# ANNOgesic: A Swiss army knife for the RNA-Seq based annotation of bacterial/archaeal genomes

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 and [2]Core Unit Systems Medicine, University of Würzburg, 97080 Würzburg, Germany

January 27, 2018

**Abstract**

To understand the gene regulation of an organism of interest, a comprehensive genome annotation is essential. While some features, such as coding sequences, can be computationally predicted with high accuracy based purely on the genomic sequence, others, such as promoter elements or non-coding RNAs are harder to detect. RNA-Seq has proven to be an efficient method to identify these genomic features and to improve genome annotations. However, processing and integrating RNA-Seq data in order to generate high-resolution annotations is challenging, time consuming and requires numerous different steps. We have constructed a powerful and modular tool called ANNOgesic that provides the required analyses and simplifies RNA-Seq-based bacterial and archaeal genome annotation. It can integrate data from conventional RNA-Seq and dRNA-Seq, predicts and annotates numerous features, including small non-coding RNAs, with high precision. The software is available under an open source license (ISCL) at https://pypi.org/project/ANNOgesic/.

## 1 Background

As the number of available genome sequences has rapidly expanded in the data bases, numerous tools have been developed that can detect genomic features of interest based on the genome sequence. Prominent representatives are Glimmer to identify open reading frames (ORFs) [13], tRNAscan-SE [56] to spot tRNAs and RNAmmer to find rRNAs [35]. Pipelines like Prokka [57] or ConsPred [68] combine

*https://orcid.org/0000-0001-7955-8645

†https://orcid.org/0000-0003-2220-1404

‡https://orcid.org/0000-0002-1481-2996; To whom correspondence should be addressed. Tel: +49-931 31-80403 ; Email: konrad.foerstner@uni-wuerzburg.de

such tools and are able to search multiple features in bacterial and archaeal genomes. Still, these tools that make their predictions purely based on the genome sequences can predict features like transcriptional start sites and non-coding RNAs, if at all, only with low confidence.

Recent developments in high-throughput sequencing offer solutions to this problem. RNA-Seq has revolutionized how differential gene expression can be measured and is widely used for this purpose [47]. Besides this it has also been applied in numerous cases to improve the genome annotation of bacteria [3, 6, 58], archaea [69] and eukaryotes [23]. RNA sequencing based protocols like differential RNA-Seq (dRNA-Seq) [4, 59], Term-seq [12] and ribosome profiling [29, 66] have been applied to globally detect transcriptional start sites (TSSs), small non-coding RNAs (ncRNAs/sRNAs), terminators, ORFs and sRNA regulatory target but require dedicated data processing. While there are tools that can process RNA-Seq data in order to predict genome-wide features like TSSs based on dRNA-Seq data [1, 16, 30] or based on conventional RNA-Seq data [42, 54], there has been, so far, no solution that combines different predictions of genomic features and compiles them into a consistent annotation.

Here we present ANNOgesic - a modular, command-line tool that can integrate data from different RNA-Seq protocols like dRNA-Seq as well as conventional RNA-Seq performed after transcript fragmentation and generate high-quality genome annotations that include features missing in most bacterial annotations (e.g. small non-coding RNAs, UTRs, TSSs, operons). The central approach is to detect transcript boundaries and then subsequently attach further information about type as well as function to the predicted features and also to infer interactions between them. Several of ANNOgesic's core functions represent new implementations that are not found in other programs. Those are accompanied by third-party tools that are embedded into ANNOgesic and are due to that accessible via a consistent command-line interface. Furthermore their results are improved e.g. by dynamic parameter-optimizations or by removing false positives. Numerous visualizations and statistics help the user to quickly evaluate the feature predictions. The tool is modular and was intensively tested with several RNA-Seq data sets from bacterial as well as from archaeal species.

# 2 Materials and Methods

## 2.1 Modules of ANNOgesic

ANNOgesic consists of the following modules - their names indicate their functions: Sequence modification, Annotation transfer, SNP/Mutation, Transcript, TSS, Terminator, UTR, Processing site (PS), Promoter, Operon, sRNA, sRNA target, sORF, GO term, Protein-protein interaction network, Subcellular localization, Riboswitch, RNA thermometer, Circular RNA, and CRISPR. Several potential workflows connecting these modules are displayed in Figure S1. An overview of the novelties and improvements of the modules in ANNOgesic are listed in Table S1.

Depending on the task to one wish to perform, ANNOgesic requires a specific set

of input information - either as coverage information in wiggle format or alignments in BAM format. This can be generated by short read aligners like STAR [15], segemehl [27], or a full RNA-Seq pipeline like READemption [18]. Certain modules additionally require annotations in GFF3 format. In case a sufficient genome annotation is not available, ANNOgesic can perform an annotation transfer from a closely related strain.

## 2.2 Implementation and installation

ANNOgesic's source code is implemented in Python 3 and hosted at `https://pypi.org/project/ANNOgesic/`. The comprehensive documentation can be found at http://annogesic.readthedocs.io/ and releases are automatically submitted to Zenodo (https://zenodo.org/) to guarantee a long term availability. It can be easily installed using pip (https://pip.pypa.io). In order to provide a frictionless installation including non-Python dependencies, we additionally offer a Docker image (https://hub.docker.com/r/silasysh/annogesic/) [44].

## 2.3 Optimization of the parameter set for TSSpredator

For several parts of ANNOgesic the selection of parameters has a strong impact on the final results. Especially the TSS prediction – building on TSSpredator [16] – requires a sophisticated fine-tuning of several parameters (namely height, height reduction, factor, factor reduction, enrichment factor, processing site factor and base height). To overcome the hard task of manual parameter selection, ANNOgesic optimized the parameters by applying a genetic algorithm, a machine learning approach, [21] which is trained based on a small user curated set of TSS predictions. This approach has the advantage of being able to find global, not only local, optima. The process of optimization is composed of three parts - random change, large change, and small change (Figure 1). In this context, a global change means a random allocation of values to all parameters, a large change is a random allocation of values to two parameters, while a small change is adding or subtracting a small fraction to or from one parameter value. The result of each iteration is evaluated by a decision statement (Equation 1).

## 2.4 Test data sets

In order to test ANNOgesic's performance, we applied it to RNA-Seq data sets originating from *Helicobacter pylori* 26695 [4, 58] and *Campylobacter jejuni* 81116 [16]. The dRNA-Seq data sets were retrieved from NCBI GEO where they are stored under the accession numbers GSE67564 and GSE38883, respectively. For *H. pylori* conventional RNA-Seq data – i.e. without TEX treatment (which degrades transcripts without a 5'-triphosphate) and with fragmentation of the transcript before the library preparation – was also retrieved from NCBI SRA (accession number SRR031126). Moreover, for assessing the performance of ANNOgesic, dRNA-Seq and conventional RNA-Seq data sets of *Escherichia coli* K12 MG1655 were downloaded from NCBI GEO (accession number: GSE55199 and GSE45443 (only the data of the wild type

strain was retrieved) [42, 65]. The ANNOgesic predictions generated using these data sets of *Escherichia coli* K12 MG1655 were compared to the databases RegulonDB, EcoCyc and DOOR[2] [20, 22, 24, 31, 41, 51].

# 3 Results

## 3.1 Correction of genome sequences and annotations

All genomic features that can be detected by ANNOgesic are listed in Table 1. In order to demonstrate and test ANNOgesic's performance we have analyzed RNA-Seq data of *H. pylori* 26695 and *C. jejuni* 81116 and discuss the prediction results as examples in the following sections.

Table 1: Overview of feature predictions for *H. pylori* 26695 and *C. jejuni* 81116

| | | *H. pylori* 26695 | *C. jejuni* 81116 |
|---|---|---|---|
| Gene | | 1560 | 1685 |
| CDS | Total | 1448 | 1630 |
| | Expressed | 1406 | 1513 |
| Transcript | | 1716 | 1147 |
| TSS | Total | 2458 | 1242 |
| | Primary | 703 | 565 |
| | Secondary | 156 | 92 |
| | Internal | 719 | 360 |
| | Antisense | 1161 | 510 |
| | Orphan | 111 | 30 |
| Processing site | | 281 | 345 |
| Terminator | Total | 820, (437) | 874, (375) |
| | TransTermHP | 631, (314) | 655, (269) |
| | Convergent genes | 229, (151) | 276, (145) |
| UTR | 5' UTR | 693 | 560 |
| | 3' UTR | 325 | 286 |
| sRNA | Total | 184 | 40 |
| | Intergenic | 60 | 16 |
| | Antisense | 85 | 21 |
| | 5' UTR-derived | 10 | 0 |
| | 3' UTR-derived | 23 | 2 |
| | InterCDS-derived | 6 | 1 |
| Operon | Total | 1716 | 1147 |
| | Monocistronic | 269 | 386 |
| | Polycistronic | 285 | 324 |
| | No CDS associated | 1162 | 437 |
| sORF | | 150 | 25 |
| Riboswitch | | 11 | 14 |
| RNA thermometer | | 4 | 8 |
| circular RNA | | 0 | 1 |
| CRISPR | | 0 | 1, (8) |

The numbers in brackets for terminator and CRISPR represent occurances of terminators with coverage drop and repeat units of CRISPR, respectively. For the prediction of terminators ANNOgesic only keeps the high confidence ones in case a CDS is associated with multiple terminators.

#### 3.1.1 Genome sequence improvement and SNP/mutation calling

Conventionally, differences in the genome sequence of a strain of interest and the reference strain are determined by DNA sequencing. However, RNA-Seq reads can

also be re-purposed to detect such SNPs or mutations that occur in transcribed regions which can help to save the resources required for dedicated DNA sequencing or DNA SNP microarray measurements. The two drawbacks of this method are that only locations which are expressed can be analyzed and that, due to RNA editing, changes could be present only in the RNA level and are not found in the genome. On the other hand, it has been shown to be a valid approach for eukaryotic species and that the majority of SNPs are found in the expressed transcripts [10, 11]. In conclusion, such an analysis could be useful to generate hypotheses that then need to be tested with complementary methods. ANNOgesic offers the user to perform the SNP/mutation calling via SAMtools [36] and BCFtools [36] applying read counting-based filtering.

#### 3.1.2 Annotation transfer

ANNOgesic integrates RATT [50], which can detect the shared synteny and mutations between a reference and query genome to transfer annotation (i.e. genes, CDSs, tRNAs, rRNAs) by applying MUMmer [34]. For the chosen strains, *H. pylori* 26695 and *C. jejuni* 81116 annotation files in GFF3 format were be obtained from NCBI RefSeq. Due to this there was no need to transfer the annotation from a closely related strain.

### 3.2 Detection of transcripts

Knowing the exact boundaries and sequence of a transcript is crucial for a comprehensive understanding of its behaviour and function. For example, UTRs can be the target of regulation by sRNAs or small molecules (e.g. riboswitches) [7, 67] or even sources of sRNAs [9]. Unfortunately, most bacterial annotations only cover the protein coding regions while the information about TSSs, terminators and UTRs is lacking. To address this issue, ANNOgesic combines several feature predictions for a reliable detection of transcripts and their boundaries (Figure 2).

#### 3.2.1 Coverage based transcript detection

There are numerous tools available for the detection of transcripts (e.g. [17]), but most of them are optimized for the assembly of eukaryotic transcripts. Due to this, we combined several heuristics to perform such predictions. Nucleotide coverage data is used for define the expressed regions, and genome annotations are applied for extending or merging the gene expressed regions to form complete transcripts. Several parameters like the threshold of coverage values can be set by the user to fine-tune the predictions (Figure 3).

By running ANNOgesic's subcommand for transcript prediction, we detected 1715 transcripts in *H. pylori* 26695 and 1147 transcripts in *C. jejuni* 81116. These transcripts cover 1520 and 1568 genes which shows that 97% and 93% of the known genes are expressed in at least one condition, respectively.

### 3.2.2 Transcriptional start sites

For the prediction of TSSs, ANNOgesic builds on TSSpredator [16], which takes dRNA-Seq coverage data as input. The outcome of TSSpredator's predictions depends strongly on the setting of numerous parameters and fine-tuning those can be time consuming. Due to this, a parameter optimization was implemented in ANNOgesic that builds on a small, manually curated set of TSSs to find optimal values.

In order to test the performance of ANNOgesic, we manually annotated TSSs in the first 200 kb of the genome of *H. pylori* 26695 and *C. jejuni* 81116 (Table S4 and S5). This set was used to benchmark the prediction of TSSpredator with default settings as well with the parameters optimized by ANNOgesic. For the test set, we manually annotated TSSs from first 200 kb to 400 kb in the genome of *H. pylori* 26695 and *C. jejuni* 81116 (Table S4 and S5). As displayed in Table 2, the optimization had minor sensitivity improvements in *H. pylori* 26695 (from 96.8% to 99.6%), while strongly increased the sensitivity for the TSS prediction for *C. jejuni* 81116 (67.1% to 98.7%) while keeping the same level of specificity. To underpin those findings, we looked at the overlap of the predicted TSS and predicted transcripts. This was nearly the same for *H. pylori* 26695 (82% for default and 83% for optimized setting) but also increased significantly for *C. jejuni* 81116 from 81% for default parameters to 96% with optimized parameters.

Moreover, TSSs are classified depending on their relative positions to genes by TSSpredator. Based on these classifications, Venn diagrams representing the different TSS classes are automatically generated (Figure S2).

### 3.2.3 Processing sites

Several transcripts undergo processing, which influences their biological activity [9,25]. In order to detect PSs based on dRNA-Seq data, ANNOgesic facilitates the same approach as described for TSS detection but searches for the reverse enrichment pattern (i.e. a relative enrichment in the library not treated with TEX). As done for the TSSs, we manually annotated the PSs in the first 200 kb of the genomes. Based on these manually curated sets, we performed parameter optimization on the test set (manually-curated from first 200 kb to 400 kb, Table S6 and S7, Table 2) and could improve the prediction of PSs by TSSpredator [16]. With optimized parameters 281 and 345 PSs were detected in *H. pylori* 26695 and *C. jejuni* 81116, respectively.

Table 2: Comparison of default and optimized parameters of TSSpredator for TSS and PS prediction

| Strains | Parameters | Sensitivity (TP) | Specificity (FP) |
|---|---|---|---|
| | TSS | | |
| *H. pylori* 26695 | Default | 96.8% (244) | 99.98% (32) |
| | Optimization | 99.6% (251) | 99.98% (32) |
| *C. jejuni* 81116 | Default | 67.1% (104) | 99.98% (31) |
| | Optimization | 98.7% (153) | 99.99% (7) |
| | PS | | |
| *H. pylori* 26695 | Default | 92.9% (26) | 99.99% (7) |
| | Optimization | 92.9% (26) | 99.99% (7) |
| *C. jejuni* 81116 | Default | 61.3% (19) | 99.99% (2) |
| | Optimization | 93.5% (29) | 99.99% (6) |

The numbers in brackets represent true positive or false positive.

#### 3.2.4 $\rho$-independent terminators

While the TSSs are in general clearly defined borders, the 3'-end of a transcript is often not very sharp. A commonly used tool for the prediction of the 3'-end of a transcript is TransTermHP [33], which detects $\rho$-independent terminators based on genome sequences. Manual inspection showed us that TransTermHP predictions are not always supported by the RNA-data (Figure S3E and S3F). This could be due to the lack of expression in the chosen conditions. Additionally, certain locations in 3'-ends that may be $\rho$-independent were not detected by TransTermHP. Due to this, we extended the prediction by two further approaches based on RNA-Seq coverage and the given genome sequence. At first, terminators predicted by TransTermHP that show a significant decrease of coverage are marked as high-confidence terminators. For this, the drop of coverage inside the predicted terminator region plus 30 nucleotides up and downstream is considered as sufficient if the ratio of the lowest coverage value and the highest coverage value is at a user-defined value (the default value is 0.5, and the schemes and examples are shown in Figure S3). In order to improve the sensitivity, an additional heuristic for the detection of $\rho$-independent terminators was developed. In this approach, only converging gene pairs (i.e. the 3'-end are facing to each other) are taken into account (Figure S4). In case the region between the two genes is A/T-rich and a stem-loop can be predicted in there, the existence of a $\rho$-independent terminator is assumed. As default, the region should consist of 80 or less nucleotides, the T-rich region should contain more than 5 Thymines, the stem-loop needs to be 4 - 20 nucleotides, the length of the loop needs to be between 3 and 10 nucleotides and maximum 25% of the nucleotides in the stem should be unpaired.

### 3.2.5 UTRs

Based on the CDS locations and the above described detection of TSSs, terminators and transcripts, 5' UTR and 3' UTR can be annotated by ANNOgesic. Additionally, it visualizes the distribution of UTR lengths in a histogram (as shown in Figure S5).

### 3.2.6 Promoters

ANNOgesic integrates MEME [2] which detects ungapped motifs and GLAM2 [19] which discovers gapped motifs for the detection and visualization of promoter motifs. The user can define the number of nucleotides upstream of TSSs that should be screened and the length of potential promoter motifs. The motifs can be generated globally or for the different types of TSSs (example in Figure S6).

### 3.2.7 Operons

Based on the TSS and transcript prediction, ANNOgesic can generate statements regarding the organization of genes in operons and suboperons as well as report the number of monocistronic operons and polycistronic operons (Figure 4).

## 3.3 Detection of sRNAs and their targets

The detection of sRNAs based on RNA-Seq data is a non-trivial task. While numerous sRNAs are found in intergenic regions, several cases of 5' and 3' UTR-derived sRNAs are reported [9, 28, 45, 60]. ANNOgesic offers the detection of all classes, combined with a detailed characterization of the sRNA candidates (Figure 5).

In order to classify newly detected intergenic transcripts as sRNAs, ANNOgesic tests several of their features. If a BLAST+ [8] search of a transcript finds homologous sequences in BSRD [37] – a database that stores experimentally confirmed sRNAs – the transcript gets the status of an sRNA. The user can also choose further databases for searching homologous sequences. In case a search against the NCBI non-redundant protein database leads to a hit it is marked as potentially protein-coding. Otherwise, a transcript must have a predicted TSS, form a stable secondary structure (i.e. the folding energy change calculated with RNAfold from Vienna RNA package [38] must be below a user defined value) and their length should be in the range of 30 to 500 nt in order to be tagged as an sRNA. All these requirements are used per default but can be modified or removed via ANNOgesic's command line parameters. ANNOgesic stores the results of all analyses and generates GFF3 files, fasta files, secondary structural figures, dot plots, as well as mountain plots based on those predictions.

For sRNAs that share a transcript with CDSs – 5' UTR, inter-CDS, or 3' UTR located sRNAs – we implemented several detection heuristics (Figure 5B / C). 5' UTR-derived sRNAs must start with a TSS and show a sharp drop of coverage or a PS in the 3'-end. The requirement for the detection of inter-CDS located sRNAs is either a TSS or a PS as well as a coverage drop at the 3'-end or a PS. Small RNAs derived from the 3' UTR are expected to have a TSS or a PS and either end with the transcript or at a PS. After the detection of a *bona fide* sRNA, the

above described quality filters (length range, secondary structure etc.) are applied to judge the potential of a candidate (examples are shown in Figure S7, S8). For the validation of sRNA candidates in our test case, the described sRNAs of two publications were chosen. Sharma *et al.* [58] described 63 sRNAs of which 4 were not expressed in the condition of the test data set (removed from the dataset) (Figure S9). Of these 59 53 (90%) were detected by ANNOgesic. In the *C. jejuni* 81116 set, 31 sRNAs were described by Dugar *et al.* [16], and ANNOgesic could recover 26 (84%). The sRNA ranking system provided by ANNOgesic is displayed in Figure S10 and Equation S1.

In order to deduce potential regulatory functions of newly-predicted sRNAs, ANNOgesic performs prediction of interaction between them and mRNAs using RNAplex [38, 63], RNAup [38, 46], and IntaRNA [40]. The user can choose if only interactions supported by all tools are reported.

## 3.4 Detection of sORFs

All newly detected transcripts that do not contain a previously described CDS as well all 5' UTRs and 3' UTRs are scanned for potential sORFs [61] (Figure 6). For this, ANNOgesic searches for start and stop codons (non-canonical start codons are not included, but can be assigned by the user) that constitute potential ORFs of 30 to 150 base-pairs. Furthermore, ribosomal binding sites (based on the Shine-Dalgarno sequence, but different sequences can be assigned as well) between the TSS and 3 to 15 bp upstream of the start codon are required for a *bona fide* sORF.

## 3.5 Detection of functional related attributes

In order to facilitate a better understanding of the biological function of known and newly detected transcripts, ANNOgesic predicts several attributes for these features.

This includes the allocation of GO as well as GOslim [64] terms to CDSs via searching of protein ids in Uniprot [39]. The occurrence of groups is visualized for expressed and non-expressed CDSs (Figure S11). Furthermore the subcellular localization is predicted by PSORTb [70] for the proteins (Figure S12). Additionally, the protein entries are enriched by protein-protein interaction information retrieved from STRING [62] and PIE [32] (examples in Figure S13).

## 3.6 Circular RNAs

ANNOgesic integrates the tool "testrealign.x" from the segemehl package for the detection of circular RNAs [26] and adds a filter to reduce the number of false positive. Candidates for circular RNAs must be located in intergenic regions and exceed a given number of reads.

## 3.7 CRISPRs

CRISPR/Cas systems represent a bacterial defence system against phages and consist of repeat units and spacers sequences as well as Cas proteins [55]. The

CRISPR Recognition Tool (CRT) [5] is integrated into ANNOgesic and extended by comparison of CRISPR/Cas candidates to other annotations to remove false positive (Figure S14).

### 3.8 Riboswitches and RNA thermometers

Riboswitches and RNA thermometers are regulatory sequences that are part of transcripts and influence the translation based on the concentration of selected small molecules and temperature change, respectively. For the prediction of these riboswitches and RNA thermometers, ANNOgesic searches [48] the sequences which are between TSSs (or starting point of a transcript if no TSS was detected) and downstream CDSs, as well as associated with ribosome binding site in the Rfam database using Infernal [49].

### 3.9 Comparison between ANNOgesic predictions and published databases for *Escherichia coli* K12

In order to assess the performance of ANNOgesic, we compared its predictions based on a dRNA-Seq data set and conventional RNA-Seq of *Escherichia coli* K12 MG1655 by Thomason *et al.* [65] and McClure *et al.* [42] with the entries in several databases [20, 22, 24, 31, 41, 51]. Most of the benchmarking features can be precisely detected (80% or more) (Table S2). Moreover, the predicted features not found in published databases have high possibility to be novel features which are strongly supported by RNA-Seq data (Figure S7B and S7D). TSSs represent an exception with lower success rates and we assume this is mostly due to the higher sensitivity of the dRNA-Seq method in comparison to older protocols. To test this assumption and to investigate the quality of the TSS entries in RegulonDB, we compared the three deposited TSS data sets (Salgado *et al.* generated with Illumina RNA-Seq as well as Mendoza-Vargas *et al.* generated with Roche 454 high-throughput pyrosequencing and generated with Roche 5'RACE [43, 53]) to each other and found very small overlap (Figure S15). Additionally, the 50 nucleotides upstream of TSSs were extracted and scanned with MEME [2] for common motifs that are similar to the ones described for promoters. Only for a small number – 0 to 7% – of TSSs such motifs were found (Table S3) while 80% of the TSS predictions from ANNOgesic have such a promoter motif located upstream (Figure S6C). The same analysis could not be performed with EcoCyc [31] which is lacking TSS information and provide only positions but no strand information for promoters. Due to these results we doubt that the data in those data bases represents a solid ground for benchmarking the accuracy of ANNOgesic's TSS predictions.

## 4 Discussion

While RNA-Seq has become a powerful method to annotate genomes, the integration of the data is usually very laborious and time-consuming. It requires bioinformatic expertise and involves the application of different programs to perform the different

required steps. Here we presented ANNOgesic, a modular, user-friendly annotation tool for the analysis of bacterial RNA-Seq data that integrated several tools, optimizes their parameters, and includes novel prediction methods for several genomic features. With the help of this command-line tool, RNA-Seq data can be efficiently used to generate high-resolution annotations of bacterial genomes with very little manual effort. Besides the annotation files in standard formats, it also returns numerous statistics and visualizations that help the user to explore and to evaluate the results. While it ideally integates conventional RNA-Seq (beneficial for detecting 3'-ends of transcripts) as well as dRNA-Seq (required for the efficiently detection of internal TSSs) as input together (see Figure S16 and S17), it can also perform sufficient predictions with only one class of data for the majority of the genomic features (Table S2).

The performance of ANNOgesic has been here demonstrated by applying it on two published data sets and comparing the results to manually-conducted annotations. ANNOgesic could detect 90% and 83% of the manually-annotated sRNAs *H. pylori* 26695 and *C. jejuni* 81116, respectively. The sRNAs missed by ANNOgesic can be explained by low coverage, not being associated with TSSs, or lack of expression in the assayed conditions (see Figure S18 and S19).

Besides the analyses presented as examples in this study (*H. pylori* 26695 and *C. jejuni* 81116), ANNOgesic was meanwhile successfully applied for detecting transcripts, sRNAs, and TSSs in additional annotation projects (e.g. *Pseudomonas aeruginosa* [14] and *Rhodobacter sphaeroides* [52]. Despite the fact that the program was developed mainly with a focus on bacterial genomes, it has also been used to annotate archaeal genomes (namely *Methanosarcina mazei* (Lutz *et al.*, unpublished)) and eukaryotic genomes which have no introns (*Trypanosoma brucei* (Müller *et al.*, unpublished)).

ANNOgesic is freely available under an OSI compliant open source license (ISCL) and an extensive documentation has been generated in order to guide the novice and advanced users.

# 5 Conclusion

ANNOgesic is a powful tool for annotating numerous available genome annotations based on RNA-Seq data from multiple protocols. ANNOgesic not only integrated the available tools, but also improve their performance by optimizing parameters, and removing false positives. For the genomic features which can not be detected by currently available tools several novel methods were developed and implemented as part of ANNOgesic. A comprehensive documentation and useful statistics as well as visualization are also provided by ANNOgesic.

# 6 Availability of supporting source code and requirements

- Project name: ANNNOgesic

- Project home page: GitHub - https://github.com/Sung-Huan/ANNOgesic.
  PyPI - https://pypi.org/project/ANNOgesic/.
  DockerHub - https://hub.docker.com/r/silasysh/annogesic/.
- Operating system(s): Linux, Mac OS
- Programming language: Python
- Other requirements: Please check the documentation (http://annogesic.readthedocs.io/en/latest/required.html).
- License: ISC (Internet Systems Consortium license  simplified BSD license).

# 7 Availability of supporting data

Shell scripts that implement the data analysis workflows as well as the results of the analyses were deposited in Zenodo (https://doi.org/10.5281/zenodo.1161115).

# 8 Additional files

Figure S1: Workflow charts of ANNogesic modules.

Figure S2: Distribution of TSS classes.

Figure S3: Concept and examples for detecting coverage decrease of terminators.

Figure S4: Terminator prediction approach based on convergent genes.

Figure S5: Length distribution of UTRs.

Figure S6: The promoter motifs detected in *Helicobacter pylori* 26695, *Campylobacter jejuni* 81116, and *Escherichia coli* K12 MG1655.

Figure S7: Examples of known and novel intergenic sRNAs that ANNOgesic can detect.

Figure S8: Examples of detected antisense and UTR derived sRNAs.

Figure S9: The coverage plots of the benchmarking sRNA HPnc4620.

Figure S10: Histograms of ranking number of the sRNA benchmarking set.

Figure S11: The distributions of GO term.

Figure S12: The distributions of subcellular localizations.

Figure S13: Visualization of protein-protein interactions.

Figure S14: The example of CRISPR in *Campylobacter jejuni* 81116.

Figure S15: The overlap of three previously published TSS datasets in RegulonDB.

Figure S16: The predicted sRNA which can be detected only in data RNA-Seq after transcript fragmentation.

Figure S17: The comparison between dRNA-Seq and RNA-Seq after transcript fragmentation for detecting transcript

Figure S18: The lowly expressed sRNA - HPnc4610.

Figure S19: An example of known sRNA – CJnc230 – which is not associated with a TSS.

Equation S1: The ranking system of sRNA prediction.

Table S1: The novelties and improvements of genomic feature detection in ANNOgesic.

Table S2: The comparison between ANNOgesic predictions and several databases.

Table S3: The number of TSSs and their associated promoter motifs in RegulonDB.

Table S4: The manual-curated TSS set of *Helicobacter pylori* 26695 (1-400bp).

Table S5: The manual-curated TSS set of *Campylobacter jejuni* 81116 (1-400bp).

Table S6: The manual-curated PS set of *Helicobacter pylori* 26695 (1-400bp).

Table S7: The manual-curated PS set of *Campylobacter jejuni* 81116 (1-400bp).

# 9 Abbreviations

- TSS: transcriptional start site
- PS: processing site
- sRNA: small RNA
- sORF: small open reading frame
- dRNA-Seq: differential RNA Sequencing
- kb: kilo base
- CDS: coding sequence
- CRISPR: clustered regularly interspaced short palindromic repeat
- bp: base pair
- nt: nucleotide
- GO: Gene Ontology

# 10 Competing interests

The authors declare that they have no competing interests.

# 11 Funding

The project was funded by the German Research Foundation (DFG) as part of the Transregio 34 (CRC-TRR34).

## 12 Acknowledgements

We would like to thank several colleagues for fruitful discussion, especially Sarah Svensson, Lars Barquist and Thorsten Bischler for comments regarding the manuscript, Diarmaid Tobin and Till Sauerwein for giving feedback regarding code and documentation.

## REFERENCES

[1] F Amman, M T Wolfinger, R Lorenz, I L Hofacker, P F Stadler, and S Findeiß. *TSSAR*: *TSS* annotation regime for *dRNA*-seq data. *BMC bioinformatics*, 15:89, 2014.

[2] T L Bailey, N Williams, C Misleh, and W W Li. *MEME*: discovering and analyzing *DNA* and protein sequence motifs. *Nucleic Acids Res.*, 34:W369–373, 2006.

[3] A Beauregard, E Smith, B Petrone, N Singh, C Karch, K McDonoughm, and J T Wade. Identification and characterization of small rnas in *Yersinia pestis*. *RNA Biol*, 10:397–405, 2013.

[4] T Bischler, H S Tan, K Nieselt, and C M Sharma. Differential *RNA*-seq (*dRNA*-seq) for annotation of transcriptional start sites and small *RNAs* in *Helicobacter pylori*. *Methods*, 86:89–101, 2015.

[5] C Bland, T L Ramsey, F Sabree, M Lowe, K Brown, N C Kyrpides, and P Hugenholtz. Crispr recognition tool (crt): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, 8:209, 2007.

[6] C Bohn, C Rigoulay, S Chabelskaya, C M Sharma, A Marchais, P Skorski, E Borezée-Durant, R Barbet, E Jacquet, A Jacq, D Gautheret, B Felden, J Vogel, and P Bouloc. Experimental discovery of small *RNAs* in *Staphylococcus* aureus reveals a riboregulator of central metabolism. *Nucleic Acids Res.*, 38:6620–6636, 2010.

[7] M Bouvier, C M Sharma, F Mika, K H Nierhaus, and J Vogel. Small rna binding to 5' mrna coding region inhibits translational initiation. *Mol. Cell*, 32:827–837, 2008.

[8] C Camacho, G Coulouris, V Avagyan, N Ma, J Papadopoulos, K Bealer, and T L Madden. *BLAST*+: architecture and applications. *BMC bioinformatics*, 10:421, 2009.

[9] Y Chao, K Papenfort, R Reinhardt, C M Sharma, and J Vogel. An atlas of *Hfq*-bound transcripts reveals 3' *UTRs* as a genomic reservoir of regulatory small *RNAs*. *EMBO j.*, 31:4005–4019, 2012.

[10] I Chepelev, G Wei, Q Tang, and K Zhao. Detection of single nucleotide variations in expressed exons of the human genome using rna-seq. *Nucleic Acids Res.*, 37:e106–e106, 2009.

[11] E T Cirulli, A Singh, K V Shianna, D Ge, J P Smith, J M Maia, E L Heinzen, J J Goedert, D B Goldstein, and the Center for HIV/AIDS Vaccine Immunology (CHAVI). Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biol.*, 11:R57, 2010.

[12] D Dar, M Shamir, J R Mellin, M Koutero, N Stern-Ginossar, P Cossart, and R Sorek. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science*, 352:aad9822–aad9822, 2016.

[13] A L Delcher, K A Bratke, E C Powers, and S L Salzberg. Identifying bacterial genes and endosymbiont dna with *Glimmer*. *Bioinformatics*, 23:673–679, 2007.

[14] J Dingemans, P Monsieurs, S H Yu, A Crabbé, K U Förstner, A Malfroot, P Cornelis, and R Van Houdt. Effect of shear stress on pseudomonas aeruginosa isolated from the cystic fibrosis lung. *mBio*, 7:e00813–16, 2016.

[15] A Dobin, C A Davis, F Schlesinger, J Drenkow, C Zaleski, S Jha, P Batut, M Chaisson, and T R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29:15–21, 2013.

[16] G Dugar, A Herbig, K U Förstner, N Heidrich, R Reinhardt, K Nieselt, and C M Sharma. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter* jejuni isolates. *PLoS Genet.*, 9:e1003495, 2013.

[17] S C Forster, A M Finkel, J A Gould, and P J Hertzog. *RNA-eXpress* annotates novel transcript features in *RNA*-seq data. *Bioinformatics*, 29:810–812, 2013.

[18] K U Förstner, J Vogel, and C M Sharma. *READemption*-a tool for the computational analysis of deep-sequencing-based transcriptome data. *Bioinformatics*, 30:3421–3423, 2014.

[19] M C Frith, N F W Saunders, B Kobe, and T L Bailey. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol.*, 4:e1000071, 2008.

[20] S Gama-Castro, H Salgado, A Santos-Zavaleta, D Ledezma-Tejeida, L Muñiz-Rascado, J S García-Sotelo, K Alquicira-Hernández, I Martínez-Flores, L Pannier, J A Castro-Mondragón, A Medina-Rivera, H Solano-Lira, C Bonavides-Martínez, E Pérez-Rueda, S Alquicira-Hernández, L Porrón-Sotelo, A López-Fuentes, A Hernández-Koutoucheva, V del Moral-Chávez, F Rinaldi, and J Collado-Vides. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, 44:D133–143, 2016.

[21] D E Goldberg. Genetic algorithms in search, optimization, and machine learning. *Addison-Wesley Pub. Co, Reading, Mass*, 1989.

[22] I Grissa, G Vergnaud, and C Pourcel. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*, 8(1):172, 2007.

[23] J Harrow, A Frankish, J M Gonzalez, E Tapanari, M Diekhans, F Kokocinski, B L Aken, D Barrell, A Zadissa, S Searle, I Barnes, A Bignell, V Boychenko, T Hunt, M Kay, G Mukherjee, J Rajan, G Despacio-Reyes, G Saunders, C Steward, R Harte, M Lin, C Howald, A Tanzer, T Derrien, J Chrast, N Walters, S Balasubramanian, B Pei, M Tress, J M Rodriguez, I Ezkurdia, J van Baren, M Brent, D Haussler, M Kellis, A Valencia, A Reymond, M Gerstein, R Guigó, and T J Hubbard. Gencode: the reference human genome annotation for the encode project. *Genome Research*, 22:1760–1774, 2012.

[24] M R Hemm, B J Paul, T D Schneider, G Storz, and K E Rudd. Small membrane proteins found by comparative genomics and ribosome binding site models. *Molecular Microbiology*, 70(6):1487–1501, 2008.

[25] A Hochschild. Gene-specific regulation by a transcript cleavage factor: Facilitating promoter escape. *J. Bacteriol.*, 189:8769–8771, 2007.

[26] S Hoffmann, C Otto, G Doose, A Tanzer, D Langenberger, S Christ, M Kunz, L M Holdt, D Teupser, J Hackermüller, and P F Stadler. A multi-split mapping algorithm for circular rna, splicing, trans-splicing and fusion detection. *Genome Biol.*, 15:R34, 2014.

[27] S Hoffmann, C Otto, S Kurtz, C M Sharma, P Khaitovich, J Vogel, P F Stadler, and J Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.*, 5:e1000502, 2009.

[28] E Holmqvist, P R Wright, L Li, T Bischler, L Barquist, R Reinhardt, R Backofen, and J Vogel. Global rna recognition patterns of post-transcriptional regulators hfq and csra revealed by uv crosslinking *in vivo*. *EMBO J.*, 35:991–1011, 2016.

[29] N T Ingolia. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.*, 15:205–213, 2014.

[30] H Jorjani and M Zavolan. *TSSer*: an automated method to identify transcription start sites in prokaryotic genomes from differential *RNA* sequencing data. *Bioinformatics*, 30:971–974, 2014.

[31] I M Keseler, J Collado-Vides, A Santos-Zavaleta, M Peralta-Gil, S Gama-Castro, L Muniz-Rascado, C Bonavides-Martinez, S Paley, M Krummenacker, T Altman, P Kaipa, A Spaulding, J Pacheco, M Latendresse, C Fulcher, M Sarker, A G Shearer, A Mackie, I Paulsen, R P Gunsalus, and P D Karp. EcoCyc: a comprehensive database of escherichia coli biology. *Nucleic Acids Research*, 39:D583–D590, 2011.

[32] S Kim, S Y Shin, I H Lee, S J Kim, R Sriram, and B T Zhang. *PIE*: an online prediction system for protein-protein interactions from text. *Nucleic Acids Res.*, 36:W411–415, 2008.

[33] C L Kingsford, K Ayanbule, and S L Salzberg. Rapid, accurate, computational discovery of *Rho*-independent transcription terminators illuminates their relationship to *DNA* uptake. *Genome Biol.*, 8:R22, 2007.

[34] S Kurtz, A Phillippy, A L Delcher, M Smoot, M Shumway, C Antonescu, and S L Salzberg. Versatile and open software for comparing large genomes. *Genome Biol.*, 5:R12, 2004.

[35] K Lagesen, P Hallin, E A Rodland, H H Staerfeldt, T Rognes, and D W Ussery. *RNAmmer*: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, 35:3100–3108, 2007.

[36] H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment map format and samtools. *Bioinformatics*, 25:2078–2079, 2009.

[37] L Li, D Huang, M K Cheung, W Nong, Q Huang, and H S Kwan. *BSRD*: a repository for bacterial small regulatory *RNA*. *Nucleic Acids Res.*, 41:D233–238, 2013.

[38] R Lorenz, S H Bernhart, C Höner Zu Siederdissen, H Tafer, C Flamm, P F Stadler, and I L Hofacker. *ViennaRNA Package* 2.0. *Algorithm. Mol. Biol.*, 6:26, 2011.

[39] M Magrane and Uniprot Consortium. *UniProt Knowledgebase*: a hub of integrated protein data. *Database*, page bar009, 2011.

[40] M Mann, P R Wright, and R Backofen. IntaRNA 2.0: enhanced and customizable prediction of RNAâĂŞRNA interactions. *Nucleic Acids Research*, 45:W435–W439, 2017.

[41] X Mao, Q Ma, C Zhou, X Chen, H Zhang, J Yang, F Mao, W Lai, and Y Xu. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Research*, 42:D654–659, 2014.

[42] R McClure, D Balasubramanian, Y Sun, M Bobrovskyy, P Sumby, C A Genco, C K Vanderpool, and B Tjaden. Computational analysis of bacterial rna-seq data. *Nucleic Acids Res.*, 41:e140, 2013.

[43] A Mendoza-Vargas, L Olvera, M Olvera, R Grande, L Vega-Alvarado, B Taboada, V Jimenez-Jacinto, H Salgado, K Juárez, B Contreras-Moreira, A M Huerta, J Collado-Vides, and E Morett. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in e. coli. *PLoS ONE*, 4(10):e7526, 2009-10-19.

[44] D Merkel. Docker: Lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014.

[45] M Miyakoshi, Y Chao, and J Vogel. Regulatory small rnas from the 3'regions of bacterial mrnas. *Curr. Opin. Microbiol.*, 24:132–139, 2015.

[46] U Mückstein, H Tafer, J Hackermüller, S H Bernhart, P F Stadler, and I L Hofacker. Thermodynamics of *RNA-RNA* binding. *Bioinformatics*, 22:1177–1182, 2006.

[47] K O Mutz, A Heilkenbrinker, M Lönne, J G Walter, and F Stahl. Transcriptome analysis using next-generation sequencing. *Curr. Opin. Biotech.*, 24:22–30, 2013.

[48] E P Nawrocki, S W Burge, A Bateman, J Daub, R Y Eberhardt, S R Eddy, E W Floden, P P Gardner, T A Jones, J Tate, and R D Finn. Rfam 12.0: updates to the *RNA* families database. *Nucleic Acids Res.*, 43:D130–137, 2014.

[49] E P Nawrocki and S R Eddy. Infernal 1.1: 100-fold faster *RNA* homology searches. *Bioinformatics*, 29:2933–2935, 2013.

[50] T D Otto, G P Dillon, W S Degrave, and M Berriman. Ratt: Rapid annotation transfer tool. *Nucleic Acids Res.*, 39:e57, 2011.

[51] K D Pruitt, T Tatusova, W Klimke, and D R Maglott. NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Research*, 37:D32–36, 2009.

[52] B Remes, T Rische-Grahl, T Müller, K Förstner, S H Yu, W Lennart, A Jäger, V Peuser, and G Klug. An rpohi-dependent response promotes outgrowth after extended stationary phase in the alphaproteobacterium *Rhodobacter sphaeroides*. *J. Bacteriol.*, 2017. in press.

[53] H Salgado, M Peralta-Gil, S Gama-Castro, A Santos-Zavaleta, L Muñiz-Rascado, J S García-Sotelo, V Weiss, H Solano-Lira, I Martínez-Flores, A Medina-Rivera, G Salgado-Osorio, S Alquicira-Hernández, K Alquicira-Hernández, A López-Fuentes, L Porrón-Sotelo, A M Huerta, C Bonavides-Martínez, Y I Balderas-Martínez, L Pannier, M Olvera, A Labastida, V Jiménez-Jacinto, L Vega-Alvarado, V del Moral-Chávez, A Hernández-Alvarez, E Morett, and J Collado-Vides. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41:D203–D213, 2013.

[54] E Sallet, J Gouzy, and T Schiex. *EuGene-PP*: a next-generation automated annotation pipeline for prokaryotic genomes. *Bioinformatics*, 30:2659–2661, 2014.

[55] J D Sander and J K Joung. Crispr-cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.*, 32:347–355, 2014.

[56] P Schattner, A N Brooks, and T M Lowe. The *tRNAscan-SE*, snoscan and snogps web servers for the detection of trnas and snornas. *Nucleic Acids Res.*, 33:W686–W689, 2005.

[57] T Seemann. *Prokka*: rapid prokaryotic genome annotation. *Bioinformatics*, 30:2068–2069, 2014.

[58] C M Sharma, S Hoffmann, F Darfeuille, J Reignier, S Findeiss, A Sittka, S Chabas, K Reiche, J Hackermüller, R Reinhardt, P F Stadler, and J Vogel. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, 464:250–255, 2010.

[59] C M Sharma and J Vogel. Differential *RNA*-seq: the approach behind and the biological insight gained. *Curr. Opin. in Microbiol.*, 19:97–105, 2014.

[60] A Smirnov, K U Förstner, E Holmqvist, A Otto, R Günster, D Becher, R Reinhardt, and J Vogel. Grad-seq guides the discovery of proq as a major small rna-binding protein. *P. Natl. Acad. Sci. USA*, 113:11591–11596, 2016.

[61] G Storz, Y I Wolf, and K S Ramamurthi. Small proteins can no longer be ignored. *Annu. Rev. Biochem.*, 83:753–777, 2014.

[62] D Szklarczyk, A Franceschini, S Wyder, K Forslund, D Heller, J Huerta-Cepas, M Simonovic, A Roth, A Santos, K P Tsafou, M Kuhn, P Bork, L J Jensen, and C von Mering. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43:D447–452, 2015.

[63] H Tafer and I L Hofacker. *RNAplex*: a fast tool for *RNA-RNA* interaction search. *Bioinformatics*, 24:2657–2663, 2008.

[64] The Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Res.*, 43:D1049–D1056, 2015.

[65] M K Thomason, T Bischler, S K Eisenbart, K U Förstner, A Zhang, A Herbig, K Nieselt, C M Sharma, and G Storz. Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in escherichia coli. *Journal of Bacteriology*, 197(1):18–28, 2015.

[66] J Wang, W Rennie, C Liu, C S Carmack, K Prévost, M Caron, E Massé, Y Ding, and J T Wade. Identification of bacterial srna regulatory targets using ribosome profiling. *Nucleic Acids Res.*, 43:10308–10320, 2015.

[67] L S Waters and G Storz. Regulatory RNAs in bacteria. *Cell*, 136:615–628, 2009.

[68] T Weinmaier, A Platzer, J Frank, H J Hellinger, P Tischler, and T Rattei. *ConsPred*: a rule-based (re-)annotation framework for prokaryotic genomes. *Bioinformatics*, 32:3327–3329, 2016.

[69] O Wurtzel, R Sapra, F Chen, Y Zhu, B A Simmons, and R Sorek. A single-base resolution map of an archaeal transcriptome. *Genome Research*, 20:133–141, 2010.

[70] N Y Yu, J R Wagner, M R Laird, G Melli, S Rey, R Lo, P Dao, S C Sahinalp, M Ester, L J Foster, and F S L Brinkman. *PSORTb* 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26:1608–1615, 2010.

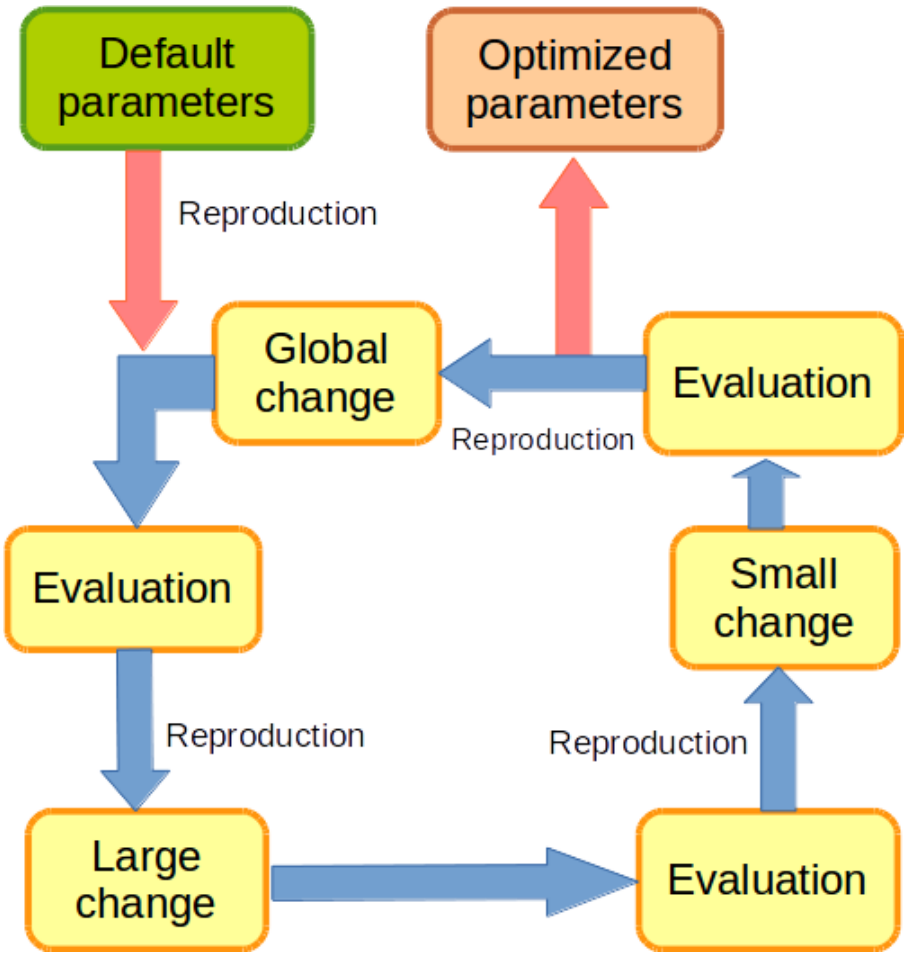


Figure 1: Schema of the genetic algorithm for optimizing the parameters of TSSpredator. It starts from the default parameters. These parameter sets will go through three steps - global change (change every parameter randomly), large change (change two of the parameters randomly), and then small change (adds/subtracts a small fraction to one of the parameters). It will then select the best parameter set for reproduction when one step is done. Usually, ANNOgesic can achieve the optimized parameters within 4000 runs.

$$TPR_c - TPR_b \geq 0.1 \quad (1)$$
$$(TPR_c > TPR_b) \wedge (FPR_c < FPR_b) \quad (2)$$
$$(TP_b - TP_c > 0) \wedge (FP_b - FP_c \geq 5 \times (TP_b - TP_c)) \quad (3)$$
$$(TP_b - TP_c < 0) \wedge (FP_c - FP_b \leq 5 \times (TP_c - TP_b)) \quad (4)$$
$$(TP_m \geq 100) \wedge (TPR_c - TPR_b \geq 0.01) \wedge (FPR_c - FPR_b \leq 5 \times 10^{-5}) \quad (5)$$
$$(TP_m \geq 100) \wedge (TPR_b - TPR_c \leq 0.01) \wedge (FPR_b - FPR_c \geq 5 \times 10^{-5}) \quad (6)$$

Equation 1: $TP_m$ is the number of manually-detected TSSs. $TP_c/TPR_c$ represents the true positive/true positive rate of the current parameters. $TP_b/TPR_b$ represents the true positive/true positive rate of the best parameters. $FP_c/FPR_c$ represents the false positive/false positive rate of the current parameters. $FP_b/FPR_b$ represents the false positive/false positive rate of the best parameters. If one of these six situations is true, it will replace the best parameters with current parameters.

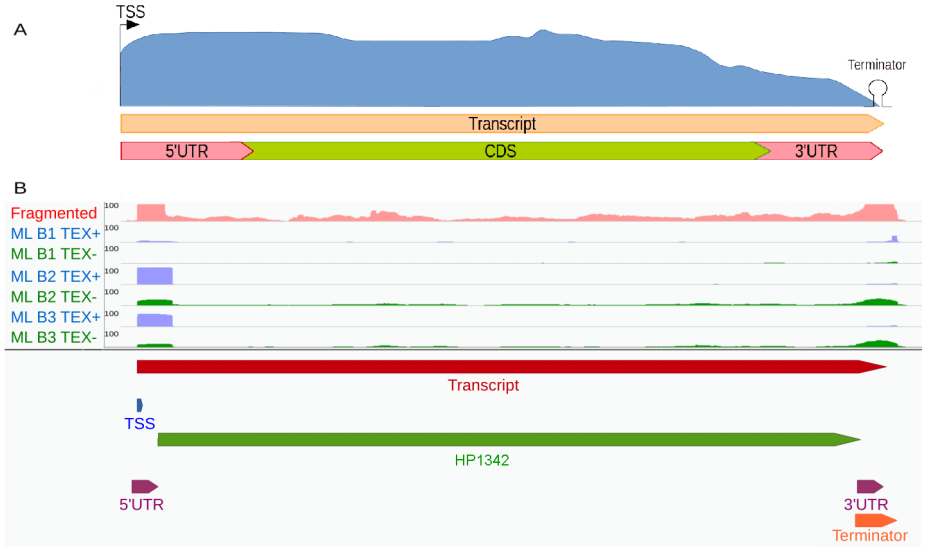


Figure 2: Transcript boundary detection. **(A)** Schema: ANNOgesic can predict TSSs, terminators, transcripts, genes and UTRs and integrate them into a comprehensive annotations. **(B)** Gene HP1342 of *H. pylori* 26695 as an example. The pink coverage plot represents RNA-Seq data of libraries after fragmentation, the blue coverage plots TEX+ libraries of dRNA-Seq, the green coverage plots TEX- libraries of dRNA-Seq. Transcript, TSS, terminator, and CDS are presented as red, blue, orange, and green bars, respectively. The figure shows how the transcript covers the whole gene location and how UTRs (presented by purple bars) can be detected based on the TSS, transcript, terminator, and gene annotations.

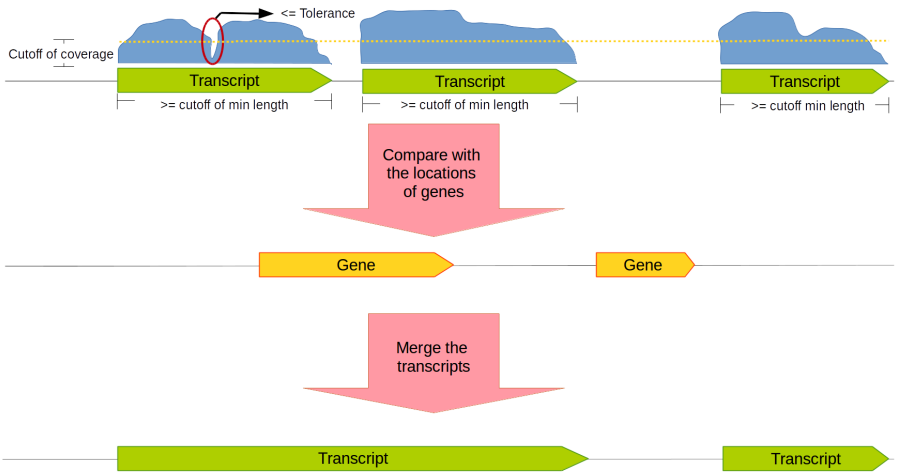


Figure 3: Coverage-based transcript detection. If the coverage (blue curve-blocks) is higher than a given coverage cut-off value (dashed line) a transcript will be called. The user can set a tolerance value (i.e. a number of nucleotides with a coverage below the cut-off) on which basis gapped transcripts are merged or are kept separated. Information of gene positions can also be used to merge transcripts in case two of them overlap with the same gene.

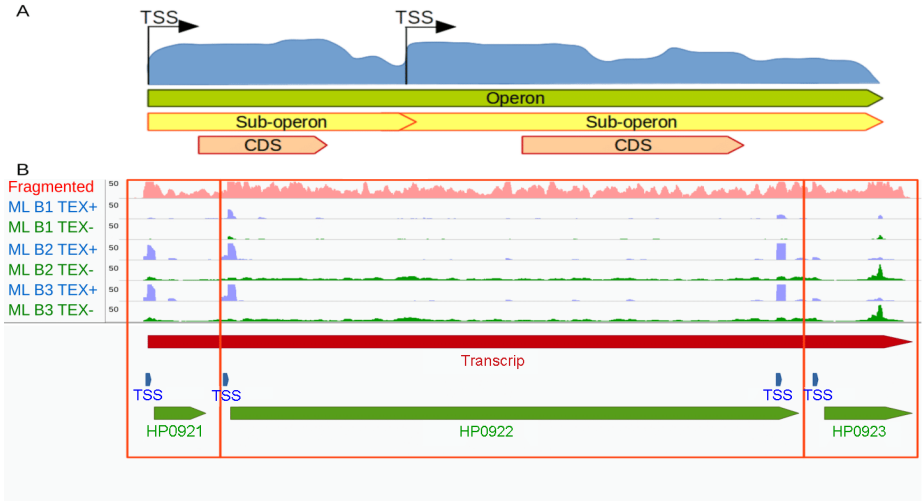


Figure 4: Operon and sub-operon detection. **(A)** If there are more than one TSSs which does not overlap with genes located within one operon, the operon can be divided to several sub-operons based on these TSSs. **(B)** An example from *H. pylori* 26695. The coverage of RNA-Seq with fragmentation, TEX+ and TEX- of dRNA-Seq are shown in pink, blue and green coverages, respectively. TSSs, transcripts/operons and genes are presented as blue, red and green bars, respectively. The two genes are located in the same operon, but also in different sub-operons (two empty red squares).

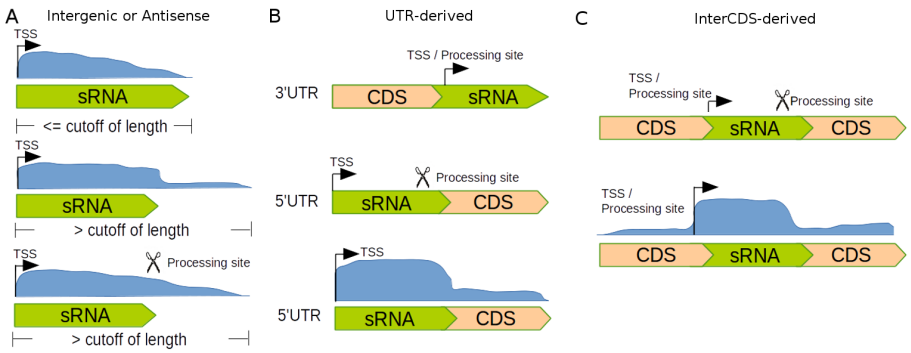


Figure 5: Detection of intergenic, antisense and UTR-derived sRNAs. Detection of intergenic, antisense and UTR-derived sRNAs. The length of potential sRNAs should be within a given range and their coverages should exceed a given cut-off coverage. **(A)** Detection of intergenic and antisense sRNAs. There are three cases shown: In the upper panel the transcript starts with a TSS, and length of the transcript is within the expected length. In the one in the middle the transcript starts with a TSS but the transcript is longer than an average sRNA; in that case ANNOgesic will search the coverage (blue region) for a point at which the coverage is decreasing rapidly. The image at the bottom is similar to the one in the middle, but the sRNA ends with a PS. **(B)** Detection of UTR-derived sRNAs. 3' UTR-derived sRNAs: if the transcript starts with a TSS or PS, it will be tagged as a 3' UTR-derived sRNA. For 5' UTR-derived sRNAs: if the transcript starts with a TSS and ends with a PS or the point where the coverage significant drops. **(C)** Detection of interCDS-derived sRNAs: Similar to the 5' UTR-derived approach but the transcript starts with a PS.

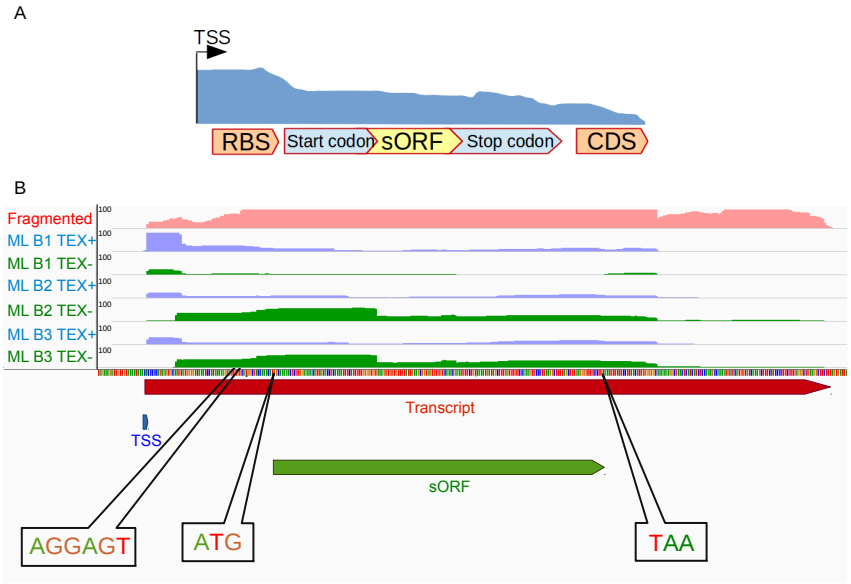


Figure 6: sORF detection. **(A)** An sORF must contain start codon and stop codon within transcript and should be inside of a given length range (default 30 - 150 nt). Additionally, a ribosomal binding site must be detected between the TSS and the start codon. **(B)** An example from *H. pylori* 26695. The coverage of RNA-Seq (fragmented libraries), TEX+ and TEX- (dRNA-Seq) are shown as pink, blue and green coverages, respectively. The TSS, transcript and sORF are presented as blue, red and green bars, respectively.