REPARATION: Ribosome Profiling Assisted (Re-)Annotation of Bacterial genomes.

Elvis Ndah[1,2,3], Veronique Jonckheere[1,2], Adam Giess[4], Eivind Valen[4,5], Gerben Menschaert[3] & Petra Van Damme[1,2,*].

[1] VIB-UGent Center for Medical Biotechnology, B-9000 Ghent, Belgium

[2]Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium

[3]Lab of Bioinformatics and Computational Genomics, Department of Mathematical Modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, B-9000 Ghent, Belgium

[4]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen 5020, Norway

[5]Sars International Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, Norway

* To whom correspondence should be addressed: VIB-UGent Center for Medical Biotechnology, Ghent University, A. Baertsoenkaai 3, B9000 Ghent, Belgium. Tel: 32 92649279 Fax: 32 92649496; E-mail: petra.vandamme@vib-ugent.be;

Keywords:

ribosome profiling, machine learning, N-terminomics, bacteria, genome (re-)annotation, open reading frames.

ABSTRACT

Prokaryotic genome annotation is highly dependent on automated methods, as manual curation cannot keep up with the exponential growth of sequenced genomes. Current automated methods depend heavily on sequence context and often underestimate the complexity of the proteome. We developed REPARATION (RibosomeE Profiling Assisted (Re-)AnnotaTION), a *de novo* algorithm that takes advantage of experimental protein translation evidence from ribosome profiling (Ribo-seq) to delineate translated open reading frames (ORFs) in bacteria, independent of genome annotation. REPARATION evaluates all possible ORFs in the genome and estimates minimum thresholds based on a growth curve model to screen for spurious ORFs. We applied REPARATION to three annotated bacterial species to obtain a more comprehensive mapping of their translation landscape in support of experimental data. In all cases, we identified hundreds of novel (small) ORFs including variants of previously annotated ORFs. Our predictions were supported by matching mass spectrometry (MS) proteomics data, sequence composition and conservation analysis. REPARATION is unique in that it makes use of experimental translation evidence to perform *de novo* ORF delineation in bacterial genomes irrespective of the sequence context of the reading frame.

INTRODUCTION:

In recent years, the advent of next generation sequencing has led to an exponential growth of sequenced prokaryotic genomes. As curation based methods cannot keep pace with the increase in the number of available bacterial genomes, researchers have reverted to the use of computational methods for prokaryotic genome annotation (Richardson & Watson 2013; Land et al. 2015). However, advanced genome annotation should entail more than simply relying on automatic gene predictions or transferred genome annotation, as these often introduce and propagate inconsistencies (Richardson & Watson 2013). Moreover, the dependence on sequence contexts of the open reading frame (ORF) by automatic methods often introduces a bias in gene prediction, as studies have shown that translation can occur irrespective of the sequence composition of the ORF (Michel et al. 2012; Fields et al. 2015). Further, gene prediction methods that depend solely on the genomic template often lack the capabilities to capture the true complexity of the translation landscape (Fields et al. 2015), overall stressing the need for non *in silico* based gene prediction approaches.

Ribosome profiling (Ingolia et al. 2009) (Ribo-seq) has revolutionized the study of protein synthesis in a wide variety of prokaryotic and eukaryotic species. Ribo-seq provides a global measurement of translation *in vivo* by capturing translating ribosomes along an mRNA. More specifically, ribosome protected mRNA footprint (RPFs) are extracted and converted into a deep sequencing cDNA library. When aligned to a reference genome, these RPFs provide a genome-wide snapshot of the positions of translating ribosomes along the mRNA at the time of the experiment (Ingolia et al. 2009). This genome-wide positional information of translating ribosomes allows for the identification of translated regions. With the advent of Ribo-seq, numerous computational methods have been developed to detect putatively translated regions in eukaryotes, all taking advantages of inherent Ribo-seq based metrics to identify translated ORFs. In the studies of Lee et al. (2012) and Crappé et al. (2014), a rule based peak detection algorithm was used to identify translation initiation sites, while Bazzini et al. (2014) and Calviello et al. (2015) take advantage of the triplet periodicity property of Ribo-seq data. Fields et al. (2015) and Chew et al. (2013) developed an ensemble classifier that aggregate multiple features to predict putative coding ORFs. No computational method has yet been reported to delineate ORFs in prokaryotic genomes based on Ribo-seq data. In this work, we aimed at developing an algorithm that makes use of experimental evidence from Ribo-seq to perform *de novo* ORF delineations in prokaryotic genomes.

Our algorithm, REPARATION (RibosomeE Profiling Assisted (Re-)AnnotaTION) trains an ensemble classifier to learn Ribo-seq patterns from a set of confident protein coding ORFs for a *de novo* delineation of ORFs in bacterial genomes. REPARATION deduces intrinsic characteristics from the data and thus can be applied to Ribo-seq experiments targeting elongating ribosomes. We evaluated the performance of REPARATION on three annotated bacterial species. REPARATION was able to identify putative coding ORFs corresponding to previously annotated protein coding and non-protein coding regions, variants of annotated ORFs (i.e. in-frame truncations or 5′ extensions) and intergenic ORFs. We validated our findings using matching proteomics, sequence composition and phylogenetic conservation analyses.

RESULTS:

To assess the performance and utility of our REPARATION algorithm (*figure 1A*), besides two publically available bacterial Ribo-seq datasets from *Escherichia coli K12 str. MG1655* and *Bacillus subtilis subsp. subtilis str. 168* (Li et al. 2014), we generated ribosome profiling data and matching RNA-seq data from a monosome and polysome enriched fraction (Heyer & Moore 2016) of *Salmonella enterica* serovar Typhimurium strain SL1344 (experimental details in *supplementary methods*).

REPARATION starts by traversing the entire prokaryotic genome sequence to generate all possible ORFs that have an arbitrary length of at least 10 codons (30nt) and initiate with either an ATG, GTG or TTG codon (the most frequently used start codons in a variety of prokaryotic species (Panicker et al. 2015)) until the next in-frame stop codon. REPARATION applies a random forest classifier trained on features derived from the meta-gene profile of known protein coding ORFs (*figure 1B*). These features encompass *1)* the start region (first 45nt of the ORF) read density, *2)* the stop region (last 21nt) read density (Fields et al. 2015) *3)* ORF RPF coverage refers to the proportion of nucleotides within the ORF covered by positional RPF reads (Chew et al. 2013), *4)* start region RPF coverage, i.e. the proportion of nucleotides within the start region covered by RPF reads, 5*)* the ratio of the average RPF read count within the start region divided by the average RPF read count within the rest of the ORF and *6)* ribosome binding site (RBS) energy (see *supplementary methods*). The classifier's training set consisted of positive examples generated using a comparative genomic approach. First we used prodigal (Hyatt et al. 2010) to generate a set of ORFs, which were subsequently BLAST searched against a curated set of bacterial protein sequences from UniProtKB-SwissProt, ORFs with e-values less than 10^{-5} and a minimum identity score of 75% were retained. Viewing their infrequent occurrence as translation starts (<0.01%) in the annotations of the interrogated species, the negative set consisted of all CTG-starting ORFs (*supplementary table T1)*. The algorithm then estimates a minimum read density and ORF RPF coverage to discard spurious ORFs by exploiting the sigmoid relationship between these features (*figure 1C*). Using a four parameter logistic regression curve on the positive set, REPARATION estimates the lower bend point of the fitted curve representing a two dimensional threshold (read density and ORF RPF coverage). All ORFs with read density and ORF RPF coverage below these thresholds (*supplementary table T2)* were discarded, including those in the training set. When trained on these sets, the random forest classifier achieved on average an 89, 90 and 92% 10-fold cross validation accuracy with area under the curve values of 0.93, 0.93 and 0.95 for the *Salmonella, E. coli* and *Bacillus* data sets respectively (*supplementary figure S1 A*).

Of the three species evaluated, REPARATION mapped putative coding ORFs corresponding to regions annotated as protein coding, as well as to non-coding and intergenic.

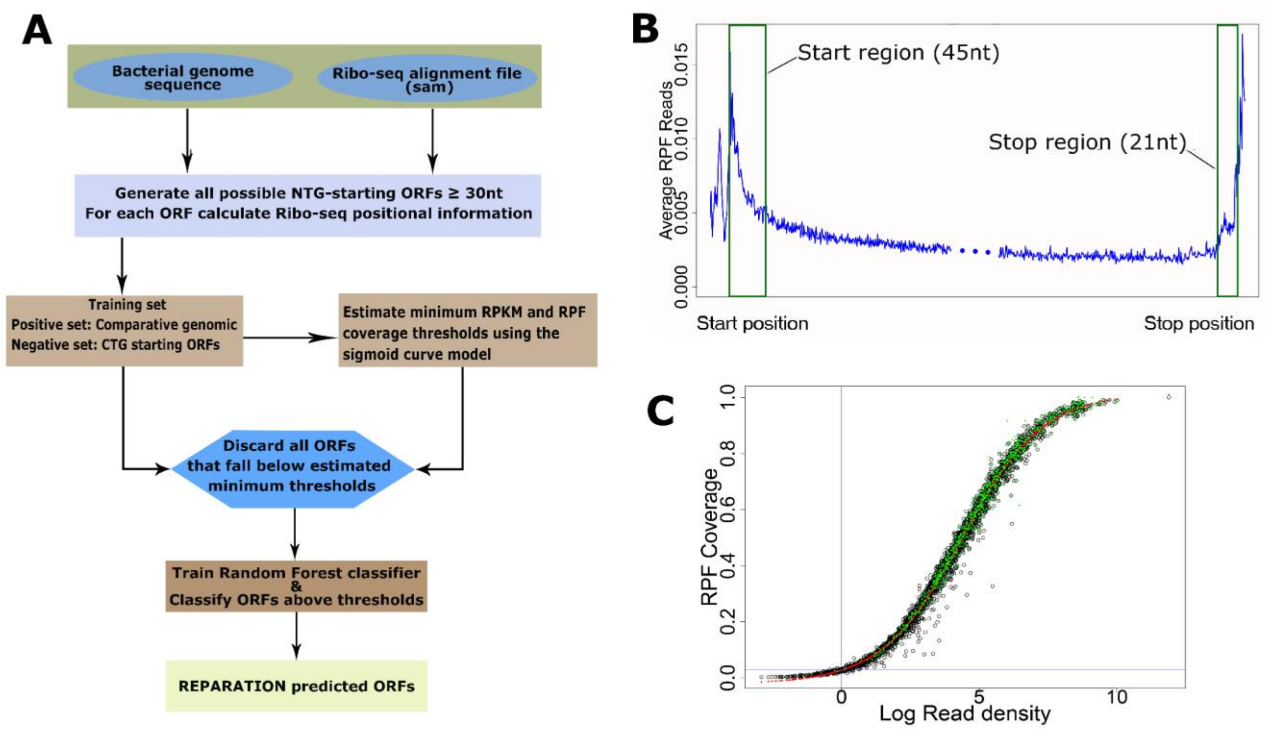

***Figure 1: REPARATION pipeline for de novo ORF delineation in prokaryotes**. (A) REPARATION workflow diagram. The entire prokaryotic genome is traversed and all possible NTG-starting ORFs are generated. Next, ORF-specific positional Ribo-seq signal information is calculated based on the metagenic gene profile (B). To discard spurious ORFs, the minimum log2 RPKM and ORF RPF coverage thresholds are estimated using a four parameter logistic S-curve (C). (B) Metagenic profile of salmonella data indicating read accumulation at the start and stop of ORFs (stitched together in the middle for visualization purposes). (C) S-curve with fitted four parameters logistic curve (red) and indication of predicted ORFs with support from N-terminal proteomics data (green) in the case of E. coli.*

REPARATION-predicted ORFs predominantly match to, or overlap with annotated ORFs and follow the reference model of start codon usage.

Viewing the previously reported similarities in the translation properties of monosomes and polysomes (Heyer & Moore (2016)) and the high correlation observed between the two samples (*supplementary figure S1 B*), we considered the *Salmonella* monosome and polysome samples as replicate samples for the purpose of translated ORF delineation.

For *Salmonella*, REPARATION predicted a total of 3868 and 3648 putative ORFs in the monosome and polysome sample respectively. Of these, 3267 (90%) ORFs found common in both datasets were considered as the high confident ORF set (*supplementary file F1*). For *E. coli*, a high confident set of 3149 (90%) was selected based on the 3518 and 3504 predicted ORFs in replicate samples 1 and 2 respectively (*supplementary file F2*). Thirdly, in the *Bacillus* sample, 3239 putative coding ORFs were predicted (*supplementary file F3*).

From the high confident set of predicted ORFs in *Salmonella* and *E. coli*, 83% (2696) and 89% (2806) correspond to previously annotated ORFs (respectively), while 84% (2734) of the *Bacillus* predicted ORFs corresponds to previously annotated ORFs (*figure 2*). 15, 8 and 14% of predicted ORFs in the *Salmonella, E. coli and Bacillus* samples (respectively), correspond to variants of previously annotated ORFs, potentially giving rise to N-terminally truncated or extended protein variants referred to as N-terminal proteoforms (Gawron et al. 2014). Consequently, in all three species, ≤ 3% belong to novel putative coding regions.

On average the truncations were 26, 25 and 19 codons downstream of the annotated starts while the extensions where 18, 18 and 15 codons upstream for *Salmonella, E. coli* and *Bacillus (*respectively). Of note, 14, 21 and 18 of the predicted variants are only 1 codon off from the annotated starts in *Salmonella, E. coli* and *Bacillus* respectively (*supplementary table T3*).

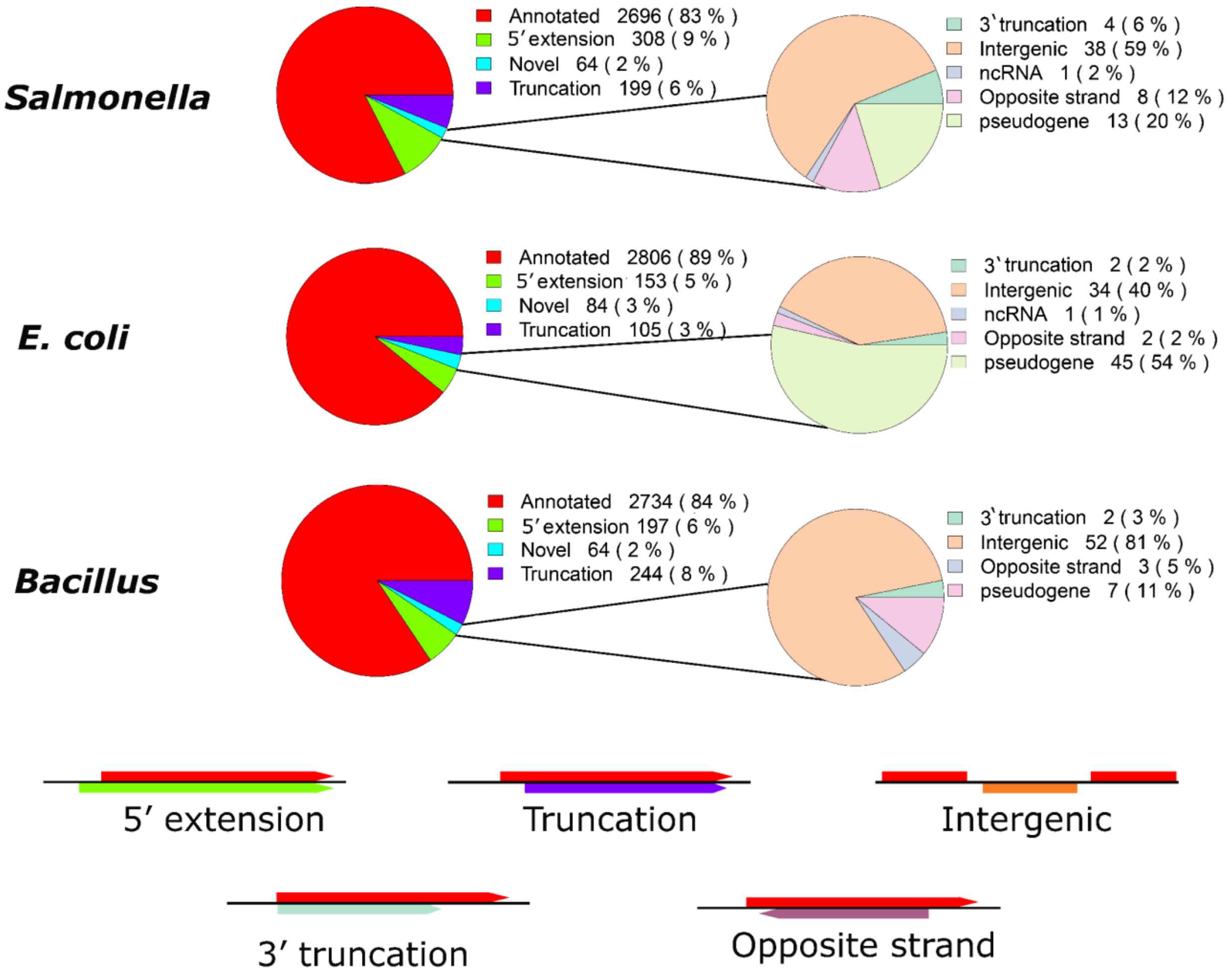

***Figure 2:** Proportion of REPARATION predicted ORFs per ORF category for the high confident ORF sets in Salmonella and E. coli as well as for the Bacillus predictions.*

Overall, 69%, 74% and 76% (including the variants) of all *ENSEMBL* annotated protein coding ORFs in *Salmonella, E. coli* and Bacillus (respectively) were predicted by REPARATION.

In our evaluation of REPARATION, we allow for the three commonly used start codons in prokaryotes ATG, GTG and TTG as translation initiation triplet. Of note however, REPARATION was designed without any bias in start codon selection for ORF prediction. The order of start codon usage over all predicted ORFs are consistent with the standard model for translation initiation in the *ENSEMBL* annotation of the corresponding species i.e. in case of *Salmonella* and *E. coli*, a preference of ATG over GTG and TTG and ATG>TTG>GTG in *Bacillus* (*table 1*) could be observed.

	ENSEMBL Annotation	**All predictions**	**Matching Annotated**	**Extensions**	**Truncations**	**Novel**
Salmonella						
ATG	4093 (88.0%)	2809 (86%)	2460 (91.2%)	168 (55%)	130(65%)	51 (80%)
GTG	429 (9.20%)	318 (10%)	199 (7.4%)	71 (23%)	44(22%)	4 (6%)
TTG	126 (2.70%)	140 (4%)	37 (1.4%)	69 (22%)	25(13%)	9 (14%)
E. coli						
ATG	3747 (90.1%)	2736 (87%)	2549 (91%)	66 (43%)	66 (63%)	55 (65%)
GTG	386 (9.2%)	273 (19%)	204 (7%)	33 (22%)	19(18%)	17 (20%)
TTG	71 (2.0%)	139 (4%)	53 (2%)	54 (35%)	20 (19%)	12 (14%)
Bacillus						
ATG	3253(77.7%)	2409 (74%)	2174 (80%)	80 (41%)	119 (49%)	36(48%)
GTG	386 (9.2%)	352 (11%)	234 (9%)	59 (30%)	50 (20%)	9 (20%)
TTG	529 (12.6%)	478 (15%)	326 (11%)	58 (29%)	75 (31%)	19(32%)

Table 1: Start codon usage distribution of the predicted putative coding ORFs. *The predicted ORFs in all three species follow the starts codon usage distributions of the corresponding species annotation. In case of Salmonella* and *E. coli, only ORFs from the high confident set were considered.*

Interestingly however, we observe that novel and variant ORFs are enriched for near-cognate start codons when compared to annotated ORFs. In case of the variants, this bias is most likely due to the preference of automatic gene prediction methods to select a neighbouring ATG as the start codon (Salzberg et al. 1998; Hyatt et al. 2010).

Novel ORFs are evolutionary conserved and display similar amino acid sequence patterns as compared to annotated ORFs.

To gain insight into the novel predictions, we analyzed and compared their evolutionary conservation pattern to that of predicted annotations. Novel and extended ORFs exhibit similar conservation patterns to annotated ORFs, with higher nucleotide conservation from the start codon onwards and within the upstream ribosomal binding site or Shine Dalgarno region positioned -15 to -5nt upstream of the predicted start (*figure 3*), a region aiding in translation initiation by its base pairing with the 3'-end of rRNA (Shultzaberger et al. 2001; Suzek et al. 2001). The higher conservation and triplet periodicity observed upstream of the truncations is likely due to the fact that in some cases, multiple forms of the gene (i.e. N-terminal proteoforms) maybe (co-)expressed (*supplementary table T4*). A manual inspection of the alignments indeed indicates that different forms of the genes are expressed across different species. Of the 66 truncations used in the *Salmonella* conservation analysis, 45% shows evidence of the existence of multiple forms across different bacterial species, while in case of *E. coli* and *Bacillus* these percentages were 40 and 28 from 26 and 25 truncations respectively.

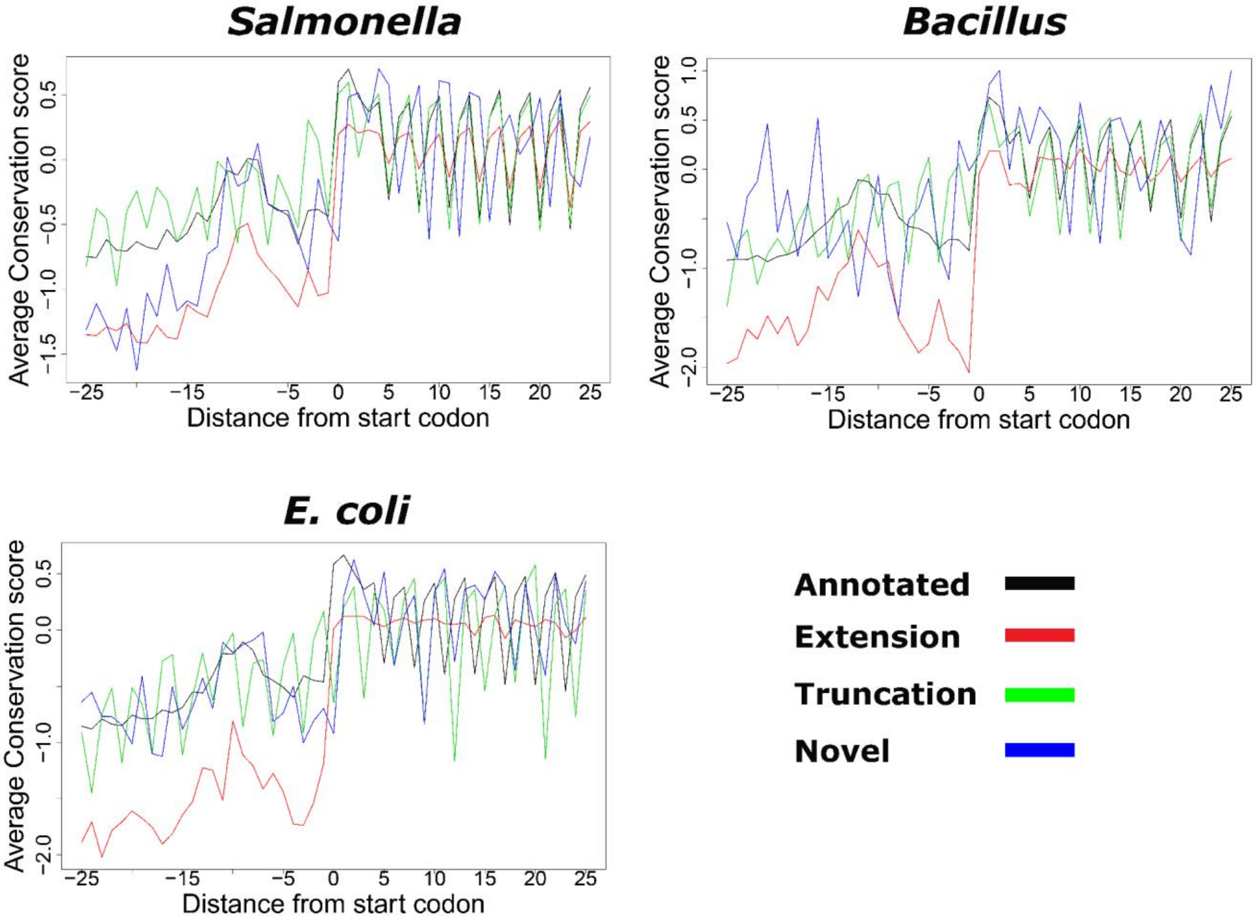

Figure 3: *Conservation pattern of REPARATION predicted ORFs. Nucleotide conservation scores are calculated using the Jukes cantor conservation matrix for nucleotides. Site conservation scores are calculated using the rate4site algorithm and*

displayed for a +/- 25nt window around the predicted start site. The site conservation score was calculated only for ORFs with at least 5 orthologous sequences from a collection of randomly selected bacteria protein sequences from species within the same family as Salmonella/E. coli and Bacillus and outside the family. 833 annotated, 203 extensions, 66 truncations and 11 novel ORFs had at least 5 orthologous sequences in case of Salmonella, while the E. coli profile consisted of 2359 annotated ORFs, 49 extensions, 26 truncations and 19 novel ORFs were considered. In the case of Bacillus there are 1886 annotated, 23 extensions, 25 truncations and 2 novel ORFs

Of the 64 novel ORFs predicted in *Salmonella*, 61% (39) had at least one reported orthologous sequence (*supplementary file F1*). While 60 (71%) and 30 (47%) in *E. coli* and *Bacillus* (respectively) had at least one orthologous sequence (*supplementary files F2 & F3*).

To further confirm that the newly identified ORFs do not represent random noise, we compared the amino acid composition of predicted annotations to that of novel putative coding ORFs and to a set of randomly generated amino acid sequences of equal lengths to predicted ORFs matching annotations. In all three species we observe a very high correlation (≥0.90) between the amino acid compositions of novel and annotated ORFs (*supplementary figure S2*). While a generally poor correlation (≤0.19) was observed when comparing novel or annotated ORFs against the random set of ORFs.

Since evolutionarily conserved significant biases in protein N- and C-termini were previously reported for pro- as well as eukaryotes, often with pronounced biases at the second amino acid positions (van Damme et al. 2011; Palenchar 2008), we next investigated whether the amino acid usage frequency at position two of the novel and re-annotated ORFs exhibited a similar pattern to that of annotated ORFs. Compared to amino acid frequency in the species proteome, clearly the overall distribution is similar for the two ORF categories. More specifically, a significant enrichment of Lys (about 3-fold) at the second amino acid position was observed in case of all three species analysed. For *Salmonella* and *E. coli*, Ser and Thr was equally enriched while in *Bacillus* Asn was slightly more frequent in the second position while other amino acids are clearly underrepresented (i.e. Trp and Tyr), all observations in line with previous N-terminal biases observed (Palenchar 2008) (*supplementary figure S3*).

Proteomics assisted validation of REPARATION predicted ORFs.

To validate our predicted ORFs we generated N-terminal and shotgun proteomics data from matching *E. coli* and *Salmonella* samples respectively. While N-terminomics enables the isolation of N-terminal peptides, making it appropriate for the validation of translation initiation events, shotgun proteomics provides a more global assessment of the expressed proteome. Three different proteome digestions were performed in the shotgun experiment to increase proteome coverage. The shotgun and N-terminal proteomics data were searched against a six frame translation database of the *E. coli* and *Salmonella* genomes. In both experiments, and based on identified peptides, longest non-redundant peptide sequences were aggregated to map onto the REPARATION predictions.

In case of Salmonella, 10751 unique peptides belonging to 2235 ORFs in the six frame translation database were identified by means of shotgun proteomics. Of these, 91% (9723) correspond to 1762 REPARATION predicted ORFs (*figure 4A*), the 9% missed by REPARATION mostly represent lowly expressed ORFs (*Supplementary figure S4 A*). While the vast majority of shotgun peptides support previously annotated regions (*figure 4B*), we additionally identified peptides in support of novel ORFs and ORF reannotations (i.e. N-terminal protein extensions). More specifically, supportive evidence was found in the case of 6 novel ORFs and 19 extensions having at least one identified peptide with a start position upstream of the annotated start (*supplementary file F1*).

For *E. coli,* N-terminal proteomics identified a total of 785 blocked N-terminal peptides that are compliant with the rules of initiator methionine processing (see *supplementary methods*) belonging to 781 ORFs. Assuming that none of these ORFs have multiple initiation sites we choose the most upstream N-terminal peptide and overlapped these with the REPARATION predictions. Of the 781 ORFs with peptide support 720 pass the S-curve estimated minimum thresholds. 86% (621) of these matched REPARATION predicted N-termini (*figure 4C & D*), while in 6% of the cases, a different translation start was predicted by REPARATION 11 downstream (with an average distance of 10 codons) and 36 upstream (with an average distance of 86 codons) from the N-termini peptides. The remaining 8%, not predicted by REPARATION, mainly represent lowly expressed ORFs (*supplementary figure S4 B*). The majority of

N-terminal supported ORFs matched annotations, while 17 correspond to re-annotations or novel ORFs (8 extensions, 7 truncations and 2 novel). We also assessed the predicted ORFs against the 917 *E. coli K-12 Ecogenes* verified protein coding sequences, a set consisting of proteins sequences with their mature N-terminal residues sequenced using Edman sequencing (Krug et al. 2013). Of these, 888 pass the estimated minimum thresholds of which 89% (788) matched REPARATION predicted ORFs (*figure 4E*). REPARATION predicted a different start sites for 40 of *Ecogene* verified ORFs, 34 upstream (average distance of 24 codons) and 6 downstream (average distance of 8 codons).

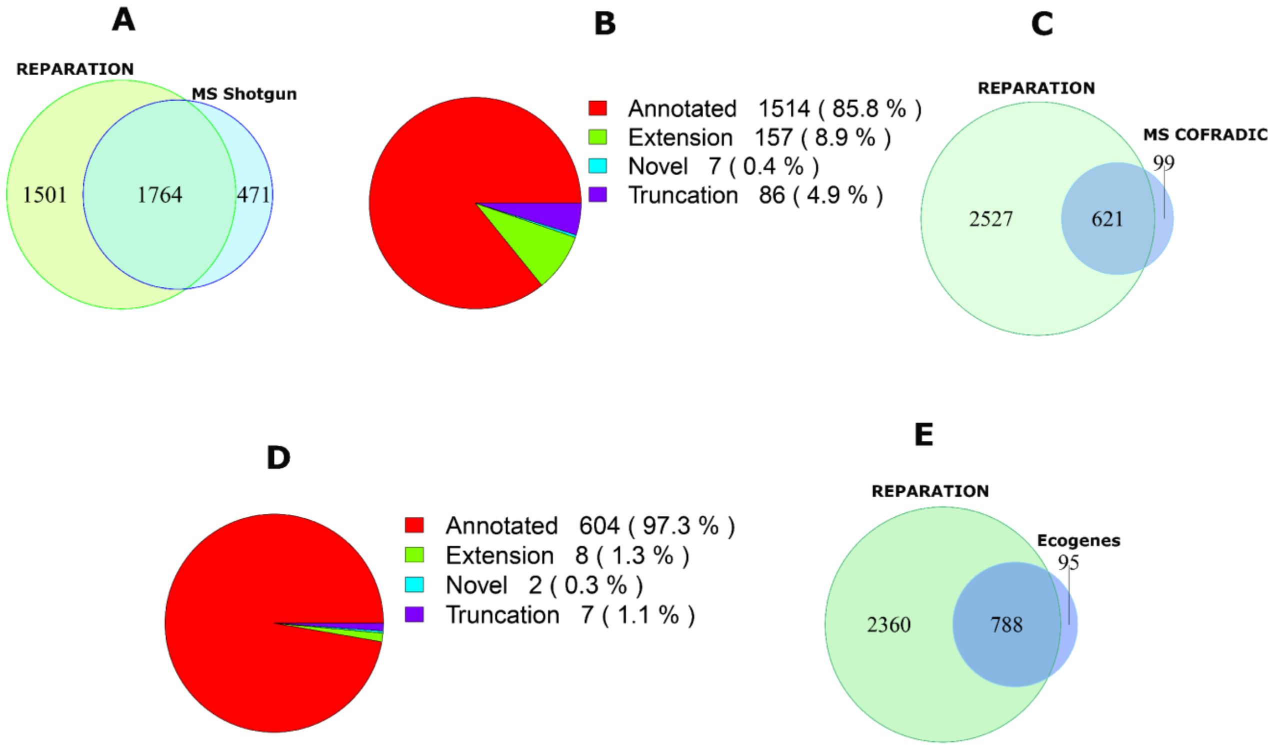

Figure 4: *MS validation of REPARATION pipeline. A) Overlap between the protein sequences identified from shotgun proteomics and the REPARATION predicted ORFs in Salmonella. B) The number of ORFs per category with at least one identified peptide for the high confident set of Salmonella predicted ORFs. C) Overlap between ORFs with N-terminal peptide support and REPARATION predicted ORFs in E. coli. D) Number of predicted ORFs for each category with N-terminal peptide support in the E. coli high confident set. E) Overlap between REPARATION predicted ORFs and the Ecogene verified E. coli ORFs.*

REPARATION in the aid of genome re-annotation.

In case of the three species-specific translatomes analyzed, REPARATION uncovered novel putative coding genes in addition to extensions and truncations of previously annotated genes with supporting proteomics and conservation evidence. More specifically, in the case of the gene *adhP (Salmonella*), REPARATION predicts that translation initiates 27 codons upstream of the annotated start, this ORF extension is supported by an N-terminal peptide identification (*figure 5A*), the corresponding sequence of which is conserved (*supplementary figure S5 A*). N-terminal peptide support, next to the clear lack of Ribo-seq reads in the region between the novel and annotated start (*figure 5B)* of gene *yidR* (*E. coli*), also points to translation initiating 11 codons downstream of the annotation start as predicted by REPARATION. A novel putative coding gene was found matching the intergenic region *Chromosome:2819729-2820319* (*Salmonella)* with Ribo-seq and RNA-seq signals complemented by two unique peptide identifications (*figure 5C*).

Of note, there are currently 72, 182 and 70 annotated pseudogenes in the current *ENSEMBL* annotations of *Salmonella, E. coli* and *Bacillus* (respectively). REPARATION predicted conserved putative coding ORFs within 12, 34 and 7 pseudogene regions leading to 13, 45 and 7 predicted ORFs for *Salmonella, E. coli* and *Bacillus* (respectively). Since pseudogenes in bacteria are typically modified/removed rapidly, coupled with the fact that only uniquely mapped reads were allowed, the observed conservation with the existence of functional orthologues points to the genuine coding potential of these loci and thus functional importance of their translation product (Goodhead & Darby 2015; Lerat & Ochman 2005). One representative example is the identified putative coding ORF in the *sugR* pseudogene (*Salmonella*) which is supported by 3 unique peptide identifications (*figure 5D*).

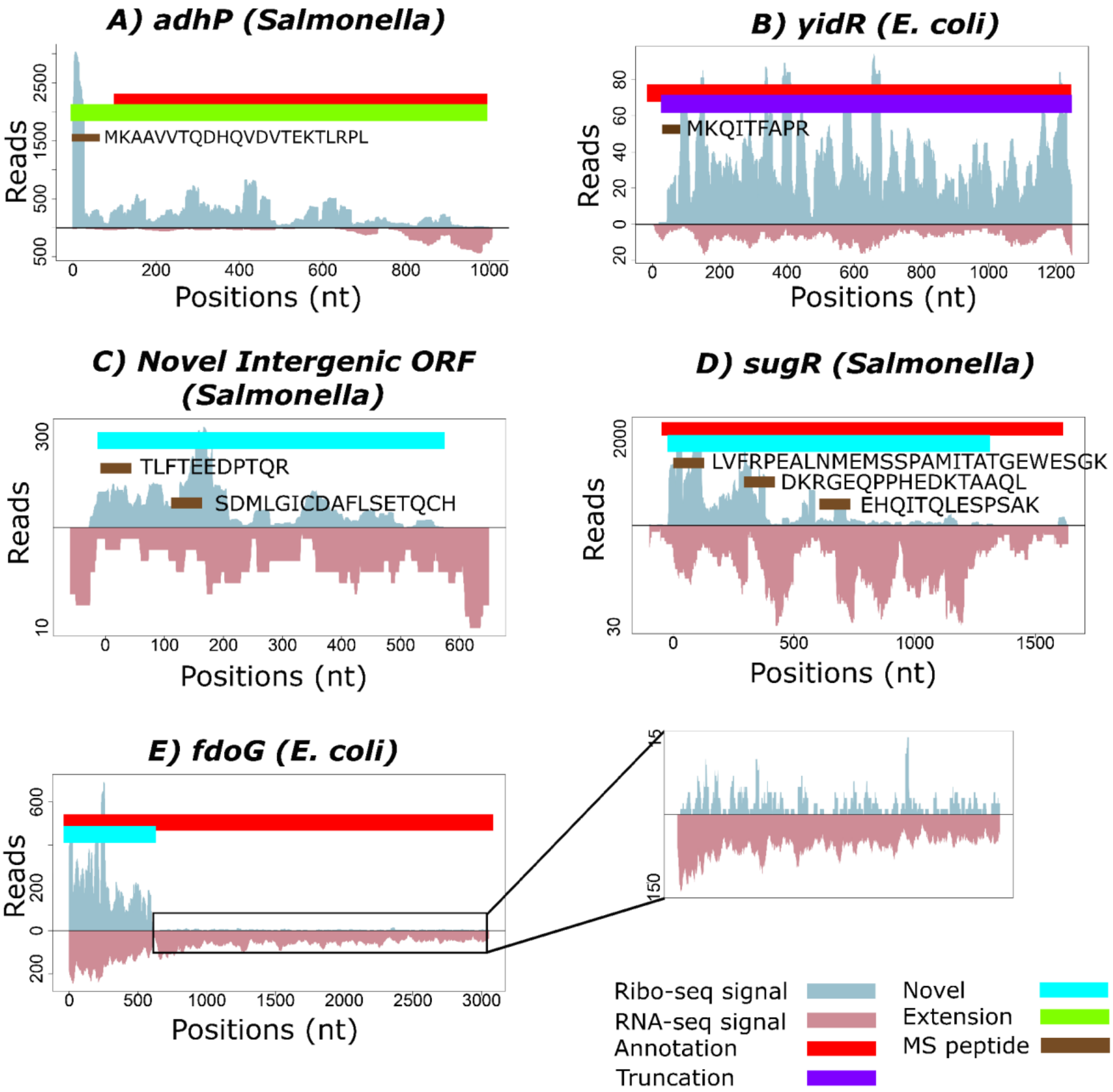

Figure 5: ***REPARATION assisted reannotation of bacterial genomes****. A) REPARATION predicted 5' extension of the adhP gene (Salmonella) with supportive peptide evidence mapping upstream and in-frame with the annotated ORF. B) Gene yidR (E. coli) predicted as a 5' truncation with N-terminal peptide support and Ribo-seq reads starting downstream of the annotated start. C) Novel putative coding intergenic ORF in the region Chromosome:2819729-2820319 (Salmonella) with supportive peptide evidence. D) Evidence of translation within pseudogene sugR (Salmonella), with two matching peptide identification. E) Putative co-expression of a 3' truncated ORF as*

well as its 3′ extended counterpart due to a frameshifting event occurring during translation of the fdoG gene (E. coli). A magnification of the region beyond the stop codon displays a continuous, though ~100-fold lower pattern of Ribo-seq reads indicative of stop codon read-through.

Interestingly, in case of *fdoG (E. coli)*, REPARATION predicted two juxtapositioned ORFs, both contained within a previously annotated ORF with a stop codon read through event (*figure 5E*). In *E. coli* three other such read through events have been reported for genes *fdnG*, *fdhF* and *prfB*. The Ribo-seq read density within the C-terminal truncated ORF is about 100 fold higher as compared to the 5′ truncated ORF while only a 3-fold difference in RNA-seq density could be observed. The RNA-seq evidence supports the presence of the stop codon TGA at the end of the C-terminal truncated ORF. The observed continuous Ribo-seq signal indicative of translation beyond the sequencing-verified stop codon most likely points to a stop codon read through event (Feng et al. 2012). The so-called 3′ and 5′ truncations of the current annotation predicted by REPARATION are likely due to the algorithm not allowing for stop codon read through. A similar trend was observed in case of its *Salmonella* orthologue, with the Ribo-seq signal and RNA-seq 30- and 2-folds (respectively) higher for the C-terminal truncated ORF than the longer N-terminal truncated ORF (*supplementary figure S5 B*).

REPARATION in the aid of small ORF annotation

Small ORFs have historically been ignored in most *in silico* predictions because of the assumption that they can easily occur by chance due to their small size (Hyatt et al. 2010). As 71 codons is the average length when considering the length of the 5% shortest annotated ORFs in the 3 species, we here arbitrarily define a sORF as a translation product with a length of ≤71 codons. In *Salmonella*, REPARATION predicted 95 putative coding sORFs. Of these, 24 (25%) represent novel ORFs including 1 extension, 12 truncations and 62% (59) matched annotations. Supportive proteomics data was found for 28 predicted sORFs. While in *E. coli* and *Bacillus the* algorithm predicted 112 (90 (80%) matching annotations, 1 extension, 5 truncations and 16 novel) and 223 (154 (69%) matching annotations, 6 extensions, 15 truncations and 48 novel) sORFs. An interesting example of a possible re-annotation of gene *yfaD* (*E. coli*) is the REPARATION predicted 56 codon sORF representing a truncated form (*figure 6A*). In line with transcriptional data pointing to transcription of an mRNA not encompassing the annotated ORF, Ribo-seq indicates expression of a smaller ORF of which the start of the gene is located 243 codons downstream of the annotated start. Other representative examples are the intergenic 47 codons long sORF *Chromosome:2470500-2470643 (E. coli) (figure 6B),* the *30* codon long sORF located on the reverse strand of the *fre* gene (Salmonella) (*figure 6C*) and a 57 codon long intergenic *Bacillus* sORF that overlaps with the CDS of the sORF-encoding *hfq* gene (*figure 6D*).

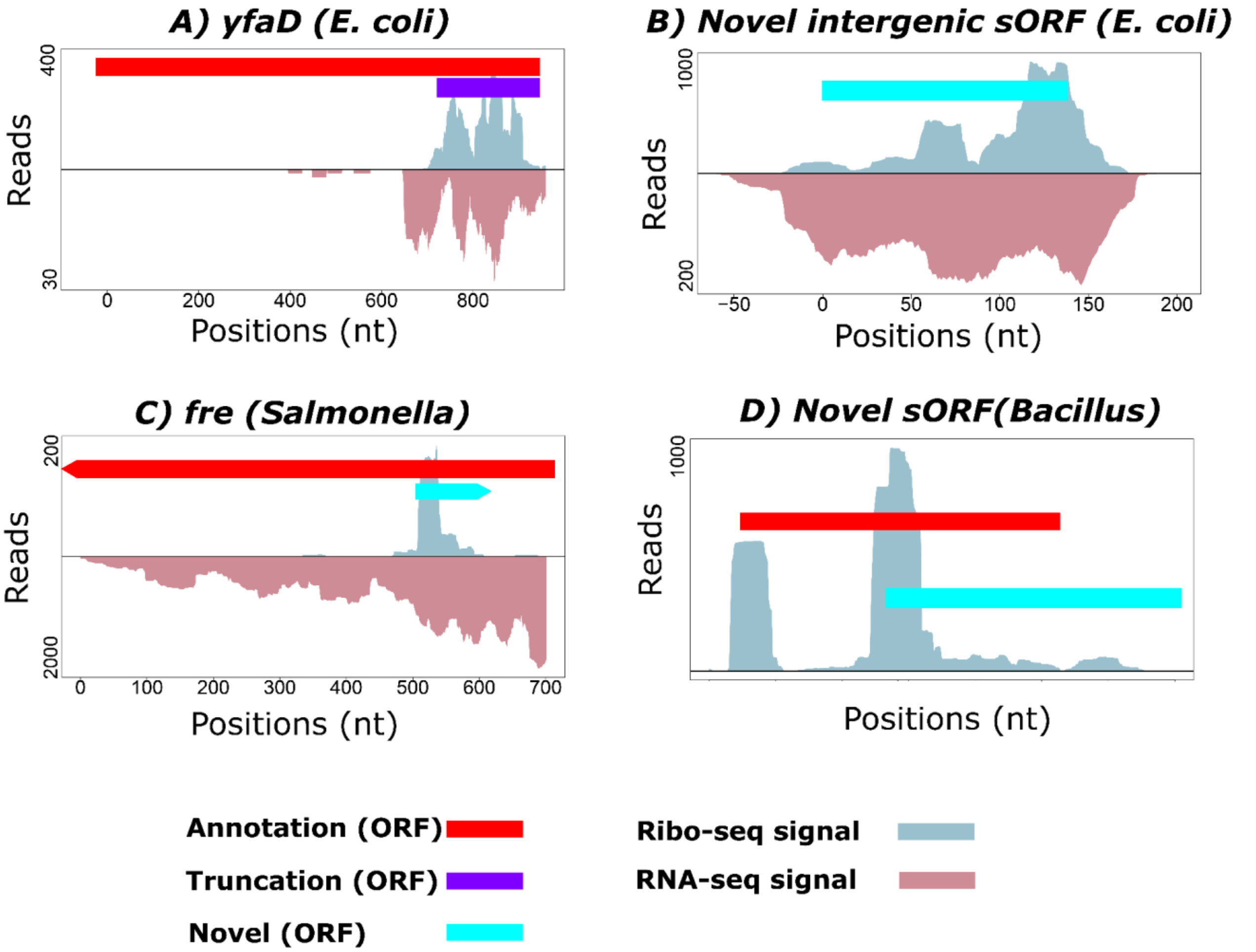

Figure 6: Novel sORFs predicted by REPARATION. *Ribo-seq and RNA-seq profiles indicate expression of A) a truncated form of the annotated yfaD (E. coli) gene B) a 47 codons sORF matching the region Chromosome:2470500-2470643 (E. coli.) C) a sORF encoded on the reverse strand encoding the fre gene (Salmonella). D) a sORF*

Chromosome:1867485-1867655 (Bacillus) that partially overlaps the annotated fhq sORF (Bacillus). The Ribo-seq profiles indicate translation initiation in another frame.

DISCUSSION

Experimental signals from ribosome profiling exhibit patterns across protein coding ORFs which can be exploited to accurately delineate translated ORFs. Although Ribo-seq is not completely standardized (Diament & Tuller 2016) and certain experimental procedures such as treatments (e.g. no treatment versus antibiotic treatment) tend to have a noticeable influence on the translation patterns observed (Calviello et al. 2015), we here developed an algorithm that enables a *de novo* delineation of translated ORFs in bacterial genomes. Our algorithm, delineates putative protein coding ORFs in bacterial genomes using experimental information deduced from Ribo-seq, aiming to minimize biases inherent to *in silico* prediction methods.

We applied REPARATION on three annotated bacterial species to illustrate its ability to predict putative coding regions. Multiple lines of evidence, including proteomics data, evolutionary conservation analysis and sequence composition suggest that the REPARATION-predicted ORFs represent *bona fide* translation events. As expected, the majority of predicted ORFs agreed with previous annotations, but additionally we were able to detect a multitude of ORF updates next to novel translated ORFs mainly within intergenic and pseudogene regions. While we clearly observed a shift towards near-cognate versus cognate start selection for the novel predictions, we nonetheless observe that the order of start codon usage follows the standard model in the respective species. Perhaps unsurprisingly viewing the difficulty to predict short ORF using classical gene predictions, the novel ORFs predicted by REPARATION are predominantly shorter than those previously annotated. Our predictions also point to possible errors in the current start site annotation of some genes, resulting in the identification of N-terminal truncations and extensions. The predicted extensions exhibit a similar conservation pattern to annotated ORFs while a higher conservation and triplet periodicity upstream of the truncated predictions (*figure 3*) is likely due to the expression of multiple proteoforms across species. The identification of multiple TIS-indicative N-termini in our *E. coli* N-terminomics dataset point to the existence of multiple translation initiation sites in at least 10 genes (*supplementary table T4*), likely an underrepresentation due to the low steady-state levels of N-terminally formylated N-termini. The former observation is in line with the recently revealed and until then highly underestimated occurrence of alternative translation events in eukaryotes (Ingolia et al. 2011; Van Damme et al. 2014). It is noteworthy that we identified 11 genes with multiple TIS-evidence from the N-terminomics data (*supplementary table T4 and figure S4 C*), in case of REPARATION however, only a single ORF is selected per ORF family.

A substantial portion of the novel ORFs, with at least one identified orthologous gene, overlaps with known pseudogene loci. By virtue of the fact that pseudogenes in bacteria tend to be (sub)genus-specific and are rarely shared even among closely related species (Goodhead & Darby 2015; Lerat & Ochman 2005), it is likely that (part of) these genes have retained their protein coding potential, a finding that is further corroborated by proteomics data. The relatively fewer peptide identifications corresponding to the translation products of novel ORFs may in part be due to the difficulty of identifying these by MS, mainly because of their predominantly shorter nature and thus likely lower number of peptides (Fields et al. 2015). An *in silico* analysis of the identifiable tryptic peptide coverage shows that on average 85% of the annotated protein sequences are covered by identifiable tryptic peptides while on average only 69% of the novel ORFs are covered by identifiable tryptic peptides. Furthermore, bacterial translation products of sORFs have previously been shown to be more hydrophobic in nature and therefore extraction biases might also (in part) contribute to their underrepresentation in our proteomics datasets (Hemm et al. 2008).

Historically sORFs have been neglected both in eukaryotes as well as prokaryotes. However, recently renewed interest has been directed toward the identification and characterization of sORFs (Andrews & Rothnagel 2014; Bazzini et al. 2014; Olexiouk et al. 2016). Small proteins represent a particularly difficult problem because they often yield weak statistics when performing computational analysis, making it difficult to discriminate protein coding from non-protein coding small ORFs (Samayoa et al. 2011; Pauli et al. 2015). Exemplified by the identification of tens of sORFs (with supportive metadata), REPARATION's utilization of Ribo-seq signal pattern at least in part alleviates the pitfalls of traditional bacterial gene prediction algorithms concerning the identification of sORFs.

Based on matching N-terminal proteomics evidence and the sequenced N-terminals from *Ecogene,* REPARATION accurately predicts 86 and 89% of the ORFs with experimental evidence. Overall, the high correlative second amino acid frequency patterns observed when comparing annotated versus re-annotated/new ORFs ORFs provide further proof of the accuracy and resolution of start codon selection in case of REPARATION predicted ORFs. Nonetheless, start site selection by REPARATION resulted in a loss of 6% of the N-terminal supported gene starts which exceeded the S-curve thresholds. While the existence of multiple N-terminal proteoforms in bacteria in contrast to the single ORF selection by REPARATION is likely the main explanatory reason for this inconsistency. The discrepancy between predicted and N-terminally supported start, might especially in the case of short truncations also be contributed (in part) by the lack of accuracy of start codon selection. REPARATION could potentially take advantage of improved measures or features to increase the prediction power of the classifier. At present REPARATION is the first attempt to perform a *de novo* putative ORF delineation in prokaryotic genomes that relies on Ribo-seq data. With automated bacterial gene prediction algorithms estimated to have false prediction rate of up to 30% (Angelova et al. 2010), machine learning algorithms that learn properties from Ribo-seq experiments such as REPARARTION pave the way for a more reliable (re-)annotation of prokaryotic genomes.

SOFTWARE

REPARATION software is available at https://github.com/Biobix/REPARATION.

ACCESSION NUMBERS

Ribo-seq and RNA-seq sequencing data reported in this paper have been deposited in NCBI's Gene Expression Omnibus with the accession number GSE91066.

All MS proteomics data and search results have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaino et al. 2016) partner repository with the dataset identifier PXD005844 for the *Salmonella typhimurium SL1344* datasets and PXD005901 for the *E. coli* K12 str. MG1655 dataset. Reviewers can access the *Salmonella* datasets by using 'reviewer61164@ebi.ac.uk' as username and 'Zg0VLXnS' as password while the *E. coli* dataset can be accessed using 'reviewer23743@ebi.ac.uk' as username and 'cn5cG4jW' as password.

ACKNOWLEDGMENTS.

We thank Prof. Kris Gevaert for financial support of this research (Research Foundation - Flanders (FWO-Vlaanderen), project number G.0440.10). P.V.D. acknowledges support from the Research Foundation - Flanders (FWO-Vlaanderen), project number G.0269.13N. G.M. is a postdoctoral fellow at the Research Foundation – Flanders (FWO-Vlaanderen).

AUTHOR CONTRIBUTIONS.

E.N., A.G., E.V., G.M. and P.V.D. conceived the study; E.N., G.M. and P.V.D. wrote the manuscript; E.N. performed the computational analysis; P.V.D. performed the proteomics experiments, E.N. and P.V.D. performed the proteomics analyses; V.J. and P.V.D. prepared the Ribo-seq libraries. G.M. and P.V.D. supervised the research.

COMPETING FINANCIAL INTERESTS.

The authors declare that they have no competing financial interests.

SUPPLEMENTARY METHODS

Experimental Procedures

Shotgun proteome analysis – *Salmonella*

Overnight stationary cultures of wild type S. Typhimurium (Salmonella enterica serovar Typhimurium - strain SL1344) grown in LB media at 37 °C with agitation (200 rpm) were diluted at 1:200 in LB and grown until they reached and OD600 of 0.5 (i.e., logarithmic (Log) phase grown cells). Bacterial cells were collected by centrifugation (6000 × g, 5 min) at 4 °C, flash frozen in liquid nitrogen and cryogenically pulverized using a liquid nitrogen cooled pestle and mortar. The frozen pellet of a 50 ml culture was re-suspended and thawed in 1 ml ice-cold lysis buffer (50 mm NH4HCO3 (pH 7.9) and subjected to mechanical disruption by 3 repetitive freeze-thaw and sonication cycles (i.e. 2 minutes of sonication on ice for 20-s bursts at output level 4 with a 40% duty cycle (Branson Sonifier 250; Ultrasonic Convertor)). The lysate was cleared by centrifugation for 15 min at 16,000 × g and the protein concentration measured using the DC Protein Assay Kit from Bio-Rad (Munich, Germany) according to the manufacturer's instructions. For all proteome analyses performed, 1 mg of protein material (corresponding to about 300 µl of lysate) was subjected to shotgun proteome analysis as described previously (Koch et al. 2014). More specifically, 3 different proteome digestions were performed at 37°C and mixing at 550 rpm using mass spectrometry grade trypsin (enzyme/substrate of 1/100, w/w; Promega, Madison, United States), chymotrypsin (1/60, w/w; Promega, Madison, United States) or endoproteinase Glu-C (1/75, w/w; Thermo Fisher Scientific, Bremen, Germany). A final set of 24 samples per proteome digest was vacuum dried, re-dissolved in 20 µl of 10 mM tris(2-carboxyethyl) phosphine (TCEP) in 2% acetonitrile and analysed by LC-MS/MS.

N-terminal proteomics – *E.coli*

Overnight stationary cultures of *E. coli tolC* CAG12148 cells ordered at the E. Coli Genetic Stock Collection (CGSC7437; F-, *λ*$^{-}$, *tolC210::Tn10*, *rph-1;* http://cgsc.biology.yale.edu/) (Singer et al. 1989) were grown in LB media at 37 °C with agitation (200 rpm) and diluted into 100 ml fresh medium until a OD_{600} of 0.02 and incubated. When the OD_{600} reached 0.55, 8 µg/ml actinonin (Sigma-Aldrich) was added. After 2 hours of cultivation (OD_{600} 1.1), cells were harvested and collected by centrifugation (3300 × g, 5 min) at 4 °C, flash frozen in liquid nitrogen and cryogenically pulverized using a liquid nitrogen cooled pestle and mortar. The frozen pellet of a 50 ml culture was re-suspended and thawed in 1 ml ice-cold lysis buffer (50 mm NH_4HCO_3 (pH 7.9) and subjected to mechanical disruption by 3 repetitive freeze-thaw and sonication cycles as described above. The lysate was cleared by centrifugation for 15 min at 16,000 × *g* and the protein concentration measured using the DC Protein Assay Kit from Bio-Rad according to the manufacturer's instructions. 4 mg of protein material (corresponding to about 1 ml of lysate) was digested overnight at 37°C and 550 rpm with sequencing-graded modified trypsin (Promega, Madison, WI, USA; enzyme/substrate, 1/200 (w/w)). The digested and modified peptides were subjected to a modified version of N-terminal COFRADIC (Staes et al. 2008) as will be described elsewhere. A final set of 90 samples were vacuum dried, re-dissolved in 20 µl of 10 mM tris(2-carboxyethyl) phosphine (TCEP) in 2% acetonitrile and analysed by LC-MS/MS.

LC-MS/MS analysis

The Salmonella shotgun samples were separated by nano-LC and analyzed with a Q Exactive instrument (Thermo Scientific) operating in MS/MS mode as previously described (Stes et al. 2014). In case of the *E. coli* N-terminal proteomics samples, LC-MS/MS analysis was performed using an Ultimate 3000 RSLC nano HPLC (Dionex, Amsterdam, the Netherlands) in-line connected to an

LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) (Gawron et al. 2016).

The generated MS/MS peak lists were searched with Mascot using the Mascot Daemon interface (version 2.5.1, Matrix Science). Searches were performed using 6-FT database of *S. typhimurium* (Salmonella enterica serovar Typhimurium - strain SL1344) genome or *E. coli* (K-12 strain MG1655), in each case combined with the Ensembl protein sequence database (assembly AMS21085v2 version 86.1 in the case of Salmonella and assembly ASM584v2version 87.1 for E. coli), which resulting to a total of 139408 and 120714 Salmonella and *E. coli* entries respectively, after removal of redundant sequences. The 6-FT databases were generated by traversing the entire genome across the six reading frames and searching for all NTG (N=A,T,C,G) start codons and extending each to the nearest in frame stop codon (TAA,TGA,TAG), with ORFs less than 30nt discarded. The Mascot search parameters were set as follows for the Salmonella samples; methionine oxidation to methionine-sulfoxide was set as fixed modifications. Variable modifications were formylation, acetylation (both at peptide level) and pyroglutamate formation of N-terminal glutamine. Mass tolerance was set to 10 ppm on the precursor ion (with Mascot's C13 option set to 1) and to 20 mmu on fragment ions. Peptide charge was set to 1+, 2+, 3+ and instrument setting was put to ESI-QUAD. Enzyme settings were set to 'no enzyme' in the case of the Chymo and Glu-C digested proteome samples (Tanco et al. 2013) and endoproteinase Trypsin/P (Trypsin specificity with Arg/Lys-Pro cleavage allowed) was set as enzyme in the case of the tryptic samples, only in case of the latter allowing for one missed cleavage.

For the E. coli N-terminal proteomics samples: Heavy acetylation at lysine side-chains (Acetyl:2H(3)C13(2) (K)) and carbamidomethylation of cysteine a were set as fixed modifications. Variable modifications were methionine oxidation to methionine-sulfoxide, formylation, acetylation and heavy acetylation of N-termini (Acetyl:2H(3)C13(2) (N-term)) all at peptide level) and pyroglutamate formation of N-terminal glutamine. Mass tolerance was set to 10 ppm on the precursor ion (with Mascot's C13 option set to 1) and to 0.5 Da on fragment ions. Peptide charge was set to 1+, 2+, 3+ and instrument setting was put to ESI-TRAP. Endoproteinase semi-Arg-C/P (Arg-C specificity with arginine-proline cleavage allowed) was be selected as enzyme allowing for 1 missed cleavages. Only peptides that were ranked one, have a minimum amino acid length of seven, scored above the threshold score, set at 95% confidence, and belonged to the category of peptides compliant with the rules of initiator methionine (iMet) processing (Martinez et al. 2006) were withheld (Supplementary File F2). More specifically, iMet processing was considered in the case of iMet-starting N-termini followed by any of the following amino acids; Ala, Cys, Gly, Pro, Ser, Thr, Met or Val and only if the iMet was encoded by ATG or any of the following near-cognate start codons; GTG and TTG.We thus allow for ambiguous double initiation codon, where the N-terminal peptide could equally support either start codons. If an ORF was predicted to start at one of these positions, the position supporting that ORF was selected.

Ribosome Profiling

Overnight stationary cultures of wild type S. Typhimurium (*Salmonella enterica serovar Thyphimurium* - strain SL1344) grown in LB media at 37°C with agitation (200 rpm) were diluted at 1:200 in LB and grown until they reached and OD600 of 0.5 (i.e., logarithmic (Log) phase grown cells). Bacterial cells were pre-treated for 5 min with chloramphenicol (Sigma Aldrich) at a final concentration of 100 µg/ml before collection by centrifugation (6000 × g, 5 min) at 4°C. Collected cells were flash frozen in liquid nitrogen. The frozen pellet of a 50 ml culture was re-suspended and thawed in 1 ml ice-cold lysis buffer for polysome isolation (10 mM $MgCl_2$, 100 mM NH_4Cl, 20 mM Tris.HCl pH 8.0, 20 U/ml of RNase-free DNase I (NEB 2U/µl), 1mM chloramphenicol (or 300µg/ml), 20 µl/ml lysozyme (50mg/ml in water) and 100u/ml SUPERase.In™ RNase Inhibitor (Thermo Fisher Scientific, Bremen, Germany)), vortexed and left on ice for 2 min with periodical agitation. Subsequently, the samples were subjected to mechanical disruption by two repetitive cycles of freeze-thawing in liquid nitrogen, added 5mM $CaCl_2$, 30µl 10% DOC and 1 × complete and EDTA-free protease inhibitor cocktail (Roche, Basel, Switzerland) and left on ice for 5 min. Lysates were clarified by centrifugation at 16,000 x g for 10 min at 4°C.

Preparation of ribosome profiling libraries: For the monosome sample, the supernatant was subjected to MNase (Roche diagnostics Belgium) digestion using 600 U MNase (about~ 1000 U per mg of protein). Digestion of polysomes proceeded for 1h at 25°C with gentle agitation at 400 rpm and the reaction was stopped by the addition of 10 mM EGTA. Next, monosomes were recovered by ultracentrifugation over a 1 M sucrose cushion in polysome isolation buffer without RNase-free DNase I and lysozyme, and with 2 mM DTT added using a TLA-120.2 rotor for 4 hr at 75,000 rpm and 4°C.

For the selective purification of monosomes from polysomes (polysome sample), the supernatant was resolved on 10-55% (w/v) sucrose gradients by centrifugation using an SW41 rotor at 35,000 rpm for 2.5 hr at 4°C. The sedimentation profiles were recorded at 260 nm and the gradient fractionated using a BioComp Gradient Master (BioComp) according to the manufacturer's instructions. Polysome-enriched fractions were pooled and subjected to MNase digestion and monosome recovery as described above.

Ribosome-protected mRNA footprints with sizes ranging from 26-34 nucleotides were selected and processed as described previously (Ingolia et al. 2012)with some minor adjustments as described in (Gawron et al. 2016). The resulting ribosome profiling cDNA libraries of the monosome and polysome sample were duplexed and sequenced on a NextSeq 500 instrument (Illumina) to yield 75-bp single-end reads.

RNA-seq

For monosome and polysome-enriched samples, part (1/10th) of the cleared lysate or pooled polysome-enriched fractions obtained using sucrose gradient centrifugation was taken for total RNA isolation, respectively making use of the TRIzol reagent (Invitrogen, Thermo Fisher Scientific Inc.) according to manufacturer’s instructions or making use of phenol/chloroform extraction after the addition of SDS at a final concentration of 1% as described before (Ingolia et al. 2011). RNA yields were determined using a NanoDrop spectrophotometer (Wilmington, Delaware, USA) and RNA quality was assessed by Agilent Bioanalyzer RNA 600 Nano Kit running the assay class 'Prokaryote Total RNA Nano'. Of note, in this case of Salmonella, rRNA transcripts carry intervening sequences that are excised during ribosome formation (Evguenieva-Hackenberg 2005). As such, the 23S and 16S rRNA components elute in multiple peaks precluding reliable analysis of the current algorithms used by the BioAnalyzer platform to analyse RIN values to calculate RNA quality scores. Nonetheless, repeated isolations demonstrated reproducible profiles in line with previous reports (Bhagwat et al. 2013) and E. coli samples processed in parallel all showed RIN values above 9. Library construction including random fragmentation, cDNA synthesis and library generation were performed Library preparation and sequencing was performed at the VIB Nucleomics Core (www.nucleomics.be) using the TruSeq stranded total RNA sample preparation kit (Illumina, San Diego, California, USA) and including a Ribo-Zero (Illumina) depletion step as to remove ribosomal RNA from total RNA by the use of biotinylated Ribo-Zero oligos. Libraries were subjected to sequencing on a NextSeq 500 instrument (Illumina) to yield 75 bp single-end reads.

Ribo-seq and RNA-seq data processing.

The E. coli and Bacillus Ribo-seq and RNA-seq data can be downloaded from the GEO repository with accessions GSM1300279 (Li et al. 2014) and GSM872395 (Li et al. 2012), respectively. In case of Salmonella, the Ribo-seq and RNA-seq datasets were generated in house.

Adapter sequences were removed from the reads using fastx_clipper and reads aligning onto rRNA and tRNA sequences were discarded. The remaining reads were aligned to the genome of using bowtie with settings –v1 –m2 –k1 allowing only for uniquely mapped reads. Since improved results were obtained this way (data not shown), ribosome occupancy positions were assigned to the 3' end of the reads in case of E. coli and Bacillus while the 5' ends were used for Salmonella. Only reads of length between 22 and 40nt were considered in the analysis (Li et al. 2014). The RNA-seq data was stripped of the adapter sequences and subsequently aligned onto the appropriate genomes using bowtie (–v1 –m2 –k1).

REPARATION.

REPARATION performs *de novo* ORF delineation by training a random forest classifier to learn patterns from Ribo-seq data. A random forest model was chosen over other algorithms for training because of its robustness to outliers, low bias and optimal performance with few parameter tuning (Hastie & Tibshirani 2009). The REPARATION pipeline (*figure 1A*) starts by first traversing the entire genome and collect all ORFs starting with NTG (N = A, T, G) across all six reading frames (selection of start codon(s) is a user definable parameter). For each possible start codon, the algorithm searches for the first in frame downstream stop codon (TAA, TAG or TGA) that is at least 10 codons apart (can be adjusted by the user).

Training Sets.

The set of positive examples is constructed by a comparative genomic approach. The algorithm uses Prodigal V2.6.3 (Hyatt et al. 2010) to generate an ORF set, this set is then BLAST searched against a database of curated bacterial protein sequences (e.g. UniprotKB-SwissProt). The BLAST search is performed using the UBLAST algorithm from the USEARCH package (Edgar 2010). ORFs that match at least one known protein coding sequence with a minimum e-value of 10^{-5} and a minimum identity of 75% are selected for the positive set. The negative set consist of ORFs starting with CTG viewing their infrequent occurrence as translation starts (<0.01%) in the annotations of the interrogated species (*supplementary table T1)* and at least as long as the minimum ORF length in the positive set. We then grouped all CTG ORFs sharing the same in frame stop codon into an "ORF family". Per ORF family we select the longest ORF as a representative member of that "ORF family". Of note, REPARATION allows the user to provide a custom list of ORFs to be used as the positive sets in training the random forest classifier.

Feature construction.

To train the random forest classifier we constructed five features based on Ribo-seq signals of translated ORFs and complemented these with the ribosome binding energy measurements (Suzek et al. 2001) (see below). The meta gene profile shown in *figure 1B* illustrates a Ribo-seq signal pattern reminiscent to patterns previously reported for protein coding transcripts in prokaryotes for Ribo-seq experiments that targets elongating ribosomes (Woolstenhulme et al. 2015). The profile exhibits read accumulation within the first 40-50nts downstream of the start and a slight increase just before the stop codon. The features used in the model are as follows:

Start and stop region read density (RPKM). We defined a start region of an ORF by taking 3nt upstream (to account for any error in P-site assignment) and 45nt downstream of the ORF start position, while the stop region constitutes the last 21nt upstream of the stop position. Of note, for ORFs shorter than 63 nucleotides we used the first 70% and last 25% of the ORF length to model the start and stop regions of the ORF. The ORF RPF read count per nucleotide position is divided by the total RPF reads within the ORF to ensure that features are comparable across different ORFs. The start and stop region RPF read densities are subsequently calculated from the proportional reads.

ORF coverage and start RPF coverage. We defined the ORF (start) RPF coverage as the proportion of nucleotide positions covered by RPF reads within a region of interest i.e within the entire ORF and the start region. RPF coverage is calculated from the positional read profile.

Read accumulation proportion. This feature is based on the positional RPF reads, it measures the ratio of the RPF reads accumulated at the start region (first 45nt) relative to RPF reads within the rest of the ORF. It is defined by

$$Accumulation\,proportion = \begin{cases} \dfrac{Average\ RPF\ count\ within\ the\ ORF\ start\ region}{Average\ RPF\ count\ on\ the\ rest\ of\ the\ ORF} \\ 0,\ \ if\ Average\ read\ on\ the\ rest\ of\ the\ ORF = 0 \end{cases}$$

We reasoned that since Ribo-seq reads tend to accumulate within the start region of a translated ORF relative to the rest of the ORF, correctly delineated ORFs will tend to have score greater than one. Spurious ORFs that overlap at the start or stop of translated ORFs will score lower as their non-overlapping regions would tend to have no reads, hence resulting to accumulation scores less than one.

Ribosome binding site (RBS) energy. The interaction between Shine-Dalgarno (SD) sequence and its complementary sequence in the 16S rRNA (anti-SD), referred to as SD ribosome binding site (RBS) was proven to be very important in the recruitment of the ribosome for translation initiation in bacteria (Shultzaberger et al. 2001). As such, and to aid in the prediction of SD/anti-SD dependent translation events, the ribosome's free binding energy or ribosome binding site (RBS) energy was included as in feature in the model. The RBS energy, representative of the probability that the ribosome will bind to a specific mRNA and thus proportional to the mRNA's translation initiation rate, was calculated using the distance dependent probabilistic method and using the anti-Shine Dalgarno (aSD) sequence GGAGG as described in Suzek et al. (2001).The inclusion of the of the RBS energy features in the prediction model as well as the aSD sequence are user defined parameters to allow for bacterial species where non aSD/SD dependent translation events have been reported (Shultzaberger et al. 2001; Hyatt et al. 2010; Omotajo et al. 2015).

Sigmoid (S)-curve model.

Since REPARATION pipeline was developed to allow for ORFs as short as 30nt, this results in an exponential increase of potential ORFs. To ensure the algorithm is traceable, we defined minimum threshold values to eliminate spurious ORFs. To do this we take advantage of the sigmoid curve (S-curve) relationship observed between ORF RPF coverage and the ORF log2 read density (RPKM) as depicted in *figure 1C and supplementary figure S6*. The fitted logistic curve (red), modelled by a 4 parameter logistic regression and describing the relationship between ribosome density and RPF coverage was used to estimate the minimum read density and ORF RPF coverage to allow for correct ORF delineation. We estimated the lower bend point of the fitted 4 parameter logistic regression using the method described in (Lutz & Lutz 2009) and implemented in the R Package *Sizer* (Sonderegger et al. 2009).

Tuning the classifier parameters

As the number of possible ORFs in the negative set vary across different bacterial genomes and are often 3 or more folds that of the positive set (*supplementary table T2*), we first ensure that the classifier is robust to class imbalance. To avoid any bias in predictions towards the majority class (Chawla et al. 2002) we evaluated four strategies to account for class imbalance; 1) over sampling the minority class i.e. by sampling the minority class with replacement to obtain a balanced training set, 2) down sampling, randomly selecting a subset from the majority class (Blagus & Lusa 2010) 3) evaluation using the SMOTE technique (Chawla et al. 2002) which is a hybrid of the above techniques and 4) the class prior setting, determining the proportion samples of each class to be used constructing the trees. The class prior technique performed the best in accounting for class imbalance with a 10-fold cross validation precision measure of 80% on the *Salmonella* and *E. coli* data sets (*supplementary figure S7*) and with optimal split of at least 25% from the minority class. We next keep all parameters constant and tune the number of samples in each terminal node, the best result was obtained with a value of four using a 10-fold cross validation from the values in the range 1 through 30 while optimal number of variable in each split was four. The number of trees in the model was optimized and the best performance was obtained with 3001 trees from the range 251 through 6001, step 500. We optimized only these parameters because they have been shown to influence the random forest model (Hastie & Tibshirani 2009). The model was implemented using the R package *randomForest* (Liaw & Wiener 2002) on a Linux Fedora R3, kernel version 4.7.2-101.

Post Processing Random Forest predicted ORFs.

We implement a rule based post processing algorithm to eliminate false positives that might be called because they share overlapping regions with actual coding ORFs (*supplementary figure S8*). First, considering the simplified assumption that bacterial genes can have only one possible translation start site, we group all predicted ORFs sharing the same in frame stop codon into an "ORF family". *Supplementary figure S8 A* depicts an ORF family with two predicted starts, if start S1 has more reads than S2 then we select S1 as the gene start. If there are no Ribo-seq reads between S1 and S2 then we select S2 as the gene start since S1 adds no extra information to the gene profile. If S1 has more reads than S2 but if S1 falls within the coding region of an out-of-frame upstream predicted ORF on the same strand, we select S1 as the most likely start if there is a peak (i.e. kurtosis > 0) within a window of -21 to +21 around S1.

Next we consider two overlapping ORFs on different frames as depicted in *supplementary figure S8 B*. If the read density and RPF coverage of the non-overlapping region of F1 are less than the S-curve estimated thresholds, then F1 is dropped in favor of F2 and vice versa. If both non-overlapping regions have a read density and RPF coverage greater than the minimum, then we assume both are expressed. Finally, we drop all internal out-of-frame ORFs falling completely within another ORF (supplementary *figure S8 C).*

Conservation Analysis.

Nucleotide conservation scores were calculated using the rate4site tool (Pupko et al. 2002). To obtain site conservation scores, we downloaded the protein and genome sequences of 165 bacteria species from *Enterobacteriaceae*, *Bacillus* and other closely related and distant genus was randomly selected from Ensembl (*supplementary file F4*). We combined all protein sequences into a Fasta database and used the *cluster_fast* algorithm in the *USEARCH* package (Edgar 2010) to remove redundant sequences within the database by clustering sequences with minimally 90% of similarity and keeping only the centroid sequence. For each predicted ORF, we searched for all possible orthologues sequences using *OrthoFinder* tool (Emms & Kelly 2015) in the non-redundant database and after adjusting the tool by replacing *blastp* with the faster *ublast* (Edgar 2010) algorithm. For each orthologue we obtain its genomic coordinates and extracted the nucleotide sequence of the ORF as well as 30nt upstream of the ORF start. We then performed a multiple sequence alignment of all orthologues ORFs using MUSCLE (Edgar 2010) and concatenated the corresponding upstream sequence with the appropriate ORF in the multiple sequence alignments. Finally, the position-specific conservation scores were calculated using the rate4site (Pupko et al. 2002) tool with the empirical Bayesian estimate and Jukes-Cantor probabilistic model for nucleotides. Only ORFs with at least five orthologues (minimum number of orthologues to properly estimate the site conservation score (Goldenberg et al. 2009)) were considered for the conservation analysis.

REFERENCES.

Andrews, S.J. & Rothnagel, J.A., 2014. Emerging evidence for functional peptides encoded by short open reading frames. *Nature reviews. Genetics*, 15(3), pp.193–204.

Angelova, M., Kalajdziski, S. & Kocarev, L., 2010. Computational Methods for Gene Finding in Prokaryotes. *ICT Innovations 2010, Web Proceedings*, (March 2016), pp.11–20.

Bazzini, A.A. et al., 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO Journal*, 33(9), pp.981–993.

Bhagwat, A. a et al., 2013. Determining RNA quality for NextGen sequencing: some exceptions to the gold standard rule of 23S to 16S rRNA ratio§. *Microbiology Discovery*, 1, p.10.

Blagus, R. & Lusa, L., 2010. Class prediction for high-dimensional class-imbalanced data.

Calviello, L. et al., 2015. Detecting actively translated open reading frames in ribosome profiling data. *Nature Methods*, 13(December), pp.1–9.

Chawla, N. V. et al., 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, pp.321–357.

Chew, G.-L. et al., 2013. Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development (Cambridge, England)*, 140(13), pp.2828–34.

van Damme, P. et al., 2011. NatF contributes to an evolutionary shift in protein N-terminal acetylation and is important for normal chromosome segregation. *PLoS Genetics*, 7(7).

Van Damme, P. et al., 2014. N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. *Molecular & cellular proteomics : MCP*, 13(5), pp.1245–61.

Diament, A. & Tuller, T., 2016. Estimation of ribosome profiling performance and reproducibility at various levels of resolution. *Biology direct*, 11, p.24.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19), pp.2460–2461.

Emms, D.M. & Kelly, S., 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1), p.157.

Evguenieva-Hackenberg, E., 2005. Bacterial ribosomal RNA in pieces. *Molecular Microbiology*, 57(2), pp.318–325.

Feng, Y. et al., 2012. Pseudogene recoding revealed from proteomic analysis of salmonella serovars. *Journal of Proteome Research*, 11(3), pp.1715–1719.

Fields, A.P. et al., 2015. A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved

Complexity to Mammalian Translation. *Molecular Cell*, 60(5), pp.816–827.

Gawron, D. et al., 2016. Positional proteomics reveals differences in N-terminal proteoform stability. *Molecular Systems Biology*, 12(2), p.858.

Gawron, D., Gevaert, K. & Damme, P. Van, 2014. The proteome under translational control. *PROTEOMICS*, 14(23–24), pp.2647–2662.

Goldenberg, O. et al., 2009. The ConSurf-DB: Pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Research*, 37(SUPPL. 1), pp.323–327.

Goodhead, I. & Darby, A.C., 2015. Taking the pseudo out of pseudogenes. *Current Opinion in Microbiology*, 23, pp.102–109.

Hastie, T. & Tibshirani, R.F., 2009. The Elements of Statistical Learning. *Methods*, 1(2), pp.305–317.

Hemm, M.R. et al., 2008. Small membrane proteins found by comparative genomics and ribosome binding site models. *Molecular Microbiology*, 70(6), pp.1487–1501.

Heyer, E.E. & Moore, M.J., 2016. Redefining the Translational Status of 80S Monosomes. *Cell*, 164(4), pp.757–769.

Hyatt, D. et al., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11, p.119.

Ingolia, N.T. et al., 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)*, 324(5924), pp.218–23.

Ingolia, N.T., Lareau, L. & Weissman, J., 2012. Ribosome Profiling of Mouse Embryonic Stem Cells Reveales Complexity of Mammalian Proteomes. *Cell*, 147(4), pp.789–802.

Ingolia, N.T., Lareau, L.F. & Weissman, J.S., 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4), pp.789–802.

Koch, A. et al., 2014. A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics*, 14(23–24), pp.2688–2698.

Krug, K. et al., 2013. Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Molecular & cellular proteomics : MCP*, 12(11), pp.3420–30.

Land, M. et al., 2015. Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*, 15(2), pp.141–61.

Lerat, E. & Ochman, H., 2005. Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Research*, 33(10), pp.3125–3132.

Li, G.W. et al., 2014. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3), pp.624–635.

Liaw, a & Wiener, M., 2002. Classification and Regression by randomForest. *R news*, 2(December), pp.18–22.

Lutz, W.K. & Lutz, R.W., 2009. Statistical model to estimate a threshold dose and its confidence limits for the analysis of sublinear dose-response relationships, exemplified for mutagenicity data. *Mutation Research - Genetic Toxicology and Environmental Mutagenesis*, 678(2), pp.118–122.

Martinez, A. et al., 2006. The Proteomics of N-terminal Methionine Cleavage * □. , pp.2336–2349.

Michel, A.M. et al., 2012. Observation of dually decoded regions of the human genome using ribosome profiling data. , pp.2219–2229.

Olexiouk, V. et al., 2016. SORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research*, 44(D1), pp.D324–D329.

Omotajo, D. et al., 2015. Distribution and diversity of ribosome binding sites in prokaryotic genomes. *BMC genomics*, 16(1), p.604.

Palenchar, P.M., 2008. Amino acid biases in the N- and C-termini of proteins are evolutionarily conserved and are conserved between functionally related proteins. *Protein Journal*, 27(5), pp.283–291.

Panicker, I.S., Browning, G.F. & Markham, P.F., 2015. The effect of an alternate start codon on heterologous expression of a PhoA fusion protein in mycoplasma gallisepticum. *PLoS ONE*, 10(5), pp.1–10.

Pauli, A., Valen, E. & Schier, A.F., 2015. Identifying (non-)coding RNAs and small peptides: Challenges and opportunities. *BioEssays*, 37(1), pp.103–112.

Pupko, T. et al., 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics (Oxford, England)*, 18 Suppl 1(1), pp.S71–S77.

Richardson, E.J. & Watson, M., 2013. The automatic annotation of bacterial genomes. *Briefings in Bioinformatics*, 14(1), pp.1–12.

Salzberg, S.L. et al., 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2), pp.544–548.

Samayoa, J., Yildiz, F.H. & Karplus, K., 2011. Identification of prokaryotic small proteins using a comparative genomic approach. *Bioinformatics*, 27(13), pp.1765–1771.

Shultzaberger, R.K. et al., 2001. Anatomy of Escherichia coli ribosome binding sites. *Journal of molecular biology*, 313, pp.215–228.

Singer, M. et al., 1989. A collection of strains containing genetically linked alternating antibiotic resistance elements for genetic mapping of Escherichia coli. *Microbiological Reviews*, 53(1), pp.1–24.

Sonderegger, D.L. et al., 2009. Using SiZer to detect thresholds in ecological data. *Frontiers in Ecology and the Environment*, 7(Cd), pp.190–195.

Staes, A. et al., 2008. Improved recovery of proteome-informative, protein N-terminal peptides by combined fractional diagonal chromatography (COFRADIC). *Proteomics*, 8(7), pp.1362–1370.

Stes, E. et al., 2014. A COFRADIC protocol to study protein ubiquitination. *Journal of Proteome Research*, 13(6), pp.3107–3113.

Suzek, B.E. et al., 2001. A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics (Oxford, England)*, 17(12), pp.1123–1130.

Tanco, S. et al., 2013. Proteome-derived Peptide Libraries to Study the Substrate Specificity Profiles of Carboxypeptidases. *Mol Cell Proteomics*, 12(8), pp.2096–2110.

Vizcaino, J.A. et al., 2016. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research*, 44(D1), pp.D447–D456.

Woolstenhulme, C.J. et al., 2015. High-Precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Reports*, 11(1), pp.13–21.