TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Life Sciences

Lehrstuhl für Mikrobielle Ökologie

Improvements to ribosome profiling analysis in *E. coli* K-12 and diverse prokaryotes, and the detection of antisense overlapping genes

Alina Sophie Glaub

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. Dmitrij Frishman
Prüfer der Dissertation:	1.	Prof. Dr. Siegfried Scherer
	2.	Priv.-Doz. Dr. Klaus Neuhaus

Die Dissertation wurde am 08.02.2021 bei der Technischen Universität München eingereicht

und durch die TUM School of Life Sciences am 12.04.2021 angenommen.

Table of Content

Abbreviations

AA - amino acid

BHI - Brain-Heart-Infusion

BLAST - basic local alignment search tool

bp - base pair

C-terminus - carboxyl-terminus

cDNA - complementary DNA

CDS - coding sequence

CM - chloramphenicol

DNA - deoxyribonucleic acid

EHEC - enterohemorrhagic *Escherichia coli*

ENA - European Nucleotide Archive

eOLG - embedded overlapping gene

eORF - embedded open reading frame

GC - guanine cytosine

n - number of samples

NGS - next generation sequencing

MNase - micrococcal nuclease

mRNA - messenger RNA

N-terminus - amino-terminus

nt - nucleotide

OLG - overlapping gene

ON - overnight

ORF - open reading frame

PCR - polymerase chain reaction

RNA - ribonucleic acid

RPKM - reads per kilobase per million mapped reads

rpm - rounds per minute

rRNA - ribosomal RNA

RT - room temperature

sas - sense antisense

SD - Shine-Dalgarno

seq - sequencing

SIHUMI - simplified human intestinal microbiota

tRNA - transfer RNA

U - Units, a measurement for enzyme activity

UTR - untranslated region

Zusammenfassung

Die Existenz von überlappenden Genen (OLGs) die dadurch beschrieben werden, dass sie in einem alternativen Frame zu einem bereits existierenden Gen lokalisiert sind, wurde kontrovers diskutiert, da dies nicht mit dem „ein Gen - ein Protein“ Prinzip übereinstimmt. Deshalb wurde die Möglichkeit von mehr als einem protein-kodierenden Gen am selben Lokus, jedoch in verschiedenen Frames, in Prokaryoten weitestgehend nicht berücksichtigt, obwohl die Beschreibung von überlappenden Genen in Viren bereits seit 1976 von Virologen anerkannt ist. Lediglich vereinzelte OLGs wurden sowohl in Eukaryoten als auch Prokaryoten unter der Verwendung der Next-Generation-Sequencing Variante RIBO-Seq identifiziert.

RIBO-Seq ist eine neue Technik, die die Momentaufnahme des Translatoms mit der bestimmten Zuordnung von Reads zu einem der beiden genomischen Stränge ermöglicht. Dabei werden lediglich ribosomal geschützte mRNA Fragmente sequenziert, wodurch aktiv translatierte Gene zum Zeitpunkt der Zellernte aufgedeckt werden. Diese Arbeit fokussiert sich auf die Detektion, Evaluation und Verifikation von OLGs in Prokaryoten unter der Verwendung einer optimierten Analyse von RIBO-Seq Datensätzen.

Zuerst wurden neun bereits publizierte RIBO-Seq Datensätze von *E. coli* K12 verglichen, um eine spezifische Größenauswahl für ribosomal geschützte mRNA Fragmente zu bestimmen. Hierbei zeigten sich Fragmente mit einer Länge zwischen 24 bis 27 Nukleotiden als am informativsten. Zusätzlich zeigte eine Analyse im 5‘-UTR, dass für diesen größtenteils längere Reads (34 Nukleotide) detektiert wurden, wodurch verdeutlicht wird, dass die Anpassung der Fragmentlängen Auswahl an die experimentelle Fragestellung essenziell ist. Generell können bei einer Größenordnung zwischen 22 bis 30 Nukleotiden Protein kodierende Fragmente gewonnen, gleichzeitig aber größtenteils jene ausgeschlossen werden, die für rRNA oder tRNA kodieren. Zusätzlich verdeutlichte eine Analyse, dass mindestens 20 Millionen Reads notwendig für eine aussagekräftige Evaluierung von RIBO-Seq Experimenten sind. Hier nicht eingerechnet sind Reads, welche entweder rRNA oder tRNA zugeschrieben werden. In Anbetracht der verbesserten Detektion von bisweilen unbekannten, überlappenden Genen, zeigt sich die Zugabe von Chloramphenicol von Vorteil, vor allem für gering exprimierte Gene durch eine verdeutlichte Startpunkt Detektion.

Des Weiteren zeigte eine Analyse von 24 verschiedenen, prokaryotischen Spezies die Verteilung von eingelassenen OLGs im phylogenetischen Stammbaum. Vorhersagen der OLGs wurden mit dem veröffentlichten Detektionstool DeepRibo oder einem internen Skript durchgeführt. Lediglich wenige OLGs wurden mit beiden Programmen detektiert, vermutlich aufgrund von abweichenden Gen-spezifischen Merkmalen, welche diese für die Vorhersagen verwenden. Ausschlaggebend für die Verifikation vorhergesagter OLGs war deshalb eine Spezies-spezifische Analyse diverser Proben. Die somit erreichte Mehrfachbestimmung desselben OLGs wies auf potenzielle Authentizität hin, weshalb diese anschließend in Bezug auf ihre evolutionäre Entwicklung analysiert wurden. 43 eingebettete OLG

Sequenzen wurde identifiziert und daraufhin auf ihre Integration im Genom und Intaktheit analysiert. Das potenzielle Alter der Gene wurde auf Grundlage ihrer Häufigkeit in der Spezies-eigenen Familie bestimmt. Allerdings scheint ein Großteil der identifizierten Varianten jung und keinem negativen Selektionsdruck ausgesetzt zu sein. Zusätzlich wurden die Länge der OLGs analysiert, basierend darauf ob diese willkürlich entstanden sein können oder sie länger sind als erwartet, wodurch deren Funktionalität angenommen werden könnte. Des Weiteren konnte eine Länge zwischen 100 - 200 Nukleotiden für einen Großteil der detektierte OLGs gezeigt werden, während deren Lokalisation vermehrt im relativen Leserahmen sas12 vom überlappenden Gen zum korrespondierenden Mutter-Gen zu finden ist. Hierbei ist die 1. Codon Position des Mutter-Gens komplementär zu der 2. Position des überlappenden Gens und umgekehrt. Daraus resultiert eine entsprechende Überlagerung der 3. Codon Position für beide Gen Varianten.

Basierend auf den Ergebnissen dieser Kriterien konnten vier potenzielle Kandidaten identifiziert werden. Auch wenn keine statistisch signifikanten Ergebnisse erzielt wurden, zeigen sich Hinweise auf Funktionalität in wenigsten zwei der drei Analysen für diese OLGs. Deshalb können evolutionäre Analysen verwendet werden, um aus der Vielzahl der Vorhersagen erste Kandidaten für anschließende experimentelle Verifikation auszuwählen.

Aufgrund der in dieser Arbeit generierten Ergebnisse werden experimentelle Empfehlungen für die Durchführung von RIBO-Seq Experimenten ausgesprochen, die möglicherweise zu einer verbesserten Detektion von überlappenden Genen, deren Existenz im phylogenetischen Stammbaum nun ebenfalls bestätigt werden konnte, beitragen.

Abstract

The existence of overlapping genes (OLGs) characterised as being encoded by an alternative reading frame within an already existing gene has been a matter of controversy since it does not conform with 'one gene - one protein' principle. Therefore, the possibility that more than one protein-coding gene can be found at the same locus in different reading frames has not been considered in prokaryotes in any detail, although the description of overlapping genes in viruses, which started as early as 1976, has been accepted among virologists. Only occasionally OLGs were suggested to exist within eukaryotes and prokaryotes, mainly based on Next-Generation-Sequencing techniques, namely RIBO-Seq.

RIBO-Seq is a new technique allowing to visualise a general snapshot of the translatome with the distinct assignment of RNA-based reads to either of the two genomic strands. As only ribosome protected mRNA fragments are subjected to sequencing, genes that are actively translated at the time of harvest are revealed. This thesis focusses on the detection, evaluation, and verification of OLGs within prokaryotic genomes using an optimized analysis of RIBO-Seq data.

First, a total of nine publicly available RIBO-Seq experiments from *E. coli* K-12 were compared to identify an appropriate size selection range to obtain solely ribosomal protected mRNA fragments. Here, fragments between 24 to 27 nucleotides in length were identified to be most informative. Further, the importance of size selection was demonstrated as an analysis focused on the 5'-UTR regions of annotated genes showed a predominant coverage of longer fragments (34 nucleotides). Hence, an adaptation of the read length analysed to the goal of the analysis is crucial. In general, a range between 22 to 30 nucleotides covers protein-coding fragments while simultaneously excluding mostly those corresponding to rRNA/tRNA. A general analysis regarding sufficient read coverage reveals that at least 20 million reads, excluding rRNA and tRNA reads, are necessary for appropriate RIBO-Seq evaluation. With reference to non-annotated overlapping genes, the application of chloramphenicol might aid in their detection as it supports start site detection especially for weakly expressed genes.

Second, an analysis of 24 prokaryotic species revealed the presence of embedded OLGs throughout the phylogenetic tree. Predictions made were based on the publicly available tool DeepRibo and an in-house prediction script. However, few OLGs were detected by both approaches, probably due to the fact that both tools focus on different gene specific features. More suitable for the verification of OLGs was species-specific analysis of different samples. Here, re-occurring determination indicated potential authenticity of the OLGs, which were subjected to a first analysis focussing on sequence evolution. On this basis, 43 embedded OLGs were identified and their sequence was analysed according to potential functionality based on maintenance in the genome and integrity. The potential genes' age was estimated based on its abundance within the species' family. Most appeared to be 'young' and no significant negative selection was detectable. Furthermore, an analysis was performed regarding the OLGs length, whether this can be explained by random creation or is longer than expected, therefore implying functionality. Additionally, the predominant length for embedded OLGs predicted is between 100 - 200

nucleotides whilst the favoured location to maintain within the genome seems to be in relative reading frame sas12 to their mother gene. Here, the 1st codon position of the mother gene is complementary to the 2nd position of the overlapping gene and *vice versa*. Consequently, the 3rd codon positions are corresponding to each other for the mother gene and overlap.

Based on these criteria, four OLG candidates of interest were identified. Though not statistically significant, those OLGs showed indications for functionality based on at least two of the three analyses performed. Thus, evolutionary analysis can be used to narrow down the number of potential OLGs to those of special interest which can then be subjected to further experimental verification.

Based on the results of this dissertation, recommendations for future RIBO-Seq experiments are given that may contribute to OLG detection, whose existence is now confirmed throughout the phylogenetic tree.

1. Introduction

1.1 Overlapping genes – fallacy or reality?

1.1.1 History of overlapping genes (OLGs)

The definition of what can be described as a gene has changed over the past 150 years (Portin & Wilkins, 2017). What defines a gene, in general, is the nucleotide structure it is based on. A seemingly random sequence consisting of bases adenine, cytosine, guanine, and thymine, attached via a sugar (deoxyribose) and phosphate backbone, forms the foundation of a functional gene (Watson & Crick, 1953). It was previously believed that genomes, no matter whether pro- or eukaryotic, are packed with genes one after the other, with some genes in such close distance, forming operon structures, resulting in both of them being transcribed at the same time (Yanofsky & Lennox, 1959). This assumption implies that one gene at a specific location in the genome stores the information for exactly one transcribed mRNA, resulting in one protein after translation. Besides factors such as alternative promotors, splicing variants and post-translational modifications, all evidence against the 'one gene - one protein' hypothesis (Hickman & Cairns, 2003; Portin & Wilkins, 2017), another counterexample is the detection of overlapping genes (OLGs).

The first overlapping genes were found in the genome of bacteriophage ΦX174 by Barrell et al. in 1976. Way back then, the discrepancy between the bacteriophages' genome size and the nucleotides needed to code for the identified proteins led to an analysis of the genome sequence itself. A first explanation for the hitherto unknown phenomenon of potential functional genes located in alternative frames and thereby overlapping already identified genes was the genome 'compression theory' (Belshaw, Pybus, & Rambaut, 2007; Brandes & Linial, 2016; Chirico, Vianelli, & Belshaw, 2010). Due to the bacteriophages' small genome size and little space left between known genes, any additional genes are required to occur in an alternative frame (Scherbakov & Garber, 2000). However, further detections of overlapping genes not only in viruses (Cassan, Arigon-Chifolleau, Mesnard, Gross, & Gascuel, 2016; Fernandes et al., 2016; Nelson, Ardern, Goldberg, et al., 2020) but also in prokaryotes (Hücker, Vanderhaeghen, Abellan-Schneyder, Scherer, & Neuhaus, 2018; Hücker, Vanderhaeghen, Abellan-Schneyder, Wecko, et al., 2018; Vanderhaeghen, Zehentner, Scherer, Neuhaus, & Ardern, 2018; Zehentner, Ardern, Kreitmeier, Scherer, & Neuhaus, 2020), plants (Terryn & Rouze, 2000), fruit flies (Henikoff, Keene, Fechtel, & Fristrom, 1986; Spencer, Gietz, & Hodgetts, 1986), mice (Williams & Fried, 1986) and even humans (Nakayama, Asai, Takahashi, Maekawa, & Kasama, 2007) are discarding the compression theory as OLGs can be detected in genomes spanning sizes from 4 Mb (Lim, Yoon, & Hovde, 2010) up to 6.2 Gb (Piovesan et al., 2019). Yet, they remained undetected due to technical limitations and their own unusual properties for quite a while. While the development of Next Generation Sequencing has contributed to the detection of new genes, overlapping ones have remained out of favour. A limiting factor in sequencing approaches was the problem that reads are most likely mapped to the open reading frame (ORF) of an annotated gene as the differentiation between frames

was not taken into account. Second, even if detected, the ORF of an OLG is often discarded as non-functional due to its short length (Sberro et al., 2019; Storz, Wolf, & Ramamurthi, 2014; Su, Ling, Yu, Wu, & Xiao, 2013). With the development of ribosome profiling (see Section 1.2.1.) the detection of overlapping genes is now possible. Still, the question of their purpose and their development, especially in prokaryotes, remains.

1.1.2 Mechanism of gene development: are overlapping genes a source of new genes through overprinting?

There are two potential explanations for the origin of new genes. One is through the modification of present genes, whereas the other is through *de novo* gene birth (Pavesi, Magiorkinis, & Karlin, 2013; Sabath, Wagner, & Karlin, 2012; Taylor & Raes, 2004). Gene fusion or transposition are just examples of possible modifications (Long, Betran, Thornton, & Wang, 2003; Q. Zhou & Wang, 2008). The relocation of an already functional gene into an alternative frame overlapping another gene is possible but highly unlikely. The nucleotide sequence change caused by such a modification would highly affect the sequence of the existing gene and more likely result in its loss of function. Therefore, this mechanism does not seem to be causing OLG development.

Overprinting, on the other hand, could explain the origin of OLGs (Hücker, Vanderhaeghen, Abellan-Schneyder, Wecko, et al., 2018; Keese & Gibbs, 1992; Pavesi et al., 2018; Rancurel, Khosravi, Dunker, Romero, & Karlin, 2009). Nucleotide sequences coding for annotated genes are quite strongly conserved as their order is the template for the resulting protein. The order of nucleotides is translated into a polypeptide sequence based on triplets, namely codons. Within a codon, the positions are of different importance, with the second position being the most decisive one for the incorporated amino acid characteristics whereas the third position is mostly inessential (Blazej, Wnetrzak, Mackiewicz, & Mackiewicz, 2018; Massey, 2006; Saier, 2019). Changes in the third codon position can oftentimes be described as synonymous mutation, defined by a nucleotide change that is still coding for the same amino acid. This factor can be described as codon degeneracy, as several codons are translated into the same amino acid (Gonzalez, Giannerini, & Rosa, 2019; Plotkin & Kudla, 2011). This synonymous mutation in the mother gene however could cause a nonsynonymous mutation in an alternative frame, leading to amino acid changes in any protein encoded

A

UUU, UUC: Phe UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr UAA, UAG: Stop	UGU, UGC: Cys UGA: Stop UGG: Trp
CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg
AUU, AUC, AUA: Ile AUG: Met	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn AAA, AAG: Lys	AGU, AGC: Ser AGA, AGG: Arg
GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly

B

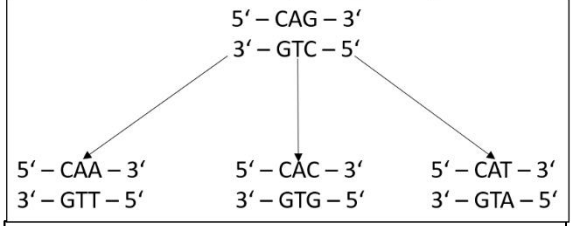

Figure 1: (A) Codon table from (Esberg, 2007), (B) Potential variations in triplets 3rd position with caused changes in both strands shown.

in the alternate frame. Additionally, it is known that the third codon position evolves faster than any of the other two, for which lower selective pressure is presumably responsible (Bofkin & Goldman, 2007). This redundancy may facilitate the development of overlapping genes.

For instance, if glutamine (codon triplet CAG) is incorporated into the mother gene, a synonymous mutation to CAA entails a synonymous mutation in the complementary strand from triplet CTG to TTG, both coding for leucine. It has been shown that TTG functions as a start codon in prokaryotes, more likely used in *Bacillus* (Belinky, Rogozin, & Koonin, 2017; Hecht et al., 2017). Additionally, a mutation in the third position of leucine, coded by CAA, could also result in an exchange to triplet CAC or CAT, both coding for more effective start codons GTG or ATG in the opposite strand. The nucleotide changes caused by mutation and the resulting changes of amino acid triplets are shown in Figure 1 accompanied by the codon table.

The introduction of a start codon could lead to a new open reading frame (ORF) in an alternative frame which could also be achieved by an emerging stop codon after an already-existing start. Here, a nucleotide change in the third codon position within the mother gene could cause the incorporation of a stop codon triplet. Additionally, depending on the localization of the OLG in an alternative frame, nucleotide exchanges in the third codon position of the mother gene can lead to base replacement in the second triplet position of the OLG. As already mentioned, the second codon position is the most relevant for the amino acid determination (Bofkin & Goldman, 2007). The combination, where the third position of the mother gene is affecting the second position in the OLG can emerge if for example the mother gene is located in frame +1 and the OLG in frame -3 (see Figure 2A, relative reading frame sas11). There are far more possibilities present in the genetic code allowing one mutation in the second position of a codon to change it into either a start or a stop codon. Due to this fact and the importance of the position itself regarding the determination of the amino acid characteristics, one analysis in this thesis focusses on the reading frame of detected OLGs in several prokaryotic species. Additionally, this analysis sheds light on whether OLGs can be found in a variety of prokaryotic species distributed over the phylogenetic tree.

To maintain functionality based on sequence integrity not only in the mother gene but also the OLG, the locus coding for both is expected to be under higher evolutionary constraint than typical non-overlapping genes. Mentioned above are just the favourable mutations that could lead to the occurrence of a second ORF at an already ‘occupied’ locus. However, another mutation within a locus coding for two gene variants (mother gene and OLG) can directly affect one or both negatively, resulting in loss of functionality for one or both. A potential explanation of only a few OLGs detected might be their integration into an intergenic region of the genome, through ‘copying out’ after a relatively short time. In this way, the selective pressure on the locus primarily coding for both would be reduced and their function can be maintained independently. Yet another possible role of genes overlapping existing genes in an alternative frame might be to function as a regulator of the mother gene (Boi, Solda, & Tenchini,

2004; Yelin et al., 2003). In this case, the complementary parts of transcribed RNAs form complexes through bonding, thereby blocking the active centre of the protein, thus regulating potential bonding (Jen, Michalopoulos, Westhead, & Meyer, 2005; Kiyosawa et al., 2003). This, however, has been shown only in higher eukaryotes, whereas this study is focussing on independently functional OLGs and their characteristics in various prokaryotes.

1.1.3 Characteristics of potential OLGs

An overlapping gene is described as a protein-coding ORF located in an alternate reading frame to another gene which is most likely already annotated. Within this explanation, two variations need further clarification: first the type of overlap and second the location in relation to the annotated gene.

The differentiation between types of overlaps is based on the length of the overlapping part. If the OLG is fully embedded within the mother gene, meaning the start and stop codon are located within the borders of the mother gene, it is called an embedded ORF (eORF). A partial antisense ORF (paORF) however, is only overlapping a part of the mother gene, either the N- or C-terminus. Overlaps ≥ 90 bp located in an alternative frame antisense to the mother gene are called non-trivial overlaps (Brandes & Linial, 2016; Vanderhaeghen et al., 2018; Zehentner et al., 2020), contrary to trivial overlaps of only a few base pairs, which for instance, enable transcriptional coupling in operon like structures (Eyre-Walker, 1995; Price, Arkin, & Alm, 2006). Even though the existence of non-trivial OLGs has recently been shown in bacteria (Fellner et al., 2014; Fellner et al., 2015; Zehentner et al., 2020), difficulties in their detection remain. While the development of ribosomal profiling (see Section 1.2.) assists OLG detection based on monitoring the translational status at the point of harvest, the resolution of prokaryotic sense overlaps in alternative frames is still challenging. The codon periodicity, enabling the exact mapping of sequencing reads, thus revealing the frame location, is precise while using, for example, RNase I in *Saccharomyces cerevisiae* (Jackson & Standart, 2015; Mohammad, Green, & Buskirk, 2019). However, when analysing prokaryotic ribosome profiling results, there are still problems assigning reads to different frames. The difficulties result from digestion enzymes used lacking precise cleaving sites (Glaub, Huptas, Neuhaus, & Ardern, 2020) resulting in nearly evenly distributed mapping of reads to all three codon positions (Gelsinger et al., 2020; Mohammad, Woolstenhulme, Green, & Buskirk, 2016). Additionally, due to their mostly shorter length, OLGs were oftentimes overlooked in sequencing approaches. ORFs shorter than 100 - 200 bp have until recently only rarely been associated with a function, and are also more difficult to exclude from ORFs present merely by chance, thus were excluded from analysis (Olexiouk, Van Criekinge, & Menschaert, 2018; Warren, Archuleta, Feng, & Setunal, 2010).

The second type of variation within OLG characterisation is their location in relation to the mother gene (pre-existing annotated gene). As detection in an alternative frame on the same strand is still difficult, only the relation of the mother gene and antisense OLG will be explained as it is the focus of the rest of the thesis. Defining the mother gene's reading frame as +1, the overlapping codon position of the OLGs can be used to classify the type of overlap. If the third codon position of the OLG is at the same sequence position as the first position of the mother gene, the classification is sense 1 antisense 3 localisations (sas13 relative reading frame). Here, the sense information is referring to the mother gene, whereas antisense is describing the OLG. Similarly, the relation can be described with sas12 or sas11 (Nelson, Ardern, & Wei, 2020; Wei & Zhang, 2014). For clarification the relative reading frames are shown in Figure 2A, complemented with the two types of overlap possible mentioned above. One question considered within this thesis is if one frame favours the development of OLGs.

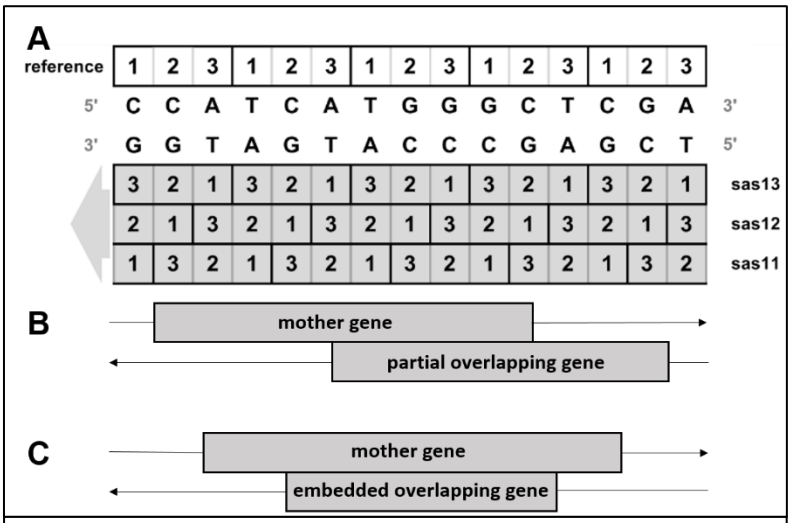

Figure 2: (A) Explanation of the relative reading frame location of a reference sequence to one of the three potential other frames on the opposite strand. Figure from (Nelson, Ardern, & Wei, 2020). (B) Visualised is a partial antisense overlap (C) Visualised is an embedded antisense overlap.

Another factor potentially influencing OLG development are the nucleotides used and the resulting codons. It has been demonstrated by Pavesi et al (2020) that in viral genomes overlapping genes are predominantly composed of dinucleotides containing one or two cytosines, whereas they have a relative lack of dinucleotide combinations of adenine and thymine. Based on these results a hypothesis can be set up stating that GC-content rich prokaryotes are more prone to the development of overlapping genes. Contradictory to this, however, the three commonly used stop codons (TAA, TGA, TAG) (Belinky, Babenko, Rogozin, & Koonin, 2018; Korkmaz, Holm, Wiens, & Sanyal, 2014; Povolotskaya, Kondrashov, Ledda, & Vlasov, 2012) consist of at least one adenine and thymine. Within high GC-content genomes these nucleotides occur less which could potentially influence the length of new ORFs in general. Additionally, results from Miravet-Verde et al. (2019) support the latter statement, reporting a connection between lower GC-content genomes containing more small proteins. This is especially of interest, as OLGs can be categorised as short genes mostly characterised due to their short length (≤ 100 amino acids) (Basrai, Hieter, & Boeke, 1997; Su et al., 2013). As both of these hypotheses are reported, one approach in this study is to compare the prediction efficiency for eORFs in low (*Staphylococcus aureus subsp. aureus* NTC8325; *Bacillus subtilis subsp. subtilis str.* 168) and high (*Pseudomonas fluorescens* F113; *Streptomyces coelicolor* A3) GC-content genomes. Is a difference detectable in the number of predicted eORFs and if so, can a nucleotide usage trend be seen in the OLGs

sequences? Furthermore, is there a difference in observed eORF length based on genome size? In the selected prokaryotes higher genome size is correlated with increased GC-content. Based on less adenine and thymine nucleotides present in these GC-rich genomes a resulting minimisation of stop codons could benefit potential for longer ORFs. All analyses made for this study are based on the performance of ribosomal profiling experiments (RIBO-Seq) in the lab. This technique and the history of next generation sequencing are the topics of the following chapter.

1.2 Ribosomal profiling as a technique to detect OLGs

1.2.1 History of next generation sequencing (NGS)

The development and improvement of nucleotide sequencing techniques span the past 50 years (Heather & Chain, 2016). Modern sequencing began in 1977 with base-per-base sequencing based on the Sanger method, where the incorporation of a new base each time stops the reaction (Sanger, Air, et al., 1977; Sanger, Nicklen, & Coulson, 1977). This was followed by the development of second generation sequencing, which can detect an attached base by generated fluorescence (J. Shendure & Ji, 2008). In recently developed third-generation sequencing methods only single molecules are necessary to perform the whole experiment (Ardui, Ameur, Vermeesch, & Hestand, 2018; Heather & Chain, 2016; Jenjaroenpun et al., 2018). As the sequencing methods have developed over the years so has the specificity of what can be found with the approach. Although methods began with whole genome shotgun sequencing approaches to unravel DNA sequences of eukaryotic and prokaryotic organisms for analysing their genome sequence, now the focus is shifting to resequencing already known species in hopes of detecting genomic variations within a species (J. A. Shendure et al., 2011). It is also now possible to characterise the transcriptional or translational status of a cell using its RNA (Ozsolak & Milos, 2011).

Transcriptome analysis started with microarray experiments in the mid-1990s (Lockhart et al., 1996; Marinov, 2017; Schena, Shalon, Davis, & Brown, 1995) but with the development of DNA-sequencing approaches, the first experiments using complementary DNA (cDNA) from RNA sequences as the input source were conducted in 2008, resulting in less noisy and more reliable data (Cloonan et al., 2008; Marinov, 2017; Nagalakshmi et al., 2008). The transcriptome of a cell is highly dependent on environmental influences or developmental stages and in multicellular eukaryotes is specific for different tissues (Qian, Ba, Zhuang, & Zhong, 2014; Z. Wang, Gerstein, & Snyder, 2009). Consequently, to explore this variability in transcriptomes many different sequencing methods for RNA-analysis arose.

Some methods are specific to eukaryotes; for instance, RNA exome sequencing which unveils just the protein-coding sequences and their variants, leaving out introns and intergenic sequences (Ng et al., 2010). Other methods are also applicable to prokaryotes, for instance, single-cell RNA-sequencing (scRNA-seq) which allows the analysis of the transcriptome of one cell at a time (Hagemann-Jensen,

Abdullayev, Sandberg, & Faridani, 2018; Imdahl, Vafadarnejad, Homberger, Saliba, & Vogel, 2020; Stegle, Teichmann, & Marioni, 2015; Svensson et al., 2017), or small RNA-sequencing (sRNA-seq) that enables detection of small non-coding RNAs (ncRNAs) such as micro RNA (miRNA) which are otherwise difficult to detect (Costa, Angelini, De Feis, & Ciccodicola, 2010; Qian et al., 2014; Raghavan, Groisman, & Ochman, 2011; Shinhara et al., 2011; Wu et al., 2017). With cappable-seq, the detection of the transcription start site is possible due to modification and enrichment of the 5'-end of transcribed RNA (Ettwiller, Buswell, Yigit, & Schildkraut, 2016; Zehentner et al., 2020). The development of ribosomal profiling allows capturing the actual status of translation (translatome) at the point of harvest. This in comparison to RNA-Seq capturing the transcriptional status of an organism allows first comments on whether an RNA sequence is coding for a translated and therefore potentially functional protein (Ingolia et al., 2014; Ingolia, Ghaemmaghami, Newman, & Weissman, 2009). The critical experimental proceedings during ribosomal profiling are described beloow.

1.2.2 Ribosomal profiling and its determining factors

RIBO-Seq, as a method to detect the actual status of translation (translatome) at the point of harvest, was first established in *S. cerevisiae* in 2009 by Ingolia et al. (Ingolia et al., 2009). Since then, the method was adapted and applied to successfully analyse other organisms, such as mammalian stem cells, bacteriophage lambda, *Drosophila melanogaster*, *E. coli* K12 MG1655, and its pathogenic relative *E. coli* O157: H7 strain Sakai (Dunn, Foo, Belletier, Gavis, & Weissman, 2013; Hücker, Ardern, et al., 2017; Ingolia, Lareau, & Weissman, 2011; Li, Burkhardt, Gross, & Weissman, 2014; Liu, Jiang, Gu, & Roberts, 2013). The workflow starts with ribosomal stalling followed by cell harvest and lysis. Captured RNA is digested and separated according to its molecular weight with gradient density centrifugation subsequently followed by fragment size selection in an urea gel. rRNA depletion should minimise the remaining rRNA present before library preparation is completing the experimental procedures followed by sequencing (Ingolia, 2010; Ingolia, Brar, Rouskin, McGeachy, & Weissman, 2012). A brief overview of these steps is shown in Figure 3.

Stalling ribosomes efficiently during translation is a crucial factor in the experimental protocol, ensuring the protection of mRNA fragments, so-called footprints, of interest. The haltered translation can be achieved by several methods all having slightly different benefits and drawbacks. One possibility is stalling induced by rapid freezing using liquid nitrogen (Ingolia, 2016) which's effectiveness is highly dependent on working efficiency. Nevertheless, this method is still recommended as it is not causing translational adaptation due to environmental changes or introducing read artefacts (Glaub et al., 2020; Mohammad et al., 2019). These can be detected by drug application, such as chloramphenicol (Cm), retapamulin (Ret), or tetracycline (Tet). An artefact detected is the ribosomal accumulation at the translational start site caused by inhibited elongation (Meydan et al., 2019; Mohammad et al., 2019;

Nakahigashi et al., 2016). Then again, this bias can be advantageous in start site detection emphasizing its location by ribosomal accumulation (Meydan et al., 2019; Nakahigashi et al., 2016).

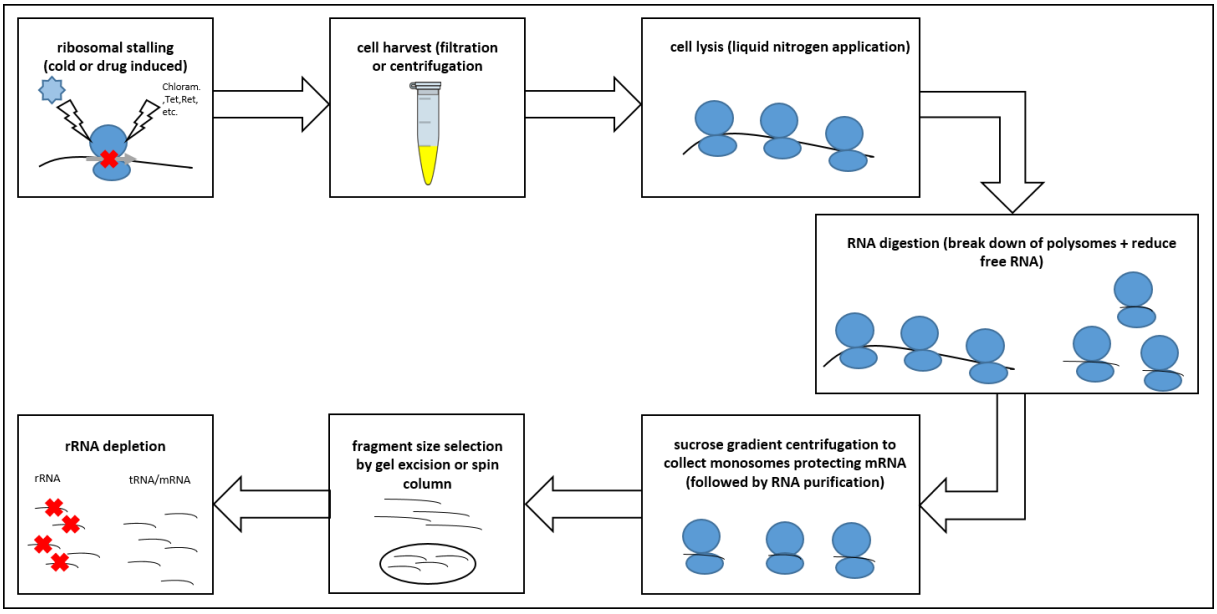

Figure 3: Overview of important steps performed in RIBO-Seq experiments. Figure obtained from (Glaub et al., 2020).

Also noteworthy is the appropriate use of RNase for the digestion of accessible RNA in the organism. Ribosomal profiling aims to only sequence ribosomal protected footprints, therefore, tRNA transporting amino acids or rRNA building new ribosomes needs to be digested. However, this is also causing the breakdown of polysome structures, characterised by several ribosomes occupying one transcript during translation, into separate ribosomes protecting parts of the mRNA, so-called monosomes. Commonly used in RIBO-Seq experiments is RNase I as it not only shows uniformly produced footprints in length but also enables high resolution reading frame detection in *S. cerevisiae* (Ingolia, 2010, 2016). However, this endoribonuclease seems unsuitable for prokaryotic experiments, especially in *E. coli*, as its 30S ribosomal subunit inhibits the enzymatic activity by actively binding it (Kitahara & Miyazaki, 2011; Zhu, Gangopadhyay, Padmanabha, & Deutscher, 1990). In this case, micrococcus nuclease (MNase) is used even though it is known to have a cleavage bias. Important for RIBO-Seq experiments is the complete digestion of any unprotected RNA up to the ribosome itself. However, as micrococcus is favouring cleavage in adenine and thymine rich regions (Dingwall, Lomonossoff, & Laskey, 1981) parts of unprotected mRNA at the boundaries of the ribosome may remain resulting in different footprint length. Thereby it is more difficult to detect a clear reading frame when MNase is used. The detection of the reading frame is dependent on the length of protected mRNA footprints and their distance relation to positions within the ribosome. If a periodicity of reads is detectable, the actual reading frame can be assigned. This seems to be more difficult in bacteria as the sequence quality is noisy, due to the different read length caused by insufficient digestion. One solution for this problem might be the use of RelE for digestion, as it precisely cleaves protected mRNA after the second nucleotide in the ribosomal A-site (Hwang & Buskirk, 2017; Pedersen et al., 2003). Another option is to mix endo- and exonucleases to

increase digestion efficiency while decreasing sequence specificity simultaneously (Hücker, Ardern, et al., 2017).

Another important aspect is the appropriate size selection of protected footprints during sequencing preparation. After digestion, samples are loaded onto a TBE-urea gel and bands containing fragments of interest are excised according to their length. This is a second step ensuring only protected footprints are used for further processing. The assumption that only protected fragments are collected from the gel is based on their size due to the limited length that can be protected by the ribosome during translation. For *S. cerevisiae* fragments of 28 to 30 nucleotides are expected after digestion, therefore gel excision will be performed according to this size (Ingolia, 2010; Ingolia et al., 2012; Ingolia et al., 2009). However, for prokaryotic experiments, the determination of a specific protected fragment length is more difficult as mentioned above. Therefore, the gel-based size selection varies more spanning ranges of 15 to 40 nucleotides, as well as aiming for fragments of 23 nucleotides in length (Burkhardt et al., 2017; Buskirk & Green, 2017; Hücker, Simon, Scherer, & Neuhaus, 2017; Li et al., 2014; Mohammad et al., 2019). A broader range is claimed to be appropriate as it covers every varying read length but narrowing the range might be more effective as it excludes reads corresponding to rRNA or tRNA. However, it is necessary to adjust the size selection range depending on the experimental aim as longer reads were found associated with 5'-UTR regions (Glaub et al., 2020).

Furthermore, rRNA depletion is another crucial step in the performance efficiency of ribosomal profiling experiments. With an amount of up to 85 - 90 % rRNA is the most prevalent type of RNA in any cell (Z. Chen & Duan, 2011; Petrova, Garcia-Alcalde, Zampaloni, & Sauer, 2017). Thus, if not depleted, RNA based approaches will result in reads covering nearly exclusively rRNA. Some depletion kits contain unique probes complementary to the targeting sequences of 16S and 23S rRNA. They are bound by covalent binding and subsequently extracted from the sample by magnetic interaction (Petrova et al., 2017). RiboZero, Illumina's depletion kit, was commonly used due to its high efficiency of rRNA reduction. Despite its success, it is no longer available resulting in the improvement and development of new kits such as RiboMinus (ThermoFisher) or Pan-riboPOOLs (siTools).

Other experiments, where depletion was neglected, need high read amounts for successful approaches. 5 to 10 million reads are claimed to be sufficient for RNA-Seq experiments whereas for RIBO-Seq at least 20 million reads are recommended (Glaub et al., 2020; Haas, Chin, Nusbaum, Birren, & Livny, 2012). These numbers are referring to the read amount passing quality control, trimming, and alignment after sequencing. If 20 million reads are covering the 5% of input reads characterised as non rRNA sequences, the starting point for sequencing coverage would be at around 400 million reads for sufficient coverage implying no pooling of samples per sequencing run. Therefore, the use of a depletion kit is a cost-efficient alternative in rRNA reduction necessary for experimental proceedings.

The combination of RIBO-Seq results with the analysis of the transcriptome (RNA-Seq) can lead to the identification of actually functional proteins. However, improvements in RIBO-Seq approaches, both during the experiment and the bioinformatic evaluation can aid in obtaining more indicative results.

1.2.3 Computational evaluation of RIBO-Seq results

Despite the influence of experimental work, the appropriate evaluation of sequencing data is similarly important. Raw reads generated during sequencing undergo different bioinformatical steps to increase their contribution to evaluation. Quality of raw reads is analysed with FastQC by inspection of read amount, sequence length distribution, adapter content present, and potential overrepresented sequences. Next, the remaining adapter, necessary for flow cell attachment and sample assignment, is trimmed at the 3'-read ends with fastp. Fastp is outperforming other tools such as Cutadapt or Trimmomatic due to its ability not only to automatically detect adapter content present in the sample but its immediate removal (S. Chen, Zhou, Chen, & Gu, 2018). Trimming of adapter sequence is especially necessary if the alignment of reads to an appropriate reference genome is performed in 'end-to-end' mode. In this case nucleotides on both read edges must be aligned to the reference sequence, whereas in 'local' mode only nucleotides on one edge have to map to the reference genome (Langmead & Salzberg, 2012). In the processing pipeline for the evaluation of all RIBO-Seq samples in these experiments, Bowtie2 was used for alignment. It was first developed for fast and accurate alignment of short reads ($\leq$ 50bp), while in the upgraded version parameters such as allowed mismatches can improve the alignment rate even further (Langmead, 2010; Langmead & Salzberg, 2012). After subsequent quality control of trimmed and aligned reads, the actual evaluation of the translatome can be performed.

A crucial factor for the evaluation efficiency of sequencing experiments, in general, is the number of reads left after read adjustments, the so-called read depth. In RNA-Seq experiments, a read depth between 5 - 10 million fragments, depleted of rRNA mapping ones, is considered sufficient to detect even low expressed genes (Haas et al., 2012). Nevertheless, the appropriate amount needed is highly dependent on the experimental procedure itself. If the focus lies on detecting primarily low expressed genes, an adjustment of the necessary read depth should be considered to guarantee sufficient coverage. In RIBO-Seq experiments the detection of highly expressed genes is in favour due to their mRNA abundance in the cell present that can be used for translation. Hence, low expressed genes are less prone to be translated. As most overlapping genes remained undetected so far, due to technology limitations and maybe lack of interest, their characterisation and therefore their potential function is unresolved. As their detection is now possible with RIBO-Seq experiments, a comparison of read depth and the amount of predicted OLGs should shed light on the number of reads necessary to enable their detection in general. The OLGs predicted in this analysis are based on the results obtained from the ribosome profiling assisted (re-) annotation (REPARATION) tool. Here, ORFs are predicted based on extracting read patterns from annotated genes in RIBO-Seq data, such as read accumulation at the start and stop

region (Ndah et al., 2017). Additionally, ORFs can be made in each of the possible six reading frames, whereas the detection of genes overlapping annotated ones is enabled. DeepRibo is another tool of interest, as it also makes use of ribosomal profiling data and additional binding site patterns to delineate open reading frames (Clauwaert, Menschaert, & Waegeman, 2019). The third option of ORF prediction is an in-house script that detects every possible open reading frame just based on the start and stop codons present. In contrast to other prediction tools it uses 11 different start codons, all summarized in the standard codon Table 11 for bacteria from NCBI. Thus, as even rare codons are used for prediction, the amount of ORFs predicted is expected to exceed the efficiency of the previously mentioned tools. Additional filtering according to thresholds such as reads per kilobase million (RPKM), coverage and length is necessary to distinguish between potentially translated ORFs and background noise. Results from each of the three different prediction techniques are then filtered for ORFs either fully located within the boundaries or partially overlapping on edge of an annotated gene but in an alternative frame.

Another analysis of special interest is the detection of the ribosome protected footprint length in prokaryotes. As mentioned, for eukaryotes the range of informative mRNA is between 28 and 30 nucleotides (Ingolia, 2010, 2014). However, for prokaryotes, no definite length has been agreed on so far. To improve RIBO-Seq experiments based on computational evaluation one aim is to analyse the length of reads mapping to mRNA sequence. Therefore, publicly available RIBO-Seq experiments performed with *E. coli* K12 are being analysed according to their length. Besides mRNA fragment length reads mapping to either rRNA or tRNA are also subjected to this type of analysis to reveal potential RNA type-specific length. If a relation between RNA type and length can be detected, this can be of interest for improvements in rRNA depletion by depleting specific length already in gel excision. However, the excision step is still highly discussed without any length unification so far for prokaryotic RIBO-Seq experiments. A comparison between the read length distribution chosen for gel excision and the actual obtained length variation after sequencing is included to potentially improve the success of RIBO-Seq experiments.

Furthermore, the last analysis regarding read length is focused on the upstream region of genes. Reads including the Shine-Dalgarno Sequence, a sequence binding motive for the ribosomal subunit, are longer. Their range between 28 - 40 nt emphasises the claim for a broader spectrum of read length in gel excision (Buskirk & Green, 2017; Li, Oh, & Weissmann, 2012). However, this could also enrich for sequences that are not of interest. Therefore, a sequence range of 25 nucleotides upstream of the start codon is analysed to detect a potential corresponding read length for regions covering the SD sequence. As this part of the analysis is partly focussing on the genes' associate start site, one more analysis is performed regarding this specific sequence part. Ribosomal stalling can be induced by drug application, which also can lead to read accumulation at the translation start site (Meydan et al., 2019; Mohammad et al., 2019). To take advantage of this, the last analysis elucidates whether the bias might also facilitate the detection of overlapping genes.

After analyses primarily focused on the improvements for the performance of RIBO-Seq experiments, the actual detection of overlapping genes and their distribution throughout the phylogenetic tree is subject of the following chapter.

1.2.4 RIBO-Seq comparison within different bacterial species: are their specific features contributing to different results?

Though ultimately presumed to be derived from one common ancestor, prokaryotes have evolved differently, i.e. in their cell wall construction which, separates them into two different groups (Errington, 2013). Within these groups, further separations take place based on genome rearrangements, which lead to various genome sizes present across the prokaryotes' phylogenetic tree. Not only the size of the genome but also its construction differs across species, as the content of guanine and cytosine (GC-content) present in the genome sequence varies throughout the different genera. This fact might be in favour for the construction of longer ORFs as the reduced amount of adenine and thymine present might be hindering in stop codon creation as they are mostly constructed out of pyrimidine bases (Korkmaz et al., 2014). A comparison of predicted eORF lengths across different species can shed light on this question after answering the general question of whether eORFs are found equally in various genera distributed across the phylogenetic tree. A broad spectrum of datasets covering eubacterial experiments ensures differences in genome size and GC-content. Two added archaeal sets allow for additional comparisons across characteristics more broadly. All eubacterial species selected for these analyses are shown in Figure 4.

A major difference between the species selected in general is the genomic construction, for instance reflected in the GC-content. Defined as the sum of guanine and cytosine (in on strand) divided by the total sequence length, variations from 20 - 70 % GC-content are reported for bacterial genomes (Bohlin, Eldholm, Pettersson, Brynildsrud, & Snipen, 2017; Hildebrand, Meyer, & Eyre-Walker, 2010; H. Q. Zhou, Ning, Zhang, & Guo, 2014). The GC-content present might be influential in the eORF length. As the three stop codons (TAA, TAG, TGA) are mostly created out of adenine and thymine (Pohl, Theissen, & Schuster, 2012; Trotta, 2016), a lack of these pyrimidine bases present in the genome might lead to lesser stop codon creation, enabling longer ORF creation. In contrast, other studies detected usage shift of stop codons based on GC-content, where stop codon TGA is clearly in favour in higher GC-content genomes (Povolotskaya et al., 2012; Wong et al., 2008). Based on these finding, an analysis between GC-content and eORF length is performed with an expected outcome of detecting longer eORFs in high GC-content genomes.

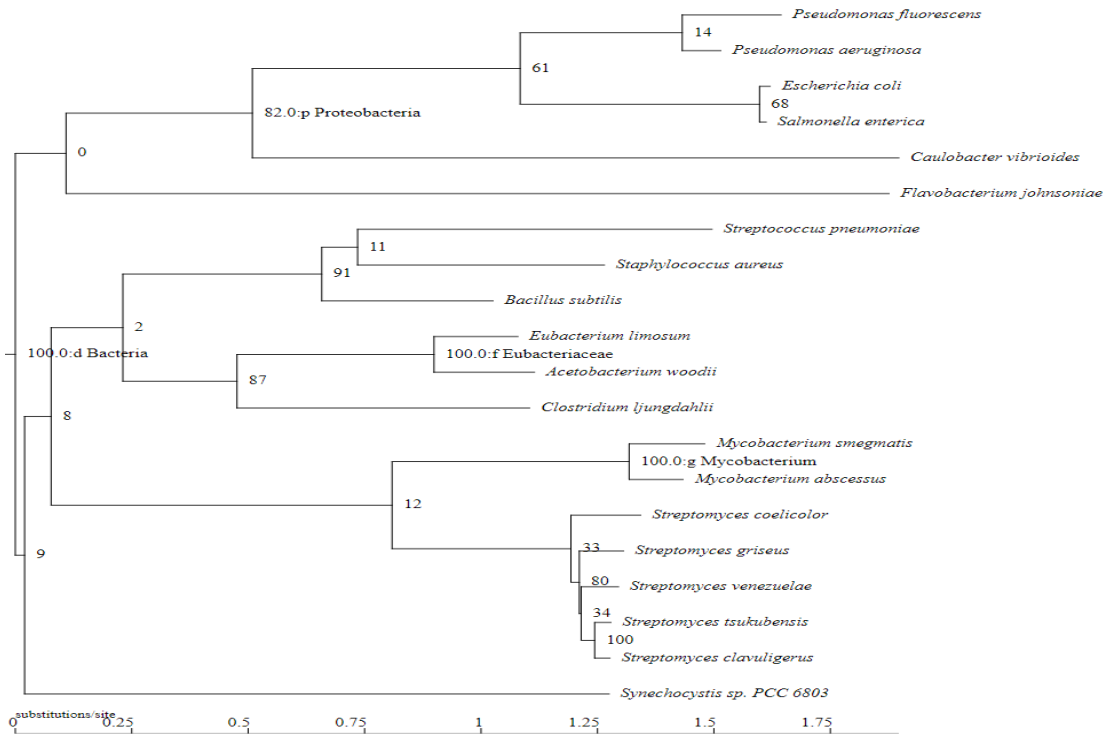

Figure 4: Self-constructed phylogenetic tree showing the species chosen for phylogenetic analysis. Note that archaeal species are missing here.

The genomic sequence by its codon triplets dictates any encoded amino acid sequence. A shift by one or two nucleotides, representing the alternative two frames per strand, is coding for completely different amino acids. Based on the dependency of one nucleotide affecting three different codon positions in the alternative frames, it is of interest if the OLG creation is favoured in one. Expected, it could be argued, is a clear trend for their detection in sas11, as the least important codon position in the mother gene affects the most important position (2^{nd}) in the overlap. Thereby, a mutation in this location does not alter the functionality of the mother gene but causes major amino acid changes in the OLG. The frame for each OLG detected is determined with subsequent comparison to predicted eORFs which do not meet the criteria as being declared translated and therefore are assumed to be 'background' ORFs and not genes.

Identified eORFs of interest, that are found to be translated in several samples per species or even with different predictions techniques, and show similar read distribution patterns as annotated genes, are the subject of further descriptive analyses. Phylostratigraphic analysis estimates a gene's age by searching homologues in the phylogenetic tree and inferring the last common ancestor of the extant homologues (Domazet-Loso, Brajkovic, & Tautz, 2007; Zhang, Tan, Fan, Zhang, & Zhang, 2019). Detection of homologues is based on the tblastn NCBI search, where a translated protein sequence is used as an input query for a nucleotide database. Codon degeneracy, in this case, can result in various nucleotide combinations for one codon which are all used within the search approach, enabling a lesser restricted search than if only one given nucleotide sequence is used as the search template (blastn). The tool OLGenie focusses on the selection pressure and therefore draws conclusions of potential functionality

of the ORF analysed (Nelson, Ardern, & Wei, 2020). Besides these two methods, Frameshift can provide information on whether the length of a predicted ORF is significantly longer than expected. Hereby, the mother gene sequence is translated into codon succession followed by random shuffling (Schlub, Buchmann, & Holmes, 2018). The length outcome of potential ORFs in the alternative frames are analysed and compared to the length of the actual eORF of interest (e.g., one predicted as translated).

1.3 Important bacteria of the SIHUMI community - *Escherichia coli* K12 MG1655 and *Bacteroides thetaiotaomicron* VPI-5482

Without overstatement *Escherichia coli (E. coli)* K12 is one of the most commonly used and best understood bacteria in biological research. A simple name search in the national centre for biotechnology information (NCBI) reveals 390.825 publications focussing on this specific prokaryote (cited 03.11.2020). Thus, *E. coli*, a gram-negative, rod-shaped bacterium with a genome size of ~ 4.6 Mb, is one of the most well-studied organisms, being one of the first for which whole genome sequencing was performed (Baba et al., 2006; Blattner et al., 1997; Kneifel & Forsythe, 2017; Lim et al., 2010). Nowadays, *E. coli* K12 MG1655 is the most commonly used lab strain with a nearly unaltered genome structure (Blattner et al., 1997; Hayashi et al., 2001), although substrains such as MC4100 or BW25113 are also of interest having slightly changed genomic structures (Grenier, Matteau, Baby, & Rodrigue, 2014; Peters, Thate, & Craig, 2003). It is a harmless inhabitant of high abundance in human and animal gut flora, while its pathogenic relative enterohemorrhagic *E. coli* O157:H7 EDL933 or Sakai are causing severe gastrointestinal tract affecting diseases (Hücker, Ardern, et al., 2017; Lim et al., 2010; Neuhaus et al., 2016). Besides, *E. coli* other inhabitants of the human gut have been identified and together are used as a model microbial community, namely *Anaerostipes caccae, Bacteroides thetaiotaomicron, Bifidobacterium longum, Blautia producta, Clostridium ramosum, Lactobacillus plantarum.* These seven species are referred to as the simplified human intestinal microbiota (SIHUMI) (Becker, Kunath, Loh, & Blaut, 2011). As *E. coli* has been analysed sufficiently, even with RIBO-Seq approaches, another member of the SIHUMI group was chosen for an own ribosome profiling experiment. *B. thetaiotaomicron* was chosen for this approach, as it makes up nearly 30 % of human gut commensals (Hooper et al., 2001; Mimee, Tucker, Voigt, & Lu, 2015).

To contribute an additional RIBO-Seq dataset to the current state of research, a RIBO-Seq experiment of a bacterial strain involved in the human gut microbiome was performed. Besides *E. coli*, *B. thetaiotaomicron* is a relatively high abundant, gram-negative gut bacterium (Colosimo et al., 2019; Mimee et al., 2015). To our knowledge, so far just one RIBO-Seq dataset focusing solely on *B. thetaiotaomicron* has been published (Sberro et al., 2019). Due to its larger proportion in the human gut and the lack of additional dataset available, *B. thetaiotaomicron* was chosen as the candidate for the following RIBO-Seq experiment. This experiment aims to potentially detect so far overlooked overlapping genes in this candidate. Simultaneously, RNA-Seq was performed for comparison purposes

and potential ribosomal coverage value (RCV) calculation. This value is obtained by the division of RPKMs obtained from RIBO-Seq experiments with those from RNA-Seq equivalents. Additionally, results from this analysis then will be compared to the ones obtained from evaluating the publicly available dataset from Sberro (Sberro et al., 2019). With this approach, the second verification of potentially detected overlapping genes could be achieved, if found in both datasets. Subsequently, a phylostratigraphy analysis should reveal the OLGs age combined with a BLAST analysis for potential function determination.

1.4 Purpose of this study

The existence of overlapping genes as an important feature of bacterial genomes is still controversial, however, with the development of RIBO-Seq an important steppingstone towards proving their existence has been made. So far, they have been detected in various model organisms across eukaryotes and prokaryotes. The first goal of this study is the comparison of available RIBO-Seq data performed on *E. coli* K-12 to potentially identify determining factors contributing to the successful detection of overlapping genes. Factors such as the read amount necessary for evaluation of data, gel excision range to obtain ribosomal covered mRNA fragments, and application of translation inhibitors which improve the detectability of the start position of genes are just a part of the analyses performed. Results gathered from the performed analyses should be seen as recommendations that are applied in experimental lab work to enhance the possibility of detecting overlapping and other unannotated genes.

Additionally, a second goal is to shed light on the existence of OLGs by analysing diverse RIBO-Seq data obtained from prokaryotic species and archaea. The detection of eORFs with homologues throughout the phylogenetic tree would support their existence and potential functionality as they are maintained in various genomes. Species-specific characteristics concerning eORFs were analysed as well as the favoured frame relation of such in relation to the mother gene. The repeated detection of the same eORF within multiple samples per species, on occasion even predicted with different methods, may indicate actual ‘functionality’. Hence, these eORFs were subjected to further characterisation based on homologue comparison, selection pressure determination, and sequence length significance analysis.

Lastly, a RIBO-Seq experiment with *B. thetaiotaomicron* was performed, not only testing the developed experimental recommendations but also to contribute to eORF detection in an additional prokaryotic species. As the main focus of analyses performed for this study was on computational evaluation and analysis, an experimental part was included to complete the RIBO-Seq based prediction of overlapping genes in multiple prokaryotic organisms.

2. Material and Methods

2.1 Computational Evaluation

2.1.1 Tools

Artemis Release 17.0.1

bedtools v2.25.0

Bowtie2 2.2.6

DeepRibo

EMBOSS:6.6.0.0

fastp 0.20.0

FastQC v0.11.4

FastQ Screen v0.14.0

genbank_to_fasta.py 1.2

GNU bash version 4.3.48

gnuplot 5.2 patchlevel 4

newick utilities V1.1

OLGenie.pl

OLGenie_bootstrap.R

ORFFinder.pl

Perl v5.22.1

Prodigal V2.6.2

Python 2.7.16

R 3.2.3

REPARATION

samtools 1.7

seqkit

2.1.2 Data set compilation

To evaluate RIBO-Seq data comparative analyses were performed on publicly available sequencing experiments from *E. coli* K12. Experiments were searched for in google scholar or on the gene expression omnibus (GEO) website from NCBI. Combinations of tags such as 'ribosomal profiling', 'Escherichia coli', 'High throughput sequencing' or 'RIBO-Seq' were used to search for suitable experiments. Results from these searches were subsequently filtered according to whether they match the requirement of a RIBO-Seq experiment performed with *E. coli* K12 in lysogeny broth medium. Data sets with the same growth medium used (LB) were chosen to ensure growth condition does not alter the translatome. In total raw reads in fastq format for 48 samples from 9 different experiments were downloaded that were available at the time of analysis (Balakrishnan, Oman, Shoji, Bundschuh, & Fredrick, 2014; Bartholomäus et al., 2016; Elgamal et al., 2014; Hwang & Buskirk, 2017; Kannan et al., 2014; Marks et al., 2016; Oh et al., 2011; J. Wang et al., 2015; Woolstenhulme, Guydosh, Green, & Buskirk, 2015). Experiments were performed with one of the three close related *E. coli* K12 substrains BW25113, MC4100 or MG1655. Differences within their genomic structure led to the differentiation from MG1655, as BW25113 and MC4100 lack operon structures and ORFs present in MG1655 (Grenier et al., 2014; Peters et al., 2003). Nevertheless, their close relation enables comparison between them, which results in a bigger data set to analyse. Samples were obtained from either the European Nucleotide Archive (ENA) or NCBI BioProject website.

Similar to the first analysis project, raw reads of different prokaryotic RIBO-Seq experiments were again obtained from ENA or the BioProject website for a second project. Here, the search of tag combinations such as 'RIBO-Seq', 'ribosome profiling', 'bacteria' or 'prokaryotes' provided many results, which were subsequently filtered to match the criteria of RIBO-Seq experiments performed in different prokaryotes with at least two samples available for reproducibility. Further, experiments were chosen according to their relations, as one goal for this analysis was to detect if eOLG can be found distributed throughout the phylogenetic tree. Raw read files were downloaded for 22 different species, 20 were bacteria whereas two were archaeal species. The species are as followed: *Acetobacterium woodii, Bacillus subtilis, Caulobacter crescentus, Clostridium ljungdahlii, E. coli, Eubacterium limosum, Flavobacterium johnsonia, Halobacter salinarium, Haloferax volcanii, Mycobacterium smegmatis, M. abscessus, Pseudomonas aeruginosa, P. fluorescens, Salmonella enterica typhimurium, Staphylococcus aureus, Streptococcus pneumonia, Streptomyces clavuligerus, S. coelicolor, S. griseus, S. tsukubensis, S.venezuelae* and *Synechocystis sp.* PCC6803 (Al-Bassam et al., 2018; Baek, Lee, Yoon, & Lee, 2017; Baez et al., 2019; Basu & Yap, 2016; Davis, Gohara, & Yap, 2014; Gelsinger et al., 2020; Giess et al., 2017; Grady et al., 2017; Grenga et al., 2017; Jeong et al., 2016; Karlsen, Asplund-Samuelsson, Thomas, Michael, & Hudson, 2018; Kim et al., 2020; Li et al., 2012; Lopez Garcia de Lomana et al., 2020; Miranda-CasoLuengo, Staunton, Dinan, Lohan, & Loftus, 2016; Ndah et al., 2017; Schrader et al., 2016; Schrader et al., 2014; Shell et al., 2015; W. Song et al., 2019; Y. Song et al., 2018; Subramaniam et al., 2013; Yang et al., 2016) (*A.woodii*: PRJEB33460, available on NCBI BioProject).

In total 364 unique samples were found, in which already 192 are belonging to *E. coli* K12 substrains, once more highlighting its role as a go-to bacterium for analysis.

2.1.3 Quality control, trimming and filtering of data

The first project was performed on the RIBO-Seq data set based on the different *E. coli* K12 substrains. Quality of reads both raw and after trimming was monitored using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). First, raw reads in fastq format were subjected to the program within the bash console and the obtained HTML file was manually inspected. Of special interest in this step were results for adapter content or overrepresented sequences present in the samples. Considering this information is available as well as extracting the adapter information from the corresponding publication if necessary, samples were subjected to the trimming step.

Here, adapter sequences present at the end of raw reads were trimmed using fastp v0.14.2 (S. Chen et al., 2018; Glaub et al., 2020). Therefore, adapter sequences were either specified as an input variable or if automated for several samples with different adapters, a tab delineated file with the information of sample number, corresponding genome number and the used adapter sequence were subjected to the program. Additionally, setting for trimming an accumulation of the same nucleotide at the 3'-end of a read (polyX structure, setting -x) was enabled, as these structures build-out of repeating adenine are oftentimes used as a kind of adapter equivalent. Also, quality filtering is disabled (setting -Q or --disable_quality_filtering) which allows keeping reads with ambiguous nucleotides (N) incorporated in the read for further processing. The amount of this undecided nucleotides within the sequence can be limited by setting --n_base_limit. This flag was set for the phylogenetic analysis (with --n_base_limit 1) and will be explained later in this chapter.

Trimmed reads were then aligned to their corresponding reference genome with Bowtie2 v2.2.6 in local alignment mode (Langmead & Salzberg, 2012). This mode was chosen to ensure that if trimming was not performed perfectly, alignment can still be conducted. In this mode (--local) exact alignment at the end of a read is not required, so remaining adapter sequence at the 3'-end is not interfering with the alignment of the read (Langmead & Salzberg, 2012). Beside standard-settings like input file and corresponding genome sequence, the length of the seed substring (-L) and the maximal number of mismatches in the seed alignment (-N) were specified (Glaub et al., 2020). A seed sequence is defined as a part of the read that is used for first aligning it to its target, which is then further extended in its length during the ongoing alignment (Ye, Meehan, Tong, & Hong, 2015). While in the alignment of the seed substring leads to multiple mapping options during the extension of the sequence the alignment becomes more precise and is at best assigned to a specific genome location after the complete read length is aligned. For the performed alignment in this project, a seed length of 19 was chosen and no mismatches within the seed sequence were allowed (-L 19, -N 0). After trimming and alignment reads

were sorted with the samtools' sort option as a sorted order is necessary for filtering the reads resulting in bam format files labelled "$input.sorted.bam". Trimming efficiency was again inspected with FastQC (Glaub et al., 2020).

Only reads obtained from ribosome protected mRNA fragments are of interest for evaluation of the translatome itself. Therefore, reads mapping either rRNA or tRNA were filtered out with the bedtools tool. Therefore, unique tables for each genome containing annotation information were collected from NCBI. Out of this locus information for tRNA and rRNA present in the respective genome were extracted. Comparison of the read containing bam files to the tRNA and rRNA locus information in bed file format was performed using bedtools intersect. Here, reads in the trimmed bam files were excluded if they map the regions specified in the bed file resulting in a sam file with reads mapping to mRNA. A workflow for the steps explained can be found in Script 1.

```
cores=8
# perform quality control for raw fastq files available
fastqc *.fastq -o /directory/of/raw_files -t 6 ;

# trimming, alignment and filtering in an automated script
accession="GCF_number" #species specific number
cat samples_info.txt | while read -r ribo adapter_sequence
do
file1="$ribo".fastq ;
genome=$(echo "$accession"*_genomic.fna) ;
ft=${genome%_*}_feature_table.txt ;
adapter="$adapter_sequence" ;

echo $genome $adapter ;

chromosome=$(cat $genome | head -1 | awk '{print $1}' | sed -e "s|>||g") ;

test ! -e $chromosome.fna && faidx -x $genome ;
test ! -e Genes-"$chromosome".txt && cat $ft |
awk -F "\t" '{if ($1=="gene" && $2=="protein_coding" &&$7=="'$chromosome'")
print $8 "\t" $9 "\t" $10}' > Genes-"$chromosome".txt ;

# BED file of tRNA and rRNA positions, from feature table
test ! -e $chromosome-excluded-RNAs.bed && cat $ft | awk -F "\t"  '{if
($1=="gene" && $7=="'$chromosome'") print}' |
awk -F "\t" '{if ( $2=="rRNA" || $2=="tRNA") print $7 "\t" ($8-1) "\t" $9
"\t" $17 "\t" "0" "\t" $10 }' > $chromosome-excluded-RNAs.bed ;

input=$(echo $file1 | awk -F "[_|.]" '{print $1}' )
# PRE-PROCESSING OF FASTQ FILE
# removing adapter sequences
# ALIGNING
test $adapter != "-" && fastp -i $file1 -x -Q -a $adapter -o ${input%.*}-
fastp.fastq ;
test ! -e $chromosome.rev.1.bt2 && bowtie2-build $chromosome.fna
$chromosome ;
bowtie2 -p $cores --local -x $chromosome -N 0 -L 19 -U ${file1%.*}-
fastp.fastq |
samtools view -bh - | samtools sort - > $input.sorted.bam ;
samtools index $input.sorted.bam  ;
```

```
# zip fastq file and remove fastp.fastq file to save space
rm ${input%.*}-fastp.fastq ;

################

# Remove rRNA & tRNA regions from BAM file, to create input SAM for
REPARATION
test ! -e ${input%.*}_RNAfree.sam && bedtools intersect -abam
${input%.*}.sorted.bam  -b $chromosome-excluded-RNAs.bed -v |
samtools view -hS > ${input%.*}_RNAfree.sam ;

# convert sam file to bam and subsequently to fastq for quality control,
each sample
samtools bam2fq $input.sorted.bam > $input.sorted.mapped.fastq ;

# perform quality control for trimmed and filtered fastq files available
fastqc $input.sorted.mapped.fastq -o /directory/of/trimmed_filtered_files -
t 6 ;
done
```

Script 1: Preprocessing pipeline of RIBO-Seq raw reads including quality control with FastQC, adapter removal using fastp and alignment to reference genome with bowtie2. Subsequently, tRNA and rRNA reads are excluded.

In the second project that focusses on the detection of eORFs throughout the phylogenetic tree, slightly different settings were used after the software was updated. In the trimming step the incorporation of only one ambiguous nucleotide was allowed (--n_base_limit 1). The complexity of adjacent bases had to be at least 30 % (--low_complexity_filter 30) and reads shorter than 15 nucleotides were discarded during analysis (--length_required 15). Additionally, for alignment the stricter mode was chosen (end-to-end¸--very-sensitive -D 20 -R 3 -N 0 -L 17 -i S,1,0.50), as the quality of trimming was improved and remaining contaminants, so-called overrepresented sequences, were removed if they exceeded 0.5 % and were not mapping to the genome. Additionally, only mapping reads were kept in the files for further processing (with the setting --no-unal). Setting changes were made in the trimming and alignment step as shown in Table 1.

Table 1: Settings used for read evaluation in analysis for detection of eORFs in several prokaryotic species.

Step	Important settings
Trimming (Fastp 0.20.1)	fastp --thread 6 --in1 .//"$sample".fastq --out1 ./"$species"/samples/"$sample"/02.Fastp/run1/"$sample".trimmed.fastq --adapter_sequence "$sequence" --disable_quality_filtering --n_base_limit 1 --low_complexity_filter --complexity_threshold 30 --length_required 15 --length_limit 0 --json ./"$species"/samples/"$sample"/02.Fastp/run1/"$sample".trimmed.json --html ./"$species"/samples/"$sample"/02.Fastp/run1/"$sample".trimmed.html 2>&1

Step	Important settings
Alignment (Bowtie2 2.3.5.1)	bowtie2 -p 6 --quiet -q --end-to-end -D 20 -R 3 -N 0 -L 17 -i S,1,0.50 --no-unal -x ./"$species"/tmp/"$genome".fna -U ./"$species"/samples/"$sample"/02.Fastp/run1/"$sample".trimmed.fastq \| samtools sort -@ 6 -O bam - 2> /dev/null \| bedtools intersect -abam stdin -b ./"$species"/tmp/RNA.bed -v | samtools sort -@ 6 -O bam -n - -o ./"$species"/samples/"$sample"/03.BedtoolsFiltering/"$sample".trimmed.filtered.bam 2> /dev/null

Output "$sample".trimmed.filtered.bam files had to be sorted again using samtools sort for further processing.

2.1.4 Read depth analysis

Evaluation of the necessary amount of reads left after trimming and filtering is based on the comparison of read amount and detected annotated genes within the sample. The total number of reads mapping to the genome after passing trimming and filtering is obtained. Therefore, reads mapping to the genome sequence are obtained with samtools' view option (setting -S forces strand specificity for the reads, -F 4 forwards only mapped reads to the output file). Their amount is calculated with samtools' depth setting. Similar, samtools' flagstat is another option to obtain the read amount information, both for total reads left but also for mapped ones.

Subsequent, a prediction of annotated genes within the first project dataset was made with REPARATION, a machine learning algorithm that is first trained on annotated ORFs to later predict potential new ORFs. A subset of data per sample is blasted against a protein database (here Uniprot Database) to analyse the pattern of read distribution across the ORFs and additional coverage estimation using only prodigal (Ndah et al., 2017). Considered start codons for detection of translated ORFs are ATG, GTG and TTG resulting in ORFs from these starts to the next stop codon (Ndah et al., 2017). The learned patterns are subsequently applied to the remaining data to predict ORFs possible. Reads shorter than 19 nucleotides were excluded from the analysis. Predicted ORFs are filtered according to specific criteria, here if they are corresponding to genome locations belonging to annotated genes. Genomic information such as locus of the gene, strand location, as well as its characteristics (described as a gene and protein-coding) are extracted from the NCBI obtained feature table. Comparison of ORF positions in both files was performed using awk programming. If positions matched the predicted ORFs were categorized as annotated genes. The amount of annotated genes identified was plotted against the read depth after trimming and filtering.

```
#!/bin/bash
# PREDICTS GENES USING REPARATION
# aORF = ORFs corresponding to annotated genes
# eaORF = embedded ORF, on opposite strand to an annotated gene
# SETTING INITIAL VARIABLES:
cores=8
date=$(LC_ALL=en_GB.utf8 date | awk '{print $3 $2}' | sed -e "s|\.||g") ;
program=prodigal ;
mkdir tmp ;

#########################
echo "GENE PREDICTION WITH REPARATION" ;
cat samples.txt | while read -r ribo accession ;
do
### Set Variables
input="$ribo";
sam=${input%.*}_RNAfree.sam ;
genome=$(echo "$accession"*_genomic.fna) ;

# uniprot database file was downloaded from the website on Feb 27, 2019
db=uniprotSP_bacteria_27022019.fasta ;
ft=$(echo "$accession"*_feature_table.txt) ;
chromosome=$(cat $genome | head -1 | awk '{print $1}' | sed -e "s|>||g") ;

test ! -e Genes-"$chromosome".txt && cat $ft |
awk -F "\t" '{if ($1=="gene" && $7=="'$chromosome'" && $2=="protein_coding"
|| $2=="pseudogene") print $8 "\t" $9 "\t" $10}' > Genes-"$chromosome".txt
;
test ! -e Genes2-"$chromosome".txt && cat Genes-"$chromosome".txt | awk
'{if ($3=="+") print $2$3 "\t" $0; else print $1$3 "\t" $0}' > Genes2-
"$chromosome".txt ;
#######################################################
echo "RUNNING REPARATION ON "$sam"" ;
perl ~/REPARATION/reparation.pl -sam $sam -g "$chromosome".fna \
-sdir ~/REPARATION/scripts/ -db $db -en "$sam"_prodigal -sd Y -mn 19 -pg 1
;
echo "CLASSIFYING PREDICTED GENES FOR "$sam"" ;
cat
"$sam"_"$program"_reparation_"$date"/"$sam"_"$program"_Predicted_ORFs.txt |

# post processing (filtering of at least three reads necessary for
evaluation is incorporated)
awk '{if ($5>=3) print $1 "\t" $2 "\t" $3 "\t" $4 "\t" $5 "\t" $6 "\t" $7
"\t" $8 "\t" $9 "\t" $10}' |
tail -n+2 | awk '{split($1,a,":"); print a[2] "\t" $2 "\t" $5 "\t" $6 "\t"
$7 }' |
awk '{split($1,a,"-"); print a[1] "\t" a[2] "\t" $2 "\t" $3 "\t" $4 "\t" $5
"\t" "'$sam'"}' |
awk '{if ($3=="+") print $1 "\t" ($2+3) "\t" $3 "\t" $5 "\t" $6 "\t"
"'$program'" "\t" "allORFs" "\t" ($2+3)$3 "\t" $4 ;
else print ($1-3) "\t" $2 "\t" $3 "\t" $5 "\t" $6 "\t" "'$program'" "\t"
"allORFs" "\t" ($1-3)$3 "\t" $4 }' | sort | uniq >
tmp/"$sam"_"$program"_ORFs.txt ;

# annotated genes (based on same stop codon and same strand)
awk -F "\t" 'NR==FNR{a[NR]=$0; next}{for (i in a){split(a[i],x,"\t"); \
if (x[1]==$8) print x[1] "\t" x[2] "\t" x[3] "\t" x[4] "\t" $0 }}' Genes2-
"$chromosome".txt tmp/"$sam"_"$program"_ORFs.txt |
awk '{print $1 "\t" $5 "\t" $6 "\t" $7 "\t" $8 "\t" $9 "\t" "||" "\t" $2
"\t" $3 "\t" $4 "\t" $(NF-3) "\t" "anORF" "\t" $NF }' | sort | uniq >
tmp/"$sam"_"$program"_aORFs-Ribo.txt ;
```

```
# embedded antisense ORFs, non-annotated
awk -F "\t" 'NR==FNR{a[NR]=$0; next}{for (i in a){split(a[i],x,"\t"); \
if (x[1]!=$7  && x[2]<$1 && x[3]>$2 && x[4]!=$3) print x[1] "\t" x[2] "\t"
x[3] "\t" x[4] "\t" $0 }}' Genes2-"$chromosome".txt
tmp/"$sam"_"$program"_ORFs.txt |
awk '{print $(NF-1) "\t" $5 "\t" $6 "\t" $7 "\t" $8 "\t" $9 "\t" "||" "\t"
$2 "\t" $3 "\t" $4 "\t" $(NF-3) "\t" "eORF" "\t" $NF }' |
awk -F"\t" 'NR==FNR{a[$1]=$1;next}{ if (!a[$1])print ;}'
tmp/"$sam"_"$program"_aORFs-Ribo.txt - |

# "$sam"_"$program"_psORFs-Ribo.txt was performed in another filtering step
(not shown)
awk -F"\t" 'NR==FNR{a[$1]=$1;next}{ if (!a[$1])print ;}'
tmp/"$sam"_"$program"_psORFs-Ribo.txt - | sort | uniq >
tmp/"$sam"_"$program"_eaORFs-Ribo.txt ;
cat tmp/$sam*Ribo.txt > $sam-genes.txt ;

# convert combined output to BED file for visualisation
# -1 to convert from 1 based to 0 based format
cat $sam-genes.txt | awk '{ print "'$chromosome'" "\t" ($2-1) "\t" $3 "\t"
$NF "\t" "0" "\t" $4 }' > $sam-genes.bed ;

# create BED file for annotated genes if not already made
test ! -e Genes-$chromosome.txt && cat Genes-$chromosome.txt |
awk '{ print "'$chromosome'" "\t" ($1-1) "\t" $2 "\t" "ANNOTATED" "\t" "0"
"\t" $3 }' > annotated_Genes-$chromosome.bed ;
echo "GENES FOR "$sam" CLASSIFIED" ;
done ;
```

Script 2: Wrapper script around REPARATION based open reading frame prediction. Subsequent categorization of predicted ORF as annotated ORFs (aORF) and embedded antisense ORFs (eaORF). Additional categories such as partial sense or antisense ORFs were performed but are not shown here. Last, all predictions per sample were combined and converted into bed format for later processing.

Additionally, as a second option to analyse a potential correlation between read depth and prediction efficiency ribosomal coverage values (RCVs) were calculated for annotated genes if RIBO- and RNA-Seq information were available for the same sample (Glaub et al., 2020). RCV is characterised as a value for translation based on the division of RPKMs obtained from RIBO-Seq experiments by RNA-Seq equivalents. Indication of translation is given if RCV exceeds 0.355 (Glaub et al., 2020; Neuhaus et al., 2017). RNA-Seq samples were available for 22 samples and were pre-processed according to Script 1 as for RCV calculation only reads covering mRNA locations were of interest. The necessary input files for both sequencing approaches included the trimmed and filtered reads which were subsequently compared to the REPARATION based ORFs (Glaub et al., 2020). Based on the read coverage per locus of interest RPKMs for genes of interest in RIBO-Seq and RNA-Seq results were calculated followed by RCV estimation. Only genes with RCVs ≥ 0.355 were considered translated and used in read depth comparison analysis (Glaub et al., 2020).

```
#!/bin/bash/

#### Predict genes of different classes, based on REPARATION output
#### Calculate RPKMs and RCVS for defined sequence regions, from BAM files
cat rcvs.txt | while read -r riboseq rnaseq accession ;
genome=$(echo "$accession"_genomic.fna) ;
chromosome=$(cat $genome | head -1 | awk '{print $1}' | sed -e "s|>||g") ;
do
cat ${riboseq%.*}_RNAfree.sam-genes.txt | awk '{print "'$chromosome'" "\t"
($2-1) "\t" $3 "\t" $1 "\t" "0" "\t" $4}' |
sort -k1,1 -k2,2n > $riboseq-positions_sorted.bed ;

for input in $rnaseq $riboseq ;
do
chr=$(samtools view ${input%.*}_RNAfree_headered.bam | awk '{print $3}' |
egrep -v '[*]' | head -1 ) ;
chrom=$(echo $chr | awk -F "|" '{print $(NF-1)}' ) ;
echo $chrom ;

# Calculate RPKMs
reads=$(samtools flagstat ${input%.*}_RNAfree_headered.bam | grep mapped |
grep -v mate | awk '{print $1}' ) ;
millions=$(echo "$reads / 1000000" | bc -l) ;
cat $genome.fai | awk '{print $1 "\t" $2}' > genome.txt ;
bedtools coverage -sorted -s -F 0.5 -a $riboseq-positions_sorted.bed -b
${input%.*}_RNAfree_headered.bam -g genome.txt |
awk '{print $2 "\t" $3 "\t" $4 "\t" $7 "\t" $7/(($9/1000)*"'$millions'")
"\t" $NF}' > ${input%.*}_coverage.txt ;
done ;

# Calculate RCVs [final column of output]
paste ${rnaseq%.*}_coverage.txt ${riboseq%.*}_coverage.txt | awk '{if
($4>0) print $0 "\t" $11/$5; else if($4==0) print $0 "\t" "0"}' > $riboseq-
RCVs.txt ;
done ;

# filter for ORFs that are above threshold of 0.355 for RCV value
cat rcvs_ribo.txt | while read -r ribo
do
awk '{if ($13 >= 0.355) print $0}' "$ribo"-RCVs.txt >
"$ribo"_RCV_Threshold.txt
done
```

Script 3:RCV calculation script. Input files were created after REPARATION prediction containing its combined results. Calculations can be made for RIBO-Seq as well as RNA-Seq files, if available. The final output is filtered according to the RCV threshold of ≥ 0.355.

Further, for three samples (SRR1734437, SRR1734439, SRR1734441, (Woolstenhulme et al., 2015)) with high read depth, an additional analysis was performed to verify the necessary read depth recommendation within samples. Therefore, variations of lesser coverage per sample were analysed in regard to the prediction efficiency of annotated genes (Glaub et al., 2020). A decreased coverage was obtained by random extraction of only a certain amount of reads (Script 4). This ‘down sampling’ was performed in triplicates per coverage step with a script provided by a bioinformatician. For each newly obtained reduced sample gene predictions were made with REPARATION with subsequent mean calculation within the triplicate per coverage step.

```
bash DownsampleSam.bash -c 50,75,100,125 -r 3 -t 8 -f assembly_size
"$sample"_RNAfree.sam "$accession_number"_genomic.gff /"$Output_Directory"
```

Script 4: Script used to randomly extract reads within a sample to test prediction efficiency in less covered subsamples. DownsampleSam.bash is an in-house written script achieving this task. -c coverage list (adapted to the coverage of each sample), -r number of replicates, -t number of threads, -f normalization factor.

Similar to the categorization of annotated genes, further classifications for ORFs predicted by REPARATION were made for later analysis. The comparison of genomic positions again was performed using awk programming, this time focussing on the identification of ORFs predicted in the same location as an annotated gene but in an alternative frame. Two types of overlaps were distinguished, where embedded ORFs are located fully within the start and stop position of an annotated gene, whereas partial overlaps are characterised by an ORF overlapping an annotated gene at one of its ends. Next, analyses were performed that focus on the different read length present after sequencing. Can specific characteristics be assigned to the different read length?

2.1.5 Evaluation of various read length

To test whether specific read length can be assigned to a type of RNA (mRNA, rRNA or tRNA) all reads per sample were categorized according to its type (Glaub et al., 2020). Here, besides the samples specific trimmed and filtered sam files that only contain reads mapping to mRNA similar files were constructed with reads stored mapping either exclusively rRNA or tRNA regions. The information for rRNA or tRNA regions were extracted from the feature table of the specific genomes. These files were then compared to bed files containing mapping information for each read within the sorted bam file, such as start and stop position, length and strand location. At least 50 % had to map in a region of interest to be considered as mapping. These read information were compared to the rRNA and tRNA regions of interest using bedtools' intersect option. The length of each read that was considered mapping to a region of interest (either rRNA or tRNA) was calculated with subsequent counting of re-occurring length. Read length analysis for mRNA mapping reads was performed differently as sample-specific trimmed and filtered sam files could be used. Here, only the length of reads that mapped the genome was calculated. Thereafter, only those with a length between 20 to 40 nucleotides were considered for analysis, as this contains most of the size selection ranges used in the experiments analysed. For each RNA type within a sample, their length distribution was calculated in percentage (Glaub et al., 2020). Therefore, the total amount of the reads was summed within the sample. The obtained read sum was used to calculate the percentage of reads representing the specific read lengths. Percentage values for all samples were combined into one file for which column and line arrangements were switched and sorted, necessary for subsequent median estimation. The median was estimated out of all corresponding length values within the 46 samples (Glaub et al., 2020). This estimation for 46 samples was based on calculating the average number from the two values at position 23 and 24 from the sorted columns. Median estimation was performed for each RNA type (mRNA, rRNA and tRNA) and plotted against each other for length

comparison between the respective types (Glaub et al., 2020). Calculation of read length for the different RNA types, followed by median estimation exemplary for rRNA is shown in Script 5.

```
# extract rRNA information of species specific feature table
cat $feature_table | awk '{if ($1=="gene" && $2=="rRNA") print $7 "\t" $8
"\t" $9 "\t" "gene" "\t" "0" "\t" $10 }' |
sort -k1,1 -k2,2n > "$feature_table"_rRNA_tmp.bed ;
#######
cat all_samples.txt | while read -r ribo accession
do
# set variables
bam_file="$ribo".sorted.bam ;
feature_table="$accession"_feature_table.txt ;
#######
echo BAM: $bam_file
bedtools bamtobed -i $bam_file | awk '{print $6}' > "$ribo"_tmp.txt  ;
samtools view -F 4 $bam_file | awk '{print $3 "\t" $4 "\t"
($4+(length($10))) "\t" $1 "\t" "0" }' > "$ribo"_tmp2.txt ;
paste "$ribo"_tmp2.txt "$ribo"_tmp.txt > "$ribo"_bam.bed ;
bedtools intersect -u -s -f 0.50 -sorted -bed -a "$ribo"_bam.bed -b
"$feature_table"_rRNA_tmp.bed | awk '{print $3-$2}' | sort | uniq -c >
"$ribo"_rRNA_reads.txt ;
rm "$ribo"_tmp2.txt "$ribo"_tmp.txt
done
# trim files containing read length to range from 20-40
for i in *_rRNA_reads.txt
do
cat "$i" | awk '{if ($2>19 && $2<41) print $0 }' > "$i"_trimmed.txt
done
# sum total amount of reads mapping to rRNA per sample
cat allsamples.txt | while read -r ribo
do
awk '{sum+=$1} END {print sum}' ${ribo%_*}_rRNA_reads.txt_trimmed.txt >
${ribo%_*}_rRNA_readsumme.txt ;
done
# calculated read distribution in percentage per sample
cat allsamples.txt | while read -r ribo ;
do
number=$(cat ${ribo%_*}_rRNA_readsumme.txt)
cat "$ribo"_rRNA_reads.txt_trimmed.txt | awk '{print
(($1/"'$number'")*100)}' > "$ribo"_rRNA_percentage.txt
done
# combine read length distribution of all samples
awk '{ a[FNR] = (a[FNR] ? a[FNR] FS : "") $1 }END{for(i=1;i<=FNR;i++) print
a[i]}' *_rRNA_percentage.txt > rRNA_percantages.txt ;
awk '
{
    for (i=1; i<=NF; i++)  {
        a[NR,i] = $i
    }
}
NF>p { p = NF }
END {
    for(j=1; j<=p; j++) {
        str=a[1,j]
        for(i=2; i<=NR; i++){
            str=str" "a[i,j];
        }
        print str
    }
}' rRNA_percantages.txt > switched_file_construction_rRNA.txt ;
```

```
cat switched_file_construction_rRNA.txt | sed -e 's/ /\t/g' >
sorted_rRNA.txt ;
# estimate median value for each read length
for i in {1..21} ;
do
awk -F '\t' -v col=$i '{print $col}' sorted_rRNA.txt | sort -b -gk1,1 | awk
'{if (NR==23 || NR==24) print}' | awk '{sum+=$1} END {print sum/NR}' ;
done > median_rRNA.dat
```

Script 5: Read length analysis according to different types of RNA (mRNA, rRNA, tRNA), here shown for rRNA. Information about rRNA sequence location is extracted from corresponding species' feature tables. Reads mapping to rRNA are analysed according to their length for each sample respectively. Within each sample, percentages were calculated for read length distributions within RNA types, and median calculation was performed over all samples to compare distribution across all.

A second analysis was only focussing on rRNA read length, as here a depletion before sequencing is crucial to obtain a higher coverage at non rRNA covering regions. As mentioned, rRNA locus information was obtained from the corresponding genome feature table. Here, reads were filtered according to which type of rRNA, 5S, 16S or 23S, they were mapped to (Glaub et al., 2020). For all three types read length distributions were calculated as already mentioned to detect if specific read lengths are referring to one type of rRNA. This would be of special interest for 5S rRNA as this type mostly is not targeted during kit-based depletion but might potentially be lower by adapted size selection (Glaub et al., 2020).

Similar to the mentioned analysis, the upstream 5'-UTR region was analysed for read length variation. As the reported length for sequences within SD like motifs is between 28 to 40 (Buskirk & Green, 2017; Li et al., 2012), an analysis was performed on reads ranging from 24 to 40 nucleotides in length mapping in a region of 25 nucleotides down- and upstream of the start position. Therefore, samples with less variation in their read length not covering the upper analysis limit were excluded, resulting in 30 samples for analysis (Glaub et al., 2020). Again, information for the location of interest, here the start position of annotated genes, was extracted from the corresponding feature table. The location for the 5'-UTR region was obtained by addition of 25 nucleotides upstream from the genes' start position (Glaub et al., 2020). For the analysis of the region directly after the start position, 25 nucleotides downstream of this location were analysed. Correspondingly to the previous read length analysis performed, reads mapping the genome position at the defined locations were analysed according to their length. These comparisons were again performed with awk scripting and bedtools' intersect option. A specific setting in bedtools was chosen for the minimum overlap required, as for this analysis a clear categorization of reads was necessary. Therefore, at least 55% of a read had to map in either of the locations to be categorized as such, hindering reads mapping in both locations to be counted twice. Additionally, to avoid interference with operon like structures and potentially biased read length due to another specific pattern at stop regions, a second filtering step was performed. Reads were analysed whether at least 55% were mapped to an adjacent gene and if so, these positions were excluded from the analysis. According to the first analysis, read length distribution was calculated within a sample, followed by median estimation per length between samples analysed (Glaub et al., 2020). The difference between the median values at each analysed length was calculated to highlight the potential alteration (Script 6). As verification for

potential location-specific read length, an analysis of reads covering the whole genes analysed and their stop regions was performed. Here, the stop region was defined spanning an area from the stop position to 25 nucleotides upstream. The space between the start and stop region is referred to as 'whole gene' (Glaub et al., 2020). These areas were analysed as start and upstream region.

```
cat samples.txt | while read -r ribo accession
do
#set variables:
bam_file="$ribo".sorted.bam
feature_table="$accession"_feature_table.txt
echo $bam_file

# extract region 25 nt upstream of gene
cat $feature_table |  awk '{if ($1=="gene" && $2=="protein_coding" &&
$10=="+") \
print $7 "\t" ($8-25) "\t" $8 "\t" "gene" "\t" "0" "\t" $10 ;
else if ($1=="gene" && $2=="protein_coding" && $10=="-") print $7 "\t" $9
"\t" ($9+25) "\t" "gene" "\t" "0" "\t" $10}' |
sort -k1,1 -k2,2n > "$ribo"_SD_region_gene.bed ;

# extract region 25 nt downstream from start
cat $feature_table | awk '{if ($1=="gene" && $2=="protein_coding" &&
$10=="+") \
print $7 "\t" $8 "\t" ($8+25) "\t" "gene" "\t" "0" "\t" $10 ;
else if ($1=="gene" && $2=="protein_coding" && $10=="-") print $7 "\t" ($9-
25) "\t" $9 "\t" "gene" "\t" "0" "\t" $10}' |
sort -k1,1 -k2,2n > "$ribo"_start_gene.bed ;

# prepare bed file out of bam file necessary for bedtools command
bedtools bamtobed -i $bam_file | awk '{print $6}' > "$ribo"_tmp.txt
samtools view -F 4 $bam_file | awk '{print $3 "\t" $4 "\t"
($4+(length($10))) "\t" $1 "\t" "0" }' > "$ribo"_tmp2.txt
paste "$ribo"_tmp2.txt "$ribo"_tmp.txt > "$ribo"_bam.bed ;
rm "$ribo"_tmp2.txt "$ribo"_tmp.txt ;

# include exclusion step from adjacent in upstream region
# creat bed file containing all gene information
cat $feature_table | awk '{if ($1=="gene" && $2=="protein_coding") print $7
"\t" $8 "\t" $9 "\t" "gene" "\t" "0" "\t" $10 }' |
sort -k1,1 -k2,2n > "$ribo"_tmp.bed

# create filtered upstream region bed file
bedtools intersect -s -f 0.55 -sorted -bed -a "$ribo"_SD_region_gene.bed -b
"$ribo"_tmp.bed -v > "$ribo"_SD_region_filtered.bed

# read distribution upstream of start region
bedtools intersect -u -s -f 0.55 -sorted -bed -a "$ribo"_bam.bed -b
"$ribo"_SD_region_filtered.bed | awk '{print $3-$2}' | sort | uniq -c >
"$ribo"_SD_region_reads.txt

# read distribution at start region
bedtools intersect -u -s -f 0.55 -sorted -bed -a "$ribo"_bam.bed -b
"$ribo"_start_gene.bed | awk '{print $3-$2}' | sort | uniq -c >
"$ribo"_start_gene_reads.txt
done
```

Script 6: Script used to compare different read lengths present at certain loci. Here, the comparison between the 5'-UTR upstream region (SD-region) is shown, as well as the start region (located 25 nucleotides downstream of the translation start point).

Analysis regarding potentially improved detection of OLGs due to chloramphenicol application was tested for differential expression types of genes. Each category includes ten annotated genes with similar RPKM values (high = 1,000 - 3,000; medium = 100 - 250; low = 10 - 20) in all eight samples analysed, where four were treated with chloramphenicol whereas the remaining ones were controls (Glaub et al., 2020). First, for each annotated gene detected RPKM and coverage were calculated to identify ten genes that showed a similar RPKM within the eight samples analysed. Again, the information for the gene locations was obtained from the corresponding feature table using awk scripting. Then, the number of mapped reads and the average read depth was calculated. With bedtools, the coverage for each gene location was calculated considering each read for analysis with an overlap of at least 50% in the locations of interest. The ten genes per expression level were obtained by comparison of RPKMs for each gene detected within the ten samples analysed. For each read the assumed p-site location defined by the position 15 nucleotides upstream of 3'- read end was calculated to ensure each read was just counted once (Script 7) (Glaub et al., 2020). Therefore, reads were obtained from the sorted bam files that were mapping in a region covering the start region of the genes of interest. Here, the start region includes the start position of the identified top ten genes per expression level to 50 nucleotides downstream of this location (Glaub et al., 2020). The assumed p-site location per read was identified and counted followed by normalization by corresponding read depth. The sum of the remaining reads per treatment (chloramphenicol vs. control) was calculated and divided by sample size (n = 4). A sum signal to detect a potential read pattern within each expression level was built by combining the pattern of all ten genes.

```
# detect p-site (subtract 15 from 3'-end from each read)
# exemplary for low expressed gene group
for bam in sample1.sorted.bam sample2.sorted.bam sample3.sorted.bam
sample4.sorted.bam ... ;
do
cat low_cov-starts.txt | while read -r chromosome first last strand ;
do
if [ $strand == "+" ] ;
then
samtools view -F 0x10 $bam -b $chromosome:$first-$last > $bam-$first-low-
fwd.bam ;
bam2=$bam-$first-low-fwd.bam ;

elif [ $strand == "-" ] ;
then
samtools view -f 0x10 $bam -b $chromosome:$first-$last > $bam-$first-low-
rev.bam ;
bam2=$bam-$first-low-rev.bam ;
fi ;
bedtools bamtobed -i $bam2 | awk '{if (($3-$2)>=20 && ($3-$2)<=40 &&
$6=="+") print $3-15; else if (($3-$2)>=20 && ($3-$2)<=40 && $6=="-") print
$2+15 }' | sort | uniq -c | sort -gk2,2 | awk '{print $2 "\t" $1}' > $bam2-
count.txt ;
if [ $strand == "+" ] ;
then
awk '{if ($1>=("'$first'"+0) && $1<=("'$first'"+50)) print }' $bam2-
count.txt > $bam2-pos-final.txt ;
elif [ $strand == "-" ] ;
then
awk '{if ($1<=("'$last'"+0) && $1>=("'$last'"-50)) print }' $bam2-count.txt
> $bam2-neg-final.txt ;
fi
done
done
```

Script 7: Script for p-site estimation (15 nt upstream of 3' read end) within each read and subsequent sequence location assignment.

All analyses mentioned so far were based on the RIBO-Seq data compilation focussing on different *E. coli* K12 experiments. The experiments were chosen based on the same analysed organism and medium to ensure that potential translatome differences were not biased by altering input. The analyses performed next are especially focussed on potential differences due to a selection of different prokaryotes.

2.1.6 Prediction of eORFs with DeepRibo and an in-house script (ORFFinder)

The subsequent analyses were performed on the subset of RIBO-Seq data collected that is used for the detection of eORFs in various prokaryotic species.

DeepRibo is designed to predict potential ORFs based on RIBO-Seq obtained results. Different to REPARATION this tool uses the combination of an area covering 30 nucleotides of the translation initiation site and the whole ORF length, exceeding 50 nucleotides up- and 20 nt downstream of the ORF (Clauwaert et al., 2019). Similar to REPARATION, the same start codons are considered in the

identification of potential ORFs. As a threshold for potential functionality minimum length of 30 nucleotides has to be exceeded for an ORF to be predicted (Clauwaert et al., 2019). Input for the prediction are trimmed, filtered and sorted bam files which are used to calculate the coverage over each possible ORF as well as the coverage only at the 5'-end. Followed by minimal RPKM and coverage calculation per sample, to which subsequently predicted ORFs are compared to and ranked according to their values. Results for predicted ORFs are stored in csv files which are then, similar to the REPARATION output, categorized with awk programming (see Script 8). The rank score assigned during processing is used in an exponential function to transform the log value of the prediction score assigned by DeepRibo, and ORFs having a higher re-calculated score greater than or equal to 0.5 are considered translated and are used in further filtering steps. Again, the genomes' feature tables are used to obtain information from annotated genes which are utilised to search for embedded OLGs in DeepRibo prediction output. Predictions made can be located at the same position with different start codons, therefore these are length variations of the same ORF. Based on the prediction rank assigned by DeepRibo per ORF just the ORF with the highest values for different length variations was chosen for further analysis. The output file from these filtering steps contains amongst other things the locus information, such as start and stop position as well as strand specificity for the eORF and the mother gene. Besides filtering the predicted eORFs from the DeepRibo output their location in relation to the mother gene was calculated. This analysis is explained in section 2.1.8.

```
## extract eORF information out of DeepRibo csv files
## $11=start_site, $13=stop_site, $6=strand, $18=prediction value
chromosome="$chromosome_number".fna
for csv in *.csv ;
do
cat $csv | awk -F "," '{print $11 "\t" $13 "\t" $6 "\t" $18}' | tail -n +2
|
awk -F "\t" '{ if ($1<$2) print $1 "\t" $2+2 "\t" $3 "\t" $4 ; else print
$2-2 "\t" $1 "\t" $3 "\t" $4 }' |
awk -F "\t" '{print $0 "\t" 1/(1+exp(-($4)))}' | awk -F "\t" '{if ($5>0.5)
print}' | tee "$csv"_predgenes.txt | while read -r first last strand pred
score ;
do
cat "$accession_number"_feature_table.txt |
awk -F "\t" '{if ($1=="gene" && $2=="protein_coding" && $7=="'$chromosome'"
&& $8<=("'$first'"+0) && $9>=("'$last'"+0) && $10!="'$strand'" &&
"'$strand'"=="+") print "'$first'" "\t" "'$last'" "\t" "'$strand'" "\t" $8
"\t" $9 "\t" $10 "\t" "||" "\t" "'$last'" "\t" "'$score'" "\t" "'$strand'"
"\t" ("'$last'"-"'$first'"); else if ($1=="gene" && $2=="protein_coding" &&
$7=="'$chromosome'" && $8<=("'$first'"+0) && $9>=("'$last'"+0) &&
$10!="'$strand'" && "'$strand'"=="-") print "'$first'" "\t" "'$last'" "\t"
"'$strand'" "\t" $8 "\t" $9 "\t" $10 "\t" "||" "\t" "'$first'" "\t"
"'$score'" "\t" "'$strand'" "\t" ("'$last'"-"'$first'")}' ;
done | tee "$csv"_eORFs.txt ; done
```

Script 8: Filtering script for DeepRibo output. Results from the prediction tool are contained in a csv file per sample. All files are then analysed according to eORFs of interest, exceeding prediction score of 0.5 to be considered.

An in-house script (called ORFFinder, written by Dr. Christopher Huptas) is the third option used for prediction of ORFs. Contrary to the other two tools, ORFFinder in general predicts every ORF possible

within the six frames of a genome. Subsequent filtering is necessary to differentiate between potential translated ORFs and background noise. Predictions of ORFs are based on a minimum ORF length of 93 nucleotides starting with a variety of start codons summarized in the bacterial, archaeal and plant plastid code (transl_Table=11; https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG11). Again, first, a comparison between all ORFs predicted to a file generated by extracting information of the location of protein-coding genes is used to filter for potential embedded ORFs within the output. Next, embedded ORFs are removed if they are adjacent to an annotated gene with less than 100 nucleotides in distance. This should prevent them from being some sort of ribosomal read through signal which is classified as background noise. Additionally, ORFs predicted must exceed a threshold of read coverage normalized to the sample itself. Therefore, the read count of mapped reads is obtained with samtools' flagstat option and divided by 1.000.000. The average read number acts as the threshold value unique for each sample analysed. Another criterion for ORFs potentially being translated is that their coverage value is greater than or equal to 0.6 with an RPKM ≥ 10. If ORFs different in length but within the same region remain, a required overlap of at least three nucleotides is chosen to classify them to be located within the same region. In that case, the longest eORF possible is chosen for further evaluation. In the following when the ORFFinder output is mentioned it refers to the results after filtering and matching their thresholds such as minimal length, read count and coverage.

```
# set variables
cat samples.txt | while read -r ribo accession ;
do
genome=${accession%_*}_genomic.fna ;
bam="$ribo".trimmed.filtered.sorted.bam ;
RPKM=10 ;
coverage_proportion=0.6 ;
filter_distance=100 ;

# ORFs within this distance from annotated genes on the same strand are not
counted as potential OLGs
# PRE-PROCESSING INPUTS:
test ! -d tmp && mkdir tmp ;
faidx -x $genome ;
chromosome=$(cat $genome | head -1 | awk '{print $1}' | sed -e "s|>||g") ;
samtools faidx $chromosome.fna ;
cat $chromosome.fna.fai | head -1 | awk '{print $1 "\t" $2}' >
tmp/"$chromosome"-genome.txt ;
feature_table=${genome%_*}_feature_table.txt ;
minimum_length=93 ;

# 30 amino acids + stop
# functions:
linear () { awk '!/^>/ { printf "%s", $0; n = "\n" } /^>/ { print n $0; n =
"" } END { printf "%s", n }'; } ;

# find annotated genes, from NCBI feature table:
cat $feature_table | awk -F "\t" '{if ($1=="gene" && $2=="protein_coding"
&& $7=="'$chromosome'" ) print}' |
awk -F "\t" '{ print $7 "\t" $8-1 "\t" $9 "\t" "gene_"NR "\t" "0" "\t"
$10}' | awk -F "\t" '{if ($2<$3) print}' > ${genome%_*.*}_genes.bed ;
```

```
# find all ORFs
# note - genome positions can be 0 or 1-based depending on file format and
genome length counting therefore adjustment may be necessary for later
comparison
perl ORFFinder.pl --code 11 --min $minimum_length "$chromosome".fna
tmp/"$chromosome"-ORFs.txt ;
cat tmp/"$chromosome"-ORFs.txt | tail -n +2 |
awk -F "," '{print $1 "\t" $2-1 "\t" $3 "\t"
"'$chromosome'"_ORF_family_"$8 "\t" "0" "\t" $4}' >
tmp/${genome%_*.*}_ORFFinder.bed ;
bedtools getfasta -s -fi "$chromosome".fna -bed
tmp/${genome%_*.*}_ORFFinder.bed > tmp/"$chromosome"-ORFs.fa ;
cat tmp/$chromosome-ORFs.fa | seqkit translate | linear > $chromosome-
ORFs_aa.fa ;

# embedded ORFs:
bedtools intersect -S -f 1 -wo -a tmp/${genome%_*.*}_ORFFinder.bed -b
${genome%_*.*}_genes.bed |
tee tmp/${genome%_*.*}_eORFs_overlaps.txt |
awk -F "\t" '{print $1 "\t" $2 "\t" $3 "\t" $4 "\t" $5 "\t" $6}' >
${genome%_*.*}_eORFs.bed ;

# filter to remove those near annotated genes, on the same strand
bedtools window -w $filter_distance -v -sm -a ${genome%_*.*}_eORFs.bed -b
${genome%_*.*}_genes.bed |
sort -k1,1 -k2,2n > tmp/${genome%_*.*}_eORFs-filtered.bed ;
reads=$(samtools flagstat $bam | grep mapped | grep -v mate | awk '{print
$1}' ) ;
millions=$(echo "$reads / 1000000" | bc -l) ;

# coverage calculation
bedtools coverage -s -F 0.5 -sorted -a tmp/${genome%_*.*}_eORFs-
filtered.bed -b $bam -g tmp/$chromosome-genome.txt |
tee tmp/tmp-cov.txt |
awk -F "\t" '{if ($7>0)
print "'$chromosome'" "\t" $2 "\t" $3 "\t" $4 "\t" $5 "\t" $6 "\t" $7 "\t"
$7/(($9/1000)*"'$millions'") "\t" $NF;
else print "'$chromosome'" "\t" $2 "\t" $3 "\t" $4 "\t" $5 "\t" $6 "\t" $7
"\t" $7 "\t" $NF}' |
sort -rgk5,5 > tmp/${bam%.*}_eORFsf_coverage.bed ;

# filtering samples:
# get out-of frame overlaps - from each full set of overlapping ORFs pick
# regions: at least 3 bp of overlap required to count as same region
cat tmp/${bam%.*}_eORFsf_coverage.bed |
awk -F "\t" '{if (($3-$2)>=93 && $7>="'$millions'" && $9>=0.6) print $1
"\t" $2 "\t" $3 "\t" $4 "\t" $5 "\t" $6}' |
sort -k1,1 -k2,2n > ${bam%.*}_candidates.bed ;
bedtools merge -d -3 -s  -c 6 -o distinct -i ${bam%.*}_candidates.bed |
awk -F "\t" '{print $1 "\t" $2 "\t" $3 "\t" "region_"NR "\t" "0" "\t" $4 }'
> ${bam%.*}-candidate-regions.bed  ;
```

```
#find all overlaps with each "region" - pick one with best start region
bedtools intersect -wao -a ${bam%.*}_candidates.bed -b ${bam%.*}-candidate-
regions.bed | tee tmp.txt |
awk -F "\t" '{print $10}' | sort | uniq | while read -r region ;
do cat tmp.txt | awk -F "\t" '{if ($10=="'$region'") print $4}' | while
read -r family ;
do cat tmp/${bam%.*}_eORFsf_coverage.bed |
awk -F "\t" '{if ($4=="'$family'") print }' ; done | sort -k7,7rg | head -1
; done |
awk -F "\t" '{if ($7>=1) print}' |
awk -F "\t" '{if ($6=="+") print $2 "\t" $4 "\t" $6; else if ($6=="-")
print $3 "\t" $4 "\t" $6}' |
while read -r start family strand ;
do cat tmp/${bam%.*}_eORFsf_coverage.bed |
awk -F "\t" '{if ($6=="+" && $2=="'$start'" && $4=="'$family'") print ;
else if ($6=="-" && $3=="'$start'" && $4=="'$family'") print}' ; done |
awk -F "\t" '{if ($7>="'$millions'") print}' > ${bam%.*}-candidates-
filtered.bed ;

#apply thresholds
awk '{if ($8>="'$RPKM'" && $9>="'$coverage_proportion'") print $0}'
${bam%.*}-candidates-filtered.bed > ${bam%.*}_RPKM_coverage_filtered.bed ;
done
```

Script 9: Script with implemented in-house ORFFinder script. Thereafter, all ORFs predicted are compared to annotated locations, ORFs embedded in these are stored into a new file (eORFs-filtered). Next, coverage for these is calculated and thresholds of RPKM ≥ 10 and coverage ≥ 0.6 are applied. Remaining eORFs are subjects of further analyses.

Additional, during the filtering of ORFFinder results, ORF families were estimated, needed for genome characteristic comparison. Within the location of an annotated gene, the possibility of various overlaps exists. These are attributed to the same ORF family if their ORFs share the same stop codon. Length variations for these overlaps are based on various start codons upstream of the shared stop codon. Even if these variations could be considered as isoforms of the same potential gene, they are not of interest for the comparison analysis. Therefore, they are considered as one ORF family.

A potential verification for eORFs being considered of interest is if they are re-occurring within multiple samples per species. Even more convenient is their prediction ability based on both methods mentioned. The re-occurrence of eORFs within each tool was analysed differently as the prediction efficiency highly varied between the two methods. DeepRibo was sparse in its efficiency in comparison to the ORFFinder output, so a manual comparison between the predicted eORFs was made. To combine the considerably higher amount of predictions based on the ORFFinder script a comparison was made with awk scripting. Therefore, all files were concatenated into one followed by sorting the file according to the numeric value in column 1. Re-occurring entries within this file representing the same ORF being predicted within the different samples were counted. Adjusted to the sample amount available per species a threshold was chosen for an ORF to be of interest depending on how many different samples the ORF was predicted. Additionally, not only the re-occurrence within the prediction of one analysis but also a comparison of eORFs predicted by both methods was performed manually. A general workflow for the detection of eORFs present in multiple samples within one species can be seen in Script 10.

```
# filter ORFFinder results for reoccurrence of eORFs within one species
cat *_all_info_eORF.txt >> combined_all.txt ;
cat combined_all.txt |  awk '{print $1 "\t" $2 "\t" $3 "\t" $4 "\t" $5 "\t"
$6 "\t" $11}' | sort | uniq -c > restricted_all.txt ;

# threshold for reoccurrence: ORF should be detected in at least specific
number of samples available per species (adapted to number of samples)
awk '{if ($1 >= "$threshold_number") print $0}' restricted_all.txt >
threshold_eORF.txt
```

Script 10: Script combining eORF predictions made within a species and subsequent filtering for the once re-occurring (exceeding threshold number which is the threshold for re-occurrence amount).

After the prediction of eORFs was made for the different prokaryotic species, another analysis focused on whether genomic differences might potentially influence characteristics of eORFs predicted and the prediction efficiency.

2.1.7 Influence of genome characteristics on OLG predictions

After eORF prediction for each sample with both DeepRibo and the ORFFinder script, a comparison of eORF length and genome characteristics were made. Based on the assumption that higher GC content may favour longer eORFs, they were analysed according to their length and compared to the corresponding GC content of their genome. As only a few predictions of eORFs were made by DeepRibo with filtered output files containing their length the results were compared manually. Results based on the in-house script and subsequent filtering were combined and analysed for ORF length computationally. First, the results from each species were combined and if the same ORF was predicted within several samples its length was only considered once. Next, the eORF lengths were split into increments of 100. The length analysis of eORFs identified and their length division can be seen in Script 11. GC-content information was obtained from the NCBI website for each genome and compared to the length variation.

```
# combine ORFFinder eORF output per species
cat *_all_info_eORF_filtered.txt | awk '{print $1 "\t" $2 "\t" $3 "\t" $4
"\t" $5 "\t" $6 "\t" $10 "\t" $11}' | sort | uniq -c >
"$species"_comparison.txt ;

# analyse eORF length in hundreds
cat "$species"_comparison.txt | awk '{print $9}' | sort -n | awk '{print
(int($1/100)+1)*100}' | sort -g | uniq -c > eORF_length_distribution.txt
```

Script 11: Combination of eORF predictions within a species, re-occurring ORFs are only counted once. Next, lengths calculated by start and stop position are categorized into length groups in increments of 100.

Prediction efficiency was compared per samples between both methods used and if there is a correlation between its amount and potential genome size. This comparison was based on the assumption that in bigger genomes more annotated genes can be found. Considerably, the number of eORFs predicted in bigger genomes was expected to be higher. First, predictions were normalized to the samples' read depth with subsequent comparison to the genome size to detect a potential correlation.

2.1.8 Relative reading frame estimation

One analysis focus on the location of the eORF predicted in relation to its mother gene. This analysis was performed on the DeepRibo predicted eORFs was well as on the output of the ORFFinder filtered output. As this type of analysis is again based on the comparison of genomic positions, awk scripting was used once again. Within the script, the distance between the nucleotides' start position of the mother gene and OLG was used to perform a modulo operation on. The numeric modulo operator is three, as there are three possible frames for location based on the three-nucleotide periodicity dictated by the codon structure. Depending on the remainder after the calculation, the relation between the mother gene and eORF is assigned to each eORF predicted. If the result is zero the relative reading frame is sas11, a remainder of one indicates the relative reading frame sas12, whereas sas13 is assigned it the modulo operation results in two. Here, sas stands for sense (s, location of mother gene) and antisense (as, location for eORF), with their respective frame location. Additionally, per sample, the localisation of eORFs was summed up to analyse a potential trend for one frame being in favour of the creation of overlapping genes.

This type of analysis was repeated on the results based on the ORFFinder script analysis. First, the start and stop position, as well as the strand specificity for the mother gene, had to be extracted from the genomic feature table. Then a modulo operator calculation was again performed of the difference between the two start positions. The same locations were assigned to the eORF according to the calculation mentioned above. The following script is showing the calculations performed with ORFFinder output as an input file.

```
#### HANDLING ORFFINDER OUTPUT in relation to reading frame location
## extract mother gene information from corresponding feature table,
accession is referring to species specifique numeric code
cat "$sample_number".trimmed.filtered.sorted-candidates-filtered.bed |
while read -r chromosome pos1 pos2 family value strand reads rpkm cov ;
do
cat "$accession"_feature_table.txt | awk -F "\t" '{if ($1=="gene" &&
$2=="protein_coding" && $7=="'$chromosome'" && $8<=("'$pos1'"+1) &&
$9>=("'$pos2'"+0) && $10!="'$strand'" && "'$strand'"=="+") print
("'$pos1'"+1) "\t" "'$pos2'" "\t" "'$strand'" "\t" $8 "\t" $9 "\t" $10 "\t"
"||" "\t" "'$pos2'" "\t" "'$cov'" "\t" "'$strand'" "\t" ("'$pos2'"-
("'$pos1'"+1)); else if ($1=="gene" && $2=="protein_coding" &&
$7=="'$chromosome'" && $8<=("'$pos1'"+1) && $9>=("'$pos2'"+0) &&
$10!="'$strand'" && "'$strand'"=="-") print ("'$pos1'"+1) "\t" "'$pos2'"
"\t" "'$strand'" "\t" $8 "\t" $9 "\t" $10 "\t" "||" "\t" "'$pos1'" "\t"
"'$cov'" "\t" "'$strand'" "\t" ("'$pos2'"-("'$pos1'"+1))}' ;
done > "$sample_number"_all_info_eORF.txt ;

for i in *_all_info_eORF.txt;
do
awk '{if ((($2-$4)% 3)==0) && $3=="-") print $0 "\t" "sas11"; else if
((($2-$4) % 3)==1) && $3=="-") print $0 "\t" "sas12" ; \
else if ((($2-$4) % 3)==2) && $3=="-") print $0 "\t" "sas13" ; else if
((($5-$1) % 3)==1) && $3=="+") print $0 "\t" "sas12" ; \
else if ((($5-$1) % 3)==2) && $3=="+") print $0 "\t" "sas13" ; else if
((($5-$1) % 3)==0) && $3=="+") print $0 "\t" "sas11" }' $i |\
tee "$i"_reading_frame.txt | awk '{print $12}' | sort | uniq -c | sort -nr
> "$i"_reading_frame_count.txt ;
done
```

Script 12: For eORFs of interest corresponding mother gene information is extracted from the corresponding feature table. Based on the start positions location relation between overlap and mother gene is calculated.

2.1.9 Phylogenetic analysis

The first analysis performed to potentially describe eOLGs of interest was focused on the detection of the last common ancestor to reveal the potential age of the gene. Re-occurring eORFs found in the comparison performed in section 2.1.6. were subject to the following analyses. First, BLAST searches were performed on the nucleotide level to identify potential homologous sequences within other species. Based on the distribution within the phylogenetic tree in relation to sequence similarity calculations the genes' approximate 'age' could be estimated. Second, protein sequence blast searches were used to potentially detect similar functional sequences in other species. Based on the sequences' similarity and an assigned functionality to the detected homologue an assumption on the eORFs functionality could be made. Additionally, the characteristics of the mother gene were analysed according to whether the function of different mother genes was similar. Sequence extraction was performed with the faidx command for which the genome in fasta format and the genomic start and stop position had to be specified. Depending on whether the nucleotide or protein sequence should be extracted, and on which strand the ORF of interest was located, different flags had to be specified. As the command is strictly working on increasing numbers, ORFs in the minus frame had to be specified according to it but with the flag 'revseq -filter' they were extracted correctly. If the amino acid sequence should be obtained the flag 'transeq -filter' had to be specified. For each of the 43 eORFs of interest both nucleotide and protein

sequence was extracted and subjected to the corresponding blast search. General extraction commands can be found in Script 13.

```
# sequence extraction of antisense location with defined locus positions
faidx "$chromosome".fna "$chromosome":"$pos1"-"$pos2" | revseq -filter >
nucleotide_sequence.fna

# obtain protein sequence using transeq filter
faidx "$chromosome".fna "$chromosome":"$pos1"-"$pos2" | revseq -filter |
transeq -filter > translated_sequence.fna
```

Script 13: Commands used to extract nucleotide and amino acid sequences from eORFs of interest, which are required for further phylostratigraphy analyses performed.

Extracted sequences were used for potential functionality characterisation of eORFs if homologues were found. Comparison of sequence on nucleotide level was performed with blastn option, whereas for protein sequences blastp was used. Both comparisons were conducted with the NCBI blast tool in Linux although the used database is stored on the NCBI server (i.e. remote search). In the tool used access to the server is implemented. Nucleotide level BLAST analysis was used to find homologue's in the phylogenetic tree for gene age estimation, whereas protein BLAST was used to analyse potential functionality based on protein sequence. Additionally, tblastn was another search query performed. Here, the input protein sequence is searched for in a nucleotide database, enabling a higher variety of sequence to search and has better sensitivity than a nucleotide-based search. Only the best 1,000 aligned sequences were obtained from this comparison which was performed within the bacterial species, family, and genus. These results were then forwarded to another evolutionary based analysis.

OLGenie, a program exclusively written for overlapping gene analysis, was used to estimate their selection and potential functionality. Analysed here is the rate of synonymous and non-synonymous nucleotide exchanges whilst considering their impact in mother and overlapping gene (Nelson, Ardern, & Wei, 2020). Necessary input information for the analysis are the aligned tblastn results and the relative reading frame of the overlap. Before analysis with OLGenie, the fasta files had to be adapted to the specific input format required. Characters that were not supported were exchanged with the stream editor sed, replacing a specified string or character by another specified input. Additionally, unique headers were required obtained by comparison with awk scripting and subsequent maintenance of unique ones. Last, an analysis with OLGenie necessitated same length sequences for comparison (to avoid having to perform an additional alignment of those sequences altering in length - a more precise analysis would involve additional alignment for each length differing sequence), hence the most common length sequences were kept for evaluation. The tool calculates the ratio of exchanges for both genes which are then subjected to an additional R script relating to OLGenie. Within this script significance of the obtained values is calculated by iterating over codons. Results from this analysis are indicators for potential functionality of the sequences analysed, as sequences with higher selection pressure are expected to be of function due to the evolutionary 'effort' required for maintaining their functionality. The command-line expression for OLGenie execution can be seen in Script 14.

```
# Script including OLGenie perl script to detect potential selection on
mother gene and overlap
cat samples.txt | while read -r species frame ;
do
OLGenie.pl --fasta_file="$species"_aligned_seq_test.fna --frame="$frame" --
output_file=OLGenie_"$species".tsv --verbose >/dev/null 2>&1 ;
Rscript /usr/bin/OLGenie_bootstrap.R OLGenie_"$species".tsv 3 1000 4 | tee
bootstrap_"$species".txt |
awk '{print $10 "\t" $12 "\t" $16 "\t" $23}' > "$species"_info.txt ;
done
```

Script 14: Wrapper Script around OLGenie perl script with subsequent bootstrap analysis performed. Here, a significance regarding potential selection pressure is calculated. Settings for OLGenie used are: minimum number of defined codons per codon position = 3, number of bootstrap replicates = 1000, number of threads used = 4.

The last analysis regarding the preliminary characterisation of the eORFs of interest is called Frameshift. In general, the length of an overlap detected is compared to possible length variations based on random eORF creation (Schlub et al., 2018). The nucleotide sequences of the respective mother genes were extracted, then randomly shuffled using a public R script (https://github.com/TimSchlub/Frameshift). Here, the triplet structure is kept during shuffling but a new random succession of these leads to changed open reading frame structures with consequently changed embedded sequence length. If those artificial created ORFs were overall significantly shorter than the originally detected one this is an indication of functionality, or selection for the long eORF.

```
# input = fasta file containing nucleotide sequence of mother genes
# mother genes are shuffled to analyse ORFs created in alternative frames
according to their length
# 'linear' converts fasta sequence to one line of sequence per header
(posted online by Pierre Lindbaum)
linear () { awk '!/^>/ { printf "%s", $0; n = "\n" } /^>/ { print n $0; n =
"" } END { printf "%s", n }'; } ;
for gene in *.fna ;
do
geneseq=$(cat $gene | linear | awk '{if (NR==2) print }') ;
Rscript Frameshift_20000_revcom0.r $geneseq |
awk '{if (NF==7 || $1~"#") print }' > ${gene%.*}-frameshift.txt ;
done ;
```

Script 15: Script shown includes Frameshift R script from (Schlub et al., 2018). Each input fna-file contains nucleotide sequence of mother gene from the embedded ORF of interest. Analysed were the same mother gene and eORFs that were subject to OLGenie analysis.

2.2 Experimental Proceedings and Equipment

2.2.1 Chemicals

A complete table containing all used chemicals for this experiment are listed in Supplementary Table S1. Ready-to-use kits needed for this experiment are shown in Supplementary Table S2 with their purpose and provider.

2.2.2 Buffers and solutions

Table 2: List of buffers necessary for RIBO-Seq and RNA-Seq preparation.

Buffer	Ingredient	End concentration
Polysome Lysis Buffer (PLP) without ions	TRIS-HCl pH8	20 mM
	NH_4Cl	100 mM
	Tergitol	0.1%
	Triton-X-100	0.4 %
Polysome Lysis Buffer (PLP)	TRIS-HCl pH8	20 mM
	$MgCl_2$	20 mM
	NH_4Cl	100 mM
	$CaCl_2$	10 mM
	Tergitol	0.1%
	Triton-X-100	0.4 %
Polysome Gradient Buffer (PGP)	TRIS-HCl pH 8	20 mM
	$MgCl_2$	10 mM
	NH_4Cl	100 mM
	DTT	2 mM
50 x Tris Acetate EDTA (TAE)	TRIS	2 M
	Acetic acid	1 M
	Na_2EDTA	50 mM
10 x Tris Borate EDTA (TBE)	TRIS	1 M
	Boric acid	1 M
	Na_2EDTA	20 mM
Gel Extraction Buffer	NaOAc pH 5.5	300 mM
	EDTA	1 mM
Sucrose solution	Sucrose	50 %
	PGP	50 %

2.2.3 Enzymes

Table 3 contains used restriction enzymes for polysome structure digestion, as well as its inhibitor (Superase IN) and the enzyme used for DNA digestion. Providers of these enzymes are Invitrogen (Carlsbad, CA, USA), Lucigen (Middleton, WI, USA), New England Biolabs (NEB; Ipswich, MA, USA) and Thermo Fisher Scientific (Waltham, MA, USA) and.

Table 3: List of enzymes used in several RIBO-Seq and RNA-Seq processing steps. The first for enzymes are used in RNA digestion for RIBO-Seq samples, which are inhibited by application of SUPERase In. TURBO DNase is used in DNA digestion.

Enzyme	**MNase**	**XRN-1**	**RNase R**	**RNase T**	**SUPERase IN**	**TURBO™ DNase**
Unit	300 U/µl	1.000 U/ml	20 U/µl	5.000 U/ml	20 U/µl	1000 U, 2U/µl
Provider	Thermo Fisher	NEB	Lucigen	NEB	Invitrogen	Invitrogen

In Table 4 the listed enzymes are used for de- and phosphorylation (phosphatase + T4 ligase), as well as the enzymes necessary for library preparation (ligases + reverse transcriptase).

Table 4: The first two enzymes listed are used to ensure the same phosphate status at all mRNA fragments before library preparation. The latter two enzymes mentioned are needed for within Illumina base library preparation.

Enzyme	**Antarctic Phosphatase**	**T4 polynucleotide (PNK) Ligase**	**T4 RNA Ligase 2, truncated**	**SuperScript II Reverse Transcriptase**
Provider	NEB	NEB	NEB	Invitrogen

2.3 Methods

2.3.1 Strain and cell harvest

Harvested culture pellets of *B. thetaiotaomicron* VPI-5482 were provided by Hannes Petruschke, Helmholtz Institute Leipzig as part of a collaboration (50 falcons each containing one pellet). Before harvest, the culture was grown anaerobically at 37 °C in Brain-Heart-Infusion (BHI) medium for 72 hours. Unfortunately, no further precautions reassuring the ribosomal stalling necessary for RIBO-Seq were performed before or during cell harvest.

2.3.2 Cell lysis

For comparison purposes, two RIBO-Seq and corresponding RNA-Seq approaches are conducted simultaneously. For each sample (n = 2), two pellets are resuspended in 375 µl PLP buffer (without ions) to increase the experiment's input material. Droplets of the obtained suspension are pipetted into liquid nitrogen followed by transferring them into a metal mortar. Under constant addition of liquid nitrogen to ensure no thawing samples are homogenized using mechanic shear force. Addition of 37 °C pre-warmth PLP buffer (750 µl, with double ion content) per sample for rapid thawing is followed by centrifugation for 5 min at 11.000 rpm and 4°C. The supernatant is transferred followed by second centrifugation (same settings) followed by splitting it into the RIBO-Seq (500 µl) and RNA-Seq (300 µl) sample. Nucleic acid extraction of the RNA-Seq samples is performed as quality control of the input material.

2.3.3 Nucleic acid extraction via Trizol/chloroform precipitation

Extraction is always performed on ice unless stated otherwise. The sample is split into 200 µl and 100 µl followed by adding 1 ml Trizol and incubation for 5 min at room temperature (RT). 200 µl (0.2 volumes of added Trizol) pre-cooled chloroform are applied with subsequent vortexing for 15 sec followed by incubation for 5 min. Centrifugation is performed for 15 min at 12.000 g and 4°C to separate the RNA. The supernatant of the top layer is transferred, 500 µl isopropanol and 1 µl glycogen are added. Slow inversion of the mixture (five times) is followed by incubation for 30 min. RNA is pelleted by centrifugation for 10 min at 12.000 g and 4°C. The supernatant is discarded whilst the pellet is washed by addition of 1 ml 80 % ethanol without resuspension. Centrifugation is repeated with the same settings followed by a second wash step with subsequent centrifugation. The supernatant is discarded, the nucleic acid pellet is dried for 10-15 min at RT, then resuspended in 30 µl RNase free H_2O (combine the split approaches).

2.3.4 Nucleic acid concentration measurement using Nanodrop

RNA quantification is measured by Nanodrop analysis based on spectrophotometer measurements using 1 µl of extract. The used tool can provide information about concentration and purification, in this case, of RNA but not value for intactness of the nucleic acid. RNA and other nucleic acids absorb light at 260 nm, whereas proteins or other contaminants absorb at 280 nm. Hence, the ratio of these two values can be used as an indicator for the pureness of the sample. The analysis is performed according to the manufacturers' instructions.

2.3.5 Size separation with gel electrophoresis

Dissolve 0.75 g agarose in 50 ml 1 x TAE buffer by heating the suspension. Add 1 µl GelRed® nucleic acid to the suspension. Mix sample with 2 x RNA loading dye, then pipet them into wells of the hardened gel. As a reference, a 1 kb and a 100 bp ladder are applied to the gel. The size separation obtained by electrophoresis is due to the slower movement of molecules through the gel based on their molecular size. Current flow is possible by covering the gel with 1 x TAE buffer. Separation takes place during 30-45 min at 110 volts followed by visualising the results under UV light. Based on the success of RNA extraction the analysis of the RIBO-Seq samples will be performed.

2.3.6 Nucleic acid quality control analysis using capillary gel electrophoresis

In comparison to concentration measurements with the Nanodrop (see 2.3.4.), the Bioanalyzer uses gel electrophoresis techniques to assess the quality of the samples. Hence, Bioanalyzer results can give information about the intactness or degradation of the input material. For RNA analysis the ribosomal subunits 16S and 23S are analysed. The proportion of these is used to estimate the RNA integrity number (RIN), where > 7 is considered intact RNA. The analysis is performed according to the manufacturers' instructions for the Bioanalyzer RNA 600 Nano Kit.

2.3.7 RNA digestion followed by density centrifugation for RIBO-Seq samples

Preparation of the RIBO-sample starts with RNA digestion to ensure polysome structures are broken down to monosomes. Digestion is performed using endo- and exonucleases for sufficient separation of the ribosomes (Gerashchenko & Gladyshev, 2017). Per sample following RNA nucleases are used:

- 375 U MNase
- 2.5 U XRN-1
- 25 U RNase R
- 6 U RNase T

Digestion is completed by adding 1 mM $CaCl_2$ and NEB4 buffer with incubation for 1 h at RT. Inhibition of digestion is ensured by first adding 0.6 µl of 0.5mM EDTA to bind MNase, followed by addition of 3.75 µl SUPERase IN to inactivate the other enzymes. The complete sample is transferred to the density centrifugation approach necessary to only collect monosomes for further proceedings. Thereby, molecules remaining in the sample are separated by their molecular weight. Different phases necessary for density centrifugation are based on different ratios of a sucrose solution and polysome gradient buffer (PGP). The ratio with its corresponding phase is listed in Table 5. Starting with the highest ratio special

centrifugation utensils are filled up in descending order. The phase of interest containing the monosomes is located between 25 % and 30 %. Therefore, the 30 % phase will be stained blue, while the 25 % phase will be dyed yellow, resulting in a green phase of interest in between. For both dyes, stock solutions are prepared by dissolving 0.03125 g of the respective colour granulate in 1.56 ml PGP. The working solutions of dyes are diluted 1:50 and will replace the PGP in the mentioned phases. Centrifugation is conducted for 3 h at 28.000 rpm and 4°C.

Table 5: Overview of sucrose density layer composition. Layers are made off differing concentration mixtures of sucrose and polysome gradient buffer (PGB). PGB is changed by dye for two layers, namely 25 % (yellow dye) and 30% (blue dye). The resulting green layer after centrifugation is of interest containing monosomes necessary for subsequent sequencing.

Gradient	**10 %**	**15 %**	**20 %**	**25 %**	**30 %**	**35 %**	**40 %**	**45 %**	**50 %**
Sucrose [ml]	0.8	1.2	1.6	2	2.4	2.8	3.2	3.6	4
PGB [ml]	3.2	2.8	2.4	-	-	1.2	0.8	0.4	-
Dye [ml]	-	-	-	2	1.6	-	-	-	-

The bottom of the centrifugation utensil is punctured with a needle (0.55 x 25 mm) allowing a slow collection of coloured phased droplets in a microtiter plate. Green droplets are united for subsequent RNA extraction, performed as described in section 2.3.3., only adapted to elution of extracted RNA in 50 µl RNase free H_2O.

2.3.8 Footprint size selection through urea gel excision

RIBO-Seq intends for sequencing only ribosome protected footprints. Due to the characteristics of this here used urea gel, RNA fragments can be separated very precisely. RNA fragments of specific sizes are used as markers on the urea gel aiding in the excision of the fragments with a length of interest. Ingredients for one 15 % urea gel can be found in Table 6.

Table 6: Ingredients for 15 % urea gel used for in size selection step during RIBO-Seq processing.

Ingredient	**Sequencing thinner [ml]**	**Sequencing concentrate [ml]**	**Sequencing buffer [ml]**	**TEMED [µl]**	**10 % APS [µl]**
Amount	5.2	12.8	2	15	150

Before transferring the samples onto the gel, a 20 min pre-run at 200 V with 1 x TBE buffer as a current flow is necessary to ensure an even distribution of heat in the gel. RNA concentration of 2.5 µg per well should not be exceeded. Specified marker aiding in the exact excision of fragments with a length of interest are also applied to the gel. One marker only contains RNA fragments with a length of 23 nt, the

second marker includes fragments in a length range between 19 to 27 nt. Separation is obtained by running the gel for 90 min at 200 V, followed by staining the nucleic acid by adding 15 µl SYBR™ Gold nucleic acid gel stain and 15 min incubation. Gel excision is performed by orientation at the marker range under UV light. Gel fragments are transferred into punctuated (0.9 x 40 mm) micro reaction vessel placed in a second one. Centrifugation for 2 min at 13.000 rpm (RT) breaks down the gel structure due to centrifugal force pressuring the gel through the punctures. For complete dissolution 400 µl gel extraction buffer and 2 µl SUPERase IN are added to the sample with incubation overnight (ON) at 800 rpm and 20 °C. Subsequently, the samples are transferred onto a cellulose acetate column filter (0.2 µm) followed by centrifugation for 2 min at 10.000 rpm (RT). RNA precipitation is obtained by adding 1 ml ethanol absolute (100 %) and 1 µl glycogen (20 µg/µl) with incubation ON at -80°C. Centrifugation for 20 min at 12.000 g and 4°C pellets the RNA, subsequent the nucleic acid is washed twice with 1 ml 80 % ethanol. After drying the pellet at RT for 10 - 15 min it is resuspended in 15 µl RNase free H_2O. Quality of purified RNA is checked via Nanodrop measurement (see Section 2.3.4.).

2.3.9 DNA digestion, a control 16S PCR and fragment shredding for RNA-Seq samples

To eliminate DNA remaining in the sample, RNA-Seq samples need special treatment. While in the processing of RIBO-Seq samples, DNA is removed during density centrifugation, an extra DNA digestion step is necessary for RNA-Seq. Input material for digestion should not exceed 10 µg per reaction. Addition of 1 µl TURBO™ DNase and 5 µl of its buffer complement the experimental approach. Incubation for 30 min at 37°C is followed by inactivation for 10 min at 65°C through adding 1.5 µl EDTA (0.5 M). EDTA binds metal ions present in the digestion enzyme therefore inhibiting its reactions. The nucleic acid is incubated ON at -80°C by 690 µl ethanol absolute, 27 µl NaOAc (3 M) and 1 µl glycogen (20 µg/µl). Precipitation is obtained by centrifugation for 15 min at 12.000g and 4°C. The nucleic acid pellet is washed twice with 1 ml 80 % ethanol, then dried and resuspended in 16 µl RNase free H_2O.

Digestion efficiency is verified performing a 16S rRNA PCR as control. With the use of specific primer (27F 5'-AGAGTTTGATCCTGGCTCAG-3'; 1492R 5'-TACGGYTACCTTGTTACGACTT-3') targeting the 16S rRNA gene, a band present on the agarose gel would indicate insufficient digestion. Per sample PCR mix is as follows:

- 1 µl Template
- 25 µl Dream Taq Mastermix (Thermo Fisher Scientific)
- 0,5 µl 10 M Forward Primer (27F)
- 0,5 µl 10 M Reverse Primer (1429R)
- 23 µl RNA-free water

Genomic DNA was used as the positive control, RNase free H_2O as negative. The PCR protocol is as follows:

Table 7: PCR protocol for 16S rRNA amplification, here used to verify the prior performed DNA digestion within RNA-Seq samples. Additionally, this protocol will be used to monitor potential contamination of redundant DNA probes after rRNA depletion.

16S rRNA PCR				
	Repeat	**Step**	**Temperature**	**Time**
Denaturation	1	1	95 °C	1 min 30 sec
	35	2	95 °C	30 sec
Annealing		3	58 °C	30 sec
Elongation		4	72 °C	1 min
	1	5	72 °C	5 min
Hold	1	6	8 °C	∞

Digestion success was visualised via gel electrophoresis (see Section 2.3.5.). From the next step on both RIBO-Seq and RNA-Seq samples are treated the same.

2.3.10 rRNA Depletion

The RNA present in a cell consists up to 85 – 90 % of rRNA, therefore sufficient depletion of this type is necessary to decrease its amount before sequencing (Z. Chen & Duan, 2011; Petrova et al., 2017). Depletion is performed using siTools Pan-Prokaryotes Kit according to the manufacturer's instructions. Here, streptavidin coat beads are used to bind rRNA present in the sample, lowering the amount present in the sample. After performing the depletion, DNA digestion was performed as described in section 2.3.9. Probes necessary during the depletion might contaminate the samples, hence an additional digestion step was required.

2.3.11 RNA quantification using Qubit Assay

The Qubit Fluorometer is used to measure nucleic acid concentration. Unlike the Nanodrop, this tool can measure concentrations in the low range from 10 pg/µl to 100 ng/µl. The fluorescent dye in the reagent mix binds to the RNA, hence the emission measurement is synonymous to the nucleic acid concentration in the sample. The sample preparation is performed according to the manufacturers' instructions. Each sample, as well as the two standards, are measured three times and results are averaged.

2.3.12 Covaris Ultrasonicator used for sheading fragment size

RNA fragments present in the RNA-Seq samples are sheared using covaris ultrasonicator to size down their length as no size selection step by gel excision was performed for these. Duration (3 min), intensity (175 W), duty cycle (10 %) and amount (200 cycles) of the ultrasonic impulse are influencing the remaining length of the fragments (here approximately 220 bp). To decrease the amount of liquid nucleic acids are resolved in and consequently increase the RNA concentration per µl they are applied to SpeedVac Vacuum Concentrators. A lower amount of input sample is necessary for the next preparation step which is performed to prepare the nucleic acids for adapter ligation.

2.3.13 Dephosphorylation and subsequent Phosphorylation

To ensure all fragments are phosphorylated at their termini uniformly dephosphorylation with subsequent phosphorylation is performed. In between the two steps and at the end, nucleic acids fragments are purified using the miRNeasy mini kit (Qiagen) according to the manufacturers' instructions. 27 µl Antarctic phosphatase buffer, 2 µl Antarctic phosphatase and 0.5 µl SUPERase IN are added to each sample followed by an incubation step for 30 min at 37°C. After purifying the samples using the mentioned kit, the phosphorylation of all termini is achieved by adding 3.5 µl T4 DNA ligase buffer, 2 µl T4 polynucleotide ligase and 0.5 µl SUPERase IN. Here, the incubation is set for 1 h at 37°C followed by a second purification step. Nanodrop is used for RNA concentration estimation (see Section 2.3.4.). Samples are then evaporated into 5 µl each, as this is the required input material amount for the TruSeq small RNA library kit (Illumina).

2.3.14 Library Preparation and Sequencing

The TruSeq small RNA library kit is used for library preparation of RIBO-Seq and RNA-Seq samples. The protocol comprises steps, such as 3'-adapter ligation, 5'-adapter ligation, reverse transcription of the library from RNA to DNA, amplification of the library including indexing of the samples followed by concentration estimation and quality verification. All steps are performed according to the manufacturers' instructions with following specified adaptations. The indices used to label the different samples before pooling are listed in Table 8. Table 9 contains the PCR protocol for library amplification.

Table 8: Indices with respective sequences used for *B. thetaiotaomicron* RIBO-Seq and RNA-Seq samples.

Sample	**RIBO-Seq I**	**RIBO-Seq II**	**RNA-Seq I**	**RNA-Seq II**
Index	RPI10	RPI11	RPI9	RPI2
Sequence	TAGCTT	GGCTAC	GATCAG	CGATGT

Table 9: PCR protocol used for library amplification in Illuminas TruSeq Small RNA Library preparation.

Library Amplification				
	Repeat	**Step**	**Temperature**	**Time**
Denaturation	1	1	98 °C	30 sec
	15	2	98 °C	10 sec
Annealing		3	60 °C	30 sec
Elongation		4	72 °C	15 sec
	1	5	72 °C	10 min
Hold	1	6	4 °C	∞

Fragments of interest are excised from the gel and subsequently purified. Within the purification step, one adaptation is the incubation ON for breaking the gel debris. The second change is made at eluting the fragments by adding glycogen, NaOAC and 100% ethanol. Here, the incubation is again performed ON. Libraries are concentrated according to following conversion formula, with values for concentration obtained from Qubit measurements and average library size from Nanodrop analysis:

$$concentration\ in\ nM = \frac{(\text{concentration in } \frac{ng}{\mu l})}{(660\frac{g}{mol} \text{ x average library size in bp})}\ x\ 10^6$$

Libraries are tested regarding their quality of covering mostly mRNA fragments in a first sequencing experiment. Therefore, 5 µl of each library (0.5 mM) are pooled and analysed in a 2 x 300 bp paired-end sequencing run.

2.3.15 Data Evaluation

Quality control of raw reads is analysed using FastQC, followed by adapter trimming with fastp. The following adapter sequence is specified for trimming: TGGAATTCTCGGGTGCCAAGG. Reads are aligned to the reference genome (NC_004663.1) or reference assembly file (GCF_000011065) using Bowtie2. Settings for alignment were: -p 6 --quiet -q --end-to-end -D 20 -R 3 -N 0 -L 17 -i S,1,0.50 --no-unal –x. FastQ Screen is used to compare input reads against a standard set of libraries providing information about the percentage of reads mapping to either mRNA, rRNA or tRNA and their uniqueness (mapping only to one genome or multiple) (Wingett & Andrews, 2018). Important settings for FastQ Screen are as follows: --threads 6 --aligner bowtie2 --bowtie2 '-p 6' --subset 0 --tag --force --filter 00003. Afterwards, a second quality control step is performed on the trimmed and filtered reads. Results before and after trimming and filtering are compared to evaluate the number of usable reads. Additionally, the FastQ Screen output is analysed regarding the distribution of reads mapping to either RNA of interest or rRNA and tRNA. An already available RIBO-Seq experiment on *Bacteroides thetaiotaomicron* (Sberro et al., 2019) was analysed identically to compare outputs obtained.

3. Results

3.1 Experimental adjustments in RIBO-Seq experiments

A compilation of sample information, such as their reference, corresponding experiment number and experimental criteria can be found in Table 10.

Table 10: Overview of available RIBO-Seq experiments chosen for comparative analysis. Table from (Glaub et al., 2020).

GEO/Experiment number	Stalling method	Harvesting method	Size selection	Publication
GSE85540	-	rapid filtration	20-40 nt	Hwang et al. 2017
GSE68762	chloramphenicol	centrifugation	n.d.	Bartholomäus et al. 2016
GSE86536	chloramphenicol/linezolid	rapid filtration	28-42 nt	Marks et al. 2016
E-MTAB-2903	chloramphenicol	rapid filtration	~28 nt	Wang et al. 2015
GSE64488	-	rapid filtration	20-40 nt	Woolstenhulme et al. 2015
SRP048921	-	rapid filtration	20-30 nt	Balakrishnan et al. 2014
SRP040142	-	rapid filtration	28-42 nt	Elgamal et al. 2014
GSE61619	erythromycin/telithromycin	rapid filtration	25-42 nt	Kannan et al. 2014
GSE33671	chloramphenicol/ -	centrifugation/filtration	25-31 nt	Oh et al. 2011

Supplementary Table S3 gives an overview of sample-specific characteristics such as raw read number, used adapter sequence for trimming and amount of effective reads after filtering. Reference genomes used for the three *E. coli* K-12 substrains with their respective accession numbers are specified in Supplementary Table S4.

3.1.1 Estimation of necessary read depth for sufficient ORF detection

Based on the results published for RNA-Seq read depth analysis by Haas et al (Haas et al., 2012), here the necessary amount of reads left after adapter removal, alignment and filtering out reads corresponding to rRNA and tRNA was estimated for RIBO-Seq results.

The evaluation of data was based on the comparison of detected annotated genes to the effective read amount left for the prediction (Figure 5). Annotated genes were considered for comparison if they showed coverage of at least 3 reads required for predictions based on REPARATION (Glaub et al., 2020). Effective reads are defined as those that remain for evaluation after adapter sequence removal and successful alignment to the corresponding reference genome not covering either rRNA or tRNA. Not only the ORF predictions made by REPARATION were used for this comparison, but also genes assumed to be translated due to their RCV value. Locations of genes predicted by REPARATION were used to analyse processed RIBO-Seq and RNA-Seq data to obtain read coverage information from both

sequencing approaches for samples where those two techniques were performed. Accepted for comparison to their read depth were only those genes with RCV values equal to or above 0.355 (Glaub et al., 2020). Based on the two comparison approaches a necessary read amount to detect annotated genes sufficiently in RIBO-Seq experiments seems to be at least 20 million effective reads (Glaub et al., 2020). A higher amount of effective reads does not seem to contribute to more predictions made. The increase of reads may give rise to spurious detections which is why the evaluation of sufficient read amount is of especial importance (Glaub et al., 2020; Haas et al., 2012).

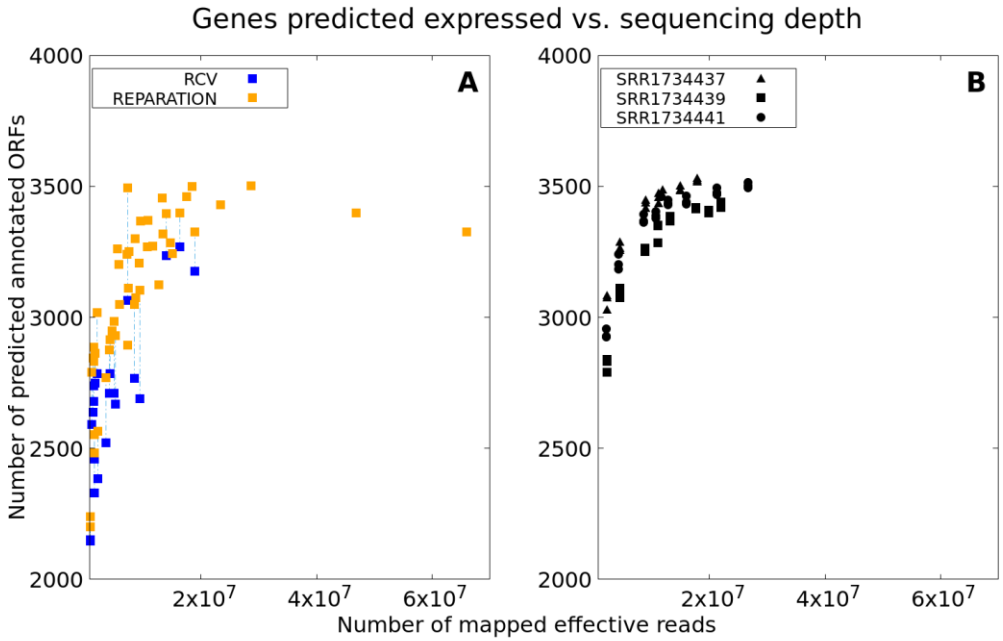

Figure 5: (A) REPARATION based prediction efficiency of annotated genes compared to used effective reads. The threshold for genes being accepted as potentially translated are a minimum of three reads needed (only REPARATION based, orange) or additionally exceeding an RCV ≥ 0.355 (blue). Available analysis results for both criteria representing one sample are connected via dashed lines. (B) Subsampling of high sequencing depth samples (SRR1734437; SRR1734439; SRR1734441) performed in triplicates for each. A comparison of reduced sequencing depth and the number of annotated genes predicted (REPARATION based) was performed. Figure from (Glaub et al., 2020).

To verify the recommended read amount necessary, three samples (SRR1734437, SRR1734439, SRR1734441, (Woolstenhulme et al., 2015)) with remaining high read coverage were used for a second determination. Here, a reduction of reads left used for gene prediction leads to a decreased amount of annotated genes detectable. Nonetheless, in question was if still 20 million reads are needed for sufficient detection or if a certain decreased amount is already sufficient. Again, as can be seen in Figure 5B, if 20 million reads are available for prediction this is a sufficient amount for gene detection. Additional reads are not contributing to more genes being predicted.

However, even if at least 20 million reads were available for prediction, only around 3,500 annotated genes could be detected. These are around 82 % of possible annotated genes known for *E. coli* K-12 (Glaub et al., 2020). For RNA-Seq analysis, the detection efficiency was as high as only 2 out of the hitherto described 4,149 annotated genes were not detected (Haas et al., 2012). Nevertheless, for the

mentioned analyses only one read was sufficient for ORF detection, whereas in this study, a minimum of at least three reads was required (Glaub et al., 2020).

Next, reads used for prediction were analysed according to their length and whether there is a trend detectable for specific read length regarding RNA type.

3.1.2 Different RNA types have specific read lengths

The primary aim of the RNA type-specific read length analysis was to potentially deplete for types of disinterest already during experimental proceedings due to adapted size selection. Samples were chosen for the analysis that showed a read length distribution from 20 to 40 nucleotides resulting in 46 samples for comparison (Glaub et al., 2020). First, percentage values for each specific length were calculated within the different RNA types. These values were then compared between the multiple samples with median estimation to avoid the high influential impact of potential outliers (Glaub et al., 2020).

A differentiation between the RNA types (Figure 6) is possible with a dominant read length between 24 to 27 nt for mRNA reads whereas rRNA and tRNA corresponding reads tend to be longer (Glaub et al., 2020). rRNA read lengths peak at 26 nt respectively 31 nt. tRNA reads tend to be even longer with peak values at 32 and 35 nt of length (Glaub et al., 2020). These results lead to the assumption that reads ranging in length between 24 to 27 nt are of especial interest to obtain ribosomal protected mRNA fragments. Although a proportion of rRNA is also included in this range an additional part seems to correspond to longer reads. These, along with most of the tRNA reads could be depleted due to an adapted size selection of 22 to 30 nt during experimental processing (Glaub et al., 2020). Even though the excision step can never be of absolute accuracy, a narrower selection range could already aid in depletion of undesirable fragments. Still, rRNA depletion is highly recommended to decrease the general amount of rRNA fragments present, that cannot be targeted with the adapted size selection range (Glaub et al., 2020).

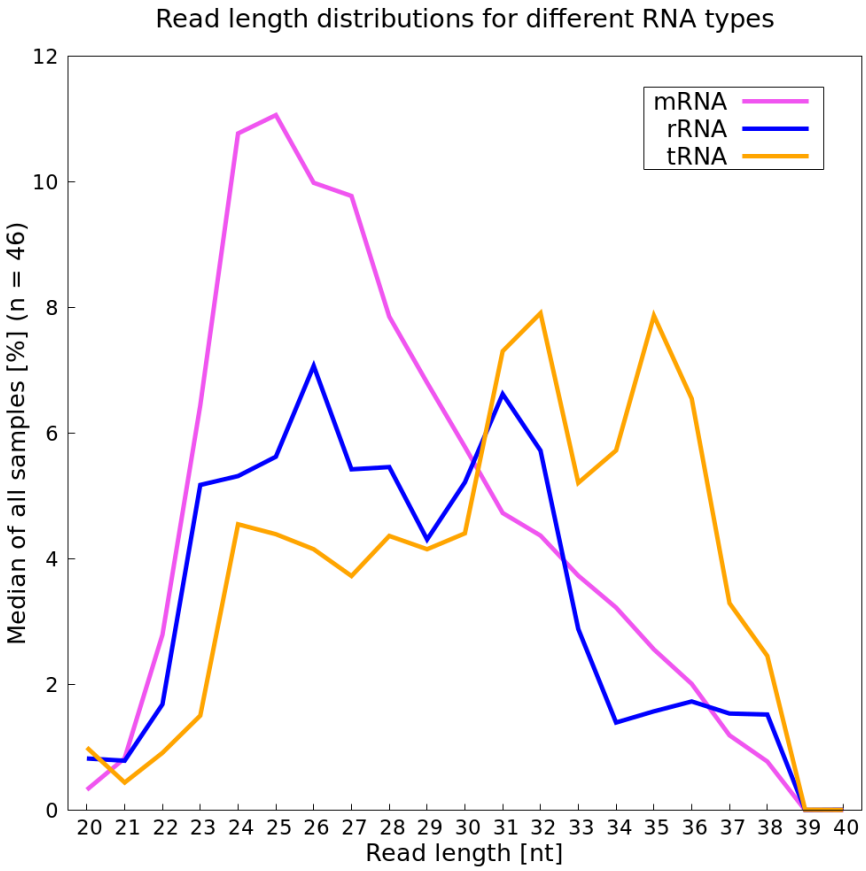

Figure 6: RNA type-specific read length analysis. Percentage values for each unique read length were compared between analysed samples (n = 46) with subsequent median estimation shown here. Colour code: pink = mRNA, blue = rRNA, orange = tRNA. Figure from (Glaub et al., 2020).

An additional analysis solely focused on rRNA corresponding reads was performed, as these are the main intern 'contaminant' within RNA type sequencing approaches. To lower the amount of rRNA present within the sample kit-based depletion is performed during experimental proceedings. However, the technique is not 100 % successful, with remaining rRNA fragments present. Therefore, the approach of potential rRNA type (5S, 16S, 23S) assessment to a specific read length was made to identify representative lengths per type (Glaub et al., 2020). Reads mapping to 23S locations are showing a pretty similar read length distribution as for the combined rRNA type analysis. For the other two types, clearer trends could be detected. Fragments obtained from 16S regions are in general 26 nucleotides in length, whereas those mapped to 5S are mostly represented by longer reads of 32 nucleotides in length (Figure 7).

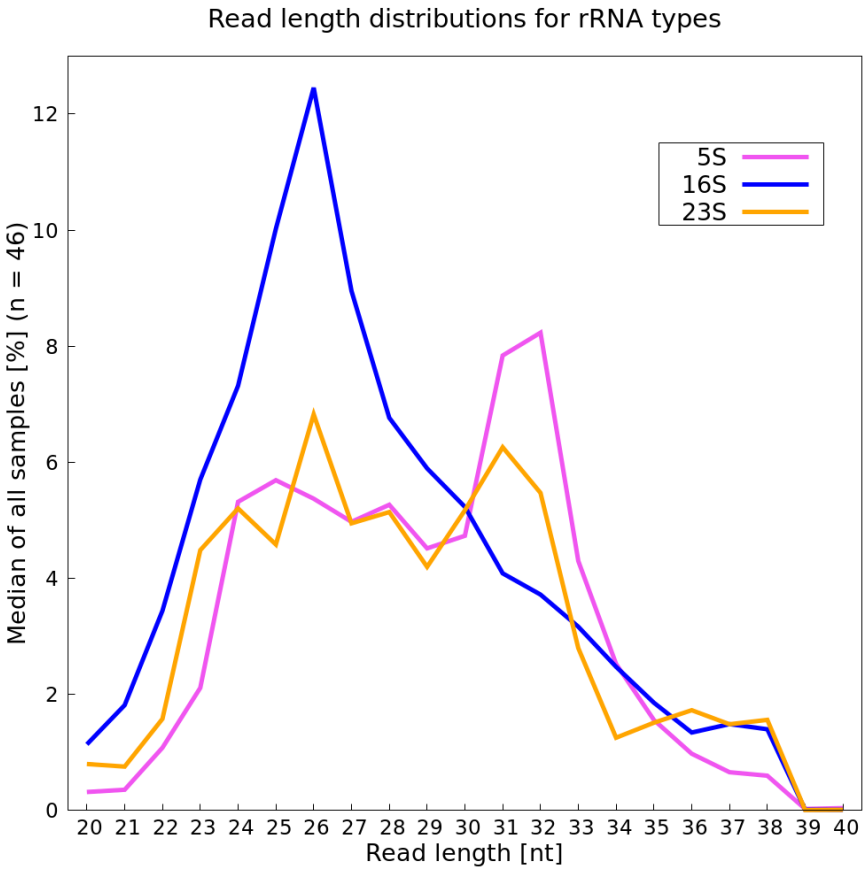

Figure 7: Analysis of read length for specific types of rRNA (pink = 5S, blue = 16S, orange = 23S). Median estimation was obtained from percentage distribution according to various length per type and sample (n = 46). Figure from (Glaub et al., 2020).

However, a modification of the selection range has always to be matched to the outcome of interest. If longer reads are of interest, the recommended size selection should be neglected due to a relatively low higher limit of 30 nt. Longer reads were the focus of the next analysis performed.

3.1.3 Longer reads mapping in 5'-UTR region

As the analysis of longer reads required a read length range up to 40 nt, samples not matching the needed range were excluded from the comparative analysis of the 5'-UTR region resulting in 30 samples.

Previous studies showed that Shine-Dalgarno like motifs are more likely to be detected in reads ranging from 28 up to 40 nt in length (Buskirk & Green, 2017; Glaub et al., 2020). As this sequence, in general, is known to be located upstream of the translational start, a range of 25 nucleotides upstream of a start codon presumably containing the SD sequence was analysed according to the mapped reads, especially their length. Here, this region is defined as the 5'-UTR region. A comparison of read length within this region was made to a region 25 nucleotide downstream of the translational start, the here called start region (Glaub et al., 2020). An expected read length for the start region corresponds to the length

identified for mRNA fragments (24 - 27 nt), whereas reads mapping in the 5'-UTR region are assumed to be longer.

Indeed, a differentiation in the read length can be made based on the two regions analysed (Figure 8). Reads that map in the start region show a dominant length of 27 nt, which is consistent with the prior detected mRNA associated read length range. The most frequently represented length for reads in the 5'-UTR region is 34 nt, emphasizing the hypothesis that reads mapping upstream of the translational start, potentially containing an SD motif, are longer (Glaub et al., 2020). This result is underlining the previously made statement that size selection during experimental proceedings has to be adjusted to the experimental goal. There is no unique read length detectable in prokaryotes that could represent all possible fragments of interest.

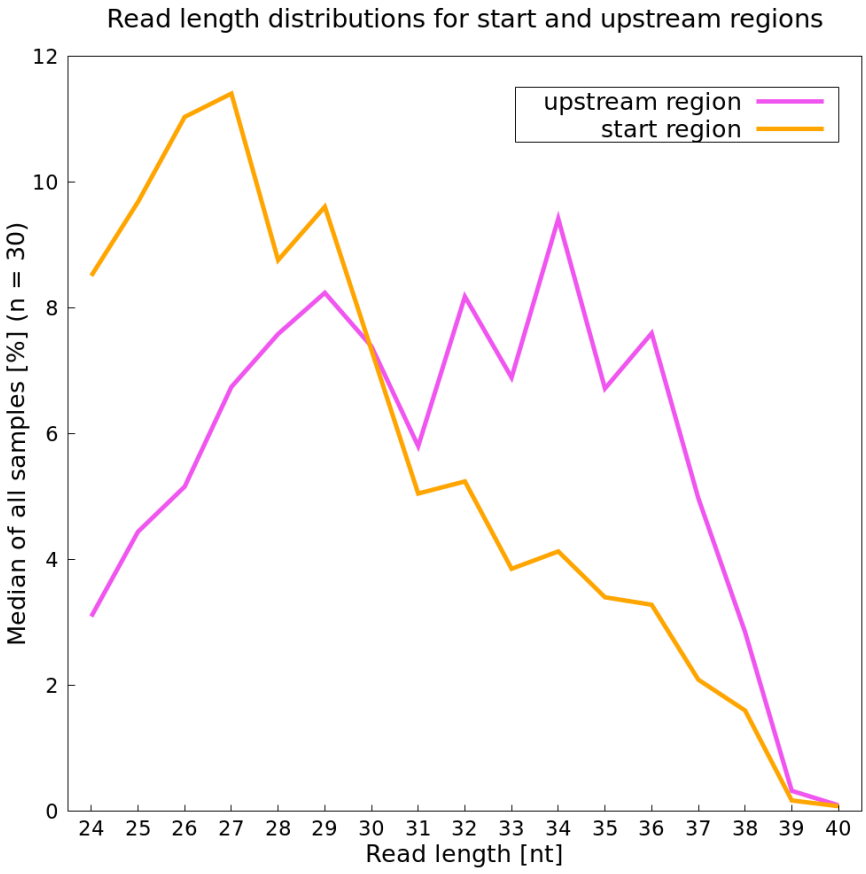

Figure 8: Comparative analysis of median calculation for read lengths of two specific regions. Start region covers 25 nt downstream of the translational start (orange), whereas the 5'-UTR region is located 25 nt upstream of the start position (pink), where the Shine-Dalgarno sequence is expected. Figure from (Glaub et al., 2020).

The additional analyses over the whole gene and stop regions (Figure 9) lead to similar results as the start region, with a most common length of 27 nucleotides for all of these gene coding regions (Glaub et al., 2020). These outcomes emphasize that reads with a length of 27 nucleotides are crucial to obtain during RIBO-Seq experiments as they most likely represent ribosome covered fragments and therefore are potential of protein-coding nature.

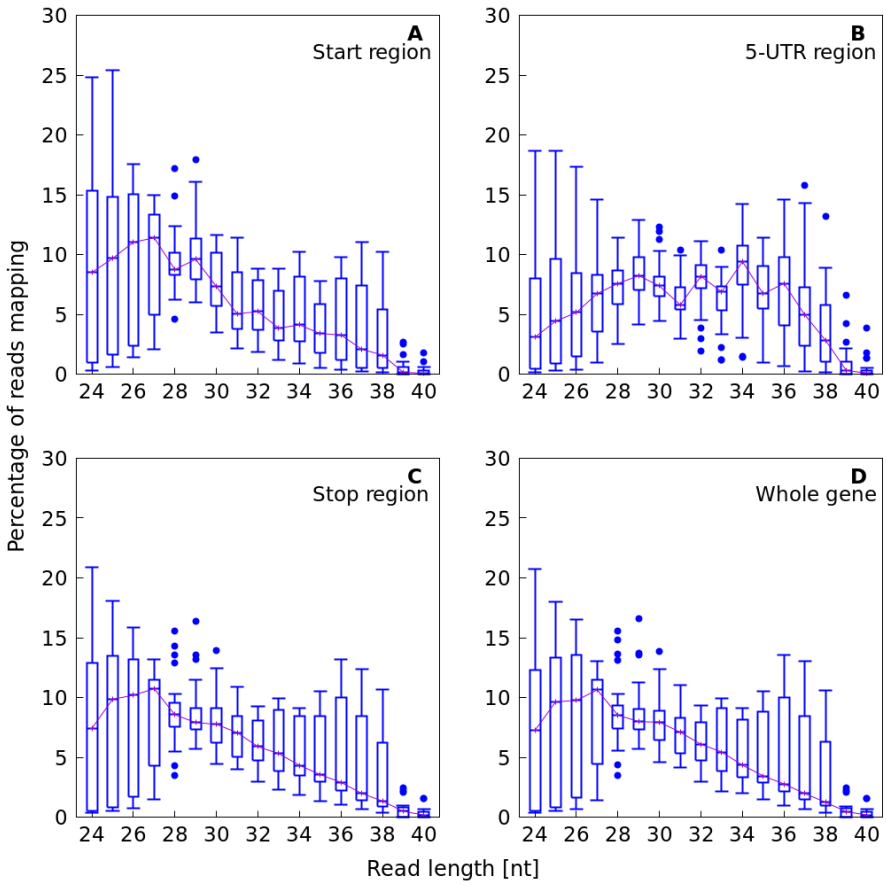

Figure 9: Region-specific read length analysis for (A) start region, (B) 5'-UTR region, (C) stop region, (D) covering the whole gene. Figure from (Glaub et al., 2020).

The next results are also focussing on the start site or more precisely on its detection.

3.1.4 Chloramphenicol addition assists translation start site detection

Ribosomal stalling can oftentimes be achieved by the application of translation inhibitors such as chloramphenicol. Accompanied by the stopped translation is an accumulation of ribosomes at the translational start site, as the attachment of ribosomes to mRNA is still possible (Mohammad et al., 2019). One chosen experiment (Oh et al., 2011) was analysed in regard to the caused accumulation artefact and its potential to aid in the detection of overlapping genes (Glaub et al., 2020).

The comparison between untreated samples and those where Cm was applied shows a detectable benefit in start site detection due to the added antibiotic in all three expressional levels analysed (Figure 10) (Glaub et al., 2020). However, it seems that the impact of ribosomal stalling at the beginning of genes is especially helpful in the detection of weakly expressed genes. Even though, the start position is more highlighted in all three expression status' analysed, the difference between mean RPKM is the highest in weakly expressed genes (Figure 10C). This finding is of special interest as the functions for many

overlapping genes in the organisms' metabolism are still unknown making it difficult to enhance their expression status due to targeted environmental conditions. Still, with the addition of in this case Cm, detection of even weakly expressed genes is possible if an accumulation of reads at the translational start site is detectable (Glaub et al., 2020).

Even though Cm is claimed to cause translational arrest read accumulations at other positions than the start site can be detected primarily in high and moderate expressed genes.

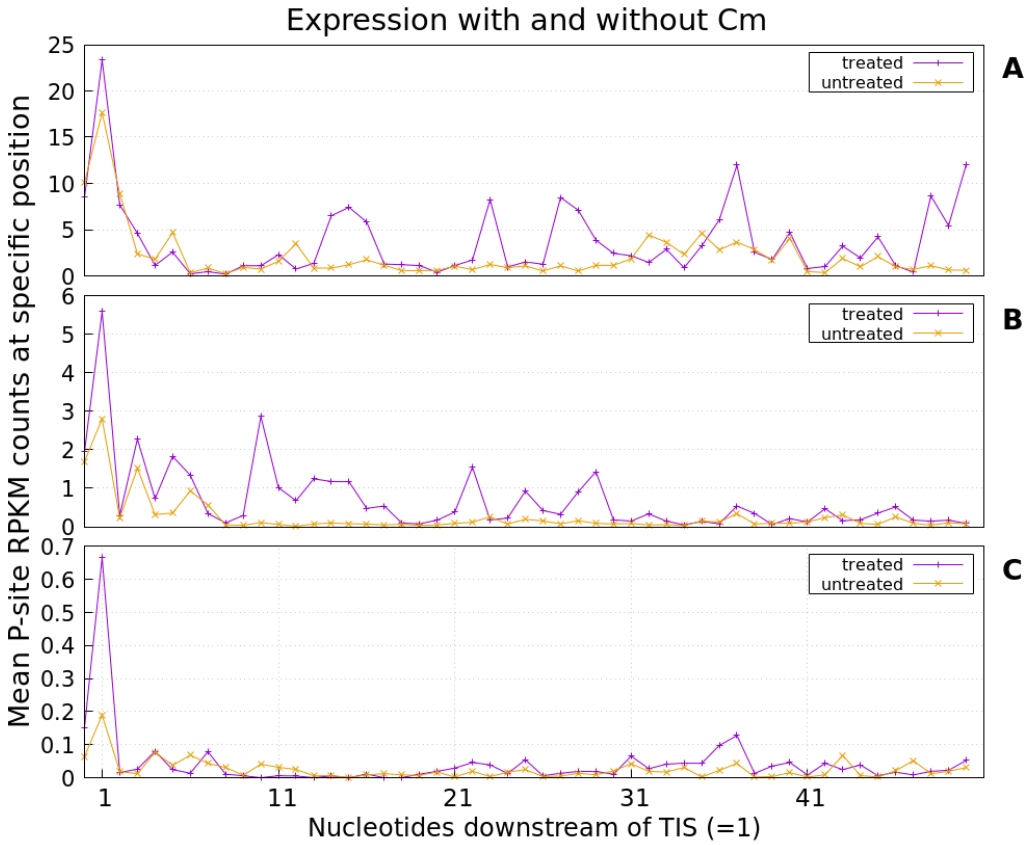

Figure 10: Translational start site analysis based on averaged read accumulation within (A) highly, (B) moderate, (C) weakly expressed genes. Chloramphenicol application (purple) is compared to non-treatment (orange). Figure from (Glaub et al., 2020).

The results obtained from analyses in this chapter lead to recommendations for an adapted size selection range for different regions of interest, a potential advanced rRNA depletion due to a narrower targeted size selection, sufficient read depth of at least 20 million reads for successful evaluation of RIBO-Seq experiments and supportive feature of Cm addition of the detection of weakly expressed genes (Glaub et al., 2020). The next chapter focusses on results obtained from overlapping gene detection in various prokaryotes and the potential influence of their genomes in detection efficiency. Furthermore, a general localisation of eOLGs detected was made with additional further analyses of selected eOLGs detected in multiple samples.

3.2 Influence of genomic features on OLG prediction amount

The number of eORFs predicted with both DeepRibo and the filtered results from the general ORFFinder script can be found in Supplementary Table S5. Similarities in prediction efficiency for both techniques used could not be identified. In general, the amount of potential eORFs is slightly bigger in the filtered ORFFinder results as it requires only exceeding an RPKM of 10 and coverage of 0.6 after being normalized to the samples read depth. Whereas DeepRibo also considers read distribution across ORFs predicted. Presumably, the stricter thresholds DeepRibo uses for prediction evaluation were contributing factors in the low prediction efficiency. A failure within the informatics implementation and evaluation of results can be denied as the prediction of annotated genes, in general, was possible, as well as the identification of overlaps within the prediction results. Therefore, further analyses were primarily based on the detected eORFs obtained from the in-house ORFFinder script. Nevertheless, for the identification of re-occurring eORFs DeepRibo results, if available, were also considered. From here on out, when referred to eORFs those are considered translated due to thresholds matched (RPKM, coverage).

First, the number of detected eORFs within the in-house prediction script (ORFFinder) were compared to the effective read amount. Again, effective reads are characterised as those remaining after alignment against a reference genome and are not mapped to either rRNA or tRNA loci. This analysis is similar to the one performed for section 3.1.1., however here the detection efficiency of eORFs is analysed. The previous claim of 20 million effective reads necessary for sufficient ORF detection should be verified whether this threshold also applies to eORF detection. Per sample number of effective reads is compared to the number of eORFs.

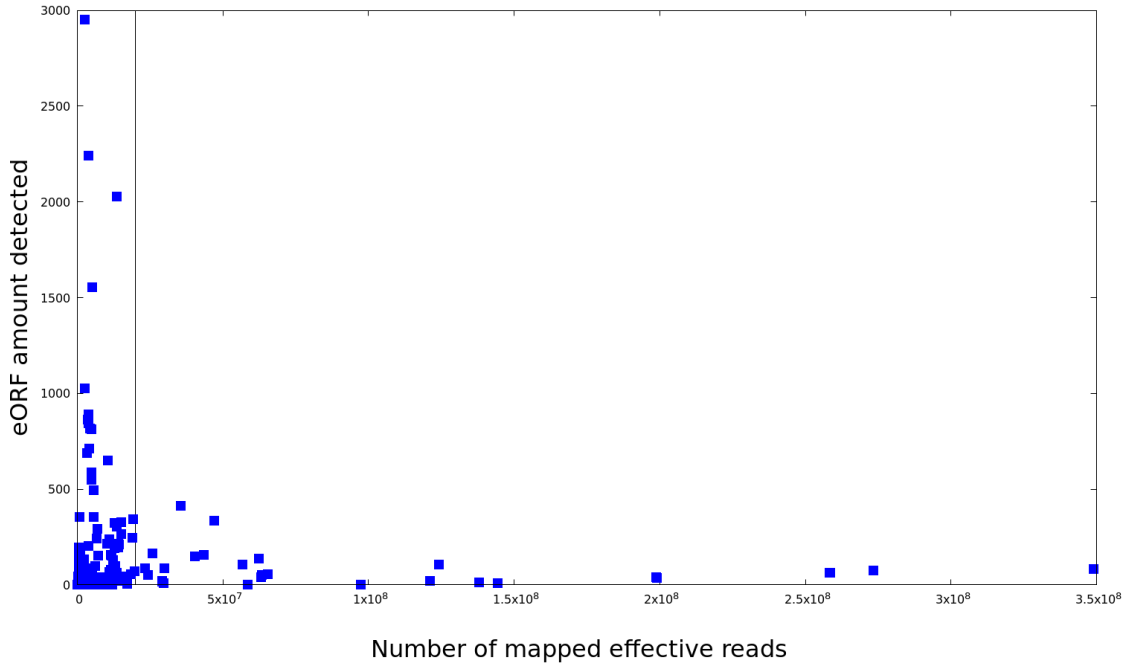

Figure 11: Comparison of eORFs identified with ORFFinder script and threshold application and the number of effectively mapped reads per sample (n = 164; without *E.coli*). Previously claimed threshold of 20 million reads for sufficient gene detection is included visualised by drawn line.

The threshold of 20 million effective reads, visualised in Figure 11 as a vertical line, can be confirmed as being sufficient for eORF detection in general. Higher read depth is not necessarily contributing to the detection of more predictions. However, the number of effective reads for samples belonging to *S. venezuelae* ranging from around $1.5*10^8$ to $3.5*10^8$ (8 highest read numbers) could be considered problematic. Here, quality control categorized up to 30 % of the sequence (sample SRR1021845) as overrepresented which were subjected to BLAST analysis before additional trimming. Sequences detected mapped to the reference genome of *S. venezuelae*, therefore, were not considered as potential contamination. Further evaluation of these reads was not performed, as they were not contributing to a higher eORF detection.

In general, even fewer read numbers seem already sufficient for eORF detection. Still, it needs to be emphasized that the threshold of 20 million effective reads for annotated genes was based on REPARATION predictions. For these stricter thresholds were applied by the tool as are used in the in-house prediction. Here, only the RPKM and coverage threshold was applied compared to read distribution patterns. Nevertheless, the threshold can be confirmed with an additional evaluation of eORFs obtained from the in-house script necessary.

The next analysis was performed to account the potential correlation of species-specific characteristics like genome size and GC-content and the amount of eORF families respectively.

3.2.1 General genome characteristics comparison

A first analysis focused on genomic feature comparison in general. The GC-content was compared to the genomic size per species respectively to detect a potential correlation based on these features. Results of this comparison were of interest to propose a hypothesis for expected eOLGs occurrence within the different species. In general, bigger genome size is accompanied by more annotated genes, as can be seen in Supplementary Table S6 for the species analysed. In comparison to this, for higher GC-content within a genome, it could be speculated that there are fewer but longer genes present. This assumption is based on the formation of stop codons mostly out of pyrimidine bases which are present in a lower proportion in high GC-content genomes. The comparison between these two features is shown in Figure 12 with the calculated R-squared value to identify a potential correlation. A slight trend can be detected implying that a bigger genome is more likely accompanied by a higher GC-content. The two clustering outliers circled represent the two archaeal species analysed. Interestingly here is that both have small genomes but unexpectedly high GC-contents. These results emphasize their phylogenetic distinction from bacteria at least within the features analysed.

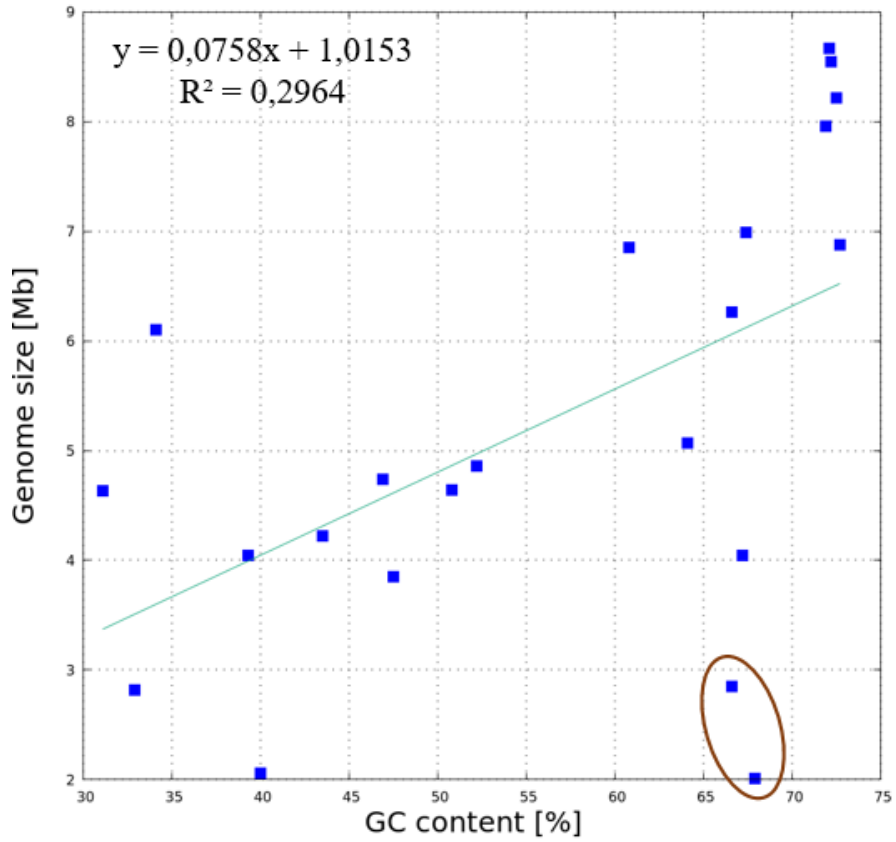

Figure 12: Correlation between genome size and GC-content (n = 22). Linear regression with the corresponding function is shown, as well as the calculated R-squared value. Clustered dots circled represent the two archaeal species analysed.

Additional comparisons were then made against the number of embedded ORFFamilies for each feature respectively. An ORFFamiliy comprises length variations of the same ORF based on a shared stop codon. The eORFs predicted in general are based on the ORFFinder output. As for these comparisons, the actual length of the translated ORF itself is not important but rather the possibility of the creation of an embedded ORF, the length variations are condensed into the ORFFamilies. Also not considered for the estimation of an ORFFamily is an exceedance of RPKM or coverage threshold. Here, the simple creation possibility of an eORF is subject of the analysis and subsequent comparison to the genomic features. Of interest here is whether GC-content or genome size influence the number of embedded ORFFamilies. The analysis shows that genome size seems to have a slightly higher impact (Figure 13A) on the amount of ORFFamilies possible than GC-content (Figure 13B) based on R-squared values calculated. ORFFamily amount and GC-content do not correlate at all. Again, for both analyses performed the clustering of the two archaeal species can be shown once more emphasizing their difference to bacteria.

From these first general comparisons, a hypothesis for eORFs based on genomic features is derived and subjected to further analysis. Genome size seems to have a slightly higher impact on embedded ORFFamily occurrence. Additionally, bigger genome size is more likely accompanied by higher numbers of annotated genes. Therefore, an assumption of more eORFs present in bigger genomes was

made but had to be rejected after analysing the number of eORFs detected within the samples (see Supplementary Table S5).

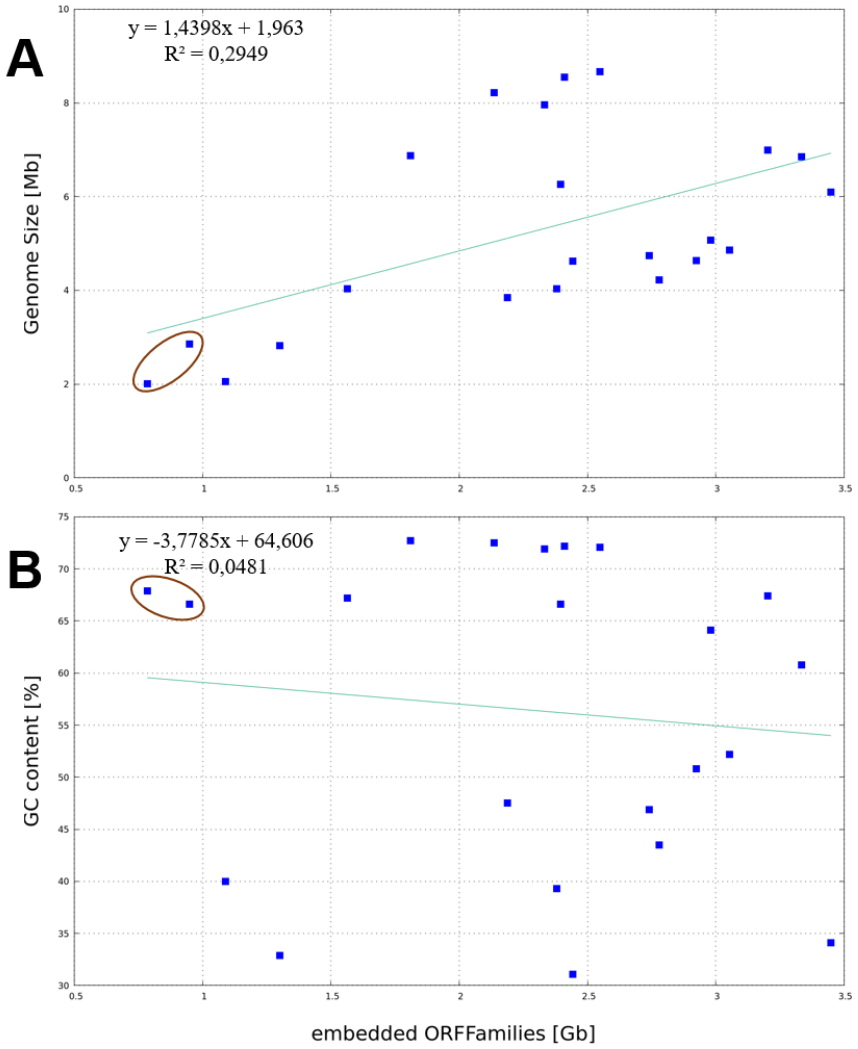

Figure 13: (A) Correlation between genome size and embedded ORF families; (B) Relation between GC-content and embedded ORF families. For both correlations (n = 22) linear regression lines with their function and R-squared values are added. Clustered circled dots again represent the two archaeal species.

Next, the focus was on whether GC-content could influence eORFs in general. Here, a hypothesis was made that higher GC-content could potentially enable the creation of longer eORFs.

3.2.2 Length analysis of eORFs detected

For each species, eORFs predicted were combined and re-occurring ones were condensed to eliminated multiple counting. Then, those identified were categorized according to their length and their abundance was calculated in percentage. Length distribution for an excerpt of species is shown in Figure 14, with species chosen as representatives for different GC-contents. Supplementary Figure S2 shown eORF length distributions of all species analysed.

Segment A shows the eORF length distribution for *S. aureus* having slightly more eORF with a length range between 93 to 100 nucleotides (Figure 14A). Shorter ORFs were in general excluded as a threshold for minimal length was set to 93 for consideration of an actual ORF. From there a shift for the most abundant length can be detected starting with *B. subtilis* having an equivalent amount of eORF predictions for both lengths. For all species with GC-contents above ~ 44 %, the predominant eORF lengths are between 100 to 200 nucleotides. Thereafter, no shift to even longer ranges as the most frequent one can be detected.

Nevertheless, in higher GC-content genomes predictions of eORFs spanning 500 nt or more can still be found. Occasionally also eORF detections were made of length up to 1,300 nt in length. Such long eORFs were identified in *P. aeruginosa*, *S. typhimurium* and *H. volcanii*, matching all criteria to be considered translated.

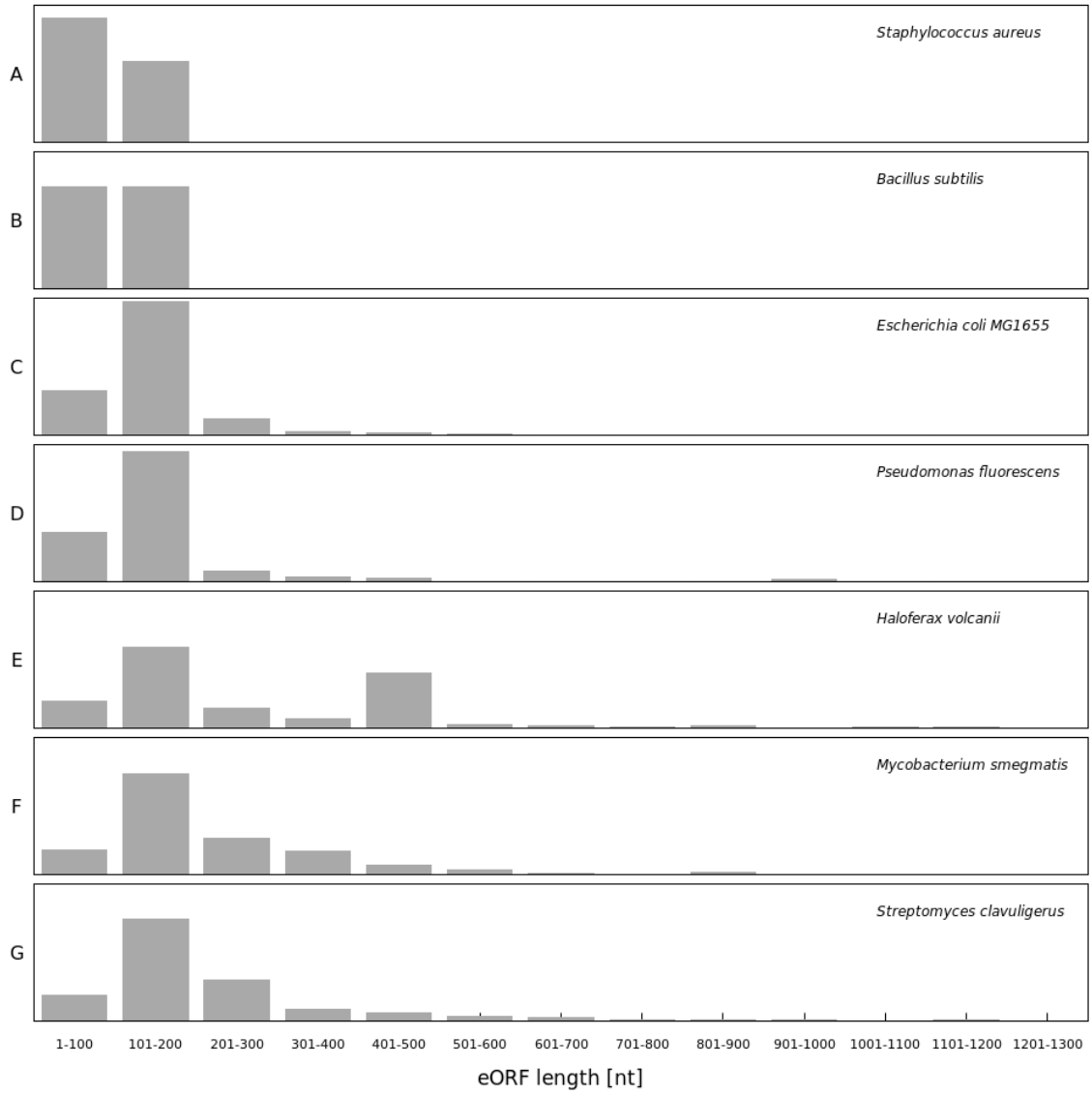

Figure 14: eORF length distribution within a selection of species analysed (n = 7). Species shown were selected based on their increasing GC-content (A) 32.9, (B) 43.5, (C) 50.8, (D) 60.8, (E) 66.6, (F) 67.4, (G) 72.7.

The length analysis revealed that the predominant length for eORFs in general ranges between 100 to 200 nt. Therefore, the hypothesis of proportional longer eORF in higher abundance present in higher GC-content genomes had to be rejected. However, in general, the detection of long eORFs even up to 1,300 nt in length, is possible in high GC-content genomes.

After length analysis of eORFs predicted another interesting point is the location relation of the embedded overlapping gene with its mother gene.

3.2.3 Relative reading frame relation between the mother gene and eORF

The maintenance of a nucleotide sequence is highly dependent on its function. Protein-coding parts of the genomic sequence are therefore under negative selection which implies that mutation resulting in functionality loss should be prevented. Hence, the conservation of the known functional sequence is affecting potential new open reading frames located in an alternative frame as it restricts potential nucleotide exchanges. The relation of annotated genes and potential eORFs was analysed regarding one frame being favoured in the creation of overlapping genes.

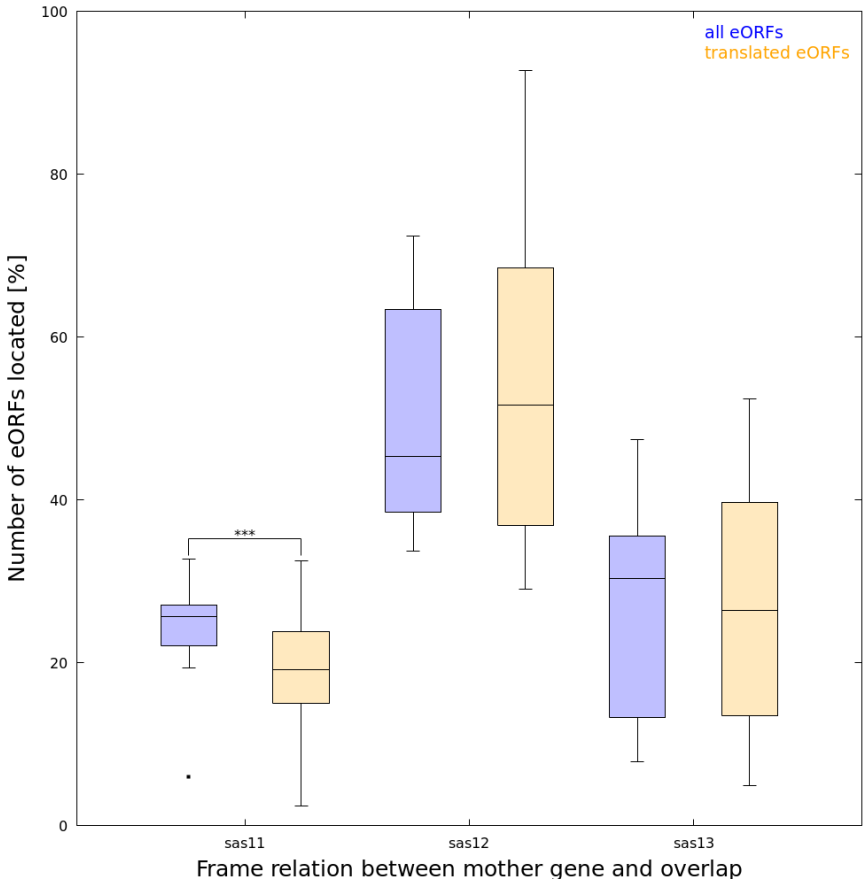

Figure 15: Location relation analysis for mother gene and embedded ORF (n = 24). Additionally shown is the comparison of all possible eORFs predicted and eORFs considered translated. A trend (p = 0.07) for the formation of translated eORFs is found in sas12, whereas their occurrence in sas11 seems even statistically significant unlikely (p = 0.0009).

Expected to be in favour is the relative reading frame sas11 as here, the third codon position of the mother gene is overlapped by the second position of the alternative ORF. An exchange in the most variable position for the annotated genes is likely to cause no amino acid exchange in the protein sequence, whereas an exchange in the most informative 2nd position would lead to major rearrangements for the overlapping gene. This hypothesis however cannot be confirmed by the results from the performed analysis (Figure 15). Rather sas12 seems to be most likely for the localisation of translated eORFs ($p = 0.07$) and sas11 significantly confirmed as unlikely for their occurrence ($p = 0.0009$). A two-sided t-test analysis performed was used for significance calculations. Numbers used for the calculation can be found in Supplementary Table S7. Still, sas12 was not expected to be the most likely relative reading frame relation as here for both annotated and overlapping ORF the third codon positions are complementary to each other. Exchanges that might happen more frequently here are most likely not impacting any of the two sequences. On the other hand, this might be in favour to maintain functional sequences. Hence, it is necessary to differentiate between the localisation of an overlapping gene to an annotated one dependent on the creation of a functional overlap versus maintaining its functionality.

3.2.4 Re-occurring eORFs and their BLAST analysis

The analysis of eORF re-occurrence within a species was first tested on a subset of species available, namely *C. crescentus, E. coli, H. salinarium, H. volcanii, M. smegmatis, M. abscessus, P. aeruginosa, P. fluorescens, S. typhimurium, S. aureus, S. clavuligerus, S. coelicolor, S. griseus* and *S. tsukubensis*. For each species, an individual frequency of re-occurrence was estimated based on the sample amount and prediction efficiency. Based on their re-occurrence within species-specific samples and additional visual inspection of read distribution across the loci, a total of 43 eORFs was chosen for further analyses. Supplementary Table S8 contains location information for eORFs and corresponding mother genes, the location relation between the two genes and additional genomic information. For all loci nucleotide as well as protein sequences were extracted from the associated genome.

First, for both mother gene and overlap an approximate age determination was performed based on homologue sequence detection within the associated species family. Therefore, protein sequences were used for a tblastn approach. Here, the protein query sequence is translated into a nucleotide sequence. For amino acids within the protein sequence that can be encoded by several nucleotide triplets, all possibilities are considered. Thus, this approach is not restricted to exact matching homologues but allows to search for sequence similarities as long as the same protein sequence is encoded.

The maximal target sequence was set to 1,000, however, if more alignments during BLAST analysis were found, the total number is displayed. Additionally, an e-value of $1*e^{-10}$ was set for sequence alignment similarity to be considered. The number of homologues found for the mother gene and overlap was compared. Therefore, if more than 500 hits were obtained the query gene was categorized as ‘old’,

whereas fewer hits indicate a ‘young’ gene. This is only a preliminary analysis as it does not take into account the number of genomes sequenced for a taxonomic group. Analysis output is summarized in Table 11.

Table 11: Obtained tblastn output performed on the mother gene and overlap protein sequence. Parameters set for database search were: maximum target sequence number 1,000, e-value threshold of sequence similarity 1*e-^{10}, search for homologues restricted to the species-specific family level.

eORF Name	BLAST hits for mother gene		BLAST hits for detected eORF	
Caulobacter1	25	+	8	+
Caulobacter2	298	+	5	+
Escherichia1	9953	*	4030	*
Escherichia2	1010	*	1001	*
Escherichia3	1000	*	1000	*
Escherichia4	21333	*	15366	*
Escherichia5	1036	*	1000	*
Escherichia6	1017	*	422	+
Escherichia7	1036	*	1000	*
Escherichia8	1017	*	422	+
Halobacter	21	+	4	+
Haloferax	18	+	1	+
Mycobacterium1	1408	*	117	+
Mycobacterium2	249	+	8	+
Mycobacterium3	661	+	9	+
Mycobacterium4	9631	*	11	+
Mycobacterium5	113	+	16	+
Mycobacterium6	15	+	10	+
Mycobacterium7	1550	*	44	+
Mycobacterium8	30	+	21	+
Pseudomonas1	898	*	895	*
Pseudomonas2	360	+	262	+
Salmonella1	1000	*	1002	*
Salmonella2	1015	*	871	*
Staphylocuccus1	1000	*	686	*
Streptomyces1	10	+	6	+
Streptomyces2	628	*	105	+
Streptomyces3	260	+	193	+
Streptomyces4	322	+	84	+
Streptomyces5	100	+	15	+

eORF name	BLAST hits for mother gene		BLAST hits for detected eORF	
Streptomyces6	263	+	225	+
Streptomyces7	684	*	16	+
Streptomyces8	743	*	331	+
Streptomyces9	314	+	15	+
Streptomyces10	29	+	36	+
Streptomyces11	408	+	15	+
Streptomyces12	156	+	52	+
Streptomyces13	28	+	15	+
Streptomyces14	398	+	16	+
Streptomyces15	508	*	19	+
Streptomyces16	325	+	315	+
Streptomyces17	335	+	3	+
Streptomyces18	263	+	203	+
Total	* = 19; + = 24		* = 10; + = 33	

A clear clustering according to the two possible categories can be detected. Except for nine sequences where the mother gene is categorized ‘old’ and the overlap new, all 34 remaining pairing hits per locus analysed correspond to the same category. For ten combinations mother gene and overlap are categorized as ‘old’. Interestingly here, seven belong to the family *Enterobacteriaceae*, more specific, six are assigned to *E. coli* and two to *S. typhimurium*. Additionally, the two pairings of *S. aureus* and the one in *P. fluorescens* are categorized ‘old’. Within the remaining 24 mother gene and overlap combinations, both genes were categorized as ‘young’. Still, for most mother genes contained in this category, more homologue hits can be found. Thus, in general, the mother gene seems to be ‘older’ as expected.

Next, the re-occurring eORFs sequences were subject of selection pressure analysis. Here, the aim was to establish if the sequences were more likely constructed due to selection or are of random origin. Additionally, these results could underlie if an eORF detected could be of interest with this second verification.

3.2.5 Indication of functionality based on selection pressure estimation

For the performed analysis the tblastn approach was repeated, however setting changes were made matching the OLGenie tools input criteria. The search for similar sequences was limited to the genus level and e-value was set to $1*e^{-5}$. Based on the narrower genus level search a less strict e-value was

chosen. If available, the 1,000 best aligned sequenced were obtained with additional filtering necessary. Sequences used for OLGenie analysis are required to have a matching length as otherwise the substitution calculation for each specific location cannot be performed appropriately (Nelson, Ardern, & Wei, 2020). Albeit the threshold adaptations for the tblastn approach, the detection of 1,000 similar sequences was not possible for each eORF of interest. For 9 eORFs however, no sequence hits were detected, hence, they were excluded from subsequent OLGenie analysis. Nevertheless, for the remaining candidates, the amount available was obtained and OLGenie analysis was performed on all 34 sequences of interest with subsequent significance analysis also provided by extending OLGenie scripts available (OLGenie_bootstrap.R obtained from https://github.com/chasewnelson/OLGenie).

For all but six eORFs analysis was successfully performed. Sequence comparison is performed regarding synonymous and nonsynonymous nucleotide exchanges at each position within the sequence and their number of occurrences. The tool detects selection for the mother genes' sequence based on the input eORF sequence and the frame relation variable which is necessary for processing. Calculated are four different substitution possibilities, namely a synonymous (dS) exchange in both mother gene and overlap, a nonsynonymous (dN) exchange within both, a synonymous exchange in the mother genes' sequence is accompanied by a nonsynonymous in the overlap and vice versa (Nelson, Ardern, & Wei, 2020).

Significance values for the mentioned substitution relation were estimated with an additional available OLGenie associated R-script. Low values detected imply purifying selection for the eORF analysed. Therefore, the functionality of the eORF is assumed based on the necessity to maintain its sequence order. The p-value results are listed in Table 12. For 16 mother genes' sequences, significant results were obtained indicating a high purifying selection for these as even synonymous substitutions are very unlikely to occur. However, only for two eORF sequences analysed p-values < 0.05 were detected. Unfortunately, these were previously not categorized as 'old' which could be considered as a second verification. Interestingly, however, both, Streptomyces3 and Streptomyces6 were detected in the genus *Streptomyces*. Nevertheless, for six additional eORF sequences' nearly significant values were detectable. Of those, two, Pseudomonas1 and Salmonella1, were categorized in the previous analysis as 'old'. As both analyses performed so far are considered as first indicators to identify eORFs of potential functionality based on the sequence's evolution, these correspondences are of interest. However further analysis, especially experimental is needed to verify their functionality conclusively.

Also addressed here should be the mother gene sequence of Streptomyces16. Noticeable, the p-value calculated implies no selection for this sequence at all. This indicates that the mother gene itself has no functionality. Potentially, the mother gene, in this case, is a pseudo gene that at some point might have been of actual function but due to unknown reason lost its functionality during evolution, or is incorrectly annotated as a gene, or is simply under relaxed selection.

Table 12: OLGenie based selection pressure analysis for 28 eORFs of interest obtained from re-occurrence analysis (sample names are connected). P-values were calculated indicating to which extend purifying selection is working on maintaining the sequences' order. Significant results are labelled according to following categories: 0.05 = *; 0.01 = **; 0.001= ***. Six marked p-values indicate further interesting eORFs nearly matching a significant threshold.

eORF name	p-value eORF	p-value mother gene	eORF length
Caulobacter1	0.71	0.00***	209
Escherichia1	0.99	0.20	290
Escherichia2	0.25	0.01**	335
Escherichia3	0.93	0.25	290
Escherichia4	0.93	0.26	290
Escherichia5	0.24	0.00***	116
Escherichia6	0.08	0.53	242
Escherichia7	0.73	0.93	176
Escherichia8	0.08	0.52	242
Halobacter	0.29	0.66	230
Mycobacterium1	0.08	0.00***	248
Pseudomonas1	0.07	0.00***	548
Pseudomonas2	0.33	0.34	221
Salmonella1	0.07	0.32	233
Salmonella2	0.45	0.00***	191
Streptomyces3	0.03*	0.15	299
Streptomyces4	0.41	0.00***	131
Streptomyces5	0.49	0.01**	215
Streptomyces6	0.03*	0.00***	542
Streptomyces7	0.10	0.05*	224
Streptomyces8	0.43	0.00***	1058
Streptomyces12	0.83	0.00***	131
Streptomyces13	0.95	0.00***	158
Streptomyces14	0.23	0.00***	683
Streptomyces15	0.16	0.00***	278
Streptomyces16	0.22	7.09	176
Streptomyces17	0.12	0.27	230
Streptomyces18	0.78	0.00***	374

First descriptive analyses of selected eORFs were completed by the following Frameshift analysis. Here, the corresponding mother gene nucleotide sequence is analysed regarding the creation of random overlaps possible. Especially their length is of interest here and will be compared to the length of eORFs detected. Thus, conclusions can be drawn whether their length is considered significant or random.

3.2.6 Probability analysis of eORF creation based on codon permutation

The nucleotide order of an overlapping gene is dictated by its mother gene sequence. Hence, for this analysis, the mother genes' sequence is analysed not only concerning the possibility of overlap formation but rather whether its length is longer than expected by chance. Here, an available tool named Frameshift is used (Schlub et al., 2018). An entered nucleotide sequence is translated according to the three-base periodicity into codon structure. After random shuffling the resulted overlapping ORFs possible within the input sequence are analysed considering their length. An also provided R script is calculating the length outcome and whether this length is longer than expected based on random codon permutation (Schlub et al., 2018). An obtained p-value ≤ 0.001 indicates that the created eORFs length is highly unlikely of random origin. Therefore, this analysis is also used for additional verification of the eORFs of interest.

Again, the 43 eORFs of interest identified in section 3.2.4. were chosen for this analysis. Therefore, their assumed functionality could be additionally verified based on their sequence length probability. Unfortunately, for none of the eORFs analysed a statistically significant p-value was calculated. Thus, their respective length is not considered valid to indicate potential functionality. Nevertheless, for three candidates nearly significant values were estimated (Escherichia1, Escherichia3, Escherichia4, Streptomyces18). None of these corresponds to sequences for which a potential selection pressure was estimated (see Section 3.2.5.). However, sequences of candidates Escherichia1, Escherichia3 and Escherichia4 were categorized as old in the analysis explained in Section 3.2.4. Even though p-values calculated here do not reach a significance level, a correspondence of eORFs analysed between the different analyses performed is always of interest. A comparison between these candidates obtained from the same species (*E. coli* MG1655) revealed identical sequences for Escherichia3 and Escherichia4. The sequence of Escherichia1 showed 2 different amino acid exchanges shown in Supplementary Figure S3. Despite differences, all three of them are located complementary to the same mother gene, the IS1 transposase B. However, a comparison for this gene also reveals exchanges in the amino acid sequences at different locations. Here, even six different amino acids incorporated were detected (see Figure S4).

A potential bias introduced on a sequences' length itself can also be neglected. For the longest eORF analysed (Streptomyces8, 1,058 nt) no significance was determined. Here, the speculation was, that just based on its length the sequence might show significance. However, an even longer mother gene simultaneously enables the possibility of long overlaps – depending also on codon usage. Based on the results obtained and the variety of eORF length tested (see Supplementary Table S8) a potential bias due to query length cannot be confirmed.

Sequence evolution can be considered as an indicator for its actual functionality. The performed analyses were focussing on different aspects in evolution enabling a comparison for obtained results. These were

used as the first evaluation of potentially assigned functionality based on sequences' integrity and distribution. Those analyses should assist in the identification of potentially functional eORF sequences in terms of narrowing down candidates of interest for subsequent experimental verification.

Table 13: List with calculated p-values for codon permutation analysis. Analysed were the 43 eORFs of interest obtained from the re-occurrence analysis mentioned in Section 3.2.4. None of the eORFs respective lengths was longer than expected based on the random shuffling of the mother genes' codons.

eORF name	**p-value for codon permutation**	**eORF name**	**p-value for codon permutation**
Caulobacter1	0.942559654	Salmonella1	0.421985380
Caulobacter2	0.999917694	Salmonella2	0.1472487392
Escherichia1	0.053546052	Staphylococcus1	0.620273717
Escherichia2	0.103011241	Streptomyces1	0.685593909
Escherichia3	0.051226354	Streptomyces2	0.723092694
Escherichia4	0.055928932	Streptomyces3	0.310695902
Escherichia5	0.996650014	Streptomyces4	0.989725877
Escherichia6	0.387908731	Streptomyces5	0.9996148381
Escherichia7	0.404562304	Streptomyces6	0.76954544
Escherichia8	0.391049067	Streptomyces7	0.9544756478
Halobacterium	0.593421465	Streptomyces8	0.391375103
Haloferax	0.66591771	Streptomyces9	0.957758014
Mycobacterium1	0.462857695	Streptomyces10	0.903352171
Mycobacterium2	0.982964929	Streptomyces11	0.761793043
Mycobacterium3	0.432052136	Streptomyces12	0.696437949
Mycobacterium4	0.999182019	Streptomyces13	0.878540392
Mycobacterium5	0.337132782	Streptomyces14	0.665218059
Mycobacterium6	0.70030499	Streptomyces15	0.677845211
Mycobacterium7	0.995674588	Streptomyces16	0.973468532
Mycobacterium8	0.911797303	Streptomyces17	0.670263340
Pseudomonas1	0.24504758	Streptomyces18	0.053114782
Pseudomonas2	0.573251756		

The results obtained from a RIBO-Seq and RNA-Seq experiment for *B. thetaiotaomicron* are the topic of the following chapter. The experiment was performed to potentially identify eORFs within the analysed species.

3.3 *B. thetaiotaomicron* experimental proceedings

Harvested cell pellets in multiple falcon tubes all obtained from the same bacterial culture were received from another research institute. For evaluation and comparison purposes for each technique (RIBO-Seq and RNA-Seq) at least duplicates were processed. Based on the labelling of falcons received, samples were named I and II respectively in the following evaluation.

3.3.1 RIBO-Seq and RNA-Seq preparation

Experimental proceedings after homogenisation started with RNA extraction and quality estimation to decide whether to directly start a RIBO-Seq experiment or not. Right from the start extracted RNA showed signs of degradation seen as smeared bands on the agarose gel (Figure 16A). Expected were two clear bands at around 1 kb and 1.5 kb representing 16S and 23S, respectively. Bands that proceeded further are marks for degradation as these represent smaller rRNA fragments. Changes in homogenisation such as quicker thawing or adapted lysis buffer composition were tried to prevent samples from degradation. Unfortunately, these improvements did not result in non-degradation. Nonetheless, the ribosomal assembly should still be intact to some extend even while degradation is proceeding. Thus, mRNA fragments embedded within the two ribosomes' subunits should be protected from degradation and the experiment was continued.

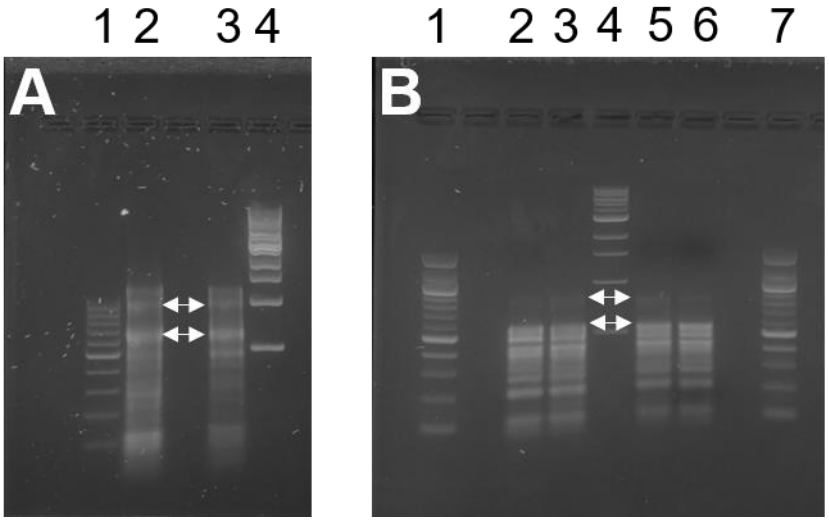

Figure 16: Samples are shown after RNA extraction (A) RNA-Seq samples: 1) 100 bp ladder, 2) RNA I, 3) RNA II, 4) 1 kb ladder; (B) RIBO-Seq samples: 1+7) 1kb ladder, 2+3) RIBO I, 5+6) RIBO II. In all samples, RNA degradation can be seen as multiple additional smaller bands visual on the gel. Arrows mark the expected localisation for 16S and 23S bands.

Interestingly, RNA samples showed a wrong size assignment on the agarose gel (Figure 16A). Expected sizes for 16S and 23S were not matched. Here, the corresponding 23S band in comparison to both the 1 kb and 100 bp ladder was at around 1 kb whereas the 16S band can be matched to the 500 bp band. To verify the right sizes for the band's gels were performed in other stations where this mislabelling did not occur. A second verification with additional quality control was done by a Bioanalyzer analysis

(Figure 17). Here, the two RNA-Seq samples shown in Figure 16A were subjected to the analysis. Again, a clear sign of degradation can be detected in both samples. Nevertheless, even though the size estimation on the agarose gel does not correspond to the ribosomal subunit sizes, the Bioanalyzer results are assigned correctly, confirming that the bands are correct.

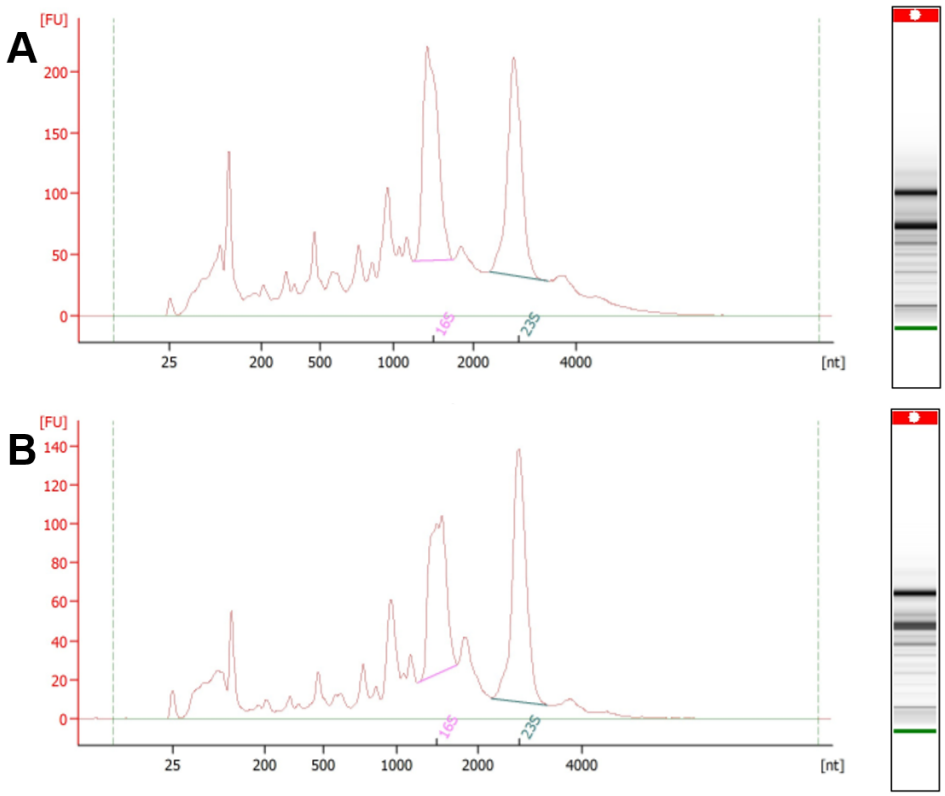

Figure 17: Results are shown for a bioanalyzer RNA 600 Nano Chip run of RNA samples (A) RNA I (same sample as shown in Figure 16A, lane 2), (B) RNA II (same sample as shown in Figure 16A, lane 3). The RNA degradation that is already seen on the agarose gel (Figure 16) can be confirmed with the performed bioanalyzer analysis. Given are the length of fragments detected in nucleotide [nt] and their abundance by detected fluorescence units [FU].

Next, remaining DNA within RNA-Seq samples was digested with subsequent performance verification by 16S PCR. No bands corresponding to 16S fragments were detectable on the agarose gel implying degradation success (Supplementary Figure S1). Then rRNA depletion was performed simultaneously with RIBO-Seq samples.

The first experimental step for RIBO-Seq samples after homogenisation is footprint generation based on a treatment with a mixture of RNA digestion enzymes. A combination of endo- and exonucleases should ensure complete digestion of unprotected mRNA. While endonucleases are causing the cleavage within their targeted sequence exonucleases are causing digestion of fragments from their edges on forward (Artymiuk, Ceska, Suck, & Sayers, 1997; Bernad, Blanco, Lkaro, Martin, & Salas, 1989; Roberts, 1978). During the process of RNA digestion mRNA whose sequence is protected by multiple ribosomes (polysome structure) is disassembled into parts solely protected by one ribosome

(monosome). To obtain only monosomes for further processing a density gradient centrifugation is performed where ribosomal protected mRNA fragments within the samples are separated according to their molecular weight (Oster & Yamamoto, 1963). A band between the 25 % and 30 % gradient levels was collected containing monosome protected mRNA of interest. RNA extraction was performed on the collected samples which also showed degradation with a slightly clearer band pattern (Figure 16B). Nevertheless, subsequent footprint extraction out of an UREA polyacrylamide gel (15 %) was performed. Aimed fragment size was around 24 to 27 nucleotides, therefore one marker of 23 nucleotides in length was chosen as a lower boundary, whereas a ladder stretched a range from 19 to 27 nucleotides. According to these guidance bands were excised from gel and RNA was subsequently purified thereof.

Depletion on RIBO-Seq and RNA-Seq samples were handled simultaneously. rRNA depletion was performed with Pan-Prokaryotes riboPOOLs to decrease the amount of remaining rRNA present in the samples. Next, dephosphorylation of 3' fragment ends were performed with subsequent phosphorylation of the same 3'-ends. Hereby, the same status of phosphorylation for all fragments is achieved necessary for the following adapter ligation during library preparation.

Concentration determination of extracted RIBO-Seq and RNA-Seq samples can be found in Supplementary Table S9. As expected, the RNA concentration within RIBO-Seq samples was lower due to targeted extraction of only ribosome protected mRNA fragments. During sample processing, a loss of RNA material was measurable, with only very low amounts before library preparation (Supplementary Table S9). Nevertheless, samples were prepared for sequencing as in this technique amplification steps are performed increasing the material concentration.

Library preparation was performed according to the Illumina TruSeq Small RNA Kit. Index (Supplementary Table S10) and adapter attachment were performed to enable multiplexing samples for test sequencing. After transcription from input RNA to DNA, libraries were amplified by a polymerase chain reaction to enhance input material for sequencing. After purification samples were separated on a polyacrylamide gel (10%) with a size selection aimed for between the custom RNA ladders' bands of 145 bp and 160 bp respectively. DNA was extracted from the gel and subsequently, concentration was measured with Qubit (Supplementary Table S9). Fragment length detection is based on a high sensitivity DNA Chip Bioanalyzer analysis (Figure 18). Concentration as well as fragment length values were then used to calculate necessary input material per sample to result in 2 nM libraries each.

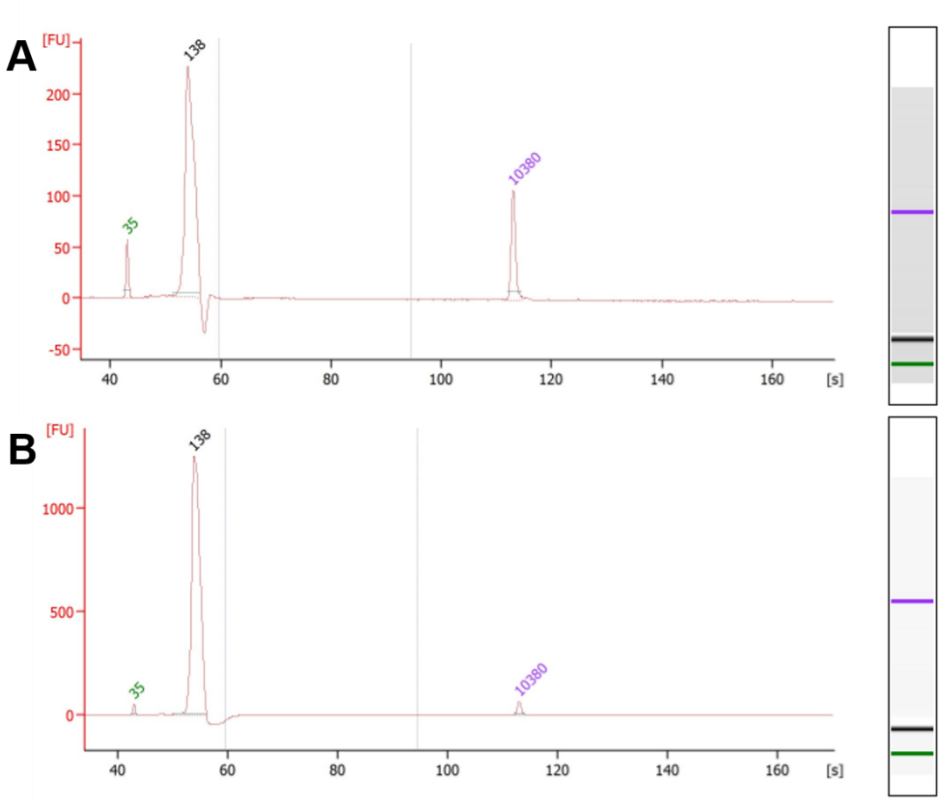

Figure 18: Bioanalyzer High Sensitivity DNA run for NGS library fragment length detection. Shown are samples (A) RIBO I and (B) RNA I. An average of all four libraries' length was calculated for subsequent library input concentration.

First, the measured concentrations [ng/µl] had to be converted considering fragment length. The length value was averaged considering all four library values (x = 140). The calculation was performed as follows according to manufacturers' instructions.

$$\frac{measured\ concentration\ [\frac{ng}{\mu l}]}{660{*}140} * 10^6 = concentration\ [nM]$$

With calculated concentrations input amounts of prepared libraries were estimated in this manner:

$$x * concentration\ [nM] = 2\ nM * 10\ \mu l$$

$$x = sample\ input\ [\mu l]$$

0.5 mM dilutions were pooled (5 µl per library) for test sequencing purposes. Raw reads were generated and processed as mentioned before including quality control, adapter removal, reference genome alignment and filtering out of 'undesired' rRNA and tRNA reads.

Again, during sample processing loss of material was recorded which is expected to some amount due to gel excision. However, concentrations before sequencing were very low (Supplementary Table S9). Nevertheless, samples processed were used for a first sequencing test run to determine the quality of samples and efficiency of rRNA depletion.

3.3.2 RIBO-Seq data evaluation of test sequencing approach compared to the available dataset

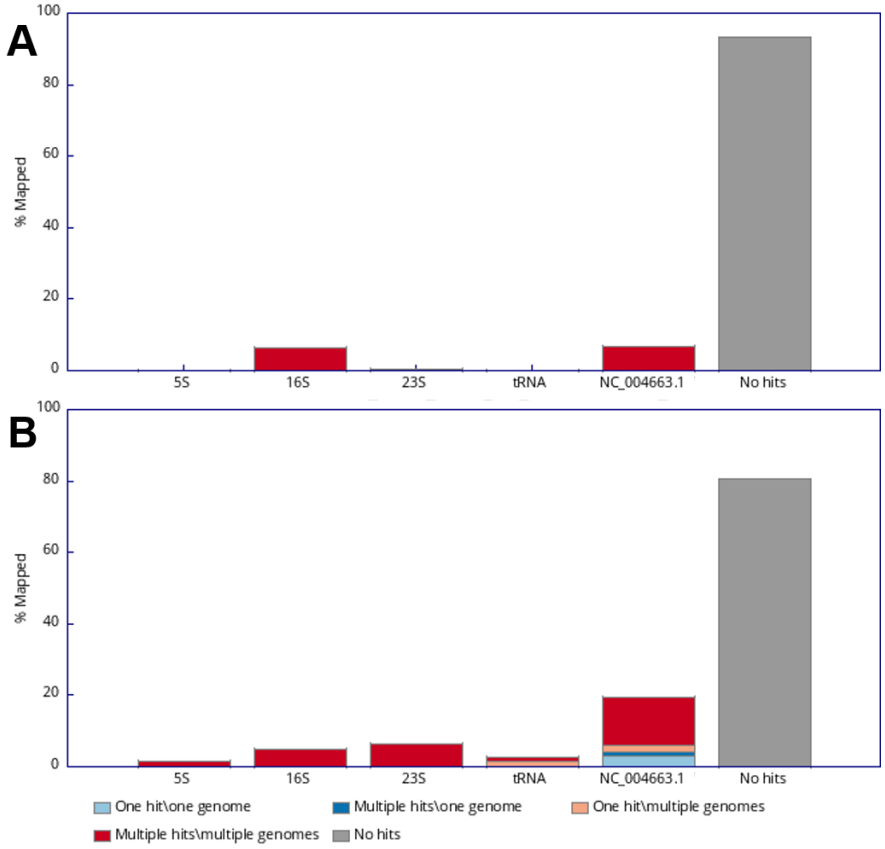

Figure 19: Custom FastQ Screen output for self-performed *B. thetaiotaomicron* experiment, shown one sample per approach with (A) RIBO I, (B) RNA I. Notable, nearly 90 % of reads are not aligned to the reference genome. However, when forwarded to BLAST analysis the reads can be aligned to the reference genome and genus. Presumably, shortened reads due to RNA degradation caused the non-alignment.

All four samples were subjected to computational evaluation according to the pipeline mentioned, the only difference here was the categorization of reads with FastQ Screen before filtering out rRNA and tRNA ones. With this upstream visualisation of reads, the efficiency of rRNA depletion can be analysed. Quality control after adapter removal and filtering reveals at most only 1 % of reads left for evaluation which will not be sufficient even if samples are sequenced unmated. Notably, FastQ Screen output of sample I RIBO (Figure 19A) and RNA (Figure 19B) shows 80 % and more reads cannot be aligned to *B. thetaiotaomicrons'* genome (NC_004663.1). FastQ Screen was set up only report reads mapping to the reference genome. Reads categorized as either one hit or multiple hits to multiple genomes are referring to the different rRNA variations (as multiple genomes). Reads for which no hits could be identified were extracted into a separate file and subsequently blasted. Most reads mapped to partial 16S rRNA sequences within *Bacteroides* genus, *Enterococcus*, *Staphylococcus* or *Escherichia*. As the sequence specificity within 16S rRNA is very high slight changes within can lead to mismatches during alignment. These changes might be sequencing errors caused by the machine. For sample II similar results were reported by FastQ Screen. Thus, the test sequencing revealed insufficient library

composition for an appropriate RIBO-Seq and RNA-Seq experiment. For both samples, a decision was made rejecting a stand-alone sequencing experiment as it would be too expensive. Only if single samples were sequenced with a starting coverage of at least 2 billion reads and outcome of 1 % (20 million reads) could be used for evaluation if rRNA is not sufficient. The insufficiency of the performed rRNA depletion can be noted here as nearly all reads, whether categorized as mapping to the reference genome or not can be matched to partial 16S rRNA sequences.

Available RIBO-Seq data for *B. thetaiotaomicron* were obtained from ENA and analysed consistent with the self-generated samples. Interestingly, again nearly none of the reads (≤ 1 %) were left after trimming and filtering. Again, FastQ Screen was used to analyse the categorization of reads.

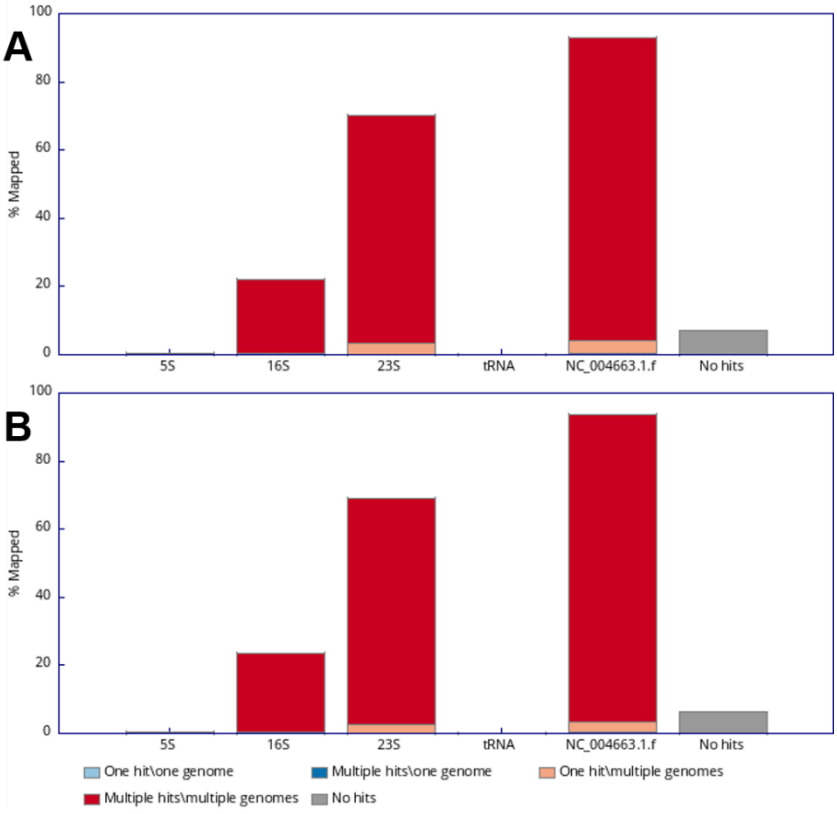

Figure 20: Custom FastQ Screen output for publicly available *B. thetaiotaomicron* samples A) RIBO I, B) RIBO II (Sberro et al., 2019). Contrary to self-performed experiment (Figure 19) most reads were aligned to the reference genome. However, those alignments showed a high abundance of reads corresponding to 16S and 23S respectively. Those are excluded before subsequent read evaluation, with only a low number of reads left. Therefore, an evaluation was also not possible.

Inspection revealed nearly all reads that did not remain for evaluation were mapped to rRNA variants for the first RIBO-Seq (Figure 20A) and second RIBO-Seq (Figure 20B) sample. Unfortunately, a downstream evaluation regarding the detection of overlapping genes was not possible for any of the two samples.

Therefore, additional experimental proceedings will be needed, at best starting with a new cultivation. Changes made during downstream sample handling did not prevent RNA degradation in total. Thus, the

assumption was made that RNA degradation already started during harvest. Additionally, an appropriate ribosomal stalling could be induced to prevent ribosomal run-off from actually translated genes.

The various analyses performed as well as the just mentioned experiment were performed to verify the existence of overlapping genes present in a variety of prokaryotic genomes and to give recommendations to further improve these findings both experimentally and during evaluation. The results will be discussed in their scientific context in the following chapter.

4. Discussion

4.1 Improvements during the experimental RIBO-Seq processing

4.1.1 Necessary read coverage

Sufficient read coverage is a crucial factor in any performed sequencing experiment, as it ensures the potential detection ability for every transcript or genomic variation present within the analysed sample. Still, prediction tools such as REPARATION require a minimum amount of reads (here 3) to consider a potential ORF being of interest (Ndah et al., 2017). This implies not only the sufficient coverage of every position within the genome for example necessary in *de novo* sequencing approaches but to have multiple reads covering a locus of interest in the mentioned scenario. Nevertheless, a sufficient amount of read coverage is highly dependent on the scientific issue (Haas et al., 2012; Sims, Sudbery, Ilott, Heger, & Ponting, 2014). Two options can be used to increase read coverage, namely, to increase the general read coverage used for sequencing which is also dependent on the machine used. Second, to enrich for fragments of interest before sequencing either by targeted enrichment of those or depletion of abundant fragments (Haas et al., 2012).

Experiments focused on RNA-Seq have shown that sufficient coverage for result evaluation is given at around 5 Mio. reads (Haas et al., 2012). Furthermore, exceeding a coverage of around 50 million is associated with the generation of spurious reads interfering with the correct interpretation of obtained results (Haas et al., 2012). The performed analysis within this thesis is contributing to the determination of a necessary read coverage for RIBO-Seq results. Compared to the amount sufficient for RNA-Seq, for RIBO-Seq experiments, the recommendation is to obtain at least 20 million reads after alignment and excludes reads covering tRNA or rRNA (see Section 3.1.1.). With this amount the detection of ~ 82 % of annotated genes for *E. coli* K 12 was possible (Glaub et al., 2020). Within an RNA-Seq experiment, the detection for nearly 99 % of the annotated genes was possible, but only one read was used for ORF validation, whereas in this thesis at least three reads were necessary for ORF detection (Glaub et al., 2020; Haas et al., 2012). Only one read for validation is considered insufficient as it cannot be excluded from being spurious and therefore wrongly aligned to a position of interest.

Still, the need for higher coverage in RIBO-Seq was not expected as more transcripts are assumed to be available for sequencing in RNA-Seq samples. Here, also sequences that are not translated such as non-coding RNA needed to be covered to obtain the whole transcriptome. A major difference between the here performed analysis and the one focussing on RNA-Seq it is compared to is the number of reads necessary for ORF evaluation. Therefore, here at least the triple number of reads is necessary to consider an ORF as of interest. Furthermore, even with the higher coverage, the reproducibility of annotated gene detection success of 99 % in the RNA-Seq experiment was not possible. Responsible here may be the fact that not every transcript present is subsequently translated, e.g., long non-coding RNA or short interfering RNA (siRNA) (Waters & Storz, 2009; Zur Bruegge, Einspanier, & Sharbati, 2017). Additional, even in RNA-Seq a detection of nearly all known genes was not expected as not all genes

present in the genome are assumed to be transcribed simultaneously. Rather the transcription status in bacteria is reflecting the actual condition and phase the organism is present in. Environmental or growth phase changes lead to altered transcription states with different transcripts present (Calviello & Ohler, 2017; Jiang et al., 2015). The low threshold of only one read considered for ORF evaluation in RNA-Seq most likely caused the high detection efficiency.

The decreased amount of detection in RIBO-Seq experiments potentially is also affected by the required read distribution across an ORF. Not only do prediction tools such as DeepRibo or REPARATION consider an ORF as being of interest only with a certain read coverage but also require a distinct pattern of reads covering an ORF of interest. Particularly, the focus is on start and stop region coverage as well as on a proportion of start region compared to the rest of a gene (Clauwaert et al., 2019; Ndah et al., 2017). This read pattern evaluation should ensure the detection of only actual translated ORFs but may also be responsible for a decreased prediction efficiency. These patterns might not be given under circumstances such as drug-induced ribosomal stalling. As mentioned, some drugs are inhibiting translation elongation, therefore might hinder read pattern construction in the stop region if ribosomes were not covering this gene proportion at the point of application. Additional, internal structures might cause ribosomal pausing during translation (Li et al., 2012) again preventing read distribution necessary that is recognized by the prediction tool used for ORF estimation. Therefore, these ORFs might be missed even if translated.

Another crucial factor influencing the read coverage is rRNA presence. The recommendation of at least 20 million reads is excluding ones' mapping to rRNA. As 85-95 % of RNA present in an organism is considered rRNA a sufficient depletion is crucial for successful RNA focused sequencing experiments (Z. Chen & Duan, 2011; Petrova et al., 2017). Arguable, depletion might be even more important in RIBO-Seq experiments as targeting ribosomal protected mRNA fragments is simultaneously accompanied by the enrichment of rRNA. For every mRNA fragment kept all three types of rRNA (5S, 16S and 23S) are also retained once as they are part of the ribosome (Bhavsar, Makley, & Tsonis, 2010; Melnikov et al., 2012). Hence, the proportion of mRNA in comparison to rRNA is low emphasizing the need of rRNA depletion before sequencing. RNA-Seq is also mostly focussing on mRNA to obtain information about the transcriptome present but a major difference here is that whole mRNA transcripts are not affected by ribosomal coverage. Therefore, the amount of rRNA is still high but not only are several fragments per transcript necessary but these are not protected by the triple amount of rRNA. Thus, the rRNA depletion performed might be more successful as the input amount of rRNA could be lower compared to RIBO-Seq experiments.

Besides necessary depletion, another potential factor already reducing the amount of rRNA present could be the adapted range of fragment size selection during RIBO-Seq experiments.

4.1.2 Appropriate size selection for mRNA corresponding fragments

First and foremost, size selection is performed to only target monosome protected fragments. This should ensure to obtain mRNA fragments being translated at the point of harvest. Polysome structures are digested before size selection resulting in several monosomes also present to obtain for further processing. The appropriate size selection range is of especial interest to aim for protected fragments and simultaneously reduce the amount of potential faulty fragments, e.g., with remained undigested sequence at the edge of a ribosome. To identify the range of interest length of sequenced mRNA fragments was analysed. Additionally to that, lengths corresponding to either rRNA or tRNA were evaluated to potentially detect representative sizes which could be used for size-specific depletion.

The results obtained in this thesis indicate that prokaryotic protected mRNA fragments range between 24 - 27 nucleotides in length (see Figure 6) whereas eukaryotic fragments are slightly longer ranging from 28 - 30 nt (Glaub et al., 2020; Ingolia et al., 2012; Ingolia et al., 2009). With this result, a recommendation for an adapted selection range from original protocols is made for prokaryotic RIBO-Seq experiments in general. This recommendation contrasts with previously published size selections, as these are ranging from 15 - 40 nt. This wide range should ensure capturing all reads of interest, but still, the most informative length is claimed to be of 24 nt in length (Li et al., 2014; Mohammad et al., 2019). With the here proclaimed range targeting a smaller variety of fragment lengths, the informative ones are also obtained with a potential reduction of unwanted fragments corresponding to either rRNA or tRNA (Glaub et al., 2020).

In comparison to mRNA, the length of fragments representing tRNA is longer, with the most dominant length of 32 and 35 nt respectively (Glaub et al., 2020). Size selection is achieved by excision of a gel band from a denaturing gel. Thus, as fragment size is orientated on a corresponding ladder with subsequent manual excision, it is highly unlikely to exactly cut at the border of an aimed length. Therefore, it is expected to always include a minor amount of slightly shorter and longer fragments as is aimed for. With the here recommended narrower selection range the additionally incorporated fragments could potentially be kept at a minimum. This claim was validated with an analysis focused on the amount of tRNA present in samples (see Supplementary Table S11) which were obtained from a range between 20 - 30 nt (Glaub et al., 2020). Here, especially the upper border of the selected range was of interest, as the most dominant tRNA corresponding length were identified as exceeding at this point. Indeed, for samples obtained from this range, less than 1 % of tRNA remained. With a shift of the upper border a successive increase of tRNA amount present could be detected (Glaub et al., 2020). Therefore, an aimed for cut off at around 30 nt or even lower should aid in the reduction of tRNA present in the analysed sample already before sequencing.

However, a similar trend of size selection-based depletion cannot be identified for remaining rRNA. One contributing factor might be that the most dominant length for rRNA corresponding fragments were 26 and 31 nt, respectively (Glaub et al., 2020). Longer fragments might be excluded with a narrow size

selection but reads with a length of 26 nt will always be kept due to their length similarity to fragments of interest. Additionally, as rRNA is the highest abundant RNA type present (Petrova et al., 2017) sufficient depletion is necessary for reduction. Here, it might be of especial interest if a specific type of rRNA is mostly represented by a certain length. Various available depletion kits target different types of rRNA; therefore, the selection of the appropriate kit might aid in the performance of a sufficient rRNA depletion.

The analysis revealed that, even to some extent similar, the various rRNA types can be assigned to different read length. Reads corresponding to 5S are generally longer with most of them having a length of 31 or 32 nt respectively. Likewise, a proportion of fragments obtained from 23S sequences are also 31 nt in length, whereas a second dominant length for this type of rRNA is 26 nt. The majority of reads belonging to 16S sequences show a length of 26 nt (Glaub et al., 2020). Common depletion kits such as RiboMinus (ThermoFisher) or MICROBExpress (ThermoFisher) only target 16S and 23S sequences. Here, it is especially important to also reduce the amount of 5S fragments present within a sample. With the performed analysis fragments with a length of 31 and 32 nt respectively can be assigned to 5S rRNA sequences. With this information, a slightly lower set upper size selection border can be used to potentially exclude a proportion of 5S rRNA fragments from samples. The development of siTOOLs rRNA depletion kits also targeting 5S sequences according to the manufacturer leads to redundancy of prior 5S reduction based on size selection. Nevertheless, a narrow range from 22 - 30 nt includes most likely all fragments of interest and simultaneously excludes longer fragments corresponding to tRNA (Glaub et al., 2020).

It is important to emphasize that size selection highly depends on the aim of the study. A narrower selection can aid in the reduction of unwanted fragments but could potentially also exclude fragments of interest hitherto unknown. An analysis focussed on the upstream region of translation start site revealed such longer fragments that are subject of the next chapter.

4.1.3 Read length variation upstream of the translation start

So far there has been no consensus about ribosomal protected fragment length within prokaryotes. The first results mentioned showed that there is already variation within different types of RNA present. Nevertheless, a certain range (24 - 27 nt) in fragment length can be assigned to mRNA (Glaub et al., 2020). However, other studies have shown that internal sequences motifs by interaction with ribosomes cause the protection of longer RNA fragments (Li et al., 2012; Mohammad et al., 2019; Mohammad et al., 2016). First studies assumed that an SD like motif within a sequence interferes with the ribosome causing translational pausing (Li et al., 2012), whereas more recent studies specifically named a glycine codon as responsible for slowed down translation (Mohammad et al., 2016). Still, the interaction between SD motif and anti-SD motif within a ribosome is identified to protected mRNA fragments

against nuclease digestion resulting in longer mRNA fragments (Mohammad et al., 2019; Mohammad et al., 2016). Due to these findings, an analysis performed focused on potential longer fragments mapped upstream of annotated genes.

The model organism *E. coli* K-12 chosen in this study is known for its presence of SD motifs upstream of annotated genes (Amin, Yurovsky, Chen, Skiena, & Futcher, 2018; Shine & Dalgarno, 1974). Therefore, a comparison of their upstream region where the supposed SD motif is located and the same length region downstream of the translation start was analysed according to the length of mapped fragments. The initial analysis (see Section 3.1.3.) revealed that reads mapped in the upstream region tend to be longer (~ 34 nt) than reads aligned downstream of translation start site (~ 27 nt). Furthermore, additional analyses focussed on fragment length for reads covering the whole genes' length and a certain area upstream of the stop region showed a similar length distribution as in the start region (see Figure 8) (Glaub et al., 2020). As hypothesized fragments assigned to the 5'-UTR region of annotated genes were found to be longer, presumably based on an interaction with the SD motif located in this area (Glaub et al., 2020).

However, the detected representative length for protein-coding areas seems to be longer than initially reported. As claimed obtained from the general length comparison within the different RNA types, the mRNA corresponding fragments range between 24 to 26 nucleotides. Here, the most frequently represented length for coding sequence is 27 nt (Glaub et al., 2020). This length discrepancy can be explained by study design primarily. For the general RNA type analysis, a broader variety of samples could be used for analysis as the only restriction chosen was a read length range from 20 to 40 nucleotides. However, for the analysis specifically focused on potential longer reads present fewer samples (n = 30) were chosen (Glaub et al., 2020). This restriction was based on the characteristics within the samples, as they did not include reads up to 40 nucleotides in length, therefore, were excluded from the analysis. This read length, however, was of especial interest, as it has been published that SD like motif caused interactions result in longer fragments (Li et al., 2012; Mohammad et al., 2016).

Certainly, it is arguable if the length range representing mRNA fragments is biased by samples included with a narrower length distribution in general not covering longer fragments or size selection in general. That is why for the general RNA type analysis all samples were kept showing a wide spectrum of size selection aimed for. Interestingly, accordance between aimed for and actual fragment length was every rarely detectable (Glaub et al., 2020). A minor proportion on fragment length variation can be assigned to size selection as it is possible to exclude the majority of certain lengths if the experimental step is performed appropriate and exact. Nevertheless, the excision will never be 100 % accurate, therefore, a small proportion of fragments slightly smaller or longer as aimed for might be present. The most determining factor on fragment length, however, is the ribosome itself. But even here variation within protection capability is seen as so far, the identification of one correct length for protected mRNA

fragments was not possible. Potentially contributing to length variations is insufficient digestion of unprotected fragments.

In eukaryotic RIBO-Seq experiments, the most commonly used nuclease is RNase I as it lacks a nucleotide cleavage bias (Ingolia, 2010; Ingolia et al., 2012). Due to its claimed inactivity within bacteria based on an interaction with the 30S ribosome subunit for prokaryotic RIBO-Seq experiments another enzyme is needed (Glaub et al., 2020; Kitahara & Miyazaki, 2011). Oftentimes used as an alternative is micrococcus nuclease despite its known sequence specificity which might cause the monitored sequence length variability (Dingwall et al., 1981; Glaub et al., 2020). Also, a mixture of endo- and exonucleases can be used to ensure sufficient digestion of unprotected mRNA sequence (Hücker, Ardern, et al., 2017). Nevertheless, ensuring a 100 % sufficient digestion might never be possible as internal mechanisms or structural folding could prevent parts from digestion. Even as the field of sequencing with its steady development is improving our understanding of many mechanisms might still be unknown.

A potential other factor resulting in different fragment length might be translation performed by alternative ribosomes. These have already been described for *E. coli* and *M. smegmatis*, with differences to canonical ribosomes mostly based on lack of specific proteins or alternative assembled ribosomal subunits (Y.-X. Chen et al., 2020; van de Waterbeemd et al., 2017). Additionally, it has been shown that artificially enriched alternative ribosomes based RIBO-Seq results are different compared to an identical performed standard RIBO-Seq experiment. Not only were different read accumulations observed but also shifted codon usage patterns and even a seemingly selective translation bias for only a proportion of the genes (Y.-X. Chen et al., 2020). Especially the finding of divergent codon usage might lead to a bias in protected mRNA length. Ribosomal pausing is associated with the incorporation of certain amino acids (Buskirk & Green, 2017). Conformation changes due to the paused translation could result in length variations within the mRNA fragment proportions protected. In general, it could also be speculated that the different assembly itself within alternative ribosomes might lead to a changed protected fragment length. This could be subject of further analysis using the results obtained from the comparison analysis performed by Chen et al. A counterargument for longer fragments in this analysis caused by alternative ribosomes is their specific activity. Presumably, they are particularly involved in stress adaptation (Y.-X. Chen et al., 2020), whereas samples used for these comparison analyses were all grown in standard LB medium. Therefore, it is highly unlikely that this type of ribosomes, in general, was active within the experiments used.

Nevertheless, results from this analysis once more are underlining the importance of size selection adaption. A range between 24 to 27 nt is the most appropriate one to obtain mRNA fragments for prokaryotic RIBO-Seq experiments. In parallel, due to the narrow range, unwanted fragments present can already be decreased even if the selection is not completely precise. However, if an analysis focuses on 5'-UTR regions it is necessary to aim for longer fragments as it has been shown that mapped reads

in this particular area tend to be longer. The last analysis performed focused on the improved start site detection by chloramphenicol application.

4.1.4 Start site detection improvements due to chloramphenicol application

Previous studies have shown that chloramphenicol, in general, is suitable in translation start site detection as it causes ribosome accumulation at this position. The reagent binds within the peptidyl-transferase centre, therefore, hindering peptide bond formation resulting in stopped translation elongation (Wilson 2009). Still, as can be seen in the obtained results (see Section 3.1.4.) ribosomes can be located at subsequent sequence position indicating that elongation is not stopped completely. One reported explanation might be that stalling efficiency is dependent on the amino acids located within the second to last position of the translated protein (Wilson 2009). Nevertheless, application of Cm is clearly causing improved start site detection due to ribosomal stalling.

Interestingly, this bias might even contribute to improved detection of overlapping genes in general. Results obtained in this thesis show, those weakly expressed genes (RPKM between 10 - 20) (Glaub et al., 2020) are benefitting the most from the drug application (see Figure 10). This might be explained only by the number of ribosomes available for translation based on the different expression levels that were determined. Highly expressed genes are most likely to be translated by a multitude of ribosomes at the same time. Therefore, it is most likely that within all samples tested despite Cm application the translation start site is always occupied by a ribosome. In the experiment chosen for this comparison the time between stalling reagent application and subsequent cell harvest and general translation inhibition was two minutes (Oh et al., 2011). As mentioned, chloramphenicol hinders translation elongation, but initiation can still proceed (Mohammad et al., 2019; Mohammad et al., 2016; Oh et al., 2011). A slight increase in ribosomal occupied fragments is detectable for highly expressed genes but even without drug application, it is highly likely that the vast majority of transcripts is subjected to translation initiation just due to its increased requirement within the bacteria tested. Therefore, it is not surprising that the emphasis on start site detection for highly expressed genes is low as this location, in general, is oftentimes occupied by ribosomes.

Likewise, the slightly improved start site detection for medium expressed genes by chloramphenicol addition can be explained. Here, the detection is more enhanced in comparison to highly expressed genes but not as good as for weakly categorized genes. The number of transcripts present for genes within this category is lower than for highly expressed but higher as for weakly expressed genes. In general, translation is performed to an extent but not as efficiently as for highly expressed genes. Therefore, a still possible translation initiation after chloramphenicol application results in the detection of more

transcripts as the general initiation amount is lower in medium expressed genes compared to highly expressed ones.

The increased detection efficiency within weakly expressed genes can also be explained due to transcripts available and the expression status. For weakly expressed genes only a few transcripts are present within a sample and are only rarely translated. If translation elongation within an organism is stopped in general, it is speculated if ribosomal subunits present to assemble at highly expressed genes now be available for other transcripts. Due to the stopped elongation most of the highly expressed genes are occupied by ribosomes in general but under normal circumstances, namely, without induced stalling, several ribosomes would proceed at one transcript for a highly expressed gene. Nevertheless, the ribosomal subunits necessary for the ribosomal assembly at the mRNA are already constructed for translation purposes. These might then even assemble at rare transcripts present of weakly expressed genes. Under normal conditions, the same transcript might be translated but at the point of harvest, elongation could be ongoing therefore not emphasizing the translation start site. Therefore, based on the expression status itself and the point of harvest weakly expressed start sites are difficult to detect but are especially emphasized by chloramphenicol application. This improved detection is of especial interest as the involvement of undetected overlapping genes within the organisms' metabolism is unknown. Thus, the expression status remains unknown but even if only weakly expressed they could be detected by ribosomal stalling inducing drugs.

The recommendations for RIBO-Seq experiments obtained from these first analyses are not only improving gene detection in general but are assumed to also promote the detection of overlapping genes. The next chapter is primarily focusing on the verification of unknown OLGs within a broad spectrum of prokaryotic species.

4.2 Influence of genomic characteristics on eORFs and their phylogenetic analysis

4.2.1 Genomic features involvement in prediction efficiency

eORF values used for the following comparisons and evaluations are containing only those considered translated. This status is assumed based on thresholds matched (RPKM and coverage) during prediction. Neither genome size nor GC-content seems influential for eORF prediction efficiency as no correlation between ORF amount and either of the features was detectable. Read coverage normalization was performed to exclude the introduction of a bias. Nevertheless, no pattern for prediction efficiency based on either of the genomic features is detectable. Further, even within a genus, here *Streptomyces,* no correlation can be detected. Exemplary calculated median values for this genus is shown in Supplementary Table S12. And even within one species prediction efficiency per samples is highly different (see Supplementary Table S5). The comparison of the genomic features against eORFFamilies predicted could imply a slight influence of genome size on the numbers of eORFFamilies. However, an

R-squared value of 0.29 does not indicate a strong relation between the two variables (see Figure 13A). Still, as bigger genome sizes are usually accompanied by an increased number of genes present (see Supplementary Table S6) this correlation, in general, might be an enabling factor. Simply as more know genes are available an increased potential of locations to form embedded genes in is obtainable. The assumption that bigger genomes are more prone to the occurrence of overlapping genes just based on their size therefore holds up but cannot be statistically verified.

On the hypothesis of an increased GC-content being accompanied by fewer eORFs present, this correlation was analysed, but had to be neglected. There was no negative relation detectable between these variables which most likely can be explained due to stop codon adaptation within higher GC-content genomes. Commonly used stop codons within bacteria are TAA, TAG, TAC (Belinky et al., 2018; Wong et al., 2008) and just based on their construction a lower GC-content could be more suitable for their formation. Especially the most common TAA codon is solely created of pyrimidine bases, therefore the formation of this specific stop codon is more unlikely in high GC-content genomes. However, it has been shown that an increment in GC-content is accompanied by stop codon usage shift towards purine bases included triplets (Belinky et al., 2018; Povolotskaya et al., 2012). Thus, the limited presence of a particular stop codon, even if it is the most common one in bacteria, does not impact ORF creation.

Nevertheless, the total eORF numbers were used in regard to verify the recommended read coverage of effective reads needed for proper RIBO-Seq evaluation (see 3.1.1.). For each sample analysed the number of reads was plotted against the eORF amount predicted with the in-house ORFFinder script. Values for *E. coli* K12 were excluded as an overlap to the analysis focused on estimating the threshold necessary should be prevented. Additionally, as 192 samples are associated with the named species, a potential bias introduced to the number of samples should also be avoided.

Even if GC-content is not influential of the eORF prediction efficiency, its potential contribution in eORF length is subject of the next chapter.

4.2.2 eORF length distribution within species analysed

For this analysis, the hypothesis formulated focused on a potential relation between detected eORFs' lengths and the GC-content of the respective species' genome. An assumption was made that high GC-content could be accompanied by longer eORFs present, as the formation of the most commonly used stop codon TAA is more likely within a low GC-content. A counterargument here is the stop codon usage shift from TAA to TAG or TGA with an increase in GC-content (Belinky et al., 2018; Povolotskaya et al., 2012). Therefore, the amount of stop codon present should not be significantly lower compared to low GC-content genomes. Based on these facts it is not surprising that there is no significant eORF length difference across the genomes analysed (see Figure 14, Supplementary Figure S2). Within

the *Staphylococcus* genome eORFs in length up to 100 nt are predominant, compared to the remaining 23 genomes. Therefore, a slight trend for shorter (93 - 100 nt) eORFs in genomes with a GC-content up to ~ 43 % is detected, whereas then a shift from equivalent length distributions to a length between 100 - 200 nt as the most dominant occurs (see Figure 14). However, for *B. subtilis* only four eORFs within eight samples were detectable as translated (see Supplementary Table S3), implying potentially non-conclusive results for the length analysis solely based on a low prediction bias within this species.

Even if the proposed hypothesis of a significantly shifted length distribution had to be rejected, still a general detection of longer eORFs in higher GC-content genomes could be noted. Occasionally, even eORFs up to length spanning 1,300 nt were found in several samples within a species (in *P. aeruginosa*, *S. typhimuirum* and *H. volcanii*). Even without more descriptive analysis revealing its potential function, these ORFs could be associated with functionality due to their length. In general, the longer a sequence proportion is without an incorporated stop codon it is more likely to deduce some functionality. However, to maintain the functionality of beneficial sequences is a cost-effective evolutionary 'effort' for an organism. More precise this refers to negative selection, characterised by the evolutionary 'effort' to reverse spontaneous mutations within coding sequences to keep them functional (Cvijovic, Good, & Desai, 2018). Especially the formation of a stop codon within a translated ORF needs to be prevented in avoidance of functionality loss. With an increased ORF length, the possibility of such malfunction causing mutations is higher, especially the incorporation of additional stop codons. Thus, if long ORFs are detected at best even in other species, this, in general, implies functionality. Even if not translated this ORF might represent a long non-coding RNA, which are mostly involved in regulatory processes (Zur Bruegge et al., 2017). One re-occurring long eORF detected in *S. clavuligerus* was even subject of analysis discussed in the following chapters to potential obtain the first hint of its functionality.

As correlations between genome size and GC-content did not yield to significant results in regard to eORF detection efficiency or length, the next analysis focused on frame location relation of mother gene and overlap and whether one frame is in favour for its creation.

4.2.3 Relative reading frame analysis of mother gene and overlap

The positions within a codon are of different importance determining specific characteristics of the amino acid they represent. For a long time, it was believed that the 1st position is the most descriptive one, whereas the 3rd position was referred to as the "wobble-base" (Saier, 2019). Within this position, a nucleotide exchange most likely does not lead to an alternative amino acid incorporation into the translated protein sequence. This fact can be underlined by the codon degeneracy phenomenon, which states that there are more codon possibilities than amino acid they could code for. Therefore, several codons, which mostly vary within the third codon position, code for the same amino acid (see Figure 1A) (Esberg, 2007). However, more recent studies showed that the main determinant of the

characteristics of the amino acids within a codon seems to be the 2nd position (Bofkin & Goldman, 2007; Saier, 2019). In particular, the nucleotide located within this position is specifying the amino acids' charge which can highly influence the resulting proteins' formation (Blazej et al., 2018; Saier, 2019).

Based on these different importances of codon positions the relative reading frame relation of sas11 is expected to be in favour of the creation of an overlap in general. Here, each 3rd codon position within the mother gene is complementary to the 2nd position within the overlap (see Figure 2A). The lower evolutionary pressure for the third position could allow exchanges there potentially not affecting the sequence of the mother gene (Bofkin & Goldman, 2007). However, a nucleotide exchange at the second position, here in the overlap, definitely causes an amino acid exchange. Either the formation of a start or stop codon could lead to the creation of an eORF with a difference in length or a substitution within the sequence can result in an alternative conformation. Therefore, a single exchange could cause major changes leading to actual translation and/or functionality of an overlap.

Results from the frame relation analysis (see Figure 15) do not support the hypothesis of sas11 being favoured. Moreover, for this relation, a significant difference can be detected compared to the general possibility of eORF creation. Therefore, it is more likely to detect a none translated eORF than an actual translated one. Two facts might be utilised to explain this result. First, as mentioned the 3rd codon position is less determining for the amino acid incorporated and, therefore, under lower evolutionary pressure (Bofkin & Goldman, 2007; Saier, 2019). An exchange can remain unnoticed due to the codon degeneracy and the same amino acid being incorporated for the mother gene (Gonzalez et al., 2019). Eight amino acids are not affected by a nucleotide exchange in the third position (see Figure 1A; Ala, Arg, Gly, Leu, Pro, Ser, Thr, Val). However, as this is beneficial for maintaining the mother genes' sequence, the most important codon position within the overlap is highly affected. Whilst one exchange might be beneficial for the formation of a functional protein sequence, another one could reverse this mutational caused positive effect resulting in loss of functionality again. The evolutionary pressure for the 2nd codon position is assumed to be more decisive than for the 3rd position within the mother gene. Of course, to maintain an additional potentially functional protein (here the embedded overlap) can be beneficial. However, if the necessity of the proteins' function is limited to distinct processes and not ubiquitous its benefits might remain undetected at the point of creation. Nevertheless, the functionality of the mother gene regardless of nucleotide exchanges within certain amino acids in the 3rd codon position maintains. Therefore, cost-effective mechanisms to change the nucleotide sequence to obtain a certain structure are not necessary. Here, the functionality of the mother gene is more important whilst less energy-consuming than maintaining a potentially selectively functional gene.

Another factor why sas11 could not be identified as being in favour of the mother gene and overlap formation might be caused by difficulties of actual detection. Results obtained show a snapshot of the mother gene and overlap relation. By analysing different species all at once a species-specific bias can be excluded. However, differentiation between creation and maintenance regarding frame location

relation needs to be discussed. The perspective of evolution is more likely in favour within the sas11 relation based on the complementary positions explained for mother and overlapping gene. However, the survival of an eORF is presumably the strongest within the relative reading frame sas12. Here, not only the complementary position of the 3rd codons for both mother gene and overlap is decisive but also location relation of the 2nd position to the 1st in both orientations (see Figure 2A). The nucleotide-based exchange in position three in sas11 might cause the incorporation of a different amino acid in the overlap. However, in sas12 whilst exchanged in the 3rd position the functionality can maintain in both sequences due to the complementary location of both third codon positions. Again, the 3rd codon position is more likely to be exchanged due to lower evolutionary pressure (Bofkin & Goldman, 2007; Saier, 2019), however, the exchange sometimes does not affect the amino acid sequence, therefore no loss in functionality is caused. Furthermore, in the sas12 frame location, this does not affect the mother gene as well as the overlap. This would be favoured to maintain the functionality of both sequences. Additionally, the most decisive 2nd position within the mother gene is complementarily located to the 1st position in the overlap and vice versa (see Figure 2A). Thus, as nucleotide exchanges at the second position are most likely prevented or reversed to preserve the amino acid incorporated at this position in this relation are presumably protected by each other. Therefore, this relation is highly appropriate to preserve both gene sequences and consequently the resulting amino acids.

Results from the analysis performed support the hypothesis that sas12 could be in favour to maintain both genes as the detection of translated eORF is significantly higher in this frame location relation (see Figure 15). However, still, the assumption remains of sas11 being in favour of eORF creation. This relation might be considered as a gene nursery, where the creation of a potential functional gene arises. Yet, to potentially maintain its functionality it is highly likely to be integrated by insertion into another genomic position in the +1 frame. Further nucleotide exchanges at the former overlap location might cause the potential loss of functionality there. However, the integrated previously overlapping gene will not be affected.

In conclusion, the creation of eORFs seems more likely in sas11, whereas they maintain especially in the relative reading frame relation sas12 thus are most likely detected there. Another potential indicator to verify an eORF of interest is its detected occurrence within several samples of one species. The repeated detection of the same eORF is assumed to distinguish it from spurious signals. In the following chapter, such eORFs were detected and forwarded to further analyses.

4.2.4 BLAST-based age categorisation of eORFs of interest

The general detection of eORFs within a genome is the first implication of their existence. However, more descriptive analysis, both informatics and experimental, are necessary to distinguish between authentic and spurious predictions. Criteria like RPKM and coverage were matched by all referred to

eORFs within this thesis. Still, further analyses were performed to narrow down the number of candidates predicted based on criteria used to verify the reliability of their detection. A subset of eORFs was determined based on their re-occurrence within several samples of one species. Hereby, the random occurrence of an eORF is eliminated thus it is more likely authentic. Those identified candidates were the content of further analyses to potentially determine reliable eORFs that could subsequently be experimentally verified.

The first hint of a genes' age can be obtained by analysing its occurrence within the phylogenetic tree. In general, this is of interest as the context of a genes' age can be used to draw a first conclusion regarding its functionality. Essential genes that are necessary to maintain an organisms' metabolism are found throughout the phylogenetic tree due to their vital importance (Jordan, Rogozin, Wolf, & Koonin, 2002; Luo, Gao, & Lin, 2015). Here, a commonly known example is the 16S rRNA gene whose occurrence within each living organism can be used to precisely distinguish e.g., bacteria of the same species (Johnson et al., 2019). Based on its important functionality accompanied by its ubiquitous occurrence, it is clearly defined as being 'old'. This means by implication, the further distributed a gene is within the phylogenetic tree, it is more likely to be of important functionality. Hence, it is kept in the genome of diverse organisms with a less close relation.

An analysis to categorize the subset of eORFs identified within 'older' or 'younger' genes was performed. As OLGs are more likely of recent origin implied by a smaller distribution within the phylogenetic tree search parameters were adapted to this. The tblastn database used for homologues identification contained only family-specific genomes. Sequence similarity with an e-value of at least $1*e^{-10}$ was required. Results for both mother gene and overlap were categorized into 'old' or 'young' gene with especial interest in eORFs categorized as 'old'.

For ten pairings both genes are considered 'old' based on the number of family-specific homologues found (see Table 11). Interestingly, eight of these are associated with the *Enterobacteriaceae* family. One possible explanation might be the well-studied variety of *E. coli* genomes included in this family. Each analysis based on database comparison will be biased by its compilation. Here, with family-specific genomes, a fairly broad cut-off was chosen to ensure non-species-specific identification of homologues. Simultaneously, a bias is introduced based on sequenced genome availability within the family to search against. Nevertheless, with a broader spectrum, the possibility of homologue detection can be given even for less well-studied species. Anyhow, the eORFs categorized as 'old' should be subject of further analysis, as their ubiquitous occurrence implies functionality. To maintain a sequence in its original order is cost-effective in terms of purifying (negative) selection (Cvijovic et al., 2018; Rogozin et al., 2002). Therefore, similar coding sequences found in a variety of species distributed across the phylogenetic tree are assumed to be 'older' and of functionality.

eORFs categorized as 'young' whereas the mother gene is 'old' are presumably indeed of recent origin. Within the families *Mycobacteriaceae* and *Streptomycetaceae* mother genes for both categories were

identified. Therefore, the genomic diversity within the database is assumed sufficient as a variety of homologues at least for a proportion of mother genes were found. Hence, speculation of insufficient database coverage responsible for low homologues numbers can be rejected. However, even if categorized as ‘younger’ this does not imply non-functional eORFs. They solely occurred later within the phylogenetic tree as they potentially were created more recently. Additionally, it is important to state that there is no correlation between mother genes’ and eORFs’ occurrence within the phylogenetic tree.

If both genes of a pairing were categorized ‘young’ here the database compilation might be the limiting factor. Unfortunately, for the candidates within this analysis, this factor cannot be ruled out as there are no detections made for an ‘old’ gene within neither *Caulobacteraceae* nor *Halobacteriaceae*. Nevertheless, as these eORFs also matched the prediction threshold and were detected multiple times they are still of interest. A pairing where the mother gene would be categorized ‘young’ whereas the overlap is assigned ‘old’ was not detected. Even if not detected as was expected, this could be explained by potential miss-annotation of the mother gene during whole genome sequencing.

The same candidates from this analysis were once again forwarded to a tblastn analysis, however, this time to a genus-specific database. Purifying selection of the eORF sequence was analysed with OLGenie based on a comparison to the best analysis obtained from the BLAST analysis.

4.2.5 OLGenie based detection of purifying selection on eORFs of interest

A genes’ nucleotide sequence is dictating the emerging amino acid sequence. Based on a codon’s triplet structure a succession of three nucleotides is determined for an incorporated amino acid. Within this nucleotide triplet, as mentioned the positions are individually decisive for the amino acids’ characteristics (Saier, 2019). Nucleotide substitutions within a triplet can result in two different outcomes, either the original amino acid is still incorporated, or a different amino acid takes its place. Synonymous substitutions are causing no alteration in the resulting protein sequence, whereas nonsynonymous substitutions result in a different, potentially non-functional protein (Ina, 1996). Therefore, nonsynonymous substitutions are required to be reversed in functional sequences if they cause loss of functionality (Cvijovic et al., 2018). This process is also known as purifying (negative) selection (Rogozin et al., 2002).

An emerged beneficial sequence is kept within a genome based on two different evolutionary processes. As described one possibility is purifying selection, characterised by hindering the incorporation of deleterious mutations. Here, the functional sequence is already implemented into a genome and whilst evolution is proceeding its functionality is sustained by purifying selection (Cvijovic et al., 2018). Contrary, positive selection is supporting the integration of a new functional gene into a genome (Tan & Riley, 1997). Either a new gene can be introduced by e.g., horizontal gene transfer or a sequence

change within an already existing sequence can cause its functionality. In those cases, when the obtained sequences are beneficial for the organism their integration into the genome is supported by positive selection (Tan & Riley, 1997). Hence, positive selection is associated with new gene integration whereas purifying selection is working on maintaining 'older' genes functionality.

OLGenie is a tool used to calculate the ratio of synonymous and nonsynonymous nucleotide exchanges within a sequence (Nelson, Ardern, & Wei, 2020). With significance estimation conclusions can be drawn whether purifying selection is working on the query sequence, hence it is maintained due to functionality. However, characterising for this tool is its ability to calculate these ratios for overlapping genes considering the substitution differences within the alternative frames (Nelson, Ardern, & Wei, 2020). Thus, an analysis regarding purifying selection on the eORFs of interest was performed, revealing that for only two of the 28 analysed eORFs significant p-values were detectable. Therefore, only for those two purifying selection is assumed simultaneously implying that they are potentially functional. However, six additional eORFs were identified whose p-values are close to the significance level (see Table 12). For two of those, their sequence was categorized as 'old' in the previous analysis performed. Those correspondences are of especial interest, as both types of analyses are considered as potential first indicators for functional gene sequences. However, in general, eight eORFs were identified that would be of interest for further descriptive analysis.

In comparison to the previously performed analysis, selection detection is considered more conclusive based on significance determination. Nevertheless, as these analyses are performed *in silico* and can be biased by database sizes and compilation it is always recommended to not exclusively rely on one approach only. In consideration of their significance, however, here the results would be narrowed down to eight eORFs of interest. Still, to discuss the result in a broader context, all 43 eORFs obtained from re-occurrence analysis were used for the following Frameshift analysis.

Briefly, the corresponding mother gene to Streptomyces16 should be mentioned here, as its calculated p-value clearly shows no selection for its sequence at all (see Table 12). A potential explanation might be this locus representing a pseudo gene. Thus, an annotation is available, but no functionality is associated with this locus. However, a reconciliation with the species genome (here *S. clavuligerus*) revealed an annotation for a vitamin B12-dependent ribonucleotide reductase. Experimental analysis should be performed regarding the coded proteins' functionality. Additional, targeted sequencing for this gene could be performed to check whether the annotation was made correctly in this case. Based on the obtained p-value there is no indication for selection on this sequence indicating that maintaining the sequence order is not vital. Therefore, it is highly unlikely to be beneficial or functional at all.

4.2.6 Probability of creation based on eORFs length

This last analysis performed is analysing whether the length of an eORF present can be explained just based on random codon structure or if it is significant in its totality. The sequence of an overlapping gene is highly dependent on the genomic structure of the mother gene (Krakauer, 2000). Thus, random shuffling of the mother genes' codon structure can lead to the creation of random eORFs possible. Now, with a tool called Frameshift a comparison between these randomly created OLGs can be made (Schlub et al., 2018). Based on the mother gene codons' permutation and the subsequently resulting eORFs a significance value can be estimated revealing if the resulting length is longer than expected (Schlub et al., 2018). The results obtained are assumed to be related to potential functionality as otherwise a necessity for this particular ORF length is not given.

Here, for none of the eORFs analysed statistically significant p-values were calculated. This was surprising as a variety of length was tested (see Supplementary Table S8) and at least for longer eORFs a significance level was expected. This expectation was based on the assumption that long ORFs are associated with functionality, whether they are actually translated (Xu et al., 2006) or are only involved in transcription regulation (Guttman, Russell, Ingolia, Weissman, & Lander, 2013; Harris & Breaker, 2018). Both possibilities are of interest for the analysis of overlapping genes. However, as the analysed data is obtained from RIBO-Seq results eORFs identified here should even be associated with functionality in terms of a resulting translated protein due to ribosomal occupation (Ingolia, 2014).

Candidate Streptomyces8 has a length of 1,058 nt, however, the calculated p-value is far off significance. Solely based on its length it was assumed to be significant but interestingly, the analysis revealed even longer variants possible within the mother genes' permuted sequence. Nevertheless, for three candidates nearly significant p-values were estimated. Moreover, three of them (Escherichia1, Escherichia3, Escherichia4) were already classified as 'older' genes in Section 3.2.4. Thus, functionality for these eORFs can be assumed as their sequences are distributed in the family *Enterobacteriaceae* and more likely to arise directed than randomly. Interestingly, two of the three sequences were exact matches, whereas in the third nucleotide exchanges lead to two different amino acids incorporated (see Figure S3). All of the candidates are located within the same genome, however, as their respective mother gene is the IS1 transposase B a re-occurrence within the genome is not surprising. As transposases are used for genomic rearrangements, they can commonly be located multiple times within a genome (Sekine & Ohtsubo, 1989).

Yet, the detected amino acid exchange for the identified transposase the identified overlap is of interest. A comparison of the two sequences is shown in Figure S3, whereas the top one corresponds to candidate Escherichia1. Selection detection (see Section 3.2.5.) as well as length significance analysis (see Section 3.2.6.) for this candidate indicate a lesser probability to code for a functional protein. One amino acid exchange can be detected in the sixth position where cysteine is replaced by serine. Here, no major difference in the proteins' secondary structure was expected as an exchange between polar and

uncharged acids should not affect the conformation (Barnes & Gray, 2003). However, claims were made that cysteine-to-serine substitutions reduce protein activity (Pavlin et al., 2019; Smith & Marnett, 1996) whereas contrary statements report no alteration in this type of exchange (Barnes & Gray, 2003). In general, from the obtained results here a cysteine-to-serine exchange could at least be associated with functionality as those sequences with serine incorporated show a preference in terms of evolution (see Sections 3.2.5. & 3.2.6.).

A second amino acid substitution is located at position 92. Here, while in candidates' Escherichia1 sequence a serine is detected, for Escherichia3 an arginine is incorporated at this position (see Figure S3). The last-mentioned variant seems more likely to code a functional protein as indicated by the sequences' evolution. A broader spectrum of homologues detected within the *Enterobacteriaceae* family as well as nearly significant values for a non-random creation of the overlap support this speculation. However, no selection was detectable for this sequence (see Section 3.2.5.). Nevertheless, the characteristics of the substituted amino acid might explain the potentially assumed functionality. Arginine is often detected within the active centre of a protein or its binding site (Barnes & Gray, 2003; Cotton, la Cour, Hazen, & Legg, 1977), whereas serine is known for its high mutation rate (Creixell, Schoof, Tan, & Linding, 2012). Therefore, the probability of a more stable protein is assumed solely based on the amino acid incorporated. Additionally, compared to serine arginine is positively charged (Barnes & Gray, 2003) which potentially could also contribute to beneficial conformation changes in the proteins' secondary structure. Yet, it should be stated again that the actual functionality of the eORF detected still requires verification. The potential involvement of the amino acid substitutions in sequence improvement discussed here is therefore speculative.

Again, it should be stated that the overlap discussed here was identified within a transposase gene. As they naturally can be located multiple times within a genome (Sekine & Ohtsubo, 1989) the identification of the overlap although in varying sequences is not surprising. Based on the transposases' own capacity to change its location within the genome frequently (Vigil-Stenman, Ininbergs, Bergman, & Ekman, 2017) the location of the overlap requires precise analysis. An overlap located at a transposons' edge needs more critical evaluation as the integration of the gene at a different location is accompanied by changing nucleotide sequence at the integrated positions. However, an embedded overlap should not be affected by the surrounding location of the integration. Therefore, this overlap relation should still be considered interesting. Nevertheless, an ideal overlap in terms of interpretation would include a potential relation between mother gene and overlap where one might be regulatory involved in the transcription of the other or would counteract to its function.

Based on results obtained from these first comparative analyses performed, the most promising candidates for further experimental descriptive analysis would be eORFs Escherichia1, Escherichia3 (Escherichia4 is identical), Pseudomonas1 and Salmonella1. A first bioinformatic evaluation of RIBO-Seq data available concludes that embedded overlapping genes can be found distributed across a

variety of prokaryotes and even two archaeal species. Faced by *in silico* verification limitations these analyses are considered as a first step to narrow down the eORF predictions made. Of especial interest was to potentially determine functionality based on eORF sequence evolution. Nonetheless, experimental verification is indispensable for the informatical based recommendations. It was shown that the tools tested should not be considered as stand-alone indicators as no eORF was estimated significant in each approach individually. However, their combined results are useful in the verification of identified candidates.

To conclude this chapter, the detection of overlapping genes in multiple prokaryotic species, in general, is possible. Interpretation and selection of eORFs of interest are based on several performed analysis focussing on sequences evolution. In the last chapter a performed RIBO-Seq experiment will be discussed and how it is contributing to the verification of potential OLGs.

4.3 RIBO-Seq of *B. thetaiotaomicron*

4.3.1 Mapping unmapped reads of *B. thetaiotaomicron*

Right from the start samples showed RNA degradation after RNA extraction (see Figure 17). Degradation is caused by ubiquitous RNases whose activity is inhibited at very low temperatures (Fabre, Colotte, Luis, Tuffet, & Bonnet, 2014; Seelenfreund et al., 2014). Therefore, experimental processes with a focus on RNA always take place on ice. Experimental changes during homogenisation such as using even more liquid nitrogen to prevent samples from thawing under any circumstances did not result in intact RNA. Additional changes in buffer composition to exclude divalent ions, which support increased RNases activity (Hsieh et al., 2010; Thompson, Zong, & Mackie, 2015) did not prevent RNA degradation. Thus, RNA degradation might even start before homogenisation, already during cell harvest. The assumption, that RNA degradation is inhibited as long as the cells, in this case, the bacteria, are intact would not explain potential degradation even before homogenisation. But as a matter of fact, even without cell lysis degradation is possible as RNases are present in prokaryotes as they are involved in various mechanisms (Deutscher, 2015).

The RIBO-Seq experiment of *B. thetaiotaomicron* was performed not only to obtain information about its translatome and potential overlapping genes but also to compare these finding, especially eORFs with mass spectrometry data to potentially verify predicted eORFs. In mass spectrometry proteins that are present in the sample can be detected, if their sequence is available for data comparison purposes. Thus, the predicted eORFs would be used as part of a database for mass spectrometry evaluation. To ensure no translatome changes due to different growth conditions, bacterial cell culture was provided from the lab that will perform mass spectrometry analysis. One major problem might already be the cell harvest step special for RIBO-Seq experiments. Here, it is crucial not only to stall the translation to obtain a snapshot of the ongoing translatome at the point of harvest but also to stabilize RNA necessary

for sequencing. As RNA is less stable than DNA or even proteins (S. Wang & Kool, 1995) specific precautions should be taken.

One option is the use of RNAlater, a non-toxic reagent used to stabilize RNA. Its water-based characteristics allow permeation into still intact cells, therefore preventing intracellular RNA degradation (Passow et al., 2019). Besides, another advantage is its aid in simplified sample handling as it keeps RNA stable even without permanent cooling (Auer et al., 2014; Passow et al., 2019). Unfortunately, this solution is not appropriate for RIBO-Seq experiments, as it has been shown to alter the expression status in RNA-Seq experiments (Passow et al., 2019), hence is believed to also change the translatome impeding the adequate detection of potential additional ORFs. As mentioned, either the application of e.g., chloramphenicol or tetracycline can result in ribosomal stalling, as well as flash freezing the cells in liquid nitrogen (Glaub et al., 2020; Ingolia, 2016; Mohammad et al., 2019). The rapid cooling of the sample with either direct processing or storage at -80°C is also considered to prevent RNA degradation (Passow et al., 2019).

A miscommunication affecting cell harvest might be the causative factor for RNA degradation in this RIBO-Seq experiment. As for mass spectrometry experiments, no additional translation inhibitory or RNA stabilizing precautions are necessary cell culture was harvested without any addition. This includes general centrifugation of cell culture from which the cell pellet is obtained for further analysis as in mass spectrometry only amino acids present in the sample are of interest which are more stable than the RNA they were constructed of. Obtained cell pellets were frozen and stored at -80°C which normally would prevent RNA degradation but here, this mechanism was already started by insufficient cooling during cell harvest. Nevertheless, even if RNA degradation occurs, a RIBO-Seq experiment could still be successful, as mostly due to its high abundance in each cell rRNA is more prone to be affected by RNases. Therefore, if rRNA structures potentially protecting mRNA of interest are broken down mRNA might get available for degradation. Nevertheless, proportions of secured mRNA can still be subjected to RIBO-Seq analysis. Hence, even if RNA degradation was detected the experiment was processed and a special focus was on rRNA depletion. As the RNA degradation might already result in less mRNA available for sequencing successful rRNA depletion was crucial to minimize its abundance. Therefore, available sequencing capacity would cover almost exclusively mRNA of interest.

4.3.2 Sequencing results evaluation

After sequencing was performed for both RIBO-Seq and RNA-Seq samples, they were subjected to subsequent evaluation according to the pipeline used for analysis in the above-mentioned chapters. A comparison of raw read amount and quantity after alignment with additional categorization according to RNA type (mRNA, rRNA or tRNA) only a fraction of reads (≤ 1%) was left. To determine if reads were not aligned to the reference genome or were identified as belonging to excluded RNA types results

from FastQ Screen (see Figure 19) were evaluated. Here, the output showed that nearly ≤ 95% of reads available after alignment to the reference were categorized as unmappable reads. These results would implicate that either a wrong reference genome was chosen for alignment or potential contamination during experimental proceedings.

The appropriate genome chosen for alignment was checked and was correctly from the beginning. To detect if there was a contamination in the culture, unmappable reads were extracted and subjected to BLAST analysis potentially identifying contamination. Nevertheless, most reads were mapped to partial 16S rRNA sequences within genera *Bacteroides, Escherichia* or *Enterococcus* all belonging to the gut flora (Rinninella et al., 2019). As the experiment started with a pure culture, it is highly unlikely that two other bacterial genera were brought into the samples analysed as all necessary purity standards were applied before sample handling. Additionally, if another bacterium was present in the sample, exact species hits during BLAST analysis were expected, not only hits to the genus.

Reads categorized as unmappable to the *B. thetaiotaomicron* genome still were mapped to the genus implying a high sequence similarity. Potentially RNA degradation caused the phenomenon of nearly only unmappable reads. For alignment, a very high sensitive approach was chosen with a seed length of 17 nucleotides. Due to degradation and consequently shorter fragments for sequencing, this might cause insufficient alignment resulting in nearly all reads categorized as unmappable. Nevertheless, even if reds were considered aligned properly to the reference genome nearly all mapped to partial 16S rRNA fragments according to the BLAST analysis. Hence, they would still not contribute to the detection of potential embedded overlapping genes in *B. thetaiotaomicron.*

The number of reads corresponding to 16S rRNA fragments is surprising as rRNA depletion was performed during sequencing preparation. siTOOLs Pan-Prokaryotic Kit was used as it is not genome-specific claiming sufficient rRNA depletion of up to 90 % in general (according to manufacturer). Additionally, here not only 16S and 23S are targeted as common for most depletion kits but also 5S rRNA (according to manufacturer). Nevertheless, depletion was not successful for any of the four samples treated. As depletion was performed two times for two samples simultaneously a handling mistake can be excluded for the lack of depletion efficiency. Additionally, probes from rRNA depletion could be potentially identified due to a small proportion of unmappable reads. These mapped eukaryotic genomes and according to the manufacturer form the basis of rRNA depletion probes. The DNA digestion performed after rRNA depletion generally should ensure minimising the number of unused probes left. Therefore, depletion insufficiency cannot be explained by experimental mishandling during rRNA depletion. Potentially the ongoing RNA degradation within the samples is responsible for insufficient rRNA depletion. The hybridization between fragmented rRNA and added DNA probes necessary for the extraction of targeted rRNA fragments might be ineffective. Indeed, the incompatibility with fragmented RNA has been shown for the used kit explaining the lack of performed rRNA depletion (Huang, Sheth, Kaufman, & Wang, 2020).

Interestingly, when compared to the evaluation of the available *B. thetaiotaomicron* RIBO-Seq samples a similar result for insufficient depletion was detected.

4.3.3 Analysis of publicly available *B. thetaiotaomicron* RIBO-Seq data

The released dataset for *B. thetaiotaomicron* was analysed according to the mentioned processing pipeline. Nevertheless, the obtained data is just as inadequate for the detection of eORFs in the used species as the own generated data. Even though the number of unmappable reads in the two evaluated samples from Sberro (Sberro et al., 2019) is much lower than from the own data (7 % compared to ≤ 95 %), nearly 89 % of their mappable reads are aligned to rRNA (see Section 3.3.3.). This might implicate almost only rRNA fragments present in these samples before sequencing as well even though rRNA depletion was also performed. In fact, here the RiboZero Removal Kit from Illumina was used with a slight adaptation of the protocol only using half the amount of input recommended (Sberro et al., 2019). The decreased amount used for depletion might have caused the insufficient rRNA depletion. From the obtained results, it is not possible to detect, whether the rRNA depletion was insufficient due to decreased input material or potential other factors such as degradation as well. As the used Illumina kit for rRNA depletion is discontinued by the manufacturer, it was not available for the self-performed RIBO-Seq experiment. Thus, it is not possible to test whether the recommended amount of input would have beneficial effects on rRNA depletion efficiency for this *Bacteroides* species.

Finally, for the experimental part of this thesis, there is to say that it, unfortunately, did not contribute to the detection of potential eORFs in *B. thetaiotaomicron.* Nevertheless, the performance of the experimental proceedings with the obtained results once more showed the importance of intact RNA as input material for RIBO-Seq experiments. Additionally, it demonstrates the significance of rRNA depletion to minimise the amount present in analysed samples to reduce the read coverage lost to this type of RNA. Furthermore, the necessity of the appropriate rRNA depletion kits is shown, as the chosen one does not seem to be capable of depletion of fragmented rRNA. However, the RiboZero kit having the best reputation for depletion performance was no longer available, therefore could not be chosen for this experiment. But even if available, the detection of potential eORFs in *B. thetaiotaomicron* is questionable as even a sufficient depletion does still not alter the RNA degradation in general.

After interpretation and discussion of all results obtained in this thesis, the following conclusion will complete this thesis.

5. Conclusion

The general aim of this thesis was to detect potential improvements for prokaryotic RIBO-Seq experiments based on the comparison of already available data. With identified experimental changes a RIBO-Seq experiment was performed which should contribute to the acceptance of overlapping genes not only being present in the genome but being of actual functionality due to RIBO-Seq signals. Furthermore, the detection of antisense OLGs within a variety of prokaryotic species widely distributed across the phylogenetic tree should once and for all demonstrate the existence of OLGs in bacteria.

First, crucial changes for prokaryotic-focused RIBO-Seq experiments were identified which led to recommendations for subsequent studies. A sufficient coverage necessary was identified after analysing the prediction success of annotated genes. In comparison to published RNA-Seq guideline values, higher coverage should be achieved for RIBO-Seq experiments. Not only should more than one read be used to verify a potential ORF, hence requiring an increased amount, but also prediction tools require specific read distribution patterns to identify ORFs as being of interest. This requires multiple reads per ORF of interest. Therefore, a recommendation of at least 20 million reads without rRNA and tRNA mapped reads is proposed. Further, the importance of appropriate size selection during RIBO-Seq experiments was shown. Based on a read length comparison for distinct RNA types present, a general selection of mRNA fragments is achieved when aimed for a size range roughly between 24 to 27 nt. Of course, an exact gel excision is difficult due to sometimes blurry gel band borders. Hence, with a narrower selection range even if excision is inaccurate to some extent the addition of unwanted fragment length is kept lower compared to a broader arrange that includes unwanted fragment sizes. However, size selection must always be adapted to the scientific question of interest. Within this thesis, it could be shown that reads mapped to the 5'-UTR upstream region of genes tend to be longer than those covering protein-coding regions. Thus, if a strived for analysis should focus on e.g., the SD sequence located within the here analysed 5'-UTR region, a size selection range adaption is crucial to cover the longer fragments needed. Lastly, the application of ribosomal stalling inducing additive chloramphenicol was found especially useful for start site detection of weekly expressed genes. This is particularly of interest considering the potential low-level expression of potential OLGs. However, despite their expression level, their detection could be improved by the application of chloramphenicol. Unfortunately, these established recommendations, even if to some extent applied, did not result in a successful performed RIBO-Seq experiment for *B. thetaiotaomicron.*

Here, the aim was to potentially identify so far unknown overlapping genes for this species. The failure of the experimental proceedings is attributed to the RNA degradation that presumably is caused due to insufficient enzyme activity suppression during harvest. Subsequent procedures of the RIBO-Seq protocol with additional changes to prevent further degradation were performed. Nevertheless, results showed alignment difficulties mostly caused by shortened reads due to RNA degradation. A published RIBO-Seq experiment for the chosen species could not be used for evaluation of potential OLGs either.

Conclusion

Analysis performed on the available samples showed nearly solely coverage of rRNA corresponding fragments. Based on the reported sample preparation it is speculated that decreased rRNA depletion volume is responsible for the remaining amount of rRNA within the samples. Thus, so far, the detection of OLGs within *B. thetaiotaomicron* remains concealed. However, the general existence of OLGs within a variety of species could be shown within this thesis.

In general, the detection of eORFs throughout the phylogenetic tree is possible although lacks clear influence of the genomic features' genome size or GC-content on prediction efficiency. Furthermore, the predominant length for eORFs detected as actually translated spans from 100 - 200 nt, even though occasional eORFs can be found with length up to 1,300 nt however only in high GC-content genomes. Detected eORFs are most likely found in sas12 relation to the mother gene, which is in accordance to be beneficial for maintaining their functionality. However, it is still believed that the creation of a new eORF is more likely to be in favour in frame relation sas11. 43 eORFs within 14 different species were found re-occurring in several samples analysed allowing to state that they are of interest for further analysis. Due to their multiple detection success, it is assumed that these eORFs are genuinely translated and are not just of spurious origin. Age determination performed on those mentioned revealed that ten eORFs can be categorized as 'old' due to the number of tblastn hits within their families. The sequences' distributions in various bacteria can be considered as an indication of functionality as well. Here, it is assumed that only beneficial sequences are maintained in a genome due to purifying selection. Thus, all sequences were submitted to OLGenie to detect potential selection on the nucleotide order. Only for two eORF sequences selection was statistical significantly detectable. Unfortunately, those two did not correspond to an eORF categorized as 'old', which was desired to have several indicators for functionality. However, for six additional eORF sequences' nearly significant values were obtained within selection detection analysis. Two of those were corresponding to eORFs categorized as 'old' in the previous analysis and therefore of especial interest for the last analysis performed. The last analysis performed focused on eORF lengths' significance based on the simulation of random ORFs. Here, three sequences were identified as nearly significant while three also corresponded to sequences of interest based on age determination. Here, both candidates had identical sequences whereas within the third two amino acid exchanges were identified. Nevertheless, all of them were located complementary to the same mother gene. Different similarity values or length analysis results are presumably caused by these amino acid variants incorporated. Thus, if the actual functionality of the eORF would be identified, an analysis of interest could be to analyse the potential influence of the amino acid exchanges in regard to the enzymes' activity.

All in all, a comparison of the analyses performed shows that none of the eORFs was validated in all three approaches. Thus, it is essential to not rely on one but multiple analyses to verify sequences of potential interest. These tools and methods assist in narrowing down the number of eORFs predicted to choose those used for first experimental verification. Candidates Escherichia1, Escherichia3, Escherichia4, Pseudomonas1 and Salmonella1 could be chosen for following experimental verification

based on their results in at least two of the three analyses performed. Here, performance analysis of those identified in any lab experiments such as competitive growth, growth rates in various environmental or stress conditions or the detection of applied antibiotics' minimal inhibitory concentration would be of interest. Additionally, a detection analysis of the corresponding protein with mass spectrometry would verify the translation status of the eORF.

Bioinformatic evaluation of RIBO-Seq data regarding eORF detection is only a first step. Applied methods such as BLAST, OLGenie and Frameshift are used to narrow down predictions to potential eORFs of interest. Nevertheless, the necessity of comparing various tools available is shown as their stand-alone results are rarely corresponding. Thus, sequences chosen for further analysis solely based on results from one analysis could be misleading. Based on the combination of first descriptive analyses performed eORFs can be selected for experimental evaluation to verify or neglect hypothesis from informatically obtained results. This first analysis of eORF detection throughout the phylogenetic tree gives a first glimpse of what still remains uncovered but now is available for experimental verification to further analyse the actual function of overlapping genes.

Literature

Al-Bassam, M. M., Kim, J. N., Zaramela, L. S., Kellman, B. P., Zuniga, C., Wozniak, J. M., et al. (2018). Optimization of carbon and energy utilization through differential translational efficiency. *Nat Commun, 9*(1), 4474. doi:10.1038/s41467-018-06993-6

Amin, M. R., Yurovsky, A., Chen, Y., Skiena, S., & Futcher, B. (2018). Re-annotation of 12,495 prokaryotic 16S rRNA 3' ends and analysis of Shine-Dalgarno and anti-Shine-Dalgarno sequences. *PLoS One, 13*(8), e0202767. doi:10.1371/journal.pone.0202767

Ardui, S., Ameur, A., Vermeesch, J. R., & Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res, 46*(5), 2159-2168. doi:10.1093/nar/gky066

Artymiuk, P. J., Ceska, T. A., Suck, D., & Sayers, J. R. (1997). Prokaryotic 5′-3′ exonucleases share a common core structure with gamma-delta resolvase. *Nucleic Acid Res, 25*(21), 5.

Auer, H., Mobley, J. A., Ayers, L. W., Bowen, J., Chuaqui, R. F., Johnson, L. A., et al. (2014). The effects of frozen tissue storage conditions on the integrity of RNA and protein. *Biotech Histochem, 89*(7), 518-528. doi:10.3109/10520295.2014.904927

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., et al. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol, 2*, 2006 0008. doi:10.1038/msb4100050

Baek, J., Lee, J., Yoon, K., & Lee, H. (2017). Identification of Unannotated Small Genes in Salmonella. *G3 (Bethesda), 7*(3), 983-989. doi:10.1534/g3.116.036939

Baez, W. D., Roy, B., McNutt, Z. A., Shatoff, E. A., Chen, S., Bundschuh, R., & Fredrick, K. (2019). Global analysis of protein synthesis in Flavobacterium johnsoniae reveals the use of Kozak-like sequences in diverse bacteria. *Nucleic Acids Res, 47*(20), 10477-10488. doi:10.1093/nar/gkz855

Balakrishnan, R., Oman, K., Shoji, S., Bundschuh, R., & Fredrick, K. (2014). The conserved GTPase LepA contributes mainly to translation initiation in Escherichia coli. *Nucleic Acid Res, 42*(21), 13.

Barnes, M. R., & Gray, I. C. (2003). Bioinformatics for Geneticists. doi:DOI:10.1002/0470867302

Bartholomäus, A., Fedyunin, I., Feist, P., Sin, C., Zhang, G., Valleriani, A., & Ignatova, Z. (2016). Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philos Trans A Math Phys Eng Sci, 374*(2063).

Basrai, M. A., Hieter, P., & Boeke, J. D. (1997). Small open reading frames: beautiful needles in the haystack. *Genome Res, 7*(8), 768-771. doi:10.1101/gr.7.8.768

Basu, A., & Yap, M. N. (2016). Ribosome hibernation factor promotes Staphylococcal survival and differentially represses translation. *Nucleic Acids Res, 44*(10), 4881-4893. doi:10.1093/nar/gkw180

Becker, N., Kunath, J., Loh, G., & Blaut, M. (2011). Human intestinal microbiota: characterization of a simplified and stable gnotobiotic rat model. *Gut Microbes, 2*(1), 25-33. doi:10.4161/gmic.2.1.14651

Belinky, F., Babenko, V. N., Rogozin, I. B., & Koonin, E. V. (2018). Purifying and positive selection in the evolution of stop codons. *Sci Rep, 8*(1), 9260. doi:10.1038/s41598-018-27570-3

Belinky, F., Rogozin, I. B., & Koonin, E. V. (2017). Selection on start codons in prokaryotes and potential compensatory nucleotide substitutions. *Sci Rep, 7*(1), 12422. doi:10.1038/s41598-017-12619-6

Belshaw, R., Pybus, O. G., & Rambaut, A. (2007). The evolution of genome compression and genomic novelty in RNA viruses. *Genome Res, 17*(10), 1496-1504. doi:10.1101/gr.6305707

Bernad, A., Blanco, L., Lkaro, J. Y., Martin, G., & Salas, M. (1989). A Consewed 3'->5' Exonuclease Active Site in Prokaryotic and Eukaryotic DNA Polymerases *Cell, 59*.

Bhavsar, R. B., Makley, L. N., & Tsonis, P. A. (2010). The other lives of ribosomal proteins. *Human Genomics, 4*(5), 18.

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

Blazej, P., Wnetrzak, M., Mackiewicz, D., & Mackiewicz, P. (2018). Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm. *PLoS One, 13*(8), e0201715. doi:10.1371/journal.pone.0201715

Bofkin, L., & Goldman, N. (2007). Variation in evolutionary processes at different codon positions. *Mol Biol Evol, 24*(2), 513-521. doi:10.1093/molbev/msl178

Bohlin, J., Eldholm, V., Pettersson, J. H., Brynildsrud, O., & Snipen, L. (2017). The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics, 18*(1), 151. doi:10.1186/s12864-017-3543-7

Boi, S., Solda, G., & Tenchini, M. (2004). Shedding Light on the Dark Side of the Genome: Overlapping Genes in Higher Eukaryotes. *Current Genomics, 5*(6), 509-524. doi:10.2174/1389202043349020

Brandes, N., & Linial, M. (2016). Gene overlapping and size constraints in the viral world. *Biol Direct, 11*, 26. doi:10.1186/s13062-016-0128-3

Burkhardt, D. H., Rouskin, S., Zhang, Y., Li, G. W., Weissman, J. S., & Gross, C. A. (2017). Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. *eLife, 6*.

Buskirk, A. R., & Green, R. (2017). Ribosome pausing, arrest and rescue in bacteria and eukaryotes. *Philos Trans R Soc Lond B Biol Sci, 372*(1716). doi:10.1098/rstb.2016.0183

Calviello, L., & Ohler, U. (2017). Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. *Trends Genet, 33*(10), 728-744. doi:10.1016/j.tig.2017.08.003

Cassan, E., Arigon-Chifolleau, A. M., Mesnard, J. M., Gross, A., & 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*(41), 11537-11542. doi:10.1073/pnas.1605739113

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*(17), i884-i890. doi:10.1093/bioinformatics/bty560

Chen, Y.-X., Xu, Z.-Y., Ge, X., Hong, J.-Y., Sanyal, S., Lu, Z. J., & Javid, B. (2020). Selective translation by alternative bacterial ribosomes. *PNAS, 117*(32), 9. doi:10.1101/605931

Chen, Z., & Duan, X. (2011). Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. *Methods Mol Biol, 733*, 93-103. doi:10.1007/978-1-61779-089-8_7

Chirico, N., Vianelli, A., & Belshaw, R. (2010). Why genes overlap in viruses. *Proc Biol Sci, 277*(1701), 3809-3817. doi:10.1098/rspb.2010.1052

Clauwaert, J., Menschaert, G., & Waegeman, W. (2019). DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. *Nucleic Acids Research*. doi:10.1093/nar/gkz061

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods, 5*(7), 613-619. doi:10.1038/nmeth.1223

Colosimo, D. A., Kohn, J. A., Luo, P. M., Piscotta, F. J., Han, S. M., Pickard, A. J., et al. (2019). Mapping Interactions of Microbial Metabolites with Human G-Protein-Coupled Receptors. *Cell Host Microbe, 26*(2), 273-282 e277. doi:10.1016/j.chom.2019.07.002

Costa, V., Angelini, C., De Feis, I., & Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotechnol, 2010*, 853916. doi:10.1155/2010/853916

Cotton, F. A., la Cour, T., Hazen, E. E., Jr., & Legg, M. J. (1977). The role of arginine residues at enzyme active sites. The interaction between guanidinium ions and p-nitro-phenyl phosphate and its effect on the rate of hydrolysis of the ester. *Biochim Biophys Acta, 481*(1), 1-5. doi:10.1016/0005-2744(77)90131-0

Creixell, P., Schoof, E. M., Tan, C. S. H., & Linding, R. (2012). Mutational properties of amino acid residues: implications for evolvability of phosphorylatable residues. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1602), 2584-2593. doi:10.1098/rstb.2012.0076

Cvijovic, I., Good, B. H., & Desai, M. M. (2018). The Effect of Strong Purifying Selection on Genetic Diversity. *Genetics, 209*(4), 1235-1278. doi:10.1534/genetics.118.301058

Davis, A. R., Gohara, D. W., & Yap, M. N. (2014). Sequence selectivity of macrolide-induced translational attenuation. *Proc Natl Acad Sci U S A, 111*(43), 15379-15384. doi:10.1073/pnas.1410356111

Deutscher, M. P. (2015). How bacterial cells keep ribonucleases under control. *FEMS Microbiol Rev, 39*(3), 350-361. doi:10.1093/femsre/fuv012

Dingwall, C., Lomonossoff, G. P., & Laskey, R. A. (1981). High sequence specificity of micrococcal nuclease. *Nucleic Acid Res, 9*(12).

Domazet-Loso, T., Brajkovic, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet, 23*(11), 533-539. doi:10.1016/j.tig.2007.08.014

Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R., & Weissman, J. S. (2013). Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife, 2*, e01179. doi:10.7554/eLife.01179

Elgamal, S., Katz, A., Hersch, S. J., Newsom, D., White, P., Navarre, W. W., & Ibba, M. (2014). EF-P Dependent Pauses Integrate Proximal and Distal Signals during Translation. *PLoS Genet*., *10*(8), e1004553

Errington, J. (2013). L-form bacteria, cell walls and the origins of life. *Open Biol, 3*(1), 120143. doi:10.1098/rsob.120143

Esberg, A. (2007). Functional aspects of wobble uridine modifications in yeast tRNA. *Biology*

Ettwiller, L., Buswell, J., Yigit, E., & Schildkraut, I. (2016). A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. *BMC Genomics, 17*, 199. doi:10.1186/s12864-016-2539-z

Eyre-Walker, A. (1995). The Distance between Escherichia coli Genes Is Related to Gene Expression Levels. *J. Bacteriol, 177*(18).

Fabre, A. L., Colotte, M., Luis, A., Tuffet, S., & Bonnet, J. (2014). An efficient method for long-term room temperature storage of RNA. *Eur J Hum Genet, 22*(3), 379-385. doi:10.1038/ejhg.2013.145

Fellner, L., Bechtel, N., Witting, M. A., Simon, S., Schmitt-Kopplin, P., Keim, D., et al. (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*(1), 57-64. doi:10.1111/1574-6968.12288

Fellner, L., Simon, S., Scherling, C., Witting, M., Schober, S., Polte, C., et al. (2015). Evidence for the recent origin of a bacterial protein-coding, overlapping orphan gene by evolutionary overprinting. *BMC Evol Biol, 15*, 283. doi:10.1186/s12862-015-0558-z

Fernandes, J. D., Faust, T. B., Strauli, N. B., Smith, C., Crosby, D. C., Nakamura, R. L., et al. (2016). Functional Segregation of Overlapping Genes in HIV. *Cell, 167*(7), 1762-1773 e1712. doi:10.1016/j.cell.2016.11.031

Gelsinger, D. R., Dallon, E., Reddy, R., Mohammad, F., Buskirk, A. R., & 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*(10), 5201-5216. doi:10.1093/nar/gkaa304

Gerashchenko, M. V., & Gladyshev, V. N. (2017). Ribonuclease selection for ribosome profiling. *Nucleic Acids Res, 45*(2), e6. doi:10.1093/nar/gkw822

Giess, A., Jonckheere, V., Ndah, E., Chyzynska, K., Van Damme, P., & Valen, E. (2017). Ribosome signatures aid bacterial translation initiation site identification. *BMC Biol, 15*(1), 76. doi:10.1186/s12915-017-0416-0

Glaub, A., Huptas, C., Neuhaus, K., & Ardern, Z. (2020). Recommendations for bacterial ribosome profiling experiments based on bioinformatic evaluation of published data. *J Biol Chem, 295*(27), 12. doi:10.1074/jbc.RA119.012161

Gonzalez, D. L., Giannerini, S., & Rosa, R. (2019). On the origin of degeneracy in the genetic code. *Interface Focus, 9*(6), 20190038. doi:10.1098/rsfs.2019.0038

Grady, S. L., Malfatti, S. A., Gunasekera, T. S., Dalley, B. K., Lyman, M. G., Striebich, R. C., et al. (2017). A comprehensive multi-omics approach uncovers adaptations for growth and survival of Pseudomonas aeruginosa on n-alkanes. *BMC Genomics, 18*(1), 334. doi:10.1186/s12864-017-3708-4

Grenga, L., Chandra, G., Saalbach, G., Galmozzi, C. V., Kramer, G., & Malone, J. G. (2017). Analyzing the Complex Regulatory Landscape of Hfq - an Integrative, Multi-Omics Approach. *Front Microbiol, 8*, 1784. doi:10.3389/fmicb.2017.01784

Grenier, F., Matteau, D., Baby, V., & Rodrigue, S. (2014). Complete Genome Sequence of Escherichia coli BW25113. *Genome Announc, 2*(5). doi:10.1128/genomeA.01038-14

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., & Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell, 154*(1), 240-251. doi:10.1016/j.cell.2013.06.009

Haas, B. J., Chin, M., Nusbaum, C., Birren, B. W., & Livny, J. (2012). How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics, 2012*(13), 11.

Hagemann-Jensen, M., Abdullayev, I., Sandberg, R., & Faridani, O. R. (2018). Small-seq for single-cell small-RNA sequencing. *Nat Protoc, 13*(10), 2407-2424. doi:10.1038/s41596-018-0049-y

Harris, K. A., & Breaker, R. R. (2018). Large Noncoding RNAs in Bacteria. *Microbiol Spectr, 6*(4). doi:10.1128/microbiolspec.RWR-0005-2017

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishi, K., Yokoyama, K., et al. (2001). Complete Genome Sequence of Enterohemorrhagic Escherichia coli O157:H7 and Genomic Comparison with a Laboratory Strain K-12.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics, 107*(1), 1-8. doi:10.1016/j.ygeno.2015.11.003

Hecht, A., Glasgow, J., Jaschke, P. R., Bawazer, L. A., Munson, M. S., Cochran, J. R., et al. (2017). Measurements of translation initiation from all 64 codons in E. coli. *Nucleic Acids Res, 45*(7), 3615-3626. doi:10.1093/nar/gkx070

Henikoff, S., Keene, M. A., Fechtel, K., & Fristrom, J. W. (1986). Gene within a gene: nested Drosophila genes encode unrelated proteins on opposite DNA strands. *Cell, 44*(1), 33-42. doi:10.1016/0092-8674(86)90482-4

Hickman, M., & Cairns, J. (2003). The centenary of the one-gene one-enzyme hypothesis. *Genetics, 163*(3), 839-841.

Hildebrand, F., Meyer, A., & Eyre-Walker, A. (2010). Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet, 6*(9), e1001107. doi:10.1371/journal.pgen.1001107

Hooper, L. V., Wong, M. H., Thelin, A., Hansson, L., Falk, P. G., & Gordon, J. I. (2001). Molecular analysis of commensal host-microbial relations in the intestine. *Science, 291*(5505), 881-884. doi:10.1126/science.291.5505.881

Hsieh, J., Koutmou, K. S., Rueda, D., Koutmos, M., Walter, N. G., & Fierke, C. A. (2010). A divalent cation stabilizes the active conformation of the B. subtilis RNase P x pre-tRNA complex: a role for an inner-sphere metal ion in RNase P. *J Mol Biol, 400*(1), 38-51. doi:10.1016/j.jmb.2010.04.050

Huang, Y., Sheth, R. U., Kaufman, A., & Wang, H. H. (2020). Scalable and cost-effective ribonuclease-based rRNA depletion for transcriptomics. *Nucleic Acids Res, 48*(4), e20. doi:10.1093/nar/gkz1169

Hücker, S. M., Ardern, Z., Goldberg, T., Schafferhans, A., Bernhofer, M., Vestergaard, G., et al. (2017). Discovery of numerous novel small genes in the intergenic regions of the Escherichia coli O157:H7 Sakai genome. *PLoS One, 12*(9), e0184119. doi:10.1371/journal.pone.0184119

Hücker, S. M., Simon, S., Scherer, S., & Neuhaus, K. (2017). Transcriptional and translational regulation by RNA thermometers, riboswitches and the sRNA DsrA in Escherichia coli O157:H7 Sakai under combined cold and osmotic stress adaptation. *FEMS Microbiol Lett, 364*(2). doi:10.1093/femsle/fnw262

Hücker, S. M., Vanderhaeghen, S., Abellan-Schneyder, I., Scherer, S., & Neuhaus, K. (2018). The Novel Anaerobiosis-Responsive Overlapping Gene ano Is Overlapping Antisense to the Annotated Gene ECs2385 of Escherichia coli O157:H7 Sakai. *Front Microbiol, 9*, 931. doi:10.3389/fmicb.2018.00931

Hücker, S. M., Vanderhaeghen, S., Abellan-Schneyder, I., Wecko, R., Simon, S., Scherer, S., & Neuhaus, K. (2018). A novel short L-arginine responsive protein-coding gene (laoB) antiparallel overlapping to a CadC-like transcriptional regulator in Escherichia coli O157:H7 Sakai originated by overprinting. *BMC Evol Biol, 18*(1), 21. doi:10.1186/s12862-018-1134-0

Hwang, J. Y., & Buskirk, A. R. (2017). A ribosome profiling study of mRNA cleavage by the endonuclease RelE. *Nucleic Acids Res, 45*(1), 327-336. doi:10.1093/nar/gkw944

Imdahl, F., Vafadarnejad, E., Homberger, C., Saliba, A. E., & Vogel, J. (2020). Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. Nat Microbiol, 5(10), 1202-1206. doi:10.1038/s41564-020-0774-1

Ina, Y. (1996). Pattern of synonymous and nonsynonymous substitutions: an indicator of mechanisms of molecular evolution *J. Genet., 75*(1), 91-115.

Ingolia, N. T. (2010). Genome-Wide Translational Profiling by Ribosome Footprinting. In *Guide to Yeast Genetics: Functional Genomics, Proteomics, and Other Systems Analysis* (pp. 119-142).

Ingolia, N. T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Genetics, 15*.

Ingolia, N. T. (2016). Ribosome Footprint Profiling of Translation throughout the Genome. *Cell, 165*(1), 22-33. doi:10.1016/j.cell.2016.02.066

Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M., & Weissman, J. S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc, 7*(8), 1534-1550. doi:10.1038/nprot.2012.086

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., et al. (2014). Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Gene. *Cell Rep, 8*(5), 14.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science, 324*(5924), 5.

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell, 147*(4), 789-802. doi:10.1016/j.cell.2011.10.002

Jackson, R., & Standart, N. (2015). The awesome power of ribosome profiling. *RNA, 21*(4), 652-654. doi:10.1261/rna.049908.115

Jen, C. H., Michalopoulos, I., Westhead, D. R., & Meyer, P. (2005). Natural antisense transcripts with coding capacity in Arabidopsis may have a regulatory role that is not linked to double-stranded RNA degradation. *Genome Biol, 6*(6), R51. doi:10.1186/gb-2005-6-6-r51

Jenjaroenpun, P., Wongsurawat, T., Pereira, R., Patumcharoenpol, P., Ussery, D. W., Nielsen, J., & Nookaew, I. (2018). Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of Saccharomyces cerevisiae CEN.PK113-7D. *Nucleic Acids Res, 46*(7), e38. doi:10.1093/nar/gky014

Jeong, Y., Kim, J. N., Kim, M. W., Bucca, G., Cho, S., Yoon, Y. J., et al. (2016). The dynamic transcriptional and translational landscape of the model antibiotic producer Streptomyces coelicolor A3(2). *Nat Commun, 7*, 11605. doi:10.1038/ncomms11605

Jiang, Z., Zhou, X., Li, R., Michal, J. J., Zhang, S., Dodson, M. V., et al. (2015). Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. *Cell Mol Life Sci, 72*(18), 3425-3439. doi:10.1007/s00018-015-1934-y

Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun, 10*(1), 5029. doi:10.1038/s41467-019-13036-1

Jordan, I. K., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res, 12*, 962.

Kannan, K., Kanabar, P., Schryer, D., Florin, T., Oh, E., Bahroos, N., et al. (2014). The general mode of translation inhibition by macrolide antibiotics. *Proc Natl Acad Sci U S A, 111*(45), 5.

Karlsen, J., Asplund-Samuelsson, J., Thomas, Q., Michael, J., & Hudson, E. P. (2018). Ribosome Profiling of Synechocystis Reveals Altered Ribosome Allocation at Carbon Starvation. *mSystems, 3*(5).

Keese, P. K., & Gibbs, A. (1992). Origins of genes: "big bang" or continuous creation? *Proc Natl Acad Sci U S A, 89*(20), 9489-9493. doi:10.1073/pnas.89.20.9489

Kim, W., Hwang, S., Lee, N., Lee, Y., Cho, S., Palsson, B., & Cho, B. K. (2020). Transcriptome and translatome profiles of Streptomyces species in different growth phases. *Sci Data, 7*(1), 138. doi:10.1038/s41597-020-0476-9

Kitahara, K., & Miyazaki, K. (2011). Specific inhibition of bacterial RNase T2 by helix 41 of 16S ribosomal RNA. *Nat Commun, 2*, 549. doi:10.1038/ncomms1553

Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S., Hayashizaki, Y., Group, R. G., & Members, G. S. L. (2003). Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res, 13*(6B), 1324-1334. doi:10.1101/gr.982903

Kneifel, W., & Forsythe, S. (2017). Editorial: The many facets of Escherichia coli: from beneficial bug and genetic workhorse to dangerous menace for plant and creature. *FEMS Microbiol Lett, 364*(10). doi:10.1093/femsle/fnx103

Korkmaz, G., Holm, M., Wiens, T., & Sanyal, S. (2014). Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J Biol Chem, 289*(44), 30334-30342. doi:10.1074/jbc.M114.606632

Krakauer, D. C. (2000). Stability and Evolution of Overlapping Genes. *Evolution, 54*(3). doi:10.1554/0014-3820(2000)054[0731:Saeoog]2.3.Co;2

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics, Chapter 11*, Unit 11 17. doi:10.1002/0471250953.bi1107s32

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Li, G.-W., Burkhardt, D., Gross, C., & Weissman, J. S. (2014). Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell, 157*(3), 624-635. doi:10.1016/j.cell.2014.02.033

Li, G.-W., Oh, E., & Weissmann, J. S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature, 484*(7395), 3.

Lim, J. Y., Yoon, J. W., & Hovde, C. J. (2010). A Brief Overview of Escherichia coli O157:H7 and Its Plasmid O1557. *J Microbiol Biotechnol., 20*(1), 9.

Liu, X., Jiang, H., Gu, Z., & Roberts, J. W. (2013). High-resolution view of bacteriophage lambda gene expression by ribosome profiling. *Proc Natl Acad Sci U S A, 110*(29), 5.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol, 14*(13), 1675-1680. doi:10.1038/nbt1296-1675

Long, M., Betran, E., Thornton, K., & Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nat Rev Genet, 4*(11), 865-875. doi:10.1038/nrg1204

Lopez Garcia de Lomana, A., Kusebauch, U., Raman, A. V., Pan, M., Turkarslan, S., Lorenzetti, A. P. R., et al. (2020). Selective Translation of Low Abundance and Upregulated Transcripts in Halobacterium salinarum. *mSystems, 5*(4). doi:10.1128/mSystems.00329-20

Luo, H., Gao, F., & Lin, Y. (2015). Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. *Sci Rep, 5*, 13210. doi:10.1038/srep13210

Marinov, G. K. (2017). On the design and prospects of direct RNA sequencing. *Brief Funct Genomics, 16*(6), 326-335. doi:10.1093/bfgp/elw043

Marks, J., Kannan, K., Roncase, E. J., Klepacki, D., Kefi, A., Orelle, C., et al. (2016). Context-specific inhibition of translation by ribosomal antibiotics targeting the peptidyl transferase center. *Proc Natl Acad Sci U S A, 113*(43), 5.

Massey, S. E. (2006). A sequential "2-1-3" model of genetic code evolution that explains codon constraints. *J Mol Evol, 62*(6), 809-810. doi:10.1007/s00239-005-0222-0

Melnikov, S., Ben-Shem, A., Garreau de Loubresse, N., Jenner, L., Yusupova, G., & Yusupov, M. (2012). One core, two shells: bacterial and eukaryotic ribosomes. *Nat Struct Mol Biol, 19*(6), 560-567. doi:10.1038/nsmb.2313

Meydan, S., Marks, J., Klepacki, D., Sharma, V., Baranov, P. V., Firth, A. E., et al. (2019). Retapamulin-Assisted Ribosome Profiling Reveals the Alternative Bacterial Proteome. *Mol Cell*. doi:10.1016/j.molcel.2019.02.017

Mimee, M., Tucker, A. C., Voigt, C. A., & Lu, T. K. (2015). Programming a Human Commensal Bacterium, Bacteroides thetaiotaomicron, to Sense and Respond to Stimuli in the Murine Gut Microbiota. *Cell Syst, 1*(1), 62-71. doi:10.1016/j.cels.2015.06.001

Miranda-CasoLuengo, A. A., Staunton, P. M., Dinan, A. M., Lohan, A. J., & 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. doi:10.1186/s12864-016-2868-y

Mohammad, F., Green, R., & Buskirk, A. R. (2019). A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *eLife*.

Mohammad, F., Woolstenhulme, C. J., Green, R., & Buskirk, A. R. (2016). Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Rep, 14*(4), 686-694. doi:10.1016/j.celrep.2015.12.073

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science, 320*(5881), 1344-1349. doi:10.1126/science.1158441

Nakahigashi, K., Takai, Y., Kimura, M., Abe, N., Nakayashiki, T., Shiwa, Y., et al. (2016). Comprehensive identification of translation start sites by tetracycline-inhibited ribosome profiling. *DNA Res, 23*(3), 8.

Nakayama, T., Asai, S., Takahashi, Y., Maekawa, O., & Kasama, Y. (2007). Overlapping of Genes in the Human Genome *Int J Biomed Sci, 3*(1), 5.

Ndah, E., Jonckheere, V., Giess, A., Valen, E., Menschaert, G., & Van Damme, P. (2017). REPARATION: ribosome profiling assisted (re-)annotation of bacterial genomes. *Nucleic Acid Res, 45*(20), e168.

Nelson, C. W., Ardern, Z., Goldberg, T. L., Meng, C., Kuo, C. H., Ludwig, C., et al. (2020). Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic. *eLife, 9*. doi:10.7554/eLife.59633

Nelson, C. W., Ardern, Z., & Wei, X. (2020). OLGenie: Estimating Natural Selection to Predict Functional Overlapping Genes. *Mol Biol Evol*. doi:10.1093/molbev/msaa087

Neuhaus, K., Landstorfer, R., Fellner, L., Simon, S., Schafferhans, A., Goldberg, T., et al. (2016). Translatomics combined with transcriptomics and proteomics reveals novel functional, recently evolved orphan genes in Escherichia coli O157:H7 (EHEC). BMC Genomics, 17, 133. doi:10.1186/s12864-016-2456-1

Neuhaus, K., Landstorfer, R., Simon, S., Schober, S., Wright, P. R., Smith, C., et al. (2017). Differentiation of ncRNAs from small mRNAs in Escherichia coli O157:H7 EDL933 (EHEC) by combined RNAseq and RIBOseq - ryhB encodes the regulatory RNA RyhB and a peptide, RyhP. *BMC Genomics, 18*(1), 216. doi:10.1186/s12864-017-3586-9

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet, 42*(1), 30-35. doi:10.1038/ng.499

Oh, E., Becker, A. H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., et al. (2011). Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell, 147*(6), 1295-1308. doi:10.1016/j.cell.2011.10.044

Olexiouk, V., Van Criekinge, W., & Menschaert, G. (2018). An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res, 46*(D1), D497-D502. doi:10.1093/nar/gkx1130

Oster, G., & Yamamoto, G. (1963). Density Gradient Techniques. *Chemical Reviews, 63*(3), 11.

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet, 12*(2), 87-98. doi:10.1038/nrg2934

Passow, C. N., Kono, T. J. Y., Stahl, B. A., Jaggard, J. B., Keene, A. C., & McGaugh, S. E. (2019). Nonrandom RNAseq gene expression associated with RNAlater and flash freezing storage methods. *Mol Ecol Resour, 19*(2), 456-464. doi:10.1111/1755-0998.12965

Pavesi, A., Magiorkinis, G., & Karlin, D. G. (2013). Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of Deltaretroviruses. *PLoS Comput Biol, 9*(8), e1003162. doi:10.1371/journal.pcbi.1003162

Pavesi, A., Vianelli, A., Chirico, N., Bao, Y., Blinkova, O., Belshaw, R., et al. (2018). Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLoS One, 13*(10), e0202513. doi:10.1371/journal.pone.0202513

Pavlin, M., Qasem, Z., Sameach, H., Gevorkyan-Airapetov, L., Ritacco, I., Ruthstein, S., & Magistrato, A. (2019). Unraveling the Impact of Cysteine-to-Serine Mutations on the Structural and Functional Properties of Cu(I)-Binding Proteins. *Int J Mol Sci, 20*(14). doi:10.3390/ijms20143462

Pedersen, K., Zavialov, A. V., Pavlov, M. Y., Elf, J., Gerdes, K., & Ehrenberg, M. (2003). The bacterial toxin RelE displays codon-specific cleavage of mRNAs in the ribosomal A site. *Cell, 112*(1), 131-140. doi:10.1016/s0092-8674(02)01248-5

Peters, J. E., Thate, T. E., & Craig, N. L. (2003). Definition of the Escherichia coli MC4100 Genome by Use of a DNA Array. *Journal of Bacteriology, 185*(6), 2017-2021. doi:10.1128/jb.185.6.2017-2021.2003

Petrova, O. E., Garcia-Alcalde, F., Zampaloni, C., & Sauer, K. (2017). Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. *Sci Rep, 7*, 41114. doi:10.1038/srep41114

Piovesan, A., Pelleri, M. C., Antonaros, F., Strippoli, P., Caracausi, M., & Vitale, L. (2019). On the length, weight and GC content of the human genome. *BMC Res Notes, 12*(1), 106. doi:10.1186/s13104-019-4137-z

Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet, 12*(1), 32-42. doi:10.1038/nrg2899

Pohl, M., Theissen, G., & Schuster, S. (2012). GC content dependency of open reading frame prediction via stop codon frequencies. *Gene, 511*(2), 441-446. doi:10.1016/j.gene.2012.09.031

Portin, P., & Wilkins, A. (2017). The Evolving Definition of the Term "Gene". *Genetics, 205*(4), 1353-1364. doi:10.1534/genetics.116.196956

Povolotskaya, I. S., Kondrashov, F. A., Ledda, A., & Vlasov, P. K. (2012). Stop codons in bacteria are not selectively equivalent. *Biology Direct, 7*(30).

Price, M. N., Arkin, A. P., & Alm, E. J. (2006). The life-cycle of operons. *PLoS Genet, 2*(6), e96. doi:10.1371/journal.pgen.0020096

Qian, X., Ba, Y., Zhuang, Q., & Zhong, G. (2014). RNA-Seq technology and its application in fish transcriptomics. *OMICS, 18*(2), 98-110. doi:10.1089/omi.2013.0110

Raghavan, R., Groisman, E. A., & Ochman, H. (2011). Genome-wide detection of novel regulatory RNAs in E. coli. *Genome Res, 21*(9), 1487-1497. doi:10.1101/gr.119370.110

Rancurel, C., Khosravi, M., Dunker, A. K., Romero, P. R., & Karlin, D. (2009). Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *J Virol, 83*(20), 10719-10736. doi:10.1128/JVI.00595-09

Rinninella, E., Raoul, P., Cintoni, M., Franceschi, F., Miggiano, G. A. D., Gasbarrini, A., & Mele, M. C. (2019). What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms, 7*(1). doi:10.3390/microorganisms7010014

Roberts, R. J. (1978). Restriction endonucleases: a new role in vivo? . *Nature, 271*.

Rogozin, I. B., Spiridonov, A. N., Sorokin, A. V., Wolf, Y. I., Jordan, I. K., Tatusov, R. L., & Koonin, E. V. (2002). Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet, 18*(5), 228-232. doi:10.1016/s0168-9525(02)02649-5

Sabath, N., Wagner, A., & Karlin, D. (2012). Evolution of viral proteins originated de novo by overprinting. *Mol Biol Evol, 29*(12), 3767-3780. doi:10.1093/molbev/mss179

Saier, M. H., Jr. (2019). Understanding the Genetic Code. *J Bacteriol, 201*(15). doi:10.1128/JB.00091-19

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., et al. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature, 265*(5596), 687-695. doi:10.1038/265687a0

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A, 74*(12), 4.

Sberro, H., Fremin, B. J., Zlitni, S., Edfors, F., Greenfield, N., Snyder, M. P., et al. (2019). Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. *Cell, 178*(5), 1245-1259 e1214. doi:10.1016/j.cell.2019.07.016

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, 270*(5235), 467-470. doi:10.1126/science.270.5235.467

Scherbakov, D. V., & Garber, M. B. (2000). Overlapping Genes in Bacterial and Phage Genomes. *Molecular Biology, 34*(4).

Schlub, T. E., Buchmann, J. P., & Holmes, E. C. (2018). A simple method to detect candidate overlapping genes in viruses using single genome sequences. *Mol Biol Evol*. doi:10.1093/molbev/msy155

Schrader, J. M., Li, G. W., Childers, W. S., Perez, A. M., Weissman, J. S., Shapiro, L., & McAdams, H. H. (2016). Dynamic translation regulation in Caulobacter cell cycle control. *Proc Natl Acad Sci U S A, 113*(44), E6859-E6867. doi:10.1073/pnas.1614795113

Schrader, J. M., Zhou, B., Li, G. W., Lasker, K., Childers, W. S., Williams, B., et al. (2014). The coding and noncoding architecture of the Caulobacter crescentus genome. *PLoS Genet, 10*(7), e1004463. doi:10.1371/journal.pgen.1004463

Seelenfreund, E., Robinson, W. A., Amato, C. M., Tan, A. C., Kim, J., & Robinson, S. E. (2014). Long term storage of dry versus frozen RNA for next generation molecular studies. *PLoS One, 9*(11), e111827. doi:10.1371/journal.pone.0111827

Sekine, Y., & Ohtsubo, E. (1989). Frameshifting is required for production of the transposase encoded by insertion sequence 1. *Proc Natl Acad Sci U S A, 86*(12), 4609-4613. doi:10.1073/pnas.86.12.4609

Shell, S. S., Wang, J., Lapierre, P., Mir, M., Chase, M. R., Pyle, M. M., et al. (2015). Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS Genet, 11*(11), e1005641. doi:10.1371/journal.pgen.1005641

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol, 26*(10), 1135-1145. doi:10.1038/nbt1486

Shendure, J. A., Porreca, G. J., Church, G. M., Gardner, A. F., Hendrickson, C. L., Kieleczawa, J., & Slatko, B. E. (2011). Overview of DNA sequencing strategies. *Curr Protoc Mol Biol, Chapter 7*, Unit7 1. doi:10.1002/0471142727.mb0701s96

Shine, J., & 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*(4), 4.

Shinhara, A., Matsui, M., Hiraoka, K., Nomura, W., Hirano, R., Nakahigashi, K., et al. (2011). Deep sequencing reveals as-yet-undiscovered small RNAs in Escherichia coli *BMC Genomics, 12*(428).

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet, 15*(2), 121-132. doi:10.1038/nrg3642

Smith, C. J., & Marnett, L. J. (1996). Effects of cysteine-to-serine mutations on structural and functional properties of prostaglandin endoperoxide synthase. *Arch Biochem Biophys, 335*(2), 342-350. doi:10.1006/abbi.1996.0515

Song, W., Joo, M., Yeom, J. H., Shin, E., Lee, M., Choi, H. K., et al. (2019). Divergent rRNAs as regulators of gene expression at the ribosome level. *Nat Microbiol, 4*(3), 515-526. doi:10.1038/s41564-018-0341-1

Song, Y., Shin, J., Jin, S., Lee, J. K., Kim, D. R., Kim, S. C., et al. (2018). Genome-scale analysis of syngas fermenting acetogenic bacteria reveals the translational regulation for its autotrophic growth. *BMC Genomics, 19*(1), 837. doi:10.1186/s12864-018-5238-0

Spencer, C. A., Gietz, R. D., & Hodgetts, R. B. (1986). Overlapping transcription units in the dopa decarboxylase region of Drosophila. *Nature, 322*(6076), 279-281. doi:10.1038/322279a0

Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet, 16*(3), 133-145. doi:10.1038/nrg3833

Storz, G., Wolf, Y. I., & Ramamurthi, K. S. (2014). Small proteins can no longer be ignored. *Annu Rev Biochem, 83*, 753-777. doi:10.1146/annurev-biochem-070611-102400

Su, M., Ling, Y., Yu, J., Wu, J., & Xiao, J. (2013). Small proteins: untapped area of potential biological importance. *Front Genet, 4*, 286. doi:10.3389/fgene.2013.00286

Subramaniam, A. R., Deloughery, A., Bradshaw, N., Chen, Y., O'Shea, E., Losick, R., & Chai, Y. (2013). A serine sensor for multicellularity in a bacterium. *eLife, 2*, e01501. doi:10.7554/eLife.01501

Svensson, V., Natarajan, K. N., Ly, L. H., Miragaia, R. J., Labalette, C., Macaulay, I. C., et al. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat Methods, 14*(4), 381-387. doi:10.1038/nmeth.4220

Tan, Y., & Riley, M. A. (1997). Positive selection and recombination: major molecular mechanisms in colicin diversification. *Trends Ecol Evol, 12*(9), 348-351. doi:10.1016/s0169-5347(97)01127-0

Taylor, J. S., & Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet, 38*, 615-643. doi:10.1146/annurev.genet.38.072902.092831

Terryn, N., & Rouze, P. (2000). The sense of naturally transcribed antisense RNAs in plants. *Trends Plant Sci, 5*(9), 394-396. doi:10.1016/s1360-1385(00)01696-4

Thompson, K. J., Zong, J., & Mackie, G. A. (2015). Altering the divalent metal ion preference of RNase E. *J Bacteriol, 197*(3), 477-482. doi:10.1128/JB.02372-14

Trotta, E. (2016). Selective forces and mutational biases drive stop codon usage in the human genome: a comparison with sense codon usage. *BMC Genomics, 17*, 366. doi:10.1186/s12864-016-2692-4

van de Waterbeemd, M., Fort, K. L., Boll, D., Reinhardt-Szyba, M., Routh, A., Makarov, A., & Heck, A. J. (2017). High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. *Nat Methods, 14*(3), 283-286. doi:10.1038/nmeth.4147

Vanderhaeghen, S., Zehentner, B., Scherer, S., Neuhaus, K., & Ardern, Z. (2018). The novel EHEC gene asa overlaps the TEGT transporter gene in antisense and is regulated by NaCl and growth phase. *Sci Rep, 8*(1), 17875. doi:10.1038/s41598-018-35756-y

Vigil-Stenman, T., Ininbergs, K., Bergman, B., & Ekman, M. (2017). High abundance and expression of transposases in bacteria from the Baltic Sea. *ISME J, 11*(11), 2611-2623. doi:10.1038/ismej.2017.114

Wang, J., Rennie, W., Liu, C., Carmack, C. S., Prévost, K., Caron, M. P., et al. (2015). Identification of bacterial sRNA regulatory targets using ribosome profiling. *Nucleic Acid Res, 43*(21), 12.

Wang, S., & Kool, E. T. (1995). Origins of the large differences in stability of DNA and RNA helices: C-5 methyl and 2'-hydroxyl effects. *Biochemistry, 34*(12), 4125-4132. doi:10.1021/bi00012a031

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet, 10*(1), 57-63. doi:10.1038/nrg2484

Warren, A. S., Archuleta, J., Feng, W. C., & Setunal, J. C. (2010). Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics, 11*(131).

Waters, L. S., & Storz, G. (2009). Regulatory RNAs in bacteria. *Cell, 136*(4), 615-628. doi:10.1016/j.cell.2009.01.043

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature, 171*(4356), 737-738. doi:10.1038/171737a0

Wei, X., & Zhang, J. (2014). A simple method for estimating the strength of natural selection on overlapping genes. *Genome Biol Evol, 7*(1), 381-390. doi:10.1093/gbe/evu294

Williams, T., & Fried, M. (1986). A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3' ends. *Nature, 322*(6076), 275-279. doi:10.1038/322275a0

Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res, 7*, 1338. doi:10.12688/f1000research.15931.2

Wong, T. Y., Fernandes, S., Sankhon, N., Leong, P. P., Kuo, J., & Liu, J. K. (2008). Role of premature stop codons in bacterial evolution. *J Bacteriol, 190*(20), 6718-6725. doi:10.1128/JB.00682-08

Woolstenhulme, C. J., Guydosh, N. R., Green, R., & Buskirk, A. R. (2015). High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Rep, 11*(1), 13-21. doi:10.1016/j.celrep.2015.03.014

Wu, X., Kim, T. K., Baxter, D., Scherler, K., Gordon, A., Fong, O., et al. (2017). sRNAnalyzer-a flexible and customizable small RNA sequencing data analysis pipeline. *Nucleic Acids Res, 45*(21), 12140-12151. doi:10.1093/nar/gkx999

Xu, L., Chen, H., Hu, X., Zhang, R., Zhang, Z., & Luo, Z. W. (2006). Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Mol Biol Evol, 23*(6), 1107-1108. doi:10.1093/molbev/msk019

Yang, X. Y., He, K., Du, G., Wu, X., Yu, G., Pan, Y., et al. (2016). Integrated Translatomics with Proteomics to Identify Novel Iron-Transporting Proteins in Streptococcus pneumoniae. *Front Microbiol, 7*, 78. doi:10.3389/fmicb.2016.00078

Yanofsky, C., & Lennox, E. S. (1959). Transduction and recombination study of linkage relations among the genes controlling tryptophan synthesis in Escherichia coli. *Virology, 8*, 425-447. doi:10.1016/0042-6822(59)90046-7

Ye, H., Meehan, J., Tong, W., & Hong, H. (2015). Alignment of Short Reads: A Crucial Step for Application of Next-Generation Sequencing Data in Precision Medicine. *Pharmaceutics, 7*(4), 523-541. doi:10.3390/pharmaceutics7040523

Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Shoshan, A., et al. (2003). Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol, 21*(4), 379-386. doi:10.1038/nbt808

Zehentner, B., Ardern, Z., Kreitmeier, M., Scherer, S., & Neuhaus, K. (2020). A Novel pH-Regulated, Unusual 603 bp Overlapping Protein Coding Gene pop Is Encoded Antisense to ompA in Escherichia coli O157:H7 (EHEC). *Front Microbiol, 11*, 377. doi:10.3389/fmicb.2020.00377

Zhang, L., Tan, Y., Fan, S., Zhang, X., & Zhang, Z. (2019). Phylostratigraphic analysis of gene co-expression network reveals the evolution of functional modules for ovarian cancer. *Sci Rep*, *9*(1), 2623. doi:10.1038/s41598-019-40023-9

Zhou, H. Q., Ning, L. W., Zhang, H. X., & Guo, F. B. (2014). Analysis of the relation between genomic GC Content and patterns of base usage, codon usage and amino acid usage in prokaryotes: similar GC content adopts similar compositional frequencies regardless of the phylogenetic lineages. *PLoS One*, *9*(9), e107319. doi:10.1371/journal.pone.0107319

Zhou, Q., & Wang, W. (2008). On the origin and evolution of new genes--a genomic and experimental perspective. *J Genet Genomics*, *35*(11), 639-648. doi:10.1016/S1673-8527(08)60085-5

Zhu, L. Q., Gangopadhyay, T., Padmanabha, K. P., & Deutscher, M. P. (1990). Escherichia coli rna gene encoding RNase I: cloning, overexpression, subcellular distribution of the enzyme, and use of an rna deletion to identify additional RNases. *J Bacteriol*, *172*(6), 3146-3151. doi:10.1128/jb.172.6.3146-3151.1990

Zur Bruegge, J., Einspanier, R., & Sharbati, S. (2017). A Long Journey Ahead: Long Non-coding RNAs in Bacterial Infections. *Front Cell Infect Microbiol*, *7*, 95. doi:10.3389/fcimb.2017.00095

Acknowledgements

Mein ausdrücklicher Dank gilt meinem Doktorvater Herrn Prof. Dr. Siegfried Scherer, der mir die Anfertigung dieser Arbeit an seinem Lehrstuhl ermöglicht hat. Dadurch konnte ich mir endlich den Traum meiner Promotion verwirklichen.

Für die Übernahme des Zweitgutachtens bedanke ich mich herzlich bei Herrn PD. Dr. Klaus Neuhaus. Neben dieser Aufgabe standst du, lieber Klaus, mir auch als Mentor zur Seite, an den ich mich mit Fragen aller Art wenden konnte.

Ein ganz besonderer Dank gilt Dr. Zachary Ardern, der sich nie über mein Englisch beschwert hat. Durch unsere ständige Zusammenarbeit hat sich nicht nur dieses, sondern auch meine gesamte Arbeit verbessert. Sein wissenschaftlicher Input zu meiner Arbeit ist nicht von der Hand zu weisen, und führte zu interessanten Diskussionen von denen diese Arbeit profitiert hat. Dafür möchte ich mich herzlich bei ihm bedanken. Danke auch an Dr. Christopher Huptas für seine informatische Unterstützung.

Und dann gilt all meine Dankbarkeit meinen Kollegen, die mir nicht nur die Arbeit, sondern auch das Leben versüßt haben. Anna und Babsi, ihr habt mich zum ersten Mal in meinem Leben in die Berge entführt. Ein kleines Stück meines Herzen habe ich an diese, aber auch an euch verloren! Und wenn es mal hart auch hart kam, ward ich stets da um mir bei kleineren oder größeren Missgeschicken zu helfen. Ich habe in dieser Zeit einen unfassbaren Road Trip durch Schottland mit Anna, Annemarie und Chris erleben dürfen. Dieses Gefühl, als würden wir uns schon ewig kennen, möchte ich für immer im Herzen tragen, denn eure Freundschaft, den Spaß den wir hatten und haben und die gute Gesellschaft, die ich mit jedem Einzelnen von euch verbinde, wird mich immer schmunzeln lassen! Natürlich wird auch Micha nicht vergessen und die tolle Zeit die ich durch sie erleben durfte, ebenfalls gipfelnd in einem gemeinsamen Urlaub. Mit dir konnte ich einfach über alles reden und auch wenn man sagt, dass Schweigen Gold wäre, so wisst ihr doch, dass das Schweigen nicht so mein Ding ist. Isabel hat mich durch unsere geteilten Hobbies immer wieder durchatmen und die Welt kurz vergessen lassen. Für diese stets willkommene Ablenkung möchte ich mich bedanken, hat sie mich doch immer wieder geerdet.

Auch möchte ich mich bei Stefan, Franzi Giehren, Franzi Graf, Anika, Kathi, Etienne, Genia und allen TA’s für die herzliche Zusammenarbeit und die netten Begegnungen bedanken.

Zuletzt möchte ich mich bei meiner Familie bedanken. Mama, Papa ihr habt mir die Welt ermöglicht, euch stolz zu machen und euch dadurch auch nur ansatzweise etwas zurückzugeben, wird mir immer ein Ziel im Leben sein. Isi und Julian, ihr seid mir, genau wie Mama und Papa Vorbilder. Ihr seid für mich da, in guten, wie in schweren Tagen. Und ihr habt mein Leben zusätzlich durch den kleinen Leo bereichert, für den ich mir jetzt und in alle Ewigkeit ein Bein ausreißen werden, soll auch er stolz auf sein Tantchen sein können. Ihr alle gebt mir den Antrieb jeden Tag mehr aus meinem Leben zu machen. Dafür bin ich euch allen, meiner Familie und meinen Freunden, auf ewig dankbar.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die bei der promotionsführenden Einrichtung TUM School of Life Sciences der TUM zur Promotionsprüfung vorgelegte Arbeit mit dem Titel

Improvements to ribosome profiling analysis in *E. coli* K-12 and diverse prokaryotes, and the detection of antisense overlapping genes

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

Ich habe keine Organisation eingeschaltet, die gegen Entgelt Betreuerinnen und Betreuer für die Anfertigung von Dissertationen sucht, oder die mir obliegenden Pflichten hinsichtlich der Prüfungsleistungen für mich ganz oder teilweise erledigt. Ich habe die Dissertation in dieser oder ähnlicher Form in keinem anderen Prüfungsverfahren als Prüfungsleistung vorgelegt.

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

Die öffentlich zugängliche Promotionsordnung der TUM ist mir bekannt, insbesondere habe ich die Bedeutung von § 28 (Nichtigkeit der Promotion) und § 29 (Entzug des Doktorgrades) zur Kenntnis genommen. Ich bin mir der Konsequenzen einer falschen Eidesstattlichen Erklärung bewusst. Mit der Aufnahme meiner personenbezogenen Daten in die Alumni-Datei bei der TUM bin ich einverstanden.

List of publication

Recommendations for bacterial ribosome profiling experiments based on bioinformatic evaluation of published data

J. Biol. Chem. 2020 vol 295 no.27; 8999-9011

Glaub A., Huptas C., Neuhaus, K., Ardern Z. (2020)

Curriculum vitae

2018 – 2021	Promotion am Lehrstuhl für mikrobielle Ökologie, Technische Universität München, DE Promotionsthese: „ Improvements to ribosome profiling analysis in *E. coli* K-12 and diverse prokaryotes, and the detection of antisense overlapping genes”
2013 – 2016	Masterstudium in molekularer Biomedizin, Westfälische Wilhelms Universität Münster, DE Masterthese „Analyse der genetischen Diversität im codierenden Bereich zweier Gene des Phase-II-Metabolismus von Tetrahydrocannabinol (THC).“
2010 – 2013	Bachelorstudium in Biologie, Universität Bielefeld, DE Bachelorthese „Genomsequenzierung und Annotation von *Corynebacterium auriscanis* und *C. falsenii* zur Identifikation von Virulenzfaktoren”
2009 – 2010	Staatsexamen in Rechtswissenschaften, Universität Osnabrück, DE
2000 – 2009	Abitur, Friedrichs-Gymnasium Herford, DE

Supplementary Files

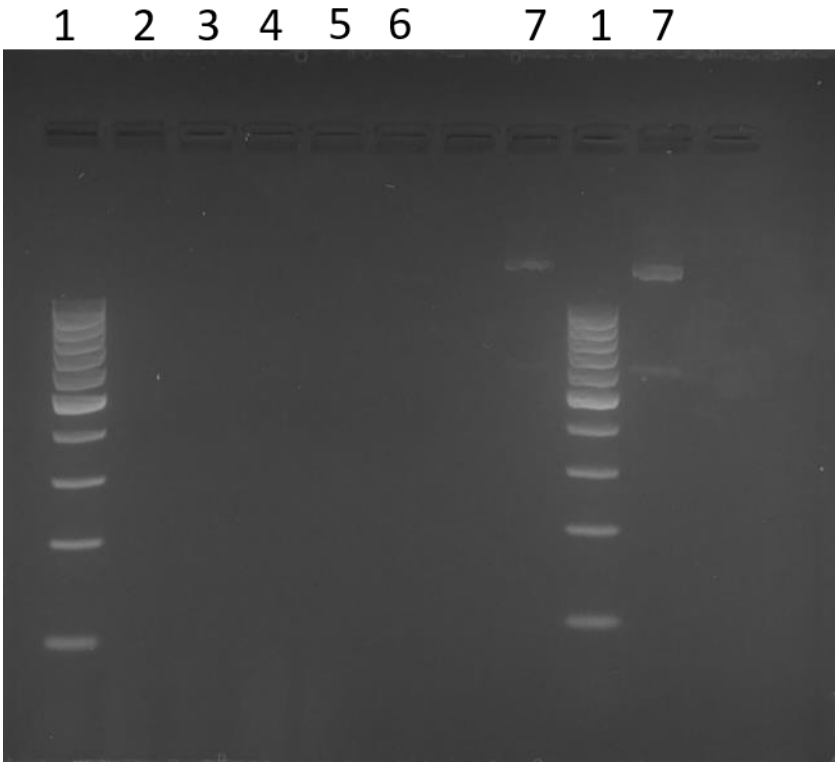

Figure S1: Gel picturing showing DNA digestion success based on performed 16S PCR after digestion. 1) 100bp ladder, double application of samples 2) + 3) RNA I, 4) + 5) RNA II, 6) negative control (RNA-free H_20 as PCR template); 7) double application of positive control (1 µl and 2µl tested as PCR template).

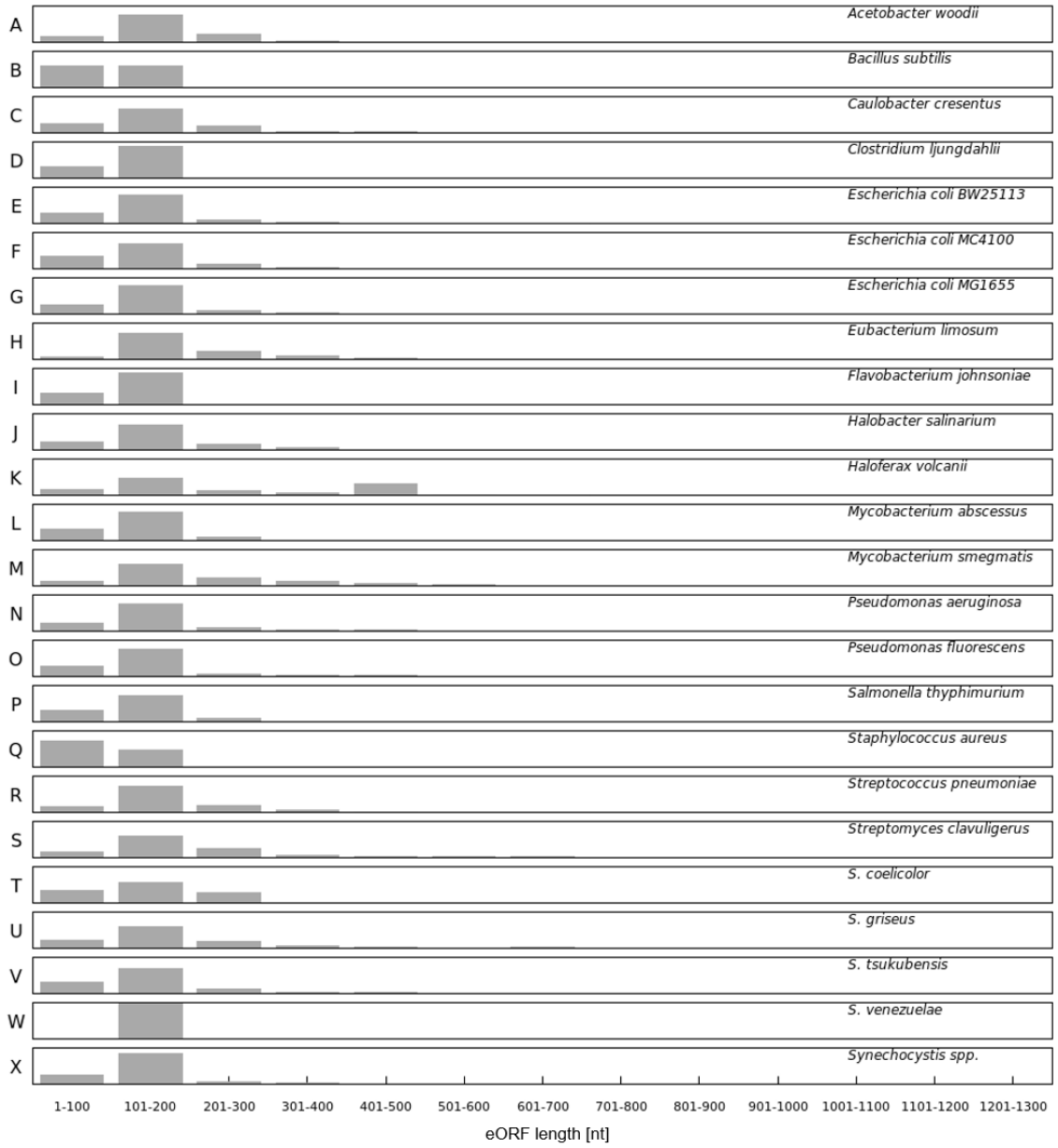

Figure S2: eORF length distribution for all species analysed in alphabetical order, genome specific GC-contents are as followed: (A) 39.3, (B) 43.5, (C) 67.2, (D) 31.1, (E) 50.8, (F) 50.6, (G) 50.6 , (H) 46.9 , (I) 34.1, (J) 67.9, (K) 66.6, (L) 64.1, (M) 67.4, (N) 66.6, (O) 60.8, (P) 52.2, (Q) 32.9, (R) 40, (S) 72.7, (T) 72.1, (U) 72.2, (V) 71.9, (W) 72.5, (X) 47.5 (n = 24). Shown is the percentage calculation for abundance of eORFs within one length category.

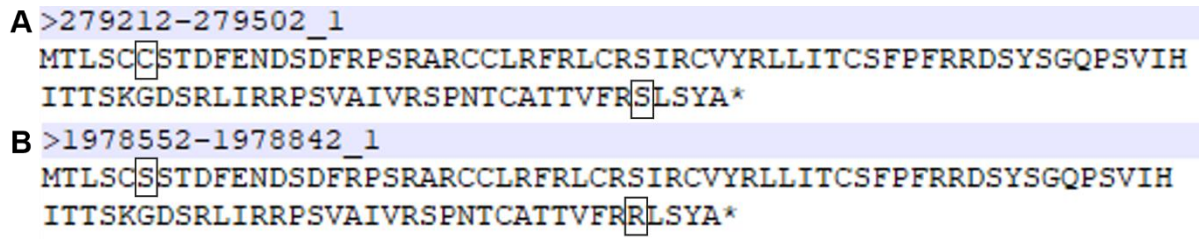

Figure S3: Shown are the protein sequences for the eORFs of (A) candidate Escherichia1 and (B) Escherichia3 (Escherichia4 respectively due to exact sequence accordance). Overlaps are similar in all but two amino acids incorporated. Substitutions are highlighted by squares at the positions of interest.

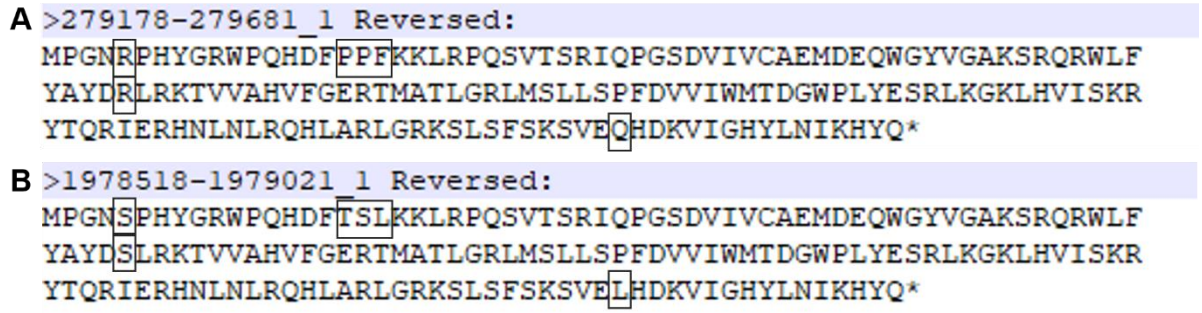

Figure S4: Shown are the protein sequences for mother genes of (A) candidate Escherichia1 and (B) Escherichia3 (Escherichia4 respectively due to exact sequence accordance). Mother gene sequences (coding for IS1 transposase B) are similar in all but six amino acids incorporated. Substitutions are highlighted by squares at the positions of interest.

Table S1: List of chemicals used in performed RIBO-Seq and RNA-Seq experiment with corresponding providers.

Chemicals	**Provider**
Agarose	Sigma-Aldrich, St. Louis, MO, USA
Ammonium chloride (NH_4Cl)	Roth, Karlsruhe, Germany
Ammonium persulfate (APS)	Roth, Karlsruhe, Germany
Boric acid	Sigma-Aldrich, St. Louis, MO, USA
Calcium chloride ($CaCl_2$)	Roth, Karlsruhe, Germany
Chloroform	Roth, Karlsruhe, Germany
D(+)-Saccharose ≥99,5%, p.a.	Roth, Karlsruhe, Germany
DTT (Dithiothreitol)	Roche, Basel, Switzerland
DreamTaq PCR Master Mix (2 x)	Thermo Fisher Scientific, Waltham, MA, USA
ethanol absolute	VWR International, Darmstadt, Germany
ethylenediaminetetraacetic acid (EDTA)	Roth, Karlsruhe, Germany
Fluorescin-Na	Roth, Karlsruhe, Germany
GelRed® Nucleic Acid 10.000X	Biotium, Fremont, CA, USA
Glycogen RNA Grade	Thermo Fisher Scientific, Waltham, MA, USA
Hydrophilic Streptavidin Magnetic Beads	NEB, Ipswich, MA, USA
Isopropanol (2-Propanol)	Sigma-Aldrich, St. Louis, MO, USA
Magnesium chloride ($MgCl_2$)	Sigma-Aldrich, St. Louis, MO, USA
NEB4 Buffer	NEB, Ipswich, MA, USA
Noves TBE Urea Sample Buffer (2x)	Thermo Fisher Scientific, Waltham, MA, USA
NP-40	Sigma-Aldrich, St. Louis, MO, USA
Patent Blue V	Sigma-Aldrich, St. Louis, MO, USA
RNA Loading Dye (2x)	NEB, Ipswich, MA, USA
Rotiphorese® sequencing gel buffer concentrate	Roth, Karlsruhe, Germany
Rotiphorese® sequencing gel concentrate	Roth, Karlsruhe, Germany
Rotiphorese® sequencing gel thinner	Roth, Karlsruhe, Germany
Sodium acetate (NaOAc)	Sigma-Aldrich, St. Louis, MO, USA
SYBR™ Gold nucleic acid gel stain	Thermo Fisher Scientific, Waltham, MA, USA
TEMED	Roth, Karlsruhe, Germany
TRIS- hydrochloric acid (HCl)	Roth, Karlsruhe, Germany
Triton X-100	Roth, Karlsruhe, Germany
Trizol	Thermo Fisher Scientific, Waltham, MA, USA
TURBO™ DNase buffer (10 x)	Thermo Fisher Scientific, Waltham, MA, USA
1 kb DNA ladder	NEB, Ipswich, MA, USA
100 bp DNA ladder	NEB, Ipswich, MA, USA

Table S2: List of kits used during RIBO-Seq and RNA-Seq experimental proceedings. Additional information contains purposes of named kits plus their provider.

Kit name	Purpose	Provider
Bioanalyzer High Sensitivity DNA Kit + Chip	Library quality control	Agilent Technologies, Santa Clara, CA, USA
Bioanalyzer RNA 600 Nano Kit + Chip	RNA quality control	Agilent Technologies, Santa Clara, CA, USA
miRNeasy Mini Kit	RNA purification	Qiagen, Hilden, Germany
MiSeq Reagent Kit v3	Next Generation Sequencing	Illumina, San Diego, CA, USA
Pan-Prokaryotes (Pan-riboPOOLs)	rRNA depletion	siTooLs Biotech GmbH, Martiensried, Germany
Qubit RNA High Sensitivity Assay	RNA concentration measurement	Thermo Fisher Scientific, Waltham, MA, USA
Qubit™ 1X dsDNA HS Assay-Kit	DNA concentration measurement	Thermo Fisher Scientific, Waltham, MA, USA
TruSeq Small RNA Library Prep Kit -Set A	Library preparation	Illumina, San Diego, CA, USA
TruSeq Small RNA Library Prep Kit, Core Solutions	Library preparation	Illumina, San Diego, CA, USA

Table S3: Table summarising samples used (with their ENA assigned numbers) with additional raw read numbers, reads left after adapter trimming and filtering and respective adapter sequence as removal template.

Sample	Raw read number	Read number after trim-/filtering	Adapter sequence (published or detected)
SRR1734437	64,469,767	18,393,118	CTGTAGGCACCATCAAT
SRR1734438	51,236,059	17,506,641	CTGTAGGCACCATCAAT
SRR1734439	47,811,573	23,341,030	CTGTAGGCACCATCAAT
SRR1734440	54,023,827	5,549,778	CTGTAGGCACCATCAAT
SRR1734441	70,560,079	28,603,430	CTGTAGGCACCATCAAT
SRR1734442	24,162,556	4,537,688	CTGTAGGCACCATCAAT
SRR1734443	33,378,397	7,471,395	CTGTAGGCACCATCAAT
SRR1734444	27,920,647	10,773,198	CTGTAGGCACCATCAAT
SRR4023274	18,237,413	9,379,697	CTGTAGGCACCATCAAT
SRR4023280	27,160,645	7,269,556	CTGTAGGCACCATCAAT
ERR618770	194,819,312	65,991,623	GAGGCTGAGGCGTGATGACGAGGCAC
ERR618771	161,912,593	46,832,304	GAGGCTGAGGCGTGATGACGAGGCAC
SRR1613268	3,640,574	1,571,362	CTGTAGGCACCATCAAT

SRR1613269	14,545,092	5,200,100	CTGTAGGCACCATCAAT
SRR1613270	5,358,746	2,086,650	CTGTAGGCACCATCAAT
SRR1613272	8,779,929	3,467,170	CTGTAGGCACCATCAAT
SRR1613277	6,811,320	1,572,969	CTGTAGGCACCATCAAT
SRR1613278	20,537,706	4,256,485	CTGTAGGCACCATCAAT
SRR1613280	3,642,225	807,692	CTGTAGGCACCATCAAT
SRR1613281	18,767,441	4,065,713	CTGTAGGCACCATCAAT
SRR1613283	3,666,535	802,984	CTGTAGGCACCATCAAT
SRR1613285	23,299,777	4,965,105	CTGTAGGCACCATCAAT
SRR1613287	23,648,253	8,415,606	CTGTAGGCACCATCAAT
SRR1200730	10,558,976	1,975,001	GGCTGAGGCGTGATGACGAGGCAC
SRR1200731	8,589,645	1,343,820	GGCTGAGGCGTGATGACGAGGCAC
SRR1200738	7,263,033	1,417,420	GGCTGAGGCGTGATGACGAGGCAC
SRR1200739	10,452,140	1,689,198	GGCTGAGGCGTGATGACGAGGCAC
SRR1200750	7,139,565	1,028,411	GGCTGAGGCGTGATGACGAGGCAC
SRR1200751	8,769,881	1,335,550	GGCTGAGGCGTGATGACGAGGCAC
SRR1613263	23,558,119	18,917,577	CTGTAGGCACCATCAAT
SRR1613265	20,020,207	16,302,283	CTGTAGGCACCATCAAT
SRR1613266	18,140,727	13,945,969	CTGTAGGCACCATCAAT
SRR1583082	22,259,168	13,406,859	CTGTAGGCACCATCAAT
SRR1583083	26,096,770	13,509,892	CTGTAGGCACCATCAAT
SRR1583084	23,465,502	12,799,823	CTGTAGGCACCATCAAT
SRR4190324	19,701,724	7,126,787	CTGTAGGCACCATCAAT
SRR4190325	15,283,224	7,315,769	CTGTAGGCACCATCAAT
SRR4190326	17,699,028	5,717,589	CTGTAGGCACCATCAAT
SRR364363	20,151,889	9,224,533	CTGTAGGCACCATCAAT
SRR364364	19,274,111	8,560,579	CTGTAGGCACCATCAAT
SRR364365	19,888,911	10,649,863	CTGTAGGCACCATCAAT
SRR364366	19,279,468	9,503,165	CTGTAGGCACCATCAAT
SRR364367	24,137,567	11,589,789	CTGTAGGCACCATCAAT
SRR364368	24,490,400	14,705,508	CTGTAGGCACCATCAAT
SRR364369	23,384,961	7,427,654	CTGTAGGCACCATCAAT
SRR364370	24,706,067	15,049,873	CTGTAGGCACCATCAAT
SRR2016457	35,352,233	10,726,959	GATCTCGTATGCCGTCTTCTG
SRR2016465	29,770,749	7,303,237	GATCTCGTATGCCGTCTTCTG

Table S4: *E. coli* K12 substrains used for RIBO-Seq data evaluation. Genome specific NCBI reference sequence number and accession number are listed.

***E. coli* K-12 substrains**	**BW25113**	**MC4100**	**MG1655**
NCBI RefSeq Genome	NZ_CP009273.1	NZ_HG738867.1	NC_000913.3
NCBI RefSeq Accession	GCF_000750555.1	GCF_000499485.1	GCF_000005845.2

Table S5: Detection numbers of multiple prokaryotic species tested for distribution analysis. Detections are based on an in-house script (ORFFinder) or the available detection tool DeepRibo. Sample numbers are adopted from ENA, with read number detected in raw sequencing file and read amount left after adapter trim and rRNA/tRNA removal.

Species	**Sample number**	**Read amount before trim**	**Read amount after trim/filter**	**eORFs in-house**	**eORFs DeepRibo**
Acetobacterium woodii DSM 1030					
	ERR3428526	249,172,796	35,488,175	414	0
	ERR3428527	69,476,513	10,580,971	651	0
	ERR3428528	131,703,279	13,641,574	303	0
	ERR3428529	139,370,873	18,840,528	247	0
Bacillus subtilis subsp. subtilis str. 168					
	SRR987023	6,567,480	1,528,968	1	0
	SRR987022	3,692,648	1,530,526	0	0
	SRR987020	8,433,879	3,199,021	1	0
	SRR407279	15,991,160	4,880,082	1	0
	SRR987018	9,660,224	5,115,255	1	0
	SRR407278	18,230,224	6,686,023	0	0
	SRR407280	35,407,892	11,988,286	0	0
	SRR407281	176,894,215	58,443,192	0	0
Caulobacter crescentus (*C. vibrioides* NA1000)					
	SRR1167750	2,656,902	1,995,944	58	1
	SRR1167751	4,274,091	4,073,158	31	1
	SRR1991280	13,538,820	5,663,671	35	9
	SRR1991278	17,278,060	6,483,702	22	8
	SRR1991279	20,443,469	9,454,087	0	7
	SRR1991277	22,993,484	9,509,804	2	7
	SRR1991275	20,703,090	11,464,530	155	18
	SRR1991276	80,717,688	19,773,223	72	26

Clostridium ljungdahlii DSM 13528					
	SRR6286686	31,815,773	10,495,830	2	0
	SRR6286687	31,226,831	8,757,358	0	0
	SRR6286690	95,785,654	17,027,233	5	1
	SRR6286691	74,932,501	10,973,071	2	0
	SRR6286692	44,431,801	4,585,435	6	0
	SRR6286693	50,618,624	3,856,272	9	0
Eubacterium limosum ATCC8486					
	SRR5442631	129,922,465	97,381,460	0	0
	SRR5442632	31,128,811	23,304,526	87	0
	SRR5442633	156,504,637	124,161,794	107	0
	SRR5442634	37,692,055	29,855,609	85	0
	SRR5442635	141,679,752	121,011,679	19	0
	SRR5442636	34,849,334	29,649,625	9	0
	SRR5442637	142,722,087	29,159,603	22	0
	SRR5442638	34,892,020	29,159,603	17	0
Flavobacterium johnsoniae UW101					
	SRR10100140	19,043,505	760,599	2	0
	SRR10100141	14,410,982	722,297	1	0
	SRR10100142	26,216,828	1,352,373	5	0
Halobacterium salinarum					
	SRR2583990	19,815,990	623,873	197	0
	SRR2583992	20,619,543	356,811	44	0
	SRR2583993	10,916,190	201,457	29	0
	SRR2583995	10,325,125	395,820	138	0
	SRR2583998	17,025,028	198,763	28	0
	SRR2583999	17,474,431	408,640	94	0
	SRR2584009	21,546,550	408,605	150	0
	SRR2584010	9,974,124	156,456	45	0
	SRR2584012	24,563,352	771,967	356	0
Haloferax volcanii DS2					
	SRR10294592	24,325,319	7,148,962	154	0
	SRR10294593	108,269,604	62,282,106	136	0

	SRR10294594	4,480,382	2,298,388	133	0
	SRR10294595	8,990,526	3,836,447	205	0
	SRR10294596	9,421,109	1,032,894	174	0
	SRR10294597	16,070,885	6,581,436	240	0
	SRR10294598	15,432,004	5,689,637	353	0
Mycobacterium abscessus ATCC 19977					
	SRR2392990	24,739,218	2,039,285	52	4
	SRR2392989	33,032,537	3,252,765	71	2
Mycobacterium smegmatis MC2 155					
	ERR599190	99,829,689	7,770,333	25	9
	ERR599192	80,612,378	13,624,091	2028	42
Pseudomonas aeruginosa ATCC33988/AO1					
	SRR5356894	85,820,972	2,543,266	1028	5
	SRR5356904	95,133,754	3,316,735	690	3
	SRR5356888	65,592,402	3,547,592	864	11
	SRR5356908	98,145,010	3,808,674	844	5
	SRR5356898	106,010,338	3,911,912	890	6
	SRR5356892	111,201,922	4,033,941	710	0
	SRR5356902	130,257,169	4,236,850	817	10
	SRR5356900	145,025,585	4,758,652	812	16
	SRR5356896	135,235,702	4,836,495	586	11
	SRR5356906	124,715,511	4,939,063	547	7
	SRR5356893	90,928,652	5,229,127	87	1
	SRR5356886	94,572,638	5,726,344	494	17
	SRR5356907	73,952,429	6,253,598	98	3
	SRR5356890	136,181,328	6,998,736	292	29
	SRR5356897	101,520,011	7,997,013	38	4
	SRR5356903	71,767,127	9,632,036	4	2
	SRR5356887	62,093,655	12,741,784	324	12
	SRR5356885	92,012,581	12,777,043	191	11
	SRR5356905	90,030,430	13,969,059	214	3
	SRR5356901	94,948,217	14,385,074	212	8
	SRR5356891	83,809,807	14,982,012	267	11

	SRR5356899	83,560,401	15,036,413	329	12
	SRR5356889	100,563,944	19,122,838	341	20
	SRR5356895	106,331,232	25,860,477	166	15
Pseudomonas fluorescens F113 (SBW25)					
	ERR1797531	13,719,271	3,375,883	166	0
	ERR1797530	27,730,412	5,573,534	106	0
	ERR1797529	37,747,215	7,943,240	84	0
	ERR1797532	24,135,343	7,618,436	69	0
Salmonella enterica subsp. enterica serovar typhimurium str. LT2					
	SRR4417735	32,704,873	18,517,302	56	2
	SRR4417736	17,593,796	9,729,040	0	1
	SRR4417737	31,420,973	13,462,771	65	3
	SRR4417738	27,566,313	13,911,214	28	2
	SRR5090708	191,796,596	63,046,643	42	66
	SRR5090709	174,834,326	63,385,584	51	56
Staphylococcus aureus subsp. aureus NCTC 8325					
	SRR1265836	13,081,432	592,248	1	0
	SRR1265837	9,430,867	1,463,752	1	0
	SRR1265839	10,954,130	749,828	0	0
	SRR1265840	14,696,988	428,145	0	0
	SRR1265842	12,124,389	1,804,100	0	0
	SRR1265843	9,563,522	622,701	1	0
	SRR1265846	12,906,461	1,742,443	0	0
	SRR1265847	16,000,000	718,000	0	0
	SRR1265848	22,788	1,021	0	0
	SRR2733429	16,000,000	481,370	0	0
	SRR2733430	16,000,000	477,822	0	0
	SRR2733431	13,928,237	416,958	0	0
	SRR2733432	16,000,000	1,271,621	4	0
	SRR2733433	16,000,000	1,276,102	5	0
	SRR2733434	16,000,000	1,261,907	4	0
	SRR2733435	16,000,000	1,264,061	4	0
	SRR2733436	6,285,308	500,716	0	0

	SRR2733445	16,000,000	482,054	1	0
	SRR2733446	16,000,000	482,892	1	0
	SRR2733447	16,000,000	483,453	0	0
	SRR2733448	4,091,470	121,372	0	0
	SRR2733449	16,000,000	1,293,590	2	0
	SRR2733450	16,000,000	1,301,908	2	0
	SRR2733451	16,000,000	1,306,651	1	0
	SRR2733452	2,132,733	167,316	0	0
	SRR2733460	16,000,000	669,493	1	0
	SRR2733461	13,895,704	585,110	1	0
	SRR2733462	16,000,000	1,903,635	1	0
	SRR2733463	16,000,000	1,862,912	1	0
	SRR2733464	16,000,000	1,904,336	0	0
	SRR2733465	4,313,498	472,712	0	0
	SRR2733471	16,000,000	3,099,260	0	0
	SRR2733472	16,000,000	3,107,825	2	0
	SRR2733473	9,946,206	1,888,373	1	0
	SRR2733474	16,000,000	2,027,267	3	0
	SRR2733475	16,000,000	2,040,915	5	0
	SRR2733476	9,098,581	1,133,138	2	0
Streptococcus pneumoniae D39					
	SRR2992164	6,399,746	5,227,238	1554	0
	SRR3031488	4,723,164	3,815,913	2242	0
	SRR3031489	5,969,494	2,482,818	2952	0
Streptomyces clavuligerus ASM169367v1					
	SRR8718525	295,724,334	16,486,892	43	0
	SRR8718526	307,178,979	24,196,090	50	0
	SRR8718527	253,508,213	12,335,014	128	0
	SRR8718528	278,275,008	11,777,616	152	0
	SRR8718529	270,412,414	10,246,762	213	0
	SRR8718530	247,353,047	9,277,273	5	0
	SRR8718531	238,332,934	10,865,692	237	0
	SRR8718532	265,467,800	14,027,593	194	0

Streptomyces coelicolor A3(2)					
	SRR2043967	21,467,666	361,454	5	0
	SRR2043969	22,247,790	559,374	26	1
	SRR2043966	19,101,972	601,934	4	0
	SRR2043968	23,725,610	651,147	8	1
Streptomyces griseus subsp. griseus NBRC 13350					
	SRR10212831	143,788,818	4,699,624	10	0
	SRR10212832	187,573,750	11,039,479	66	1
	SRR10212833	155,225,446	3,769,262	22	1
	SRR10212834	162,820,587	8,578,068	3	0
	SRR10212835	178,011,502	11,851,509	60	2
	SRR10212836	191,018,264	10,834,240	40	1
	SRR10212837	207,591,209	13,867,575	20	0
	SRR10212838	146,481,080	6,828,396	10	1
Streptomyces tsukubensis					
	SRR5443325	125,024,824	2,171,449	20	0
	SRR5443326	124,524,054	1,990,028	58	0
	SRR5443327	132,160,059	3,971,304	61	0
	SRR5443328	113,065,267	1,648,784	31	0
	SRR5443329	162,942,510	12,977,918	97	0
	SRR5443330	166,664,595	11,422,798	80	0
	SRR5443331	146,033,026	2,573,723	49	0
	SRR5443332	199,958,654	7,648,370	13	0
Streptomyces venezuelae ATCC 10712					
	SRR1021839	319,268,737	258,412,481	0	0
	SRR1021840	214,371,450	198,785,737	0	0
	SRR1021841	329,759,627	273,382,674	1	0
	SRR1021842	166,241,490	138,028,609	1	0
	SRR1021843	408,127,842	348,866,580	2	0
	SRR1021844	304,119,238	258,439,907	0	0
	SRR1021845	238,259,689	198,953,437	1	0
	SRR1021846	168,577,692	144,243,170	0	0

Synechocystis sp. PCC 6803					
	ERR2736130	67,594,046	40,391,662	147	0
	ERR2736131	102,553,513	65,302,494	55	0
	ERR2736132	74,184,736	43,438,600	158	0
	ERR2736133	96,384,647	56,788,237	105	0
	ERR2736134	105,335,418	47,130,902	335	0

Table S6: Species-specific genome information with genome size, number of annotated proteins within a genome, GC-content, cell wall based gram category, NCBI reference genome sequence number and corresponding assembly number.

Species	**Genome Size [Mb]**	**Proteins**	**GC Content**	**Gram**	**NCBI Reference Sequence**	**Assembly**
Halobacterium salinarum	2.01	2,095	67.9	gram-negative	NC_002607.1	GCF_000006805.1
Streptococcus pneumoniae D39	2.06	1,861	40	gram-positive	NC_003098.1	GCF_000014365.2
Staphylococcus aureus subsp. aureus NCTC 8325	2.82	2,767	32.9	gram-positive	NC_007795.1	GCF_000013425.1
Haloferax volcanii DS2	2.85	2,883	66.6	gram-positive	NC_013967.1	GCF_000025685.1
Synechocystis sp. PCC 6803	3.85	3,559	47.5	gram-negative	NC_000911.1	GCF_000009725.1
Acetobacterium woodii DSM 1030	4.04	3,564	39.3	gram-positive	NC_016894.1	GCA_000247605.1
Caulobacter crescentus (*C. vibrioides* NA1000)	4.04	3,886	67.2	gram-negative	NC_011916.1	GCF_000022015.1
Bacillus subtilis subsp. subtilis str. 168	4.22	4,237	43.5	gram-positive	NC_000964.3	GCF_000009055.1
Clostridium ljungdahlii DSM 13528	4.63	4,116	31.1	gram-positive	NC_014328.1	GCF_000143685.1
Escherichia coli str. K-12 substr. MG1655	4.64	4,242	50.8	gram-negative	NC_000913.3	GCF_000005845.2

Eubacterium limosum ATCC8486	4.74	4,419	46.9	gram-positive	NZ_LR215983.1	GCF_000807675.2
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	4.86	4,446	52.2	gram-negative	NC_003197.2	GCF_000006945.2
Mycobacterium abscessus ATCC 19977	5.07	4,920	64.1	gram-positive	NC_010397.1	GCF_000069185.1
Flavobacterium johnsoniae UW101	6.10	5,091	34.1	gram-negative	NC 009441.1	GCF_000016645.1
Pseudomonas aeruginosa ATCC33988/AO1	6.26	5,572	66.6	gram-negative	NC_002516.2	GCF_000006765.1
Bacteroides thetaiotaomicron VPI-5482	6.26	4,646	42.8	gram-negative	NC_004663.1	GCF_000011065.1
Pseudomonas fluorescens F113 (SBW25)	6.85	5,919	60.8	gram-negative	NC_012660.1	GCF_000009225.2
Streptomyces clavuligerus ASM169367v1	6.88	5,529	72.7	gram-positive	NZ_CP016559.1	GCF_001693675.1
Mycobacterium smegmatis MC2 155	6.99	6,480	67.4	gram-positive	NC_008596.1	GCF_000015005.1
Streptomyces tsukubensis	7.96	6,239	71.9	gram-positive	NZ_CP020700.1	GCF_003932715.1
Streptomyces venezuelae ATCC 10712	8.22	7,141	72.5	gram-positive	NZ_CP029197.1	GCF_008639165.1
Streptomyces griseus subsp. griseus NBRC 13350	8.55	6,959	72.2	gram-positive	NC_010572.1	GCF_000010605.1
Streptomyces coelicolor A3(2)	8.67	7,767	72.1	gram-positive	NC_003888.3	GCF_000203845.1

Table S7: Frame relation of mother gene and overlap calculated in percentage per species. Compared is the distribution of all eORFs possibly predicted per species to those considered actual translated (based on matching thresholds RRKM and coverage).

Species	**All eORFs**	**Translated**	**All eORFs**	**Translated**	**All eORFs**	**Translated**
	sas11		sas12		sas13	
B. subtilis	26.3	0	36.6	50	37.1	50
C. cresentus	19.4	12.6	64.7	68.8	16	18.6
H. salinarium	29.1	24.3	54.2	61.9	16.7	13.7
H. volcanii	21.8	17.9	63	64.4	15.2	17.7
M. smegamtis	30.9	15	51.9	60.7	17.2	24.4
M. abscessus	29.6	32.5	43.4	53.2	27	14.3
S. aureus	28.3	28.6	45.7	42.9	26	28.6
P. aeruginosa	25.3	19.7	63.2	68.1	11.5	12.2
P. fluorescens	6	13.2	46.6	69.1	47.4	17.6
S. typhimurium	32.7	29.3	33.7	31.5	33.6	39.2
C. ljungdahlii	21.5	26.3	44.9	36.8	33.7	36.8
A. woodi	23.2	20.4	38.5	29	38.4	50.6
S. coelicolor	22.6	20	67.9	66.7	9.5	13.3
S. venezuelae	19.7	0	72.4	100	7.9	0
S. tsukubensis	27.1	17.7	63.5	73.5	9.4	8.8
S. griseus	22.7	13	68.5	79.5	8.8	7.5
S. clavuligerus	21.7	17.8	69.3	70.8	9	11.4
E. coli BW25513	26.2	23.7	38.4	37	35.5	39.3
E. coli MC4100	26.1	17.3	38.3	43.8	35.6	38.9
E. coli MG1655	26.4	23.9	38.3	38.4	35.3	37.8
Synechocystis spp.	23.3	22.7	39.3	36.4	37.4	40.9
S. pneumoniae	26.2	18.7	38.6	32	35.1	49.4
E. liosum	22.4	15.1	40.8	32.5	36.8	52.4
F. johnsoniae	27	20	39	40	34	40

Table S8: Location information for eORFs of interest based on re-occurrence analysis. Also provided are corresponding mother gene information, their frame relation, the assigned eORF name, length [nt] and the genome ID they were identified in.

		eORF info			**mother gene info**				
Genome	**Name**	**Position 1**	**Position 2**	**Strand**	**Position 1**	**Position 2**	**Strand**	**eORF length [nt]**	**Localisation relation**
NC_011916.1	Caulobacter1	486162	486371	+	486159	487127	-	209	sas13
NC_011916.1	Caulobacter2	3294604	3294696	+	3294602	3296371	-	92	sas11
NC_002607.1	Halobacterium	1060409	1060639	+	1060402	1060698	-	230	sas12

NC_013967.1	Haloferax	1904214	1904441	+	1903778	1905196	-	227	sas12
NC_008596.1	Mycobacterium1	4054253	4054501	-	4053599	4054810	+	248	sas13
NC_008596.1	Mycobacterium2	5456131	5456256	+	5456127	5456711	-	125	sas12
NC_008596.1	Mycobacterium3	6147012	6147326	-	6146474	6147355	+	314	sas11
NC_008596.1	Mycobacterium4	6248395	6248556	-	6247672	6250494	+	161	sas13
NC_008596.1	Mycobacterium5	6417181	6417426	-	6416588	6417490	+	245	sas12
NC_010397.1	Mycobacterium6	236032	236127	-	235976	236137	+	95	sas12
NC_010397.1	Mycobacterium7	362749	362841	-	362594	363481	+	92	sas12
NC_010397.1	Mycobacterium8	436440	436646	+	436073	437515	-	206	sas12
NC_016830.1	Pseudomonas1	2527588	2528136	+	2527546	2528220	-	548	sas13
NC_002516.2	Pseudomonas2	897654	897875	+	897335	898423	-	221	sas12
NC_003197.2	Salmonella1	2096306	2096539	+	2096293	2097222	-	233	sas11
NC_003197.2	Salmonella2	2165530	2165721	+	2165290	2166240	-	191	sas13
NC_007795.1	Staphylococcus1	710601	710693	-	710153	712093	+	92	sas11
NC_007795.1	Staphylococcus2	1457446	1457586	+	1456670	1457623	-	140	sas11
NC_003888.3	Streptomyces1	3636411	3636536	-	3636329	3636847	+	125	sas11
NC_003888.3	Streptomyces2	3669722	3669895	+	3669717	3670676	-	173	sas11
NC_003888.3	Streptomyces3	6418707	6419006	-	6418515	6419081	+	299	sas13
NC_003888.3	Streptomyces4	7518673	7518804	-	7518030	7519466	+	131	sas11
NZ_CP016559.1	Streptomyces5	154160	154375	+	153172	156768	-	215	sas12
NZ_CP016559.1	Streptomyces6	864676	865218	+	864673	866106	-	542	sas13
NZ_CP016559.1	Streptomyces7	1424833	1425057	+	1424772	1427657	-	224	sas12
NZ_CP016559.1	Streptomyces8	2730822	2731880	+	2730801	2732177	-	1058	sas13
NZ_CP016559.1	Streptomyces9	2772254	2772346	-	2771679	2772449	+	92	sas12
NZ_CP016559.1	Streptomyces10	3713295	3713516	-	3712075	3713586	+	221	sas12
NZ_CP016559.1	Streptomyces11	3818543	3818716	+	3818520	3819389	-	173	sas11
NZ_CP016559.1	Streptomyces12	3945269	3945400	+	3945130	3945879	-	131	sas12
NZ_CP016559.1	Streptomyces13	4688749	4688907	+	4688082	4689164	-	158	sas12
NZ_CP016559.1	Streptomyces14	4819691	4820374	-	4819169	4820410	+	683	sas13
NC_010572.1	Streptomyces15	5043512	5043790	+	5043490	5044134	-	278	sas12
NC_010572.1	Streptomyces16	5911631	5911807	-	5910855	5911820	+	176	sas12
NZ_CP020700.1	Streptomyces18	2724669	2724899	+	2724335	2725168	-	230	sas12
NZ_CP020700.1	Streptomyces19	4152999	4153373	+	4152980	4153381	-	374	sas12
NC_000913.3	Escherichia1	279212	279502	+	279178	279681	-	290	sas12
NC_000913.3	Escherichia2	906421	906522	+	905740	906753	-	101	sas13
NC_000913.3	Escherichia3	1978552	1978842	+	1978518	1979021	-	290	sas12
NC_000913.3	Escherichia4	3583856	3584146	-	3583677	3584180	+	290	sas12
NZ_HG738867.1	Escherichia5	293575	293691	-	292234	293694	+	116	sas13
NZ_HG738867.1	Escherichia6	2983427	2983669	+	2983234	2984493	-	242	sas12
NZ_CP009273.1	Escherichia7	386795	386971	-	385707	387167	+	176	sas12
NZ_CP009273.1	Escherichia8	3791132	3791374	-	3790308	3791567	+	242	sas12

Table S9: List of concentration measurements during RIBO-Seq and RNA-Seq duplicate sample processing. Experimental steps are provided to assigned values during progressing processing.

Experimental step	**Sample I**	**Sample II**
RNA extraction of RNA samples	20883.3 ng/µl	986.7ng/µl
RNA extraction of RIBO samples	199.4 ng/µl	110.4 ng/µ
RIBO samples after size selection	52.2 ng/µl	21.4 ng/µl
RNA samples after DNA digestion	27.9 ng/µl	94 ng/µl
RNA samples after 16S control PCR	220 ng/µl	272.3 ng/µl
RIBO samples after rRNA depletion (Nanodrop)	5.6 ng/µl	5.4 ng/µl
RNA samples after rRNA depletion (Nanodrop)	7.6 ng/µl	15.9 ng/µl
RIBO samples after rRNA depletion (Qubit)	not detectable	not detectable
RNA samples after rRNA depletion (Qubit)	3.63 ng/µl	6.99 ng/µl
RIBO samples after phosphorylation	4.5 ng/µl	4.4 ng/µl
RNA samples after phosphorylation	2.6 ng/µl	6.9 ng/µl
Library Prep RIBO samples (Qubit)	1.57 ng/µl	2.93 ng/µl
Library Prep RNA samples (Qubit)	6.38 ng/µl	13.53 ng/µl

Table S10: TruSeq Small RNA Library Prep Kit -Set A (Illumina) specific Indices used to label the four samples prepared differently to ensure appropriate read assignment after sample multiplexing.

Sample	Index number	Index sequence
RIBO I	RPI10	TAGCTT
RIBO II	RPI11	GGCTAC
RNA I	RPI9	GATCAG
RNA II	RPI2	CGATGT

Table S11: Percentage calculation of reads present within samples assigned to their respective RNA category. The original table is obtained from (Glaub et al., 2020).

Sample	Genes	tRNA	rRNA	Sample	Genes	tRNA	rRNA	Sample	Genes	tRNA	rRNA
SRR1734437	29.04	23.35	45.42	SRR4023280	27.94	22.3	44.04	SRR1200739	20.49	35.33	41.84
SRR1734438	35.16	27.89	34.57	SRR1613263	81.76	0.41	14.64	SRR1200750	16.36	34.66	46.52
SRR1734439	52.05	16.59	29.42	SRR1613265	82.41	0.44	13.89	SRR1200751	18.35	39.41	41.13
SRR1734440	9.54	68.82	20.34	SRR1613266	77.73	0.39	18.84	SRR1583082	61.92	23.19	11.01
SRR1734441	42.12	17.63	37.69	SRR1613268	84.34	0.42	11.54	SRR1583083	53.8	23.75	17.38
SRR1734442	41.29	14.97	41.8	SRR1613269	35.3	0.18	63.2	SRR1583084	58.44	21.15	17.24
SRR1734443	36.6	18.05	43.01	SRR1613270	83.64	0.39	12.41	SRR4190324	33.84	12.72	50.49
SRR1734444	48.09	15.37	34.89	SRR1613272	84	0.42	12.01	SRR4190325	43.07	19.28	32.08
SRR4023274	59.17	16.6	22.73	SRR1613277	72.1	0.72	13.03	SRR4190326	28.46	15.25	51.71
SRR1613278	79.97	0.68	16.55	SRR1200730	25.13	33.65	38.45	SRR364363	51.95	7.93	36.79
SRR1613280	82.73	0.67	13.5	SRR1200731	18.09	29.53	50.38	SRR364364	45.11	18.2	32.11
SRR1613281	78.78	0.62	17.9	SRR1200738	25.84	18.61	52.64	SRR364365	55.6	8.37	30.36
SRR1613283	85.17	0.7	10.95	SRR2016457	42.69	10.83	40.66	SRR364366	48.46	19.03	27.52
SRR1613285	61.99	0.48	35.39	SRR2016465	29.75	1.47	66.06	SRR364367	54.2	9.2	33.84

SRR1613287	73.35	0.36	23.54	SRR364370	62.91	12.15	20.76	SRR364368	63.72	11.34	19.7
								SRR364369	38.67	5.81	53.95

Table S12: Comparison of eORF prediction efficiency (calculated as the median) to genome-specific size or GC-content within genus *Streptomyces*.

Species	**Median eORF amount**	**GC-content**	**Genome Size**
S. clavuligerus	140	72.7	6.88
S. coelicolor	6.5	72.1	8.67
S. griseus	21	72.2	8.55
S. tsukubensis	53.5	71.9	7.96
S. venezuelae	1	72.5	8.22

List of Figures

List of Tables

List of Tables

List of Scripts