frontiers
in Molecular Biosciences

HYPOTHESIS AND THEORY
published: 14 August 2020
doi: 10.3389/fmolb.2020.00187

Are Antisense Proteins in Prokaryotes Functional?

Zachary Ardern, Klaus Neuhaus and Siegfried Scherer*

Chair for Microbial Ecology, Technical University of Munich, Munich, Germany

Many prokaryotic RNAs are transcribed from loci outside of annotated protein coding genes. Across bacterial species hundreds of short open reading frames antisense to annotated genes show evidence of both transcription and translation, for instance in ribosome profiling data. Determining the functional fraction of these protein products awaits further research, including insights from studies of molecular interactions and detailed evolutionary analysis. There are multiple lines of evidence, however, that many of these newly discovered proteins are of use to the organism. Condition-specific phenotypes have been characterized for a few. These proteins should be added to genome annotations, and the methods for predicting them standardized. Evolutionary analysis of these typically young sequences also may provide important insights into gene evolution. This research should be prioritized for its exciting potential to uncover large numbers of novel proteins with extremely diverse potential practical uses, including applications in synthetic biology and responding to pathogens.

Keywords: overlapping gene, antisense transcription, antisense translation, function, selected effects, gene annotation

OPEN ACCESS

Edited by:
Eleonora Leucci,
KU Leuven, Belgium

Reviewed by:
Konrad Ulrich Förstner,
Leibniz Information Centre for Life Sciences (LG), Germany
Gabriele Fuchs,
University at Albany, United States

****Correspondence:***
Zachary Ardern
zachary.ardern@tum.de

Specialty section:
This article was submitted to Protein and RNA Networks, a section of the journal Frontiers in Molecular Biosciences

Received: *20 February 2020*
Accepted: *16 July 2020*
Published: *14 August 2020*

Citation:
Ardern Z, Neuhaus K and Scherer S (2020) Are Antisense Proteins in Prokaryotes Functional? Front. Mol. Biosci. 7:187. doi: 10.3389/fmolb.2020.00187

INTRODUCTION

The Many Functions of Antisense RNAs

A wide range of non-coding RNAs have been characterized in bacterial genomes. Among these putatively non-coding sequences are many antisense transcripts. Indeed, up to 75% of all prokaryotic genes are associated with antisense RNAs – though the number differs significantly between species and according to the methods used (Georg and Hess, 2018). Their functions, if any, are poorly understood in most cases. The characteristics of antisense RNAs range widely in terms for instance of length, location in relation to the sense gene, and mechanisms of regulation (Lejars et al., 2019). In studies so far they are usually associated with reducing transcription of the sense gene, but they can also increase it, for instance by changing the structure of the sense transcript – various mechanisms are known in each case (Lasa et al., 2012). They can influence single genes, or have global effects for instance through a target involved in general translation. Other known effects relate to functions including virulence, motility, various mechanisms of gene transfer, and biofilm formation (Lejars et al., 2019). The numerous examples of antisense transcription which have been investigated do not just include short antisense RNAs, though these are well-known; the many longer examples include a 1200 nucleotide antisense RNA in *Salmonella enterica*, AmgR (Courtney and Chatterjee, 2014). Antisense transcripts have been shown to be co-expressed within a single cell with the use of an antibody against double-stranded RNA in various studies, including in *Escherichia coli* and *Streptomyces coelicolor*, as reviewed in Georg and Hess (2018). Relatively little attention, however, has been paid to the possibility that RNA in antisense to protein coding genes may also frequently encode proteins (Georg and Hess, 2018). Rather than short, trivial overlaps, which are well known (Saha et al., 2016), here we focus on cases where an antisense (or "antiparallel") ORF with evidence of translation is fully embedded within a known protein coding gene.

The existence of substantially overlapping gene pairs has been known since the beginning of modern genome sequencing, when the proteins directly detected in the bacteriophage phiX174 were shown to not be able to fit into the sequenced genome without the translation of overlapping open reading frames (ORFs; Barrell et al., 1976). Since then, overlapping genes have typically been assumed to be fairly common only in viruses and extremely rare in other taxa, with the possibility of there being multiple examples in other taxa only sporadically discussed, e.g., Chou et al. (1996). However, their occurrence in bacteriophage in particular should raise the suspicion that they may be common in bacteria as well, given for instance the large amounts of genetic material transferred from temperate phage genomes to bacterial genomes (Harrison and Brockhurst, 2017; Owen et al., 2020). The properties of same-strand overlaps between viral genes have been studied (Pavesi et al., 2018; Willis and Masel, 2018), but even in viruses, relatively little attention has been given to antisense overlaps. There is, however, increasing evidence for functional translated antisense ORFs too, notably the antisense protein Asp in HIV-1 (Cassan et al., 2016; Affram et al., 2019; Nelson et al., 2020). In general it can be said that small ncRNAs are well recognized but their coding potential has been overlooked. Many might be protein-coding (i.e., mRNA), some are indeed ncRNA, and several will be dual-functional (Wadler and Vanderpool, 2007; Gimpel and Brantl, 2017; Neuhaus et al., 2017). The same trichotomy of functional categories applies in the case of antisense RNAs.

In bacteria, a number of individual antisense proteins have been discovered; the lines of evidence for some of these will be discussed below. High throughput analyses of ribosome profiling data, which uncovers the part of the transcriptome associated with ribosomes (Ingolia et al., 2009) thus revealing the "translatome," have begun to suggest that many more may be present. Friedman et al. (2017) found evidence for approximately 17 antisense ORFs, previously thought to be non-coding sRNAs, translated over above the level expected by chance in *E. coli* K12. The 10 sRNAs these belong to are shown in **Figure 1A**. Weaver et al. (2019) found ribosome profiling evidence, including evidence specifically for translation initiation (using retapamulin), for nine antisense overlapping gene candidates in *E. coli* K12, also shown, combined with the data from Weaver et al., in **Figure 1A**. As reported in a recent pre-print, Smith et al. (2019) found many overlapping ORFs in *Mycobacterium tuberculosis* associated with ribosomes using retapamulin. From 355 novel ORFs expressed in two replicates they report 241 overlapping and embedded in annotated genes, including both sense, and antisense overlaps, of which many were very short. Of those encoding at least 20 amino acids, 51 are antisense embedded. These antisense ORFs are shown in **Figure 1B**. From **Figure 1** we see that translated antisense overlapping genes are distributed roughly evenly across the genome, and in both frames, in the best-studied example genomes to-date; *E. coli* K-12 and *M. tuberculosis*. We have good reason to expect many similar overlapping genes across prokaryotic genomes.

It has been claimed that as a class the novel ORFs in *M. tuberculosis* are not under selection, and the association with ribosomes was attributed to non-functional pervasive translation (Smith et al., 2019) – this is discussed further in the section on evolution and constraint below. Whatever their selective status, other claims of translation in antisense to known genes continue to accumulate in prokaryotes. Jeong et al. (2016) reported ribosome profiling in *S. coelicolor* – although this result was not highlighted, examining the supplementary data showed 10 antisense putative sRNAs with ribosome profiling evidence. No doubt many more such discoveries await systematic analysis of published ribosome profiling data. There are also many putative same-strand overlapping genes, as discussed early on by Ellis and Brown (2003) showing that alternate frame translation is likely a general phenomenon – but these have also been claimed to not be under selection (Meydan et al., 2019). This increasing evidence for translation of both sense and antisense alternate frame ORFs, currently only typically acknowledged as ncRNAs, should push the question of "pervasive function" and how to categorize the range of translated ORFs to the forefront of microbiology, but it is yet to receive substantial attention. The evidence of expression in alternate frames is generally ignored, and when acknowledged it is generally presumed to be non-functional –, however, we argue this inference is made too quickly on insufficient grounds. Here we explore how to ascertain function and present a few examples of antisense genes with evidence for functionality.

"Function" and Natural Selection

The question of what counts as "function" in a biological context is not straightforward. An interdisciplinary group of researchers have recently discussed the issue in relation specifically to *de novo* gene origin (Keeling et al., 2019) and proposed five categories of meanings of function, pertaining to expression, capacities, interactions, physiology, and evolution. As they helpfully note, "Separating these meanings from one another enables communicating with increased precision about what the findings are, thereby helping to [avoid] fallacious logical shortcuts such as 'this protein is expressed therefore it is functional therefore it is under selection."' Interestingly, they had limited success in actually applying their categorisation, with most instances in a test set of article abstracts not uniformly assigned to a category by different team members. This suggests that biologists should write with more precision to clarify the sense of function intended. In this article we will focus on the senses relating to the biochemical "capacity" of the products of genetic elements and their evolutionary history, although the other senses will also come into play. The unifying general concept we use is that an element is functional if it does something useful for the organism in ecologically relevant circumstances.

The important philosophical questions have been reviewed elsewhere (Brandon, 2013). Here we summarize some established methods for determining function in the molecular biosciences and how they have been or could be applied to antisense proteins. It has become popular to adopt an etiological account of function,

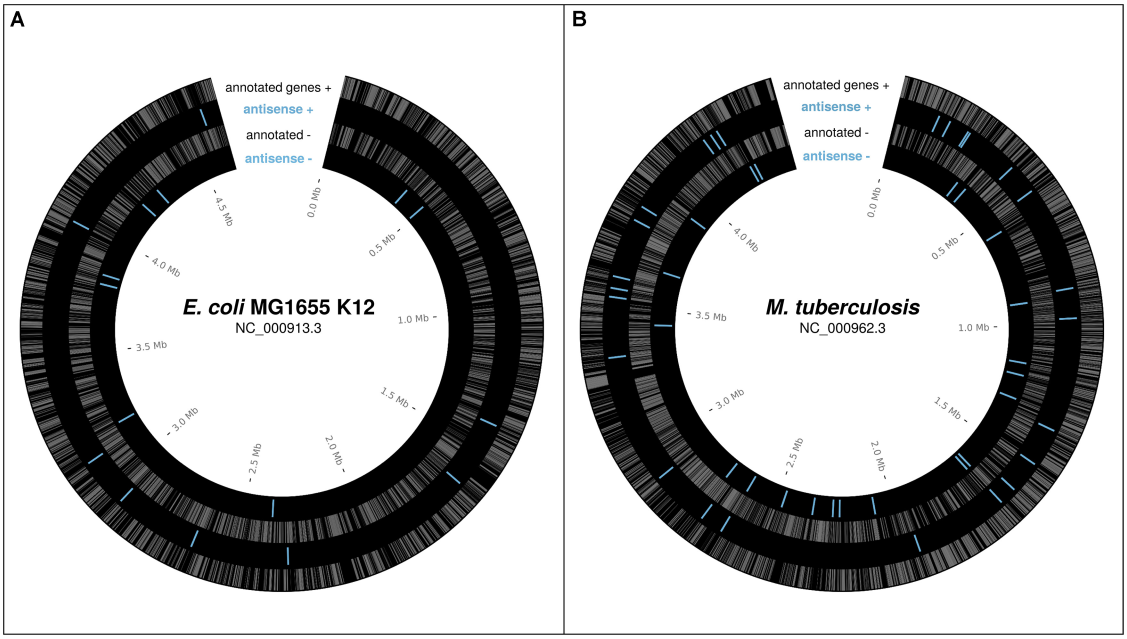

FIGURE 1 | Reported potential protein-coding ORFs overlapping in antisense, based on ribosome profiling experiments. **(A)** Reported antisense OLGs in *E. coli* K12 (NC_000913.3) – Weaver et al. (2019), Friedman et al. (2017; sRNAs with evidence of translation). **(B)** Reported antisense OLGs in *M. tuberculosis* (NC_000962.3) –Smith et al. (2019). Annotated genes gray. Antisense blue.

i.e., that an element's function depends on its selective history, particularly in relation to the dispute over how to assign function to elements in the human genome following the ENCODE project (Graur et al., 2013, 2015; Doolittle et al., 2014; Doolittle, 2018). However, the evolutionary etiology of biological systems is not always fully accessible to us (Ardern, 2018) and sometimes the history of selection in a lineage or for a particular gene may be inaccessible or the accessible parts incomplete in important ways. The genomic influence of different kinds of selection on bacterial genomes, including selective sweeps, background selection, positive selection, and purifying selection, remains a point of contention (Takeuchi et al., 2015; Bendall et al., 2016; Gibson and Eyre-Walker, 2019; Sela et al., 2019). Perhaps the most difficult issue here is how to characterize function in young genes, which may be subject to evolutionary forces lying anywhere along a spectrum between positive selection and purifying selection. Positive selection may be acting to modify a sequence which has only recently evolved or only recently become useful, for instance due to new environmental conditions. At some point, however, modifications are overwhelmingly selected against, i.e., purifying selection dominates. This fascinating transition region to our knowledge has received little study, but it is plausible that most young genes fall within it (Vishnoi et al., 2010). As such, many young genes are likely to be missed by methods seeking clear signatures of either purifying or positive selection. A recent study has shown that embedded overlapping genes in viruses usually evolve faster than the gene they are embedded in (Pavesi, 2019) such cases will tend to be missed by tests of purifying selection.

Additional relevant complexities include recombination, horizontal gene transfer, varying evolutionary rates, and unknown past environmental conditions. Evolutionary analyses certainly can provide strong evidence for function in cases of strong selection, but appropriate lower thresholds for determining that an element is functional while minimizing false negatives are much harder to determine. Arguably of much greater relevance than etiology for molecular biologists is what a genetic element does in the current system, and whether it contributes to the goals or life-conducive activities of that system. That is, as the etiological theorists correctly emphasize, function is not just about "causal role," it concerns a contribution to a wider system which is in some sense goal-directed. However, given complex histories of multiple evolutionary forces this does not necessarily imply anything directly about a particular canonical signature of natural selection being observable in the existing sequence. A good example of these complexities is the prevalence of translation and likely functions in putative "pseudogenes" (Goodhead and Darby, 2015; Cheetham et al., 2019).

EVIDENCE AND OBJECTIONS

High-Throughput Experimental Evidence

The "gold-standard" proof of the active translation of a gene has traditionally been direct evidence from proteomics experiments, a technology which precedes modern genome sequencing by a few years. However, evidence from current proteomics methods is inherently limited even after decades of improvements. For instance, small proteins are notoriously difficult to detect by mass spectrometry, because upon proteolytic digestion they tend to generate no suitable peptides or just a small number.

Another issue for detecting proteins by mass spectrometry is high hydrophobicity (Lescuyer et al., 2004) for example, proteins that contain *trans*-membrane domains are often underrepresented in proteomic data sets. Finally, also factors like a low protein abundance, only context-specific expression, a high turnover rate or protein secretion might all hamper a successful detection of proteins (Elguoshy et al., 2016). Nonetheless, despite these barriers there are a few examples of translated overlapping genes with proteomic evidence. Notably, a large-scale study of 46 bacterial genomes found up to 261 cases of annotation "conflict," i.e., overlaps greater than 40 base pairs with either proteomic evidence for both, or the unevidenced gene being annotated as something other than "hypothetical" (Venter et al., 2011). A more recent study of 11 bacterial transcriptomes (Miravet-Verde et al., 2019) found 185 antisense transcripts previously annotated as non-coding could in fact code for proteins based on a random forest classifier (RanSEPs). A study in *Pseudomonas putida* found proteomic evidence for 44 ORFs embedded in antisense to annotated ORFs (Yang et al., 2016). An improved proteogenomics pipeline reported in a recent pre-print manuscript found numerous gene candidates in *S. enterica* serovar Typhimurium, including a 199 amino acid long protein antisense to the annotated gene CBW18741 (Willems et al., 2019). It is interesting given the previous comment concerning the rate of phage to bacterial gene transfer that a BLAST search shows that this is likely a bacteriophage protein. The same study also found 18 antisense ORFs in *Deinococcus radiodurans* supported by at least two peptides. A search in *Helicobacter pylori* mass spectrometry data from a previously published study designed to find small proteins (Müller et al., 2013) found evidence for a protein encoded by an ORF antisense to a proline/betaine transporter gene (Friedman et al., 2017). A recent discussion paper presented proteomic evidence for many small proteins (sORFs) and overlapping genes ("altORFs"), but did not specifically consider antisense overlaps (Orr et al., 2020).

Aside from proteomics datasets there is extensive publicly available high throughput RNA sequencing data which can be mined for further indicators of specific reproducible regulation of antisense ORFs. There are approximately 1500 relevant RNAseq studies from prokaryotes in the NCBI GEO database (Edgar et al., 2002) each with multiple samples; over 100 ribosome profiling studies, and a number of more bespoke methods which may also provide relevant information. Cappable-seq data, which discovers transcriptional start sites (Ettwiller et al., 2016), helps to delineate the borders of operons and their expression under different conditions. The new method SEnd-seq, through circularisation of transcripts, is able to detect both transcriptional start and termination sites with single nucleotide resolution (Ju et al., 2019). CHiPseq datasets indicate whether known transcription factors are associated with a particular operon of interest (Wade, 2015) other TF-binding assays also have potential for testing hypotheses concerning TF binding, e.g., DNAse footprinting (Haycocks and Grainger, 2016). Each of these methods is yet to be fully utilized in searching for the transcriptional regulation of overlapping genes. At the level of translation, there are a number of variations on ribosome profiling now available, including accurate prediction of translation initiation sites. The first study of ribosome profiling in bacteria used chloramphenicol in one of the two methods presented (Oh et al., 2011), which has since been shown to stall the ribosome at initiation and, thereby, can assist in inferring translation initiation site positions (Mohammad et al., 2019; Glaub et al., 2020). More precise stalling has been achieved with the use of tetracycline (Nakahigashi et al., 2016), retapamulin (Meydan et al., 2019), and the antibacterial peptide Onc112 (Weaver et al., 2019). Properties of ribosomes at different stages of translation, including initiation, have recently been studied in *E. coli* K12 with TCP-seq; translation complex profiling (Sharma and Anand, 2019). Translation stop sites have also been specifically explored (Baggett et al., 2017). Most of these methods, outside the analysis of ribosome profiling discussed above, have not yet been applied to the detection or investigation of protein coding alternate frame ORFs, and any RNAs at these sites are assumed to be non-coding. Perhaps particularly useful will be ribosome profiling experiments conducted for cells grown in different conditions – many relevant contexts may, however, not be able to be surveyed due to technical limitations.

Phenotypes of Antisense Proteins

An important indicator of functionality is specific regulation in response to defined environmental conditions. Some key canonical work in molecular genetics (Jacob and Monod, 1961; Ames and Martin, 1964) has been concerned with the differential induction of genetic elements under varying environmental conditions. Specific differential induction is widely assumed in this kind of literature to be equivalent to function – how precisely to draw a line between functional and non-functional, given the inherent noisiness of biology, is not, however, entirely clear.

In general, what kind of phenotype is a good indicator of functionality? The most obvious case perhaps is an improvement in growth associated with expression of a genetic element. This could be either through improved growth following overexpression, or decreased growth following a deletion in the genomic sequence. Within an evolutionary context, a growth advantage effectively just is what it is to be "useful" or "functional." An example of this for antisense proteins is *citC*, discussed below. However, less intuitively, a decrease in growth associated with expression, as seen in the cases of *asa*, *laoB*, and *ano* is also an indicator of functionality in the right context. Most simply, the gene might literally function as a toxin. More generally though, overexpression of many functional genes is deleterious – in fact in *E. coli* the majority of annotated genes have a deleterious effect on growth in overexpression constructs (Kitagawa et al., 2005). Similarly, a condition-specific positive growth phenotype following knockout of an expressed gene is also indicative of function. Such a phenotype could be simply because the protein is not required in this environment and so losing it decreases the cost of expression. Or it could be because losing a gene with, for instance, a regulatory or inhibitory function is beneficial under certain conditions where regulation of a process is not useful. Indeed, whatever the underlying mechanisms, adaptation in bacteria following loss of function is pervasive (Behe, 2010; Hottes et al., 2013; Albalat and Cañestro, 2016).

Possible reasons for the high tendency toward deleterious over-expression phenotypes in bacteria compared with organisms such as yeast are discussed in Bhattacharyya et al. (2016). This situation should perhaps not be surprising given the extreme optimality of bacterial metabolism (Schuetz et al., 2012) significant disturbance of such a finely tuned system is unlikely to be beneficial under most conditions. This general principle follows from, for instance, Fisher's Geometric Model, in which random changes are less likely to be beneficial when a population is close to a fitness optimum (Tenaillon, 2014). Overexpression of a non-functional "junk" genetic sequence, however, is also likely to be deleterious (Weisman and Eddy, 2017; Knopp and Andersson, 2018) so such a phenotype does not by itself provide evidence for functionality. What is important in the examples discussed above is that the deleterious growth phenotypes are observed as a significant difference between environmental (media) conditions. This implies a specificity of interaction which appears improbable under the "junk" hypothesis, and so constitutes evidence of function.

A number of antisense overlapping genes in *E. coli* have been analyzed regarding expression and phenotypes across different environmental conditions. The gene *nog1* is almost fully embedded in antisense to *citC*. A strand-specific deletion mutant has a growth advantage over the wildtype in LB, and a stronger advantage in medium supplemented with magnesium chloride (Fellner et al., 2015). The gene *asa*, embedded in antisense to a transcriptional regulator in *E. coli* O157:H7 strains, was found to be regulated in response to arginine, sodium, and different growth phases. Overexpression resulted in a negative growth phenotype in both excess sodium chloride and excess arginine, and no phenotype in LB medium (Vanderhaeghen et al., 2018). The gene *laoB* is embedded in antisense to a CadC-like transcriptional regulator (**Figure 2A**). A strand-specific genomic knock-out mutant was shown to provide a growth advantage specific to media supplemented with arginine. Further, the differential phenotype was replicated with the addition of inducible plasmid constructs bearing Δ*laoB* and WT *laoB*, showing that the phenotype is removed through complementation (Hücker et al., 2018b). How to mechanistically interpret such a growth advantage following gene knockout is unclear, but the condition-specific clear phenotype implies a functional role. The gene *ano* is nearly fully embedded antisense to an L,D-transpeptidase (**Figure 2B**). Similarly to *laoB*, a knock-out mutant showed a condition-specific phenotype. In this case it occurs in anaerobic conditions, and could be partially complemented with a plasmid construct (Hücker et al., 2018a). The putative protein-coding gene *aatS* was found in the pathogenic *E. coli* strain ETEC H10407 fully embedded in antisense to the ATP transporter ATB binding protein AatC (Haycocks and Grainger, 2016). It was shown to be transcribed, to have a functional ribosome binding sequence, and to have widespread homologs including a conserved domain of unknown function. **Figure 2** illustrates the expression of three examples of antisense genes, with the gene in *Staphylococcus aureus* (**Figure 2C**) a special case, as discussed below.

Other than the high-level phenotypes (e.g., expression under particular conditions) determined for some candidates, very little is known about the possible roles or mechanisms of action of antisense proteins. Signaling or interactions between cells will be a significant area to investigate regarding possible functions. This suggestion is based on both the evidence gained so far for small proteins (Neuhaus et al., 2016; Hücker et al., 2017a; Sberro et al., 2019) and the particular importance of signaling or interaction proteins, meaning that this hypothesis derived from findings on small proteins more generally, deserves special attention This area is a crucial field of research as infectious disease continues to be a major health burden and the long natural history of interactions between microbes has been a fruitful source of new antimicrobial strategies.

Simultaneous Transcription?

In response to the evidence for overlapping genes, the question is often raised concerning how two genes could be simultaneously expressed from opposite strands. Indeed, the phenomenon of RNA polymerase collision is a real barrier to antisense transcription in at least some instances and is involved in transcriptional silencing or reduction via various mechanisms (Courtney and Chatterjee, 2014). Bypass of sense and antisense RNA polymerases has been shown for bacteriophage RNA polymerases (Ma and McAllister, 2009) but *in vitro* experiments have shown no such bypass in bacterial systems (Crampton et al., 2006). The role of accessory helicases in removing barriers to replication due to the presence of RNA polymerases has recently been highlighted (Hawkins et al., 2019) expanding on knowledge of simultaneous transcription and replication (Helmrich et al., 2013). It is conceivable that transcribing alongside the formation of a replication fork could facilitate antisense transcription, but this would restrict antisense transcription to the replication process. However, even in cases of collision of RNA polymerases operating in antisense, transcriptional stalling is not guaranteed. A recent study argues on the basis of simulations and careful assays with reporter constructs that RNA polymerases trailed by an active ribosome are, remarkably, about 13-times more likely to resume transcription following collision than those without the translation apparatus following (Hoffmann et al., 2019). This finding follows on from a range of similar work in recent years showing multiple mechanisms involved in ensuring that RNA polymerases stall and are subsequently released less in protein coding than non-coding RNAs (Proshkin et al., 2010; Brophy and Voigt, 2016; Ju et al., 2019). We suggest that this phenomenon likely applies to antisense embedded protein-coding genes as much as to convergent antisense transcripts and thereby facilitates antisense protein expression.

Recent detailed elucidation showed the working of an operon in *S. aureus* with a functional gene encoded in antisense to a contiguous set of co-transcribed genes (Sáenz-Lahoya et al., 2019). The authors showed that despite being encoded on opposite strands (although not directly overlapping in this case), these elements comprised a single transcriptional unit (**Figure 2C**). This study highlights a mechanism which may be widespread and may apply to genes which are directly antiparallel as well. Results from Weaver et al. (2019) obtained

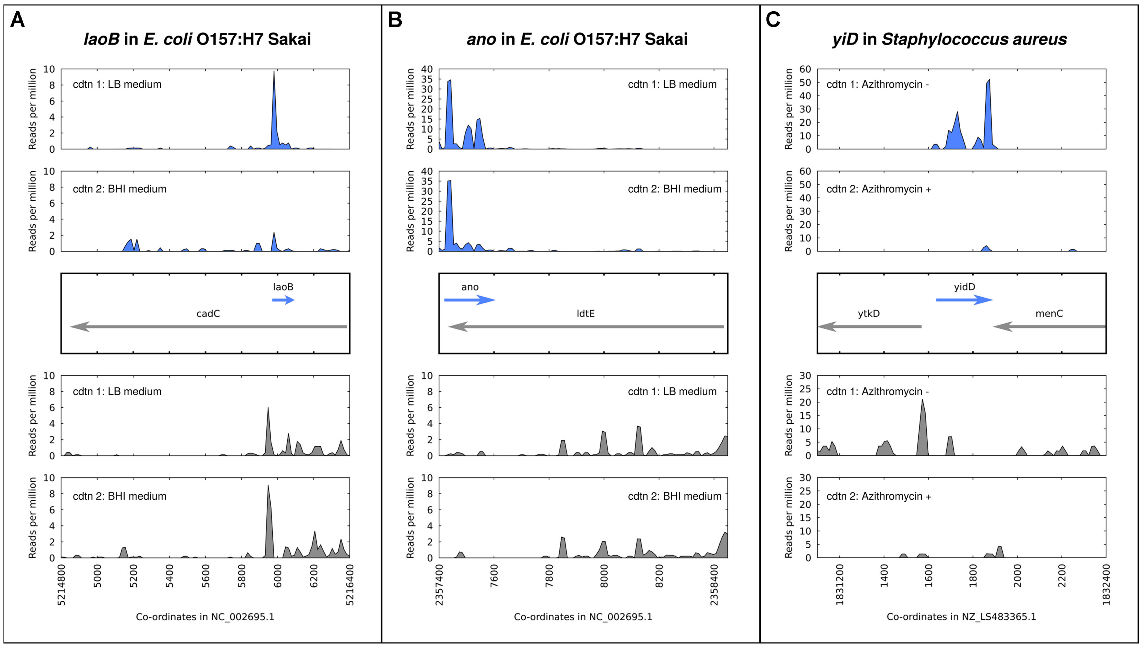

FIGURE 2 | Expression and regulation of antisense genes as demonstrated by ribosome profiling experiments; aligned ribosome protected fragments shown as reads per million at each site after removal of rRNA and tRNA reads. **(A)** *laoB* gene in *E. coli* O157:H7 Sakai; expression in LB medium is higher than in BHI. **(B)** *ano* gene in *E. coli* O157:H7 Sakai; expression in LB medium versus BHI is constant. **(C)** *yidD* (MW1733) gene, part of non-contiguous (antisense) operon, in *Staphylococcus aureus*. Expression in positive and negative strands is positively correlated – high in the absence of the antibiotic azithromycin, low when it is added.

by chromosomal tagging of three antisense proteins show that proteins encoded in antisense can be expressed simultaneously, i.e., under the same growth conditions. All this evidence for simultaneous antiparallel gene expression notwithstanding, it may be that antiparallel overlapping genes are generally translated under different conditions, or separated in time – this is yet to be determined.

Evolution and Constraint in Antisense Proteins

The evolutionary analysis of function at the nucleotide sequence level is a fairly recent development (Robinson-Rechavi, 2019) so we should not be surprised at unexpected results in this rapidly developing field. While the evolutionary analysis of antisense proteins in prokaryotes awaits further investigation of strong overlapping gene candidates, those discovered so far are typically relatively young (Fellner et al., 2014, 2015; Hücker et al., 2018a). This may be seen as a point against their functionality, particularly for candidates limited to just one species. However, a number of genome elements with undisputed functionality are also evolutionarily young. Various functional putatively ncRNA elements are known to have high evolutionary turnover, see Dutcher and Raghavan (2018). For instance, an sRNA found only in *E. coli* was shown to be derived from a pseudogenized bacteriophage gene (Kacharia et al., 2017). Also relevant here is the large literature on the functions of "orphan" or taxonomically restricted genes restricted to a single genome or small clade (Satoshi and Nishikawa, 2004; Tautz and Domazet-Lošo, 2011) and orphan genes may play diverse important roles in bacteria (Hu et al., 2009).

It appears likely that antisense proteins are often less constrained in sequence than most protein-coding genes currently known. For one, antisense proteins are typically quite small and hence unlikely to fold into complex structures. Secondly, given initial evidence from viruses that protein domains in overlapping genes may be situated so as to not overlap (Fernandes et al., 2016), it seems that overlapping gene sequences are unlikely to be comprized of a high proportion of constrained sequence domains. While our previous analyses of individual prokaryotic overlapping genes have shown that they are typically young compared to the genes in which they are embedded (Hücker et al., 2018a) many embedded ORFs are quite well conserved beyond the genus. As a conservative example, we take a subset of the Enterobacteriaceae family, the smallest clade including both *Citrobacter rodentium* and *E. coli* (**Figure 3A**). We find that out of the 3391 antisense embedded ORFs predicted as having single homologs in all 13 representative genomes assessed, 29.5% exceed the conservation level of the lower quartile of annotated genes (**Figures 3B,C**). Here, conservation is judged by median pairwise amino acid similarity between genomes. Given the conservative nature of this analysis and that less than half of even annotated genes met the criterion of having single homologs in all of these genomes, we posit that thousands of embedded antisense ORFs are sufficiently conserved beyond the *Escherichia* genus to be candidates for functional genes in this particular respect. Factors affecting these conservation statistics,

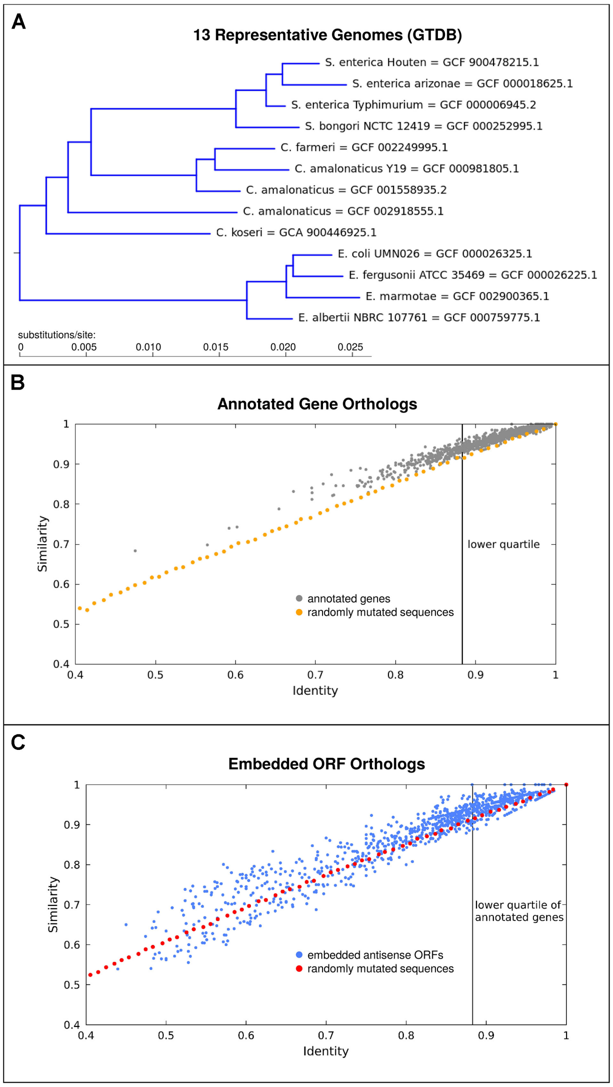

FIGURE 3 | A subset of embedded antisense ORFs are well conserved in the genomes situated taxonomically between *E. coli* and *Citrobacter rodentium*. **(A)** Phylogenetic tree showing 13 representative genomes used, derived from the genome taxonomy database (GTDB). **(B)** Median pairwise amino acid identity and similarity (gray) among orthologs for 1000 annotated genes with single copy orthologs in all of the 13 genomes. As a comparison, the effect on identity and similarity of adding random mutations to simulated sequences is shown (orange). There is a clear bias toward variants which result in higher "similarity." **(C)** Conservation of antisense embedded ORFs (blue), as compared to the identity-similarity relationship observed for control randomly mutated sequences (red) simulated as before but translated in antisense. Many antisense embedded ORFs are highly conserved and a subset also shows a bias toward similarity.

and additional criteria for gene-likeness which distinguish coding from non-coding antisense sequences deserve further study.

The orange and red lines in **Figures 3B,C** show the effect of randomly mutating a sequence created based on the codon usage in the annotated genes in *E. coli* K12. The points plotted represent median identities and similarities in comparison to originally simulated sequences, following successive rounds of random mutation, approximately mimicking the mutational distances observed between the orthologs of annotated genes and embedded ORFs. We suggest that two main results should be taken from **Figure 3**. Firstly, the blue cluster in the top right of **Figure 3C** shows that many embedded antisense ORFs are highly conserved across a significant evolutionary distance – they are not all immediately degraded following mutations in the alternate frame as might be naively assumed. Secondly, the bias above the orange and red lines shows that nearly all annotated genes and many embedded antisense ORFs tend toward fixing more "similar" mutations than might be predicted based on amino acid identity statistics alone. This result may be partly due to the structure of the genetic code, i.e., when a "mother gene" in the reference frame is conserved there is some tendency for conservation in the alternative strand (Wichmann and Ardern, 2019), but it is also suggestive of a kind of purifying selection where mutations to biochemically similar amino acids are preferred in a subset of embedded antisense ORFs. It has previously been shown that long antisense ORFs appear more often in natural genomes than expected based on codon composition of annotated coding genes (Mir et al., 2012), another hint of selective processes preserving some antisense ORFs.

A recent, currently unpublished, study in *M. tuberculosis* (Smith et al., 2019) has claimed that novel ORFs identified by ribosome profiling typically do not illustrate the strong codon bias evident in annotated mycobacterial genes and therefore cannot be expected to be functional. Given that many of these ORFs are situated in antisense to annotated genes, where the genetic code limits the possibilities for achieving optimal codon usage, this result is not surprising, and we suggest provides little evidence for the claim that they are nonfunctional. There is a problematic circularity here as well, as annotation of prokaryotic ORFs is based on models which take into account codon usage, based on usage in long ORFs – so short ORFs with "abnormal" codon usage will likely remain unannotated, reinforcing any bias in codon usage statistics in annotated genes. In general, short and weakly expressed genes should not be expected to match "canonical" highly expressed genes in terms of codon usage (Gupta and Ghosh, 2001), although the relationship between expression and codon usage is not straightforward (Dos Reis et al., 2003). Careful evolutionary sequence analyses are required here. A study of some putative same-strand overlapping genes also suggested that they are not under constraint (Meydan et al., 2019). However, more biologically nuanced analyses of sequence constraint, for instance after partitioning the homologs into phylostrata, would be useful. Further, a fundamental assumption of methods for detecting selection (e.g., Firth, 2014; Wei and Zhang, 2015), is the neutrality of synonymous mutations, but this assumption has been shown to be false, with the rate of synonymous mutations varying widely across sites (Wisotsky et al., 2020). The extent to which this affects conclusions regarding dN/dS as calculated with the various available methods remains unexplored. More generally, to our knowledge, there has been no demonstration of any synthesis of non-functional protein in prokaryotes. The high bioenergetic cost of protein production (Lynch and Marinov, 2015) would seem to militate against such a phenomenon being widespread in bacteria, where costs are minimized through gene loss (Koskiniemi et al., 2012). As such, we argue that the default assumption following demonstration of a clear signal of translation should be that the product plays a functional role.

DISCUSSION

The Context: Unexpected Complexity

The historical trajectory in bacterial genomics has been toward finding previously unappreciated layers of complexity (Grainger, 2016). In particular, the number of different kinds of functional elements recognized has continued to increase in recent years. Examples of genetic elements previously ignored or written off as background noise which are now known to be functional in some or many cases include antisense transcription, small RNAs, and microRNAs, proteins with alternative start sites, small proteins (Storz et al., 2014), and micropeptides. Antisense transcription has been widely disregarded as noise (Raghavan et al., 2012; Lloréns-Rico et al., 2016). However, despite these generalizations, these elements have recently been found to at least sometimes have physiological roles (Wade and Grainger, 2014; Lejars et al., 2019). Functional RNAs which are not yet well understood include structured noncoding RNAs such as riboswitches (Hücker et al., 2017b; Stav et al., 2019). Proteins with alternative start sites, designated isoforms or "proteoforms," have also been reported in a few bacterial systems (Berry et al., 2016; Nakahigashi et al., 2016; Meydan et al., 2019). These genetic elements are all yet to be incorporated into genome annotation files and gene prediction algorithms. As such, genome annotation is years behind the leading edge of research in bacterial genetics, and various functional elements remain unannotated.

Recommendations for Further Research

Even recent attempts at comprehensive studies of small proteins have tended to ignore antisense proteins or to use methods unintentionally biased against them – perhaps unsurprising given the reigning paradigm in genome annotation, which excludes substantive overlaps as a matter of principle. As an example, the NCBI prokaryotic genome annotation standards include among the minimum standards that there can be "[no] gene completely contained in another gene on the same or opposite strand" (NCBI 2020). For instance, a recent study investigated small proteins in the human microbiome (Sberro et al., 2019) finding hundreds of previously unknown small proteins with evidence from evolutionary sequence constraint, and many also with evidence of transcription and/or translation. Two key steps were the use of MetaProdigal for gene prediction and RNAcode for inference of conservation. Both of these are implicitly biased against overlapping genes, in that Prodigal explicitly excludes long overlaps, and RNAcode looks for patterns of sequence constraint associated with normal non-overlapping genes, which are unlikely to be found in overlapping genes.

The bacteriological research community ought to relinquish the common assumption that unannotated functional elements are only to be found in intergenic regions. We must also be aware that antisense regions often need to be treated differently from intergenic regions, for instance in analyses of sequence constraint. Developing appropriate corrections to take into account the sequence context of antisense overlapping ORFs is an important area for further work. A major emphasis should be on high-throughput functional studies. For in-depth laboratory studies dissecting the details of an overlapping gene's regulation and function, the focus should be on the strongest candidates as determined with sequence and expression data. One key criterion here is evidence of reproducible regulated translation from one of the various ribosome profiling methods now available. Sequence properties determined from such sets should help to find strong candidates which are not expressed under already-assayed conditions. It is also clear that further advances in proteomics for small proteins should result in proteomic evidence for the translation of many more antisense proteins in bacteria and other systems. Following on from this, the discovery of any protein structures would be a major step forward toward understanding the molecular mechanisms of function. Finally, studying the evolutionary history of antisense proteins may provide useful insights on function. In this aspect these genes have a significant advantage over others in that their genomic context is relatively fixed by the gene in which they are embedded. This study has focused on eubacteria, but the same principles conceivably apply in archaea. A recent study, for instance, chose to only consider same-frame overlapping ORFs (proteoforms) on account of an absence of proteomics results and reliable BLAST hits for out-of-frame overlapping ORFs (Ten-Caten et al., 2018). Neither of these negative results are surprising, however, given the limitations of proteomics discussed above and the current bias against annotating out-of-frame overlaps; as such, archaeal datasets ought also be re-examined for functional overlapping genes.

In summary, what is required in order to assign the descriptor 'functional' to a putative gene, such as a gene encoded in antisense to a known gene? Regarding evolutionary evidence, a codon-level pattern of sequence constraint is sufficient to guarantee function, as constraint matching expectations for amino acids is unexpected in coding sequences. Detecting such constraint is a challenge for antisense sequences, however. Regarding evidence from wet-lab experiments, a condition-specific phenotype is also sufficient to establish functionality. The "gold standard" in this area would be a condition-specific negative growth phenotype in a genomic knock-out mutant, which could be complemented in *trans* (e.g., with a plasmid construct). Regarding high-throughput evidence, significant protein expression is evidence of functionality in highly optimized bacterial genomes, particularly if shown to be consistent across species or highly diverged strains. Appropriate thresholds for significant expression and sufficient evolutionary divergence in order to be able to confidently infer function are yet to be established. While each of these three lines of evidence is arguably sufficient to establish function, none is necessary, as there are functional elements which fail to meet at least one of these criteria.

We have collated evidence from diverse bacteria (including the genera *Escherichia*, *Pseudomonas*, and *Mycobacterium*) for protein coding ORFs embedded in antisense to annotated genes, discussed reasons to believe that they are biologically functional, and responded to common objections, informed by the most recent work in bacterial molecular genetics. We suggest that a pro-function attitude regarding antisense prokaryotic transcripts and the antisense translatome is both more useful for research and justified by multiple lines of evidence. How many of these elements are functional and what they do remain contentious, however, and worthy of significant further investigation.

METHODS

For **Figure 1**, positions of previously discovered putative antiparallel genes in *E. coli* K12 and *M. tuberculosis* were extracted from the supplementary data of previous studies (Friedman et al., 2017; Smith et al., 2019; Weaver et al., 2019); information on those with ribosome profiling reads was provided by Robin Friedman. Positions are shown visualized with Circos (Krzywinski et al., 2009).

For **Figure 2**, ribosome profiling ("RIBO-seq") data was visualized to show examples of antisense overlapping genes. In each case, adapter sequences were predicted using DNApi.py (Tsuji and Weng, 2016), trimmed with cutadapt (Martin, 2011) using a minimum length of 19 and quality score of 10, and aligned (local alignment) with bowtie2 (Langmead and Salzberg, 2012). Fastq data from SRR5874479 (LB) and SRR5874484 (BHI) for *E. coli* O157:H7 Sakai was aligned against the genome GCF_000008865.1_ASM886v1. Fastq data from SRR1265839 (without azithromycin) and SRR1265836 (with azithromycin) for *S. aureus* was aligned against genome GCF_900475245.1_43024_E01. Reads mapping at each site per million total mapped reads (RPM) are calculated from aligned bam files with reads mapping to rRNA and tRNA locations removed, using samtools (Li et al., 2009). Images of RPM in the region around the putative antisense gene are drawn in gnuplot with "smooth csplines."

For **Figure 3**, the relationship between similarity and identity in comparisons of different ORF homologs was compared. Representative genomes from release 89 of the genome taxonomy database (GTDB; Parks et al., 2018) in the smallest clade uniting *E. coli* and *C. rodentium* (**Figure 3A**) were chosen. Of these 23 strains, 13 had a GenBank genome and feature table with the same accession version available. These were downloaded, and annotated ORFs in each compared to each other using OrthoFinder (Emms and Kelly, 2015). Genes with a single copy ortholog present in all 13 genomes were extracted, and members of each ortholog family were aligned against each other using the EMBOSS (Rice et al., 2000) program needleall to determine median similarity and identity at the amino acid level (**Figure 3B**). As a control, 50 sequences of 333 codons length were created based on codon usage in *E. coli* K12, using EMBOSS programs cusp and makenucseq. These were then mutated through 70 rounds of point mutation (10 mutations per round) using the EMBOSS program msbar, and translated in order to determine the relationship between varying levels of amino acid identity and similarity. In each case, the mutated sequences were compared to the original simulated sequence they were derived from, using EMBOSS needle. For each percent decrease in identity observed, results were collated and the median values of identity and similarity reported. The procedure used initially for annotated genes was repeated using all antisense embedded ORFs (using the bacterial genetic code, NCBI code table 11), found using a Perl script "ORFFinder" (available from Christopher Huptas) and Bedtools (Quinlan and Hall, 2010). A negative control for these sequences was also created similarly, to before, but using an antisense reading frame. As the particular antisense frame used had no significant effect on the sequence similarities obtained in the simulation, for the data shown the initial sequences based on codon usage in *E. coli* K12 were directly reverse complemented with no further frame-shift, prior to the 70 rounds of random mutation.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

ZA drafted the manuscript and prepared the figures. KN and SS assisted with drafting the manuscript and conceiving of the project. All authors read and approved the final version of the manuscript.

FUNDING

ZA was supported by the Bavarian State Government and the National Philanthropic Trust.

ACKNOWLEDGMENTS

Thanks to Christina Ludwig for advice on proteomics references, Christopher Huptas for development of an ORF finder Perl script, and Robin Friedman for providing ribosomal profiling information for putative sRNAs.

REFERENCES

Affram, Y., Zapata, J. C., Zhou, W., Pazgier, M., Iglesias-Ussel, M., Ray, K., et al. (2019). PJ-1 The HIV-1 antisense protein ASP is a structural protein of the viral envelope. *J. Acquir. Immune Defic. Syndr.* 81:79. doi: 10.1097/01.qai.0000558040.82718.71

Albalat, R., and Cañestro, C. (2016). Evolution by gene loss. *Nat. Rev. Genet.* 17, 379–391.

Ames, B. N., and Martin, R. G. (1964). Biochemical aspects of genetics: the operon. *Annu. Rev. Biochem.* 33, 235–258. doi: 10.1146/annurev.bi.33.070164.001315

Ardern, Z. (2018). Dysfunction, disease, and the limits of selection. *Biol. Theory* 13, 4–9. doi: 10.1007/s13752-017-0288-0

Baggett, N. E., Zhang, Y., and Gross, C. A. (2017). Global analysis of translation termination in *E. coli*. *PLoS Genet.* 13:e1006676. doi: 10.1371/journal.pgen.1006676

Barrell, B. G., Air, G., and Hutchison, C. (1976). Overlapping genes in bacteriophage φX174. *Nature* 264, 34–41. doi: 10.1038/264034a0

Behe, M. J. (2010). Experimental evolution, loss-of-function mutations, and "the first rule of adaptive evolution". *Q. Rev. Biol.* 85, 419–445. doi: 10.1086/656902

Bendall, M. L., Stevens, S. L., Chan, L.-K., Malfatti, S., Schwientek, P., Tremblay, J., et al. (2016). Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* 10, 1589–1601. doi: 10.1038/ismej.2015.241

Berry, I. J., Steele, J. R., Padula, M. P., and Djordjevic, S. P. (2016). The application of terminomics for the identification of protein start sites and proteoforms in bacteria. *Proteomics* 16, 257–272. doi: 10.1002/pmic.201500319

Bhattacharyya, S., Bershtein, S., Argun, T., Gilson, A. I., Trauger, S. A., and Shakhnovich, E. I. (2016). Transient protein-protein interactions perturb *E. coli* metabolome and cause gene dosage toxicity. *eLife* 5:e20309.

Brandon, R. N. (2013). "A general case for functional pluralism," in *Functions: Selection and Mechanisms*, ed. P. Huneman, (Dordrecht: Springer), 97–104. doi: 10.1007/978-94-007-5304-4_6

Brophy, J. A., and Voigt, C. A. (2016). Antisense transcription as a tool to tune gene expression. *Mol. Syst. Biol.* 12:854. doi: 10.15252/msb.20156540

Cassan, E., Arigon-Chifolleau, A. M., Mesnard, J. M., Gross, A., and Gascuel, O. (2016). Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic. *Proc. Natl. Acad. Sci. U.S.A.* 113, 11537–11542. doi: 10.1073/pnas.1605739113

Cheetham, S. W., Faulkner, G. J., and Dinger, M. E. (2019). Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nat. Rev. Genet.* 21, 191–201. doi: 10.1038/s41576-019-0196-1

Chou, K.-C., Zhang, C.-T., and Elrod, D. W. (1996). Do "antisense proteins" exist? *J. Protein Chem.* 15, 59–61. doi: 10.1007/bf01886811

Courtney, C., and Chatterjee, A. (2014). cis-Antisense RNA and transcriptional interference: coupled layers of gene regulation. *J. Gene Ther.* 1, 1–9.

Crampton, N., Bonass, W. A., Kirkham, J., Rivetti, C., and Thomson, N. H. (2006). Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. *Nucleic Acids Res.* 34, 5416–5425. doi: 10.1093/nar/gkl668

Doolittle, W. F. (2018). We simply cannot go on being so vague about 'function'. *Genome Biol.* 19:223.

Doolittle, W. F., Brunet, T. D., Linquist, S., and Gregory, T. R. (2014). Distinguishing between "function" and "effect" in genome biology. *Genome Biol. Evol.* 6, 1234–1237. doi: 10.1093/gbe/evu098

Dos Reis, M., Wernisch, L., and Savva, R. (2003). Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res.* 31, 6976–6985. doi: 10.1093/nar/gkg897

Dutcher, H. A., and Raghavan, R. (2018). Origin, evolution, and loss of bacterial small RNAs. *Microbiol. Spectr*. 6:RWR-0004-2017.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Elguoshy, A., Magdeldin, S., Xu, B., Hirao, Y., Zhang, Y., Kinoshita, N., et al. (2016). Why are they missing?: Bioinformatics characterization of missing human proteins. *J. Proteomics* 149, 7–14. doi: 10.1016/j.jprot.2016.08.005

Ellis, J. C., and Brown, J. W. (2003). Genes within genes within bacteria. *Trends Biochem. Sci.* 28, 521–523. doi: 10.1016/j.tibs.2003.08.002

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157.

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

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, 57–64.

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, 1762–1773.e12. doi: 10.1016/j.cell.2016.11.031

Firth, A. E. (2014). Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. *Nucleic Acids Res.* 42, 12425–12439. doi: 10.1093/nar/gku981

Friedman, R. C., Kalkhof, S., Doppelt-Azeroual, O., Mueller, S. A., Chovancova, M., Von Bergen, M., et al. (2017). Common and phylogenetically widespread coding for peptides by bacterial small RNAs. *BMC Genomics* 18:553. doi: 10.1186/s12864-017-3932-y

Georg, J., and Hess, W. R. (2018). Widespread antisense transcription in prokaryotes. *Microbiol. Spectr*. 6:RWR-0029-2018.

Gibson, B., and Eyre-Walker, A. (2019). Investigating evolutionary rate variation in bacteria. *J. Mol. Evol.* 87, 317–326. doi: 10.1007/s00239-019-09912-5

Gimpel, M., and Brantl, S. (2017). Dual-function small regulatory RNAs in bacteria. *Mol. Microbiol.* 103, 387–397. doi: 10.1111/mmi.13558

Glaub, A., Huptas, C., Neuhaus, K., and Ardern, Z. (2020). Recommendations for bacterial ribosome profiling experiments based on bioinformatic evaluation of published data. *J. Biol. Chem.* 295, 8999–9011. doi: 10.1074/jbc.ra119.012161

Goodhead, I., and Darby, A. C. (2015). Taking the pseudo out of pseudogenes. *Curr. Opin. Microbiol.* 23, 102–109. doi: 10.1016/j.mib.2014.11.012

Grainger, D. C. (2016). The unexpected complexity of bacterial genomes. *Microbiology* 162, 1167–1172. doi: 10.1099/mic.0.000309

Graur, D., Zheng, Y., and Azevedo, R. B. (2015). An evolutionary classification of genomic function. *Genome Biol. Evol.* 7, 642–645. doi: 10.1093/gbe/evv021

Graur, D., Zheng, Y., Price, N., Azevedo, R. B., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets:"function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* 5, 578–590. doi: 10.1093/gbe/evt028

Gupta, S., and Ghosh, T. (2001). Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. *Gene* 273, 63–70. doi: 10.1016/s0378-1119(01)00576-5

Harrison, E., and Brockhurst, M. A. (2017). Ecological and evolutionary benefits of temperate phage: what does or doesn't kill you makes you stronger. *Bioessays* 39:1700112. doi: 10.1002/bies.201700112

Hawkins, M., Dimude, J. U., Howard, J. A. L., Smith, A. J., Dillingham, M. S., Savery, N. J., et al. (2019). Direct removal of RNA polymerase barriers to replication by accessory replicative helicases. *Nucleic Acids Res.* 47, 5100–5113. doi: 10.1093/nar/gkz170

Haycocks, J. R., and Grainger, D. C. (2016). Unusually situated binding sites for bacterial transcription factors can have hidden functionality. *PLoS One* 11:e0157016. doi: 10.1371/journal.pone.0157016

Helmrich, A., Ballarino, M., Nudler, E., and Tora, L. (2013). Transcription-replication encounters, consequences and genomic instability. *Nat. Struct. Mol. Biol.* 20, 412–418. doi: 10.1038/nsmb.2543

Hoffmann, S. A., Hao, N., Shearwin, K. E., and Arndt, K. M. (2019). Characterizing transcriptional interference between converging genes in bacteria. *ACS Synth. Biol.* 8, 466–473. doi: 10.1021/acssynbio.8b00477

Hottes, A. K., Freddolino, P. L., Khare, A., Donnell, Z. N., Liu, J. C., and Tavazoie, S. (2013). Bacterial adaptation through loss of function. *PLoS Genet.* 9:e1003617. doi: 10.1371/journal.pgen.1003617

Hu, P., Janga, S. C., Babu, M., Díaz-Mejía, J. J., Butland, G., Yang, W., et al. (2009). Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS Biol.* 7:e96. doi: 10.1371/journal.pbio.1000096

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

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

Hücker, S. M., Vanderhaeghen, S., Abellan-Schneyder, I., Scherer, S., and Neuhaus, K. (2018a). The novel anaerobiosis-responsive overlapping gene *ano* is overlapping antisense to the annotated gene ECs2385 of *Escherichia coli* O157:H7 Sakai. *Front. Microbiol.* 9:931. doi: 10.3389/fmicb.2018.00931

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

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218–223. doi: 10.1126/science.1168978

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3, 318–356. doi: 10.1016/s0022-2836(61)80072-7

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.

Ju, X., Li, D., and Liu, S. (2019). Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. *Nat. Microbiol.* 4, 1907–1918. doi: 10.1038/s41564-019-0500-z

Kacharia, F. R., Millar, J. A., and Raghavan, R. (2017). Emergence of new sRNAs in enteric bacteria is associated with low expression and rapid evolution. *J. Mol. Evol.* 84, 204–213. doi: 10.1007/s00239-017-9793-9

Keeling, D. M., Garza, P., Nartey, C. M., and Carvunis, A.-R. (2019). The meanings of'function'in biology and the problematic case of de novo gene emergence. *eLife* 8:e47014.

Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H., et al. (2005). Complete set of ORF clones of *Escherichia coli* ASKA library (a complete set of *E. coli* K-12 ORF archive): unique resources for biological research. *DNA Res.* 12, 291–299. doi: 10.1093/dnares/dsi012

Knopp, M., and Andersson, D. I. (2018). No beneficial fitness effects of random peptides. *Nat. Ecol. Evol.* 2, 1046–1047. doi: 10.1038/s41559-018-0585-4

Koskiniemi, S., Sun, S., Berg, O. G., and Andersson, D. I. (2012). Selection-driven gene loss in bacteria. *PLoS Genet.* 8:e1002787. doi: 10.1371/journal.pgen.1002787

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

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

Lasa, I., Toledo-Arana, A., and Gingeras, T. R. (2012). An effort to make sense of antisense transcription in bacteria. *RNA Biol.* 9, 1039–1044. doi: 10.4161/rna.21167

Lejars, M., Kobayashi, A., and Hajnsdorf, E. (2019). Physiological roles of antisense RNAs in prokaryotes. *Biochimie* 164, 3–16. doi: 10.1016/j.biochi.2019.04.015

Lescuyer, P., Hochstrasser, D. F., and Sanchez, J. C. (2004). Comprehensive proteome analysis by chromatographic protein prefractionation. *Electrophoresis* 25, 1125–1135. doi: 10.1002/elps.200305792

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H., et al. (2016). Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci. Adv.* 2:e1501363. doi: 10.1126/sciadv.1501363

Lynch, M., and Marinov, G. K. (2015). The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15690–15695. doi: 10.1073/pnas.1514974112

Ma, N., and McAllister, W. T. (2009). In a head-on collision, two RNA polymerases approaching one another on the same DNA may pass by one another. *J. Mol. Biol.* 391, 808–812. doi: 10.1016/j.jmb.2009.06.060

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12.

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* 74, 481–493.e6. doi: 10.1016/j.molcel.2019.02.017

Mir, K., Neuhaus, K., Scherer, S., Bossert, M., and Schober, S. (2012). Predicting statistical properties of open reading frames in bacterial genomes. *PLoS One* 7:e45103. doi: 10.1371/journal.pone.0045103

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

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

Müller, S. A., Findeiß, S., Pernitzsch, S. R., Wissenbach, D. K., Stadler, P. F., Hofacker, I. L., et al. (2013). Identification of new protein coding sequences and signal peptidase cleavage sites of *Helicobacter* pylori strain 26695 by proteogenomics. *J. Proteomics* 86, 27–42. doi: 10.1016/j.jprot.2013.04.036

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, 193–201. doi: 10.1093/dnares/dsw008

NCBI (2020). *NCBI Prokaryotic Genome Annotation Standards*. Available online at: https://www.ncbi.nlm.nih.gov/genome/annotation_prok/standards/ (accessed 20.02.2020).

Nelson, C. W., Ardern, Z., and Wei, X. (2020). OLGenie: estimating natural selection to predict functional overlapping genes. *Mol. Biol. Evol.* msaa087. 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* 18:216. doi: 10.1186/s12864-017-3586-9

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, 1295–1308. doi: 10.1016/j.cell.2011.10.044

Orr, M. W., Mao, Y., Storz, G., and Qian, S.-B. (2020). Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res.* 48, 1029–1042. doi: 10.1093/nar/gkz734

Owen, S. V., Canals, R., Wenner, N., Hammarlöf, D. L., Kröger, C., and Hinton, J. C. (2020). A window into lysogeny: revealing temperate phage biology with transcriptomics. *Microb. Genomics* 6:e000330.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36, 996–1004. doi: 10.1038/nbt.4229

Pavesi, A. (2019). Asymmetric evolution in viral overlapping genes is a source of selective protein adaptation. *Virology* 532, 39–47. doi: 10.1016/j.virol.2019.03.017

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:e0202513. doi: 10.1371/journal.pone.0202513

Proshkin, S., Rahmouni, A. R., Mironov, A., and Nudler, E. (2010). Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* 328, 504–508. doi: 10.1126/science.1184939

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Raghavan, R., Sloan, D. B., and Ochman, H. (2012). Antisense transcription is pervasive but rarely conserved in enteric bacteria. *mBio* 3:e00156-12.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/s0168-9525(00)02024-2

Robinson-Rechavi, M. (2019). Molecular evolution and gene function. *arXiv*. Availble at: https://arxiv.org/abs/1910.01940#:~{}:text=Functional%20data%20provides%20information%20on,e.g.%2C%20substitutions%20or%20duplications (accessed July 31, 2020).

Sáenz-Lahoya, S., Bitarte, N., García, B., Burgui, S., Vergara-Irigaray, M., Valle, J., et al. (2019). Noncontiguous operon is a genetic organization for coordinating bacterial gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 116, 1733–1738. doi: 10.1073/pnas.1812746116

Saha, D., Podder, S., Panda, A., and Ghosh, T. C. (2016). Overlapping genes: a significant genomic correlate of prokaryotic growth rates. *Gene* 582, 143–147. doi: 10.1016/j.gene.2016.02.002

Satoshi, F., and Nishikawa, K. (2004). Estimation of the number of authentic orphan genes in bacterial genomes. *DNA Res.* 11, 219–231. doi: 10.1093/dnares/11.4.219

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, 1245–1259.e14. doi: 10.1016/j.cell.2019.07.016

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., and Sauer, U. (2012). Multidimensional optimality of microbial metabolism. *Science* 336, 601–604. doi: 10.1126/science.1216882

Sela, I., Wolf, Y. I., and Koonin, E. V. (2019). Selection and genome plasticity as the key factors in the evolution of bacteria. *Phys. Rev. X* 9:031018.

Sharma, H., and Anand, B. (2019). Ribosome assembly defects subvert initiation Factor3 mediated scrutiny of bona fide start signal. *Nucleic Acids Res.* 47, 11368–11386. doi: 10.1093/nar/gkz825

Smith, C., Canestrari, J., Wang, J., Derbyshire, K., Gray, T., and Wade, J. (2019). Pervasive translation in *Mycobacterium tuberculosis*. *bioRxiv* [Preprint]. doi: 10.1101/665208

Stav, S., Atilho, R. M., Arachchilage, G. M., Nguyen, G., Higgs, G., and Breaker, R. R. (2019). Genome-wide discovery of structured noncoding RNAs in bacteria. *BMC Microbiol.* 19:66. doi: 10.1186/s12866-019-1433-7

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

Takeuchi, N., Cordero, O. X., Koonin, E. V., and Kaneko, K. (2015). Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. *BMC Biol.* 13:20. doi: 10.1186/s12915-015-0131-7

Tautz, D., and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12, 692–702. doi: 10.1038/nrg3053

Tenaillon, O. (2014). The utility of Fisher's geometric model in evolutionary genetics. *Annu. Rev. Ecol. Evol. Syst.* 45, 179–201. doi: 10.1146/annurev-ecolsys-120213-091846

Ten-Caten, F., Vêncio, R. Z., Lorenzetti, A. P. R., Zaramela, L. S., Santana, A. C., and Koide, T. (2018). Internal RNAs overlapping coding sequences can drive the production of alternative proteins in archaea. *RNA Biol.* 15, 1119–1132.

Tsuji, J., and Weng, Z. (2016). DNApi: a de novo adapter prediction algorithm for small RNA sequencing data. *PLoS One* 11:e0164228. doi: 10.1371/journal.pone.0164228

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

Venter, E., Smith, R. D., and Payne, S. H. (2011). Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *PLoS One* 6:e27587. doi: 10.1371/journal.pone.0027587

Vishnoi, A., Kryazhimskiy, S., Bazykin, G. A., Hannenhalli, S., and Plotkin, J. B. (2010). Young proteins experience more variable selection pressures than old proteins. *Genome Res.* 20, 1574–1581. doi: 10.1101/gr.109595.110

Wade, J. T. (2015). Mapping transcription regulatory networks with ChIP-seq and RNA-seq. *Adv. Exp. Med. Biol.* 883, 119–134. doi: 10.1007/978-3-319-23603-2_7

Wade, J. T., and Grainger, D. C. (2014). Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat. Rev. Microbiol.* 12, 647–653. doi: 10.1038/nrmicro3316

Wadler, C. S., and Vanderpool, C. K. (2007). A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc. Natl. Acad. Sci. U.S.A.* 104, 20454–20459. doi: 10.1073/pnas.0708102104

Weaver, J., Mohammad, F., Buskirk, A. R., and Storz, G. (2019). Identifying small proteins by ribosome profiling with stalled initiation complexes. *mBio* 10:e02819-18.

Wei, X., and Zhang, J. (2015). A simple method for estimating the strength of natural selection on overlapping genes. *Genome Biol. Evol.* 7, 381–390. doi: 10.1093/gbe/evu294

Weisman, C. M., and Eddy, S. R. (2017). Gene evolution: getting something from nothing. *Curr. Biol.* 27, R661–R663.

Wichmann, S., and Ardern, Z. (2019). Optimality in the standard genetic code is robust with respect to comparison code sets. *Biosystems* 185:104023. doi: 10.1016/j.biosystems.2019.104023

Willems, P., Fijalkowski, I., and Van Damme, P. (2019). Lost and found: re-searching and re-scoring proteomics data aids the discovery of bacterial proteins and improves proteome coverage. *bioRxiv* [Preprint]. doi: 10.1101/2019.12.18.881375

Willis, S., and Masel, J. (2018). Gene birth contributes to structural disorder encoded by overlapping genes. *Genetics* 210, 303–313. doi: 10.1534/genetics.118.301249

Wisotsky, S. R., Kosakovsky Pond, S. L., Shank, S. D., and Muse, S. V. (2020). Synonymous site-to-site substitution rate variation dramatically inflates false positive rates of selection analyses: ignore at your own peril. *Mol. Biol. Evol.* msaa037.

Yang, X., Jensen, S. I., Wulff, T., Harrison, S. J., and Long, K. S. (2016). Identification and validation of novel small proteins in *Pseudomonas putida*. *Environ. Microbiol. Rep.* 8, 966–974. doi: 10.1111/1758-2229.12473

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.