Comparison of Stranded and Non-stranded RNA-Seq in Predicting Small RNAs in a Non-model Bacterium

SEDLÁŘ, K.; ZIMMER, R.

Bioinformatics and Biomedical Engineering
9th International Work-Conference, IWBBIO 2022, Maspalomas, Gran Canaria, Spain, June 27–30, 2022, Proceedings, Part II

pp 45-56

eISBN: 978-3-031-07802-6

DOI: https://doi.org/10.1007/978-3-031-07802-6_4

Accepted manuscript

Comparison of Stranded and Non-Stranded RNA-Seq in Predicting Small RNAs in a Non-Model Bacterium

Karel Sedlar[1,2][0000-0002-8269-4020], Ralf Zimmer[1][0000-0003-1439-2327]

[1]Institute of Bioinformatics, Department of Informatics, Ludwig-Maximilians-Universität München, Munich, Germany
[2]Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Brno, Czechia

sedlar@bio.ifi.lmu.de

Abstract. Thanks to their diversity, non-model bacteria represent an inexhaustible resource for microbial biotechnology. Their utilization is only limited by our lack of knowledge regarding the regulation of processes they are capable to perform. The problem lies in non-coding regulators, for example small RNAs, that are not so widely studied as coding genes. One possibility to overcome this hurdle is to use standard RNA-Seq data, gathered primarily to study gene expression, for the prediction of non-coding elements. Although computational tools to perform this task already exist, they require the utilization of stranded RNA-Seq data that must not be available for non-model organisms. Here, we showed that *trans*-encoded small RNAs can be predicted from non-stranded data with comparable sensitivity to stranded data. We used two RNA-Seq datasets of non-type strain *Clostridium beijerinckii* NRRL B-598, which is a promising hydrogen and butanol producer, and obtained comparable results for stranded and non-stranded datasets. Nevertheless, the non-stranded approach suffered from lower precision. Thus, the results must be interpreted with caution. In general, more benchmarking for tools performing direct prediction of small RNAs from standard RNA-Seq data is needed so these techniques could be adopted for automatic detection.

Keywords: small non-coding RNA; *Clostridium beijerinckii* NRRL B-598; RNA-Seq; genome annotation

1 Introduction

It has been almost half a century since small non-coding RNAs (sRNAs) were discovered in bacteria [1]. During years, sRNAs were shown to play important regulatory roles in diverse cellular processes by participating in post-transcriptional regulation of gene expression [2]. This is the reason why sRNAs are drawing more attention than ever before. While the first experiments were done with a model bacterium, *Escherichia coli*, primarily its non-pathogenic strain K-12, later studies showed the role of sRNA in the virulence of pathogenic bacteria [3–5]. Besides their role in medicine, sRNAs can

The final authenticated version is available online at https://doi.org/10.1007/978-3-031-07802-6_4.

be used in general biotechnologies for their involvement in other processes, for example, degradation of toxic compounds [6]. Finally, the latest research shows that the engineering of a novel sRNA can improve bacterial phenotype, for example, tolerance to acids [7], which could be utilized in various fermentation processes for the production of bio-based chemicals.

As the former widely used title small non-coding RNA suggests, it is a small molecule that is not translated into a protein. Although this is true in a majority of cases, it has been proved that some sRNAs can encode small proteins [8]. Therefore, it is common that these short regulatory RNAs are simply referred to as small RNA. Its length can vary but it typically spans within the interval 40–500 nucleotides [9]. Most commonly, sRNAs can be divided according to the locations of sRNA genes and their targets into two groups, *cis*-encoded and *trans*-encoded sRNAs [8]. A *cis*-encoded sRNA overlaps with a regulated gene but is coded by the antisense strand and during its regulation binds to the target mRNA by perfect base pairing. Binding can occur at any location depending on the location of sRNA expression [10]. There are three mechanisms that *cis*-encoded sRNAs use for regulation. They can act as transcription terminators, potential inhibitors of translation initiation, or modulators of mRNA degradation. A *trans*-encoded sRNA interacts with its target mRNA by imperfect base pairing because such sRNA is coded by an intergenic region (ITR) and its coding sequence does not overlap with a sequence of the target gene [11]. This also means that *trans*-encoded sRNAs can be coded by the same strand as target genes and they have a wider range of regulatory mechanisms. They can act as repressors of expression but also as activators. They can increase as well as block mRNA degradation.

Some of the early experiments showed that sRNA genes identified in *E. coli* were found in *Salmonella enterica* and vice versa [2]. This suggested their conservation across the bacterial domain and made them ideal targets for computational prediction. A wide range of tools has been proposed. In general, they can be divided into two groups: comparative genomics-based and machine learning-based techniques [8]. While the former techniques rely on sequence alignment and cluster analysis with phylogenetic profiling, the latter are taking advantage of widely used machine learning methods such as neural networks, support vector machines, and genetic algorithms. Nevertheless, these techniques can only predict a location of sRNA but cannot predict its target site, which can be cumbersome, primarily for *trans*-encoded sRNAs that pair imperfectly to target sites. Besides computational solution lying in *in silico* prediction of sRNA-target mRNA interaction, e.g., sRNATarget [12], IntaRNA [13], or RNApredator [14], there is a plethora of techniques based on RNA-Seq to reveal these interactions experimentally [15]. The main disadvantage of these specialized techniques such as GRIL-Seq [16], RIP-Seq [17], RIL-Seq [18], and many others, is their difficult implementation in non-model bacteria that limits their utilization to model organisms, mainly *E. coli* [15]. On the other hand, even standard RNA-Seq that became a commonly used technique in bacterial research, can be used to discover sRNA genes.

Despite existing algorithms as well as experimental techniques, identification of sRNA genes is still not a common procedure during annotation of non-model bacterial genomes. While there is currently more than a million bacterial genome assemblies in

the GenBank database (27th January 2022), the number of annotated sRNAs for particular genomes is very limited, usually in units of genes. The most commonly used tool for genome annotation, the PGAP pipeline [19], uses homology-based annotation by scanning the Rfam database [20] with infernal's cmsearch [21]. This suggests that computational prediction of sRNAs in non-model bacteria might be limited by low sequence similarity to model organisms whose sRNAs were discovered experimentally. This opens a door to the utilization of standard RNA-Seq data which is available for many non-model bacteria. Nevertheless, a systematic pipeline for such predictions is missing and various authors use different techniques. Zhu et al. [22] predicted approximately ten sRNAs in *Bifidobacterium animalis* by combining prediction using TargetRNA2 [23] with RNA-Seq data used to calculate RPKM (Reads per kilobase per million) values summarizing expression of identified sRNAs. Liu et al. [24] found 263 sRNAs candidates in *Mycobacterium neoaurum* by combining RPKM and IntaRNA predictions. On the contrary, Wang et al. [25] used RNA-Seq data itself for searching sRNAs in *Mycobacterium tuberculosis* by examining coverage of unannotated regions. Thanks to the utilization of strand-specific RNA-Seq, 192 sRNAs candidates were found in intergenic regions and additional 664 candidates coded by antisense strand in regions overlapping to target genes. Although their study is presented as an automated approach, it brings no computational tool that could be used for another organism.

It is the unavailability of computational tools that prevents the wider utilization of RNA-Seq data in the prediction of sRNA genes in non-model bacteria. There are only a few tools that suffer from various drawbacks. For example, APERO [26] needs paired-end reads which are usually not available for bacterial RNA-Seq data, Rockhopper [27] is very hard to be implemented to other pipelines due to its graphical user interface nature and utilization of obsolete formats such as protein table for genome annotation, and baerhunter [28] is no longer working with the current version of R/Bioconductor. Moreover, benchmarking for different tools is missing and a comparison of prediction possibilities regarding input data was never performed before. In this paper, we got inspired by current tools and performed sRNAs prediction in the non-model bacterium *Clostridium beijerinckii* NRRL B-598 [29] using two different RNA-Seq datasets taken under the same conditions. We showed that the current approach in sRNAs prediction can be, with some limitations, applied to both, stranded as well as non-stranded RNA-Seq data and that more benchmarking is needed to establish functional pipelines for sRNAs prediction using standard RNA-Seq data.

2 Materials and methods

2.1 Genome and annotation

To examine sRNAs prediction in a non-model bacterium, we selected *C. beijerinckii* NRRL B-598, a non-type strain, which is a promising butanol and hydrogen producer. Most importantly, it is a non-type strain with the highest number of RNA-Seq-based transcriptomic studies among solventogenic clostridia [30]. In this study, we used its third complete genome assembly, available at DDBJ/EMBL/GenBank under accession No. CP011966.3, which was constructed using a combination of Roche 454 GS Junior,

PacBio RSII, and Illumina NextSeq500 reads [31]. The genome annotation was performed with PGAP v4.6 [19] and genome features are summarized in **Table 1**.

Table 1. Genome features of *Clostridium beijerinckii* NRRL B-598.

Feature	Chromosome
Length (bp)	6,186,993
GC content (%)	29.8
Protein coding genes	5,128
Pseudogenes	166
rRNAs (5S,16S, 23S)	17, 16, 16
tRNAs	94
Non-coding RNAs	5
Riboswitches	31

2.2 Transcriptomic data

RNA-Seq data used in this study comes from a publicly available study performing transcriptional profiling of the butanol fermentation using glucose as a substrate [32]. Two particular samples, A and B, from the exponential growth phase, after 3.5h from the start of fermentation, were selected. These samples are available from the NCBI Sequence Read Archive (SRA) under the project accession number PRJNA229510. Cell samples for isolation of total RNA were collected from 3 ml of culture broth (OD600 0.9–1.0) by centrifugation at 10000 rpm for two minutes, washed with RNase free water and cell pellets were immediately stored at − 70 °C. RNA from the cell pellet was isolated using High Pure RNA Isolation Kit (Roche). Isolated total RNA was stored frozen at − 70 °C. The total RNA concentration was determined on DS-11 FX+ Spectrophotometer (DeNovix). Quality and integrity of the samples were assessed using the Agilent RNA 6000 Nano Kit (Agilent) with the Agilent 2100 Bioanalyzer (Agilent). RNA integrity number was measured using 2100 Bioanalyzer Expert software. Frozen total RNA samples were thawed on ice and an aliquot of each sample containing 10 µg of RNA was taken for 16S and 23S ribosomal RNAs removal using The MICROBExpress™ Bacterial mRNA Enrichment Kit (Ambion). Efficiency of ribosomal RNA depletion and concentration of RNA samples were checked on the Agilent 2100 Bioanalyzer (Agilent) with the Agilent RNA 6000 Nano Kit (Agilent).

For sample A, library construction and sequencing was performed by BGI Europe A/S (Copenhagen, Denmark). During the library preparation, cDNA was synthesized by using a random hexamer-primer and the sample was sequenced on Illumina HiSeq 4000, single-end, 50 bp. This means that resulting reads are non-stranded, i.e., it is not possible to determine a strand of DNA that codes genes producing sequence transcript as reads mapping to analyzed loci have both orientations.

For sample B, library construction and sequencing was performed by CEITEC Genomics core facility (Brno, Czechia). NEBNext Ultra II stranded kit was used for library preparation and the sample was sequenced on Illumina NextSeq500, single-end,

75 bp. This resulted in reads that are reversely stranded, i.e., the reads have the opposite orientation to the locus producing sequenced transcripts.

2.3 Data preprocessing

Adapter and quality trimming was performed using Trimmomatic v0.36 [33]. Two different settings were used for comparison. In the first settings, parameters LEADING and TRAILING specifying minimum qualities (PHRED score) to keep a base, were both set to three. The length of the SLIDINGWINDOW parameter was set to four and required average quality of 15. Finally, only reads reaching the length of 36 bases were kept by setting up a parameter MINLEN. In the second settings, the parameters were stricter. Minimum qualities were both set to 10 and a sliding window of the length four required at least a quality of 25. On the other hand, reads of length 20 nucleotides and more were preserved.

Although laboratory ribodepletion was performed prior to sequencing, the step of computational rRNA filtering was done for comparison. This step was done with SortMeRNA v2.1 [34] using the SILVA database [35] of known bacterial 16S and 23S rRNA genes. Finally, the mapping to the reference genome was performed with STAR v2.5.4b [36]. Reads mapping to more than three loci were filtered out by setting up a parameter outFilterMultimapNmax.

Quality assessment after particular steps was performed using FastQC in combination with MultiQC [37] to summarize the reports. The resulting SAM (Sequence Read Alignment/Map) files were indexed and transformed into more compact BAM (Binary Read Alignment/Map) format using SAMtools v1.7 [38].

2.4 sRNAs prediction

The prediction of sRNA loci was performed in R v4.1.2 and Bioconductor v3.14. The whole pipeline was inspired by baerhunter [28] that uses thresholding of coverage. Baerhunter itself cannot be used due to erroneous functions for counting sRNAs and untranslated regions (UTRs). Nevertheless, the pipeline was reproduced by rewriting these functions to be compatible with the current Biocondutor. Although baerhunter requires stranded RNA-Seq data, the whole pipeline can be reproduced by similar custom-made code that works also with non-stranded data. The main steps of the pipeline are summarized in **Fig. 1**.

The main idea of thresholding coverage requires coverage to be counted across the whole reference sequence in the first step. This can be achieved using samtools depth or by loading BAM files into R/Bioconductor and calculating coverage with suitable functions, for example “coverage” from the GenomicAlignments [39] package. In the case of non-stranded data, the coverage is calculated for the chromosome at once. However, for stranded data, coverages of particular strands of DNA have to be calculated separately. Before thresholding is performed, only ITRs are selected. This again requires selecting these regions separately for particular strands in the case of stranded RNA-Seq.

Selecting putative sRNAs inspired by baerhunter requires three input parameters. The first parameter “low coverage cutoff” is used to select potential sRNA loci. Once the coverage exceeds the threshold, the start of a potential sRNA is marked. The region is being continually expanded until coverage falls under the threshold again. Other parameters are used for additional filtering. The parameter “high coverage cutoff” sets another threshold for coverage. Only previously selected regions in which at least one base is covered by more reads than the thresholds are preserved. The last parameter “min sRNA length” simply filters out regions that are shorter than the selected length.

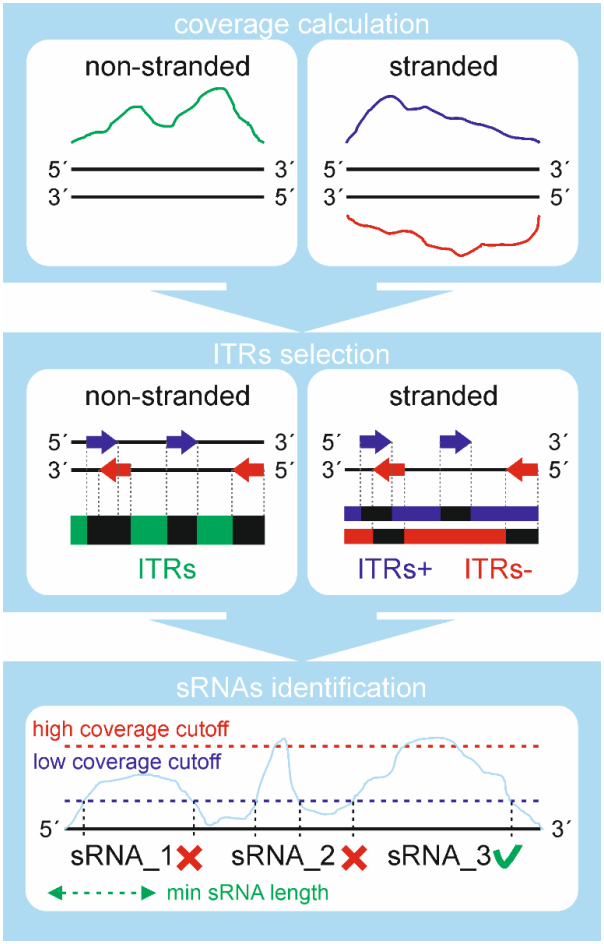

Fig. 1. A schema of coverage-based identification of sRNAs. Coverage of ITRs in examined. Here, only a sRNA_3 candidate is returned as a putative sRNA as it meets high coverage and min sRNA length cutoff value criteria.

The thresholds used in this study were, 10 for the low coverage cutoff, 50 for the high coverage cutoff, and 40 for the min sRNA length. The values were set empirically based on benchmarking study of baerhunter [28].

3 Results and discussion

3.1 Data preprocessing

Sample A contained 21 million and sample B had 53 million raw sequences. The initial quality assessment showed high GC content suggesting remaining rRNA contamination. The resulting numbers of reads after filtering and mapping steps are summarized in **Table 2**. Particular parameters settings for quality trimming can be found in materials and methods.

Table 2. Results of data preprocessing

Sample	Trimming settings	rRNA removal	No. of reads in a sample (million)	No. of mapped reads (million)
A1	1	No	21.0	11.9
A2	2	No	20.6	11.7
A1r	1	Yes	12.3	11.8
A2r	2	Yes	12.2	11.6
B1	1	No	52.5	15.3
B2	2	No	48.9	14.3
B1r	1	Yes	15.2	14.6
B2r	2	Yes	15.7	13.7

The results showed very high, up to 73%, contamination by rRNA. Although rRNA is filtered during mapping as multi-mapped reads, numbers of mapped reads for samples with and without computational ribodepletion are different, therefore, this step may affect the final identification of sRNA genes.

3.2 sRNAs prediction in stranded data

Before comparison of stranded and non-stranded data, we performed prediction of sRNAs by the same procedure that is used in baerhunter to identify putative sRNA genes as they have never been reported in *C. beijerinckii* NRRL B-598 genome before. The sensitivity of baerhunter was tested against more complex tools, particularly Rockhopper, APERO, and ANNOgesic, using simulated as well as real datasets [28]. Thus, we used its predictions, summarized in **Table 3**, to estimate sRNAs counts.

Table 3. Numbers of sRNAs predicted by baerhunter

Sample	No. of sRNA genes		
	trans-encoded	*cis*-encoded	total number
B1	121	115	236
B2	115	99	214
B1r	121	101	222
B2r	115	87	202

Although baerhunter was benchmarked in comparison to other tools, our result showed that its prediction is influenced by data preprocessing as the total number of predicted sRNAs ranged from 202 to 236. While the detection of *cis*-encoded sRNAs was influenced by quality trimming and rRNA removal, only quality trimming affected the identification of *trans*-encoded elements. The predicted *trans*-encoded sRNAs for B1 and B1r and for B2 and B2r were the same. More benchmarking would be needed to reveal the origin of these differences. Nevertheless, it is evident that direct prediction of sRNAs from RNA-Seq data is affected by computational data preprocessing and should be investigated in detail to ensure reliable prediction of non-coding genomic elements in bacteria.

3.3 Comparison of stranded and non-stranded data

Because non-stranded RNA-Seq does not preserve information about the orientation of genomic elements producing sequenced transcripts, it cannot be used for the identification of elements that overlap. Thus, only *trans*-encoded sRNAs can be predicted using non-stranded data. Since the pipeline for non-stranded data is a little bit different (see **Fig. 1**), we recalculated the results for sample B using the pipeline for non-stranded data. The results are summarized in **Table 4**.

Table 4. Numbers of sRNAs predicted by approach for non-stranded RNA-Seq

Sample	A	B	A∩B
X1	76	109	32
X2	75	108	30
X1r	76	109	32
X2r	75	108	30

Computational ribodepletion again did not affect the results. The sensitivity of detection by non-stranded approach was a little bit lower as the numbers of predicted sRNAs in B samples was slightly lower. The detection was not completely the same but very similar when only three sRNAs identified in the non-stranded approach were different from those detected by the stranded approach in samples B1/B1r and six in samples B2/B2r. If baerhunter predictions of *trans*-encoded sRNA were considered as a reference, the sensitivity (or recall) and precision of the non-stranded approach could have been calculated, see **Table 5**.

Table 5. Precision and recall of approach for non-stranded RNA-Seq

Sample	A		B	
	Precision	Recall	Precision	Recall
X1/X1r	44.7%	28.1%	97.2%	87.6%
X2/X2r	42.7%	27.8%	94.4%	88.7%

Unfortunately, the prediction using non-stranded data from sample A was considerably worse. Not only was the total number of detected sRNAs lower, more than half of predicted loci did not match those predicted using data from sample B. Such a difference between both samples is surprising. Direct detection of sRNAs from RNA-Seq data can only capture those loci that are currently being transcribed [28]. Nevertheless, both samples, A and B, come from the biological replicates taken under the same conditions, and the data were preprocessed in the same manner. Thus, the prediction should be very similar. On the other hand, there is plenty of other parameters that could be responsible for the difference: sequencing depth, preparation of library, or platform used for sequencing, etc.

The only parameter whose influence can be examined computationally is the sequencing depth. Considering the number of mapped reads and their length, sample A contains only half of the sequenced bases in comparison to B. Therefore, we set the high coverage cutoff parameter to 25 for the following detection. This resulted in 180 identified sRNAs for both quality trimming settings. The number of sRNAs that were previously detected by baerhunter was considerably higher, 113 for A1/A1r and 114 for A2/A2r. This means that the resulting recall, 93.4% for A1/A1r and 99.1% for A2/A2r, was even higher than recall for B samples processed by the non-stranded approach. The improvement of precision was lower, resulting in 62.8% for A1/A1r and 63.3% for A2/A2r.

The results showed that non-stranded RNA-Seq can be used for the prediction of *trans*-encoded sRNAs with very high sensitivity, however, the results must be interpreted carefully due to lower precision. Detection by direct processing of RNA-Seq is also heavily influenced by the sequencing depth and detection thresholds must be adjusted according to it. Moreover, the results suggested that thresholds for achieving the same sensitivity in stranded and non-stranded data might be different even if the sequencing depth correction is performed.

4 Conclusions

Prediction of small RNAs in bacterial genomes can be performed by several computational as well as laboratory techniques. Direct prediction from standard RNA-Seq data seems to be advantageous. Unlike fully computational approaches, it brings experimental evidence while recalculating data that are easily obtainable even for non-model bacterial genomes for the simplicity of technique that is widely used to measure expression on a genome-wide scale. Unfortunately, computational tools to perform such predictions are not widely adopted. Although current tools require the utilization of stranded RNA-Seq, we demonstrated that sRNAs can also be identified using non-stranded RNA-Seq with comparable sensitivity to the stranded approach. Nevertheless, only *trans*-encoded sRNAs can be identified. Moreover, we demonstrated that the prediction from non-stranded as well as stranded RNA-Seq is highly influenced by sequencing depth. Since the results depend on a threshold that has to be set up manually in current tools, more benchmarking is needed to ensure reliable and fully automatic prediction of small RNAs in bacterial genomes.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 101023766.

References

1. Ikemura, T., Dahlberg, J.E.: Small ribonucleic acids of Escherichia coli. I. Characterization by polyacrylamide gel electrophoresis and fingerprint analysis. J. Biol. Chem. 248, 5024–5032 (1973). https://doi.org/10.1016/S0021-9258(19)43666-1.
2. Hör, J., Matera, G., Vogel, J., Gottesman, S., Storz, G.: Trans-Acting Small RNAs and Their Effects on Gene Expression in Escherichia coli and Salmonella enterica. EcoSal Plus. 9, (2020). https://doi.org/10.1128/ecosalplus.esp-0030-2019.
3. Bhatt, S., Egan, M., Jenkins, V., Muche, S., El-Fenej, J.: The tip of the iceberg: On the roles of regulatory small RNAs in the virulence of enterohemorrhagic and enteropathogenic Escherichia coli. Front. Cell. Infect. Microbiol. 6, (2016). https://doi.org/10.3389/fcimb.2016.00105.
4. Koeppen, K., Hampton, T.H., Jarek, M., Scharfe, M., Gerber, S.A., Mielcarz, D.W., Demers, E.G., Dolben, E.L., Hammond, J.H., Hogan, D.A., Stanton, B.A.: A Novel Mechanism of Host-Pathogen Interaction through sRNA in Bacterial Outer Membrane Vesicles. PLoS Pathog. 12, e1005672 (2016). https://doi.org/10.1371/journal.ppat.1005672.
5. Padalon-Brauch, G., Hershberg, R., Elgrably-Weiss, M., Baruch, K., Rosenshine, I., Margalit, H., Altuvia, S.: Small RNAs encoded within genetic islands of Salmonella typhimurium show host-induced expression and role in virulence. Nucleic Acids Res. 36, 1913–1927 (2008). https://doi.org/10.1093/nar/gkn050.
6. Peng, T., Kan, J., Hu, J., Hu, Z.: Genes and novel sRNAs involved in PAHs degradation in marine bacteria Rhodococcus sp. P14 revealed by the genome and transcriptome analysis. 3 Biotech. 10, 1–10 (2020). https://doi.org/10.1007/s13205-020-2133-6.
7. Lin, Z., Li, J., Yan, X., Yang, J., Li, X., Chen, P., Yang, X.: Engineering of the Small Noncoding RNA (sRNA) DsrA Together with the sRNA Chaperone Hfq Enhances the Acid Tolerance of Escherichia coli. Appl. Environ. Microbiol. 87, 1–15 (2021). https://doi.org/10.1128/AEM.02923-20.
8. Li, W., Ying, X., Lu, Q., Chen, L.: Predicting sRNAs and Their Targets in Bacteria. Genomics, Proteomics Bioinforma. 10, 276–284 (2012). https://doi.org/10.1016/j.gpb.2012.09.004.
9. Huang, H.Y., Chang, H.Y., Chou, C.H., Tseng, C.P., Ho, S.Y., Yang, C.D., Ju, Y.W., Huang, H. Da: sRNAMap: Genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. Nucleic Acids Res. 37, (2009). https://doi.org/10.1093/nar/gkn852.
10. Cho, K.H., Kim, J.H.: Cis-encoded non-coding antisense RNAs in streptococci and other low GC Gram (+) bacterial pathogens. Front. Genet. 6, 110 (2015). https://doi.org/10.3389/fgene.2015.00110.
11. Rath, E.C., Pitman, S., Cho, K.H., Bai, Y.: Identification of streptococcal small RNAs that are putative targets of RNase III through bioinformatics analysis of RNA sequencing data. BMC Bioinformatics. 18, 111–120 (2017). https://doi.org/10.1186/s12859-017-1897-0.

12. Cao, Y., Zhao, Y., Cha, L., Ying, X., Wang, L., Shao, N., Li, W.: sRNATarget: a web server for prediction of bacterial sRNA targets. Bioinformation. 3, 364–366 (2009). https://doi.org/10.6026/97320630003364.
13. Busch, A., Richter, A.S., Backofen, R.: IntaRNA: Efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. Bioinformatics. 24, 2849–2856 (2008). https://doi.org/10.1093/bioinformatics/btn544.
14. Eggenhofer, F., Tafer, H., Stadler, P.F., Hofacker, I.L.: RNApredator: Fast accessibility-based prediction of sRNA targets. Nucleic Acids Res. 39, (2011). https://doi.org/10.1093/nar/gkr467.
15. Saliba, A.E., C Santos, S., Vogel, J.: New RNA-seq approaches for the study of bacterial pathogens. Curr. Opin. Microbiol. 35, 78–87 (2017). https://doi.org/10.1016/j.mib.2017.01.001.
16. Han, K., Tjaden, B., Lory, S.: GRIL-seq provides a method for identifying direct targets of bacterial small regulatory RNA by in vivo proximity ligation. Nat. Microbiol. 2, 1–10 (2016). https://doi.org/10.1038/nmicrobiol.2016.239.
17. Chao, Y., Papenfort, K., Reinhardt, R., Sharma, C.M., Vogel, J.: An atlas of Hfq-bound transcripts reveals 3′ UTRs as a genomic reservoir of regulatory small RNAs. EMBO J. 31, 4005–4019 (2012). https://doi.org/10.1038/emboj.2012.229.
18. Melamed, S., Peer, A., Faigenbaum-Romm, R., Gatt, Y.E., Reiss, N., Bar, A., Altuvia, Y., Argaman, L., Margalit, H.: Global Mapping of Small RNA-Target Interactions in Bacteria. Mol. Cell. 63, 884–897 (2016). https://doi.org/10.1016/j.molcel.2016.07.026.
19. Li, W., O'Neill, K.R., Haft, D.H., Dicuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M.K., Durkin, A.S., Gonzales, N.R., Gwadz, M., Lanczycki, C.J., Song, J.S., Thanki, N., Wang, J., Yamashita, R.A., Yang, M., Zheng, C., Marchler-Bauer, A., Thibaud-Nissen, F.: RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res. 49, D1020–D1028 (2021). https://doi.org/10.1093/nar/gkaa1105.
20. Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J., Finn, R.D.: Rfam 12.0: Updates to the RNA families database. Nucleic Acids Res. 43, D130–D137 (2015). https://doi.org/10.1093/nar/gku1063.
21. Nawrocki, E.P., Eddy, S.R.: Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 29, 2933–2935 (2013). https://doi.org/10.1093/bioinformatics/btt509.
22. Zhu, D.Q., Liu, F., Sun, Y., Yang, L.M., Xin, L., Meng, X.C.: Genome-wide identification of small RNAs in Bifidobacterium animalis subsp. lactis KLDS 2.0603 and their regulation role in the adaption to gastrointestinal environment. PLoS One. 10, e0117373 (2015). https://doi.org/10.1371/journal.pone.0117373.
23. Kery, M.B., Feldman, M., Livny, J., Tjaden, B.: TargetRNA2: Identifying targets of small regulatory RNAs in bacteria. Nucleic Acids Res. 42, W124–W129 (2014). https://doi.org/10.1093/nar/gku317.
24. Liu, M., Zhu, Z.T., Tao, X.Y., Wang, F.Q., Wei, D.Z.: RNA-Seq analysis uncovers non-coding small RNA system of Mycobacterium neoaurum in the metabolism of sterols to accumulate steroid intermediates. Microb. Cell Fact. 15, 1–17 (2016). https://doi.org/10.1186/s12934-016-0462-2.
25. Wang, M., Fleming, J., Li, Z., Li, C., Zhang, H., Xue, Y., Chen, M., Zhang, Z., Zhang, X.E., Bi, L.: An automated approach for global identification of sRNA-encoding regions in RNA-Seq data from Mycobacterium tuberculosis. Acta Biochim. Biophys. Sin. (Shanghai). 48, 544–553 (2016). https://doi.org/10.1093/abbs/gmw037.

26. Leonard, S., Meyer, S., Lacour, S., Nasser, W., Hommais, F., Reverchon, S.: APERO: A genome-wide approach for identifying bacterial small RNAs from RNA-Seq data. Nucleic Acids Res. 47, e88–e88 (2019). https://doi.org/10.1093/nar/gkz485.
27. Tjaden, B.: A computational system for identifying operons based on RNA-seq data. Methods. 176, 62–70 (2020). https://doi.org/10.1016/j.ymeth.2019.03.026.
28. Ozuna, A., Liberto, D., Joyce, R.M., Arnvig, K.B., Nobeli, I.: Baerhunter: An R package for the discovery and analysis of expressed non-coding regions in bacterial RNA-seq data. Bioinformatics. 36, 966–969 (2020). https://doi.org/10.1093/bioinformatics/btz643.
29. Sedlar, K., Kolek, J., Skutkova, H., Branska, B., Provaznik, I., Patakova, P.: Complete genome sequence of Clostridium pasteurianum NRRL B-598, a non-type strain producing butanol. J. Biotechnol. 214, 113–114 (2015). https://doi.org/10.1016/j.jbiotec.2015.09.022.
30. Patakova, P., Branska, B., Vasylkivska, M., Jureckova, K., Musilova, J., Provaznik, I., Sedlar, K.: Transcriptomic studies of solventogenic clostridia, Clostridium acetobutylicum and Clostridium beijerinckii. Biotechnol. Adv. 107889 (2021). https://doi.org/10.1016/j.biotechadv.2021.107889.
31. Sedlar, K., Kolek, J., Gruber, M., Jureckova, K., Branska, B., Csaba, G., Vasylkivska, M., Zimmer, R., Patakova, P., Provaznik, I.: A transcriptional response of Clostridium beijerinckii NRRL B-598 to a butanol shock. Biotechnol. Biofuels. 12, 1–16 (2019). https://doi.org/10.1186/s13068-019-1584-7.
32. Sedlar, K., Koscova, P., Vasylkivska, M., Branska, B., Kolek, J., Kupkova, K., Patakova, P., Provaznik, I.: Transcription profiling of butanol producer Clostridium beijerinckii NRRL B-598 using RNA-Seq. BMC Genomics. 19, 1–13 (2018). https://doi.org/10.1186/S12864-018-4805-8/TABLES/4.
33. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 30, 2114–2120 (2014). https://doi.org/10.1093/bioinformatics/btu170.
34. Kopylova, E., Noé, L., Touzet, H.: SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 28, 3211–3217 (2012). https://doi.org/10.1093/bioinformatics/bts611.
35. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O.: The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596 (2013). https://doi.org/10.1093/nar/gks1219.
36. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29, 15–21 (2013). https://doi.org/10.1093/BIOINFORMATICS/BTS635.
37. Ewels, P., Magnusson, M., Lundin, S., Käller, M.: MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 32, 3047–3048 (2016). https://doi.org/10.1093/BIOINFORMATICS/BTW354.
38. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.: The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25, 2078–2079 (2009). https://doi.org/10.1093/bioinformatics/btp352.
39. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J.: Software for Computing and Annotating Genomic Ranges. PLoS Comput. Biol. 9, e1003118 (2013). https://doi.org/10.1371/journal.pcbi.1003118.