ANNOgesic: A genome annotation pipeline for bacterial RNA-Seq data

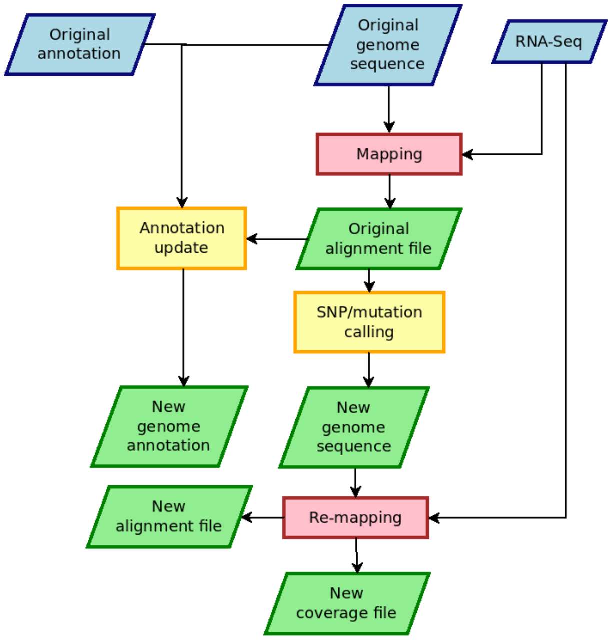
Sung-Huan Yu [1], Jörg Vogel [1], and Konrad U. Förstner [1,2*]

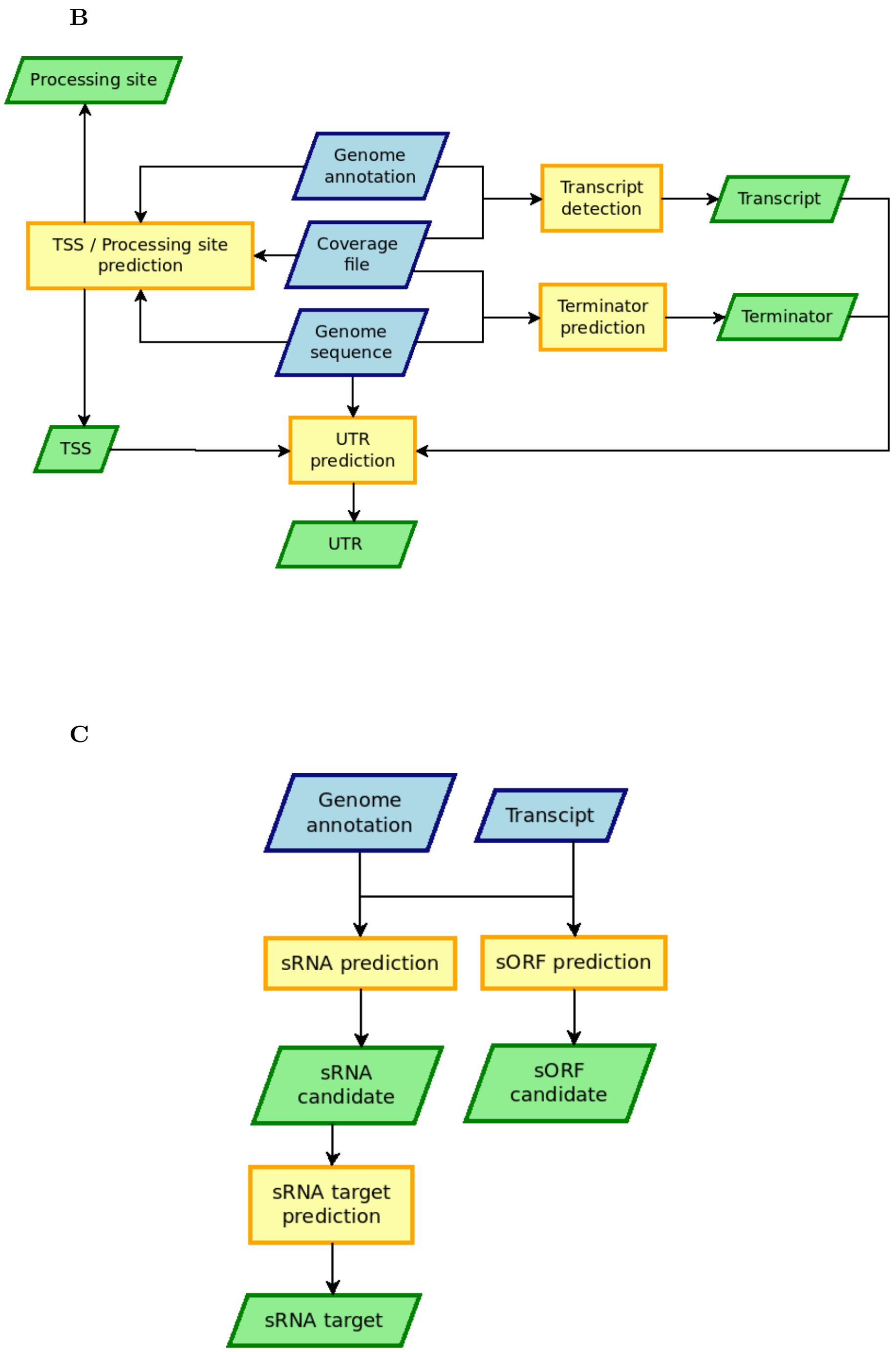
[1]Institute of Molecular Infection Biology (IMIB), University of Würzburg, 97080 Würzburg, Germany
[2]Core Unit Systems Medicine, University of Würzburg, 97070 Würzburg, Germany

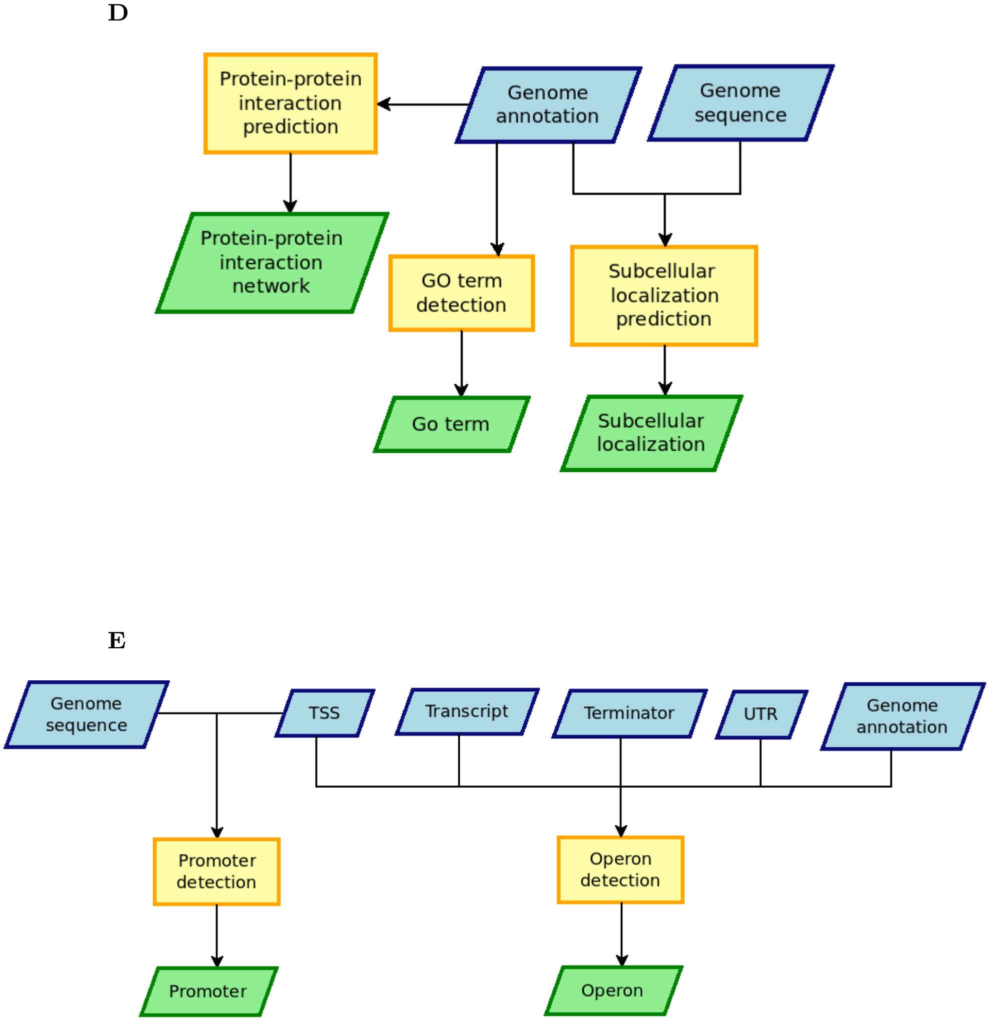
To whom correspondence should be addressed. Tel: +49-931/31-84279 ; Email: konrad.foerstner@uni-wuerzburg.de

Supplementary Figures

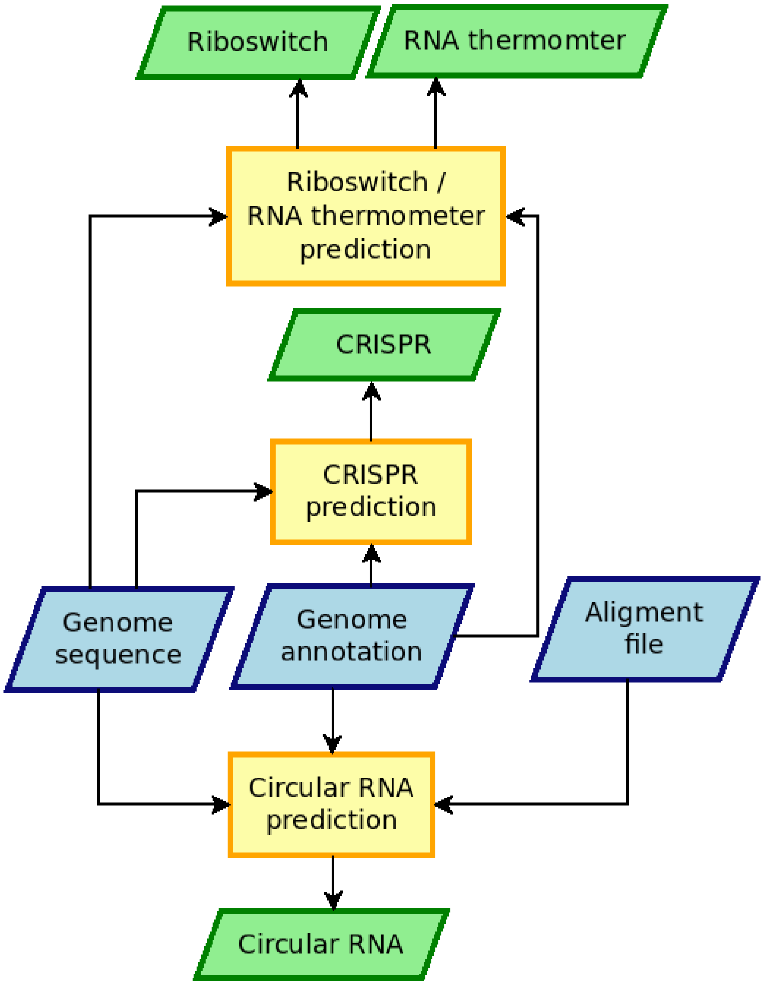
A



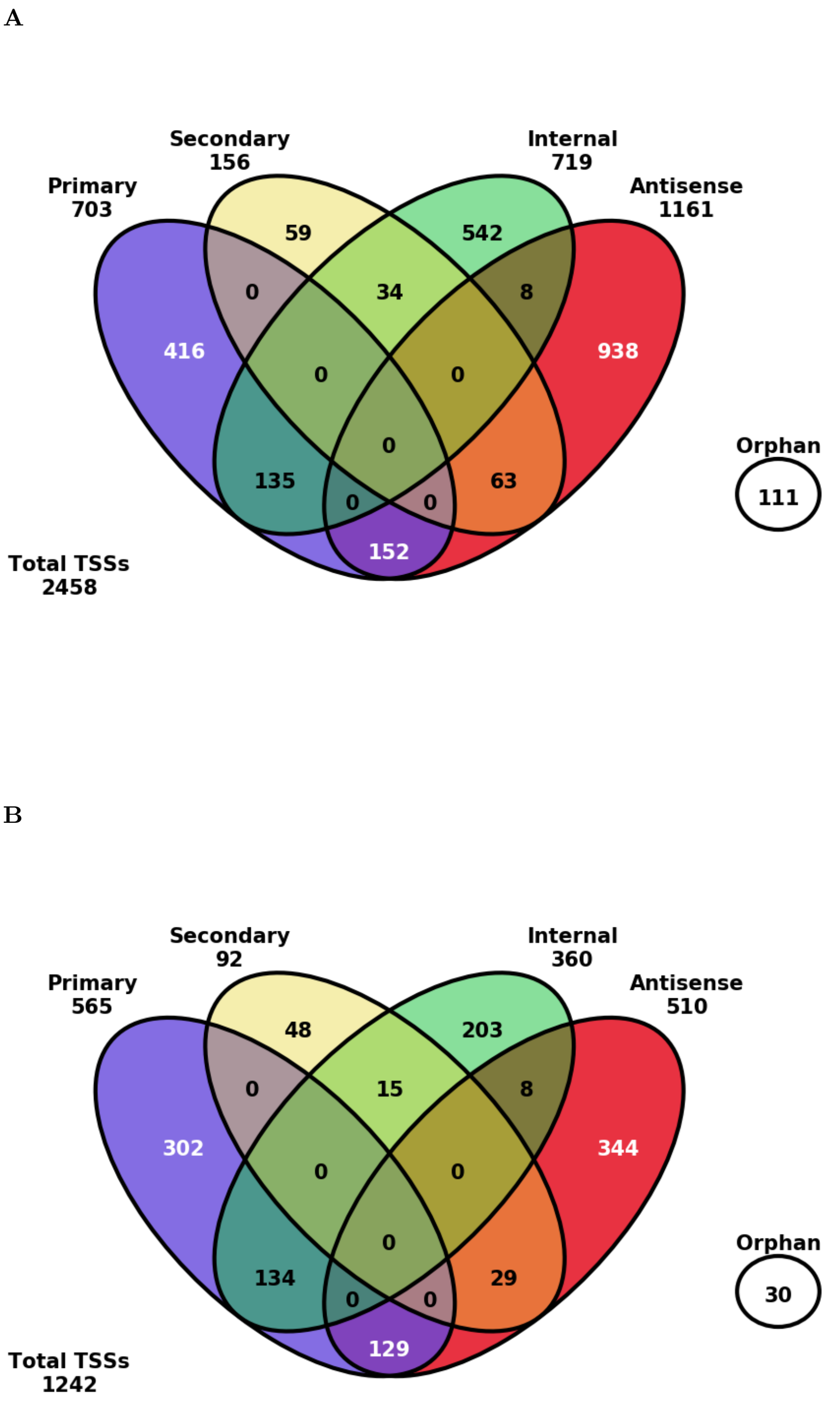

B
Processing site
Genome annotation
Transcript detection
Transcript
TSS / Processing site prediction
Coverage file
Terminator prediction
Terminator
Genome sequence
TSS
UTR prediction
UTR
C
Genome annotation
Transcipt
sRNA prediction
sORF prediction
sRNA candidate
sORF candidate
sRNA target prediction
sRNA target


D
Protein-protein interaction prediction
Genome annotation
Genome sequence
Protein-protein interaction network
GO term detection
Subcellular localization prediction
Go term
Subcellular localization
E
Genome sequence
TSS
Transcript
Terminator
UTR
Genome annotation
Promoter detection
Operon detection
Promoter
Operon

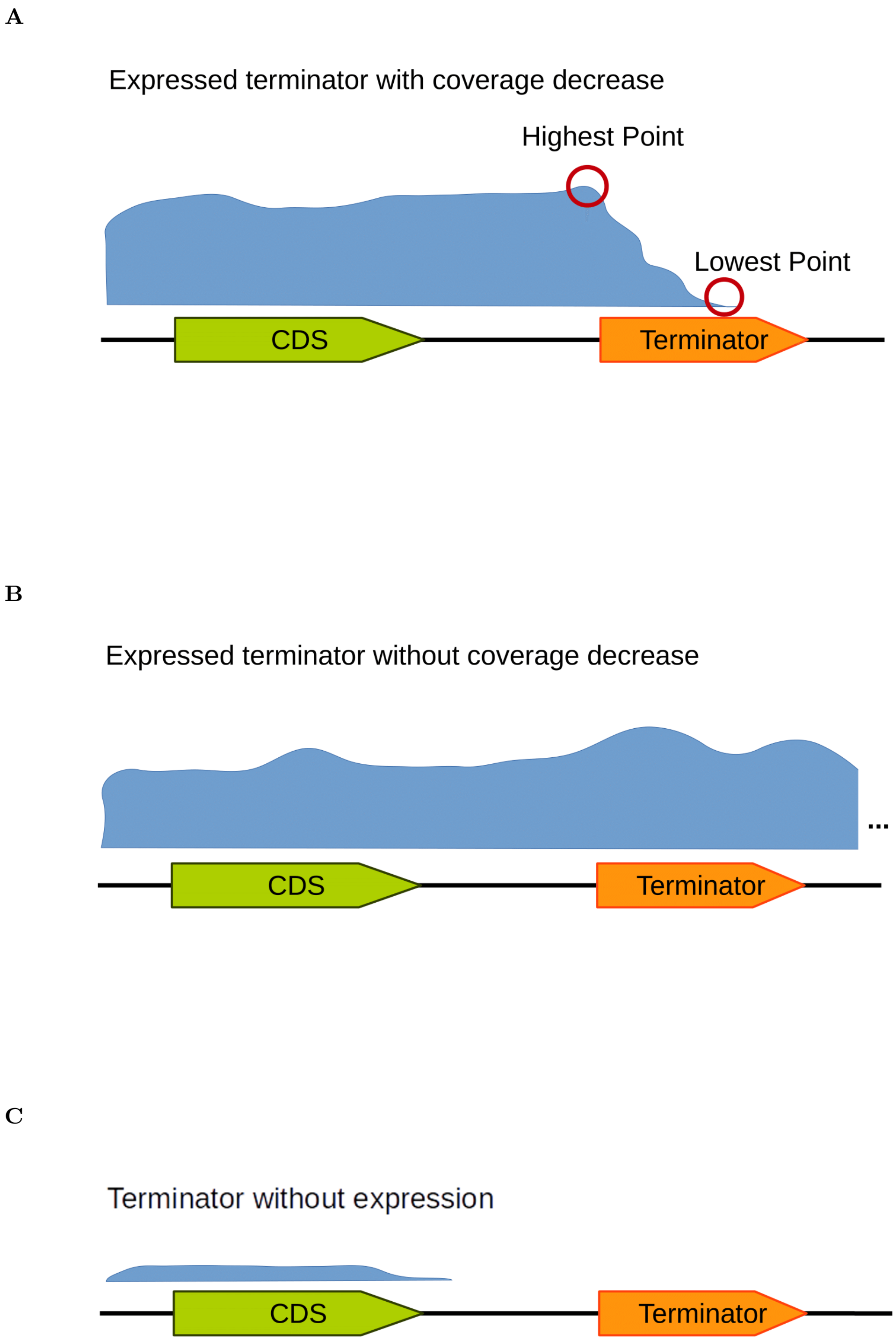
F



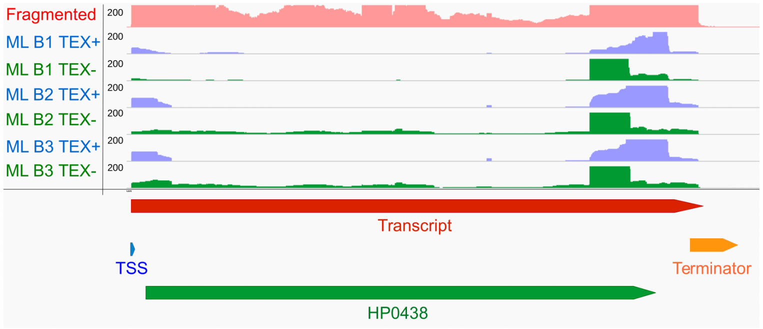
Supplementary Figure 1: Workflow charts of ANNogesic modules. The yellow blocks represent the tools or methods of detection. The red blocks indicate that it is performed by the third-party tools. The blue parallelograms and the green parallelograms are input and output, repectively. **(A)** Reference genome improvement, **(B)** Transcript boundary, **(C)** Small RNA and small ORF, **(D)** Regulatory feature, **(E)** Promoter and Operon and **(F)** Other features.



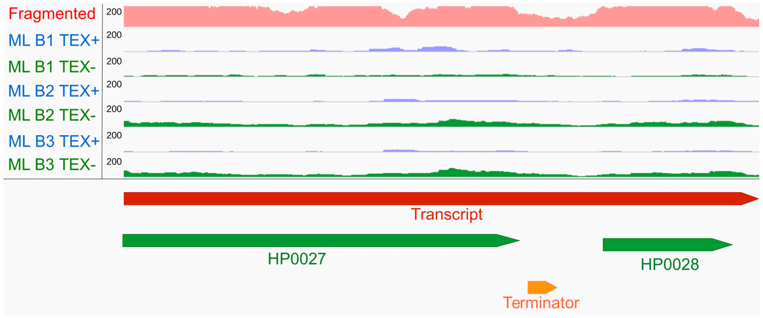
Supplementary Figure 2: The distribution of TSS classes. **(A)** *Helicobacter pylori* 26695. **(B)** *Campylobacter jejuni* 81116.


A
Expressed terminator with coverage decrease
Highest Point
Lowest Point
CDS
Terminator
B
Expressed terminator without coverage decrease
...
CDS
Terminator
C
Terminator without expression
CDS
Terminator

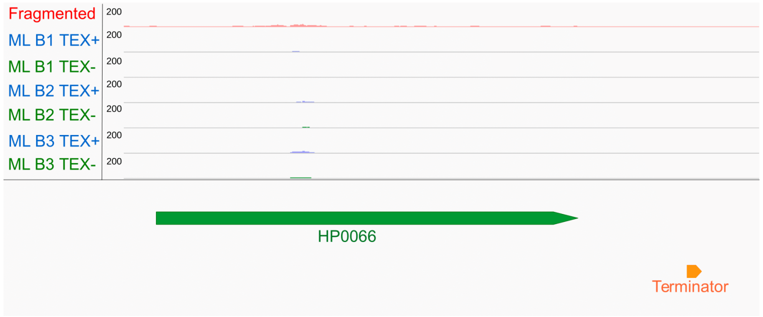
D



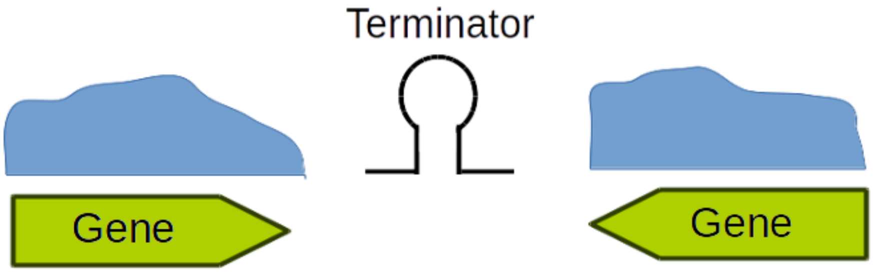
E



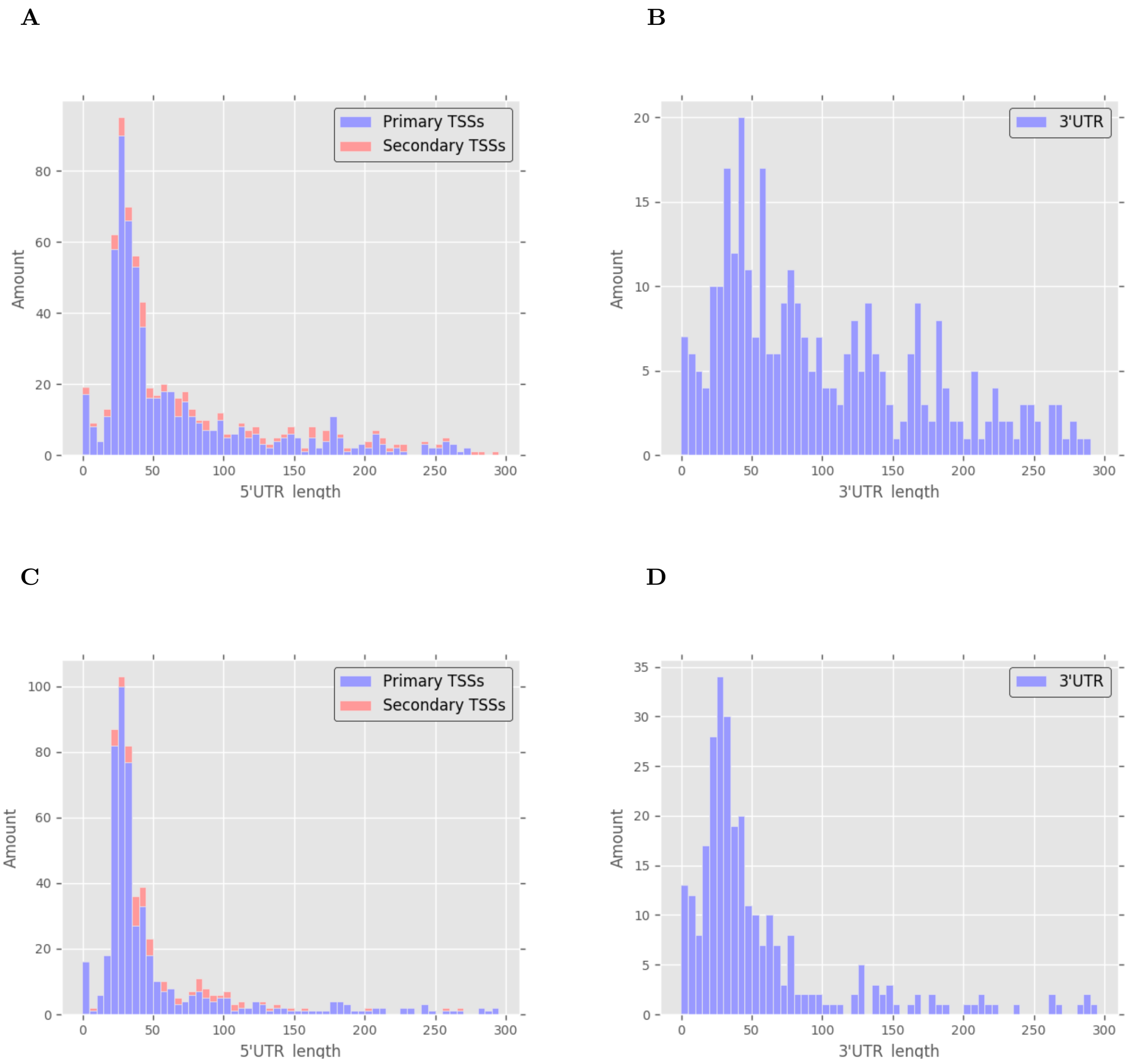
F



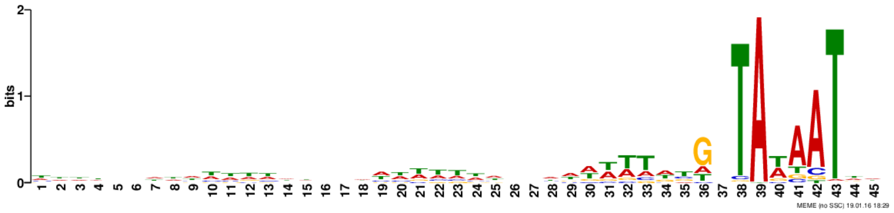
Supplementary Figure 3: The method and examples for detecting coverage decrease of terminators. (A) and (D) An expressed terminator with coverage significant drop. The ratio of the lowest coverage value and the highest coverage values is lower than 0.5 (default). (B) and (E) An expressed terminator without coverage decrease. (C) and (F) A terminator without expression. In (D), (E), and (F), the coverage of RNA-Seq with transcript fragmentation, TEX+ and TEX- of dRNA-Seq are presented as pink, blue and green coverages, respectively. Terminators, TSSs, CDSs and transcripts are showed as orange, blue, green and red bars, respectively.



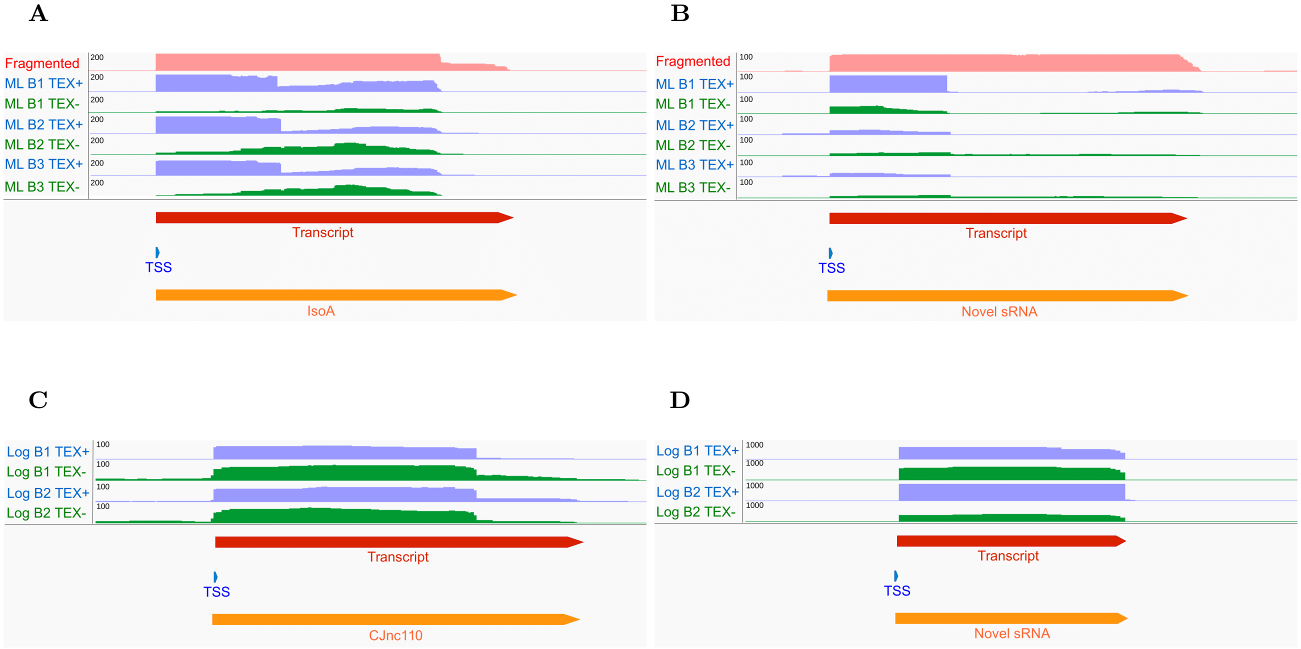
Supplementary Figure 4: Terminator prediction approach based on convergent genes. The blue curve-blocks represent the coverages; the green arrows show two genes from different strands. Ideally, there should be a ρ-independent terminator within the region of two converging genes.



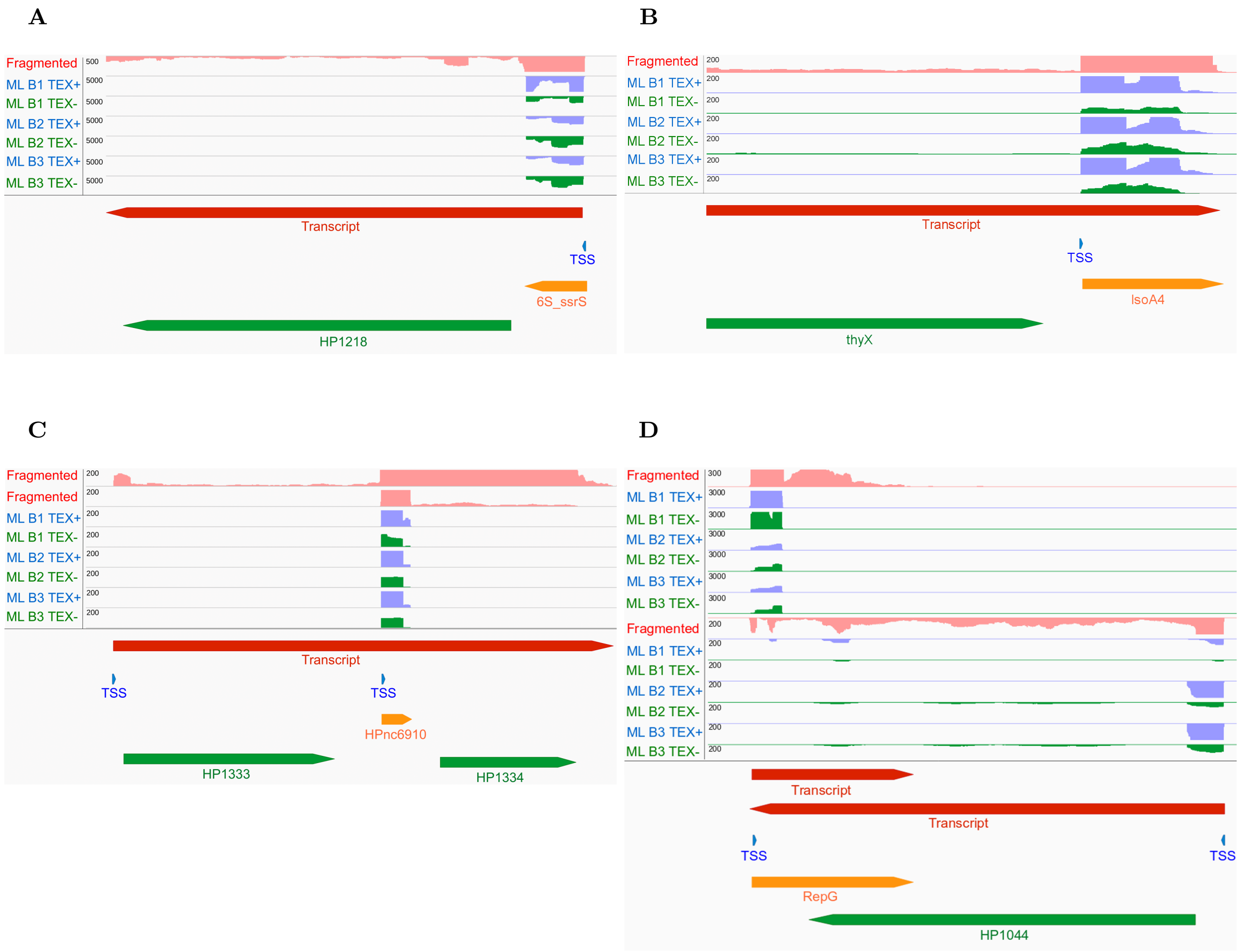
Supplementary Figure 5: The length distribution of UTRs. For 5'UTR the blue bars represent primary TSSs and the pink bars represent secondary TSSs. **(A)** 5'UTRs of *Helicobacter pylori* 26695. **(B)** 3'UTRs of *Helicobacter pylori* 26695. **(C)** 5'UTRs of *Campylobacter jejuni* 81116. **(D)** 3'UTRs of *Campylobacter jejuni* 81116.



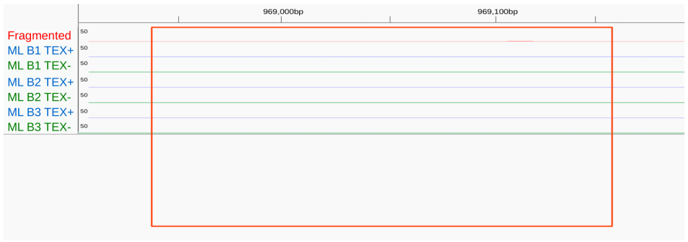
Supplementary Figure 6: The promoter motif detected in *Helicobacter pylori* 26695 (detected in front of 2297 TSSs i.e. 93.4%).



Supplementary Figure 7: The examples of known and novel intergenic sRNAs that ANNOgesic can detect. The coverage of RNA-Seq with fragmentation, TEX+ and TEX- of dRNA-Seq are presented as pink, blue and green coverages, respectively. In the annotation track sRNAs, TSSs, CDSs and transcripts are showed as orange, blue, green and red bars, respectively. **(A)** IsoA (HPnc7630) of *Helicobacter pylori* 26695 **(B)** Novel sRNA in *Helicobacter pylori* 26695 **(C)** CJnc110 in *Campylobacter jejuni* 81116 **(D)** novel sRNA in *Campylobacter jejuni* 81116



Supplementary Figure 8: The examples of antisense and UTR derived sRNAs that ANNOgesic can detect. The coverage of RNA-Seq with fragmentation, TEX+ and TEX- of dRNA-Seq are presented as pink, blue and green coverages, respectively. In the annotation track sRNAs, TSSs, CDSs and transcripts are showed as orange, blue, green and red bars, respectively. **(A)** 5'UTR-derived sRNA – the sRNA and CDS are in the same transcript and the sRNA is located in the 5'UTR. **(B)** 3'UTR-derived sRNA – the sRNA and CDS are in the same transcript and sRNA is located in 3'UTR. **(C)** InterCDS-derived sRNA – the sRNA and CDSs are in the same transcript, and sRNA is located in the non-annotated region between two CDS. The two pink coverages are from the same fragmented library, but presented by different scales. **(D)** Antisense sRNA.



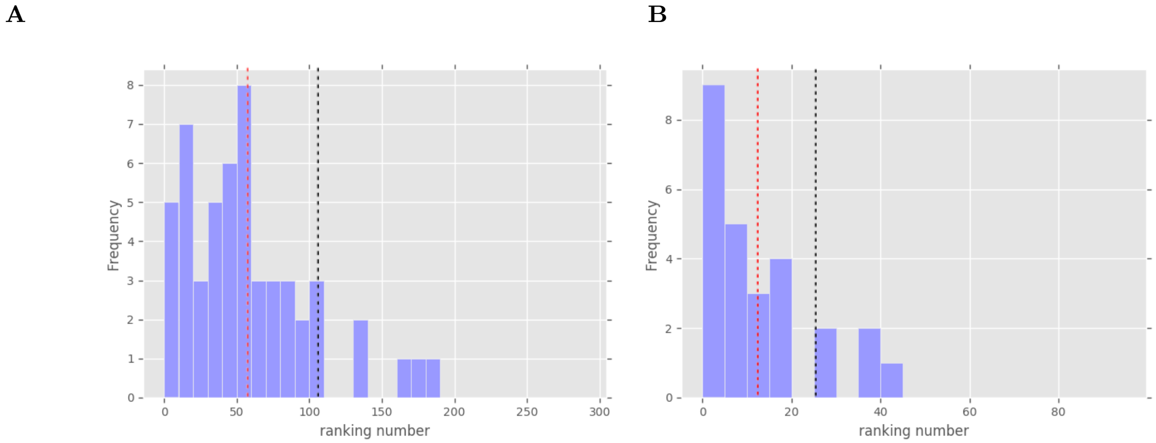
Supplementary Figure 9: The coverage plots of the sRNA HPnc4620 which was excluded from the benchmarking set. It is located at region from base 968980 to 969164 (marked by the orange hollow square) of *Helicobacter pylori* 26695 and has no expression.

Ranking of sRNA

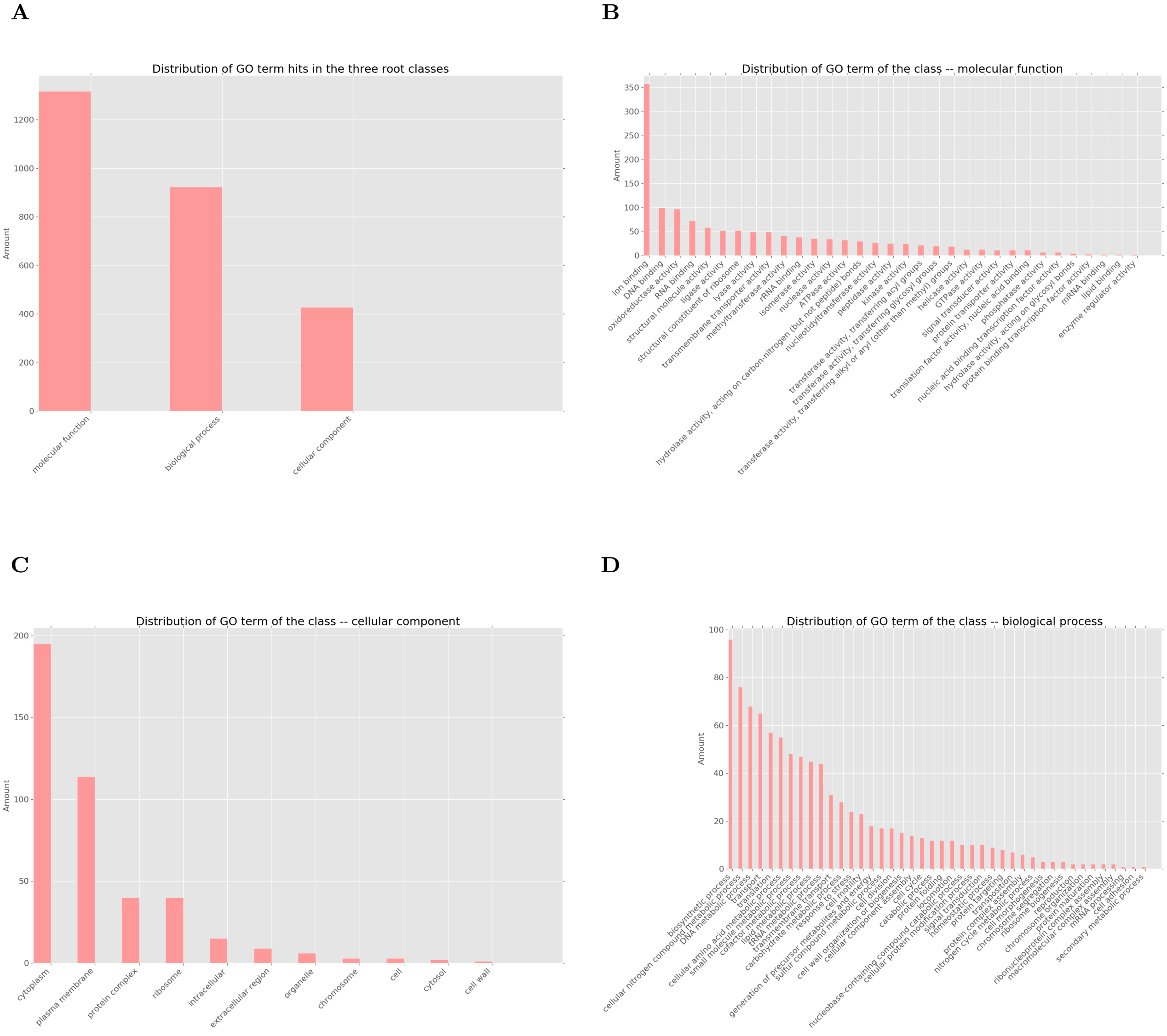
For providing the reliability of sRNA candidates, a ranking system which is based on average coverage and promoter information was implemented. (Supplementary Equation 1). In case a Pribnow box was detected in front of a sRNA, the score is the average coverage value multiplied by 2. If this is not the case, the score is simply the average coverage. The distribution of scores is shown in Supplementary Figure 10. Previously described sRNA show in general a high scoring value. The p-values of t-test between the list of benchmarking sets and the rest population are 1.631e-09 and 4.629e-04 for *Helicobacter pylori* 26695 and *Campylobacter jejuni* 81116, respectively. The results show the ranking system in ANNOgesic is reliable and useful for selection of experimental validation.

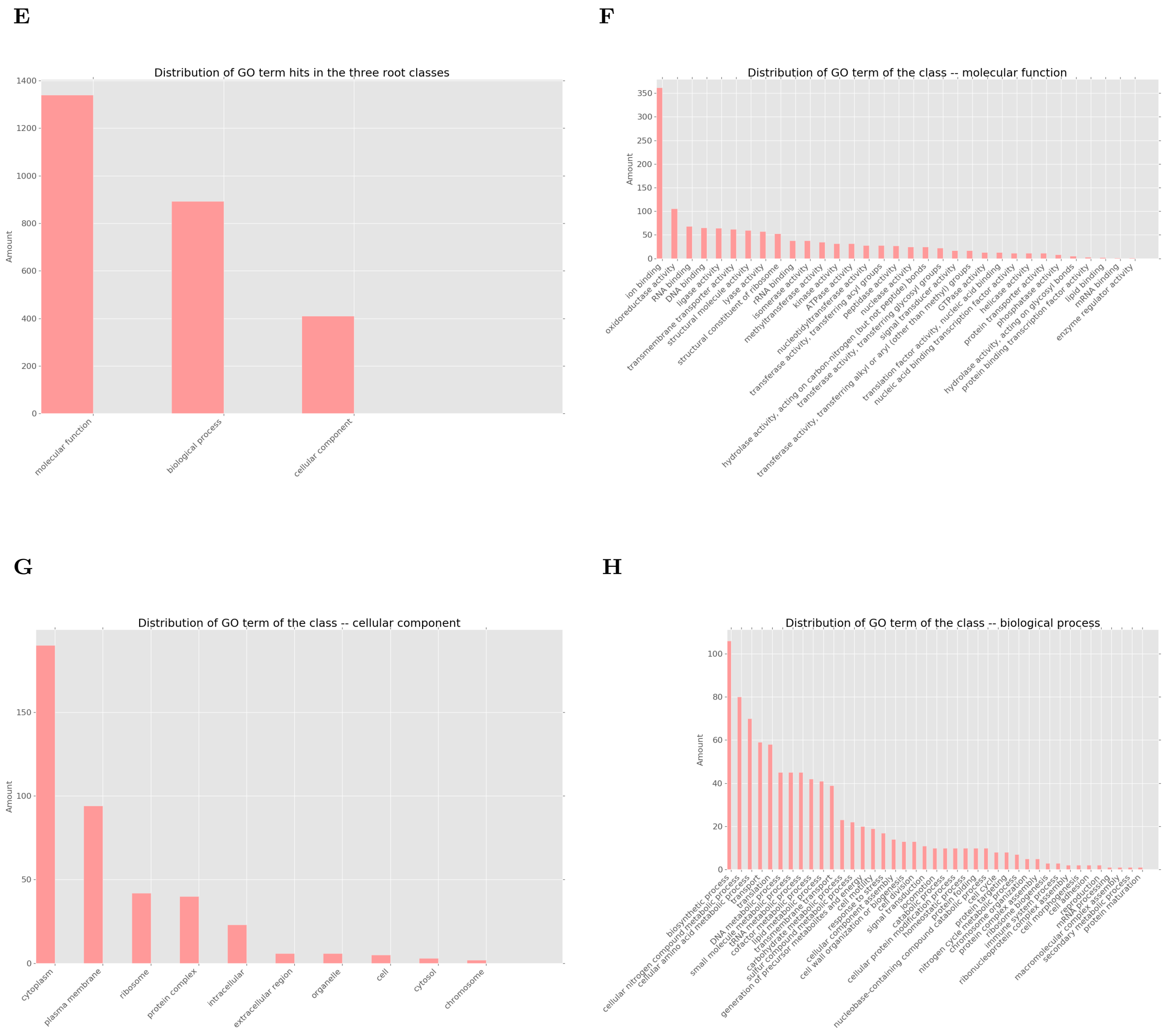
$$
\begin{aligned}
& if\ sRNA\ associated\ with\ promoter:\\
& \quad S = C \times P\\
& else:\\
& \quad S = C
\end{aligned}
$$

Supplementary Equation. 1: S is the score for ranking, C is average coverage, P is the times if sRNA associated with promoter.

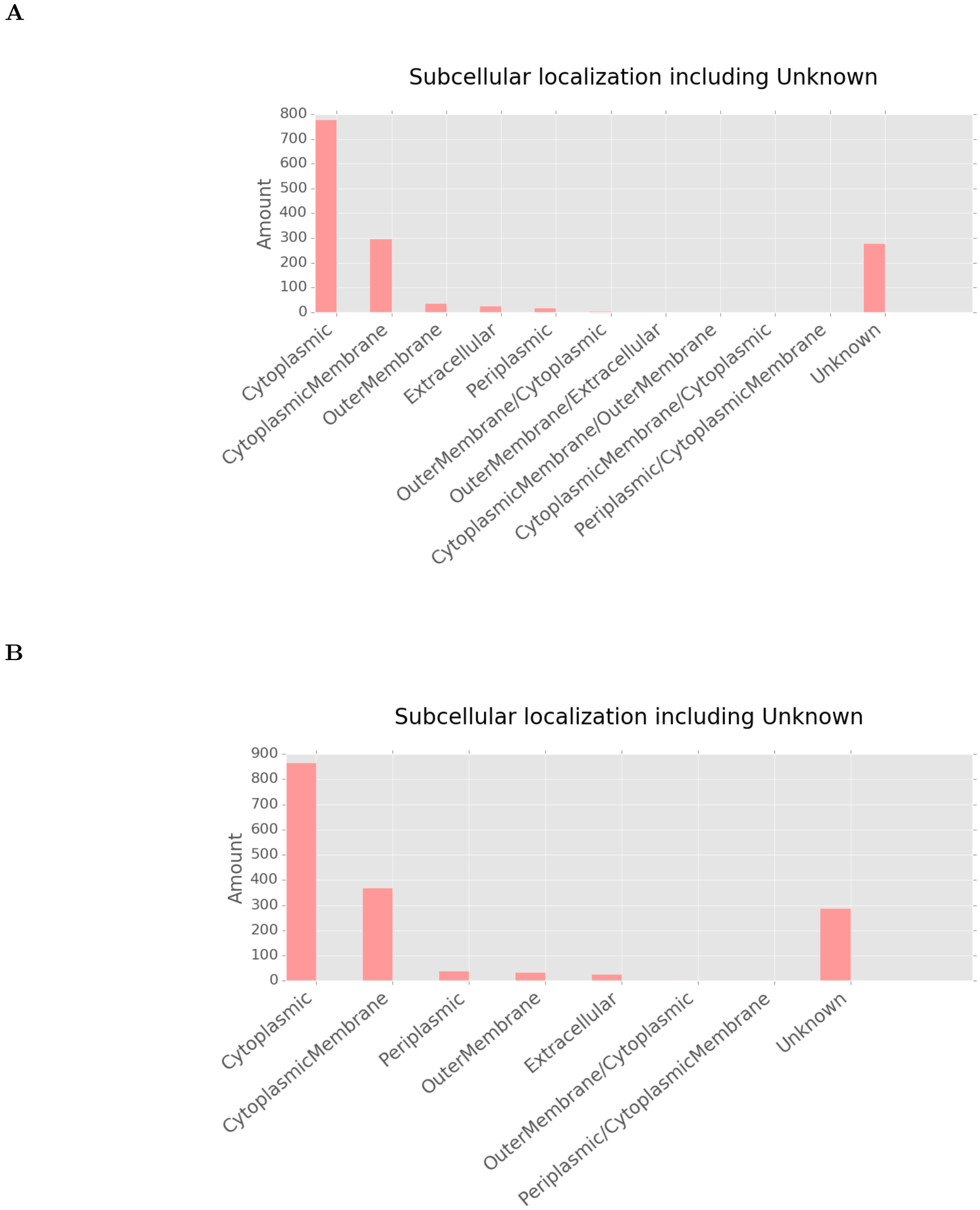


Supplementary Figure. 10: Histograms of ranking number of the sRNA benchmarks. The red dash line represents the average ranking number of benchmarking sets (57.25 and 13.19 of *Helicobacter pylori* 26695 and *Campylobacter jejuni* 81116, respectively), and black dash line shows the average ranking number of the rest populations (106.17 and 25.05 of *Helicobacter pylori* 26695 and *Campylobacter jejuni* 81116, respectively). **(A)** The histogram of *Helicobacter pylori* 26695 and **(B)** The histogram of *Campylobacter jejuni* 81116.

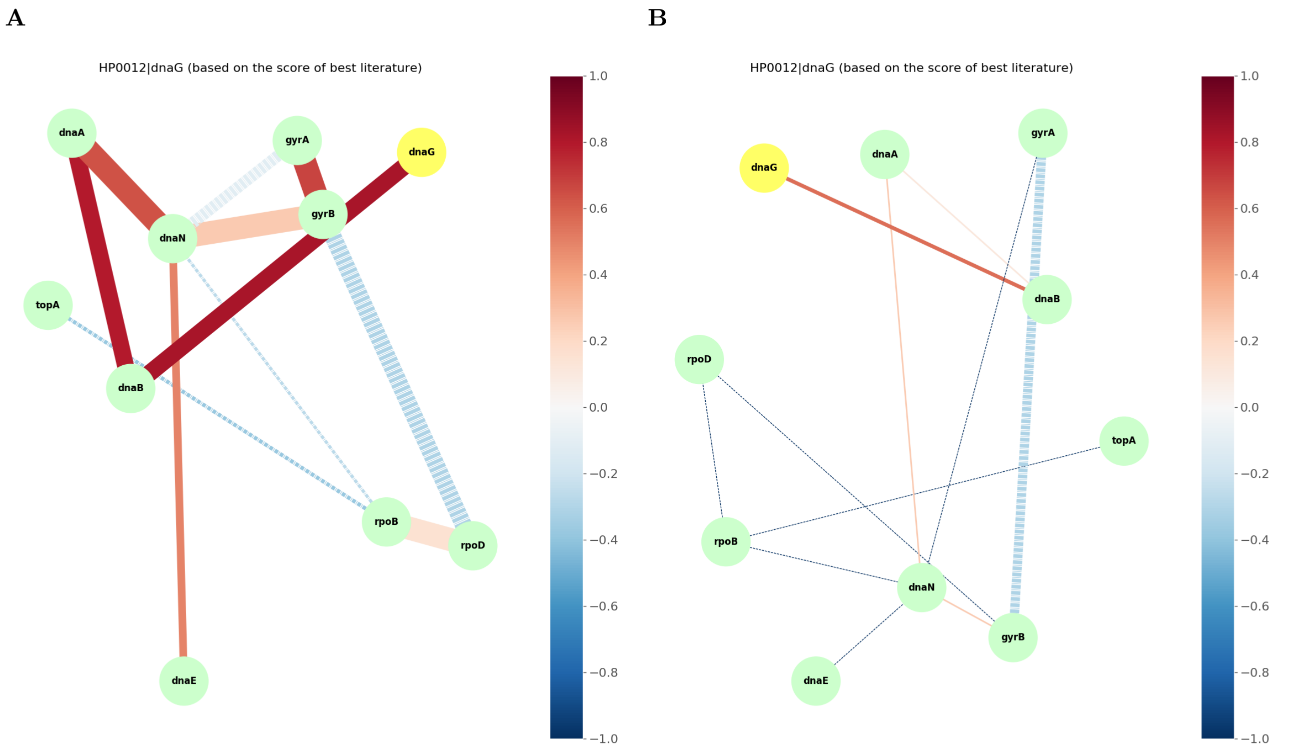

A
Distribution of GO term hits in the three root classes
Amount
0
200
400
600
800
1000
1200
molecular function
biological process
cellular component
B
Distribution of GO term of the class -- molecular function
Amount
0
50
100
150
200
250
300
350
ion binding
DNA binding
oxidoreductase activity
RNA binding
structural molecule activity
ligase activity
structural constituent of ribosome
lyase activity
transmembrane transporter activity
methyltransferase activity
rRNA binding
isomerase activity
nuclease activity
ATPase activity
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
nucleotidyltransferase activity
peptidase activity
kinase activity
transferase activity, transferring acyl groups
transferase activity, transferring glycosyl groups
transferase activity, transferring alkyl or aryl (other than methyl) groups
helicase activity
GTPase activity
signal transducer activity
protein transporter activity
translation factor activity, nucleic acid binding
phosphatase activity
nucleic acid binding transcription factor activity
hydrolase activity, acting on glycosyl bonds
protein binding transcription factor activity
mRNA binding
lipid binding
enzyme regulator activity
C
Distribution of GO term of the class -- cellular component
Amount
0
50
100
150
200
cytoplasm
plasma membrane
protein complex
ribosome
intracellular
extracellular region
organelle
chromosome
cell
cytosol
cell wall
D
Distribution of GO term of the class -- biological process
Amount
0
20
40
60
80
100
biosynthetic process
cellular nitrogen compound metabolic process
DNA metabolic process
transport
translation
cellular amino acid metabolic process
small molecule metabolic process
cofactor metabolic process
lipid metabolic process
tRNA metabolic process
transmembrane transport
carbohydrate metabolic process
response to stress
cell motility
generation of precursor metabolites and energy
sulfur compound metabolic process
cell division
cell wall organization or biogenesis
cellular component assembly
cell cycle
catabolic process
protein folding
locomotion
nucleobase-containing compound catabolic process
cellular protein modification process
signal transduction
homeostatic process
protein targeting
transposition
protein complex assembly
nitrogen cycle metabolic process
cell morphogenesis
chromosome segregation
ribosome biogenesis
reproduction
chromosome organization
protein maturation
ribonucleoprotein complex assembly
macromolecular complex assembly
mRNA processing
cell adhesion
secondary metabolic process



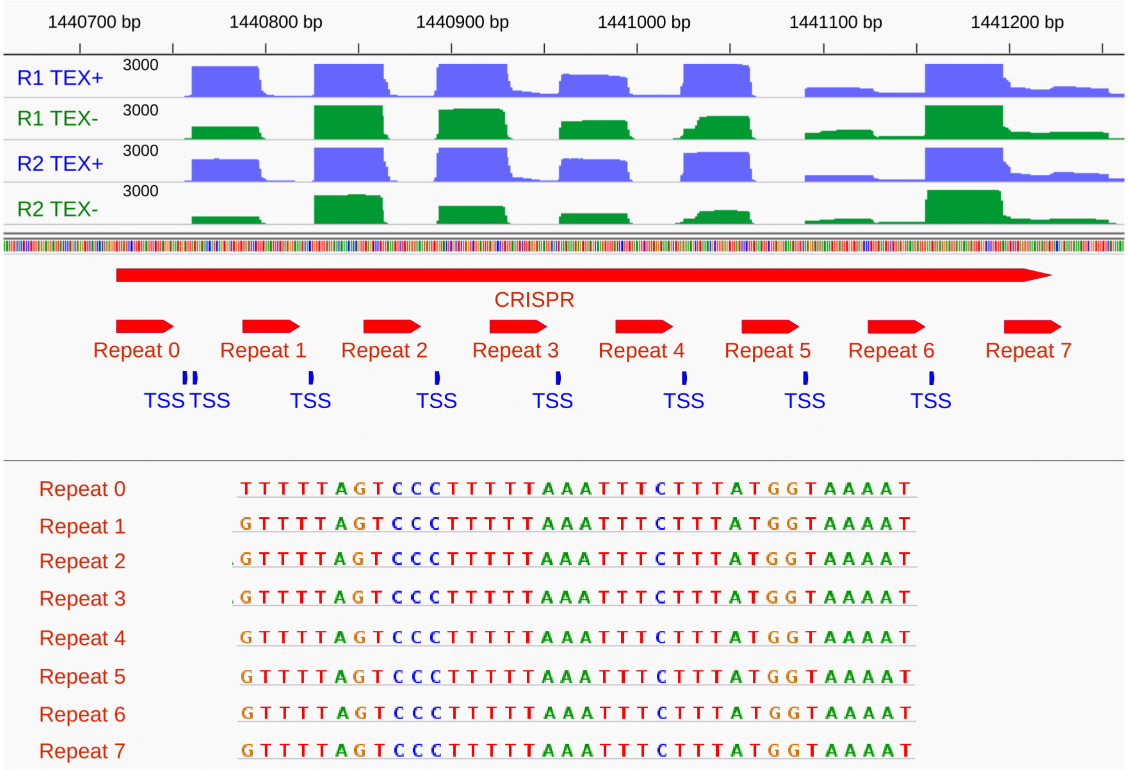
Supplementary Figure. 11: The distributions of GO term. From **(A)** to **(D)** are the distributions of three main domains, molecular function, cellular component and biological process of *Helicobacter pylori* 26695. From **(E)** to **(H)** are the distributions of three main domains, molecular function, cellular component and biological process of *Campylobacter jejuni* 81116.



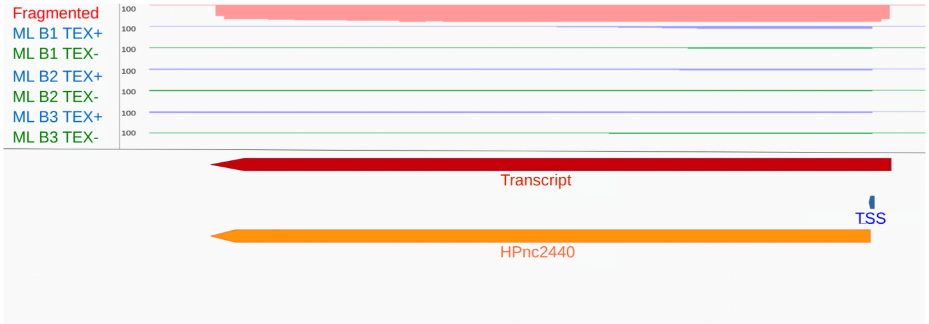
Supplementary Figure. 12: The distributions of subcellular localization of proteins for **(A)** *Helicobacter pylori* 26695, and **(B)** *Campylobacter jejuni* 81116.



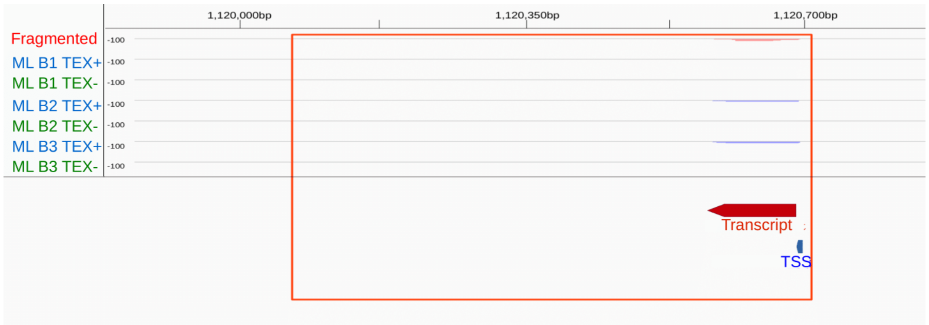
Supplementary Figure. 13: Visualization of protein-protein interactions. The yellow circles represent the query protein (dnaG) in *Helicobacter pylori* 26695. The other proteins are related to the query one showed as green circles. The dotted lines represent the interactions without support in the literature; the dashdot lines represent the interactions with literature support but scores (given by PIE) below 0; a solid lines indicate that the interactions are supported in the literatures with high PIE score (higher than 0); the thickness of the lines is proportional to the number of articles that report the interaction; the color of connections encode score reported by PIE. **(A)** The result of search with the text "Helicobacter pylori" **(B)** The result of search with only protein names (without the strain name).


1440700 bp
1440800 bp
1440900 bp
1441000 bp
1441100 bp
1441200 bp
R1 TEX+
3000
R1 TEX-
3000
R2 TEX+
3000
R2 TEX-
3000
CRISPR
Repeat 0
Repeat 1
Repeat 2
Repeat 3
Repeat 4
Repeat 5
Repeat 6
Repeat 7
TSS TSS
TSS
TSS
TSS
TSS
TSS
TSS
Repeat 0 TTTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT
Repeat 1 GTTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT
Repeat 2 GTTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT
Repeat 3 GTTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT
Repeat 4 GTTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT
Repeat 5 GTTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT
Repeat 6 GTTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT
Repeat 7 GTTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT

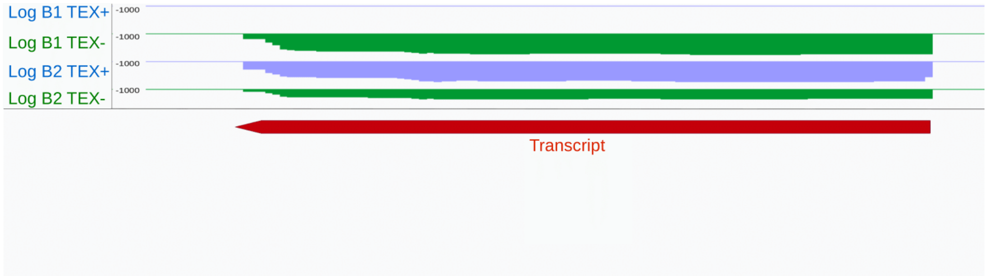
Supplementary Figure. 14: The example of CRISPR in *Campylobacter jejuni* 81116. The coverage of TEX+ and TEX- libraries of dRNA-Seq are presented as blue and green coverages, respectively. Red Bars represent CRISPR with repeat units, and Blue spots mean TSSs. Moreover, the repeat sequences are showed at the bottom.



Supplementary Figure. 15: The sRNA which can be detected only in data RNA-Seq after transcript fragementation. The coverage of RNA-Seq with fragmentation, TEX+ and TEX- libraries of dRNA-Seq are presented as pink, blue and green coverages, respectively.



Supplementary Figure. 16: The lowly expressed sRNA (HPnc4610 – located in the region 968583 to 968616, orange hollow square) cannot be detected by ANNOgesic. The coverage of RNA-Seq with fragmentation, TEX+ and TEX- libraries of dRNA-Seq are presented as pink, blue and green coverages, respectively. TSS, CDS, and transcript are represented as blue, green and red bars, respectively. The average coverage of the low expressed benchmark is around 8 in the RNA-Seq data of the fragmentated librara and lower than 1 in the dRNA-Seq library.



Supplementary Figure. 17: The example of benchmark (CJnc230 of *Campylobacter jejuni* 81116) which is not associated with a TSS. The blue coverages shows the TEX+ libraries of dRNA-Seq and green coverages represents the TEX- libraries of dRNA-Seq.

Supplementary Tables

Supplementary Table 1: The default cutoffs of coverage for sRNA prediction

Methods of RNA-Seq	sRNA types	TSS types	Coverage
dRNA-Seq (TEX+)	intergenic	primary	not included
		secondary	not included
		internal	not included
		antisense	40 reads
		orphan	20 reads
	5'UTR	all	80 percentile
	3'UTR	all	60 percentile
	interCDS	all	70 percentile
dRNA-Seq (TEX-)	intergenic	primary	not included
		secondary	not included
		internal	not included
		antisense	30 reads
		orphan	10 reads
	5'UTR	all	70 percentile
	3'UTR	all	50 percentile
	interCDS	all	60 percentile
fragmenetation RNA-Seq	intergenic	primary	400 reads
		secondary	200 reads
		internal	not included
		antisense	50 reads
		orphan	20 reads
	5'UTR	all	70 percentile
	3'UTR	all	50 percentile
	interCDS	all	60 percentile

The minimum coverage of UTR-derived sRNA must be higher than 50.