BATTER: Accurate Prediction of Rho-dependent and Rho-independent Transcription Terminators in Metagenomes

Yunfan Jin[1,2], Hongli Ma[1,2], Zhenjiang Zech Xu[3,4,5*], Zhi John Lu[1,2*]

1 MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China

2 Institute for Precision Medicine, Tsinghua University, Beijing 100084, China.

3 State Key Laboratory of Food Science and Technology, Nanchang University, Nanchang 330047, China

4 Shenzhen Stomatology Hospital (Pingshan), Southern Medical University, Shenzhen 518118, China

5 Microbiome Medicine Center, Department of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong 510280, China

*To whom correspondence should be addressed: Zhi John Lu Tel: +86 10 62789217; Email: zhilu@tsinghua.edu.cn Correspondence may also be addressed to Zhenjiang Zech Xu. Tel: 0791-88304447-8323; Email: zhenjiang.xu@gmail.com

Abstract

Transcription terminators mark the 3' ends of both coding and noncoding transcripts in bacteria and play crucial roles in gene regulations (such as controlling the stoichiometry of gene expression and conditionally switching off gene expression by inducing premature termination). Recently developed experimental 3' end mapping techniques greatly improved the current understanding of bacteria transcription termination, but these methods cannot detect transcripts that are unexpressed

in the limited experimental conditions and cannot utilize the vast amount of information embedded in the rapidly growing metagenome data. Computational approaches can relieve these problems, but the development of such *in-silico* methods lags behind the experimental techniques. Previous computational tools are limited to predicting rho-independent terminators (RITs) and are primarily optimized for a few model species. The prediction of rho-dependent terminators (RDTs) which lack obvious consensus sequence patterns, and terminators in diverse non-model bacteria species still presents significant challenges.

To address these challenges, we introduce BATTER (**BA**cteria **T**ranscript **T**hree prime **E**nd **R**ecognizer), a computational tool for predicting both RITs and RDTs in diverse bacteria species that allows metagenome-scale scanning. We developed a data augmentation pipeline by leveraging available high throughput 3' end mapping data in 17 bacteria species, and a large collection of 42,905 species-level representative bacteria genomes. Taking advantage context sensitive natural language processing techniques, we trained a BERT-CRF model, using both local features and context information for tagging terminators in genomic sequences.

Systematic evaluations demonstrated our model's superiority: at a false positive rate of 0.1/kilobase, BATTER achieves a sensitivity of 0.924 for predicting *E. coli* RDTs; and a sensitivity of 0.756 for predicting terminators on term-seq dataset of oral microbiome, outperforming the best existing tool by 0.153. Based on BATTER's predictions, we systematically analyzed the clade-specific properties of bacteria terminators. The practical utility of BATTER was exemplified through two case studies: identifying functional transcripts from metatranscriptome data and discovering candidate noncoding RNAs related to antimicrobial resistance. As far as we know, BATTER is the first tool simultaneously predicting RITs and RDTs in diverse bacteria species.

BATTER is available at https://github.com/lulab/BATTER.

Keywords: bacteria terminator, transcription unit prediction, bacteria noncoding RNA prediction, BERT, natural language processing

Background

Bacteria transcription terminators are noncoding RNA elements that dissociate RNA polymerase complexes and induce transcription termination. They are traditionally classified into two categories: rho-independent terminators and rho-dependent terminators[1, 2]. Rho-independent terminators (RITs) induce transcription termination independent of protein factors, while the function of rho-dependent terminators (RDTs) depends on rho protein. RDTs were reported to account for 20%-30% of transcription termination events in *E. coli*, and are frequently associated with mildly expressed genes[3]. RITs are typically GC-rich RNA hairpins followed by a stretch of uracils, while RDTs are RNA sequences rich in cytosines, depleted of guanines and secondary structures, and are believed to lack compact, discernible sequence patterns, and challenging for computational prediction.

Bacteria transcription terminators, including both RITs and RDTs, play important regulatory roles. The strengths of many terminators are tunable, allow finetuning of gene expression, and maintaining specific protein stoichiometry[4, 5]. The cis-regulatory attenuator sequences, such as metabolite-binding riboswitches[6] and ribosomal protein leaders[7], encompass regulatory terminators. The attenuator sequences adapt switchable RNA secondary structures to conditionally regulate gene expression. Besides, as transcription terminators define the boundaries of bacteria transcripts and signify regulatory noncoding RNAs, they have been utilized as important features to annotate bacteria transcription architectures and identify novel regulatory noncoding RNAs[8-10]. Given the

crucial roles of bacteria terminators in gene regulation and their practical utility in genome annotation, there is significant interest in developing experimental and computational tools to identify RITs and RDTs from genomes.

Recently, several high-throughput experimental techniques, such as term-seq[11], Rend-Seq[4], and SEnd-seq[12], were developed. These techniques could detect 3' ends of RNAs, regardless of whether the transcripts employ RITs or RDTs for transcription termination. These sequencing-based methods have been successfully applied to bacteria isolates and microbial communities[11]. 3' end mapping data suggests the traditional RITs/RDTs dichotomy seems to be over-simplified. It was reported that transcript ends of many experimentally validated RDTs can fold into hairpin structures reminiscent of that in RITs with comparable folding energy but typically lack U-rich tracts[3]. A recent study also suggested that many terminators with canonical RIT patterns in *B. subtilis* are strongly regulated by protein factors[13]. High-throughput 3' mapping techniques have made great progress in understanding bacteria transcriptome architecture, but several limitations remain to be addressed. First, many bacteria transcripts, especially noncoding RNAs, are only expressed in specific conditions[14] or have low abundance, and 3' ends of these transcripts are elusive for experimental detection. Second, although 3' end mapping techniques can be applied to microbial communities in environmental samples, as metatranscriptome reads are often dominated by few species, identifying transcript ends of less abundant species from sequencing data becomes challenging. Third, experimental methods cannot leverage the vast amount of information embedded in the rapidly accumulating metagenome data.

For the reasons above, computational methods are imperative to supplement the experimental methods for bacteria terminator prediction. Several computational tools have been developed in the

last two decades. However, almost all of them focused on RITs only because RDTs are believed to be poorly conserved. In addition, they are only optimized for a few model bacteria species like *E. coli and B. subtilis*. These methods can be roughly divided into two categories: RNA folding-based methods and motif searching-based methods. RNA folding-based methods, such as transterm-HP[15], search for U-rich sequences, locally fold these candidates, and use folding energy for scoring. Assuming most RITs have relatively small variations across different genes and species, motif search-based tools utilized established RNA descriptors, such as the profile stochastic context-free grammar[16] (pSCFG), RNAmotif [17], and ERPIN[18], to fit manually curated known RIT stem-loops. RNIE[19] and ARNold[20] both use this approach. Despite their importance, the development of computational methods lags behind experimental techniques, and there is still room for improvement. Both RNA folding-based methods and motif searching-based methods cannot predict RDTs, and the relatively simple scoring schema may not generalize well to diverse non-model species. Almost all available terminator prediction tools were developed before the development of 3'end mapping techniques, and the data generated by experimental methods is not utilized for both model training and performance evaluation. Moreover, the function of terminators is highly relevant to its genomic context, but most available methods assume terminators are local motifs embedded in random sequences ignoring the context information, which could be improved.

Modeling context information is central to many natural language processing (NLP) tasks, and advances in NLP offer a unique opportunity for improving terminator prediction. Recently, given its superior representation power, BERT (Bidirectional Encoder Representations from Transformers) has emerged as a popular model. Terminator prediction is a token classification task in NLP (classify each nucleotide as either terminator or background) similar to protein secondary structure

prediction[21] and signal peptide prediction[22], which have been effectively handled by stacking a conditional random field (CRF) layer on the top of a neural encoder. Hence, BERT-CRF is a promising tool for modeling the local and context features of transcription terminators in diverse bacteria species.

Exploiting advancement in both experimental data and computational techniques, we present BATTER, a machine-learning model for accurately predicting transcription terminators. Leveraging published 3' end mapping data and a large bacteria genome collection[23], we developed a data augmentation pipeline and generated ~2.5 million transcription terminator instances for model training. Taking advantage of recent progress in natural language processing, we trained BATTER, a BERT-CRF model, which uses both local features and context information for terminator prediction. To our knowledge, it is the first computational tool for the simultaneous prediction of RITs and RDTs in bacteria species of diverse taxa, allowing metagenome-wide scanning. Systematic evaluation on diverse metrics suggested BATTER achieves unprecedented accuracy. We further showcased BATTER's practical utility on two tasks: identifying transcript ends in metatranscriptome, and predicting noncoding RNAs (ncRNAs) with potential regulatory roles in the expression of antimicrobial resistance (AMR) genes.

Results

Common local patterns of RITs and RDTs and their covariation with sequence context

As we've already mentioned, previous term-seq study suggests that both RITs and RDTs of *E. coli* are often composed of stem-loop structure [3], while the conventional computational models assume that RDTs are depleted of RNA secondary structure[1, 2]. A possible explanation for this apparent

133 discrepancy is that, for transcripts with RDTs, 3' ends are rapidly degraded by intracellular
134 exoribonuclease after transcription termination, until reaching stem-loop structures which prevent
135 further degradation; therefore, the transcript ends detected by term-seq are not real RDT sites, but
136 protective upstream RNA structures[3, 24]. In other words, the RDT sites were usually far away from
137 the stem-loop structures, making them hard to be detected/predicted computationally.

138 The upstream stem-loop structures of RDTs are crucial for transcripts' integrity and are likely
139 positively selected in evolution. This indicates RITs and RDTs may exist on an evolutionary
140 continuum, namely, an RIT may evolve into an RDT by changing its downstream sequence while
141 preserving the stem-loop structure, and vice-versa[3].

142 To validate this hypothesis, we gathered annotated RITs and RDTs in *E. coli*[3], identified
143 homologs of their upstream genes within taxa *Enterobacteriaceae*, and searched conserved hairpins
144 near stop codons (upstream 16 nt, and downstream 128 nt) with cmfinder[25]. For homolog search,
145 we used Genomic catalog of Earth's Microbiomes (GEMs), a diverse prokaryotic genome collection,
146 which clustered metagenome-assembled genomes (MAGs) and RefSeq genomes into 45,599
147 species-level representatives, of which 42,905 are bacteria species[23]. In line with our assumption,
148 the terminators of some genes show RDT-like properties in some species, while resembling
149 canonical RIT in other species. For instance, as shown in Figure 1 A, in *E. coli*, the terminator
150 associated with the transcript coding transaldolase is annotated as RDT (experimentally determined
151 with rho inhibitor), and species within *Escherichia* genus as well as *Citrobacter*, *Leclercia*,
152 *Enterobacter*, and *Pseudescherichia*, also have RDT-like properties (lack in U tract, and elevated
153 C/G ratio in downstream sequence); while in *Serratia*, *Yersinia* and *Pectobacterium* species, the
154 transaldolase transcript is ended with a typical RIT-like motif. R-scape[26] analysis revealed strong

155 covariation of the predicted base pairing (Figure 1B), suggesting that despite the seemingly distinct
156 rho factor dependency, all of these transcripts are associated with a conserved hairpin. We stratified
157 all stem-loops by their maximum length of U runs, inspected the C/G ratio within 100 nt upstream
158 and 100 nt downstream flanking sequences, and observed a significant negative correlation (Figure
159 1 C, ANOVA P=7.61×10^{-19}), suggesting that these two features are co-evolving. Notably, species
160 in genera *Pantoea* and *Cedecea* exhibit properties resembling both RIT and RDT, evident in the
161 presence of transition states between RDTs and RITs. Such phenomena are observed in multiple
162 genes, suggesting transadolase is not a special case (Figure S1).
163 Taken together, RITs and at least a subset of RDTs are characterized by conserved stem-loops and
164 the stem-loop structures might be more conserved than factor dependency to protect 3' ends from
165 exonuclease degradation. Thus, it's possible to predict both RITs and a substantial fraction of RDTs
166 in a common framework: predicting stem-loop structures embedded in certain sequence context
167 (Figure 1 D).
168
169 ***Data collection, augmentation, and context-aware model training***
170 We collected transcription termination sites (TTSs) defined by published genome-wide studies[4, 27-
171 35] (Table S1). Naively training models on experimentally verified TTS presented several challenges.
172 First, the available dataset only contains 17 species spanning only 4 phyla with strong phylogeny
173 distribution bias (Figure S2), insufficient for enabling prediction across diverse taxa. Additionally,
174 the sample size is inadequate for training modern neural network models effectively. Second, term-
175 seq and related techniques cannot distinguish actual transcription termination sites from processed
176 transcript ends: the 3' ends can either derive from either exonuclease processing of RDTs

177 (informative), or other exonuclease processing events and endonuclease cleavage (not informative).
178 Third, TTSs identified by different studies may have unwanted variations due to different
179 experimental settings, which may lead to suboptimal performance.

180 To address these challenges, we developed a data augmentation pipeline (Figure 2A). Briefly,
181 protein products of genes supported by downstream term-seq signals were clustered to 40%
182 sequence identity, and we expected that homologous genes spanning multiple phyla within each
183 cluster are more likely associated with terminators in species without experimental data compared
184 to random protein-coding genes. For 115 gene clusters supported by term-seq signals across at least
185 3 phyla (Table S2), we searched for their homologous genes in representative bacteria genomes in
186 GEMs, and identified conserved stem-loops within the downstream sequences (upstream 16 nt, and
187 downstream 128 nt sequences flanking stop codon) of these homologs for each gene cluster
188 separately with cmfinder[25]. To consider potential clade-specific properties of terminators, and
189 reduce the computational burden, candidate sequences of each gene cluster are grouped by
190 taxonomy relationship (see Methods), to make sure that for each cmfinder run, the input contains
191 no more than 128 candidate sequences. This process produced 768 covariance models on average
192 for each gene cluster.

193 Except for transcription terminators, 3' ends of protein-coding genes may encompass other
194 conserved RNA structures; cmfinder can also produce false positives. To select confident candidates,
195 we compared all covariance models discovered from each gene cluster against the downstream
196 sequence of other gene clusters, and motifs that detected in at least 15 gene clusters were reserved.
197 395 covariance models per gene cluster on average passed the filtering. To further improve the
198 diversity of training data, the obtained motif instances were augmented by searching diverse

sequences that are known to be enriched in transcription terminators with identified covariance model, including candidate sequences that were used for terminator motif discovery, downstream sequences of a random subset of CDSs in GEMs bacteria genomes, and Rfam regulatory RNA (known bacteria sRNA and attenuator sequences, see Table S2) hits in GEMs. The consensus hits were subjected to multiple steps of filtering and refinement (see Methods), and around 2.5 million putative terminator instances associated with diverse protein-coding genes and noncoding RNA were identified (Figure S2).

The stem-loop structures are pervasive in bacteria genomes, to reduce false positives, it's crucial to model context sequences, which can be addressed by neural networks designed for natural language processing (NLP) tasks. We aim to scan both strands of bacteria DNA sequences to predict the coordinates of terminators, and each nucleotide at the given strand is assigned a label (either terminator or background). This task is referred to as token classification or named entity recognition (NER) in the NLP community. Conditional random field (CRF) is a classical solution for NER, which models the joint distribution of the sequence context and the labels to predict as an undirect graph model.

Here we used longformer[36] (a variant of BERT encoder with sliding window attention to reduce computational burden) to predict the likelihood of whether each input token belongs to terminators, followed by the post-processing of the prediction with a CRF layer (Figure 2B). The encoder part of this model was pre-trained with masked language modeling (MLM) loss using sequences sampled from GEMs' representative genomes. The classification layer and conditional random field layer were appended, and the model was optimized on augmented terminator instances (with 250 nt flanking sequences on both sides) and background instances sampled from bacteria genomes.

221

222 ***Performance evaluation and comparison with existing methods***

223 For benchmarking model performance for terminator prediction, defining true negatives on real
224 genomic sequences is non-trivial: experimental data can only detect a subset of transcription
225 terminators, and sampling random sequences from genome regions without experimentally
226 supported terminators as background instances would lead to false negatives. For this reason, we
227 used simulated data for model false positive rate (FPR) calibration. The tetra-nucleotide frequencies
228 (TNFs) of different parts of a bacteria genome are relatively stable, but vary across different bacteria
229 species, and are widely used for binning metagenomic contigs[37]. To preserve some species-specific
230 properties, for a given bacteria genome, we calculated the TNF, and sampled background sequences
231 from this distribution with a 3-order Markov model (given the last 3 nucleotides, sample the next
232 one, and generate sequences autoregressively). Predicted terminators on these sequences were
233 considered false positives. Using this approach, sensitivity at a given FPR cutoff was utilized as the
234 performance measure. The widely used terminator prediction tool, transterm-HP, uses a default
235 cutoff, corresponding to an FPR of ~0.16/KB on sequences simulated from the TNF of *E. coli*. We
236 used a slightly more stringent cutoff of 0.1/KB throughout this study. For both BATTER and
237 published methods, we've observed a single cutoff leads to considerable variations in FPR on
238 background sequences simulated from different TNF distributions (Figure S3), hence using a
239 species-specific cutoff is necessary. At a specific cutoff, lower GC content generally leads to a higher
240 false positive rate, this phenomenon is particularly noticeable in RNIE and BATTER.

241 We first evaluated the performance of BATTER on an *E. coli* terminator set with RIT/RDT
242 annotation[3] (Figure 3 A). Compared with three published methods, transterm-HP, RNIE, and

ARNold, BATTER has better performance on both RITs and RDTs. For RITs, all of the three published methods have good performance, and BATTER performs exceptionally well. For RDTs, at 0.1/KB FPR cutoff, ARNold could correctly predict less than 25% of the terminators, and performances of transterm-HP and RNIE are comparable (sensitivity of less than 0.45). BATTER outperforms the other three methods by a large margin (sensitivity of 0.924). Since ARNold is a web tool prohibitive for large-scale evaluation, and the other two tools both outperform ARNold on *E. coli* data, we considered only transterm-HP and RNIE for further performance comparisons.

The performances of BATTER and published tools were then evaluated on term-seq data of human oral microbiome[11]. Its sequences were stringently held out from model training. Term-seq reads were mapped to species with abundance > 1% (Table S3), and downstream TTSs of protein-coding genes were extracted. When FPR is controlled at 0.1/KB, BATTER consistently detects more TTSs (Figure 3B). This improvement is particularly noticeable for phyla like *Actinobacteria* and *Fusobacteria*, highlighting the advantage of our expansion of training data from limited model organisms to diverse taxa.

To further evaluate the terminator prediction systematically on the diverse GEM representative genomes that lack high throughput RNA sequencing data, here we used the approach proposed by developers of transterm-HP[15]: There should be transcription terminators in intergenic regions between convergently transcribed ("tail to tail") gene pairs. We performed genome-wide scanning on a random subset of ~2000 representative bacteria genomes in GEMs with transterm-HP, RNIE, and BATTER. At an FPR cutoff of 0.1/KB, for all the IGRs flanked by tail-to-tail gene pairs, the fraction of IGRs with predicted terminators measures model performance. Notably, BATTER consistently achieves the best performance. The improvement varies across different bacteria clades

compared to the other two methods. For instance, the improvement is mild in *Firmicutes*, but noticeable in some *Proteobacteria* species and most species in *Actinobacteria* phylum. This may be explained by the intra-species variations of the transcription terminator: transterm-HP and RNIE were optimized toward the canonical terminator model. For clades in which terminators deviate from the "GC rich stem-loop followed by consecutive Us" pattern, they would have suboptimal performance.

We further investigated the genomic distribution of predicted terminators. RNIE, transterm-HP, and BATTER take sequences alone as input without prior knowledge of CDS locations. Assuming terminators within CDS are not prevalent, a high fraction of predicted terminators inside CDS regions indicates high FPR. For all three tools, terminators predicted within CDSs make up comparable proportions (around 35%), with BATTER showing the smallest inter-species variations (Figure 3 D), indicating its superior performance. This result also confirms the reliability of our TNF-based methodology for simulating negative sequences. It's notable that although on average 90% of a bacteria genome are coding sequences, terminators predicted by all three tools are strongly enriched in intergenic regions.

It's desirable to predict the strand of terminators accurately. This feature is particularly useful for distinguishing strands in ncRNA prediction, which proved to be challenging[38]: CG and AU base pairings are symmetric in two strands, leading to comparable folding energy, and previous methods have to rely on the mild difference introduced by the asymmetric effect of GU pairs[39]. We defined a predicted terminator as "concordant" when located on the same strand as nearby CDS 3' end (within 32 nt upstream and 128 nt downstream, relative to the stop codon). "Discordant" terminators occur on the opposite strand. Presumably, the correctly predicted terminators typically share strand

287 orientation with their upstream CDSs; thus, the fraction of discordant predictions indicates the
288 accuracy of strand prediction. A smaller discordant fraction represents better performance. Based
289 on this metric, BATTER achieved the lowest discordant fraction, outperforming transterm-HP and
290 RNIE (Figure 3E).

291 As mentioned previously, transcription terminators have been utilized as a signature for
292 identifying novel bacteria noncoding RNAs. Differential RNA-seq (dRNA-seq) is an experimental
293 method for annotating bacteria transcription start sites (TSSs) and is widely used for sRNA
294 identification[40, 41]. Assuming the majority of sRNA with TSS detected by dRNA-seq have
295 terminators, predicting more terminators between intergenic TSS and the next CDS should indicate
296 better performance. We curated 13 published dRNA-seq-defined TSS sets[40, 42-49] (Table S4), and
297 identified intergenic TSSs (transcripts with > 16 nt to the closest CDS) that are followed by a trailing
298 terminator. At an FPR of 0.1/KB, a large proportion of the predictions are supported by multiple
299 methods. Notably, a considerable fraction is unique to BATTER, suggesting it is capable of
300 predicting terminators that can't be identified by transterm-HP and RNIE (Figure 3F). For each
301 dataset, three tools were ranked by the number of intergenic TSSs that were assigned with predicted
302 terminators. BATTER ranked as top 1 in 9 of the 13 datasets, and in most cases, RNIE and transterm-
303 HP ranked 2nd and 3rd, respectively (Figure 3G).

304 Providing the superior performance of BATTER, we further sought to interpret which features
305 contribute to the model's prediction. Taking the RIT-like *Yersinia* terminators and RDT-like
306 *Enterobacter* terminators that associate with transaldolase as examples, using the integrated gradient
307 algorithm[50], we evaluated the importance of each position in input sequences for terminator
308 prediction. As expected, for RIT-like terminators, the local features of terminators themselves are

309 the most important, while for RDT-like terminators, the context features, such as the presence of
310 stop codons, and nucleotide compositions of the flanking sequence become pronounced (Figure S4).
311
312 ***Inter-gene and inter-clade variations of bacteria terminators***
313 The suboptimal performance of transterm-HP and RNIE may be partially attributable to the
314 deviation of some RITs from the canonical pattern in certain bacterial taxa. To illustrate this, we
315 summarized stem lengths and lengths of trailing consecutive U-rich sequences of BATTER-
316 predicted terminators (only include concordant terminator predicted near stop codon) on scanned
317 GEMs genomes (Figure 4 A). *Actinobacteria* terminators are characterized by long stems and
318 frequently lacking in downstream U-rich sequences. In a previous study, it was noticed that *M.*
319 *tuberculosis* genome harbors stem-loops that deviate from the canonical pattern, and these stem-
320 loops are termed TRIT[19] (*Tuberculosis* Rho-independent Terminator). Here we showed such a
321 terminator motif is not restricted to *M. tuberculosis*, but is a common observation in *Actinobacteria*
322 phylum. Most *Firmicutes* and *Bacteroidota* terminators have small deviations from typical RITs
323 (this explains why transterm-HP and RNIE perform relatively well in these two phyla), and the
324 stems are generally longer in *Firmicutes*. *Proteobacteria* terminators are relatively short and have
325 relatively large intra-phylum variations. When only considering concordant terminators near stop
326 codon, in most bacteria species, 1/5 to 1/2 of CDSs are predicted to have trailing terminators, and
327 on average, 1-5 terminators are predicted in every kilobase of IGR sequences (Figure 4 A). This
328 result is within expectation since multiple prokaryotic CDSs in the same operon often share a
329 common terminator.
330 Bacteria terminators have inter-gene variations (i.e. variations between terminators of different

genes in the same species) and inter-species variation (i.e. variations between homologous genes across species). Specifically, homologous genes in different species, originating from a common ancestral sequence, may encompass terminators with similar features. On the other hand, different genes in a bacteria species share the same transcriptional machinery, therefore favoring the emergence of similar terminators across different genes due to evolutionary pressure. To quantify the two types of variations, we trained a neural network that projects RNA sequences to vector embeddings, supervised by a triplet loss that pushes RNAs belonging to the same Rfam family closer to each other in embedding space, while pushing RNAs of different families apart (see Methods and Figure S5 A for details). The reliability of the neural embedding was validated by t-SNE (t-distributed stochastic neighbor embedding) visualization on holdout Rfam families. The encoder performed well in representing both inter-RNA family variations and finer-grained within RNA family variations (Figure S5 B-C). Then we considered predicted terminators associated with several widely distributed genes (5 protein-coding gene families and 2 riboswitches) in all GEMs representative bacteria genomes and projected them into embeddings with the neural network encoder. The neural embeddings of predicted terminators were visualized with t-SNE (Figure 4 B-C). Terminators of homolog genes in closely related species are similar, but for phylogenetically distant species, covariation between terminators and the species-specific transcription machinery becomes the dominant effect, and the inter-gene variations are generally smaller than inter-species variations. We further analyzed the variance explained by taxonomies (at levels of phylum, class, and order, respectively) and gene families for each of the top 36 principal components (explained ~80% of total variance) and took average across principal components (weighted by their fractions of variance explained). As shown in Figure 4D, compared to gene families, taxonomy constantly

353 explained a larger fraction of variance. To exemplify the terminators predicted across clades, we
354 clustered the neural embeddings of terminators associated with cysteine synthase into 10 clusters
355 with K-means and visualized the predicted secondary structures of centroid instances (Figure 4E).
356
357 ***Application on metatranscriptome analysis***
358 Metatranscriptome sequencing is helpful for understanding the function of microbial communities.
359 Most of the current analysis methods for metatranscriptome data focus on protein-coding genes.
360 Noncoding transcripts may also be captured by metatranscriptome sequencing, but except for
361 constitutively expressed structural RNAs (rRNA, tRNA, RNAase P, etc.), bacteria ncRNAs are
362 rarely detected intact due to limited sequencing coverage. Additionally, pervasive transcription is
363 common in bacteria, requiring differentiation of functional ncRNAs from transcriptional noise.
364 Therefore, searching metagenomes and evaluating structural covariation remains the primary
365 method for identifying ncRNAs in uncultivated bacteria[14]. BATTER's prediction could help
366 address these challenges: terminator prediction can be utilized to refine the boundary of assembled
367 transcripts, and transcripts containing terminators are more likely to represent functional ncRNAs,
368 rather than products of pervasive transcription.
369 Here, we mapped the metatranscriptome reads of 42 samples in the Tara Ocean project (Table
370 S5)[51] to contigs assembled from metagenome data of matching samples, and assembled them into
371 transcript fragments. Terminators for IGRs containing orphan transcripts (transcripts with > 32 nt to
372 the closest CDS) were predicted with BATTER, RNIE, and transterm-HP. Each transcript fragment
373 was annotated to the closest terminator if possible. Consistent with the results on dRNA-seq datasets
374 shown in the previous section, more transcript fragments can be annotated with terminators from

375 BATTER's prediction than RNIE and transterm-HP (Figure 5A, B). Transcripts assembled from
376 different samples were pooled and clustered to 90% identity, each transcript was labeled as either
377 conserved (the corresponding cluster contains sequences from multiple samples) or singleton (the
378 cluster only contains sequences from a single sample). Given products of pervasive transcription
379 usually lack sequence conservation[52], transcripts in the conserved group are less likely to represent
380 transcriptional noise. As expected, the conserved fraction was significantly higher for transcripts
381 with terminators than those without (P=6.38×10^{-28}, Figure 5C), supporting the hypothesis that
382 terminator-containing transcripts are more likely to be functional ncRNAs.
383
384 ***Application on AMR gene-related noncoding RNA identification***
385 Antibiotic resistance is becoming a severe problem in public health, and understanding
386 antimicrobial resistance (AMR) gene regulation is becoming ever more urgent. Previous studies
387 suggested premature transcription termination is relevant to bacteria transcription regulation in
388 response to antibiotic challenges[11], terminator scanning on the AMR genes might identify novel
389 ncRNA involved in the regulation of AMR genes. Candidate AMR genes in GEMs representative
390 bacteria genomes were predicted with AMRfinder[53], and their upstream sequences were scanned
391 with BATTER. The fraction of antimicrobial resistance (AMR) genes containing predicted upstream
392 terminator structures varied substantially across different AMR gene families. Upstream terminators
393 are prevalent in ribosomal protection proteins, 23S rRNA methyltransferases, and efflux transporters
394 but were rare in other AMR families (Figure 5D).
395 For the top 15 AMR gene families that most frequently contain predicted upstream terminators,
396 we searched for conserved structures and evaluated their conservation. Sequence covariations

397 supported conserved secondary structures identified upstream of translation-related AMRs
398 (ribosomal protection proteins and 23S rRNA methyltransferase), but not those related to efflux
399 transporter (Figure 5E, Figure S6). In agreement with our findings, previous studies suggested some
400 ribosomal protection proteins are regulated by structured leader sequences via transcription
401 attenuation[54]. For instance, we identified a conserved RNA structure upstream of Lsa family ABC-
402 F proteins. Cacofold[55] predictions suggested the structure motif could adopt two conformations
403 (Figure 5F, see Methods for details). We hypothesized it may regulate translation similarly to
404 reported ABC-F leaders[54]: in physiology conditions, active translation of uORF makes the leader
405 sequence adopt a loop conformation with its terminator exposed, which leads to premature
406 termination. Upon challenges of antibiotics that act as translation inhibitors, ribosomes stall on
407 uORF, the leader sequence adapts the compact conformation with terminator hindered, and
408 transcription of downstream gene is switched on. As for efflux transporters, whose leader sequences
409 do not have obvious structure conservation, an Rfam search suggested the upstream sequence of
410 macrolide efflux MFS transporter Mef(A) hits to *Enterococcus* sRNA_1300 (RF02862), a known
411 sRNA related to antibiotic response[31]. sRNA target prediction with CopraRNA[56] further
412 confirmed its relevance to drug efflux (Table S6).
413 In summary, our analysis identified several putative AMR-related regulatory RNA, either acting
414 in *cis* (ribosome protection protein leader) or in *trans* (sRNAs that may regulate expression of efflux
415 transporters), suggesting BATTER is a useful tool for bacteria ncRNA discovery.
416
417 **Discussion**
418 In this study, we introduced BATTER, a tool for terminator prediction. We systematically

419 demonstrated its effectiveness and practical utility. As a terminator prediction tool capable of
420 predicting both RITs and RDTs optimized across diverse bacteria species, BATTER can be very
421 helpful in understanding transcriptome architecture and RNA regulation in non-model bacteria. We
422 believe that BATTER is a powerful tool for exploring the bacterial RNA world.

423 We have demonstrated multiple advantages of BATTER, including predicting RDTs, utilizing
424 context information up to 250 nt flanking the termination site, optimizing for broad bacteria
425 phylogeny, and allowing metagenome-wide scanning, but it still has some limitations:

426 First, our data augmentation pipeline is based on the assumption that RDTs have upstream stem-
427 loops that are essential to transcript integrity and are positively selected during evolution. However,
428 not all prokaryote terminators follow this paradigm. As for RDTs that generate premature
429 termination, in most cases, the prematurely terminated transcripts are not functional, and a stem-
430 loop structure may not be positively selected. In addition, archaea terminators usually do not contain
431 stem-loop structures[57]. BATTER cannot effectively predict these terminators. That being said, our
432 methodology, data augmentation pipeline, and model architecture here can be generalized to the
433 prediction of premature RDTs and archaea terminators with sufficient 3' end annotations. Second,
434 BATTER was trained on sequences of ~500 nt for context information to aid prediction. It may
435 suffer from low sensitivity on short reads or contigs from highly fragmented assemblies. This is
436 often not a problem for practical usage, since we are generally interested in the sequence context
437 (e.g. CDS or ncRNA) associated with the terminator rather than the terminator itself. Third, for
438 efficient inference, BATTER requires GPUs, which are less accessible than CPUs.

439 In this study, for data augmentation, we've focused on the commonality of terminators and
440 assumed most bacteria terminators have stem-loops. For predicted terminators, we've also analyzed

the specificity of terminators in different clades and genes. A potential caveat is, as only motifs that occurred in multiple gene clusters were used in data augmentation, gene-specific terminators are underrepresented in training data, which can lead to underestimation of between-gene variations. However, as BATTER empirically achieved the best performance on multiple metrics defined on diverse gene sets without preselection, the bias may have little impact on our conclusion.

RNA polymerase complex is present in virtually all bacteria species, despite conserved properties such as subunit composition and promoter architecture, the transcription machinery itself and the mechanism of its regulation also gain some clade-specific properties in evolution[58]. For instance, β and β' subunits of RNA polymerase contain clade-specific insertions, deleting some of these regions leads to various deficiencies. Multiple clade-specific transcription factors have also been reported to regulate bacteria transcription[58]. Assuming terminators' sequence and structure are optimized for stalling the RNA polymerase complex of the same species, inter-species variations of terminators may reflect co-evolutions between terminators and the transcription machinery.

On the other hand, the gene-specific property is also pronounced. In many scenarios, terminators play important regulatory roles in addition to stalling RNA polymerase, such as serving as Hfq binding sites in some sRNA[59] or being involved in the formation of alternative structures in some metabolite-binding riboswitches[6]. Such regulations are often gene-specific, imposing additional constraints on these genes and becoming a source of inter-genes variation of the terminators. We believe in-depth study of the evolutionary force that drives the inter-species and intra-species variations of terminators is an interesting direction for future investigation

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (32170671，8237061277), Tsinghua University Guoqiang Institute Grant (2021GQG1020), Tsinghua University Initiative Scientific Research Program of Precision Medicine (2022ZLA003), Bioinformatics Platform of National Center for Protein Sciences (Beijing) (2021-NCPSB-005). This study was also supported by Bayer Micro-funding, Beijing Advanced Innovation Center for Structural Biology, Bio-Computing Platform of Tsinghua University Branch of China National Center for Protein Sciences.

CONFLICT OF INTEREST

The authors declare no potential conflicts of interest.

References

1. Roberts, J.W., *Mechanisms of Bacterial Transcription Termination.* Journal of Molecular Biology, 2019. **431**(20): p. 4030-4039.
2. Peters, J.M., A.D. Vangeloff, and R. Landick, *Bacterial Transcription Terminators: The RNA 3'-End Chronicles.* Journal of Molecular Biology, 2011. **412**(5): p. 793-813.
3. Dar, D. and R. Sorek, *High-resolution RNA 3'-ends mapping of bacterial Rho-dependent transcripts.* Nucleic Acids Research, 2018. **46**(13): p. 6797-6805.
4. Lalanne, J.-B., et al., *Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry.* Cell, 2018. **173**(3): p. 749-761.e38.
5. Wu, S., et al., *Internal Transcription Terminators Control Stoichiometry of ABC Transporters in Cellulolytic Clostridia.* Microbiology Spectrum, 2022. **10**(2): p. e01656-21.
6. Breaker, R.R., *Riboswitches and the RNA World.* Cold Spring Harbor Perspectives in Biology, 2012. **4**(2).
7. Fu, Y., et al., *Most RNAs regulating ribosomal protein biosynthesis in Escherichia coli are narrowly distributed to Gammaproteobacteria.* Nucleic Acids Research, 2013. **41**(6): p. 3491-3503.
8. Pichon, C. and B. Felden, *Small RNA gene identification and mRNA target predictions in bacteria.* Bioinformatics, 2008. **24**(24): p. 2807-2813.
9. Pichon, C., et al., *An in silico model for identification of small RNAs in whole bacterial genomes: characterization of antisense RNAs in pathogenic Escherichia coli and Streptococcus agalactiae strains.* Nucleic Acids Research, 2012. **40**(7): p. 2846-2861.
10. Naville, M. and D. Gautheret, *Premature terminator analysis sheds light on a hidden world of bacterial transcriptional attenuation.* Genome Biology, 2010. **11**(9): p. R97.

11. Dar, D., et al., *Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria.* Science, 2016. **352**(6282): p. aad9822.
12. Ju, X., D. Li, and S. Liu, *Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria.* Nature Microbiology, 2019. **4**(11): p. 1907-1918.
13. Mandell, Z.F., et al., *Comprehensive transcription terminator atlas for Bacillus subtilis.* Nature Microbiology, 2022. **7**(11): p. 1918-1931.
14. Weinberg, Z., et al., *Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic regions.* Nucleic Acids Research, 2017. **45**(18): p. 10811-10823.
15. Kingsford, C.L., K. Ayanbule, and S.L. Salzberg, *Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake.* Genome Biology, 2007. **8**(2): p. R22.
16. Nawrocki, E.P., D.L. Kolbe, and S.R. Eddy, *Infernal 1.0: inference of RNA alignments.* Bioinformatics, 2009. **25**(10): p. 1335-1337.
17. Macke, T.J., et al., *RNAMotif, an RNA secondary structure definition and search algorithm.* Nucleic Acids Research, 2001. **29**(22): p. 4724-4735.
18. Gautheret, D. and A. Lambert, *Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles11Edited by J. Doudna.* Journal of Molecular Biology, 2001. **313**(5): p. 1003-1011.
19. Gardner, P.P., et al., *RNIE: genome-wide prediction of bacterial intrinsic terminators.* Nucleic Acids Research, 2011. **39**(14): p. 5845-5852.
20. Naville, M., et al., *ARNold: A web tool for the prediction of Rho-independent transcription terminators.* RNA Biology, 2011. **8**(1): p. 11-13.
21. Wang, S., et al., *Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields.* Scientific Reports, 2016. **6**(1): p. 18962.
22. Teufel, F., et al., *SignalP 6.0 predicts all five types of signal peptides using protein language models.* Nature Biotechnology, 2022. **40**(7): p. 1023-1025.
23. Nayfach, S., et al., *A genomic catalog of Earth's microbiomes.* Nature Biotechnology, 2021. **39**(4): p. 499-509.
24. Wang, X., et al., *Processing generates 3' ends of RNA masking transcription termination events in prokaryotes.* Proceedings of the National Academy of Sciences, 2019. **116**(10): p. 4440-4445.
25. Yao, Z., Z. Weinberg, and W.L. Ruzzo, *CMfinder—a covariance model based RNA motif finding algorithm.* Bioinformatics, 2006. **22**(4): p. 445-452.
26. Rivas, E., J. Clements, and S.R. Eddy, *A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs.* Nature Methods, 2017. **14**(1): p. 45-48.
27. Thomason Maureen, K., et al., *A rhlI 5' UTR-Derived sRNA Regulates RhlR-Dependent Quorum Sensing in Pseudomonas aeruginosa.* mBio, 2019. **10**(5): p. 10.1128/mbio.02253-19.
28. Lee, Y., et al., *The Transcription Unit Architecture of Streptomyces lividans TK24.* Frontiers in Microbiology, 2019. **10**.
29. Warrier, I., et al., *The Transcriptional landscape of Streptococcus pneumoniae TIGR4 reveals a complex operon architecture and abundant riboregulation critical for growth and virulence.* PLOS Pathogens, 2018. **14**(12): p. e1007461.
30. Vera Jessica, M., et al., *Genome-Scale Transcription-Translation Mapping Reveals Features of Zymomonas mobilis Transcription Units and Promoters.* mSystems, 2020. **5**(4): p. 10.1128/msystems.00250-20.

31. Sinel, C., et al., *Small RNAs in vancomycin-resistant Enterococcus faecium involved in daptomycin response and resistance.* Scientific Reports, 2017. **7**(1): p. 11067.
32. Hwang, S., et al., *Elucidating the Regulatory Elements for Transcription Termination and Posttranscriptional Processing in the Streptomyces clavuligerus Genome.* mSystems, 2021. **6**(3): p. 10.1128/msystems.01013-20.
33. Jeong, Y., et al., *Multi-Omic Analyses Reveal Habitat Adaptation of Marine Cyanobacterium Synechocystis sp. PCC 7338.* Frontiers in Microbiology, 2021. **12**.
34. Cho, S.-H., et al., *Different Regulatory Modes of Synechocystis sp. PCC 6803 in Response to Photosynthesis Inhibitory Conditions.* mSystems, 2021. **6**(6): p. e00943-21.
35. Forquet, R., et al., *Mapping the Complex Transcriptional Landscape of the Phytopathogenic Bacterium Dickeya dadantii.* mBio, 2022. **13**(3): p. e00524-22.
36. Beltagy, I., M.E. Peters, and A. Cohan, *Longformer The Long-Document Transformer.* arXiv [cs.CL], 2020.
37. Wu, Y.-W., et al., *MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm.* Microbiome, 2014. **2**(1): p. 26.
38. Gorodkin, J. and I.L. Hofacker, *From Structure Prediction to Genomic Screens for Novel Non-Coding RNAs.* PLOS Computational Biology, 2011. **7**(8): p. e1002100.
39. Reiche, K. and P.F. Stadler, *RNAstrand: reading direction of structured RNAs in multiple sequence alignments.* Algorithms for Molecular Biology, 2007. **2**(1): p. 6.
40. Sharma, C.M., et al., *The primary transcriptome of the major human pathogen Helicobacter pylori.* Nature, 2010. **464**(7286): p. 250-255.
41. Hör, J., S.A. Gorski, and J. Vogel, *Bacterial RNA Biology on a Genome Scale.* Molecular Cell, 2018. **70**(5): p. 785-799.
42. Ponath, F., et al., *RNA landscape of the emerging cancer-associated microbe Fusobacterium nucleatum.* Nature Microbiology, 2021. **6**(8): p. 1007-1020.
43. Jeong, Y., et al., *The dynamic transcriptional and translational landscape of the model antibiotic producer Streptomyces coelicolor A3(2).* Nature Communications, 2016. **7**(1): p. 11605.
44. Wilms, I., et al., *Deep sequencing uncovers numerous small RNAs on all four replicons of the plant pathogen Agrobacterium tumefaciens.* RNA Biology, 2012. **9**(4): p. 446-457.
45. Roscetto, E., et al., *Functional Characterization of the RNA Chaperone Hfq in the Opportunistic Human Pathogen Stenotrophomonas maltophilia.* Journal of Bacteriology, 2012. **194**(21): p. 5864-5874.
46. Čuklina, J., et al., *Genome-wide transcription start site mapping of Bradyrhizobium japonicum grown free-living or in symbiosis – a rich resource to identify new transcripts, proteins and to study gene regulation.* BMC Genomics, 2016. **17**(1): p. 302.
47. Ryan, D., et al., *A high-resolution transcriptome map identifies small RNA regulation of metabolism in the gut microbe Bacteroides thetaiotaomicron.* Nature Communications, 2020. **11**(1): p. 3557.
48. Mitschke, J., et al., *An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp. PCC6803.* Proceedings of the National Academy of Sciences, 2011. **108**(5): p. 2124-2129.
49. Mitschke, J., et al., *Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in Anabaena sp. PCC7120.* Proceedings of the National Academy of Sciences, 2011. **108**(50): p. 20130-20135.

50. Sundararajan, M., A. Taly, and Q. Yan, *Axiomatic Attribution for Deep Networks*. 2017.
51. Salazar, G., et al., *Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome.* Cell, 2019. **179**(5): p. 1068-1083.e21.
52. Wade, J.T. and D.C. Grainger, *Pervasive transcription: illuminating the dark matter of bacterial transcriptomes.* Nature Reviews Microbiology, 2014. **12**(9): p. 647-653.
53. Feldgarden, M., et al., *Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates.* Antimicrobial Agents and Chemotherapy, 2019. **63**(11): p. 10.1128/aac.00483-19.
54. Cai, X., et al., *Expression of ribosomal protection protein RppA is regulated by a ribosome-dependent ribo-regulator and two mistranslation products.* Environmental Microbiology, 2021. **23**(2): p. 696-712.
55. Rivas, E., *RNA structure prediction using positive and negative evolutionary information.* PLOS Computational Biology, 2020. **16**(10): p. e1008387.
56. Wright, P.R., et al., *CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains.* Nucleic Acids Research, 2014. **42**(W1): p. W119-W123.
57. Maier, L.-K. and A. Marchfelder, *It's all about the T: transcription termination in archaea.* Biochemical Society Transactions, 2019. **47**(1): p. 461-468.
58. Chen, J., H. Boyaci, and E.A. Campbell, *Diverse and unified mechanisms of transcription initiation in bacteria.* Nature Reviews Microbiology, 2021. **19**(2): p. 95-109.
59. Otaka, H., et al., *PolyU tail of rho-independent terminator of bacterial small RNAs is essential for Hfq action.* Proceedings of the National Academy of Sciences, 2011. **108**(32): p. 13059-13064.

Methods

The data augmentation pipeline

To obtain relatively confident putative terminators, for collected 3' end mapping data, protein products of genes with experimentally supported terminators were clustered to 40% sequence identity with cd-hit[1]. 115 clusters that contain protein sequences from at least 3 bacteria phyla were annotated with hmmsearch[2] (hmmer version 3.3.2) using Pfam[3] database and default parameters. To detect remote homology and minimize the taxonomy bias in available 3' end mapping data, instead of directly using protein sequences, Pfam hits of different protein clusters were utilized as surrogates for homolog search. If a protein sequence was annotated to at least 3/4 Pfam domains that were assigned to a gene cluster, it was mapped to the cluster for downstream analysis.

For protein hits, sequences of 16 nt upstream and 128 nt downstream of the stop codons were extracted and subjected to cmfinder[4] search. To consider inter-species variation, and reduce the computational burden, candidate sequences of each cluster were further grouped by the taxonomy relationship. Phyla with < 16 sequences were neglected; phyla with 16-128 sequences were grouped as cmfinder input; for phyla with > 128 sequences, we further grouped these sequences by lower taxonomy level (orders), and ran the previous process recursively (from phylum, class, order, family, genus, to species). This process guarantees each cmfinder run receives no more than 128 sequences from phylogenetically related species.

To ensure reliable data augmentation, we set several cutoffs in this process as follows: all candidate downstream sequences were searched by covariance models of non-redundant motifs with exactly one stem-loop, at least 5 sequences, and at least 5 base pairs with cmsearch program in infernal[5] (version 1.0.2). Only cm models hit >40 downstream sequences in at least 15 other

clusters were kept.

To augment terminator instances with ncRNAs, Rfam seed alignments of a list of known bacteria ncRNA (mainly attenuators and sRNAs) were searched against GEMs bacteria genomes with mmseqs[6], and cmsearch was used to refine the alignment boundary. The results sequences, together with their downstream 128 nt sequences, were searched using cmsearch with cm models passed filtering, and a similar process was applied to downstream sequences of a random subset (128 genes per genome) of protein-coding genes to identify putative terminators of diverse protein-coding genes.

The secondary structures of all putative terminators were predicted with RNAfold[7], and further filtered by the number of base pairs (>7), and the number of loops (exactly 1). The stem-loops with 5 nt flanking sequences at both sides were defined as terminators for model training.

Model implementation and training

The longformer-CRF model was implemented using pytorch[8], with longformer encoder (8 layers with hidden size of 256, intermediate size of 1024, max length of 514, and 64 nt sliding window attention at each side) implemented in transformers[9] package. The model was pre-trained on 42905 GEMs representative genomes with masked language modeling (MLM) loss[10], that is, 20% of the tokens are masked, and the model is trained to predict the masked tokens with unmasked tokens (the loss is defined as cross entropy between predicted logits and actual token).

Denoting longformer predicted score for state j (0 for background, 1 for terminator) at position i as E_{ij}, state of position i as T_i. The learnable parameters in CRF layers are transition scores $P_{T_i-1T_i}$. between two token states. Both longformer predicted scores and transition scores can be

interpreted as probabilities in log space. At the training stage, the state of each token is known, and model parameters are trained end-to-end, the object is to maximize the likelihood of a given state configuration. The score of given state assignments $T = (T_1, \ldots, T_n)$ can be written as:

$$score_T = \sum_{i=1}^{N} \left(P_{T_{i-1}T_i} + E_{iT_i}\right) + P_{T_{i,T_{i+1}}}$$

The likelihood of this state assignment is:

$$LL_T = \frac{\exp\left(score_T\right)}{\sum_{t \in all\ states} \exp(score_t)}$$

Maximizing the likelihood is equivalent to minimizing negative log-likelihood, which is the loss function:

$$loss = score_T - \log\left(\sum_{t \in all\ states} \exp(score_t)\right)$$

For both model pretraining and training, the AdamW optimizer (lr:5e-5, eps:1e-8, betas:(0.9,0.999)) was used with a weight decay of 0.01 and linear decayed learning rate scheduling. A batch size of 128 was used. For model training, 70% of the instances contain putative terminators, and the remaining 30% were either random genomic sequences (25%) or random genomic sequences shuffled by 4-mer with ushuffle[11] (75%). For each instance, a subsequence with a minimal length of 64, and a maximal length of 510 were cropped as model input.

In the inference stage, state paths with top k $score_T$ were determined by Viterbi decoding. Spans of consecutive tokens that were tagged as a terminator that is longer than 32 nt were identified, and averaged predicted probability of the tokens were utilized to reflect the confidence of the prediction:

$$C(s..e) = \frac{1}{e - s + 1} \sum_{i=s}^{e} \frac{\exp\left(E_{i1}\right)}{\exp(E_{i0}) + \exp\left(E_{i1}\right)}$$

Input genome sequences were scanned with a window size of 500 nt, and a stride size of 100 nt. The top 10 predictions were considered. For overlapping predictions on the same strand, only the

66 best-scoring one is selected as the final prediction.

67

68 **Metaterm-seq and metatranscriptome data processing**

69 Metaterm-seq and paired RNA-seq reads of 2 human oral metagenome samples were downloaded
70 from PRJEB12568. Metaphlan2[12] was used for species profiling, species with abundance > 1% in
71 at least one sample were considered for downstream analysis. Metaterm-seq reads were mapped to
72 corresponding genomes downloaded from Genbank (Table S3). 5' end coverage of the reads was
73 summarized with HTSeq[13]. Local max positions within 50 nt windows with read coverage >=6 are
74 considered as transcripts 3' ends with experimental support.

75 Metatranscriptome reads of Tara Ocean are accessed from PRJEB6608, and contigs assembled
76 from paired DNA-seq data were downloaded from https://www.ebi.ac.uk/biostudies/files/S-
77 BSST297/OM-RGC_v2_assemblies.tar.gz. CDSs were predicted by prodigal[14] in anonymous
78 mode. RNA reads of each sample were mapped to their paired contigs with bowtie2[15] using the
79 default parameter, and transcripts were defined by stringtie[16].

80

81 **Performance evaluation and model interpretation**

82 For performance evaluation, the false positive rate was estimated by scanning 1000 1 KB sequences
83 sampled from the tetramer frequency distribution of the corresponding genome. For term-seq data,
84 predicted intervals that overlap with the upstream 40 nt of term-seq identified 3' ends on the same
85 strand were considered as true positive.

86 We ran BATTER with the default setting. For RNIE, the parameter "--thresh 0 --gene -m gene.cm"
87 was used. For transterm-HP, the parameter "--min-conf=50 --all-context" was used. For arnold, the

web server http://rssf.i2bc.paris-saclay.fr/toolbox/arnold/ was used with default setting. For overlapped predictions of existing tools, the intervals with the highest score were selected for performance evaluation.

For model interpretation, we used the LayerIntegratedGradients modules implemented in captum[17] package, to attribute the contribution of token embeddings at each position for predicting terminators. For each position, we take the sum of scores assigned to each of the 256 embedding dimensions as an importance measure.

Visualization of predicted terminators

The prediction performance and terminator characteristics were visualized with the web tool interactive Tree of Life [18] (https://itol.embl.de/). Predicted secondary structures of representative terminators were visualized with varna[19].

For t-SNE visualization of terminators, a BERT encoder was trained to project each RNA sequence to a numeric vector on a 256-dimensional unit sphere. The model was implemented with transformers[9] library, pre-trained on RNAcentral[20] database with MLM loss, and fine-tuned on Rfam supervised by a triplet loss. Given an instance in a minibatch as anchor a, within the same minibatch, the instance that shares the same label with the anchor while having the largest cosine distance d_{pa} is considered as the positive instance, and the instance with the different label while having the smallest cosine distance d_{na} is considered as the negative instance. Then triplet loss of the anchor instance is defined as:

$$L(p, n, a) = max\ (d_{pa} - d_{na} + margin, 0)$$

$margin$ is a hyperparameter (set to 0.32 in this study). Within a minibatch, based on the

similarity metric learned by the neural network, if the most dissimilar positive instance is not closer than the most similar negative instance by a margin, it will contribute to the loss. The overall effect is that the embeddings of similar RNAs were attracted closer to each other while distinct ones were pushed away. The qualities of RNA embeddings were validated on holdout RNA families. The vector embeddings were visualized with Multicore-TSNE[21].

PCA and K-means clustering were implemented with decomposition.PCA and cluster.KMeans modules in scikit-learn package[22], respectively. The statsmodels[23] package was used for the linear model that decomposes the variance of the principal components.

Scanning of AMR gene leader sequences

AMR genes in GEMs representative genomes were annotated with AMRfinder[24]. 800 nt sequences (upstream 400 nt and downstream 400 nt) flanking start codons of detected AMR genes are subjected to terminator scanning. Predicted terminator located within upstream 128 nt and downstream 16 nt of stop codons are retained. These upstream 128 nt and downstream 16 nt terminator-containing sequences were grouped by AMR families and searched for conserved motifs with cmfinder[4]. The secondary structural covariations of resulting motifs were evaluated with R-scape[25]. Targets of sRNA_1300 (RF02862) were predicted using CopraRNA[26] with default parameters.

References

1. Fu, L., et al., *CD-HIT: accelerated for clustering the next-generation sequencing data.* Bioinformatics, 2012. **28**(23): p. 3150-3152.

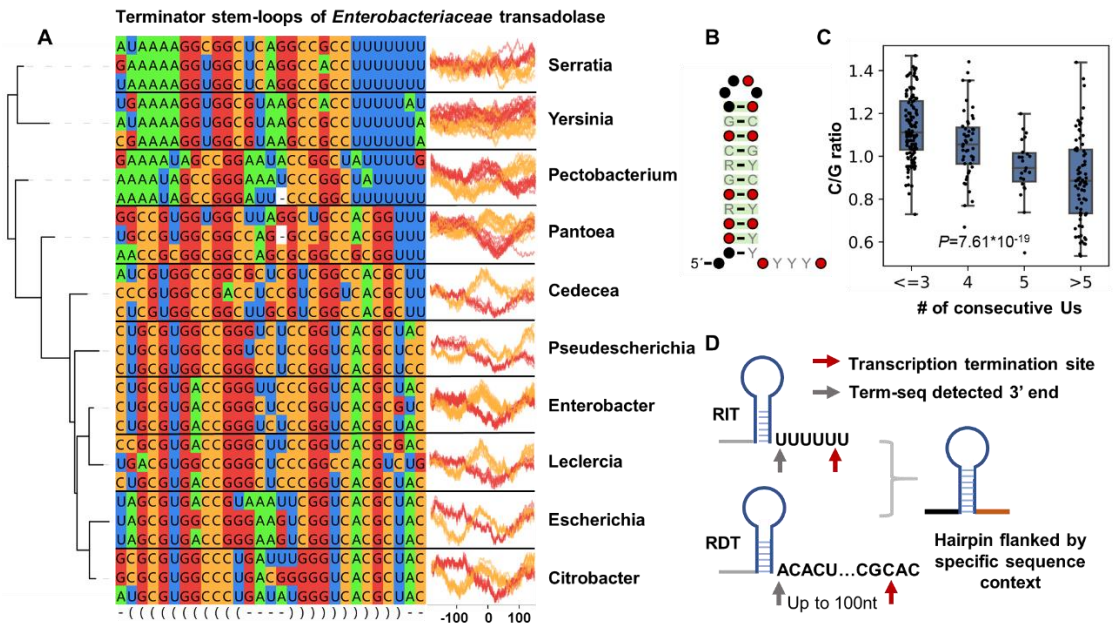
2. Eddy, S.R., *Accelerated Profile HMM Searches.* PLOS Computational Biology, 2011. **7**(10): p. e1002195.
3. Mistry, J., et al., *Pfam: The protein families database in 2021.* Nucleic Acids Research, 2021. **49**(D1): p. D412-D419.
4. Yao, Z., Z. Weinberg, and W.L. Ruzzo, *CMfinder—a covariance model based RNA motif finding algorithm.* Bioinformatics, 2006. **22**(4): p. 445-452.
5. Nawrocki, E.P., D.L. Kolbe, and S.R. Eddy, *Infernal 1.0: inference of RNA alignments.* Bioinformatics, 2009. **25**(10): p. 1335-1337.
6. Steinegger, M. and J. Söding, *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.* Nature Biotechnology, 2017. **35**(11): p. 1026-1028.
7. Lorenz, R., et al., *ViennaRNA Package 2.0.* Algorithms for Molecular Biology, 2011. **6**(1): p. 26.
8. Paszke, A., et al., *PyTorch: An Imperative Style, High-Performance Deep Learning Library.* CoRR, 2019. **abs/1912.01703**.
9. Wolf, T., et al., *HuggingFace's Transformers: State-of-the-art Natural Language Processing.* arXiv [cs.CL], 2020.
10. Devlin, J., et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv e-prints, 2018: p. arXiv:1810.04805.
11. Jiang, M., et al., *uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts.* BMC Bioinformatics, 2008. **9**(1): p. 192.
12. Truong, D.T., et al., *MetaPhlAn2 for enhanced metagenomic taxonomic profiling.* Nature Methods, 2015. **12**(10): p. 902-903.
13. Anders, S., P.T. Pyl, and W. Huber, *HTSeq—a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-169.
14. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC Bioinformatics, 2010. **11**(1): p. 119.
15. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biology, 2009. **10**(3): p. R25.
16. Pertea, M., et al., *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.* Nature Biotechnology, 2015. **33**(3): p. 290-295.
17. Kokhlikyan, N., et al., *Captum: A unified and generic model interpretability library for PyTorch*. 2020.
18. Letunic, I. and P. Bork, *Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation.* Nucleic Acids Research, 2021. **49**(W1): p. W293-W296.
19. Darty, K., A. Denise, and Y. Ponty, *VARNA: Interactive drawing and editing of the RNA secondary structure.* Bioinformatics, 2009. **25**(15): p. 1974-1975.
20. Consortium, R.N., *RNAcentral 2021: secondary structure integration, improved sequence search and new member databases.* Nucleic Acids Research, 2021. **49**(D1): p. D212-D220.
21. Ulyanov, D., *Multicore-TSNE*. 2016, GitHub.
22. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**(85): p. 2825-2830.
23. Seabold, S. and J. Perktold. *statsmodels: Econometric and statistical modeling with python*. in *9th Python in Science Conference*.
24. Feldgarden, M., et al., *Validating the AMRFinder Tool and Resistance Gene Database by Using*

Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. Antimicrobial Agents and Chemotherapy, 2019. **63**(11): p. 10.1128/aac.00483-19.

25. Rivas, E., J. Clements, and S.R. Eddy, *A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs.* Nature Methods, 2017. **14**(1): p. 45-48.

26. Wright, P.R., et al., *CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains.* Nucleic Acids Research, 2014. **42**(W1): p. W119-W123.

Figure 1. ***Common local structures of RITs' and RDTs' upstream termination sites and their covariation with sequence context***



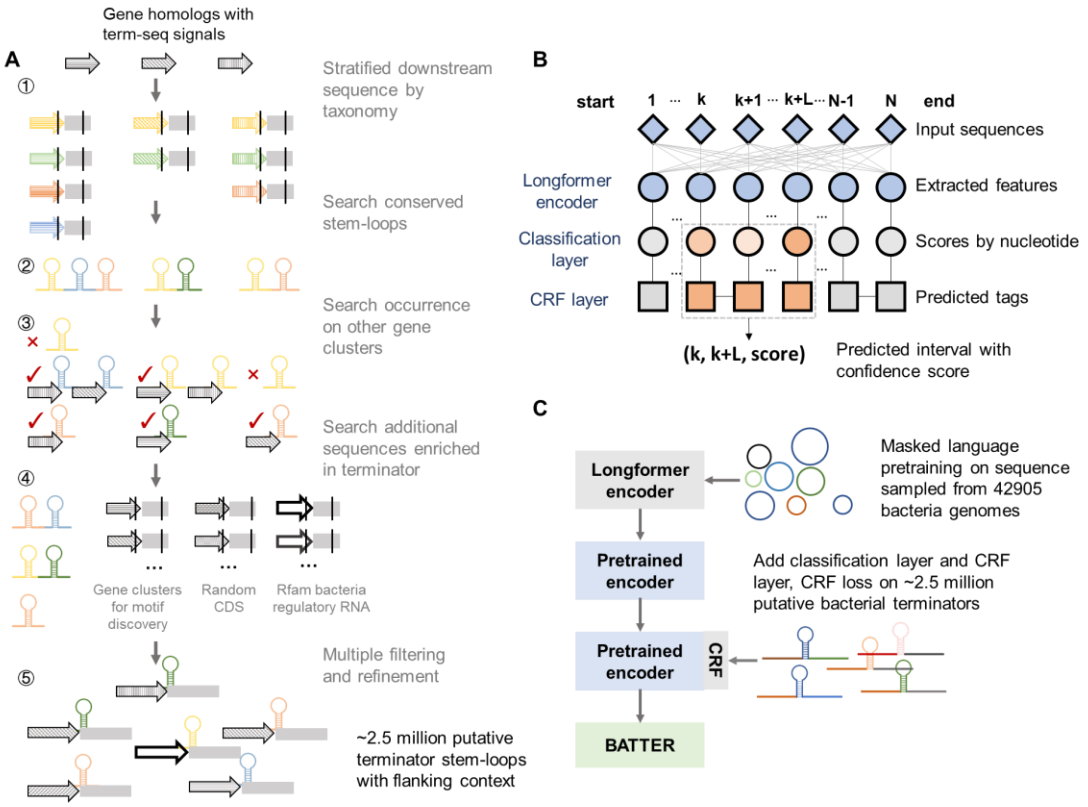
A. Multiple sequence alignment (MSA) of the conserved stem-loop structure in downstream sequences of CDSs encoding transaldolase in the *Enterobacteriaceae* family. For each genus, sequences from 3 representative species are shown. In the right panel, for all species of each genus in GEMs representative genomes, guanodine (G) ratios and cytosine (C) ratios (labeled in red and orange, respectively) of 101 nt sequences centered at each nucleotide from 128 nt upstream to 128 nt downstream relative to ends of the stem-loops are shown.

B. Secondary structure covariation of the identified stem-loop. Base pairs with R-scape *E* value < 0.05 are labeled in green, and nucleotides with sequence identity > 97% and > 90% are labeled in red and black, respectively. R indicates purine, Y indicates pyrimidine.

C. Correlation between maximum lengths of consecutive uracils (Us) and C/G ratio within 128 nt sequences flanking ends of the stem-loops downstream transaldolase CDSs in all considered *Enterobacteriaceae* genomes.

D. Schematic illustration of common and distinctive features of RITs and RDTs. For RITs, term-seq sites indicate the location of terminator stem-loops, while for RDTs, term-seq sites indicate stem-loops upstream of real termination sites. Both RITs and RDTs are characterized by stem-loops flanked by particular sequence contexts.

Figure 2. ***Data augmentation, model architecture, and the context-aware model training pipeline***

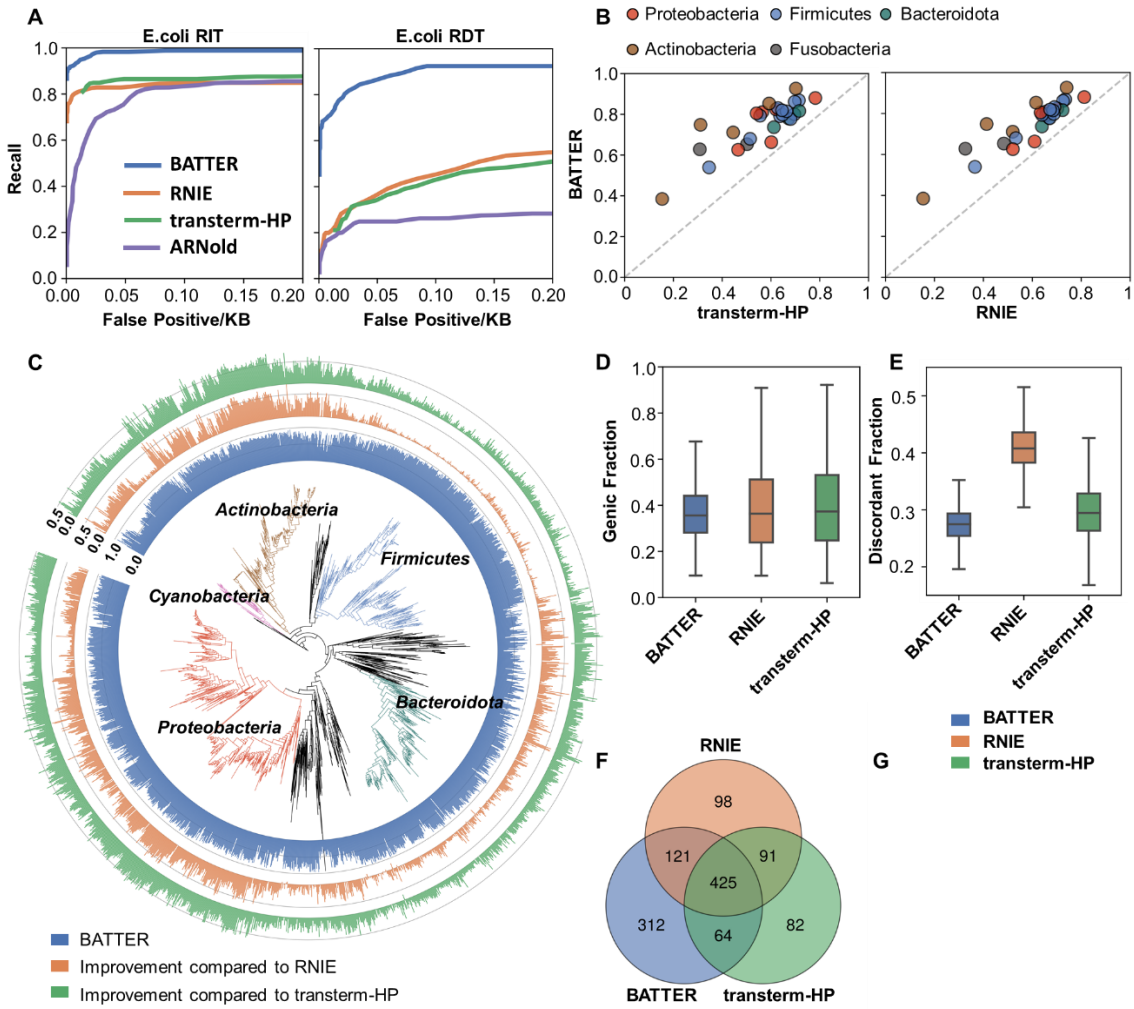


A. A schematic overview of the data augmentation pipeline. For each protein-coding gene cluster (patterns indicate different gene clusters) supported by term-seq signal from at least 3 phyla, 16 nt upstream and 128 nt downstream sequences relative to stop codons were extracted and stratified by taxonomy (step ①, colors indicate different clades), then subjected to cmfinder search (step ②). Stem-loop motifs discovered from each cluster were searched against other gene clusters, and motifs specific to too few gene clusters were discarded (step ③). Motifs passed the filter were further searched against diverse sequences known to be enriched in terminators (step ④). After multiple steps of filtering and refinement, a final set of putative terminators was identified (step ⑤).

B. The model architecture of BERT-CRF for terminator prediction. The model is composed of a longformer encoder, a dense layer for classification, and a CRF layer. The logits predicted by longformer model are post-processed by a CRF layer to make the final prediction. For predicted intervals, the average of probabilities by nucleotide (predicted by the classification layer) across the interval is utilized as a confidence measure.

C. The model training pipeline. The longformer encoder is pre-trained on bacteria genomic sequences sampled from GEMs bacteria genomes with masked language modeling (MLM) loss in an unsupervised manner. The encoder parameters, together with parameters of the appended classification layer and CRF layer, are jointly finetuned to predict transcription terminators.

Figure 3. ***Performance evaluation and comparison with existing methods***



A. Receiver operating characteristic (ROC) curve of BATTER and 3 published methods on *E. coli* term-seq dataset with RIT/RDT annotations (left panel for RIT, right panel for RDT).

B. Performance comparisons for BATTER vs. RNIE (left panel) and BATTER vs. transterm-HP (right panel) on an oral microbiome term-seq dataset. Recalls of two methods at an FPR of 0.1/KB on different species are plotted against each other, colors indicate different phyla.

C. Performance comparisons on predicting terminators between tail-to-tail gene pairs in diverse bacteria species. The inner cycle shows recall of BATTER (at an FPR of 0.1/KB), and the outer two cycles show recall improvement relative to RNIE and transterm-HP, respectively.

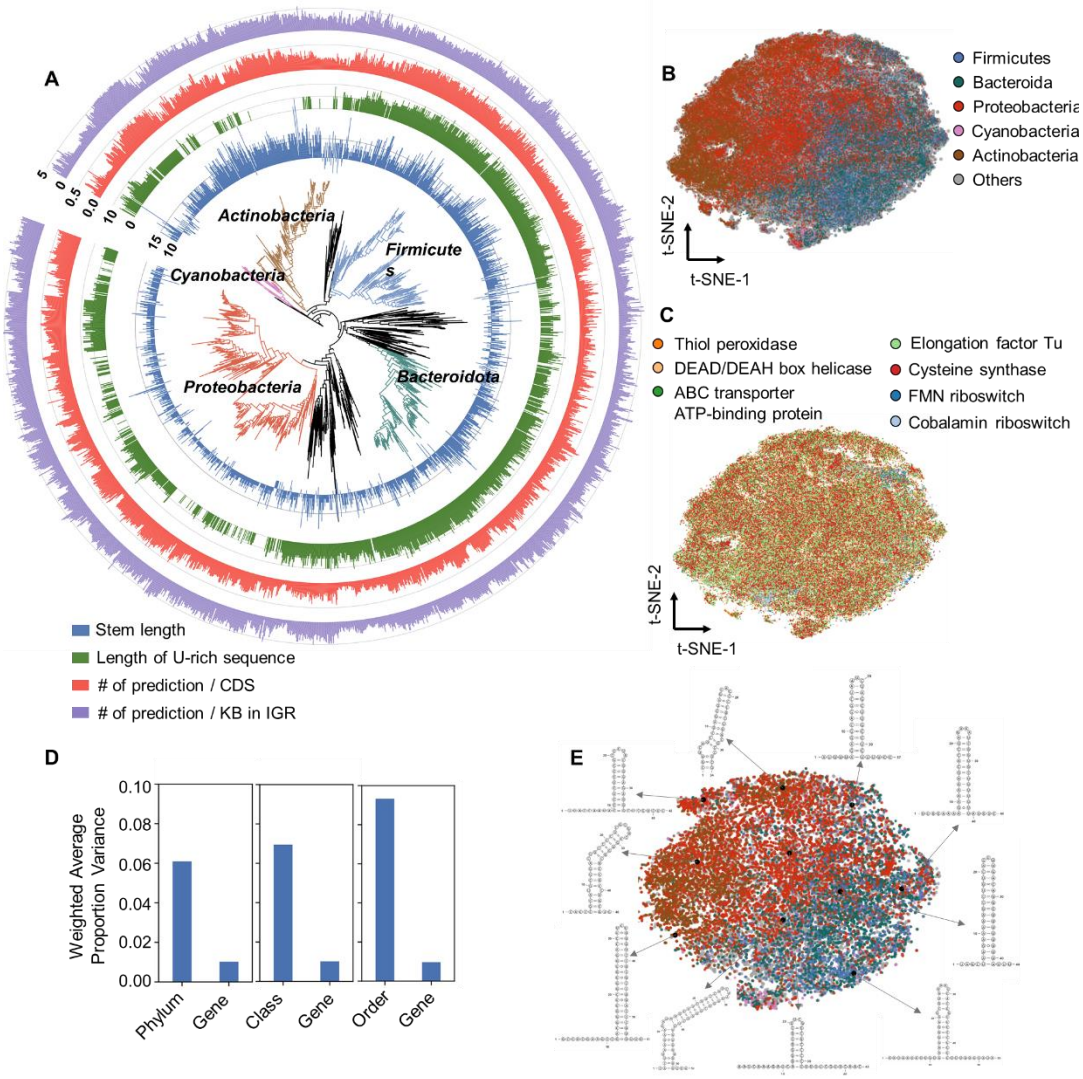
D. Fraction of predictions made by 3 methods that overlap with CDSs.

E. Discordant fractions (fraction of antisense terminators in all terminators predicted near 3' ends of CDSs) in predictions made by 3 methods.

F. Overlaps between RNIE, BATTER, and transterm-HP's predictions on curated differential RNA-seq datasets, different studies are pooled together.

G. Frequency of BATTER, RNIE, and transterm-HP's performance ranks on differential dRNA-seq datasets.

Figure 4. ***Inter-gene and inter-clade variations of bacteria terminators***



A. Characteristics of predicted terminators on different bacteria species. From the innermost cycle to the outermost cycle, the mode of stem lengths (blue), U-rich sequence lengths (green), number of prediction/CDS (red), and number of predictions/KB (purple) are shown.

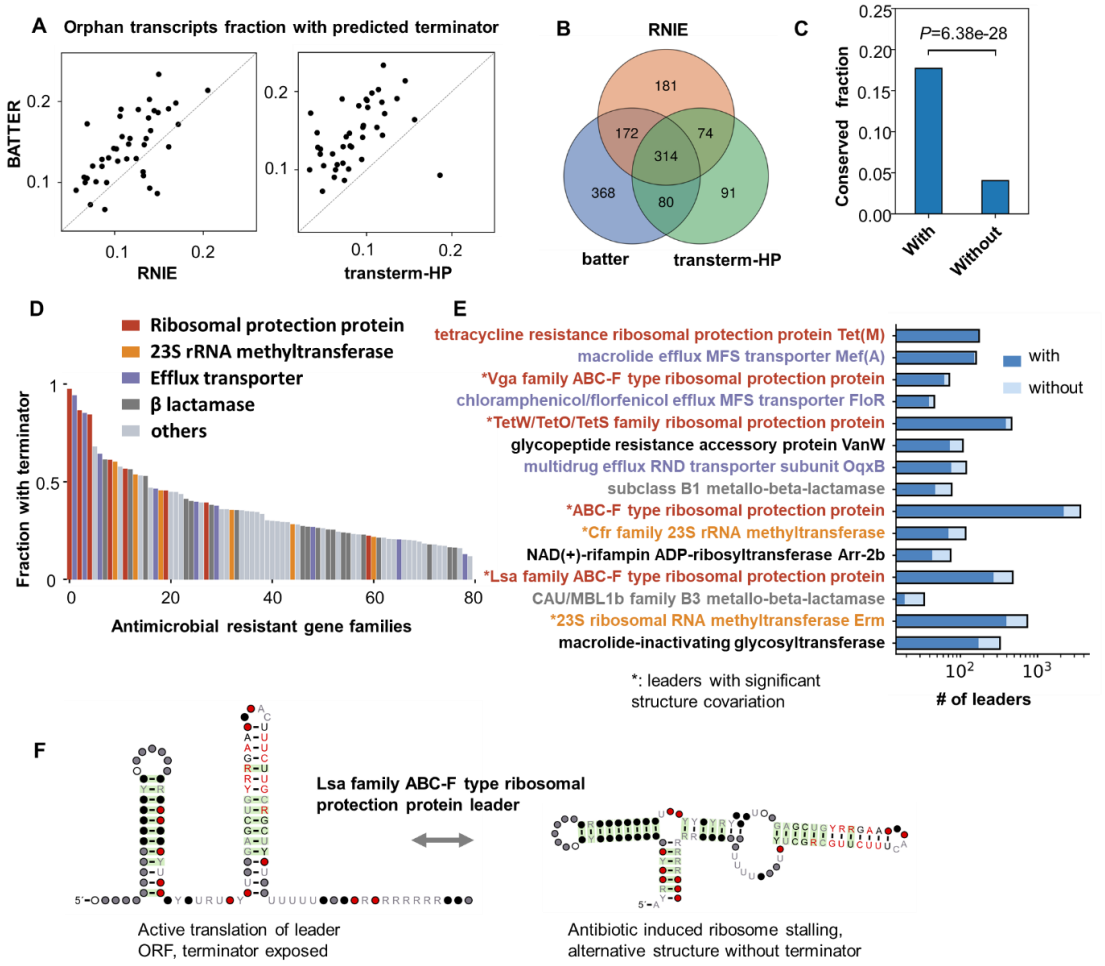
B. t-SNE (t-distributed stochastic neighbor embedding) on vector embeddings of terminators downstream several widely distributed genes, coloring by phyla.

C. t-SNE visualization same as Figure 4B, coloring by upstream gene families

D. Weighted average variance proportion of top 36 principal components explained by different genes and different clades. Taxonomies at three ranks were used as proxies for bacteria clades (left, middle, and right panel for phylum, class, and order, respectively).

E. Visualization of terminators associated with cysteine synthase CDS. The vector embeddings were clustered into 10 clusters with K-means, and the predicted secondary structures of cluster centers (black dot) were shown.

Figure 5. ***Application on metatranscriptome analysis and AMR-related ncRNA identification***



A. Comparison of the fraction of orphan transcripts with terminators predicted by different methods. Left panel: BATTER vs. RNIE, right panel: BATTER vs. transterm-HP.

B. Venn diagram of terminators predicted by different methods on Tara Ocean dataset.

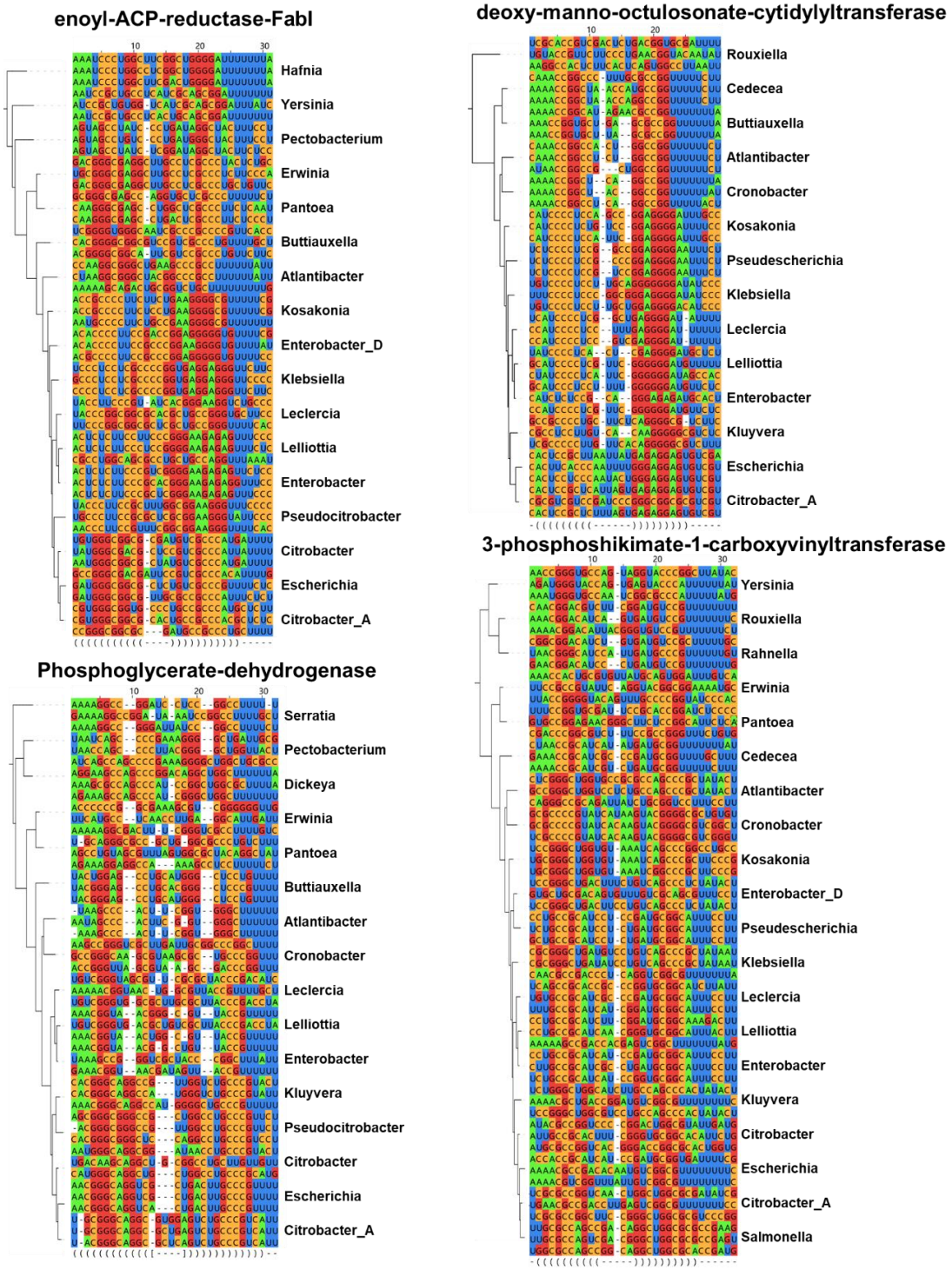
C. Non-singleton fraction in orphan transcripts with and without terminators predicted by BATTER.

D. Antimicrobial resistance (AMR) gene families are ranked by fractions with predicted terminators, each bar indicates an AMR gene family. Gene families that share similar mechanisms of action are labeled with the same color.

E. Fraction of leader sequence preceding AMR gene families that are frequently associated with predicted transcription terminators.

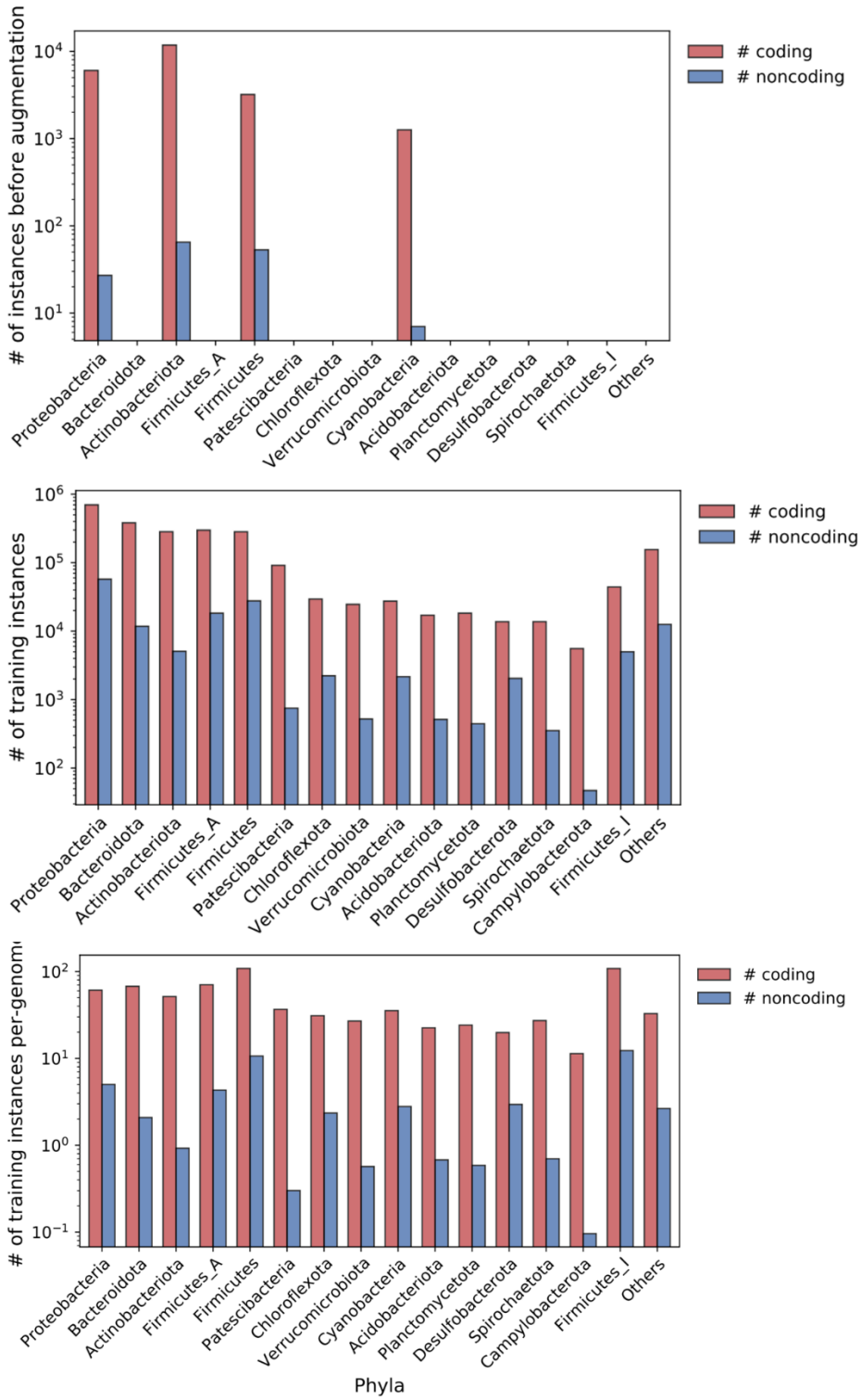
F. Two alternative secondary structures of a conserved element upstream Lsa family ABC-F type ribosomal protection proteins predicted by Cacofold. Base pairs supported by significant sequence covariation (R-scape *E* value < 0.05) are labeled in green.

Supplementary Figure 1. ***Gene 3' end homologs with seemingly lineage-specific factor dependencies but conserved secondary structures***



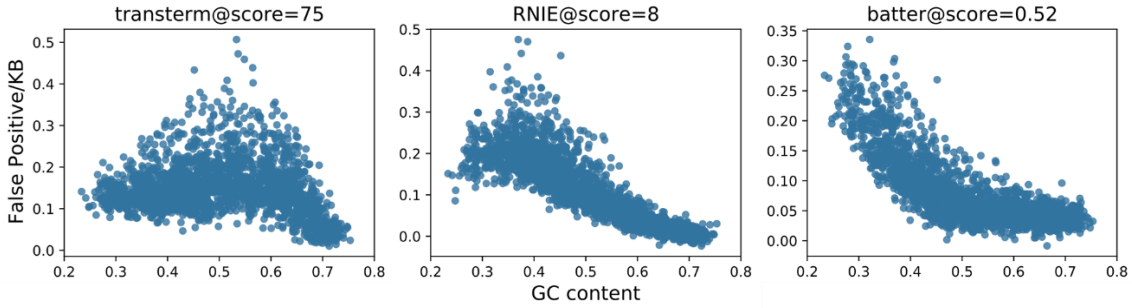
Multiple sequence alignments of conserved stem-loops downstream four genes, enoyl-ACP-reductase-FabI (RDT in *E. coli*), deoxy-manno-octulosonate-cytidylyltransferase (RDT in *E. coli*), phosphoglycerate dehydrogenase (RIT in *E. coli*) and 3-phosphoshikimate-1-carboxyvinyltransferase (RIT in *E. coli*) in *Enterobacteriaceae* are shown. In each of the four genes, terminators in some species have RIT-like properties, while the others have RDT-like properties.

Supplementary Figure 2. ***Taxonomy distribution of terminator instances before and after data augmentation***



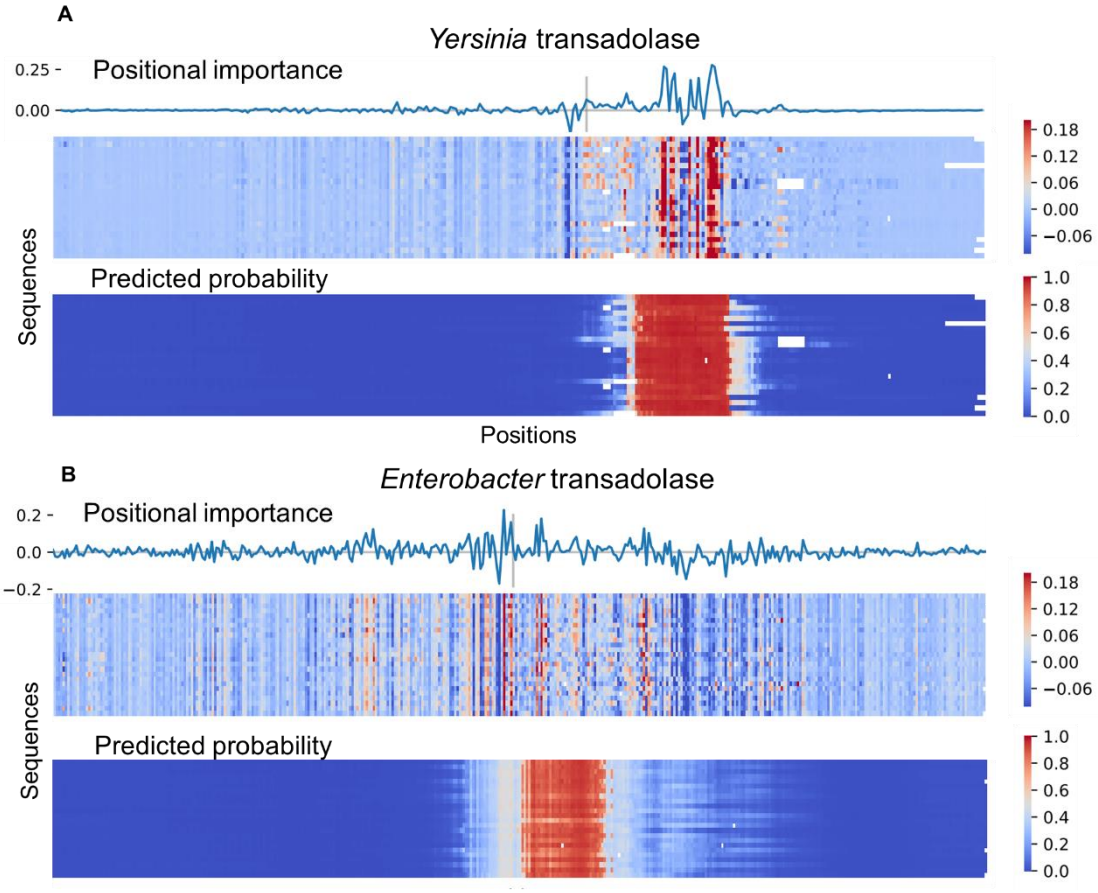
Total number of experimentally supported terminators (upper panel), total number of augmented putative terminators (middle panel), and number of putative terminators per genome (bottom panel) associated with coding genes (red) and noncoding RNAs (blue) in each phylum.

Supplementary Figure 3. ***Correlations between GC-content and FPR at fixed model score cutoffs***



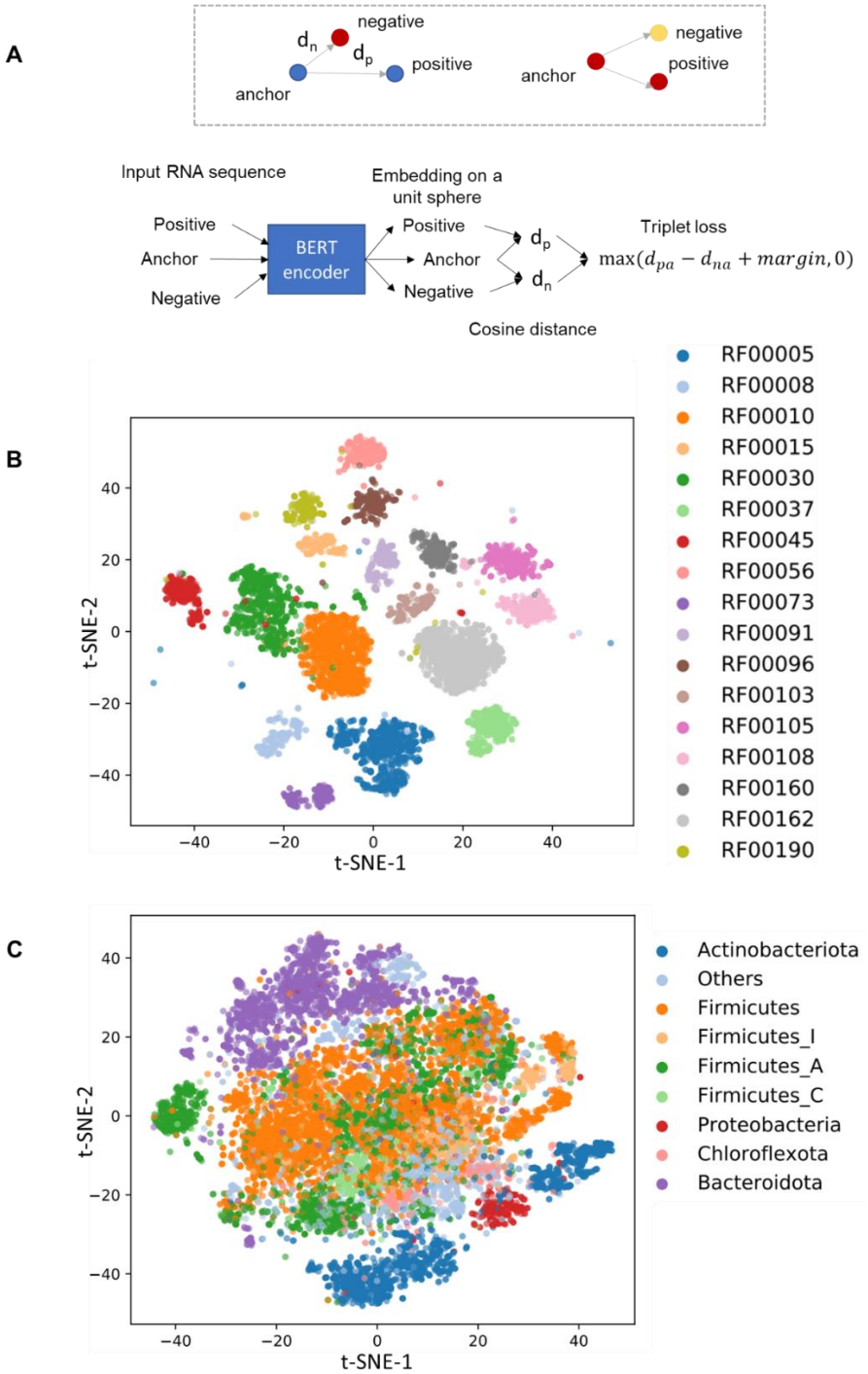
GC content is a proxy for variations in tetramer frequency across different bacteria genomes. For transterm-HP, RNIE, and batter, the false positive rates at a prespecified cutoff on sequence simulated from tetramer frequency of different genomes are plotted against the GC content of corresponding genomes.

Supplementary Figure 4. ***Positional importance of nucleotides for terminator prediction measured by integrated gradient***



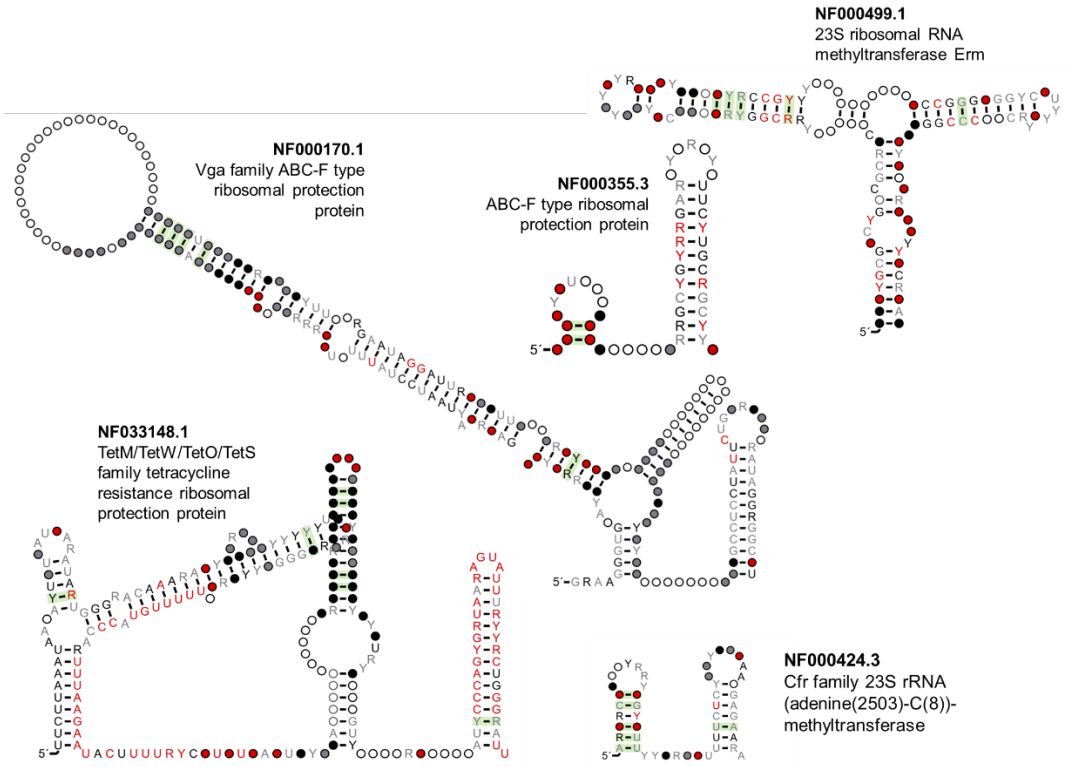
The contribution of context information for terminator prediction in downstream sequences of transaldolase in *Yersinia* (A, encompass RIT-like terminators) and *Enterobacter* (B, encompass RDT-like terminators), respectively. For nucleotides in each sequence, the probabilities of being transcription terminators were predicted by BATTER, and the importance of each position was evaluated with the integrated gradient algorithm. Scores were mapped to corresponding nucleotides in the multiple sequence alignments and columns that contained gaps in > 20% of the sequences were excluded from visualization. Upper panel: positional importance averaged across sequences. Middle panel: positional importance by sequences. Bottom panel: the predicted probability of each nucleotide.

Supplementary Figure 5. ***Training and validation of neural encoder for terminator visualization***



A. Schematic illustration of RNA encoder. A BERT model pre-trained on RNAcentral database is finetuned using Rfam data under the supervision of triplet loss that pushes RNA sequence of the same family closer to each other
B. t-SNE visualization of the embedding of RNA in the holdout set, colors indicate different RNA families.
C. t-SNE visualization of SAM riboswitches (holdout from training), colors indicate different phyla.

Supplementary Figure 6. S*tructured RNA upstream AMR genes*



Predicted consensus RNA secondary structure of conserved structural element upstream several ARM gene families.