# Conserved 3' Stem-Loop Structures Enable Comprehensive Analysis of Bacterial Transcription Termination in Metagenomes, Regardless of Rho Factor Dependency

Yunfan Jin[1,2,†], Jiyun Cui[3,†], Hongli Ma[4,1,2,5], Fei Gan[3*], Zhenjiang Zech Xu[5,6*], 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 Hubei Key Laboratory of Cell Homeostasis, College of Life Science, TaiKang Center for Life and Medical Sciences, Wuhan University, Wuhan, 430072, China

4 School of Mathematics, Harbin Institute of Technology, Harbin 150001, China

5 State Key Laboratory of Food Science and Resources, Nanchang University, Nanchang 330047, China

6 School of Mathematics and Computer Sciences, Nanchang University

*To whom correspondence should be addressed: Zhi John Lu, Email: urluzhi@gmail.com; Correspondence may also be addressed to Zhenjiang Zech Xu, Email: zhenjiang.xu@gmail.com or Fei Gan, Email: feigan@whu.edu.cn.

†These authors contributed equally to this work.

**Competing Interest Statement:** The authors declare no potential conflicts of interest.

**Keywords:** bacteria terminator, transcription termination, transcription unit prediction

**Abstract**

Bacterial transcription termination is a critical yet underexplored mechanism of gene regulation in microbial ecosystems. Existing computational tools, however, primarily focus on predicting Rho-independent terminators (RITs) in model species, leaving significant gaps in understanding Rho-dependent termination (RDTs) and termination mechanisms in non-model species. To address these limitations, we developed BATTER (BActeria Transcript Three Prime End Recognizer), a comprehensive computational tool for bacterial transcript 3' termini prediction. BATTER builds on the observation that conserved stem-loop structures are frequently associated with 3' ends of primary transcripts generated by both RIT and RDT mechanisms across distantly

related bacterial species. By leveraging Longformer (a transformer-based neural network model) with a CRF (Conditional Random Field) layer, BATTER demonstrated superior performance compared to existing tools. It enabled a comprehensive analysis of 42,905 representative bacterial genomes, uncovering that stem-loop structures exhibit clade-specific properties with greater variations between species than between gene families. Notably, BATTER uncovered that certain *Cyanobacteria* lineages, despite lacking Rho homologs, harbor Rho utilization (RUT) site-like sequences near 3' ends, with preliminary experimental validation in *E. coli* suggesting their partial functionality in transcription termination. Additionally, BATTER systematically identified pervasive premature termination events in antimicrobial resistance (AMR) genes, highlighting their regulatory roles in translation protection and drug efflux. This study advances our understanding of transcription termination across diverse bacterial lineages and provides a robust computational approach for exploring transcription regulation in complex microbial ecosystems.

## Introduction

Bacteria transcription terminators are RNA elements that dissociate RNA polymerase complexes and induce transcription termination. They are traditionally classified into two categories: Rho-independent terminators (RITs) and Rho-dependent terminators (RDTs)[1, 2]. In model species *Escherichia coli* and *Bacillus subtilis*, RITs are GC-rich RNA hairpins followed by a stretch of uracils. RDTs typically contain a ~60 nt subsequence called Rho utilization (RUT) sites, which are the primary binding sites of Rho protein[3]. RUT sites are rich in YC (a pyrimidine followed by a cytosine) dimers that are regularly spaced by distances of around 10 nt[2, 4]. They are also characterized by features like high C/G ratios (referred to as C>G “bubbles” [3]) and depletion of secondary structures[2].

RDTs were reported to account for 20%-30% of the primary transcription termination events in *E. coli* [5], but play a less important role in *B. subtilis*[6]. Computational prediction algorithms for RIT have been established in *E. coli* and *B. subtilis*[7-9], yet predicting RDTs remains challenging. This difficulty arises because the known sequence patterns of RDTs are less compact and distinct than those of RITs, even in well-studied model organisms [2, 3, 10]. Available tools predict a paucity of RITs in some bacteria species[7, 9], raising questions as to whether these species mainly use Rho-dependent mechanisms, or their RITs have alternative sequence patterns. This gap in computational predictions poses a significant challenge for studying transcription termination mechanisms across bacterial lineages, especially in metagenomic datasets where diverse bacterial species are often present. Accurate terminator prediction is essential for understanding gene regulation and community dynamics in these complex microbial environments. Limitations in

current computational methods thus present a major barrier to comparative analyses of transcription terminators across diverse bacterial clades, impeding insights into transcriptional regulation in diverse microbial communities.

Recently, several high-throughput 3' ends mapping experimental techniques, such as term-seq[11], Rend-Seq[12], and SEnd-seq[13] were developed. These techniques detect 3' ends of RNAs, regardless of whether the transcripts employ RITs or RDTs for transcription termination. Although a 3' ends mapping experiment alone cannot determine whether a transcript end is generated by RIT or RDT mechanism, the rough classification can be achieved by evaluating changes in RNA-seq coverage flanking detected transcript 3' ends upon inactivation of Rho protein by gene knockout[14], inducible protein product depletion[15] or the small molecule inhibitor BCM (bicyclomycin)[16, 17].

Available term-seq data suggested stem loop structures are pervasively associated with transcript 3' ends in bacteria species belongs to *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Spirochaetota,* and *Cyanobacteria* lineages[12, 17-26]. While RUT sites are thought to be depleted of RNA secondary structure[1, 2], based on 3' end mapping techniques, it was reported that in *E. coli*, transcript ends generated by the Rho-dependent mechanism can fold into stem loop structures reminiscent of that in RITs with comparable folding energies, but typically lack the U-rich tracts[5]. In 3' end mapping data of *Streptomyces lividans* TK24, stem loop structures are also frequently observed in 3' ends that not associated with U rich sequences[19]. A possible explanation for this apparent discrepancy is that, for transcripts terminated by RDTs, 3' ends are rapidly degraded by intracellular exoribonuclease such as PNPase after transcription termination, until reaching the stem loop structures which prevent further degradation[5, 27].

The stem loop structures are crucial for maintaining transcript integrity, as a result, for functional transcripts, a 3' end associated stem loop might be positively selected in evolution[5]. Although the steady-state 3' ends may not correspond to real transcription termination sites[28], predicting 3' ends is sufficient for many applications, such as operon prediction and sRNA prediction. If this property is broadly conserved in bacteria, as a stem loop structure is a stronger feature compared to sequence patterns of RUT sites, this observation can be leveraged for comprehensive prediction of transcript 3' ends in diverse bacteria lineages.

In this study, through term-seq data analysis, we validated that at least for primary transcription terminations, 3' ends generated by both RITs and RDTs are frequently associated with conserved stem loop structures in diverse bacteria species. Based on this observation, exploiting advancement in both experimental data and computational techniques, we present BATTER (https://github.com/lulab/BATTER), a computational tool for bacteria transcription termination analysis. BATTER is composed of two modules: a machine learning model based on BERT-CRF architecture for predicting steady-state RNA 3' ends (BATTER-TPE, Three Prime End) and a rule-

based algorithm for detecting putative RUT sites (BATTER-RUT). Systematic evaluation on diverse metrics suggested BATTER outperformed existing tools for both 3' ends detection and RUT sites prediction.

Using BATTER, we comprehensively predicted transcript 3' ends and putative RUT sites near the predicted 3' ends in representative bacteria genomes. We characterized the variations of these conserved stem loops across diverse bacterial lineages and evaluated their coevolution with the presence of Rho proteins and putative RUT sites. Our analysis suggests a competing role between RDT and RIT in different species. Unexpectedly, we also detected RUT sites-like signatures in *Cyanobacteria* species, which lack Rho homologs. This signature is correlated with more recently evolved traits. Preliminary experimental analysis suggested some of these RUT site-like sequences are capable of inducing transcription terminations in *E. coli*. As another application, by predicting premature terminations of known antimicrobial resistance (AMR) genes, we systematically identified premature termination-based regulations and related noncoding RNAs. Our analysis suggests premature termination-based regulations are pervasive in AMR genes related to translation protection and drug efflux.

## Results

### A conserved stem loop is associated with transcript 3' end regardless of the factor dependency

To inspect the secondary structures of detected 3' ends, we curated available 3' ends annotations determined by term-seq, and RNA-seq data with and without Rho inhibition from 4 phylogenetically diverse species (belonging to 4 distinct phyla), including *Escherichia coli* [5, 16], *Bacillus subtilis*[14, 27], *Borrelia burgdorferi*[17] and *Mycobacterium tuberculosis*[15](Table S1, Table S2). It has been reported RNA-seq is more effective in identifying Rho-dependent regions compared to term-seq, as term-seq tends to detect steady-state 3' end instead of RUT sites[29]. Using RNA-seq data, the detected 3' ends were classified as RIT or RDT according to changes in read-through upon Rho inhibition (see Methods). Here we define "primary" termination sites as transcript 3' ends located within 128 nt downstream of the CDSs, slightly more flexible compared to the 100 nt cutoff used by a previous study[30] as it has been reported 3' UTRs longer than 100 nt are not rare in some species[31]. In all four species, 3' ends detected by term-seq exhibit lower folding energies compared to the genomic background, particularly at primary termination sites. Notably, the folding energies of primary 3' ends assigned to RDTs and that assigned to RITs are similar (Figure 1A).

As an example, the RNA-seq read coverages near 3' end of *E. coli* talB gene were illustrated (Figure 1B). Upon BCM treatment, talB has a higher level of transcription read-through. The presence of regularly spaced YC dimers and a high C/G ratio confirmed that the Rho-dependent mechanism is involved. There is a stem loop structure immediately upstream of its 3' end. To

135 analyze the conservation of the stem loop, we identified homologs of talB gene within taxa
136 *Enterobacteriaceae* in representative genomes of GEMs (Genomic catalog of Earth's Microbiomes)
137 dataset [32] and searched conserved hairpins near stop codons (upstream 16 nt and downstream
138 128 nt) with CMfinder[33]. As shown in Figure 1C, species within *Escherichia* genus, as well as
139 *Citrobacter*, *Leclercia*, *Enterobacter*, and *Pseudescherichia*, exhibit RDT-like properties including
140 the presence of C>G bubbles and the absence of U-rich tracts, while in *Serratia*, *Yersinia*, and
141 *Pectobacterium* species, the homolog transcripts are ended with a typical RIT-like motif,
142 characterized by a stretch of uracils (Us) and lacking a C>G bubble in downstream sequences. R-
143 scape[34] analysis revealed strong covariation of the predicted base pairing (Figure 1D). We also
144 observed a significant negative correlation between the maximum length of U runs near the stem
145 loops and the C/G ratio within 100 nt upstream and 100 nt downstream flanking sequences, (Figure
146 1E, ANOVA $P$=7.61$\times$10$^{-19}$). These results suggested the hairpin is structurally conserved in all
147 homolog transcripts, despite the apparent differences in Rho factor dependency. talB is not a
148 special case, and similar phenomena are observed in other genes, 3 of such genes were
149 exemplified in Figure S1.
150 Taken together, primary transcription terminations conducted by both RIT and RDT are
151 characterized by conserved stem loops, and the stem loop structures can be more conserved than
152 Rho factor dependency. This feature is likely to be conserved in diverse bacteria lineages. Thus, it
153 is possible to predict transcript 3' ends generated by both RITs and RDTs in a common framework,
154 that is predicting stem loop structures embedded in certain sequence contexts (Figure 1F).
155
156 **Models for transcript 3' ends and RUT sites prediction**
157 We collected transcript 3' ends defined by published genome-wide studies[12, 18-26] (Table S1).
158 Naively training models on experimentally verified 3' ends presented several challenges, including
159 phylogeny distribution bias (Figure S2 B), limited sample size, and variations due to different
160 experimental settings. Based on the conservation of primary 3' ends associated stem loop
161 structures, we developed a data augmentation pipeline (Figure S2 A) and generated ~2.5 million
162 putative 3' ends associated stem loop instances (Figure S2 C). Then we trained a BERT-CRF
163 model called BATTER-TPE, for bacterial transcript 3' ends prediction (Figure S2 D-E). While
164 BATTER-TPE predicts the 3' ends of primary transcripts, it cannot locate potential RUT sites around
165 these 3' ends. To address this, we also developed a rule-based post-processing module, BATTER-
166 RUT, which identifies putative Rho binding sites by searching YC dimers with optimal spacing
167 (Figure S3). See Methods for the implementation details of the data augmentation pipeline,
168 BATTER-TPE, and BATTER-RUT.
169 We further benchmarked the performance of BATTER-TPE and BATTER-RUT, and compared
170 with 4 existing tools on diverse metrics. For primary transcription termination, at a false positive

171 rate (FPR) cutoff of 0.1 per kilobase, stem loop prediction-based methods outperform Rho binding
172 sites prediction-based methods by a large margin, even for RDT associated 3' ends (Figure 2A).
173 This is within expectation, as features like a high local C/G ratio and the presence of regularly
174 spaced YC dimers are pervasive in random sequences. For both RDT- and RIT-associated primary
175 3' ends, BATTER-TPE outperforms transterm-HP and RNIE. As shown in Figure S4 A, stem-loop
176 prediction methods, including BATTER-TPE, capture a subset of non-primary 3' ends in *E. coli*
177 (recall of around 40%) and *B. subtilis* (recall of around 90% for RIT, 75% for RDT), but perform
178 poorly on *M. tuberculosis* and *B. burgdorferi* (recall below 10%). Nonetheless, these stem-loop-
179 based approaches generally yield better or comparable results to RUT site-based prediction
180 methods. This aligns with our observation that folding energy drops near non-primary transcript
181 ends are less pronounced than at primary transcript ends, indicating that non-primary termination
182 events may involve complex, clade-specific properties.

183 The performances of BATTER-TPE and published tools on 3' ends prediction were also
184 evaluated on term-seq data of the human oral microbiome[11]. For species with abundance > 1%
185 (Table S3), BATTER-TPE consistently detects more 3' ends (Figure 2B). This improvement is
186 particularly noticeable for phyla like *Actinobacteria* and *Fusobacteria*.

187 Since RUT site prediction-based methods performed poorly in detecting transcript 3' ends, we
188 focused on evaluating their effectiveness in identifying RUT sites. Given the challenge of defining
189 precise RUT locations and the lack of comprehensive genome-wide annotation[3], we used Rho-
190 dependent regions identified by RNA-seq as positive instances, as these regions are likely enriched
191 for RUT sites. In four species, BATTER-RUT mildly outperformed the other two methods (Figure
192 S4 B) at FPR cutoff of 1/KB. For all 4 species, predictions of BATTER-RUT are enriched
193 downstream of transcript 3' ends, peaking at around 50 nt (Figure S4 C).

194 For genomes that lack experimental data, here we used a previously proposed approach[9]: There
195 should be transcript 3' ends in intergenic regions between convergently transcribed ("tail to tail")
196 gene pairs. In a random subset of ~2000 representative bacteria genomes in GEMs, at an FPR
197 cutoff of 0.1/KB, for all the IGRs flanked by tail-to-tail gene pairs, the fraction of IGRs with predicted
198 3' ends measures model performance (Figure S5 A). BATTER-TPE consistently achieves the best
199 performance across diverse bacterial phylogeny compared to the other two methods.

200 The performance of BATTER-TPE on detecting 3' ends of noncoding RNAs was also evaluated.
201 Assuming the majority of noncoding RNAs with TSS detected by dRNA-seq have terminators, a
202 higher number of predicted terminators located between intergenic TSS and the next CDS would
203 indicate better performance. We curated TSSs from 14 bacteria species that are representative of
204 major bacteria lineages (Table S4) defined by dRNA-seq, and defined intergenic TSSs as TSSs
205 that do not overlap with protein-coding genes, and have distances larger than 128 nt to closest start
206 codons. We annotate such putative 3' ends with transterm-HP, RNIE, and BATTER-TPE at an FPR

207 of 0.1/KB. For *Firmicutes* and *Bacteroidota*, the predictions by the three methods largely overlapped.
208 However, for *Cyanobacteria*, *Proteobacteria*, and *Actinobacteria* species, a considerable portion of
209 such predictions is unique to BATTER-TPE, suggesting its ability to detect terminators overlooked
210 by transterm-HP and RNIE (Figure S5 B).
211
212 **Inter-gene and inter-clade variations of 3' end associated stem loops**
213 We next explored the variations in the stem loop structures associated with transcription 3' ends
214 predicted by BATTER-TPE (only including primary 3' ends) across all the GEMs genomes. As
215 illustrated in Figure 3A, stem loops of *Actinobacteria* are characterized by long stems and often
216 lack downstream U-rich sequences. Stem loops of most *Firmicutes* and *Bacteroidota* show only
217 minor deviations from typical RITs. Stem loops of *Proteobacteria* tend to be shorter and have
218 relatively large intra-phylum variations. When only considering primary 3' ends in the predictions,
219 20% to 50% of CDSs in most bacteria species are predicted to have trailing terminators, and on
220 average, 1 to 5 terminators are predicted in every kilobase of IGR sequences (Figure 3A).
221 The 3' end associated stem loops show inter-gene and inter-species variations. To quantify these,
222 we trained a neural network to project RNA sequences to numeric vectors, using a triplet loss to
223 bring RNAs from the same Rfam family closer in embedding space while pushing RNAs from
224 different families apart (Figure S6 A, see Methods for details). We validated the embeddings with
225 t-SNE visualization on holdout Rfam families (Figure S6 B-C). We then analyzed terminators
226 associated with several widely distributed genes (5 protein-coding gene families and 2 riboswitches)
227 in all 42,905 GEMs representative bacterial genomes, projecting them into embeddings with the
228 neural network encoder. t-SNE visualizations of the terminator embeddings (Figure 3B-C) showed
229 that terminators of homologous genes in closely related species were similar, while in
230 phylogenetically distant species, inter-gene variations were generally smaller than inter-species
231 variations. To assess the contributions of taxonomy (at phylum, class, and order levels) and gene
232 families to variations of terminators, we analyzed each of the top 36 principal components
233 (representing ~80% of the total variance) and averaged the variance explained by each factor. As
234 shown in Figure 3D, taxonomies consistently explained a larger fraction of variance than gene
235 families. Taking cysteine synthase as an example, we clustered neural embeddings of associated
236 stem loops into ten clusters using K-means and visualized the predicted secondary structures of
237 centroid instances (Figure 3E), illustrating such cross-clade diversity of a single gene family.
238
239 **Differential contribution of Rho-dependent termination in different clades**
240 As an application of BATTER, here we sought to quantify the relative contribution of RITs and RDTs
241 to transcription termination in different bacterial lineages. For primary 3' ends predicted by
242 BATTER-TPE in each genome, we used the median length of U-rich sequences associated with

243 these 3' ends as a proxy to measure the contribution of RITs. We also calculated a "Rho
244 dependency score" based on BATTER-RUT's predictions (See Materials and Methods) to measure
245 the contribution of RDTs. The phylogenetic distribution of Rho proteins determined by an
246 established homolog search strategy[35] was also visualized (Figure 4A). To avoid impact from
247 genome incompleteness and potential binning error, only isolate genomes with genome
248 completeness > 95% in GEMs representative gnomes were considered.

249 In agreement with previous findings [35], we found homologs of Rho proteins are present in most
250 of the bacterial lineages, with *Cyanobacteria* and several *Firmicutes* species as exceptions. The
251 presence of Rho proteins and shorter U-rich sequences is strongly associated with higher Rho
252 dependency scores (Figure 4B-C). *Actinobacteria* species have the highest average Rho
253 dependency scores, followed by *Proteobacteria*, *Spirochaetota*, *Bacteroidota*, and *Firmicutes*
254 (Figure 4A, Figure 4D). These findings are concordant with pervasive Rho-dependent terminations
255 in *M. tuberculosis*[36] and a diminished role of Rho protein in transcription termination in *Firmicutes*
256 clade[6]. In most cases, phylogenetically related species tend to have similar Rho dependency
257 scores. *Bartonella* species in the *Rhizobiaceae* family, which have a sharp drop in Rho dependency
258 scores compared to other genera in the same family, is an exception. In *Bartonella* species, a basic
259 amino acid conserved across diverse bacterial lineage (the counterpart of the R87 residue in *E.*
260 *coli* Rho protein) is substituted into glutamine (Figure S7). In *E. coli*, R87 is directly involved in Rho
261 protein-substrate interaction[4], hence we hypothesized this substitution may lead to functional
262 deficiency, driven preference toward usage of RIT in this clade.

263 Taking a closer look at genomes that lack homolog of Rho protein, surprisingly, certain
264 *Cyanobacteria* species have unexpectedly high Rho dependency scores (Figure 4A, Figure 4E).
265 Concurring with that, metagene plots for *Cyanobacteria* model species *Nostoc sp.* PCC 7120
266 suggests the predicted RUT-like sequences are also enriched downstream of predicted 3' ends,
267 resembling the distribution of RUT sites in species with functional Rho homologs, such as *E. coli*
268 and *B. subtilis* (Figure 4F). However, there are notable differences: in *Cyanobacteria*, the RUT-like
269 sequences overlap more extensively with stem loop structures, with metagene density peaks
270 located closer to the predicted 3' ends.

271

272 **A RUT-like signature presents in some *Cyanobacteria* species**

273 We further investigated the phylogenetic distribution of the RUT-like sequences within
274 *Cyanobacteria* clade and their association with traits of corresponding species. Interestingly,
275 *Cyanobacteria* species characterized by ancient traits[37]—such as unicellularity and marine
276 habitats,—tend to lack RUT-like signatures. In contrast, species with more recently evolved traits,
277 such as the nitrogen fixation capability, far-red light photoacclimation (FarLip), and low-light
278 photoacclimation (LoLip)[38], tend to have stronger RUT-like signatures (Figure 5A-B, Table S5).

279 As *Cyanobacteria* genomes lack functional Rho homologs, here we used a fluorescent reporter
280 assay to test the functionality of these RUT-like sequences in *E. coli* (Figure 5C). In a sfGFP-rrnB
281 T1 terminator-mCherry construct, the rrnB T1 terminator sequence was replaced with either
282 predicted RUT-like sequences or a non-terminating sequence from the lacI gene as a negative
283 control. A low mCherry/sfGFP ratio indicates high transcription termination efficiency. 7 predicted
284 RUT site-like sequences associated with nitrogen fixation-related genes in *Nostoc sp.* PCC 7120
285 and 8 predicted RUT site-like sequences associated with FarLip and LoLip-related genes in *C.*
286 *fritschii* PCC 6912 were tested. Among these tested sequences, 7/7 *Nostoc sp.* PCC 7120
287 sequences and 7/8 *C. fritschii* PCC 6912 sequences have lower mCherry/sfGFP ratios compared
288 to the negative control (Figure 5D and 5E), indicating the predicted RUT sequences are at least
289 partially functional for terminating transcription in *E. coli*.
290
291 **Application on identification of AMR genes related premature terminations**
292 As another application of BATTER, we systematically predict premature termination events
293 associated with 5' UTRs of antimicrobial resistance (AMR) genes. Previous studies have suggested
294 premature transcription termination plays an important role in bacterial transcriptional regulation
295 under antibiotic stress [11]. Premature terminations in bacteria could act in *cis* (through attenuation
296 mechanism) or in *trans* (premature transcription product serve as sRNA to regulate target genes[18]
297 or sponge other sRNAs[29]), and in some cases, both mechanisms are simultaneously involved[39].
298 For *cis*-regulation, BATTER-TPE could predict RIT-based premature termination, BATTER-RUT
299 could provide clues for RDT-based premature termination. For *trans*-regulation, as the
300 transcriptional products are functional, their 3' ends are likely to contain protective stem loops, and
301 could be effectively detected by BATTER-TPE. As a result, although BATTER is designed for
302 predicting primary transcription termination, it can also predict premature termination in many cases.
303 We thus predicted AMR genes in GEMs representative bacteria genomes AMRfinder[40] and
304 identified the premature termination sites in their upstream sequences with BATTER-TPE. The
305 prevalence of predicted upstream terminator structures varied substantially across different AMR
306 gene families, being notably frequent in ribosomal protection proteins, 23S rRNA
307 methyltransferases, and efflux transporters, while rare in other AMR families (Figure 6A-B). It was
308 reported the RppA gene (annotated to ABC-F type ribosomal protection protein by AMRfinder) in
309 *Bacillus thuringiensis* BMB171 is regulated by an attenuator-induced premature termination
310 event[41], our analysis suggests similar regulation is conserved in other ribosome protection protein
311 families. It was well characterized that the ermC gene (annotated to 23S rRNA methyltransferase
312 Erm(C) family by AMRfinder) is regulated by a attenuator that shelters ribosomal binding sites for
313 translational regulation[42], here our predictions suggest in other 23S rRNA methyltransferase
314 families, such as Cfr family 23S rRNA methyltransferase and 23S ribosomal RNA

315 methyltransferase Erm, premature termination-based regulation can be more pervasive. In
316 agreement with our prediction, a previous study also suggested premature termination of multiple
317 antibiotic efflux-related genes are suppressed upon antibiotic treatment[11]. An Rfam search
318 suggested the upstream sequence of macrolide efflux MFS transporter Mef(A) associated
319 premature terminator hits to Enterococcus sRNA_1300 (RF02862), a known sRNA related to
320 antibiotic response[22], suggesting premature termination of these genes may generate *trans*-acting
321 sRNA.
322 For the top 15 AMR gene families with the most predicted upstream terminators, we evaluated
323 the conservation of their secondary structures. Sequence covariation analysis supported
324 conserved secondary structures in upstream sequences of translation protection-related AMRs
325 (ribosomal protection proteins and 23S rRNA methyltransferase), but not those related to efflux
326 transporters (Figure 6B and Figure S8). A potential explanation is that most translation-related
327 premature terminations detected here may act in *cis*, in which RNA structures play a more
328 prominent role; while premature terminations of efflux transporters may generate *trans*-acting
329 sRNAs, which are relatively unstructured.
330

331 ## Discussion

332 In this study, by term-seq data analysis, we confirmed a conserved stem loop structure is often
333 associated with transcript 3' ends generated by primary termination events in diverse bacteria
334 species, regardless of the dependency of Rho factor. Based on this finding, we developed a data
335 augmentation pipeline and trained BATTER-TPE, a BERT-CRF model for predicting transcript 3'
336 ends. We also developed an accessory module, BATTER-RUT, for detecting putative RUT sites.
337 Benchmarking BATTER on diverse datasets demonstrated its superior performance compared to
338 existing tools. Using BATTER, we systematically characterized the properties of 3' ends associated
339 stem loops and evaluated the role of Rho-dependent termination across major bacterial lineages.
340 Additionally, we systematically predicted putative transcript 3' ends located upstream of AMR
341 genes, identifying several AMR gene families potentially regulated by premature termination.
342 While BATTER-TPE outperforms existing tools, it does have some limitations. BATTER-TPE is
343 optimized to detect 3' ends from transcription termination events in which the resulting products are
344 functional and hence should be protected from RNase degradation, as in the case of primary
345 transcription termination of protein-coding genes and functional noncoding RNAs. Many secondary
346 termination events, however, do not follow this paradigm. In addition, archaea terminators usually
347 do not contain stem loop structures[43], and BATTER-TPE cannot effectively predict transcript 3'
348 ends in archaea. Thirdly, although BATTER-RUT mildly outperformed existing tools for RUT
349 prediction, and achieved reasonable performance for detecting 3' end associated RUTs in
350 combination with BATTER-TPE, exhaustive genome-wide prediction of RUTs remains challenging.

351 Our systematic prediction of 3' end-associated stem loops revealed that inter-species variations
352 generally exceed inter-genes variations. The RNA polymerase complex, while universally present
353 in bacteria with conserved features such as subunit composition, has evolved clade-specific
354 variations that affect transcription regulation[44]. For instance, β and β' subunits of RNA polymerase
355 contain clade-specific insertions, deletions of these regions lead to various functional
356 deficiencies[45]. Multiple clade-specific transcription factors have also been reported to regulate
357 bacteria transcription[44]. Assuming sequences and structures of 3' end-associated stem loops are
358 optimized for stalling the RNA polymerase complex, inter-species variations of terminators may
359 reflect co-evolutions between terminators and the transcription machinery.
360 The discovery of RUT-like signatures near transcript 3' ends in certain *Cyanobacteria* species is
361 very intriguing. One potential explanation is that those *Cyanobacteria* species may harbor an
362 undetected Rho homolog. It has been reported that in *Arabidopsis thaliana*, RHON1, a Rho-like
363 protein, is responsible for some transcription termination events in plastids[46]. It was well accepted
364 that plastids in plants evolve from endosymbiotic *Cyanobacteria*[47]; hence some *Cyanobacteria*
365 may harbor a Rho-like protein homolog. Following a previously proposed strategy[35], in this study,
366 proteins with both Rho RNA binding domain and ATPase domains are considered as functional
367 Rho, and no functional Rho was detected in *Cyanobacteria* clade. We could not completely exclude
368 the probability that some *Cyanobacteria* species may harbor homologs of Rho protein, as if the
369 sequence divergence is too large, Rho protein homologs may elude detection. An alternative
370 hypothesis is that Rho protein was lost in the common ancestor of existing *Cyanobacteria*, in some
371 descendants, other protein factors may have adapted to use the original Rho binding sites for
372 transcription termination, while in others, RUT sites may have been lost entirely. This may explain
373 the association found between RUT-like signatures and recently evolved traits, such as nitrogen
374 fixation, FarLip, and LoLip. Both two hypotheses require further experimental validation.
375 BATTER-TPE is primarily optimized for predicting 3' ends of the primary transcript, we have
376 shown in practice it can also be applied to predict a considerable fraction of premature termination
377 events, including termination events that generate functional products (such as trans-acing sRNAs
378 generated by both RITs and RDTs), and *cis*-regulatory RITs. Although BATTER-RUT could detect
379 some RUT sites in leader sequences, comprehensive prediction of *cis*-regulatory premature RDTs
380 remains an open problem. Despite such limitations, BATTER successfully detects variants of
381 known and novel riboregulators and provides insights for the clade-specific regulation of RDTs.
382 These riboregulators can serve as potential drug targets for antibiotic-resistant strains.
383 As BATTER is optimized for 3' ends of functional transcript in very diverse bacteria species,
384 taking genome sequences as the solitary input, it is particularly useful for discovering 3' ends
385 functional transcripts in contigs assembled from metagenome samples, where very diverse
386 bacteria lineages, including uncultivated ones, are often present. We hope BATTER will be a useful

387 tool for understanding the diversity of noncoding RNAs across the bacteria domain.

388 # Methods

389 **Curation of experimentally supported 3' ends**

390 Known transcript 3' ends with experimental supports used in this work were curated from previous
391 studies (Table S1). The transcript 3' ends of the metaterm-seq dataset[11], which are not provided
392 by the original publication, were identified by a customized pipeline. In brief, metaterm-seq reads
393 and RNA-seq reads of 2 human oral metagenome samples were downloaded from PRJEB12568.
394 Metaphlan2[48] was used for species profiling, species with abundance > 1% in at least one sample
395 (either RNA-seq or term-seq) were considered for downstream analysis. Reads were then mapped
396 to corresponding reference genomes downloaded from genbank (Table S3). 5' end coverage of
397 the reads was summarized with HTSeq[49]. Local max positions within 50 nt windows covered by at
398 least 6 reads are considered as transcripts 3' ends with experimental support.

399 **Identification of Rho-dependent regions**

400 The RNA-seq reads of *E. coli* (GSE41939), *B. subtilis* (GSE195579), and *B. burgdorferi*
401 (GSE222085) were mapped to the corresponding reference genomes (see Table S2) with
402 bowtie2[50] using the default parameters. For *M. tuberculosis* data (E-MTAB-11753), bam files were
403 downloaded from http://ftp.ebi.ac.uk/biostudies/fire/E-MTAB-/753/E-MTAB-11753/Files. The
404 identification of Rho-dependent regions was implemented using a customized python script. In brief,
405 using a stride of 50 nt, for each considered genome location, the numbers of RNA-seq fragments
406 assigned within upstream 200 nt ($U_{200}$) and downstream 200 nt ($D_{200}$) were summarized, replicates
407 were combined by summation, and the ratio $D_{200}/U_{200}$ measures the level of read-through at this
408 location. Single-side Fisher exact test was used to evaluate whether the read-through level in Rho
409 protein inhibition group was greater than the control group. For *B. subtilis*, data of WT and Rho
410 knockout strains in the exponential growth stage were considered. For *M. tuberculosis*, samples at
411 6h after Rho protein depletion were used. Regions with odds ratio > 3 and P-value < 0.001 were
412 classified as Rho-dependent regions, those with odds ratio < 1 were classified as Rho-independent
413 regions, and the remaining ones were classified as "others". 3' ends that overlap with Rho-
414 dependent regions, Rho-independent regions, and "others" are categorized as RDT, RIT, and
415 unclassified, respectively. The folding energy of upstream 45 nt, and downstream 5 nt sequences
416 flanking annotated 3' ends were evaluated with RNAfold utility in ViennaRNA package[51].

417 **Implementation and benchmarking of BATTER algorithm**

418 See Supplementary Information.

419 **Calculation of Rho dependency scores**

420 To quantify the contribution of Rho-dependent termination, for BATTER-TPE predicted primary
421 transcript 3' ends at an FPR cutoff of 0.1/KB, the upstream 100nt and downstream 300nt
422 sequences are extracted and scanned by BATTER-TPE with default parameters. For each 3' ends,

423 the best scoring RUT site is extracted. If no RUT site is predicted, a score of 0 is assigned. The
424 average score across all primary 3' ends was utilized as a measure of Rho dependency.

425 ***Cyanobacteria* traits analysis**

426 For *Cyanobacteria* traits analysis, the trait data was curated based on several previous studies[35,
427 52, 53]. *Cyanobacteria* genomes with PsaA2, PsaB2, and PsbA4 gene orthologs were annotated as
428 FarLip, genomes with apcD4, apcB3, and isiX gene orthologs were annotated as LoLip. Species
429 with complete genomes were considered. The species tree is built with FastTree[54] using multiple
430 sequence alignment of 120 universal marker genes defined by gtdb-tk[55].

431 **Fluorescence reporter assay**

432 All cloning was performed in *E. coli* DH5α strain. Transformed cells were cultured in LB medium
433 and on 1.5% (w/v) LB agar plates supplemented with 15 μg/mL chloramphenicol as required.
434 Sequences encoding sfGFP, the rrnB T1 terminator, and mCherry were assembled into a vector
435 backbone containing the replication origin of pSC101 to yield recombinant plasmid pXG10SFM. To
436 generate the negative control plasmid pXG10SFM0, the rrnB T1 terminator of pXG10SFM was
437 replaced with a 96 bp-region from the lacI coding sequence. To construct plasmids containing
438 predicted RUT sites from nitrogen-fixation gene cluster in cyanobacterium *Nostoc sp.* PCC 7120
439 and FaRLiP and LoLiP gene clusters in *Chlorogloeopsis fritschii* PCC 6912, pXG10SFM was first
440 digested with XhoI and KpnI (Thermo Fisher Scientific). RUT site-like sequences were generated
441 by PCR using two synthesized oligonucleotide primers, and the amplicons were assembled into
442 the digested pXG10SFM, followed by transformation into *E. coli* DH5α. Transformants were verified
443 by DNA sequencing. The recombinant plasmids and primers used in this study are listed in Table
444 S6 and Table S7.

445 To measure fluorescence in *E. coli*, transformants were first inoculated into 1 mL LB medium
446 supplemented with 15 μg/mL chloramphenicol in 24 well-plates and grown overnight at 37°C with
447 constant shaking at 200 rpm. Cells were washed and resuspended in PBS buffer (10 mM
448 Phosphate, 137 mM NaCl, 2.7 mM KCl, pH=7.4). To initiate the assay, samples were diluted 1:10
449 into a 96-well Black/Clear bottom microplate to a final volume of 200 μL. Culture density (OD600)
450 and fluorescence were measured by multi-mode microplate reader Cytation3 (BioTek).
451 Fluorescence intensity of sfGFP and mCherry were read at excitation/emission wavelengths 488
452 nm/515 nm and 579 nm/616 nm, respectively.

453 **Scanning premature terminators of AMR genes**

454 AMR genes in GEMs representative genomes were annotated with AMRfinder[40]. 800 nt
455 sequences (upstream 400 nt and downstream 400 nt) flanking start codons of detected AMR genes
456 are subjected to terminator scanning. Predicted terminators located within upstream 128 nt and
457 downstream 16 nt of start codons are retained. These upstream 128 nt and downstream 16 nt
458 terminator-containing sequences were grouped by AMR families and searched for conserved

motifs with CMfinder[33]. The secondary structural covariations of resulting motifs were evaluated with R-scape[34].

## Code availability

BATTER is deposited at https://github.com/lulab/BATTER. Codes and data for model benchmarking and comparative analysis of bacteria transcript 3' ends are available from https://github.com/uaauaguga/terminator-prediction-scripts.


## Acknowledgments

This work is supported by National Key RD Program of China (2022YFA1304200), 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.


## 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. Nadiras, C., et al., *A multivariate prediction model for Rho-dependent termination of transcription.* Nucleic Acids Research, 2018. **46**(16): p. 8245-8260.
4. Molodtsov, V., et al., *Structural basis of Rho-dependent transcription termination.* Nature, 2023. **614**(7947): p. 367-374.
5. 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.
6. Johnson, G.E., et al., *Functionally uncoupled transcription–translation in Bacillus subtilis.* Nature, 2020. **585**(7823): p. 124-128.
7. Gardner, P.P., et al., *RNIE: genome-wide prediction of bacterial intrinsic terminators.* Nucleic Acids Research, 2011. **39**(14): p. 5845-5852.
8. de Hoon, M.J.L., et al., *Prediction of Transcriptional Terminators in Bacillus subtilis and Related Species.* PLOS Computational Biology, 2005. **1**(3): p. e25.
9. 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.
10. Di Salvo, M., et al., *RhoTermPredict: an algorithm for predicting Rho-dependent transcription terminators based on Escherichia coli, Bacillus subtilis and Salmonella enterica databases.* BMC Bioinformatics, 2019. **20**(1): p. 117.
11. Dar, D., et al., *Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria.* Science, 2016. **352**(6282): p. aad9822.

12. Lalanne, J.-B., et al., *Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry.* Cell, 2018. **173**(3): p. 749-761.e38.
13. 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.
14. Bidnenko, V., et al., *Termination factor Rho mediates transcriptional reprogramming of Bacillus subtilis stationary phase.* PLOS Genetics, 2023. **19**(2): p. e1010618.
15. D'Halluin, A., et al., *Premature termination of transcription is shaped by Rho and translated uORFS in Mycobacterium tuberculosis.* iScience, 2023. **26**(4): p. 106465.
16. Peters, J.M., et al., *Rho and NusG suppress pervasive antisense transcription in Escherichia coli.* Genes & Development, 2012. **26**(23): p. 2621-2633.
17. Petroni, E., et al., *Extensive diversity in RNA termination and regulation revealed by transcriptome mapping for the Lyme pathogen Borrelia burgdorferi.* Nature Communications, 2023. **14**(1): p. 3931.
18. 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.
19. Lee, Y., et al., *The Transcription Unit Architecture of Streptomyces lividans TK24.* Frontiers in Microbiology, 2019. **10**.
20. 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.
21. 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.
22. Sinel, C., et al., *Small RNAs in vancomycin-resistant Enterococcus faecium involved in daptomycin response and resistance.* Scientific Reports, 2017. **7**(1): p. 11067.
23. 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.
24. Jeong, Y., et al., *Multi-Omic Analyses Reveal Habitat Adaptation of Marine Cyanobacterium Synechocystis sp. PCC 7338.* Frontiers in Microbiology, 2021. **12**.
25. 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.
26. Forquet, R., et al., *Mapping the Complex Transcriptional Landscape of the Phytopathogenic Bacterium Dickeya dadantii.* mBio, 2022. **13**(3): p. e00524-22.
27. Mandell, Z.F., et al., *Comprehensive transcription terminator atlas for Bacillus subtilis.* Nature Microbiology, 2022. **7**(11): p. 1918-1931.
28. 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.
29. Adams, P.P., et al., *Regulatory roles of Escherichia coli 5' UTR and ORF-internal RNAs detected by 3' end mapping.* eLife, 2021. **10**: p. e62438.
30. Bar, A., et al., *TRS: a method for determining transcript termini from RNAtag-seq sequencing data.* Nature Communications, 2023. **14**(1): p. 7843.
31. Ren, G.-X., X.-P. Guo, and Y.-C. Sun, *Regulatory 3′ Untranslated Regions of Bacterial mRNAs.* Frontiers in Microbiology, 2017. **8**.

547 32. Nayfach, S., et al., *A genomic catalog of Earth's microbiomes.* Nature Biotechnology,
548 2021. **39**(4): p. 499-509.
549 33. Yao, Z., Z. Weinberg, and W.L. Ruzzo, *CMfinder—a covariance model based RNA motif*
550 *finding algorithm.* Bioinformatics, 2006. **22**(4): p. 445-452.
551 34. Rivas, E., J. Clements, and S.R. Eddy, *A statistical test for conserved RNA structure shows*
552 *lack of evidence for structure in lncRNAs.* Nature Methods, 2017. **14**(1): p. 45-48.
553 35. Moreira, Sofia M., et al., *Diversification of the Rho transcription termination factor in*
554 *bacteria.* Nucleic Acids Research, 2024. **52**(15): p. 8979-8997.
555 36. Ahmad, E., et al., *Rho-dependent transcription termination is the dominant mechanism*
556 *in Mycobacterium tuberculosis.* Biochimica et Biophysica Acta (BBA) - Gene Regulatory
557 Mechanisms, 2023. **1866**(2): p. 194923.
558 37. Hammerschmidt, K., et al., *The Order of Trait Emergence in the Evolution of*
559 *Cyanobacterial Multicellularity.* Genome Biology and Evolution, 2021. **13**(2): p. evaa249.
560 38. Gan, F. and D.A. Bryant, *Adaptive and acclimative responses of cyanobacteria to far-red*
561 *light.* Environmental Microbiology, 2015. **17**(10): p. 3450-3465.
562 39. Loh, E., et al., *A trans-Acting Riboswitch Controls Expression of the Virulence Regulator*
563 *PrfA in Listeria monocytogenes.* Cell, 2009. **139**(4): p. 770-779.
564 40. Feldgarden, M., et al., *Validating the AMRFinder Tool and Resistance Gene Database by*
565 *Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of*
566 *Isolates.* Antimicrobial Agents and Chemotherapy, 2019. **63**(11): p. 10.1128/aac.00483-
567 19.
568 41. Cai, X., et al., *Expression of ribosomal protection protein RppA is regulated by a*
569 *ribosome-dependent ribo-regulator and two mistranslation products.* Environmental
570 Microbiology, 2021. **23**(2): p. 696-712.
571 42. Koch, M., et al., *Critical 23S rRNA interactions for macrolide-dependent ribosome stalling*
572 *on the ErmCL nascent peptide chain.* Nucleic Acids Research, 2017. **45**(11): p. 6717-6728.
573 43. Maier, L.-K. and A. Marchfelder, *It's all about the T: transcription termination in archaea.*
574 Biochemical Society Transactions, 2019. **47**(1): p. 461-468.
575 44. Chen, J., H. Boyaci, and E.A. Campbell, *Diverse and unified mechanisms of transcription*
576 *initiation in bacteria.* Nature Reviews Microbiology, 2021. **19**(2): p. 95-109.
577 45. Chlenov, M., et al., *Structure and Function of Lineage-specific Sequence Insertions in the*
578 *Bacterial RNA Polymerase β′ Subunit.* Journal of Molecular Biology, 2005. **353**(1): p.
579 138-154.
580 46. Chi, W., et al., *RHON1 Mediates a Rho-Like Activity for Transcription Termination in*
581 *Plastids of Arabidopsis thaliana* The Plant Cell, 2014. **26**(12): p. 4918-4932.
582 47. McFadden, G.I., *Origin and Evolution of Plastids and Photosynthesis in Eukaryotes.* Cold
583 Spring Harbor Perspectives in Biology, 2014. **6**(4): p. a016105.
584 48. Truong, D.T., et al., *MetaPhlAn2 for enhanced metagenomic taxonomic profiling.* Nature
585 Methods, 2015. **12**(10): p. 902-903.
586 49. Anders, S., P.T. Pyl, and W. Huber, *HTSeq—a Python framework to work with high-*
587 *throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-169.
588 50. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature
589 Methods, 2012. **9**(4): p. 357-359.
590 51. Lorenz, R., et al., *ViennaRNA Package 2.0.* Algorithms for Molecular Biology, 2011. **6**(1):
591 p. 26.

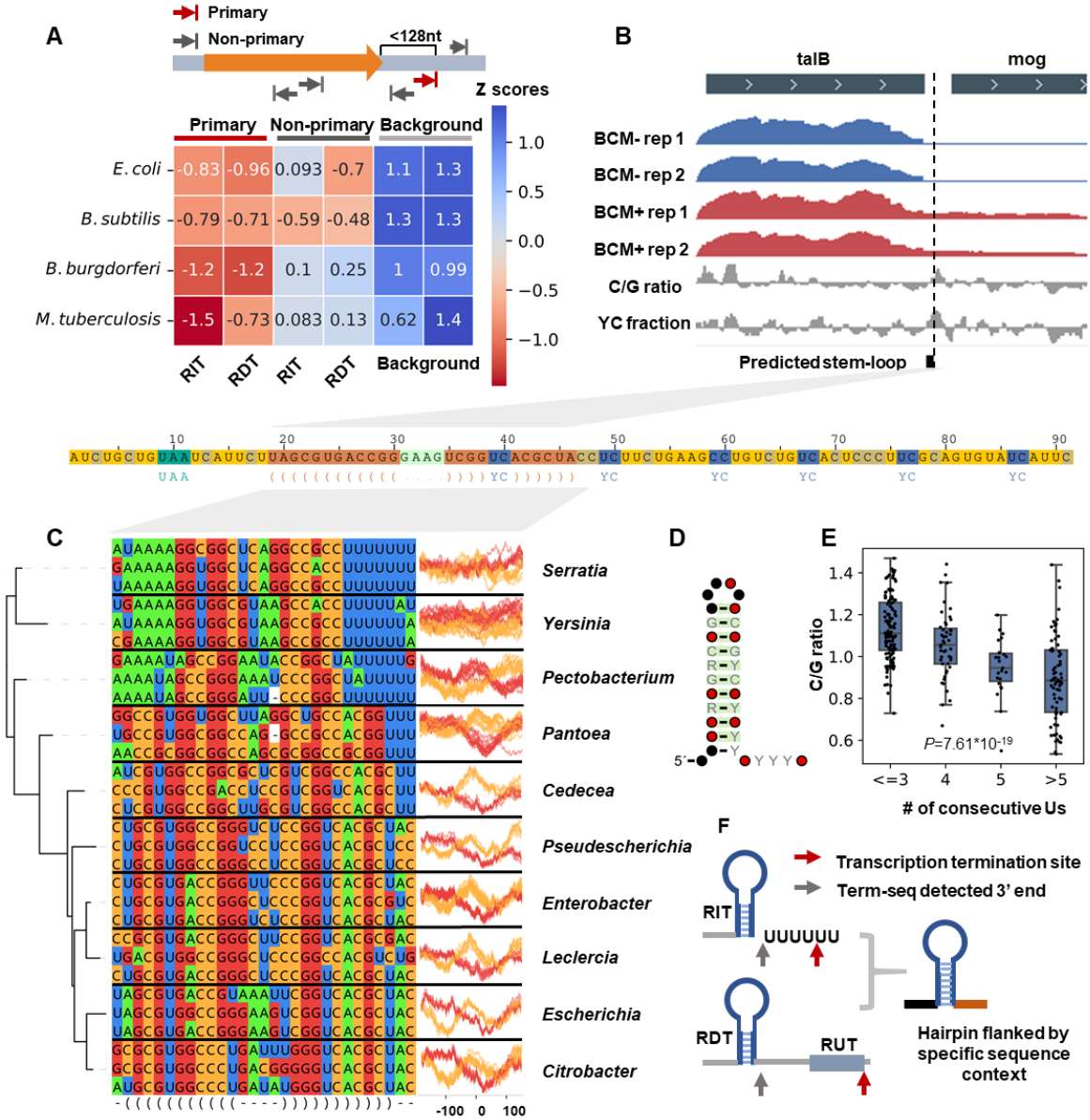
592 52. Chen, M.-Y., et al., *Phylogenomics Uncovers Evolutionary Trajectory of Nitrogen Fixation*
593 *in Cyanobacteria.* Molecular Biology and Evolution, 2022. **39**(9): p. msac171.
594 53. Shih, P.M., et al., *Improving the coverage of the cyanobacterial phylum using diversity-*
595 *driven genome sequencing.* Proceedings of the National Academy of Sciences, 2013.
596 **110**(3): p. 1053-1058.
597 54. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2 – Approximately Maximum-Likelihood*
598 *Trees for Large Alignments.* PLOS ONE, 2010. **5**(3): p. e9490.
599 55. Chaumeil, P.-A., et al., *GTDB-Tk: a toolkit to classify genomes with the Genome*
600 *Taxonomy Database.* Bioinformatics, 2020. **36**(6): p. 1925-1927.

601

602

# Figures

**Figure 1.** A conserved stem loop is associated with transcript 3' end regardless of the Rho factor dependency



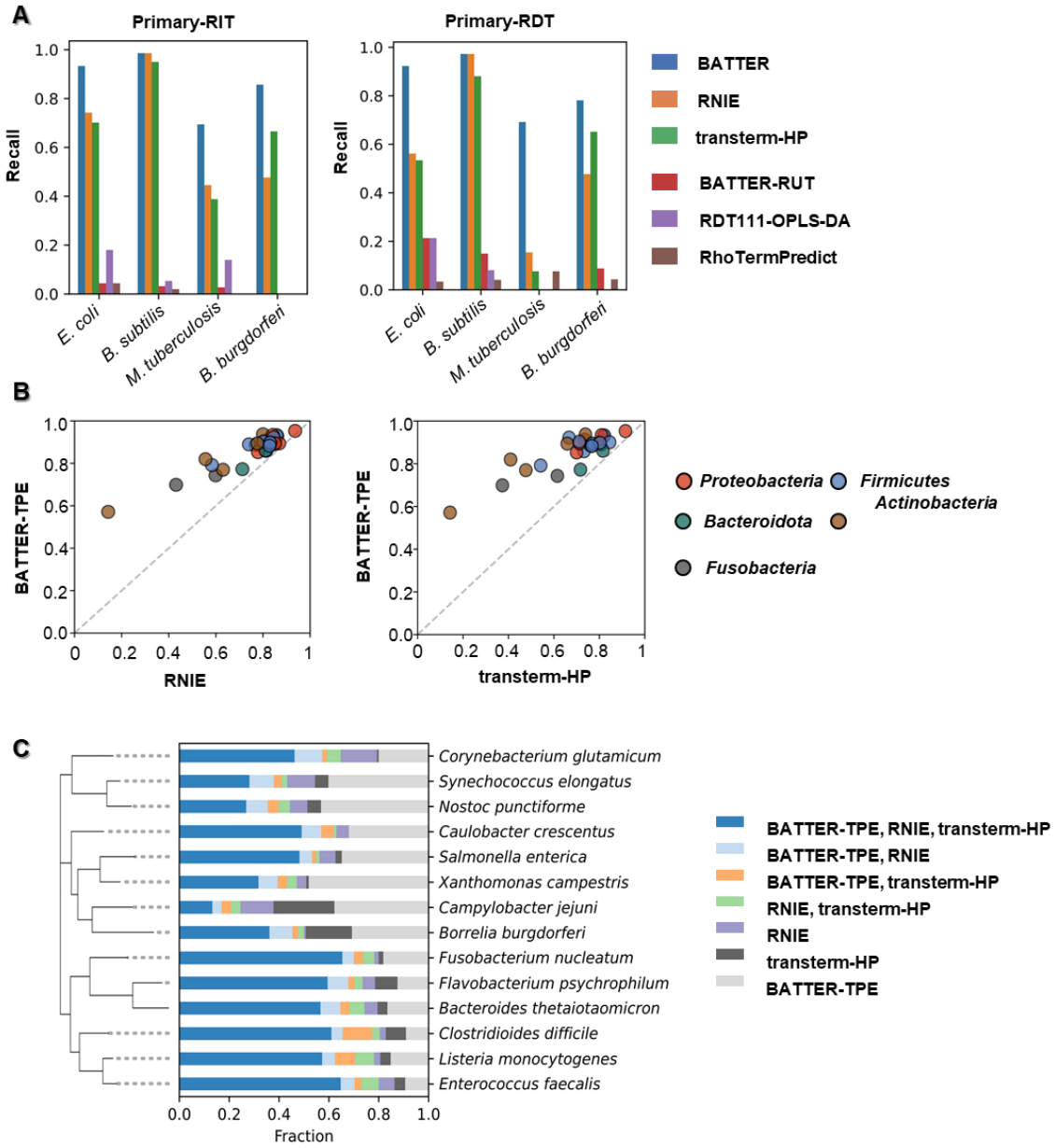

A. Z scores of average folding energy in 4 distantly related species, stratified by Rho factor dependency (RITs and RDTs) and genomic context (primary 3' ends, non-primary 3' ends, and background)

B. RNA-seq coverage near the 3' end of E. coli talB under BCM treatment (BCM+ rep 1, BCM+ rep 2) and control (BCM- rep 1, BCM- rep 2) conditions. Local C/G ratio and density of YC dimers are also shown.

C. Multiple sequence alignment (MSA) of the conserved stem loop structure in downstream sequences of talB gene 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, guanidine(G) ratios and cytosine (C) ratios (labeled in red and orange,

617 respectively) of 101 nt sequences centered at each nucleotide from 128 nt upstream to 128 nt
618 downstream relative to ends of the stem loops are shown.
619 D. Secondary structure covariation of the identified stem loop. Base pairs with R-scape E value <
620 0.05 are labeled in green, and nucleotides with sequence identity > 97% and > 90% are labeled in
621 red and black, respectively. R indicates purine, Y indicates pyrimidine.
622 E. Correlation between maximum lengths of consecutive uracils (Us) and C/G ratio within 128 nt
623 sequences flanking ends of the stem loops downstream talB in all considered Enterobacteriaceae
624 genomes.
625 F. Schematic illustration of common and distinctive features of RITs and RDTs. For RITs, term-seq
626 sites indicate the location of terminator stem loops, while for RDTs, term-seq sites indicate stem
627 loops upstream of real termination sites. Both RITs and RDTs are characterized by stem loops
628 flanked by particular sequence contexts.
629

**Figure 2. Performance evaluation of the 3' end prediction model**

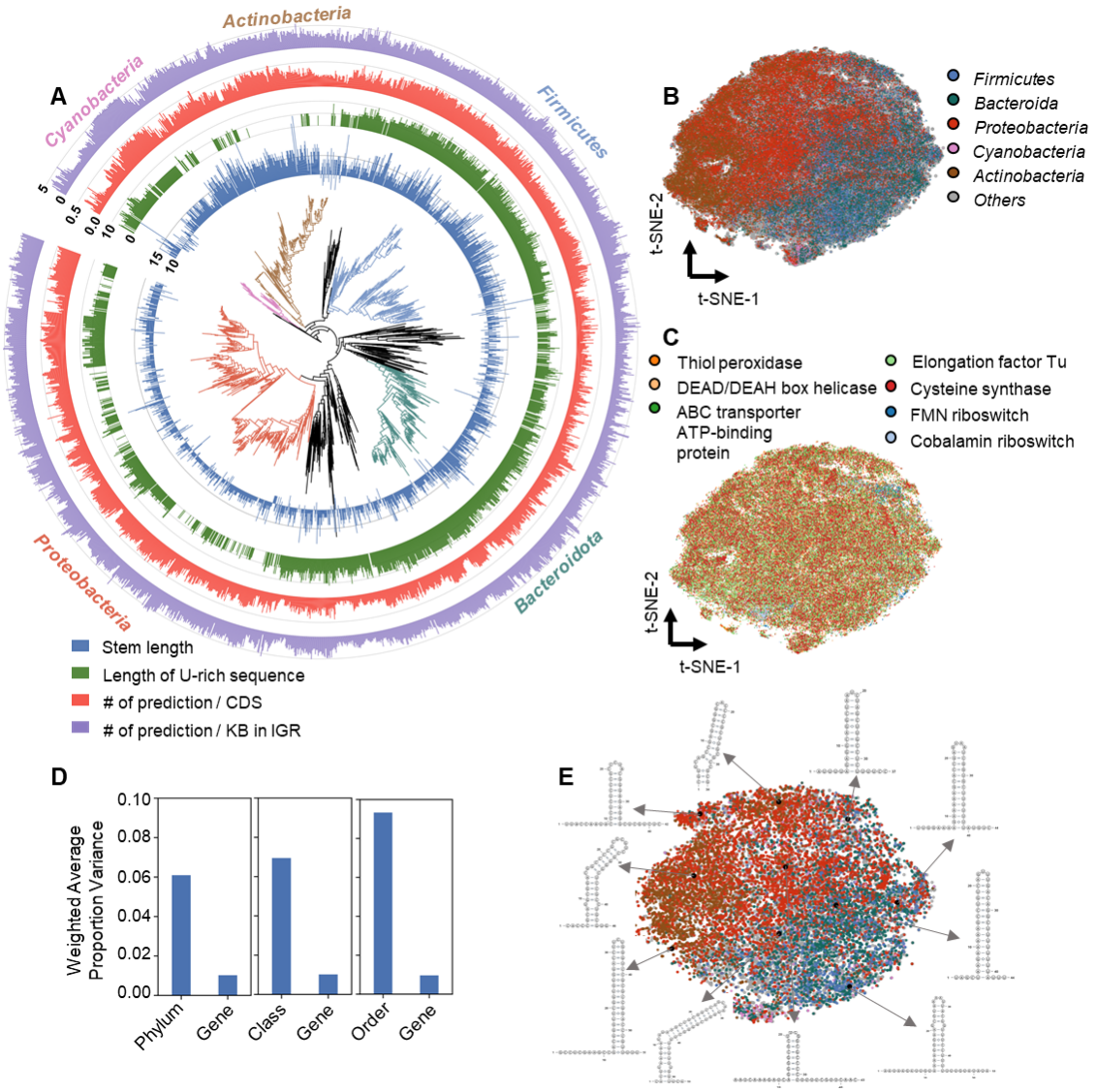



A. Recalls of stem loop prediction-based methods (BATTER-TPE, RNIE, and transterm-HP) and RUT sites prediction-based methods (BATTER-RUT, RDT111-OPLS-DA and RhoTermPredict) at FPR of 0.1/KB.

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

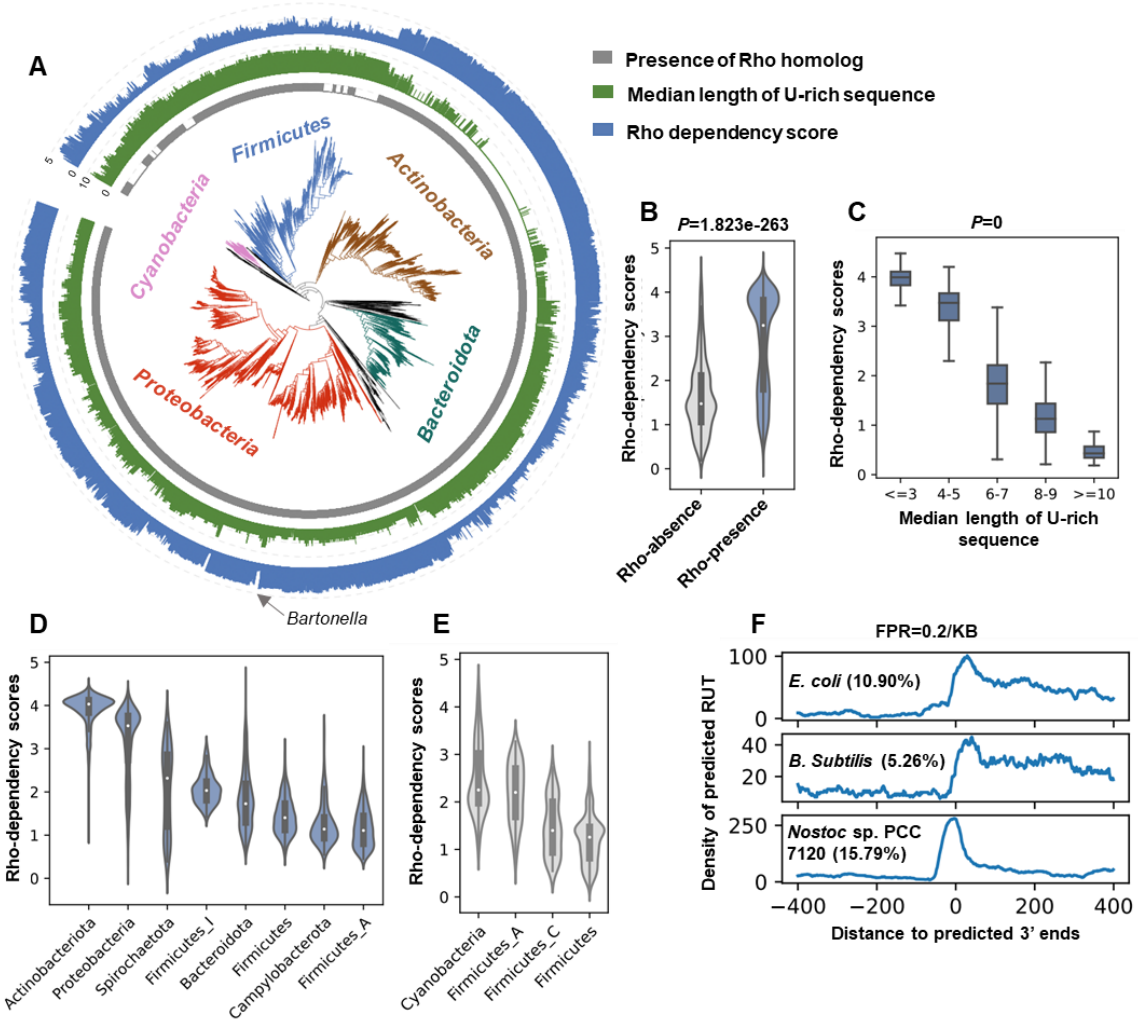
C. Overlaps between RNIE, BATTER, and transterm-HP's predictions on curated differential RNA-seq datasets.

641 **Figure 3.** Characterization of 3' ends associated stem loops across diverse bacteria species



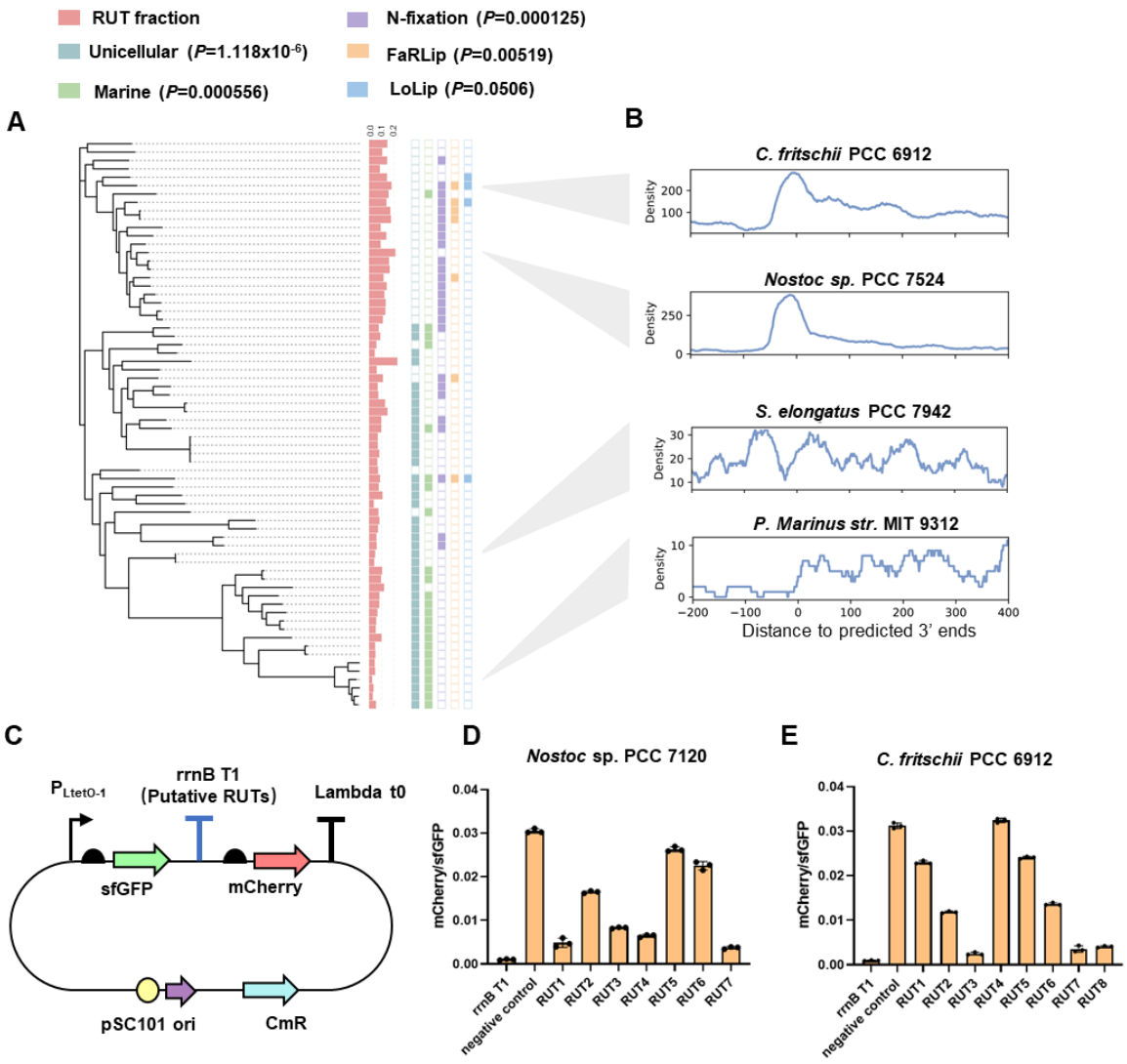

642
643 A. Characteristics of predicted terminators on different bacteria species. A subset of ~2000 species
644 were sampled for visualization. From the innermost cycle to the outermost cycle, the stem lengths
645 (blue), U-rich sequence lengths (green), number of prediction/CDS (red), and number of
646 predictions/KB (purple) are shown.
647 B. t-SNE (t-distributed stochastic neighbor embedding) on vector embeddings of terminators
648 downstream several widely distributed genes, coloring by phyla.
649 C. t-SNE visualization same as Figure 4B, coloring by the gene families
650 D. Weighted average variance proportion of top 36 principal components explained by different
651 genes and different clades at the taxonomy levels of phylum, class, and order.
652 E. Visualization of terminators associated with cysteine synthase CDS. The vector embeddings
653 were clustered into 10 clusters with K-means, and the predicted secondary structures of cluster
654 centers (black dot) were shown.
655

**Figure 4.** Evaluation of Rho dependency in diverse bacteria lineages



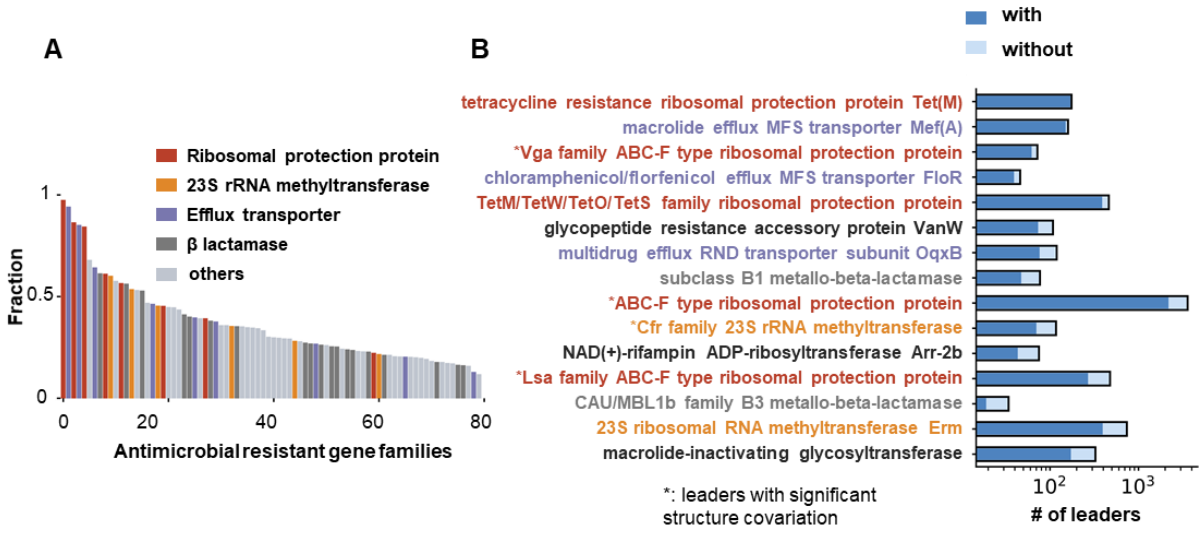

A. The association between the presence of Rho protein homolog (innermost gray circle), the median length of U-rich sequences (green), and the Rho dependency score (outer circle in blue).
B. Rho dependency score distribution in species with and without Rho protein homologs. *P*-value was determined by Wilcoxon test.
C. The correlation between median length of U-rich sequences and Rho dependency scores. P-value was determined by Kruskal-Wallis test.
D-E. Rho dependency score distribution in species with and without Rho protein homologs, stratified by phyla.
F. Metagene plot for BATTER-RUT predicted RUT sites (FPR=0.2/KB) near 3' ends predicted by BATTER-TPE.

**Figure 5.** The presence of RUT-like sequences correlates with recently evolved traits in *Cyanobacteria*




A. Association between Rho dependency scores and *Cyanobacteria* traits. Wilcoxon test *P*-values were shown in the brackets.
B. Metagene plot for predicted RUT sites near predicted transcript 3' ends at FPR=0.2/KB. Four representative *Cyanobacteria* species are shown.
C. Diagram of the reporter plasmid for fluorescence assay.
D. mCherry/sfGFP ratio in tested *Nostoc sp.* PCC 7120 RUT sites-like sequences.
E. mCherry/sfGFP ratio in tested *C. fritschii* PCC 6912 RUT sites-like sequences. In D and E, error bars show the mean ± SEM of 3 biological replicates.

681 **Figure 6.** Terminator prediction reveal pervasive riboregulations in some antibiotic resistance
682 gene families




683

684 A. Antimicrobial resistance (AMR) gene families are ranked by fractions with predicted upstream
685 terminators, each bar indicates an AMR gene family. Gene families that share similar
686 mechanisms of action are labeled with the same color.
687 B. Fraction of leader sequence preceding AMR gene families that are frequently associated with
688 predicted transcription terminators.