TaRTLEt: Transcriptionally-active Riboswitch Tracer Leveraging Edge deTection

Sachit Kshatriya[1] and Sarah C. Bagby[1]

[1]**Case Western Reserve University Department of Biology**

Corresponding author:

Sarah C. Bagby[1]

Email address: sarah.bagby@case.edu

ABSTRACT

Structured RNAs have emerged as a major component of cellular regulatory systems, but their mechanism of action is often poorly understood. Riboswitches are structured RNAs that allosterically regulate gene expression through any of several different mechanisms. *In vitro* approaches to characterizing mechanism are costly, low-throughput, and must be repeated for each individual riboswitch locus of interest. Bioinformatic methods promise higher throughput; despite robust computational identification of riboswitches, however, computational classification of riboswitch mechanism has so far been both model-bound, relying on identification of sequence motifs known to be required for specific models of riboswitch activity, and empirically untested, with predictions far outpacing biological validation. Here, we introduce `TaRTLEt` (Transcriptionally-active Riboswitch Tracer Leveraging Edge deTection), a new high-throughput tool that recovers *in vivo* patterns of riboswitch-mediated transcription termination from paired-end RNA-seq data using edge detection methods. `TaRTLEt` successfully extracts transcription termination signals despite numerous sources of biological and technical noise. We tested the effectiveness of `TaRTLEt` on riboswitches identified from a wide range of sequenced bacterial taxa by utilizing publicly available paired-end RNA-seq readsets, finding broad agreement with previously published *in vitro* characterization results. In addition, we use `TaRTLEt` to infer the *in vivo* regulatory mechanism of uncharacterized riboswitch loci from existing public data. `TaRTLEt` is available on GitHub and can be applied to paired-end RNA-seq datasets from isolates or complex communities.

INTRODUCTION

Riboswitches are highly conserved, widespread gene regulatory elements that exploit RNA's potential for complex structure to build ligand-responsive differential regulation of gene expression directly into an mRNA. Encoded in the untranslated regions (UTRs) of genes, riboswitches are highly specific aptamers for a wide variety of small-molecule ligands. In the two decades since the discovery of the first riboswitch (Mironov et al., 2002), we have learned that riboswitches are compact, specific, and responsive to transient changes; that they often regulate critical organism functions; and that they operate via diverse and complex modes (Mandal et al., 2003; Amadei et al., 2023; Sudarsan et al., 2005). Given these properties, riboswitches have been investigated for uses ranging from synthetic sensors (Wachsmuth et al., 2013; Boussebayle et al., 2019), to logic elements in complex gene expression systems (Hanson et al., 2003; Groher and Suess, 2014), to potential targets for antimicrobials and novel therapeutics (Blount and Breaker, 2006; Ellinger et al., 2023; Giarimoglou et al., 2022). Yet our understanding of natural riboswitch function remains patchy, with high-confidence computational identification of candidate riboswitch sequences and ligands (Nawrocki, 2014; Nawrocki and Eddy, 2013a; Chang et al., 2009; Eddy and Durbin, 1994; Yao et al., 2007; Stav et al., 2019) far outpacing our understanding of their activity *in vivo*.

This gap is widened by the breadth of riboswitches' regulatory repertoire. Biochemical methods such as transcript fragment length and ribosome binding assays (Winkler et al., 2003; Nou and Kadner, 2000; Welz and Breaker, 2007; Hollands et al., 2012) have identified an overall pattern of riboswitch regulation: riboswitch-mediated allosteric regulation of expression begins with selective binding of a ligand to the non-coding aptamer domain, stabilizing the domain's conformation (Gilbert et al., 2006) and altering the structure of the expression platform. But this broad pattern includes riboswitches that affect any of several different levels of gene expression. Most straightforwardly, riboswitch ligand binding can sequester or expose conventional expression control logic elements like intrinsic terminators/antiterminators (Mironov et al., 2002; Mandal and Breaker, 2004) or rho-binding sites (Hollands et al., 2012) to alter mRNA production, or translation start sites to alter protein synthesis (Breaker, 2018). However, riboswitch mechanisms can also go beyond interaction with direct control architectures. TPP riboswitches in the fungus *Neurospora crassa* alter gene expression by controlling mRNA splicing (Cheah et al., 2007); the *glmS* riboswitch-ribozyme in *Bacillus subtilis* recruits exonucleases for transcript degradation by catalyzing upstream mRNA cleavage (Collins et al., 2007; Klein and Ferré-D'Amaré, 2006); the *Escherichia coli lysC* riboswitch both sequesters the Shine-Dalgarno (SD) sequence and exposes an RNase E binding site (Caron et al., 2012); and a SAM riboswitch in *Listeria monocytogenes* is capable of acting in *trans* on a distal target (Loh et al., 2009), among numerous other examples documented to date (Bédard et al., 2020; Ariza-Mateos et al., 2021).

To understand the role riboswitches play in a given microbe's physiology, we need to understand at what regulatory level each acts. Regulatory mechanisms have been established for at least one member of the riboswitch families binding roughly a dozen ligands, so that in principle we might expect to infer the mechanism used by other members of the same families. But even in the small set of families with ≥ 2 biochemically characterized members, we find cases where different members of a family act by different mechanisms (Barrick and Breaker, 2007). Thus, robust identification of riboswitch ligands is insufficient to predict regulatory mechanism even for comparatively well-studied riboswitch families; while for many more families, no representative riboswitch has yet been characterized *in vitro*. *In silico* methods seeking to fill this gap have largely relied on DNA-level signals, using the computation of folding energies (Gong et al., 2017) coupled with the identification of SD sequences, U-rich terminator motifs, or rho-binding sites to determine states that could alter transcription/translation efficiency (Barrick and Breaker, 2007; Sun and Rodionov, 2014). While these analytical methods can produce *a priori* predictions about riboswitch regulatory modes, they leave unexplored a more direct readout of the realized activity of riboswitches within living cells: the distribution of fragments captured in RNA-seq data.

Here, we describe the new tool `TaRTLEt` (Transcriptionally-active Riboswitch Tracer Leveraging Edge deTection), which applies high-throughput computational approaches to RNA-seq data to determine which of the known riboswitches identified in a (meta)omic dataset show evidence of regulating gene expression by altering transcription termination efficiency. We hypothesized that different experimental conditions would, directly or indirectly, evoke changes in riboswitch ligand concentrations that might alter riboswitch regulatory state. Under this hypothesis, we predicted that, when the range of conditions tested included both above- and below-threshold ligand concentrations, transcription-attenuating riboswitches would produce a distinctive coverage pattern at riboswitch loci in paired-end RNA-seq datasets; and, further, that this pattern could be robustly identified using computational approaches developed for edge detection (Canny, 1986). Importantly, this pattern should support distinguishing riboswitch-mediated transcription termination from the broader set of changes in gene expression that could be evoked by other transcriptional regulators (e.g., transcription factor proteins) acting at these loci. We show that the occurrence of this pattern at riboswitch loci identified in publicly available transcriptomic datasets is in good agreement with existing *in vitro* characterization of riboswitch regulatory modes. By extracting information from the traces of riboswitch activity that are left in the RNA pool of cells actively regulating gene expression, our approach complements and extends earlier methods based on inference from DNA sequence. Our high-throughput approach can identify signatures of riboswitch-mediated transcription termination in existing or new (meta)transcriptomic data, greatly expanding not only the set of riboswitch loci with data-driven predictions of regulatory mode but also our understanding of the range of growth conditions under which each transcription-terminating riboswitch is active.

MATERIALS & METHODS

Theoretical framework

Predicted transcript pools for transcriptionally active riboswitches

Riboswitches are often associated with transcription start sites (TSS), allowing transcription start site analysis to serve as a strategy for identifying novel riboswitches and regulatory RNA (Rosinski-Chupin et al., 2014, 2019; Adams et al., 2017; Yu et al., 2018). We reasoned that a transcription-terminating riboswitch in the 5′ UTR of a gene should produce different patterns of RNA-seq read coverage in the ON vs. OFF state (Figure 1a, b). In either case, transcription should begin at the TSS and proceed through the riboswitch itself. In the OFF state, formation of the transcription-terminator structure should truncate transcripts near the riboswitch 3′ end (Figure 1a top). By contrast, in the ON state, RNA polymerase should read through, producing transcripts that extend well into the downstream gene (Figure 1a bottom). A riboswitch that produces these two patterns under different conditions exhibits behavior consistent with riboswitch-mediated conditional regulation of transcription termination.

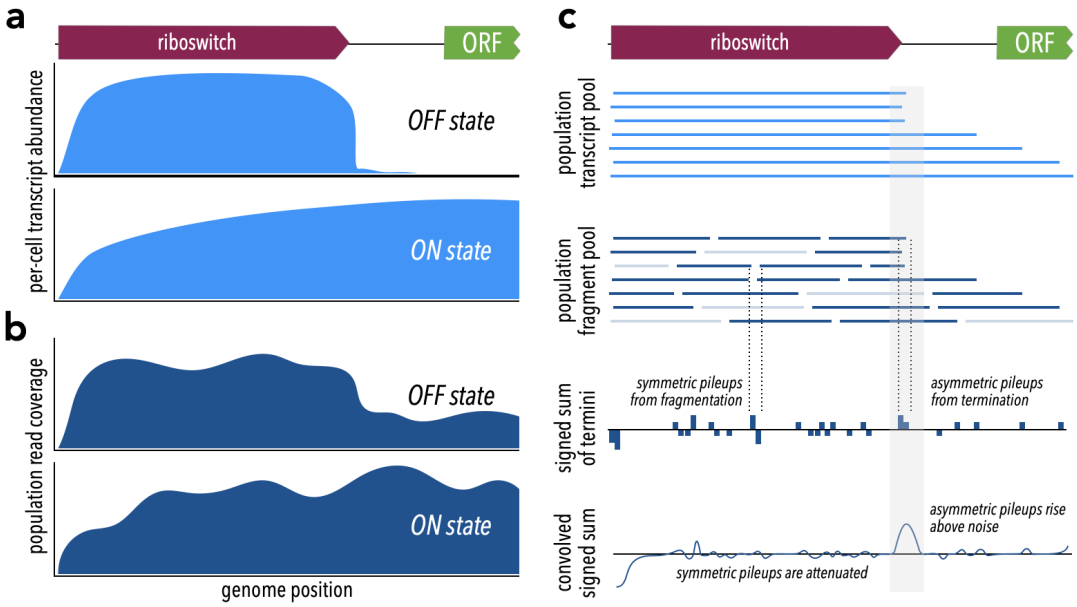

Figure 1. Theoretical framework for `TaRTLEt's` approach. (a) At riboswitches that act by modulating transcription termination, a single cell's transcript abundance profile differs sharply between the riboswitch OFF and ON states. (b) When RNA is extracted from a population of cells for sequencing, coverage profiles are made noisier by a combination of biological (e.g., subpopulations in ON states alongside others in OFF states) and technological factors (e.g., transcript fragmentation, sequencing of a subset of the fragment pool). (c) Weighted convolution on the signed sum of fragment termini attenuates the noise signal, making transcription termination signals readily detectable even when the population-level read coverage does not fall to zero.

Fragment distributions and terminal pileups

What relation does the coverage map of sequencing reads bear to the pool of transcripts? Each RNA-seq read represents some region of a transcript, but because transcripts are typically fragmented prior to sequencing we know only that the 5′ end of a read must align with a fragment terminus and not whether that fragment is internal or terminal to a transcript. Fragmentation often, as in the validation dataset used here, results in a random biased distribution (Fig. S1) of fragment sizes. Importantly, however, the new 3′ and 5′ termini produced by each fragmentation event come in neighboring pairs. Even though the total number of fragment termini, and thus the total number of read termini, may far exceed the number of original transcript termini, we can still extract a transcript size signal by considering pileups of read termini. Because paired-end sequencing generates reads that represent both ends of each sequenced fragment, the resulting data offers a higher signal:noise ratio than single-end sequencing: there, fragmentation is more extensive *and* at most one terminus of each fragment is represented in the read set, so that

comparable sequencing depth yields greater stochastic fluctuations in the set of termini captured. We therefore designed `TaRTLEt` v1.0 to restrict analysis to paired-end datasets. Future versions could be extended to analyze appropriately deep single-end datasets.

We reasoned that a convolution approach would allow us to both extract the needed transcript size signal and smooth the overall signal. Convolution is heavily used in applications like edge-detection algorithms (Canny, 1986; Maini and Aggarwal, 2009; Albawi et al., 2017), where it can act as an attenuating filter for neighboring signals. Here, we want neighboring pileups of $3'$ and $5'$ read termini to drop out, filtering out transcript-internal fragmentation to reveal the transcript-terminal signal. Tracking $3'$ ends as $+1$ and $5'$ ends as -1, we can first compute the signed sum of read termini for each position in the reference (Fig. 1c), then calculate a weighted average of surrounding sums for each position. We call the resulting convolution's features "peaks", and we seek to identify peaks at which a pileup of $3'$ termini is not accompanied by a pileup of $5'$ termini immediately downstream.

The choice of kernel function for computation of the weighted average affects the filter's selectivity for changes in coverage. Rather than a flat uniform distribution, we chose to use kernels constructed by discretizing a Gaussian distribution, such that tuning the spread of this distribution (by the standard deviation σ) tunes the filter's selectivity for how sharply coverage must change. Because sites of transcription termination, especially intrinsic termination, are constrained (Roberts, 2019; Gusarov and Nudler, 1999), riboswitch-mediated transcription termination should produce a sharp drop in coverage. To match this expectation, we might use a tight convolution (small σ) kernel that attenuates peaks arising from changes over $>$1–2 nt. However, both biological and technical factors (e.g., thermodynamic fluctuations; processes in library preparation, sequencing, and quality control) are expected to broaden the biological signal to some degree. Through trial and error, we chose a default σ of 1.5 for kernel generation, corresponding to a filter that passes coverage changes localized within a 6-nt region.

Regardless of the number of fragmentation events, the convolution always returns non-zero peaks at positions with asymmetric pileups of $5'$ or $3'$ termini. These surviving peaks should represent a range of events: transcription termination should be relatively rare and should cause a large, tightly centered coverage drop, giving rise to high, narrow peaks; sporadic processes (e.g., artifacts in library prep or sequencing) should be relatively common and yield lower, broader peaks. While these differences in peak shapes should provide an opportunity to further refine the set of candidate peaks, the probability and details of the processes generating background peaks may be sequence-dependent in as yet poorly understood ways, complicating comparison to a global baseline. Instead, we reasoned that we could collect a per-locus "noise set" of peaks within a defined neighborhood scaled to riboswitch size. Fitting a 2-dimensional Gaussian noise distribution to this collection of peaks provides a basis for comparison, allowing us to robustly identify unusually sharp peaks at each locus as candidate sites of riboswitch-mediated transcription termination.

Probing fragment-end distributions across treatments

For any of these sites to reveal conditional regulation by a transcription-terminating riboswitch, the efficiency of termination (that is, the proportion of transcripts that terminate) near the riboswitch $3'$ end must vary across treatments that evoke ON and OFF states. For a riboswitch that acts by transcription termination, termination efficiency should be low (low or no coverage drop) in the ON state and high (large coverage drop) in the OFF state. We bound the region of interest for this search to a range of $0.5\times$riboswitch size both upstream and downstream of the riboswitch $3'$ end and examine the fractional coverage change (that is, (change in coverage across peak) / (average coverage depth across entire riboswitch)) at each convolution peak in the region under each treatment condition. To compare results at coincident peaks across treatment conditions, we account for the possibility that peak location might shift slightly due to small variations in process noise by clustering peak locations across treatment conditions (Fig. 2). Any cluster that captures both ON and OFF states of a transcription-terminating riboswitch should exhibit fractional coverage change with a mean that is significantly more negative and a variance that is significantly larger than baseline (Fig. 2, 3).

`TaRTLEt` Implementation

`TaRTLEt` is implemented in `python` v3.11 with a dedicated command-line interface (CLI). Its usage does not require knowledge of python and the included `conda` environment contains installations for any

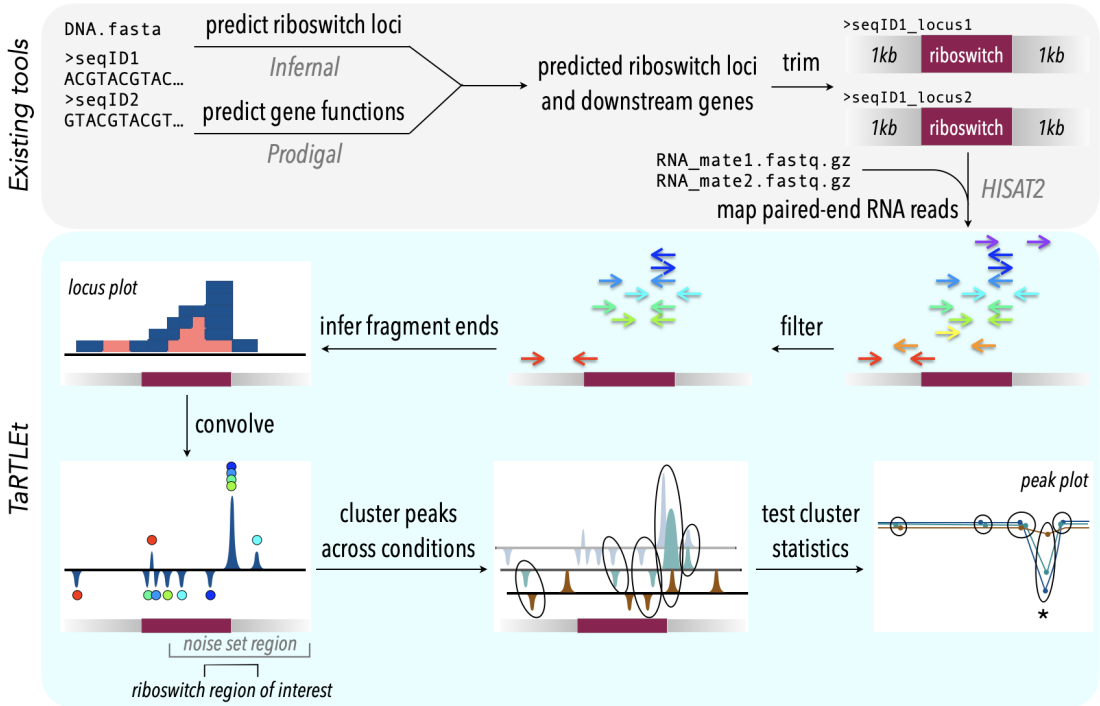

Figure 2. `TaRTLEt` pipeline for detection of transcription-terminating riboswitches in paired-end RNA-seq data. Existing tools (top) are used to identify known riboswitch families in (meta)genomic data, predict downstream open reading frames, and map RNA-seq reads to a reduced dataset of riboswitch loci and their downstream regions. `TaRTLEt` then filters mapped reads by orientation and analyzes the distribution of fragment termini, generating a signed sum of termini at each position and applying a weighted convolution to attenuate the symmetric signals arising from mRNA fragmentation before sequencing. Peaks in the convolution are clustered across conditions by genome position; then `TaRTLEt` analyzes changes in fractional read coverage across each cluster to identify sites of condition-dependent transcription termination. Locus plots are generated for each riboswitch locus under each experimental condition represented in the RNA-seq data, to support per-locus per-condition calling of riboswitch ON or OFF state; peak plots capture all experimental conditions, to support per-locus calling of riboswitch regulatory mode.

external CLI tools. For widespread utility, we chose to incorporate a riboswitch identification pipeline into our approach, so that inputs may be unannotated genomes/sequences; identified riboswitch loci are then submitted to the mechanistic inference pipeline. A complete walk-through of `TaRTLEt` usage and the analysis of data generated by `TaRTLEt` is available at the tool GitHub repository (`https://github.com/lepton-7/tartlet`). The pipeline is summarized in Fig. 2.

Identification of riboswitches and open reading frames

Input sequences in FASTA format are processed through `Infernal` v1.1.5 (`http://eddylab.org/infernal`) (Nawrocki and Eddy, 2013b) and `Prodigal` v2.6.3 (`https://github.com/hyattpd/Prodigal`) (Hyatt et al., 2010). For `Infernal`, each input sequence is passed to `cmscan` using the `--cut_ga --rfam --nohmmonly --noali --tblout` options against a comprehensive riboswitch covariance model of families compiled from the latest Rfam collection (version 14.9 as of writing) (`https://rfam.org/`) (Griffiths-Jones et al., 2003). `cmscan` output is parsed to identify riboswitch loci for further processing. `Prodigal` is called with the default output and translation options. For each identified riboswitch, the downstream gene is identified from the `Prodigal` output and recorded for downstream plot annotations. The downstream gene is defined as the first open reading frame past the riboswitch 5′ end and on the same strand. This definition captures ORFs that overlap with part of the identified riboswitch sequence, in order to maximize the likelihood of capturing ORFs that are biologically relevant to the riboswitch.

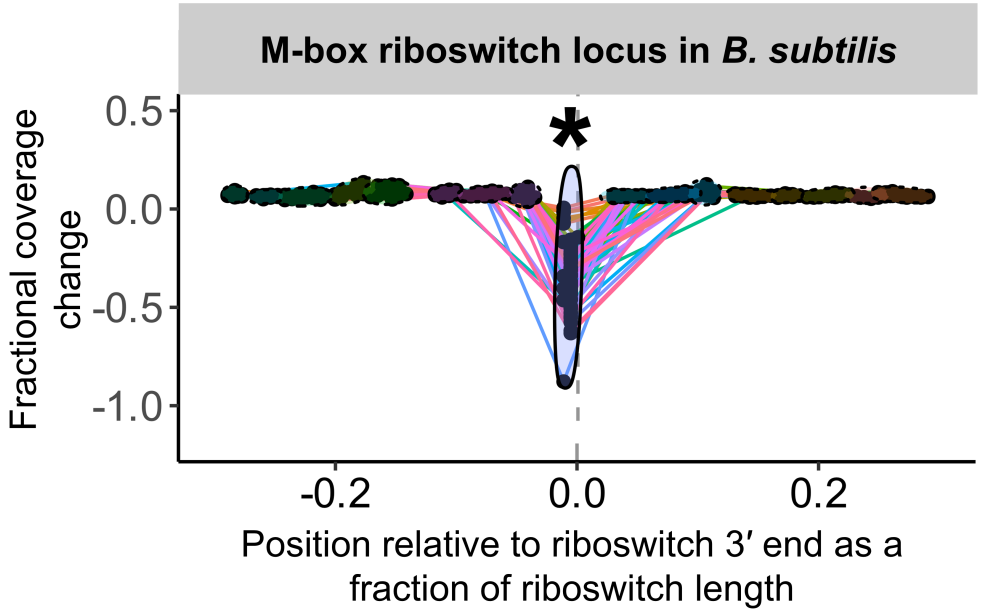

Figure 3. Sample peak plot output of `TaRTLEt`, showing evidence of conditional transcription termination at a riboswitch 3′ end. For each riboswitch locus, the locus plot arising from each experimental condition is convolved and fragment terminus peaks are detected. Each point shows the fractional coverage change across one such peak. Colored lines connect peaks detected under the same experimental conditions. Peak locations are shown relative to riboswitch size and 3′ end; that is, a peak at 0.1 on the x-axis lies 10% of the riboswitch length downstream of the riboswitch 3′ end. Peaks are clustered across experimental conditions based on x-position; ellipses mark peak clusters. All detected peaks are clustered, but only clusters containing at least one transcription-termination peak are viable candidates. Significance testing of the mean and variance of each such cluster's fractional coverage drop identifies clusters consistent with conditional transcription termination by the riboswitch locus (asterisk).

Riboswitch reference generation

To reduce computational requirements for transcript mapping, `TaRTLEt` trims input genomic sequences back to riboswitch neighborhoods, producing reference sequences that stretch from 1000 nt upstream of each identified riboswitch locus to 1000 nt downstream. (For metagenomic inputs, where riboswitch loci may lie less than 1000 nt from contig ends, neighborhoods are permitted to be smaller or asymmetric.) We chose this neighborhood size on the basis of common practice in paired-end sequencing library preparation, where the target fragmentation size is often less than 1000 nt to preserve base quality (Tan et al., 2019). Consequently, the generated reference neighborhoods are typically (96.61% of riboswitch-aligned reads in our validation dataset) large enough to capture both a riboswitch-aligned read and its paired-end mate. All references generated by `TaRTLEt` are oriented $5' \rightarrow 3'$; references derived from (-) strand riboswitches are reverse complemented before saving to file.

Reference alignment

RNA-seq reads are aligned to the generated riboswitch reference using `HISAT2` (`http://daehwankimlab.github.io/hisat2/`) (Kim et al., 2019). `HISAT2` was chosen for its fast per-read alignment time, good parallel scaling, and acceptable alignment rates (Musich et al., 2021). `HISAT2` calls are made using the default options, but any additional valid alignment options may be passed through the `TaRTLEt` interface. The current `TaRTLEt` implementation requires that the input RNA-seq datasets be paired-end. The resulting SAM files are converted to sorted BAMs for downstream processing.

We recommend running `HISAT2` using the options `--no-unal --score-min L,0,-0.4`. The use of `--no-unal` dramatically reduces output file sizes by omitting unaligned reads. The use of `--score-min L,0,-0.4` decreases stringency to allow reads with alignment mismatches/soft-clipping to be included in the output; any alignment errors are handled during downstream processing.

The validation dataset presented here was run using these parameters in addition to `-p 40` (performance tuning with multiple threads) and `-t` (wall-time logging).

Processing mapped reads

`TaRTLEt` heavily utilizes the `pysam` (`https://github.com/pysam-developers/pysam`) library (a python interface for `samtools` (Li et al., 2009) and `htslib` (Bonfield et al., 2021)) to parse BAM files and generate alignment data objects for downstream data transformation. Once mapped reads/read pairs have been collected from the sorted BAMs, they are filtered to discard pairs in invalid orientations (RF, TANDEM, FFR, RRF), which may reflect sequencing errors. Only pairs in the EQ and FR orientations, and pairs with only one mapped read, are retained as valid. The EQ ("equal") orientation denotes reads aligned in opposite directions that overlap across their entire length, as is often the case when the fragment sequenced is $\sim$150bp, comparable in scale to individual short reads. In the FR orientation, the left-most aligned base (looking $5' \rightarrow 3'$ in the reference) is either shared between the F and R oriented reads or belongs only to the F oriented read, and the right-most aligned base belongs exclusively to the R read.

Next, `TaRTLEt` generates fragment coverage pileups using the filtered set of reads for every position in the reference sequence. Single reads (with unmapped mates) are discarded during counting unless the `--allow-single-reads` flag is passed to `TaRTLEt`. `TaRTLEt` tracks five types of coverage. Two can incorporate data from single reads if included: "read coverage" counts bases directly mapped by a read, while "clipped fragment coverage" counts bases mapped by `HISAT2`'s soft-clipped read regions. The remaining three types require both mates to be mapped. "Overlapped fragment coverage" counts bases overlapped by both mates; "inferred fragment coverage" counts unmapped bases that lie between the mapped regions of the two mates; and "fragment termini coverage" counts only the mapped base that is the origin of read synthesis, i.e., looking $5' \rightarrow 3'$ in the reference, the left-most mapped base for an F read and right-most mapped base for an R read. This last category captures the ends of sequenced fragments to feed into the signed sum of termini for convolution analysis; we count the fragment terminus captured by the F read as -1 and the fragment terminus captured by the R read as $+1$. Reads that have soft-clipped regions are discarded from this step unless the `--allow-soft-clips` flag is specified. Fragment termini counts are removed from the read coverage and/or overlapped coverage counts, and read coverage counts are removed from the overlapped coverage counts, such that no mapped base is counted more than once. This ensures that counts across all coverage types at a given position sum to the read depth at that position.

Finally, to ensure sufficient coverage for further analysis, riboswitch loci are filtered based on the average read coverage across the length of the riboswitch. The coverage threshold can be adjusted with the `--min-cov-depth` option; the validation presented here uses the default value of 15.

Transformation and feature identification

Fragment termini coverage pileups are transformed using a convolution to extract information about transcription starts and stops. `TaRTLEt` generates a 51-element 1-dimensional weighted kernel by discretizing a Gaussian probability density function with $\sigma = 1.5$. The weighted kernel is then applied to the fragment termini coverage array through the convolution to generate an array of peaks corresponding to transcription events. The convolution transforms coverage at each position into convolution amplitude (arbitrary units).

`TaRTLEt` conducts simple peak detection as follows for a positive peak: a local convolution amplitude maximum is defined as the peak "summit". The peak region covers all the bases on either side of the summit at which convolution amplitude (a) is monotonically decreasing and (b) does not enter or cross a zero-region, defined as the interval $(-0.1, 0.1)$. To delineate ambiguous shoulder regions between peaks, peak bounds follow pythonic inclusive-exclusive convention where the left peak bound represents the left-most valid base according to the monotonicity and zero-crossing criteria and the right peak bound represents the first base that fails the criteria. Negative and positive amplitude peaks are considered candidate transcription start and termination events, respectively.

Next, `TaRTLEt` assesses candidate transcription-termination peaks individually to determine which rise above background. Working with per-locus, per-condition convolution plots, `TaRTLEt` compares each positive peak within a user-defined region to a 2-dimensional multivariate normal (MVN) distribution

fitted to the widths and amplitudes of each other peak in the region (the "noise set"). The region that feeds into the noise set is defined relative to riboswitch length, with parameters defining the proportion of the riboswitch that should be excluded at the 5′ end and how far past the riboswitch 3′ end the region should stretch. (Our validation analysis used `--ext-prop -0.3 1.0` to define a noise set region that excludes the first 30% of the riboswitch, includes the remainder, and continues beyond the riboswitch 3′ end to include a region equal in length to the complete riboswitch.) The likelihood of drawing a given peak from the MVN is used as a pseudo p-value, with a significance threshold of < 0.05. Peaks that pass this test and exceed a user-configurable coverage change threshold (by default, a drop of $\geq$20% of the mean coverage across the locus) are called as transcription termination events.

To facilitate biological interpretation of riboswitch activity, `TaRTLEt` then asks, for each condition and each locus, whether any passing transcription-termination peaks were observed. This within-condition check does not in itself offer evidence for riboswitch regulation by transcription termination. Instead, it (1) feeds into the cluster-level calls described below; (2) supports manual curation of results; and (3) allows the user to ask, among the transcription-terminating riboswitches identified below, what set of experimental conditions evoke the ON state vs. the OFF state.

Peak clustering and statistics

To detect differential transcription termination, `TaRTLEt` compares data from all treatment conditions at each riboswitch locus. All peaks that fall within the riboswitch region of interest (user-configurable; by default, within $\pm 0.5\times$riboswitch size of the riboswitch 3′ end) under any condition are clustered by position to enable cross-condition comparisons. The riboswitch 3′ end defines position 0, and peak positions are given relative to this site as a fraction of the riboswitch size. For example, a position of -0.15 would indicate that the peak is located 15% of the riboswitch size upstream of the riboswitch 3′ end. Clustering is implemented via `fclusterdata` from the `SciPy` package and follows a cophenetic distancing algorithm with complete linkage. The default (user-configurable) cophenetic distance threshold of 0.04 was chosen through trial and error (Fig. S6).

Once peaks are assigned to clusters, cross-cluster comparisons at each locus are used to test for condition-dependent transcription termination. Each peak is characterized by the change in total fragment coverage across its bounds. The fractional coverage change is defined as the difference in total coverage between the left and right bounds of the peak expressed as a fraction of the average total read coverage between the 5′ and 3′ ends of the riboswitch. `TaRTLEt` performs a Mann-Whitney one-tailed U-test (`scipy.stats.mannwhitneyu`) to compare each cluster's fractional coverage change to the mean fractional coverage change of all other peaks, to test whether the cluster mean is more negative than the mean of the set of all other peaks.

Testing for differences in cluster variance must account for bias from differences in cluster sizes. `TaRTLEt` uses Levene's test (`scipy.stats.levene`) to run pairwise comparisons between the variance of a given cluster enclosing n peaks and that of random samples of n peaks chosen without replacement from all other peaks at the same riboswitch locus until the superset of peaks is exhausted or at least 60 comparisons have been made. If the superset is exhausted before 60 comparisons are made, all peaks are replaced and the clusters are randomly resampled to reach 60 comparisons. `TaRTLEt` returns the median of the resulting 60 p-values for each peak cluster.

Finally, `TaRTLEt` combines peak- and cluster-level tests to assess significance. To be called as showing evidence of condition-dependent transcription termination, a riboswitch locus must have at least one cluster that (i) includes at least one significant transcription-termination peak (i.e., distinct from the noise-set MVN at $p < 0.05$ *and* coverage drop greater than or equal to the user-defined threshold; see previous section); (ii) passes the Mann-Whitney U-test for mean fractional coverage change; and (iii) passes the Levene test for increased variance.

Description of outputs

`TaRTLEt` outputs data on two layers of abstraction: single-condition convolution peaks, and cross-condition peak clusters. At each level, `TaRTLEt` generates both tabular output on each feature and plots of each riboswitch locus. For each candidate peak in the convolution, the single-condition tabular output (`peak_log.csv`) captures the riboswitch locus, the RNA-seq dataset, the location and shape of the peak, the "noise set" of comparator peaks, `TaRTLEt`'s pass/fail call for that peak as a candidate transcription

termination event, and the reason for that call. Single-condition locus plots are sorted into `pass/` and `fail/` subdirectories, where `pass/` contains locus plots for all conditions in which any peak, regardless of location within the locus, was called as a transcription termination event. A sample locus plot of the MOCO riboswitch from *E. coli* is shown in Fig. 4; for a sample peak log, see Table S1.

At the level of cross-condition cluster comparisons, `TaRTLEt` records each cluster's mean and variance test statistics in `cluster_stats.csv`, then uses this file in conjunction with `peak_log.csv` to generate a summary peak plot that allows easy inspection of condition-dependent modulation of transcription termination efficiency near the 3′ ends of all riboswitches in the dataset. Fig. 5 shows a sample summary peak plot. The user may simply scan the summary peak plot to identify loci in which a cluster was marked significant (*), optionally using the locus plots to identify particular experimental conditions that evoked a change in riboswitch transcription termination efficiency (at loci identified as modulating transcription termination) or to spot-check peak call quality.

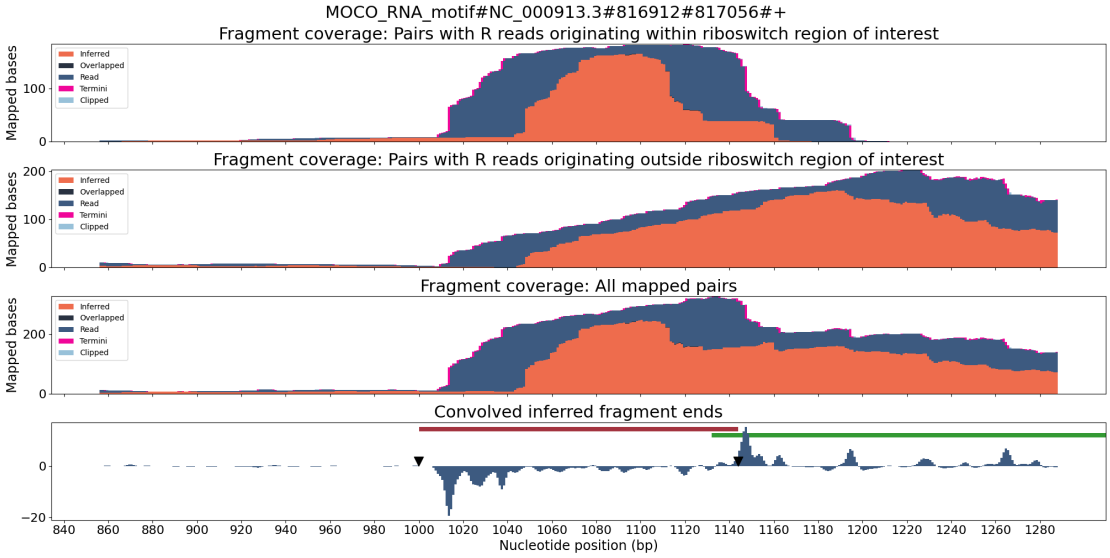

Figure 4. Locus plot of the MOCO riboswitch from *E. coli* for RNA-seq sample SRR7154624. The top panel collects fragment coverage for read pairs with R-oriented reads starting in the riboswitch region of interest ($\pm 0.5\times$riboswitch size). The second panel collects fragment coverage for read pairs with R-oriented reads starting outside the region of interest. Coverage counts in the third panel are the union of the top two panels. The bottom panel records the results from applying the Gaussian convolution to the fragment termini coverage. Vertical axes for all panels are not normalized and are scaled to fit the maximum displayed value.

Validation datasets

All the paired-end RNA-seq datasets used to test and characterize `TaRTLEt` here are presented in Table 1. These datasets were obtained from the Gene Expression Omnibus (GEO) (Barrett et al., 2012). We used the Ribocentre-switch database (Bu et al., 2024) to facilitate identification of previously characterized riboswitches in this dataset.

RESULTS

`TaRTLEt` performs well on known transcriptionally-active riboswitches

To validate this new tool, we ran `TaRTLEt` on publicly available paired-end RNA-seq datasets for 31 microbial genomes (Table 1). Of these, datasets for 29 genomes were successfully processed through `TaRTLEt`. This collection of microbes is phylogenetically diverse, with good representation of both Gram-positive and Gram-negative bacteria (Fig. 6a), and includes 21 riboswitch loci that have previously been experimentally characterized.

While this low count underscores the need for high-throughput methods to supplement *in vitro* approaches, it also presents a practical impediment to robust determination of error rates. In light of this limitation,

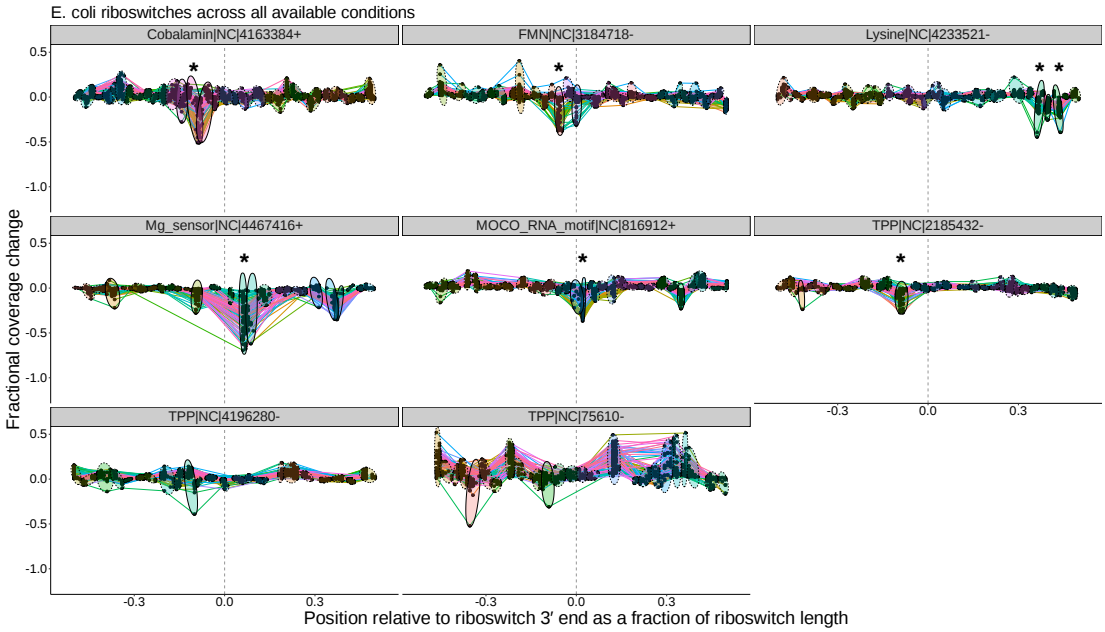

Figure 5. Peak plots (see Fig. 3) for the eight *E. coli* riboswitch loci identified by `Infernal` in the *E. coli* genome. Peak clusters are grouped by colored ellipses; clusters with at least one significant transcription-termination peak are bounded by solid ellipses, and those with no significant peaks are bounded by dashed ellipses. Asterisks mark peak clusters that (i) contain at least one significant transcription-termination peak, (ii) have significantly lower mean ($p < 0.05$, Mann-Whitney one-tailed U-test) than the set of all other peak clusters, and (iii) have significantly higher variance ($p < 0.05$, Levene's test) when compared to randomly sampled sets of the same number of peaks.

`TaRTLEt` uses conservative analytical and statistical approaches and thresholds, as well as making intermediate analysis steps available to the user for inspection. Our 29-genome validation dataset included 405 riboswitch loci and a total of 655 RNA-seq readsets. Manual curation of `TaRTLEt`'s locus plots led us to exclude one genome (*P. fluorescens*, 10 loci, 40 readsets) as poorly mapped. Among the remaining 28 genomes and 395 loci, `TaRTLEt` identified 115,130 coverage-change features as potential transcription-termination peaks; comparison to transcriptomic noise allowed us to reject all but 3437 candidates. In cross-condition significance testing of peak clusters, 1765 of these candidates lay within the 104 clusters at 95 riboswitch loci that passed our mean and variance test criteria. `TaRTLEt` calls riboswitch loci as showing evidence of condition-dependent transcription termination based on this latter, most conservative, most data-rich set, while the within-condition single-peak calls (and supporting data) are made available to the user to support manual curation.

Overall, our results are in good agreement with literature findings (Fig. 6b). `TaRTLEt`'s peak plots (Fig. 5, S2–S5) showed clear condition-dependent transcription termination at 10 of the 19 loci known to use this regulatory mode: in *B. anthracis*, the upstream tandem TPP riboswitch (Welz and Breaker, 2007); in *B. subtilis*, the *lysC* lysine, *mgtE* M-box, *xpt* purine, and *samT* and *metK* SAM riboswitches (Sudarsan et al., 2003; Dann et al., 2007; Mandal and Breaker, 2004; Winkler et al., 2003); and in *E. coli* the *btuB* cobalamin, *ribB* FMN, *lysC* lysine, and *thiM* TPP riboswitches (Nou and Kadner, 1998; Hollands et al., 2012; Sedlyarova et al., 2016; Bastet et al., 2017). In addition, `TaRTLEt` detected no indication (Fig. S2, S4) of transcription-termination activity at either of the two negative-control loci experimentally characterized as acting strictly through mechanisms other than transcription termination (the *B. subtilis glmS* and *Enterococcus faecalis* S_{MK} riboswitches (Collins et al., 2007; Fuchs et al., 2007; **?**)).

We also examined the validation dataset's performance across the pipeline to identify the range of situations that can give rise to inconclusive or negative results, as observed at the nine remaining loci in our dataset with previous experimental characterization. First, `TaRTLEt` cannot provide insight into riboswitch loci with read coverage too low to support statistical analysis. We observed this outcome

Table 1. List of GEO accessions for paired-end RNA-seq read sets used for characterizing riboswitches from each organism.

Microbe	GEO Accession	Citation
Acinetobacter baumannii	GSE183334	Unpublished
Aliivibrio fischeri ES114	GSE237189	Vander Griend et al. (2024)
Apilactobacillus kunkeei	GSE205998	Seeger et al. (2023)
Bacillus anthracis	GSE152356	Corsi et al. (2021)
Bacillus subtilis str. 168	GSE219221	Guo and Herman (2023)
Bacillus subtilis str. 168	GSE226559	Unpublished
Bacteroides fragilis	GSE220692	Fiebig et al. (2024)
Bacteroides thetaiotaomicron	GSE129572	Briliūtė et al. (2019)
Bacteroides thetaiotaomicron	GSE169260	Lewis and Gui (2023)
Bacteroides xylanisolvens	GSE74379	Despres et al. (2016)
Burkholderia pseudomallei	GSE152295	Avican et al. (2021)
Caulobacter vibrioides	GSE241057	McLaughlin et al. (2023)
Clostridioides difficile	GSE173804	Weiss et al. (2021)
Desulfovibrio vulgaris str. Hildenborough	GSE101911	Chen et al. (2019)
Desulfovibrio vulgaris str. Hildenborough	GSE78834	Gao et al. (2016)
Enterococcus faecalis	GSE152295	Avican et al. (2021)
Escherichia coli	GSE114358	Guzman et al. (2018)
Escherichia coli	GSE122211	Sastry et al. (2019)
Escherichia coli	GSE122295	Sastry et al. (2019)
Escherichia coli	GSE122296	Sastry et al. (2019)
Escherichia coli	GSE122779	Anand et al. (2019)
Eubacterium limosum	GSE149269	Jeong et al. (2020)
Klebsiella pneumoniae	GSE229867	Unpublished
Mycobacterium tuberculosis str. H37Rv	GSE218354	Martini et al. (2023)
Mycolicibacterium smegmatis MC2 155	GSE232901	Grigorov et al. (2023)
Neisseria gonorrhoeae	GSE191020	Ray et al. (2022)
Piscirickettsia salmonis	GSE235725	Carril et al. (2023)
Pseudomonas chlororaphis subsp. Aureofaciens 30-84	GSE61200	Wang et al. (2016)
Pseudomonas fluorescens	GSE200822	Hasnain et al. (2023)
Salmonella enterica sv. Typhimurium	GSE203342	Kant et al. (2023)
Sinorhizobium meliloti	GSE151705	Fagorzi et al. (2021)
Staphylococcus epidermidis	GSE152295	Avican et al. (2021)
Streptococcus sanguinis SK36	GSE174672	Puccio et al. (2022)
Streptomyces coelicolor	GSE234439	Li et al. (2023)
Synechococcus elongatus	GSE227397	Lu et al. (2023)
Synechocystis sp. PCC 6803	GSE132275	Unpublished
Xanthomonas albilineans	GSE229478	Wimmer et al. (2024)
Xanthomonas oryzae pv. Oryzae KACC 10331	GSE89651	Park et al. (2022)

in 29 of the 395 riboswitch loci investigated here, leaving 366 loci tractable for further investigation. Second, `TaRTLEt` might detect no candidate transcription-termination events in the search space at a riboswitch locus (132 of 366 tractable loci); this could happen either because there is no terminator present in the region of interest, or because under the conditions tested the terminator is always inactive. In our validation set, this group included two loci previously characterized as acting via transcription termination, the *B. subtilis metI* SAM riboswitch (Fig. S2) and the *E. faecalis pduQ* AdoCbl riboswitch (Fig. S4) (Winkler et al., 2003; DebRoy et al., 2014).

Third, a locus with transcription-termination peaks might have no clusters that pass the mean fold coverage drop test, if the conditions examined yield consistently high rates of transcriptional read-through (whether because the conditions used sample only the ON state or because the riboswitch acts downstream of

transcription). In our validation dataset, 20 of the 234 loci with passing peaks failed this test. This group includes two riboswitch loci (the *E. coli thiC* TPP and *B. subtilis pbuE* purine riboswitches) that have previously been shown to modulate transcription (Winkler et al., 2002; Chauvier et al., 2017; Mandal and Breaker, 2004) but show no mean fold coverage drop here, suggesting that the range of conditions examined was too narrow to capture the OFF state for these riboswitches (Fig. 5). Interestingly, the purine ligand does vary enough across the conditions tested in *B. subtilis* to cause a detectable switch in transcription termination at the *xpt* purine riboswitch (Fig. 6b, S2), suggesting that these two *B. subtilis* purine riboswitches are responsive to different ligand concentrations.

Finally, among the clusters at a locus that both include transcription-termination peaks and pass the mean test, none may pass the variance test (119 of the remaining 214 loci). A cluster might fail the variance test either if the variance of other peak clusters is large (e.g., due to shallow coverage) or if the variance of the candidate cluster is small (due to a set of conditions that capture only one riboswitch regulatory state). In our validation set, five loci previously found to mediate transcription termination fell into this group: the *B. anthracis tenA* TPP (Fig. S3), *B. subtilis ribD* FMN, *gcvT* glycine and *yoaD* SAM (Fig. S2), and *S. enterica mgtA* Mg^{2+} sensor riboswitches (Fig. S5) (Welz and Breaker, 2007; Wickiser et al., 2005; Mandal et al., 2004; Winkler et al., 2003; Cromie et al., 2006). Two of these, the *B. subtilis ribD* FMN riboswitch (Fig. S2, panel `FMN|NC|2431617-`) and the *S. enterica mgtA* Mg^{2+} sensor riboswitch (Fig. S5, panel `Mg_sensor|4699497+`), appear to fail the variance test by being consistently OFF under the conditions tested; two more (*B. subtilis gcvT* glycine and *yoaD* SAM; fig. S2, panels `Glycine|NC|2549606-` and `SAM|NC|2025251-`) are clearly consistent with conditional transcription termination but fall just shy of significance. Thus while our mean- and variance-based calling approach can in general be applied to existing datasets, the noisiness and the scope of the datasets examined can limit robust inference. However, where regulatory mechanism is already established by other means, the signal `TaRTLEt` exploits may still show riboswitch regulatory state even when the conditions sampled capture only ON or only OFF.

`TaRTLEt` gives new insights into undercharacterized riboswitches

Thiamine pyrophosphate (TPP) riboswitches are the most abundant riboswitches in genome databases (McCown et al., 2017) and were identified in nearly all of the genomes in our validation set. *In vitro* characterization of this class has variously pointed to regulation at the level of transcription termination, translation initiation, or even both (Sedlyarova et al., 2016; Winkler et al., 2002; Chauvier et al., 2017). In one notable case, *B. anthracis* encodes two TPP riboswitches in tandem upstream of the *tenA* gene, and each has been shown *in vitro* to be able to regulate transcription termination independently (Welz and Breaker, 2007). Of the 57 TPP riboswitches in the validation dataset, only six (two from *B. anthracis* and one each from *B. subtilis*, *S. epidermidis*, *E. coli*, and *C. vibrioides*) TPP representatives show transcriptional activity under the conditions tested. Intriguingly, the peak plot for *B. anthracis* showed transcriptional activity by the upstream but not the downstream *tenA* TPP riboswitch (Fig. S3, panels `TPP|752099+` and `TPP|752272+`), suggesting that the two tandem riboswitches are responsive under different conditions *in vivo*.

A second very widespread riboswitch class binds cobalamin, with at least one example in all but four genomes in our validation dataset. Previous research has predicted that cobalamin riboswitches in Gram-positive taxa will typically work at the transcriptional level, while those in Gram-negative taxa will regulate translation (Lennon and Batey, 2022). Contra this expectation, we see clear evidence of transcriptionally-active cobalamin riboswitches in the majority of Gram-negative taxa we investigated (Fig. 6; 15 Gram-negative taxa with cobalamin riboswitch loci, of which ten have at least one positive `TaRTLEt` result). Interestingly, however, in nearly every case, these taxa also contained one or more additional cobalamin riboswitches that showed no evidence of transcriptional regulation. These other riboswitches could either regulate other aspects of mRNA production, stability, or translation, or regulate transcription termination at a different range of ligand concentrations than this dataset captures. While `TaRTLEt` cannot distinguish between these possibilities with the data used here, the observation underscores the potential variability of riboswitches within a ligand class and highlights the utility of high-throughput analyses for identifying interesting candidates for experimental characterization.

`TaRTLEt` also offers insights into the less-studied riboswitch classes. For example, the molybdenum cofactor ("MOCO") riboswitch in *E. coli* regulating the *moaA* operon is well-characterized structurally

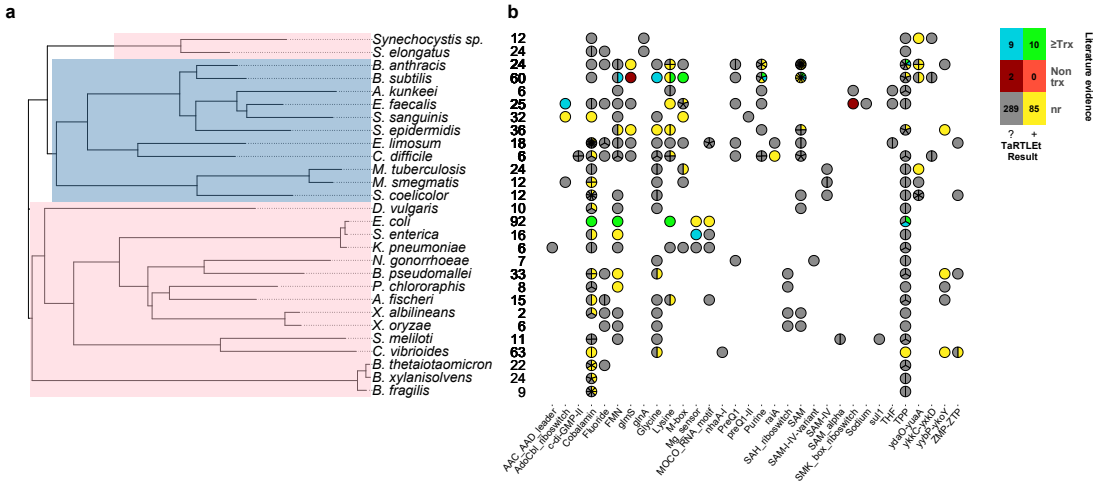

Figure 6. Compilation of TaRTLEt results for a collection of commonly studied bacteria, using RNA-seq data obtained from public GEO repositories. (a) Phylogenetic relationships between bacteria studied with TaRTLEt. The tree is based on GTDB (Parks et al., 2022; Rinke et al., 2021; Parks et al., 2020, 2018) bacterial phylogenies (release 214.1) and built using the ggtree R package (Yu et al., 2017). (b) Evidence for condition-dependent transcription termination for every riboswitch representative, as found in the literature and by TaRTLEt. Counts at left give the number of experimental conditions for which paired-end RNA-seq data was available; a wider range of conditions increases the likelihood that the dataset will capture condition-dependent transcription termination. At each locus, evidence reported in the literature may support regulation by transcription termination, potentially among other mechanisms ("$\geq$Trx"), or by non-transcription-termination mechanisms only ("Non trx"), or the locus may not have been characterized *in vitro* ("nr"); and TaRTLEt may be inconclusive ("?") or provide positive evidence ("+") for regulation by transcription termination. Where a species encodes multiple riboswitches of a given family, plotting symbols are divided into wedges for the different loci. Counts in the color key are tallies of riboswitch loci in each group.

(Amadei et al., 2023) but not functionally. Our results indicate that this MOCO riboswitch modulates transcription termination. Across the 32 riboswitch families represented in our validation dataset, we were able to identify previous *in vitro* characterization for members of only 12 families, leaving 20 families completely uncharacterized. We find clear evidence for riboswitch-mediated transcription termination at 12 loci across five of these understudied families (MOCO, *raiA*, *ydaO-yuaA*, *yybP-ykoY*, and ZMP-ZTP families; Fig. 6). That is, TaRTLEt-enabled opportunistic analysis of data originally gathered for other purposes expands by nearly 50% the range of riboswitch families for which we have data-driven insights into regulatory mechanism.

DISCUSSION

TaRTLEt's built-in visualizations facilitate user assessment of positive-call reliability. Clusters of peaks at multiple loci under multiple conditions can be rapidly inspected in peak plots, with the underlying coverage plots available to check results at lower degrees of abstraction. Thus, TaRTLEt serves as a much-needed high-throughput tool for preliminary assessment of riboswitch mechanism and can readily identify strong candidates for *in vitro* validation.

As currently implemented, TaRTLEt can analyze paired-end RNA-seq data to detect positive evidence of condition-dependent changes in transcription termination at riboswitch loci. The absence of such evidence cannot be used to infer a different mechanism of riboswitch action, because TaRTLEt's approach depends on having a dataset that captures both ON and OFF states. Thus, among the eight species in our validation dataset for which the RNA-seq collection included ≤ 9 conditions, TaRTLEt could detect evidence for transcriptional regulation at any riboswitch in only four (Fig. 6b; compare to positive results in 19 of 21 species with data on ≥ 10 conditions). Importantly, our approach supports inference from RNA-seq data

collected across multiple different experiments, such that marginal additional data collection for the less well-studied species could make detection rates rise substantially.

CONCLUSIONS

We identified a transcriptomic signal that should arise from riboswitch-mediated conditional regulation of transcription termination, developed `TaRTLEt` to apply convolution methods to extract the signal from existing public RNA-seq data, and demonstrated that this tool supports confident identification of riboswitches that control transcription termination. Our results with this method are in good agreement with the limited number of *in vitro* characterizations available, offer new insights into previously uncharacterized riboswitches, and emphasize that riboswitches within the same ligand class, even within the same organism, can exhibit different ligand sensitivities and act by different mechanisms. Whereas detailed *in vitro* characterization can establish which of several regulatory modes a single riboswitch uses, our high-throughput approach can ask which of any number of riboswitch loci appear to act by transcription termination and broadly survey the conditions under which this activity is invoked *in vivo*. Application of `TaRTLEt` to microbial transcriptomic and metatranscriptomic datasets therefore offers a path forward to understanding the regulation of a wide range of biosynthetic and transport pathways in biological contexts ranging from pure culture to natural communities *in situ*.

CODE AVAILABILITY

The `tartlet` GitHub repository houses the tool source, while the `tartlet-pub` GitHub repository tracks the data collection, processing, and analysis code used for this work.

ACKNOWLEDGMENTS

We thank the members of EMERGE-BII, the VirSoil collaboration, and the Bagby lab for helpful discussions.

REFERENCES

Adams, P. P., Flores Avile, C., Popitsch, N., Bilusic, I., Schroeder, R., Lybecker, M., and Jewett, M. W. (2017). In vivo expression technology and 5′ end mapping of the *Borrelia burgdorferi* transcriptome identify novel RNAs expressed during mammalian infection. *Nucleic Acids Research*, 45(2):775–792.

Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. IEEE.

Amadei, F., Reichenbach, M., Gallo, S., and Sigel, R. K. (2023). The structural features of the ligand-free *moaA* riboswitch and its ion-dependent folding. *Journal of Inorganic Biochemistry*, 242:112153.

Anand, A., Olson, C. A., Yang, L., Sastry, A. V., Catoiu, E., Choudhary, K. S., Phaneuf, P. V., Sandberg, T. E., Xu, S., Hefner, Y., Szubin, R., Feist, A. M., and Palsson, B. O. (2019). Pseudogene repair driven by selection pressure applied in experimental evolution. *Nature Microbiology*, 4(3):386–389.

Ariza-Mateos, A., Nuthanakanti, A., and Serganov, A. (2021). Riboswitch mechanisms: new tricks for an old dog. *Biochemistry (Moscow)*, 86:962–975.

Avican, K., Aldahdooh, J., Togninalli, M., Mahmud, A. F., Tang, J., Borgwardt, K. M., Rhen, M., and Fällman, M. (2021). RNA atlas of human bacterial pathogens uncovers stress dynamics linked to infection. *Nature Communications*, 12(1):3282.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995.

Barrick, J. E. and Breaker, R. R. (2007). The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biology*, 8(11):1–19.

Bastet, L., Chauvier, A., Singh, N., Lussier, A., Lamontagne, A.-M., Prévost, K., Massé, E., Wade, J. T., and Lafontaine, D. A. (2017). Translational control and Rho-dependent transcription termination are intimately linked in riboswitch regulation. *Nucleic Acids Research*, 45(12):7474–7486.

Bédard, A.-S. V., Hien, E. D., and Lafontaine, D. A. (2020). Riboswitch regulation mechanisms:

RNA, metabolites and regulatory proteins. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(3):194501.

Blount, K. F. and Breaker, R. R. (2006). Riboswitches as antibacterial drug targets. *Nature Biotechnology*, 24(12):1558–1564.

Bonfield, J. K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., and Davies, R. M. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience*, 10(2):giab007.

Boussebayle, A., Torka, D., Ollivaud, S., Braun, J., Bofill-Bosch, C., Dombrowski, M., Groher, F., Hamacher, K., and Suess, B. (2019). Next-level riboswitch development—implementation of capture-SELEX facilitates identification of a new synthetic riboswitch. *Nucleic Acids Research*, 47(9):4883–4895.

Breaker, R. R. (2018). Riboswitches and translation control. *Cold Spring Harbor Perspectives in Biology*, 10(11):a032797.

Briliūtė, J., Urbanowicz, P. A., Luis, A. S., Baslé, A., Paterson, N., Rebello, O., Hendel, J., Ndeh, D. A., Lowe, E. C., Martens, E. C., Spencer, D. I. R., Bolam, D. N., and Crouch, L. I. (2019). Complex *N*-glycan breakdown by gut *Bacteroides* involves an extensive enzymatic apparatus encoded by multiple co-regulated genetic loci. *Nature Microbiology*, 4(9):1571–1581.

Bu, F., Lin, X., Liao, W., Lu, Z., He, Y., Luo, Y., Peng, X., Li, M., Huang, Y., Chen, X., Xiao, B., Jiang, J., Deng, J., Huang, J., Lin, T., Miao, Z., and Huang, L. (2024). Ribocentre-switch: a database of riboswitches. *Nucleic Acids Research*, 52(D1):D265–D272.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 679–698.

Caron, M.-P., Bastet, L., Lussier, A., Simoneau-Roy, M., Massé, E., and Lafontaine, D. A. (2012). Dual-acting riboswitch control of translation initiation and mRNA decay. *Proceedings of the National Academy of Sciences*, 109(50):E3444–E3453.

Carril, G., Winther-Larsen, H. C., Løvoll, M., and Sørum, H. (2023). Cohabitation of *Piscirickettsia salmonis* genogroups (LF-89 and EM-90): synergistic effect on growth dynamics. *Frontiers in Cellular and Infection Microbiology*, 13.

Chang, T.-H., Huang, H.-D., Wu, L.-C., Yeh, C.-T., Liu, B.-J., and Horng, J.-T. (2009). Computational identification of riboswitches based on RNA conserved functional sequences and conformations. *RNA*, 15(7):1426–1430.

Chauvier, A., Picard-Jean, F., Berger-Dancause, J.-C., Bastet, L., Naghdi, M. R., Dubé, A., Turcotte, P., Perreault, J., and Lafontaine, D. A. (2017). Transcriptional pausing at the translation start site operates as a critical checkpoint for riboswitch regulation. *Nature Communications*, 8(1):13892.

Cheah, M. T., Wachter, A., Sudarsan, N., and Breaker, R. R. (2007). Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature*, 447(7143):497–500.

Chen, Z., Gao, S.-h., Jin, M., Sun, S., Lu, J., Yang, P., Bond, P. L., Yuan, Z., and Guo, J. (2019). Physiological and transcriptomic analyses reveal CuO nanoparticle inhibition of anabolic and catabolic activities of sulfate-reducing bacterium. *Environment International*, 125:65–74.

Collins, J. A., Irnov, I., Baker, S., and Winkler, W. C. (2007). Mechanism of mRNA destabilization by the *glmS* ribozyme. *Genes & Development*, 21(24):3356–3368.

Corsi, I. D., Dutta, S., Van Hoof, A., and Koehler, T. M. (2021). AtxA-controlled small RNAs of *Bacillus anthracis* virulence plasmid pXO1 regulate gene expression *in trans*. *Frontiers in Microbiology*, 11:610036.

Cromie, M. J., Shi, Y., Latifi, T., and Groisman, E. A. (2006). An RNA sensor for intracellular Mg^{2+}. *Cell*, 125(1):71–84.

Dann, C. E., Wakeman, C. A., Sieling, C. L., Baker, S. C., Irnov, I., and Winkler, W. C. (2007). Structure and mechanism of a metal-sensing regulatory RNA. *Cell*, 130(5):878–892.

DebRoy, S., Gebbie, M., Ramesh, A., Goodson, J. R., Cruz, M. R., van Hoof, A., Winkler, W. C., and Garsin, D. A. (2014). A riboswitch-containing sRNA controls gene expression by sequestration of a response regulator. *Science*, 345(6199):937–940.

Despres, J., Forano, E., Lepercq, P., Comtet-Marre, S., Jubelin, G., Yeoman, C. J., Miller, M. E. B., Fields, C. J., Terrapon, N., Le Bourvellec, C., Renard, C. M. G. C., Henrissat, B., White, B. A., and Mosoni, P. (2016). Unraveling the pectinolytic function of *Bacteroides xylanisolvens* using a RNA-seq approach and mutagenesis. *BMC Genomics*, 17:1–14.

Eddy, S. R. and Durbin, R. (1994). Rna sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088.

Ellinger, E., Chauvier, A., Romero, R. A., Liu, Y., Ray, S., and Walter, N. G. (2023). Riboswitches as therapeutic targets: promise of a new era of antibiotics. *Expert Opinion on Therapeutic Targets*, 27(6):433–445.

Fagorzi, C., Bacci, G., Huang, R., Cangioli, L., Checcucci, A., Fini, M., Perrin, E., Natali, C., Dicenzo, G. C., and Mengoni, A. (2021). Nonadditive transcriptomic signatures of genotype-by-genotype interactions during the initiation of plant-rhizobium symbiosis. *mSystems*, 6(1):10–1128.

Fiebig, A., Schnizlein, M. K., Pena-Rivera, S., Trigodet, F., Dubey, A. A., Hennessy, M. K., Basu, A., Pott, S., Dalal, S., Rubin, D., Sogin, M. L., Eren, A. M., Chang, E. B., and Crosson, S. (2024). Bile acid fitness determinants of a *Bacteroides fragilis* isolate from a human pouchitis patient. *mBio*, 15(1):e02830–23.

Fuchs, R. T., Grundy, F. J., and Henkin, T. M. (2007). S-adenosylmethionine directly inhibits binding of 30S ribosomal subunits to the SMK box translational riboswitch RNA. *Proceedings of the National Academy of Sciences*, 104(12):4876–4880.

Gao, S.-H., Ho, J. Y., Fan, L., Richardson, D. J., Yuan, Z., and Bond, P. L. (2016). Antimicrobial effects of free nitrous acid on *Desulfovibrio vulgaris*: implications for sulfide-induced corrosion of concrete. *Applied and Environmental Microbiology*, 82(18):5563–5575.

Giarimoglou, N., Kouvela, A., Maniatis, A., Papakyriakou, A., Zhang, J., Stamatopoulou, V., and Stathopoulos, C. (2022). A riboswitch-driven era of new antibacterials. *Antibiotics*, 11(9):1243.

Gilbert, S. D., Stoddard, C. D., Wise, S. J., and Batey, R. T. (2006). Thermodynamic and kinetic characterization of ligand binding to the purine riboswitch aptamer domain. *Journal of Molecular Biology*, 359(3):754–768.

Gong, S., Wang, Y., Wang, Z., and Zhang, W. (2017). Co-transcriptional folding and regulation mechanisms of riboswitches. *Molecules*, 22(7):1169.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. (2003). Rfam: an rna family database. *Nucleic Acids Research*, 31(1):439–441.

Grigorov, A. S., Skvortsova, Y. V., Bychenko, O. S., Aseev, L. V., Koledinskaya, L. S., Boni, I. V., and Azhikina, T. L. (2023). Dynamic transcriptional landscape of *Mycobacterium smegmatis* under cold stress. *International Journal of Molecular Sciences*, 24(16):12706.

Groher, F. and Suess, B. (2014). Synthetic riboswitches—a tool comes of age. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1839(10):964–973.

Guo, T. and Herman, J. K. (2023). Magnesium modulates *Bacillus subtilis* cell division frequency. *Journal of Bacteriology*, 205(1):e00375–22.

Gusarov, I. and Nudler, E. (1999). The mechanism of intrinsic transcription termination. *Molecular Cell*, 3(4):495–504.

Guzman, G. I., Sandberg, T. E., LaCroix, R. A., Nyerges, A., Papp, H., Raad, M. d., King, Z. A., Northen, T. R., Notebaart, R. A., Pál, C., Palsson, B. O., Papp, B., and Feist, A. M. (2018). Enzyme promiscuity shapes evolutionary innovation and optimization.

Hanson, S., Berthelot, K., Fink, B., McCarthy, J. E., and Suess, B. (2003). Tetracycline-aptamer-mediated translational regulation in yeast. *Molecular Microbiology*, 49(6):1627–1637.

Hasnain, A., Balakrishnan, S., Joshy, D. M., Smith, J., Haase, S. B., and Yeung, E. (2023). Learning perturbation-inducible cell states from observability analysis of transcriptome dynamics. *Nature Communications*, 14(1):3148.

Hollands, K., Proshkin, S., Sklyarova, S., Epshtein, V., Mironov, A., Nudler, E., and Groisman, E. A. (2012). Riboswitch control of Rho-dependent transcription termination. *Proceedings of the National Academy of Sciences*, 109(14):5376–5381.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:1–11.

Jeong, J., Kim, J.-Y., Park, B., Choi, I.-G., and Chang, I. S. (2020). Genetic engineering system for syngas-utilizing acetogen, *Eubacterium limosum* KIST612. *Bioresource Technology Reports*, 11:100452.

Kant, S., Till, J. K. A., Liu, L., Margolis, A., Uppalapati, S., Kim, J.-S., and Vazquez-Torres, A. (2023). Gre factors help *Salmonella* adapt to oxidative stress by improving transcription elongation and fidelity of metabolic genes. *PLoS Biology*, 21(4):e3002051.

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment

and genotyping with hisat2 and hisat-genotype. *Nature biotechnology*, 37(8):907–915.

Klein, D. J. and Ferré-D'Amaré, A. R. (2006). Structural basis of *glmS* ribozyme activation by glucosamine-6-phosphate. *Science*, 313(5794):1752–1756.

Lennon, S. R. and Batey, R. T. (2022). Regulation of gene expression through effector-dependent conformational switching by cobalamin riboswitches. *Journal of Molecular Biology*, 434(18):167585.

Lewis, J. P. and Gui, Q. (2023). Iron deficiency modulates metabolic landscape of *Bacteroidetes* promoting its resilience during inflammation. *Microbiology Spectrum*, 11(4):e04733–22.

Li, C., Urem, M., Du, C., Zhang, L., and van Wezel, G. P. (2023). Systems-wide analysis of the ROK-family regulatory gene *rokL6* and its role in the control of glucosamine toxicity in *Streptomyces coelicolor*. *Applied and Environmental Microbiology*, 89(12):e01674–23.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Loh, E., Dussurget, O., Gripenland, J., Vaitkevicius, K., Tiensuu, T., Mandin, P., Repoila, F., Buchrieser, C., Cossart, P., and Johansson, J. (2009). A *trans*-acting riboswitch controls expression of the virulence regulator PrfA in *Listeria monocytogenes*. *Cell*, 139(4):770–779.

Lu, K.-J., Chang, C.-W., Wang, C.-H., Chen, F. Y., Huang, I. Y., Huang, P.-H., Yang, C.-H., Wu, H.-Y., Wu, W.-J., Hsu, K.-C., Ho, M.-C., Tsai, M.-D., and Liao, J. C. (2023). An ATP-sensitive phosphoketolase regulates carbon fixation in cyanobacteria. *Nature Metabolism*, 5(7):1111–1126.

Maini, R. and Aggarwal, H. (2009). Study and comparison of various image edge detection techniques. *International Journal of Image Processing (IJIP)*, 3(1):1–11.

Mandal, M., Boese, B., Barrick, J. E., Winkler, W. C., and Breaker, R. R. (2003). Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell*, 113(5):577–586.

Mandal, M. and Breaker, R. R. (2004). Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nature Structural & Molecular Biology*, 11(1):29–35.

Mandal, M., Lee, M., Barrick, J. E., Weinberg, Z., Emilsson, G. M., Ruzzo, W. L., and Breaker, R. R. (2004). A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science*, 306(5694):275–279.

Martini, B. A., Grigorov, A. S., Skvortsova, Y. V., Bychenko, O. S., Salina, E. G., and Azhikina, T. L. (2023). Small RNA MTS1338 configures a stress resistance signature in *Mycobacterium tuberculosis*. *International Journal of Molecular Sciences*, 24(9):7928.

McCown, P. J., Corbino, K. A., Stav, S., Sherlock, M. E., and Breaker, R. R. (2017). Riboswitch diversity and distribution. *RNA*, 23(7):995–1011.

McLaughlin, M., Fiebig, A., and Crosson, S. (2023). XRE transcription factors conserved in *Caulobacter* and φCbK modulate adhesin development and phage production. *PloS Genetics*, 19(11):e1011048.

Mironov, A. S., Gusarov, I., Rafikov, R., Lopez, L. E., Shatalin, K., Kreneva, R. A., Perumov, D. A., and Nudler, E. (2002). Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell*, 111(5):747–756.

Musich, R., Cadle-Davidson, L., and Osier, M. V. (2021). Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider. *Frontiers in Plant Science*, 12:657240.

Nawrocki, E. P. (2014). *Annotating functional RNAs in genomes using Infernal*, volume 1097 of *Methods in Molecular Biology*, chapter 9, pages 163–197. Humana Press, Totowa, NJ.

Nawrocki, E. P. and Eddy, S. R. (2013a). Computational identification of functional RNA homologs in metagenomic data. *RNA Biology*, 10(7):1170–1179.

Nawrocki, E. P. and Eddy, S. R. (2013b). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935.

Nou, X. and Kadner, R. J. (1998). Coupled changes in translation and transcription during cobalamin-dependent regulation of *btuB* expression in *Escherichia coli*. *Journal of Bacteriology*, 180(24):6719–6728.

Nou, X. and Kadner, R. J. (2000). Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. *Proceedings of the National Academy of Sciences*, 97(13):7190–7195.

Park, J., Jung, H., Mannaa, M., Kim, N., Han, G., Lee, S.-W., Lee, S.-W., and Seo, Y.-S. (2022). Genome-guided comparative *in planta* transcriptome analyses for identifying cross-species common virulence factors in bacterial phytopathogens. *Frontiers in Plant Science*, 13:1030720.

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A

complete domain-to-species taxonomy for bacteria and archaea. *Nature biotechnology*, 38(9):1079–1086.

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., and Hugenholtz, P. (2022). Gtdb: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic acids research*, 50(D1):D785–D794.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology*, 36(10):996–1004.

Puccio, T., An, S.-S., Schultz, A. C., Lizarraga, C. A., Bryant, A. S., Culp, D. J., Burne, R. A., and Kitten, T. (2022). Manganese transport by *Streptococcus sanguinis* in acidic conditions and its impact on growth in vitro and in vivo. *Molecular Microbiology*, 117(2):375–393.

Ray, J. C., Smirnov, A., Maurakis, S. A., Harrison, S. A., Ke, E., Chazin, W. J., Cornelissen, C. N., and Criss, A. K. (2022). Adherence enables *Neisseria gonorrhoeae* to overcome zinc limitation imposed by nutritional immunity proteins. *Infection and Immunity*, 90(3):e00009–22.

Rinke, C., Chuvochina, M., Mussig, A. J., Chaumeil, P.-A., Davín, A. A., Waite, D. W., Whitman, W. B., Parks, D. H., and Hugenholtz, P. (2021). A standardized archaeal taxonomy for the genome taxonomy database. *Nature microbiology*, 6(7):946–959.

Roberts, J. W. (2019). Mechanisms of bacterial transcription termination. *Journal of Molecular Biology*, 431(20):4030–4039.

Rosinski-Chupin, I., Sauvage, E., Fouet, A., Poyart, C., and Glaser, P. (2019). Conserved and specific features of *Streptococcus pyogenes* and *Streptococcus agalactiae* transcriptional landscapes. *BMC Genomics*, 20:1–15.

Rosinski-Chupin, I., Soutourina, O., and Martin-Verstraete, I. (2014). Riboswitch discovery by combining RNA-seq and genome-wide identification of transcriptional start sites. In *Methods in Enzymology*, volume 549, pages 3–27. Elsevier.

Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A., and Palsson, B. O. (2019). The *Escherichia coli* transcriptome mostly consists of independently regulated modules. *Nature Communications*, 10(1):5536.

Sedlyarova, N., Shamovsky, I., Bharati, B. K., Epshtein, V., Chen, J., Gottesman, S., Schroeder, R., and Nudler, E. (2016). sRNA-mediated control of transcription termination in *E. coli*. *Cell*, 167(1):111–121.

Seeger, C., Dyrhage, K., Näslund, K., and Andersson, S. G. (2023). *Apilactobacillus kunkeei* releases RNA-associated membrane vesicles and proteinaceous nanoparticles. *Microlife*, 4:uqad037.

Stav, S., Atilho, R. M., Mirihana Arachchilage, G., Nguyen, G., Higgs, G., and Breaker, R. R. (2019). Genome-wide discovery of structured noncoding RNAs in bacteria. *BMC Microbiology*, 19(1):1–18.

Sudarsan, N., Cohen-Chalamish, S., Nakamura, S., Emilsson, G. M., and Breaker, R. R. (2005). Thiamine pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine. *Chemistry & Biology*, 12(12):1325–1335.

Sudarsan, N., Wickiser, J. K., Nakamura, S., Ebert, M. S., and Breaker, R. R. (2003). An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes & Development*, 17(21):2688–2697.

Sun, E. I. and Rodionov, D. A. (2014). Computational analysis of riboswitch-based regulation. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1839(10):900–907.

Tan, G., Opitz, L., Schlapbach, R., and Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports*, 9(1):2856.

Vander Griend, J. A., Isenberg, R. Y., Kotla, K. R., and Mandel, M. J. (2024). Transcriptional pathways across colony biofilm models in the symbiont *Vibrio fischeri*. *mSystems*, 9(1):e00815–23.

Wachsmuth, M., Findeiß, S., Weissheimer, N., Stadler, P. F., and Mörl, M. (2013). *De novo* design of a synthetic riboswitch that regulates transcription termination. *Nucleic Acids Research*, 41(4):2541–2551.

Wang, D., Yu, J. M., Dorosky, R. J., Pierson III, L. S., and Pierson, E. A. (2016). The phenazine 2-hydroxy-phenazine-1-carboxylic acid promotes extracellular DNA release and has broad transcriptomic consequences in *Pseudomonas chlororaphis* 30–84. *PLoS One*, 11(1):e0148003.

Weiss, A., Lopez, C. A., Beavers, W. N., Rodriguez, J., and Skaar, E. P. (2021). *Clostridioides difficile* strain-dependent and strain-independent adaptations to a microaerobic environment. *Microbial Genomics*, 7(12):000738.

Welz, R. and Breaker, R. R. (2007). Ligand binding and gene control characteristics of tandem riboswitches in *Bacillus anthracis*. *RNA*, 13(4):573–582.

Wickiser, J. K., Winkler, W. C., Breaker, R. R., and Crothers, D. M. (2005). The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Molecular Cell*, 18(1):49–60.

Wimmer, F., Englert, F., Wandera, K. G., Alkhnbashi, O. S., Collins, S. P., Backofen, R., and Beisel, C. L. (2024). Interrogating two extensively self-targeting type I CRISPR-Cas systems in *Xanthomonas albilineans* reveals distinct anti-CRISPR proteins that block DNA degradation. *Nucleic Acids Research*, 52(2):769–783.

Winkler, W., Nahvi, A., and Breaker, R. R. (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, 419(6910):952–956.

Winkler, W. C., Nahvi, A., Sudarsan, N., Barrick, J. E., and Breaker, R. R. (2003). An mRNA structure that controls gene expression by binding S-adenosylmethionine. *Nature Structural & Molecular Biology*, 10(9):701–707.

Yao, Z., Barrick, J., Weinberg, Z., Neph, S., Breaker, R., Tompa, M., and Ruzzo, W. L. (2007). A computational pipeline for high-throughput discovery of *cis*-regulatory noncoding RNA in prokaryotes. *PLoS Computational Biology*, 3(7):e126.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36.

Yu, S.-H., Vogel, J., and Förstner, K. U. (2018). ANNOgesic: a Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. *Gigascience*, 7(9):giy096.