Published online 31 May 2019

Nucleic Acids Research, 2019, Vol. 47, No. 15 ***e88***
doi: 10.1093/nar/gkz485

APERO: a genome-wide approach for identifying bacterial small RNAs from RNA-Seq data

Simon Leonard[1], Sam Meyer[1], Stephan Lacour[2], William Nasser[1], Florence Hommais[1,*] and Sylvie Reverchon[1,*]

[1]Université de Lyon, INSA-Lyon, Université Claude Bernard Lyon1, CNRS UMR5240, Laboratoire de Microbiologie, Adaptation, Pathogénie, 11 avenue Jean Capelle, F-69621 Villeurbanne, France and [2]Univ. Grenoble Alpes, CNRS, Inria, LiPhy (UMR5588), 38000 Grenoble, France

Received September 18, 2018; Revised May 06, 2019; Editorial Decision May 17, 2019; Accepted May 20, 2019

ABSTRACT

Small non-coding RNAs (sRNAs) regulate numerous cellular processes in all domains of life. Several approaches have been developed to identify them from RNA-seq data, which are efficient for eukaryotic sRNAs but remain inaccurate for the longer and highly structured bacterial sRNAs. We present APERO, a new algorithm to detect small transcripts from paired-end bacterial RNA-seq data. In contrast to previous approaches that start from the read coverage distribution, APERO analyzes boundaries of individual sequenced fragments to infer the 5′ and 3′ ends of all transcripts. Since sRNAs are about the same size as individual fragments (50–350 nucleotides), this algorithm provides a significantly higher accuracy and robustness, e.g., with respect to spontaneous internal breaking sites. To demonstrate this improvement, we develop a comparative assessment on datasets from *Escherichia coli* and *Salmonella enterica*, based on experimentally validated sRNAs. We also identify the small transcript repertoire of *Dickeya dadantii* including putative intergenic RNAs, 5′ UTR or 3′ UTR-derived RNA products and antisense RNAs. Comparisons to annotations as well as RACE-PCR experimental data confirm the precision of the detected transcripts. Altogether, APERO outperforms all existing methods in terms of sRNA detection and boundary precision, which is crucial for comprehensive genome annotations. It is freely available as an open source R package on https://github.com/Simon-Leonard/APERO

INTRODUCTION

In recent years, small non-coding RNAs (sRNAs) have been identified as major regulators of gene expression in all three domains of life (1–3). In bacteria, quantitative comparisons of sRNA-based and protein-based gene regulation suggest that sRNAs allow cells to switch quickly yet reliably between distinct states, while protein regulators are better suited for quantitative adjustment of protein level (4). Accordingly, sRNAs have largely been found in circuits responding to strong environmental cues (e.g. extreme nutrient limitation, stress response) (5). Most sRNAs range from 50 to 350 nucleotides in length, are generally highly structured, and alter the translation of mRNAs and/or modulate transcript turnover through base-pairing with their mRNA targets (6–9). They can be broadly divided into two categories: (i) *cis*-antisense sRNAs, expressed from the strand opposite to their target gene (10,11) and (ii) *trans*-acting sRNAs, expressed either from intergenic regions (12) or from untranslated regions close to a CDS (13,14).

The development of high-throughput RNA sequencing opened the way for genome-wide detection of sRNAs. However, their small length in comparison to mRNAs and their distinct properties give rise to specific identification difficulties. As a result, the number of small transcripts detected in bacteria is extremely variable, ranging from several hundred (15,16) to several thousand (17,18) depending on the species but also strongly on the RNA library preparation protocol and analysis method used, and the reported small transcript lengths are equally variable. In view of their importance in gene regulation, the robust detection and mapping of sRNAs thus remains a significant problem in microbial genetics. The definition of their boundaries (in both 5′ and 3′ directions) is especially delicate, as 5′ ends can be processed by RNase E and be capped with a monophosphate instead of a triphosphate (19), whereas 3′ ends are often rapidly degraded by polynucleotide phosphorylase (PNPase) (20). And yet the detected boundaries are then the main input given to functional annotation algorithms, which look for sRNA targets and predict their biological function. These further steps thus crucially rely on the precision of this analysis.

*To whom correspondence should be addressed. Tel: +33 472 43 85 68; Fax: +33 472 43 26 86; Email: sylvie.reverchon-pescheux@insa-lyon.fr
Correspondence may also be addressed to Florence Hommais. Email: florence.hommais@univ-lyon1.fr

Almost all existing sRNA detection algorithms from RNA-Seq data are based on the same principle. Single-end or paired-end reads are first mapped to a reference genome, and then converted into a genome coverage distribution, from which sRNAs are defined as regions with sufficient and/or uniform coverage. This method was initially inspired by algorithms designed for eukaryotic sRNAs (miRNA, siRNA), which are much smaller in size (20–25 nucleotides), and where the coverage remains relatively uniform along the transcript. In contrast, bacterial sRNAs are longer and exhibit very strong coverage inhomogeneities. The latter constitute a well-known problem in usual RNA-Seq analysis, but they are even stronger in sRNAs, because (i) the latter are highly structured and consequently display stronger spontaneous breaking/cleavage sites and (ii) they can be specifically processed by RNase E, which exhibits strong sequence selectivity (21). Consequently, algorithms searching for uniform coverages, such as sRNA-Detect (22), tend to cut the detected transcripts at their spontaneous cleavage sites, resulting in excessive numbers of small transcripts (see Results). Other algorithms usually use a coverage threshold to define a small transcript or find its boundaries (23–26), but here also, spontaneous breaking/cleavage sites tend to produce coverage drops that are difficult to robustly distinguish from the actual transcript end. Recently, ANNOgesic (27) obtained a gain in robustness by tolerating several nucleotides below the chosen coverage threshold. Whatever the algorithm, utilization of coverage distribution certainly contributes in the heterogeneity of results reported on bacterial small transcripts.

Our algorithm differs from all those previously mentioned in that it avoids the step of converting the reads into a coverage distribution where significant information is lost, resulting in the issue mentioned above. Instead, the extensions of identified small transcripts are directly analysed from the sequenced pairs of reads, taking full advantage of the precision of paired-end sequencing. Indeed, in contrast to mRNAs, the size of sRNAs is of the same order as the sequenced fragments; in many cases, a few fragments are thus sufficient to cover the entire sRNA in length, and allow locating its boundaries at high resolution. In the aforementioned example of a spontaneous breaking site inside a sRNA, even if the coverage signal drops sharply, a small number of fragments extending on both sides of the site is sufficient to prove that the sRNA extends to their respective ending points or even further. The rationale of the algorithm thus consists in (a) detecting 5′ ends of small transcripts and (b) several iterative extensions of the transcripts in the 5′ → 3′ direction based on the previous operation, where the conserved information of sequenced fragments start/end pairing provides an increased statistical power compared to methods based on the coverage where this information is lost, hence the name of the algorithm (**A**nalysis of **P**aired-**E**nd **R**NA-Seq **O**utput).

It must be noted that one existing method (DETR'PROK) also directly deals with reads rather than with coverage (28), but since it only considers single-end reads, the benefit of our approach is lacking, and the results are indeed very comparable to coverage-based methods (see below). In fact, several of the latter were developed when paired-end data were still scare, and the improvement offered by APERO was therefore not yet technically relevant. Existing algorithms were also tested either on RNA-Seq data where the library preparation includes a fragmentation step (thereby adding an additional and artificial source of noise), or more recently on specific datasets of non-fragmented, size-selected RNAs. Although our method can also be applied on both types of data, it is more specifically suited to the latter case, where many small transcripts should then be directly sequenced as intact fragments with high precision, even when partial spontaneous fragmentation results in inhomogeneous coverage values.

In the following, we present the APERO algorithm and apply it on RNA-Seq datasets from *Escherichia coli* and *Salmonella enterica*. To evaluate its performance, we develop a comparative assessment between APERO and seven existing methods (Rockhopper (24), DETR'PROK (28), TLA from RNA-eXpress (23), sRNA-Detect (22), the two custom-made scripts developed by Gómez-Lozano *et al.* (25) and Nuss *et al.* (26), and ANNOgesic (27) using sets of known sRNAs annotated in these two species. We show that APERO outperforms all previous methods in terms of sRNA detection and boundary precision. In addition, it is able to detect different isoforms of a single sRNA. We also analyze a new set of RNA-Seq data from the phytopathogenic bacterium *Dickeya dadantii*, in presence or absence of Terminator EXonuclease (TEX) treatment that enriches primary transcripts. APERO allows detecting 1703 primary small transcripts, including intergenic RNAs, 5′ UTR or 3′ UTR-derived RNA products and antisense RNAs. 23 of which were already annotated by similarity with *E. coli*. Eight sRNAs whose boundaries are annotated differently depending on the algorithm were analyzed using RACE-PCR, which experimentally validated the boundaries detected by APERO and confirmed the accuracy of our algorithm in identifying novel full-length small transcripts.

MATERIALS AND METHODS

APERO determines a set of small transcripts from paired-end RNA-seq data, preferably obtained with size-selected but un-fragmented RNAs. As input, it requires a sequence alignment file in BAM format, which is filtered using bitwise flags in order to select paired-reads that are correctly oriented and positioned with respect to each other. APERO is written in R and depends on the Rsamtools and Reshape2 packages to extract the genomic positions of the sequenced fragments, as well as on the Snowfall package for parallel computations. We also implemented the algorithm as a tool on a local Galaxy server instance (bioinfo.insa-lyon.fr), where users can upload data and run the workflows without any software installation. APERO is composed of two modules: the first one infers small transcript 5′ ends from the ends of mapped fragments (i.e. paired reads), and the second then determines the length of small transcripts by extending them iteratively in the 5′→ 3′ direction following sequenced fragments (Figure 1). These modules are described qualitatively at the beginning of the Results section; in the two following paragraphs, we provide the precise definitions of the parameters and statistical methods used. The

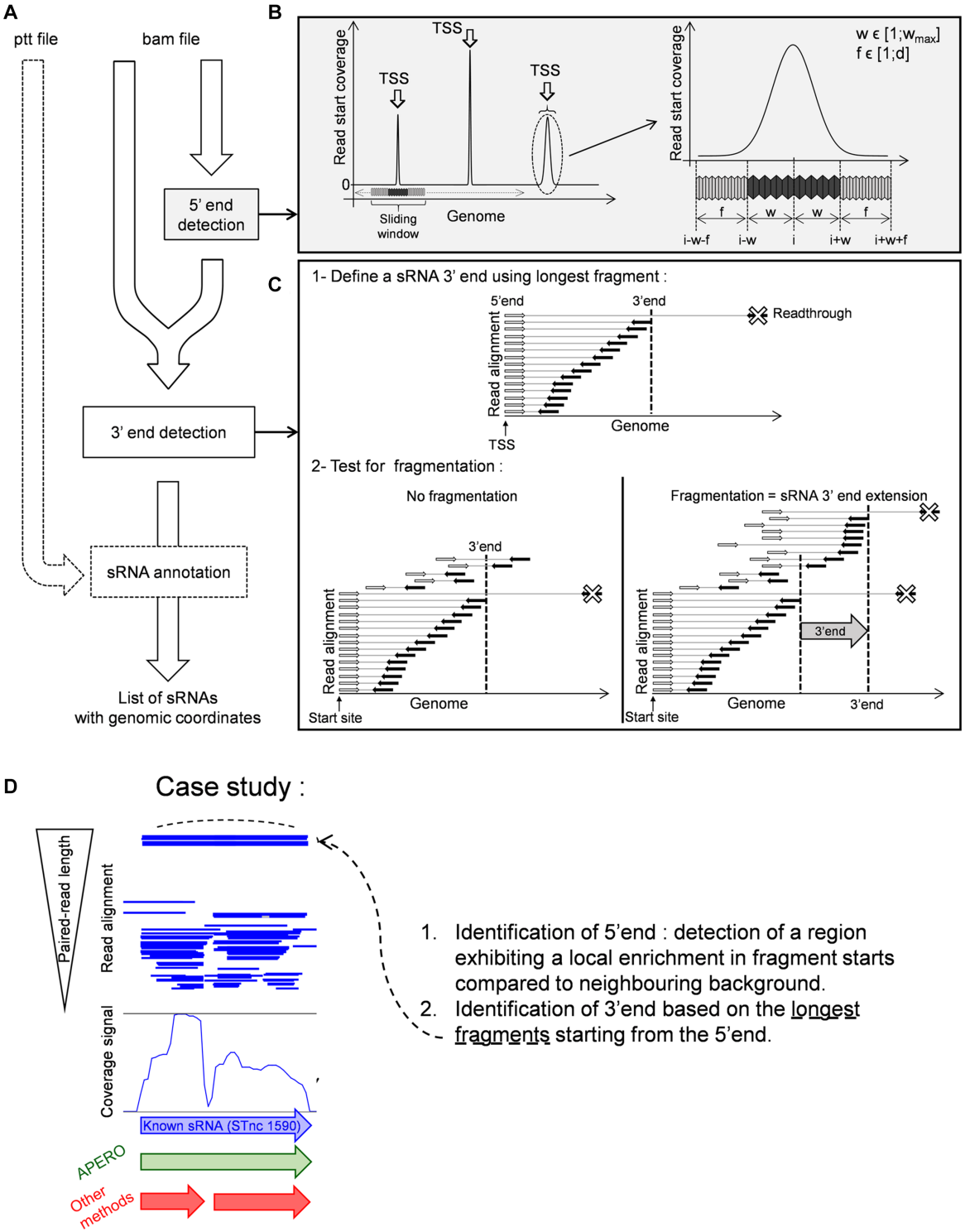


Figure 1. Description of the APERO algorithm. (**A**) Workflow. Starting from an input BAM sequence alignment file, a first module (**B**) detects the 5′ end of small transcripts (hereafter called *start site*), and a second module (**C**) then identifies the 3′ end of small transcripts. The output is a list of annotated transcripts with their genomic coordinates and further information, including small transcript annotation if an optional genomic annotation file (ptt format) is provided. (B) A sliding window is used to evaluate read start coverage at each position. A *start site* (or *start* region) is identified as a region exhibiting a local enrichment in fragment starts, compared to neighboring background. w is the sliding window length, f is the length of the region chosen to calculate the neighboring background. The parameters w_{max} and d (upper values) are fixed by the user, and control the spatial resolution of the algorithm (see materials and methods). (C) The 3′-ends of transcripts are computed by iteratively extending the transcripts based on the longest fragments observed at the current position. Starting from the *start site* (1), a first putative 3′ end is tested (2) by counting the number of fragments extending beyond this position. If this number is significant (right panel), a new putative 3′ end is computed from these fragments, and tested iteratively. Analyzing individual paired-end fragments rather than a read coverage improves the robustness and statistical power of the analysis, especially with respect to spontaneous RNA degradation. (**D**) APERO screen shot of STnc 1590 detection.

remaining paragraphs describe the datasets, benchmarking methods, and experimental procedures.

Determination of 5′ ends

The first APERO module aims at identifying 5′ ends of small transcripts, i.e. transcriptional start sites or 5′ processed ends (Figure 1B). The number of read starts is computed at each position along both strands of the genome, and 5′ ends (hereafter called *start sites*) are defined as regions exhibiting local enrichment in read starts compared to neighboring genomic positions. 5′ ends are not always defined at a single position, but can extend over a few (usually up to three) nucleotides. The user can set the maximal width $w_{\max}$, and *start sites* of width $1 \leq w \leq w_{\max}$ are then iteratively analyzed by the algorithm. For a current genomic position i and a current *start site* width w, the number of read starts in the considered region ($i \pm w$) is compared to that of neighboring regions of increasing sizes $1 \leq f \leq d$, where d is the second parameter provided by the user. With this choice, d represents the minimal distance for distinguishing two separate 5′ ends, i.e. the spatial resolution of the algorithm.

The local enrichment $E_{i,w,f}$ is defined as:

$$E_{i,w,f} = \frac{\frac{\sum_{j\epsilon[i-w-f,i-w-1]} c_j + \sum_{j\epsilon[i+w+1,i+w+f]} c_j}{2f}}{\frac{\sum_{j\epsilon[i-w,i+w]} c_j}{2w+1}}$$

where c_j is the number of reads starting at position j. A region $i \pm w$ is then considered as a *start site* if $E_{i,w,f}$ is above a threshold value $E_{\min}$ whatever the f value. If several *start sites* are identified at the same position i (but with different widths w), only the narrowest (and most precise) *start site* is retained. *Start sites* identified at different positions i within the same window are merged and define a *start* cluster. The enrichment threshold value was chosen equal to $E_{\min} = 1/(2d)$.

The output of this APERO module is a list of 5′ ends containing the central position of the *start* region, its half-width w, the strand from which the transcription is initiated and the number of reads starting from this *start* region. This output file can optionally be filtered according to the number of read starts depending on the dataset read depth. In the applications presented here, we used a filter equal to (20 × total number of reads)/genome size to eliminate weakly expressed transcripts.

Determination of 3′ ends

The second APERO module determines the length of small transcripts (i.e. their 3′ end) by extending them iteratively in the 5′→ 3′ direction from the 5′ end, as long as sequenced fragments are present in sufficient numbers (Figure 1C). Fragments starting at a 5′ end *start site* are first selected and ranked according to their length. In order to rule out potentially irrelevant transcriptional readthrough, the 1% longest fragments are discarded. We call e the 3′ end of the remaining longest fragment, C_e the coverage value at position e, $C_{start,e}$ the total number of fragments between *start* and e, and $L_{start,e}$ the distance between the *start site* and e. The coverage ratio at position e, $F_{start,e}$, is then defined as:

$$F_{start,e} = \frac{C_e}{\frac{C_{start,e}}{L_{start,e}}}$$

$F_{start,e}$ represents the number of fragments that overlap e, normalized by the expression strength of the whole transcript. If $F_{start,e}, < F_{\min}$, e is then considered as the 3′ end of the transcript. If $F_{start,e} \geq F_{\min}$, the small transcript is extended further (see Figure 1C): the longest fragment overlapping position e is identified using the procedure described above, and its end e' is tested as a new putative 3′ end. This operation is repeated until the $F_{start,e'}$ value is below the threshold $F_{\min}$.

The threshold value $F_{\min}$ controls the tendency of the algorithm to increase transcript lengths (with lower values resulting in longer transcripts). Tests showed that the optimal value depends on the dataset, maybe due to differences in library preparation, protocols, sequencing depth, etc. We determined this optimal value by analyzing the distributions of $F_{start,e}$ after one iteration of the algorithm on the experimentally validated sRNAs: threshold values $F_{\min} = 3$ (for *S. enterica*) and $F_{\min} = 6$ (for *E. coli*) efficiently discriminated between the transcripts where the 3′ end was already found after a single iteration, and those which required further iterations (Supplementary Figure S1). These values correspond approximately to the first quartile of the $F_{start,e}$ distribution for both datasets; consequently, a default $F_{\min}$ value is computed automatically in the same way when handling a new dataset. This procedure also gives a value $F_{\min} = 3$ for the *D. dadantii* dataset.

Annotation/Classification

As shown in Figure 1, if the user provides an annotation file of the reference genome to APERO (in ptt format), the detected small transcripts are classified according to their position with respect to annotated CDSs and their strand (Supplementary Figure S2). Orphan RNAs are transcribed from intergenic regions and could be regulatory sRNAs. RNAs transcribed on the same strand as the CDS are classified as: Primary RNA (P) when located in the 250 nucleotides upstream of the CDS without overlapping it; 5′-UTR and 3′-UTR RNAs when overlapping the start and stop codon respectively. RNAs transcribed on the opposite strand are classified as: Ai (antisense internal) when transcribed inside CDS; Div (divergent) when starting in the 250 nucleotides upstream of the CDS; 5′Ai and 3′Ai RNAs when overlapping the start and stop codon respectively.

Size-selected RNA libraries

We considered three paired-end sequencing data obtained from size-selected and non-fragmented RNAs: one from *S. enterica* (accession number SRX1036363) (18) and two new datasets from *E. coli* (accession number SRX4670654) and *D. dadantii* (accession number SRX4664132).

For *S. enterica*, the total RNAs from the log-phase were isolated using Trizol® (Life Sciences) standard manufacture protocol. The small RNA fractions were separated from the large fractions using the RNeasy MinElute

Cleanup Kit (Qiagen) and submitted to Otogenetics (http://www.otogenetics.com/) for commercial RNA-Seq on an Illumina MiSeq genome sequencer according to the Illumina TruSeq Small RNA Sample Prep protocol.

For *E. coli*, the total RNAs from the early stationary phase were isolated using the hot-phenol procedure and RNAs of size ranging from 50 and 200 nucleotides were extracted from a denaturing polyacrylamide electrophoresis gel. The strand-specific library was prepared and sequenced by Fasteris SA (https://www.fasteris.com/dna/) according to the Illumina TruSeq Small RNA Sample Prep protocol.

For *D. dadantii*, the total RNAs from previously described conditions (29) were extracted using the frozen acid-phenol procedure (30) and the small RNA fractions (<500 nucleotides) were separated from the large fractions using the RNeasy MinElute Cleanup Kit (Qiagen). The strand-specific libraries were prepared and sequenced by Vertis Biotechnologie AG (http://www.vertis-biotech.com/) according to the Illumina TruSeq Small RNA Sample Prep protocol.

Sequenced reads were trimmed using Trim Galore (on a Galaxy server, version 0.4.3.1), checked for sufficient quality with FastQC (bioinformatics.babraham.ac.uk/projects/fastqc/), filtered, and mapped to respective reference genomes (*S. enterica* serovar Typhimurium strain SL1344, NC_016810.1; *E. coli* str. K-12 substr. MG1655, NC_000913.3; *D. dadantii* 3937, NC_014500.1) using Bowtie2 (31) with the local alignment mode (Galaxy Version 2.3.2.2).

These RNA-seq datasets were chosen for benchmarking because (i) the size-selection and absence of fragmentation of RNAs should remove a noise-generating step in the library preparation and improve the statistical coverage of small transcripts; (ii) many sRNAs were experimentally validated in *S. enterica* and *E. coli*; (iii) they exhibit differences in sequencing depth and read length: 2 × 2.6 million mapped paired-end 100-bp reads for *S. enterica*, 2 × 15.5 million mapped paired-end 125-bp reads for *E. coli*, 2 × 40 million paired-end 75-bp reads for *D. dadantii*. Furthermore, *D. dadantii* RNA-seq was performed with (+TEX) and without (–TEX) Terminator Exonuclease treatment, resulting in an enrichment of reads starting at Transcriptional Start Site positions (TSS) in the +TEX dataset.

Performance evaluation of sRNA detection algorithms

Predictions of APERO were compared to those of different programs dedicated to the identification of small transcripts in bacteria (listed in Results). All methods (except Rockhopper) were used with the same minimum height/coverage parameter, fixed to (20 × total number of reads)/genome size. This value approximately matches that suggested by authors for the sequencing depth of the *S. enterica* dataset (15,32). To evaluate their performance, we computed two alternate estimators: recall and Jaccard index. Recall indicates the proportion of annotated small transcripts detected by a given approach (true positives divided by the total number of positive). Jaccard index is defined as the number of intersecting base pairs between an annotated sRNA and a detected small transcript divided by the number of base pairs in the union of the two sRNAs. All known sRNAs were considered as positives (true positives or false negatives). Since we could not evaluate the number of true negatives and false positives in our predicted small transcripts, no further statistical criteria were considered.

RACE-PCR validation of predicted small transcripts

The 5′ and 3′ ends of small transcripts were validated by RACE-PCR experiments as previously described (17). RNAs were extracted from *D. dadantii* bacterial cells grown in exponential phase in minimal medium in the presence of sucrose as carbon source. Primers listed in Supplementary File S1 are used for reverse transcription and PCR amplification. The RACE-PCR products were cloned into the pGEMT-easy. Cloned DNA fragments were then sequenced using the M13 primers (Invitrogen).

RESULTS

Detection and mapping of small transcripts from paired-end reads

Paired-end RNA-seq data from *S. enterica* (18) and from *E. coli* (this work) were used as training datasets to compare APERO with other methods and evaluate their performance in the identification of known sRNAs. In all RNA libraries presented in this paper, small transcripts were separated according to their size and sequenced without any fragmentation step. Technical details, as well as methods and parameters of the algorithm, can be found in Materials and Methods; in the following paragraphs, we only give a short description of the program before evaluating its performance on the data.

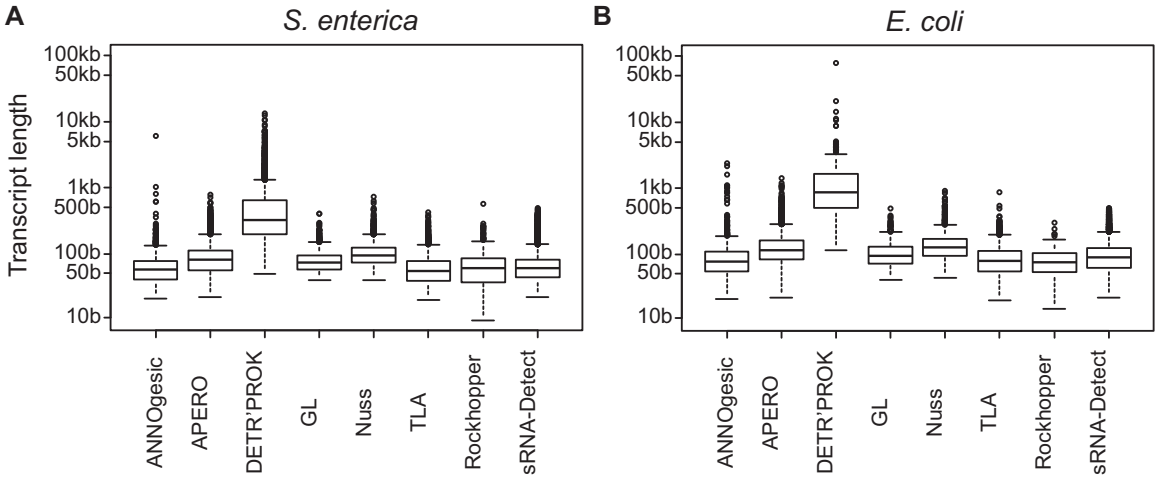
APERO is written in R and can be used on all operating systems after installing several R packages (Rsamtools, Reshape2, Snowfall). A webserver version is also available on a local instance of the Galaxy bioinformatics platform (bioinfo.insa-lyon.fr), which provides an installation-free access to all users. The workflow of the program is described in Figure 1, together with a screenshot of the STnc_1590 sRNA from *S. enterica* (Figure 1D). The required input is a BAM sequence alignment file from a RNA-Seq dataset, and the output is a text file giving the list of detected small transcripts, with different information fields (genomic coordinates, strand, value of the width parameter used, intensity of small transcript start and number of iterations of the second module). Optionally, the user can provide a genome annotation (in ptt format), which is then used to classify each small transcript with respect to neighboring or overlapping CDS. Detected small transcripts could correspond to intergenic RNAs, 3′UTR RNAs, 5′ UTR-derived RNA products or antisense RNAs.

APERO is composed of two separate modules. The first one (Figure 1B) is dedicated to the identification and mapping of the 5′ ends of small transcripts, *i.e.* the putative transcriptional start sites (TSSs) or 5′ processed ends. Inspired by previous algorithms specifically dedicated to TSS detection (26,33), 5′ ends are defined as regions exhibiting a local enrichment in read starts compared to neighbor regions. Two common difficulties are encountered: (i) due to wobbling of the RNA polymerase, TSSs peaks are not always defined sharply, but can extend over several nucleotides; (ii)

adjacent peaks distant of several nucleotides are sometimes observed, reflecting alternate 5′ *starts* of the same transcript which prevent the identification of individual peaks. The analysis procedure was designed to handle these issues and optimize precision, by varying the scanning window sizes and favoring the most well-defined *start sites*. The user can play on two parameters illustrated on Figure 1B: the maximal accepted width of a *start* peak (w_{max}), and the minimal distance between separate *start sites* (i.e. the spatial resolution of the detection method, *d*). We tested different values reported in the literature (34), resulting in differences in the number of detected *start sites*, as shown in Supplementary Figure S1. In practice, it is convenient to use the same value for these two parameters, which characterize the spatial precision of the method. With a value of 10 nucleotides (retained for all upcoming analyses), 16 123 5′-ends were identified from the *S. enterica* dataset and 10 991 from the *E. coli* dataset. As a benchmark of APERO's ability to accurately detect 5′ ends, we first applied its 5′ module on previously published +TEX sequencing data of total (small + large) RNAs from *E. coli* (GSE55199) (35) and *Salmonella* transcripts (GSE49829) (32), where TSSs were already analyzed. For *E. coli*, 57% of TSSs identified by Thomason *et al.* (35) are also identified by APERO within 10 nucleotides resolution (among those, 95% are identified within 3 nucleotides resolution) and for *Salmonella* 80% are in common with those described by Kröger *et al*. (32). These controls show that the first module of APERO can safely be applied to determine the 5′ ends of small transcripts (from new datasets). An advantage of our method compared to previous ones is that it can be applied either on usual RNA-Seq datasets or on RNAs treated with Terminal Exonuclease (TEX) for primary transcript enrichment. In the former case, the user would be unable to distinguish a TSS from a 5′ processed end. In the latter case, transcript starts are found with increased precision; in both cases, the precision is better than that of previous sRNA detection algorithms based on read coverage, as shown below. An interesting example in terms of 5′ end detection is that of sRNAs exhibiting different isoforms. For example, ArcZ, MicL and RaiZ sRNAs undergo maturation to generate their functional isoforms. In *E. coli,* the primary 120-nucleotides ArcZ transcript is processed to form a low-abundance 88-nucleotides form and a stable 55-nucleotides form derived from the 3′ end of the primary transcript (36), MicL is synthesized as a 308-nucleotides primary transcript that is processed to an 80-nucleotides form (37). In *S. enterica,* two major RaiZ species, a 160-nucleotides form and a 122-nucleotides processed sRNA, were detected (38). As illustrated in Supplementary Figure S3, APERO correctly identified both TSS and 5′ processed end of these sRNAs from datasets obtained without TEX treatment and is the only algorithm to do so.

The second module of APERO starts from the identified 5′ ends, and localizes the 3′ end of each small transcript, as shown in Figure 1C. Theoretically, since the small transcripts were not fragmented, the paired-end sequencing should immediately allow identifying the 3′ boundary from the aligned reads. However, due to experimental bias, spontaneous fragmentation and degradation of 3′ ends by polynucleotide phosphorylase (PNPase), small transcript lengths are generally underestimated by this approach. Starting from an identified 5′ end, the method thus consists in iteratively extending the identified small transcript by locating the 3′ end of the longest fragments (if their number is sufficient). This end position can then be either the actual 3′ end of the transcript, or a mere intermediate point or spontaneous breaking site within the transcript; to distinguish between these two scenarios, the program computes the number of fragments that overlap this position and were not yet counted in the previous iteration (see Figure 1C). If this number is significant (as compared to the number of fragments starting at the 5′ end, i.e. the expression strength of this transcript), the transcript is further extended toward the 3′ end of these overlapping fragments and the operation is repeated; otherwise, the iteration stops and this point is defined as the 3′ end. The tendency of the algorithm to increase the transcript size is set by a threshold parameter, the optimal value of which can be obtained automatically from the analysis of the dataset provided by the user (see Materials and Methods and Supplementary Figure S1). The rationale of this method is that the typical size of s is of the same order as that of typical fragments, so that only a small number of iterations should be required to recover the full-length transcripts if spontaneous fragmentation is not too strong. This is indeed what we observe: most small transcripts are found with less than three iterations (85% in the *D. dadantii* dataset). We used several sources of information on transcript ends in *E. coli* to control the accuracy of the resulting 3′ ends, detected by APERO from the small transcript dataset (SRX4670654). We first compared the 3′ end positions of the 9 riboswitches, 15 attenuators and 312 transcriptional terminators described in EcoCyc, 66% of these are accurately detected by APERO, i.e. with a median distance less than 10 nucleotides (Supplementary File S3). The 3′ end positions of 49 sRNAs determined by Term-seq in *E. coli* (39) were also analyzed and 80% of these positions are detected by APERO with a distance less than 10 nucleotides (Supplementary File S3). In addition, we used data from *E. coli* obtained by SMRT-cappable-seq (40). 94% of 3′ ends identified by this approach were also detected by APERO with a median distance to 3′ end of 10 nucleotides (Supplementary Figure S4). These controls show that the module can accurately detected 3′ end RNA boundaries.

For the following analyses, we discarded internal small transcripts located inside CDSs, which could correspond to mRNA degradation products. Using the default parameters, we finally obtained 5347 and 4507 small transcripts for *S. enterica* and *E. coli* respectively. These large amounts of detected small transcripts by APERO in *S. enterica* and *E. coli* also include presumably cleaved, processed or attenuated UTRs (5′UTR represent 12–13% of the total, 3′UTR 15.5–24.6%), antisense RNAs (34.2–22%) and intergenic RNAs (which included orphans, divergent and primary small transcripts) that could be present in several isoforms (27.5–25%) (Supplementary Figure S5). Taking into account all these categories, the number of small transcripts detected by APERO is close to previous estimations (as an example >1000 antisense RNAs are detected in *S. enterica* and *E. coli*) (39,41,42). Obviously, some of these small transcripts could correspond to transcriptional noise and



	ANNOgesic	APERO	DETR'PROK	GL	Nuss	TLA	Rockhopper	sRNA-Detect
Total	1971	3302	2281	941	1442	4084	464	17662
Models	64	67	31	57	62	64	38	71
Models/Total	0.0325	0.0203	0.0136	0.0606	0.043	0.0157	0.0819	0.004
Recall	0.75	0.78	0.36	0.67	0.73	0.75	0.45	0.83

	ANNOgesic	APERO	DETR'PROK	GL	Nuss	TLA	Rockhopper	sRNA-Detect
Total	1456	2150	376	799	1201	2471	99	12477
Models	66	77	2	62	63	71	2	84
Models/Total	0.0453	0.0358	0.0053	0.0776	0.0524	0.0287	0.0202	0.0067
Recall	0.65	0.76	0.02	0.61	0.62	0.70	0.02	0.83

Figure 2. Comparison of small transcripts identified by different methods. Total number of intergenic and antisense small transcripts as well as their length distribution are indicated for *S. enterica* (**A**) and *E. coli* (**B**). Experimentally validated sRNAs (84 for *S. enterica*. 101 for *E. coli*) are used as models. The number of detected models is indicated, as well as the recall (proportion of models detected) for each method. GL = Method from Gómez-Lozano *et al.* (25); Nuss = Method from Nuss *et al.* (26); TLA = TLA from RNA-eXpress (23). Models are considered detected if Jaccard index is not null.

may be false positives. However, among small transcripts detected by APERO in *S. enterica*, the highest quartile (271 highly expressed small transcripts) are expressed more than the large majority of the Kröger *et al.* (32). Among these 271 small transcripts, 83% were not annotated by Kröger *et al.* (Supplementary file S4) and we annotated 58 of them by querying Rfam or by searching similarity to *E. coli* (attenuators, riboswitches, etc.). In addition, we evaluate the expression of these small transcripts in independent experiments by using the RNA-seq data from Kröger (Supplementary file S4, supplementary Figure S6): 64 are expressed at a higher level than adjacent genes. These highly expressed small transcripts are likely more than mere noise and may have important functions in gene regulation (43). Identified transcripts have a median length of 181 nucleotides in *E. coli* (111–410 nucleotides, first and third quartiles) and 114 nucleotides in *S. enterica* (73–231 nucleotides, first and third quartiles) (Supplementary Figure S7). Some candidates have lengths higher than 500 nucleotides and can be eliminated as suggested by previous studies (32,44). The observed difference in transcript size between the two species may be explained by (i) the use of different RNA extraction methods (Trizol®, hot phenol) that can affect RNA integrity (45), (ii) different size selection procedures (denaturing acrylamide gel or column) and (iii) RNAs from log phase (*S. enterica*) and from early stationary phase (*E. coli*), have different spontaneous fragmentation profiles due to the increased expression of PNPase in stationary phase (32).

Comparison of small transcript detection methods

We next compared APERO's performance with that of other available methods, including DETR'PROK (28), sRNA-Detect (22), TLA from RNA-eXpress (23), Rockhopper (24), ANNOgesic (27) and the two in-house GL (25) and Nuss (26) methods, using the entire *S. enterica* and *E. coli* datasets and a list of experimentally validated sRNAs (models) from these organisms (from ASAP/Ecogene, Ecocyc and RegulonDB databases for *E. coli* and BSRD database for *S. enterica*). We used these sRNAs as models for performance evaluation. Since several methods focused on sRNA identification from intergenic regions and antisense RNAs, only these two classes of small transcripts were analyzed (intergenic RNAs correspond to Orphans, Divergent and Primary RNAs in our annotation and antisense RNAs to Ai, 5′Ai and 3′Ai), i.e. 3302 remaining RNAs in *S. enterica* and 2150 in *E. coli*. The corresponding experimental sets contain 84 sRNA models in *S. enterica* and 101 in *E. coli* (Supplementary File S2).

We first compared the total number of small RNA candidates (intergenic and antisense) predicted by each method. This number varies widely according to the algorithm used, as described in Figure 2, from a hundred (Rockhopper) to almost 20 000 (sRNA-Detect), with APERO providing intermediate figures. Identified transcripts have similar median sizes of 75 ± 20 nucleotides for *S. enterica* and 100±25 nucleotides for *E. coli*, except those of DETR'PROK, which are considerably longer. The transcript lengths found by APERO are generally longer than those of the other methods, which is expected if the latter tend to incorrectly cut the transcripts at internal sites where the coverage drops, as will be seen below. Before looking at the transcript boundaries, we test how many experimentally validated sRNAs were detected by each method, by computing the recall (i.e. the fraction of detected sRNAs). Results are similar for both species, and APERO ranks second in terms of detec-

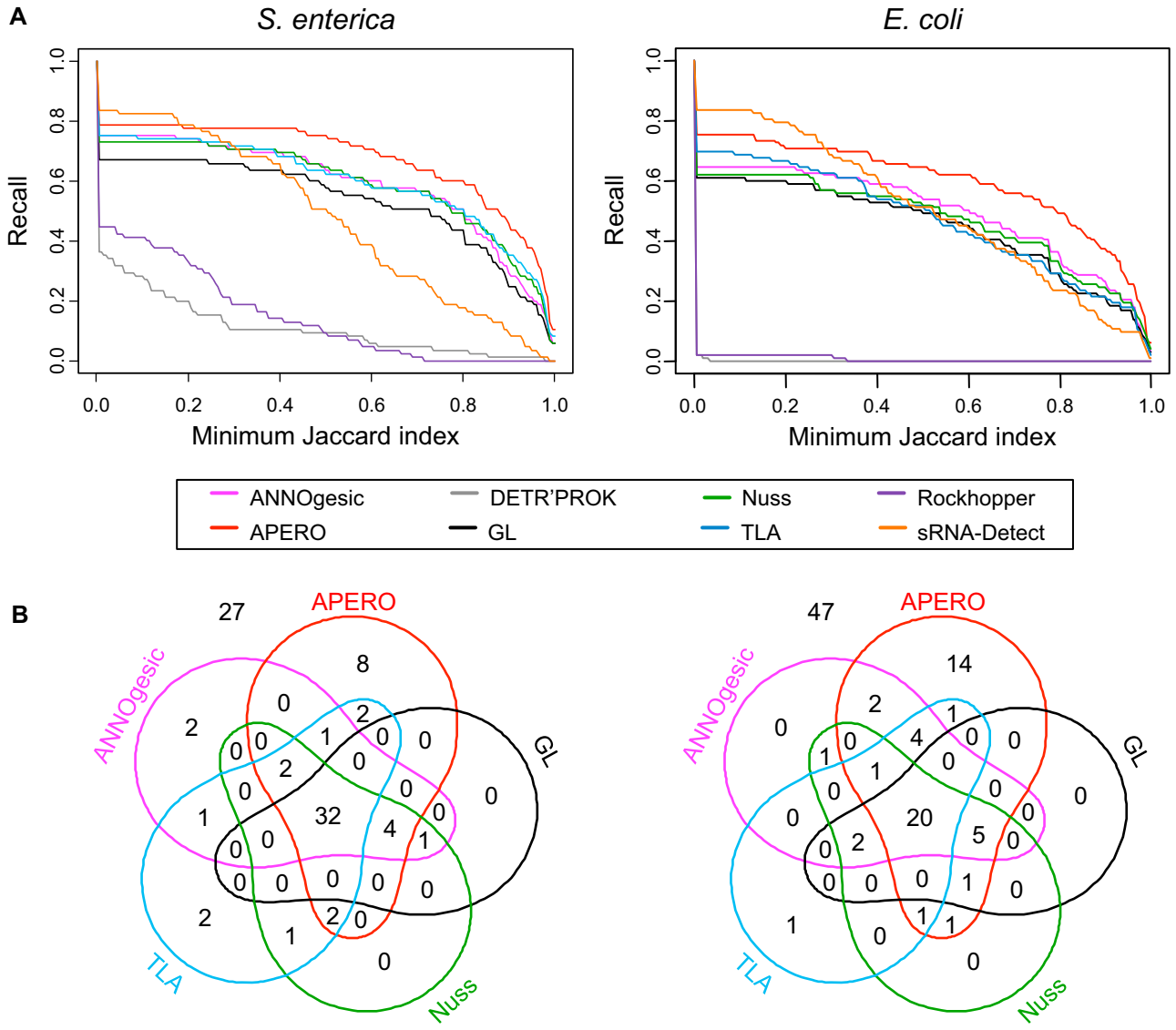


Figure 3. Performance assessment on different datasets. (**A**) Recall of each method is plotted against the Jaccard index (indicating the accuracy in boundary identification) for *S. enterica* (left) and *E. coli* (right) data. (**B**) Venn diagrams of models detected by APERO, ANNOgesic (27), TLA from RNA-eXpress (23), GL (25) and Nuss (26) algorithms, with a Jaccard index ≥0.8. 8 models in *S. enterica* (out of 84 models) and 14 models in *E. coli* (out of 101 entities) are detected only by APERO.

tion ability, behind sRNA-Detect. However, the very high number of small transcripts detected by the latter indicates that many of them could be false positives; since we do not have the possibility to quantify the prediction specificity on these datasets, we now test the accuracy of the predicted sRNA boundary positions to evaluate the precision of each method, which was the primary objective of APERO.

Accuracy in the sRNA boundaries detection

The mapping accuracy can be quantified from the distances between the annotated and predicted (5′ and 3′) ends of each transcript. Here, APERO clearly outperforms all other methods, with 5′ ends deviated from <20 nucleotides (third quartile) and 3′ ends from <9 nucleotides from the annotated positions for both bacterial species, whereas the four next best methods (ANNOgesic, GL and Nuss methods and TLA from RNA-eXpress) exhibit deviations between 26 and 46 nucleotides at 5′-ends and between 14 and 44 nucleotides at 3′-ends (Supplementary Figure S8). In addition, APERO makes significantly fewer very large errors (>100 nucleotides), especially for the 3′ end, which may typically correspond to erroneous internal stops. To quantify the combined prediction and accuracy of each method, we computed the Jaccard indexes for all model sRNAs. This index represents the proportion of the predicted sRNA length that matches the corresponding annotation (see Materials and Methods), and thus allows representing the precision of each method over all model sRNAs in a single plot. Figure 3A shows the recall values as a function of this number (i.e. retaining only the best predicted transcripts with an increasing threshold), which clearly demonstrates the accuracy of APERO. At a minimum Jaccard index of 0.8,

i.e. keeping only the most accurately predicted transcripts, it is the only method to recover well over 50% of the annotated sRNAs. Analyzing them individually (Figure 3B) confirms that 8 *S. enterica* and 14 *E. coli* annotated sRNAs (out of 84 and 101 models respectively) are detected by APERO alone, well over all other methods. Among the larger *S. enterica* dataset of 203 sRNAs identified by Kröger *et al.* (15,32) (not experimentally verified), 79 were detected by APERO with a Jaccard Index ≥0.8, more than any other method (including ANNOgesic, Supplementary File S2)

Taken together, these data demonstrate the preeminence of the APERO algorithm for full-length sRNA detection with accurate boundary identification and for detection of different isoforms of a single sRNA. A test of robustness was also carried by varying the parameters in APERO and other methods (Supplementary Figures S9 and S10), which demonstrated its lower sensitivity to parameter variations.

Identifying and annotating new small transcripts

We finally used the APERO method to identify sRNAs in *D. dadantii* from new sRNA-seq data (accession number SRX4664132). Using the same parameter values as above on a standard dataset of size-selected RNA (<500 nucleotides), we detected 3974 small transcripts, with a median length of 132 nucleotides (Figure 4A). We then tested the effect of Terminator EXonuclease (TEX) treatment on the detection of small transcripts. TEX was introduced to improve the detection of primary TSS positions in bacteria by selectively digesting fragments with monophosphate 5′ ends, and thus enriching triphosphate 5′ ends of RNA transcripts. Therefore, by using both +TEX and –TEX libraries, it is possible to discriminate between primary small RNA transcripts and processed small transcripts and to potentially identify the different isoforms of a single small transcript.

Since some small transcripts are known to be capped with a monophosphate 5′ end because of RNase E (19), we expected a partial loss of detected small transcripts, which is indeed observed. Using the +TEX library for TSS identification, and the –TEX library for the second step of the algorithm, only 1703 primary small transcripts are detected, with a similar median length of 141 nucleotides. However, the detected TSSs are defined much more precisely, and 70% of them even at a single nucleotide position (Figure 4B). Among the ∼2300 small transcripts not detected with TEX treatment, approximately 500 display the consensus motif of RNase E at their 5′ end and may thus correspond to bona fide processed sRNAs or to mere degradation products of longer RNAs (especially mRNAs), in which case the TEX treatment may improve the specificity of the analysis. To quantify this amount, we reasoned that such degradation products would then tend to overlap one another more frequently than actual small transcripts, and we compared the effect of merging overlapping small transcripts in the output lists obtained from both datasets (+TEX/–TEX or –TEX). Indeed, only a small fraction (14%) of the 1703 primary small transcripts identified using TEX are eliminated by this procedure, this fraction is considerably higher (50%) in the 3974 small transcripts obtained in absence of TEX. In the following, we therefore consider that TEX increases the accuracy of small transcript 5′ end detection and use the former dataset, keeping in mind that some true sRNAs produced by the endonucleolytic cleavage of mRNAs are discarded by this treatment.

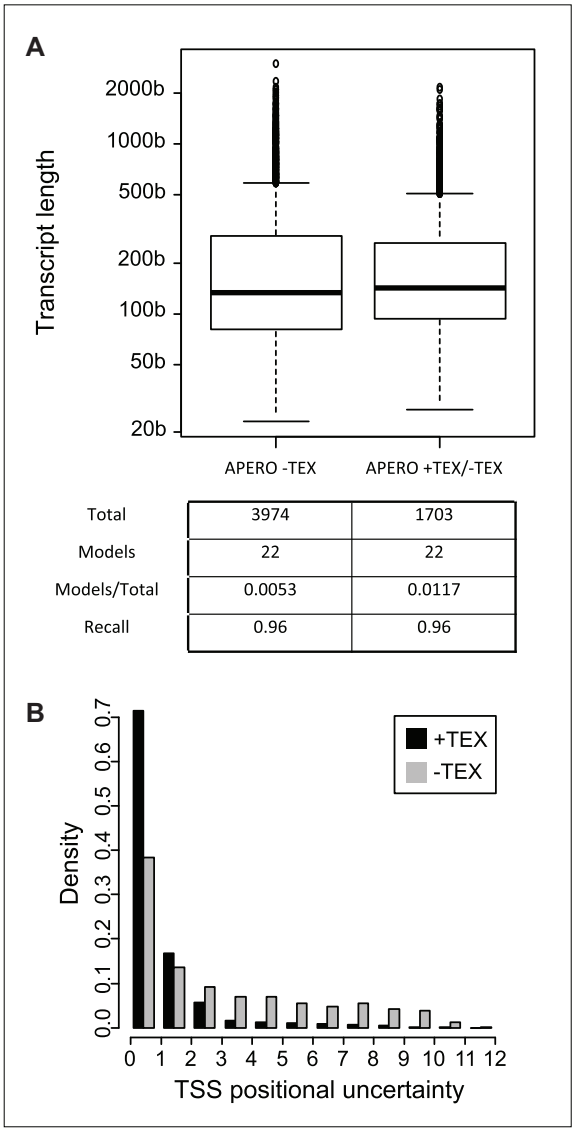
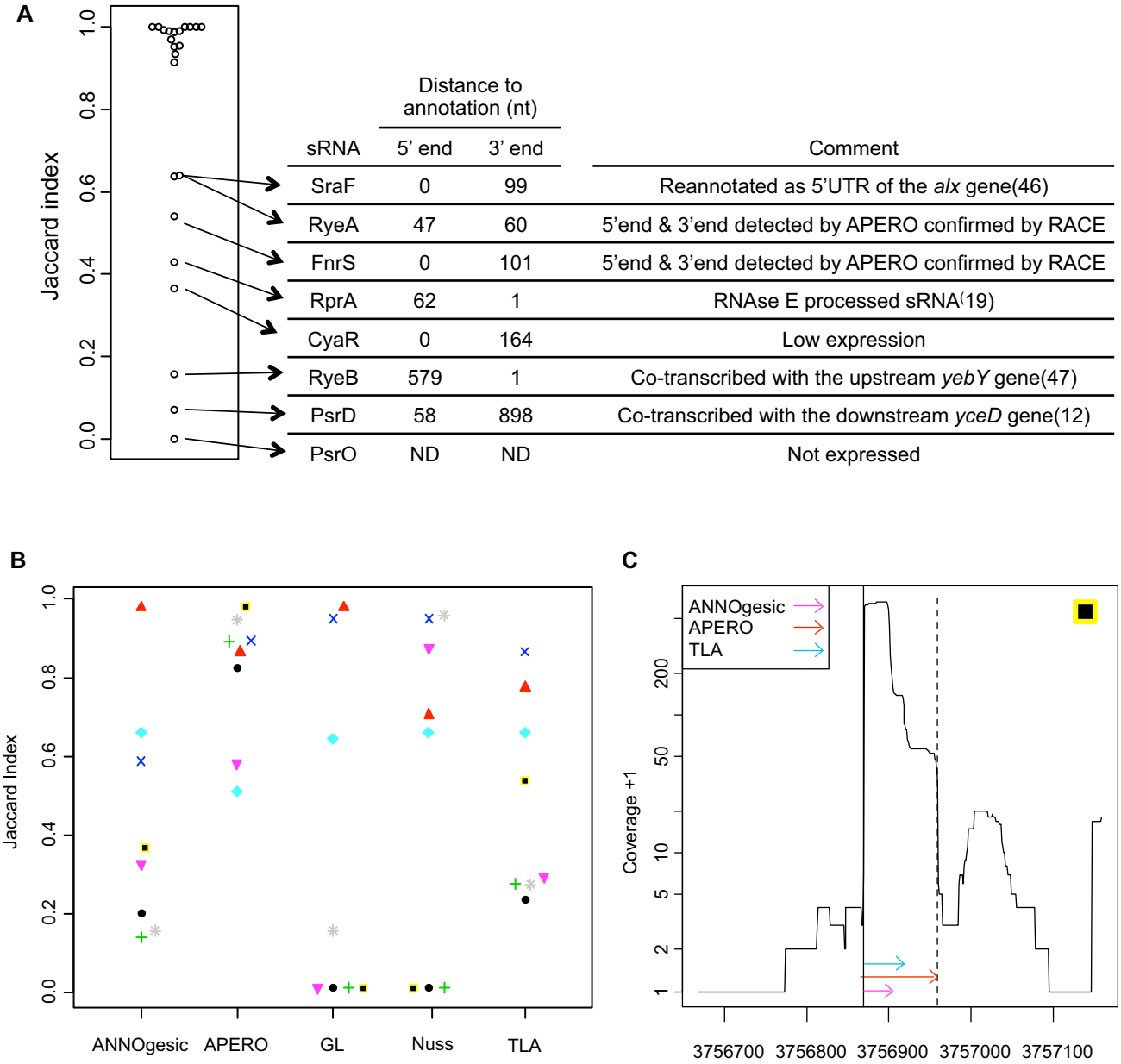


Figure 4. Identification of novel putative sRNAs in *Dickeya dadantii*. (**A**) Transcript length distributions of small transcripts identified from RNAs untreated (–TEX) or treated (+TEX) with Terminator Exonuclease. Number of small transcripts, number of detected models (out of 23 annotated sRNAs) and recall are indicated. (**B**) Evaluation of TEX treatment effect on the precision of identified TSSs.

The 1703 remaining primary small transcripts could correspond to intergenic RNAs, 5′ UTR or 3′ UTR-derived RNA products (which may or may not be functional regulatory sRNAs in addition to attenuation or processing products), and antisense RNAs. Hence, the APERO annotation module classified 526 of these 1703 as intergenic (div +P + O), 680 as antisense RNAs, 311 as 5′ UTR and 108 as 3′ UTR and 78 as both antisense and UTR (Supplementary Figure S5). We now examine if some of these can be validated using existing annotations or experiments.



sRNA	Distance to annotation (nt) 5' end	3' end	Comment
SraF	0	99	Reannotated as 5'UTR of the *alx* gene(46)
RyeA	47	60	5'end & 3'end detected by APERO confirmed by RACE
FnrS	0	101	5'end & 3'end detected by APERO confirmed by RACE
RprA	62	1	RNAse E processed sRNA(19)
CyaR	0	164	Low expression
RyeB	579	1	Co-transcribed with the upstream *yebY* gene(47)
PsrD	58	898	Co-transcribed with the downstream *yceD* gene(12)
PsrO	ND	ND	Not expressed

Figure 5. Performance in the detection of *D. dadantii* sRNAs. (**A**) Jaccard index of the 23 annotated sRNAs present in *D. dadantii*. 15 are detected with a Jaccard index >0.9. PsrO is not detected (Jaccard index = 0). For the seven others, data analysis and previous observations suggest that the differences are due to discrepancies between the annotations inferred from *E. coli* and our data, rather than a failure of the algorithm (see text and Supplementary Figure S11). ND = not detected. (**B**) Evaluation of APERO performance by RACE-PCR using 8 new predicted sRNAs. Jaccard indexes of the 8 sRNA candidates are shown. APERO clearly outperforms all other algorithms, with 6 accurate sRNAs (Jaccard $\geq$ 0.8) and no strong failure (Jaccard $\leq$ 0.5). (**C**) Example RNA-Seq profile of a sRNA candidate chosen for RACE-PCR validation. This sRNA is detected by APERO, ANNOgesic (27) and TLA from RNA-eXpress (23) only (red, pink and blue arrow, respectively). sRNA start and end positions obtained by RACE-PCR are shown as black vertical lines (solid and dashed, respectively). GL (25) and Nuss (26) methods failed to detect any sRNA in this region.

Twenty-three sRNAs have been annotated in the *Dickeya* genome based on sequence similarities to *E. coli*. Twenty-two of these are detected by APERO (Recall > 96%) except PsrO/SraG, which is expressed after heat shock or cold shock treatment in *E. coli* (12) and not expressed in our conditions. Among the 22 detected sRNAs, 15 (68%) are identified with accurate 5′ and 3′-ends as deduced from *E. coli* annotations (Jaccard index>0.9), whereas the remaining seven exhibit significant deviations. We examined each of these seven individuals, to understand if these deviations result from a failure of the algorithm or reflect a discrepancy between the annotation and our data. The summary of this analysis is given in Figure 5, and snapshots of the RNA-Seq data are shown in Supplementary Figure S11. One sRNA (CyaR) has a very low coverage, and thus presumably also a very low expression in our conditions. Another (RprA) is a RNAse E processed sRNA with a monophosphate 5′ end (19); accordingly, its 5′ end was not detected after TEX

treatment, but the shorter transcript detected by APERO is also present in the –TEX dataset and might thus correspond to an alternate form (Supplementary Figure S11). In three cases, in our data, the annotated sRNA is not transcribed alone, but co-transcribed with a neighbor gene: SraF is a pH-responsive riboregulator that resides in the 5′ UTR region of the *alx* gene (46), and should thus be re-annotated; PsrD and RyeB are co-transcribed with their downstream and upstream genes respectively (Supplementary Figure S11), as previously observed (12,47), and the boundaries given by APERO match those of the reported transcription units. For the two remaining sRNAs (RyeA and FnrS), the boundaries provided by APERO were validated by RACE-PCR. Altogether, the identified 5′ and 3′-ends were accurate for more than 90% sRNA models (with Jaccard index > 0.9). Discrepancies with the annotation deduced from *E. coli* thus indicate that the latter does not always match the transcriptional patterns present in our RNA-seq data and highlight the complexity of sRNA biogenesis in the living cell.

As a final test, we examine the validity of newly annotated small transcripts by RACE-PCR. We chose eight new candidates, whose boundaries were annotated differently depending on the algorithm used (Figure 5B and Supplementary Figure S12). Compared to RACE-PCR data, APERO annotations are accurate for six out of eight new candidates (Jaccard index > 0.8), against two to three for other methods (Figure 5B, Supplementary File S5). Also, in contrast to APERO, all other algorithms strongly missed the boundaries of at least one small transcript (Jaccard < 0.5). Distances between RACE-PCR and APERO annotations are mostly less than one nucleotide at the 5′-end and around twenty nucleotides at the 3′end (Supplementary File S5, Figure 5C).

Taken together, these validations clearly confirm the ability of APERO to accurately identify bacterial sRNAs.

DISCUSSION AND CONCLUSION

In this paper, we presented a novel tool named APERO to detect small bacterial transcripts from paired-end size-selected RNA-seq data. Rather than focusing on coverage depth as most of the available state-of-art tools, APERO takes advantage of paired-end sequencing to localize the ends of fragments derived from the analysed transcript, with high precision. Because of this improvement, the method is less impacted by the strong coverage variations appearing within a transcript, and it identifies the 5′ and especially 3′ ends with more accuracy. An additional specificity of APERO is its capacity to identify precursor and matured sRNA isoforms of a single sRNA.

Among the existing tools, Rockhopper (24) and DETR'PROK (28) were the first published methodologies. Rockhopper was not specifically designed for sRNA detection and DETR'PROK was designed for the use of short unpaired reads, and consequently exhibit the worst performances for the latter. In particular, they were designed for RNA-seq libraries without size-selection but prepared with a fragmentation step. This step is required for the sequencing of long RNAs but artificially increases the identification noise for short RNAs. The performance of the other best tool, namely sRNA-Detect (22), is also skewed because it searches for uniform coverage, resulting in a high number of incorrectly short RNAs. The performances of the three remaining outstanding methods are lowered by their sensitivity to the minimum height/coverage parameter (Supplementary Figures S9, S10). A minimum height parameter chosen too close to the background level results in a lot of false positives, and in arbitrarily extending highly expressed small transcripts, while a too high value results in missing actual sRNAs, but also incorrectly cutting them at internal breaking/cleavage sites where the coverage drops. This last issue is especially corrected by the APERO algorithm, leading to the significantly better performances exhibited in the Results section.

As an example, APERO not only identified 144 sRNAs of the 185 experimentally validated sRNAs in both *E. coli* and *S. enterica*, but in addition, for >75% of these sRNAs, the distances between annotated boundaries and APERO detected boundaries are <20 bp. The APERO algorithm was designed to be applicable on standard RNA-Seq data after selection of small transcripts (and no fragmentation step), but also preferentially after Terminator Exonuclease (TEX) for improved TSS identification (as demonstrated using the *D. dadantii* dataset). It is also less sensitive to variation in parameter values, and altogether, can thus be considered as the most robust and accurate method for the detection of bacterial small RNAs.

DATA AVAILABILITY

APERO R package: https://github.com/Simon-Leonard/APERO; https://zenodo.org/record/2536767#.XDc6xlxKhPY

APERO Galaxy module: http://bioinfo.insa-lyon.fr

S. enterica dataset: SRA accession number SRX1036363.

E. coli dataset: SRA accession number SRX4670654.

D. dadantii dataset: SRA accession number SRX4664132.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

S. Leonard received a doctoral grant from the French Ministère de l'Education nationale de l'Enseignement Supérieur et de la Recherche. The authors thank Vincent Lacroix and Hubert Charles for helpful discussion and Camille Villard for her helpful technical assistance. S. Lacour would like to thank Fasteris SA (Geneva) for its valuable technical assistance in the preparation of the sequencing library of small RNAs.

FUNDING

ANR Combicontrol grant [ANR-15-CE21-0003-01]; FR BioEnviS and using the computing facilities of the Computing Cluster of the INSA Bioinfo plateform. Funding for open access charge: CNRS grant.

Conflict of interest statement. None declared.

REFERENCES

1. Wagner,E.G.H. and Romby,P. (2015) Small RNAs in bacteria and archaea: who they are, what they do, and how they do it. *Adv. Genet.*, **90**, 133–208.
2. Patil,V.S., Zhou,R. and Rana,T.M. (2014) Gene regulation by non-coding RNAs. *Crit. Rev. Biochem. Mol. Biol.*, **49**, 16–32.
3. Shin,S.-Y. and Shin,C. (2016) Regulatory non-coding RNAs in plants: potential gene resources for the improvement of agricultural traits. *Plant Biotechnol. Rep.*, **10**, 35–47.
4. Mehta,P., Goyal,S. and Wingreen,N.S. (2008) A quantitative comparison of sRNA-based and protein-based gene regulation. *Mol. Syst. Biol.*, **4**, 221.
5. Gottesman,S., McCULLEN,C.A., Guillier,M., Vanderpool,C.K., Majdalani,N., Benhammou,J., Thompson,K.M., FitzGERALD,P.C., Sowa,N.A. and FitzGERALD,D.J. (2006) Small RNA regulators and the bacterial response to stress. *Cold Spring Harbor Symp. Quant. Biol.*, **71**, 1–11.
6. Storz,G., Vogel,J. and Wassarman,K.M. (2011) Regulation by small RNAs in bacteria: expanding frontiers. *Mol. Cell*, **43**, 880–891.
7. Desnoyers,G., Bouchard,M.-P. and Massé,E. (2013) New insights into small RNA-dependent translational regulation in prokaryotes. *Trends Genet.*, **29**, 92–98.
8. Papenfort,K. and Vanderpool,C.K. (2015) Target activation by regulatory RNAs in bacteria. *FEMS Microbiol. Rev.*, **39**, 362–378.
9. Lalaouna,D., Simoneau-Roy,M., Lafontaine,D. and Massé,E. (2013) Regulatory RNAs and target mRNA decay in prokaryotes. *Biochim. Biophys. Acta*, **1829**, 742–747.
10. Dornenburg,J.E., Devita,A.M., Palumbo,M.J. and Wade,J.T. (2010) Widespread antisense transcription in Escherichia coli. *MBio*, **1**, e00024-10.
11. Sesto,N., Wurtzel,O., Archambaud,C., Sorek,R. and Cossart,P. (2013) The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. *Nat. Rev. Microbiol.*, **11**, 75–82.
12. Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr. Biol.*, **11**, 941–950.
13. Kawano,M., Reynolds,A.A., Miranda-Rios,J. and Storz,G. (2005) Detection of 5′- and 3′-UTR-derived small RNAs and cis-encoded antisense RNAs in Escherichia coli. *Nucleic Acids Res.*, **33**, 1040–1050.
14. Ren,G.-X., Guo,X.-P. and Sun,Y.-C. (2017) Regulatory 3′ untranslated regions of bacterial mRNAs. *Front. Microbiol.*, **8**, 1276
15. Kröger,C., Dillon,S.C., Cameron,A.D.S., Papenfort,K., Sivasankaran,S.K., Hokamp,K., Chao,Y., Sittka,A., Hébrard,M., Händler,K. *et al.* (2012) The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E1277–E1286.
16. Barquist,L. and Vogel,J. (2015) Accelerating discovery and functional analysis of small RNAs with new technologies. *Annu. Rev. Genet.*, **49**, 367–394.
17. Dequivre,M., Diel,B., Villard,C., Sismeiro,O., Durot,M., Coppée,J.Y., Nesme,X., Vial,L. and Hommais,F. (2015) Small RNA Deep-Sequencing analyses reveal a new regulator of virulence in *Agrobacterium fabrum* C58. *Mol. Plant-Microbe Interactions*, **28**, 580–589.
18. Amin,S.V., Roberts,J.T., Patterson,D.G., Coley,A.B., Allred,J.A., Denner,J.M., Johnson,J.P., Mullen,G.E., O'Neal,T.K., Smith,J.T. *et al.* (2016) Novel small RNA (sRNA) landscape of the starvation-stress response transcriptome of *Salmonella enterica* serovar Typhimurium. *RNA Biol.*, **13**, 331–342.
19. Chao,Y., Li,L., Girodat,D., Förstner,K.U., Said,N., Corcoran,C., Śmiga,M., Papenfort,K., Reinhardt,R., Wieden,H.-J. *et al.* (2017) In vivo cleavage map Illuminates the central role of RNase E in coding and non-coding RNA pathways. *Mol. Cell*, **65**, 39–51.
20. Briani,F., Carzaniga,T. and Dehò,G. (2016) Regulation and functions of bacterial PNPase. *Wiley Interdiscip. Rev. RNA*, **7**, 241–258.
21. Göpel,Y., Papenfort,K., Reichenbach,B., Vogel,J. and Görke,B. (2013) Targeted decay of a regulatory small RNA by an adaptor protein for RNase E and counteraction by an anti-adaptor RNA. *Genes Dev.*, **27**, 552–564.
22. Peña-Castillo,L., Grüell,M., Mulligan,M.E. and Lang,A.S. (2016) Detection of bacterial small transcripts from RNA-seq data: a comparative assessment. *Pac. Symp. Biocomput.*, **21**, 456–467.
23. Forster,S.C., Finkel,A.M., Gould,J.A. and Hertzog,P.J. (2013) RNA-eXpress annotates novel transcript features in RNA-seq data. *Bioinformatics*, **29**, 810–812.
24. McClure,R., Balasubramanian,D., Sun,Y., Bobrovskyy,M., Sumby,P., Genco,C.A., Vanderpool,C.K. and Tjaden,B. (2013) Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res.*, **41**, e140.
25. Gómez-Lozano,M., Marvig,R.L., Molin,S. and Long,K.S. (2012) Genome-wide identification of novel small RNAs in Pseudomonas aeruginosa: Small RNAs in Pseudomonas aeruginosa. *Environ. Microbiol.*, **14**, 2006–2016.
26. Nuss,A.M., Heroven,A.K., Waldmann,B., Reinkensmeier,J., Jarek,M., Beckstette,M. and Dersch,P. (2015) Transcriptomic profiling of yersinia pseudotuberculosis reveals reprogramming of the Crp regulon by temperature and uncovers Crp as a master regulator of small RNAs. *PLOS Genet.*, **11**, e1005087.
27. 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**, doi:10.1093/gigascience/giy096.
28. Toffano-Nioche,C., Luo,Y., Kuchly,C., Wallon,C., Steinbach,D., Zytnicki,M., Jacq,A. and Gautheret,D. (2013) Detection of non-coding RNA in bacteria and archaea using the DETR'PROK Galaxy pipeline. *Methods*, **63**, 60–65.
29. Jiang,X., Sobetzko,P., Nasser,W., Reverchon,S. and Muskhelishvili,G. (2015) Chromosomal 'stress-response' domains govern the spatiotemporal expression of the bacterial virulence program. *MBio.*, **6**, e00353-15.
30. Maes,M. and Messens,E. (1992) Phenol as grinding material in RNA preparations. *Nucleic Acids Res.*, **20**, 4374.
31. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
32. Kröger,C., Colgan,A., Srikumar,S., Händler,K., Sivasankaran,S.K., Hammarlöf,D.L., Canals,R., Grissom,J.E., Conway,T., Hokamp,K. *et al.* (2013) An Infection-Relevant transcriptomic compendium for salmonella enterica serovar Typhimurium. *Cell Host Microbe.*, **14**, 683–695.
33. Schlüter,J.-P., Reinkensmeier,J., Barnett,M.J., Lang,C., Krol,E., Giegerich,R., Long,S.R. and Becker,A. (2013) Global mapping of transcription start sites and promoter motifs in the symbiotic α-proteobacterium Sinorhizobium meliloti 1021. *BMC Genomics*, **14**, 156.
34. Prados,J., Linder,P. and Redder,P. (2016) TSS-EMOTE, a refined protocol for a more complete and less biased global mapping of transcription start sites in bacterial pathogens. *BMC Genomics*, **17**, 849.
35. Thomason,M.K., Bischler,T., Eisenbart,S.K., Förstner,K.U., Zhang,A., Herbig,A., Nieselt,K., Sharma,C.M. and Storz,G. (2015) Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in Escherichia coli. *J. Bacteriol.*, **197**, 18–28.
36. Mandin,P. and Gottesman,S. (2010) Integrating anaerobic/aerobic sensing and the general stress response through the ArcZ small RNA. *EMBO J.*, **29**, 3094–3107.
37. Guo,M.S., Updegrove,T.B., Gogol,E.B., Shabalina,S.A., Gross,C.A. and Storz,G. (2014) MicL, a new σE-dependent sRNA, combats envelope stress by repressing synthesis of Lpp, the major outer membrane lipoprotein. *Genes Dev.*, **28**, 1620–1634.
38. Smirnov,A., Wang,C., Drewry,L.L. and Vogel,J. (2017) Molecular mechanism of mRNA repression in trans by a ProQ-dependent small RNA. *EMBO J.*, **36**, 1029–1045.
39. Dar,D. and Sorek,R. (2018) High-resolution RNA 3′-ends mapping of bacterial Rho-dependent transcripts. *Nucleic Acids Res.*, **46**, 6797–6805.
40. Yan,B., Boitano,M., Clark,T.A. and Ettwiller,L. (2018) SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat. Commun.*, **9**, 3676.
41. Lybecker,M., Bilusic,I. and Raghavan,R. (2014) Pervasive transcription: detecting functional RNAs in bacteria. *Transcription*, **5**, e944039.
42. Wade,J.T. and Grainger,D.C. (2014) Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat. Rev. Microbiol.*, **12**, 647–653.

43. Georg,J. and Hess,W.R. (2018) Widespread antisense transcription in prokaryotes. *Microbiol. Spectr.*, **6**, doi:10.1128/microbiolspec.RWR-0029-2018.
44. Wang,M., Fleming,J., Li,Z., Li,C., Zhang,H., Xue,Y., Chen,M., Zhang,Z., Zhang,X.-E. and Bi,L. (2016) An automated approach for global identification of sRNA-encoding regions in RNA-Seq data from Mycobacterium tuberculosis. *Acta Biochim. Biophys. Sin.*, **48**, 544–553.
45. Jahn,C.E., Charkowski,A.O. and Willis,D.K. (2008) Evaluation of isolation methods and RNA integrity for bacterial RNA quantitation. *J. Microbiol. Methods*, **75**, 318–324.
46. Nechooshtan,G., Elgrably-Weiss,M. and Altuvia,S. (2014) Changes in transcriptional pausing modify the folding dynamics of the pH-responsive RNA element. *Nucleic Acids Res.*, **42**, 622–630.
47. Vogel,J., Bartels,V., Tang,T.H., Churakov,G., Slagter-Jäger,J.G., Hüttenhofer,A. and Wagner,E.G.H. (2003) RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res.*, **31**, 6435–6443.