


METHODOLOGY ARTICLE

Open Access



# Common and phylogenetically widespread coding for peptides by bacterial small RNAs

Robin C. Friedman<sup>1,2,3\*</sup> , Stefan Kalkhof<sup>5,6</sup>, Olivia Doppelt-Azeroual<sup>4</sup>, Stephan A. Mueller<sup>5,8</sup>, Martina Chovancová<sup>5</sup>, Martin von Bergen<sup>5,7</sup> and Benno Schwikowski<sup>1,3</sup>

## Abstract

**Background:** While eukaryotic noncoding RNAs have recently received intense scrutiny, it is becoming clear that bacterial transcription is at least as pervasive. Bacterial small RNAs and antisense RNAs (sRNAs) are often assumed to be noncoding, due to their lack of long open reading frames (ORFs). However, there are numerous examples of sRNAs encoding for small proteins, whether or not they also have a regulatory role at the RNA level.

**Methods:** Here, we apply flexible machine learning techniques based on sequence features and comparative genomics to quantify the prevalence of sRNA ORFs under natural selection to maintain protein-coding function in 14 phylogenetically diverse bacteria. Importantly, we quantify uncertainty in our predictions, and follow up on them using mass spectrometry proteomics and comparison to datasets including ribosome profiling.

**Results:** A majority of annotated sRNAs have at least one ORF between 10 and 50 amino acids long, and we conservatively predict that  $409 \pm 191.7$  unannotated sRNA ORFs are under selection to maintain coding (mean estimate and 95% confidence interval), an average of 29 per species considered here. This implies that overall at least  $10.3 \pm 0.5\%$  of sRNAs have a coding ORF, and in some species around 20% do.  $165 \pm 69$  of these novel coding ORFs have some antisense overlap to annotated ORFs. As experimental validation, many of our predictions are translated in published ribosome profiling data and are identified via mass spectrometry shotgun proteomics. *B. subtilis* sRNAs with coding ORFs are enriched for high expression in biofilms and confluent growth, and *S. pneumoniae* sRNAs with coding ORFs are involved in virulence. sRNA coding ORFs are enriched for transmembrane domains and many are predicted novel components of type I toxin/antitoxin systems.

**Conclusions:** We predict over two dozen new protein-coding genes per bacterial species, but crucially also quantified the uncertainty in this estimate. Our predictions for sRNA coding ORFs, along with predicted novel type I toxins and tools for sorting and visualizing genomic context, are freely available in a user-friendly format at <http://disco-bac.web.pasteur.fr>. We expect these easily-accessible predictions to be a valuable tool for the study not only of bacterial sRNAs and type I toxin-antitoxin systems, but also of bacterial genetics and genomics.

**Keywords:** sRNAs, Type I toxin/antitoxin, Short ORFs, Machine learning, Ribosome profiling, Mass spectrometry

\*Correspondence: [robin.friedman@gmail.com](mailto:robin.friedman@gmail.com)

<sup>1</sup>Systems Biology Laboratory, Department of Genomes and Genetics, Institut Pasteur, Paris, France

<sup>2</sup>Molecular Microbial Pathogenesis Unit, Department of Cell Biology and Infection, Institut Pasteur, Paris, France

Full list of author information is available at the end of the article

## Background

Recent technological advances such as tiling microarrays and deep RNA sequencing have led to a new appreciation of bacterial transcription, identifying thousands of new bacterial small RNAs (sRNAs) [1–4]. Single strains can contain hundreds of sRNAs, including both independent transcripts and extensive transcription antisense to annotated open reading frames (ORFs) [2, 5, 6]. In virtually all cases, sRNAs do not contain an annotated coding sequence (CDS, or coding ORF) and it is therefore assumed that their primary function is to act as antisense RNAs modulating the expression of other genes [4, 7]. However, ORFs occur frequently by chance, and although there are many reasons that an ORF would not be translated (for example, strong secondary structure of the RNA), it is difficult to prove that an ORF is not translated in vivo.

In particular, gene annotations for short ORFs (usually defined as shorter than 100 or 50 amino acids) are notoriously incomplete and thousands of protein-coding genes remain unannotated in bacteria [8], so many sRNAs could in theory code for functional small proteins. There are several examples of dual-function sRNAs having a regulatory role that also code for an experimentally-validated functional small protein: *E. coli* SgrS encodes the protein SgrT [9], *S. aureus* RNAIII encodes  $\delta$ -hemolysin [10], *B. subtilis* SR1 encodes SR1P [11], and *P. aeruginosa* PhrS encodes an unnamed protein [12]. However, because no antisense regulatory function has been found so far for most known sRNAs [3], it is possible that the primary function of many could be simply coding for functional peptides.

Small proteins play important roles in bacteria, including quorum sensing, transcription, translation, stress response, metabolism, and sporulation [13, 14]. However, they are difficult to identify by computational or experimental methods. The short sequences have less space for evidence of natural selection, resulting in high levels of statistical noise and false positives, making computational discrimination of coding ORFs smaller than about 50 amino acids difficult [8, 15]. Standard proteomics methods usually utilize gel electrophoresis or other chromatography methods, which bias towards proteins larger than about 30 kDa and preclude detection of very small proteins [16, 17]. Proteolytic cleavage of some small proteins also results in no peptides of a length detectable by mass spectrometers.

Nevertheless, efforts to identify bacterial short coding sequences have had some success. Proteogenomics, the reannotation of genomes using mass-spectrometry-based proteomics, is a powerful tool for identifying protein-coding genes but still suffers from false negatives, especially for small proteins [17–19]. Most computational methods applied so far have not taken advantage of sRNA

annotations and either used comparative genomics information exclusively [8, 20] or were applied only to a single species [15, 21, 22]. No existing method is ideal for determining the overall number of sRNA coding ORFs. Some comparative genomics methods take into account more information than the  $D_n/D_s$  test, but more complexity can make algorithms more brittle. For PhyloCSF [23], a greater number of parameters to fit can be problematic for small bacterial genomes and this method remains untested on prokaryotes. RNAcode [24] handles multiple alignment issues like insertions and deletions intelligently, but because it does not take into account phylogenetic structure it relies on careful selection of orthologous species to yield relevant results, making it difficult to apply on a large scale. Warren et al. [8] used a clever BLAST-based approach to quickly find new genes, but this is less sensitive than  $D_n/D_s$ , which is aware of phylogeny and mutations at the DNA level. Other methods are either ad-hoc and difficult to apply to other species [22] and/or do not incorporate both sequence features and comparative genomics [21].

Short proteins can rarely be predicted with nearly 100% confidence because of limited evidence, but most standard gene annotation tools do not provide an estimated false discovery rate (FDR) for marginal predictions, instead choosing ad-hoc cutoffs for amino acid length or coding score. However, even without confident individual predictions, statistically sound conclusions can be made when considering short ORFs in aggregate; for example, the overall number of ORFs under natural selection to maintain protein-coding potential can be estimated.

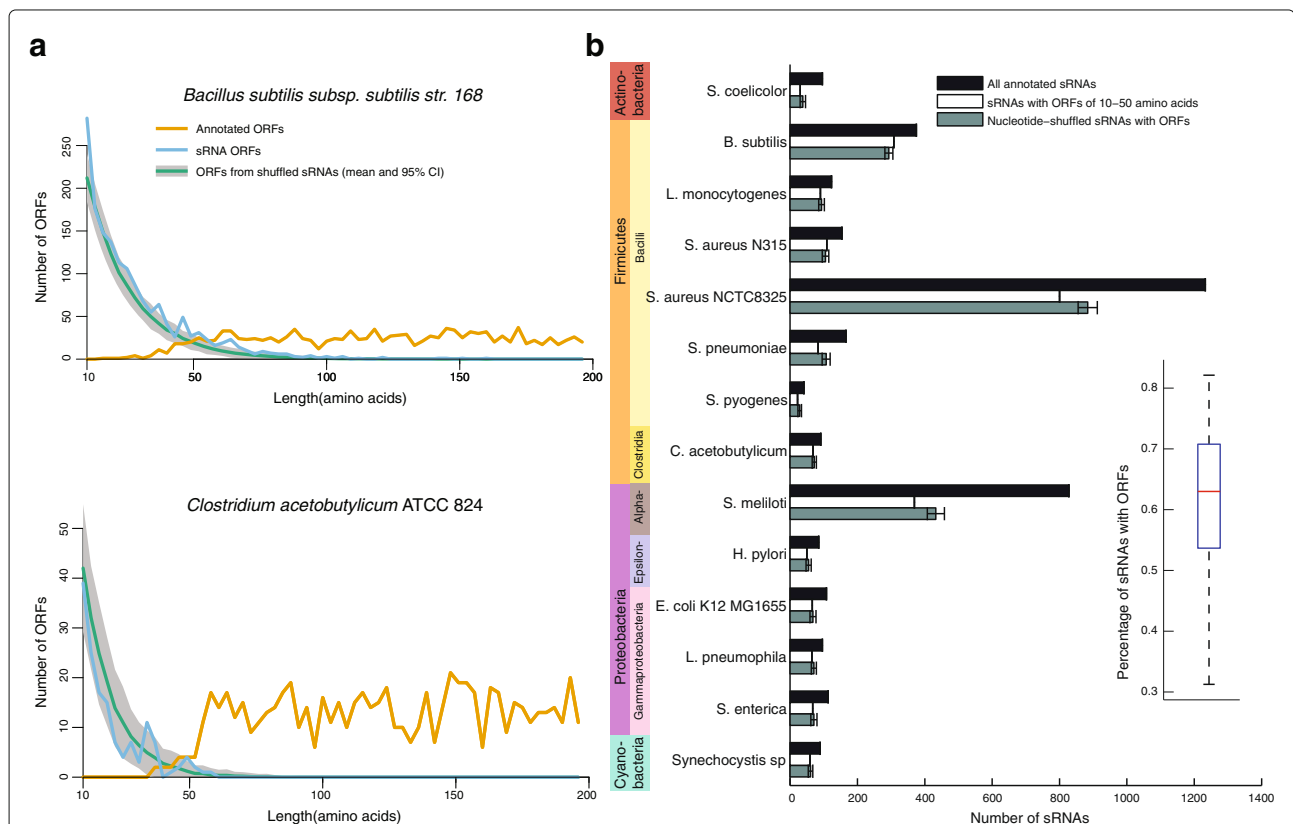
To identify short coding sequences in diverse species with high fidelity, algorithms must adapt to composition biases such as GC content, the strength and frequency of Shine-Dalgarno sequence motifs, the availability of closely-related genomes, and the structure of the phylogenetic tree relating these species. We set out to reexamine the assumption that most sRNAs are noncoding by applying simple and adaptable computational and statistical methods to a broad range of bacterial species, paying special attention to controlling for several biases in sRNA ORF sequence properties. We developed a computational method to predict coding ORFs called Discovery of sRNA Coding ORFs in Bacteria (DiSCO-Bac). We then validate the translation of predicted coding sRNA ORFs with experimental data from previously published ribosome profiling experiments [25] and mass spectrometry. We also mine experimental data from various sources to show that many of the resulting small proteins are likely to be functional, and a surprising number may be encoded antisense to other RNAs, many of which represent predicted or known toxin-antitoxin systems.

**Results**

**Most bacterial sRNAs have open reading frames**

Hypothesizing that many bacterial sRNAs contain unannotated short ORFs that could code for small proteins, we retrieved sRNA annotations from BSRD [3], supplemented by recent published datasets not included in the BSRD database [26, 27], and existing ORF annotations from Genbank. sRNAs were defined as any RNA expressed outside of annotated ORFs, including those found by small-scale experimental validation, RNA-seq, microarrays, or by sequence and structural conservation to known sRNAs. The sRNA definitions could be heterogeneous depending on the species, as they were identified by diverse methods. We focused first on a Gram-positive model species with particularly good sRNA annotations, *Bacillus subtilis* 168. We observed a sharp decrease in annotated coding ORFs shorter than around 50 amino acids (Fig. 1a, top). Because many gene finding algorithms use an arbitrary cutoff of 50 amino acids or greater, the

handful of shorter annotated ORFs are generally added by hand after their discovery via small-scale experiments. Many less intensely studied bacterial species have an even sharper cutoff around 50 amino acids, as in *Clostridium acetobutylicum* ATCC 824 (Fig. 1a, bottom). Presumably, this sharp cutoff reflects a technical artifact of automated annotation pipelines rather than an aspect of the true underlying length distribution. For example, reannotation of coding ORFs in yeast based on ribosome profiling recently found translation of 2869 non-internal ORFs shorter than 50 amino acids [28]. The presence of an artificial cutoff in bacterial ORF lengths suggests that many short coding ORFs remain unannotated. We next asked how many potentially coding ORFs in sRNAs were unannotated, limiting our search to ORFs of between 10 and 50 amino acids in length. Smaller proteins have a lower chance to be functional (for example, only 0.6% of antimicrobial peptides are less than 10 amino acids in length [29]), and longer proteins would likely be identified by



**Fig. 1** Most sRNAs have at least one potential protein-coding ORF. **a** Top: Length distribution of *B. subtilis* 168 ORFs annotated in Genbank (orange) compared to those between 10 and 50 amino acids in length in annotated sRNAs (blue) or those arising by chance in shuffled sRNAs (green). 95% confidence limits based on one thousand shuffles of sRNA sequence are shown in grey. Bottom: The same for *Clostridium acetobutylicum* ATCC 824 ORFs, which have a sharper drop-off around 50 amino acids. **b** Number of annotated sRNAs and sRNAs with at least one ORF for 14 species, representing 4 phyla. Phyla are represented as colors on the left, and colors on the right indicate when multiple taxonomic classes are represented within one phylum. As in (a), shuffled sRNA sequences generate a number of ORFs comparable in length to the observed amount. Inset: For each species, the percentage of sRNAs having at least one ORF of between 10 and 50 amino acids in length. Box represents median and first and third quartiles, and whiskers extend to the most extreme values. Most species have ORFs in more than 50% of sRNAs

modern gene-finding algorithms. We found 1365 ORFs in this size range in the 375 annotated *B. subtilis* sRNAs, with 82.1% of sRNAs having at least one ORF (Fig. 1a, b). This was roughly the number of ORFs expected by chance based on shuffles of the sRNA sequences (i.e. based on the length and nucleotide composition), suggesting that their occurrence is not strongly avoided by negative evolutionary pressure, and conversely there is not a large number of coding ORFs easily identified by length alone (Fig. 1a, b). We next examined 14 bacterial species representing 4 phyla and 7 classes, as a diverse representation of species having well-annotated transcriptomes. Each bacterial species had between 22 and 800 sRNAs having at least one ORF (Fig. 1b). The number of ORFs meeting our size cutoffs in sRNAs was roughly the same for shuffled control sequences in each species, so simply having many ORFs does not provide strong evidence for coding function in sRNAs. This result suggests that whatever the number of sRNA ORFs under selection to maintain coding function, it is approximately balanced by selection against maladaptive coding ORFs. Nevertheless, over a broad phylogenetic range, a majority of bacterial sRNAs have at least one ORF with the potential to code for a protein.

#### Bacterial sRNA ORFs have sequence features predictive of protein-coding function

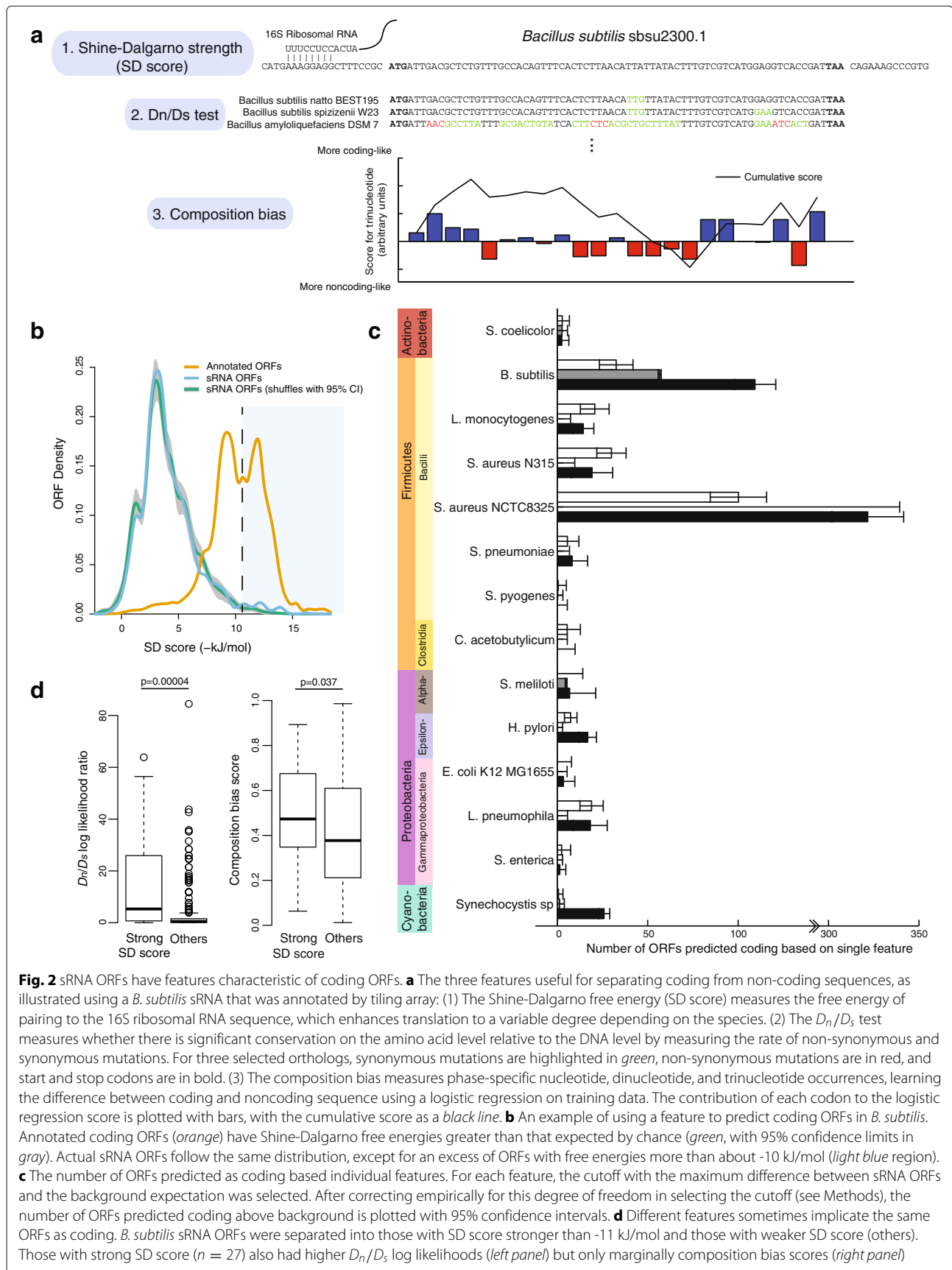
Most sRNA ORFs are thought not to be translated into protein products, because the presence of an ORF is necessary but not sufficient for expression at the protein level. For example, the Shine-Dalgarno sequence must be accessible to the ribosome for robust translation of most ORFs, and strong secondary structure may preclude translation. Several sequence features have been used to identify ORFs under natural selection to maintain protein coding, separating them from those not likely to be protein-coding [8, 15, 21, 22, 24, 30]. We use three sequence features with wide generality and applicability: The strength of the Shine-Dalgarno sequence (the SD score), conservation at the amino acid level (the  $D_n/D_s$  test), and phase-specific mono- and oligo-nucleotide bias (composition bias) (Fig. 2a). Each has varying predictive power depending on the species considered or on the available close orthologs, but is applicable to a wide phylogenetic range of species. We compared each of these features to carefully selected negative controls, which we call “mock ORFs”. Mock ORFs were regions selected from intergenic space to match the length distribution of sRNA ORFs, were required to not contain in-frame stop codons, and when appropriate, were matched for overlap with full-length annotated coding ORFs (Methods).

The free energy of pairing between the 16S rRNA and the Shine-Dalgarno sequence (the SD score) is a commonly used predictor of the protein-coding potential

of ORFs based on its requirement in some species for strong translation [31]. For some species this feature alone can separate annotated coding ORFs from mock ORFs remarkably well, as in *B. subtilis* (Fig. 2b, orange and green lines). sRNA ORFs typically have similar SD scores to randomly shuffled sequences, but for many species a substantial number comprise a tail with stronger SD scores than expected by chance, i.e. an excess of sRNA ORFs that “look like” coding ORFs by SD score (Fig. 2b, blue line). This tail of SD scores stronger than expected by chance can be attributed to natural selection acting to preserve translation; therefore, in *B. subtilis*, based only on their SD scores, we can predict that  $33 \pm 9.3$  sRNA ORFs code for proteins (95% confidence interval; Fig. 2b, shaded region). The SD score alone was able to predict at least one coding ORF for 10 of the 14 bacterial species tested (Fig. 2c, white bars; Additional file 1: Figure S1).

A more direct test for natural selection maintaining protein function is the classic  $D_n/D_s$  log likelihood test, which compares the mutation rate at the amino acid level to that at the nucleotide level. Although there are more sophisticated methods that can perform better in some circumstances [23, 24] we use the classic  $D_n/D_s$  test because it explicitly controls for phylogeny, making it applicable in many contexts, and it is independent of codon bias and nucleotide composition, which can then be explicitly captured in an orthogonal measure. Also, this test is relatively robust to missing data, the choice of orthologous species, and the nature of the selection acting on the sequences. Sequences lost in orthologous species or diverged too far away to align are generally treated as missing data and therefore count neither for nor against an ORF. Therefore it is a conservative test that can be applied in an automated manner to species with diverse phylogenetic tree structures. Applying this test to 14 bacterial species yielded predictions for coding sRNA ORFs above background in 4 species (Fig. 2c, grey bars; Additional file 1: Figure S2).

Phase-specific nucleotide bias has been used to successfully predict protein-coding potential in both short and long ORFs [21, 32]. ORFs have biased nucleotide content overall, in specific phases, and other subtle biases such as codon bias. For example, there is a universal bias for purines at the first codon position [33]. Because composition biases differ qualitatively and quantitatively in each species, they must be learned from training data, i.e. real coding ORFs and noncoding sequences. We use a logistic regression method to learn these biases from a subsets of annotated ORFs and mock ORFs in noncoding regions with properties matching the sRNA ORFs in each species (see Methods). The regression then outputs a score for each sRNA ORF representing its likelihood of being coding. Again, sRNA ORFs tend to have higher scores than mock ORFs, with 12 of the 14 bacterial species



tested having an excess (Fig. 2c, black bars; Additional file 1: Figure S3).

One concern with the nucleotide bias method is that sequences recently acquired via horizontal gene transfer may have different biases and thus may suffer from drastically reduced prediction accuracy. To evaluate the extent of this problem, we examined the *SPβ* prophage in *B. subtilis*, a 134 kilobase region having substantially lower GC content than the overall genome (34.6% compared to 43.5%). After training the only on ORF subsets and mock ORFs outside of the *SPβ* region, we calculated the nucleotide bias score for 5941 ORF subsets and mock ORFs coming from 234 full-length ORFs within the region, resulting in 81.0% of instances classified correctly (AUC=0.804). By contrast, training on the ORF subsets and mock ORFs in the *SPβ* region itself resulted in only a marginal increase to 83.0% of instances classified correctly (AUC=0.877) when using 10-fold cross-validation. This indicates that the nucleotide bias measure should be robust to recent horizontal gene transfer.

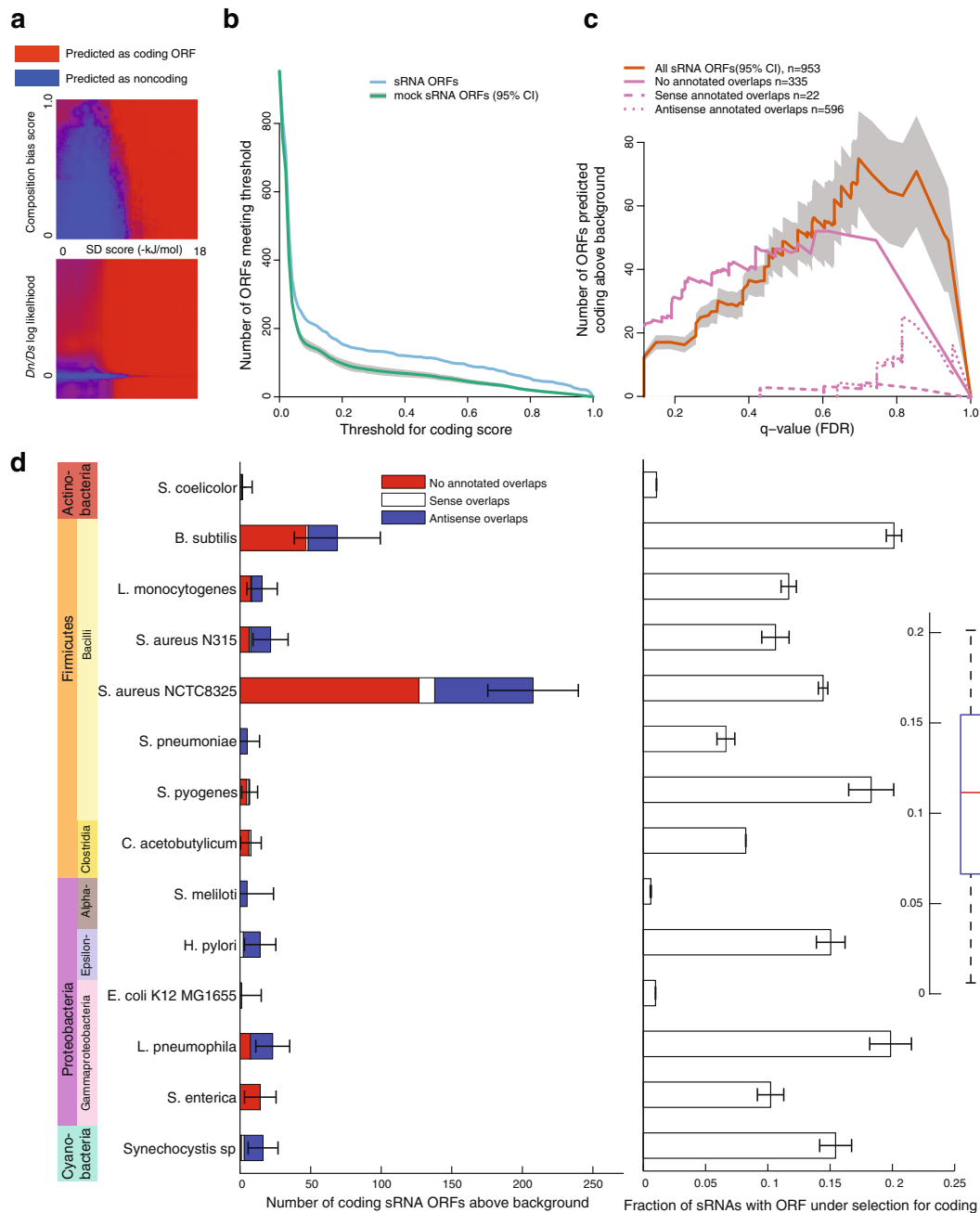
The three sequence features used here are complementary because they rely on different information, so it is not surprising that they predict different numbers of coding ORFs for each species. However, presumably many bona fide coding ORFs should score well in multiple features at once, as they are all signs of natural selection for protein-coding potential. Indeed, *B. subtilis* sRNA ORFs with SD scores better than  $-11$  kJ/mol had strongly and significantly higher  $D_n/D_s$  log-likelihood scores as well as barely higher composition bias scores compared to sRNA ORFs with worse SD scores (Fig. 2d,  $p = 0.00004, 0.037$  respectively, two-tailed Mann-Whitney test). This indicates that the three sequence features may be differently complementary in different combinations, and there may not be a trivial way to combine all three into a single score.

#### Machine learning predicts numerous protein-coding sRNA ORFs in several bacterial phyla

Although the three features may be correlated, they provide complementary evidence, so they must be combined intelligently to maximize their predictive power. Machine learning provides a natural solution for this problem, i.e. training binary classifiers to distinguish between coding ORFs and noncoding sequence. Because each species differs in the phylogeny of available orthologs, history of horizontal gene transfer, codon bias, GC content, and other factors affecting sequence evolution, no single classification scheme can accurately separate coding from non-coding sequence across species. Therefore, we trained individual classifiers on each species, yielding tailored predictions taking into account the predictive power of each sequence feature in each genomic context. We constructed positive training sets based on subsets of annotated coding ORFs, and negative sets based on

mock ORFs in likely noncoding sequence, controlling for relevant properties of sRNA ORFs (see Methods). We trained several types of classifiers on these datasets and selected a bootstrap aggregated (bagged) decision tree classifier, which had consistently high performance on all species, for further analysis (Additional file 1: Figure S4). In general, higher scores for each feature increased the likelihood of predicting an ORF to be coding (Fig. 3a). However, the classifier takes into account some complexities, such as the fact that an SD sequence is not always required for coding, and that negative log likelihoods for the  $D_n/D_s$  test (i.e. more non-synonymous than synonymous mutations) can be evidence for coding as well.

When this classifier is applied to sRNA ORFs, it outputs a “coding score” for each ORF between zero and one, which can be interpreted as a probability that an ORF is coding under the assumption that the prior probability of an ORF to be coding is 50%. We estimate the background distribution of coding scores (i.e. scores for noncoding sRNA ORFs) by applying the classifier to the mock ORFs under 10-fold cross-validation, because they are matched for several relevant properties of the sRNA ORFs. At any selected threshold for coding scores, if more sRNA ORFs meet the cutoff than are expected based on the background distribution, we can attribute the excess to natural selection to maintain coding. Few *B. subtilis* sRNA ORFs meet very stringent thresholds for the coding score, but even fewer mock ORFs do (Fig. 3b, right side). As the threshold is relaxed, the number of sRNA ORFs remaining increases faster than the mock ORFs, so more sRNA ORFs can be predicted as coding. However, most ORFs meet the threshold as it approaches zero and the separation between sRNA ORFs and mocks disappears (Fig. 3b, left side). Therefore both the number of sRNA ORFs predicted as coding and the proportion of predictions that are expected to be false positives (the FDR) depend on the coding score threshold. If we want to be more confident in the predictions of individual coding ORFs, we will make fewer overall predictions, while our best estimate of the total number of coding ORFs will be associated with lower confidence in each individual prediction (higher FDR). Figure 3c illustrates this tradeoff for *B. subtilis*, with the estimated false discovery rate ( $q$ -value) plotted against the number of coding ORFs predicted above background expectation. If we choose the threshold that maximizes sensitivity (i.e. the maximum number of sRNA ORFs predicted as coding, black dashed line in Fig. 3c) and apply a correction for the degree of freedom that this adds (see Methods), we predict that there are in total  $69 \pm 30.5$  *B. subtilis* sRNA coding ORFs (95% confidence interval; Fig. 3c). When we perform this calculation separately for sRNA ORFs that overlap annotated coding ORFs (either in the sense or antisense direction), we find less evidence in support of their coding in *B. subtilis* (Fig. 3c, broken lines).



**Fig. 3 a** Visualization of a machine-learning classifier for combining features into a single predictive score. A bagged decision tree classifier was trained on *B. subtilis* ORF subsets and mock ORFs, and its output is plotted for each value of SD score and composition bias score (above) or  $D_n/D_s$  (below). For each position, the hidden third feature is subsampled and the classifier output is averaged over these possibilities. **b** Number of sRNA ORFs and mock ORFs classified as coding as a function of the coding score threshold in *B. subtilis*. Gray band represents 95% confidence intervals based on 20 mock ORF sets. **c** Number of ORFs predicted as coding above background expectation for *B. subtilis*. For each coding score threshold, a false-discovery rate  $q$ -value is calculated using the ratio between the sRNA ORFs and mock ORFs plotted in **(b)**. The difference between these two, i.e. the number of ORFs predicted coding above background, is plotted on the y-axis in red, with 95% confidence intervals plotted in gray. The cutoff with the highest sensitivity is marked (dashed black line). The calculation is also made separately for subsets of ORFs having no overlaps with annotated coding ORFs, or those with sense or antisense overlaps. **d Left**: Estimated number of ORFs under selection for coding for all species. The numbers predicted for each species were calculated as illustrated in **(c)** at the most sensitive cutoff and a correction was applied for random fluctuations. Error bars represent 95% confidence intervals on the total estimate, and the breakdown by overlap with annotated coding ORFs is represented by colors. **Right**: The predicted coding ORFs are sampled to estimate the fraction of sRNAs having at least one coding ORF, and error bars represent 95% confidence intervals. Inset: The fraction of sRNAs having a predicted coding ORF for each species in box plot form. Box represents first and third quartiles and median; whiskers extend to most extreme values

However, because we do not perform the  $D_n/D_s$  test for these ORFs and because they have other differing properties, we cannot conclude that there are fewer bona fide sRNA coding ORFs that overlap annotated coding ORFs.

Training bagged decision tree classifiers on other bacterial genomes and applying the same test yields between 1 and 207 predicted coding sRNA ORFs, depending on the species (with the high outlier being *S. aureus* NCTC8325, having the most annotated sRNAs by far). Because the predictive power of the sequence features used varies, we cannot directly compare the number of predictions across species, i.e. we cannot conclude that any species has more bona fide coding sRNA ORFs than another. In addition, some species have more short ORFs previously annotated than others, which are excluded from the analysis. Combining the 14 species considered here, we predict that  $409 \pm 191.7$  (95% confidence intervals) previously unannotated small proteins are coded by sRNA ORFs, an average of 29 per species (15 per species excluding the *S. aureus* NCTC8325 outlier). We believe this to be a very conservative estimate even without taking into account the limited nature of current sRNA annotations, for two main reasons: first, we use only a limited set of sequence features that can at best detect a fraction of cases of natural selection; second, the classifier expects coding ORFs to have similar properties to full-length annotated coding ORFs, while they may be expressed at a lower level on average and conserved in a different manner. For example, the estimate for protein-coding sRNA ORFs based only on the composition bias feature (Fig. 2c) is higher for some species than the combined estimate; this is because the SD score and/or  $D_n/D_s$  features hurt the sRNA ORFs in the machine learning classifier more than they help them. Therefore, we can also report a simple, slightly less conservative estimate for the number of coding sRNA ORFs by using only the composition bias feature of  $548 \pm 117$ , an average of 39 per species.

To test the sensitivity of our method, we deleted the annotated ORFs less than 50 amino acids in length from *B. subtilis* 168 and in *E. coli* K12 MG1655. We then ran our analysis on these genomes and evaluated our methods on how well they separated the known short ORFs from intergenic mock ORFs of equivalent size. Out of 101 annotated ORFs less than 50 amino acids in length in *B. subtilis* 168, 82 ORFs had a  $q$ -value of at most 0.05, meaning our machine learning approach separates 81% of annotated ORFs from intergenic mock ORFs with at most a 5% false positive rate (Additional file 2: Table S1). In *E. coli* K12 MG1655, 65/120 (54%) had  $q$ -value of at most 0.05. Thus, our methods were able to separate true coding ORFs from background with appreciable sensitivity in multiple species with varying signals of selection (e.g. the Shine-Dalgarno score is highly informative in *B. subtilis*, but not in *E. coli*).

Because many sRNAs have multiple ORFs, it is not obvious what fraction contain at least one ORF that is under selection to be protein-coding. To estimate this fraction, we used the estimate for the total number of coding sRNA ORFs and randomly chose 100 sets of particular ORFs as coding (Fig. 3c, right). For most species, 5-15% of sRNAs had at least one predicted coding ORF, with  $10.3 \pm 0.5\%$  of sRNAs across all species.

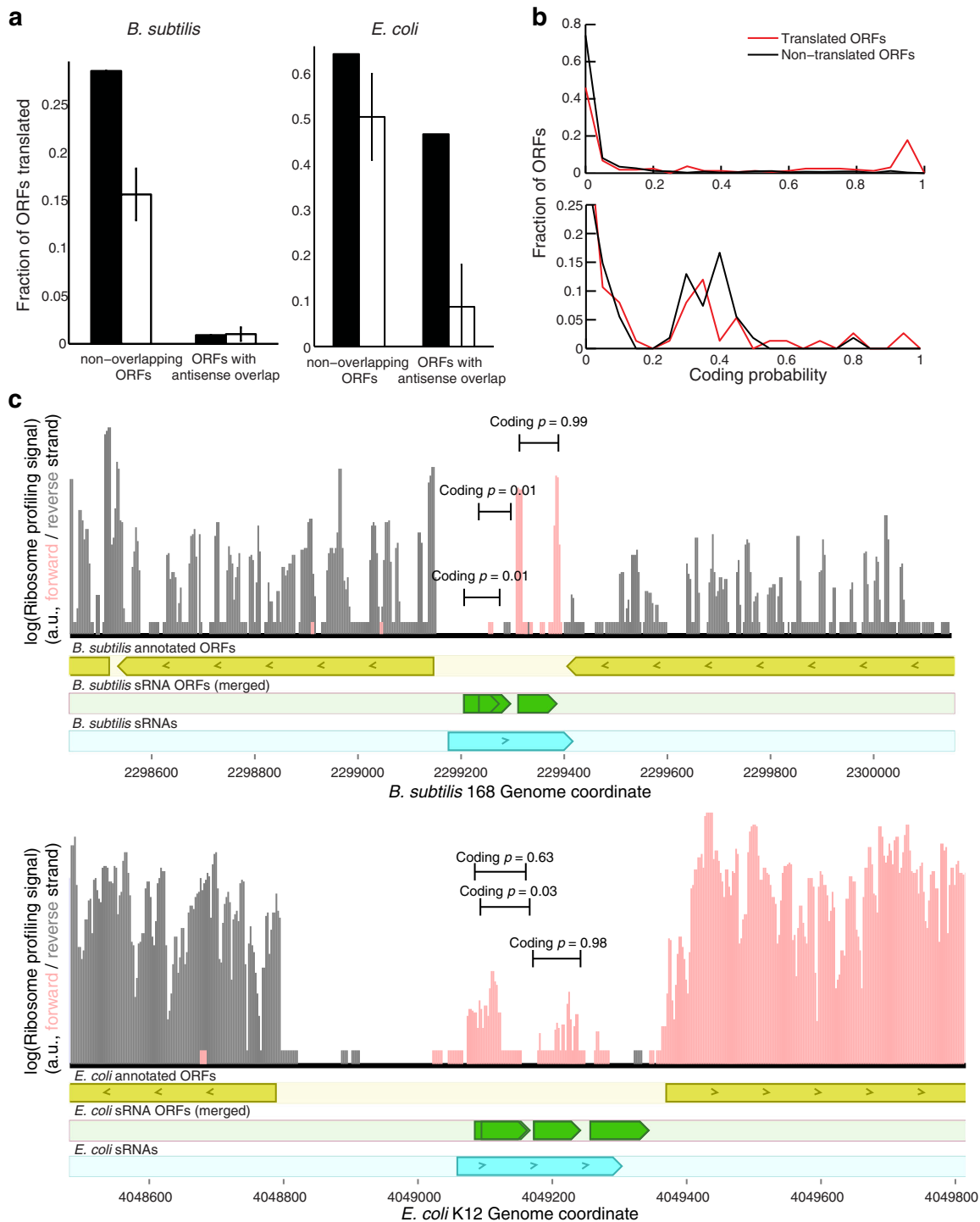
We find almost no evidence in any species for coding of sRNA ORFs overlapping with annotated coding ORFs on the sense strand (Fig. 3c, white bars), but there is a substantial number of predictions overlapping coding ORFs on the antisense strand (blue bars). For example, *L. pneumophila*, which has 40 sRNAs with antisense overlap to annotated coding ORFs [34], has  $16 \pm 5.0$  predicted coding sRNA ORFs in this orientation (Fig. 3c). Overall,  $164 \pm 70$  of the predicted coding ORFs had antisense overlap to annotated ORFs, compared to  $222 \pm 82$  with no overlaps.

#### Many sRNA ORFs are bound by ribosomes and expressed as peptides

The predictions of protein-coding ORFs reflect evidence for conservation of a protein-coding function, which implies as a pre-requisite expression at the RNA level and protein level. Any annotated sRNA must have evidence for its expression at the RNA level, so all sRNA ORFs have the potential to be translated into peptides. Therefore, as an independent experimental validation of our predictions, we looked for experimental evidence of translation from two types of data: ribosome profiling, showing the binding of ribosomes (which correlates in most cases with translation of transcripts), and mass spectrometry, showing the accumulation of protein to detectable levels.

We used ribosome profiling data for *B. subtilis* 168 and *E. coli* K12 from [25] to annotate translated sRNA ORFs. Looking for signal accumulating on either the start or stop codon of ORFs not overlapping annotated coding ORFs (see Methods), we found evidence for ribosome binding in 156 out of 546 *B. subtilis* sRNA ORFs, compared to on average  $85 \pm 15$  expected by chance (based on the translation of mock ORFs in regions not annotated as coding); in *E. coli* 54 out of 84 ORFs had evidence for ribosome binding compared to  $42 \pm 8.0$  by chance (Fig. 4a). Because the ribosome profiling data was strand-specific, we could also test the binding of ribosomes in ORFs in the antisense strand to annotated coding ORFs. In this case, the numbers were 7 of 774 ORFs compared to  $7.8 \pm 5.5$  by chance in *B. subtilis*, and 21 of 45 *E. coli* ORFs compared to  $3.9 \pm 4.2$  by chance. In all, this corresponds to 83 and 17 more unannotated sRNA ORFs bound by ribosomes than expected by chance, respectively. Interestingly, antisense ORFs were enriched for a translation signal in





**Fig. 4** sRNA ORFs with evidence for translation are preferentially predicted to be protein-coding. Ribosome profiling data from [25] was mapped to *B. subtilis* 168 and *E. coli* K12. sRNA ORFs were annotated as translated based only on ribosome profiling signal within three codons of the start or stop codon. **a** sRNA ORFs were separated into those independent of annotated ORFs and those antisense to annotated ORFs. The fraction of ORFs translated (black bar) is compared to mock ORFs matched for length and overlap properties (white bars with 95% confidence intervals). **b** The predicted coding probability of translated sRNA ORFs is compared to non-translated sRNA ORFs in a histogram for *B. subtilis* (top) and *E. coli* (bottom). **c** Top: *B. subtilis* sRNA sbsu2300.1 has three potential coding ORFs, but only one is predicted to be coding. The start and stop codons of this ORF correspond to ribosome profiling peaks, while the ORFs predicted as noncoding do not. Bottom: *E. coli* sRNA seco4050.1 (CsrC) has four ORFs, two of which are predicted as coding and overlap with ribosome profiling peaks. The coding probability can help distinguish between the coding frame for overlapping ORFs, as in the first two in this sRNA

*E. coli* but not *B. subtilis*, but the reverse is true for standalone ORFs.

Some sRNA ORFs with a protein-coding function may not be found by these ribosome profiling experiments because they are only be transcribed and translated under certain conditions. Conversely, translation does not necessarily imply function. Nevertheless, there should be significant overlap between proteins translated in specific conditions and ORFs under natural selection to maintain protein-coding function, since translation is a prerequisite for protein-encoded function. Indeed, more ribosome-bound ORFs had more high coding scores than non-ribosome-bound ORFs, meaning the coding score statistic was able to predict which sRNA ORFs were bound by ribosomes (Fig. 4b). This predictive power was not due to the strength of the Shine-Dalgarno sequence alone, implying that selection for protein-coding function reflected in the  $D_n/D_s$  test and composition bias was correlated with translation (Additional file 1: Figure S5A). For *E. coli* K12 MG1655, we only predicted 1-2 sRNA ORFs under selection to maintain protein coding, but there was still significant evidence for ribosome binding of sRNA ORFs that appeared to be correlated to the coding score. Our methods may be unable to predict these protein products with statistical confidence even if they are functional, for example because the SD score has little predictive power in *E. coli* (Additional file 1: Figure S1), and because translated sRNA ORFs may have different expression profiles and different sequence features compared to annotated ORFs. In other words, even though sRNA ORFs may not rise above mock ORFs in terms of coding features such as the  $D_n/D_s$  test, those coding features are still predictive of translation.

Ribosome-bound ORFs typically had peaks of ribosome profiling signal concentrated at the start or stop codons, as was the case with many full-length annotated ORFs (Fig. 4c). Many sRNAs that were previously identified only by high-throughput screens had evidence for ribosome binding, such as the *B. subtilis* sRNA sbsu2300.1, which was defined based on tiling array data ([35], Fig. 4c, top). Some sRNAs have multiple ORFs between 10 and 50 amino acids long, making the assignment of ribosome profiling coverage to individual ORFs ambiguous. For example, the CsrC noncoding RNA in *E. coli* has ORFs overlapping in different frames (Fig. 4c, bottom). In this case, only one ORF under the coverage peak had a coding score of greater than 0.5, showing that the coding score can help to prioritize ORFs for follow-up experiment even when evidence for translation is ambiguous. Although these analyses were performed by dividing sRNA ORFs by the mere presence or absence of ribosome profiling reads (see Methods), similar results were obtained when quantifying the number of reads in each ORF (Additional file 1: Figure S5B).

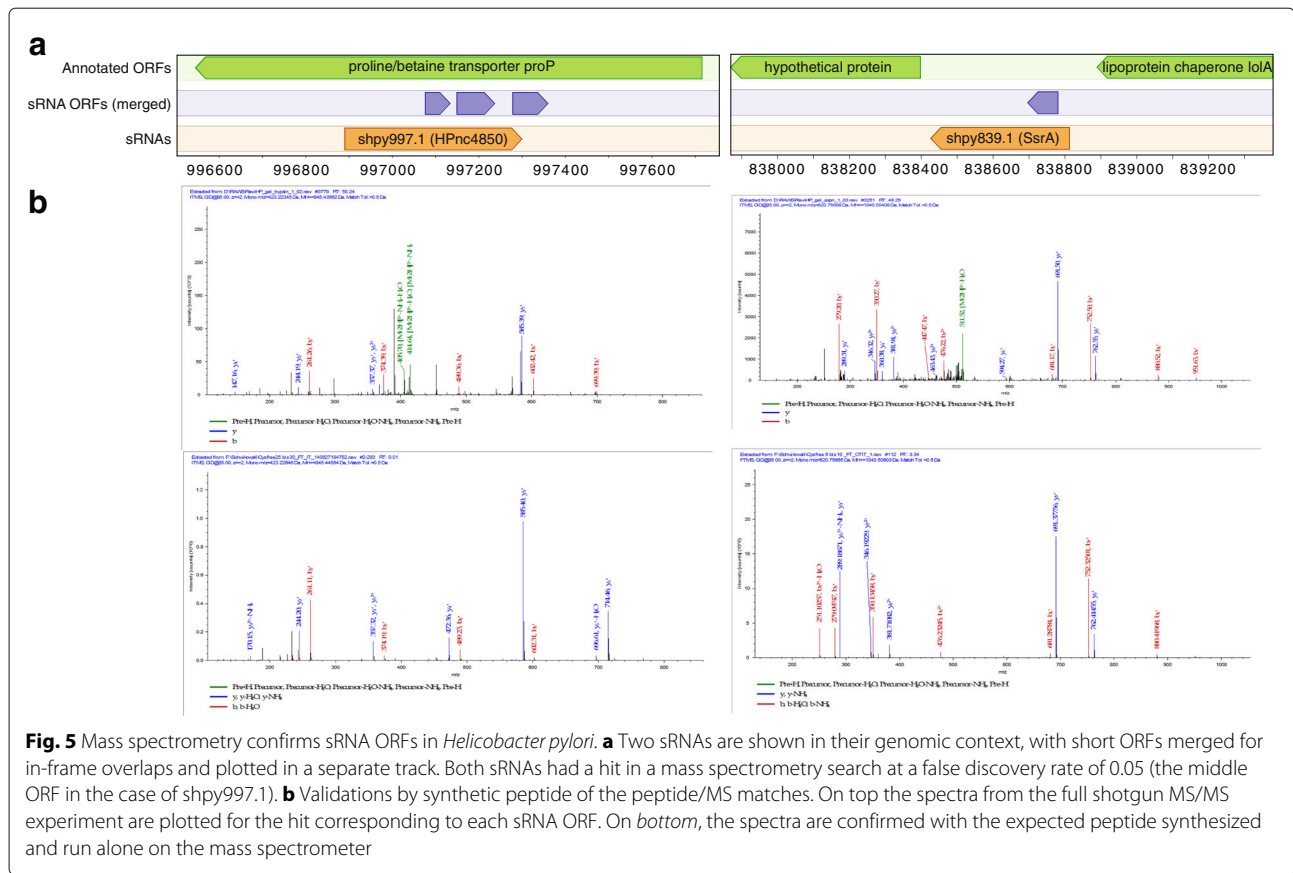
### Mass spectrometry confirms small proteins from sRNA ORFs

Many sRNA short ORFs could be occasionally translated without accumulation of a protein product to appreciable levels if the protein were quickly degraded. Conversely, detection of sRNA ORF protein products in cells by methods with limited sensitivity would be strong evidence against this possibility. Therefore, we took advantage of a mass spectrometry dataset from an experiment specifically designed to find small proteins to search for sRNA ORF products in *Helicobacter pylori*. For the purposes of this search, we included the widest set of sRNA ORF predictions possible, filtering neither for predicted coding score nor potential overlaps with annotated full-length ORFs. This search resulted in 25 peptide hits from 17 sRNA ORFs at FDR less than 0.05, with 6 hits having FDR less than 0.01. Despite the statistical significance of these matches, most novel proteins were identified by only a single peptide, making their identification unreliable. The ultimate confirmation of protein identification must be made by matching the observed spectrum to that of a synthetic peptide with the expected sequence. Therefore, we synthesized each of our putative matching peptides. 17 of the synthetic peptides resulted in useable mass spectra, of which 6 tested peptides were validated with a spectrum match. One of the hits, shpy580.1.10, was likely due to an alternative start site of a previously-annotated ORF, i.e. a misannotated or alternative start codon (Additional file 3: Table S2). Another peptide hit in shpy1027.1.1 is for a short sequence shared by an annotated ORF, so may not be considered strong evidence for a novel protein-coding sRNA. However, another peptide comes from an sRNA labeled here as shpy839.1, which is a tmRNA with a known (but unannotated) coding peptide that helps to recycle stalled ribosomes (Fig. 5, Additional file 3: Table S2). One other, shpy997.1.2 appears to be a bona fide novel coding peptide arising from an sRNA antisense to an annotated ORF (Fig. 5).

### sRNA ORFs are enriched for certain functional annotations

Some of the sRNAs considered in our study are already known to encode functional peptides in addition to their non-coding functions. For example, the transfer-messenger RNA gene SsrA has tRNA-like properties but also codes for a short peptide, which is rarely annotated. We predicted the coding peptide for SsrA in *S. pneumoniae* at a  $q$ -value of 0.10, and in two other species at a coding score  $p \geq 0.5$ . Most other sRNAs with known coding ORFs were already annotated correctly, for example SgrT in *E. coli* and SR1P/YkzW in *B. subtilis*, so they were skipped by our methods.

To systematically find other annotations for our predicted sRNA coding ORFs, we searched their translated sequences against known proteins using protein BLAST.



**Fig. 5** Mass spectrometry confirms sRNA ORFs in *Helicobacter pylori*. **a** Two sRNAs are shown in their genomic context, with short ORFs merged for in-frame overlaps and plotted in a separate track. Both sRNAs had a hit in a mass spectrometry search at a false discovery rate of 0.05 (the middle ORF in the case of shpy997.1). **b** Validations by synthetic peptide of the peptide/MS matches. On top the spectra from the full shotgun MS/MS experiment are plotted for the hit corresponding to each sRNA ORF. On bottom, the spectra are confirmed with the expected peptide synthesized and run alone on the mass spectrometer

43 sRNA ORFs with coding score  $p \geq 0.5$  had a significant match with a similar length and an informative description, including 12 with coding  $q \leq 0.25$  (Additional file 4: Table S3). Many of these BLAST hits were annotated in the same genus as the sRNA ORF, suggesting that they are direct orthologs that have escaped annotation so far. In one case (sbsu22741.1), a *B. subtilis* sRNA ORF matched a type I toxin-antitoxin system annotated in the same species but added to genbank after we performed our analysis. Another notable example is a gene with BLAST hits to many proteins annotated as phenol-soluble modulins, that was nevertheless unannotated in *S. aureus* N315 (Additional file 4: Table S3). Phenol-soluble modulins are a crucial virulence factor for methicillin-resistant *S. aureus* [36], so it is surprising that the protein in *S. aureus* N315 was never annotated as such based on homology alone. Still, the 43 sRNA ORFs with informative BLAST hits leave more than 350 new predicted coding ORFs with no known homolog.

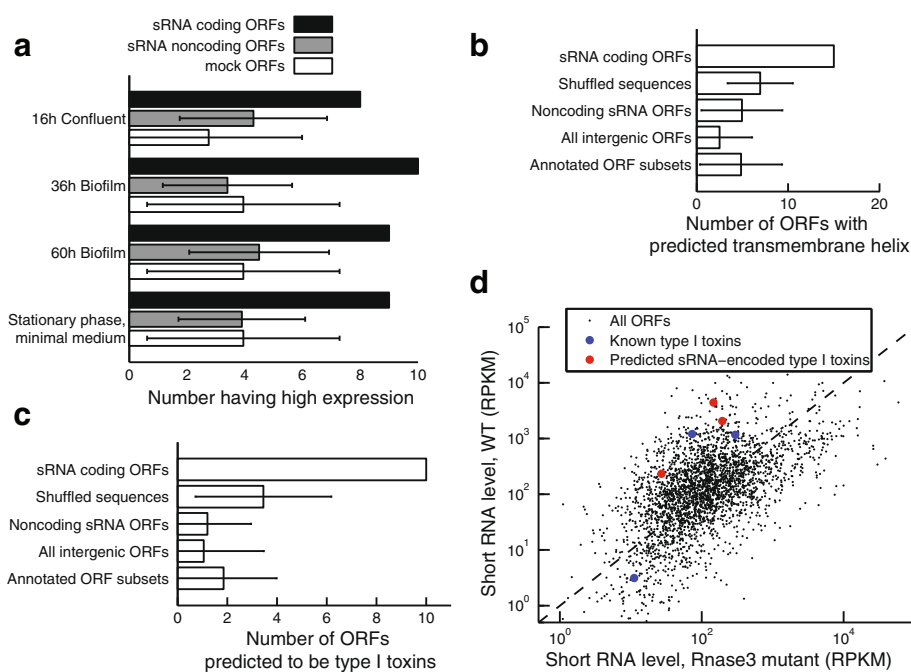
To identify potential functions of our sRNA predicted coding ORFs lacking annotated homologs, we examined a large compendium of *B. subtilis* tiling array data for expression patterns [26]. sRNAs that contained a predicted coding ORF were enriched for having high expression in biofilms, stationary phase in minimal media, and other confluent conditions, when compared to sRNAs

having ORFs not predicted to be coding, or to mock ORFs (Fig. 6).

One of the species we analyzed, *Streptococcus pneumoniae*, was the subject of a recent genome-wide screen for sRNA function in virulence [37]. We searched for predicted coding ORFs in the *S. pneumoniae* sRNAs as defined by Mann et al., resulting in two having coding  $q \leq 0.25$ , the R12 and F32 sRNAs. Both had a phenotype when knocked out in the study, with the R12 mutant having reduced fitness in blood and reduced nasopharynx colonization, and the F32 mutant having reduced fitness in Lung infection. The F32 RNA is also known as SsrA, the tmRNA gene with known peptide-encoding function.

**Many sRNA ORFs are part of type I toxin/antitoxin systems**

An obvious trend in the annotations of the homologs of our predicted coding sRNA ORFs is that 27 were membrane-associated, including 6 matches to holins or type I toxin-antitoxin systems, which typically encode membrane-associated small protein toxins (Additional file 4: Table S3). Therefore we hypothesized that a common function for many sRNA coding ORFs would be to act as membrane-binding proteins and possibly also as toxins. Translated sRNA ORFs with coding  $q \leq 0.25$  had a predicted transmembrane helix more often than shuffled sequences or equivalent-length regions of annotated



**Fig. 6** sRNA coding ORFs are enriched for predicted functions including type I toxins. **a** *B. subtilis* 168 sRNAs containing at least one predicted coding ORF were compared to sRNAs with only noncoding ORFs or mock ORFs. Four growth conditions were enriched for high expression of predicted sRNA coding ORFs (1.5-fold increase in coding ORFs over both controls, minimum 5 sRNA coding ORFs). **b** For all predicted coding ORFs in all species ( $q \leq 0.25$ ,  $n = 120$ ), the number of ORFs with at least one predicted transmembrane helix is compared to the number expected by chance based on their shuffled amino acid sequences, sRNA ORFs with  $q \geq 0.25$ , all intergenic ORFs, or size-matched subsets of annotated ORFs. Error bars represent 95% confidence intervals. **c** As in (**b**), except that the number of ORFs predicted as type I toxins by the physicochemical classifier are plotted. **d** Degradation levels of *S. aureus* ORFs as measured by sequencing of short RNA fragments are plotted. ORFs subject to degradation of double-stranded RNAs have higher levels in wild-type bacteria (y-axis) than in RNase III mutant bacteria (x-axis). Known and predicted type I toxins are highlighted (blue and red, respectively)

ORFs, notably in *B. subtilis*, *S. aureus* N315, and *S. meliloti* (Fig. 6b, Additional file 1: Figure S6). Altogether, the number of coding sRNA ORFs with at least one transmembrane helix in excess of controls was 8.0-13, depending on the type of control, 6.7-11% of the 120 predicted coding ORFs.

To more directly find unannotated toxins, we developed a machine-learning classifier to differentiate type I toxin peptides from non-toxic short proteins based on the physicochemical characteristics of their amino acid sequences, a method adapted from Torrent et al. [38]. We trained a Random Forest classifier on type I toxin peptides found via an exhaustive PSI-BLAST search [39], with length-matched Uniprot sequences as negative controls. This classifier achieved a sensitivity (true positive rate) of 64.3% with a false positive rate of only 3.6% in 10-fold cross-validation of the training data. When applied to predicted sRNA ORFs with coding  $q \leq 0.25$ , 10 were predicted type I toxins compared to only 2-4 expected by chance based on shuffled sequences or controls from other ORF types (Fig. 6c). Of these 10 predicted toxins, three have annotated holin or type I toxin BLAST hits (sbsu2274.3:2273534-2273824.2, ssau1857.1:1856223-

1856978.5, sbsu2679.1:2678645-2679017.7). This result implies that coding sRNA ORFs are enriched for predicted type I toxin-antitoxin systems compared to chance.

If these predicted sRNA coding ORFs are really part of type I toxin-antitoxin systems, their expression should be controlled by the formation of a double-stranded RNA followed by degradation mediated by RNase III. We examined short RNA fragment sequencing data from a recent study for signs of these degradation products in *S. aureus* [40]. Of the three type I toxins reported by Fozo et al. [39], two had very high levels of RNA degradation signal (93rd percentile or above, Fig. 6d, Additional file 1: Figure S6). The third was poorly expressed. This signal could not be accounted for by background degradation, as it did not persist in an RNase III mutant (78th percentile or below). We reasoned that novel type I toxins should have similar signal for RNase III dependent degradation. Because we did not predict any confident type I toxins in *S. aureus* sRNAs, we expanded the search to all short ORFs not overlapping annotated ORFs. Three of these ORFs had coding  $p \geq 0.5$  and were predicted type I toxins. All three had high RNA degradation signal that was dependent on RNase III to a similar extent (Fig. 6d, Additional

file 1: Figure S6). Expanding this search to *B. subtilis*, 3/5 predicted type I toxins in sRNA ORFs had high RNA degradation signal (85th percentile and above, Additional file 1: Figure S6), although the lack of data from an RNase III mutant precludes their confident confirmation.

### Web Server

We anticipate that this collection of predicted bacterial sRNA coding ORFs with quantified uncertainty will be of broad utility for experimental microbiology. To make these results as accessible as possible, we created a user-friendly web site providing all of our coding ORF predictions for searching or browsing. A parser was written in Python to browse the results of all analyses, formatting and storing them in a couchdb database (<http://couchdb.apache.org/>). NoSQL technologies were used to dynamically display bioinformatics results. The resulting website is available at the URL <http://disco-bac.web.pasteur.fr>.

For each species (cf. Table 1), a summary of the coding sRNA ORF search statistics is available. From this first page, the user may view the detailed characteristics including coding score for each sRNA ORF, or may expand the search to all intergenic ORFs. In each case, summary figures, for example displaying the ROC curves for prediction accuracy, are also shown. Furthermore, an interactive genome browser [41] was embedded to visualize the genomic context of sRNA ORFs. Tracks show genome position, annotated sRNAs, all sRNA ORFs (with and without in-frame overlaps merged), as well as full-length annotated ORFs. We expect that the accessibility

of this data should empower both computationalists and experimentalists to follow up on these results with ease.

### Discussion

Our survey was enabled by DiSCO-Bac, a flexible machine learning method to find coding sRNA ORFs based on simple sequence features and comparative genomics. Although relatively straightforward, the method is conservative and versatile, making as few assumptions as possible while still being able to incorporate a wide range of evidence. Our use of “mock ORFs” to measure the empirical background distributions of our sequence features is crucial for controlling for several biases that would otherwise strongly skew the analysis, for example, the length distribution and number of ORFs, the GC content of the upstream sequence, higher-order oligonucleotide frequencies, the length of runs of conservation, the effects of overlap with annotated coding ORFs, and the frequency of horizontally transferred sequence. We are careful to guard against double-counting by merging all overlapping ORFs before estimating the number that are coding, to correct for the freedom in fitting the coding score threshold to the data, and to use only a single machine learning algorithm with a single parameter set to guard against researcher degrees of freedom — all of which are common problems in this type of bioinformatic analysis.

Although ultimate proof of protein-coding function for these predicted proteins will await shotgun mass spectrometry targeted towards finding small proteins coupled to genetic and molecular validation, we note that validation at the protein level is extremely difficult. In *H. pylori*, only a single novel short ORF was confidently validated by mass spectrometry using a synthetic peptide, despite dozens of putative hits in the shotgun search. Small proteins are notoriously difficult to detect with shotgun mass spectrometry, because protein purification, electrophoresis, and chromatography all bias towards larger proteins, and because small proteins have few tryptic peptides available to search against. However, we believe more attention paid towards this topic by the proteomics community will yield advances through improved techniques, such as the incorporation of sRNA ORF coding predictions with selected reaction monitoring.

Our evidence suggests that even with proof of proteins encoded by sRNA ORFs, many will likely be nonfunctional. Our ribosome profiling analysis suggests widespread translation at some low level in sRNA ORFs, and our sRNA ORF abundance result suggests weak or no selection against having sRNA ORFs. Therefore, the default state of sRNAs may be to have short ORFs coming into and out of existence, with occasional translation but little functional impact. However, because of our comparison to mock ORFs, we have quantified how often natural selection has put pressure on the Shine-Dalgarno

**Table 1** Species used in this study and associated Genbank accession numbers for their chromosomal genome

Species	Genbank version
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	AL009126.3
<i>Clostridium acetobutylicum</i> ATCC 824	AE001437.1
<i>Escherichia coli</i> str. K12 substr. MG1655	U00096.2
<i>Helicobacter pylori</i> 26695	CP003904.1
<i>Legionella pneumophila</i> subsp. <i>pneumophila</i> str. Philadelphia	AE017354.1
<i>Listeria monocytogenes</i> strain EGD	AL591824.1
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Typhimurium SL1344	FQ312003.1
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> N315	NC_002745.2
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> NCTC 8325	CP000253.1
<i>Sinorhizobium meliloti</i> 1021	AL591688.1
<i>Streptococcus pneumoniae</i> TIGR4	AE005672.3
<i>Streptococcus pyogenes</i> MGAS5005	CP000017.1
<i>Streptomyces coelicolor</i> A3(2)	NC_003888.3
<i>Synechocystis</i> sp. PCC 6803	BA000022.2

sequence, the amino acid sequence, or the coding composition to be more coding-like than expected by chance. Taking these ideas together, it is likely that widespread coding ORFs are preferentially newly evolved and conserved over shorter distances than full-length proteins.

No pre-existing method is ideal for estimating the number of sRNA coding ORFs as we have. Some comparative genomics methods take into account more information than the  $D_n/D_s$  test, but more complexity can make algorithms more brittle. For PhyloCSF [23], a greater number of parameters to fit can be problematic for small bacterial genomes and this method remains untested on prokaryotes. RNaCode [24] handles multiple alignment issues like insertions and deletions intelligently, but because it does not take into account phylogenetic structure it relies on careful selection of orthologous species to yield relevant results, making it difficult to apply on a large scale. Warren et al. [8] used a clever BLAST-based approach to quickly find new genes, but this is less sensitive than  $D_n/D_s$ , which is aware of phylogeny and mutations at the DNA level. Other methods are either ad-hoc and difficult to apply to other species [22] and/or do not incorporate both sequence features and comparative genomics [21].

A key advantage of our analysis is the effective aggregation of predictions with weak confidence (high FDR) to make statistically accurate statements about the set of sRNA ORFs as a whole. At first glance, it might seem strange that one can have reasonable confidence in the number of coding ORFs when the confidence in any particular coding ORF prediction is only 50% or less. The situation is analogous to estimating the mean of a probability distribution with a high variance. Each individual observation from the distribution is likely to be far from the mean, but when aggregating over multiple observations, the sample mean converges rapidly to the true expected value as sample size increases, as proven by the law of large numbers and the central limit theorem. There are many instances of biological insight gained by aggregating noisy predictions; for example, the number of human genes that are conserved targets of microRNAs can be quantified despite having low confidence in most individual predictions [42].

One caveat with our method is that it depends on the sRNA annotations, which reduce the amount of ORFs to consider and therefore reduce the background noise and the problem of multiple test correction. For example, *S. aureus* N315 has 154 annotated sRNAs, compared to 1233 in NCTC8325. As a result, we predict  $21.6 \pm 12.4$  *S. aureus* N315 coding ORFs compared to  $207 \pm 32$  in *S. aureus* NCTC8325 (Fig. 3d). Future additions to sRNA annotations will likely reduce these differences. Additionally, *S. meliloti* appears as an outlier in this analysis, with a large number of sRNAs annotated but a small number of coding ORFs predicted. This could be attributed to sRNA

annotations, as a large number of its *S. meliloti* sRNA annotations overlap an annotated coding ORF in the sense orientation, and we discard sRNA ORFs with in-frame overlap to annotated ORFs. We found very little evidence supporting coding of ORFs with out-of-frame sense overlap to annotated ORFs in any species, explaining the *S. meliloti* outlier.

We expect that the true number of coding sRNA ORFs is much higher than what we predicted for several reasons: 1) sRNA annotations are based on a subset of published studies which have profiled a subset of potential conditions, and therefore are likely incomplete, 2) the sequence features we used have limited predictive power, for example very few ORFs have strong Shine-Dalgarno sequences in some species, 3) we only consider AUG start codons, while 10.0% of annotated *E. coli* ORFs begin with GUG or UUG, 4) our mock ORFs used for non-coding background calculations likely contain some real ORFs because of missed annotations and nonstandard start codons, 5) we miss unannotated sRNA ORFs smaller than 10 or larger than 50 amino acids, and 6) the machine-learning classifier assumes that sRNA ORFs have similar properties to full-length ORFs, but many bona fide small proteins are likely poorly expressed and/or have different evolutionary histories and thus will have different sequence properties. For example, some sRNA ORFs may have recently evolved *de novo* and therefore would not have orthologs. For all these reasons, we expect that the true number of small proteins encoded by sRNA ORFs is much larger than what we reported, and future improvements to methodology and sRNA annotations will increase these estimates substantially.

## Conclusions

We present here a broad survey of protein coding in bacterial sRNAs, to our knowledge, the first of its kind. We combined phylogenetics and known biological effects into a machine learning classifier, a method we call DiSCO-Bac, Discovery of sRNA Coding ORFs in Bacteria. We found that more than half of sRNAs contain canonical ORFs between 10 and 50 amino acids in length, and conservatively at least 10% of sRNAs contain ORFs under selection to maintain protein-coding function. In each species considered here, an average of 29 new protein-coding ORFs were predicted in annotated sRNAs. We showed experimental evidence that many of these ORFs were bound by ribosomes (using ribosome profiling data) and that some protein products accumulated to detectable levels (using mass spectrometry). Although few of the predicted protein products had orthologs with annotations, we nonetheless found clues to potential commonly encoded functions. In *B. subtilis*, sRNAs with predicted coding ORFs were preferentially expressed in biofilms and confluent conditions.

Overall, sRNA predicted coding ORFs more often had transmembrane domains than expected by chance. Building a machine learning classifier to predict novel type I toxins, we found that predicted sRNA coding ORFs were enriched for predicted toxins, which were associated in *S. aureus* and *B. subtilis* with small RNA degradation products characteristic of control of expression by an RNA antitoxin (Fig. 6). Chromosomal type I toxins are often involved in forming bacterial persisters and biofilms [43], consistent with their expression tendencies in *B. subtilis*. We expect that type I toxins constitute a substantial minority of our novel coding ORFs.

Although many are careful to point out that bacterial sRNAs may code for small peptides [1], the term “noncoding” is sometimes used interchangeably, and a trans-acting function for sRNAs is sometimes assumed [4]. Given that more than 50% of sRNAs contain ORFs and up to 60% of *E. coli* sRNA ORFs have some evidence for translation in a single condition (Fig. 4a), it is clear that sRNAs should not be assumed noncoding and should not be assumed to have only a trans-acting antisense function. Underscoring this point, for many species, a high fraction of antisense ORFs are predicted to be coding (Additional file 1: Figure S4C). The fact that two sRNAs involved in *S. pneumoniae* virulence, R12 and F32, have confident predicted coding ORFs underscores that the possibility of protein-coding function must always be considered. Of course, there are many examples of sRNAs that have both protein-coding and RNA-level functions, so one does not preclude the other [9–12]. Antisense transcription is prevalent in many bacteria [4, 40] but these RNAs are even more commonly assumed to function via antisense binding. However, almost half of our newly-predicted coding sRNA ORFs had some antisense overlap with annotated ORFs, and these ORFs were found to be commonly bound by ribosomes (Fig. 4). For many species, over 20% of antisense ORFs are predicted to be coding (Additional file 1: Figure S4C). This should serve as another reminder that apparent antisense RNA function does not preclude protein-coding potential.

Although our ORF predictions are certainly not perfect, it is important to provide predictions with quantified uncertainty, both to define a lower limit on the size of this class of proteins, and to provide realistic expectations for experimental follow-up. In addition to presenting our detailed predictions here (Additional file 5: Table S4), we distributed our predictions in a user-friendly web database (<http://disco-bac.web.pasteur.fr>) combining summary results, individual ORF characteristics, and tools for sorting and visualizing genomic context. These easily-accessible predictions should hasten the experimental characterization of sRNA ORFs. With an average of over a dozen new protein-coding genes predicted in each bacterial species, we expect that future experiments

will elucidate novel functions for sRNA ORFs for years to come.

## Methods

### Genomic data collection

Bacterial genomes and ORF annotations were downloaded from Genbank, with accession numbers found in Table 1. Bacterial sRNA annotations were downloaded from BSRD version 1.2 [3]. BSRD annotations were supplemented by sRNA annotations that were not in BSRD from two recently published studies on *B. subtilis* 168 (Table S6.2 of [26]) and *S. aureus* NCTC8325 (Table S6 of [27]).

### Sequence features

sRNA ORFs between 10 and 50 amino acids in length were found assuming ATG as the only start codon. The 50 amino acid maximum length was applied to account for full-length ORFs that would be found by conventional algorithms but were not included in a particular genome's Genbank annotations. When an ORF extended past the 3' end of an annotated sRNA, it was extended until the nearest in-frame stop codon, allowing for inaccurate sRNA 3' end annotations. sRNA ORFs having in-frame overlap in the same sense as an annotated coding ORF were filtered out. When all intergenic ORFs were considered, no overlap was allowed with annotated coding ORFs, regardless of the frame. The number of ORFs expected by chance (Fig. 1b) was generated by shuffling the sRNA sequences one thousand times and counting the resulting ORFs using the same procedure as above. In many instances, multiple start codons could be matched to the same stop codon, corresponding to overlapping in-frame protein sequences. In all predictions of coding ORFs (Figs. 3, 4, and 6, these in-frame overlaps were merged, resulting in only the longest ORF, which avoids double-counting coding ORFs. The ORF with the highest coding score was retained in these cases.

Shine-Dalgarno sequence strength (the SD score) was calculated as in Ma & Karlin [31] using their published anti-SD sequences or when unavailable, the corresponding section of the species' 16S rRNA sequence downloaded from Genbank. For each ORF or mock ORF, ensemble free energies of SD binding were calculated by hybridizing the twenty basepairs upstream of the start codon to the anti-SD sequence using the “hybrid” program from the UNAFold package [44]. Energies reported are at 37 degrees Celsius and in units of -kJ/mol.

Composition bias was calculated using the 69-parameter Z-curve transformation, i.e. phase-specific mononucleotide frequency at each position, and di- and tri-nucleotide frequency at frame zero only [32]. The Z-curve transformation is a numerical representation of the codon bias of an ORF, but also of amino acid frequencies

and nucleotide biases such as GC content and purine bias. Z-curve transformations of ORFs and matching intergenic controls were then used to train a logistic regression classifier in Weka version 3-7-9 [45] with a ridge parameter of  $10^{-8}$  yielding scores ranging between zero and one, representing the estimated probability that a sequence is protein-coding. We found that this classifier was more robust and more interpretable than the original Fisher discriminant method used by Gao and Zhang [32]. We skipped any ATG codons when scoring any sequence, because they are, by definition, almost never found in the intergenic mock ORFs.

Density plots in Fig. 2 and Additional file 1: Figures S1 and S3 were smoothed using a gaussian kernel using the *density* function in R with *adjust=0.7*. 95% confidence intervals were  $\pm 1.96$  times the standard deviation of the control sets.

### Comparative genomics

Orthologous sequences were selected with a blastn search against all complete bacterial genomes in Genbank with an E-value cutoff of 0.05 and parameters *-word\_size 7 -gapopen 5 -gapextend 2 -penalty -3 -reward 2*. After running blastn for all full-length annotated ORFs in the reference species, species with at least 50% of the maximum number of hits were included as orthologous species; in other words, 50% of the maximum was our ad-hoc threshold for the number of total genes shared with species to be included in the tree. For this analysis, plasmids were included in the search and counted towards the number of hits for a species. To construct a phylogenetic tree, annotated ORFs in the reference sequence were truncated to a maximum of 5000 nucleotides, re-blasted against the orthologous species, and aligned using Clustal Omega [46] with default parameters. The alignments were then concatenated and analyzed with the dnaml program of the PHYLIP package using default options [47]. To limit memory usage and running time, we randomly subsampled the sites to yield a maximum of 5 million total nucleotides including all species. For species having more than fifty members of the tree at this stage, we pruned the tree using a greedy algorithm to remove redundant leaves and decrease computational complexity, i.e. we iteratively removed one random species from the pair with the shortest intervening branch length until only fifty leaves remained. At this point, multiple alignments were recreated as above and the phylogenetic tree recalculated using the fifty remaining orthologous species. Any gap positions in the reference organism sequence were spliced out of the multiple alignment, and in-frame stop codons in non-reference species were replaced by gaps. These steps were intended to maximize the amount of data to the  $D_n/D_s$  tests while remaining robust to indels and substitutions in the genome assembly and alignment

errors. Misaligned sequences statistically add equally to  $D_n$  and  $D_s$ , thus providing evidence against protein-coding potential, and therefore making the method more conservative with respect to predicting functional coding sequences. Likewise, we did not explicitly test for the presence of start or stop codons in the orthologous sequences. This choice was also deliberately made to provide more information, while making the method more conservative with respect to predicting coding.  $D_n/D_s$  tests were performed using the codeml program of the PAML package using equal amino acid distance and one site type [48]. Codeml was run once with a fixed omega of one and once with a variable omega, and the reported log likelihood is the difference between the log likelihood of the two models (or zero if the fixed-omega model was more likely than the variable-omega model). We did not perform the  $D_n/D_s$  test on sRNA ORFs overlapping with annotated coding ORFs on either strand, as, in these cases, any signal for selection could not be solely attributed to the sRNA ORF.

To search for homologs to amino acid sequences, command-line BLASTP was used to search the non-redundant protein database (Nr) with the parameters *-task blastp-short -evalue 0.25*. The search was performed both with and without composition-based statistics (*-comp\_based\_stats 0*), and all matches were filtered to be within a two-fold difference in length to the predicted short protein. Any matches with E-value  $< 0.25$  in either search were considered homologs. Homologs with descriptions containing “hypothetical”, “uncharacterized”, “unknown”, “unnamed”, “predicted”, “undefined function”, or beginning with “conserved domain protein” were considered to be hypothetical proteins.

### Machine learning

All machine learning was performed with Weka version 3-7-9 [45]. When positive and negative sets were not perfectly matched for size, a cost-sensitive classifier was used to ensure an equal weighting of positive and negative training data. For each sequence feature, a unique set of negative controls (which we refer to as “mock ORFs”) was selected to best match relevant properties, and positive controls were sampled from annotated ORFs. These controls were used for statistical tests, plotting figures, and as training data for the machine learning step of coding prediction.

sRNA ORFs having any overlap with an annotated coding ORF were treated separately and were given their own mock ORF sets as described below. The conservation analysis was not performed for these ORFs and was treated as missing data, because the conservation of the annotated ORF would typically have overwhelmed signal for conservation out-of-frame or in the antisense orientation.



For Shine-Dalgarno sequences, the scores of the 20 nucleotides upstream of all annotated ORFs were used as positive controls, as in Ma & Karlin [31]. These sequences were shuffled (i.e. randomly reordered) twenty times and re-scored to generate negative control sets controlled for nucleotide composition in each upstream region. The same was done for sRNA ORFs.

For conservation analysis, positive and negative control sets were matched for ORF length and were not shuffled, therefore preserving frame-specific composition and spatial clustering of conservation. Instead, each positive control set consisted of in-frame subsets of annotated ORFs matching the length distribution of sRNA ORFs. Each negative control set consisted of mock ORFs with the same length distribution as sRNA ORFs but randomly taken from “intergenic” regions between annotated coding ORFs, also excluding all sRNA ORFs between 10 and 50 amino acids long. Mock ORFs did not start with an ATG codon, but were constrained not to contain an in-frame stop codon. Twenty sets of positive and negative controls were generated, with each set equal in size and length distribution to the sRNA ORFs. The resulting background estimate of noncoding background is conservative, because the mock ORFs may contain coding ORFs not starting with ATG codons.

Positive and negative controls for the composition bias analysis were the same as those used for the conservation analysis except in the case of sRNA ORFs having overlap with annotated coding ORFs. For each of these sRNA ORFs, twenty mock ORFs were randomly selected with the same overlap length with an annotated ORF in the same orientation (and by implication the same reading frame). For example, an sRNA ORF of 50 nucleotides with its last 10 nucleotides overlapping the 3' end of an annotated coding ORF would have a negative control set of 20 corresponding mock ORFs having 40 nucleotides in intergenic regions and the last 10 3' nucleotides overlapping the 3' end of a randomly selected annotated coding ORF. These mock ORF sets are also required to not contain sRNA ORFs between 10 and 50 amino acids long or in-frame stop codons.

To generate the coding score, we used a two-step classification procedure. First, a logistic regression classifier was trained for composition bias as described above, yielding the composition bias score for each ORF. Next, the SD score, log likelihood of the  $D_n/D_s$  test, and the composition bias score were combined into a training set. As a starting point, the 20 control sets from the conservation and nucleotide composition analysis were used. The positive control sets were matched with the actual SD score of the annotated ORF containing the ORF subset, while the negative control sets were randomly matched to one of the shuffled SD scores. These data were used to train a bagged decision tree classifier using the Weka

Bagging class with REPTree base classifiers with default options, but with 100 iterations instead of the default 10. This classifier was used to generate the final coding score of each sRNA ORF, mock ORF (for the calculation of the background distribution), and intergenic ORF. AUC values and ROC curves on training data are for 10-fold cross-validated evaluation. When confidence limits are stated, they represent  $\pm 1.96$  times the standard deviation of the negative control sets, i.e., a 95% confidence interval.

Plots in Fig. 3a were generated using Weka's BoundaryVisualizer class using the training data with parameters  $r = 2$  and  $k = 5$ .

### Number of sRNA ORFs coding above background

The coding score represents an estimate of the probability that a given ORF is protein-coding, given the assumptions that the training sets accurately represent the parameter distributions for ORFs and noncoding sequences, and that there is an equal number of coding and noncoding ORFs being tested. The latter assumption is problematic, because when we test all sRNA ORFs, far fewer than half may be coding, meaning the coding score is not an accurate probability estimate.

Here, we used the coding score to calculate an empirical false discovery rate estimate. For each species, all in-frame overlapping sRNA ORFs were merged and assigned the largest coding score of the individual ORFs. Let  $S$  and  $M$  be the number of sRNA ORFs and mock sRNA ORFs, and  $S_T$  and  $M_T$  be an estimate of the number of ORFs having coding score greater than a threshold  $T$ . With the assumption that the distribution of coding scores for noncoding sRNA ORFs is the same as for mock ORFs, we calculated the empirical false discovery rate (FDR) as

$$FDR_T = \frac{\text{Expected false positives at } T}{\text{All positives at } T} = \frac{M_T/M}{S_T/S}$$

and the  $q$ -value, the FDR analog of the  $p$ -value, was by definition

$$q_T = \min_{T' \geq T} FDR_{T'}$$

We then estimated the raw number of true coding ORFs,  $C$ , by

$$\begin{aligned} C &= \max_T (\text{All positives} - \text{Expected false positives}) \\ &= \max_T (S_T - q_T S_T) \end{aligned}$$

This maximization over all thresholds leaves open the possibility that random fluctuations due to limited sample sizes yield an erroneous signal. This is because any random set of noncoding ORFs will have a transient positive difference  $S_T/S > M_T/M$  at some threshold  $T$ . Therefore, we made two modifications to this estimate. First we reasoned that the true coding ORFs would be found preferentially at the highest coding scores, so we did not

consider any thresholds below the first point at which the coding signal above background was negative, i.e.:

$$T' = \max(t) \text{ such that } S_t - \frac{M_t}{M} \times S \leq -1$$

$$C = \max_{T > T'}(S_T - q_T S_T)$$

The second modification was to select twenty random subsets of mock ORFs matching the number of true sRNA ORFs to estimate the effect of this random fluctuation, correcting the estimate of coding sRNA ORFs by

$$C' = C - \frac{\sum_{i=1}^{20} M_T^i - q_T^i M_T^i}{20}$$

where  $M^i$  is subset  $i$  of the mock ORFs and  $q_T^i$  is the calculated as above considering the remainder of the mock ORFs as the false positive distribution. This corrected  $C'$  is plotted in Fig. 3d, with the confidence intervals estimated from the standard deviation of the  $M^i$  mock ORF subsets.

In species having sRNA ORFs overlapping annotated coding ORFs, the FDR calculation was performed separately for ORFs with no annotated overlap, ORFs with out-of-frame overlap in the sense direction, and ORFs with antisense overlap, yielding three  $C'$  values. The confidence intervals on sum of the three are based on the standard deviation of the sum of the three  $M^i$  mock ORF subsets, one for each category of sRNA ORF.

To estimate the number of sRNAs with at least one coding ORF, we randomly sampled sets of ORFs to be coding 100 times, with each set equal in size to the estimate for the total number of coding sRNA ORFs. This was done separately for ORFs with overlap to annotated coding ORFs in sense and in antisense orientations, and the union of the three sets of coding ORFs were counted.

### Intergenic ORFs

When the analysis was expanded to all intergenic ORFs rather than limited to annotated sRNAs, two additional controls were added to eliminate false positives from pseudogenes. First, all ORFs having a  $D_n/D_s > 1$  with a log likelihood ratio of at least 3, i.e. those having signal for positive selection, were removed as they are likely out-of-frame ORFs in unannotated pseudogenes. Also, all ORFs having a BLAST hit larger than 100 amino acids in length were also eliminated, as they are likely in-frame degenerations of full-length protein.

### Ribosome profiling data

Processed ribosome profiling data from [25] was downloaded in processed WIG format from the NCBI Gene Expression Omnibus (Accession GSE35641). Analysis shown is for samples GSM872393 and GSM872397, although similar results were obtained from the other replicates. We began by visualizing the rates of reads mapping in all sRNA ORFs and mock ORFs. We found that a

large fraction of ORFs had no reads mapping at all, and when reads were mapped they preferentially accumulated at the start and stop codons. Therefore, a normalization by ORF length would disadvantage longer ORFs. Instead, we counted ORFs as being ribosome-bound if they had ribosome profiling signal in the correct strand within 3 codons of the start or stop codon, the range reported by Li et al. [25]. ORFs were compared to 20 sets of mock ORFs chosen in the same way as the negative controls for the nucleotide composition analysis. ORFs with out-of-frame overlap to annotated ORFs in the same sense were excluded, because the ribosome profiling data did not have sub-codon resolution.

### Proteogenomic analysis of *H. pylori*

Parts of the data and methods were previously published [49], but for convenience we include a complete summary here:

#### Sample preparation

For validation of predicted novel proteins of *H. pylori* strain 26695 we reanalyzed our in-depth proteome analysis which was especially focused on low molecular weight proteins [Mueller et al., submitted]. Briefly, *H. pylori* strain 26695 was cultured in Ham's F12 medium (without arginine, Biosera, UK) supplemented with either "light" (12C6, 14N4), "heavy" (13C6, 15N4) or "medium" (13C6, 14N4) isotopically labeled arginine (Cambridge Isotope Laboratories, USA) and 5% (v/v) dialyzed fetal calf serum (FCS) (Thermo Scientific, USA). Cells were harvested and the extracted proteins of the heavy (repG deletion mutant, spiral morphology ([50]), medium (wild type *H. pylori* strain 26695, spiral morphology) and light arginine (wild type *H. pylori* strain 26695, coccoid morphology) labeled samples were mixed 1:1:1 (w/w). The protein mixtures were either separated by 1D-SDS-PAGE or, for enrichment of low molecular weight, by gel-eluted liquid fraction entrapment electrophoresis (GELFREE) (5 fractions collecting between 0 and 50 kDa. Proteins were reduced and alkylated and 50% of each sample was digested by endoproteases trypsin and 50% by AspN. Samples were reconstituted with 0.1% (v/v) formic acid for LC-MS/MS analysis.

#### Nano-uHPLC/nano-ESI analysis

Briefly, proteolytic peptide mixtures were separated on a nano-uHPLC system (nanoAcquity, Waters, Milford, MA, USA) coupled online with an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA). Peptides were trapped and washed for 5 min with 2% acetonitrile containing 0.1% formic acid. Peptide separation was performed using a gradient of 94 min (SDS-PAGE fractions) or 154 min (off-gel fractions) ramping from 2 to 40% acetonitrile, 0.1% formic acid on a C18

column (nanoAcquity UPLC column, C18, 75  $\mu\text{m} \times 150$  mm, 1.7  $\mu\text{m}$ , Waters) with a flow rate of 300 nl/min. The mass spectrometer automatically switched between full scan MS mode ( $m/z$  300 – 1600,  $R = 60,000$ ) and tandem MS acquisition. Peptide ions exceeding an intensity of 2000 counts were fragmented within the linear ion trap by collision induced fragmentation. Dynamic precursor exclusion for MS/MS measurements was set to 2 min.

#### Data Analysis

Proteome Discoverer (version 1.4.1.14, Thermo Scientific., Bremen, Germany) was utilized for peptide identification. The database search engines MS Amanda [51] and Sequest [52] were applied for peptide and protein identification using a concatenated database containing all proteins of *H. pylori* strain 26695 from NCBI (1596 entries) as well as 89 predicted proteins from an sRNA short ORF with coding  $p > 0.5$ . The search was conducted allowing a precursor mass tolerance of 10 ppm and a fragment mass tolerance of 0.5 Da. Up to two proteolytic missed cleavages were allowed. Carbamidomethylation of cysteine was defined as fixed modification, whereas oxidation of methionine was set as variable modification for both proteases. AspN specificity was defined to cleave at the N-terminal side of aspartic acid and glutamic acid. A FDR of 5% was applied for peptide and 1% was tolerated for protein identifications. Minimum one unique peptide was required for protein identifications.

#### Confirmation using synthetic peptides

Thirty three proteolytic peptides of the predicted proteins were synthesized (Thermo Scientific., Bremen, Germany), measured by nano-uHPLC/nano-ESI MS/MS as been described above using a 94 min gradient, and were manually compared to spectra obtained by the analyses of the *H. pylori* samples. Finally, only proteins being identified with a protein FDR below 1% with at least one unique peptide which could be confirmed by a synthetic peptide were reported as novel proteins.

#### Functional annotations

*B. subtilis* expression data was taken from table S5 of [26]. ORFs in sRNAs or size-matched mock ORFs from intergenic regions were matched to array features if they had any overlap with the annotated feature coordinates. Array features were counted as coding if they contained any sRNA ORF with coding  $q \leq 0.25$ . Predicted coding ORFs were then matched to conditions of high expression as annotated by [26] through their array feature. Number of coding ORFs with high expression in a condition expected by chance was calculated based on the mean and standard deviation of 20 sets of mock ORFs or noncoding ORFs equal in size to the predicted coding ORFs. Categories were considered significantly enriched if there

was 1.5-fold more predicted coding ORFs than expected based on either control set, with a minimum of 5 ORFs. 95% confidence intervals were calculated using the normal distribution.

Transmembrane domains were predicted by TMHMM [53]. Two negative control sets were used for comparison: first, sRNA ORFs with coding  $q \leq 0.25$  were shuffled 20 times for one control set; second, mock ORFs were created by selecting random fragments of annotated ORFs matching the sRNA ORFs with coding  $q \leq 0.25$  in length.

Type I toxins were taken from BLAST hits in [39]. Toxin amino acid sequences were filtered for only those 10-50 amino acids in length and clustered using CD-HIT v4.6.1 [54], to select representative sequences not more than 90% similar to each other, yielding 114 positive training examples. Negative training examples were sampled from Uniref50 filtered to remove fragments and proteins containing nonstandard amino acids, with length distribution matching the positive control set exactly. 20 negative training examples were selected for each positive example. For each amino acid sequence, physicochemical properties were predicted using various methods: In vitro aggregation,  $\alpha$ -helix,  $\beta$ -strand, and  $\beta$ -turn conformation, and helical aggregation propensity were calculated using TANGO version 2.2 [55] with the parameters  $ct = N$   $nt = N$   $ph = 7$   $te = 298$   $io = 0.1$ . Isoelectric point and hydrophobicity were calculated using Bio::Tools::pI Calculator and Bio::Tools::SeqStats from Bioperl 1.006923. A random forest classifier was trained on this data using Weka 3-7-9 with parameters -I 10 -K 0 -S 1 -num-slots 1.

Data from [40] were downloaded from the sequence read archive (SRA, accessions SRR064320, SRR064321, and SRR064325). Short RNA reads were trimmed to remove 3' adapter sequences as in [40], and mapped to the *S. aureus* NCTC 8325 or *B. subtilis* 168 genomes using Bowtie 1.1.0 using the parameters  $-v 2 -m 1 -seed 2 -strata -best$ . Alignments less than 10 bases long were discarded, and alignments less than 20 bases long had at most one allowed mismatch. Coverage of each annotated ORF was calculated using coverageBed from bedtools 2.19.1 [56]. Coverage was normalized to RPKM using the length of each annotated ORF and the total number of mapping reads.

#### Additional files

**Additional file 1:** Supplemental figures. (PDF 1870 kb)

**Additional file 2:** Machine learning validation on annotated ORFs. (TXT 8 kb)

**Additional file 3:** Mass spectrometry results. (XLSX 21 kb)

**Additional file 4:** BLAST hits for sRNA ORFs. (TXT 29 kb)

**Additional file 5:** ORF prediction details. (XLSX 1020 kb)

**Additional file 6:** mock ORF sequences. (ZIP 3000 kb)

## Abbreviations

FDR: False discovery rate; ORF: Open Reading Frame; sRNA: Small RNA (ribonucleic acid); SD score: Shine-Dalgarno score

## Acknowledgements

The authors wish to thank Dr. Jeffrey Mellin for inspiration of the project and useful input into the method. Olivia Doppelt-Azeroual is grateful to Bertrand Néron for his technical help on the couchdb database used for the web interface and Jean-Baptiste Denis for his help setting up the virtual machine which hosts <http://disco-bac.web.pasteur.fr>.

## Funding

RCF was funded by a Pasteur Foundation Postdoctoral Fellowship and ERC grant ERC-2009-AG-232798-HOMEOPITH. MvB acknowledges funding from DFG Priority programme 2002. MC acknowledges funding from DFG CRC Aquadiva. None of the funding bodies had a role in study design, analysis, interpretation of data, or in writing the manuscript.

## Availability of data and materials

The sRNA ORF calculations and predictions generated during the current study are available on our webserver at <http://disco-bac.web.pasteur.fr>. The mock sequences are available as Additional file 6.

## Authors' contributions

RF and BS conceived the project. RF performed all computational analysis and wrote the manuscript. SK, SM, MC, and MB performed mass spectrometry experiments and analysis. ODA wrote the web interface for sharing the predictions and other data. All authors have read and approved the manuscript for publication.

## Ethics approval and consent to participate

All databases used are freely accessible and did not require special permission to access.

## Consent for publication

Not applicable.

## Competing interests

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.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Author details

<sup>1</sup>Systems Biology Laboratory, Department of Genomes and Genetics, Institut Pasteur, Paris, France. <sup>2</sup>Molecular Microbial Pathogenesis Unit, Department of Cell Biology and Infection, Institut Pasteur, Paris, France. <sup>3</sup>Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France. <sup>4</sup>Bioinformatics and Biostatistics Hub, C3BI, USR 3756 IP CNRS, Institut Pasteur, Paris, France. <sup>5</sup>Department of Molecular Systems Biology, Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany. <sup>6</sup>Current Address: Department of Bioanalytics, University of Applied Sciences and Arts of Coburg, Coburg, Germany. <sup>7</sup>Institute of Biochemistry, University of Leipzig, Leipzig, Germany. <sup>8</sup>Current Address: Neuroproteomics, German Center for Neurodegenerative Diseases (DZNE), Munich, Germany.

Received: 5 November 2016 Accepted: 9 July 2017

Published online: 21 July 2017

## References

- Waters LS, Storz G. Regulatory RNAs in Bacteria. *Cell*. 2009;136(4):615–28. doi:10.1016/j.cell.2009.01.043.
- Irnov I, Sharma CM, Vogel J, Winkler WC. Identification of regulatory RNAs in *Bacillus subtilis*. *Nucleic Acids Res*. 2010;38(19):6637–51. doi:10.1093/nar/gkq454.
- Li L, Huang D, Cheung MK, Nong W, Huang Q, Kwan HS. BSRD: a repository for bacterial small regulatory RNA. *Nucleic Acids Res*. 2012;41(D1):233–8. doi:10.1093/nar/gks1264.
- Wade JT, Grainger DC. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat Rev Micro*. 2014;12(9):647–53. doi:10.1038/nrmicro3316.
- Sharma CM, Hoffmann S, Darfeuille F, Reigier J, Findeiß S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*. 2010;464(7286):250–5. doi:10.1038/nature08756.
- Rasmussen S, Nielsen HB, Jarmer H. The transcriptionally active regions in the genome of *Bacillus subtilis*. *Mol Microbiol*. 2009;73(6):1043–57. doi:10.1111/j.1365-2958.2009.06830.x.
- Storz G, Vogel J, Wassarman K. Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell*. 2011;43(6):880–91. doi:10.1016/j.molcel.2011.08.022.
- Warren AS, Archuleta J, Feng W-c, Setubal JC. Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics*. 2010;11(1):131. doi:10.1186/1471-2105-11-131.
- Wadler CS, Vanderpool CK. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc Natl Acad Sci*. 2007;104(51):20454–9. doi:10.1073/pnas.0708102104.
- Williams REO, Harper GJ. Staphylococcal haemolysins on sheep-blood agar with evidence for a fourth haemolysin. *J Pathol Bacteriol*. 1947;59(1-2):69–78. doi:10.1002/path.1700590109.
- Gimpel M, Heidrich N, Mäder U, Krügel H, Brantl S. A dual-function sRNA from *B. subtilis*: SR1 acts as a peptide encoding mRNA on the gapA operon. *Mol Microbiol*. 2010;76(4):990–1009. doi:10.1111/j.1365-2958.2010.07158.x.
- Sonnleitner E, Sorger-Domenigg T, Madej MJ, Findeiss S, Hackermüller J, Huttenhofer A, Stadler PF, Blasi U, Moll I. Detection of small RNAs in *Pseudomonas aeruginosa* by RNomics and structure-based bioinformatic tools. *Microbiology*. 2008;154(10):3175–87. doi:10.1099/mic.0.2008/019703-0.
- Zuber P. A peptide profile of the *Bacillus subtilis* genome. *Peptides*. 2001;22(10):1555–77. doi:10.1016/s0196-9781(01)00492-2.
- Hobbs EC, Fontaine F, Yin X, Storz G. An expanding universe of small proteins. *Curr Opin Microbiol*. 2011;14(2):167–73. doi:10.1016/j.mib.2011.01.007.
- Samayoa J, Yildiz FH, Karplus K. Identification of prokaryotic small proteins using a comparative genomic approach. *Bioinformatics*. 2011;27(13):1765–71. doi:10.1093/bioinformatics/btr275.
- Garbis S, Lubec G, Fountoulakis M. Limitations of current proteomics technologies. *J Chromatogr A*. 2005;1077(1):1–18. doi:10.1016/j.chroma.2005.04.059.
- Hemm MR, Paul BJ, Miranda-Rios J, Zhang A, Soltanzad N, Storz G. Small stress response proteins in *Escherichia coli*: proteins missed by classical proteomic studies. *J Bacteriol*. 2010;192(1):46–58. doi:10.1128/jb.00872-09.
- Tinoco AD, Saghatelian A. Investigating endogenous peptides and peptidases using peptidomics. *Biochemistry*. 2011;50(35):7447–61. doi:10.1021/bi200417k.
- Müller SA, Findeiß S, Pernitzsch SR, Wissenbach DK, Stadler PF, Hofacker IL, von Bergen M, Kalkhof S. Identification of new protein coding sequences and signal peptidase cleavage sites of *Helicobacter pylori* strain 26695 by proteogenomics. *J Proteome*. 2013;86:27–42. doi:10.1016/j.jprot.2013.04.036.
- Washietl S, Findeiss S, Müller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N. RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*. 2011;17(4):578–94. doi:10.1261/rna.2536111.
- Ibrahim M, Nicolas P, Bessieres P, Bolotin A, Monnet V, Gardan R. A genome-wide survey of short coding sequences in streptococci. *Microbiology*. 2007;153(11):3631–44. doi:10.1099/mic.0.2007/006205-0.
- Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol*. 2008;70(6):1487–501. doi:10.1111/j.1365-2958.2008.06495.x.
- Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*. 2011;27(13):275–82. doi:10.1093/bioinformatics/btr209.
- Washietl S, Findeiss S, Müller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N. RNAcode: Robust discrimination of coding and

- noncoding regions in comparative sequence data. *RNA*. 2011;17(4):578–94. doi:10.1261/ma.2536111.
25. Li GW, Oh E, Weissman JS. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*. 2012;484(7395):538–41. doi:10.1038/nature10965.
  26. Nicolas P, Mader U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S, Becher D, Bisicchia P, Botella E, Delumeau O, Doherty G, Denham EL, Fogg MJ, Fromion V, Goelzer A, Hansen A, Hartig E, Harwood CR, Homuth G, Jarmer H, Jules M, Klipp E, Chat LL, Lecointe F, Lewis P, Liebermeister W, March A, Mars RAT, Nannapaneni P, Noone D, Pohl S, Rinn B, Rugheimer F, Sappa PK, Samson F, Schaffer M, Schwikowski B, Steil L, Stulke J, Wiegert T, Devine KM, Wilkinson AJ, van Dijl JM, Hecker M, Volker U, Bessieres P, Noirot P. Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science*. 2012;335(6072):1103–6. doi:10.1126/science.1206848.
  27. Mäder U, Nicolas P, Depke M, Pané-Farré J, Debarbouille M, van der Kooi-Pol MM, Guérin C, Dérozier S, Hiron A, Jarmer H, Leduc A, Michalik S, Reilman E, Schaffer M, Schmidt F, Bessières P, Noirot P, Hecker M, Msadek T, Völker U, van Dijl JM. *Staphylococcus aureus* transcriptome architecture: From laboratory to infection-mimicking conditions. *PLOS Genet*. 2016;12(4):1–32. doi:10.1371/journal.pgen.1005962.
  28. Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science*. 2011;335(6068):552–7. doi:10.1126/science.1215110.
  29. Wang G, Li X, Wang Z. APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Res*. 2009;37(Database):933–7. doi:10.1093/nar/gkn823.
  30. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*. 2007;23(6):673–9. doi:10.1093/bioinformatics/btm009.
  31. Ma J, Campbell A, Karlin S. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol*. 2002;184(20):5733–45. doi:10.1128/JB.184.20.5733-5745.2002.
  32. Gao F, Zhang CT. Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinformatics*. 2004;20(5):673–81. doi:10.1093/bioinformatics/btg467.
  33. Carels N, Frias D. Classifying Coding DNA with Nucleotide Statistics. *Bioinforma Biol Insights*. 2009;3:141–54. doi:10.4137/BBI.S3030, <http://insights.sagepub.com/classifying-coding-dna-with-nucleotide-statistics-article-a1718>.
  34. Weissenmayer BA, Prendergast JGD, Lohan AJ, Loftus BJ. Sequencing Illustrates the Transcriptional Response of *Legionella pneumophila* during Infection and Identifies Seventy Novel Small Non-Coding RNAs. *PLoS ONE*. 2011;6(3):17570. doi:10.1371/journal.pone.0017570.
  35. Rasmussen S, Nielsen HB, Jarmer H. The transcriptionally active regions in the genome of *Bacillus subtilis*. *Mol Microbiol*. 2009;73(6):1043–57.
  36. Wang R, Braughton KR, Kretschmer D, Bach T-HL, Queck SY, Li M, Kennedy AD, Dorward DW, Klebanoff SJ, Peschel A, et al. Identification of novel cytolytic peptides as key virulence determinants for community-associated mrsa. *Nature*. 2007;200:7.
  37. Mann B, van Opijnen T, Wang J, Obert C, Wang YD, Carter R, McGoldrick DJ, Ridout G, Camilli A, Tuomanen EI, Rosch JW. Control of Virulence by Small RNAs in *Streptococcus pneumoniae*. *PLoS Pathogens*. 2012;8(7):1002788. doi:10.1371/journal.ppat.1002788.
  38. Torrent M, Andreu D, Nogués VM, Boix E. Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PLoS ONE*. 2011;6(2):16968. doi:10.1371/journal.pone.0016968.
  39. Fozo EM, Makarova KS, Shabalina SA, Yutin N, Koonin EV, Storz G. Abundance of type I toxin-antitoxin systems in bacteria: searches for new candidates and discovery of novel families. *Nucleic Acids Res*. 2010;38(11):3743–759. doi:10.1093/nar/gkq054.
  40. Lasa I, Toledo-Arana A, Dobin A, Villanueva M, de los Mozos IR, Vergara-Irigaray M, Segura V, Fagegaltier D, Penades JR, Valle J, Solano C, Gingeras TR. Genome-wide antisense transcription drives mRNA processing in bacteria. *Proc Natl Acad Sci*. 2011;108(50):20172–7. doi:10.1073/pnas.1113521108.
  41. Down TA, Piiipari M, Hubbard TJP. Dalliace: interactive genome viewing on the web. *Bioinformatics*. 2011;27(6):889–90. doi:10.1093/bioinformatics/btr020.
  42. Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*. 2008;19(1):92–105. doi:10.1101/gr.082701.108.
  43. Wang X, Wood TK. Toxin-Antitoxin Systems Influence Biofilm and Persister Cell Formation and the General Stress Response. *Appl Environ Microbiol*. 2011;77(16):5577–83. doi:10.1128/aem.05068-11.
  44. Markham NR, Zuker M. UNAFold In: Keith JM, editor. *Bioinformatics: Structure, Function and Applications*. Totowa: Humana Press; 2008. p. 3–31. doi:10.1007/978-1-60327-429-6\_1.
  45. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software. *ACM SIGKDD Explor News*. 2009;11(1):10. doi:10.1145/1656274.1656278.
  46. Sievers F, Higgins DG. Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences In: Russell DJ, editor. *Multiple Sequence Alignment Methods*. Totowa: Humana Press; 2014. p. 105–16. doi:10.1007/978-1-62703-646-7\_6.
  47. Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*. 1989;5(2):163–6. doi:10.1111/j.1096-0031.1989.tb00562.x.
  48. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24(8):1586–91. doi:10.1093/molbev/msm088.
  49. Müller SA, Pernitzsch SR, Haange SB, Uetz P, von Bergen M, Sharma CM, Kalkhof S. Stable isotope labeling by amino acids in cell culture based proteomics reveals differences in protein abundances between spiral and coccoid forms of the gastric pathogen *Helicobacter pylori*. *J Proteome*. 2015;126:34–45. doi:10.1016/j.jprot.2015.05.011.
  50. Pernitzsch S, Tirier S, Beier D, Sharma C. A variable homopolymeric G-repeat defines small RNA-mediated posttranscriptional regulation of a chemotaxis receptor in *Helicobacter pylori*. *Proc Natl Acad Sci U S A*. 2014;111:501–10.
  51. Dorfer V, Pichler P, Stranzl T, Stadlmann J, Taus T, Winkler S, Mechtler K. MS Amanda a universal identification algorithm optimized for high accuracy tandem mass spectra. *J Proteome Res*. 2014;13(8):3679–84. doi:10.1021/pr500202e.
  52. Washburn MP. The H-Index of 'An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database'. *J Am Soc Mass Spectrom*. 2015. doi:10.1007/s13361-015-1181-3.
  53. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol*. 2001;305(3):567–80. doi:10.1006/jmbi.2000.4315.
  54. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28:3150–2.
  55. Fernandez-Escamilla A, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol*. 2004;22:1302–6.
  56. Quinlan A, Hall I. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

