ANNOgesic: A pipeline to translate bacterial/archaeal RNA-Seq data into high-resolution genome annotations

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

May 27, 2017

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 pipeline called ANNOgesic that provides the required analyses and simplifies RNA-Seq-based bacterial and archaeal genome annotation. It 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://pythonhosted.org/ANNOgesic/.

1 INTRODUCTION

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) [1], tRNAscan-SE [2] to spot tRNAs and

*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

RNAmmer to find rRNAs [3]. Pipelines like Prokka [5] or ConsPred [6] combine 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 [7]. Besides this it has also been applied in numerous cases to improve the genome annotation of bacteria [8–10], archaea [11] and eukaryotes [12]. RNA sequencing based protocols like differential RNA-Seq (dRNA-Seq) [13, 14], Term-seq [15] and ribosome profiling [16, 17] 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 [18–20] or based on conventional RNA-Seq data [21, 22], 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 different types of RNA-Seq data like dRNA-Seq as well as RNA-Seq generated after transcript fragmentation (or conventional RNA-Seq) and generate high-quality genome annotations. It can detect several genomic features including genes, CDSs (coding DNA sequence), tRNAs, rRNAs, TSSs, and processing sites (PSs), transcripts, terminators, untranslated regions (UTRs) as well as sRNAs, small open reading frames (sORFs), circular RNAs, CRISPR-related RNAs, riboswitches, and RNA-thermometers. It can also perform RNA-RNA and protein-protein interaction predictions on detected features. Furthermore, it groups genes into operons as well as sub-operons and can generate promoter motifs that are found in front of transcriptional start sites. It can also allocate GO (Gene Ontology) terms and subcellular localizations to genes. Several of ANNOgesic's data processing steps are new implementations, while others are performed by third-party tools after dynamic parameter-optimizations through ANNOgesic itself. Numerous visualizations and statistics help the user to quickly evaluate the feature predictions. The pipeline 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 Implementation and installation

ANNOgesic is implemented in Python 3 and requires the third-party libraries Biopython [23], numpy [24], matplotlib [25], as well as networkx [26]. Its source code and comprehensive documentation are hosted at https://pythonhosted.org/ANNOgesic/ 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 guarantee a frictionless installation including non-Python dependencies, we additionally offer a docker image (https://hub.docker.com/r/silasysh/annogesic/) [27].

2.2 Modules and input data of ANNOgesic

ANNOgesic consists of the following twenty modules - their names indicate their functions: Sequence modification, Annotation transfer, SNP/Mutation, Transcript, TSS, Terminator, UTR, Processing site, 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 Supplementary Figure 1.

Depending on the task to one wish to perform, ANNOgesic requires a specific set of input information - either as in coverage information in wiggle, or alignments in BAM format. This can be generated by a mapper like BWA [28], STAR [29], segemehl [30], or a full RNA-Seq pipeline like READemption [31]. 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.

2.3 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 [18] – 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, [32] 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).

$$TPR_c - TPR_b \geq 0.1 \tag{1}$$

$$(TPR_c > TPR_b) \wedge (FPR_c < FPR_b) \tag{2}$$

$$(TP_b - TP_c > 0) \wedge (FP_b - FP_c \geq 5 \times (TP_b - TP_c)) \tag{3}$$

$$(TP_b - TP_c < 0) \wedge (FP_c - FP_b \leq 5 \times (TP_c - TP_b)) \tag{4}$$

$$(TP_m \geq 100) \wedge (TPR_c - TPR_b \geq 0.01) \wedge (FPR_c - FPR_b \leq 5 \times 10^{-5}) \tag{5}$$

$$(TP_m \geq 100) \wedge (TPR_b - TPR_c \leq 0.01) \wedge (FPR_b - FPR_c \geq 5 \times 10^{-5}) \tag{6}$$

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.

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 [8, 14] and *Campylobacter jejuni* 81116 [18]. The dRNA-Seq data sets were retrieved from NCBI GEO where they are stored under the accession numbers GSE67564 and GSE38883, respectively. For *Helicobacter* 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).

3 RESULTS

3.1 Correction of genome sequences and annotations

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 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 [33, 34]. 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 [35] and BCFtools [35] applying read counting-based filtering.

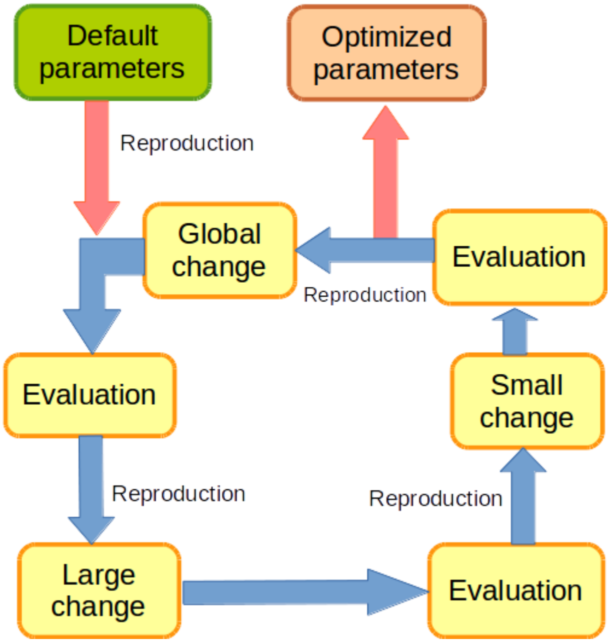

Figure 1: The genetic algorithm that ANNOgesic uses for optimizing the parameters of TSSpredator. It starts from the default parameters. These parameter sets will go through three steps - global change (change every parameter randomly), large change (change two of the parameters randomly), and then small 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.

3.1.2 Annotation transfer.

ANNOgesic integrates RATT [36], 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 [37]. For the chosen strains, *H. pylori* 26695 and *C. jejuni* 81116 annotation files in GFF3 format could 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 transcript boundaries

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) [38, 39] or even sources of sRNAs [40]. Unfortunately, most bacterial annotations only cover the protein coding regions while the information about TSSs, terminators and UTRs is

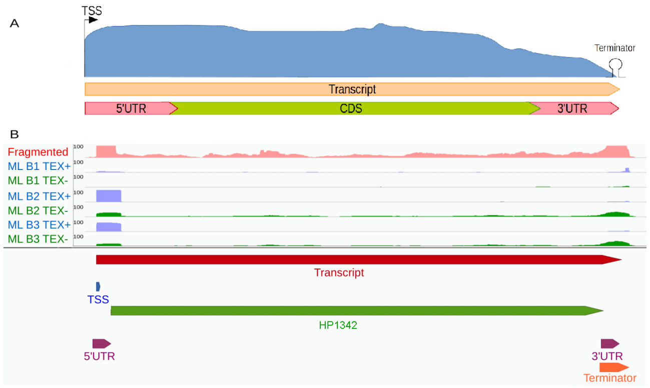

Figure 2: Transcript boundary detection. **(A)** ANNOgesic can integrate TSSs, terminators, transcripts, genes and UTRs, for defining transcript boundary if those features were predicted before. **(B)** An example from *H. pylori* 26695. The pink coverage represents RNA-Seq data of libraries after fragmention, the blue coverages TEX+ libraries of dRNA-Seq, the green coverages TEX- libraries of dRNA-Seq. Transcript, TSS, terminator, and CDS are presented as red, blue, orange, and green bars, respectively. This figure shows that the transcript covers the whole gene location, and that UTRs (presented by purple bars) can be detected based on the TSS, transcript, terminator, and gene annotations.

lacking. To address this issue, ANNOgesic combines several feature predictions for the reliable detection of transcript boundaries (Figure 2).

3.2.1 Transcript detection.

The primary step for the detection of transcript boundaries is transcript detection. For this purpose numerous tools are available (e.g. [41]), but most of them are optimized for the assembly of eukaryotic transcripts. Due to this, we combined several heuristics to perform such predictions based on the nucleotide coverage data, given gene annotations and several parameters that can be set by the user (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 Optimization of TSS prediction parameters.

For the prediction of TSSs, ANNOgesic builds on TSSpredator [18], 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

Table 1: The features of annotation in *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	935, (540)	987, (471)
	TransTermHP	614, (310)	631, (265)
	Convergent gene	397, (289)	464, (274)
UTR	5’ UTR	693	560
	3’ UTR	325	286
sRNA	Total	183	40
	Intergenic	60	16
	Antisense	84	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		119	14
Riboswitch		11	14
RNA thermometer		4	8
circular RNA		0	1
CRISPR		0	1, (8)
SNP / mutation		55	89

The numbers in brackets for Terminator and CRISPR mean the amount of terminators with coverage drop and repeat units of CRISPR, respectively. For Terminators, if a CDS associated with multiple terminators, ANNOgesic will only keep the high confidence one.

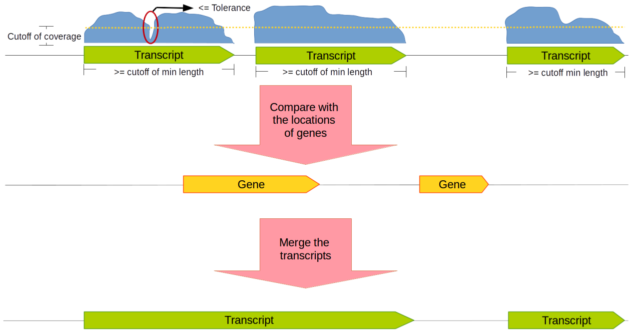

Figure 3: Transcript detection. If the coverage (blue curve-blocks) is higher than a given coverage cut-off value (dash line) a transcript will be called. The user can set a tolerance value (i.e. a number of nucleotides with a coverage below the cut-off) on which basis gapped transcripts are merged or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

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 (Supplementary Table 2 and 3). 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 or first 400 kb in the genome of *H. pylori* 26695 and *C. jejuni* 81116 (Supplementary Table 2 and 3), respectively. 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%) at 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 (Supplementary Figure 2).

3.2.3 Processing sites.

Several transcripts undergo processing, which influences their biological activity [40, 42]. In order to detect processing sites 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

Table 2: Comparison of default and optimized parameters of TSSpredator for TSS and processing site 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)
		processing site	
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 percentages in the table are the sensitivity or specificity. The numbers in brackets are true positive or false positive.

TEX). As done for the TSSs, we manually annotated the processing sites in first 200 kb of the genomes and found in *H. pylori* 26695, 281 and in *C. jejuni* 81116, 345 processing sites, respectively. Based on these, we performed parameter optimization on the test set (manually-curated from first 200 kb to 400 kb, Supplementary Table 4 and 5, Table 2) and could improve the prediction of processing sites via TSSpredator [18].

3.2.4 ρ-independent terminators.

While the transcriptional start sites are in general clearly defined boarders, 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 [43], which detects ρ-independent terminators based on genome sequences. Manual inspection showed us that TransTermHP predictions are not always supported by the RNA-data (Supplementary Figure 3e and f). This could be due to the lack of expression in the chosen conditions. Additionally, certain locations in 3'-ends that may be ρ-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 (see Supplementary Figure 3). In order to improve the sensitivity, an additional heuristic for the detection of ρ-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 (Supplementary Figure 4). 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 ρ-independent terminator is assumed. As default, the region should consist of 80 or less nucleotides, the A/T-rich region should be longer than 3 nucleotides, 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 Supplementary Figure 5).

3.2.6 Promoters.

ANNOgesic integrates MEME [44] (which detects ungapped motifs) and GLAM2 [45] (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 Supplementary Figure 6).

3.2.7 Operon.

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, there are also examples of 3' UTR-derived sRNAs [40, 46–48]. ANNOgesic offers the detection of both classes, combined with a detailed characterization of the sRNA candidates.

In order to classify newly detected intergenic transcripts as sRNAs, ANNOgesic tests several of their features (Figure 5A). If a BLAST+ [49] search of a transcript finds homologous sequences in BSRD [50] – 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 [51] must be below 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

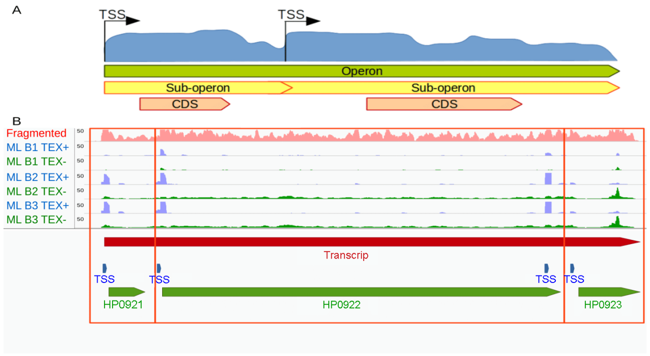

Figure 4: Operon and sub-operon detection. **(A)** If there are more than one TSSs which does not overlap with genes located within one operon, the operon can be divided to several sub-operons based on these TSSs. **(B)** An example from *H. pylori* 26695. The coverage of RNA-Seq with fragmentation, 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).

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 Supplementary Figure 7, 8). For the validation of sRNA candidates in our test case, the described sRNAs of two publications were chosen. Sharma *et al.* [8] described 63 sRNAs of which 4 were not expressed in the condition of the test data set (removed from the dataset) (Supplementary Figure 9). Of these 59, 53 were detected by ANNOgesic. In the *C. jejuni* 81116 set, 31 sRNAs were found by Dugar *et al.* [18], and ANNOgesic could recover 26 (84%) (Supplementary Figure 10).

In order to deduce potential regulatory functions of newly-predicted sRNAs, ANNOgesic performs prediction of interaction between them and mRNAs using RNAplex [51, 52] and RNAup [51, 53]. The user can chooose if only interactions supported by both tools are reported.

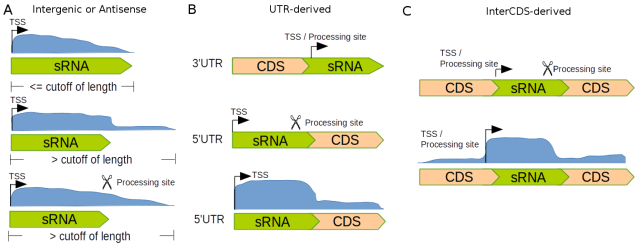

Figure 5: Detection of intergenic and UTR-derived sRNAs. The length of potential sRNAs should be within a given range and their coverages should exceed a given cut-off coverage. **(A)** Detection of intergenic and antisense sRNAs. There are three cases shown: In the upper panel the transcript starts 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 processing site. **(B)** Detection of UTR-derived sRNAs. 3' UTR-derived sRNAs: if the transcript starts with a TSS or processing site, 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 processing site 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 processing site.

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 [54] (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 [55] terms to CDSs via searching of protein ids in Uniprot [56]. The occurrence of groups is visualized for expressed and non-expressed CDSs (Supplementary Figure 11). Furthermore the subcellular localization is predicted by PSORTb [57] for the proteins (Supplementary Figure 12). Additionally, the protein entries are enriched by protein-protein interaction information retrieved from STRING [58] and PIE [59] (examples in Supplementary Figure 13).

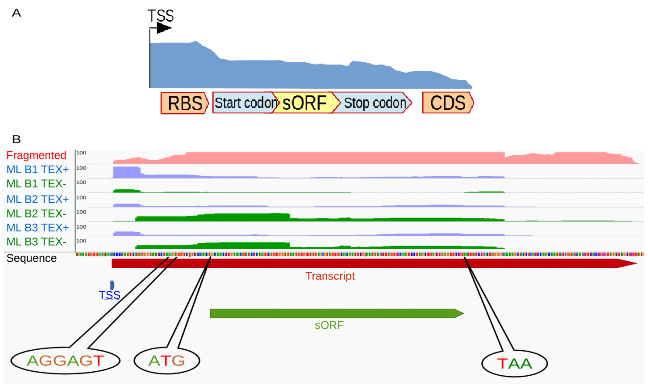

Figure 6: sORF detection. **(A)** An sORF must contain start codon and stop codon within transcript and should be inside of a given length range (default 30 - 150 nt). Additionally, a ribosomal binding site must be detected between the TSS and the start codon. **(B)** An example from *H. pylori* 26695. 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.

3.6 Circular RNAs

ANNOgesic integrates the tool "testrealign.x" from the segemehl package for the detection of circular RNAs [60] 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 [61]. The CRISPR Recognition Tool (CRT) [62] is integrated into ANNOgesic and extended by comparison of CRISPR/Cas candidates to other annotations to remove false positive (Supplementary Figure 14).

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 [63] 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 [64].

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 pipeline 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 has conventional (fragmentation) RNA-Seq as well as dRNA-Seq as input (see Supplementary Figure 15), it can also perform sufficient predictions with only one class of data for the majority of the genomic features.

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 Supplementary Figure 16 and 17).

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* [65] and *Rhodobacter sphaeroides* [66]. 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 FUNDING

The project was funded by the German Research Foundation (DFG) as part of the Transregio 34 (CRC-TRR34). This publication was funded by the German Research Foundation and the University of Würzburg in the funding program Open Access Publishing.

6 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] Delcher A.L., Bratke K.A., Powers E.C. and Salzberg S.L. (2007) Identifying bacterial genes and endosymbiont DNA with *Glimmer Bioinformatics*, **23**, 673–679.

[2] Schattner P., Brooks A.N. and Lowe T.M. (2005) The *tRNAscan-SE*, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs *Nucleic Acids Res.*, **33**, W686–W689.

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

[4] Richardson E. J. and Watson M. (2013) The automatic annotation of bacterial genomes *Brief. Bioinform.*, **14**, 1–12.

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

[6] Weinmaier T., Platzer A., Frank J., Hellinger, H.J., Tischler P. and Rattei T. (2016) *ConsPred*: a rule-based (re-)annotation framework for prokaryotic genomes *Bioinformatics*, btw393.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[23] Cock P., Antao T., Chang J.T., Chapman B.A., Cox C.J., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B. and de Hoon M.J.L. (2009) Biopython: freely available *Python* tools for computational molecular biology and bioinformatics *Bioinformatics*, **25**, 1422–1423.

[24] van der Walt S., Colbert S.C. and Varoquaux G. (2011) The *NumPy* Array: A Structure for Efficient Numerical Computation *Comput. Sci. Eng.*, **13**, 22–30.

[25] Hunter J.D. (2007) Matplotlib: *A* 2D *Graphics Environment Comput. Sci. Eng.*, **9**, 90–95.

[26] Hagberg A.A., Schult D.A. and Swart P.J. (2008) Exploring *Network Structure*, *Dynamics*, and *Function* using *NetworkX Proceedings of the 7th Python in Science conference*, 11–15.

[27] Merkel D. (2014) Docker: *Lightweight Linux Containers* for *Consistent Development* and *Deployment*. *Linux Journal*.

[28] Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform *Bioinformatics*, **25**, 1754–1760.

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

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

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

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

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

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

[35] Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The *Sequence Alignment/Map* format and *SAMtools Bioinformatics*, **25**, 2078–2079.

[36] Otto T.D., Dillon G.P., and Degrave W.S. and Berriman M. (2011) RATT: Rapid Annotation Transfer Tool *Nucleic Acids Res.*, **39**, e57.

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

[38] Waters L.S. and Storz G. (2009) Regulatory RNAs in Bacteria *Cell*, **136**, 615–628.

[39] Bouvier M., Sharma C.M., Mika F., Nierhaus K.H. and Vogel J. (2008) Small RNA Binding to 5âĂš mRNA Coding Region Inhibits Translational Initiation *Mol. Cell*, **32**, 827–837.

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

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

[42] Hochschild A. (2007) Gene-Specific Regulation by a Transcript Cleavage Factor: Facilitating Promoter Escape *J. Bacteriol.*, **189**, 8769–8771.

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

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

[45] Frith M.C., Saunders N.F.W., Kobe B. and Bailey T.L. (2008) Discovering Sequence Motifs with Arbitrary Insertions and Deletions *PLoS Comput. Biol.*, **4**, e1000071.

[46] Holmqvist E., Wright P. R, Li L., Bischler T., Barquist L., Reinhardt R., Backofen R. and Vogel J. (2016) Global RNA recognition patterns of postâĂŘtranscriptional regulators Hfq and CsrA revealed by UV crosslinking *in vivo EMBO J.*, **35**, 991–1011

[47] Miyakoshi M., Chao Y. and Vogel J. (2015) Regulatory small RNAs from the 3âĂš regions of bacterial mRNAs *Curr. Opin. Microbiol.*, **24**, 132–139

[48] Smirnov A., Förstner K. U., Holmqvist E., Otto A., Günster R., Becher D., Reinhardt R. and Vogel J. (2016) rad-seq guides the discovery of ProQ as a major small RNA-binding protein *P. Natl. Acad. Sci. USA*, **113**, 11591–11596

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

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

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

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

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

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

[55] The Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward *Nucleic Acids Res.* **43**, D1049–D1056.

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

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

[58] Franceschini A., Szklarczyk D., Frankild S., Kuhn M., Simonovic M., Roth A., Lin J., Minguez P., Bork P., von Mering C. and Jensen L.J. (2013) *STRING* v9.1: protein-protein interaction networks, with increased coverage and integration *Nucleic Acids Res.*, **41**, D808–815.

[59] Kim S., Shin S.Y., Lee I.H., Kim S.J., Sriram R. and Zhang B.T. (2008) *PIE*: an online prediction system for protein-protein interactions from text

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

[61] Sander J.D., and Joung J.K. (2014) CRISPR-Cas systems for editing, regulating and targeting genomes *Nat. Biotechnol.*, **32**, 347–355.

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

[63] 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. and Finn R.D. (2014) Rfam 12.0: updates to the *RNA* families database *Nucleic Acids Res.*.

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

[65] Dingemans J., Monsieurs P., Yu S.H., Crabbé A., Förstner K.U., Malfroot A., Cornelis P. and Van Houdt R. (2016) Effect of Shear Stress on Pseudomonas aeruginosa Isolated from the Cystic Fibrosis Lung *mBio*, **7**, e00813–16.

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