon (i.e., ATA, ATC, ATT & CTG) was detected.

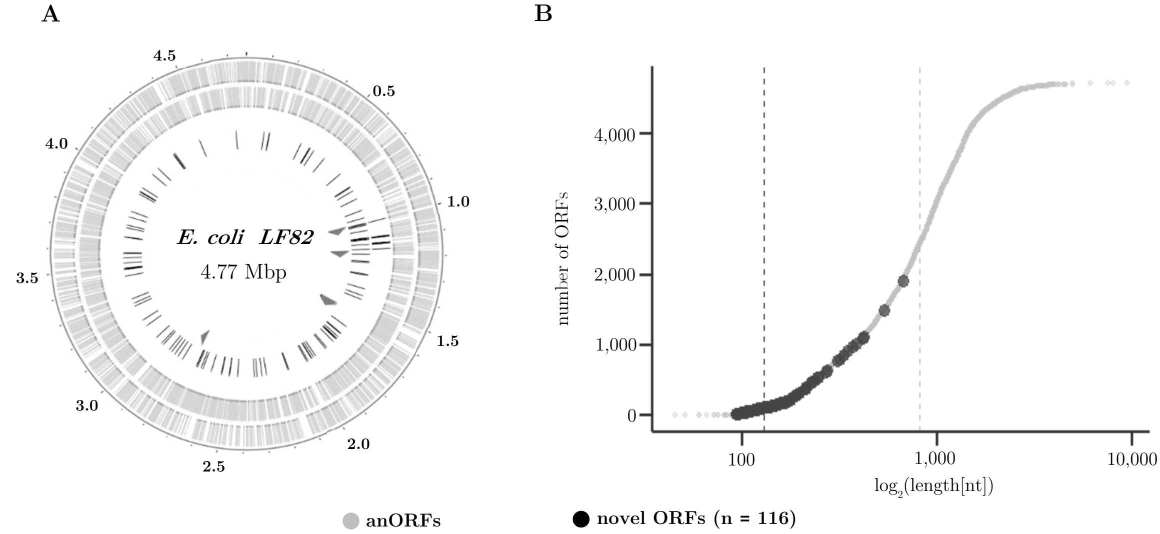


Figure 24. (**A**) Distribution and (**B**) length of novel gene candidates in the *E. coli* LF82 genome. (**A**) The circles show the annotated genes (grey, n = 4,586) located at the sense and the antisense strand, the novel gene candidates (black, n = 116) located at the sense and the antisense strand as well as phage regions as indicated by grey arrows (from outside to inside). (**B**) The length distribution of all novel gene candidates and all annotated genes with their respective median values (dashed line) are indicates by grey (anORFs) or black (novel ORFs) dots.

All novel ORFs were categorised according to their type of overlap, if present (**Table 15**). Twelve ORFs showed no (iORF) or only a trivial overlap (<30 nt, OLG_TL), whereas the remaining ORFs (n = 104) overlapped non-trivially with anORFs. For the latter, the overlap region covered 34.9% to 98.2% of the entire ORF length. In 103 out of 104 cases, the overlapping mother gene was functionally annotated, implying that the novel ORF encoded in another reading frame can be considered as a "real" gene overlap. Unless otherwise stated, all novel ORFs were treated as one group for the subsequent analyses independently of their classification status.

Table 15. Number of novel gene candidates identified in *E. coli* LF82. All hits exceeding a prediction score of 9 are divided into the categories intergenic ORF (iORF), trivial OLG (OLG_TL), antisense embedded OLG (OLG_EA), sense embedded OLG (OLG_ES), partial antisense OLG with overlap at the 3´ (OLG_PA3) or at the 5´ end (OLG_PA5 as well as partial sense OLG with overlap at the 3´ (OLG_PS3) or at the 5´ end (OLG_PS5), respectively. Hits matching to more than one non-trivial overlap type were classified as "multiple types".

ORF_type	iORF	OLG_TL	OLG_EA	OLG_ES	OLG_PA3	OLG_PA5	OLG_PS3	OLG_PS3	Multiple types
Number	7	5	33	50	4	4	5	5	3

All novel ORFs were analysed regarding the presence of structural features necessary for gene expression. Three hundred base pairs of the upstream sequence of all candidates were subjected to promoter analysis using BPROM (Solovyev & Salamov, 2011). For all ORF except one, a σ^{70}-dependent promoter was predicted in an average distance of 131 bp (**Supplementary Figure 11A**). The LDF value, which is a measure for the prediction accuracy and specificity, ranged from 0.34 to 9.13 (median = 2.42), and thus, clearly exceeded the minimum threshold value of 0.2. Termination of transcription at ρ-independent terminator structures was analysed in a 300 bp region downstream of the stop codon using FindTerm (Solovyev & Salamov, 2011). Forty-four of all novel gene candidates harboured a putative terminator in a distance ranging from 2 to 272 bp (**Supplementary Figure 11B**). A SD sequence was predicted for 59.5% of all novel ORFs according to the method described by Hyatt *et al.* (2010). However, a clear SD consensus pattern as obtained for protein-coding anORFs was not observed (**Supplementary Figure 12**) since the SD sequence motif and the distance from the start codon were highly variable (**Supplementary File S1**).

Expression of the novel gene candidates was evaluated based on the results obtained for RNA-seq and Ribo-seq in two biological replicates (Exp1_aerobic_RepI+II). RPKM values were used to estimate transcriptional and translation strength. As seen in **Figure 25A**, RPKM values of the translatome (median = 19.5) of the novel ORFs were comparable to those of all protein-coding anORFs (median = 21.7) indicating transcriptional activity. RPKM values of the translatome ranged from 0.38 to 134,149 for the novel ORFs (median = 19.2) and were on average slightly larger than those of the anORFs (median = 14.8; **Figure 25B**). Read coverage values of the novel ORFs (median = 0.53) were also in a similar scale as the median coverage values of anORFs (0.50; **Figure 25C**). Translatability of the novel gene candidates as indicated by the RCV was marginally lower for the novel gene candidates (median = 0.68) compared to those of the anORFs (median = 0.8; **Figure 25D**). No significant differences were obtained when dividing the novel candidates in intergenic and overlapping ORFs for expression analysis (results not shown). The expression values obtained for the anaerobic condition essentially confirmed similar transcriptional and translational signals for both the novel ORFs and the anORFs (results not shown).

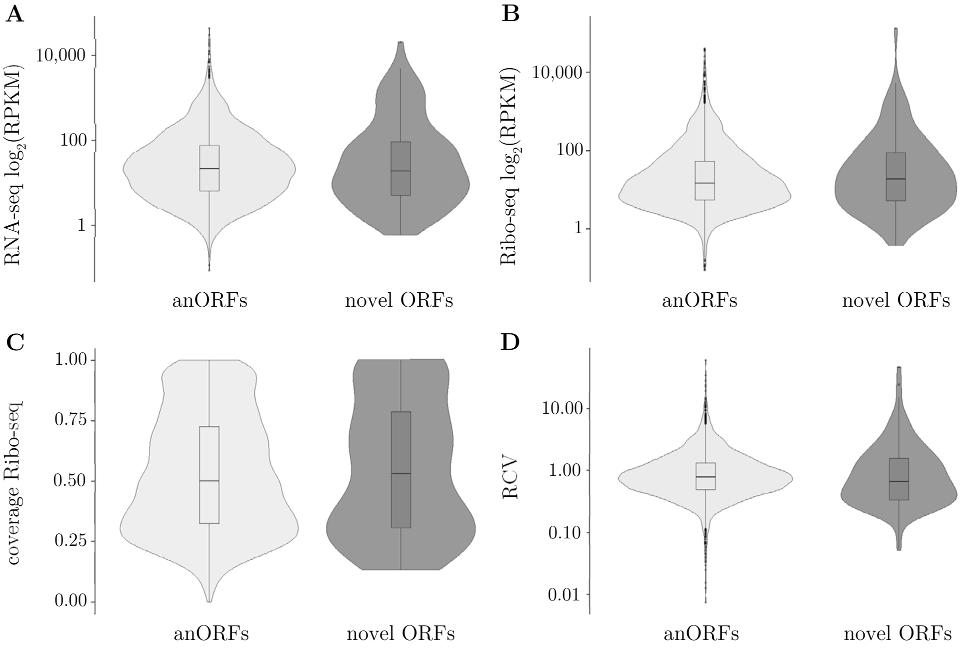


Figure 25. Expression metrics of the novel gene candidates (n = 116) in comparison to protein-coding, annotated genes (n = 4,586) in *E. coli* LF82. Violin plots displaying the mean reads per million mapped reads (RPKM) values of (**A**) RNA-seq and (**B**) Ribo-seq, the (**C**) mean read coverage of ORFs in Ribo-seq and (**D**) the mean ribosome coverage value (RCV) of two biological replicates (Exp1_aerobic_RepI+II) are shown.

83.6% of all novel gene candidates, including the candidates LF82_14 and LF82_71 (**Figure 23**, top row), showed an accumulation of reads equivalent to more than 1 RPM at the start codon in the Ribo-RET experiment, and thus, were successfully predicted by the applied algorithm. Metagene analysis (**Figure 26A**) confirmed the redistribution of the ribosomes at the putative start position by a factor of five compared to the ND experiment. However, in comparison to the metagene plot derived from highly expressed anORFs (**Figure 15B**), the peak was slightly less pronounced and shifted by one nucleotide at the start codon region. Nevertheless, the median height of the RET peak of the novel ORFs was even slightly higher than those of all anORFs (**Figure 26B**).

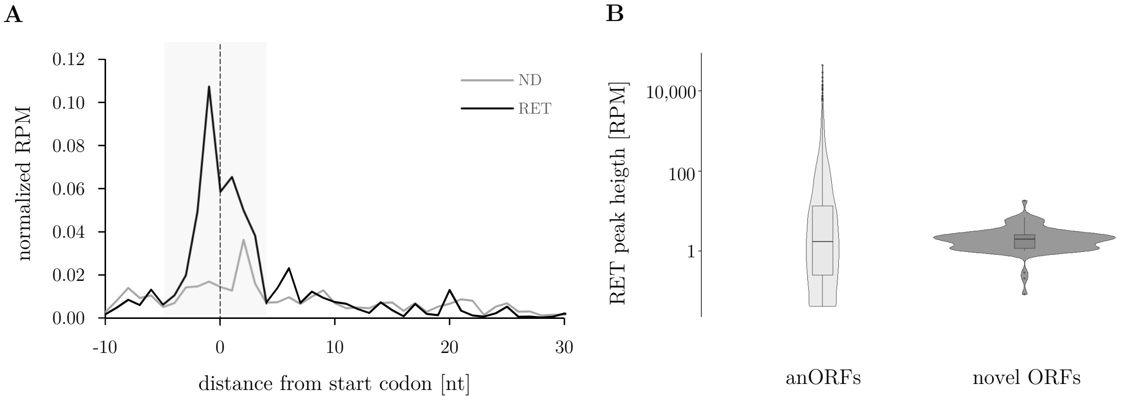


Figure 26. Translation initiation site analysis of novel gene candidates in *E. coli* LF82. (**A**) The metagene plot shows the normalized RPM values of each position in a -10 to 30 nt window around the start codon (dashed line, distance 0 nt) after determining the P site position of reads mapping to the novel gene candidates (n = 116) in the no drug (ND) and retapamulin (RET) Ribo-seq experiment. (**B**) Violin plots display the RET peak height measured in RPM for the novel gene candidates in comparison to protein-coding, annotated genes (n = 4,586).

RF analysis of the novel ORFs revealed a clear periodicity signal at the third sub-codon position. With a percentage of 59%, the read density at this position was similar to the RF signal obtained for anORFs (57%; **Figure 21B**) after NNC shifting. However, instead of an accumulation of reads at the second sub-codon position as expected based on RelE´s cleavage characteristics, the highest reads density was obtained for sub-codon position 3 (**Figure 27A**). This observation suggested the occurrence of translation of an additional overlapping frame located in -1 relative to the novel ORFs, e.g., those of a sense overlapping mother gene. Indeed, the majority of all novel gene candidates showed such a sense overlap with annotated genes. Considering sense overlaps may result in a shift in the periodicity signal. To test for this hypothesis, sense overlapping ORFs were separated from the remaining ORFs, and both groups were analysed individually. For the first, a distribution of 23%, 12% and 65% was obtained (**Figure 27B**), which was quite similar to the signal obtained for all novel ORFs regardless of their overlapping status (**Figure 27A**). Since the RF analysis is based on absolute read numbers and not on normalized read numbers, ORFs with low read counts contribute less to the sum signal than ones with high read counts. Some of the novel candidates indeed overlapped with highly expressed mother genes e.g., ribosomal proteins or flagellar proteins (**Supplementary File S1**), thereby offering an explanation for the high resolution of the periodicity signal obtained. As a result, the reading frame signal of a weakly expressed novel ORF located in frame +2 might be completely concealed by the signal of a sense overlapping anORF (e.g., encoded in frame +3; **Figure 27B**) due to higher read counts mapping to frame +3 instead of to frame +2. In contrast, intergenic ORFs as well as antisense overlapping ORFs showed a clear preference for sub-codon position 2 (67%; **Figure 27C**), which provided evidence for their genuine translation. RNA-seq control analyses confirmed either an entire lack or a reduction of the periodicity signal of the novel ORFs in the absence of translation (**Figure 27**).

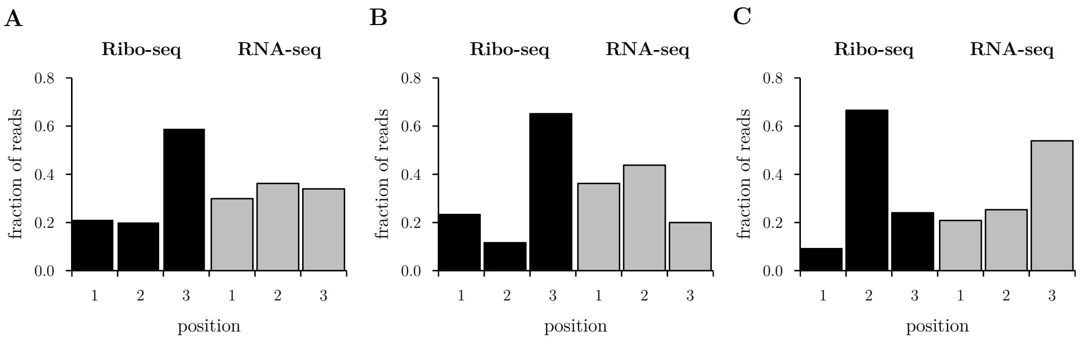


Figure 27. Reading frame sum signal of all novel gene candidates in *E. coli* LF82. Shown are the results obtained for (**A**) all novel candidates (n = 116), (**B**) for all ORFs overlapping sense with anORFs (n = 62) and (**C**) all ORFs without the sense overlapping ORFs (n = 54) of all reads in the Ribo-seq and RNA-seq datasets of Exp3 after shifting of NNC reads.

Evidence for the functionality of some of the novel ORF candidates was provided by differential gene expression analysis. In total, six of the novel ORFs were differentially regulated in the presence (Exp1_aerobic_RepI+II) and absence of oxygen (Exp1_anaerobic_RepI+II); two at the transcriptional level and four at the translational level. Both genes regulated at the transcriptional level showed increased expression under anaerobic conditions, whereas two of the ORFs found to be regulated at the translatome level were upregulated and the remaining two were downregulated depending on the cultivation conditions (**Supplementary File S1**). Further hints for a potential functionality of the novel ORFs was delivered by blastp analysis. For 70 out of 116 ORFs, homologous proteins with a minimum identity percentage of 70% and an e-value $\leq 1\times10^{-3}$ were detected when searching in the non-redundant protein database nr. Application of a more stringent e-value of $\leq 1\times10^{-10}$ aided in the detection of homologous proteins for 62 novel gene candidates. Approximately half of the detected homologues were either classified as "hypothetical", "uncharacterized" or "putative"; for the remaining a functional description was available. When searching against the RefSeq Select proteins database, high confident hits were obtained for two intergenic ORFs. All relevant details on the novel gene candidates including their blastp results are listed in **Supplementary File S1**.

3.2 Identification of protein-coding genes in *P. aeruginosa* PAO1

In addition to *E. coli* LF82, *P. aeruginosa* PAO1 was also subjected to novel ORF analysis. *P. aeruginosa* PAO1 has a high GC genome with a size of 6,264,404 bp and encodes for 5,572 anORFs according to RefSeq (GCF_000006765.1). Remarkably, 40.5% of all anORFs are annotated as hypothetical and lack functional characterization. Besides the mentioned anORFs, the genome of *P. aeruginosa* PAO1 also harbours 57,752 additional ORFs exceeding a size of 93 bp. Among these, 45,068 overlap either trivially (<30 nt; 1,657) or non-trivially (≥30 nt; 43,411) while showing different types of overlaps (**Table 16**). Further 8,785 ORFs lack a clear allocation and belong to multiple categories; the remaining 3,899 ORFs are located in intergenic regions.

Table 16. Number and classification of non-trivial overlapping genes (OLGs) predicted for the *P. aeruginosa* PAO1 genome. All ORFs exceeding 93 nt and overlapping more than 30 nt with an annotated gene were determined and classified according to the type of overlap and their strand location.

strand	overlap type	designation	number
sense	embedded	OLG_ES	12,977
	partially at 5`end	OLG_PS5	2,014
	partially at 3`end	OLG_PS3	1,846
antisense	embedded	OLG_EA	19,904
	partially at 5`end	OLG_PA5	3,561
	partially at 3`end	OLG_PA3	3,109

For the analysis of the protein-coding capacity of these alternative ORFs, eleven NGS datasets and one MS dataset were generated, which were:

- **Experiment 1 (Exp1) – four datasets:** Ribo-seq and RNA-seq at optimal growth conditions (37 °C, LB, aerobic) in two biological replicates (section 3.2.1). Optimal experimental conditions were evaluated in pretests prior to implementation of the main experiments.
- **Mass spectrometry experiment (Exp1_MS) – one dataset:** Proteomic analysis using DDA-MS with ultra-deep sample fractionation to gain a global snapshot of the translated ORF products (section 3.2.2). The analysed sample matched the sample of Exp1.
- **Experiment 2 (Exp2) – two datasets:** Ribo-RET and the respective Ribo-ND dataset were implemented in a *P. aeruginosa* sextuple mutant to aid TIS identification (section 3.2.3).
- **Experiment 3 (Exp3) – two datasets:** *E. coli*-originating RelE was used for reading frame analysis in RelE-supported Ribo-seq using RelE and MNase for nuclease footprinting. A Ribo-seq experiment with MNase only served as an internal control for RelE functionality in *P. aeruginosa* (section 3.2.4).
- **Experiment 4 (Exp4) – three datasets:** Transcriptome samples matching to Exp1 were analysed using Cappable-seq in order to capture 5´ ends of primary transcripts thereby facilitating the global identification of TSS (section 3.2.5). This experiment was performed in three biological replicates.

Analogous to *E. coli* LF82, all experiments conducted in *P. aeruginosa* were analysed separately with regard to both annotated as well as unannotated ORFs and evaluated using different prediction tools. Finally, all prediction results were combined and high confident OLGs and iORFs were selected for further bioinformatical characterization. Two exceptionally long ORFs and their respective mother genes identified in *P. aeruginosa* PAO1 were analysed experimentally and bioinformatically in more detail.

3.2.1 Conventional RNA-seq & Ribo-seq

The aim of this part was to analyse the transcriptional and translation landscape of *P. aeruginosa* using conventional RNA-seq and Ribo-seq. For this purpose, the Ribo-seq procedure had to be adapted and optimized for this organism. After determining the optimal procedure, RNA-seq and Ribo-seq were conducted in two biological replicates and sequencing results were evaluated with respect to data quality, read length and reproducibility.

3.2.1.1 Adaptation and optimization of the Ribo-seq protocol

A temperature of 37 °C and LB broth were chosen for *P. aeruginosa* PAO1 in order to ensure optimal growth for Ribo-seq experiments. Samples were harvested after 180 min of cultivation at the transition from exponential to stationary phase ($OD_{600nm} = 1$) and after 15 h of cultivation in late stationary phase ($OD_{600nm} = 6$). In a first approach, Ribo-seq of these samples was conducted following the protocol described for *E. coli* LF82 (Exp1) using five nucleases for footprinting. However, sequencing resulted in mRNA yields of 5.8% and 1.9% and rRNA yields of 71.3% and 78% for the sample harvested at $OD_{600nm} = 1$ and $OD_{600nm} = 6$, respectively. The reduced mRNA content in the late stationary sample can most likely be traced back to the impairment of the RNA integrity, which was already visible after RNA extraction (**Supplementary Figure 13A**). For this reason, late stationary phase samples were deemed not appropriate for Ribo-seq experiments. Although RNA quality of the $OD_{600nm} = 1$ sample was by far higher than those of the $OD_{600nm} = 6$ sample (**Supplementary Figure 13A**), the low mRNA percentage also indicates sub-optimal reaction conditions. In order to increase mRNA yield by decreasing rRNA reads, all nucleases used in the experiment were evaluated regarding their effect on rRNA integrity. Pretests suggested that RNase I -even when present in small amounts- leads to massive RNA degradation, and thus, is unsuitable for Ribo-seq in *P. aeruginosa* PAO1 (**Supplementary Figure 13B**). By omitting RNase I from the nuclease mixture as well as by reducing the concentration of all other nucleases by 30%, higher yields of mRNA (i.e., effective reads) were obtained (**Table 17**).

3.2.1.2 Sequencing output and data reproducibility

Two biological-independent RNA-seq and Ribo-seq experiments were conducted according to the optimized Ribo-seq protocol for *P. aeruginosa* PAO1. After sequencing, read numbers of 61.4 to 220.3 mio were obtained. Although the percentage of reads which mapped to genome was higher for *P. aeruginosa* PAO1 compared to *E. coli* LF82 (**Table 12**), the yield of mRNA-mapping reads was clearly reduced. Nevertheless, 14.9 to 35.2 mio and 3.5 to 11.0 mio effective reads were obtained for both Ribo-seq and RNA-seq experiments, respectively (**Table 17**). Analogous to *E. coli* LF82, the percentage of mRNA reads was significantly higher for Ribo-seq than for RNA-seq experiments. The vast majority of all read mapping to mRNA regions in the Ribo-seq datasets had a length of 21 to 26 nt suggesting that this size corresponds to the ribosomal footprint in *P. aeruginosa* (**Figure 28**).

Table 17. Overview of Ribo-seq and RNA-seq reads obtained for Exp1 in *P. aeruginosa* PAO1 after sequencing. Shown are the total number of sequenced reads as well as the number of reads mapping to the *P. aeruginosa* PAO1 genome in millions for two biological replicates. Percentages in brackets indicate the proportion of reads mapping to rRNA and tRNA as well as mRNA regions of the *P. aeruginosa* PAO1 genome.

Experiment	Replicate	Total	Mapped	rRNA & tRNA		mRNA	
Ribo-seq	I	133.6	112.2	97.3	(86.7%)	14.9	(13.3%)
	II	191.8	163.5	128.3	(78.5%)	35.2	(21.5%)
RNA-seq	I	61.4	46.0	42.5	(92.4%)	3.5	(7.6%)
	II	220.3	179.9	168.8	(93.9%)	11.0	(6.1%)

When calculating RPKM values of all anORFs for the two biological replicates, Pearson correlation coefficients r of 0.99 and 0.81 were obtained for RNA-seq and Ribo-seq (see **Figure 29**), respectively. This observation points to a very strong linear relationship between the biological replicates suggesting excellent reproducibility. Experimental reproducibility was even better than observed for *E. coli* LF82.

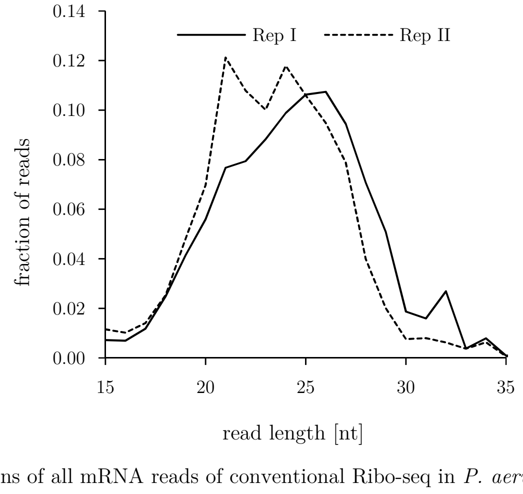


Figure 28. Read length distributions of all mRNA reads of conventional Ribo-seq in *P. aeruginosa* PAO1. The proportion of reads mapping to mRNA regions is displayed for the first (RepI) and second (RepII) replicate of the experiment Exp1.

3.2.2 Proteomic verification of anORF translation by mass spectrometry

In addition to Ribo-seq, a sample harvested at $OD_{600nm} = 1$ was subjected to DDA-MS to identify the global proteome at this time point. In order to increase sensitivity of detection, the sample was separated into 48 fractions prior to MS measurement. Proteins encoded by anORFs were identified by searching the measured spectra against all entries of a RefSeq-derived database. After deleting all peptides, which were of low quality or were assigned to be contaminants or reverse hits, 54,947 peptides mapping to anORF-encoded proteins remained. Proteins with more than two unique mapping peptides were considered to be successfully detected. Overall, 3,992 out of 5,572 proteins (71.6%) were detected using MS with a total number 54,884 peptides. 1,463 of all identified proteins were annotated as 'hypothetical' or 'uncharacterized' constituting 64.9% of all known 'hypothetical' anORFs. This result confirms the genuine protein-coding nature of this type of anORFs.

When comparing translatome and proteome data, a correlation between the expression strength as indicated by RPKM values and the success of MS detection was observable. Genes encoding for proteins which were detected by MS showed on average higher RPKM values than genes without MS signal (**Figure 29**). Significance of this observation was confirmed by a Wilcoxon test with a p-value of 2.2×10^{-16}. A moderate linear relationship of 0.45 according to Pearson was obatined when correlating RPKM values with MS intensity. Correlation strength was increased to 0.69 when calculating the Spearman's rank correlation coefficient.

In addition to the RefSeq reference protein sequence file, MS2 spectra were additionally searched against a six-frame translation of the *P. aeruginosa* PAO1 genome to enable the identification of novel protein-coding regions. All peptides obtained for this analysis were combined with the results of the Ribo-seq predictions as described in section 3.2.6.

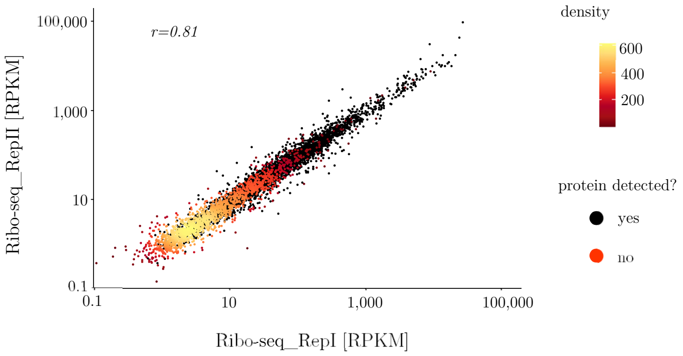


Figure 29. Comparison of translatome and proteome data in *P. aeruginosa* PAO1. Shown are the RPKM values obtained for all annotated genes (n = 5,572) of two biological Ribo-seq experiments with their Pearson correlation coefficient *r*. Genes encoding for proteins which were detected by MS are displayed in black; genes lacking MS detection of their protein products are shaded according to their density (yellow ≙ high density; dark red ≙ low density).

3.2.3 Ribo-RET for translation initiation site detection

Ribo-RET (i.e., using retapamulin) was applied to *P. aeruginosa* in order to detect TISs in this species. For a proper RET effect, the gene *oprM* encoding the homologue to the TolC outer membrane factor of the $AcrA^{MFP}$-$AcrB^{RND}$-$TolC^{OMF}$ efflux pump had to be deleted. Due to the high number of intrinsic efflux systems, a deletion mutant strain, *P. aeruginosa* PAO397, which lacks multiple genes involved in antibiotic efflux was used for this experiment. The MIC of both strains was determined and RET was added in 100-fold excess during the Ribo-seq cultivation.

3.2.3.1 Determination of the minimum inhibitory concentration of retapamulin

The wild type strain PAO1 and its deletion mutant strain PAO397 were cultivated in LB as described for *E. coli* LF82 (section 3.1.2.2). One hundred microliters of a cell suspension with 1.35×10^6 CFU/mL (±standard deviation 0.9×10^4) or 5.1×10^6 CFU/mL (±standard deviation 3.7×10^6) of strain PAO1 and PAO397 were 1:2 diluted with LB containing varying concentrations of RET, respectively. The final inoculum concentration was approximately 6.8×10^5 CFU/mL for the wild type and 2.6×10^6 CFU/mL for the deletion mutant strain.

Analogous to *E. coli* LF82, the wild type *P. aeruginosa* strain showed a MIC ≥32 µg/mL, which was the highest RET concentration used in this experiment. In contrast, strain PAO397 showed a reduced MIC value of 0.5 µg/mL after 24 h of cultivation (**Figure 30**). This value was twice as high as the respective MIC value obtained for *E. coli* LF82Δ*tolC*. Again, slight fluctuation were observable throughout the cultivation process (results not shown), but all three independent replicates confirmed the final MIC values.

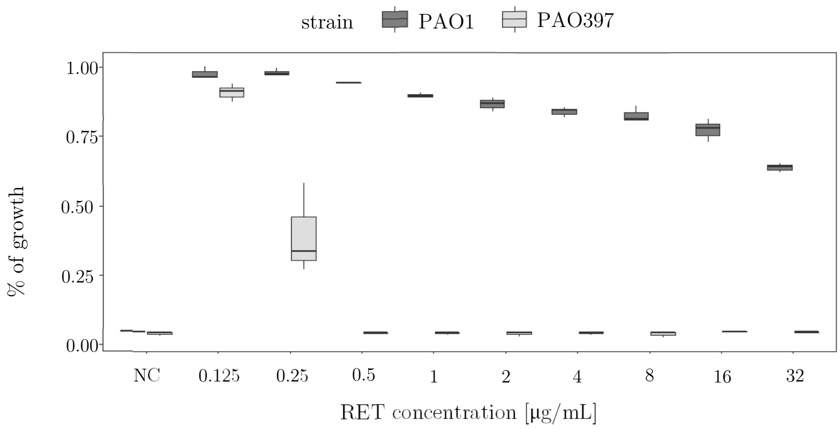


Figure 30. Results of minimum inhibitory concentration (MIC) testing of retapamulin (RET) for *P. aeruginosa* PAO1 and *P. aeruginosa* PAO397. The overall growth of both strains in the presence of different RET concentrations ranging from 0.125 to 32 µg/mL was calculated after 24 h by dividing the measured optical density by the respective value of bacterial cultures without RET (positive control). Plain broth without cell inoculum was used as a negative control (NC).

3.2.3.2 Cut-off specification for genome wide TISs analysis

Ribo-seq of strain PAO397 after treatment with 100× MIC of RET when reaching an $OD_{600nm} = 1$ resulted in 17.3 mio (13.8%) and 29.2 mio (18.1%) mRNA reads for Ribo-RET and the ND control. Yields were in a similar order of magnitude as those obtained for the first experiment, whereby reads were on average larger (33-36 nt; **Supplementary Figure 14A**) than reads of the conventional Ribo-seq.

For RET data evaluation, P site mapping with an offset of 15 nt was performed and RPM and RPKM values were calculated. In contrast to *E. coli* LF82, subtraction of 15 nt from the 3´end was sufficient for P site mapping, resulting in a peak of normalized reads at the expected position (i.e., 0 nt distance from start) in the metagene analysis of 184 highly expressed anORFs (**Figure 31A**). Unexpectedly, a second, distinct peak was observable at position +2 suggesting slight inaccuracies of P site mapping. However, since this peak was located within the ±3 nt window, which is used for ORF prediction, an adjustment was not performed. Overall, RET treatment resulted in clear accumulation of reads at start codons (distance to start codon 0 nt) as observed by the eightfold higher normalized reads counts in the RET dataset compared to the ND experiment. This effect was even more pronounced in *P. aeruginosa* PAO397 than in *E. coli* LF82 (see **Figure 15B**). Ribosome redistribution caused by RET in *P. aeruginosa* PAO397 was also confirmed on single gene level, for instance, when analysing genes encoded by the S10 ribosomal protein operon (**Supplementary Figure 15**). These results suggest successful implementation of Ribo-RET in *P. aeruginosa* PAO397. Although the overall correlation between normalized read counts at start codons in the RET dataset and the RPKM values in the ND dataset was limited (Pearson´s *r*: 0.61), highly significant results were obtained when comparing RET peak heights of highly expressed anORFs (RPKM ≥ 100) with those of medium to low expressed anORFs (RPKM < 100; **Figure 31B**).

In order to determine an appropriate value for the detection of translated ORFs, the algorithm published by Meydan *et al.* (2019) was applied using variable threshold ranging from 0.2 to 5. Pairwise fold change calculations suggested an optimal value of about 0.8 (**Supplementary Figure 16**). A further rise of the threshold did not result in a severe drop in the number of predictions. However, application of this threshold yielded the detection of 73% of all highly expressed anORFs with RPKM values larger than 100. The observed number was substantially lower than the number of predictions obtained for *E. coli* LF82 (88%) after applying a threshold value of 1. To increase the number of detectable ORFs in *P. aeruginosa*, a threshold of 0.63 was used for the final prediction, thereby enabling the identification of 33% of all anORFs as well as 76% of all highly expressed anORFs (RPKM ≥ 100).

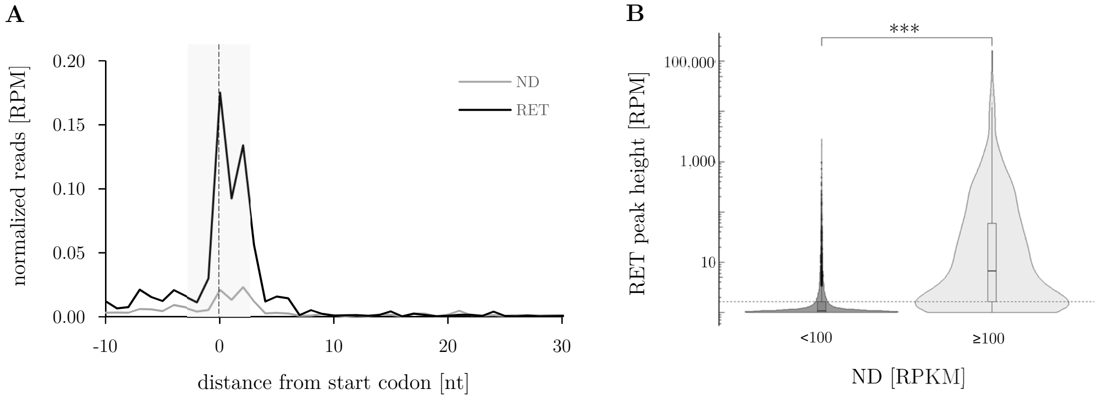


Figure 31. (**A**) Metagene and (**B**) threshold analysis of all anORFs (n = 5,572) in *P. aeruginosa* PAO397. (**A**) Shown are normalized RPM values of each position in a -10 to 30 nt window around the start codon (dashed line, distance 0 nt) after determining the P site position by subtracting 15 nt of all genes with RPM values ≥100 in the ND and RET datasets (n = 184). (**B**) RET peak heights at start codons are displayed as a function of expression strength in the no drug (ND) control Ribo-seq dataset. A significant difference (Wilcoxon test; p-value < 0.001) was obtained for anORFs with RPKM values below 100 (n = 4,715) and equal or larger than 100 (n = 857). The threshold used for ORF prediction (≥0.63) is indicated by the dotted line.

3.2.4 RelE-enhanced Ribo-seq to visualize triplet periodicity

E. coli-originating RelE was also used for RelE-supported Ribo-seq in *P. aeruginosa*. To ensure sufficient cleavage activity in this strain, the amount of RelE was increased by factor two and a half compared to the experiment in *E. coli* LF82. Footprint lengths of 10-40 nt were isolated after nuclease digestion and subjected to Illumina sequencing. However, low similarity and identity percentages of 17.2% and 9.7% between the *E. coli*-derived RelE and the corresponding homologue of *P. aeruginosa* necessitated a careful evaluation of the potential functionality of the foreign toxin in the target species as discussed in section 3.2.4.1. For this purpose, an additional Ribo-seq experiment without RelE was conducted as a negative control.

3.2.4.1 Evaluation of RelE functionality in PAO1

Sequencing reads were processed and mapped as described for *E. coli* LF82 resulting in 40.7 and 52.1 mio mRNA reads in the Ribo-seq experiment with and without RelE, respectively. Regarding read length, both experiments showed similar results with most reads having a size of 26 to 27 nt (**Supplementary Figure 14B**). However, a slight increase in the number of short reads (15 to 21 nt) in the RelE-supported dataset could point towards cleavage by RelE. In order to gain further information about RelE activity in *P. aeruginosa* PAO1, the sequence bias at the 3´ ends of all RFPs was analysed. For the dataset generated with MNase only, a clear sequence preference was observable (**Figure 32A**). The majority of reads ended with the nucleobase C or G at position -1, whereas downstream of the cleavage site (position +1) nucleobases A and T were enriched. This observation is in concordance with the well-known sequence specificity of MNase, cleaving predominantly before A and T nucleobases (Dingwall *et al.*, 1981). Hwang & Buskirk (2017) also obtained a similar result upon analysis of a MNase-only Ribo-seq experiment. Reads of the RelE-supported Ribo-seq dataset, in contrast, showed a completely different sequence composition (**Figure 32B**) indicating a substantial difference to the negative control thereby suggesting that the 3´end cleavage was affected by RelE. However, a clear bias for nucleobase C at position -1 and nucleobase G at position +1 as observed by Hwang & Buskirk (2017) was neither apparent for *P. aeruginosa* PAO1 (**Figure 32B**) nor for *E. coli* LF82 (**Supplementary Figure 17**). Separation of reads according to their length and repetition of this analysis did not unravel a sequence-specific effect as a function of length (results not shown).

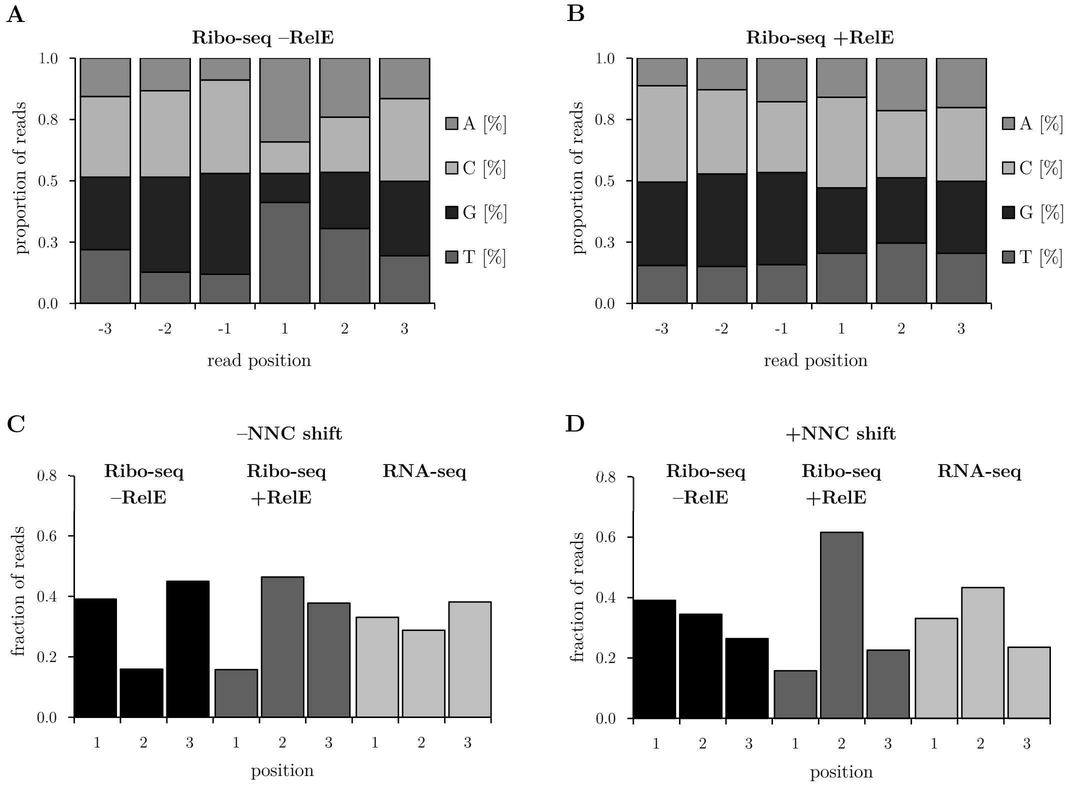


Figure 32. Results of (**A, B**) RelE sequence bias and (**C, D**) periodicity analysis in *P. aeruginosa* PAO1. Figure **A** and **B** pictures the sequence bias at the 3´ end of all mRNA reads obtained in the Ribo-seq experiment after digestion (**A**) with MNase only or (**B**) with a combination of MNase and RelE. Reading frame sum signal of all annotated genes (n = 5,572) are shown for the Ribo-seq experiment with (+RelE) and without RelE (-RelE) as well as for an independent RNA-seq dataset (Exp1_RepII). Subtraction of 30 nt from start and stop positions was performed. Figure **C** displays the results of the raw data; Figure **D** shows the reading frame signal after shifting of reads arising from codons ending in C.

For reading frame analysis, 3´ends of mRNA reads were mapped to the first, second or third sub-codon position of all anORFs after subtraction of 30 nt at the gene boundaries. After counting all reads at each position, 46% of all read end mapped to the second sub-codon position, 38% to the third and 16% to the first position (**Figure 32C**). When performing an NNC shift, a pronounced reading frame signal of 62% to 23% to 16% was obtained (**Figure 32D**). In contrast, the control experiment without RelE did not show an accumulation of reads at the second sub-codon position, neither in the raw data nor in the shifted data. This result suggests that the difference observed in the periodicity signals were caused by RelE and that the toxin is able to cleavage mRNA in *P. aeruginosa*. A second control analysis using an independent RNA-seq dataset of Exp1_RepII exhibited a uniform read distribution in the raw data, and shifting of NNC reads did not lead to a substantial improvement of the reading frame signal.

3.2.5 Cappable-seq for transcription start site (TSS) identification

Cappable-seq was performed in biological triplicates to facilitate the detection of transcription start site. In a first step, the data quality and reproducibility of the sequencing output of three replicates was compared. Afterwards, optimal parameters for global TSS identification including threshold values as well as the length of the region to be analysed were determined based on the results obtained for all protein-coding anORFs in *P. aeruginosa* PAO1.

3.2.5.1 Sequencing output and reproducibility of biological replicates

RNA of samples harvested at $OD_{600nm} = 1$ was isolated and handed over to vertis Biotechnology AG which performed the experimental Cappable-seq procedure. Sequencing resulted in at least 9.8 mio reads per sample (**Table 18**). The proportion of reads mapping to rRNA and tRNA regions was up to a maximum of 7.5% of all mappable reads, indicating successful depletion of rRNA and tRNA reads due to their monophosphorylated 5`ends. In return, more than 92.5% of all reads mapped to mRNA regions thereby confirming efficient enrichment of 5` triphosphorylated mRNA reads without requiring further rRNA depletion.

Table 18. Overview of Cappable-seq reads after sequencing. The total number of sequenced reads as well as the number of reads mapping to the *P. aeruginosa* PAO1 genome in millions are shown for three biological replicates. Percentages in brackets indicate the proportion of reads mapping to rRNA and tRNA as well as mRNA regions of the *P. aeruginosa* PAO1 genome.

Replicate	Total	Mapped	rRNA & tRNA		mRNA	
I	10.8	10.6	0.8	(7.5%)	9.8	(92.5%)
II	9.8	9.6	0.7	(7.3%)	8.9	(92.7%)
III	10.4	10.2	0.6	(5.9%)	9.6	(94.1%)

The reproducibility of the sequencing data of biological replicates was investigated by calculating relative read scores (RRS) per genome position. For this purpose, the script provided by Ettwiller *et al.* (2016) was applied to trim reads to their 5´ ends, thereby yielding the outermost base of each sequencing read. All genome positions covered by at least one read (RRS > 0) in all biological replicates were used for calculation of pairwise Pearson correlation coefficients *r*. For all combination, very strong linear relationships with *r* ≥0.998 were obtained (**Table 19**). This result confirmed excellent biological reproducibility although cultivation, RNA isolation and further processing of the RNA for Cappable-seq were conducted independently.

Table 19. Calculation of Pearson correlation coefficients *r* for Cappable-seq experiments in *P. aeruginosa* PAO1. Genomic positions with a relative read score exceeding zero in three biological replicates (I-III) were used for the calculation of Pearson´s *r*.

Replicates	I + II	I + III	II + III
Pearson´s *r*	0.998	0.998	1.000

3.2.5.2 Cut-off analysis for global detection of TSS

The script for RRS calculation determines all possible TSS at each position of the genome which exceed a certain minimum RRS. This value is a variable, user-defined threshold and should be adjusted to the respective dataset. In order to determine the optimal threshold for global TSS prediction, multiple values ranging from 0 to 20 were specified and the number of predicted TSS of all three replicates was evaluated (**Figure 33A**). An inverse relationship between the RRS and the number of predictions was observable: the higher the minimum RRS, the lower the number of obtained TSS. When calculating fold changes between the number of TSS predicted by using two consecutive threshold values, a flattening with increasing RRS was observable. While thresholds of 0 and 0.5 amounted, for instance, to a 2.6-fold difference in the number of predictions, a further increase of the RRS from 0.5 to 1 or from 1 to 1.5 resulted only in a 1.5-fold and 1.3-fold reduction of predicted TSS. A RRS value of 1.5 was chosen for the subsequent analyses because above this threshold no significant reduction in the number of predictions was obtained. The same RRS value was also chosen by Ettwiller *et al.* (2016) and Zehentner (2020) for global TSS analysis in *E. coli* K-12 MG1655 and *E. coli* O157:H7 EDL933, respectively. Application of this threshold in *P. aeruginosa* PAO1 Cappable-seq data resulted in 11,101 TSS for replicate 1, 11,739 TSS for replicate 2 and 12,230 TSS for replicate 3; 9,205 reliable TSS exceeding a RRS of 1.5 were found in all three replicates.

Since the TSS marks the start of the 5´UTR and is located upstream of a gene´s coding region, a further analysis of the distance between the TSS and start codon was necessary for a reliable assignment. For this purpose, all anORFs were analysed regarding the presence of a TSS with an RRS ≥1.5 in a 500 bp region upstream of the coding region. This distance was previously defined to be the maximum 5´UTR length in *P. aeruginosa* PA14 (Wurtzel *et al.*, 2012). As reproducibility was shown to be excellent (section 3.2.5.1), only values found in all three replicates were considered for 5´UTR length analysis. In total, 2,369 out of 5,572 anORFs showed a reproducible TSS matching the applied criteria. Among these, 1,207 were also reported for the homologues present in PA14 by Wurtzel *et al.* (2012), whereby 55.5% of them were located at the exact same position in both strains. In strain PAO1, the TSS was on average located 131 bp (median = 69 bp) upstream of the start codon (**Figure 33B**). However, a variety of shorter as well as larger 5´UTRs were obtained (1st quartile = 30 bp; 3rd quartile = 193 bp). Based on this analysis, a conservative distance of 200 bp upstream of novel gene candidates, which corresponds to detection of 75% of all reproducible anORF-associated TSS, was searched for the presence of a TSS as described in section 3.2.6.2.

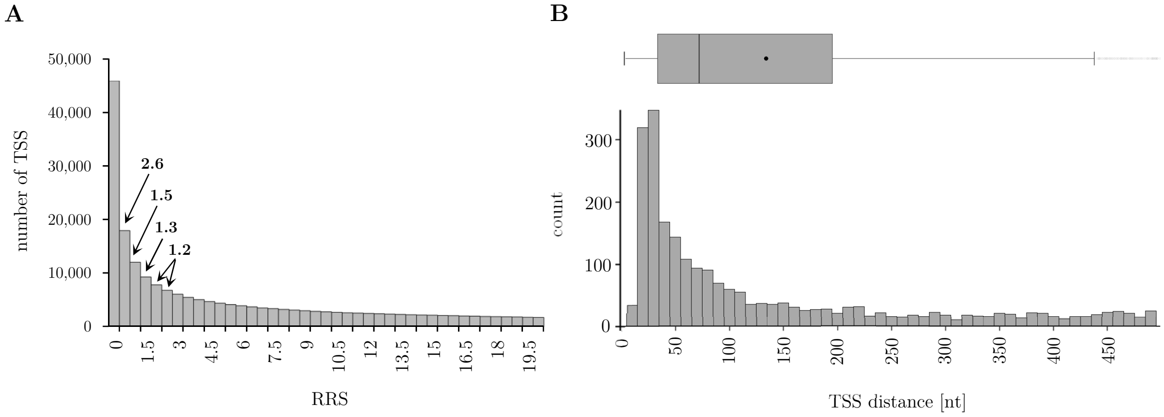


Figure 33. Results of the (**A**) cut-off and (**B**) 5`UTR length analysis of transcription start sites (TSS) in *P. aeruginosa* PAO1. (**A**) Number of predicted TSS after applying relative read scores (RRS) of 0 to 20. Numbers above bars represent fold changes for TSS counts obtained with a certain threshold in comparison to the respective value of the previous applied threshold. (**B**) Distance of TSS from the start codon of all anORFs for which a reliable TSS exceeding a RRS of 1.5 was detected in three biological replicates. The boxplot indicates first, second and third quartile values; the dot represents the mean value.

3.2.6 Detection and analyses of novel intergenic and overlapping genes

Analogous to *E. coli* LF82, all six Ribo-seq datasets generated for *P. aeruginosa* PAO1 were evaluated with regard to novel, translated ORFs by using different prediction tools. Individual results as well as the reproducibility of the predictions across biological replicates are discussed in section 3.2.6.1. The most reliable gene candidates with high prediction probability scores were subsequently characterized in more detail (section 3.2.6.2). Finally, two examples of exceptional long OLGs were further investigated experimentally and bioinformatically (section 3.2.6.3).

3.2.6.1 Application of prediction algorithms for individual and collective datasets

Retapamulin-supported Ribo-seq resulted in 22,517 novel gene candidates after applying the scripts by Meydan *et al.* (2019) using a threshold value of 0.63. This number was slightly higher than the number predicted for *E. coli* LF82, probably due to the reduced threshold value. The remaining Ribo-seq datasets of Exp1_RepI+II, Exp2_ND and both datasets of Exp3 were analysed using DeepRibo, REPARATION, and the scripts by Giess *et al.* (2017) yielding on average 6,389, 4,033 and 79,086 predictions per dataset, respectively. DeepRibo predicted a 2.2-fold higher and REPARATION a 2.1 lower number of translated ORFs in *P. aeruginosa* PAO1 compared to *E. coli* LF82. In contrast, scripts by Giess *et al.* (2017) predicted similar numbers for both organisms. DeepRibo as well as the scripts by Giess *et al.* (2017) showed the highest number of predictions for the dataset Exp1_RepII, whereas REPARATION predicted the most hits for the datasets of Exp3.

When calculating the percentage of ORFs predicted by one tool in multiple datasets (**Table 20**), the by far highest prediction percentages ranging from 56.8% to 76.8% were again obtained for the scripts by Giess *et al.* (2017). DeepRibo and REPRATION, in contrast, had lower prediction overlaps. However, compared to *E. coli* LF82, the percentage of identically predicted ORFs was on average higher for *P. aeruginosa* PAO1. The largest overlap of hits predicted by DeepRibo and REPRATION was obtained for biological replicates, thereby suggesting a better prediction reproducibility for *P. aeruginosa* PAO1 compared to *E. coli* LF82.

Table 20. Reproducibility of Ribo-seq prediction results for *P. aeruginosa* PAO1. Shown are the percentages of identical predictions obtained by the tool DeepRibo, REPARATION and the scripts by Giess *et al.* (2017) for pairwise comparisons of all experiments performed in *P. aeruginosa* PAO1.

			DeepRibo	REPARATION	Scripts by Giess *et al.* (2017)
Exp1_RepI	—	**Exp1_RepII**	67.81	51.91	61.41
Exp1_RepI	—	**Exp2_ND**	40.20	37.71	56.78
Exp1_RepI	—	**Exp3_MNase**	60.74	37.93	60.60
Exp1_RepI	—	**Exp3_RelE**	59.30	40.51	62.50
Exp1_RepII	—	**Exp2_ND**	36.90	40.01	74.23
Exp1_RepII	—	**Exp3_MNase**	62.32	40.11	74.00
Exp1_RepII	—	**Exp3_RelE**	58.59	42.04	72.05
Exp2_ND	—	**Exp3_MNase**	38.78	41.94	76.24
Exp2_ND	—	**Exp3_RelE**	44.87	41.31	76.78
Exp3_MNase	—	**Exp3_RelE**	67.11	45.25	75.14

Analogous to the results obtained for *E. coli* LF82, the prediction overlap of all three tools on one Ribo-seq dataset in *P. aeruginosa* PAO1 was only marginal and ranged from 1.4 to 2.5% of the entire number of hits (**Supplementary Figure 18**). For most reliable ORF identification, all prediction results obtained for the Ribo-seq datasets (Exp1_RepI+II, Exp2_ND and Exp3 with and without RelE) were merged with the results obtained for the Ribo-Ret dataset (Exp2_RET) evaluated according to Meydan *et al.* (2019).

In addition, all peptides detected in the DDA-MS experiment after searching MS2 spectra against a six-frame translation were added. After merging of the 16 different prediction results, 81,222 ORFs predicted in at least one dataset by one of the mentioned possibilities were obtained. A score between 1 and 16 was assigned to each hit which represents the absolute number of positive predictions (score 1 ≙ ORF predicted by one tool in one dataset; score 16 ≙ ORF predicted by all tools in all datasets). Predictions found in more than a half of all possibilities (score > 8) as well as all hits with a lower score but with MS evidence by minimum two peptides were further analysed. After classification and visual inspection, 124 novel gene candidates with scores between 2 and 15 remained (**Supplementary Table 5**). Three examples of the identified OLG candidates with their respective Ribo-seq signals of Exp2 are displayed in **Figure 34**.

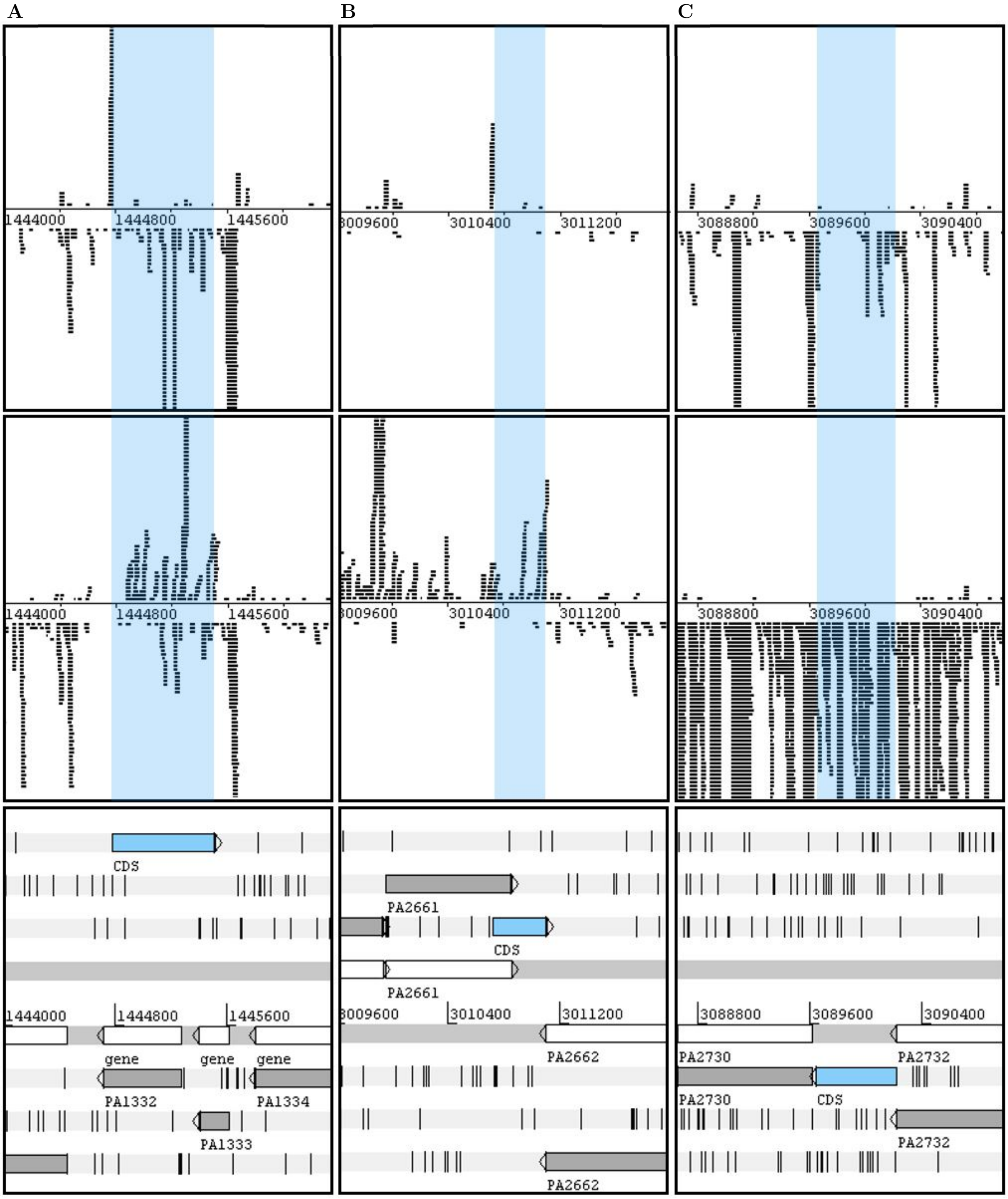


Figure 34. Ribo-seq signals of three novel gene examples identified in *P. aeruginosa* PAO1. Strand-specific Ribo-seq signals of the RET dataset (Exp2_RET, top row) and the ND dataset (Exp2_ND, middle row) were visualized using the Artemis genome browser (Rutherford *et al.*, 2000). Translatome reads mapping to the forward strand are displayed above the centre line and reads mapping to the antisense strand are plotted below the centre line. The bottom row represent a six-frame translation of the genomic loci coding for the novel overlapping gene candidates (**A**) PAO1_109 (antisense overlap with two anORFs), (**B**) PAO1_108 (sense overlap) as well as (**C**) PAO1_1 (trivial overlap). All ORFs are displayed in their respective frame and anORFs are highlighted in grey. Black bars indicate stop codon positions. The genomic regions coding for the novel ORFs are shaded in blue.

3.2.6.2 Bioinformatic characterization of the selected ORFs

Seventy-three of the identified gene candidates were encoded on the antisense strand and fifty-one candidates on the sense strand. Both groups were widely disseminated across the genome and six gene candidates were located in putative phage regions (**Figure 35A**). With a median size of 339 bp, the novel ORFs were shorter than annotated, protein-coding genes in *P. aeruginosa* PAO1 (median = 873 bp) but substantially larger than the novel gene candidates identified in *E. coli* LF82. The largest novel gene candidate was a 2,037 bp long ORF overlapping non-trivially with three anORFs. Overlaps with three anORFs were also detected for two further ORFs spanning 1,221 and 1,533 bp. ATG was the most used start codon (82.3%), followed by GTG (13.7%) and TTG (4.0%). The anORFs showed a comparable start codon usage, with 89.1% of all anORFs starting with ATG, 9.4% with GTG, 1.4% with TTG and 0.02% with rare start codons (i.e., ATC). Many of the ORFs detected were located in intergenic regions (n = 41) or overlapped only trivially with anORFs (n = 20), while the remaining candidates showed substantial overlaps (n = 63; **Supplementary Table 5**) covering minimum 8.8% of the entire ORF length. In 51 cases, the mother ORFs, which overlapped either trivially or non-trivially with a novel ORF, were annotated as "hypothetical".

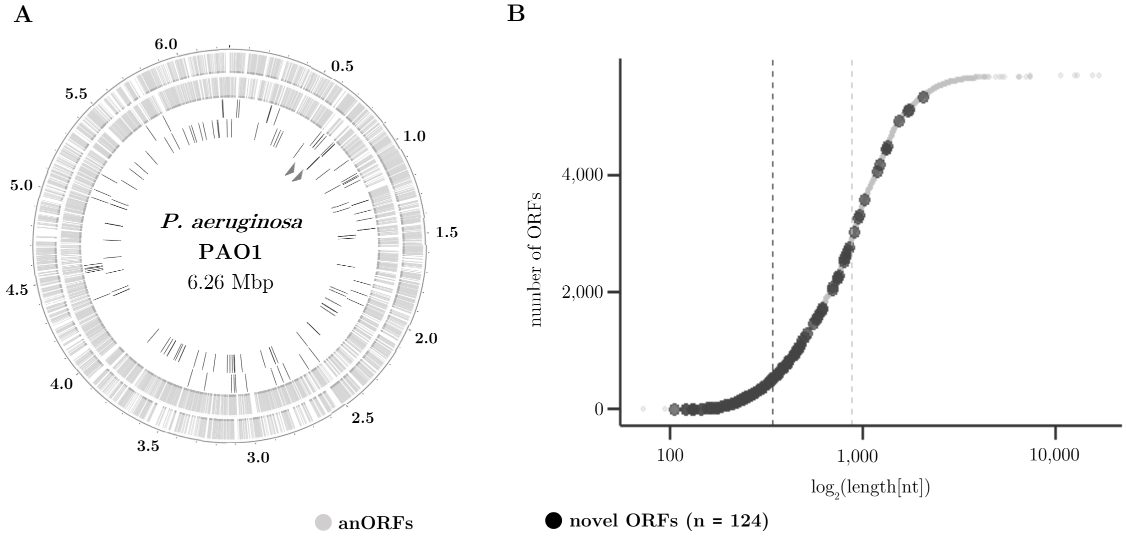


Figure 35. (**A**) Distribution and (**B**) length of novel gene candidates in the *P. aeruginosa* PAO1 genome. (**A**) The circles show the annotated genes (grey, n = 5,572) located at the sense strand and the antisense strand, the novel gene candidates (black, n = 124) located at the sense and the antisense strand as well as phage regions as indicated by grey arrows (from outside to inside). (**B**) The length distribution of all novel gene candidates and all annotated genes with their respective median values (dashed line) are indicates by grey (anORFs) or black (novel ORFs) dots.

For 49.2% of the novel gene candidates, a reproducible TSS with a RRS score ≥1.5 was detectable within a 200 bp region upstream of the putative start position in two of three biological replicates. This percentage decreased only marginally to 45.2% when the TSS had to be present in all three replicates. With a mean length of 58 bp, the novel ORF candidates had similar 5`UTR lengths as reported for the encoded anORFs (**Supplementary Figure 19A**). The RRS of the novel TSS found in two of three replicates ranged from 1.5 to 1,157.2. However, a correlation between the RRS of the TSS and the PRKM values of the transcriptome was not detectable, neither for the novel ORFs (Pearson´s r = 0.02) nor for the anORFs (Pearson´s r = 0.19). In order to detect conserved promoter motifs, a 20 bp region upstream of the TSS was extracted and sequence logos were created using WebLogo (Crooks *et al.*, 2004). Both the novel ORFs (**Figure 36A**) as well as the anORFs (**Figure 36B**) with reproducible TSS showed a highly conserved -10 promoter region with A and T being the most prevalent nucleobases. In addition to TSS, 83.9% of all novel gene candidates exhibited a SD sequence in an average distance of 8 bp upstream of the start codon. When analysing a region 20 bp around of each start codon, a clear SD consensus pattern was observable which resembled the SD pattern obtained for all anORFs (**Figure 36CD**). Furthermore, ρ-independent terminator structures were detected in a median distance of 103 bp (**Supplementary Figure 19B**) downstream of the stop codon for 22.6% of all novel gene candidates.

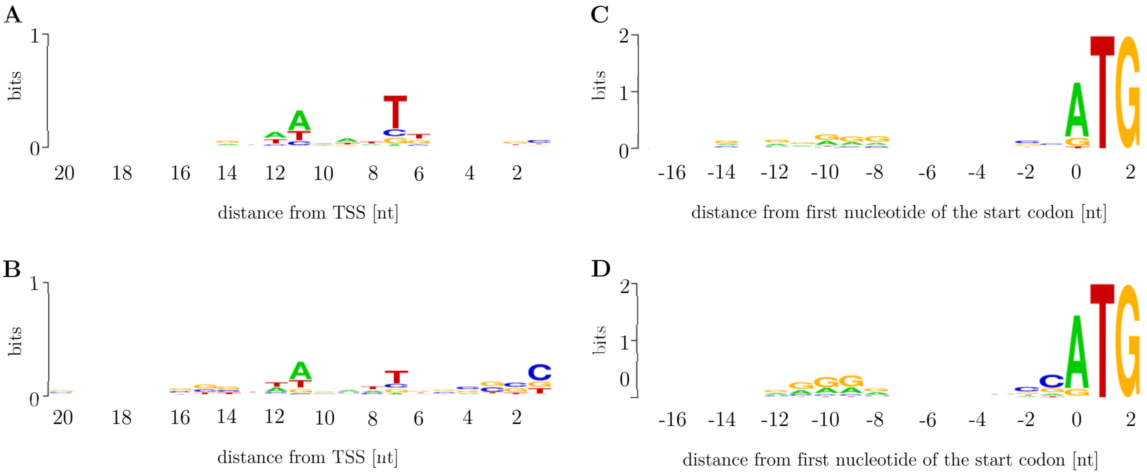


Figure 36. Sequence logos of (**A, B**) promoter regions and (**C, D**) putative Shine-Dalgarno sequences in *P. aeruginosa* PAO1. A region of 20 bp upstream of reproducible TSS with RRS ≥1.5 are shown for (**A**) novel ORFs (n = 61) and (**B**) anORFs (n = 2,369). The Shine Dalgarno sequence motif is displayed within a 20 bp sequence region around the start codon of (**C**) all novel ORFs (n = 124) and (**D**) all anORFs (n = 5,572).

Despite having structural features associated with protein coding, the novel ORF also showed profound transcription and translation signals (**Figure 37**). Their values of RPKM, RCV and read coverage used for evaluation of expression capability and strength were similar to those of protein-coding anORFs. When comparing RPKM values of the transcriptome (**Figure 37A**), the novel ORF candidates showed substantial higher values (median = 71.7) compared to all anORFs (median = 37.5). A similar observation was also made for the RPKM values of the translatome (median novel ORF = 61.7; median anORF = 27.4; **Figure 37B**). Both the novel ORFs as well as the anORFs exhibited on average an excellent read coverage (**Figure 37C**). In addition, the novel ORFs exhibited on average higher RCV values (median = 0.96; **Figure 37D**), and thus, showed a better translatability compared to the anORFs (median = 0.74). Analogous to *E. coli* LF82, separating overlapping from intergenic gene candidates did not lead to deviating results (results not shown).

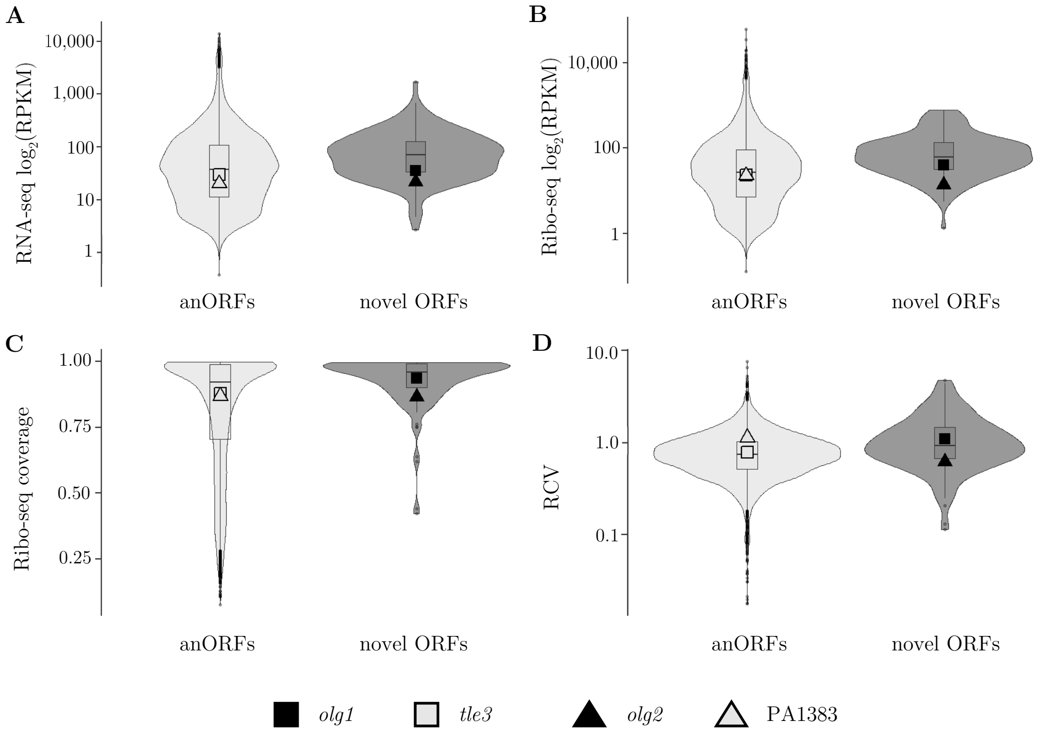


Figure 37. Expression metrics of the novel gene candidates (n = 124) in comparison to protein-coding, annotated genes (n = 5,572) in *P. aeruginosa* PAO1. Violin plots display the mean reads per million mapped reads (RPKM) values of (**A**) RNA-seq and (**B**) Ribo-seq, the (**C**) mean read coverage of ORFs in Ribo-seq and (**D**) the mean ribosome coverage value (RCV) of two biological replicates (Exp1_RepI+II). The values of *olg1* and *olg2* are indicated by a black rectangle and triangle; those of their mother genes *tle3* and PA1383 by the respective grey-filled symbol.

Ribo-RET analysis revealed translation initiation sites with a RET peak value larger than 0.63 RPM for 33.9% of all novel gene candidates. This number was 2.4-fold lower than the respective value of the novel ORFs identified in *E. coli* LF82 after applying a threshold of 1 RPM. The peak heights obtained for the novel gene candidates (median = 0.06) were slightly lower than those of all anORFs (median = 0.11) and substantially lower compared to those obtained in *E. coli* LF82 Ribo-RET suggesting reduced RET efficacy in *P. aeruginosa* PAO397. Nevertheless, metagene analysis confirmed the selective accumulation of ribosomes at start codons of the novel gene candidates in comparison to the ND control (**Supplementary Figure 20**).

A clear RF signal was not obtained when analysing all novel ORFs in sum, neither in the Ribo-seq experiment with RelE nor in the experiment prepared without RelE (**Figure 38A**). For the first, 44% of all reads mapped to the expected sub-codon position 2, whereas this percentage was slightly lower (37%) for the latter. A similar observation was made when analysing all sense overlapping ORFs separately (**Figure 38B**). However, after omission of all sense overlapping ORFs from the novel ORFs, a pronounced periodicity signal of 55.5%, 26.4% and 18.1% for positions 2,1 and 3 were observed in the RelE-supported Ribo-seq dataset (**Figure 38C**). The resolution of this signal was greatly enhanced compared to the control dataset prepared without RelE suggesting effectiveness of RelE in *P. aeruginosa* PAO1. The latter also showed a slight periodicity signal (36% to 44% to 20%); however, this signal arises from the NNC shift and was not present in the unshifted raw data (25% of all read mapping to position 2). In contrast, unshifted reads of the dataset prepared with RelE already showed a preference of position 2 (41%). In sum, these results were similar to those obtained for novel ORFs in *E. coli* LF82. In this organism, a RF signal was only detectable when leaving sense overlapping ORFs aside because their mother genes encoded at the same strand caused a distortion of the RF signal.

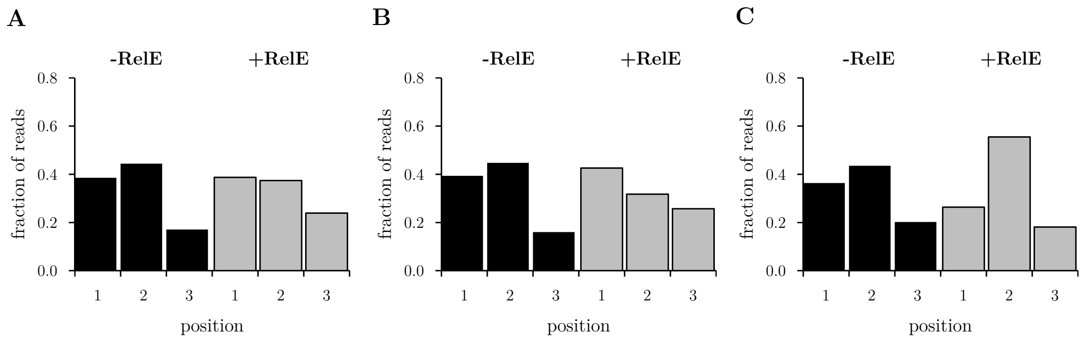


Figure 38. Reading frame sum signal of all novel gene candidates in *P. aeruginosa* PAO1. All reads obtained for Exp3 generated with (+RelE) or without RelE (-RelE) were mapped to each sub-codon position and a NNC shift was performed. The results obtained (**A**) for all novel candidates (n = 124), (**B**) for all ORFs overlapping sense with anORFs (n = 47) and (**C**) all ORFs minus sense overlapping ORFs (n = 77) are shown.

The proteinaceous nature of 47 out of 124 novel gene candidates was verified by the detection of up to 13 peptides by DDA-MS. The majority of these ORFs (n = 31) was located either in intergenic regions or overlapped only trivially with anORFs. However, proteomic evidence was also obtained for 16 overlapping ORFs, often even by multiple peptides (for details see **Supplementary File S2**). Among the latter were also the candidates exemplarily shown in **Figure 34** as well as all candidates exhibiting an overlap with three anORFs each, whereby their encoded proteins were detected by minimum two peptides. In addition, MS confirmed that in 54 of 63 cases at least one of the annotated mother genes overlapping non-trivially with novel ORFs were protein coding. Of all overlapping ORFs with peptide evidence, 87.5% of the protein products encoded by their respective mother genes were also detected via MS.

Blastp analysis using the non-redundant protein database revealed protein homologues of 84 novel gene candidates with an e-value $\leq 1\times10^{-3}$ and a minimum identity percentage of 70%. The homologues were primarily found within the genus *Pseudomonas*. 55 of them were either assigned as "hypothetical", "putative" or "uncharacterized", for the remaining a function, but often only the presence of a certain functional domain or the affiliation to a specific protein family was specified. Searching against the RefSeq Select proteins database resulted in the detection of 23 homologous proteins with high confidence, seven of them were classified as non-trivial overlapping ORFs in this study. However, four of them were homologous to unknown proteins or proteins containing domains of unknown functions. One of the remaining candidates had a significant hit with a virulence-associated protein of the RhuM family. According to the Pfam database (Mistry *et al.*, 2021), proteins belonging to that family have not been experimentally validated so far, but are considered to be involved in pathogenicity due to the location of homologues within the *Salmonella* pathogenicity island 3 (Amavisit *et al.*, 2003, Blanc-Potard *et al.*, 1999). However, the homologous region between the protein encoded by the overlapping ORF found in *P. aeruginosa* and the blastp hit was only 58% and was exclusively restricted to the non-overlapping region. A second hit was obtained for a protein encoded by an ORF fully embedded within the gene *mexT*. This protein showed an alignment with a multispecies LysR-family transcriptional regulator protein. Interestingly, *mexT* also encodes for a transcriptional regulator belonging to the LysR family and is known to have a high mutational rate (LoVullo & Schweizer, 2020, Chandler *et al.*, 2019, Maseda *et al.*, 2000) due to its regulatory function of diverse processes including multidrug resistance (Köhler *et al.*, 1999). The protein product of the last overlapping ORF showed 75.5% similarity to a multispecies class I SAM-dependent methyltransferase with an e-value of 2×10^{-104}. This gene overlapped with two anORFs in *P. aeruginosa*, whereby the overlapping ORF as well as one of the hypothetical mother ORFs were confirmed to be protein coding by DDA-MS. All results of the novel ORF candidates described in this chapter are detailly listed in **Supplementary File S2**.

3.2.6.3 Characteristics of two exceptionally long OLGs

Two overlapping gene candidates, later designated as *olg1* and *olg2*, showed profound transcriptional and translational signals as well as multiple peptides detected via DDA-MS. In addition, their exceptional sizes made them promising candidates for further experimental and bioinformatic characterization. All results in this section are published as a joint preprint of Michaela Kreitmeier, Zachary Ardern, Miriam Abele, Christina Ludwig, Siegfried Scherer, and Klaus Neuhaus (Kreitmeier *et al.*, 2021). Michaela Kreitmeier conducted ribosome profiling and analysis, Zachary Ardern performed evolutionary analyses, Miriam Abele provided MS data, Christina Ludwig, Siegfried Scherer, and Klaus Neuhaus provided resources for the experiments conducted, supervised the publication and helped with writing.

3.2.6.3.1 Genomic localization

The 957 nt long OLG *olg1* is encoded at the genome coordinates 291556-292512(+) with ATG$_{291556}$ as the putative start codon (**Figure 39A**). A N-terminal extension of the coding region by up to 261 nt would be conceivable due to the existence of five alternative start codons (CTG$_{291295}$, TTG$_{291370}$, CTG$_{291436}$, ATG$_{291508}$ & GTG$_{291535}$), which were located in-frame upstream of ATG$_{291556}$. However, several aspects emphasize that ATG$_{291556}$ is the correct start codon of the ORF under the condition tested. These aspects will be detailly explained in the following sections and include correct spacing to gene-like elements and corresponding Ribo-seq signals. Regardless of the correct start position, *olg1* overlaps entirely in frame -1 with the annotated mother gene *tle3* (PA0260). *Tle3* is part of the *vgrG2b-tli3-tle3-tla3* operon and encodes for an antibacterial type VI lipase effector 3, which is delivered to prey bacteria via a T6SS (Berni *et al.*, 2019). The remaining genes encoded in the operon are functional related and are necessary for the secretion of Tle3 or confer Tle3 immunity. Remarkably, *olg1* overlaps with two structural domains of *tle3*, sharing 39 nt with a N-terminal α/β hydrolase fold domain as well as 594 nt with an C-terminal domain of unknown function (DUF3274; Berni *et al.*, 2019).

The novel gene *olg2* has a length of 1,728 nt and is located at the genomic position 1501875-1503602(-). Its coding region most likely starts at the ATG$_{1503602}$, and it overlaps partially with two annotated genes (**Figure 39B**). *Olg2* is encoded in frame -1 and shares 1,536 nt with the mother gene PA1383, a hypothetical gene putatively associated with exopolysaccharide synthesis (Matsukawa & Greenberg, 2004) and predicted to harbour a N-terminal type I signal sequence (Lewenza *et al.*, 2005). A second overlap of 34 nt is shared with the UDP-glucose 4-epimerase encoding gene *galE* (PA1384) in frame +2.

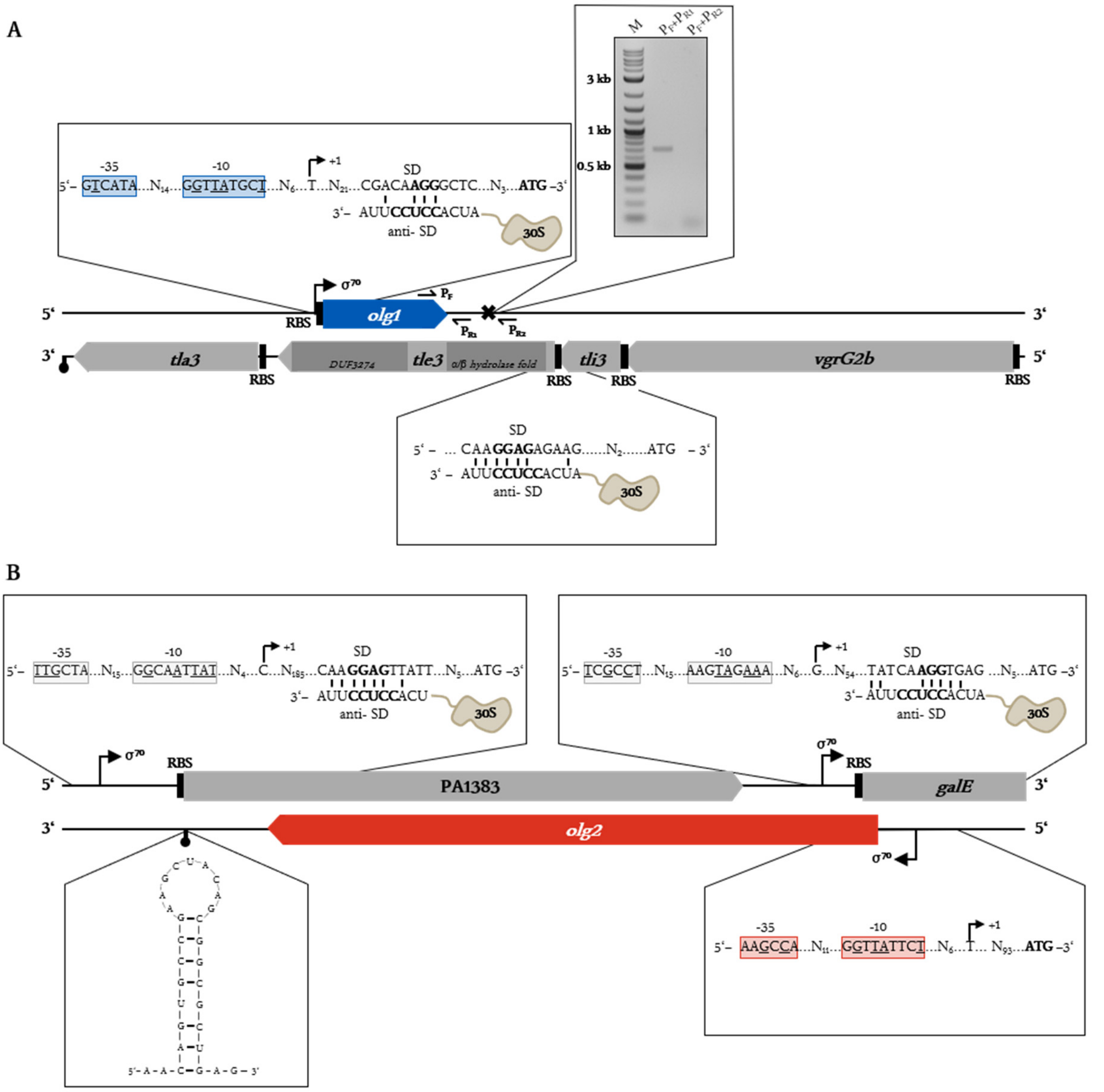

A
M
PF+PR1
PF+PR2
3 kb
1 kb
0.5 kb
-35
-10
+1
SD
5'– GTCATA...N14...GGTTATGCT...N6...T...N21...CGACAAGGGCTC...N3...ATG –3'
3'– AUUCCUCCACUA
anti- SD
30S
σ70
PF
5'
olg1
3'
RBS
PR1
PR2
3'
tla3
RBS
DUF3274
tle3
α/β hydrolase fold
RBS
tli3
RBS
vgrG2b
RBS
5'
SD
5' – ... CAAGGAGAGAAG......N2......ATG – 3'
3' – AUUCCUCCACUA
anti- SD
30S
B
-35
-10
+1
SD
5'– TTGCTA...N15...GGCAATTAT...N4...C...N185...CAAGGAGTTATT...N3...ATG –3'
3'– AUUCCUCCACU
anti- SD
30S
-35
-10
+1
SD
5'– TCGCCT...N15...AAGTAGAAA...N6...G...N54...TATCAAGGTGAG ...N5...ATG –3'
3'– AUUCCUCCACUA
anti- SD
30S
σ70
RBS
σ70
RBS
5'
PA1383
galE
3'
3'
olg2
5'
σ70
5'-A-A-C-G-A-G-3'
-35
-10
+1
5'– AAGCCA...N11...GGTTATTCT...N6...T... N93...ATG –3'

Figure 39. Genomic location of (**A**) *olg1* and (**B**) *olg2*. *Olg1* is fully embedded in the annotated gene *tle3,* which has two functional domains (α/β hydrolase fold domain and DUF3274 domain) and is encoded within the *vgrG2b-tli3-tle3-tla3* operon. *Olg2* overlaps partially with the annotated gene PA1383 as well as trivially by 34 nt with an UDP-glucose 4-epimerase encoding gene (*galE*). Both OLGs have structural features associated with protein-coding including -35 and -10 consensus elements of a σ^{70} promoter, a SD sequence (only *olg1*) and a rho-independent terminator (only *olg2*) as indicated. In addition, RT-PCR using the primer pairs P$_{F}$+P$_{R1}$ and P$_{F}$+P$_{R2}$ identified cessation of mRNA synthesis between 219 and 349 nt downstream of *olg1*´s stop codon. Structural features of the annotated mother genes are displayed. Figure adapted from Kreitmeier *et al.* (2021).

3.2.6.3.2 Gene-like structural features of *olg1* and *olg2*

In silico analysis using BROM (Solovyev & Salamov, 2011) predicted σ^{70} promoter sequences in a distance of 37 nt and 94 nt for *olg1* and *olg2*, respectively (**Figure 39**). Obtained LDF values of 1.94 (*olg1*) and 1.37 (*olg2*) were significantly larger than the threshold value of 0.2, and thus, indicated high accuracy and specificity of the predicted promoters. Similar results were obtained by Cappable-seq when analysing an optimal region of 200 bp upstream of the OLGs´ coding regions. For *olg1*, a maximum TSS with a mean RRS of 3.0 was detected in a distance of 149 bp in three biological replicates. In addition, a second reproducible TSS with a slightly reduced RRS of 2.0 was located 36 nt upstream of *olg1*´s start codon (**Supplementary Figure 21A**), which endorsed the results predicted by BROM.

A TSS with a mean RRS of 2.2 located 94 nt upstream of the start codon as well as an additional TSS 129 nt upstream of the start codon (RRS = 4.93) were also identified for *olg2* in at least two biological replicates (**Supplementary Figure 21B**).

Terminator regions within a 300 bp region downstream of the OLGs´ stop codon were identified using FindTerm (Solovyev & Salamov, 2011). This program predicted a ρ-independent terminator 218 to 247 nt downstream of *olg2*, whereas no terminator could be detected for *olg1*. However, termination of transcription was experimentally verified by RT-PCR to take place in a region between 219 and 349 nt downstream of the stop codon (**Figure 39A**). In addition, the upstream vicinity of both OLGs was subjected to SD analysis. The optimal aSD sequence in *P. aeruginosa* is located 7-9 nt upstream of the start codon, has the sequence pattern CCUCC and a ΔG_{SD} of −6.5 kcal/mol (Ma *et al.*, 2002). *Olg1* exhibited a SD sequence with the motif AGG and a ΔG_{SD} of −3.6 kcal/mol in an optimal spacing of 8 nt to the start codon ATG_{291556}. In contrast, *olg2* did not harbour a SD sequence. The mother genes *tle3*, PA1383 and *galE* also shown profound structural elements indicating their genuine protein-coding nature. Both mother genes of *olg2*, for instance, had a SD sequence with a ΔG_{SD} of −6.1 (PA1383) and −4 kcal/mol (*galE*) 8 nt upstream of their start codon (**Figure 39B**). Furthermore, TSSs were predicted by BROM to be localized 203 nt (PA1383) and 72 nt (*galE*) upstream of their start codon, respectively. For the former, a TSS with a mean RRS of 9.2 was also confirmed in a similar distance (197 bp) by Cappable-seq. As a member of the *vgrG2b* operon (Berni *et al.*, 2019), *tle3* neither possesses a σ^{70} promoter nor a ρ-independent terminator. However, a strong SD sequence (ΔG_{SD} −5.1 kcal/mol) with the sequence pattern AGGAG was detected 5 nt upstream of the start codon.

Programs used for gene delineation, e.g., Prodigal, often rely on diverse sequence features including start codon and SD motif usage, gene size, GC bias, hexamer coding statistics etc. for gene prediction (Hyatt *et al.*, 2010). However, dynamic programming prohibits prediction of large overlapping genes and chooses the overlapping ORF with the highest score for annotation. In total, 5,681 protein-coding genes were predicted by Prodigal to be encoded in the *P. aeruginosa* PAO1 genome with scores ranging from 0.1 (PA2732 encoding for a hypothetical protein) to 3,037.7 (PA2424 encoding for a peptide synthase). The mother genes *tle3* and PA1383 were also classified as protein-coding genes with high scores of 132.6 and 225.9, respectively (**Supplementary Figure 22**, **Supplementary File S3**). To test whether Prodigal classifies the two overlapping ORFs as protein-coding genes, all start codon nucleotides within the coding region as well as in the upstream vicinity of the annotated genes *tle3* and PA1383 were replaced by Ns, thus concealing the mother genes. As a result, both overlapping ORFs were predicted to be protein-coding genes with scores of 4.63 (*olg1*) and 23.63 (*olg2*). Although their overall scores were rather low in comparison to those of the other predicted genes, both OLGs display values comparable to anORFs for some features, for instance the GC content score (gc_cont) and the start sequence region score (uscore; **Supplementary Figure 22**, **Supplementary File S3**). Clearly both overlapping ORFs, despite having sequence features associated with protein-coding, missed annotation due to their long antisense overlaps.

3.2.6.3.3 Transcriptome and translatome analysis

Both OLGs were covered with a substantial number of reads when analysing the RNA-seq datasets of Exp1. The reads mainly mapped to the region between the predicted TSS and the putative transcription termination site (**Figure 40**, first track) resulting in $RPKM_{RNA\text{-}seq}$ values of 35.9 and 22.1 for *olg1* and *olg2*, and values of 29.7 and 20.2 for the mother genes *tle3* and PA1383 (**Supplementary Table 6**). These values were in a medium range compared to all protein-coding anORFs (**Figure 37A**) suggesting that all four ORFs are genuinely transcribed.

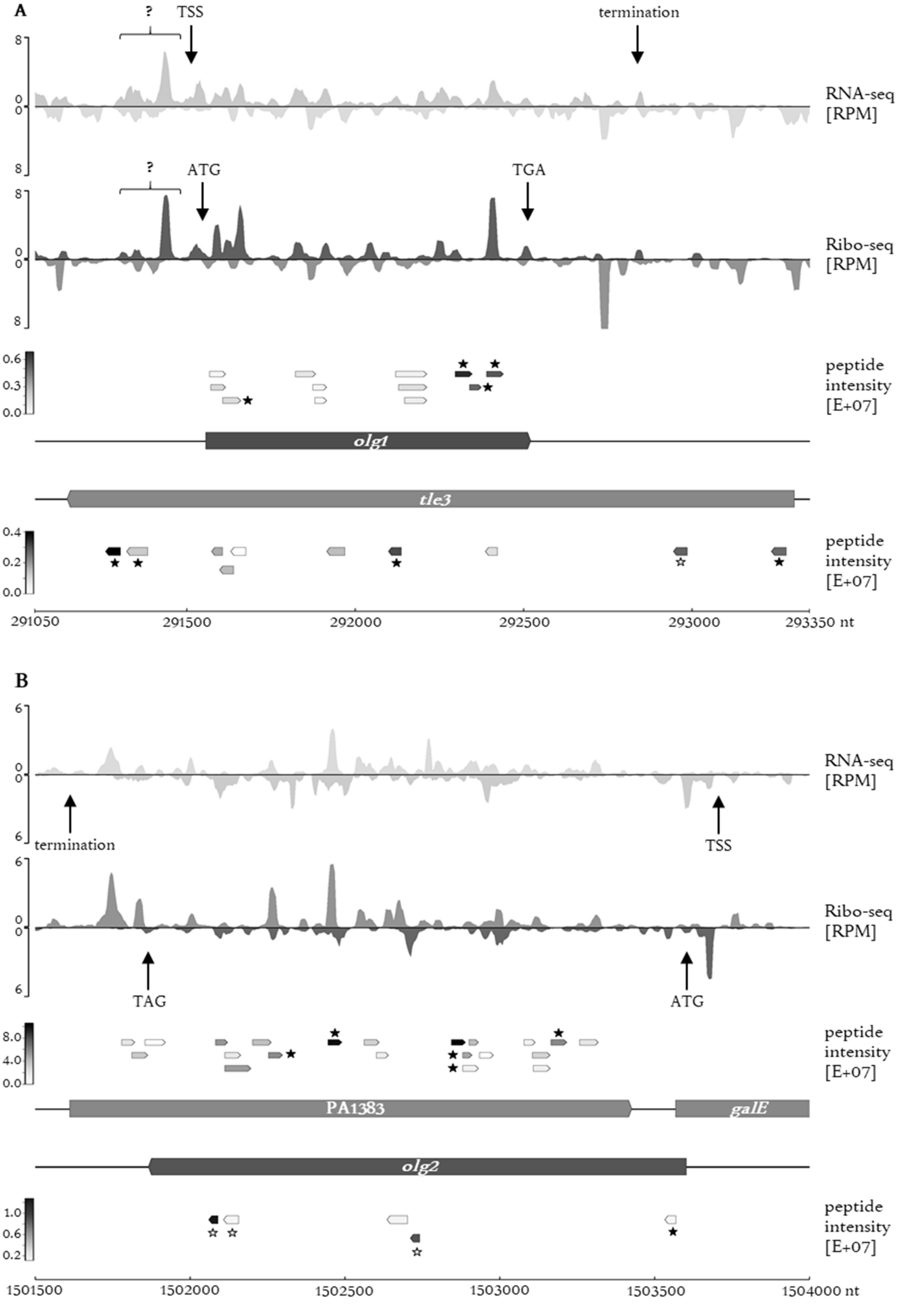


Figure 40. Results of RNA-seq, Ribo-seq and proteome analysis at the (**A**) *olg1/tle3* and (**B**) *olg2*/PA1383 locus. Strand-specific RPM values of RNA-seq (first track) and Ribo-seq reads (second track) averaged over the biological replicates of Exp1 are displayed for *olg1* and *olg2* as well as their mother genes *tle3* and PA1383. Arrows indicate the position of transcription start (TSS) and stop sites (termination) and question marks highlight regions with unexpected signals. The position and intensity of all peptides measured by DDA-MS are displayed in track 3 and 4. Peptides marked with an asterisk were chosen for targeted mass spectrometry; and those verified by the latter are illustrated by a filled asterisk. Figure adapted from Kreitmeier *et al.* (2021).

In addition to RNA-seq signals, RT-PCR using primers binding at the beginning and at the end of both ORFs also confirmed transcription of the entire coding region of *olg1* and *olg2* (**Supplementary Figure 23**). Furthermore, both OLGs showed profound translational signals which were reproducibly detected in both biological replicates. The reads mainly covered the region between the start and stop codon (**Figure 40**, second track). However, for *olg1* an accumulation of RNA-seq as well as Ribo-seq reads upstream of the start codon indicated either that *olg1* is N-terminally enlarged or that an additional ORF located in a different reading frame is transcribed and translated. With an $\mathrm{RPKM}_{\text{Ribo-seq}}$ value of 40.3, the expression of *olg1* was even higher than those of many anORFs (**Figure 37B**). *Olg2*, in contrast, had a lower, but nevertheless indubitable $\mathrm{RPKM}_{\text{Ribo-seq}}$ value of 14.2. The $\mathrm{RPKM}_{\text{Ribo-seq}}$ values of the mother ORFs (*tle3*: 23.2; PA1383: 22.8) were in a similar range as the medium value obtained for all anORFs (**Figure 37B**). The Ribo-seq signals observed at both gene loci did not seem to originate from pervasive translation, as for instance observed in *Mycobacterium tuberculosis* (Smith *et al.*, 2019), as the genes encoded upstream of *tle3* in the *vgrG2b* operon did not shown any antisense signals (**Supplementary Figure 24**). In addition, all four target ORFs exhibited a high read coverage (**Figure 37C**) and a medium to high translatability compared to all anORFs as indicated by their RCV values (**Figure 37D**). Based on the observed Ribo-seq signals, *olg1* and *olg2* were predicted to be protein-coding genes with translation starting from ATG_{291556} and $\mathrm{ATG}_{1503602}$, respectively. *Olg1* was predicted by multiple tools in multiple datasets in 10 out of 16 possible combination (score = 10), whereas *olg2* had a slightly lower prediction score of 9. Prediction tools also confirmed the protein-coding nature of the mother genes *tle3* and PA1383 with overall scores of 13 and 9, respectively.

For *olg2* and *tle3*, the exact position of the start codon was also supported by the result of Ribo-RET. The first showed a TIS with a peak height of 2.3 RPM at position 1503601 nt, and the latter an even more pronounced peak (4.6 RPM) at position 293302 nt (**Supplementary Figure 21**, first track). No RET peak was detected for *olg1* and PA1383. A clear RF signal of 31% to 63% to 5% was obtained for all read mapping to *olg1* after performing an NNC shift (**Supplementary Figure 25A**). In contrast, the absence of such a signal in the dataset prepared without RelE (43% to 41% to 16%) confirmed the authenticity of the translational signals. Even in the absence of a NNC shift, a read accumulation at the second sub-codon position, which amounted to 45%, was observable for all reads of the RelE-prepared dataset, whereas reads of the control experiment without RelE did not preferentially map to sub-codon position 2 (**Supplementary Figure 25B**). Conversely, *olg2* showed no preference for reads mapping to sub-position 2, neither in the shifted nor in the unshifted data (**Supplementary Figure 25CD**). In this case, the majority of reads mapped to the first sub-codon position regardless of the presence or absence of RelE during nuclease footprinting.

Further transcriptional and translational signals were obtained when analysing sequencing datasets published by Grady *et al.* (2017) who performed RNA-seq and Ribo-seq in *P. aeruginosa* PAO1 and *P. aeruginosa* ATCC33988 after cultivation in M9 both supplemented with different carbon sources (**Supplementary Figure 26**). Reanalysis of these datasets suggested expression of *olg1* in *P. aeruginosa* PAO1 as well as in *P. aeruginosa* ATCC33988. Altogether, the expression strength as indicated by RPKM values was higher when cultivating with glycerol as carbon source compared to *n*-alkanes, but nevertheless lower than after cultivation in LB broth (**Supplementary File S4**). *Olg2,* which is absent in strain ATCC33988, also showed slight transcriptional and translational signals in the data published for *P. aeruginosa* PAO1 (**Supplementary Figure 26B**). Comparison of the $\mathrm{RPKM}_{\text{RNA-seq}}$ values revealed a similar transcription strength after cultivation in LB and M9+glycerol, whereas the $\mathrm{RPKM}_{\text{RNA-seq}}$ values obtained after cultivation in the presence of *n*-alkanes were clearly lower (**Supplementary Figure 27**). This observation points to a potential regulation of *olg1* and *olg2* at the transcriptional level as a function of the carbon source present in the cultivation broth.

Further hits for a regulated expression of both OLGs were provided by differential expression analysis of the Ribo-seq datasets. Significant differences were obtained for *olg1* and *olg2* when calculating log fold changes between the M9+*n*-alkane and M9+glycerol datasets at a FDR of 0.05 (log_FC_{olg1} = 1.15; log_FC_{olg2} = -0.55). These results point towards a possible regulation, and thus, functionality of both OLGs.

3.2.6.3.4 Detection of translated peptides by mass spectrometry

For the initial discovery of translated peptides, a sample taken at OD_{600nm} = 1 was fractionated into 48 fractions and analysed by DDA-MS. Obtained MS spectra were searched against an organism-specific database of all protein-coding anORFs complemented by the AA sequences of *olg1* and *olg2*. In total, twelve and five different peptides were detected for *olg1* and *olg2,* respectively, which covered a wide area of the encoded proteins (**Figure 40**, track 3 and 4). Additionally, the expression of 4,076 anORFs was confirmed by this experiment. Among the proteins detected were also the proteins encoded by the mother genes *tle3* and PA1383. For the first, 10 peptides were identified, for the latter translation was proven by 21 peptides (**Figure 40**, track 3 and 4). 46 of the 48 detected target peptides had high dot product scored when comparing the observed fragment ion spectra with the spectra predicted by the algorithm Prosit (Gessulat *et al.*, 2019) indicating their high-confident identification (**Supplementary File S5**). The normalized iBAQ intensities of Olg1 and PA1383 were in a medium range compared the values measured for anORF-encoded proteins, whereas Olg2 and Tle3 were of rather low abundance (**Figure 41**).

In order to validate and quantify the detected proteins, targeted proteomic using PRM-MS was performed on several samples harvested throughout different growth phases (**Supplementary Figure 28A**). For this purpose, these samples were supplemented with four to five isotopically labelled reference peptides per gene. The selection of the peptides to be targeted was made based the results of the DDA-MS experiment (**Figure 40**, peptides highlighted by an asterisk). One peptide for Olg2, four peptides for each Olg1 and Tle3 as well as five peptides for PA1383 were successfully verified by this approach (**Figure 40**, peptides highlighted by a filled asterisk).

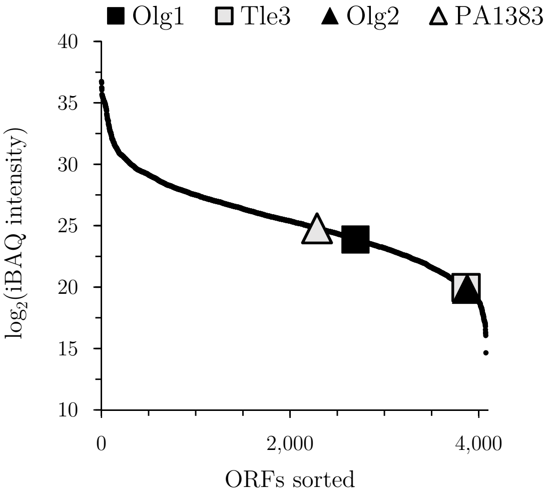


Figure 41. Normalized protein intensities measured by DDA-MS. Intensity Based Absolute Quantification (iBAQ) values of all proteins (n = 4,080) detected in a 48-fold fractionated sample harvested at OD_{600nm} = 1 are shown. Values of Olg1 and Olg2 are highlighted by a black rectangle and triangle, respectively. The proteins encoded by the mother genes *tle3* and PA1383 are illustrated by the respective grey-filled symbol. Figure adapted from Kreitmeier *et al.* (2021).

Quantification of the detected peptides revealed differences in protein abundance of all four targeted proteins as a function of growth (**Figure 42**). Olg1, for instance, was highly abundant in exponential phase (1 h and 2 h) as well as at the transition from exponential to stationary phase, whereas the abundance decreased sharply in stationary phase (6 h to 24 h). The by far highest intensity of Olg2 was measured at $OD_{600nm} = 1$. In late stationary phase, however, this protein was either not expressed or only in marginal amounts, which were below the detection limit. In contrast, the respective protein encoded by the gene PA1383 was stably expressed during the entire cultivation process. Increasing amount of the Tle3 protein were quantified from the beginning of cultivation until early stationary phase (4 h), but not in mid (6 h, 8 h) and late stationary phase (24 h). In sum, both OLGs showed a different time course of protein abundance compared to the proteins encoded by the mother genes located at the antisense strand. This result suggests that expression of the OLGs was regulated autonomously. A loading control of the summarized intensities of all peptides measured for the each of the samples confirmed that the observed differences in protein abundance were not merely caused by deviating amounts of input sample (**Supplementary Figure 28B**). In addition, qPCR analysis of *olg1* mRNA essentially confirmed the time course of protein abundance as seen in PRM-MS on the transcriptome level (**Supplementary Figure 29**).

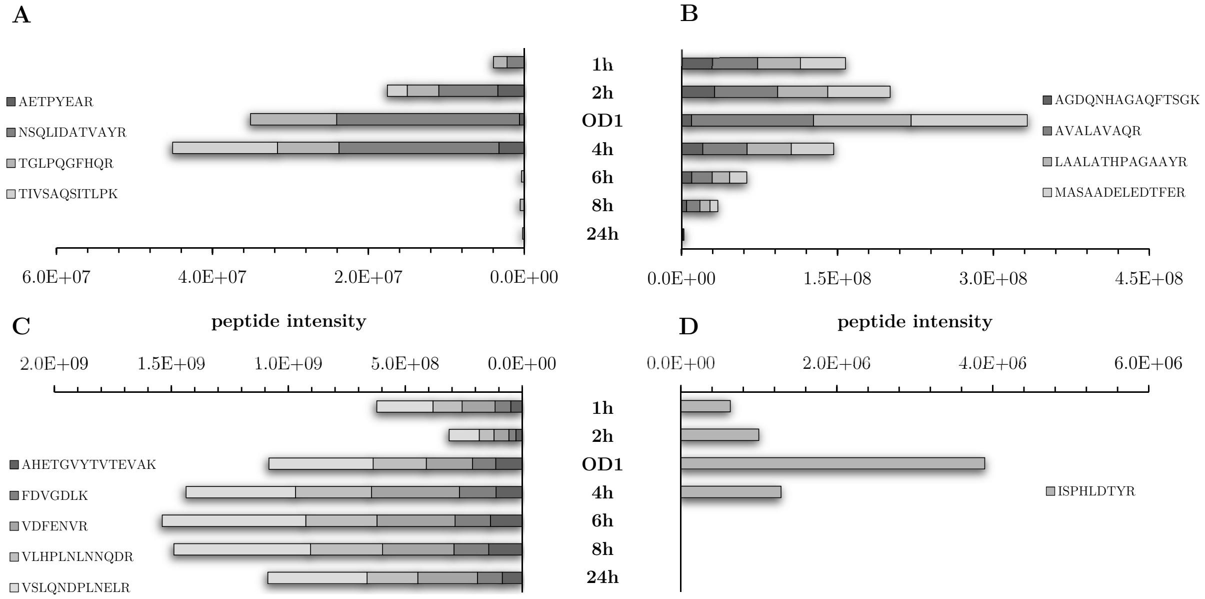


Figure 42. Intensities of all peptides measured for (**A**) Tle3, (**B**) Olg1, (**C**) PA1383 and (**D**) Olg2 using targeted proteomics. Summarized intensities of all peptides validated by PRM-MS in *P. aeruginosa* PAO1 samples harvested after 1 h, 2 h, 4 h, 6 h, 8 h and 24 h of cultivation as well as at $OD_{600nm} = 1$ are shown.

3.2.6.3.5 Evolutionary analyses of *olg1* and *olg2*

Blastp analysis of *olg1* and *olg2* revealed homologues which were mainly restricted to the genus *Pseudomonas*. A similar observation was made for PA1383, the mother gene of *olg2*. In contrast, homologues of *tle3* were detected in diverse phyla including Proteobacteria. The presence within the genus *Pseudomonas* and the absence in other genera of the order *Pseudomonadales* suggests that the *tle3* gene was horizontally transferred from a different order. The full-length ORFs of *olg1* and *olg2* as found in strain *P. aeruginosa* PAO1 were predominantly limited to the species with few exceptions (**Supplementary Figure 30**) suggesting that both OLGs are of comparatively young age and evolved rather recently.

Olg1 and *olg2* were longer than expected by random chance, according to the results of codon permutation and synonymous mutation tests following the methods of Schlub *et al.* (2018). For the first, the codons of the mother gene are randomly shuffled across the gene and the resulting ORF lengths in alternative frames were determined. The synonymous mutation test, in contrast, is based on the exchange of synonymous mutations at each site in the mother gene, thereby altering the sequence and length of the alternative frame ORFs when new stops are introduced in alternative reading frames. When applying both tests to the mother genes *tle3* and PA1383, considerably shorter ORF lengths for alternative frames were obtained than expected based on the AA composition of the mother gene. With p-values less than $10^{-10,}$ the ORFs of *olg1* and *olg2* were significantly larger than expected according to the synonymous mutation test (**Figure 43A**). The codon permutation test confirmed these results (**Figure 43B**), however, p-values of 0.163 (*olg1*) and 0.0635 (*olg2*) indicated that the obtained differences were not significant. Altogether, these findings suggest that the observed ORF lengths of *olg1* and *olg2* are the result of purifying selection acting on stop codons rather than being a sole side effect of the mother ORFs sequence.

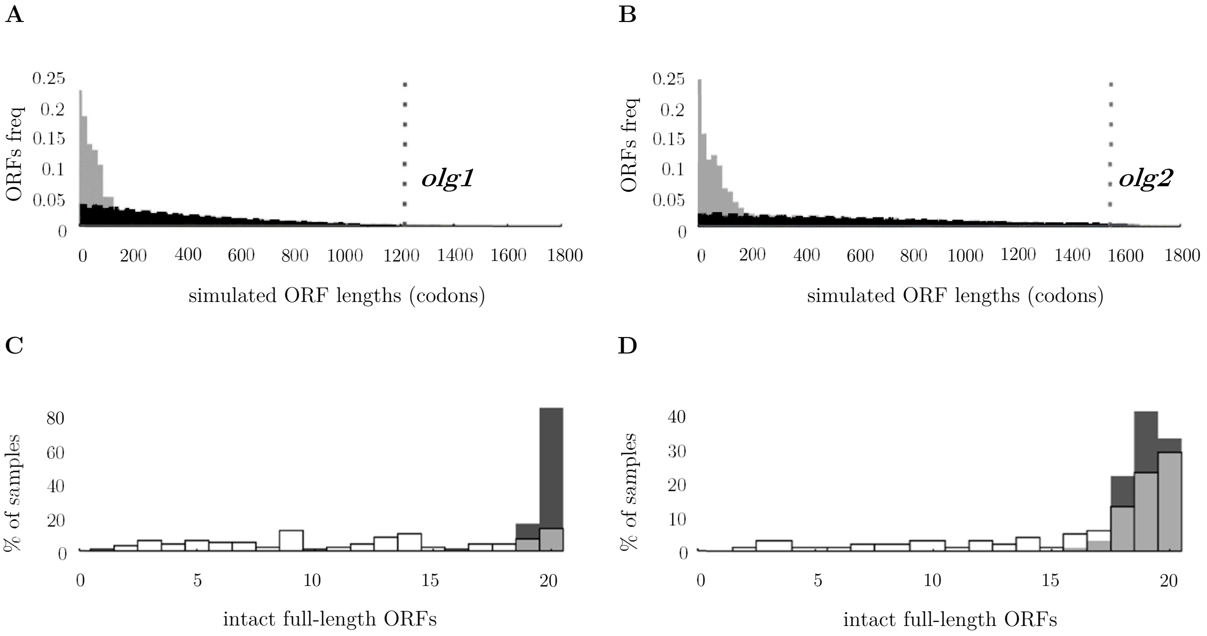


Figure 43. Stop codon depletion as indicator for purifying selection acting on *olg1* and *olg2*. Lengths obtained for the -1 frame after introduction of synonymous mutation into the mother genes (**A**) *tle3* or (**B**) PA1383 are indicated by grey bars; those of the permutation test after shuffling of mother gene codons are highlighted in black, respectively. The dotted line represents the length of the overlapping ORFs of (**A**) *olg1* and (**B**) *olg2* as observed in the *P. aeruginosa* PAO1 genome measured from stop codon to stop codon. Length of alternative frames after simulation of evolution of the mother genes *tle3* and PA1383 using an empirical codon model are seen for (**C**) *olg1* and (**D**) *olg2* . Evolution was simulated along the phylogenetic tree for the OLG clade, which was rooted on an outgroup harbouring a full-length ORF. Simulated sequences (unfilled bars) were on average shorter than the naturally observed sequences (black-filled bars) indicating evolutionary conservation of the overlapping ORFs. Figure adapted from Kreitmeier *et al.* (2021).

Another indication of selection against stop codons at the *olg1*/*tle3* and *olg2*/PA1383 loci was provided by simulations of the mother genes` evolution while disregarding any evolutionary forces acting on the overlapping ORFs. This provides a negative control, showing the result of the absence of selection. This method was originally applied by Cassan *et al.* (2016) in order to confirm a selection pressure to maintain an ORF overlapping the *env* gene in HIV-1. When applying this method to the mother genes *tle3* and PA1383 and simulate evolution along phylogenetic trees (**Supplementary Figure 30**), stop codons emerged more often in sequences with simulated evolution compared to those with natural evolution. As a result, the overlapping ORFs *olg1* and *olg2* were more frequently interrupted by stop codons, and thus, a lower number of simulated sequences showed the full-length ORFs than observed in the natural sequences of *P. aeruginosa* PAO1 (**Figure 43CD**).

In addition to stop codon depletion, synonymous variability within the mother genes *tle3* and PA1383 was analysed using the framework FRESCo (Sealfon *et al.*, 2015). Reduced synonymous variation was detected within those regions of the mother genes which overlapped with *olg1* (**Figure 44A**) and *olg2* (**Figure 44B**). In contrast, genomes containing the mother genes but lacking the intact OLGs exhibited a synonymous rate of approximately 1 indicating the absence of a synonymous constraint. A paired two tailed t-test confirmed that the difference between the results obtained for genomes harbouring the full-length ORF and genomes lacking the intact ORF was statistically significant for the *olg1* region (p-value = 0.086). For *olg2*, a significantly increased constraint was observed within a 350 codon region at the 3` end of the ORF (p-value = 0.03), but not in the upstream region. The tool OLGenie (Nelson *et al.*, 2020) also suggested purifying selection at the OLG loci by calculating an OLG-adapted ratio of non-synonymous and synonymous variants, referred to as dNN/dNS ratio. This ratio specifies the number of nucleotide changes with non-synonymous effect on both the mother and the overlapping gene (dNN) divided by the number of substitutions with a non-synonymous effect on the mother and a synonymous effect on overlapping gene (dNS), with both measures normalized by the number of sites of the respective class. OLGenie analysis resulted in dNN/dNS ratios below 1 over large parts of the intact OLG region within the mother genes *tle3* (**Figure 44C**) and PA1383 (**Figure 44D**) implying that synonymous mutations were favoured. In contrast, genomes without intact *olg1* and *olg2* were shown to have higher dNN/dNS ratios of around 1 (*olg1*: 1.02; *olg2*: 0.92), and thus, were not under purifying selection. Pairwise sequence comparisons of the *olg1*/*tle3* as well as of the *olg2*/PA1383 loci between different genomes showed that the strongest evidence for purifying selection was within the *P. aeruginosa* clade (**Supplementary Figure 31**). In sum, all results combined indicate an evolutionary sequence constraint and purifying selection on *olg1* and *olg2*.

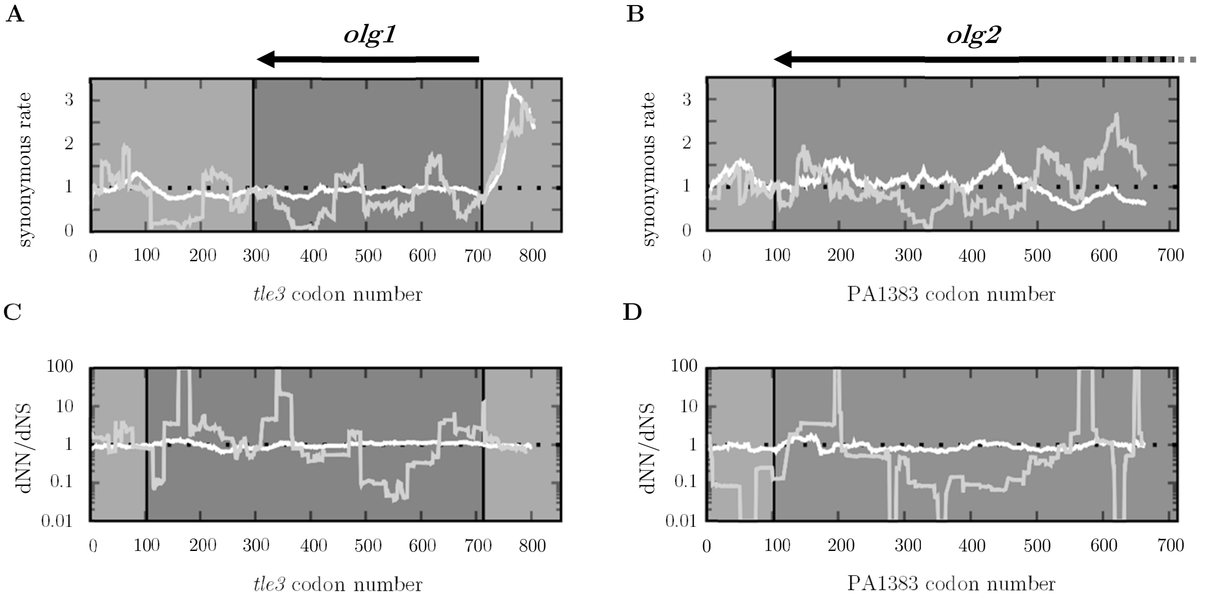


Figure 44. Sequence constraint analysis of *olg1* and *olg2* using (**A, B**) FRESCo and (**C, D**) OLGenie. The synonymous rates of the mother genes (**A**) *tle3* and (**B**) PA1383 highlighted by the grey lines were reduced within the OLG region (indicated by dark grey boxes and arrows) implying that these regions are under synonymous constraint. In contrast, genomes lacking full-length *olg1* and *olg2* exhibited synonymous rates (white line) equivalent to the expected value of 1 indicative of neutral evolution (dotted line). The dNN/dNS ratio (grey line) as measured by OLGenie was reduced within the OLG region (dark grey boxes) of genomes harbouring an intact (**C**) *olg1* and (**D**) *olg2*, whereas non-OLG genomes exhibited a ratio (white line) similar to neutral evolution (dotted line). Synonymous rates as well as dNN/dNS ratios were calculated in regions of 50 codons, each. Figure adapted from Kreitmeier *et al.* (2021).

4. Discussion

4.1 High-throughput discovery of novel genes – strengths and limitations

Knowledge about the entire coding capacity of bacterial genomes is crucial to improve our understanding of genome complexity and the ongoing biological processes including evolution. However, the detection and characterization of novel genes and their encoded protein products can be challenging, especially when genes are short (Warren *et al.*, 2010) and encode for low abundant proteins (Hemm *et al.*, 2020). Overlapping genes in prokaryotes are often associated with low gene expression (e.g., Fellner *et al.*, 2015) and short length (e.g., Zehentner *et al.*, 2020a). The latter was also confirmed by the results of this study since the detected gene candidates were on average shorter than annotated genes of *E. coli* LF82 (**Figure 24**) and *P. aeruginosa* PAO1 (**Figure 35**). As such, it is not surprising that many overlapping genes have escaped annotation. In addition to their small size, limited evolvability of two or more ORFs encoded at the same locus further contributed to the general disregard of overlapping genes. NGS-based methods offer the possibility for a highly sensitive and size-independent detection of gene expression, and thus, work well for the identification of such genes. In contrast to many low-throughput methods, NGS techniques do not rely on prior knowledge about the sequence to be detected and are less labour-intensive while enabling a genome-wide analysis of gene expression. In the following, the assets and drawbacks of all high-throughput methods used in this study are detailly discussed regarding their success in novel gene delineation as described in this thesis.

4.1.1 Transcriptome sequencing & Cappable-seq

Sequencing experiments using NGS platforms have revolutionized the field of molecular biotechnology by enabling cost-effective access to a variety of possible application including whole genome-analysis, detection of phenotypic variation, epigenetic analysis and so on (for review see Lee *et al.*, 2013). Especially RNA-seq has become particularly important in whole transcriptome studies to discover RNA transcripts and quantify their abundance. One major aspect to consider in RNA-seq experiments is the sequencing depth necessary for meaningful data analysis. A moderate read coverage is usually sufficient for the analysis of highly expressed genes, whereas sequencing depth has to be increased for accurate identification and quantification genes with low expression (Sims *et al.*, 2014, Mortazavi *et al.*, 2008). Haas *et al.* (2012) proposed a sequencing depth of 5 to 10 mio mRNA reads to optimally capture genes of variable expression strength; a further increase in sequencing coverage up to 50 mio mRNA reads did not result in a significant improvement in the detection of novel transcripts. In this study, we followed the recommendations by Haas *et al.* (2012) and obtained RNA-seq sequencing depths of 5.9 to 29.2 mio mRNA reads for *E. coli* LF82 (**Table 12**) and 3.5 to 11 mio mRNA reads for *P. aeruginosa* PAO1 (**Table 17**). Despite enhancing sequencing depth, high mRNA coverage values can also be achieved by the efficient depletion of rRNA prior to sequencing. For the experiments of this study, rRNA depletion was either carried out using the Ribo-Zero Kit by Illumina or the *Pseudomonas*-specific riboPOOL probes by siTOOLs Biotech. Depletion was successful since reads mapping to rRNA regions were reduced compared to the theoretically expected range of 80-95% which was previously reported for transcriptome sequencing in bacteria (Haas *et al.*, 2012, Giannoukos *et al.*, 2012, He *et al.*, 2010). However, substantial numbers of reads mapped to regions opposite to *rRNA* genes (**Supplementary Figure 1**), thereby suggesting a carryover of rRNA removal probes during depletion. Despite negatively affecting mRNA read coverage, rRNA depletion had no major effect on data quality as indicated by a high reproducibility of RPKM values measured for all anORFs. With Pearson's *r* ranging from 0.84 to 0.99, reproducibility between biological replicates was excellent.

As mentioned earlier, the global discovery of novel transcripts is a major application of RNA-seq and has already been applied to a variety of different bacteria, e.g., *P. aeruginosa* (Gómez-Lozano *et al.*, 2012), *Salmonella typhi* (Perkins *et al.*, 2009) and *Mycoplasma pneumoniae* (Güell *et al.*, 2009). Several possibilities have been used to delineate functional transcriptional events including the application of pre-defined threshold. Beaume *et al.* (2010), for instance, used a certain coverage threshold to detect actively transcribed regions within the genome of *S. aureus*, whereas others defined a minimum RPKM cut-off to differentiate genuine signals from background noise (Landstorfer *et al.*, 2014, Mortazavi *et al.*, 2008). Applying a certain threshold, though, has the disadvantage of a limited detection of transcripts which do not comply with this threshold, e.g., transcripts of low abundance. Alternatively, bioinformatic tools like Rockhopper (McClure *et al.*, 2013) or READemption (Förstner *et al.*, 2014) were specifically developed for bacterial RNA-seq analysis and they often combine multiple functionalities including read processing, mapping, differential expression analysis, etc. in one program. With increasing number of conducted RNA-seq experiments in bacteria, the phenomenon of pervasive transcription became of special interest. Pervasive transcription, designated as transcription from non-coding regions, was initially thought to be non-functional noise arising either from spurious promoters or from a transcriptional read through at terminators (Wade & Grainger, 2014). As pervasive transcription is frequently – but not exclusively – detected antisense to annotated genes in bacteria, the term pervasive transcription is often synonymously used with the term antisense transcription (for review see Lybecker *et al.*, 2014). Nowadays, though, several examples of antisense, yet functional transcripts have been reported in diverse organisms (reviewed in Lejars *et al.*, 2019, Thomason & Storz, 2010) and contribute to the complexity of bacterial transcriptomes. The widespread occurrence of pervasive transcription raises the question how many of the RNA-seq signals detected outside of annotated regions are indeed non-functional. A regulated expression of a novel transcript, either coding or not, provides a first clue for a potential functionality and helps to discriminate between functional and non-functional transcripts. Differential expression analysis has the power to identify genes, which differ significantly in their expression across different conditions based on RNA-seq raw data. The typical workflow for the elucidation of differential expressed transcripts usually starts with mapping of the raw data, quantification of read counts and sample normalization followed by a statistic test to identify genes with varying abundance (Oshlack *et al.*, 2010). In this study, statistical testing was performed using edgeR (Robinson *et al.*, 2010) in order to detect annotated gens in *E. coli*, which were differentially regulated as a function of oxygen availability. Some of the genes shown to be of changing abundance under the conditions tested in this study were also identified in a similar study by Bayramoglu *et al.* (2017), confirming the reliability of the observed results and the suitability of the generated datasets for the expression analysis of novel transcripts.

Further developments in the field of transcriptome sequencing also contributed to precise transcript mapping, and thus, helped to differentiate between genuine transcriptional signals and non-functional signals. Developed by Sharma *et al.* in 2010, the method of differential RNA-seq (dRNA-seq), for instance, facilitated nucleotide-precise mapping of 5´ transcript boundaries, thereby supporting the identification of novel transcripts in a variety of different organisms including archaea (e.g., Laass *et al.*, 2019) and bacteria (e.g., Thomason *et al.*, 2015, Albrecht *et al.*, 2011, Filiatrault *et al.*, 2011). Of central importance in the dRNA-seq procedure is the terminator 5'-phosphate-dependent exonuclease (TEX). This enzyme degrades processed transcripts with monophosphorylated 5`ends like rRNAs and tRNAs while preserving primary transcripts including mRNAs or small RNAs carrying a tri-phosphate group at the 5´end. After sequencing, mapping of the primary transcripts, indirectly enriched by the TEX treatment, should aid the selective mapping of TSS in a global manner. However, as RNA secondary structures are known to resist the exonuclease treatment (Jäger *et al.*, 2014, Zhelyazkova *et al.*, 2012), TEX libraries require an undigested control library to evaluate signal specificity. The recently established method Cappable-seq was shown to accurately determine TSS without the need for a control library (Ettwiller *et al.*, 2016). In contrast to dRNA-seq, the Cappable-seq procedure directly enriches primary transcripts by enzymatic biotinylation and subsequent streptavidin bead capture, resulting in an increased specificity for primary transcripts compared to other TSS methods (Ettwiller *et al.*, 2016). As a side effect of primary transcript enrichment, rRNA and tRNA species are specifically depleted. Ettwiller *et al.* (2016) observed rRNA and tRNA percentages as low as 4% for *E. coli*, whereas rRNA and tRNA reads accounted for 86% in an untreated control library. Analysis of the Cappable-seq data in this study also revealed a pronounced depletion of rRNA and tRNA transcripts as indicated by low mapping percentages of 5.9 to 7.5% for all three biological Cappable-seq replicates of *P. aeruginosa* PAO1. The small fraction of residual rRNA and tRNA reads could possibly tracked back to a limited specificity of the enzymes used for capping as proposed for Cappable-seq in EHEC (unpublished data; personal communication with Dr. Barbara Zehentner). The slightly higher rRNA and tRNA read percentages could also originate from organism-specific differences. This assumption, though, could not be confirmed with experimental data since Cappable-seq has not been implemented for *Pseudomonas* strains so far. However, since differences in the rRNA and tRNA percentages between the original data published by Ettwiller *et al.* (2016) and the data of this study were only marginal and Cappable-seq libraries were subjected to ultra-deep sequencing (**Table 18**), the amount of mRNA reads sufficiently aided the detection of even weak TSS. With pairwise Pearson´s *r* values of at least 0.99, reproducibility of genome-wide RRS values for all three biological replicates was excellent and even higher than the technical reproducibility observed in the original publication (Ettwiller *et al.*, 2016). After determining an optimal RRS score of 1.5 for the reliable identification of TSS, 9,205 TSS were detected in all three replicates. This number was substantial lower than the number described by Ettwiller *et al.* (2016) for *E. coli* (16,539). However, comparability of the datasets might be limited due to the use of different organisms and the absence of TSS validation in *E. coli* by biological replicates. Filiatrault *et al.* (2011), in contrast, identified a substantial lower number of 2,510 TSS in a TEX-supported RNA-seq experiment in *Pseudomonas syringae*. However, since 2,117 TSS were already associated with transcription units solely spanning anORFs in *P. aeruginosa* PA14 (Wurtzel *et al.*, 2012), it seems likely that the actual number of TSS in *P. aeruginosa* PAO1 is somewhat higher. In this study, 2,369 TSS were reproducibly predicted in the upstream proximity of anORFs in all three biological replicates after applying a minimum RRS threshold of 1.5. This number was only marginally increased by 7.5% when TSS detection in two out of three biological replicates was sufficient for reliable TSS determination, which again confirmed the high robustness of the Cappable-seq datasets. A subset of 1,207 anORF-associated TSS identified in this study were also detected by Wurtzel *et al.* (2012) for homologous genes in *P. aeruginosa* PA14. About 55% of the TSS shared by both *P. aeruginosa* strains were located at the exact same distance to the start codon, and further 14% showed no more than 10 nt deviation from the position of the TSS obtained for the respective homologue in the other strain. The high correspondence between the results of this study and the results by Wurtzel *et al.* (2012) suggests conservation of TSS across different *P. aeruginosa* strains, and thus, confirms that a substantial percentage of TSS have been identified correctly by Cappable-seq. In addition, when analysing the nucleotide sequence upstream of anORF-associated TSS, a clear TA-rich sequence motif was observed indicating the presence of -10 elements of σ^{70} promoters. Due to the high precision and reproducibility of the Cappable-seq results obtained for the anORFs of *P. aeruginosa* PAO1, these datasets seem to be highly suitable for TSS detection of novel transcripts as well.

4.1.2 Classical ribosome profiling

Ultra-deep RNA-seq in combination with accurate TSS mapping give a precise insight into the transcriptional landscape of an organism but is not necessarily suggestive of whether a transcribed genomic region is translated or not. The detection of thousands of TSS located within or antisense to annotated regions across diverse bacteria genomes (e.g., Thomason *et al.*, 2015, Sharma *et al.*, 2010) called the biological relevance and specificity of pervasive transcription into question (Lloréns-Rico *et al.*, 2016, Raghavan *et al.*, 2012). Although antisense transcription has frequently been associated with regulatory non-coding RNAs (Eckweiler & Häussler, 2018, Dornenburg *et al.*, 2010), recent studies demonstrated translation of some antisense transcripts. (de Almeida *et al.*, 2019, Weaver *et al.*, 2019, Friedman *et al.*, 2017). Ribosome profiling (Ingolia *et al.*, 2009) made a substantial contribution to the deciphering of translated RNA species by selective sequencing of ribosome-covered mRNAs. Using this method, a multitude of antisense transcripts have been reported to be ribosome-associated implying their translation (e.g., Zehentner *et al.*, 2020a, Ardern *et al.*, 2020). This study also focused on Ribo-seq data for the identification of novel, translated genes. Multiple Ribo-seq experiments were carried out under varying conditions followed by deep sequencing. As Glaub *et al.* (2020) observed a saturation in the number of detectable genes when read counts exceeded 20 mio after rRNA and tRNA removal, we strived for a sequencing depth of approximately 20 mio mRNA reads in our Ribo-seq experiments. For this purpose, we increased sequencing depth compared to the RNA-seq experiments and yielded 6.9 to 21.2 mio and 14.9 to 35.2 mio mRNA reads for *E. coli* LF82 and *P. aeruginosa* PAO1 Ribo-seq experiments, respectively. The reproducibility between biological replicates of the Ribo-seq experiments was on average lower than those observed for RNA-seq replicates. A similar finding was reported by Hücker (2018) upon analysis of RNA-seq and Ribo-seq data in EHEC. Despite biological variance, differences in sequencing depth (Diament & Tuller, 2016) and biases arising from multiple experimental steps including translational inhibition and monosome fractionation (Aeschimann *et al.*, 2015, Hussmann *et al.*, 2015) were reported to affect data quality, and thus, may be accounted for the overall lower reproducibility of Ribo-seq data. Of central importance in Ribo-seq is also the choice and quantity of the nuclease used for the generation of RFPs (Aeschimann *et al.*, 2015). RNase I, the most common nuclease in eukaryotic ribosome profiling, delivers precise cleavage without sequence specificity (delCardayré & Raines, 1995). Although this nuclease has been successfully used in Ribo-seq experiments in EHEC EDL933 (Neuhaus *et al.*, 2016, Neuhaus *et al.*, 2017), it was claimed to be inactive in *E. coli* (Kitahara & Miyazaki, 2011, Datta & Burma, 1972), resulting in a need for alternative nucleases in this bacterial species. MNase isolated from *Staphylococcus aureus* is commonly used for nuclease footprinting in bacterial Ribo-seq (e.g., Grady *et al.*, 2017, Li *et al.*, 2012, Oh *et al.*, 2011), but also other nucleases or nuclease mixtures have been applied in the past (Hücker *et al.*, 2018a, Neuhaus *et al.*, 2017, Gerashchenko & Gladyshev, 2017). A careful choice of the nuclease to be used in Ribo-seq as suggested by Gerashchenko & Gladyshev (2017) as well as an adjustment of the enzyme concentration and the digestion period is advisable to yield robust data while preserving the structural integrity of the ribosome. The importance of nuclease selection was also emphasized in this study: The five-nuclease mixture including RNase I, which was previously used for Ribo-seq in EHEC (Hücker *et al.*, 2018a) and successfully implemented for *E. coli* LF82 in this study, led to a massive degradation of rRNA in preliminary Ribo-seq experiments in *P. aeruginosa.* As a result, the number of reads mapping to mRNA was substantially reduced although rRNA removal was performed (see section 3.2.1.1). Incubation of *Pseudomonas* RNA with each of the nucleases separately revealed that even small concentrations of the endoribonuclease RNase I severely affected rRNA integrity (**Supplementary Figure 13B**) indicating incompatibility of this nuclease with *P. aeruginosa* ribosomes. Gerashchenko & Gladyshev (2017) also observed varying tolerance of ribosomal populations originating from different organisms as a function of nuclease selection; however, they mainly focused on different eukaryotic species in their analysis. By removing RNase I from the nuclease mixture and reducing the concentration of the other nucleases by 30%, we increased the mRNA yield of *P. aeruginosa* Ribo-seq experiments by a factor of 2 to 3. In addition, as rRNA quality worsened during stationary cultivation, samples harvested at late exponential/early stationary phase were used for Ribo-seq, which again resulted in an increase in mRNA yields by factor 3. Despite these improvements, overall mRNA percentages of mapped read were lower for *P. aeruginosa* PAO1 than *E. coli* LF82, even when cultivated under optimal conditions. However, this could have been influenced by many factors, e.g., deviating rRNA depletion efficiencies, which were not subject of this study.

Data evaluation is another key step in the Ribo-seq procedure. Up to now, numerous eukaryotes have been subjected to Ribo-seq (Bazzini *et al.*, 2014, Dunn *et al.*, 2013, Ingolia *et al.*, 2011, Ingolia *et al.*, 2009) and methods for the delineation of translated ORFs based on the resulting data have been developed (e.g., Erhard *et al.*, 2018, Xiao *et al.*, 2018, Crappé *et al.*, 2015, Fields *et al.*, 2015, Chew *et al.*, 2013, Lee *et al.*, 2012). Many of the developed tools rely on certain Ribo-seq signatures which are characteristic for eukaryotic Ribo-seq data, for instance, triplet periodicity (e.g., Calviello *et al.*, 2016, Michel *et al.*, 2012). Since these signatures are often less pronounced in prokaryotes, those programs are unsuitable for translated ORF delineation in prokaryotes. In addition to deviating Ribo-seq-based characteristics, variation in the experimental procedure (Mohammad *et al.*, 2016) as well as differences in gene architecture between pro- and eukaryotes limit the applicability of these tools on bacterial Ribo-seq data (Ndah *et al.*, 2017). As Ribo-seq more and more became a powerful tool to study translational events in bacteria, the programs REPARATION (Ndah *et al.*, 2017) and DeepRibo (Clauwaert *et al.*, 2019) as well as the algorithm described by Giess *et al.* (2017) were developed in order to identify protein-coding genes within prokaryotes. These tools take advantage of different metrics inherent to bacterial Ribo-seq data. REPARATION, for instance, utilizes start and stop RPKM values, coverage values of the entire ORF and of the start region, the proportion of read accumulation as well as the ribosome binding site energy for translated ORF prediction (Ndah *et al.*, 2017). DeepRibo, in contrast, uses a recurrent neural network to process Ribo-seq data of candidate ORFs and combines the results with information derived from a 30 nt long DNA stretch around the SD sequence processed by a convolutional neural network (Clauwaert *et al.*, 2019). The method described by Giess *et al.* (2017) relies on multiple features for TIS prediction including calculation of normalized 5`read counts, the proportion of 5´reads in the up- and downstream vicinity as well as the ratio thereof (Giess *et al.*, 2017). All tools predict translated ORFs without *a priori* knowledge about the genome annotation, and thus, are not biased towards non-canonical signals, e.g., the translation of overlapping ORFs. REPARATION (Ndah *et al.*, 2017), DeepRibo (Clauwaert *et al.*, 2019) as well as the scripts by Giess *et al.* (2017) were also applied to the Ribo-seq datasets of this study. As the three tools cover different functions and exhibited limited reproducibility and coherence of the predicted results (see sections 3.1.4.1 and 3.2.6.1), all outputs were combined in order to achieve a maximum identification of novel genes.

Numerous novel ORFs were predicted to be translated based on Ribo-seq data in bacteria (Smith *et al.*, 2019, Weaver *et al.*, 2019, Jeong *et al.*, 2016). However, the ubiquity of translational signals outside annotated regions across many bacterial species as well as the surprising finding of ribosome-covered RNAs, which were formerly thought to be non-coding (Friedman et al., 2017, Neuhaus et al., 2017), led to a controversial debate about the specificity of the translational signals observed. Structured RNAs (Fremin & Bhatt, 2020) as well as RNA associated with non-ribosomal proteins (Ji *et al.*, 2016, Ingolia *et al.*, 2014) were discussed to be the source of artificial Ribo-seq reads, probably protected from RNase cleavage due to a reduced nuclease accessibility. Alternatively, Smith *et al.* (2019) suggested that bacterial transcripts are subjected to pervasive translation, a process equivalent to pervasive transcription assumed to be noise rather than bearing an function. The absence of clear ribosome signatures e.g., start and stop codon peaks or a triplet periodicity (Fields *et al.*, 2015, Bazzini *et al.*, 2014) as reported for translated ORFs in eukaryotes further hampers differentiation between functional and non-functional translational signals in prokaryotes. However, recent modifications of classical Ribo-seq like Ribo-RET (Meydan *et al.*, 2019) or RelE-supported Ribo-seq (Hwang & Buskirk, 2017) enhance resolution and validity of bacterial Ribo-seq experiments, and thus, are an asset for the detection of true cellular translational events.

4.1.3 Modified variants of ribosome profiling

4.1.3.1 RelE-assisted Ribo-seq for reading frame analysis

As mentioned earlier, detection of a triplet periodicity in the mapping of Ribo-seq reads is a hallmark for genuine translation in eukaryotes and facilitates distinction between RFPs originating from translating ribosomes and signals arising from nonspecific artefacts. The detection of a trinucleotide periodicity in eukaryotes can mainly be traced back to the use of RNase I for RFP generation. This nuclease offers precise cleavage without sequence specificity (delCardayré & Raines, 1995). In prokaryotes, though, the reading frame signal seems to be blurred by multiple factors including the cutting preference of the nuclease used for digestion. MNase, the most frequently used nuclease in bacterial Ribo-seq, exhibits significantly higher cleavage rates before A and T nucleobases (Dingwall *et al.*, 1981), and thus, does not trim accurately to the 5´ and 3´ ends of the ribosome-covered footprints. Consequently, a sequence bias at both ends of the RFPs is introduced limiting the resolution and power of such experiments. Other nucleases like RNase A (Gerashchenko & Gladyshev, 2017) or RNase I, as observed in this study, did not preserve structural integrity of the bacterial ribosome or showed an even more pronounced sequence preference. The attempt to use multiple nucleases in order to get precise RFP ends was also only partly successful in the detection of a periodicity signal in EHEC (Hücker *et al.*, 2017). A further factor which hampers reading frame determination in bacteria is the heterogeneous length of RFPs. RNase I yields a robust 28 nt long RFP population in eukaryotic ribosome profiling (Hsu *et al.*, 2016, Ingolia, 2010), whereas bacterial RFPs are less uniform in size (Mohammad *et al.*, 2019). Up to now, no consensus about the actual size of the bacterial RFP has been reached in literature. Some studies included shorter RPFs starting from ~18 nt (Andreev *et al.*, 2017), other studies isolated RFPs with a length of 21 nt for *Mycobacterium bovis* (Ngan *et al.*, 2021), 20-30 nt (Balakrishnan *et al.*, 2014) or even longer (Li *et al.*, 2012, Oh *et al.*, 2011). On the one hand, the heterogeneity in bacterial RFP length might be attributed to the impreciseness of the nuclease used for mRNA digestion. Cleavage with MNase, for instance, does not only introduce a sequence bias at the 5´ and 3´ ends of RPFs but also results in broad distribution of RPFs which further restricts reading frame analysis (Hwang & Buskirk, 2017). On the other hand, deviating read lengths may arise from the bacterial ribosome itself due to its distinct conformational flexibility. O'Connor *et al.* (2013), for example, showed that the interaction between rRNA and mRNA alters the length of RFPs. In addition, Mohammad *et al.* (2016) confirmed that longer RFPs originate from initiating ribosomes through the interaction between SD and aSD motifs. Due to their variable lengths, it was suggested to select a broader range of fragments between 10 and 40 nt in order to capture the whole variety of RFPs (Mohammad *et al.*, 2019, Mohammad *et al.*, 2016).

The usage the nuclease RelE was recommended by Hwang & Buskirk (2017) to get more precise information about the reading frame in bacteria. In an *in vitro* Ribo-seq experiment with MNase and RelE, a highly resolved reading frame signal in the sum of all anORFs in *E. coli* K-12 was achieved by exploiting RelE´s unique feature of cleaving translating mRNA within the ribosomal A site codon (Pedersen *et al.*, 2003). In order to implement RelE-assisted Ribo-seq in *E. coli* LF82 and *P. aeruginosa* PAO1, the RelE toxin had to be overexpressed and purified. However, preparation of this toxin turned out to be challenging due to the intrinsic properties of RelE. Sole overexpression of RelE, for instance, failed in a first approach due to the inhibitory effect of RelE on cell growth (Gotfredsen & Gerdes, 1998) and necessitated the co-expression of RelE with its cognate antitoxin RelB. Furthermore, as both RelE as well as RelB constitute small proteins of 11.2 (Gotfredsen & Gerdes, 1998) and 9 kDa (Bech *et al.*, 1985), respectively, several methodological adaptions were necessary to detected these proteins via SDS-PAGE and Western blot (see section 2.2.4). In addition, the blurred band pattern and the discrepancy between the expected and the observed size of RelE on SDS gels (**Figure 16 & 17**) impeded the examination of the expression and purification success. However, discrepancies between the theoretical molecular mass and the protein size as determined by SDS gels are frequently described in literature. One prominent example for such a discrepancy is the tumour suppressor protein p53, which was given this name because SDS-PAGE analysis indicated a molecular mass of 53 kDa, whereby the real mass of the p53 protein is 43.7 kDa. A possible explanation for this observation is that a proline-rich region within p53 decreased gel mobility and therefore the molecular mass was overestimated (Levine & Oren, 2009). In general, the amino acid composition of the protein to be analysed seems to have a great impact on the gel mobility. Guan *et al.* (2015), for example, were able to show that a high content of acidic amino acids, e.g., glutamate and aspartate, is linked to a retarded mobility of the protein in SDS gels. Other possible explanation regarding discrepancies between SDS-displayed and predicted masses are differences in the amount of SDS bound by the protein or conformational differences, e.g., aggregation of proteins due to overloading (Rath *et al.*, 2009, Reynolds & Tanford, 1970). Despite these difficulties, the presence of RelE in cell lysates and purification samples was successfully confirmed by MS analysis (e.g., **Figure 17B**). Another drawback of the co-expression procedure of RelE and RelB was the tight protein-protein interaction between both (Overgaard *et al.*, 2009) which necessitated multiple attempts and harsh denaturing conditions to adequately separate both proteins.

After successful preparation of RelE following the procedure described by Dunican *et al.* (2015) and Griffin *et al.* (2013), RelE was used for nuclease footprinting in *E. coli* LF82 alongside MNase. Reading frame analysis of the sequencing data (**Figure 21**) revealed a similar periodicity signal for the sum of all anORFs as shown by (Hwang & Buskirk, 2017). This result in combination with the observed accumulation of short reads (**Figure 19**) as previously reported for RelE (Hwang & Buskirk, 2017) indicated effective cleavage of mRNA, and thus, functionality of RelE. However, the resolution of the periodicity signal was reduced compared to the results described in the original publication. Possible reasons for a lower resolution of the reading frame signal in the datasets of this study could include experimental artefacts, e.g., arising from biases during library preparation. Adapter ligation (Fuchs *et al.*, 2015, Zhuang *et al.*, 2012, Hafner *et al.*, 2011), reverse transcription (Hansen *et al.*, 2010) and PCR amplification (Fu *et al.*, 2018) can favour the unspecific enrichment of reads depending on read sequence and structure, and thus, introduce biases during library preparation. Several studies showed that the Illumina TruSeq Small RNA Kit, which was used for library preparation in this study, is prone to such biases (Wright *et al.*, 2019b, Dard-Dascot *et al.*, 2018, Baran-Gale *et al.*, 2015). For example, T4 RNA Ligase 2 was shown to ligate adapter in a sequence-dependent manner (Jayaprakash *et al.*, 2011) and PCR amplification is also heavily influenced by the efficiency of the Phusion Polymerase (Quail *et al.*, 2012). In addition, multiple rounds of PCR amplification were carried out in order to increase the amount of material needed for sequencing, which might also have led to a disproportional increase of PCR duplicates. However, Fu *et al.* (2018) showed that the frequency of PCR duplicates strongly depends on the amount of input material used rather than on the number of PCR cycles. In future, indexing reads with additional barcodes as proposed by Shiroguchi *et al.* (2012) and Mir *et al.* (2014) might improve differentiation between reads duplicated during PCR and read arising from physiological events. Reanalysis of the data provided by Hwang & Buskirk (2017) also resulted in a reading frame signal which was less pronounced compared to the signal described in the original publication, suggesting that data processing (e.g., mapping or trimming parameters) or data analysis (e.g., reading frame analysis) must have been divergent. Despite lower resolution, the summarized reading frame signal for all annotated genes was distinct and significantly better than those previously reported for *E. coli* O157:H7 Sakai by Hücker *et al.* (2017). However, in this publication periodicity analysis was based on 5´ends which are known to deliver less precise information about the bacterial reading frame than the 3´ ends (Mohammad *et al.*, 2019, Woolstenhulme *et al.*, 2015). Evidence of RelE functionality in *E. coli* LF82 was provided by several control experiments as well. For instance, a comparable dataset prepared with a mixture of five nucleases showed no reading frame signal at all, neither when analysing NNC-shifted reads nor upon raw read analysis (**Figure 21**). The absence of a periodicity signal was also confirmed by RNA-seq data analysis. In this case, illegitimate read shifting resulted in a negligibly small accumulation of reads at the sub-codon position two, whereby the resolution was comparable to the results obtained for other RNA-seq datasets of this study and the study by Hwang & Buskirk (2017). The pronounced difference in the resolution of the signal obtained for Ribo-seq and RNA-seq confirmed that biases introduced during library preparation cannot solely explain the observed results. On those grounds, we have no doubt that the signal differences observed for our dataset prepared with RelE are due to mRNA cleavage by RelE. Hwang & Buskirk (2017) also performed periodicity analysis for the gene *prfB* and were able to visualize the well-reported +1 frame shift in this gene (Craigen & Caskey, 1986). To test whether the datasets of this study are also suitable for reading frame determination of single genes, we analysed the distribution of reads covering *ompA,* which was the gene with the highest expression in the Ribo-seq dataset. Indeed, a clear reading frame signal was observable when analysing the entire ORF as well as individual sections (**Figure 20**). However, a magnitude of reads mapping to position 1 or 3 indicated either an absent or an insufficient cleavage by RelE. The source of these reads could not be clarified unequivocally but could include the accumulation of non-specific reads or a skewed cleavage inherent to the sequence composition of the ORF. Contradictory to the results by Hwang & Buskirk (2017), Hurley *et al.* (2011) reported the highest frequency for RelE cleavage after the third position instead of after the second, which could further contribute to a distortion of the periodicity signal. Another aspect to consider in this context is the usage of not normalized raw reads for reading frame analysis. On the one hand, transcript-specific properties may have a greater impact on the resolution of single gene reading frames and less influence when analysing the reading frame in the sum signal of multiple genes. On the other hand, sum signals of multiple genes are always biased towards genes with high read counts and, therefore, with high expression, whereas the effects of low expressed genes covered by a small read number only are highly undervalued.

The gene encoding RelE is widely distributed across bacteria and archaea (Gerdes, 2000) and is also present in *P. aeruginosa* (Williams *et al.*, 2011). Although similarity and identity of RelE homologues in *E. coli* and *P. aeruginosa* as identified by pairwise sequence alignment using EMBOSS Needle (Needleman & Wunsch, 1970) was shown to be low (see section 3.2.4), Goeders *et al.* (2013) observed that overexpression of RelE originating from diverse bacterial phyla led to an inhibition of growth in *E. coli* due to the cleavage of translation-associated mRNA. In addition, the RelE-characteristic cleavage pattern including preferential cleavage before purine bases and after the second or the third nucleotide of the A site codon (Pedersen *et al.*, 2003, Christensen *et al.*, 2001) further suggested that RelE-like toxins, despite having low similarity and identity with the *E. coli*-originating RelE, exhibit a relaxed specificity and can exert their function even in distantly related bacterial species (Goeders *et al.*, 2013). Based on these observations, the *E. coli*-originating RelE protein prepared in this study was also applied to Ribo-seq in *P. aeruginosa* PAO1 and the resulting data was evaluated regarding RelE´s suitability for reading frame determination in this strain. Analogous to the results of *E. coli* LF82, analysis of raw reads in the dataset prepared with RelE delivered a predominant reading frame signal at sub-codon position 2 for the sum of all anORFs, which was even more pronounced upon NNC read shifting (**Figure 32**). In contrast, a control experiment prepared with MNase only showed a deviating pattern with sub-codon position 3 being the most prevalent read position. Hwang & Buskirk (2017) reported that minor periodicity signals obtained by using MNase only are exclusively caused by the sequence specificity of MNase in combination with the nucleotide bias of the ORF and consequently do not resemble the genuine reading frame signal obtained by translating ribosomes. In addition, NNC shifting of the control datasets completely erased any positional differences. Although these results are suggestive of RelE functionality in *P. aeruginosa*, it must be noted that the clear cleavage specificity of RelE described by Hwang & Buskirk (2017) was not detectable as indicated by the absence of an accumulation of nucleobase C before and nucleobase G after cleavage at the 3´end of RFPs (**Figure 32**). Consequently, the lawfulness of NNC read shifting severely affecting resolution of the periodicity signal can be challenged. However, no cleavage bias was seen in the *E. coli* datasets of this study either. In contrast, digestion with MNase in the control experiment confirmed preferential cleavage before the nucleobases A and T as expected (**Figure 32**; Hwang & Buskirk, 2017, Dingwall *et al.*, 1981). Further experiments like overexpression of the *E. coli*-derived *relE* gene in combination with Northern blot analysis of mRNA transcripts before and after *relE* induction would be conceivable to unequivocally confirm RelE activity in *P. aeruginosa.*

4.1.3.2 Retapamulin-supported Ribo-seq for TIS detection

Accurate mapping of TISs by exploiting ribosome-stalling antibiotics like tetracycline (Nakahigashi *et al.*, 2016), Onc112 (Weaver *et al.*, 2019) or retapamulin (Meydan *et al.*, 2019) enhances sensitivity and specificity of classical Ribo-seq, especially when methods are combined. As tetracycline was shown to arrest elongating ribosomes (Nagalakshmi *et al.*, 2008) and Onc112 was reported to create more variable read peaks in deviating distances to the start codons (Weaver *et al.*, 2019), retapamulin was used for TIS detection in this study. After constructing an *E. coli* LF82Δ*tolC* deletion mutant as well as receiving the strain *P. aeruginosa* PAO397, both strains were subjected to MIC testing alongside their wild type strains LF82 and PAO1. For both wild type strains, the MIC of RET exceeded the value of 32 µg/mL. Corbett *et al.* (2017) observed a lower MIC value of 8 µg/mL for the RET-treated *E. coli* K-12 derivate strain BW25113. However, MIC results were reported to be highly variable as the preciseness of the measurement is affected by several factors including strain and laboratory variability (Mouton *et al.*, 2018). As expected, *E. coli* LF82Δ*tolC* was more suspensible to RET treatment with an MIC value of 0.25 µg/mL, which was in a similar range as the values reported by Meydan *et al.* (2019) for two *E. coli tolC* deletion mutants (0.0125 µg/mL and 0.05 µg/mL). The MIC of *P. aeruginosa* PAO397 was twice as high as those of *E. coli,* which is in concordance with the finding that *P. aeruginosa* has in general a higher resistance against the pleuromutilin basic substance indicted by higher MIC values as measured by Kavanagh *et al.* (1951). After MIC determination, cultures of *E. coli* LF82Δ*tolC* and *P. aeruginosa* PAO397 were incubated with the 100-fold MIC for effective stalling of ribosomes. Metagene analysis of highly expressed genes revealed a pronounced redistribution of ribosomes upon RET treatment in both strains (**Figure 15 & 31**). An slight accumulation of ribosomes at start codons was also observed in the ND datasets which was hypothesized to be the result of either a reduced translational efficiency at the beginning of an coding sequence (Tuller *et al.*, 2010) or of a higher ribosome occupancy caused by increased ribosomal initiation rates (Shah *et al.*, 2013). Remarkably, the peak height and sharpness was reduced in the datasets of this study compared to the results published by Meydan *et al.* (2019). This observation was not only limited to the RET-treated sample but occurred in the ND datasets of both organisms as well, suggesting that experimental factors rather than biological aspects were causative for the less specific ribosome accumulation. Ongoing translation in the bacterial lysate, for instance, seems to be a possible source of reduced resolution. Thawing of the lysate for already 15 min was shown to result in a measurable translational activity in the absence of any drug-based inhibitor (Mohammad *et al.*, 2019). In addition, ribosomes are known to have varying conformations during different stages of translation resulting in variable RFP lengths (Lareau *et al.*, 2014). The entering of the elongation phase by stalled ribosomes could result in an additional population of reads with alternative length, which may impair with the correct positioning of the ribosomal P site after subtraction of a fixed offset from the 3´end. Consequently, imprecisely positioned reads could have led to a peak broadening at start positions. In future experiments, the application of elongation inhibitors like cycloheximide or high salt concentrations in the lysis buffer are recommended to efficiently arrest translation in the lysate (Mohammad *et al.*, 2019). Another aspect known to heavily influence translational dynamics, and thus, may lead to a blurred resolution of bacterial Ribo-seq data is the type of cell harvest. Commonly, cultures are either harvested by centrifugation after introduction of a translational arrest, e.g., by chloramphenicol or by dry ice, or are subjected to rapid filtration (for review see Glaub *et al.*, 2020). However, both methods are associated with sequence-specific stalling of ribosomes (Mohammad *et al.*, 2019). Bacterial cultures of this study were harvested by dry ice-treatment followed by centrifugation. This procedure could have influenced data quality by affecting translation initiation rates characteristic of severe bacterial stress (Gerashchenko & Gladyshev, 2014). Direct freezing of bacterial cultures in liquid nitrogen seems to be a promising alternative to obtain an optimal resolution in future Ribo-seq experiments (Mohammad *et al.*, 2019).

Possible caveats associated with the RET approach became clear when analysing RET and ND signals of the genes encoded by the S10 ribosomal protein operon. Only a limited correlation between the overall expression strength of the genes in the ND datasets and the normalized peak heights at the start codons in the RET datasets was observable, not only for the aforementioned genes but also for the entirety of anORFs in *E. coli* LF82 and *P. aeruginosa* PAO1. However, as the use of translational inhibitors constitutes stress for the cells, such inhibitors were frequently reported to alter ribosome coverage profiles (Gerashchenko & Gladyshev, 2014), and thus, interfere with expression strength measurements. In addition, further RET peaks located within the coding regions of annotated genes, e.g., in gene *rpIC* (**Figure 14**), questioned the specificity of the signals observed after RET treatment. Meydan *et al.* (2019) were able to assign a biological function to some of the internal TIS detected by Ribo-RET, suggesting that translation of ORFs starting at internal TISs may result in real proteins with varying functionalities. However, the authors did not rule out the possibility that cryptic TIS located both in inter- as well as intragenic regions represent unspecific noise due to an imprecise recognition of start codons by ribosomes. Alternatively, spurious TIS peaks could also be the results of an increased number of free ribosomes binding unspecifically to untranslated start codons, or of a ribosomal run-off upon RET treatment, which reveals mRNA regions typically covered by elongating ribosomes in the absence of RET. Therefore, careful evaluation of the Ribo-RET data is necessary for differentiating between biological relevant RET signals and artificial signals without function.

4.1.4 Mass spectrometry

Ribo-seq experiments provide a good insight into which transcripts are translated but cannot answer the question whether these transcripts encode stable peptides. MS is a widely used technique for whole proteome analysis and aided the detection of numerous novel genes in diverse organism (e.g., Zhang *et al.*, 2019, Brosch *et al.*, 2011, Bitton *et al.*, 2010, Gupta *et al.*, 2007, Kruft *et al.*, 2001) while showing better correlation with Ribo-seq than with RNA-seq data (Blevins *et al.*, 2019, Ingolia *et al.*, 2009). However, MS reaches its limits when analysing proteins without or insufficient proteolytic cleavage sites. In this context, small proteins are difficult to detect due to their limited number or the entire absence of tryptic cleavage sites (Petruschke *et al.*, 2020, Müller *et al.*, 2010). Despite methodical accessibility, abundance also plays a critical role in successful protein detection (Baldwin, 2004). Standard shotgun proteomics, for instance, is restricted to the detection of the most abundant proteins in a sample and often fails to capture less abundant proteins. Several technical as well as experimental approaches were implemented in the past in order to improve the identification of low abundance proteins, e.g., enhanced detection of MS1 precursor ions (Meier *et al.*, 2018), prolonged LC gradients (Hsieh *et al.*, 2013) and increased sample fractionation (Faca *et al.*, 2007). In this study, we also performed ultra-deep sample fractionation in order to increase the sensitivity of protein detection in *P. aeruginosa* PAO1. In total, 48 fractions were measured using DDA-MS and yielded the confident identification of 3,772 protein products encoded by known anORFs. This number was comparable to other *P. aeruginosa* proteome studies, which yielded detection of maximum 4,000 proteins (Wright *et al.*, 2019a, Erdmann *et al.*, 2019, Kamath *et al.*, 2017, Kumari *et al.*, 2014, Hare *et al.*, 2012). However, it must be noted that most of the studies mentioned analysed the *P. aeruginosa* proteome under deviating cultivation conditions, used different MS techniques or reported the summarized number of proteins detected for multiple *P. aeruginosa* strains, thereby restricting comparability with the results obtained in this study. Remarkably, ultra-deep proteomics also verified the proteinaceous nature of about 65% of all hypothetical ORFs in *P. aeruginosa* PAO1 suggesting that the majority of these ORFs are indeed functional and are not only the result of genome over-prediction as proposed for bacterial genomes (Yu *et al.*, 2011). The increasing number of hypothetical genes experimentally confirmed to be expressed and functional in various bacteria (Tian et al., 2019, Yang et al., 2019, Prava et al., 2018, Landstorfer et al., 2014) further supports this assumption. Nevertheless, even deep fractionation could not solve the problem of selection against low abundant proteins entirely as indicated by the significant increased proportion of MS detected proteins which were encoded by highly expressed genes (**Figure 29**). Alternative MS data acquisition strategies offer higher sensitivity at the expense of limited throughput. Targeted proteomics, for example, performs better in detecting low abundant proteins and is able to precisely quantify pre-specified proteins as shown in this study for selected overlapping gene pairs (discussed in section 4.3). However, targeted MS cannot measure the whole proteome in a global fashion (reviewed by Shi *et al.*, 2016). Several strategies including immunoaffinity enrichment, depletion of high abundant proteins, and others were successfully implemented to further enhance the detection of low abundant proteins (Shi *et al.*, 2012). In the last few years, specialized MS techniques addressing the issue of detecting low abundant and small proteins were successful in detecting such proteins, including overlapping ones, encoded by bacterial genes (D'Lima *et al.*, 2017, Impens *et al.*, 2017).

4.2 Cumulative evidence for protein-coding capacity of identified gene candidates

Integration of “omics” datasets contributes decisively to the deep understanding of biological systems and the underlying processes. By combining multiple datasets picturing different molecular layers, possible limitations of single “omic” techniques, like background noise in transcript- and translatomics or biased detection of low-abundant proteins in proteomics, are compensated, allowing to draw clearer and more comprehensive conclusions (for review see Kumar *et al.*, 2016). One of the various possible applications of multi-“omics” dataset analysis is genome annotation, which aided the detections of novel genes in eukaryotic (Koch *et al.*, 2014, Wu *et al.*, 2014) as well as bacterial cells (Miravet-Verde *et al.*, 2019, Neuhaus *et al.*, 2016, Schrimpe-Rutledge *et al.*, 2012). In this study, multiple “omic” technologies including transcriptomics, translatomics and proteomics were used for most reliable delineation of novel intergenic and overlapping genes. In a first step, all datasets arising from conventional and modified Ribo-seq experiments were evaluated using three independent ORF prediction tools and the results were subsequently merged and scored to minimize the effect of experimental, biological, and methodological fluctuations. The obtained results were then combined with the results of transcriptomics, Cappable-seq and proteomics to identify high confident gene candidates. This procedure aided the detection of 116 and 124 promising novel gene candidates for *E. coli* LF82 and *P. aeruginosa* PAO1, respectively.

The number of 116 novel intergenic and overlapping ORFs identified in *E. coli* LF82 was substantially lower than the number of novel ORFs reported for *E. coli* strains in literature after applying conventional RNA-seq and Ribo-seq. Hücker *et al.* (2017), for instance, claimed the detection of 465 putative novel genes in intergenic regions of *E. coli* O157:H7 strain Sakai, 72 intergenic novel genes were proposed by Neuhaus *et al.* (2016) for *E. coli* O157:H7 strain EDL933, and a recent preprint identified between 84 and 190 ORFs solely embedded in antisense to annotated genes in four different *E. coli* strains (Zehentner *et al.*, 2020a). It seems likely that the decreased number of ORFs observed in this study resulted from more conservative delineation criteria, e.g., the usage of multiple tools for ORF prediction combined with subsequent scoring, and the higher number of Ribo-seq experiments including modified variants like Ribo-RET and RelE-supported Ribo-seq. With a median length of 129 bp (mean = 162.7 bp; **Figure 24**), the ORFs detected were rather short compared to the known anORFs. This observation was consistent with the findings by Hücker *et al.* (2017) and Zehentner *et al.* (2020a) who reported mean lengths of 127 to 172 bp (depending on the absence or presence of annotated homologues) and 250 bp for the novel ORFs, respectively. As the detection of small proteins is technically challenging (discussed in 4.1), the short length of some of the identified ORFs might be one explanation why they have escaped discovery for a long time. The novel ORFs detected in *P. aeruginosa* PAO1 were also on average shorter than the encoded annotated genes (**Figure 35**); however, a lack of Ribo-seq-based studies addressing the topic of novel ORF annotation in *P. aeruginosa* prohibited comparison with published results. On average, the novel ORFs of *P. aeruginosa* were larger than those of *E. coli* LF82, which was expected based on the increased GC content of the *P. aeruginosa* genome resulting in the occurrence of statically larger ORFs (Mir *et al.*, 2012).

Despite size, properties considered to be "anomalous" for protein-coding genes, e.g., an aberrant start codon usage, may also hamper the detection of novel translated ORFs. In both *E. coli* as well as *P. aeruginosa*, ATG is the mostly common used codon for translational initiation, distantly followed by GTG and TTG (Villegas & Kropinski, 2008, West & Iglewski, 1988). In this study, 87.7% and 89.1% of all annotated genes in *E. coli* LF82 and *P. aeruginosa* PAO1 possessed an ATG start codon, and thus, confirmed the high utilization percentages of this start codon described in literature. Although the majority of the novel ORFs also started translation from a putative ATG codon, the overall percentage of the alterative start codons was clearly higher (39.6% and 17.7% for *E. coli* LF82 and *P. aeruginosa* PAO1, respectively). However, an altered start codon usage was frequently reported for small as well as overlapping ORFs. Meydan *et al.* (2019), for instance, observed conserved RET peaks for 74 ORFs overlapping out-of-frame with annotated genes, whereby more than 40% of them started with a non-ATG codon. Recently characterized overlapping genes in bacteria as well as mammals, e.g., *pop* in *E. coli* (Zehentner *et al.*, 2020b) or *POLGARF* in human cells (Loughran *et al.*, 2020) were also shown to utilize the alternative start codon CTG to guide translation. Although not subject of this study, even rare start codons with suboptimal translational efficiency (Nie *et al.*, 2006) like TTG were reported for short translated ORFs (Neuhaus *et al.*, 2016, Hücker *et al.*, 2017). Ribo-RET analysis provided further evidence for translation initiation at the selected start codons of 97 and 42 novel ORFs identified in *E. coli* LF82 and *P. aeruginosa* PAO1, among them also many non-ATG start codons (**Supplementary Table 1 & 2**).

In addition to the choice of start codon, the presence and type of SD sequence can also influence translational initiation and efficiency (Vimberg *et al.*, 2007, Shine & Dalgarno, 1974). For approximately 60% of all novel ORFs in *E. coli* LF82, a SD sequence was detected in an optimal spacing of 9 bp. The percentage of novel ORFs harbouring a SD sequences was even higher for *P. aeruginosa* PAO1 (~84%) which is in concordance with its increased percentage of annotated genes possessing a SD sequence (69% versus 57% for *P. aeruginosa* and *E. coli*; Ma *et al.*, 2002). Furthermore, a clear conservation pattern of the SD sequence was observed for *P. aeruginosa* PAO1 (**Figure 36**). A lack of a SD sequence, though, does not necessarily exclude the occurrence of translation of the remaining novel gene candidates, since a SD sequence is not obligatory for start codon selection (Saito *et al.*, 2020) and even transcripts lacking any leader have been reported (Dötsch *et al.*, 2012, Zheng *et al.*, 2011). For the vast majority of ORFs in *E. coli* LF82, a σ^{70} promoter was predicted to be localized in a median distance of 131 bp upstream of the start codon (**Supplementary Figure 11**). This distance was far longer than the 5´UTR lengths experimentally validated for *E. coli* mRNAs in literature (25 - 35 bp; Kim *et al.*, 2012) suggesting that some of the promoters identified in this study were false positive predictions. Experimental evidence, e.g., by Cappable-seq (Ettwiller *et al.*, 2016), could contribute to precise TSSs mapping and determination of 5´UTR lengths in *E. coli* LF82. By exploiting this technique, reproducible TSS were detected for 61 of the 124 novel gene candidates in *P. aeruginosa* PAO1 in a median distance of 58 bp upstream of the respective start codon. The 5`UTRs of the novel candidates showed a similar length as the 5´UTRs of all anORFs in strain PAO1 (median = 69 bp) and were also comparable to the results described in literature for *P. aeruginosa* PA14 (median = 47 nt; Wurtzel *et al.*, 2012) and *P. syringae* pv. tomato str. DC3000 (mean = 78 nt; Filiatrault *et al.*, 2011). In addition, sequence conservation of the -10 promoter element as observed for both the novel ORFs as well as the anORFs in *P. aeruginosa* PAO1 (**Figure 36**) further suggested specificity of TSS detection.

ρ-independent terminator structures were predicted for 44 (*E. coli* LF82) and 28 (*P. aeruginosa* PAO1) of the novel gene candidates using FindTerm (Solovyev & Salamov, 2011). Analogous to promoter and SD elements, the presence of a ρ-independent terminator provides evidence for the protein-coding capacity of an ORF, whereas the absence of such elements is not indicative of an absence of protein coding. Possible reasons for a lack of ρ-independent terminator include the involvement of other termination mechanisms like ρ-dependent termination or the location of an ORF within an operon. Term-seq, recently developed to map 3'-termini of RNA transcripts (Dar *et al.*, 2016), seems to be a promising experimental approach to unravel transcriptional events in a global fashion, and thus, also to determine the exact terminator location of the novel gene candidates.

Since several of the novel ORFs fulfilled the structural qualifications associated with protein coding, their genuine transcription and translation was assumed. Indeed, the novel ORFs exhibited gene expression metrics including RCV, RPKM and coverage values which were in a similar range as those of the respective protein-coding anORFs (**Figure 25 & 37**) and largely consistent across biological replicates. Furthermore, when applying the threshold Hücker *et al.* (2017) used for the Ribo-seq-based delineation of novel ORFs (RPKM ≥ 1, coverage ≥ 0.5 & RCV ≥ 0.25) in *E. coli* O157:H7 Sakai, 87% of all ORFs identified in *E. coli* LF82 complied with this criteria in at least one of the Ribo-seq experiments performed. Although transferability of these thresholds might be limited, more than 97% of all ORFs identified in *P. aeruginosa* PAO1 also exhibited expression metrics, which were above those suggested by Hücker *et al.* (2017). However, in this context it must be noted that the expression metrics of translated ORFs overlapping in sense with anORFs must be interpreted with caution as a differentiation between reads belonging to the novel ORFs and reads arising from translation of the mother ORFs cannot be distinguished. For this purpose, a distortion of the aforementioned values might be possible, especially when the mother ORF shows a high expression. In this study, the programs REPARATION (Ndah *et al.*, 2017), DeepRibo (Clauwaert *et al.*, 2019) and the scripts by

Giess *et al.* (2017) were used for most objective translated ORF delineation. Although those programs are not devoid of false positives (Clauwaert *et al.*, 2019, Giess *et al.*, 2017, Ndah *et al.*, 2017), it seems highly unlikely that the selected ORFs were predicted erroneously considering the fact that they had high overall prediction scores of at least 9 (*P. aeruginosa* PAO1) or 10 (*E. coli* LF82), implying that they were identified in multiple different Ribo-seq datasets by independent tools. However, since those programs delineate ORFs *de novo* without prior knowledge about the existing annotation, signals of anORFs might be interpreted mistakenly as novel ORFs in other frames. Therefore, the prediction of sense embedded ORFs might be particularly prone to errors. In addition, reading frame analysis of the novel ORF candidates was also mainly restricted to intergenic as well as antisense overlapping ORFs because ORFs overlapping in sense often overlapped with highly expressed genes. As a result, the reading frames of the novel sense overlapping ORFs were often concealed by the periodicity signal of their mother genes` frame. For the remaining ORFs without sense overlap, RelE-supported Ribo-seq confirmed genuine translation as indicates by a pronounced reading frame sum signal at sub-codon position 2 (**Figure 27 & 38**) which portrayed the codon-wise movement of the ribosomes on the respective mRNAs.

Regulation of gene expression is an indicator for functionality of an ORF (Ardern *et al.*, 2020) and an argument against non-functional pervasive transcription and translation as discussed previously. For *E. coli* LF82, some novel ORFs, among them four sense overlapping ORFs, were shown to be differentially regulated depending on the presence or the absence of oxygen during cultivation, either at the transcriptional or at the translational level. The percentage of novel ORFs shown to be regulated (5.2%) was comparable to the percentage of regulated anORFs (4.6%) in the same datasets. The large percentage of detectable homologues after blastp search further indicates evolutionary conservation, and thus, functionality of the detected ORFs. In many cases, though, the type of functionality could not be inferred from blastp analysis as the majority of hits were characterized as hypothetical genes. Nevertheless, the status "hypothetical" does not equate to a lack of functionality in many instances as shortly mentioned in section 4.1.4.

For more than a third of the novel ORFs detected in *P. aeruginosa* PAO1, the proteinaceous nature was verified by MS. Among the detected ones were also 16 proteins encoded by non-trivially overlapping ORFs; eight of them were overlapping in antisense, seven of them in sense and one showed multiple overlaps both in sense as well as in antisense to anORFs. All protein products except one were identified by two or more peptides confirming their confident detection. This is an extraordinary result considering that the native proof of OLG-encoded proteins has rarely been achieved, presumably due to their intrinsic properties like short size or weak expression limiting MS detection (see section 4.1.4). Notwithstanding, a few studies successfully detected OLG-encoded proteins, but often by serendipity. In a large-scale study, Venter *et al.* (2011), for instance, identified 245 loci conflicts in 37 different bacterial genera with overlaps of more than 40 bp between novel and annotated regions with proteomic support. However, the authors noted that conflicting peptides could point to a high false positive rate. Proteomic studies in other bacterial genera like *Helicobacter* (Friedman *et al.*, 2017), *Salmonella* (Willems *et al.*, 2020) or *Pseudomonas* (Kim *et al.*, 2009) further provided evidence for the proteinaceous character of OLGs. In the latter, Yang *et al.* (2016) also detected peptides for 44 small antisense sequences in the species *P. putida*, whereby the recording of one peptide was sufficient for identification.

DDA-MS also confirmed that in more than 85% of all cases, at least one of the mother genes overlapping non-trivially with a novel ORFs in this study was protein coding. This result indicates that the novel overlapping ORFs were designated to be overlapping with justification as they share an overlap with genuine protein-coding anORFs.

As summarized in **Figure 45**, we believe that the novel ORFs identified in *E. coli* LF82 and *P. aeruginosa* PAO1 are protein coding, and thus, can be considered as "real" genes due to

- the existence of structural features associated with protein coding as validated experimentally or predicted *in silico*
- their expression patterns which were comparable to those of protein-coding anORFs resulting in their successful prediction my multiple tools in multiple Ribo-seq datasets
- a pronounced triplet periodicity signal reflecting those obtained for anORFs
- the presence of RET peaks at many start codons
- evidence for functionality provided by differential gene expression as well as blastp analysis
- & the confident detection of many protein products encoded by the novel ORFs.

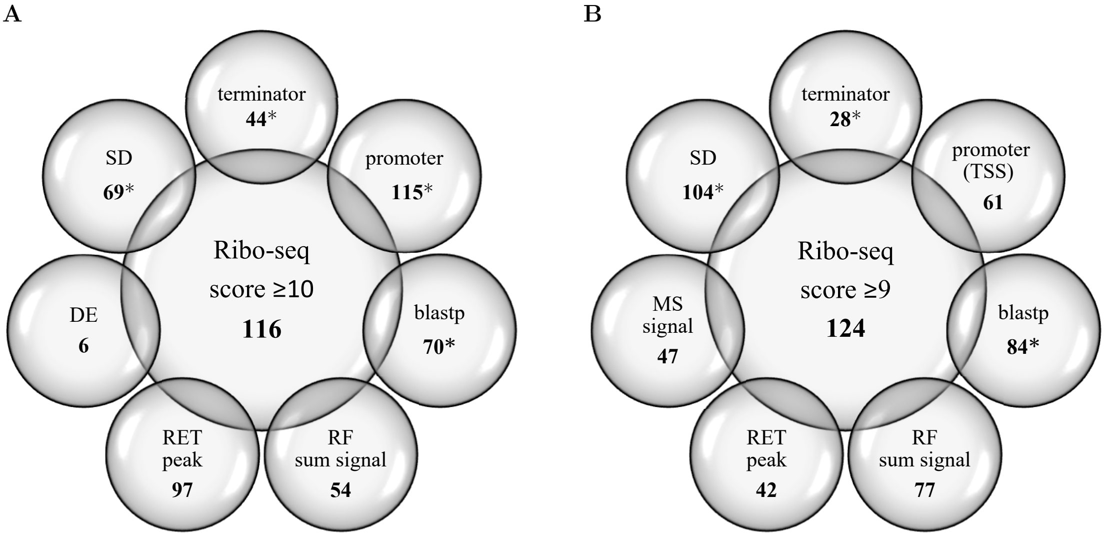


Figure 45. Overview of all novel gene candidates identified in (**A**) *E. coli* LF82 and (**B**) *P. aeruginosa* PAO1. The Venn diagrams show the absolute number of ORFs delineated based on Ribo-seq data with absolute prediction scores (**A**; LF82) ≥10 or (**B**; PAO1) ≥9 supported by different kinds of experimental or computational evidence for protein-coding. These include analysis of structural gene features like presence of a promoter, a transcription start site (TSS), a terminator and a Shine-Dalgarno (SD) sequence, the detection of homologous hits using a blastp search, analysis of differential gene expression (DE), proteomic evidence by mass spectrometry (MS) as well as the detection of a retapamulin (RET) start peak or a reading frame (RF) signal in Ribo-RET or RelE-assisted Ribo-seq datasets, respectively. Results which lack experimental evidence and were solely predicted bioinformatically are highlighted by an asterisk.

4.3 *olg1* and *olg2* as prime examples for overlooked OLGs in *Pseudomonas*

Upon reviewing the overlapping gene candidates identified in *P. aeruginosa* PAO1, two ORFs attracted our attention due to their convincing experimental results and their striking lengths. These two OLGs, named *olg1* and *olg2*, spanned 957 and 1,728 nt and exhibited large antisense overlaps, covering at least 100% and 88.8% of their entire sequence, with the mother genes *tle3* and PA1383. Interestingly, both OLGs were located in frame -1 relative to their mother genes. An overlap in frame -1 is considered to be highly conserved, thus limiting the flexibility for mutational changes (Wichmann & Ardern, 2019). However, some bacterial OLGs with this type of overlap have been reported in literature (e.g., Zehentner *et al.*, 2020b, Rogozin *et al.*, 2002b).

Several aspects emphasized that *olg1* and *olg2* are protein coding and give rise to the synthesis of functional protein products. Firstly, both OLGs exhibited multiple structural features necessary for gene expression. σ^{70} promoters were predicted computationally in the upstream vicinity of *olg1* and *olg2* and were also confirmed experimentally by the results of Cappable-seq. However, additional Cappable-seq peaks indicated that transcription could be also initiated at a different position, probably also guided by one of the other 23 sigma factors known for *P. aeruginosa* (Potvin *et al.*, 2008). Transcription was shown to cease at a ρ-independent terminator in case of *olg2*, and via a different mechanism, e.g., by exploiting protein Rho (Mitra *et al.*, 2017), in case of *olg1*. Furthermore, *olg1* possessed a strong SD sequence in an optimal aligned spacing to the ATG start codon. The ORF of *olg2* also started with an ATG codon, which was additionally verified by a Ribo-RET peak. The gene annotation program Prodigal confirmed the protein-coding potential of both OLGs, but only upon hiding of the mother genes since dynamic programming restricts prediction of large overlapping genes and chooses the overlapping ORF with the highest score for annotation (Hyatt *et al.*, 2010). Secondly, both OLGs were covered with a substantial number of RNA-seq as well as Ribo-seq reads implying that they were transcribed and translated under the conditions tested. However, an accumulation of reads upstream of *olg1´*s start codon raised the possibility of an N-terminal extension of the coding region. Alternatively, the expression of two isoform might be possible, whereby various functions of such isoform have been discovered and discussed for bacteria (Fijalkowska *et al.*, 2020, Meydan *et al.*, 2019, Meydan *et al.*, 2018). Due to the absence of a RET peak as well as further contradictory results, the exact start position of *olg1´*s coding region could not be proven, and therefore the most conservative assumption is that translation starts at ATG_{291556}, which was also suggested by Prodigal. Further experiments like reporter gene fusions, usage of alternative initiation inhibitors like Onc112 in Ribo-seq (Weaver *et al.*, 2019) or N-terminomics (Impens *et al.*, 2017) could be implemented to shed light on the correct translation initiation site. Despite some uncertainties about the correct start position, both OLGs showed high overall prediction scores of 10 (*olg1*) and 9 (*olg2*) providing strong evidence for their genuine translation. Further evidence for translation was provided by regulated transcriptional and translational signals obtained after reanalysis of data published by Grady *et al.* (2017) and by the reading frame signal of *olg1* in this study. For *olg2*, the lack of a periodicity signal at sub-codon position 2 was unexpected due to *olg2´*s clear transcriptional and translational signals. However, absolute read counts of *olg2* were substantially lower than those of *olg1*. Furthermore, a biased distribution of the nucleobases present in this ORF or an accumulation of ribosomes at certain positions, e.g., pause sites, could also have led to anomalies in the reading frame signal as proposed by Hwang & Buskirk (2017). The fact that multiple peptides were confidently discovered for *olg2* using DDA-MS leaves no doubt about its coding potential. For *olg1*, an even higher number of twelve translated peptides were detected. As translation of even short genes is associated with high bioenergetic costs (Lynch & Marinov, 2015), and expression of random sequences was shown to have an inhibitory effect on bacterial growth (Neme *et al.*, 2017), a tight regulation of translation is essential. With respect to these aspects, it seems highly unlikely that proteins of 318 and 575 AAs in length represent non-functional products arising from pervasive translation (Smith *et al.*, 2019). An extensive regulation of the expression of both OLGs throughout the exponential and stationary growth of *P. aeruginosa* PAO1 was also confirmed by PRM-MS, which is considered as the gold standard for protein quantification (Peterson *et al.*, 2012) and further supported validity and functionality of the encoded proteins. Compared to all anORFs, the absolute abundance of Olg1 and Olg2 was in a medium to low range, which is in concordance with published results as some of the hitherto experimentally characterized OLGs were confirmed to have a weak expression (e.g., Hücker *et al.*, 2018a, Fellner *et al.*, 2015). Due to their low expression and small size, OLGs were often hypothesized to be evolutionary young genes and in an initial stage of adaptation (Fellner *et al.*, 2015, Fellner *et al.*, 2014, Donoghue *et al.*, 2011). Based on the results of evolutionary analyses, we also suggest that *olg1* and *olg2* are of rather young age. The AA sequence of both OLGs is evolving more rapidly than those of their mother genes (slower by factor 2 and 12 for *tle3* and PA1383, respectively) indicating a lower conservation of the OLG sequence, a property highly associated with a recent emergence (Carvunis *et al.*, 2012). Furthermore, both OLGs showed signs of purifying selection as the sequence constraint in the mother gene was reduced in the OLG region and stop codons were depleted. Both OLGs were predominately found within the species whereas their mother genes were more broadly distributed, suggesting that *olg1* and *olg2* constitute taxonomically restricted genes, which probably arose by overprinting (Grassé, 1977, Ohno, 1970) within their older mother genes. Such genes, also called orphans, are hypothesized to be involved in the development of taxon-specific morphology and in adaptation processes in response to changing conditions (Colbourne *et al.*, 2011, Khalturin *et al.*, 2009). In addition, they usually have low expression levels (Prabh & Rödelsperger, 2016), are shorter as well as less conserved compared to older genes (Palmieri *et al.*, 2014) and often encode for disordered proteins (Heames *et al.*, 2020). Assuming taxonomic restriction and an evolutionary young age, it comes as no surprise to find many OLG candidates (described in section 4.2) which lack protein features like mature promoter, terminator, etc. elements or have suboptimal and less efficient start codons. However, the exact evolutionary origin of orphans remains a mystery and requires further efforts for clarification. In this context, OLGs offer some interesting approaches to study the underlying processes as their mother genes fix their genomic region. Therby, OLGs could help to solve some common issues associated with homology inference, e.g., detection failures (Weisman *et al.*, 2020, Vakirlis *et al.*, 2020).

In sum, these results provide compelling evidence that both OLGs are protein coding and translated into bioactive proteins with cellular function. As shortly discussed in a recent preprint (Zehentner *et al.*, 2020a), the designation of a gene as "overlapping" is highly dependent on the annotation status of the mother gene. If a mother gene would be the result of a mis-annotation and lacks protein-coding capacity, the translated ORF located in another frame cannot be rightly termed "overlapping". However, in case of *tle3* and PA1383, the mother genes of *olg1* and *olg2*, its beyond doubt that they encode for proteins due to multiple reasons: Firstly, the functionality of Tle3 was confirmed by Berni *et al.* (2019) who characterized the antibacterial effect and the secretion of Tle3 via a T6SS by performing intra-species bacterial competition assays. In contrast, PA1383 has an "hypothetical" annotation status and lacks experimental characterization, but functionality is implied by the presence of an N-terminal type I signal sequence for cellular export (Lewenza *et al.*, 2005), purifying selection acting on PA1383 as well as a broad distribution of PA1383 homologues across the genus *Pseudomonas*. Furthermore, PA1383 but also *tle3* were confirmed to be transcribed and translated as suggested by the RNA-seq, Ribo-seq and MS experiments of this study. As such, in light of the data regarding expression and evolutionary sequence analysis it is clear that both ORFs are correctly annotated and represent translated genes with functionality. Consequently, *olg1* and *olg2* can be considered genuinely as overlapping genes, which have most likely been overlooked due to the striking gene-like characteristics of their mother genes.

Although NGS-based methods facilitated the increasing identification of translated overlapping, the vast majority of the ORFs detected were typically short end encoded for small proteins, e.g., as shown for different *E. coli* strains (Zehentner *et al.*, 2020a, Weaver *et al.*, 2019, Meydan *et al.*, 2019). A similar observation was made by Smith *et al.* (2019) who detected 274 novel ORFs in *Mycobacterium tuberculosis*, many of them being short. Ribo-seq in *Salmonella enterica* Typhimurium also resulted in the detection of translated ORFs, some of them overlapping and encoding for protein with a maximum length of 144 AAs (Baek *et al.*, 2017). Congruently, the length of OLG candidates with experimental characterization seldomly exceeded 200 codons (e.g., Vanderhaeghen *et al.*, 2018, Hücker *et al.*, 2018a, Hücker *et al.*, 2018b, Fellner *et al.*, 2015, Fellner *et al.*, 2014, Behrens *et al.*, 2002). Rare exceptions of larger OLGs with experimental evidence constitute the genes *pop* recently detected in EHEC (Zehentner *et al.*, 2020b), the gene *adm* putatively involved in the secondary metabolism of *Streptomyces coelicolor* (Tunca *et al.*, 2009) and *cosA* identified in *P. fluorescens* Pf0-1 (Silby & Levy, 2008), which encode for proteins with a length of 200, 233 and 338 AAs, respectively. However, public available Ribo-seq data raise doubts about the authenticity of *adm* as a genuine overlapping gene (personal communication with Dr. Klaus Neuhaus). For the OLG *cosA*, a proteomic proof of the encoded protein in a follow up study was unfeasible presumably due to its weak expression under the conditions tested (Kim *et al.*, 2009). However, in the same study the authors succeeded in finding peptides of in total nine protein-coding ORFs, which were located antisense to annotated genes in *P. fluorescens*. The by far largest ORF reported in this study encoded for a 530 AA large protein; the proteins encoded by the other antisense OLGs had a length less or equal 195 AAs. A 1,644 nt long ORF encoding for a theoretical protein of 548 AAs was identified in the genome of *Deinococcus radiodurans* and constitutes the hitherto largest OLG with proteomic evidence which has been described in literature so far (Willems *et al.*, 2020). However, the utmost number of OLGs identified via mass spectrometry lack further targeted studies to confirm the validity of the peptides detected. For *olg1* and *olg2*, targeted proteomics confirmed their protein-coding capacity with high confidence, and therefore these OLGs are currently the best attested ones in terms of proteomic evidence while simultaneously constituting two of the largest -if not the largest- OLGs detected in prokaryotes. *Olg1* and *olg2* represent unusually large OLGs even when compared with OLGs reported for other non-prokaryotic genomes. In the model archaeon *Haloferax volcanii*, Gelsinger *et al.* (2020) for instance detected 160 novel TSS, whereby 75% of them encoded for proteins shorter than 129 AAs. Schlub & Holmes (2020) analysed in total 5,976 viral genomes and predicted 83,722 instances of overlapping regions, of which 84% were shorter than 50 nt. Although a large proportion of overlaps were located on the same strand, antisense overlaps up to 2,351 nt (median = 212 nt) were detected in all viral groups except +ssRNA viruses. An exception to this constitutes a recently discovered ~1,000 AA long antisense ORF encompassing almost the entire genome length of some “ambigrammatic” +ssRNA viruses belonging to the family *Narnaviridae* (DeRisi *et al.*, 2019). Although initially suggested to be a protein-coding sequence (Dinan *et al.*, 2020), the author of a recent preprint reported contradictory results and claimed that reverse open reading frames encoded by “ambigrammatic” viruses are not translated into proteins (Dudas *et al.*, 2021). To conclude, the exceptional sizes of *olg1* and *olg2* in combination with their profound experimental and evolutionary evidence leave no doubt about their protein-coding nature and functionality. However, the type and scope of their function has to await further studies.

4.4 Concluding remarks and outlook

This study provided cumulative evidence for the existence and proteinaceous nature of OLGs, gene constructs which are a matter of controversy that still lack full acceptance in the scientific community (e.g., Wade & Grainger, 2014, Palleјà *et al.*, 2008). However, the increasing detection of transcriptional and translational events in non-canonical regions associated with non-coding RNAs, micropeptides, small proteins or OLGs (Orr *et al.*, 2019, Storz *et al.*, 2014) raised the question whether these signals represent non-functional background noise as formerly proposed (Smith *et al.*, 2019, Lloréns-Rico *et al.*, 2016). Multiple research groups made a substantial effort to unveil the “dark proteome” of prokaryotes by using Ribo-seq and recent modifications of it, thereby detecting a variety of such “alternative”, yet functional ORFs (e.g., Weaver *et al.*, 2019, Meydan *et al.*, 2019). In this study, classical Ribo-seq was combined with Ribo-RET, RelE-assisted Ribo-seq, RNA-seq, Cappable-seq and mass spectrometry to aid maximum confident delineation of novel OLGs in a high-throughput fashion. The candidates identified in this study expand the short, but growing list of experimentally validated OLGs and provide a solid basis for their further study. Of special interest might be the functional characterization of *olg1* and *olg2*, which represent the longest yet known OLGs in prokaryotes.

Experiments like competitive growth experiments at various conditions, phenotypic analysis using genomic knockout mutants, or co-immunoprecipitation with OLG-associated proteins could shed light on the physiological roles of Olg1 and Olg2 in *P. aeruginosa*. Follow-up studies are also advisable for all sense overlapping genes as they are experimentally difficult to access, leading to a void of studies addressing such constructs, as well as all OLG candidates with proteomic evidence as the detection and validation of a native protein should leave no doubt about the proteinaceous nature, and thus, also about the functionality of an OLG. Despite biological reproduction of the existing datasets, additional experiments like Term-seq (Dar *et al.*, 2016) or apidaecin-supported Ribo-seq (Mangano *et al.*, 2020) could further contribute not only to the precise mapping of ORF and transcript boundaries of the OLGs identified in this study, but also facilitate the detection of unknown genes which are still hiding in bacterial genomes. In this context, it must be noted that this study solely focused on the most reliable ORF candidates, which showed multiple lines of evidence for being protein coding. However, as the entirety of novel ORFs was not analysed due their decreased confidence as indicates by lower prediction scores, it might be possible that many more promising OLG and intergenic gene candidates could be derived from this data. In conclusion, the results of this study give a first impression of how many "unexpected" functional elements still await discovery and how much work has to be done in order to unravel the entire complexity of bacterial genomes (Kirchberger *et al.*, 2020, Grainger, 2016).

5. References

Aeschimann, F., Xiong, J., Arnold, A., Dieterich, C., and Großhans, H. (2015) Transcriptome-wide measurement of ribosomal occupancy by ribosome profiling. *Methods* **85**: 75-89.

Albrecht, M., Sharma, C.M., Dittrich, M.T., Müller, T., Reinhardt, R., Vogel, J., and Rudel, T. (2011) The transcriptional landscape of *Chlamydia pneumoniae*. *Genome Biol* **12**: R98.

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

Amavisit, P., Lightfoot, D., Browning, G.F., and Markham, P.F. (2003) Variation between pathogenic serovars within *Salmonella* pathogenicity islands. *J Bacteriol* **185**: 3624-3635.

Andreev, D.E., O'Connor, P.B., Loughran, G., Dmitriev, S.E., Baranov, P.V., and Shatsky, I.N. (2017) Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Res* **45**: 513-526.

Andrews, S. (2010) FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Angus, B.L., Carey, A.M., Caron, D.A., Kropinski, A.M., and Hancock, R.E. (1982) Outer membrane permeability in *Pseudomonas aeruginosa*: comparison of a wild-type with an antibiotic-supersusceptible mutant. *Antimicrob Agents Chemother* **21**: 299-309.

Aoyama, T., Takanami, M., Ohtsuka, E., Taniyama, Y., Marumoto, R., Sato, H., and Ikehara, M. (1983) Essential structure of *E. coli* promoter: effect of spacer length between the two consensus sequences on promoter function. *Nucleic Acids Res* **11**: 5855-5864.

Arber, W. (2014) Horizontal gene transfer among bacteria and its role in biological evolution. *Life* **4**: 217-224.

Ardern, Z., Neuhaus, K., and Scherer, S. (2020) Are Antisense Proteins in Prokaryotes Functional? *Front Mol Biosci* **7**: 187-187.

Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. (2016) PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* **44**: W16-21.

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

Balaban, N.Q., Gerdes, K., Lewis, K., and McKinney, J.D. (2013) A problem of persistence: still more questions than answers? *Nat Rev Microbiol* **11**: 587-591.

Balabanov, V.P., Kotova, V.Y., Kholodii, G.Y., Mindlin, S.Z., and Zavilgelsky, G.B. (2012) A novel gene, *ardD*, determines antirestriction activity of the non-conjugative transposon Tn5053 and is located antisense within the *tniA* gene. *FEMS Microbiol Lett* **337**: 55-60.

Balakrishnan, R., Oman, K., Shoji, S., Bundschuh, R., and Fredrick, K. (2014) The conserved GTPase LepA contributes mainly to translation initiation in *Escherichia coli*. *Nucleic Acids Res* **42**: 13370-13383.

Baldwin, M.A. (2004) Protein Identification by Mass Spectrometry. *Mol Cell Proteomics* **3**: 1.

Balloy, V., Verma, A., Kuravi, S., Si-Tahar, M., Chignard, M., and Ramphal, R. (2007) The role of flagellin versus motility in acute lung disease caused by *Pseudomonas aeruginosa*. *J Infect Dis* **196**: 289-296.

Baran-Gale, J., Kurtz, C.L., Erdos, M.R., Sison, C., Young, A., Fannin, E.E., Chines, P.S., and Sethupathy, P. (2015) Addressing Bias in Small RNA Library Preparation for Sequencing: A New Protocol Recovers MicroRNAs that Evade Capture by Current Methods. *Front Genet* **6**: 352.

Barne, K.A., Bown, J.A., Busby, S.J., and Minchin, S.D. (1997) Region 2.5 of the *Escherichia coli* RNA polymerase sigma70 subunit is responsible for the recognition of the 'extended-10' motif at promoters. *EMBO J* **16**: 4034-4040.

Barnich, N., Carvalho, F.A., Glasser, A.-L., Darcha, C., Jantscheff, P., Allez, M., Peeters, H., Bommelaer, G., Desreumaux, P., Colombel, J.-F., and Darfeuille-Michaud, A. (2007) CEACAM6 acts as a receptor for adherent-invasive *E. coli,* supporting ileal mucosa colonization in Crohn disease. *J Clin Invest* **117**: 1566-1574.

Barrell, B.G., Air, G.M., and Hutchison, C.A., 3rd (1976) Overlapping genes in bacteriophage phiX174. *Nature* **264**: 34-41.

Basrai, M.A., Hieter, P., and Boeke, J.D. (1997) Small open reading frames: beautiful needles in the haystack. *Genome Res* **7**: 768-771.

Bassetti, M., Vena, A., Croxatto, A., Righi, E., and Guery, B. (2018) How to manage *Pseudomonas aeruginosa* infections. *Drugs Context* **7**: 212527-212527.

Baumgart, M., Dogan, B., Rishniw, M., Weitzman, G., Bosworth, B., Yantiss, R., Orsi, R.H., Wiedmann, M., McDonough, P., Kim, S.G., Berg, D., Schukken, Y., Scherl, E., and Simpson, K.W. (2007) Culture independent analysis of ileal mucosa reveals a selective increase in invasive *Escherichia coli* of novel phylogeny relative to depletion of Clostridiales in Crohn's disease involving the ileum. *ISME J* **1**: 403-418.

Bayramoglu, B., Toubiana, D., and Gillor, O. (2017) Genome-wide transcription profiling of aerobic and anaerobic *Escherichia coli* biofilm and planktonic cultures. *FEMS Microbiol Lett* **364**.

Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C., and Giraldez, A.J. (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**: 981-993.

Beaume, M., Hernandez, D., Farinelli, L., Deluen, C., Linder, P., Gaspin, C., Romby, P., Schrenzel, J., and Francois, P. (2010) Cartography of Methicillin-Resistant *S. aureus* Transcripts: Detection, Orientation and Temporal Expression during Growth Phase and Stress Conditions. *PLOS ONE* **5**: e10725.

Bech, F.W., Jørgensen, S.T., Diderichsen, B., and Karlström, O.H. (1985) Sequence of the *relB* transcription unit from *Escherichia coli* and identification of the *relB* gene. *EMBO J* **4**: 1059-1066.

Becker, A.H., Oh, E., Weissman, J.S., Kramer, G., and Bukau, B. (2013) Selective ribosome profiling as a tool for studying the interaction of chaperones and targeting factors with nascent polypeptide chains and ribosomes. *Nat Protoc* **8**: 2212-2239.

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

Belshaw, R., Pybus, O.G., and Rambaut, A. (2007) The evolution of genome compression and genomic novelty in RNA viruses. *Genome Res* **17**: 1496-1504.

Berni, B., Soscia, C., Djermoun, S., Ize, B., and Bleves, S. (2019) A Type VI Secretion System Trans-Kingdom Effector Is Required for the Delivery of a Novel Antibacterial Toxin in *Pseudomonas aeruginosa*. *Front Microbiol* **10**: 1218.

Besemer, J., and Borodovsky, M. (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* **33**: W451-W454.

Bianconi, I., D'Arcangelo, S., Esposito, A., Benedet, M., Piffer, E., Dinnella, G., Gualdi, P., Schinella, M., Baldo, E., Donati, C., and Jousson, O. (2019) Persistence and Microevolution of *Pseudomonas aeruginosa* in the Cystic Fibrosis Lung: A Single-Patient Longitudinal Genomic Study. *Front Microbiol* **9**: 3242-3242.

Bitton, D.A., Smith, D.L., Connolly, Y., Scutt, P.J., and Miller, C.J. (2010) An Integrated Mass-Spectrometry Pipeline Identifies Novel Protein Coding-Regions in the Human Genome. *PLOS ONE* **5**: e8949.

Blair, J.M.A., Webber, M.A., Baylay, A.J., Ogbolu, D.O., and Piddock, L.J.V. (2015) Molecular mechanisms of antibiotic resistance. *Nat Rev Microbiol* **13**: 42-51.

Blanc-Potard, A.B., Solomon, F., Kayser, J., and Groisman, E.A. (1999) The SPI-3 pathogenicity island of *Salmonella enterica*. *J Bacteriol* **181**: 998-1004.

Blanco, P., Hernando-Amado, S., Reales-Calderon, J.A., Corona, F., Lira, F., Alcalde-Rico, M., Bernardini, A., Sanchez, M.B., and Martinez, J.L. (2016) Bacterial Multidrug Efflux Pumps: Much More Than Antibiotic Resistance Determinants. *Microorganisms* **4**: 14.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., and Shao, Y. (1997) The Complete Genome Sequence of *Escherichia coli* K-12. *Science* **277**: 1453.

Blevins, W.R., Tavella, T., Moro, S.G., Blasco-Moreno, B., Closa-Mosquera, A., Díez, J., Carey, L.B., and Albà, M.M. (2019) Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. *Sci Rep* **9**: 11005.

Blount, Z.D. (2015) The unexhausted potential of *E. coli*. *Elife* **4**: e05826.

Boekhorst, J., Wilson, G., and Siezen, R.J. (2011) Searching in microbial genomes for encoded small proteins. *Microb Biotechnol* **4**: 308-313.

Borodovsky, M., and McIninch, J. (1993) GENMARK: Parallel gene recognition for both DNA strands. *Computers & Chemistry* **17**: 123-133.

Boto, L. (2010) Horizontal gene transfer in evolution: facts and challenges. *Proc Biol Sci* **277**: 819-827.

Boudeau, J., Barnich, N., and Darfeuille-Michaud, A. (2001) Type 1 pili-mediated adherence of *Escherichia coli* strain LF82 isolated from Crohn's disease is involved in bacterial invasion of intestinal epithelial cells. *Mol Microbiol* **39**: 1272-1284.

Boudeau, J., Glasser, A.L., Masseret, E., Joly, B., and Darfeuille-Michaud, A. (1999) Invasive ability of an *Escherichia coli* strain isolated from the ileal mucosa of a patient with Crohn's disease. *Infect Immun* **67**: 4499-4509.

Brandes, N., and Linial, M. (2016) Gene overlapping and size constraints in the viral world. *Biol Direct* **11**: 26.

Brar, G.A., and Weissman, J.S. (2015) Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* **16**: 651-664.

Breaker, R.R. (2018) Riboswitches and translation control. *Cold Spring Harb Perspect Biol* **10**: a032797.

Brennan, F.P., Grant, J., Botting, C.H., O'Flaherty, V., Richards, K.G., and Abram, F. (2013) Insights into the low-temperature adaptation and nutritional flexibility of a soil-persistent *Escherichia coli. FEMS Microbiol Ecol* **84**: 75-85.

Bringer, M.A., Barnich, N., Glasser, A.L., Bardot, O., and Darfeuille-Michaud, A. (2005) HtrA stress protein is involved in intramacrophagic replication of adherent and invasive *Escherichia coli strain* LF82 isolated from a patient with Crohn's disease. *Infect Immun* **73**: 712-721.

Bringer, M.A., Rolhion, N., Glasser, A.L., and Darfeuille-Michaud, A. (2007) The oxidoreductase DsbA plays a key role in the ability of the Crohn's disease-associated adherent-invasive *Escherichia coli* strain LF82 to resist macrophage killing. *J Bacteriol* **189**: 4860-4871.

Brosch, M., Saunders, G.I., Frankish, A., Collins, M.O., Yu, L., Wright, J., Verstraten, R., Adams, D.J., Harrow, J., Choudhary, J.S., and Hubbard, T. (2011) Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res* **21**: 756-767.

Browning, D.F., and Busby, S.J.W. (2004) The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2**: 57-65.

Buchfink, B., Xie, C., and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59-60.

Bush, K. (2010) Bench-to-bedside review: The role of beta-lactamases in antibiotic-resistant Gram-negative infections. *Crit Care* **14**: 224-224.

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., and Ohler, U. (2016) Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* **13**: 165-170.

Camprubí-Font, C., Bustamante, P., Vidal, R.M., O'Brien, C.L., Barnich, N., and Martinez-Medina, M. (2020) Study of a classification algorithm for AIEC identification in geographically distinct *E. coli* strains. *Sci Rep* **10**: 8094.

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

Casari, G., Daruvar, D., Sander, C., and Schneider, R. (1996) Bioinformatics and the discovery of gene function. *Trends Genet* **128**: 244-245.

Cassan, E., Arigon-Chifolleau, A.-M., Mesnard, J.-M., Gross, A., and Gascuel, O. (2016) Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic. *Proc Natl Acad Sci U S A* **113**: 11537-11542.

Castellani, A., and Chalmers, A.J., (1919) *Manual of Tropical Medicine*. Williams Wood and Co, New York.

Céspedes, S., Saitz, W., Del Canto, F., De la Fuente, M., Quera, R., Hermoso, M., Muñoz, R., Ginard, D., Khorrami, S., Girón, J., Assar, R., Rosselló-Mora, R., and Vidal, R.M. (2017) Genetic Diversity and Virulence Determinants of *Escherichia coli* Strains Isolated from Patients with Crohn's Disease in Spain and Chile. *Front Microbiol* **8**: 639-639.

Chandler, C.E., Horspool, A.M., Hill, P.J., Wozniak, D.J., Schertzer, J.W., Rasko, D.A., and Ernst, R.K. (2019) Genomic and Phenotypic Diversity among Ten Laboratory Isolates of *Pseudomonas aeruginosa* PAO1. *J Bacteriol* **201**: e00595-00518.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884-i890.

Chen, W.-H., Trachana, K., Lercher, M.J., and 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. *Mol Biol Evol* **29**: 1703-1706.

Chen, X., Zhou, L., Tian, K., Kumar, A., Singh, S., Prior, B.A., and Wang, Z. (2013) Metabolic engineering of *Escherichia coli:* a sustainable industrial platform for bio-based chemical production. *Biotechnol Adv* **31**: 1200-1223.

Chengguang, H., Sabatini, P., Brandi, L., Giuliodori, A.M., Pon, C.L., and Gualerzi, C.O. (2017) Ribosomal selection of mRNAs with degenerate initiation triplets. *Nucleic Acids Res* **45**: 7309-7325.

Cheregi, O., Vermaas, W., and Funk, C. (2012) The search for new chlorophyll-binding proteins in the cyanobacterium *Synechocystis* sp. PCC 6803. *J Biotechnol* **162**: 124-133.

Cherny, I., Overgaard, M., Borch, J., Bram, Y., Gerdes, K., and Gazit, E. (2007) Structural and Thermodynamic Characterization of the *Escherichia coli* RelBE Toxin− Antitoxin System: Indication for a Functional Role of Differential Stability. *Biochemistry* **46**: 12152-12163.

Chew, G.-L., Pauli, A., Rinn, J.L., Regev, A., Schier, A.F., and Valen, E. (2013) Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development* **140**: 2828-2834.

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

Christensen, S.K., Mikkelsen, M., Pedersen, K., and Gerdes, K. (2001) RelE, a global inhibitor of translation, is activated during nutritional stress. *Proc Natl Acad Sci U S A* **98**: 14328-14333.

Chuanchuen, R., Karkhoff-Schweizer, R.R., and Schweizer, H.P. (2003) High-level triclosan resistance in *Pseudomonas aeruginosa* is solely a result of efflux. *Am J Infect Control* **31**: 124-127.

Chuanchuen, R., Murata, T., Gotoh, N., and Schweizer, H.P. (2005) Substrate-dependent utilization of OprM or OpmH by the *Pseudomonas aeruginosa* MexJK efflux pump. *Antimicrob Agents Chemother* **49**: 2133-2136.

Cianciotto, N.P. (2005) Type II secretion: a protein secretion system for all seasons. *Trends Microbiol* **13**: 581-588.

Clauwaert, J., Menschaert, G., and Waegeman, W. (2019) DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. *Nucleic Acids Res* **47**: e36-e36.

Clermont, O., Christenson, J.K., Denamur, E., and Gordon, D.M. (2013) The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ Microbiol Rep* **5**: 58-65.

Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H., Tokishita, S., Aerts, A., Arnold, G.J., Basu, M.K., Bauer, D.J., Cáceres, C.E., Carmel, L., Casola, C., Choi, J.-H., Detter, J.C., Dong, Q., Dusheyko, S., Eads, B.D., Fröhlich, T., Geiler-Samerotte, K.A., Gerlach, D., Hatcher, P., Jogdeo, S., Krijgsveld, J., Kriventseva, E.V., Kültz, D., Laforsch, C., Lindquist, E., Lopez, J., Manak, J.R., Muller, J., Pangilinan, J., Patwardhan, R.P., Pitluck, S., Pritham, E.J., Rechtsteiner, A., Rho, M., Rogozin, I.B., Sakarya, O., Salamov, A., Schaack, S., Shapiro, H., Shiga, Y., Skalitzky, C., Smith, Z., Souvorov, A., Sung, W., Tang, Z., Tsuchiya, D., Tu, H., Vos, H., Wang, M., Wolf, Y.I., Yamagata, H., Yamada, T., Ye, Y., Shaw, J.R., Andrews, J., Crease, T.J., Tang, H., Lucas, S.M., Robertson, H.M., Bork, P., Koonin, E.V., Zdobnov, E.M., Grigoriev, I.V., Lynch, M., and Boore, J.L. (2011) The Ecoresponsive Genome of *Daphnia pulex*. *Science* **331**: 555.

Conte, M.P., Aleandri, M., Marazzato, M., Conte, A.L., Ambrosi, C., Nicoletti, M., Zagaglia, C., Gambara, G., Palombi, F., De Cesaris, P., Ziparo, E., Palamara, A.T., Riccioli, A., and Longhi, C. (2016) The Adherent/Invasive *Escherichia coli* Strain LF82 Invades and Persists in Human Prostate Cell Line RWPE-1, Activating a Strong Inflammatory Response. *Infect Immun* **84**: 3105.

Corbett, D., Wise, A., Langley, T., Skinner, K., Trimby, E., Birchall, S., Dorali, A., Sandiford, S., Williams, J., Warn, P., Vaara, M., and Lister, T. (2017) Potentiation of Antibiotic Activity by a Novel Cationic Peptide: Potency and Spectrum of Activity of SPR741. *Antimicrob Agents Chemother* **61**: e00200-00217.

Costa, V., Angelini, C., De Feis, I., and Ciccodicola, A. (2010) Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotechnol* **2010**: 853916-853916.

Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**: 1367-1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**: 1794-1805.

Craigen, W.J., and Caskey, C.T. (1986) Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature* **322**: 273-275.

Crappé, J., Ndah, E., Koch, A., Steyaert, S., Gawron, D., De Keulenaer, S., De Meester, E., De Meyer, T., Van Criekinge, W., Van Damme, P., and Menschaert, G. (2015) PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res* **43**: e29-e29.

Craven, M., Egan, C.E., Dowd, S.E., McDonough, S.P., Dogan, B., Denkers, E.Y., Bowman, D., Scherl, E.J., and Simpson, K.W. (2012) Inflammation drives dysbiosis and bacterial invasion in murine models of ileal Crohn's disease. *PLOS ONE* **7**: e41594-e41594.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res* **14**: 1188-1190.

D'Lima, N.G., Khitun, A., Rosenbloom, A.D., Yuan, P., Gassaway, B.M., Barber, K.W., Rinehart, J., and Slavoff, S.A. (2017) Comparative Proteomics Enables Identification of Nonannotated Cold Shock Proteins in *E. coli*. *J Proteome Res* **16**: 3722-3731.

Dar, D., Shamir, M., Mellin, J.R., Koutero, M., Stern-Ginossar, N., Cossart, P., and Sorek, R. (2016) Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science* **352**: aad9822.

Dard-Dascot, C., Naquin, D., d'Aubenton-Carafa, Y., Alix, K., Thermes, C., and van Dijk, E. (2018) Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *BMC Genomics* **19**: 118.

Darfeuille-Michaud, A., Boudeau, J., Bulois, P., Neut, C., Glasser, A.L., Barnich, N., Bringer, M.A., Swidsinski, A., Beaugerie, L., and Colombel, J.F. (2004) High prevalence of adherent-invasive *Escherichia coli* associated with ileal mucosa in Crohn's disease. *Gastroenterology* **127**: 412-421.

Darfeuille-Michaud, A., and Colombel, J.-F. (2008) Pathogenic *Escherichia coli* in inflammatory bowel diseases: Proceedings of the 1st International Meeting on *E. coli* and IBD, June 2007, Lille, France. *J Crohns Colitis* **2**: 255-262.

Darfeuille-Michaud, A., Neut, C., Barnich, N., Lederman, E., Di Martino, P., Desreumaux, P., Gambiez, L., Joly, B., Cortot, A., and Colombel, J.F. (1998) Presence of adherent *Escherichia coli* strains in ileal mucosa of patients with Crohn's disease. *Gastroenterology* **115**: 1405-1413.

Datsenko, K.A., and Wanner, B.L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci U S A* **97**: 6640-6645.

Datta, A.K., and Burma, D.P. (1972) Association of ribonuclease I with ribosomes and their subunits. *J Biol Chem* **247**: 6795-6801.

Davidovich, C., Bashan, A., Auerbach-Nevo, T., Yaggie, R.D., Gontarek, R.R., and Yonath, A. (2007) Induced-fit tightens pleuromutilins binding to ribosomes and remote interactions enable their selectivity. *Proc Natl Acad Sci U S A* **104**: 4291-4296.

Davies, K.J., Lloyd, D., and Boddy, L. (1989) The effect of oxygen on denitrification in *Paracoccus denitrificans* and *Pseudomonas aeruginosa*. *J Gen Microbiol* **135**: 2445-2451.

de Almeida, J.P.P., Vêncio, R.Z.N., Lorenzetti, A.P.R., Caten, F.T., Gomes-Filho, J.V., and Koide, T. (2019) The Primary Antisense Transcriptome of *Halobacterium salinarum* NRC-1. *Genes* **10**.

de Winter, J.C., Gosling, S.D., and Potter, J. (2016) Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychol Methods* **21**: 273-290.

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

delCardayré, S.B., and Raines, R.T. (1995) The extent to which ribonucleases cleave ribonucleic acid. *Anal Biochem* **225**: 176-178.

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

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**: 4636-4641.

DeRisi, J.L., Huber, G., Kistler, A., Retallack, H., Wilkinson, M., and Yllanes, D. (2019) An exploration of ambigrammatic sequences in narnaviruses. *Sci Rep* **9**: 17982.

Diament, A., and Tuller, T. (2016) Estimation of ribosome profiling performance and reproducibility at various levels of resolution. *Biol Direct* **11**: 24.

Díaz, E., Ferrández, A., Prieto, M.A., and García, J.L. (2001) Biodegradation of aromatic compounds by *Escherichia coli*. *Microbiol Mol Biol Rev* **65**: 523-569.

Diggle, S.P., and Whiteley, M. (2020) Microbe Profile: *Pseudomonas aeruginosa*: opportunistic pathogen and lab rat. *Microbiology* **166**: 30-33.

Dinan, A.M., Lukhovitskaya, N.I., Olendraite, I., and Firth, A.E. (2020) A case for a negative-strand coding sequence in a group of positive-sense RNA viruses. *Virus Evol* **6**: veaa007.

Dingwall, C., Lomonossoff, G.P., and Laskey, R.A. (1981) High sequence specificity of micrococcal nuclease. *Nucleic Acids Res* **9**: 2659-2674.

Doellinger, J., Schneider, A., Hoeller, M., and Lasch, P. (2020) Sample Preparation by Easy Extraction and Digestion (SPEED) - A Universal, Rapid, and Detergent-free Protocol for Proteomics Based on Acid Extraction. *Mol Cell Proteomics* **19**: 209.

Dogan, B., Scherl, E., Bosworth, B., Yantiss, R., Altier, C., McDonough, P.L., Jiang, Z.D., Dupont, H.L., Garneau, P., Harel, J., Rishniw, M., and Simpson, K.W. (2013) Multidrug resistance is common in *Escherichia coli* associated with ileal Crohn's disease. *Inflamm Bowel Dis* **19**: 141-150.

Dogan, B., Suzuki, H., Herlekar, D., Sartor, R.B., Campbell, B.J., Roberts, C.L., Stewart, K., Scherl, E.J., Araz, Y., Bitar, P.P., Lefébure, T., Chandler, B., Schukken, Y.H., Stanhope, M.J., and Simpson, K.W. (2014) Inflammation-associated Adherent-invasive *Escherichia coli* Are Enriched in Pathways for Use of Propanediol and Iron and M-cell Translocation. *Inflamm Bowel Dis* **20**: 1919-1932.

Domazet-Loso, T., and Tautz, D. (2003) An evolutionary analysis of orphan genes in *Drosophila. Genome Res* **13**: 2213-2219.

Dong, T., and Schellhorn, H.E. (2009) Control of RpoS in global gene expression of *Escherichia coli* in minimal media. *Mol Genet Genom* **281**: 19-33.

Dong, T., Yu, R., and Schellhorn, H. (2011) Antagonistic regulation of motility and transcriptome expression by RpoN and RpoS in *Escherichia coli. Mol Microbiol* **79**: 375-386.

Donoghue, M.T.A., Keshavaiah, C., Swamidatta, S.H., and Spillane, C. (2011) Evolutionary origins of *Brassicaceae* specific genes in *Arabidopsis thaliana. BMC Evol Biol* **11**: 47.

Dornenburg, J.E., Devita, A.M., Palumbo, M.J., and Wade, J.T. (2010) Widespread antisense transcription in *Escherichia coli. mBio* **1**: e00024-00010.

Dötsch, A., Eckweiler, D., Schniederjans, M., Zimmermann, A., Jensen, V., Scharfe, M., Geffers, R., and Häussler, S. (2012) The *Pseudomonas aeruginosa* Transcriptome in Planktonic Cultures and Static Biofilms Using RNA Sequencing. *PLOS ONE* **7**: e31092.

Dudas, G., Huber, G., Wilkinson, M., and Yllanes, D. (2021) Polymorphism of Genetic Ambigrams. *bioRxiv*: 2021.2002.2016.431493.

Dujon, B. (1996) The yeast genome project: what did we learn? *Trends Genet* **12**: 263-270.

Dunican, B.F., Hiller, D.A., and Strobel, S.A. (2015) Transition State Charge Stabilization and Acid–Base Catalysis of mRNA Cleavage by the Endoribonuclease RelE. *Biochemistry* **54**: 7048-7057.

Dunn, J.G., Foo, C.K., Belletier, N.G., Gavis, E.R., and Weissman, J.S. (2013) Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster. Elife* **2**: e01179.

Eaves-Pyles, T., Allen, C.A., Taormina, J., Swidsinski, A., Tutt, C.B., Jezek, G.E., Islas-Islas, M., and Torres, A.G. (2008) *Escherichia coli* isolated from a Crohn's disease patient adheres, invades, and induces inflammatory responses in polarized intestinal epithelial cells. *Int J Med Microbiol* **298**: 397-409.

Eckweiler, D., and Häussler, S. (2018) Antisense transcription in *Pseudomonas aeruginosa. Microbiology* **164**: 889-895.

Engel, J., and Balachandran, P. (2009) Role of *Pseudomonas aeruginosa* type III effectors in disease. *Curr Opin Microbiol* **12**: 61-66.

Erdmann, J., Thöming, J.G., Pohl, S., Pich, A., Lenz, C., and Häussler, S. (2019) The Core Proteome of Biofilm-Grown Clinical *Pseudomonas aeruginosa* Isolates. *Cells* **8**: 1129.

Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D.J., Weekes, M.P., Stevanovic, S., Zimmer, R., and Dölken, L. (2018) Improved Ribo-seq enables identification of cryptic translation events. *Nat Methods* **15**: 363-366.

Ermolaeva, M.D., White, O., and Salzberg, S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res* **29**: 1216-1221.

Eschbach, M., Schreiber, K., Trunk, K., Buer, J., Jahn, D., and Schobert, M. (2004) Long-term anaerobic survival of the opportunistic pathogen *Pseudomonas aeruginosa* via pyruvate fermentation. *J Bacteriol* **186**: 4596-4604.

Escherich, T., and Bettelheim, K. (1988) The Intestinal Bacteria of the Neonate and Breast-Fed Infant. *Rev Infect Dis* **10**: 1220-1225.

Estepa, V., Rojo-Bezares, B., Torres, C., and Saenz, Y. (2014) Faecal carriage of *Pseudomonas aeruginosa* in healthy humans: antimicrobial susceptibility and global genetic lineages. *FEMS Microbiol Ecol* **89**: 15-19.

Estrem, S.T., Gaal, T., Ross, W., and Gourse, R.L. (1998) Identification of an UP element consensus sequence for bacterial promoters. *Proc Natl Acad Sci U S A* **95**: 9761-9766.

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

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., Krymova, E., Gelfand, M.S., Govorun, V.M., Bogdanov, A.A., Sergiev, P.V., and Dontsova, O.A. (2017) Application of sorting and next generation sequencing to study 5´-UTR influence on translation efficiency in *Escherichia coli*. *Nucleic Acids Res* **45**: 3487-3502.

Faca, V., Pitteri, S.J., Newcomb, L., Glukhova, V., Phanstiel, D., Krasnoselsky, A., Zhang, Q., Struthers, J., Wang, H., Eng, J., Fitzgibbon, M., McIntosh, M., and Hanash, S. (2007) Contribution of protein fractionation to depth of analysis of the serum and plasma proteomes. *J Proteome Res* **6**: 3558-3565.

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

Fellner, L., Simon, S., Scherling, C., Witting, M., Schober, S., Polte, C., Schmitt-Kopplin, P., Keim, D.A., Scherer, S., and Neuhaus, K. (2015) Evidence for the recent origin of a bacterial protein-coding, overlapping orphan gene by evolutionary overprinting. *BMC Evol Biol* **15**: 283-283.

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., and Whitehouse, C.M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**: 64.

Fernandez, L., and Hancock, R.E. (2012) Adaptive and mutational resistance: role of porins and efflux pumps in drug resistance. *Clin Microbiol Rev* **25**: 661-681.

Fernando, D.M., and Kumar, A. (2013) Resistance-Nodulation-Division Multidrug Efflux Pumps in Gram-Negative Bacteria: Role in Virulence. *Antibiotics* **2**: 163-181.

Fields, A.P., Rodriguez, E.H., Jovanovic, M., Stern-Ginossar, N., Haas, B.J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S.A., Ingolia, N.T., Regev, A., and Weissman, J.S. (2015) A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol Cell* **60**: 816-827.

Fijalkowska, D., Fijalkowski, I., Willems, P., and Van Damme, P. (2020) Bacterial riboproteogenomics: the era of N-terminal proteoform existence revealed. *FEMS Microbiol Rev* **44**: 418-431.

Filiatrault, M.J., Stodghill, P.V., Bronstein, P.A., Moll, S., Lindeberg, M., Grills, G., Schweitzer, P., Wang, W., Schroth, G.P., Luo, S., Khrebtukova, I., Yang, Y., Thannhauser, T., Butcher, B.G., Cartinhour, S., and Schneider, D.J. (2010) Transcriptome Analysis of *Pseudomonas syringae* Identifies New Genes, Noncoding RNAs, and Antisense Activity. *J Bacteriol* **192**: 2359.

Filiatrault, M.J., Stodghill, P.V., Myers, C.R., Bronstein, P.A., Butcher, B.G., Lam, H., Grills, G., Schweitzer, P., Wang, W., Schneider, D.J., and Cartinhour, S.W. (2011) Genome-wide identification of transcriptional start sites in the plant pathogen *Pseudomonas syringae* pv. tomato str. DC3000. *PLOS ONE* **6**: e29335-e29335.

Filloux, A. (2011) Protein Secretion Systems in *Pseudomonas aeruginosa:* An Essay on Diversity, Evolution, and Function. *Front Microbiol* **2**: 155-155.

Fitzgerald, D.M., Bonocora, R.P., and Wade, J.T. (2014) Comprehensive mapping of the *Escherichia coli* flagellar regulatory network. *PLoS Genet* **10**: e1004649-e1004649.

Förstner, K.U., Vogel, J., and Sharma, C.M. (2014) READemption—a tool for the computational analysis of deep-sequencing–based transcriptome data. *Bioinformatics* **30**: 3421-3423.

Fremin, B.J., and Bhatt, A.S. (2020) Structured RNA Contaminants in Bacterial Ribo-Seq. *mSphere* **5**: e00855-00820.

Fremin, B.J., Sberro, H., and Bhatt, A.S. (2020) MetaRibo-Seq measures translation in microbiomes. *Nat Commun* **11**: 3268.

Friedman, R.C., Kalkhof, S., Doppelt-Azeroual, O., Mueller, S.A., Chovancová, M., von Bergen, M., and Schwikowski, B. (2017) Common and phylogenetically widespread coding for peptides by bacterial small RNAs. *BMC Genomics* **18**: 553-553.

Frimmersdorf, E., Horatzek, S., Pelnikevich, A., Wiehlmann, L., and Schomburg, D. (2010) How *Pseudomonas aeruginosa* adapts to various environments: a metabolomic approach. *Environ Microbiol* **12**: 1734-1747.

Fu, Y., Wu, P.-H., Beane, T., Zamore, P.D., and Weng, Z. (2018) Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics* **19**: 531.

Fuchs, R.T., Sun, Z., Zhuang, F., and Robb, G.B. (2015) Bias in Ligation-Based Small RNA Sequencing Library Construction Is Determined by Adaptor and RNA Structure. *PLOS ONE* **10**: e0126049.

Fukuchi, S., and Nishikawa, K. (2004) Estimation of the number of authentic orphan genes in bacterial genomes. *DNA Res* **11**: 219-231, 311-313.

Fukuda, H., Hosaka, M., Iyobe, S., Gotoh, N., Nishino, T., and Hirai, K. (1995) *nfxC*-type quinolone resistance in a clinical isolate of *Pseudomonas aeruginosa. Antimicrob Agents Chemother* **39**: 790-792.

Galvani, C., Terry, J., and Ishiguro, E.E. (2001) Purification of the RelB and RelE Proteins of *Escherichia coli:* RelE Binds to RelB and to Ribosomes. *J Bacteriol* **183**: 2700.

Gay, P., Le Coq, D., Steinmetz, M., Ferrari, E., and Hoch, J.A. (1983) Cloning structural gene *sacB*, which codes for exoenzyme levansucrase of *Bacillus subtilis*: expression of the gene in *Escherichia coli. J Bacteriol* **153**: 1424-1431.

Gelsinger, D.R., Dallon, E., Reddy, R., Mohammad, F., Buskirk, Allen R., and DiRuggiero, J. (2020) Ribosome profiling in archaea reveals leaderless translation, novel translational initiation sites, and ribosome pausing at single codon resolution. *Nucleic Acids Res* **48**: 5201-5216.

Gerashchenko, M.V., and Gladyshev, V.N. (2014) Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res* **42**: e134-e134.

Gerashchenko, M.V., and Gladyshev, V.N. (2017) Ribonuclease selection for ribosome profiling. *Nucleic Acids Res* **45**: e6-e6.

Gerdes, K. (2000) Toxin-antitoxin modules may regulate synthesis of macromolecules during nutritional stress. *J Bacteriol* **182**: 561-572.

Gerdes, K., Christensen, S.K., and Løbner-Olesen, A. (2005) Prokaryotic toxin–antitoxin stress response loci. *Nat Rev Microbiol* **3**: 371-382.

Gessard, C. (1882) Sur les colorations bleue et verte des linges a pansements. *C R Acad Sci* **94**: 536-538.

Gessulat, S., Schmidt, T., Zolg, D.P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, H.-C., Aiche, S., Kuster, B., and Wilhelm, M. (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods* **16**: 509-518.

Giamarellou, H., and Antoniadou, A. (2001) Antipseudomonal antibiotics. *Med Clin North Am* **85**: 19-42.

Giannoukos, G., Ciulla, D.M., Huang, K., Haas, B.J., Izard, J., Levin, J.Z., Livny, J., Earl, A.M., Gevers, D., Ward, D.V., Nusbaum, C., Birren, B.W., and Gnirke, A. (2012) Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol* **13**: R23-R23.

Giess, A., Jonckheere, V., Ndah, E., Chyżyńska, K., Van Damme, P., and Valen, E. (2017) Ribosome signatures aid bacterial translation initiation site identification. *BMC Biology* **15**: 76.

Gillet, L.C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012) Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Mol Cell Proteomics* **11**: O111.016717.

Glasser, A.L., Boudeau, J., Barnich, N., Perruchot, M.H., Colombel, J.F., and Darfeuille-Michaud, A. (2001) Adherent invasive *Escherichia coli* strains from patients with Crohn's disease survive and replicate within macrophages without inducing host cell death. *Infect Immun* **69**: 5529-5537.

Glaub, A., Huptas, C., Neuhaus, K., and Ardern, Z. (2020) Recommendations for bacterial ribosome profiling experiments based on bioinformatic evaluation of published data. *J Biol Chem* **295**: 8999-9011.

Goeders, N., Drèze, P.-L., and Van Melderen, L. (2013) Relaxed cleavage specificity within the RelE toxin family. *J Bacteriol* **195**: 2541-2549.

Gómez-Lozano, M., Marvig, R.L., Molin, S., and Long, K.S. (2012) Genome-wide identification of novel small RNAs in *Pseudomonas aeruginosa. Environ Microbiol* **14**: 2006-2016.

Gotfredsen, M., and Gerdes, K. (1998) The *Escherichia coli relBE* genes belong to a new toxin-antitoxin gene family. *Mol Microbiol* **29**: 1065-1076.

Gotoh, N., Itoh, N., Tsujimoto, H., Yamagishi, J., Oyamada, Y., and Nishino, T. (1994) Isolation of OprM-deficient mutants of *Pseudomonas aeruginosa* by transposon insertion mutagenesis: evidence of involvement in multiple antibiotic resistance. *FEMS Microbiol Lett* **122**: 267-273.

Govan, J.R., and Deretic, V. (1996) Microbial pathogenesis in cystic fibrosis: mucoid *Pseudomonas aeruginosa* and *Burkholderia cepacia. Microbiol Rev* **60**: 539-574.

Grady, S.L., Malfatti, S.A., Gunasekera, T.S., Dalley, B.K., Lyman, M.G., Striebich, R.C., Mayhew, M.B., Zhou, C.L., Ruiz, O.N., and Dugan, L.C. (2017) A comprehensive multi-omics approach uncovers adaptations for growth and survival of *Pseudomonas aeruginosa* on n-alkanes. *BMC Genomics* **18**: 334.

Graf, F., (2019) Analysis of translational start sites of overlapping genes in *Escherichia coli* LF82 using ribosome profiling. Master Thesis. Technische Universität München.

Grainger, D.C. (2016) The unexpected complexity of bacterial genomes. *Microbiology* **162**: 1167-1172.

Grassé, P.P., (1977) *Evolution of Living Organisms: Evidence for a New Theory of Transformation.* Academic Press, New York, San Francisco, London.

Green, S.K., Schroth, M.N., Cho, J.J., Kominos, S.K., and Vitanza-jack, V.B. (1974) Agricultural plants and soil as a reservoir for *Pseudomonas aeruginosa. Appl Microbiol* **28**: 987-991.

Griffin, M.A., Davis, J.H., and Strobel, S.A. (2013) Bacterial toxin RelE: a highly efficient ribonuclease with exquisite substrate specificity using atypical catalytic residues. *Biochemistry* **52**: 8633-8642.

Guan, Y., Zhu, Q., Huang, D., Zhao, S., Jan Lo, L., and Peng, J. (2015) An equation to estimate the difference between theoretically predicted and SDS PAGE-displayed molecular weights for an acidic peptide. *Sci Rep* **5**: 13370.

Gudyś, A., and Deorowicz, S. (2017) QuickProbs 2: Towards rapid construction of high-quality alignments of large protein families. *Sci Rep* **7**: 41553.

Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., Rode, M., Suyama, M., Schmidt, S., Gavin, A.-C., Bork, P., and Serrano, L. (2009) Transcriptome Complexity in a Genome-Reduced Bacterium. *Science* **326**: 1268.

Gupta, N., Tanner, S., Jaitly, N., Adkins, J.N., Lipton, M., Edwards, R., Romine, M., Osterman, A., Bafna, V., Smith, R.D., and Pevzner, P.A. (2007) Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res* **17**: 1362-1377.

Gusarov, I., and Nudler, E. (1999) The mechanism of intrinsic transcription termination. *Mol Cell* **3**: 495-504.

Haas, B.J., Chin, M., Nusbaum, C., Birren, B.W., and Livny, J. (2012) How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics* **13**: 734.

Hafner, M., Renwick, N., Brown, M., Mihailović, A., Holoch, D., Lin, C., Pena, J.T.G., Nusbaum, J.D., Morozov, P., Ludwig, J., Ojo, T., Luo, S., Schroth, G., and Tuschl, T. (2011) RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* **17**: 1697-1712.

Han, X., Aslanian, A., and Yates, J.R., 3rd (2008) Mass spectrometry for proteomics. *Curr Opin Chem Biol* **12**: 483-490.

Hansen, K.D., Brenner, S.E., and Dudoit, S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38**: e131-e131.

Hare, N.J., Solis, N., Harmer, C., Marzook, N.B., Rose, B., Harbour, C., Crossett, B., Manos, J., and Cordwell, S.J. (2012) Proteomic profiling of *Pseudomonas aeruginosa* AES-1R, PAO1 and PA14 reveals potential virulence determinants associated with a transmissible cystic fibrosis-associated strain. *BMC Microbiol* **12**: 16.

Haycocks, J.R.J., and Grainger, D.C. (2016) Unusually Situated Binding Sites for Bacterial Transcription Factors Can Have Hidden Functionality. *PLOS ONE* **11**: e0157016.

He, S., Wurtzel, O., Singh, K., Froula, J.L., Yilmaz, S., Tringe, S.G., Wang, Z., Chen, F., Lindquist, E.A., Sorek, R., and Hugenholtz, P. (2010) Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat Methods* **7**: 807-812.

Heames, B., Schmitz, J., and Bornberg-Bauer, E. (2020) A Continuum of Evolving *De Novo* Genes Drives Protein-Coding Novelty in *Drosophila. J Mol Evol* **88**: 382-398.

Hecht, A., Glasgow, J., Jaschke, P.R., Bawazer, L.A., Munson, M.S., Cochran, J.R., Endy, D., and Salit, M. (2017) Measurements of translation initiation from all 64 codons in *E. coli. Nucleic Acids Res* **45**: 3615-3626.

Hemm, M.R., Weaver, J., and Storz, G. (2020) *Escherichia coli* Small Proteome. *EcoSal Plus* **9**: 10.1128/ecosalplus.ESP-0031-2019.

Holloway, B.W. (1955) Genetic recombination in *Pseudomonas aeruginosa. J Gen Microbiol* **13**: 572-581.

Housseini B Issa, K., Phan, G., and Broutin, I. (2018) Functional Mechanism of the Efflux Pumps Transcription Regulators From *Pseudomonas aeruginosa* Based on 3D Structures. *Front Mol Biosci* **5**.

Hsieh, E.J., Bereman, M.S., Durand, S., Valaskovic, G.A., and MacCoss, M.J. (2013) Effects of column and gradient lengths on peak capacity and peptide identification in nanoflow LC-MS/MS of complex proteomic samples. *J Am Soc Mass Spectrom* **24**: 148-153.

Hsu, P.Y., Calviello, L., Wu, H.-Y.L., Li, F.-W., Rothfels, C.J., Ohler, U., and Benfey, P.N. (2016) Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proc Natl Acad Sci U S A*: 201614788.

Hu, P., Janga, S.C., Babu, M., Díaz-Mejía, J.J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P., Chandran, S., Christopoulos, C., Nazarians-Armavil, A., Nasseri, N.K., Musso, G., Ali, M., Nazemof, N., Eroukova, V., Golshani, A., Paccanaro, A., Greenblatt, J.F., Moreno-Hagelsieb, G., and Emili, A. (2009)

Global Functional Atlas of *Escherichia coli* Encompassing Previously Uncharacterized Proteins. *PLOS Biol* **7**: e1000096.

Hücker, S.M., Ardern, Z., Goldberg, T., Schafferhans, A., Bernhofer, M., Vestergaard, G., Nelson, C.W., Schloter, M., Rost, B., Scherer, S., and Neuhaus, K. (2017) Discovery of numerous novel small genes in the intergenic regions of the *Escherichia coli* O157:H7 Sakai genome. *PLOS ONE* **12**: e0184119.

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

Hücker, S.M., Vanderhaeghen, S., Abellan-Schneyder, I., Wecko, R., Simon, S., Scherer, S., and 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 Evol Biol* **18**: 21-21.

Hücker, S.M.M., (2018) RIBOseq-based discovery of non-annotated genes in *Escherichia coli* O157:H7 Sakai and their functional characterization. Doctoral Thesis. Technische Universiät München.

Hurley, J.M., Cruz, J.W., Ouyang, M., and Woychik, N.A. (2011) Bacterial toxin RelE mediates frequent codon-independent mRNA cleavage from the 5' end of coding regions in vivo. *J Biol Chem* **286**: 14770-14778.

Hussmann, J.A., Patchett, S., Johnson, A., Sawyer, S., and Press, W.H. (2015) Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLoS Genet* **11**: e1005732.

Hwang, J.-Y., and Buskirk, A.R. (2017) A ribosome profiling study of mRNA cleavage by the endonuclease RelE. *Nucleic Acids Res* **45**: 327-336.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119-119.

Impens, F., Rolhion, N., Radoshevich, L., Bécavin, C., Duval, M., Mellin, J., García Del Portillo, F., Pucciarelli, M.G., Williams, A.H., and Cossart, P. (2017) N-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in *Listeria monocytogenes. Nat Microbiol* **2**: 17005-17005.

Ingolia, N.T. (2010) Genome-wide translational profiling by ribosome footprinting. *Methods Enzymol* **470**: 119-142.

Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7**: 1534-1550.

Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J., Jackson, S.E., Wills, M.R., and Weissman, J.S. (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8**: 1365-1379.

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

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**: 789-802.

Ishii, S., and Sadowsky, M.J. (2008) *Escherichia coli* in the Environment: Implications for Water Quality and Human Health. *Microbes Environ* **23**: 101-108.

Iyer, R., and Erwin, A.L. (2015) Direct measurement of efflux in *Pseudomonas aeruginosa* using an environment-sensitive fluorescent dye. *Res Microbiol* **166**: 516-524.

Jacob, F. (1977) Evolution and tinkering. *Science* **196**: 1161.

Jäger, D., Förstner, K.U., Sharma, C.M., Santangelo, T.J., and Reeve, J.N. (2014) Primary transcriptome map of the hyperthermophilic archaeon *Thermococcus kodakarensis. BMC Genomics* **15**: 684.

Jakics, E.B., Iyobe, S., Hirai, K., Fukuda, H., and Hashimoto, H. (1992) Occurrence of the *nfxB* type mutation in clinical isolates of *Pseudomonas aeruginosa. Antimicrob Agents Chemother* **36**: 2562-2565.

Jangam, D., Feschotte, C., and Betrán, E. (2017) Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends Genet* **33**: 817-831.

Jayaprakash, A.D., Jabado, O., Brown, B.D., and Sachidanandam, R. (2011) Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res* **39**: e141.

Jensen, K.T., Petersen, L., Falk, S., Iversen, P., Andersen, P., Theisen, M., and Krogh, A. (2006) Novel overlapping coding sequences in *Chlamydia trachomatis. FEMS Microbiol Lett* **265**: 106-117.

Jeong, Y., Kim, J.N., Kim, M.W., Bucca, G., Cho, S., Yoon, Y.J., Kim, B.G., Roe, J.H., Kim, S.C., Smith, C.P., and Cho, B.K. (2016) The dynamic transcriptional and translational landscape of the model antibiotic producer S*treptomyces coelicolor* A3(2). *Nat Commun* **7**: 11605.

Ji, Z., Song, R., Huang, H., Regev, A., and Struhl, K. (2016) Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat Biotechnol* **34**: 410-413.

Johnson, Z.I., and Chisholm, S.W. (2004) Properties of overlapping genes are conserved across microbial genomes. *Genome Res* **14**: 2268-2272.

Jyot, J., Balloy, V., Jouvion, G., Verma, A., Touqui, L., Huerre, M., Chignard, M., and Ramphal, R. (2011) Type II secretion system of *Pseudomonas aeruginosa:* in vivo evidence of a significant role in death due to lung infection. *J Infect Dis* **203**: 1369-1377.

Kahvejian, A., Quackenbush, J., and Thompson, J.F. (2008) What would you do if you could sequence everything? *Nat Biotechnol* **26**: 1125-1133.

Kamath, K.S., Krisp, C., Chick, J., Pascovici, D., Gygi, S.P., and Molloy, M.P. (2017) *Pseudomonas aeruginosa* Proteome under Hypoxic Stress Conditions Mimicking the Cystic Fibrosis Lung. *J Proteome Res* **16**: 3917-3928.

Kaniga, K., Delor, I., and 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**: 137-141.

Kans, J., (2013) Entrez Direct: E-utilities on the Unix Command Line [Updated 2021 Apr 29]. In: Entrez Programming Utilities Help [Internet]. Bethesda: National Center for Biotechnology Information (US).

Kaper, J.B., Nataro, J.P., and Mobley, H.L.T. (2004) Pathogenic *Escherichia coli. Nat Rev Microbiol* **2**: 123-140.

Karas, M., and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **60**: 2299-2301.

Karlin, S., Mrázek, J., Campbell, A., and Kaiser, D. (2001) Characterizations of highly expressed genes of four fast-growing bacteria. *J Bacteriol* **183**: 5025-5040.

Kavanagh, F., Hervey, A., and Robbins, W.J. (1951) Antibiotic Substances From Basidiomycetes: VIII. *Pleurotus Multilus* (Fr.) *Sacc.* and *Pleurotus Passeckerianus Pilat. Proc Natl Acad Sci U S A* **37**: 570-574.

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

Keilty, S., and Rosenberg, M. (1987) Constitutive function of a positively regulated promoter reveals new sequences essential for activity. *J Biol Chem* **262**: 6389-6395.

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., and Bosch, T.C.G. (2009) More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* **25**: 404-413.

Kim, D., Hong, J.S.-J., Qiu, Y., Nagarajan, H., Seo, J.-H., Cho, B.-K., Tsai, S.-F., and Palsson, B.Ø. (2012) Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling. *PLoS Genet* **8**: e1002867-e1002867.

Kim, W., and Levy, S.B. (2008) Increased Fitness of *Pseudomonas fluorescens* Pf0-1 Leucine Auxotrophs in Soil. *Appl Environ Microbiol* **74**: 3644.

Kim, W., Silby, M.W., Purvine, S.O., Nicoll, J.S., Hixson, K.K., Monroe, M., Nicora, C.D., Lipton, M.S., and Levy, S.B. (2009) Proteomic Detection of Non-Annotated Protein-Coding Genes in *Pseudomonas fluorescens* Pf0-1. *PLOS ONE* **4**: e8455.

Kimata, N., Nishino, T., Suzuki, S., and Kogure, K. (2004) *Pseudomonas aeruginosa* isolated from marine environments in Tokyo Bay. *Microb Ecol* **47**: 41-47.

Kirchberger, P.C., Schmidt, M.L., and Ochman, H. (2020) The Ingenuity of Bacterial Genomes. *Annu Rev Microbiol* **74**: 815-834.

Kitahara, K., and Miyazaki, K. (2011) Specific inhibition of bacterial RNase T2 by helix 41 of 16S ribosomal RNA. *Nat Commun* **2**: 1-7.

Koch, A., Gawron, D., Steyaert, S., Ndah, E., Crappé, J., De Keulenaer, S., De Meester, E., Ma, M., Shen, B., Gevaert, K., Van Criekinge, W., Van Damme, P., and Menschaert, G. (2014) A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics* **14**: 2688-2698.

Köhler, T., Epp, S.F., Curty, L.K., and Pechère, J.C. (1999) Characterization of MexT, the regulator of the MexE-MexF-OprN multidrug efflux system of *Pseudomonas aeruginosa. J Bacteriol* **181**: 6300-6305.

Komarova, E.S., Chervontseva, Z.S., Osterman, I.A., Evfratov, S.A., Rubtsova, M.P., Zatsepin, T.S., Semashko, T.A., Kostryukova, E.S., Bogdanov, A.A., Gelfand, M.S., Dontsova, O.A., and Sergiev, P.V. (2020) Influence of the spacer region between the Shine–Dalgarno box and the start codon for fine-tuning of the translation efficiency in *Escherichia coli. Microb Biotechnol* **13**: 1254-1261.

Koronakis, V., Sharff, A., Koronakis, E., Luisi, B., and Hughes, C. (2000) Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature* **405**: 914-919.

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

Kreitmeier, M., Ardern, Z., Abele, M., Ludwig, C., Scherer, S., and Neuhaus, K. (2021) Shadow ORFs illuminated: long overlapping genes in *Pseudomonas aeruginosa* are translated and under purifying selection. *bioRxiv*: 2021.2002.2009.430400.

Kremer, F.S., Eslabão, M.R., Dellagostin, O.A., and Pinto, L.d.S. (2016) Genix: a new online automated pipeline for bacterial genome annotation. *FEMS Microbiol Lett* **363**.

Kruft, V., Eubel, H., Jänsch, L., Werhahn, W., and Braun, H.P. (2001) Proteomic approach to identify novel mitochondrial proteins in *Arabidopsis*. *Plant Physiol* **127**: 1694-1710.

Kumar, D., Bansal, G., Narang, A., Basak, T., Abbas, T., and Dash, D. (2016) Integrating transcriptome and proteome profiling: Strategies and applications. *Proteomics* **16**: 2533-2544.

Kumari, H., Murugapiran, S.K., Balasubramanian, D., Schneper, L., Merighi, M., Sarracino, D., Lory, S., and Mathee, K. (2014) LTQ-XL mass spectrometry proteome analysis expands the *Pseudomonas aeruginosa* AmpR regulon to include cyclic di-GMP phosphodiesterases and phosphoproteins, and identifies novel open reading frames. *J Proteomics* **96**: 328-342.

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

Laass, S., Monzon, V.A., Kliemt, J., Hammelmann, M., Pfeiffer, F., Förstner, K.U., and Soppa, J. (2019) Characterization of the transcriptome of *Haloferax volcanii,* grown under four different conditions, with mixed RNA-Seq. *PLOS ONE* **14**: e0215986.

LaBauve, A.E., and Wargo, M.J. (2012) Growth and laboratory maintenance of *Pseudomonas aeruginosa. Curr Protoc Microbiol* **Chapter 6**: Unit-6E.1.

Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**: 680-685.

Lambert, P.A. (2002) Mechanisms of antibiotic resistance in *Pseudomonas aeruginosa. J R Soc Med* **95 Suppl 41**: 22-26.

Landry, C.R., Zhong, X., Nielly-Thibault, L., and Roucou, X. (2015) Found in translation: functions and evolution of a recently discovered alternative proteome. *Curr Opin Struct Biol* **32**: 74-80.

Landstorfer, R., Simon, S., Schober, S., Keim, D., Scherer, S., and 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.

Langmead, B., and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Lareau, L.F., Hite, D.H., Hogan, G.J., and Brown, P.O. (2014) Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife* **3**: e01257-e01257.

Lee, C.-Y., Chiu, Y.-C., Wang, L.-B., Kuo, Y.-L., Chuang, E.Y., Lai, L.-C., and Tsai, M.-H. (2013) Common applications of next-generation sequencing technologies in genomic research. *Transl Cancer Res* **2**: 33-45.

Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., and Qian, S.-B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A* **109**: E2424.

Lejars, M., Kobayashi, A., and Hajnsdorf, E. (2019) Physiological roles of antisense RNAs in prokaryotes. *Biochimie* **164**: 3-16.

Levine, A.J., and Oren, M. (2009) The first 30 years of p53: growing ever more complex. *Nat Rev Cancer* **9**: 749-758.

Lewenza, S., Gardy, J.L., Brinkman, F.S.L., and Hancock, R.E.W. (2005) Genome-wide identification of *Pseudomonas aeruginosa* exported proteins using a consensus computational strategy combined with a laboratory-based PhoA fusion screen. *Genome Res* **15**: 321-329.

Lewis, D.E.A., and Adhya, S. (2004) Axiom of determining transcription start points by RNA polymerase in *Escherichia coli. Mol Microbiol* **54**: 692-701.

Li, G.W., Oh, E., and Weissman, J.S. (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**: 538-541.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li, X.-T., Thomason, L.C., Sawitzke, J.A., Costantino, N., and Court, D.L. (2013) Positive and negative selection using the *tetA-sacB* cassette: recombineering and P1 transduction in *Escherichia coli. Nucleic Acids Res* **41**: e204-e204.

Li, X.Z., Plesiat, P., and Nikaido, H. (2015) The challenge of efflux-mediated antibiotic resistance in Gram-negative bacteria. *Clin Microbiol Rev* **28**: 337-418.

Lister, P.D., Wolter, D.J., and Hanson, N.D. (2009) Antibacterial-resistant *Pseudomonas aeruginosa*: clinical impact and complex regulation of chromosomally encoded resistance mechanisms. *Clin Microbiol Rev* **22**: 582-610.

Livak, K.J., and Schmittgen, T.D. (2001) Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* **25**: 402-408.

Livermore, D.M. (2009) Has the era of untreatable infections arrived? *J Antimicrob Chemother* **64 Suppl 1**: i29-36.

Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H., Bork, P., Glass, J.I., Serrano, L., and Lluch-Senar, M. (2016) Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci Adv* **2**: e1501363.

Loh, E., Righetti, F., Eichner, H., Twittenhoff, C., and Narberhaus, F. (2018) RNA thermometers in bacterial pathogens. *Microbiol Spectr* **6**.

Lomovskaya, O., Lee, A., Hoshino, K., Ishida, H., Mistry, A., Warren, M.S., Boyer, E., Chamberland, S., and Lee, V.J. (1999) Use of a genetic approach to evaluate the consequences of inhibition of efflux pumps in *Pseudomonas aeruginosa. Antimicrob Agents Chemother* **43**: 1340-1346.

Loughran, G., Zhdanov, A.V., Mikhaylova, M.S., Rozov, F.N., Datskevich, P.N., Kovalchuk, S.I., Serebryakova, M.V., Kiniry, S.J., Michel, A.M., O'Connor, P.B.F., Papkovsky, D.B., Atkins, J.F., Baranov, P.V., Shatsky, I.N., and Andreev, D.E. (2020) Unusually efficient CUG initiation of an overlapping reading frame in *POLG* mRNA yields novel protein POLGARF. *Proc Natl Acad Sci U S A* **117**: 24936.

LoVullo, E.D., and Schweizer, H.P. (2020) *Pseudomonas aeruginosa mexT* is an indicator of PAO1 strain integrity. *J Med Microbiol* **69**: 139-145.

Lukjancenko, O., Wassenaar, T.M., and Ussery, D.W. (2010) Comparison of 61 sequenced *Escherichia coli* genomes. *Microb Ecol* **60**: 708-720.

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

Lynch, M., and Marinov, G.K. (2015) The bioenergetic costs of a gene. *Proc Natl Acad Sci U S A* **112**: 15690.

Ma, D., Cook, D.N., Alberti, M., Pon, N.G., Nikaido, H., and Hearst, J.E. (1993) Molecular cloning and characterization of *acrA* and *acrE* genes of *Escherichia coli. J Bacteriol* **175**: 6299-6313.

Ma, D., Cook, D.N., Alberti, M., Pon, N.G., Nikaido, H., and Hearst, J.E. (1995) Genes *acrA* and *acrB* encode a stress-induced efflux system of *Escherichia coli. Mol Microbiol* **16**: 45-55.

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

MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., Kern, R., Tabb, D.L., Liebler, D.C., and MacCoss, M.J. (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**: 966-968.

Macnab, R.M. (1992) Genetics and biogenesis of bacterial flagella. *Annu Rev Genet* **26**: 131-158.

Mahajan-Miklos, S., Rahme, L.G., and Ausubel, F.M. (2000) Elucidating the molecular mechanisms of bacterial virulence using non-mammalian hosts. *Mol Microbiol* **37**: 981-988.

Mahajan-Miklos, S., Tan, M.-W., Rahme, L.G., and Ausubel, F.M. (1999) Molecular Mechanisms of Bacterial Virulence Elucidated Using a *Pseudomonas aeruginosa-Caenorhabditis elegans* Pathogenesis Model. *Cell* **96**: 47-56.

Makalowska, I., Lin, C.F., and Makalowski, W. (2005) Overlapping genes in vertebrate genomes. *Comput Biol Chem* **29**: 1-12.

Manadas, B., Mendes, V.M., English, J., and Dunn, M.J. (2010) Peptide fractionation in proteomics approaches. *Expert Rev Proteomics* **7**: 655-663.

Mangano, K., Florin, T., Shao, X., Klepacki, D., Chelysheva, I., Ignatova, Z., Gao, Y., Mankin, A.S., and Vázquez-Laslop, N. (2020) Genome-wide effects of the antimicrobial peptide apidaecin on translation termination in bacteria. *Elife* **9**: e62655.

Manoil, C., and Beckwith, J. (1985) TnphoA: a transposon probe for protein export signals. *Proc Natl Acad Sci U S A* **82**: 8129-8133.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**: 1509-1517.

Martinez-Medina, M., Aldeguer, X., Lopez-Siles, M., González-Huix, F., López-Oliu, C., Dahbi, G., Blanco, J.E., Blanco, J., Garcia-Gil, L.J., and Darfeuille-Michaud, A. (2009) Molecular diversity of *Escherichia coli* in the human gut: new ecological evidence supporting the role of adherent-invasive *E. coli* (AIEC) in Crohn's disease. *Inflamm Bowel Dis* **15**: 872-882.

Maseda, H., Saito, K., Nakajima, A., and Nakae, T. (2000) Variation of the *mexT* gene, a regulator of the MexEF-oprN efflux pump expression in wild-type strains of *Pseudomonas aeruginosa. FEMS Microbiol Lett* **192**: 107-112.

Masuda, N., Sakagawa, E., Ohya, S., Gotoh, N., Tsujimoto, H., and Nishino, T. (2000) Substrate Specificities of MexAB-OprM, MexCD-OprJ, and MexXY-OprM Efflux Pumps in *Pseudomonas aeruginosa. Antimicrob Agents Chemother* **44**: 3322.

Matsukawa, M., and Greenberg, E.P. (2004) Putative exopolysaccharide synthesis genes influence *Pseudomonas aeruginosa* biofilm development. *J Bacteriol* **186**: 4449-4456.

May, T.B., Shinabarger, D., Maharaj, R., Kato, J., Chu, L., DeVault, J.D., Roychoudhury, S., Zielinski, N.A., Berry, A., Rothmel, R.K., and *et al.* (1991) Alginate synthesis by *Pseudomonas aeruginosa*: a key pathogenic factor in chronic pulmonary infections of cystic fibrosis patients. *Clin Microbiol Rev* **4**: 191-206.

McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C.A., Vanderpool, C.K., and Tjaden, B. (2013) Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res* **41**: e140-e140.

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

Meconi, S., Vercellone, A., Levillain, F., Payré, B., Al Saati, T., Capilla, F., Desreumaux, P., Darfeuille-Michaud, A., and Altare, F. (2007) Adherent-invasive *Escherichia coli* isolated from Crohn's disease patients induce granulomas in vitro. *Cell Microbiol* **9**: 1252-1261.

Meier, F., Geyer, P.E., Virreira Winter, S., Cox, J., and Mann, M. (2018) BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat Methods* **15**: 440-448.

Méric, G., Kemsley, E.K., Falush, D., Saggers, E.J., and Lucchini, S. (2013) Phylogenetic distribution of traits associated with plant colonization in *Escherichia coli. Environ Microbiol* **15**: 487-501.

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

Metzker, M.L. (2005) Emerging technologies in DNA sequencing. *Genome Res* **15**: 1767-1776.

Meydan, S., Marks, J., Klepacki, D., Sharma, V., Baranov, P.V., Firth, A.E., Margus, T., Kefi, A., Vazquez-Laslop, N., and Mankin, A.S. (2019) Retapamulin-Assisted Ribosome Profiling Reveals the Alternative Bacterial Proteome. *Mol Cell* **74**: 481-493.e486.

Meydan, S., Vázquez-Laslop, N., and Mankin, A.S. (2018) Genes within Genes in Bacterial Genomes. *Microbiol Spectr* **6**.

Michel, A.M., Choudhury, K.R., Firth, A.E., Ingolia, N.T., Atkins, J.F., and Baranov, P.V. (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res* **22**: 2219-2229.

Migula, W. (1894) Über ein neues System der Bakterien. Arbeiten aus dem Bakteriologischen Institut der Technischen Hochschule zu Karlsruhe, 1:235-238.

Mir, K., Neuhaus, K., Bossert, M., and Schober, S. (2014) Short Barcodes for Next Generation Sequencing. *PLOS ONE* **8**: e82933.

Mir, K., Neuhaus, K., Scherer, S., Bossert, M., and Schober, S. (2012) Predicting Statistical Properties of Open Reading Frames in Bacterial Genomes. *PLOS ONE* **7**: e45103.

Miranda-CasoLuengo, A.A., Staunton, P.M., Dinan, A.M., Lohan, A.J., and Loftus, B.J. (2016) Functional characterization of the *Mycobacterium abscessus* genome coupled with condition specific transcriptomics reveals conserved molecular strategies for host adaptation and persistence. *BMC Genomics* **17**: 553.

Miravet-Verde, S., Ferrar, T., Espadas-García, G., Mazzolini, R., Gharrab, A., Sabido, E., Serrano, L., and Lluch-Senar, M. (2019) Unraveling the hidden universe of small proteins in bacterial genomes. *Mol Syst Biol* **15**: e8290.

Missiakas, D., Georgopoulos, C., and Raina, S. (1993) The *Escherichia coli* heat shock gene *htpY*: mutational analysis, cloning, sequencing, and transcriptional regulation. *J Bacteriol* **175**: 2613-2624.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, Gustavo A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., and Bateman, A. (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**: D412-D419.

Mitra, P., Ghosh, G., Hafeezunnisa, M., and Sen, R. (2017) Rho Protein: Roles and Mechanisms. *Annu Rev Microbiol* **71**: 687-709.

Miyata, T., and Yasunaga, T. (1978) Evolution of overlapping genes. *Nature* **272**: 532-535.

Mohammad, F., Green, R., and Buskirk, A.R. (2019) A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *Elife* **8**: e42591.

Mohammad, F., Woolstenhulme, C.J., Green, R., and Buskirk, A.R. (2016) Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Reports* **14**: 686-694.

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

Morrison, A.J., Jr., and Wenzel, R.P. (1984) Epidemiology of infections due to *Pseudomonas aeruginosa. Rev Infect Dis* **6 Suppl 3**: S627-642.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621-628.

Moshensky, D., and Alexeevski, A. (2019) Long antiparallel open reading frames are unlikely to be encoding essential proteins in prokaryotic genomes. *bioRxiv*: 724807.

Mouton, J.W., Meletiadis, J., Voss, A., and Turnidge, J. (2018) Variation of MIC measurements: the contribution of strain and laboratory variability to measurement precision. *J Antimicrob Chemother* **73**: 2374-2379.

Mulcahy, L.R., Burns, J.L., Lory, S., and Lewis, K. (2010) Emergence of *Pseudomonas aeruginosa* strains producing high levels of persister cells in patients with cystic fibrosis. *J Bacteriol* **192**: 6191-6199.

Müller, S.A., Kohajda, T., Findeiss, S., Stadler, P.F., Washietl, S., Kellis, M., von Bergen, M., and Kalkhof, S. (2010) Optimization of parameters for coverage of low molecular weight proteins. *Anal Bioanal Chem* **398**: 2867-2881.

Murakami, S., Nakashima, R., Yamashita, E., and Yamaguchi, A. (2002) Crystal structure of bacterial multidrug efflux transporter AcrB. *Nature* **419**: 587-593.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344-1349.

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

Nataro, J.P., and Kaper, J.B. (1998) Diarrheagenic *Escherichia coli. Clin Microbiol Rev* **11**: 142.

Ndah, E., Jonckheere, V., Giess, A., Valen, E., Menschaert, G., and Van Damme, P. (2017) REPARATION: ribosome profiling assisted (re-)annotation of bacterial genomes. *Nucleic Acids Res* **45**: e168.

Needleman, S.B., and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443-453.

Nelson, C.W., Ardern, Z., and Wei, X. (2020) OLGenie: Estimating Natural Selection to Predict Functional Overlapping Genes. *Mol Biol Evol* **37**: 2440-2449.

Neme, R., Amador, C., Yildirim, B., McConnell, E., and Tautz, D. (2017) Random sequences are an abundant source of bioactive RNAs or peptides. *Nat Ecol Evol* **1**: 0127.

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

Neuhaus, K., Landstorfer, R., Fellner, L., Simon, S., Schafferhans, A., Goldberg, T., Marx, H., Ozoline, O.N., Rost, B., Kuster, B., Keim, D.A., and 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**: 133-133.

Neuhaus, K., Landstorfer, R., Simon, S., Schober, S., Wright, P.R., Smith, C., Backofen, R., Wecko, R., Keim, D.A., and Scherer, S. (2017) Differentiation of ncRNAs from small mRNAs in *Escherichia coli* O157:H7 EDL933 (EHEC) by combined RNAseq and RIBOseq - *ryhB* encodes the regulatory RNA RyhB and a peptide, RyhP. *BMC Genomics* **18**: 216-216.

Neuhaus, K., Oelke, D., Fürst, D., Scherer, S., and Keim, D.A., (2010) Towards Automatic Detecting of Overlapping Genes - Clustered BLAST Analysis of Viral Genomes. In: Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. C. Pizzuti, M.D. Ritchie & M. Giacobini (eds). Springer, Berlin Heidelberg.

Ngan, J.Y.G., Pasunooti, S., Tse, W., Meng, W., Ngan, S.F.C., Jia, H., Lin, J.Q., Ng, S.W., Jaafa, M.T., Cho, S.L.S., Lim, J., Koh, H.Q.V., Abdul Ghani, N., Pethe, K., Sze, S.K., Lescar, J., and Alonso, S. (2021) HflX is a GTPase that controls hypoxia-induced replication arrest in slow-growing mycobacteria. *Proc Natl Acad Sci U S A* **118**.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**: 268-274.

Nie, L., Wu, G., and Zhang, W. (2006) Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in *Desulfovibrio vulgaris*: a quantitative analysis. *Genetics* **174**: 2229-2243.

Nissum, M., Kuhfuss, S., Hauptmann, M., Obermaier, C., Sukop, U., Wildgruber, R., Weber, G., Eckerskorn, C., and Malmström, J. (2007) Two-dimensional separation of human plasma proteins using iterative free-flow electrophoresis. *Proteomics* **7**: 4218-4227.

O'Connor, P.B., Li, G.W., Weissman, J.S., Atkins, J.F., and Baranov, P.V. (2013) rRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome profiling experiments. *Bioinformatics* **29**: 1488-1491.

Oh, E., Becker, A.H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R.J., Typas, A., Gross, C.A., Kramer, G., Weissman, J.S., and Bukau, B. (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor *in vivo*. *Cell* **147**: 1295-1308.

Ohno, S., (1970) *Evolution by Gene Duplication*. Springer-Verlag, Berlin Heidelberg.

Oliver, J.L., and Marín, A. (1996) A relationship between GC content and coding-sequence length. *J Mol Evol* **43**: 216-223.

Oliver, S.G., van der Aart, Q.J., Agostoni-Carbone, M.L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J.P., Benit, P., and *et al.* (1992) The complete DNA sequence of yeast chromosome III. *Nature* **357**: 38-46.

Orr, M.W., Mao, Y., Storz, G., and Qian, S.-B. (2019) Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res* **48**: 1029-1042.

Oshlack, A., Robinson, M.D., and Young, M.D. (2010) From RNA-seq reads to differential expression results. *Genome Biol* **11**: 220.

Overbeek, R., Bartels, D., Vonstein, V., and Meyer, F. (2007) Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chem Rev* **107**: 3431-3447.

Overgaard, M., Borch, J., and Gerdes, K. (2009) RelB and RelE of *Escherichia coli* form a tight complex that represses transcription via the ribbon-helix-helix motif in RelB. *J Mol Biol* **394**: 183-196.

Ozer, E.A., Allen, J.P., and Hauser, A.R. (2014) Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt. *BMC Genomics* **15**: 737.

Paget, M.S.B., and Helmann, J.D. (2003) The sigma70 family of sigma factors. *Genome Biol* **4**: 203-203.

Pallejà, A., Harrington, E.D., and Bork, P. (2008) Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics* **9**: 335.

Palleroni, N., Krieg, N., and Holt, J. (1984) Bergey's manual of systematic bacteriology. The Willian and Wilkins Co., Baltimore.

Palleroni, N.J. (2010) The *Pseudomonas* story. *Environ Microbiol* **12**: 1377-1383.

Palmela, C., Chevarin, C., Xu, Z., Torres, J., Sevrin, G., Hirten, R., Barnich, N., Ng, S.C., and Colombel, J.F. (2018) Adherent-invasive *Escherichia coli* in inflammatory bowel disease. *Gut* **67**: 574-587.

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

Pasek, S., Risler, J.-L., and Brézellec, P. (2006) Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* **22**: 1418-1423.

Pavesi, A., Vianelli, A., Chirico, N., Bao, Y., Blinkova, O., Belshaw, R., Firth, A., and Karlin, D. (2018) Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLOS ONE* **13**: e0202513-e0202513.

Pedersen, K., Christensen, S.K., and Gerdes, K. (2002) Rapid induction and reversal of a bacteriostatic condition by controlled expression of toxins and antitoxins. *Mol Microbiol* **45**: 501-510.

Pedersen, K., Zavialov, A.V., Pavlov, M.Y., Elf, J., Gerdes, K., and Ehrenberg, M. (2003) The Bacterial Toxin RelE Displays Codon-Specific Cleavage of mRNAs in the Ribosomal A Site. *Cell* **112**: 131-140.

Peña, C., Suarez, C., Tubau, F., Dominguez, A., Sora, M., Pujol, M., Gudiol, F., and Ariza, J. (2009) Carbapenem-resistant *Pseudomonas aeruginosa*: factors influencing multidrug-resistant acquisition in non-critically ill patients. *Eur J Clin Microbiol Infect Dis* **28**: 519-522.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**: 3551-3567.

Perkins, T.T., Kingsley, R.A., Fookes, M.C., Gardner, P.P., James, K.D., Yu, L., Assefa, S.A., He, M., Croucher, N.J., Pickard, D.J., Maskell, D.J., Parkhill, J., Choudhary, J., Thomson, N.R., and Dougan, G. (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi. PLoS Genet* **5**: e1000569.

Peters, J.M., Mooney, R.A., Kuan, P.F., Rowland, J.L., Keles, S., and Landick, R. (2009) Rho directs widespread termination of intragenic and stable RNA transcription. *Proc Natl Acad Sci U S A* **106**: 15406-15411.

Peterson, A.C., Russell, J.D., Bailey, D.J., Westphall, M.S., and Coon, J.J. (2012) Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Mol Cell Proteomics* **11**: 1475.

Petkau, A., Stuart-Edwards, M., Stothard, P., and Van Domselaar, G. (2010) Interactive microbial genome visualization with GView. *Bioinformatics* **26**: 3125-3126.

Petruschke, H., Anders, J., Stadler, P.F., Jehmlich, N., and von Bergen, M. (2020) Enrichment and identification of small proteins in a simplified human gut microbiome. *J Proteomics* **213**: 103604.

Pitt, J.J. (2009) Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *Clin Biochem Rev* **30**: 19-34.

Plotnikova, J.M., Rahme, L.G., and Ausubel, F.M. (2000) Pathogenesis of the human opportunistic pathogen *Pseudomonas aeruginosa* PA14 in *Arabidopsis. Plant Physiol* **124**: 1766-1774.

Poole, K. (2008) Bacterial Multidrug Efflux Pumps Serve Other Functions. *Microbe* **3**: 179-185.

Poole, K. (2011) *Pseudomonas Aeruginosa:* Resistance to the Max. *Front Microbiol* **2**.

Poole, K., Krebes, K., McNally, C., and Neshat, S. (1993) Multiple antibiotic resistance in *Pseudomonas aeruginosa*: evidence for involvement of an efflux operon. *J Bacteriol* **175**: 7363-7372.

Potvin, E., Sanschagrin, F., and Levesque, R.C. (2008) Sigma factors in *Pseudomonas aeruginosa. FEMS Microbiol Rev* **32**: 38-55.

Prabh, N., and Rödelsperger, C. (2016) Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? *BMC Bioinformatics* **17**: 226.

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

Pukatzki, S., Kessin, R.H., and Mekalanos, J.J. (2002) The Human Pathogen *Pseudomonas aeruginos*a Utilizes Conserved Virulence Pathways to Infect the Social Amoeba *Dictyostelium discoideum. Proc Natl Acad Sci U S A* **99**: 3159-3164.

Quail, M.A., Otto, T.D., Gu, Y., Harris, S.R., Skelly, T.F., McQuillan, J.A., Swerdlow, H.P., and Oyola, S.O. (2012) Optimal enzymes for amplifying sequencing libraries. *Nat Methods* **9**: 10-11.

Quinlan, A.R., and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Raghavan, R., Sloan, D.B., and Ochman, H. (2012) Antisense transcription is pervasive but rarely conserved in enteric bacteria. *mBio* **3**.

Rahme, L.G., Ausubel, F.M., Cao, H., Drenkard, E., Goumnerov, B.C., Lau, G.W., Mahajan-Miklos, S., Plotnikova, J., Tan, M.W., Tsongalis, J., Walendziewicz, C.L., and Tompkins, R.G. (2000) Plants and animals share functionally common bacterial virulence factors. *Proc Natl Acad Sci U S A* **97**: 8815-8821.

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Grüning, B.A., Villaveces, J., Habermann, B., Akhtar, A., and Manke, T. (2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun* **9**: 189.

Ramos, J.-L., (2004) *Pseudomonas* Volume 1: Genomics, life style and molecular architecture. Springer.

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

Rankin, D.J., Rocha, E.P.C., and Brown, S.P. (2011) What traits are carried on mobile genetic elements, and why? *Heredity* **106**: 1-10.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* **2**: 1896-1906.

Rasko, D.A., Rosovitz, M.J., Myers, G.S.A., Mongodin, E.F., Fricke, W.F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N.R., Chaudhuri, R., Henderson, I.R., Sperandio, V., and Ravel, J. (2008) The Pangenome Structure of *Escherichia coli:* Comparative Genomic Analysis of *E. coli* Commensal and Pathogenic Isolates. *J Bacteriol* **190**: 6881.

Rath, A., Glibowicka, M., Nadeau, V.G., Chen, G., and Deber, C.M. (2009) Detergent binding explains anomalous SDS-PAGE migration of membrane proteins. *Proc Natl Acad Sci U S A* **106**: 1760-1765.

Rauniyar, N. (2015) Parallel Reaction Monitoring: A Targeted Experiment Performed Using High Resolution and High Mass Accuracy Mass Spectrometry. *Int J Mol Sci* **16**: 28566-28581.

Reisch, C.R., and Prather, K.L. (2015) The no-SCAR (Scarless Cas9 Assisted Recombineering) system for genome editing in *Escherichia coli. Sci Rep* **5**: 15096.

Ren, G.-X., Guo, X.-P., and Sun, Y.-C. (2017) Regulatory 3' Untranslated Regions of Bacterial mRNAs. *Front Microbiol* **8**: 1276-1276.

Reynolds, J.A., and Tanford, C. (1970) Binding of dodecyl sulfate to proteins at high binding ratios. Possible implications for the state of proteins in biological membranes. *Proc Natl Acad Sci U S A* **66**: 1002-1007.

Richardson, E.J., and Watson, M. (2013) The automatic annotation of bacterial genomes. *Brief Bioinform* **14**: 1-12.

Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G.D., and Gold, L. (1992) Translation initiation in *Escherichia coli:* sequences within the ribosome-binding site. *Mol Microbiol* **6**: 1219-1229.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.

Rodnina, M. (2018) Translation in Prokaryotes. *Cold Spring Harb Perspect Biol* **10**: a032664.

Rogozin, I.B., Makarova, K.S., Natale, D.A., Spiridonov, A.N., Tatusov, R.L., Wolf, Y.I., Yin, J., and Koonin, E.V. (2002a) Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res* **30**: 4264-4271.

Rogozin, I.B., Spiridonov, A.N., Sorokin, A.V., Wolf, Y.I., Jordan, I.K., Tatusov, R.L., and Koonin, E.V. (2002b) Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet* **18**: 228-232.

Rolhion, N., Barnich, N., Bringer, M.-A., Glasser, A.-L., Ranc, J., Hébuterne, X., Hofman, P., and Darfeuille-Michaud, A. (2010) Abnormally expressed ER stress response chaperone Gp96 in CD favours adherent-invasive *Escherichia coli* invasion. *Gut* **59**: 1355.

Roy-Burman, A., Savel, R.H., Racine, S., Swanson, B.L., Revadigar, N.S., Fujimoto, J., Sawa, T., Frank, D.W., and Wiener-Kronish, J.P. (2001) Type III protein secretion is associated with death in lower respiratory and systemic *Pseudomonas aeruginosa* infections. *J Infect Dis* **183**: 1767-1774.

Roymondal, U., Das, S., and Sahoo, S. (2009) Predicting Gene Expression Level from Relative Codon Usage Bias: An Application to *Escherichia coli* Genome. *DNA Res* **16**: 13-30.

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

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

Saito, K., Green, R., and Buskirk, A.R. (2020) Translational initiation in *E. coli* occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. *Elife* **9**: e55002.

Saito, K., Yoneyama, H., and Nakae, T. (1999) *nalB*-type mutations causing the overexpression of the MexAB-OprM efflux pump are located in the *mexR* gene of the *Pseudomonas aeruginosa* chromosome. *FEMS Microbiol Lett* **179**: 67-72.

Salzberg, S.L., Delcher, A.L., Kasif, S., and White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **26**: 544-548.

Sanna, C.R., Li, W.H., and Zhang, L. (2008) Overlapping genes in the human and mouse genomes. *BMC Genomics* **9**: 169.

Sarker, M.R., and Cornelis, G.R. (1997) An improved version of suicide vector pKNG101 for gene replacement in Gram-negative bacteria. *Mol Microbiol* **23**: 410-411.

Savageau, M.A. (1983) *Escherichia coli* Habitats, Cell Types, and Molecular Mechanisms of Gene Control. *Am Nat* **122**: 732-744.

Schaechter, M., (2009) *Escherichia Coli.* In: The Desk Encyclopedia of Microbiology. Academic Press, Kidlington.

Schägger, H. (2006) Tricine–SDS-PAGE. *Nat Protoc* **1**: 16-22.

Scherbakov, D.V., and Garber, M.B. (2000) Overlapping genes in bacterial and phage genomes. *Mol Biol* **34**: 485-495.

Scherer, S., Neuhaus, K., Bossert, M., Mir, K., Keim, D., and Simon, S., (2018) Finding New Overlapping Genes and Their Theory (FOG Theory). In: Information- and Communication Theory in Molecular Biology. Springer International Publishing, Cham.

Schlub, T.E., Buchmann, J.P., and Holmes, E.C. (2018) A Simple Method to Detect Candidate Overlapping Genes in Viruses Using Single Genome Sequences. *Mol Biol Evol* **35**: 2572-2581.

Schlub, T.E., and Holmes, E.C. (2020) Properties and abundance of overlapping genes in viruses. *Virus Evol* **6**: veaa009.

Schlünzen, F., Pyetan, E., Fucini, P., Yonath, A., and Harms, J.M. (2004) Inhibition of peptide bond formation by pleuromutilins: the structure of the 50S ribosomal subunit from *Deinococcus radiodurans* in complex with tiamulin. *Mol Microbiol* **54**: 1287-1294.

Schmid, K.J., and Aquadro, C.F. (2001) The evolutionary analysis of "orphans" from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes. *Genetics* **159**: 589-598.

Schneider-Poetsch, T., Ju, J., Eyler, D.E., Dang, Y., Bhat, S., Merrick, W.C., Green, R., Shen, B., and Liu, J.O. (2010) Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nat Chem Biol* **6**: 209-217.

Schrimpe-Rutledge, A.C., Jones, M.B., Chauhan, S., Purvine, S.O., Sanford, J.A., Monroe, M.E., Brewer, H.M., Payne, S.H., Ansong, C., Frank, B.C., Smith, R.D., Peterson, S.N., Motin, V.L., and Adkins, J.N. (2012) Comparative Omics-Driven Genome Annotation Refinement: Application across *Yersiniae*. *PLOS ONE* **7**: e33903.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., and Ragg, T. (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol* **7**: 3.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* **473**: 337-342.

Sealfon, R.S., Lin, M.F., Jungreis, I., Wolf, M.Y., Kellis, M., and Sabeti, P.C. (2015) FRESCo: finding regions of excess synonymous constraint in diverse viruses. *Genome Biol* **16**: 38-38.

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., and Plotkin, J.B. (2013) Rate-limiting steps in yeast protein translation. *Cell* **153**: 1589-1601.

Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiß, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P.F., and Vogel, J. (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**: 250-255.

Sharma, P., Nilges, B.S., Wu, J., and Leidel, S.A. (2019) The translation inhibitor cycloheximide affects ribosome profiling data in a species-specific manner. *bioRxiv*: 746255.

Shi, T., Song, E., Nie, S., Rodland, K.D., Liu, T., Qian, W.-J., and Smith, R.D. (2016) Advances in targeted proteomics and applications to biomedical research. *Proteomics* **16**: 2160-2182.

Shi, T., Su, D., Liu, T., Tang, K., Camp, D.G., 2nd, Qian, W.-J., and Smith, R.D. (2012) Advancing the sensitivity of selected reaction monitoring-based targeted quantitative proteomics. *Proteomics* **12**: 1074-1092.

Shine, J., and Dalgarno, L. (1974) The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A* **71**: 1342-1346.

Shiroguchi, K., Jia, T.Z., Sims, P.A., and Xie, X.S. (2012) Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci U S A* **109**: 1347.

Shultzaberger, R.K., Chen, Z., Lewis, K.A., and Schneider, T.D. (2007) Anatomy of *Escherichia coli* sigma70 promoters. *Nucleic Acids Res* **35**: 771-788.

Silby, M.W., and Levy, S.B. (2004) Use of *in vivo* expression technology to identify genes important in growth and survival of *Pseudomonas fluorescens* Pf0-1 in soil: discovery of expressed sequences with novel genetic organization. *J Bacteriol* **186**: 7411-7419.

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

Silby, M.W., Nicoll, J.S., and Levy, S.B. (2009) Requirement of polyphosphate by *Pseudomonas fluorescens* Pf0-1 for competitive fitness and heat tolerance in laboratory media and sterile soil. *Appl Environ Microbiol* **75**: 3872-3881.

Silva, J.C., Gorenstein, M.V., Li, G.Z., Vissers, J.P., and Geromanos, S.J. (2006) Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* **5**: 144-156.

Simon, R., Priefer, U., and Pühler, A. (1983) A Broad Host Range Mobilization System for *In Vivo* Genetic Engineering: Transposon Mutagenesis in Gram Negative Bacteria. *Nat Biotechnol* **1**: 784-791.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* **15**: 121-132.

Skorski, P., Leroy, P., Fayet, O., Dreyfus, M., and Hermann-Le Denmat, S. (2006) The highly efficient translation initiation region from the *Escherichia coli rpsA* gene lacks a Shine-Dalgarno element. *J Bacteriol* **188**: 6277-6285.

Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., and Saghatelian, A. (2013) Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nat Chem Biol* **9**: 59-64.

Slonczewski, J., and Foster, J.W., (2009) *Microbiology: an evolving science*. W.W. Norton & Co, New York.

Smith, C., Canestrari, J.G., Wang, J., Derbyshire, K.M., Gray, T.A., and Wade, J.T. (2019) Pervasive Translation in *Mycobacterium tuberculosis*. *bioRxiv*: 665208.

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

Spielman, S.J., and Wilke, C.O. (2015) Pyvolve: A Flexible Python Module for Simulating Sequences along Phylogenies. *PLOS ONE* **10**: e0139047.

Srikumar, R., Paul, C.J., and Poole, K. (2000) Influence of Mutations in the *mexR* Repressor Gene on Expression of the MexA-MexB-OprM Multidrug Efflux System of *Pseudomonas aeruginosa*. *J Bacteriol* **182**: 1410.

Stern-Ginossar, N., and Ingolia, N.T. (2015) Ribosome profiling as a tool to decipher viral complexity. *Annu Rev Virol* **2**: 335-349.

Stickland, H.G., Davenport, P.W., Lilley, K.S., Griffin, J.L., and Welch, M. (2010) Mutation of *nfxB* causes global changes in the physiology and metabolism of *Pseudomonas aeruginosa*. *J Proteome Res* **9**: 2957-2967.

Storz, G., Vogel, J., and Wassarman, K.M. (2011) Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell* **43**: 880-891.

Storz, G., Wolf, Y.I., and Ramamurthi, K.S. (2014) Small proteins can no longer be ignored. *Annu Rev Biochem* **83**: 753-777.

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K., Wu, Z., Paulsen, I.T., Reizer, J., Saier, M.H., Hancock, R.E., Lory, S., and Olson, M.V. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* **406**: 959-964.

Studier, F.W. (1991) Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. *J Mol Biol* **219**: 37-44.

Suyama, M., Torrents, D., and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**: W609-W612.

Szoboszlay, S., Atzel, B., and Kriszt, B. (2003) Comparative biodegradation examination of *Pseudomonas aeruginosa* (ATCC 27853) and other oil degraders on hydrocarbon contaminated soil. *Commun Agric Appl Biol Sci* **68**: 207-210.

Takase, H., Nitanai, H., Hoshino, K., and Otani, T. (2000) Impact of siderophore production on *Pseudomonas aeruginosa* infections in immunosuppressed mice. *Infect Immun* **68**: 1834-1839.

Takatsuka, Y., and Nikaido, H. (2009) Covalently linked trimer of the AcrB multidrug efflux pump provides support for the functional rotating mechanism. *J Bacteriol* **191**: 1729-1737.

Tamboli, C.P., Neut, C., Desreumaux, P., and Colombel, J.F. (2004) Dysbiosis in inflammatory bowel disease. *Gut* **53**: 1-4.

Tang, H., Kays, M., and Prince, A. (1995) Role of *Pseudomonas aeruginosa* pili in acute pulmonary infection. *Infect Immun* **63**: 1278-1285.

Tang, H.Y., Ali-Khan, N., Echan, L.A., Levenkova, N., Rux, J.J., and Speicher, D.W. (2005) A novel four-dimensional strategy combining protein and peptide separation methods enables detection of low-abundance proteins in human plasma and serum proteomes. *Proteomics* **5**: 3329-3342.

Tautz, D. (2014) The discovery of *de novo* gene evolution. *Perspect Biol Med* **57**: 149-161.

Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E. (2010) The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol* **8**: 207-217.

Thomason, M.K., Bischler, T., Eisenbart, S.K., Förstner, K.U., Zhang, A., Herbig, A., Nieselt, K., Sharma, C.M., and Storz, G. (2015) Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in *Escherichia coli*. *J Bacteriol* **197**: 18.

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

Tian, M., Lian, Z., Bao, Y., Bao, S., Yin, Y., Li, P., Ding, C., Wang, S., Li, T., Qi, J., Wang, X., and Yu, S. (2019) Identification of a novel, small, conserved hypothetical protein involved in *Brucella abortus* virulence by modifying the expression of multiple genes. *Transbound Emerg Dis* **66**: 349-362.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., Karoui, M.E., Frapy, E., Garry, L., Ghigo, J.M., Gilles, A.M., Johnson, J., Le Bouguénec, C., Lescat, M., Mangenot, S., Martinez-Jéhanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M.A., Pichon, C., Rouy, Z., Ruf, C.S., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Médigue, C., Rocha, E.P.C., and Denamur, E. (2009) Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths. *PLoS Genet* **5**: e1000344.

Tsiatsiani, L., and Heck, A.J. (2015) Proteomics beyond trypsin. *FEBS J* **282**: 2612-2626.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010) An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* **141**: 344-354.

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

Tyanova, S., Temu, T., and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* **11**: 2301-2319.

Vakirlis, N., Carvunis, A.R., and McLysaght, A. (2020) Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *Elife* **9**.

Valles, J., Mariscal, D., Cortes, P., Coll, P., Villagra, A., Diaz, E., Artigas, A., and Rello, J. (2004) Patterns of colonization by *Pseudomonas aeruginosa* in intubated patients: a 3-year prospective study of 1,607 isolates using pulsed-field gel electrophoresis with implications for prevention of ventilator-associated pneumonia. *Intensive Care Med* **30**: 1768-1775.

Valot, B., Guyeux, C., Rolland, J.Y., Mazouzi, K., Bertrand, X., and Hocquet, D. (2015) What It Takes to Be a *Pseudomonas aeruginosa*? The Core Genome of the Opportunistic Pathogen Updated. *PLOS ONE* **10**: e0126468-e0126468.

Van Oss, S.B., and Carvunis, A.-R. (2019) *De novo* gene birth. *PLoS Genet* **15**: e1008160.

Vander Wauven, C., Piérard, A., Kley-Raymann, M., and Haas, D. (1984) *Pseudomonas aeruginosa* mutants affected in anaerobic growth on arginine: evidence for a four-gene cluster encoding the arginine deiminase pathway. *J Bacteriol* **160**: 928-934.

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. *Sci Rep* **8**: 17875-17875.

Veeramachaneni, V., Makałowski, W., Galdzicki, M., Sood, R., and Makałowska, I. (2004) Mammalian overlapping genes: the comparative perspective. *Genome Res* **14**: 280-286.

Venable, J.D., Dong, M.-Q., Wohlschlegel, J., Dillin, A., and Yates, J.R. (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* **1**: 39-45.

Venter, E., Smith, R.D., and Payne, S.H. (2011) Proteogenomic Analysis of Bacteria and Archaea: A 46 Organism Case Study. *PLOS ONE* **6**: e27587.

Villegas, A., and Kropinski, A.M. (2008) An analysis of initiation codon utilization in the Domain *Bacteria* – concerns about the quality of bacterial genome annotation. *Microbiology* **154**: 2559-2661.

Vimberg, V., Tats, A., Remm, M., and Tenson, T. (2007) Translation initiation region sequence preferences in *Escherichia coli*. *BMC Mol Biol* **8**: 100-100.

Vincent, J.-L., Rello, J., Marshall, J., Silva, E., Anzueto, A., Martin, C.D., Moreno, R., Lipman, J., Gomersall, C., Sakr, Y., Reinhart, K., and EPIC II Group of Investigators, f.t. (2009) International Study of the Prevalence and Outcomes of Infection in Intensive Care Units. *JAMA* **302**: 2323-2329.

Wade, J.T., and Grainger, D.C. (2014) Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat Rev Microbiol* **12**: 647-653.

Wang, L.-F., Park, S.-S., and Doi, R.H. (1999a) A Novel Bacillus subtilis Gene, *antE*, Temporally Regulated and Convergent to and Overlapping *dnaE*. *J Bacteriol* **181**: 353.

Wang, N., Yamanaka, K., and Inouye, M. (1999b) CspI, the ninth member of the CspA family of *Escherichia coli*, is induced upon cold shock. *J Bacteriol* **181**: 1603-1609.

Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57-63.

Warren, A.S., Archuleta, J., Feng, W.-C., and Setubal, J.C. (2010) Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics* **11**: 131.

Watson, J.D., and Crick, F.H.C. (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**: 737-738.

Weaver, J., Mohammad, F., Buskirk, A.R., and Storz, G. (2019) Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes. *mBio* **10**: e02819-02818.

Weinstein, R.A., Gaynes, R., Edwards, J.R., and National Nosocomial Infections Surveillance, S. (2005) Overview of Nosocomial Infections Caused by Gram-Negative Bacilli. *Clin Infect Dis* **41**: 848-854.

Weisman, C.M., Murray, A.W., and Eddy, S.R. (2020) Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLOS Biol* **18**: e3000862.

Wells, J.M., and McLuckey, S.A. (2005) Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol* **402**: 148-185.

Wen, J.D., Lancaster, L., Hodges, C., Zeri, A.C., Yoshimura, S.H., Noller, H.F., Bustamante, C., and Tinoco, I. (2008) Following translation by single ribosomes one codon at a time. *Nature* **452**: 598-603.

West, S.E., and Iglewski, B.H. (1988) Codon usage in *Pseudomonas aeruginosa. Nucleic Acids Res* **16**: 9323-9335.

Westermann, A.J., Gorski, S.A., and Vogel, J. (2012) Dual RNA-seq of pathogen and host. *Nat Rev Microbiol* **10**: 618-630.

Wheater, D.W.F., Mara, D.D., Jawad, L., and Oragui, J. (1980) *Pseudomonas aeruginosa* and *Escherichia coli* in sewage and fresh water. *Water Res* **14**: 713-721.

Wichmann, S., and Ardern, Z. (2019) Optimality in the standard genetic code is robust with respect to comparison code sets. *Biosystems* **185**: 104023.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bähler, J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**: 1239-1243.

Willems, P., Fijalkowski, I., and Van Damme, P. (2020) Lost and Found: Re-searching and Re-scoring Proteomics Data Aids Genome Annotation and Improves Proteome Coverage. *mSystems* **5**: e00833-00820.

Williams, J.J., Halvorsen, E.M., Dwyer, E.M., DiFazio, R.M., and Hergenrother, P.J. (2011) Toxin–antitoxin (TA) systems are prevalent and transcribed in clinical isolates of *Pseudomonas aeruginosa* and methicillin-resistant *Staphylococcus aureus. FEMS Microbiol Lett* **322**: 41-50.

Willis, S., and Masel, J. (2018) Gene Birth Contributes to Structural Disorder Encoded by Overlapping Genes. *Genetics* **210**: 303-313.

Wilson, B.A., and Masel, J. (2011) Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol Evol* **3**: 1245-1252.

Wilson, G.A., Bertrand, N., Patel, Y., Hughes, J.B., Feil, E.J., and Field, D. (2005) Orphans as taxonomically restricted and ecologically important genes. *Microbiology* **151**: 2499-2501.

Wolfgang, M.C., Kulasekara, B.R., Liang, X., Boyd, D., Wu, K., Yang, Q., Miyada, C.G., and Lory, S. (2003) Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa. Proc Natl Acad Sci U S A* **100**: 8484-8489.

Wong, K.K., Brinkman, F.S., Benz, R.S., and Hancock, R.E. (2001) Evaluation of a structural model of *Pseudomonas aeruginosa* outer membrane protein OprM, an efflux component involved in intrinsic antibiotic resistance. *J Bacteriol* **183**: 367-374.

Wood, D.E., Lin, H., Levy-Moonshine, A., Swaminathan, R., Chang, Y.-C., Anton, B.P., Osmani, L., Steffen, M., Kasif, S., and Salzberg, S.L. (2012) Thousands of missed genes found in bacterial genomes and their analysis with COMBREX. *Biol Direct* **7**: 37-37.

Woolstenhulme, C.J., Guydosh, N.R., Green, R., and Buskirk, A.R. (2015) High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Rep* **11**: 13-21.

Wright, B.W., Kamath, K.S., Krisp, C., and Molloy, M.P. (2019a) Proteome profiling of *Pseudomonas aeruginosa* PAO1 identifies novel responders to copper stress. *BMC Microbiol* **19**: 69.

Wright, C., Rajpurohit, A., Burke, E.E., Williams, C., Collado-Torres, L., Kimos, M., Brandon, N.J., Cross, A.J., Jaffe, A.E., Weinberger, D.R., and Shin, J.H. (2019b) Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods. *BMC Genomics* **20**: 513.

Wu, P., Zhang, H., Lin, W., Hao, Y., Ren, L., Zhang, C., Li, N., Wei, H., Jiang, Y., and He, F. (2014) Discovery of Novel Genes and Gene Isoforms by Integrating Transcriptomic and Proteomic Profiling from Mouse Liver. *J Proteome Res* **13**: 2409-2419.

Wurtzel, O., Yoder-Himes, D.R., Han, K., Dandekar, A.A., Edelheit, S., Greenberg, E.P., Sorek, R., and Lory, S. (2012) The single-nucleotide resolution transcriptome of *Pseudomonas aeruginosa* grown in body temperature. *PLoS Pathog* **8**: e1002945-e1002945.

Xiao, Z., Huang, R., Xing, X., Chen, Y., Deng, H., and Yang, X. (2018) *De novo* annotation and characterization of the translatome with ribosome profiling data. *Nucleic Acids Res* **46**: e61.

Yang, X., Jensen, S.I., Wulff, T., Harrison, S.J., and Long, K.S. (2016) Identification and validation of novel small proteins in *Pseudomonas putida. Environ Microbiol Rep* **8**: 966-974.

Yang, Z., Zeng, X., and Tsui, S.K.-W. (2019) Investigating function roles of hypothetical proteins encoded by the *Mycobacterium tuberculosis* H37Rv genome. *BMC Genomics* **20**: 394.

Yoneyama, H., Ocaktan, A., Tsuda, M., and Nakae, T. (1997) The role of *mex*-gene products in antibiotic extrusion in *Pseudomonas aeruginosa. Biochem Biophys Res Commun* **233**: 611-618.

Yoshimura, F., and Nikaido, H. (1982) Permeability of *Pseudomonas aeruginosa* outer membrane to hydrophilic solutes. *J Bacteriol* **152**: 636-642.

Young, C.L., Britton, Z.T., and Robinson, A.S. (2012) Recombinant protein expression and purification: A comprehensive review of affinity tags and microbial applications. *Biotechnol J* **7**: 620-634.

Yu, J.-F., Xiao, K., Jiang, D.-K., Guo, J., Wang, J.-H., and Sun, X. (2011) An integrative method for identifying the over-annotated protein-coding genes in microbial genomes. *DNA Res* **18**: 435-449.

Zehentner, B., (2020) Experimental characterization of overlapping genes in enterohemorrhagic *E. coli*: Overexpression phenotypes and high-throughput NGS analysis of transcription start sites. Doctoral Thesis. Technische Universität München.

Zehentner, B., Ardern, Z., Kreitmeier, M., Scherer, S., and Neuhaus, K. (2020a) Evidence for Numerous Embedded Antisense Overlapping Genes in Diverse *E. coli* Strains. *bioRxiv*: 2020.2011.2018.388249.

Zehentner, B., Ardern, Z., Kreitmeier, M., Scherer, S., and Neuhaus, K. (2020b) A Novel pH-Regulated, Unusual 603 bp Overlapping Protein Coding Gene *pop* Is Encoded Antisense to *ompA* in *Escherichia coli* O157:H7 (EHEC). *Front Microbiol* **11**.

Zgurskaya, H.I., and Nikaido, H. (1999) Bypassing the periplasm: reconstitution of the AcrAB multidrug efflux pump of *Escherichia coli. Proc Natl Acad Sci U S A* **96**: 7190-7195.

Zhang, H., Liu, P., Guo, T., Zhao, H., Bensaddek, D., Aebersold, R., and Xiong, L. (2019) Arabidopsis proteome and the mass spectral assay library. *Sci Data* **6**: 278.

Zhang, J. (2003) Evolution by gene duplication: an update. *Trends Ecol Evol* **18**: 292-298.

Zhang, M.Q. (2002) Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet* **3**: 698-709.

Zhelyazkova, P., Sharma, C.M., Förstner, K.U., Liere, K., Vogel, J., and Börner, T. (2012) The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. *Plant Cell* **24**: 123-136.

Zheng, X., Hu, G.-Q., She, Z.-S., and Zhu, H. (2011) Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genomics* **12**: 361.

Zhuang, F., Fuchs, R.T., Sun, Z., Zheng, Y., and Robb, G.B. (2012) Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res* **40**: e54.

Ziha-Zarifi, I., Llanes, C., Köhler, T., Pechere, J.C., and Plesiat, P. (1999) *In vivo* emergence of multidrug-resistant mutants of *Pseudomonas aeruginosa* overexpressing the active efflux system MexA-MexB-OprM. *Antimicrob Agents Chemother* **43**: 287-291.

Zolg, D.P., Wilhelm, M., Yu, P., Knaute, T., Zerweck, J., Wenschuh, H., Reimer, U., Schnatbaum, K., and Kuster, B. (2017) PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration. *Proteomics* **17**: 1700263.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406-3415.

Zurawski, G., and Zurawski, S.M. (1985) Structure of the *Escherichia coli* S10 ribosomal protein operon. *Nucleic Acids Res* **13**: 4521-4526.

6. Supplement

6.1 Supplementary Files

Supplementary Files S1 – S5 can be found on the attached CD-ROM.

Supplementary File S1. Details about all OLG and iORF candidates identified in *E. coli* LF82 in this study. Listed are all relevant information about the novel gene candidates including their genomic position, ORF characteristics (length, start codon, type of overlap, etc.), details about their mother genes, results of blastp, differential expression and structural gene-like element (promoter, terminator, RBS) analyses, their expression metrics (RCV, RPKM and coverage values) of all RNA-seq and Ribo-seq experiment, the individual probabilities obtained by the three prediction tools (REPARATION, DeepRibo, scripts by Giess) as well as the overall prediction score as determined in this study.

Supplementary File S2. Details about all OLG and iORF candidates identified in *P. aeruginosa* PAO1 in this study. Listed are all relevant information about the novel gene candidates including their genomic position, ORF characteristics (length, start codon, type of overlap, etc.), details about their mother genes, results of blastp, differential expression, structural gene-like element (terminator, RBS), Cappable-seq and mass spectrometry (MS) analyses, their expression metrics (RCV, RPKM and coverage values) of all RNA-seq and Ribo-seq experiment, the individual probabilities obtained by the three prediction tools (REPARATION, DeepRibo, scripts by Giess) as well as the overall prediction score as determined in this study.

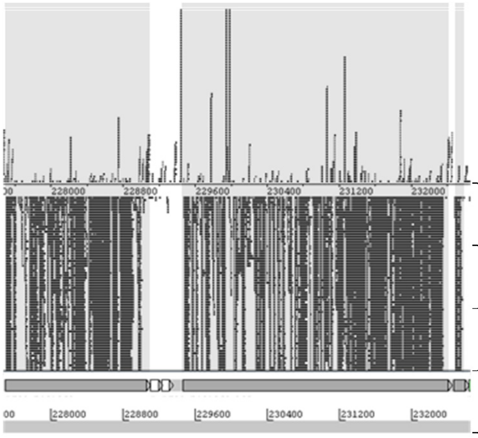
Supplementary File S3. Results of gene prediction in the genome of *P. aeruginosa* PAO1 after application of Prodigal (Hyatt *et al.*, 2010). Detailed values of the individual categories including GC content (gc_cont), confidence score (conf), overall score (score), hexamer coding proportion score (cscore), TIS score (sscore), ribosome binding site score (rscore), region score flanking the start codon (uscore) and the start codon sequence score (tscore) are listed for all predicted genes (n = 5,681) as well as for *olg1* and *olg2* after hiding of their mother genes *tle3* and PA1383.

Supplementary File S4. Expression metrics (RCV, RPKM, coverage values) obtained for *olg1* and *olg2* as well as their mother genes *tle3* and PA1383 obtained for RNA-seq and Ribo-seq data published by Grady *et al.* (2017) after cultivation of *P. aeruginosa* PAO1 and *P. aeruginosa* ATCC 33988 in M9 broth supplemented with either glycerol or *n*-alkanes as sole carbon source.

Supplementary File S5. Details about all peptides of Olg1, OLG2, Tle3 and PA1383 detected in the DDA-MS experiment. For each of the peptides, the sequence, the number of missed cleavage sites, the charge, the MaxQuant score (Cox & Mann, 2008), the intensity as well as the dotp values are listed.

6.2 Supplementary Figures

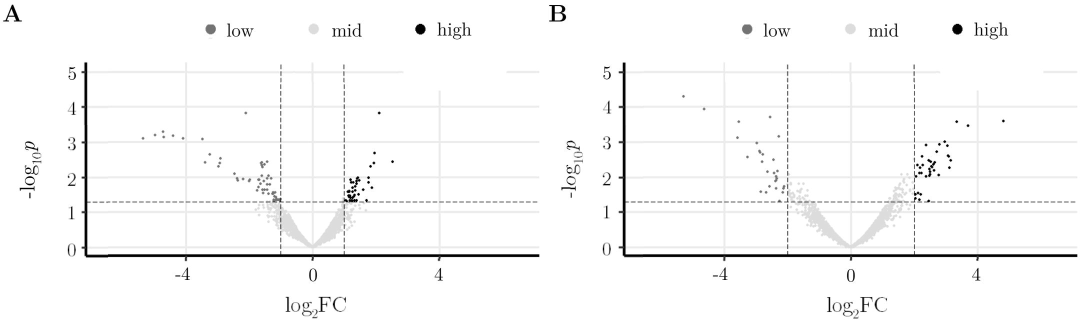
A



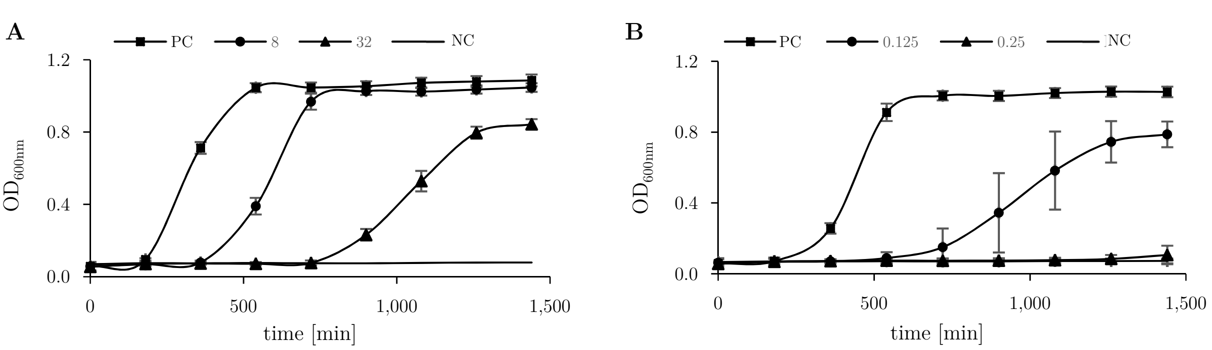
B

		Ribo-seq aerobic	Ribo-seq anaerobic	RNA-seq aerobic	RNA-seq anaerobic
Ø 23S *rRNA*	+	11.3	62.7	31.8	12.6
[RPKM]	-	5,043.3	3,333.5	332.9	658.1
Ø 16S *rRNA*	+	4.3	12.0	5.2	2.1
[RPKM]	-	7,476.7	4,776.6	423.3	689.8
Ø 5S *rRNA*	+	3.1	13.5	6.9	4.3
[RPKM]	-	4,319.0	3,835.8	835.6	764.5

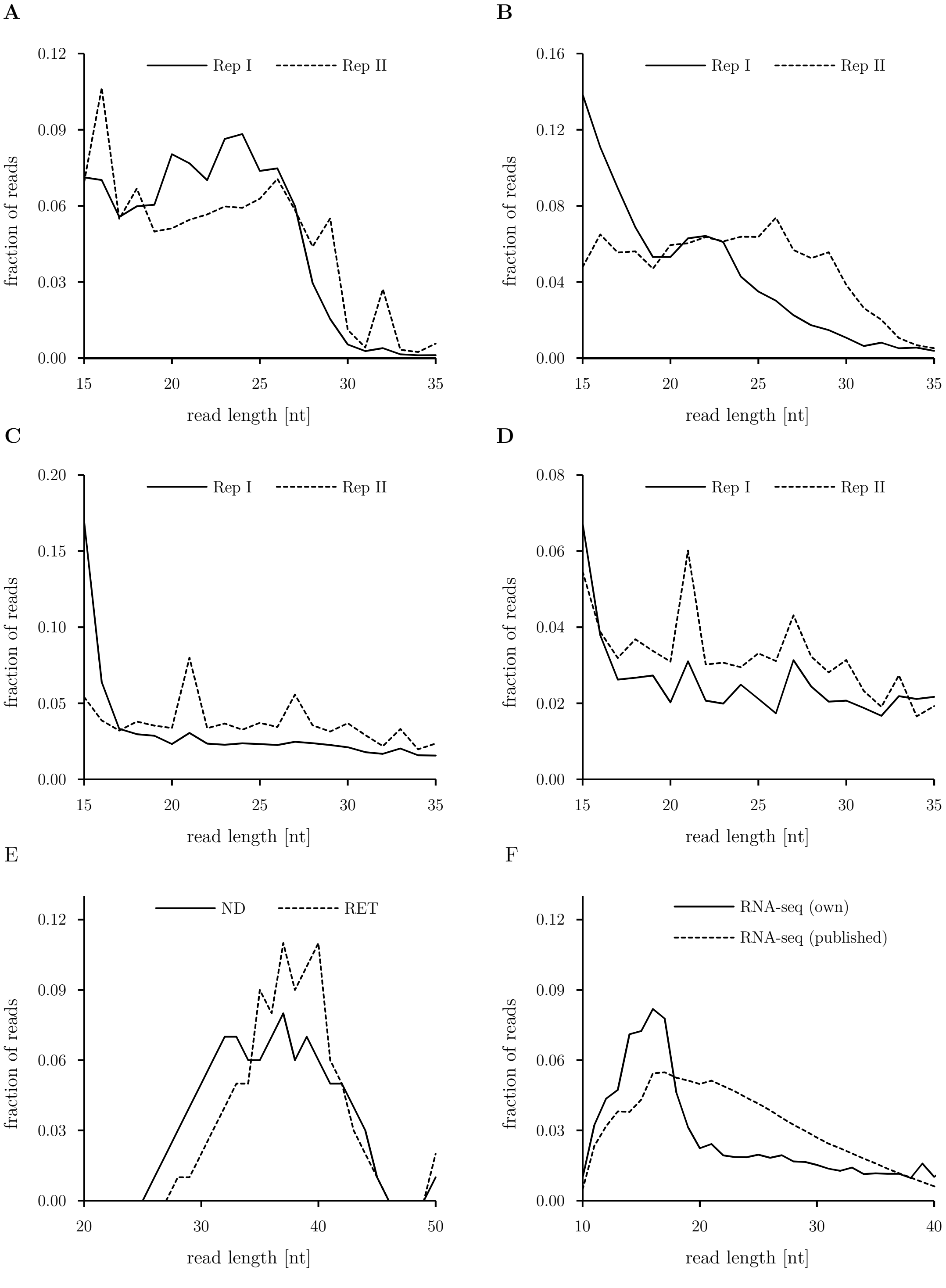
Supplementary Figure 1. Reads mapping antisense to *rRNA* genes indicate probe carryover during rRNA depletion. (**A**) Ribo-seq reads of Exp1_anaerobic_RepI visualized in the genome browser Artemis. Grey regions indicate the 16S, 23S and 5S *rRNA* genes of the *E. coli* LF82 genome with reads mapping primarily to the opposite strand (-). (**B**) Averaged RPKM values obtained for *rRNA* genes were higher for the antisense region (-) indicating probe contamination.



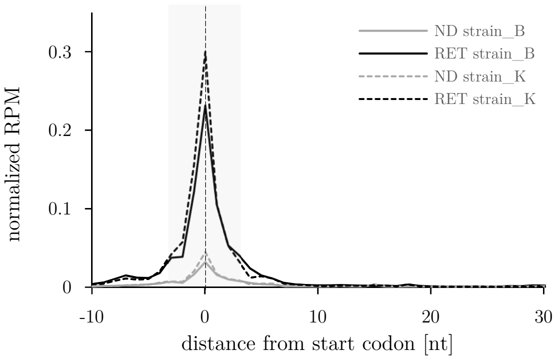
Supplementary Figure 2. Results of differential expression analysis of all anORFs and novel gene candidates using edgeR. Shown are the obatined $\log_2$ fold changes ($\log_2$FC, x-axis) and p-values (y-axis) of all genes with significant (**A**) RNA-seq and (**B**) Ribo-seq signals. At the trancriptome level, 55 anORFs were down- and 38 anORFs were upregulated in the anaerobic condition compared to the aerobic condition. In contrast, 63 and 80 anORFs exhibited down- or upregulation at the translational level, respectively.



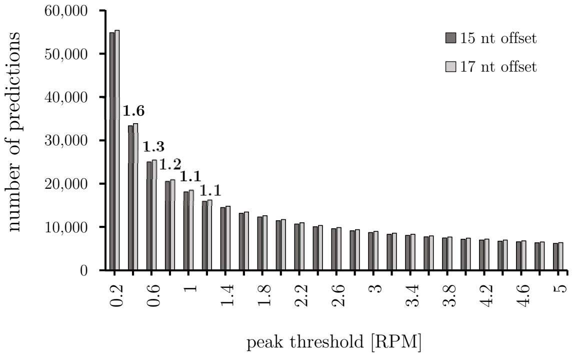
Supplementary Figure 3. Growth of (**A**) *E. coli* LF82 and (**B**) *E. coli* LF82Δ*tolC* in the presence of different RET concentrations indicated in µg/mL. At various time point during growth the optical density was measured in technical as well as biological triplicates. Bacterial cultures without RET served as a positive control (PC) and plain broth without cell inoculum was used as a negative control (NC).



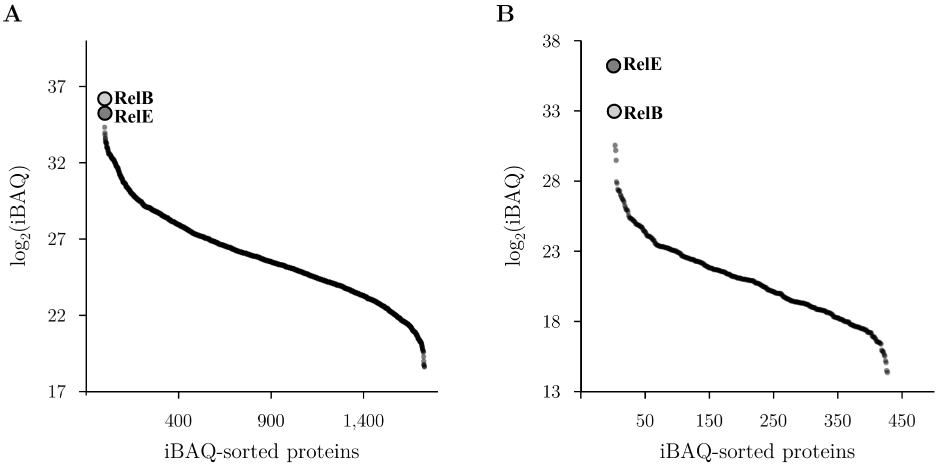
Supplementary Figure 4. Read length distributions of all *E. coli* LF82 sequencing datasets. The proportion of all reads mapping to mRNA regions is displayed for (**A**) Exp1_Ribo-seq_aerobic RepI+II, (**B**) Exp1_Ribo-seq_anaerobic RepI+II, (**C**) Exp1_RNA-seq_aerobic RepI+II, (**D**) Exp1_RNA-seq_anaerobic RepI+II, (**E**) Exp2_Ribo-seq ND & RET and (**F**) Exp3_RNA-seq ("own") as well as the respective control RNA-seq dataset published by Hwang & Buskirk (2017).



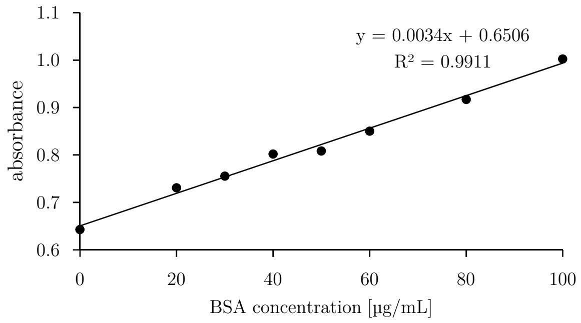
Supplementary Figure 5. Metagene analysis of highly expressed genes in the datasets published by Meydan *et al.* (2019). Shown are normalized RPM values of each position in a -10 to 30 nt window around the start codon (dashed line, distance 0 nt) after determining the P site position by subtracting 15 nt of all genes with RPM values ≥100 in the ND and RET datasets of *E. coli* strain BL21 (strain B) or *E. coli* strain BW25113 (strain K).



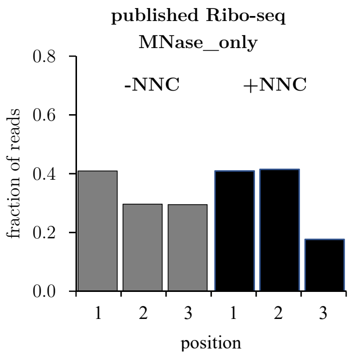
Supplementary Figure 6. Number of predictions after applying RET peak thresholds of 0.2 to 5 RPM after subtraction of a 15 nt offset (dark bars) or a 17 nt offset (light bars). Numbers above bars represent fold changes for the predictions obtained with a certain threshold in comparison to the respective value of the previous applied threshold.



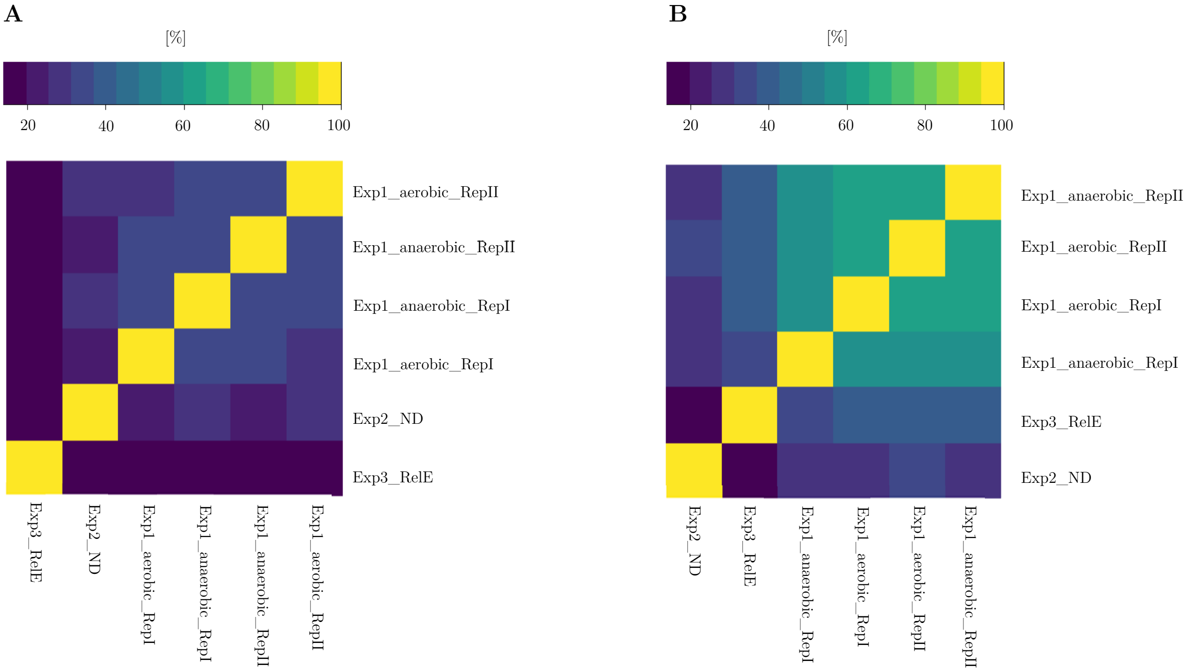
Supplementary Figure 7. Results of the mass spectrometry analysis of cells harvested (**A**) 3 h after induction of *relE* and *relB* expression and (**B**) purification using a FPLC system with a pre-packed column. The logarithmic iBAQ value of each detected protein (y-axis) was plotted against the detected proteins sorted by their iBAQ values (x-axis). Each dot represents one detected protein.



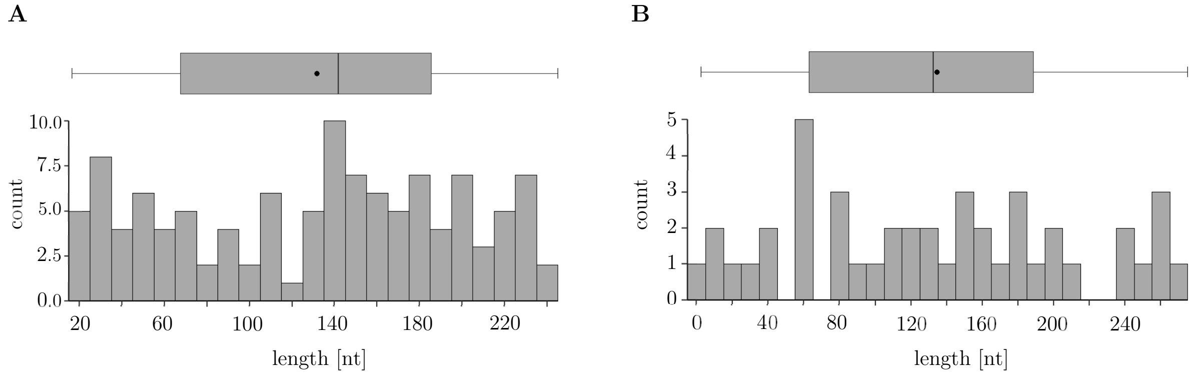
Supplementary Figure 8. Bovine serum albumin (BSA) calibration curve for RelE protein quantification using Bradford reagent. Absorbance was measured at 600 nm.



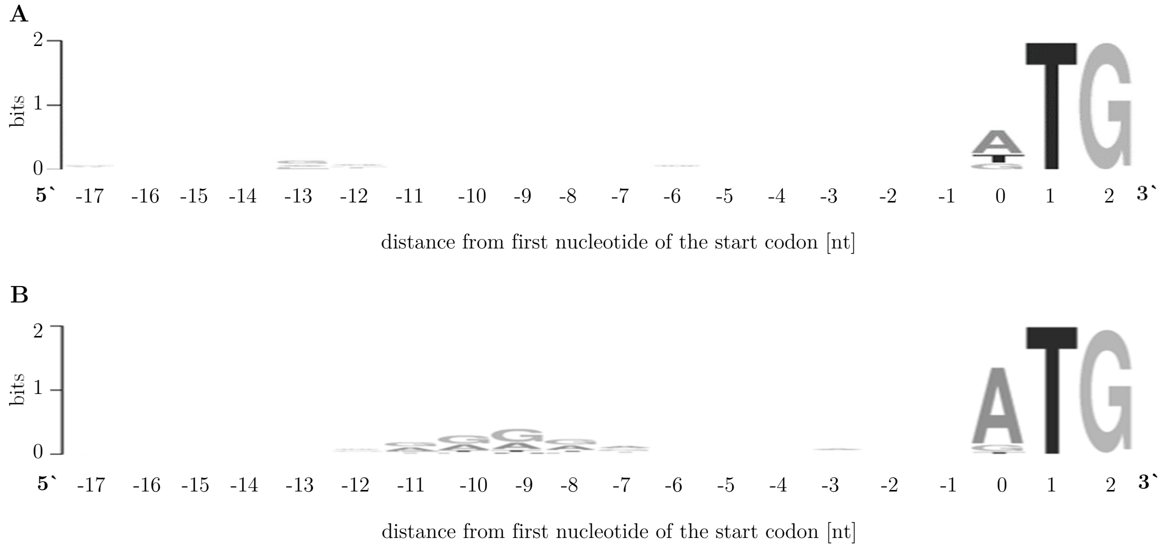
Supplementary Figure 9. Reading frame sum signal of all annotated genes for the Ribo-seq datasets published by Hwang & Buskirk (2017) using MNase only. Subtraction of 30 nt from start and stop positions was performed. The grey bars display the results of the raw data; the black bars show the reading frame signal after shifting of reads arising from codons ending in C.



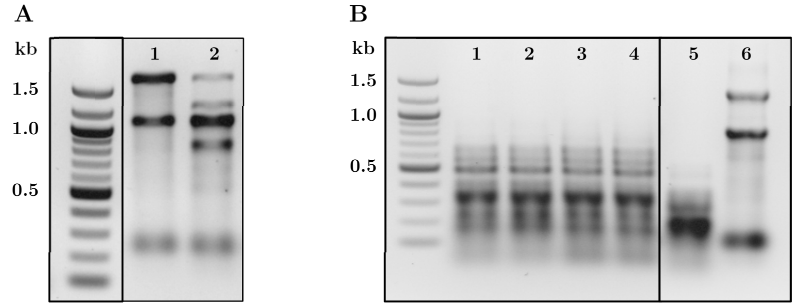
Supplementary Figure 10. Reproducibility of Ribo-seq prediction results. Shown are the percentages of identical predictions obtained by the tool (**A**) REPARATION and (**B**) the scripts by Giess *et al.* (B; 2017).



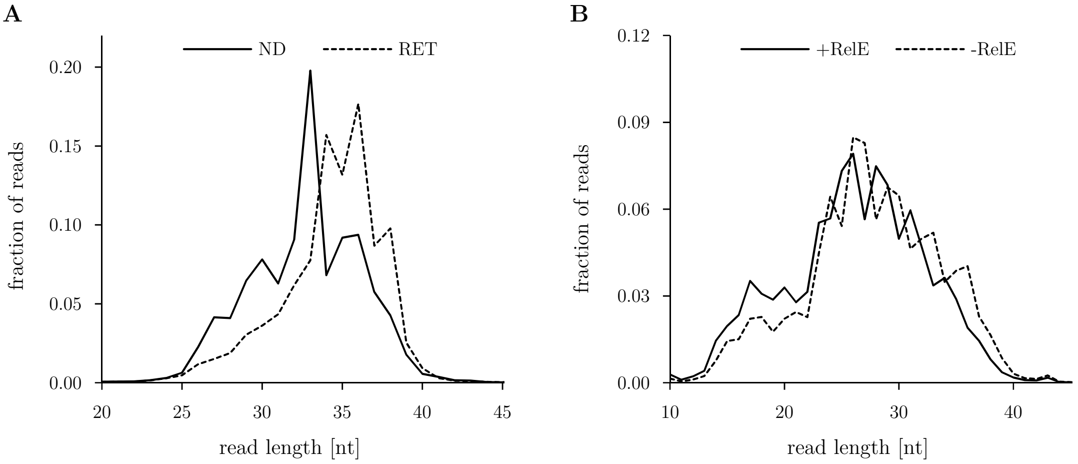
Supplementary Figure 11. Distance of (**A**) promoter and (**B**) terminator structures of the novel ORFs in *E. coli* LF82 as predicted by BROM (n = 115) and FindTerm (n = 44). The boxplot indicates first, second and third quartile values; the dot represents the mean value.



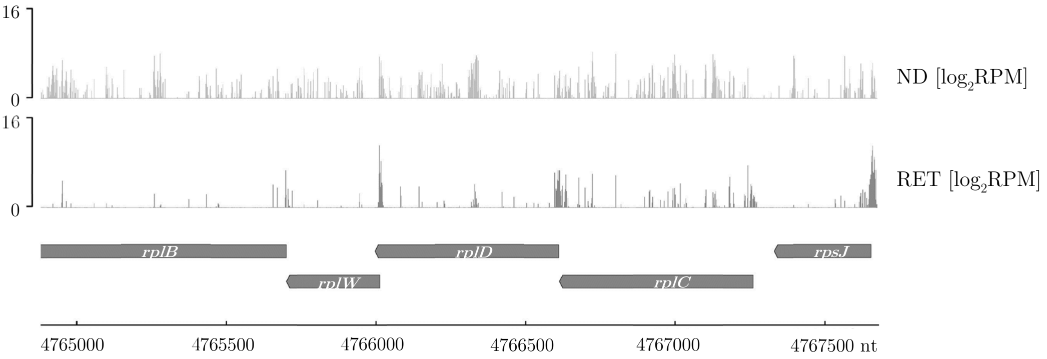
Supplementary Figure 12. Sequence logos of putative Shine-Dalgarno sequences in the region upstream of the start codon of (**A**) all novel gene candidates (n = 116) and (**B**) all protein-coding, annotated genes (n = 4,587).



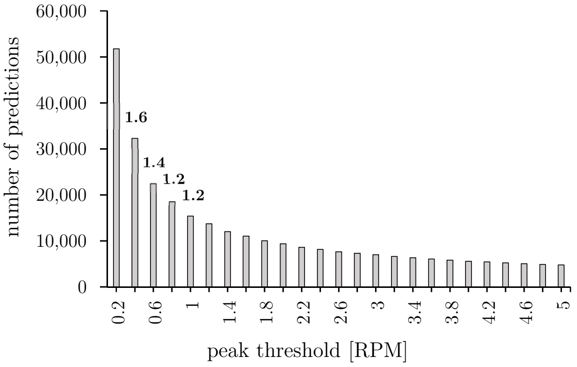
Supplementary Figure 13. Integrity of *P. aeruginosa* PAO1 RNA samples. RNA extraction and agarose gel electrophoresis of samples harvested at (**A**; lane 1) $OD_{600nm} = 1$ and (**A**; lane 2) $OD_{600nm} = 6$ for RNA-seq indicate a decrease in RNA quality with increasing optical density. (**B**) Analysis of RNA integrity after incubation with RNase I in different time-concentration combinations (**B**; lane 1: 3.75 U/AU RNase I for 45 min; lane 2: 3.75 U/AU RNase I for 60 min; lane 3: 2.5 U/AU RNase I for 45 min; lane 4: 2.5 U/AU RNase I for 60 min) suggests ineptitude of RNase I in *P. aeruginosa* Ribo-seq. Incubation with RNase A (lane 5) as well as without any nuclease (lane 6) served as positive and negative control, respectively.



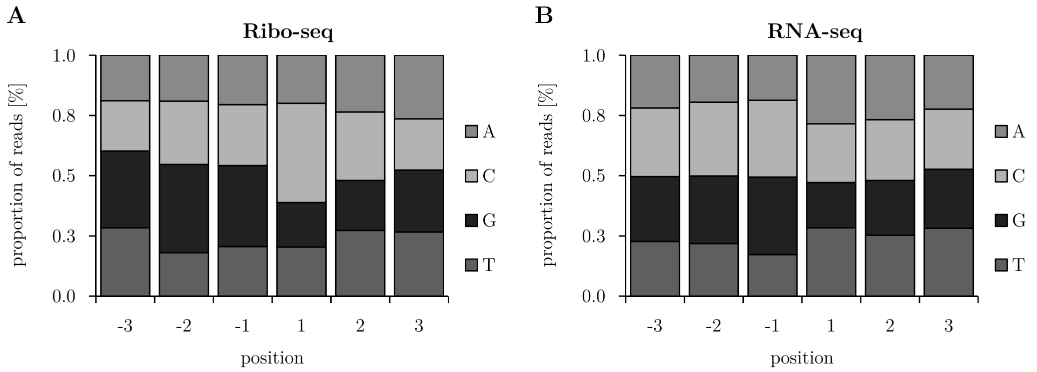
Supplementary Figure 14. mRNA read length distributions of the Ribo-seq datasets (**A**) Exp2 and (**B**) Exp3 for both main (RET or +RelE) and control experiment (ND or -RelE) in *P. aeruginosa* PAO1.



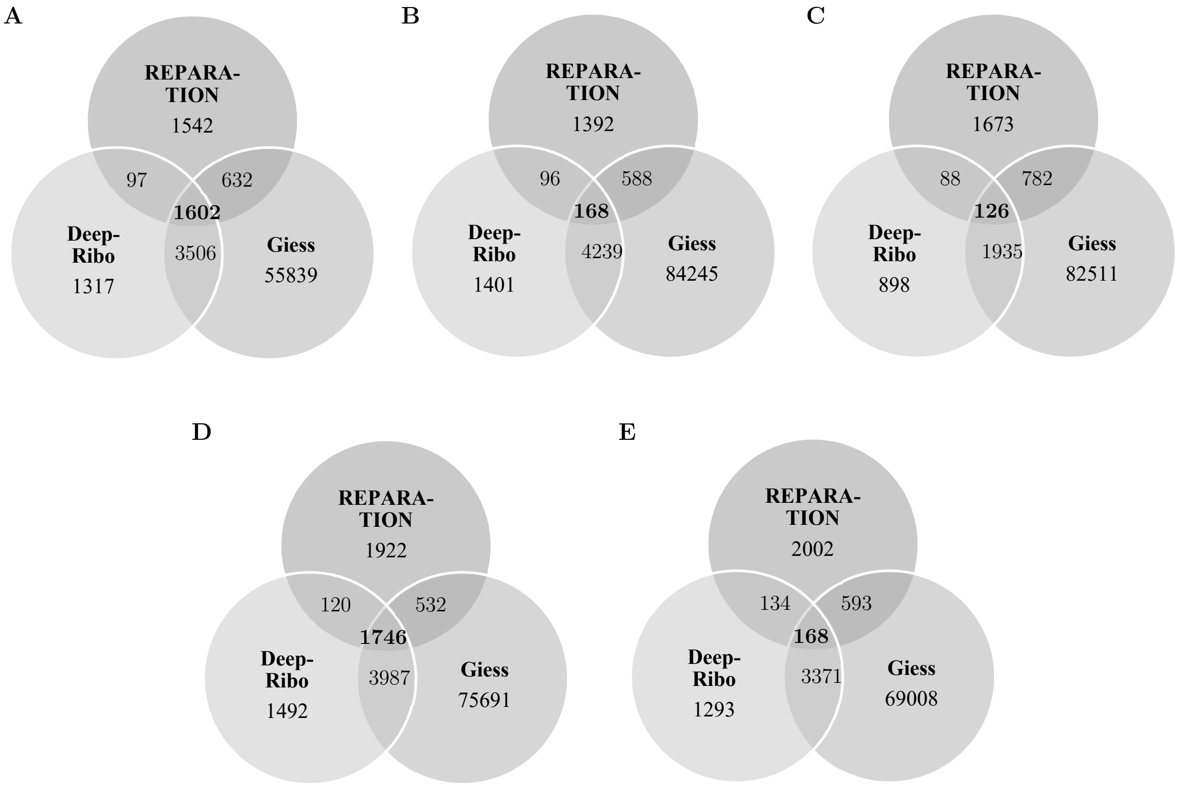
Supplementary Figure 15. Translation signals of highly expressed proteins of the S10 ribosomal protein operon in *P. aeruginosa* PAO397. Shown are the log$_2$RPM values at each position of the genes *rplB, rplW, rplD, rplC* and *rpsJ* in the ND (top panel) and the RET (bottom panel) datasets.



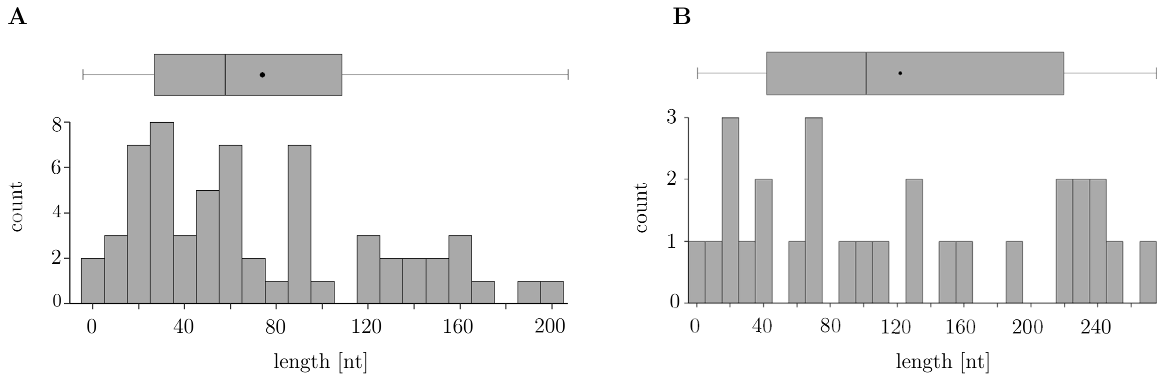
Supplementary Figure 16. Number of predictions after applying RET peak thresholds of 0.2 to 5 RPM after subtraction of a 15 nt offset in *P. aeruginosa* PAO397. Numbers above bars represent fold changes for the predictions obtained with a certain threshold in comparison to the respective value of the previous applied threshold.



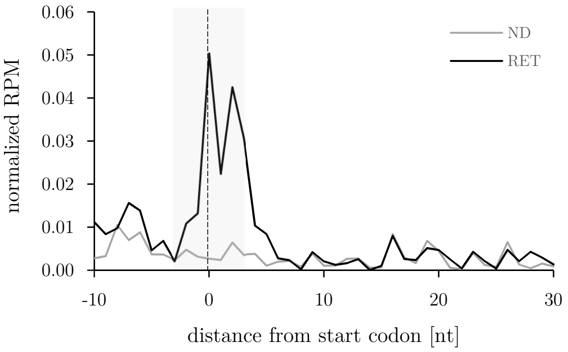
Supplementary Figure 17. Results of sequence bias at the 3´end of all mRNA reads obtained in the (**A**) Ribo-seq or (**B**) RNA-seq experiment in *E. coli* LF82.



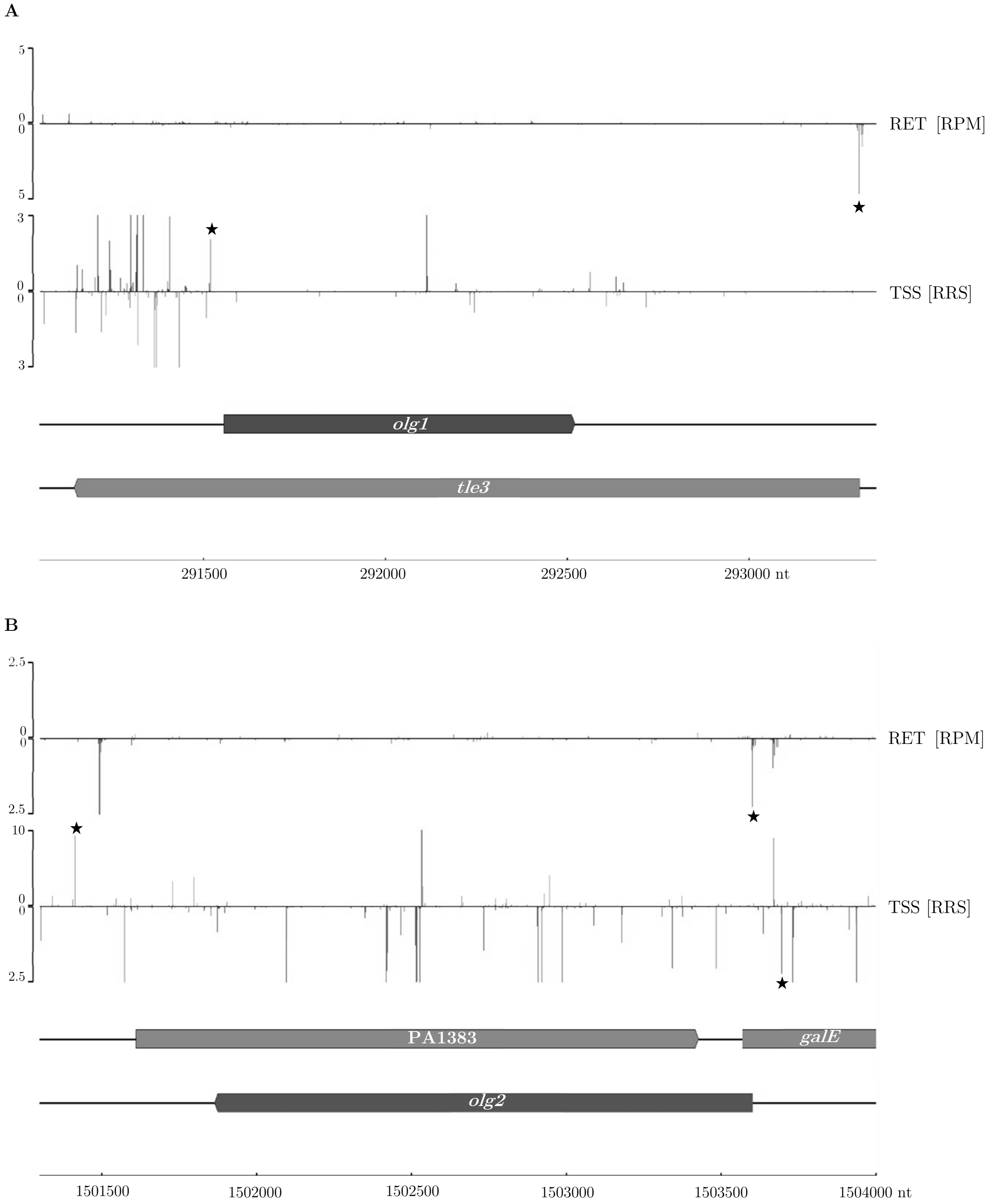
Supplementary Figure 18. Reproducibility of Ribo-seq prediction results in *P. aeruginosa* PAO1. The overlap and absolute number of predictions obtained by all three prediction tools (DeepRibo, REPARATION, scripts by Giess) for the dataset (**A**) Exp1_RepI, (**B**) Exp1_RepII, (**C**) Exp2_ND, (**D**) Exp3_MNase and (**E**) Exp3_RelE are shown.



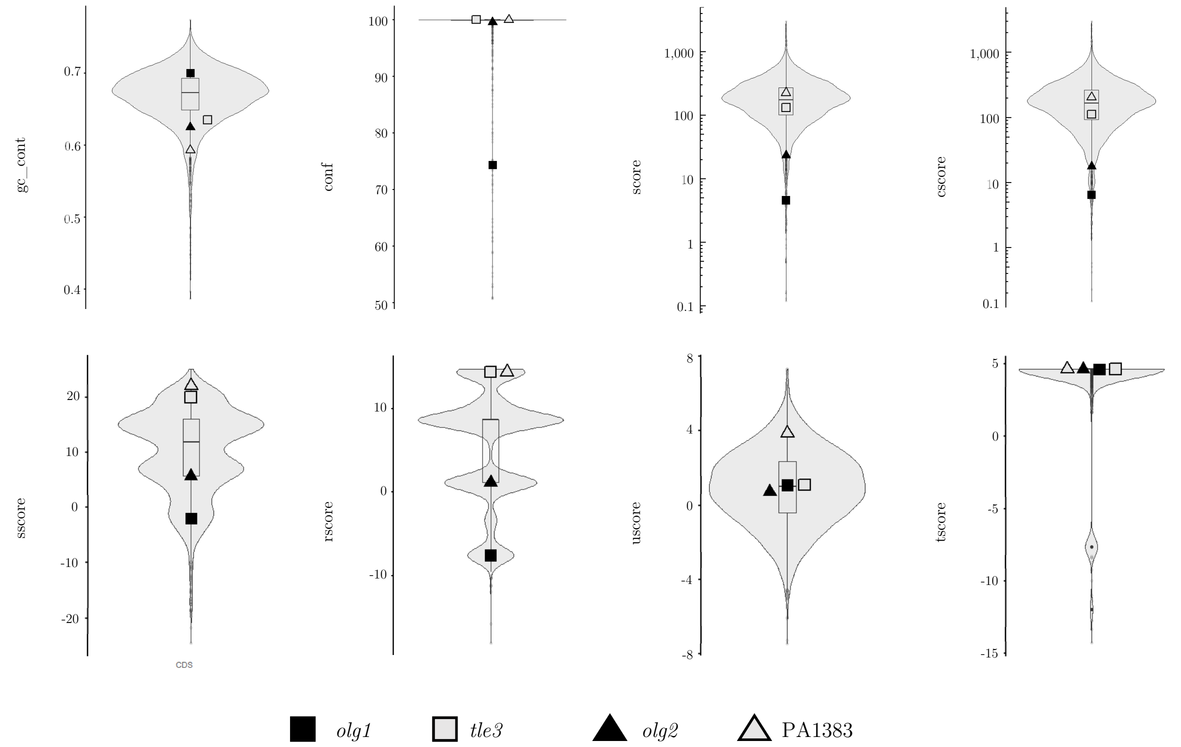
Supplementary Figure 19. Distance of (**A**) TSS and (**B**) terminator structures of the novel ORFs in *P. aeruginosa* PAO1 as indicated by Cappable-seq (n = 61) or predicted by FindTerm (n = 28). The boxplots indicate first, second and third quartile values; the dots represent the mean values.



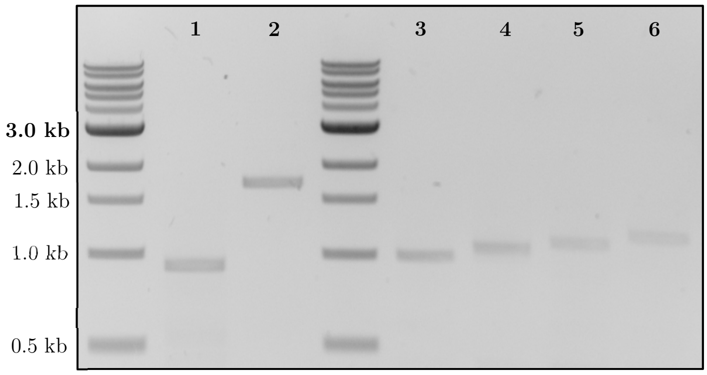
Supplementary Figure 20. Metagene analysis of novel gene candidates in *P. aeruginosa* PAO1. Normalized RPM values of each position in a -10 to 30 nt window around the start codon (dashed line, distance 0 nt) after determining the P site position of reads mapping to the novel gene candidates (n = 68) in the no drug (ND) and retapamulin (RET) Ribo-seq experiment are shown.



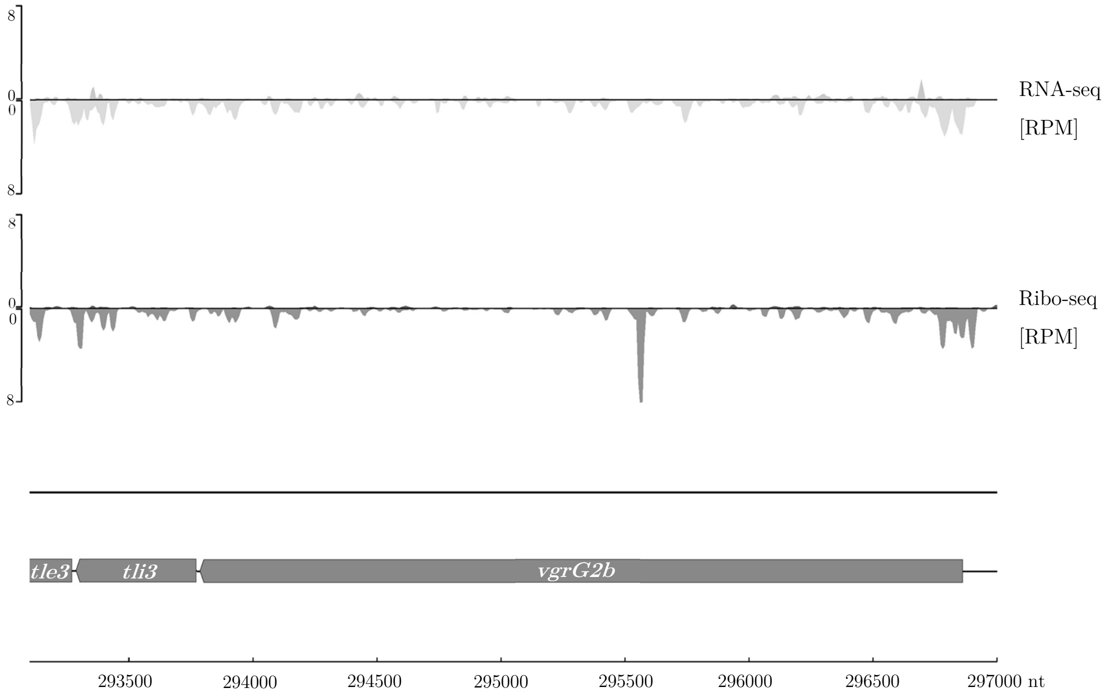
Supplementary Figure 21. Results of Ribo-RET and Cappable-seq for the (**A**) *olg1/tle3* and the (**B**) *olg2*/PA1383 locus in *P. aeruginosa* PAO1. Mean RPM and RRS values of all Ribo-RET (first track) and Cappable-seq reads (second track) of this study (n = 1 and n = 3, respectively) are shown for (**A**) *olg1* and the mother gene *tle3* (sense and antisense, respectively) as well as for (**B**) *olg2* and the mother genes PA1383 and *galE* (antisense and sense, respectively). Translation initiation (TIS) and transcription initiation sites (TSS) are indicated by asterisks.



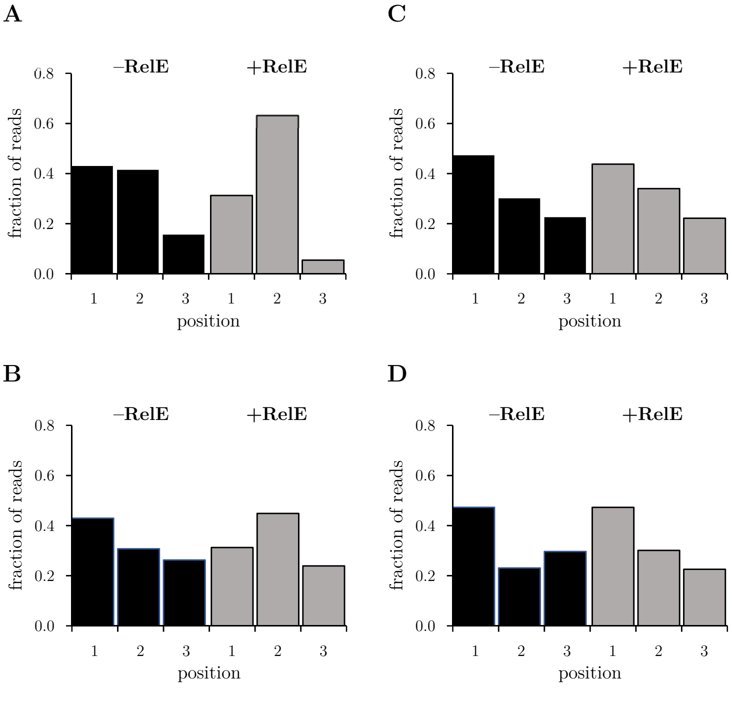
Supplementary Figure 22. Prodigal prediction results of the *P. aeruginosa* PAO1 genome with and without hiding of the mother genes *tle3* and PA1383 by replacing all possible start codons by Ns. Violin plots with included boxplots display the values of all predicted protein-coding genes (n = 5,681) for the following categories (from top left to bottom right): GC content (gc_cont), confidence score (conf), overall score (score), hexamer coding proportion score (cscore), TIS score (sscore), ribosome binding site score (rscore), region score flanking the start codon (uscore) and the start codon sequence score (tscore). All values obtained for *olg1* and *olg2* are indicated by a black rectangle and triangle, respectively. Values of their mother genes are represented by the respective grey-shaded symbol. Figure adapted from Kreitmeier *et al.* (2021).



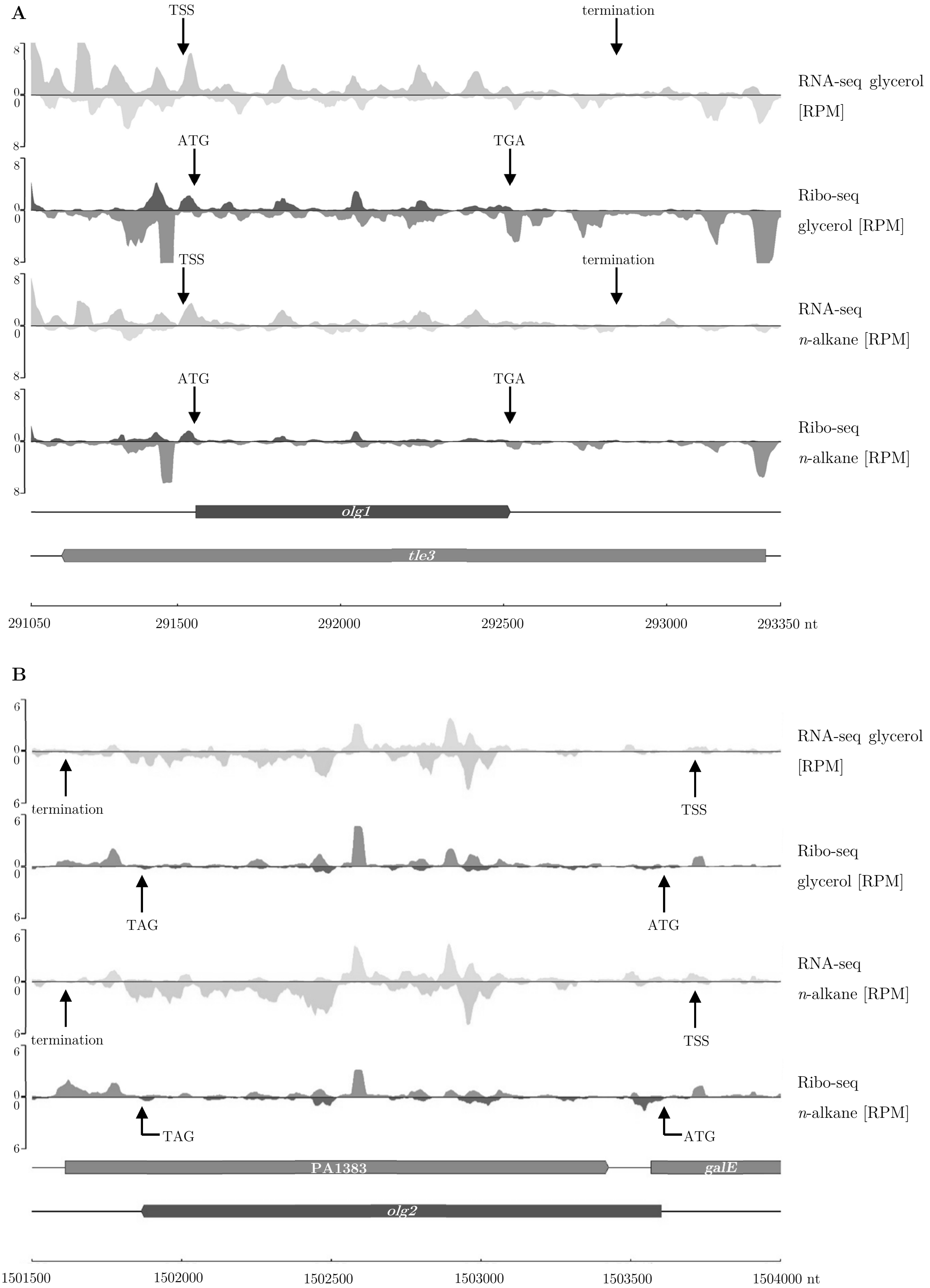
Supplementary Figure 23. Results of the RT-PCR analysis of *olg1* and *olg2*. Full-length transcription of *olg1* (lane 1, target length = 917 nt) and *olg2* (lane 2, target length = 1,696 nt) was verified by RT-PCR with primers binding at the N-terminal and C-terminal region of both OLGs. Transcription of *olg1* starts considerably upstream of the coding region as indicated by the detection of a RNA transcript using primers binding 45 nt (lane 3, target length = 995 nt), 110 nt (lane 4, target length = 1,060 nt), 161 nt (lane 5, target length = 1,111 nt) and 240 nt (lane 6, target length = 1,190 nt) upstream of the start codon. Figure adapted from Kreitmeier *et al.* (2021)



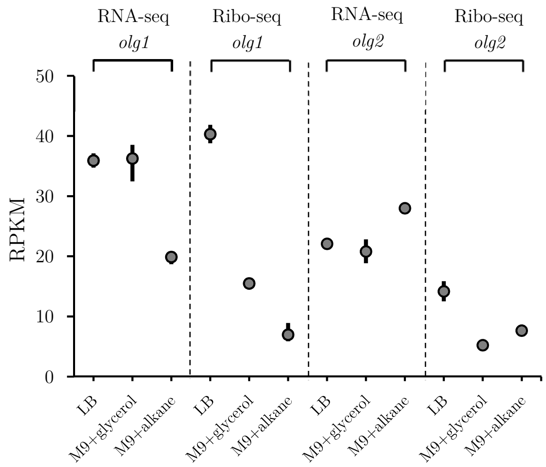
Supplementary Figure 24. Results of RNA-seq and Ribo-seq analysis of the genes *tli3* and *vgrG2b* encoded upstream of the mother gene *tle3* in *P. aeruginosa* PAO1. Strand-specific RPM values of RNA-seq (first track) and Ribo-seq reads (second track) averaged over the biological replicates of Exp1 are displayed. A lack of signals antisense to these genes indicate the absence of pervasive transcription and translation at this locus. Figure adapted from Kreitmeier *et al.* (2021).



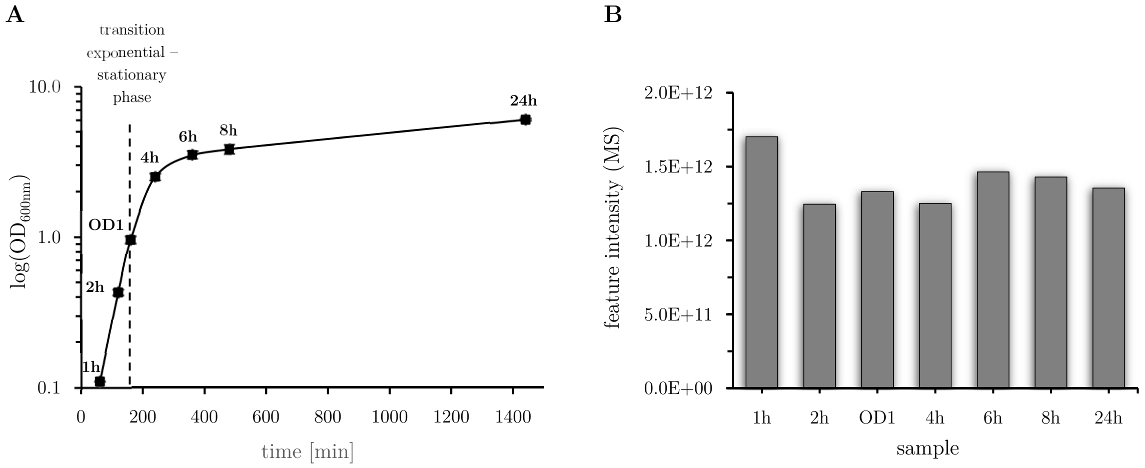
Supplementary Figure 25. Reading frame signal of novel (**A, B**) *olg1* and (**C, D**) *olg2* in *P. aeruginosa* PAO1. All reads obtained for Exp3 generated with (+RelE) or without RelE (-RelE) were mapped to each sub-codon position. Figure **A** and **D** display the distribution of all reads after NNC shift; in Figure **B** and **D** the signals arising from unshifted reads are illustrated.



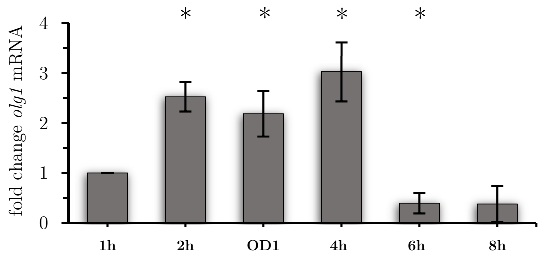
Supplementary Figure 26. Transcriptome and translatome signals at the (**A**) *olg1*/*tle3* and (**B**) *olg2*/PA1383 locus obtained for the datasets provided by Grady *et al.* (2017). Strand-specific RPM values of RNA-seq (first & third track) and Ribo-seq reads (second & fourth track) averaged over the biological triplicates of the datasets "M9+glycerol" and "M9+*n*-alkane" (n = 3, each) were shown. Arrows indicate the position of transcription start (TSS) and stop sites (termination) of *olg1* and *olg2*. Figure adapted from Kreitmeier *et al.* (2021).



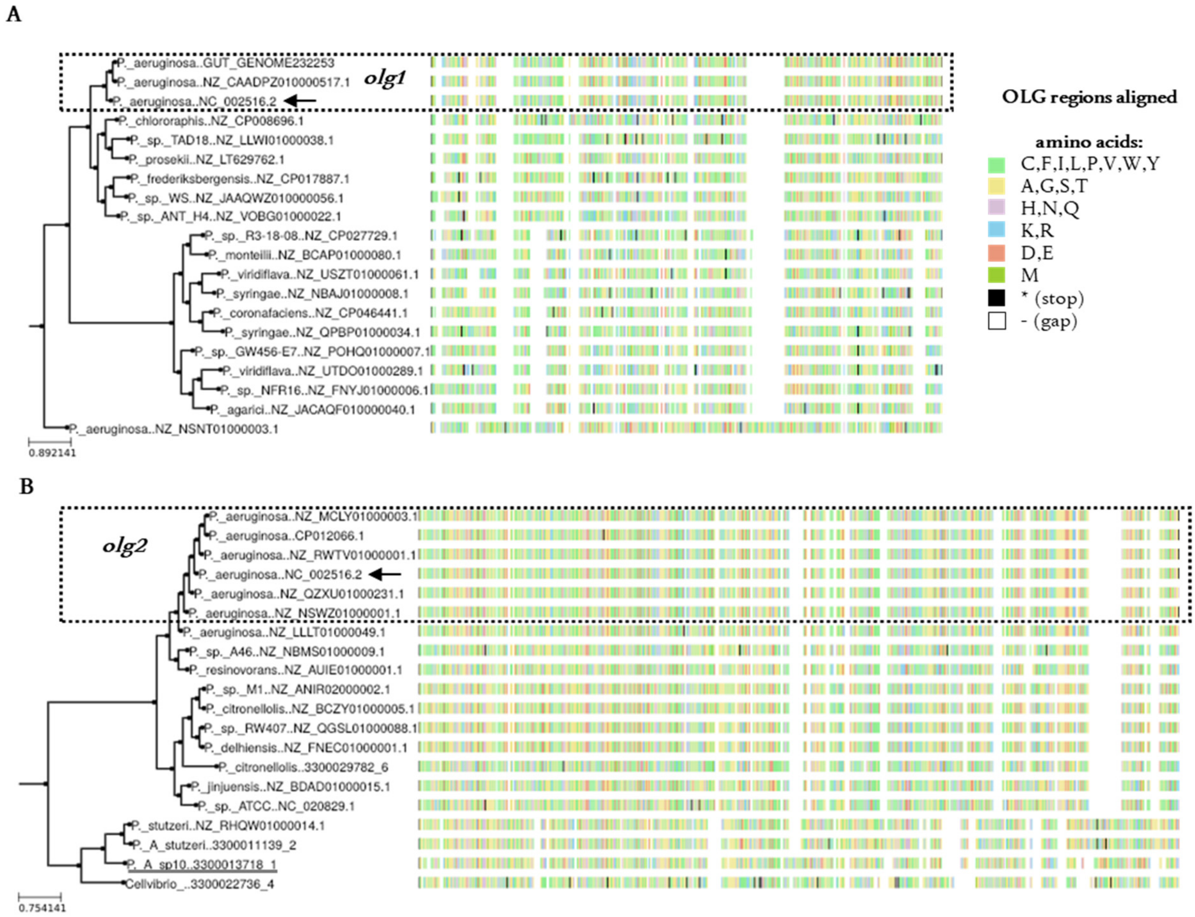
Supplementary Figure 27. Mean RNA-seq and Ribo-seq reads per reads per kilobase per million mapped reads (RPKM) for *olg1* and *olg2* of the datasets generated in this study ("LB", n = 2, each) and the datasets provided by Grady *et al.* (2017), who performed RNA-seq and Ribo-seq experiments with *P. aeruginosa* PAO1 after cultivation in M9 broth with glycerol ("M9+glycerol", n = 3) or *n*-alkanes ("M9+alkane", n = 3) as sole carbon source. Figure adapted from Kreitmeier *et al.* (2021).



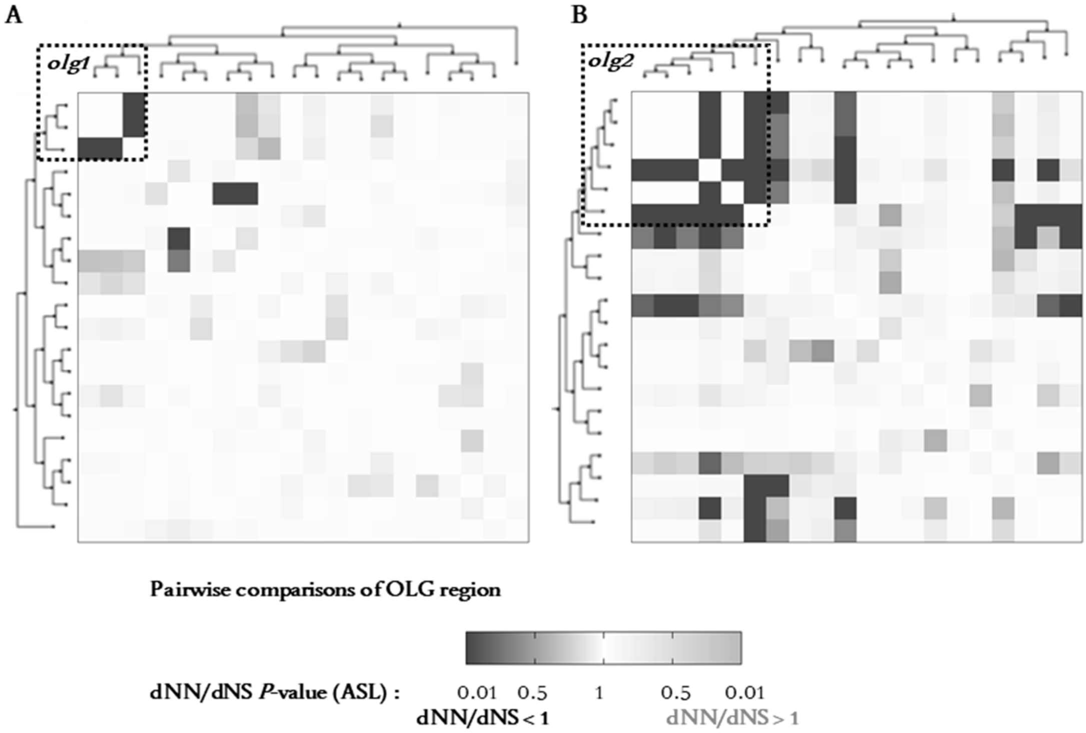
Supplementary Figure 28. (**A**) Sampling points and (**B**) loading control for PRM-MS analysis. (**A**) The growth curve of *P. aeruginosa* PAO1 in LB broth of three biological replicates is shown. Samples were taken at 1 h, 2 h, 4 h, 6 h, 8 h and 24 h as well as at OD_{600nm} = 1 (~160 min) and measured via PRM-MS. (**B**) The measured intensities of all peptides detected in each sample. Figure adapted from Kreitmeier *et al.* (2021).



Supplementary Figure 29. qPCR analysis of *olg1* mRNA in *P. aeruginosa* PAO1. Ct values were measured for samples taken 1 h, 2 h, 4 h, 6 h, 8 h and 24 h after cultivation as well as at OD_{600nm} = 1 (~160 min). All values were normalized to the values obtained for the housekeeping gene *gyrA* and compared to the expression level measured for the 1 h sample. Statistical significance was evaluated based on pairwise comparison of the $\log_2$ fold changes obtained for the 1 h sample with those of all other samples using a two-tailed Welch two sample t-test (*p ≤ 0.05). Figure adapted from Kreitmeier *et al.* (2021).



Supplementary Figure 30. Phylogenetic analysis indicates taxonomic restriction of (**A**) *olg1* and (**B**) *olg2*. The amino acid sequence of the mother genes (**A**) *tle3* and (**B**) PA1383 was used to calculate a maximum likelihood tree which was down sampled to 20 genomes for graphic representation of homologous *olg1* and *olg2* sequences. Genomes which harbour homologous ORFs sharing either (**A**) the same start and stop codon or (**B**) only the same stop codon as the reference genome of *P. aeruginosa* PAO1 (highlighted by an arrow) are indicated by a dotted box. The outgroup used for evolutionary analyses is underlined. Figure adapted from Kreitmeier *et al.* (2021).



Supplementary Figure 31. dNN/dNS analysis of the (**A**) *olg1* and (**B**) *olg2* region using OLGenie. Pairwise comparisons of dNN/dNS ratios indicate the strongest evidence for purifying selection within the *P. aeruginosa* clade. However, purifying selection on *olg2* can be also found in other species of the genus *Pseudomonas*. Figure adapted from Kreitmeier *et al.* (2021).

6.3 Supplementary Tables

Supplementary Table 1. All chemicals and reagents used in this study listed according to their manufacturer. If necessary, sub-brands and abbreviations are stated.

Manufacturer (including associated brands)	Chemicals or reagents (abbreviation)
BioVectra	isopropyl β-d-1-thiogalactopyranoside (IPTG)
Carl Roth	acetic acid, agarose, ammonium chloride (NH_4Cl), ammonium persulfate (APS), arabinose, boric acid, bovine serum albumin (BSA), calcium chloride ($CaCl_2$), cetrimonium bromide (CTAB), chloroform, diethyl pyrocarbonate (DEPC), disodium phosphate (Na_2HPO_4), ethylene glycol-bis(2-aminoethylether)-N,N,N′,N′-tetraacetic acid (EGTA), ethylenediaminetetraacetic acid (EDTA), glutaraldehyde, glycine, hydrogen chloride (HCl), iron(II) sulfate ($FeSO_4$), isopropyl, magnesium chloride ($MgCl_2$), magnesium sulfate ($MgSO_4$), methanol, milk powder, monopotassium phosphate (KH_2PO_4), monosodium phosphate (NaH_2PO_4), Ponceau S, potassium chloride (KCl), RotiPhenol, RotiPhenol/chloroform/isoamylalcohol, Rotiphorese NF-Acrylamide/Bis-solution 30 (29:1), Rotiphorese Sequencing gel buffer concentrate, Rotiphorese Sequencing gel concentrate, Rotiphorese Sequencing gel diluent, Roti-Quant 5X-staining solution, Schaedler bouillon, sodium acetate (NaOAc), sodium chloride (NaCl), sodium dodecyl sulfate (SDS), sodium hydroxide (NaOH), tetramethylethylendiamine (TEMED), tricine, tris-(hydroxymethyl)-aminomethan (Tris), Triton X-100, Tween 20, urea
iNtRON biotechnology	RedSafe Nucleic Acid Staining Solution
J.T.Baker (including Avantor)	ethanol (EtOH)
Merck	2-mercaptoethanol, bromphenolblue, chloroacetamide (CAA), glycerol, nuclease-free H_2O
Roche	Trypsin recombinant, Proteomics Grade
Sigma-Aldrich (including Fluka)	ammonium acetate, ammonium bicarbonate, ammonium formate, bicine, Bradford-Reagent B6916, CDP-Star AP substrate, Coomassie Brilliant Blue G-250, dimethyl sulfoxide (DMSO), dithiothreitol (DTT), formic acid (FA), glucose, guanosine 5′-(β-γ-imido)-triphosphate (GMP-PNP), imidazole, NP-40, sucrose, trifluoroacetic acid (TFA), Tris(2-carboxyethyl)phosphine (TCEP)
Thermo Fisher Scientific (including Invitrogen & Oxoid)	agar, dNTPs, ethylenediaminetetraacetic acid (EDTA), glycogen, sodium acetate (NaOAc), SYBR Gold Nucleic Acid Gel Stain, TRIzol Reagent, tryptone, yeast extract
VWR International	acetonitrile (ACN)

Supplementary Table 2. Overview of Ribo-seq and RNA-seq reads after sequencing of all *E. coli* LF82 experiments. Shown are the total number of reads in millions for all categories.

Experiment	**Sequenced reads**	**Too short and low-quality reads**	**Processed reads**	**Reads not mapping to target genome**	**Reads mapping to target genome**	**tRNA reads**	**rRNA reads**	**mRNA reads**
Exp1_Ribo-seq_aerobic_RepI	65.1	37.9	27.2	11.9	15.4	2.3	3.3	9.8
Exp1_Ribo-seq_anaerobic_RepI	60.6	29.3	31.3	14.2	17.2	0.4	9.9	6.9
Exp1_RNA-seq_aerobic_RepI	64.5	20.3	44.3	26.0	18.2	7.5	4.8	5.9
Exp1_RNA-seq_anaerobic_RepI	70.2	16.3	53.9	20.0	34.0	13.4	5.0	15.5
Exp1_Ribo-seq_aerobic_RepII	79.6	37.2	42.4	19.8	22.6	0.3	1.1	21.2
Exp1_Ribo-seq_anaerobic_RepII	82.8	38.7	44.1	19.4	24.7	1.0	2.7	21.0
Exp1_RNA-seq_aerobic_RepII	81.2	13.1	68.1	24.1	44.1	12.6	9.9	21.6
Exp1_RNA-seq_anaerobic_RepII	98.6	17.4	81.2	29.5	51.6	19.7	2.7	29.2
Exp2_Ribo-seq_RET	199.0	13.6	185.4	8.7	176.8	1.6	151.5	23.8
Exp2_Ribo-seq_ND	271.1	23.8	247.3	16.6	230.7	3.4	161.5	65.8
Exp3_Ribo-seq_RelE	254.1	133.1	121.0	95.6	25.3	1.9	0.0	23.4
Exp3_RNA-seq	3.1	0.1	2.9	2.2	0.7	0.3	0.0	0.4

Supplementary Table 3. Differential expressed anORFs (n = 210) showing logFC ≥|1| and a p-value ≤0.05 in transcriptome sequencing and/or ribosome profiling (Exp1_RepI+II) in *E. coli* LF82.

Locus_Tag	**Gene**	**RNA-seq logFC**	**RNA-seq logCPM**	**RNA-seq p-value**	**Ribo-seq logFC**	**Ribo-seq logCPM**	**Ribo-seq p-value**
LF82_RS05430	*flgB*	-5.34	5.38	0.00	-5.28	6.02	0.00
LF82_RS05445	*flgE*	-4.97	6.19	0.00	-3.57	5.50	0.00
LF82_RS05435	*flgC*	-4.72	5.05	0.00	-4.64	5.69	0.00
LF82_RS05440	*flgD*	-4.68	5.84	0.00	-2.42	4.87	0.01
LF82_RS05450	*flgF*	-4.40	4.81	0.00	-2.97	3.82	0.00
LF82_RS05455	*flgG*	-4.08	5.78	0.00	-3.53	7.86	0.00
LF82_RS10070	*fliG*	-3.46	4.81	0.00	-1.96	4.07	0.01
LF82_RS10080	*fliI*	-3.38	4.70	0.00	-1.84	4.72	0.04
LF82_RS05460	*flgH*	-2.97	3.25	0.01	-2.64	3.08	0.01
LF82_RS05425	*flgA*	-2.92	4.18	0.00	-2.86	3.65	0.00
LF82_RS05465	*flgI*	-2.91	3.71	0.00	-2.80	3.93	0.00
LF82_RS10115	*fliP*	-2.37	2.73	0.01	-1.79	3.56	0.04
LF82_RS20800	*metF*	-2.10	5.53	0.00	-2.54	5.36	0.00
LF82_RS25490		-2.01	8.38	0.01	-2.77	8.26	0.00
LF82_RS10105	*fliN*	-1.74	4.63	0.02	-1.64	3.04	0.02
LF82_RS10100	*fliM*	-1.71	5.20	0.02	-1.36	4.08	0.03
LF82_RS20080	*metE*	-1.70	10.73	0.01	-2.35	11.12	0.01
LF82_RS21865	*fumB*	-1.69	6.07	0.01	-1.31	5.44	0.05
LF82_RS01040	*metN*	-1.61	5.63	0.00	-1.49	4.68	0.02
LF82_RS20075	*metR*	-1.61	5.33	0.00	-2.27	5.04	0.00
LF82_RS21190	*metA*	-1.57	6.34	0.01	-1.60	5.86	0.03
LF82_RS02915	*citD*	-1.24	5.46	0.03	-2.88	7.31	0.00
LF82_RS07590	*fdnH*	-1.18	3.57	0.05	-1.56	4.31	0.02
LF82_RS07445	*ansP*	-1.12	3.87	0.04	1.96	5.43	0.01
LF82_RS02800	*entC*	1.08	4.86	0.05	1.37	5.77	0.04
LF82_RS02820	*entH*	1.26	4.22	0.02	1.61	3.66	0.01
LF82_RS20035		-1.78	7.66	0.08	4.82	6.05	0.00
LF82_RS16440	*agaV*	-1.68	4.22	0.12	1.70	2.84	0.01
LF82_RS19615	*atpH*	-1.34	7.39	0.08	-2.34	6.96	0.01
LF82_RS14280	*cysN*	-1.20	4.12	0.10	-1.68	4.35	0.02
LF82_RS10095	*fliL*	-1.13	4.45	0.09	-1.59	3.54	0.03
LF82_RS15280	*metK*	-1.11	11.28	0.11	-1.96	9.86	0.03
LF82_RS19195		-1.11	2.81	0.07	2.28	3.69	0.00
LF82_RS19185		-1.08	5.39	0.07	2.38	6.56	0.01
LF82_RS17150	*rrf*	-1.06	1.21	0.06	2.17	2.69	0.04
LF82_RS14725	*amiC*	-1.05	7.24	0.11	-1.90	7.30	0.02
LF82_RS12960	*guaB*	-0.99	5.98	0.24	-1.90	6.29	0.04
LF82_RS09460	*mntP*	-0.98	2.42	0.08	1.86	3.99	0.01
LF82_RS15440		-0.97	7.26	0.09	3.16	8.22	0.00
LF82_RS17990	*gntU*	-0.93	5.81	0.23	1.61	6.21	0.02
LF82_RS02560		-0.93	5.54	0.23	1.49	5.26	0.03
LF82_RS03845	*glnQ*	-0.93	7.61	0.20	-2.11	7.64	0.02
LF82_RS26000		-0.91	5.82	0.14	3.05	6.62	0.00
LF82_RS15230	*cmtB*	-0.89	1.00	0.20	1.71	1.58	0.03
LF82_RS18800		-0.88	2.85	0.31	1.41	2.54	0.03
LF82_RS20465		-0.88	5.24	0.31	1.38	4.63	0.05
LF82_RS05250	*csgD*	-0.87	4.92	0.10	3.09	6.13	0.00
LF82_RS18095	*livH*	-0.80	3.75	0.15	1.54	5.78	0.05
LF82_RS21925	*cadB*	-0.79	6.81	0.27	-1.40	6.14	0.05
LF82_RS10340		-0.78	10.10	0.28	3.07	11.01	0.00
LF82_RS02550		-0.77	7.85	0.35	1.97	5.36	0.02
LF82_RS14685	*tssJ*	-0.75	2.91	0.18	1.74	3.63	0.03
LF82_RS02590	*cysS*	-0.73	8.31	0.27	1.96	9.27	0.03
LF82_RS16145	*rpoD*	-0.71	11.51	0.29	2.21	12.05	0.03
LF82_RS10850	*wcaE*	-0.70	2.21	0.35	1.59	2.16	0.03
LF82_RS17030		-0.69	1.16	0.23	1.56	2.15	0.02
LF82_RS16180		-0.67	3.77	0.23	2.46	4.97	0.00
LF82_RS06620	*oppB*	-0.66	2.22	0.32	-1.62	3.46	0.03
LF82_RS22800	*pyrB*	-0.66	3.51	0.23	-1.47	4.66	0.04
LF82_RS19620	*atpF*	-0.65	8.30	0.38	-1.79	7.44	0.04
LF82_RS04475	*ompF*	-0.63	7.29	0.47	-2.32	6.22	0.01
LF82_RS09505		-0.61	4.27	0.27	2.71	5.62	0.00
LF82_RS10445		-0.60	5.64	0.26	-1.91	7.21	0.03
LF82_RS01705	*prpR*	-0.57	3.16	0.35	1.77	4.97	0.02
LF82_RS08615		-0.56	5.08	0.33	2.19	5.93	0.01
LF82_RS00080		-0.55	0.92	0.31	1.35	2.68	0.05
LF82_RS00670	*gcd*	-0.55	7.61	0.41	1.67	5.89	0.05
LF82_RS10735	*wzzB*	-0.55	7.32	0.46	3.71	8.70	0.00
LF82_RS14595	*tssB*	-0.54	2.00	0.28	1.36	3.03	0.02
LF82_RS15985		-0.54	4.33	0.28	1.62	4.97	0.02
LF82_RS19350		-0.54	6.35	0.46	1.42	4.71	0.02
LF82_RS03395	*sdhA*	-0.53	4.54	0.32	1.42	5.32	0.05
LF82_RS00365	*sgrR*	-0.51	7.35	0.48	2.51	7.38	0.01
LF82_RS17865	*malT*	-0.51	5.34	0.38	1.87	5.91	0.02
LF82_RS09840	*flhD*	-0.49	4.95	0.35	-1.37	4.54	0.02
LF82_RS04960		-0.47	7.61	0.51	-1.82	7.04	0.05
LF82_RS05190	*pgaB*	-0.46	3.72	0.39	1.54	3.65	0.03
LF82_RS19270	*adeD*	-0.44	3.68	0.41	1.53	4.28	0.02
LF82_RS12130	*hisQ*	-0.43	3.32	0.38	2.23	3.60	0.01
LF82_RS04760		-0.42	4.91	0.44	-2.12	6.79	0.02
LF82_RS09700		-0.36	4.44	0.59	3.35	5.88	0.00
LF82_RS22500	*ulaR*	-0.35	8.34	0.60	3.11	9.44	0.01
LF82_RS12350	*emrK*	-0.35	2.42	0.48	1.22	2.82	0.04
LF82_RS17700	*rpe*	-0.35	6.18	0.51	-1.93	8.47	0.03
LF82_RS11305	*preA*	-0.32	4.56	0.52	2.25	5.11	0.00
LF82_RS00450	*murF*	-0.30	7.39	0.60	-1.67	7.53	0.04
LF82_RS19935	*aslA*	-0.29	3.46	0.60	2.16	4.46	0.01
LF82_RS06625	*oppC*	-0.29	2.76	0.62	-1.62	4.70	0.04
LF82_RS06230		-0.24	2.71	0.69	-1.83	4.97	0.03
LF82_RS19550	*bglF*	-0.21	5.18	0.72	2.97	6.12	0.00
LF82_RS03680	*bioD*	-0.19	4.04	0.72	1.65	4.15	0.03
LF82_RS14920		-0.16	3.34	0.75	1.41	4.21	0.03
LF82_RS04440	*mukE*	-0.14	4.30	0.80	-1.61	4.94	0.04
LF82_RS17365	*rplV*	-0.14	8.85	0.85	-2.67	10.98	0.03
LF82_RS15635		-0.10	2.12	0.86	2.07	3.30	0.01
LF82_RS08980	*hxpB*	-0.10	3.67	0.84	1.80	5.85	0.01

LF82_RS17380	*rplW*	-0.10	7.95	0.90	-2.26	9.69	0.05
LF82_RS25245	*yoaJ*	-0.08	1.70	0.88	-1.57	3.47	0.02
LF82_RS00315	*araD*	-0.05	3.80	0.91	1.51	2.83	0.04
LF82_RS21455	*sucC*	-0.05	3.50	0.92	1.51	4.43	0.02
LF82_RS19375		-0.01	4.13	0.99	-1.52	6.19	0.04
LF82_RS12355	*evgA*	0.00	4.07	0.99	-1.90	4.85	0.02
LF82_RS04890		0.00	5.17	1.00	-2.15	7.70	0.03
LF82_RS09770	*argS*	0.01	6.93	0.99	2.62	7.19	0.01
LF82_RS08115		0.02	3.83	0.99	1.43	5.04	0.03
LF82_RS18230		0.02	2.24	1.00	1.44	1.75	0.03
LF82_RS20185		0.03	1.75	0.97	1.44	2.62	0.02
LF82_RS13505		0.04	4.10	0.96	-1.99	4.66	0.02
LF82_RS02580	*lpxH*	0.04	5.30	0.95	2.11	5.24	0.00
LF82_RS22640	*qorB*	0.06	4.94	0.92	-1.43	6.08	0.04
LF82_RS04735		0.06	4.92	0.94	2.79	5.86	0.00
LF82_RS07145	*zntB*	0.06	6.43	0.92	-1.87	8.18	0.05
LF82_RS10930	*pphC*	0.08	1.45	0.93	1.69	2.64	0.01
LF82_RS04370	*ycaL*	0.08	3.75	0.88	1.65	3.57	0.02
LF82_RS05970		0.11	2.21	0.83	1.48	3.18	0.02
LF82_RS16795	*npr*	0.12	5.74	0.83	-1.90	6.51	0.03
LF82_RS17370	*rpsS*	0.15	7.60	0.84	-2.84	10.48	0.03
LF82_RS17300	*rpmD*	0.21	7.79	0.74	-2.56	10.01	0.02
LF82_RS05655	*lolE*	0.22	4.87	0.66	1.38	4.93	0.03
LF82_RS13055	*iscX*	0.23	4.81	0.65	-2.38	6.81	0.01
LF82_RS04885		0.38	4.55	0.59	-2.35	7.62	0.02
LF82_RS02760	*ybdZ*	0.65	2.99	0.32	1.67	3.89	0.01
LF82_RS07930	*ydfK*	0.66	0.70	0.30	1.37	1.91	0.04
LF82_RS04785		0.70	3.85	0.45	-1.79	5.36	0.03
LF82_RS21415	*pspG*	0.79	1.79	0.29	2.05	4.63	0.04
LF82_RS23910		0.80	3.23	0.31	1.24	3.32	0.04
LF82_RS02815	*entA*	0.97	5.11	0.12	1.50	4.82	0.02
LF82_RS05100		1.09	6.08	0.32	2.13	6.99	0.03
LF82_RS06210		1.18	1.05	0.13	2.52	2.01	0.00
LF82_RS10060	*fliE*	NA	NA	NA	-2.44	2.18	0.00
LF82_RS10085	*fliJ*	NA	NA	NA	-2.35	3.31	0.01
LF82_RS05940	*iss*	NA	NA	NA	-1.94	1.83	0.02
LF82_RS10120	*fliQ*	NA	NA	NA	-1.59	1.76	0.05
LF82_RS10985		NA	NA	NA	1.43	1.78	0.04
LF82_RS15515	*gspH*	NA	NA	NA	1.45	2.03	0.05
LF82_RS20025		NA	NA	NA	1.47	1.19	0.05
LF82_RS23260		NA	NA	NA	1.47	2.05	0.05
LF82_RS22825	*arcC*	NA	NA	NA	1.51	2.36	0.05
LF82_RS10865	*wcaB*	NA	NA	NA	1.53	1.73	0.03
LF82_RS21260		NA	NA	NA	1.57	3.00	0.02
LF82_RS25615		NA	NA	NA	1.63	0.65	0.04
LF82_RS13910	*ygaH*	NA	NA	NA	1.87	1.70	0.03
LF82_RS17105	*acrE*	NA	NA	NA	2.38	2.70	0.00
LF82_RS10065	*fliF*	-3.24	5.10	0.00	-1.36	4.23	0.06
LF82_RS06205		1.33	1.77	0.05	-0.44	2.51	0.47
LF82_RS16125	*ttdT*	1.35	4.94	0.02	0.75	4.32	0.21
LF82_RS10455		1.35	1.22	0.02	0.68	1.50	0.37
LF82_RS25875		1.39	1.48	0.05	NA	NA	NA
LF82_RS07840	*marB*	1.39	3.95	0.01	0.95	3.91	0.11

LF82_RS05470	*flgJ*	-2.46	3.82	0.01	-0.77	3.66	0.32
LF82_RS10110	*fliO*	-2.38	3.02	0.01	-1.52	2.34	0.06
LF82_RS05275		-2.20	2.36	0.01	-1.06	2.33	0.13
LF82_RS02225	*amtB*	-1.64	4.97	0.02	-0.81	4.67	0.20
LF82_RS01460	*ecpC*	-1.60	6.11	0.01	0.58	5.13	0.40
LF82_RS06685	*ompW*	-1.60	6.96	0.00	-1.23	8.48	0.12
LF82_RS21870	*dcuB*	-1.59	4.32	0.02	-0.26	3.24	0.69
LF82_RS19165		-1.55	2.77	0.03	0.22	2.74	0.74
LF82_RS20135		-1.53	1.26	0.01	0.16	3.52	0.84
LF82_RS06985		-1.52	1.65	0.00	-0.65	2.19	0.39
LF82_RS01580		-1.51	3.05	0.02	0.12	2.33	0.86
LF82_RS15520	*gspG*	-1.43	0.94	0.01	0.28	2.87	0.68
LF82_RS02890	*citT*	-1.42	4.55	0.02	-0.74	3.55	0.23
LF82_RS09785	*flhB*	-1.42	3.84	0.02	-0.76	3.29	0.23
LF82_RS02835	*hcxA*	-1.42	3.11	0.00	-0.90	3.57	0.14
LF82_RS11705	*atoD*	-1.40	2.99	0.03	0.61	2.34	0.34
LF82_RS15495	*gspL*	-1.39	1.75	0.01	1.19	2.56	0.07
LF82_RS09490	*mgrB*	-1.36	1.77	0.02	-1.12	6.15	0.13
LF82_RS21680	*alsE*	-1.30	4.78	0.01	-0.22	4.06	0.72
LF82_RS01270		-1.25	1.04	0.03	0.01	2.21	1.00
LF82_RS15650		-1.25	3.76	0.03	1.19	3.17	0.08
LF82_RS14910	*ygeW*	-1.22	4.71	0.05	0.63	3.22	0.29
LF82_RS13690		-1.21	0.74	0.04	0.50	1.48	0.49
LF82_RS10555	*pduB*	-1.20	2.46	0.04	-0.14	1.83	0.88
LF82_RS16170	*aer*	-1.19	6.05	0.03	-1.25	5.86	0.06
LF82_RS00545	*gspE*	-1.15	3.13	0.04	0.03	2.75	0.97
LF82_RS02220	*glnK*	-1.15	4.25	0.04	-0.19	5.52	0.78
LF82_RS12760	*eutQ*	-1.07	1.21	0.05	0.83	2.44	0.21
LF82_RS09215		-1.04	2.21	0.05	-0.68	4.55	0.30
LF82_RS08015		-1.01	3.69	0.04	-1.43	4.85	0.07
LF82_RS18365	*chuT*	1.05	3.19	0.05	0.03	3.23	0.97
LF82_RS02740	*entD*	1.11	2.73	0.03	0.19	2.66	0.75
LF82_RS06670	*yciA*	1.13	3.80	0.03	0.01	3.42	0.99
LF82_RS25330	*ypdK*	1.16	3.75	0.04	0.52	5.34	0.52
LF82_RS12585	*cysZ*	1.16	4.82	0.03	0.42	3.34	0.50
LF82_RS13870	*nrdI*	1.16	3.93	0.03	1.21	4.09	0.05
LF82_RS14295		1.18	4.89	0.03	0.36	5.50	0.57
LF82_RS08080	*cspI*	1.19	1.71	0.05	-0.76	2.38	0.23
LF82_RS00025		1.19	3.13	0.04	0.41	2.51	0.50
LF82_RS00655		1.20	3.32	0.01	-0.07	3.10	0.93
LF82_RS03750		1.23	2.88	0.01	0.16	2.33	0.80
LF82_RS08760		1.23	6.62	0.05	0.35	7.78	0.67
LF82_RS02755	*fes*	1.24	6.32	0.05	0.35	5.39	0.64
LF82_RS09560		1.26	3.21	0.04	0.86	3.44	0.21
LF82_RS09290		1.27	2.68	0.03	0.38	1.78	0.57
LF82_RS01985	*acpH*	1.29	3.82	0.01	0.07	3.16	0.92
LF82_RS23150		1.30	1.94	0.04	1.09	3.01	0.10
LF82_RS06205		1.33	1.77	0.05	-0.44	2.51	0.47
LF82_RS18305	*rsmJ*	1.60	5.01	0.02	0.23	4.12	0.70
LF82_RS01490		1.70	4.00	0.05	0.46	4.27	0.52
LF82_RS25900	*mgtL*	1.77	8.06	0.01	0.06	8.53	0.94
LF82_RS01695		1.78	1.54	0.01	0.80	1.43	0.25
LF82_RS02805	*entE*	1.84	6.21	0.01	1.00	6.47	0.15

LF82__RS25800	*pheL*	1.39	6.03	0.02	1.38	6.72	0.10
LF82__RS25670		1.41	1.84	0.01	0.07	4.34	0.92
LF82__RS08105	*cspB*	1.44	4.04	0.01	1.29	5.06	0.09
LF82__RS08620	*ynhF*	1.46	7.83	0.03	0.58	9.32	0.48
LF82__RS08450	*ydgT*	1.47	6.55	0.01	0.16	4.73	0.82
LF82__RS18325		1.88	1.53	0.02	-0.01	0.81	1.00
LF82__RS02810	*entB*	1.95	5.88	0.00	0.69	6.93	0.33
LF82__RS25265	*azuC*	1.96	5.40	0.00	0.73	6.96	0.36
LF82__RS24050	*yncL*	2.52	6.84	0.00	0.25	7.52	0.75

Supplementary Table 4. Number of predictions obtained by DeepRibo (Clauwaert *et al.*, 2019), REPARATION (Ndah *et al.*, 2017) and the scripts by Giess *et al.* (2017) for the Ribo-seq experiments in *E. coli* LF82 .

	DeepRibo	**REPARATION**	**Giess**
Exp1__aerobic__RepI	1,777	8,060	79,245
Exp1__anaerobic__RepI	2,045	7,994	74,929
Exp1__aerobic__RepII	2,830	8,760	72,402
Exp1__anaerobic__RepII	3,453	10,687	76,261
Exp2__ND	5,749	11,160	60,253
Exp3__RelE	1,324	4,990	86,322
Mean	2,863	8,608.5	74,902

Supplementary Table 5. Number and score of novel gene candidates identified in *P. aeruginosa* PAO1. All hits exceeding a prediction score of 8 or showing proteomic evidence with at least two mapping peptides are divided into the categories intergenic ORF (iORF), trivial OLG (OLG_TL), antisense embedded OLG (OLG_EA), sense embedded OLG (OLG_ES), partial antisense OLG with overlap at the 3´ (OLG_PA3) or at the or 5´ end (OLG_PA5) as well as partial sense OLG with overlap at the 3´ (OLG_PS3) or at the 5´ end (OLG_PS5), respectively. Hits matching to more than one non-trivial overlap type were classified as "multiple types".

Prediction score	**15**	**14**	**13**	**12**	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**3**	**2**
iORF				1	12	9	13		1	3	2		
OLG_TL	1	1	1		3	5	5	1	2		1		
OLG_EA						2	1						
OLG_ES				2			5						
OLG_PA3						1							
OLG_PA5							2				1		
OLG_PS3						2	3	1					
OLG_PS5		1			1	3	3						
multiple		1	1	1	2	8	16	1		2	1	1	1
sum	**1**	**3**	**2**	**4**	**18**	**30**	**48**	**3**	**3**	**5**	**5**	**1**	**1**

Supplementary Table 6. Expression metrics obtained for the overlapping genes *olg1* and *olg2*, their mother genes as well as adjacent genes in *P. aeruginosa* PAO1. Reads per million mapped reads (RPKM) and read coverage values were calculated for RNA-seq and Ribo-seq of Exp1 and averaged over biological replicates. Ribosome coverage values (RCVs) indicating translatability were calculated by dividing the RPKM values of the translatome by the RPKM values of the transcriptome.

Locus tag	**gene name**	**genome_start**	**genome_stop**	**strand**	**RPKM$_{\text{RNA-seq}}$**	**coverage$_{\text{RNA-seq}}$**	**RPKM$_{\text{Ribo-seq}}$**	**coverage$_{\text{Ribo-seq}}$**	**RCV**
PA0260	*tle3*	291154	293304	-	29.69	0.84	23.17	0.88	0.79
-	*olg1*	291556	292512	+	35.91	0.88	40.31	0.94	1.13
PA1383	NA	1501611	1503416	+	20.22	0.70	22.76	0.87	1.13
-	*olg2*	1501875	1503602	-	22.06	0.75	14.17	0.87	0.64
PA0261	*til3*	293301	293798	-	27.38	0.82	28.59	0.89	1.05
PA0262	*vgrG2b*	293802	296861	-	18.39	0.70	16.72	0.84	0.91
PA1384	*galE*	1503568	1504581	+	5.82	0.47	5.48	0.73	0.96
PA0259	*tla3*	289562	291004	-	67.22	0.94	46.72	0.98	0.69

Danksagung

An dieser Stelle möchte ich mich sehr herzlich bei meinem Doktorvater Herrn Prof. Dr. Siegfried Scherer für die Überlassung des Promotionsthemas, die hervorragende fachliche Unterstützung und das kontinuierliche Vertrauen, welches er mir entgegenbrachte und mir eine freie Gestaltung des Forschungsprojektes nach meinen Vorstellungen ermöglichte, bedanken.

Besonderer Dank gilt meinem Mentor Herrn Dr. Klaus Neuhaus für die großartige experimentelle Betreuung, seine grenzenlose und ansteckende Begeisterung für wissenschaftliche Fragestellungen und dass er stets ein offenes Ohr für mich und meine Anliegen hatte.

Ein herzliches Dankeschön auch an Herrn Prof. Dr. Hanno Schäfer für die Übernahme des Prüfungsvorsitzes sowie an Herrn Prof. Dr. Wolfgang Liebl für die zügige Erstellung des Zweitgutachtens.

Des Weiteren bedanke ich mich ganz besonders bei Frau Romy Wecko, die mich durch ihr fundiertes Wissen und ihre exzellente Arbeit im Labor bestmöglichst unterstützt hat. Vielen Dank auch an meine Studentinnen Thabea Stöhr, Ursula Rambold, Franziska Graf und Lisa Pütthoff für deren tatkräftige Unterstützung im Labor und die überaus angenehme Zusammenarbeit.

Bei meinen Arbeitskollegen Dr. Sonja Vanderhaegen, Dr. Barbara Zehentner, Stefan Wichmann, Dr. Alina Glaub, Anika Wahl und Franziska Graf möchte ich mich herzlich für die lustige und freundschaftliche Atmosphäre in unserem Büro bedanken. Mein weiterer Dank gilt Dr. Christopher Huptas für seine fortwährende Unterstützung bei informatischen Angelegenheiten und die stets aufmunternden Worte, sowie Dr. Zachary Ardern für seine Hilfe bei der Bearbeitung von evolutionären Fragestellungen. Allen weiteren Kolleginnen und Kollegen möchte ich meinen Dank für deren stetige Hilfsbereitschaft und für die schöne Zeit am Lehrstuhl für Mikrobielle Ökologie aussprechen.

Lieben Dank auch an Miriam Abele und Frau Dr. Christina Ludwig für die lehrreiche Zusammenarbeit, sowie an Dr. Philipp Heckmeier für seine Hilfe bei proteinbiochemischen Experimenten. Herrn Dr. Johannes Ehrmaier danke ich für die zahlreichen Nachhilfestunden in der Programmiersprache Python. Ferner möchte ich mich recht herzlich bei Frau Elizabeth Hamzi-Schmidt für das Korrekturlesen dieser Arbeit bedanken.

Mein größter Dank gilt meiner Familie, meinen Eltern Helga und Sebastian sowie meinen Schwestern Stephanie und Barbara, die mich in allen Lebenslagen -so auch bei der Durchführung dieser Arbeit- bedingungslos und unermüdlich unterstützt und dadurch ausschlaggebend zum Erfolg und der Fertigstellung dieser Arbeit beigetragen haben.

Curriculum Vitae

MICHAELA VERONIKA KREITMEIER

Date of birth: February 18, 1993

Place of birth: Moosburg a.d.Isar

Nationality: German

EDUCATION

since 03/2017	**Doctoral candidate** Chair of Microbial Ecology (Prof. Dr. Siegfried Scherer)/ Technical University of Munich Research topic: “Identification of overlapping genes in the human pathogens *Escherichia coli* LF82 and *Pseudomonas aeruginosa* PAO1 using transcriptomics, translatomics and proteomics”
10/2014 – 11/2016	**Master of Science** Molecular Biotechnology/Technical University of Munich Master thesis at the Chair of Microbial Ecology (Prof. Dr. Siegfried Scherer): “Analysis of the AprA peptidase production in *Pseudomonas* – correlation between peptidase activity and transcription of the *aprA-lipA* operon genes”
10/2011 – 09/2014	**Bachelor of Science** Molecular Biotechnology/Technical University of Munich Bachelor thesis at the Institute of Pharmacology and Toxicology (Prof. Dr. Stefan Engelhardt): “Analysis of the β1 adrenergic receptor phosphorylation”
09/2003 – 07/2011	**Abitur** Karl-Ritter-von-Frisch Gymnasium Moosburg a.d.Isar

Eidesstaatliche Erklärung

Ich erkläre an Eides statt, dass ich die bei der TUM School of Life Sciences der TUM zur Promotionsprüfung vorgelegte Arbeit mit dem Titel:

Identification of overlapping genes in the human pathogens
Escherichia coli **LF82 and** ***Pseudomonas aeruginosa*** **PAO1**
using transcriptomics, translatomics and proteomics

am Lehrstuhl für Mikrobielle Ökologie, ZIEL – Institute for Food & Health, unter der Anleitung und Betreuung durch Herrn Professor Doktor 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