FEMS Microbiology Reviews, 2023, **47**, 1–13
DOI: 10.1093/femsre/fuad063
Advance access publication date: 27 November 2023
Review Article

Exposing the small protein load of bacterial life

Laure Simoens, Igor Fijalkowski, Petra Van Damme

iRIP Unit, Laboratory of Microbiology, Department of Biochemistry and Microbiology, Ghent University, K. L. Ledeganckstraat 35, 9000 Ghent, Belgium
*Corresponding author. iRIP Unit, Laboratory of Microbiology, Department of Biochemistry and Microbiology, Ghent University, K. L. Ledeganckstraat 35, 9000 Ghent, Belgium. E-mail: Petra.Vandamme@UGent.be

Abstract

The ever-growing repertoire of genomic techniques continues to expand our understanding of the true diversity and richness of prokaryotic genomes. Riboproteogenomics laid the foundation for dynamic studies of previously overlooked genomic elements. Most strikingly, bacterial genomes were revealed to harbor robust repertoires of small open reading frames (sORFs) encoding a diverse and broadly expressed range of small proteins, or sORF-encoded polypeptides (SEPs). In recent years, continuous efforts led to great improvements in the annotation and characterization of such proteins, yet many challenges remain to fully comprehend the pervasive nature of small proteins and their impact on bacterial biology. In this work, we review the recent developments in the dynamic field of bacterial genome reannotation, catalog the important biological roles carried out by small proteins and identify challenges obstructing the way to full understanding of these elusive proteins.

Keywords: small ORF (sORF); small ORF-encoded polypeptide (SEP); riboproteogenomics; genome (re)annotation; bacterial pathogens; proteomics

Introduction

In the continuous effort to improve bacterial genome annotations, the development of ribosome profiling by next-generation sequencing (Ingolia et al. 2009), Ribo-seq in short, permitted the recent discovery of a plethora of small open reading frames (sORFs). Classified as open reading frames built of no more than 300 nucleotides (nt), these newly discovered genes potentially give rise to their encoded small proteins; referred to as sORF-encoded polypeptides (SEPs). By providing direct translation evidence of numerous sORFs (Ingolia et al. 2009, 2011, Guo et al. 2010), Ribo-seq no longer penalized these specific ORFs for their short lengths in the process of gene prediction and genome annotation. Currently, multiple Ribo-seq datasets have been published for model bacterial species like the Gram-negative *Escherichia coli* (Li et al. 2014, Hucker et al. 2017, VanOrsdel et al. 2018, Weaver et al. 2019) and the Gram-positive *Bacillus subtilis* (Li et al. 2012). Similar efforts were also reported for specific bacterial human pathogens including the model species *Salmonella enterica* subspecies *enterica* serovar Typhimurium (Baek et al. 2017, Ndah et al. 2017, Venturini et al. 2020) and more recently for *Streptococcus pneumoniae* (Laczkovich et al. 2022), *Mycobacterium tuberculosis* (Smith et al. 2022), *Staphylococcus aureus* (Bartholomaus et al. 2021), and *Campylobacter jejuni* (Froschauer et al. 2022). A comprehensive overview of available prokaryotic ribosome profiling studies has recently been compiled by Vazquez-Laslop et al. (2022) and ribosome profiles corresponding to (some of) these and other studies can be consulted via the online genome browser GWIPS-viz (Michel et al. 2014).

Since the aforementioned studies report on the discovery of novel, putative sORFs, these recent efforts all contributed to a now exhaustive list of hypothetical bacterial SEPs. In aid of gene annotation, Ribo-seq studies can provide evidence for the translation of *in silico*-predicted sORFs by demonstrating ribosome interaction of their transcripts, making their consideration in genome (re-)annotation efforts more straightforward. However, functional characterization has only been reported for a small portion of putative sORFs and their encoding SEPs, leaving an enormous world of the sORFeome uncharted. With documented bacterial SEP functions falling within diverse categories of basic and essential bacterial physiology (Wang et al. 2017, Araújo-Bazán et al. 2019, Xu et al. 2019, Yoshitani et al. 2019, Burby and Simmons 2020, Sweet et al. 2021) as well as infection biology (Olvera et al. 2019, Williams et al. 2019, Yadavalli et al. 2020, Sur et al. 2022), the need for more large-scale validation and functional characterization efforts is high. In this context, it is noteworthy that difficulties in biochemical detection and therefore validation of SEPs are known and have been extensively documented (Fijalkowski et al. 2021, 2022, Gray et al. 2021), but that also recent bacterial SEP validation studies fail to fully address many of the challenges in small protein detection (Fijalkowski et al. 2022), such as their proposed low expression or low stability (Stringer et al. 2022). Nonetheless, as detection of protein expression is a prerequisite for functional investigations, further improvement in SEP detection might turn out to be of great value in expanding our current understanding of bacterial (infection) biology. Fundamentally, because of the lack of standard workflows to go from computationally predicted sORFs to functionally annotated SEPs, a whole piece might be missing from the puzzle the bacterial life is known to be. While independent of mRNA, small proteins may be generated by proteolytic action or nonribosomal synthesis (i.e. nonribosomal peptides), the focus of this review is on the small protein load of bacterial life encoded by sORFs. More specifically, besides reporting on their cataloged

Received 9 May 2023; **revised** 10 November 2023; **accepted** 24 November 2023

biological functions, and following an elaborate discussion on genomic sORF discovery and classification, we highlight the technical SEP-inherent validation challenges that mark the path toward full bacterial SEP annotation, ranging from experimental expression validation by standard protein detection methods [i.e. mass spectrometry and (immuno)blotting] toward uncovering their biological roles.

Studying bacterial biology in the genomics era

Sequencing revolution demands annotation evolution

To date, 25 years after the first genome was sequenced, advances in sequencing techniques under the form of high-throughput, next-generation sequencing, and supporting bioinformatics have resulted in an exponential increase in the number of bacterial genomes available by significantly lowering both sequencing time and costs (Land et al. 2015, Loman and Pallen 2015, Dorado et al. 2021). In the meantime, genome sequencing technologies already reached the third revolution (van Dijk et al. 2018), which was initiated with the advent of high-resolution, single-molecule, long-read sequencing. Genome sequences serve as a starting point for a better understanding of bacterial functioning, evolution, and interaction with the environment. The ever-augmenting number of available gene sequences evoked a switch from forward to reverse genetics approaches (Fels et al. 2020). However, the massive accumulation of genome sequences entailed a new hurdle. Data on microbial genomes are generated faster than they can be manually processed to extract valuable information from, making automatization of annotation an urging challenge for the current revolution in sequencing technology.

Automatic genome annotation requires manual curation: a frustrating paradox

The applied principles in popular, prokaryotic annotation tools (e.g. RAST, Prokka, and PGAP; Dong et al. 2021) illustrate the central idea of automatic genome annotation, which is the search for homology with database-annotated genes, proteins or domains. Hence, annotation of newly sequenced genomes is strongly influenced by the available information in databases and consequently, is steered by the annotation principles that have been used thus far. Once introduced, annotation errors propagate in these databases (Danchin et al. 2018), illustrating why the "automatic" character of widely applied annotation pipelines is not that absolute, as human intervention or metadata is still heavily required to manually curate their performance (Dziurzynski et al. 2021). Lobb et al. (2020) demonstrated the incompleteness of bacterial genome annotations by determining the percentage of the average bacterial proteome that can be functionally annotated based on homology with database-annotated proteins and domains, indicating a range between 52% and 79%, depending on the annotation tool used. Because of these inherent shortcomings to automatic genome annotation, (ribo)proteogenomics (i.e. the combination of systematic complementary ribosome profiling, proteomics, and genomics for studying translational landscapes; Willems et al. 2022) has become an invaluable approach for gene annotation as it already links gene prediction to expression (Fijalkowska et al. 2020, 2022, Cao et al. 2021, Willems et al. 2022).

sORFs: the weak spots of automated bacterial genome annotation

The annotation bias is especially challenging when it comes to the detection of sORFs. These coding sequences of arbitrarily no more than 150–300 nt encode small proteins or SEPs, with the interpretation for "small" ranging from smaller than (or equal to) 50 (Hemm et al. 2010, Storz et al. 2014) to 100 amino acids (AA) (Andrews and Rothnagel 2014). SEPs distinguish themselves from canonical small peptides by the fact that their origin is translational and no proteolytic processing step is required to make them this small. In this way, the typical length cut-offs for gene prediction and annotation, chosen based on the strong original belief that genes should be of sufficient length to be functional, turned out to be too short-sighted thereby obstructing sORF identification. Within existing genome annotations, a vast majority of sORFs encode ribosomal SEPs, which are significantly more conserved than the sORFs that are currently being discovered in large numbers because of the introduction of riboproteogenomics for the reanalysis of bacterial genomes (Gray et al. 2021). One possible explanation for the general lower degree of conservation of such translated sORFs might be due to their relatively rapid *de novo* evolvement in bacterial genomes (Gray et al. 2021).

In line with previous ORF annotations, newly discovered sORFs can be classified as intergenic sORFs, upstream (overlapping) (regulatory) sORFs [u(o)sORFs], internal (out-of-frame) sORFs (intsORFs) or downstream (out-of-frame) (overlapping) (regulatory) sORFs [d(o)sORFs]; a classification based on the relation between the genome-orientation of the newly discovered sORF and existing gene annotations (Fig. 1A) (Gray et al. 2021, Mudge et al. 2022, Stringer et al. 2022). For bacterial genomes, the genomic positioning of annotations is especially informative in the case of polycistronic mRNAs, frequently encoding gene product(s) with a strong functional interplay. Further, existing (s)ORF annotations are also frequently updated (i.e. sORF reannotations) (Fig. 1B).

sORFs were generally overlooked until increasingly more SEPs were identified – rather by chance – across all domains of life as well as viruses (Finkel et al. 2018). Moreover, sORFs and their encoding SEPs turned out to be of considerable biological importance for the respective organisms, further strengthened by Lluch-Senar et al. (2015) who identified the genomic class of sORFs as being the most frequently essential one in the case of the genome-reduced bacterium *Mycoplasma pneumoniae* (Lluch-Senar et al. 2015). Bacterial SEPs are, among other functions, known to be involved in basic (essential) processes underlying bacterial functioning, including cell division [e.g. Blr (Karimova et al. 2012), MciZ (Araújo-Bazán et al. 2019)], transport of molecules [e.g. AcrZ (Hobbs et al. 2012), KdpF (Sweet et al. 2021)], and signal transduction [e.g. MgrB (Xu et al. 2019), SafA (Yoshitani et al. 2019)] and to act as chaperones [e.g. YrhB (Ahn et al. 2012), MntS (Martin et al. 2015)] (Fig. 2). The discovery of the unexpected coding potential of bacterial sRNAs – not surprisingly – took place through mining the *E. coli* genome (Wassarman et al. 2001, Hemm et al. 2010, Hemm et al. 2020). In this regard, the bacterial operon gene structure surely deserves some credit for the initial, unintended discovery of the functional potential of small proteins. In 1999, Gaβel et al. (1999) discovered, at that time, the smallest *E. coli* protein KdpF (29 AA) through extensive examination of the K^+-transporter complex encoding *KdpABC* operon, and copurified KdpF with the complex emphasizing the possibility for small proteins to take active roles in protein complexation and bacterial functioning (Gaßel et al. 1999).

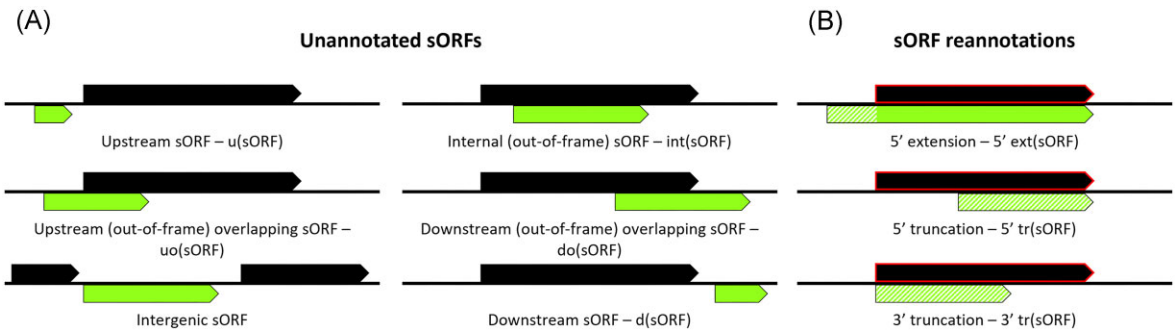

Figure 1. Categorization of unannotated sORFs and reannotations of sORFs discovered by riboproteogenomics. (A) The annotation of newly discovered sORFs is based on the relation between the genomic location of the novel sORF (green) and existing (s)ORF annotations (black). Especially for the typical bacterial polycistronic gene organization, positional ORF annotations in the context of transcripts are meaningful as the interaction of the resulting gene products can be regulatory in nature. (B) The implementation of riboproteogenomics for genome annotation can also result in the (conditional) reannotation of previously annotated ORFs. When ORFs appear as 3′ [3′ tr(sORF)] or 5′ [5′ tr(sORF)]-truncated variants, previous ORF annotations can turn into sORF annotations (green diagonal upward pattern fill), or sORFs into shorter sORF annotations. For sORFs, 5′ extensions result in N-terminally extended translation products [here 5′ ext(sORF)].

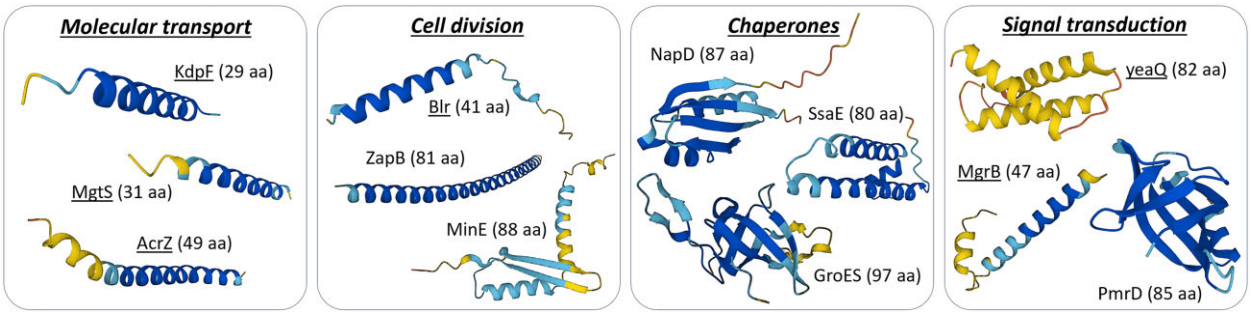

Figure 2. Examples of SEPs involved in basic (and essential) biological processes in the model species *S.* Typhimurium (strain SL1344). Some recurring functional categories can be distinguished among characterized bacterial SEPs, such as "Molecular transport" (e.g. KdpF, MgtS, and AcrZ), "Cell division" (e.g. Blr, ZapB, and MinE), "Signal transduction" (e.g. MgrB, YeaQ, and PmrD), and "Chaperones"(e.g. NapD, SsaE, and GroES). AlphaFold predicted structures are given for corresponding *S.* Typhimurium SEPs [(KdpF (A0A0H3P1C0), MgtS (A0A719A915), AcrZ (A0A0H3NEG2), Blr (A0A0H3NXY9), ZapB (A0A718Z7K1), MinE (A0A0H3NHJ8), NapD (A0A718RT46), SsaE (A0A0H3NKV4), GroES (A0A0H3NPH0), yeaQ (A0A0H3NKG6), MgrB (A0A718YD36), and PmrD (A0A0H3NDT1)]. Transmembrane helix-containing SEPs are underlined. aa = amino acids.

SEPs as a novel research hotspot for the study of bacterial biology

SEPs as accomplices in bacterial (infection) biology

With existing examples of SEPs acting as virulence factors and toxins (Fozo et al. 2008, Andresen et al. 2020, Wang et al. 2021), functional investigations of SEPs belonging to the proteome of pathogenic bacterial species paved the way for research endeavors focussing on SEPs as potential novel therapeutic targets. For *Listeria monocytogenes*, N-terminal proteomics linked the small membrane protein Prli42 (31 AA) to survival in macrophages by tethering the stressosome component RsbR – hypothesized to act as the sensor of stress signals – to the membrane (Impens et al. 2017, Williams et al. 2019). Another human pathogen, *S. aureus*, expresses the membrane peptide toxin PepA1 (31 AA) that has been postulated to be implicated in the regulation of survival after internalization into immune cells by driving selection of the slowly dividing bacterial population through the lysis of rapidly dividing bacteria (Schuster and Bertram 2016). Moreover, the 59 AA SEP Yp1 encoded by the bubonic plague pathogen *Yersinia pestis* was found to regulate virulence through expression modulation of type III secretion system components. In the same study, conservation of SEPs was shown to be high between different pathogenic *Yersinia* strains, so SEPs can be expected to confer pathogenesis-related benefits (Cao et al. 2021). Intriguingly, numerous novel SEPs are predicted transmembrane proteins (VanOrsdel et al. 2018, Miravet-Verde et al. 2019, Laczkovich et al. 2022). The increased prevalence of transmembrane localization of SEPs (Garai and Blanc-Potard 2020, Yadavalli and Yuan 2022) suggests a link with their involvement in host–pathogen interactions as these are initiated upon contact between pathogen and host membranes (Cao et al. 2021).

The SEP arsenal of *Salmonella*

Salmonella Typhimurium has served as a model species for the study of SEP expression in general and within the framework of bacteria–host interactions (Baek et al. 2017, Giess et al. 2017, Ndah et al. 2017, Venturini et al. 2020, Fijalkowski et al. 2022). The divergence of the genus *Salmonella* from *E. coli*, estimated to have taken place 160–120 million years ago (Ochman et al. 1999), was established through multiple horizontal gene transfer events, which left their marks in the genome in the shape of *Salmonella* pathogenicity islands and gave the genus the capacity to develop into the successful pathogen as we know it today. In addition to local gastroenteritis, *S.* Typhimurium leads to systemic disease in mice reminiscent of typhoid and paratyphoid fever caused

exclusively in humans by *S. enterica* subspecies *enterica* serovar Typhi and *S. enterica* subspecies *enterica* serovar Paratyphi, respectively. As such, *S.* Typhimurium infection in mice has been extensively used as a model system mimicking human typhoid fever (Chaudhuri et al. 2018).

For *Salmonella*, some of the few functionally characterized SEPs represent convincing links with virulence. As a first *Salmonella* SEP study case, MgtR (30 AA) was shown to be indirectly involved in intramacrophage survival by functioning as a regulator of the degradation of the virulence protein MgtC (Olvera et al. 2019). Further, the two cold shock proteins CspC (69 AA) and CspE (70 AA); implicated in diverse processes like response to membrane stress, motility, and biofilm formation; were proven to relate to *Salmonella* pathogenicity after observing noninfectiveness for double mutant *S.* Typhimurium (Michaux et al. 2017). For MgrB (47 AA), it has been shown that it binds and thereby inhibits the PhoQ kinase (Yadavalli et al. 2020), which takes part in the PhoPQ two-component system involved in regulation of virulence. Recently, a direct link between MgrB and virulence has been demonstrated through the creation of an Δ*mgrB* mutant, which failed in infecting macrophages and epithelial cells (Venturini et al. 2020). What's more, Venturini et al. (2020) showed more than half of the SEPs identified in their study to be differentially expressed upon infection, which is evident in the case of the type III secretion system apparatus or injectisome protein members SsaS (88 AA) and SsaI (82 AA). Accordingly, a great deal of studies found the expression of a significant part of previously unannotated sORFs of bacterial pathogens to follow infection-relevant expression patterns (Baek et al. 2017, Ndah et al. 2017, Fijalkowski et al. 2022).

Ribo-seq: a game changer for genome (re)annotation

The intriguing bacterial SEP functions reported further highlight the need for new advances to "enrich" bacterial genome annotations for sORFs. Ribo-seq revolutionized the study of translation by deep sequencing of ribosome-protected mRNA fragments (Wang et al. 2020). Ribosomes cover ∼30 nt when bound onto mRNA, causing these "ribosome protected parts" to be resistant toward nuclease degradation (Fig. 3). Sequencing of these footprints gives clues on the whereabouts of ribosomes along translated mRNAs while additionally enabling to demarcate boundaries of translation, and thus delineation of translated ORFs (Fig. 3B). Recently, Ribo-seq was tailored toward identification of prokaryotic translation initiation sites by stalling initiating ribosomes through the action of the pleuromutilin antibiotic retapamulin (Ribo-RET) (Fig. 3A) or alternatively, the newer pleuromutilin lefamulin, as especially in Gram-negative bacteria lefamulin was shown to exceed retapamulin activity (Weaver et al. 2019, Vazquez-Laslop et al. 2022). More recently, Ribo-seq protocols were developed to search the genome for ribosomal activity at stop codons (Fig. 3C) using the terminating ribosome bound release factor sequestrator apidaecin (Ribo-API) in combination with puromycin, a protein synthesis inhibitor causing premature chain termination during translation to remedy the obstacles of stop codon read through and ribosome queuing inherent to the use of apidaecin (Froschauer et al. 2022, Stringer et al. 2022, Vazquez-Laslop et al. 2022).

The more precise delineation of translation initiation sites further enabled the discovery of overlapping (u)(s)ORFs besides the discovery of (s)ORFs translated as distinct protein isoforms or N-terminal proteoforms (Fig. 4) (Weaver et al. 2019, Fijalkowska et al. 2020), features that challenge standard annotation algorithms (Pavesi et al. 2018, Wright et al. 2022) and that are widespread among sORFs. Ribo-RET together with Ribo-seq data is at the heart of (conditional) gene reannotation (Fig. 1B) (Ndah et al. 2017, Willems et al. 2020, Fijalkowski et al. 2022). Besides revealing differential expression, conditional Ribo-seq and -RET profiles (e.g. when comparing diverse bacterial growth conditions) can further disclose the existence of (conditional) gene extensions and truncations by showing differential Ribo-seq coverage patterns (3′ truncations and 5′ extensions) or alternative translations starts (5′ truncations and 5′ extensions) across the tested conditions (Fig. 4).

Since ribosomal protection does not necessarily point to translation, combining Ribo-RET with Ribo-API data (Fig. 3D) may prove valuable for the precise delineation of truly translated ORFs (Stringer et al. 2022), while additionally enabling the discovery of translational particularities such as ribosomal frameshifting events and internal sORFs [e.g. int(sORF) in *E. coli sfsA* (Meydan et al. 2019)] (Fig. 1A). Ribo-seq data, in turn, can fuel *de novo* machine learning algorithms like ribosome profiling assisted (re)annotation (REPARATION) (Ndah et al. 2017) and the modular algorithm smORFer (Bartholomaus et al. 2021) for the delineation of translated prokaryotic ORFs. In particular for sORFs, that are so difficult to find in genomes through standard annotation tools, Ribo-seq has thus been proven instrumental in uncovering their translation potential (Baek et al. 2017, Ndah et al. 2017, Fijalkowski et al. 2022, Laczkovich et al. 2022, Vazquez-Laslop et al. 2022). SmProt offers a dedicated platform for the structured database storage of SEPs from diverse model organisms, including *E. coli* SEPs, which have been experimentally or computationally identified (by Ribo-seq) (Olexiouk et al. 2018).

SEPs: the thorns in the eye of standard protein detection methods

Empirical SEP discovery is hindered by biochemical peculiarities

From a biochemical perspective, SEPs are inherently more difficult to study than average-sized proteins, a statement also applying to proteins significantly larger than average. Traditional two-dimensional gel electrophoresis (2DE), a technique profiting from the charge- and molecular weight (MW)-based separation of proteins respectively by isoelectric focusing and sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), fails in detecting proteins of extreme sizes, at both ends of the spectrum (Meleady 2018, Lee et al. 2020), while the category of larger proteins in regular gel-free shotgun proteomics is rather overrepresented by the theoretical increased peptide coverage. The relatively higher hydrophobicity indices in the class of recently discovered SEPs (Garai and Blanc-Potard 2020, Fijalkowski et al. 2022, Yadavalli and Yuan 2022) on the other hand, offers yet another explanation for their absence on 2DE gels (Meleady 2018, Lee et al. 2020, Kielkopf et al. 2021) and proteomics datasets in general. The fact that 9 out of the 10 most hydrophobic *Salmonella* proteins fall within the category of SEPs definitely supports the previous observations (Fijalkowski et al. 2022).

With respect to mass spectrometry (MS), the short size of SEPs heavily constraints the number of peptides produced after trypsin digestion (Fijalkowski et al. 2022), as this scarcity of peptides is outnumbered by the peptides originating from larger non-SEPs in the complete pool of tryptic peptides. While SEPs with lengths shorter than 100 AA are accountable for 10% of database-annotated *S.* Typhimurium proteins, the theoretically identifiable tryptic peptides originating from SEPs merely compose 2.5% of the totality of identified tryptic peptides in *Salmonella* (Fijalkowski

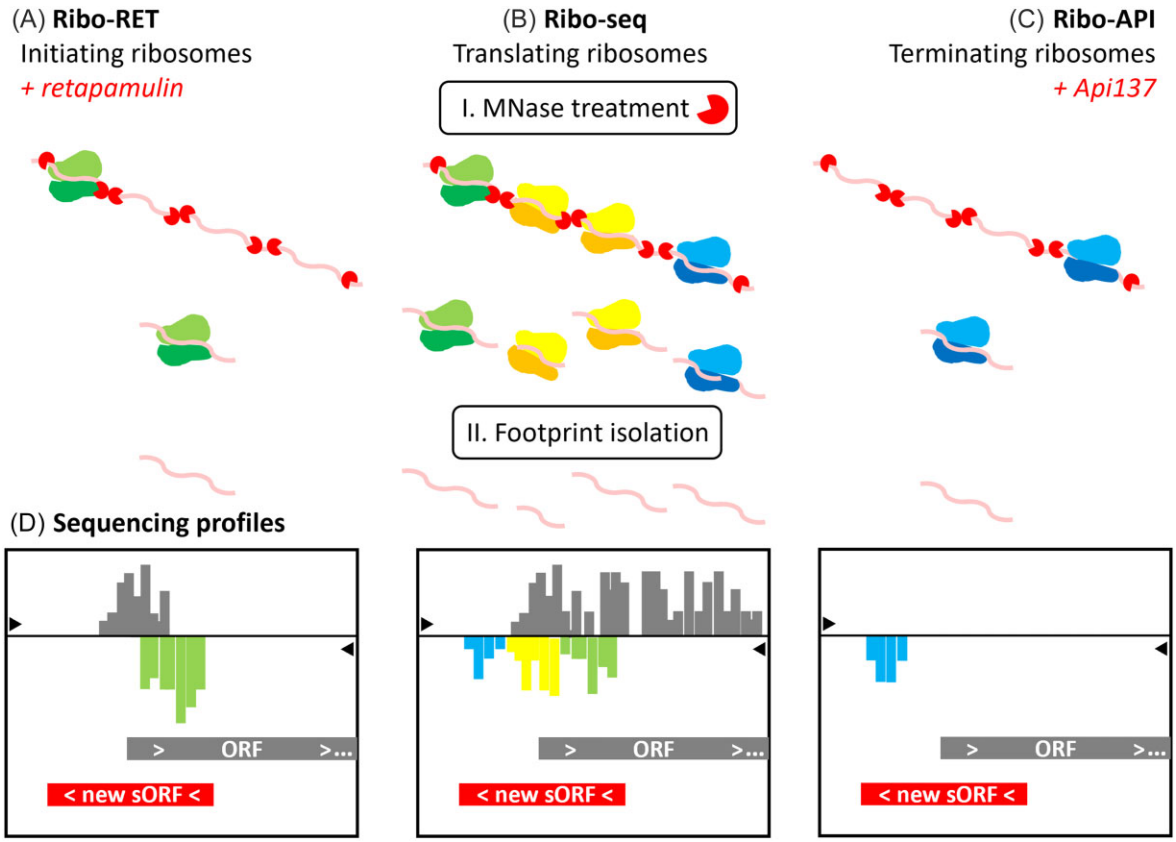

Figure 3. Ribo-seq toolset for the discovery of translated sORFs in bacterial genomes. In general, ribosome profiling by sequencing (Ribo-seq) (B) relies on the binding of ribosomes onto translated mRNA molecules as evidence for their translation into proteins (Ingolia et al. 2009). When performing bacterial Ribo-seq, micrococcal nuclease (MNase) is added to (ribosome-bound) mRNA. Nuclease digestion by MNase proceeds when no ribosomes are bound onto the mRNA molecule. Deep sequencing of isolated, ribosome-protected intact mRNA fragments enables delineation of translated genomic regions. Retapamulin-assisted Ribo-seq (Ribo-RET) (A) (Weaver et al. 2019) and apidaecin-assisted Ribo-seq (Ribo-API) (C) (Froschauer et al. 2022) are variants of the standard Ribo-seq protocol, which make use of the antibiotic retapamulin and the antimicrobial peptide apidaecin (Api137) for the specific halting of initiating (green) and terminating (blue) ribosomes by and large eliminating the signal from elongating (yellow) ribosomes, thereby allowing for the more accurate assignment of translation initiation and termination sites, respectively. Recently, the pleuromutilin lefamulin has been introduced as an alternative to retapamulin with higher activity in Gram-negative bacteria, and therefore likely more general and wider applicability (Vazquez-Laslop et al. 2022). Combining Ribo-seq deep sequencing patterns with Ribo-RET and Ribo-API-derived profiles can be used to more precisely delineate the start and stop codons of newly discovered (small) ORFs. (D) Ribo-seq traces corresponding to a hypothetical unannotated sORF (red annotation) obtained through Ribo-RET (green), Ribo-seq (yellow), and Ribo-API (blue) and with the profiles for a 5′ end of a hypothetical annotated ORF (gray annotation) indicated in gray. "…" indicates partial coding sequence annotations.

et al. 2022). The significant technical limitations that SEPs bring along for empirical protein discovery methods like 2DE and MS, are further accountable for the long-time ignorance towards SEPs, again explaining the under-representation of sORFs in genome annotations.

Experimental SEP validation suffers the same flaws

Independent from their empirical discovery, the aforementioned SEP-specific peculiarities have also hindered experimental validation, either through MS- or blotting-based detection, and even Ribo-seq-based sORF predictions. Computational analysis of riboproteogenomics data on putative translated sORFs and identified SEPs provided insights into the intrinsic MS-detectability of SEPs with specific attention to correlations with SEP size, abundance, stability, and hydrophobicity (Fijalkowski et al. 2022). Based on an *S.* Typhimurium dataset of complementary translatomics (Ribo-seq and -RET) and shotgun proteomics data (Willems et al. 2022), AP3 – an algorithm designed for the prediction of MS-detectability of theoretical peptides (Gao et al. 2019) – was implemented in an attempt to explain the obvious discrepancy between these two experimental omics datasets, i.e. the hits of the proteomic pipeline only covered 65% of the translated proteome identified by Ribo-seq. The trend observed showed a clear correlation between the number of theoretically detectable peptides and the length of the protein from which the theoretical peptide descends, a conclusion logically linking SEP detection difficulties to size.

Protocols aiming at high-molecular-weight protein depletion (Cassidy et al. 2019) or low-molecular-weight protein enrichment (Fijalkowski et al. 2021) have been proposed to increase peptide identification rate and coverage of the limited theoretically identifiable peptide arsenal originating from SEPs (Becher et al. 2020), but these technologies forego the quantitative aspect of proteomics data, and are therefore not generally applicable. Also, the use of a more diverse set of MS-sequencing proteases for proteome digestion might benefit SEP identification through increased sequence coverage as was done during a recent proteogenomic study of the *Y. pestis* genome (Cao et al. 2021). For example, for *S.* Typhimurium, choosing chymotrypsin over the standardly used trypsin, 30% more SEPs could in theory be picked up through shotgun proteomics (Fijalkowski et al. 2022). Do's and don'ts for MS-based small protein discovery were comprehensively reviewed by Ahrens et al. (2022).

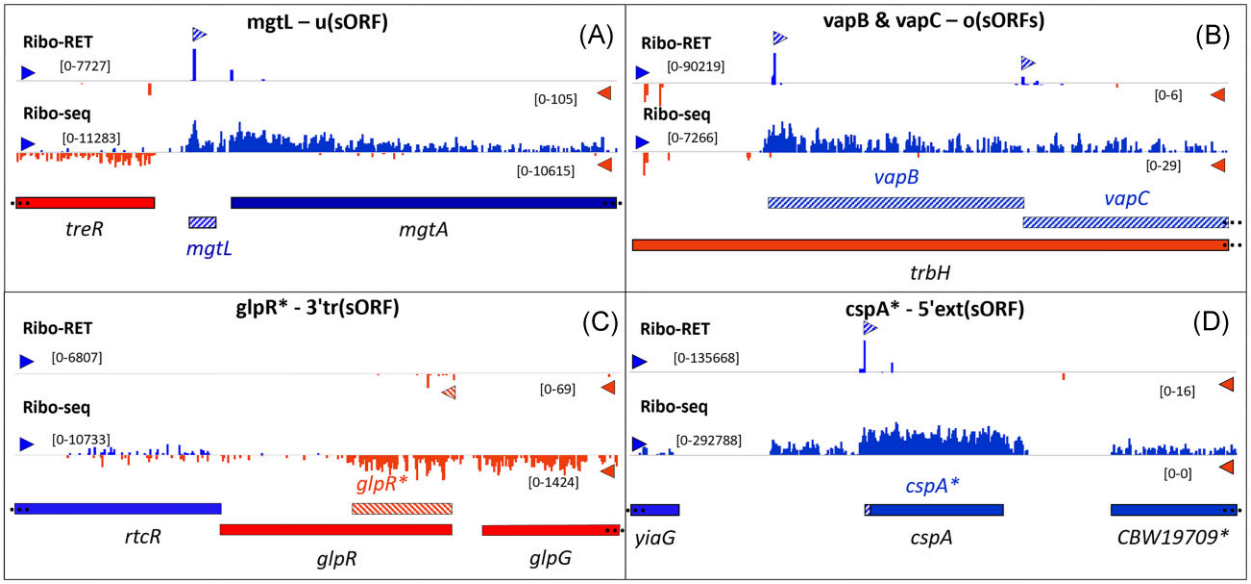

Figure 4. Riboproteogenomics-supported novel and reannotated *Salmonella* sORFs. Ribo-seq/Ribo-RET profiles of S. Typhimurium (strain SL1344) are shown. (A) *mgtA* regulatory leader peptide MgtL was delineated as a new sORF in the SL1344 genome by its matching annotations in related genomes (Ndah et al. 2017). Moreover, matching peptide evidence is available for MgtL (Ndah et al. 2017). (B) VapB and VapC, SEPs encoded by an upstream and downstream osORF, respectively, take part in the plasmid-encoded Vap toxin–antitoxin system and were only recently annotated in the SL1344 genome (Willems et al. 2020). (C) For the pseudogene *glpR*, a 3′-truncated version of GlpR (GlpR*) was predicted with the same start site (Willems et al. 2020). (D) *cspA* was found to have an in-frame upstream alternative start encoding a 5′-extended proteoform (CspA*), supported by peptide evidence (Ndah et al. 2017). * is indicative of an ORF reannotation. Sense and antisense encoded (s)ORFs are in blue and red color, respectively, with a diagonal upward pattern fill in case of newly discovered sORFs. "…" indicates partial coding sequence annotations.

Unlike the proteome-wide characteristic of MS, immunoblotting is a common go-to for the detection and quantification of epitope-tagged proteins, but here, the initial fractionation of the proteins by means of 1D SDS-PAGE, is already troublesome when dealing with SEPs (Kielkopf et al. 2021) as mentioned for 2DE (Meleady 2018, Lee et al. 2020). On top, the ensuing blotting step is also problematic as the small size of the SEPs permits them to more easily move through the blotting membrane, a phenomenon known as membrane blow-through (Kurien and Hal Scofield 2015). Moreover, the associated incubation steps for the purpose of immunodetection, which often take place under shaking conditions in voluminous incubation and washing solutions, make the conventional immunoblotting procedure far less favorable for SEP detection. Low MW proteins distinguish themselves under these conditions by the ability to easily detach from the membrane and to get lost in the discarded incubation fluids (Tomisawa et al. 2013).

All considered, SEPs can generally blame their small sizes for giving gene prediction, expression and validation analyses a hard time. Some voices, however, state that sORFs specifically come with low expression levels translating into a low SEP abundance (Olexiouk et al. 2016, Baek et al. 2017, Miravet-Verde et al. 2019, Peeters and Menschaert 2020). On top of that, SEPs are postulated to be highly unstable because of rapid (conditional-dependent) SEP degradation (Baek et al. 2017, Smith et al. 2022, Stringer et al. 2022), a feature that again links to low protein abundances, which in turn might explain why these proteins are often missed by standard protein detection methods. Experiments exploring the positive effect of the ClpP protease inhibitor bortezomib on blotting-based SEP expression validation attempts seemed to corroborate the SEP instability assumption as inhibitor usage allowed the blotting-based validation of three additional SEPs under study (Stringer et al. 2022). A more critical view of the presented data, however, should acknowledge a more general instead of a SEP-specific protein stabilization of the compound (Fijalkowski et al. 2022). Nonetheless, conformational studies of bacterial and archaeal SEPs by means of NMR spectroscopy may be in support of this decreased stability assumption as the results suggested the majority of studied SEPs to go through life without a well-defined structure (Kubatova et al. 2020) and intrinsically disordered proteins have been experimentally linked to higher proteolytic degradation susceptibility (Uversky 2017).

Contrastingly, based on bioinformatics predictions, Kubatova et al. (2020) reported that folding of the SEPs might require complexation and so not all these apparently unstructured SEPs are intrinsically disordered under physiologically relevant conditions. The fact that many SEPs have been found to engage in larger cytosolic and membrane protein complexes is further supportive of this (Storz et al. 2014, Venturini et al. 2020). New light on this discussion was moreover shed by the use of a multivariate logistic regression model for the prediction of SEP MS-detection probability including the number and detectability (AP3 score) of SEP tryptic peptides and translational abundance (Ribo-seq expression values in RPKM), stability (instability index), and hydrophobicity of the SEPs. Here, 75% of the variation in the model was explained by the scarcity of (unique) tryptic peptides and poor peptide detectability as the major factors limiting MS-detectability of SEPs (Fijalkowski et al. 2022), at least pointing toward stability not being the main driver when considering detection of the SEPs under study.

Current trends in dealing with sORFs and their encoded SEPs

State-of-the-art in the genomic discovery of sORFs

Currently, Ribo-seq is considered the most comprehensive method to scan genomes for expressed sORFs (Table 1). Ribo-seq offers a plethora of advantages over all other techniques that have

Table 1. Existing toolkit for sORF prediction/annotation, SEP expression validation, and functional characterization. MW; molecular weight, MS; mass spectrometry.

	State-of-the-art	Pros	Cons	Suggested improvements
sORF prediction and annotation	Ribosome profiling by sequencing (Ribo-seq)	• Genome-wide • Independent from existing annotations • Indicative of ribosomal activity • Broadly applicable • Improved resolution for the detection of start (retapamulin-assisted Ribo-seq; Ribo-RET) and stop codons (apidaecin-assisted Ribo-seq; Ribo-API)	• Requirement for experimental SEP validation • Computationally intensive and complex • Poor data resolution inherent to bacterial Ribo-seq	Refinement of bacterial Ribo-seq protocols and data analysis
SEP expression validation	MS	• SEP abundance data • Proteome-wide technique suited for empirical SEP discovery	• Limited number of tryptic SEP peptides • Hydrophobic character of peptides • Sensitivity of detection	• Use of alternative proteases (e.g. chymotrypsin) • High-MW protein depletion or low-MW protein enrichment
	(Immuno)blotting	• Information on MW and thus SEP integrity • Quantification of SEP expression	• Tag interference on SEP function/localization/extraction • Small SEP size • Sensitivity not adequate for low SEP abundances	• Use of smaller, charge-neutral tags (e.g. HiBiT; Schwinn et al. 2018) • SEP-specific customization [e.g. blotting membrane (type, pore size), blotting buffer and method] • In solution detection (e.g. HiBiT lytic detection; Schwinn et al. 2018)
SEP functional characterization	Conservation analysis	• First impression of SEP functioning • High-throughput screening	Lower conservation of sORFs and SEPs	• Interrogation of gene co-occurrence • RNA secondary structure analysis
	Domain prediction	First impression of SEP functioning and localization	Too short primary SEP sequences for domain prediction	Motif prediction (e.g. transmembrane motifs)
	Structural analysis	• Besides experimental (e.g. cryoelectron microscopy, nuclear magnetic resonance) also high-accuracy predictions are possible (e.g. AlphaFold) • High accuracy of AlphaFold proven for small proteins (Tejero et al. 2022)	The structure may depend on modifications and interactions	Macromolecular complexation info desirable; combine with interactomics data
	Mutation analysis Expression analysis Complexation analysis	Targeted and multiplex approach Gradient profiling by sequencing (Grad-seq) (coupled to MS) can distinguish between SEPs that engage in stable molecular interactions and SEPs that do not (Venturini et al. 2020)	Time-consuming Conditional impact of expression unknown The identity of the interaction partners remains to be elucidated	Conditional expression maps Proximity-dependent biotinylation interactomics approaches like ascorbate peroxidase-based proximity labeling (APEX; Santin et al. 2018) and proximity-dependent biotin identification (BioID; Herfurth et al. 2023) might be well-suited for the study of (transmembrane) bacterial SEPs

been exploited to improve genome annotations, with its main strengths being the genome-wide and high-throughput character complemented by its independence from existing annotations (Venturini et al. 2020). Moreover, the knowledge obtained on sORFs on behalf of Ribo-seq approaches created opportunities for the development of dedicated (s)ORF prediction algorithms (Yu et al. 2021), like REPARATION (Ndah et al. 2017). The wealth of data on putative translated sORFs, however, brings the strong need for experimental validation for Ribo-seq (and computationally) predicted sORFs (Fremin and Bhatt 2020) as resistance toward MNase degradation can be due to complex secondary structures of regulatory (small) noncoding ((s)nc)RNAs and translation of sORFs may serve a purely regulatory function as is the case for u(o)sORFs (Cabrera-Quio et al. 2016), meaning that ribosomal protection of mRNA per se cannot be considered as a proxy for the identification of functional SEPs. While viewing their generally less conserved nature, it is still not possible to distinguish intrinsically nontranslated from translated sORFs *in silico*, the inspection of the translational initiation context, dynamic transcript expression patterns in conjunction with Ribo-seq data can already be very informative in this regard (Cabrera-Quio et al. 2016). In contrast to Ribo-seq, however, only MS and alternative biochemical protein-based validation [e.g. (immuno)blotting] can truly affirm the existence of the according proteins as the ultimate products of the translation of a transcript and can therefore be used to filter out likely false positive SEP candidates (Fremin and Bhatt 2020).

Viewing the clear and noncircumventable irreconcilability between SEPs and MS- or (immuno)blotting-based expression validation (Table 1) (Tomisawa et al. 2013, Kurien and Hal Scofield 2015, Kielkopf et al. 2021, Fijalkowski et al. 2022), and because of the computational complications inherent to Ribo-seq data analysis (Gelhausen et al. 2022), it can be stated that comprehensive and robust sORF and SEP detection technologies are still lacking. This is especially so in the case of bacterial Ribo-seq data, because of the generally poorer resolution and resulting inadequacy of translation-specific features like its inherent triplet periodicity (Mohammad et al. 2019). It is well-established that prokaryotic ribosomal footprints display less-consistent lengths over their eukaryotic counterparts, a phenomenon attributed to both intrinsic properties of bacterial ribosomes and sequence specificity of employed nucleases. Clearly, both experimental and computational improvements of the technology are needed to fully address these challenges.

State-of-the-art in experimental SEP validation and functional SEP studies

Some of the molecular size-related detection obstacles for SEPs are conveniently circumvented by using protein tags or translational reporters (e.g. superfolder green fluorescent protein; Laczkovich et al. 2022), often with corresponding MWs making the translational fusions exceeding the class of SEPs. Based on the literature on bacterial SEP validation and characterization, the sequential peptide affinity (SPA) tag has been the go-to epitope tag for immunodetection of putative (bacterial) SEPs (Hemm et al. 2010, Baek et al. 2017, VanOrsdel et al. 2018, Weaver et al. 2019, Venturini et al. 2020, Froschauer et al. 2022). Studies resorting to this tag reported the epitope to permit the visualization of the expression of many reported SEPs. The tag combines the calmodulin-binding peptide (CBP) and three consecutive FLAG-tags (Hopp et al. 1988), separated by a tobacco etch virus (TEV) protease cleavage site, together accounting for an MW increase of about 6.3 kDa (Zeghouf et al. 2004). When dealing with SEPs, a tag of 6.3 kDa will often be as voluminous as – if not more voluminous than – the protein under study. However, while aiding detection, such a tag may alertly interfere with the physiological function/localization of the SEP (Vandemoortele et al. 2019). Nevertheless, epitope interference has also been reported for smaller peptide tags like the highly positively charged His-tag (Booth et al. 2018, Munadziroh et al. 2020), while contrarily, large (globular) tags were frequently shown to be innocuous (Vandemoortele et al. 2019).

The ultimate aim: functional characterization of validated bacterial SEPs

As only a few SEP validations corroborate genome-wide bacterial SEP discoveries, there is a need for more general, unbiased sORFeome-wide validation efforts. For example, an extensive study of the translational landscape of *S. pneumoniae* connected the SEP of only one of their newly discovered sORFs, *rio3*, to bacterial host colonization through targeted endogenous mutagenesis (Table 1) (Laczkovich et al. 2022), as also done in the pathogenic bacterium *Y. pestis* and in the extremophilic bacterium *Deinococcus radiodurans* for the functional characterization of SEP-yp1 and SEP-yp2 (Cao et al. 2021) and SEP068184 (Zhou et al. 2022), respectively implicating these SEPs in regulation of antiphagocytic capability and regulation of oxidative resistance.

Prior to functional studies, motif, domain, or structural prediction might provide a first hint toward the biological implication of the newly discovered SEPs (Table 1). Bioinformatics prediction of hydrophobic transmembrane motifs is relatively straightforward and widely exploited for the exploratory study of novel SEPs (Cao et al. 2021, Fijalkowski et al. 2022, Froschauer et al. 2022, Zhou et al. 2022). The short primary SEP structures are, however, no ideal subjects for functional domain searches (Weaver et al. 2019, Hemm et al. 2020, Stringer et al. 2022), which is explained by the average size of protein domains coinciding with the upper length threshold of SEPs (100 AA) (Xiong 2006). When contrasting bacterial domain annotations of SEPs versus non-SEPs for the SL1344 and LT2 *S.* Typhimurium strains as well as the K12 *E. coli* strain (Fig. 5), the domain annotation ranged from 7% to 18% for SEPs, while for non-SEPs these percentages varied from 28% to 62%. While remarkably big discrepancies in the percentages of domain annotations between these related species/strains could be observed, SEP versus non-SEP domain annotations were in each case shown to be 3- to 4-fold lower. Large-scale SEP studies reporting Pfam domain predictions for high-confidence, novel SEPs are in line with these lower percentages of domain annotation (Cao et al. 2021, Zhou et al. 2022).

Also, experimental structure determination has been explored to unravel SEP functionalities, for example through cryoelectron microscopy (Impens et al. 2017) or NMR (Kubatova et al. 2020). Importantly, in the case of small (rigid) proteins, the accuracies of structural predictions by AlphaFold were shown to equal those obtained through solution NMR (Tejero et al. 2022), making structure prediction an interesting alternative when domain prediction fails. Unfortunately, less conserved proteins – enriched in the class of newly discovered SEPs – are inherently prone to less good AlphaFold folding predictions due to poor multiple-sequence alignment, thus far hindering large-scale SEP structure prediction efforts. Conservation analysis (Hucker et al. 2017, Stringer et al. 2022) of (the genomic surrounding of) predicted sORFs (e.g. gene co-occurrence in case of polycistrons) might also help to prioritize functional conserved SEPs (Table 1) (Gray et al. 2021). For the genomic context of the putative sORF start codons, higher RNA

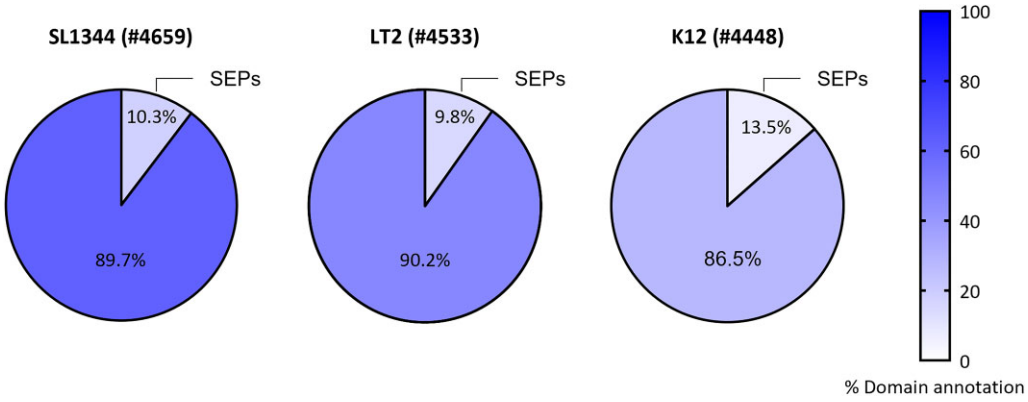

Figure 5. Domain annotations for bacterial SEPs compared to non-SEPs. Database protein entries of *S.* Typhimurium strains SL1344 (#4659, UP000008962) and LT2 (#4533, UP000001014), and *E. coli* K12 (#4448, UP000000625) were interrogated for domain annotations in the category of SEPs and non-SEPs with the % domain annotation color coded (blue scale). Categorical percentages of proteome occurrence for SEPs and non-SEPs are indicated in the corresponding pie charts.

secondary structure predictions (as compared to the start codon region itself) may further also serve as indicators of translation initiation (Stringer et al. 2022).

Specifically for sORF discovery in pathogenic bacteria, exploring putative SEP sequences for transmembrane domains and signal peptides is commonly exerted (Cao et al. 2021, Laczkovich et al. 2022) as cell contact and secreted molecules are the major interaction routes between bacterial pathogens and their host cells (Cao et al. 2021). These predictions are also used for the discovery of novel quorum-sensing systems and players in Gram-negatives, for which peptides are known to fulfil the role of quorum-sensing pheromones (Laczkovich et al. 2022). Cao et al. (2021) performed differential expression analysis of a set of putative sORFs interrogating different host environment-mimicking conditions. The same principle can be applied to all kinds of alternative stress conditions to find SEPs involved in different bacterial defense mechanisms and responses (Gray et al. 2021, Zhou et al. 2022), but also expression differences over standard growth conditions may suggest a biological impact (Hucker et al. 2017). A recent study on SEP profiling specifically focussed on the identification of stress response SEPs through the choice of the extremophilic bacterium *D. radiodurans*. Being a model organism for studying bacterial extreme stress responses, bacteria were subjected to ionizing radiation and oxidative stress resulting in the identification of 19 and 11 out of 109 newly identified SEPs as being upregulated under the respective stress condition (Zhou et al. 2022).

Discussion and conclusion

Ribo-seq-based mapping of bacterial translatomes has been drawing attention to sORFs, which still represent an underannotated class of genomic elements, as new and promising study subjects in the context of general microbiology as well as bacterial infection biology. Ribo-seq efforts should be appreciated for the contributions they made to the wealth of (putative) SEPs uncovered over the past years. However, experimental SEP validation is deemed required before including them in existing bacterial genome annotations, with epitope protein tagging and MS as the most commonly used methods. Besides the inherently small size and often hydrophobic nature of SEPs as confounding features hindering their detection, the – sometimes highly specific – expression conditions of SEPs might further complicate the validation process (Baek et al. 2017). Conditional gene expression programs of *Salmonella* are clearly illustrated in the continuous cataloging effort – the SalCom repository (Srikumar et al. 2015). Whether the common assumption of lower stability and abundance of SEPs serves as a general additional hindering factor, however, remains to be firmly established. Of note for SEP validation is that the rapid development of more sensitive high-throughput MS developments (e.g. data-independent acquisition and ion mobility MS) is expected to further aid bottom-up as well as top-down SEP detection (Kitata et al. 2022).

While nonetheless both go hand-in-hand, overcoming the obstacles of SEP validation is one thing, but framing these small proteins within the host's biology is another. Studies hunting for SEPs in diverse bacterial proteomes often focus on the functional investigation of one or few individual SEPs through targeted endogenous mutagenesis and no efforts to collectively address the functionality of the small proteome have been undertaken (Cao et al. 2021, Laczkovich et al. 2022, Zhou et al. 2022). There is, consequently, no doubt that the largest part of the sORFeome remains to be functionally explored. However, when discussing biologically meaningful SEPs, not too much emphasis must be placed on the word "function," a term used for proteins that contribute to cell fitness and that are under purifying selection (Keeling et al. 2019), as even merely the act of translating a sORF can influence the expression of its genomic context, like regulatory u(o)sORFs (e.g. the threonine operon uORF *thrL*) (Weaver et al. 2019, Hemm et al. 2020). These ORFs, which are located upstream of the coding sequence of a gene (Samal 2013) are often exclusively regulatory in nature (Cabrera-Quio et al. 2016) and only occasionally exert an encoding role. Defined based on their impact on gene regulation typically in an operon setting, leader peptides represent a special case of u(o)sORFs in case the translated peptide is shorter than 100 AA. Obtained MS-based peptide evidence of leader peptides, like for the *mgtA* leader MgtL (Fig. 4A), has changed the exclusively regulatory perception of this type of genomic element as stable protein levels appear to be produced (Ndah et al. 2017), thereby introducing leader peptides as a category within the class of SEPs.

With the functional knowledge at hand, and irrespective of ribosomal SEPs, bacterial SEPs can be concluded to be extensively engaged in protein complexation with an important fraction represented by (trans)membrane complexes (Hemm et al. 2020). Many of the known small proteins function through the binding and regulation of standard-sized proteins (Weaver et al. 2019) and according to gradient profiling by sequencing (Grad-seq) (coupled to MS), the majority of uncharacterized *Salmonella* SEPs engage in stable molecular interactions (Venturini et al. 2020) (Table 1). From this viewpoint, the current missing factor in bacterial SEP characterization is an interactomics-oriented approach. As

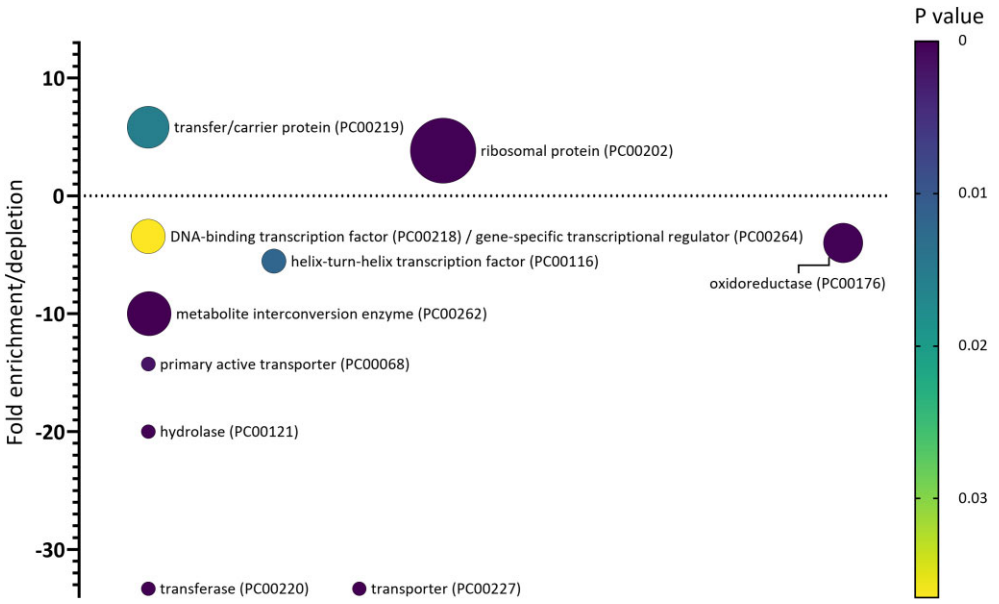

Figure 6. Multiple variable bubble plot representations of significantly over- or under-represented protein classes in the category of annotated *S.* Typhimurium SEPs. The over-representation test of SEP annotations – represented as fold enrichment/depletion – of *S.* Typhimurium strain LT2 (#4533, UP000001014) was determined using PANTHER (version 17.0, released 2022-02-22) and *P*-values were corrected for multiple hypotheses testing using Bonferroni correction for multiple testing. Only corrected *P*-values < .05 were considered. The bubble size corresponds to the number of proteins corresponding to the indicated PANTHER protein class (the smallest bubble size corresponds to seven members) and the color code represents the *P*-value scale.

protein–protein interactions might in particular play an important role for SEPs containing transmembrane domains, the underrepresentation of members of the e.g. (transmembrane) transporter protein class and domain annotations in general for UniProt SEP annotations (Fig. 5) might again indicate that the biological occupations of (transmembrane) SEPs rather goes through binding and regulation of other proteins or protein complexes, as also evident from the over-representation of transfer/carrier proteins among SEPs (Fig. 6), while the same analysis also revealed that unclassified proteins were about 2-fold overrepresented in the category of SEPs, providing an interesting niche for future functional SEP discoveries. Also, established *E. coli* SEP interactomes show the SEP players to be located on the periphery of complexes, suggesting SEPs take part in transient and differing interactions in multiple complexes, again providing evidence for SEPs as regulators (Hemm et al. 2020).

The interactome-associated SEP characteristics "hydrophobicity" and "transientness" eventually bring the concept of *in vivo* proximity-dependent biotinylation (PDB) to the forefront. Unlike affinity purification (coupled to MS), PDB approaches have the power to capture weak and transient protein–protein interactions and – equally important for the SEP protein class shown to be enriched for transmembrane proteins – are capable of handling less soluble proteins viewing its compatibility with the solubilization of membrane-(associated) proteins (Liu et al. 2020, Samavarchi-Tehrani et al. 2020). Ascorbate peroxidase (APEX)-based proximity labeling, a PDB method exploiting the enzymatic activity of peroxidases for the biotinylation of proteins (Samavarchi-Tehrani et al. 2020), was recently successfully applied in bacteria for the elucidation of the type VI secretion biogenesis process in *E. coli* (Santin et al. 2018). BioID, standing for proximity-dependent biotin identification, is another implementation of the PDB principle and requires the translational fusion of the protein of interest to a promiscuous biotin ligase for the biotinylation of proximal and interacting proteins and was very recently applied for interactome mapping in bacteria (Herfurth et al. 2023). As the overlap between APEX and BioID interactomes has been claimed to be limited, BioID is thus likely to offer interesting complementary perspectives into bacterial SEP biology (Samavarchi-Tehrani et al. 2020) (Table 1). In parallel with the discussion on the importance of the size of epitope tags used for expression analysis of SEPs, the potential harmful effect of translationally fusing small proteins to the relatively large enzymes APEX2 (MW = 27 kDa) or BirA∗ (MW = 35 kDa) on SEP localization and function needs to be borne in mind, but variant approaches derived from the original BioID protocol have been reported that may circumvent such problems [e.g. Off-the-shelf BioID (Santos-Barriopedro et al. 2021) and split-TurboID (Cho et al. 2020)].

High-throughput phenotypic screening is also emerging as an initiative to characterize gene products through the use of phenotypic microarrays interrogating the metabolization of compounds deterministic for unique molecular pathways (Guard 2022). The construction of knock-out libraries through CRISPR/Cas9 or alternative recombineering strategies could offer a valuable approach to enable sORF/SEP phenotyping at larger scales (Fels et al. 2020, Todor et al. 2021). Alternatively, profiting from the sORFs uncovered by Ribo-seq, existing transposon insertion sequencing (Tn-seq) datasets can be revisited for phenotyping and can even guide toward essential sORFs (Cain et al. 2020). In a broader context, large-scale annotation of sORFs in bacterial genome annotations permits re-exploration of available omics and other sequencing data (e.g. Chip-seq), with the aim to link features to novel sORFs (Myers et al. 2015). The combination of information from both existing and new omics data should enable small protein research to finally piece together the functionalities of bacterial SEPs encoded by newly discovered sORFs in support of full genomic ORF annotation.

Author contributions

L.S. and P.V.D. conceptualized the review. All authors contributed to the creation of illustrations. L.S. and P.V.D. wrote the manuscript

and all authors contributed to finalizing the manuscript text and gave approval to the final version of the manuscript.

Conflict of interest: The authors have declared that no competing interests exist.

Funding

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (PROPHECY grant agreement number 803972) and the Research Foundation Flanders (FWO-Vlaanderen) (project numbers G051120N and G045921N), all to P.V.D.

References

Ahn KY, Park JS, Han KY *et al.* YrhB is a highly stable small protein with unique chaperone-like activity in *Escherichia coli* BL21(DE3). *FEBS Lett* 2012;**586**:1044–8.

Ahrens CH, Wade JT, Champion MM *et al.* A practical guide to small protein discovery and characterization using mass spectrometry. *J Bacteriol* 2022;**204**. https://journals.asm.org/doi/10.1128/jb.00353-21.

Andresen L, Martínez-Burgo Y, Zangelin JN *et al.* The small toxic *Salmonella* protein TimP targets the cytoplasmic membrane and is repressed by the small RNA TimR. *mBio* 2020;**11**. https://doi.org/10.1128/mBio.01659-20.

Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* 2014;**15**:193–204.

Araújo-Bazán L, Huecas S, Valle J *et al.* Synthetic developmental regulator MciZ targets FtsZ across *Bacillus* species and inhibits bacterial division. *Mol Microbiol* 2019;**111**:965–80.

Baek J, Lee J, Yoon K *et al.* Identification of unannotated small genes in *Salmonella*. *G3 Genes Genomes Genet* 2017;**7**:983–9.

Bartholomaus A, Kolte B, Mustafayeva A *et al.* SmORFer: a modular algorithm to detect small ORFs in prokaryotes. *Nucleic Acids Res* 2021;**49**:e89–9. https://doi.org/10.1093/nar/gkab477.

Becher D, Bartel J, Varadarajan AR *et al.* Optimized proteomics workflow for the detection of small proteins. *J Proteome Res* 2020;**19**:4004–18.

Booth WT, Schlachter CR, Pote S *et al.* Impact of an N-terminal polyhistidine tag on protein thermal stability. *ACS Omega* 2018;**3**:760–8.

Burby PE, Simmons LA. Regulation of cell division in bacteria by monitoring genome integrity and DNA replication status. *J Bacteriol* 2020;**202**. https://doi.org/10.1128/JB.00408-19.

Cabrera-Quio LE, Herberg S, Pauli A. Decoding sORF translation—from small proteins to gene regulation. *RNA Biol* 2016;**13**:1051–9.

Cain AK, Barquist L, Goodman AL *et al.* A decade of advances in transposon-insertion sequencing. *Nat Rev Genet* 2020;**21**:526–40.

Cao S, Liu X, Huang Y *et al.* Proteogenomic discovery of sORF-encoded peptides associated with bacterial virulence in *Yersinia pestis*. *Commun Biol* 2021;**4**. https://doi.org/10.1038/s42003-021-02759-x.

Cassidy L, Kaulich PT, Tholey A. Depletion of high-molecular-mass proteins for the identification of small proteins and short open reading frame encoded peptides in cellular proteomes. *J Proteome Res* 2019;**18**:1725–34.

Chaudhuri D, Chowdhury AR, Biswas B *et al.* *Salmonella* Typhimurium infection leads to colonization of the mouse brain and is not completely cured with antibiotics. *Front Microbiol* 2018;**9**. https://doi.org/10.3389/fmicb.2018.01632.

Cho KF, Branon TC, Rajeev S *et al.* Split-TurboID enables contact-dependent proximity labeling in cells. *Proc Natl Acad Sci USA* 2020;**117**:12143–54.

Danchin A, Ouzounis C, Tokuyasu T *et al.* No wisdom in the crowd: genome annotation in the era of big data – current status and future prospects. *Microb Biotechnol* 2018;**11**:588–605.

Dong Y, Li C, Kim K *et al.* Genome annotation of disease-causing microorganisms. *Brief Bioinform* 2021;**22**:845–54.

Dorado G, Gálvez S, Rosales TE *et al.* Analyzing modern biomolecules: the revolution of nucleic-acid sequencing—review. *Biomolecules* 2021;**11**:1–18.

Dziurzynski M, Decewicz P, Ciuchcinski K *et al.* Simple, reliable, and time-efficient manual annotation of bacterial genomes with MAISEN. *Methods Mol Biol* 2021;**2242**:221–9.

Fels U, Gevaert K, Van Damme P. Bacterial genetic engineering by means of recombineering for reverse genetics. *Front Microbiol* 2020;**11**. https://doi.org/10.3389/fmicb.2020.548410.

Fijalkowska D, Fijalkowski I, Willems P *et al.* Bacterial riboproteogenomics: the era of N-terminal proteoform existence revealed. *FEMS Microbiol Rev* 2020;**44**:418–31.

Fijalkowski I, Peeters MKR, Van Damme P. Small protein enrichment improves proteomics detection of sORF encoded polypeptides. *Front Genet* 2021;**12**. https://doi.org/10.3389/fgene.2021.713400.

Fijalkowski I, Willems P, Jonckheere V *et al.* Hidden in plain sight: challenges in proteomics detection of small ORF-encoded polypeptides. *microLife* 2022;**3**. https://doi.org/10.1093/femsml/uqac005.

Finkel Y, Stern-Ginossar N, Schwartz M. Viral short ORFs and their possible functions. *Proteomics* 2018;**18**. https://doi.org/10.1002/pmic.201700255.

Fozo EM, Kawano M, Fontaine F *et al.* Repression of small toxic protein synthesis by the Sib and OhsC small RNAs. *Mol Microbiol* 2008;**70**:1076–93.

Fremin BJ, Bhatt AS. Structured RNA contaminants in bacterial riboseq. *mSphere* 2020;**5**. https://doi.org/10.1128/msphere.00855-20.

Froschauer K, Svensson SL, Gelhausen R *et al.* Complementary Riboseq approaches map the translatome and provide a small protein census in the foodborne pathogen *Campylobacter jejuni*. *Biorxiv* 2022, https://doi.org/10.1101/2022.11.09.515450.

Gao Z, Chang C, Yang J *et al.* AP3: an advanced proteotypic peptide predictor for targeted proteomics by incorporating peptide digestibility. *Anal Chem* 2019;**91**:8705–11.

Garai P, Blanc-Potard A. Uncovering small membrane proteins in pathogenic bacteria: regulatory functions and therapeutic potential. *Mol Microbiol* 2020;**114**:710–20.

Gaßel M, Möllenkamp T, Puppe W *et al.* The KdpF subunit is part of the K-translocating Kdp complex of *Escherichia coli* and is responsible for stabilization of the complex in vitro. *J Biol Chem* 1999;**274**:37901–7.

Gelhausen R, Müller T, Svensson SL *et al.* RiboReport - benchmarking tools for ribosome profiling-based identification of open reading frames in bacteria. *Brief Bioinform* 2022;**23**. https://doi.org/10.1093/bib/bbab549.

Giess A, Jonckheere V, Ndah E *et al.* Ribosome signatures aid bacterial translation initiation site identification. *BMC Biol* 2017;**15**. https://doi.org/10.1186/s12915-017-0416-0.

Gray T, Storz G, Papenfort K. Small proteins; big questions. *J Bacteriol* 2021;**204**. https://doi.org/10.1128/jb.00341-21.

Guard J. Through the looking glass: genome, phenome, and interactome of *Salmonella enterica*. *Pathogens* 2022;**11**:581. https://doi.org/10.3390/pathogens11050581.

Guo H, Ingolia NT, Weissman JS *et al.* Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 2010;**466**:835–40.

Hemm MR, Paul BJ, Miranda-Ríos J *et al.* Small stress response proteins in *Escherichia coli*: proteins missed by classical proteomic studies. *J Bacteriol* 2010;**192**:46–58.
Hemm MR, Weaver J, Storz G. *Escherichia coli* small proteome. *EcoSal Plus* 2020;**9**. https://doi.org/10.1128/ecosalplus.esp-0031-2019.
Herfurth M, Müller F, Søgaard-Andersen L *et al.* A miniTurbo-based proximity labeling protocol to identify conditional protein interactomes in vivo in *Myxococcus xanthus*. *STAR Protoc* 2023;**4**:102657. https://doi.org/10.1016/j.xpro.2023.102657.
Hobbs EC, Yin X, Paul BJ *et al.* Conserved small protein associates with the multidrug efflux pump AcrB and differentially affects antibiotic resistance. *Proc Natl Acad Sci USA* 2012;**109**:16696–701.
Hopp TP, Prickett KS, Price VL *et al.* A short polypeptide marker sequence useful for recombinant protein identification and purification. *Nat Biotechnol* 1988;**6**:1204–10.
Hucker SM, Ardern Z, Goldberg T *et al.* Discovery of numerous novel small genes in the intergenic regions of the *Escherichia coli* O157:H7 Sakai genome. *PLoS ONE* 2017;**12**:e0184119. https://doi.org/10.1371/journal.pone.0184119.
Impens F, Rolhion N, Radoshevich L *et al.* N-terminomics identifies Prli42 as a membrane miniprotein conserved in *Firmicutes* and critical for stressosome activation in *Listeria monocytogenes*. *Nat Microbiol* 2017;**2**. https://doi.org/10.1038/nmicrobiol.2017.5.
Ingolia NT, Ghaemmaghami S, Newman JRS *et al.* Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 2009;**324**:218–23.
Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 2011;**147**:789–802.
Karimova G, Davi M, Ladant D. The β-lactam resistance protein Blr, a small membrane polypeptide, is a component of the *Escherichia coli* cell division machinery. *J Bacteriol* 2012;**194**:5576–88.
Keeling DM, Garza P, Nartey CM *et al.* The meanings of "function" in biology and the problematic case of de novo gene emergence. *eLife* 2019;**8**. https://doi.org/10.7554/eLife.47014.001.
Kielkopf CL, Bauer W, Urbatsch IL. Sodium dodecyl sulfate-polyacrylamide gel electrophoresis of proteins. *Cold Spring Harb Protoc* 2021;**2021**:494–504.
Kitata RB, Yang JC, Chen YJ. Advances in data-independent acquisition mass spectrometry towards comprehensive digital proteome landscape. *Mass Spectrom Rev* 2022:**42**. https://doi.org/10.1002/mas.21781.
Kubatova N, Pyper DJ, Jonker HRA *et al.* Rapid biophysical characterization and NMR spectroscopy structural analysis of small proteins from bacteria and archaea. *ChemBioChem* 2020;**21**:1178–87.
Kurien BT, Hal Scofield R. Western blotting: an introduction. In: *Western Blotting: Methods and Protocols*. Vol. **1312**. Bethesda: NIH Public Access, 2015, 17–30.
Laczkovich I, Mangano K, Shao X *et al.* Discovery of unannotated small open reading frames in *Streptococcus pneumoniae* D39 involved in quorum sensing and virulence using ribosome profiling. *mBio* 2022;**13**. https://doi.org/10.1128/mbio.01247-22.
Land M, Hauser L, Jun SR *et al.* Insights from 20 years of bacterial genome sequencing. *Funct Integr Genom* 2015;**15**:141–61.
Lee PY, Saraygord-Afshari N, Low TY. The evolution of two-dimensional gel electrophoresis—from proteomics to emerging alternative applications. *J Chromatogr A* 2020;**1615**:460763. https://doi.org/10.1016/j.chroma.2019.460763.
Li GW, Burkhardt D, Gross C *et al.* Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* 2014;**157**:624–35.
Li GW, Oh E, Weissman JS. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 2012;**484**:538–41.
Liu X, Salokas K, Weldatsadik RG *et al.* Combined proximity labeling and affinity purification – mass spectrometry workflow for mapping and visualizing protein interaction networks. *Nat Protoc* 2020;**15**:3182–211.
Lluch-Senar M, Delgado J, Chen W *et al.* Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Mol Syst Biol* 2015;**11**. https://doi.org/10.15252/msb.20145558.
Lobb B, Tremblay BJM, Moreno-Hagelsieb G *et al.* An assessment of genome annotation coverage across the bacterial tree of life. *Microb Genom* 2020;**6**. https://doi.org/10.1099/mgen.0.000341.
Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. *Nat Rev Micro* 2015;**13**:787–94.
Martin JE, Waters LS, Storz G *et al.* The *Escherichia coli* small protein MntS and exporter MntP optimize the intracellular concentration of manganese. *PLoS Genet* 2015;**11**:1–31.
Meleady P. Two-dimensional gel electrophoresis and 2D-DIGE. *Methods Mol Biol* 2018;**1664**:3–14.
Meydan S, Marks J, Klepacki D *et al.* Retapamulin-assisted ribosome profiling reveals the alternative bacterial proteome. *Mol Cell* 2019;**74**:481–93.e6.
Michaux C, Holmqvist E, Vasicek E *et al.* RNA target profiles direct the discovery of virulence functions for the cold-shock proteins CspC and CspE. *Proc Natl Acad Sci USA* 2017;**114**:6824–9.
Michel AM, Fox G, Kiran AM *et al.* GWIPS-viz: development of a Ribo-seq genome browser. *Nucl Acids Res* 2014;**42**:D859–64.
Miravet-Verde S, Ferrar T, Espadas-García G *et al.* Unraveling the hidden universe of small proteins in bacterial genomes. *Mol Syst Biol* 2019;**15**:1–17.
Mohammad F, Green R, Buskirk AR. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *eLife* 2019;**8**. https://doi.org/10.7554 /eLife.42591.001.
Mudge JM, Ruiz-Orera J, Prensner JR *et al.* Standardized annotation of translated open reading frames. *Nat Biotechnol* 2022;**40**:994–9.
Munadziroh E, Ulfa EU, Labiqah A *et al.* Effect of poly-histidine tag position toward inhibition activity of secretory leukocyte protease inhibitor as candidate for material wound healing. *Avicenna J Med Biotechnol* 2020;**12**:32–6.
Myers KS, Park DM, Beauchene NA *et al.* Defining bacterial regulons using ChIP-seq. *Methods* 2015;**86**:80–8.
Ndah E, Jonckheere V, Giess A *et al.* REPARATION: ribosome profiling assisted (re-)annotation of bacterial genomes. *Nucleic Acids Res* 2017;**45**:e168–8. https://doi.org/10.1093/nar/gkx758.
Ochman H, Elwyn S, Moran NA. Calibrating bacterial evolution. *Proc Natl Acad Sci USA* 1999;**96**:12638–43.
Olexiouk V, Crappé J, Verbruggen S *et al.* sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res* 2016;**44**:D324–29.
Olexiouk V, Van Criekinge W, Menschaert G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res* 2018;**46**:D497–502.
Olvera MR, Garai P, Mongin G *et al.* Synthetic hydrophobic peptides derived from MgtR weaken *Salmonella* pathogenicity and work with a different mode of action than endogenously produced peptides. *Sci Rep* 2019;**9**:15253. https://doi.org/10.1038/s41598-019-51760-2.
Pavesi A, Vianelli A, Chirico N *et al.* Overlapping genes and the proteins they encode differ significantly in their sequence compo-

sition from non-overlapping genes. *PLoS ONE* 2018;**13**:e0202513. https://doi.org/10.1371/journal.pone.0202513.
Peeters MKR, Menschaert G. The hunt for sORFs: a multidisciplinary strategy. *Exp Cell Res* 2020;**391**:111923. https://doi.org/10.1016/j.yexcr.2020.111923.
Samal SK. Leader Sequence. *Brenner's Encyclopedia of Genetics*. 2nd edn. Amsterdam: Elsevier Inc., 2013, 203–5.
Samavarchi-Tehrani P, Samson R, Gingras AC. Proximity dependent biotinylation: Key enzymes and adaptation to proteomics approaches. *Mol Cell Proteom* 2020;**19**:757–73.
Santin YG, Doan T, Lebrun R *et al.* In vivo TssA proximity labelling during type VI secretion biogenesis reveals TagA as a protein that stops and holds the sheath. *Nat Microbiol* 2018;**3**:1304–13.
Santos-Barriopedro I, van Mierlo G, Vermeulen M. Off-the-shelf proximity biotinylation for interaction proteomics. *Nat Commun* 2021;**12**. https://doi.org/10.1038/s41467-021-25338-4.
Schuster CF, Bertram R. Toxin-antitoxin systems of *Staphylococcus aureus*. *Toxins* 2016;**8**:140. https://doi.org/10.3390/toxins8050140.
Schwinn MK, Machleidt T, Zimmerman K *et al.* CRISPR-mediated tagging of endogenous proteins with a luminescent peptide. *ACS Chem Biol* 2018;**13**:467–74.
Smith C, Canestrari JG, Wang AJ *et al.* Pervasive translation in *Mycobacterium tuberculosis*. *eLife* 2022;**11**. https://doi.org/10.7554/eLife.
Srikumar S, Kröger C, Hébrard M *et al.* RNA-seq brings new insights to the intra-macrophage transcriptome of *Salmonella* typhimurium. *PLoS Pathog* 2015;**11**:e1005262. https://doi.org/10.1371/journal.ppat.1005262.
Storz G, Wolf YI, Ramamurthi KS. Small proteins can no longer be ignored. *Annu Rev Biochem* 2014;**83**:753–77.
Stringer A, Smith C, Mangano K *et al.* Identification of novel translated small open reading frames in *Escherichia coli* using complementary ribosome profiling approaches. *J Bacteriol* 2022;**204**. https://doi.org/10.1128/JB.00294-21.
Sur VP, Simonik O, Novotna M *et al.* Dynamic study of small toxic hydrophobic proteins PepA1 and PepG1 of *Staphylococcus aureus*. *Int J Biol Macromol* 2022;**219**:1360–71. https://doi.org/10.1016/j.ijbiomac.2022.07.192.
Sweet ME, Larsen C, Zhang X *et al.* Structural basis for potassium transport in prokaryotes by KdpFABC. *Proc Natl Acad Sci USA* 2021;**118**. https://doi.org/10.1073/pnas.2105195118.
Tejero R, Huang YJ, Ramelot TA *et al.* AlphaFold models of small proteins rival the accuracy of solution NMR structures. *Front Mol Biosci* 2022;**9**. https://doi.org/10.3389/fmolb.2022.877000.
Todor H, Silvis MR, Osadnik H *et al.* Bacterial CRISPR screens for gene function. *Curr Opin Microbiol* 2021;**59**:102–9.
Tomisawa S, Abe C, Kamiya M *et al.* A new approach to detect small peptides clearly and sensitively by western blotting using a vacuum-assisted detection method. *Biophysics* 2013;**9**:79–83.
Uversky VN. Paradoxes and wonders of intrinsic disorder: stability of instability. *Intrinsic Disord Proteins* 2017;**5**:e1327757. https://doi.org/10.1080/21690707.2017.1327757.
Vandemoortele G, Eyckerman S, Gevaert K. Pick a Tag and explore the functions of your pet protein. *Trends Biotechnol* 2019;**37**:1078–90.
van Dijk EL, Jaszczyszyn Y, Naquin D *et al.* The third revolution in sequencing technology. *Trends Genet* 2018;**34**:666–81.
VanOrsdel CE, Kelly JP, Burke BN *et al.* Identifying new small proteins in *Escherichia coli*. *Proteomics* 2018;**18**. https://doi.org/10.1002/pmic.201700064.
Vazquez-Laslop N, Sharma CM, Mankin A *et al.* Identifying small open reading frames in prokaryotes with ribosome profiling. *J Bacteriol* 2022;**204**. https://doi.org/10.1128/JB.00294-21.
Venturini E, Svensson SL, Maaß S *et al.* A global data-driven census of *Salmonella* small proteins and their potential functions in bacterial virulence. *microLife* 2020;**1**. https://doi.org/10.1093/femsml/uqaa002.
Wang H, Yin X, Orr MW *et al.* Increasing intracellular magnesium levels with the 31-amino acid MgtS protein. *Proc Natl Acad Sci USA* 2017;**114**:5689–94.
Wang N, Yin Z, Duan W *et al.* sORF-encoded polypeptide SEP1 is a novel virulence factor of *Phytophthora* pathogens. *MPMI* 2021;**34**:157–67.
Wang Y, Zhang H, Lu J. Recent advances in ribosome profiling for deciphering translational regulation. *Methods* 2020;**176**:46–54.
Wassarman KM, Repoila F, Rosenow C *et al.* Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev* 2001;**15**:1637–51.
Weaver J, Mohammad F, Buskirk AR *et al.* Identifying small proteins by ribosome profiling with stalled initiation complexes. *mBio* 2019;**10**. https://doi.org/10.1128/mBio.02819-18.
Willems P, Fijalkowski I, Van Damme P. Lost and found: re-searching and re-scoring proteomics data aids genome annotation and improves proteome coverage. *mSystems* 2020;**5**. https://doi.org/10.1128/mSystems.00833-20.
Willems P, Ndah E, Jonckheere V *et al.* To new beginnings: riboproteogenomics discovery of N-terminal proteoforms in *Arabidopsis Thaliana*. *Front Plant Sci* 2022;**12**. https://doi.org/10.3389/fpls.2021.778804.
Williams AH, Redzej A, Rolhion N *et al.* The cryo-electron microscopy supramolecular structure of the bacterial stressosome unveils its mechanism of activation. *Nat Commun* 2019;**10**. https://doi.org/10.1038/s41467-019-10782-0.
Wright BW, Molloy MP, Jaschke PR. Overlapping genes in natural and engineered genomes. *Nat Rev Genet* 2022;**23**:154–68.
Xiong J. 7 - protein motifs and domain prediction. In: *Essential Bioinformatics*. Cambridge: Cambridge University Press, 2006, 85–94.
Xu J, Li T, Gao Y *et al.* MgrB affects the acid stress response of *Escherichia coli* by modulating the expression of iraM. *FEMS Microbiol Lett* 2019;**366**. https://doi.org/10.1093/femsle/fnz123.
Yadavalli SS, Goh T, Carey JN *et al.* Functional determinants of a small protein controlling a broadly conserved bacterial sensor kinase. *J Bacteriol* 2020;**202**, https://doi.org/10.1128/JB.00305-20.
Yadavalli SS, Yuan J. Bacterial small membrane proteins: the Swiss army knife of regulators at the lipid bilayer. *J Bacteriol* 2022;**204**. https://doi.org/10.1128/JB.00344-21.
Yoshitani K, Ishii E, Taniguchi K *et al.* Identification of an internal cavity in the PhoQ sensor domain for PhoQ activity and SafA-mediated control. *Biosci Biotechnol Biochem* 2019;**83**:684–94.
Yu J, Guo L, Dou X *et al.* Comprehensive evaluation of protein-coding sORFs prediction based on a random sequence strategy. *Front Biosci Landmark* 2021;**26**:272–8.
Zeghouf M, Li J, Butland G *et al.* Sequential Peptide Affinity (SPA) system for the identification of mammalian and bacterial protein complexes. *J Proteome Res* 2004;**3**:463–8.
Zhou C, Wang Q, Huang Y *et al.* Probing the sORF-Encoded Peptides of *Deinococcus radiodurans* in Response to Extreme Stress. *Mol Cell Proteomics* 2022;**21**. https://doi.org/10.1016/j.mcpro.2022.100423.

Received 9 May 2023; **revised** 10 November 2023; **accepted** 24 November 2023