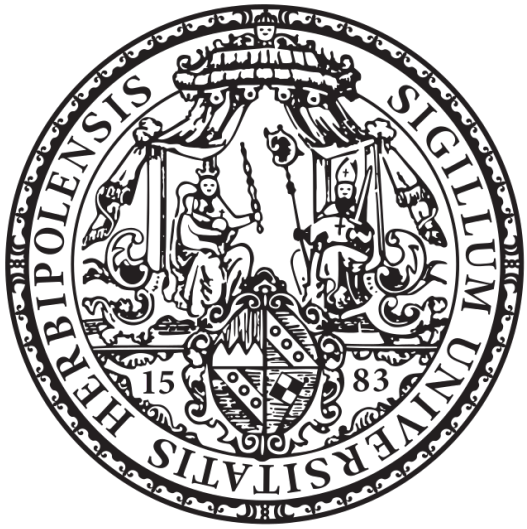


Development and application of computational tools for RNA-Seq based transcriptome annotations

Entwicklung und Anwendung bioinformatischer Werkzeuge für RNA-Seq-basierte Transkriptom-Annotationen

Doctoral thesis

Graduate School of Life Sciences,
Julius-Maximilians-Universität Würzburg,
Section: Infection and Immunity

Submitted by

Sung-Huan Yu

Würzburg, 2018

Submitted on:

Members of the Doctoral Thesis Committee:

Chairperson: Prof. Dr. Jörg Schultz

Primary Supervisor: Prof. Dr. Jörg Vogel

Supervisor (Second): Prof. Dr. Thomas Dandekar

Supervisor (Third): Prof. Dr. Cynthia Sharma

Date of Public Defence: ……

Date of Receipt of Certificates: …….

Abstract

In order to understand the regulation of gene expression in organisms, precise genome annotation is essential. In recent years, RNA-Seq has become a potent method for generating and improving genome annotations. However, this approach is time consuming and often inconsistently performed when done manually. In particular, the discovery of non-coding RNAs benefits strongly from the application of RNA-Seq data but requires significant amounts of expert knowledge and is labor-intensive. As a part of my doctoral study, I developed a modular tool called ANNOgesic that can detect numerous transcribed genomic features, including non-coding RNAs, based on RNA-Seq data in a precise and automatic fashion with a focus on bacterial and achaeal species. The software performs numerous analyses and generates several visualizations. It can generate annotations of high-resolution that are hard to produce using traditional annotation tools that are based only on genome sequences. ANNOgesic can detect numerous novel genomic features like UTR-derived small non-coding RNAs for which no other tool has been developed before. ANNOgesic is available under an open source license (ISCL) at https://github.com/Sung-Huan/ANNOgesic.

My doctoral work not only includes the development of ANNOgesic but also its application to annotate the transcriptome of *Staphylococcus aureus* HG003 - a strain which has been a insightful model in infection biology. Despite its potential as a model, a complete genome sequence and annotations have been lacking for HG003. In order to fill this gap, the annotations of this strain, including sRNAs and their functions, were generated using ANNOgesic by analyzing differential RNA-Seq

data from 14 different samples (two media conditions with seven time points), as well as RNA-Seq data generated after transcript fragmentation. ANNOgesic was also applied to annotate several bacterial and archaeal genomes, and as part of this its high performance was demonstrated. In summary, ANNOgesic is a powerful computational tool for RNA-Seq based annotations and has been successfully applied to several species.

Zusammenfassung

Exakte Genomannotationen sind essentiell für das Verständnis Genexpressionsregulation in verschiedenen Organismen. In den letzten Jahren entwickelte sich RNA-Seq zu einer äußerst wirksamen Methode, um solche Genomannotationen zu erstellen und zu verbessern. Allerdings ist das Erstellen von Genomannotationen bei manueller Durchführung noch immer ein zeitaufwändiger und inkonsistenter Prozess. Die Verwendung von RNA-Seq-Daten begünstigt besonders die Identifizierung von nicht-kodierenden RNAs, was allerdings arbeitsintensiv ist und fundiertes Expertenwissen erfordert. Ein Teil meiner Promotion bestand aus der Entwicklung eines modularen Tools namens ANNOgesic, das basierend auf RNA-Seq-Daten in der Lage ist, eine Vielzahl von Genombestandteilen, einschließlich nicht-kodierender RNAs, automatisch und präzise zu ermitteln. Das Hauptaugenmerk lag dabei auf der Anwendbarkeit für bakterielle und archaeale Genome. Die Software führt eine Vielzahl von Analysen durch und stellt die verschiedenen Ergebnisse grafisch dar. Sie generiert hochpräzise Annotationen, die nicht unter Verwendung herkömmlicher Annotations-Tools auf Basis von Genomsequenzen erzeugt werden könnten. Es kann eine Vielzahl neuer Genombestandteile, wie kleine nicht-kodierende RNAs in UTRs, ermitteln, welche von bisherigen Programme nicht vorhergesagt werden können. ANNOgesic ist unter einer Open-Source-Lizenz (ISCL) auf https://github.com/Sung-Huan/ANNOgesic verfügbar.

Meine Forschungsarbeit beinhaltet nicht nur die Entwicklung von ANNOgesic, sondern auch dessen Anwendung um das Transkriptom des *Staphylococcus aureus*-Stamms HG003 zu annotieren. Dieser ist einem Derivat von *S. aureus* NCTC8325 - ein

Stamm, Dear ein bedeutendes Modell in der Infektionsbiologie darstellt. Zum Beispiel wurde er für die Untersuchung von Antibiotikaresistenzen genutzt, da er anfällig für alle bekannten Antibiotika ist. Der Elternstamm NCTC8325 besitzt zwei Mutationen im regulatorischen Genen (*rsbU* und *tcaR*), die Veränderungen der Virulenz zur Folge haben und die in Stamm HG003 auf die Wildtypsequenz zurückmutiert wurden. Dadurch besitzt *S. aureus* HG003 das vollständige, ursprüngliche Regulationsnetzwerk und stellt deshalb ein besseres Modell zur Untersuchung von sowohl Virulenz als auch Antibiotikaresistenz dar. Trotz seines Modellcharakters fehlten für HG003 bisher eine vollständige Genomsequenz und deren Annotationen. Um diese Lücke zu schließen habe ich als Teil meiner Promotion mit Hilfe von ANNOgesic Annotationen fÃijr diesen Stamm, einschlieÃ§lich sRNAs und ihrer Funktionen, generiert. Dafür habe ich Differential RNA-Seq-Daten von 14 verschiedenen Proben (zwei Mediumsbedingungen mit sieben Zeitpunkten) sowie RNA-Seq-Daten, die von fragmentierten Transkripten generiert wurden, analysiert. Neben *S. aureus* HG003 wurde ANNOgesic auf eine Vielzahl von Bakterien- und Archaeengenome angewendet und dabei wurde eine hohe Performanz demonstriert. Zusammenfassend kann gesagt werden, dass ANNOgesic ein mächtiges bioinformatisches Werkzeug für die RNA-Seq-basierte Annotationen ist und für verschiedene Spezies erfolgreich angewandt wurde.

Abbreviations and symbols

5'-P	5' monophosphate
5'-PPP	5' triphosphate
API	Application programming interface
cDNA	Complementary DNA
CDS	Protein-coding sequence
CircRNA	Circular RNA
CLIP-Seq	Cross-linking immunoprecipitation sequencing
CRISPR	Clustered regularly interspaced short palindromic repeat
dRNA-Seq	Differenial RNA sequencing
FP	False positive
FPR	False positive rate
GO	Gene ontology
Grad-Seq	Gradient sequencing
KEGG	Kyoto Encyclopedia of Genes and Genomes
OD_{600}	Optical density of a sample measured at a wavelength of 600 nm
ORF	Open reading frame
PPI	Protein-protein interaction
ROC curve	receiver operating characteristic curve
PS	Processing site
QC	Quality control
RBS	Ribosome binding site
RIP-Seq	RNA Immunoprecipitation sequencing

RNA-Seq	RNA sequencing
RNAT	RNA thermometer
rRNA	Ribosomal RNA
SaPI	*Staphylococcus aureus* pathogenicity islands
SD	Shine-Dalgarno Sequence
Spr	Small pathogenicity island RNAs
SigB	Transcription factor sigma B
SNP	Single Nucleotide Polymorphism
sPEP	Small peptide
sORF	Small open reading frame
sRNA	Small non-coding RNA
TEX	Terminator exonuclease
TP	True positive
TPR	True positive rate
tRNA	Transfer RNA
TSS	transcriptional start site
UTR	untranslated region

Contents

List of Tables

List of Figures

Chapter 1

Introduction

RNA-Sequencing

RNA-Sequencing (RNA-Seq) is a powerful and precise approach to analyze transcriptomes in order to detect genes, and quantify their expression levels [1]. It is widely used to study bacterial, archaeal and eukaryotic species. The currently dominating mode is high-throughput sequencing of complementary DNA (cDNA), for example with platforms provided by Illumina or Ion Torrent. The resulting reads have lengths around 50-400 bp depending on the used platforms. They are used for either a mapping to a reference genome [2–5] or a *de novo* transcriptome assembly [6, 7] in case the reference genome is not available (Figure 1.1).

Before analyzing RNA-Seq data, two initial steps need to be done - adapter clipping and quality trimming. Adapters which have to be ligated to DNA sequences during library preparation contain barcoding sequences, primers, and binding sequences for connecting short reads to the flow cell. Those adapter sequences have to

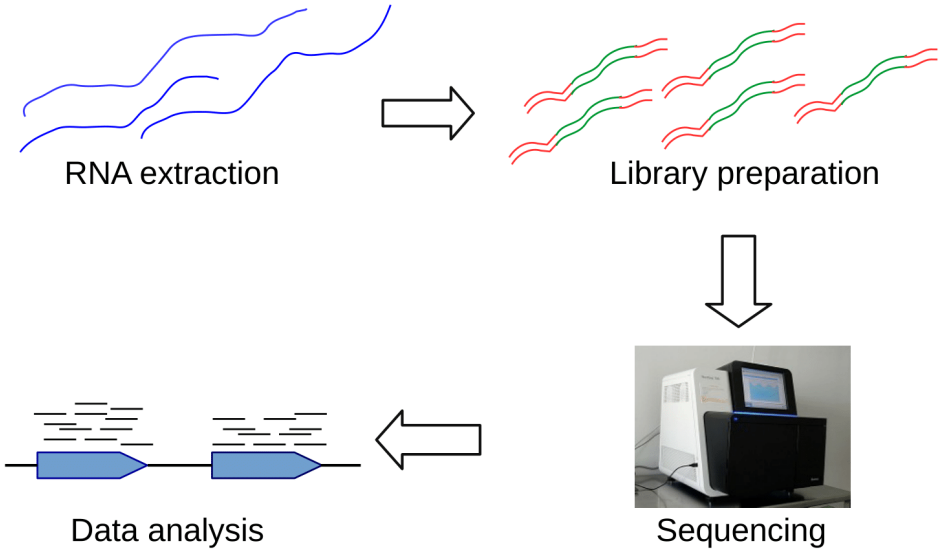


Figure 1.1: The general procedure of RNA-Seq. [8]

be removed in order to guarantee an optimal alignment of the reads. Furthermore, regions with low sequencing quality are trimmed of.

For mapping the RNA-Seq reads to reference genomes, numerous mapping tools like BWA [2], BWA-MEN [2], Bowtie2 [3], Segemehl [4], and STAR [5] were developed. Several pipelines like READemption [9] which integrates the aligners and other analysis software, such as DESeq2 [10] were also implemented.

Various RNA-Seq protocols were developed for detecting different genomic features and quantifying gene expression levels. The two protocols which were used for my doctoral work are differential RNA-Seq (dRNA-Seq) and RNA-Seq after transcript fragmentation. For the construction of dRNA-Seq libraries, the original sample is split into two different aliquots: one of them is treated by terminator exonuclease (in the following abbreviated as TEX+ library) which specifically degrades RNA molecules

with 5'- monophosphate (5'-P), while the other remains untreated (written as TEX- library in the following). Due to this procedure, primary transcripts are enriched in the TEX+ libraries, in comparison to the TEX- libraries. By the application of dRNA-Seq protocol, transcription starting sites (TSSs) can be detected through comparing the read coverage between TEX+ and TEX- libraries [11, 12] (Figure 1.2).

Read quality usually decreases towards the 3' end of reads and the bases of 3' end need to be removed in order to improve mappability. Due to this, the whole transcripts, especially the 3' end may not be able to be detected by using dRNA-Seq. RNA-Seq generated after transcript fragmentation was applied for solving this issue. The reads generated with this approach covers the whole expressed regions and help to identify transcript boundary without losing the information of the 3' end (Figure 1.3) [1].

Besides the two protocols mentioned above, some useful RNA-Seq based protocols for detecting specific genomic features were developed as well, such as Term-Seq [13] for detecting terminators and riboswitches, ribosome profiling for identifying open reading frames [14], RIL-Seq for identifying sRNA regulatory targets [15], CLIP-Seq [16] and RIP-Seq [17] for searching RNA-protein interaction, and Grad-Seq [18] for capturing RNA complexes. In order to translate these data into valuable insights, computational tools which can analyze these data with high quality performance need to be created.

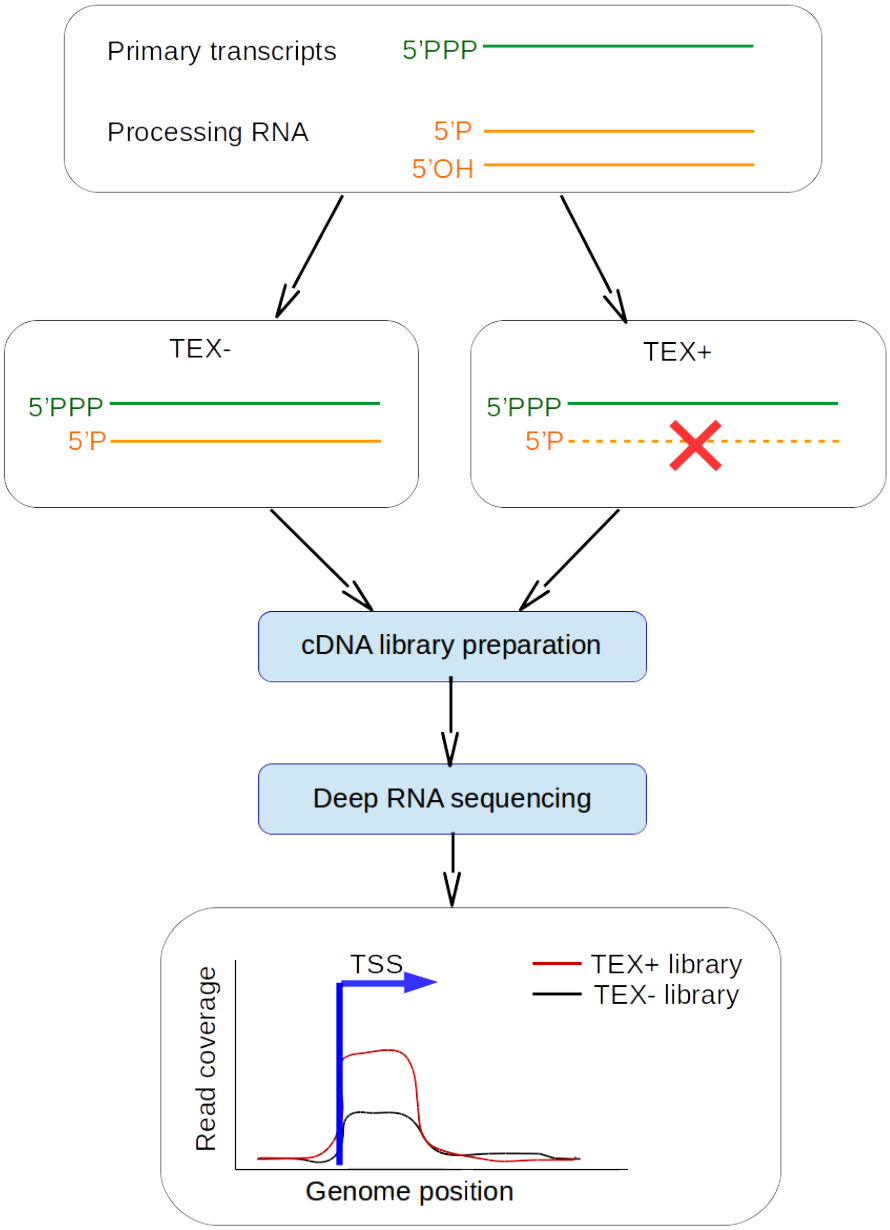


Figure 1.2: Workflow of dRNA-Seq.

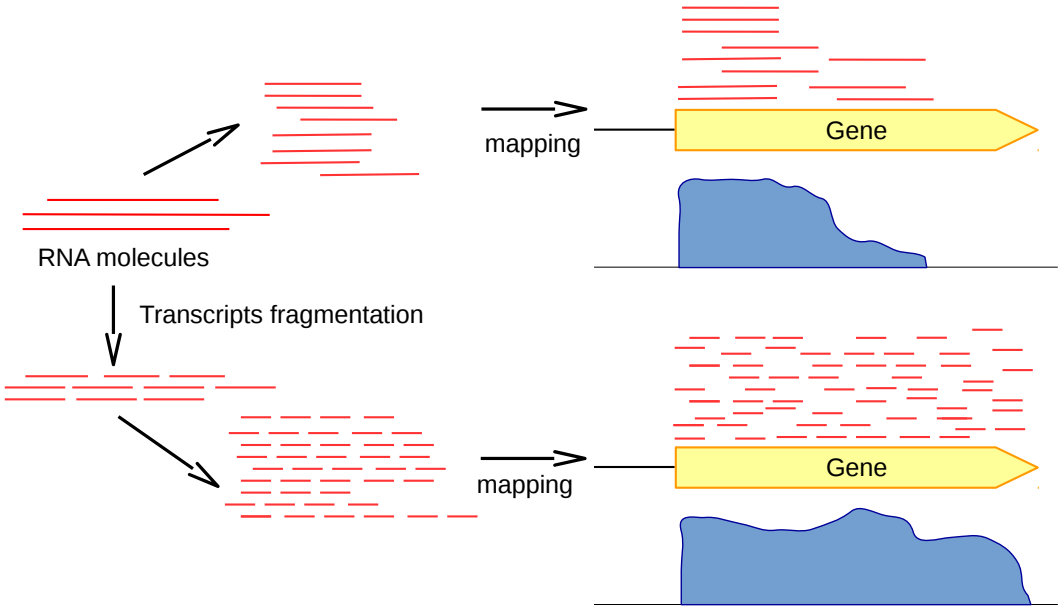


Figure 1.3: The workflow of RNA-Seq generated after transcript fragmented.

An overview of the available tools for genome annotations

A high resolution of genome annotation is essential for understanding the regulatory mechanisms of organisms. Due to the development of sequencing methods and the number of available genome sequences is increasing expeditiously, numerous tools purely based on genome sequences for detecting genomic features have been constructed. The representative tools are Glimmer for detecting open reading frames (ORFs) [19], tRNAscan-SE [20] for searching tRNAs, and RNAmmer [21] for predicting rRNAs [21]. In order to detect different genomic features, several genome annotation pipelines were created. Prominent examples are Prokka [22] and

ConsPred [23] which integrate several tools to identify multiple features in bacterial genomes. However, the predictions based only on the genome sequences are unreliable for certain features like TSSs which can only be predicted precisely by applying dRNA-Seq.

Since using RNA-Seq can significantly improve the predictions of genomic features, several methods were created in order to generate genome annotations based on RNA-Seq data, such as the computational tools for detecting TSSs [24–26] and transcripts [27–29]. EuGene-PP is a comprehensive pipeline which can generate multiple genome annotations based on genome sequence information or RNA-Seq data [30]. However, RNA-Seq data is only applied for the TSS prediction of EuGene-PP but not for other predictions like sRNA detection. Thus, the automatic integration and translation of the data from different RNA-Seq based protocols into high-quality genome annotations is still an unsolved issue.

Small non-coding RNAs

RNA-Seq is also widely used for the detection of small non-coding RNAs (sRNAs). Several thousands bacterial sRNAs have been identified by different methods. Members of non-coding bacterial RNAs are normally between 50 and 500 nucleotides long, are highly structured and usually contain several stem-loops. sRNAs can either pair with target mRNAs to regulate their translation, stability or bind to target proteins in order to modify their functions [31, 32]. Nearly all of the sRNAs are expressed under specific growth conditions like iron limitation, shear stress, nutrition starvation, oxidative stress etc [33, 34]. Based on the locations, sRNAs can be roughly split into

two classes: *cis*-encode RNAs (antisense sRNAs) which are transcribed opposite of the annotated genes, and *trans*-encode RNAs including intergenic and UTR-derived sRNAs (sRNAs that share a transcript with CDSs) [18, 34, 35].

In order to detect sRNAs, numerous sRNA prediction tools were built based on different methods which can be roughly divided to three types. The first class is based on sequence conservation of intergenic region such as QRNA [36] and Intergenic Sequence Inspector [37]. The second one is based on the information of secondary structure like RNAz [38] and sRNAPredict [39, 40]. The core methods of these tools rely on either the thermo-stability of secondary structures of conserved intergenic sequences, or the information of promoters and terminators. The final one is based on machine learning approaches, such as CoRAL [41] which using fragment length and cleavage specificity as input features to predict sRNAs. Additionally, several tools integrate more than one type of information to predict sRNAs like sRNAscanner [42] which uses both the information of sequence and structure. However, none of these tools use RNA-Seq data for their predictions.

Numerous sRNAs were recently identified by applying RNA-Seq In order to understand their functions, several sRNA target prediction tools or RNA-RNA interaction tools were constructed. Several studies compared such tools were published in the recent two years [43, 44]. In these studies, the performances of CopraRNA [45], IntaRNA [46], RNAplex [47, 48], and RNAup [48, 49] were shown to be better than their competitors. Still, these tools have several shortages. CopraRNA needs manually selected homologs from different species of an sRNA, it is no able to generate the results automatically. Usually, CopraRNA, IntaRNA and RNAup require long

computational time to search mRNA targets for one sRNA. Although executing time of RNAplex is the lowest, its performance is also the worst within these four tools. Based on the results of these analyses, the current sRNA target prediction tools still need to be improved.

Approaches for detecting the functions of genes

In order to understand the mechanisms of RNA-RNA interactions, the information of RNA secondary structures are essential requirements. For example, CsrA-binding sRNAs contain a highly conserved GGA triplet nucleotides located on the loop part of hairpin [50, 51]. Due to the importance of the secondary structure information, variant tools for predicting secondary structures of RNAs were developed such as RNAfold [48, 52], CMfinder [53], and UNAFold [54]. These tools not only predict secondary structures of RNAs precisely, but also provide visualization.

For understanding the functions and regulatory networks of genes, two representative databases were built – Kyoto Encyclopedia of Genes and Genomes (KEGG) [55] and Gene Ontology (GO) [56, 57]. KEGG contains systems information, genomic information, chemical information and health information. By applying the information stored in KEGG, the homologs of query genes and their possible regulation networks can be found. GO is another system which is widely used by numerous annotation tools for characterizing the functions of genes across all species. GO provides numerous controlled vocabularies which can be divided into three groups (biological processes, cellular components and molecular functions) to classify the functions and locations of genes. Moreover, the GO terms can be constructed to a

hierarchy tree for revealing the functions and relative amount of genes for a whole genome (Figure 1.4).

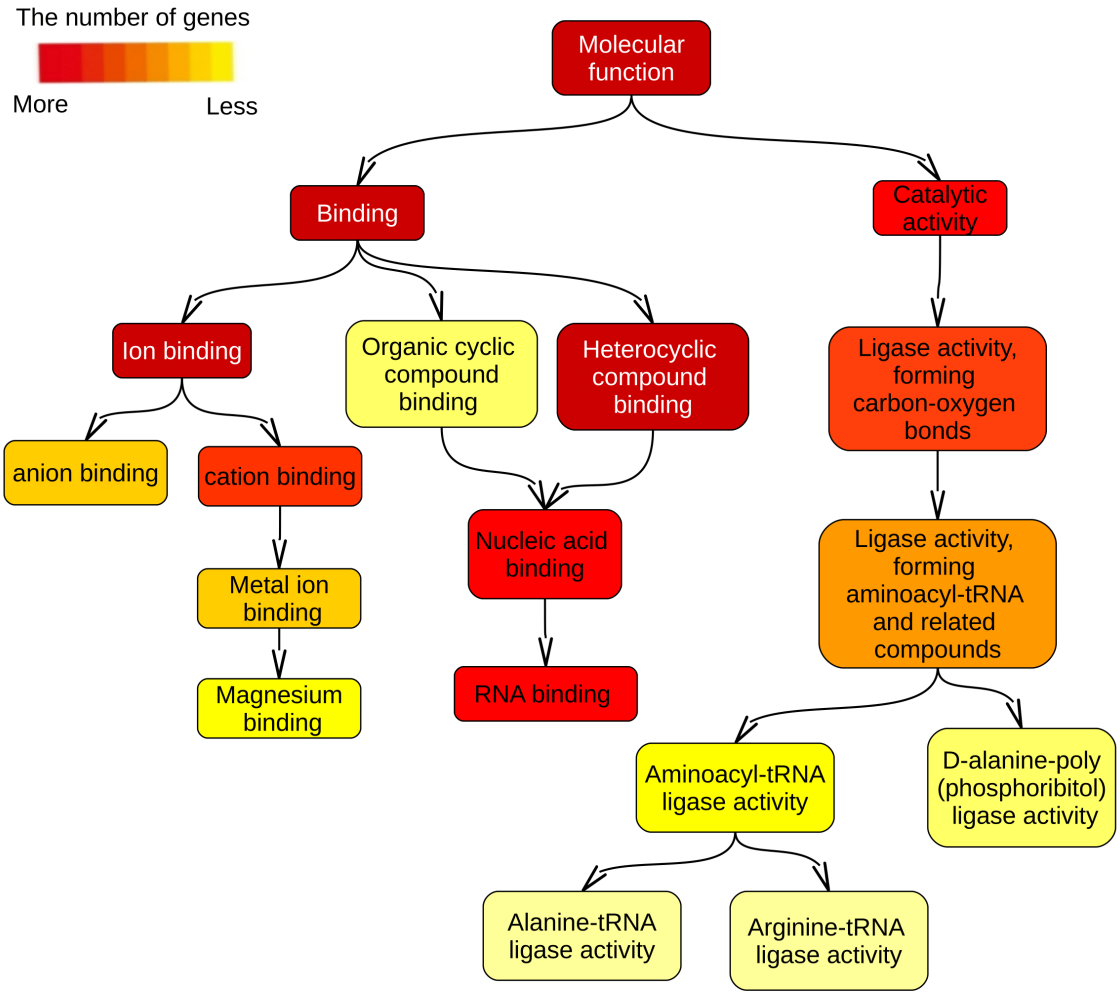


Figure 1.4: An example of the hierarchy tree for GO term. The functions and relative amount of genes for *S. aureus* HG003 can be shown in the tree.

Actually, RNA-Seq information is also a valuable resource for detecting the functions and regulation network of genes. Gene co-expression analysis is one of the commonly used method for exploring the functions of newly discovered genes by clustering the genes which show a similar expression pattern across samples or

conditions [58, 59].

Previous studies of genome annotations based on RNA-Seq

Since RNA-Seq has become a powerful tool to improve genome annotations for many organisms, numerous studies provide genome annotations in bacterial [60, 61], archaeal [62] and eukaryotic [63] genomes based on RNA-Seq.

For example, many useful genomic features, especially sRNAs were detected based on RNA-Seq data for *S. aureus* in recent years. In 2010, Bohn *et al.* successfully applied sequencing approach to identify 30 sRNAs including 14 newly discovered ones [61]. In 2015, SRD (a Staphylococcus regulatory RNA database) was constructed for providing sRNAs reported from literatures or detected by computational methods [64]. Moreover, a global sRNA identification of three strains of *S. aureus* was performed based on RNA-Seq data in 2016 [65].

However, the useful RNA-Seq protocols which can precisely detect transcript boundary were not used to these study, such as dRNA-Seq [11, 12] for TSSs and Term-Seq [13] for terminators. In addition, nearly all of the genomic features detected in these studies were manually curated, it is a time consuming and inconsistent process. Thus, developing an automatic tool which can detect sRNAs based on RNA-Seq data by experimental validation might be more consistent and reliable.

ANNOgesic: A tool for generating genome annotations

In order to fill the gap of lacking computational tools for predicting genomic features based on RNA-Seq data, I constructed ANNOgesic which can process RNA-Seq data from different protocols to automatically generate high-quality genome annotations for bacterial and archaeal genomes. It is a modular tool that is able to predict multiple genomic features via different modules. Many modules were newly developed for detecting the genomic features which cannot be detected by using the currently available tools. If the the genomic features can be predicted by the third-party tools, while others were created by integrating third-party tools with significant improvements like parameters optimization and removing false positives.

ANNOgesic was also successfully applied for many RNA-Seq data sets of bacteria and archaea, and high performance was shown. ANNOgesic can identify genes, protein-coding sequences (CDSs), tRNAs, rRNAs, TSSs, processing sites (PSs), transcripts, terminators, untranslated regions (UTRs), operons as well as sub- operons, promoter motifs, sRNAs, small open reading frames (sORFs), circular RNAs (circRNAs), CRISPRs, riboswitches, and RNA thermometers. Furthermore, it can predict RNA-RNA and protein-protein interaction as well. Additionally, ANNOgesic can allocate Gene Ontology (GO) terms and subcellular localizations to proteins. In order to help the user to analyze the genomic features, numerous statistics and visualizations are also provided. All modules of ANNOgesic are presented in Figure 1.5.

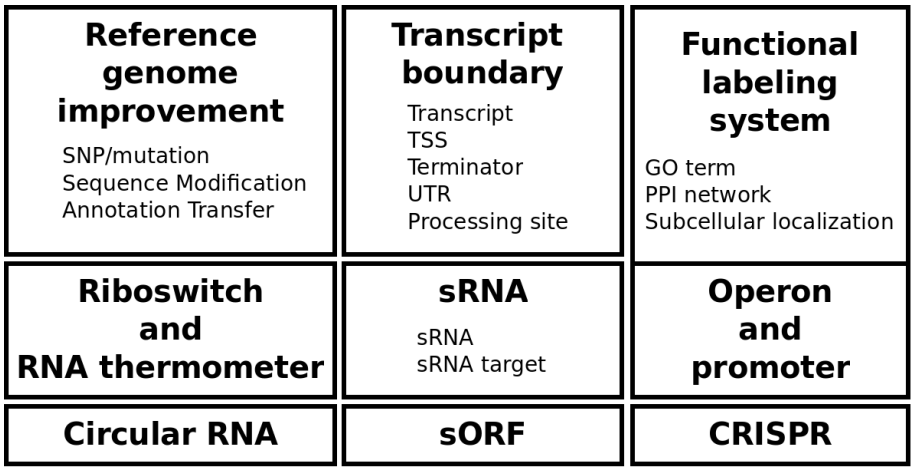


Figure 1.5: The modules of ANNOgesic

Applying ANNOgesic to *Staphylococcus aureus* HG003

As a part of my doctoral work, I not only created ANNOgesic, but also applied it to an important bacterial pathogen – *S. aureus*. *S. aureus* is a gram-positive bacterium and an intensively studied pathogen for bacterial infection. It leads to skin infections, respiratory disease, food poisoning, and septic arthritis as well as meningitis in infants [66, 67].

S. aureus produces various virulence factors such as Panton-Valentine leukocidin (PVL) which can cause leukocyte destruction and necrotizing pneumonia [67]. Moreover, the pathogenicity island of *S. aureus* (SaPI), which can be transferred by plasmids, phages, or conjugative transposons, contains virulence and antibiotic resistance genes and can promote the pathogenesis of infection [68]. SaPI is a 15-20 kb molecule occupied at constant chromosomal sites, and carries numerous genes

for superantigen toxins. SaPIs have similar attributes as bacteriophage such as genes coding for integrases, helicases and terminases, and flanking direct repeats [69]. Furthermore, the Spr (small pathogenicity island RNAs) family, which may play an important role in staphylococcal virulence, is expressed from SaPI [70].

S. aureus HG003 is a derivative strain of *S. aureus* NCTC8325 which is a relevant model strain for the studies of antibiotic resistance transfer and carriage by plasmids, as it is sensitive to all known antibiotics. However, *S. aureus* NCTC8325 is defective in two regulators, *rsbU* (deletion) which is an activator of SigB, and *tcaR* (point mutation) which is an activator of protein A transcription. In *S. aureus* HG003, these two genes are repaired and the original regulation network is preserved. Strain HG003 has further interesting characteristics including weak hemolysis, high spa transcript levels, strong biofilm formation and high virulence, all of which make it an useful strain for infection studies [66]. Henceforth, generating a complete genome sequence and annotations of this strain is a foundation for understanding the infection and gene regulation networks of *S. aureus*.

However, a complete genome sequence and genome annotations of *S. aureus* HG003 are not available currently. Although the annotations of some closely related strains like *S. aureus* NCTC8325 can be found, several important genomic features are still missing like sRNAs, TSSs, terminators, etc. In order to fill this gap, ANNOgesic was used to generate a genome sequence and annotations for *S. aureus* HG003 based on the data of dRNA-Seq and RNA-Seq generated after transcript fragmentation in this study. Furthermore, allocation of the potential functions for sRNAs were done by using gene co-expression analysis.

Data-driven research v.s. hypothesis-driven research

The massive quantities of data is an accompaniment of the application of RNA-Seq. As in many other fields of research, a paradigm shift has happened: Instead of the classical hypothesis-driven approach in which hypotheses are made and then testing it by experimentation, a data-driven research mode is performed [71,72]. Data-driven research uses the scientific methods, algorithms and tools to extract knowledge and insights from data in various forms. My doctoral work, which follows the data-driven path, is to create a tool for detecting and analyzing genomic features for bacterial or archaeal genomes based on RNA-Seq data.

Chapter 2

Methods and Materials

Used RNA-Seq data sets

The RNA-Seq data of *S. aureus* HG003 comprises 14 dRNA-Seq data sets and 1 RNA-Seq data set generated after transcript fragmentation. The samples of the 14 dRNA-Seq data were gained from two media (rich media and poor media) with seven time points (three time points are in exponential phase, another three time points are in stationary phase, and the last one is for overnight) (Table 2.1). All the samples are without replicates.

ANNOgesic has been widely applied to numerous RNA-Seq data sets including bacterial genomes (*Helicobacter pylori* 26695 [60], *Campylobacter jejuni* 81116 [24], *Pseudomonas aeruginosa* [73] and *Rhodobacter sphaeroides* [74], archaeal genomes (*Methanosarcina mazei* (Lutz *et al.*, unpublished)), and eukaryotic genomes which have no introns (*Trypanosoma brucei* (Müller *et al.*, unpublished)) (Table 2.2). In order to test several predictions of ANNOgesic like parameter optimization of TSS

Table 2.1: The time points of dRNA-Seq data for *S. aureus* HG003

Phase	Time point
Exponential phase	$OD_{600} = 0.2$ (OD 0.2) $OD_{600} = 0.5$ (OD 0.5) $OD_{600} = 1$ (OD 1)
Stationary phase	0 hour (t0) 2 hour (t1) 4 hour (t2) Overnight (ON)

prediction, sRNA detection, and CRISPR identification, RNA-Seq data sets of *H. pylori* 26695 [12, 60] and *C. jejuni* 81116 [24] were also retrieved from NCBI GEO where they are stored under the accession numbers GSE67564 and GSE38883, respectively.

Moreover, dRNA-Seq data and conventional RNA-Seq data sets of *Escherichia coli* K12 MG1655 were also retrieved from NCBI GEO (accession number: GSE55199 and GSE45443 (only the data of wild type was retrieved)) in order to assess the performances of ANNOgesic's predictions [27, 75]. The predicted features of ANNOgesic were compared to published databases like RegulonDB, EcoCyc and DOOR2 [76–81].

Read mapping, mutation detecting and genome sequence update for *S. aureus* HG003

In general, detecting SNPs or mutations is based on DNA sequencing data. However, RNA-Seq reads can also be re-used to detect the SNPs or the differences of nucleotides

Table 2.2: The strains which annotations were generated by using ANNOgesic

Strains	Annotations
Staphylococcus aureus HG003	All features
Helicobacter pylori 26695	All features
Campylobacter jejuni 81116	All features
Pseudomonas aeruginosa CF PA9	Transcript, sRNA
Rhodobacter sphaeroides 2.4.1	TSS with optimization
Staphylococcus aureus HPV107	PS with optimization, TSS with optimization, transcript, sRNA
Sinorhizobium fredii NGR234	PS, TSS, transcript, terminator
Methanosarcina mazei Goe1	Transcript, sRNA, sORF
Staphylococcus epidermidis PS2	TSS, PS, transcript, CDS, terminator, UTR
Salmonella Typhimurium SL1344	TSS, Transcript
Escherichia coli K-12	TSS, Transcript, terminator, sRNA
Trypanosoma brucei 427 and 927	Transcript

in transcribed regions. Two drawbacks of using RNA-Seq data to identify SNPs are that only the expressed regions can be analyzed and the nucleotide change may be only exists in RNA not in DNA level due to RNA-editing. However, some studies have shown that the majority of SNPs are found in the expressed transcripts in eukaryotic genomes [82, 83]. Thus, RNA-Seq data may also be used for detecting SNPs and mutations if DNA-Seq data is not available.

Since the complete genome sequence of *S. aureus* HG003 is still unknown, The reads of *S. aureus* HG003 were mapped on *S. aureus* NCTC8325 by using READemption [9] which is a full RNA-Seq analysis pipeline. Afterward, the differences of nucleotides between these two strains were detected manually for modifying the

genome sequence of NCTC8325 in order to generate the genome sequence of *S. aureus* HG003. As long as the genome sequence of *S. aureus* HG003 is available, re-mapping the reads, generating alignment and coverage files, as well as computing gene quantification for *S. aureus* HG003 can be done by using READemption.

ANNOgesic

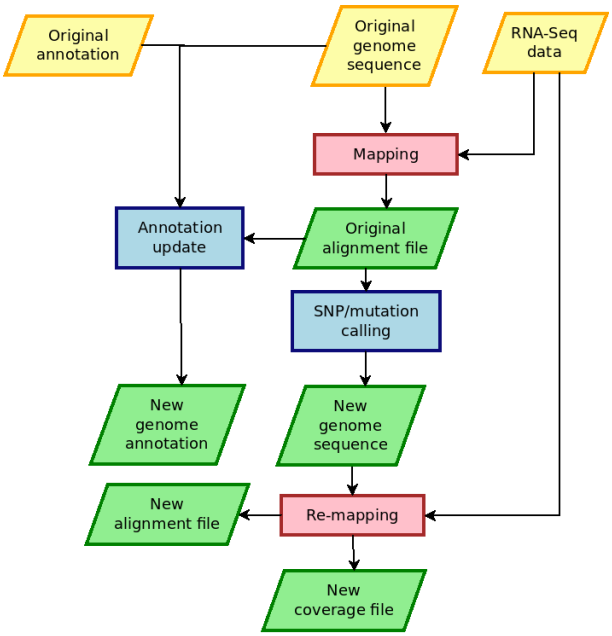
Implementation and installation

ANNOgesic is constructed in Python 3 and requires Biopython [84], numpy [85], matplotlib [86], and networkx [87]. All the source codes can be downloaded from a git repository [88], and a comprehensive documentation and tutorials are hosted at the site of "Read the Doc" [89]. ANNOgesic can be easily installed by using pip3 [90]. For installation of third-party software, a Docker image [91, 92] is provided as well.

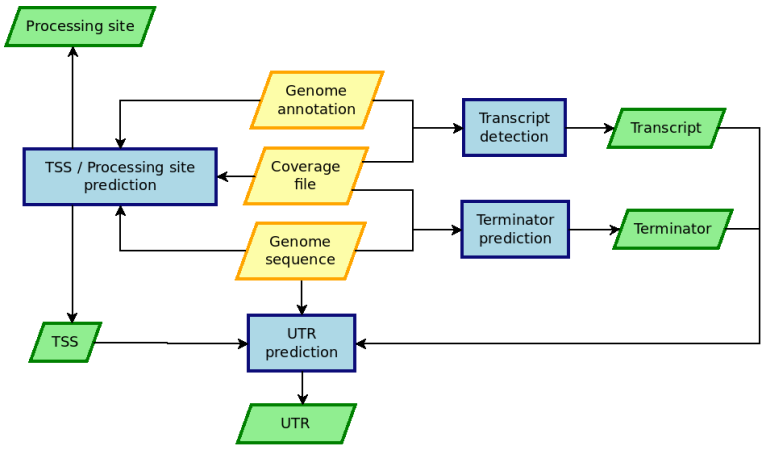
Modules and input data of ANNOgesic

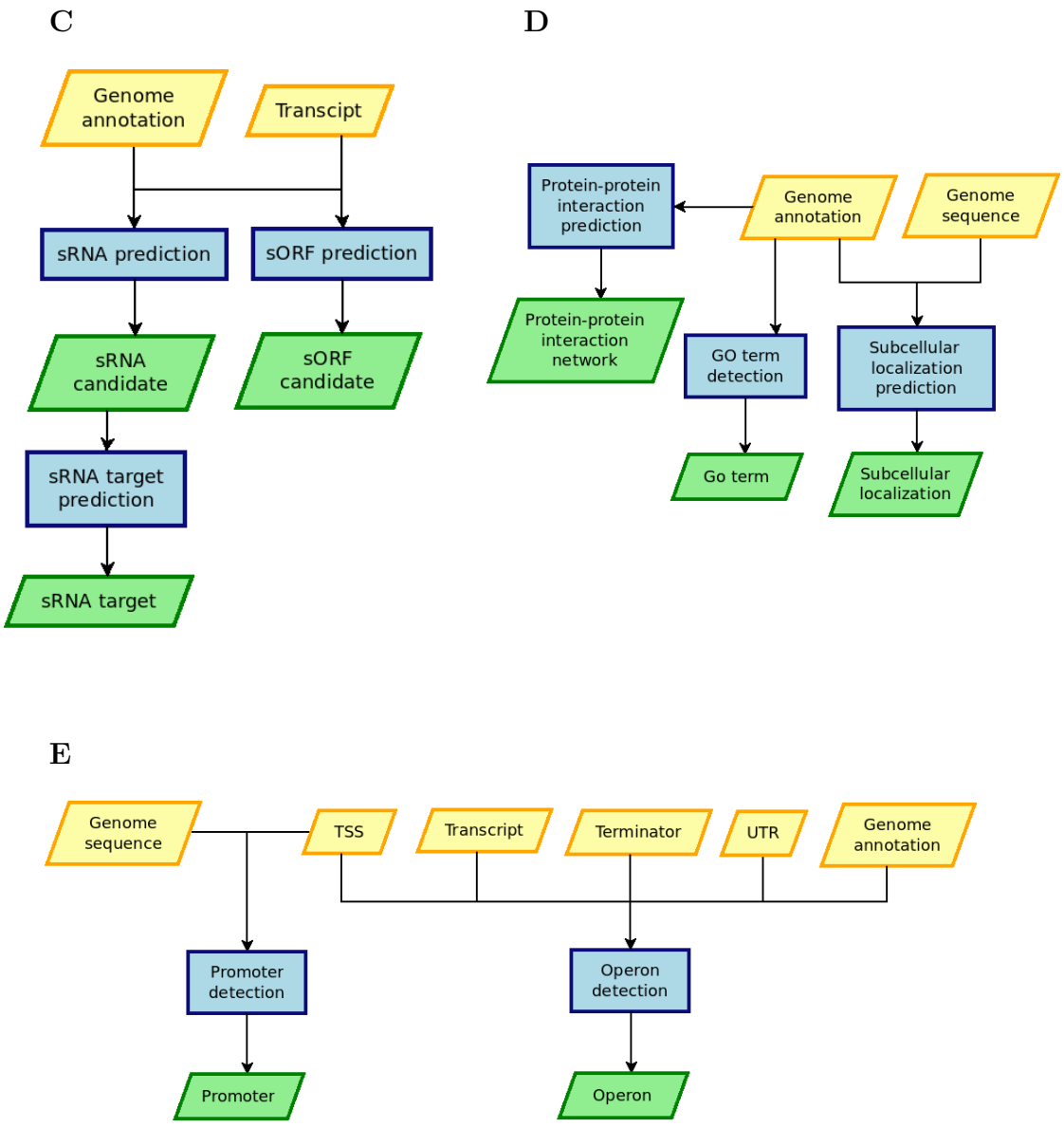
ANNOgesic is composed of the following twenty modules: Sequence modification, Annotation transfer, SNP/Mutation, Transcript, TSS, Terminator, UTR, PS, Promoter, Operon, sRNA, sRNA target, sORF, GO term, Protein-protein interaction network, Subcellular localization, Riboswitch, RNA thermometer, Circular RNA, and CRISPR. The workflows of connecting these modules are presented in Figure 2.1.

A


Original annotation
Original genome sequence
RNA-Seq data
Mapping
Annotation update
Original alignment file
SNP/mutation calling
New genome annotation
New genome sequence
New alignment file
Re-mapping
New coverage file

B


Processing site
Genome annotation
Transcript detection
Transcript
TSS / Processing site prediction
Coverage file
Terminator prediction
Terminator
Genome sequence
TSS
UTR prediction
UTR


C
Genome annotation
Transcipt
sRNA prediction
sORF prediction
sRNA candidate
sORF candidate
sRNA target prediction
sRNA target
D
Protein-protein interaction prediction
Genome annotation
Genome sequence
Protein-protein interaction network
GO term detection
Subcellular localization prediction
Go term
Subcellular localization
E
Genome sequence
TSS
Transcript
Terminator
UTR
Genome annotation
Promoter detection
Operon detection
Promoter
Operon

F

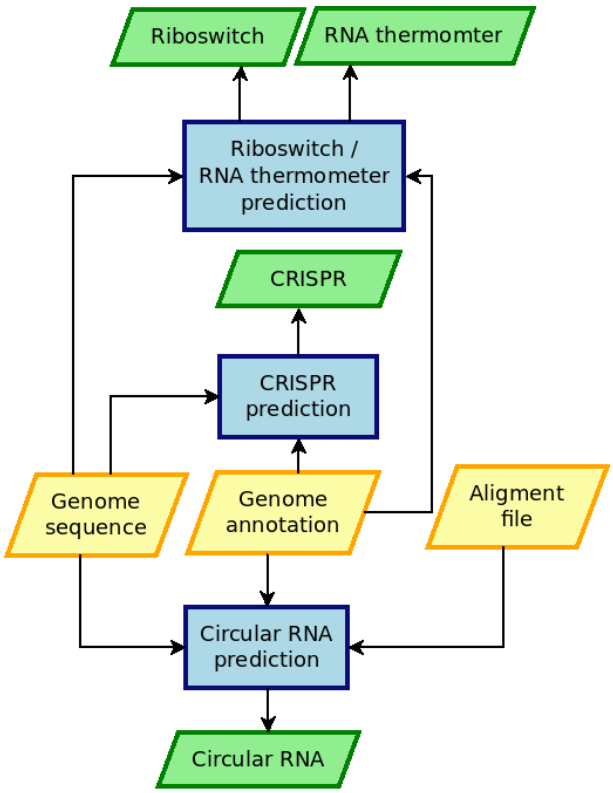


Figure 2.1: The workflows of the modules in ANNOgesic. The blue blocks represent the feature detection integrated in ANNOgesic. The red blocks represent the detection done by third-party tools. The yellow parallelograms and the green parallelograms indicate inputs and outputs, respectively. **(A)** Reference genome improvement, **(B)** Transcript boundary, **(C)** sRNA and sORF, **(D)** Functional labeling system, **(E)** Promoter and operon and **(F)** Other features.

Each module of ANNOgesic requires different input data like RNA-Seq coverage information in wiggle format, alignment data in BAM format, genome sequence in FASTA format, and annotations in GFF3 format. Wiggle files and BAM files can be generated by mapping tools such as BWA [2], STAR [5], segemehl [4], or a

full RNA-Seq analysis pipeline like READemption [9]. In case the queried genome sequences and annotations are not available, ANNOgesic can generate them from closely related strains.

Around half of the modules in ANNOgesic were newly developed for detecting the genomic features which can not be identified or not precisely detected by the currently available tools. The other modules not only integrated the third-party software for detecting the genomic features but also added improvements such as parameter optimization and removing false positives. The novelties and improvements of the available tools in ANNOgesic are listed in Table 2.3.

Detection of RNA-Seq coverage-based transcripts

Transcript detection is one of the core modules of ANNOgesic. Numerous predictions are based on the information of transcripts like the detections of sRNAs, sORFs, operons, and UTRs. Although many tools for detecting transcripts based on RNA-Seq data were created, most of the tools are optimized for the detection of eukaryotic transcripts, and only few of them can be used to bacterial species.

For the accurate detection of transcripts for bacterial genomes, a new method was created and integrated into ANNOgesic. The approach starts from searching gene expressed regions based on coverage values. Afterward, comparison between the expressed regions and gene annotations is performed in order to merge multiple transcripts located in the same gene. Additionally, several parameters can be assigned by the users to fine-tune the detection (Figure 2.2).

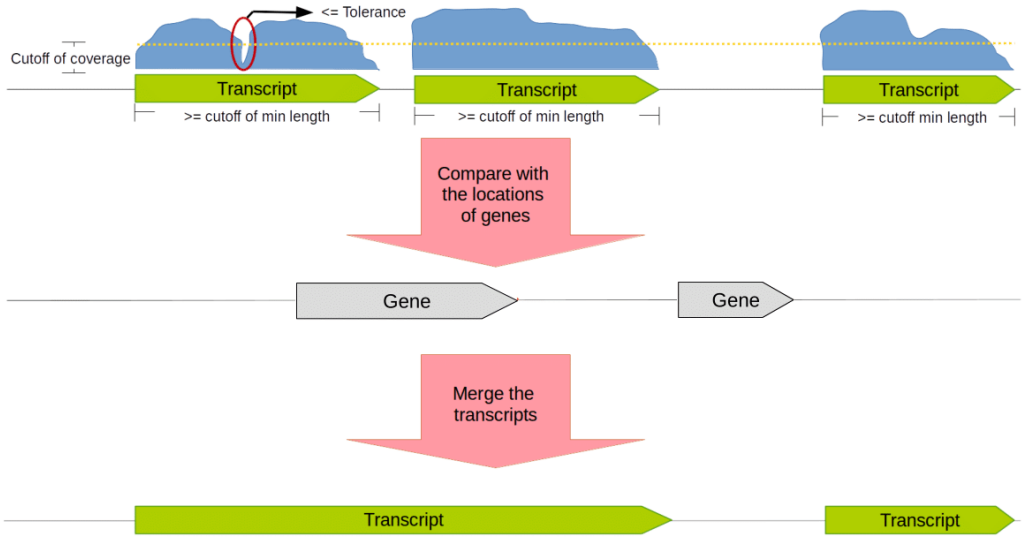


Figure 2.2: The method of RNA-Seq coverage-based transcript detection. If the coverage (blue curve- blocks) is higher than a given threshold of coverage (dash line), a transcript is defined. A tolerance value (i.e. The number of nucleotides with a coverage value below the tolerance) is set by the user for merging the gapped transcripts or keeping separated. Gene positions are applied to merge transcripts that overlap with the same gene.

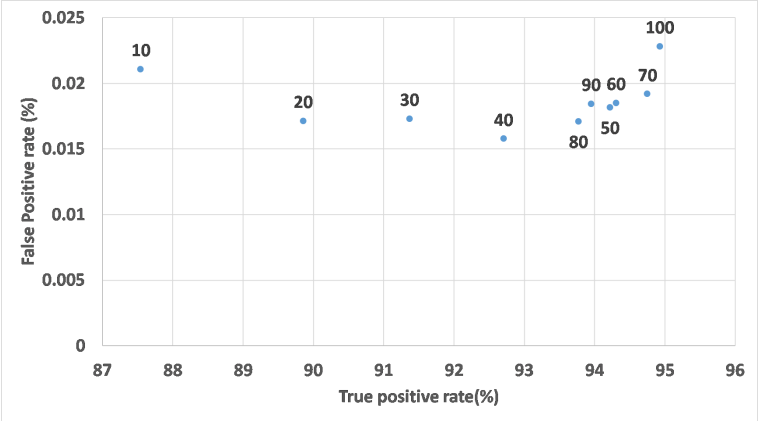
Optimization of TSSpredator's parameters

Several tools influenced by the selection of parameters were integrated into ANNOgesic, such as TSSpredator [24] which requires an experienced fine-tuning for the parameters (namely height, height reduction, factor, factor reduction, enrichment factor, processing factor and base height). In order to avoid the time-consuming manual parameter selection, ANNOgesic can search the optimized parameters by applying a genetic algorithm, a machine learning approach [93]. A small manually detected set of TSSs is used as a training set. In order to define the minimum number TSS in this set, a comparison for different number of benchmarking TSSs was

performed. The results shows that when the size of benchmarking set is larger than 50, the performance have no significant improvement (Figure 2.3). The approach of optimization is composed of three steps: a global change, a large change, and a small change that represent a random selection of values to all parameters, a random selection of values to two parameters, and adding or subtracting a small fraction to or from a parameter value, respectively (Figure 2.4). After each step of modifying the parameters, the results will be evaluated by a decision statement (Equation 2.1), and only the best parameters will be kept for the next step. In general, the optimized parameters can be obtained within 4,000 runs.

For the parameter optimization of TSSs in *S. aureus* HG003, 1,123 TSSs of the whole reference genome were detected manually. For PS, 58 PSs in the first 200 kb of the genome were identified manually. Based on the manually curated sets, the optimization of the TSS and PS predictions can be performed. After the optimization, manually detected set and computational-predicted set are merged by ANNOgesic to generate the final candidates of TSSs and PSs. The performance of the optimization are shown in the next chapter.

A



B

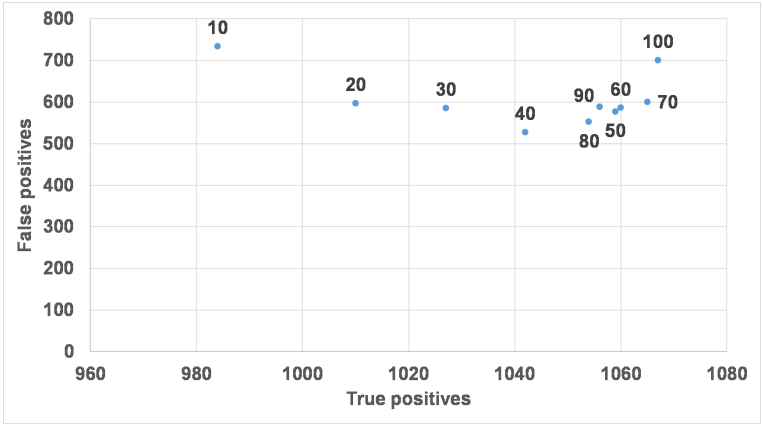


Figure 2.3: [The comparison for the number of manual TSSs in *S. aureus* HG003 parameter optimization.**(A)** shows the results of the comparison for the true postive rate and false positive rate. **(B)** is for the comparison between the number of true positives and false positives. The false positive rate is low because the amount of TSSs is relatively fewer than the number of genome nucleotides. The blue dots represent the number of benchmarking TSSs (the numbers shown near the dots) for the training.

Table 2.3: The new developed methods of the modules in ANNOgesic

Feature	Tools	New developed methods of ANNOgesic
SNP	SAMtools [94] and BCFtools [94]	Filter of QUAL and read depth
CDS/tRNA/rRNA	RATT [95]	Genbank (input) and GFF3 (output) format are acceptable
TSS and PS	TSSpredator [24]	Parameter optimization
Transcript	New approach*	Detecting expressed region and modifying transcripts based on genome annotation
Terminator	TranstermHP [96] and a New approach	Coverage drop detection and checking structures of the intergenic region between convergent genes
UTR	New approach	Comparison of TSSs, transcripts, CDSs, and terminators
Promoter	MEME [97] and GLAM2 [98]	Extraction of sequences automatically and TSS comparison
Operon	New approach	Comparison of TSSs, transcripts, CDSs, and terminators
sRNA	New approach	Detecting different types of sRNAs
sRNA target	RNAplex [47, 48] and RNAup [48, 49]	Merging RNAup and RNAplex
sORF	New approach	Searching ORFs in transcripts with a RBS
GO term	Uniprot [99, 100]	Comparison of transcripts
PPI network	STRING [101]	Network and Visualization with literature support by using PIE [102]
Subcellular localization	Psortb [103, 104]	Comparison of transcripts
Circrna	Segamehl [105]	Comparison of genome annotation
Riboswitch and RNA thermometer	New approach	Extracting sequences with a RBS in UTRs for a infernal [106] search in Rfam [107]
CRISPR	CRT [108]	Comparison of genome annotation

*"New approach" means that the approach was newly created in this work.

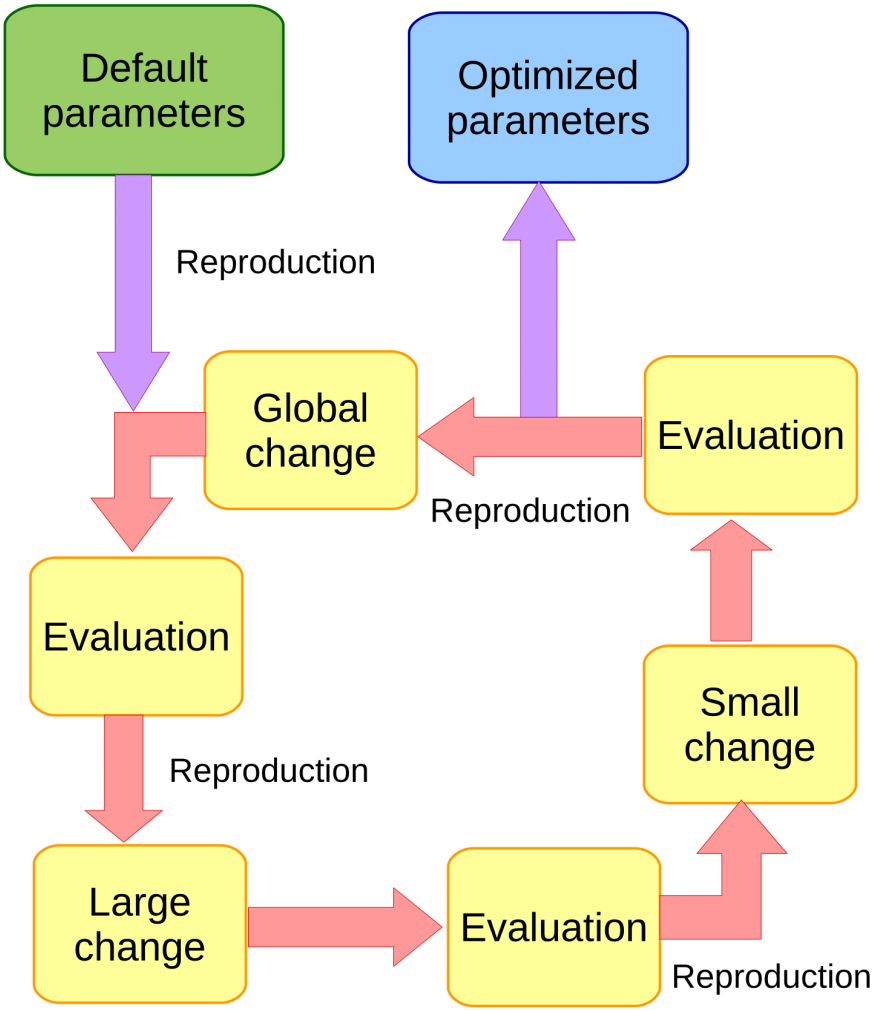


Figure 2.4: A genetic algorithm was applied for optimizing the parameters of TSSpredator. It starts from default parameters. Each iteration of this approach is composed of three steps - global change (change every parameter randomly), large change (change two of the parameters randomly), and then small change (adds/subtracts a small fraction to/from one parameter). The best parameters of each step will be selected for the next modification. Usually, ANNOgesic can achieve the optimized parameters within 4000 runs.

$$
\begin{aligned}
&TPR_c - TPR_b \geq 0.1 \\
&(TPR_c > TPR_b) \wedge (FPR_c < FPR_b) \\
&(TP_b - TP_c > 0) \wedge (FP_b - FP_c \geq 5 \times (TP_b - TP_c)) \\
&(TP_b - TP_c < 0) \wedge (FP_c - FP_b \leq 5 \times (TP_c - TP_b)) \\
&(TP_m \geq 100) \wedge (TPR_c - TPR_b \geq 0.01) \wedge (FPR_c - FPR_b \leq 5 \times 10^{-5}) \\
&(TP_m \geq 100) \wedge (TPR_b - TPR_c \leq 0.01) \wedge (FPR_b - FPR_c \geq 5 \times 10^{-5})
\end{aligned}
$$

Equation 2.1: TP_m is the number of manually detected TSSs. TP_c/TPR_c represents the true positives/true positive rate of the current parameters. TP_b/TPR_b represents the true positives/true positive rate of the best parameters. FP_c/FPR_c represents the false positives/false positive rate of the current parameters. FP_b/FPR_b represents the false positives/false positive rate of the best parameters. If one of these six statements is true, the best parameters will be replaced by the current parameters.

Allocating functions of sRNAs by using gene co-expression analysis

Although sRNAs can be detected by applying ANNOgesic, the functions of the newly discovered sRNAs are still unknown and hard to predict. Based on the data of the 14 RNA-Seq samples of *S. aureus* HG003 with different time points, the functions of sRNAs can be allocated by using gene co-expression analysis. First, gene quantification of CDSs and sRNAs was performed by READemption [9]. Afterward, tRNAs and rRNAs were removed due to their high expression which might influence the normalization. Hypothetical proteins and the non-expressed proteins (coverage $<$ 10 reads) were excluded as well in order to avoid noise. When the selection of genes and gene quantification were done, DESeq2 [10] was applied to compute $\log_2$

fold changes which were then used for expression kinetics and Spearman correlation coefficient calculation. In order to define the genes co-expressed and inversely expressed with the queried sRNAs, cutoffs of Spearman correlation coefficient are required. For *S. aureus* HG003, the cutoffs are 0.77 which is the 97.5 percentile of all-against-all correlation coefficients for positive correlation, and -0.77 which is the 2.5 percentile of all-against-all correlation coefficients for negative correlation were used (Figure 2.5). Moreover, GOATOOLS [109] was applied for extracting the enriched GO terms. Since the genes (including sRNAs and the known genes) which have a similar pattern of kinetic curves were clustered together, the potential functions of sRNAs may be related to the genes located in the same cluster (Figure 2.6).

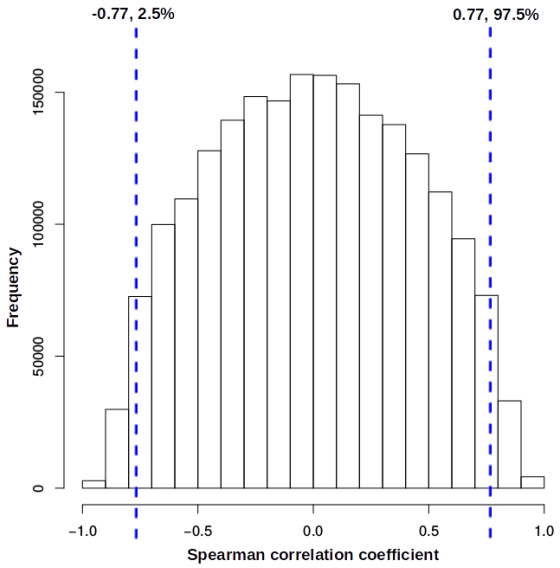


Figure 2.5: Spearman correlation coefficient of all-against-all of expression vaiues based on genes for *S. aureus* HG003. The cutoffs of correlation coefficients for correlation and anti-correlation are 0.77 (97.5 percentile) and -0.77 (2.5 percentile), respectively.

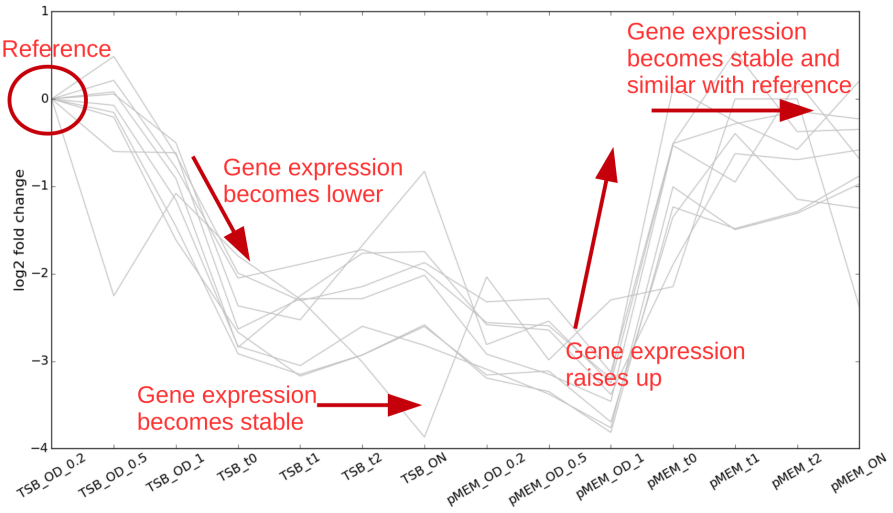


Figure 2.6: A schema of the gene co-expression analysis. The gray lines represent the kinetic curves of genes, x axis represents different conditions, and y axis shows expression values of $\log_2$ fold changes. The kinetic curve can be grouped based on the similarity of gene expression values. The example presents a group that all members are *pur* family genes.

Chapter 3

Results

An overview of the genomic features for *S. aureus* HG003

By applying ANNOgesic to *S. aureus* HG003, numerous high quality genome sequence and genomic annotations were generated. The genome sequence of *S. aureus* HG003 is composed of 2,821,354 base pairs, and the genome features include 2,872 genes, 2,778 proteins, 2,658 operons, 2,659 transcripts, 1,688 TSSs, 1,041 5' UTRs, 869 3' UTRs, 1,766 PSs, 1,359 terminators, 21 riboswitches, 11 RNA thermometers, 2 CRISPRs, 257 sRNAs and 143 sORFs (Table 3.1). Moreover, GO term, subcellular localization and promoter predictions were performed as well. Function related features such as protein- protein and RNA-RNA interactions were also predicted by ANNOgesic.

Table 3.1: Number of all detected genomic features

Genomic features	Classes	Numbers
Transcript		2,659
TSS	Total	1,688
	Intergenic	1,047
	Antisense	232
	Secondary	178
	Internal	338
	Orphan	134
Gene	Total	2,872
	Expressed	2,529
CDS	Total	2,778
	Expressed	2,433
PS		1,766
UTR	5' UTR	1,041
	3' UTR	869
Terminator		1,359
Operon		1,498
Promoter		1,547
sRNA	Total	257
	Intergenic	75
	Antisense	25
	5' UTR-derived	54
	3' UTR-derived	64
	InterCDS-derived	38
sORF		143
Riboswitch		15
RNA thermometer		7
CRISPR		1

Reference genome improvement

Reference sequence

The reference genome sequence of *S. aureus* HG003 was generated from *S. aureus* NCTC8325 which is the most closely related strain (the procedures are described in Chapter 2 - Methods and Materials). The genome sequence is composed of 2,821,354 base pairs which are 33.2% Adenines, 33.9% Thymines, 16.5% Cytosines, and 16.4% Guanines. In 2014, 19 sequence scaffolds of *S. aureus* HG003 were generated by using *de novo* transcript assembly [110]. In order to validate the genome sequence of *S. aureus* HG003 generated by ANNOgesic, a pairwise sequence alignment between the complete genome sequence and the 19 previously published scaffolds was performed. The result shows no significant difference between these two sequences (Appendix table A.1). Moreover, both sequences are repaired versions of the two mutations - *rsbU* and *tcaR* of *S. aureus* NCTC8325. Since the 19 previously published scaffolds contain some unknown base pairs and are not a complete genome sequence, the sequence generated by ANNOgesic is more reliable and contain more information.

SNP / mutation calling

For detecting SNPs and mutations, ANNOgesic integrates SAMtools [94] and BCFtools [94] which can identify the nucleotide differences between the high-throughput sequencing reads and the reference genome. If the genome sequence of the queried strain is not available, the module for detecting of SNPs and mutations

in ANNOgesic can also be used for generating the genome sequence. Since SNP detection is influenced by numerous factors like read depth and quality, ANNOgesic offers many parameters helping the users to remove false positives. The default settings for the comparison between the genome sequences of *S. aureus* HG003 and NCTC8325 are as follows: a minimum read depth is 140 (which means 10 reads per sample), minimum 140 mapped reads on variants are 140, a ratio between the reads mapped on variants and reference higher than 0.8, and minimum QUAL score of 40. Additionally, insertion and deletion need a ratio between total reads and the reads of insertion or deletion higher than 0.8. 32 nucleotide differences between *S. aureus* HG003 and NCTC832 were detected by applying ANNOgesic with those parameters (Figure 3.1). They were also confirmed by manual curation, and were used to generate the genome sequence of *S. aureus* HG003 (Figure 3.2).

Annotation transfer

ANNOgesic integrates RATT [95], which can transfer genome annotations from an annotated genome to an unannotated one by comparing the similarity of the genomes. Since the genome sequence and annotations of *S. aureus* NCTC8325 are available and the sequence identity between these two strains is higher than 99%, annotation transfer from strain NCTC8325 to HG003 can be precisely performed. In addition, *rsbU* and *tcaR* (two mutations of *S. aureus* NCTC8325) were added to the genome annotations of *S. aureus* HG003 manually. The genome annotations of *S. aureus* HG003 generated by ANNOgesic contain 2,872 genes, 2,778 proteins (1,534 hypothetical proteins), 61 tRNAs, and 16 rRNAs. 2,529 (88%) genes and

2,433 (88%) proteins are expressed (coverage $\geq$ 10).

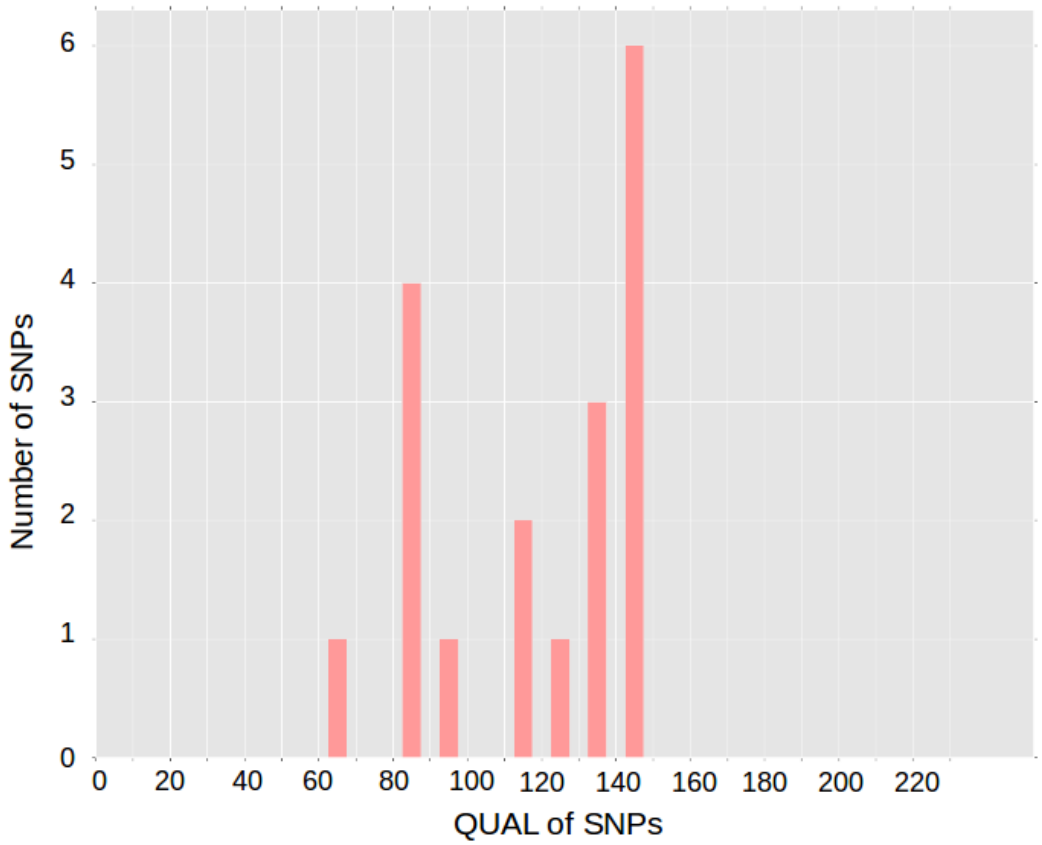
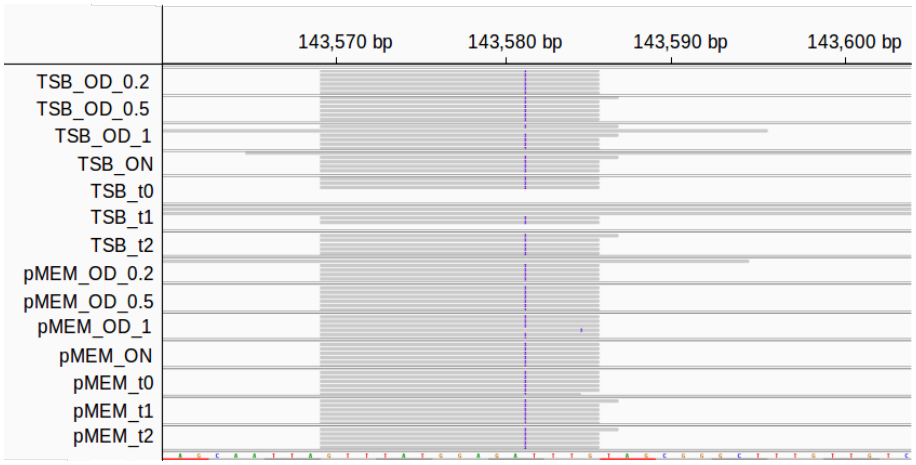
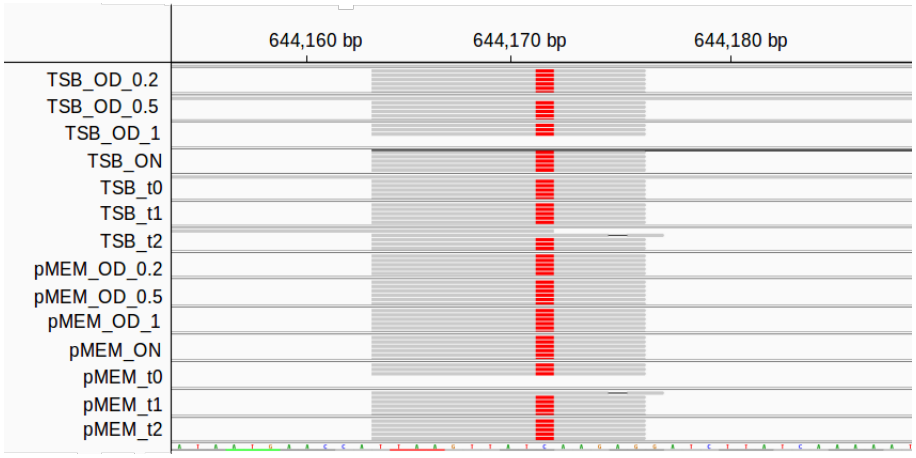


Figure 3.1: The distribution of SNPs and mutations between *S. aureus* HG003 and NCTC8325 based on QUAL scores. The minimum QUAL score is 40.

A


143,570 bp
143,580 bp
143,590 bp
143,600 bp
TSB_OD_0.2
TSB_OD_0.5
TSB_OD_1
TSB_ON
TSB_t0
TSB_t1
TSB_t2
pMEM_OD_0.2
pMEM_OD_0.5
pMEM_OD_1
pMEM_ON
pMEM_t0
pMEM_t1
pMEM_t2

B


644,160 bp
644,170 bp
644,180 bp
TSB_OD_0.2
TSB_OD_0.5
TSB_OD_1
TSB_ON
TSB_t0
TSB_t1
TSB_t2
pMEM_OD_0.2
pMEM_OD_0.5
pMEM_OD_1
pMEM_ON
pMEM_t0
pMEM_t1
pMEM_t2

C

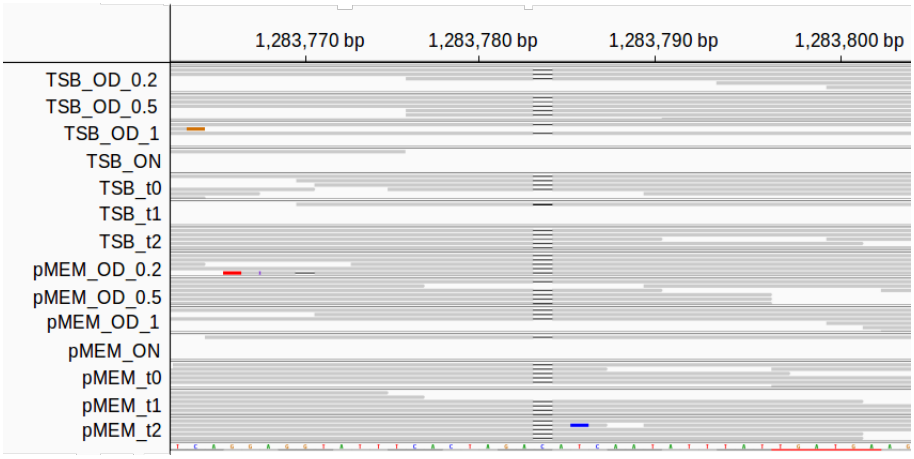


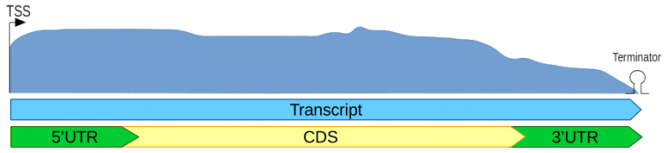
Figure 3.2: Examples of the nucleotide differences between *S. aureus* HG003 and NCTC 8325. The RNA-Seq reads are from *S. aureus* HG003, and the reference genome is *S. aureus* NCTC 8325. **(A)**: a insertion (shown by purple lines) at 143581 bp, **(B)**: a substution which represented by a read block (C to T) at 644172 bp, and **(C)**: a deletion (black lines) at 1283784 bp.

Transcripts

For the comprehensive understanding of the functions of transcripts, detecting the exact boundaries and sequences of the transcripts is crucial. For instance, UTRs may be the target of sRNAs or small molecules to perform post-transcriptional regulation or regulate the translation [32, 111], and numerous sRNAs may be found in UTRs as well [18, 35, 112, 113]. Without the information of transcript boundaries, UTRs may not be able to be detected. However, most of the available bacterial annotations only contain CDSs while information about TSSs, terminators and UTRs is not provided. In order to fill this gap, ANNOgesic provides the reliable information of transcript boundary based on RNA-Seq coverages and the predictions of TSSs, terminators,

and UTRs (Figure 3.3).

A



B

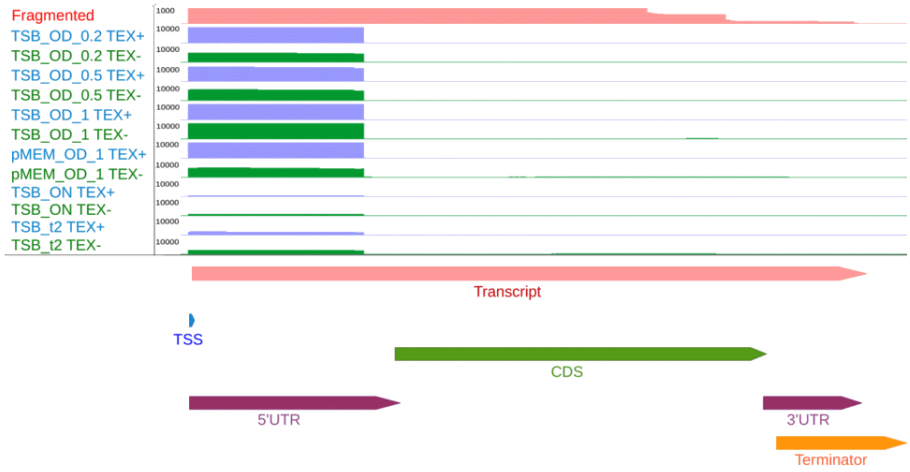


Figure 3.3: A schema and an example of transcript boundary. **(A)** ANNOgesic combines the information of TSSs, CDSs, terminators and UTRs to define transcript boundaries. **(B)** An example of transcript boundaries. The pink coverage, the blue coverages, and the green coverages represent fragmented library, TEX+ libraries of dRNA-Seq, and TEX- libraries of dRNA- Seq, respectively. Transcript, TSS, terminator, and CDS are represented by pink, blue, orange, and green bars, respectively. The transcript is from 800,959 to 801,322 bp at the forward strand.

RNA-Seq coverage-based transcript detection

In order to detect transcripts, numerous computational approaches have been developed. These tools can be classified by two types - *de nove* transcriptome assembly [6,7]

which can detect transcripts without genome sequences, and reference dependent transcriptome assembly [7, 29] which enables assembly of RNA-Seq reads based on genome sequence [28]. By applying the new method which was created and integrated into ANNOgesic, 2,659 transcripts were identified in *S. aureus* HG003. These transcripts cover 2,529 genes that show expression in at least one condition (Figure 3.3B).

TSS and PS predictions based on dRNA-Seq data

For the detection of transcripts and their boundaries, TSS is a crucial feature which may influence UTR, operon and promoter predictions. Differential RNA-Seq (dRNA-Seq) is a powerful RNA-Seq protocol which can detect TSSs in single nucleotide resolution [11]. Due to this, several tools for TSS prediction based on dRNA-Seq data were published such as TSSpredator [24], TSSer [26] and TSSAR [25]. In order to integrate the best tool for detecting TSSs into ANNOgesic, a comparison between these tools with default parameters was performed. The manually detected TSSs of whole genome (*S. aureus* HG003) were used as a benchmarking set for computing true positives and false positives. The result of the comparison shows that TSSpredator, which was integrated into ANNOgesic, is the most outstanding one (Table 3.2 and Figure 3.4).

Table 3.2: The comparison of TSS prediction tools

Methods	TP	FP	Missing
TSSpredator	1,032	2,460	92
TSSer	514	5,011	610
TSSAR	878	3,264	246

TP, FP and Missing represent true positives, false positives and TSSs not detected by the TSS prediction tool, respectively.

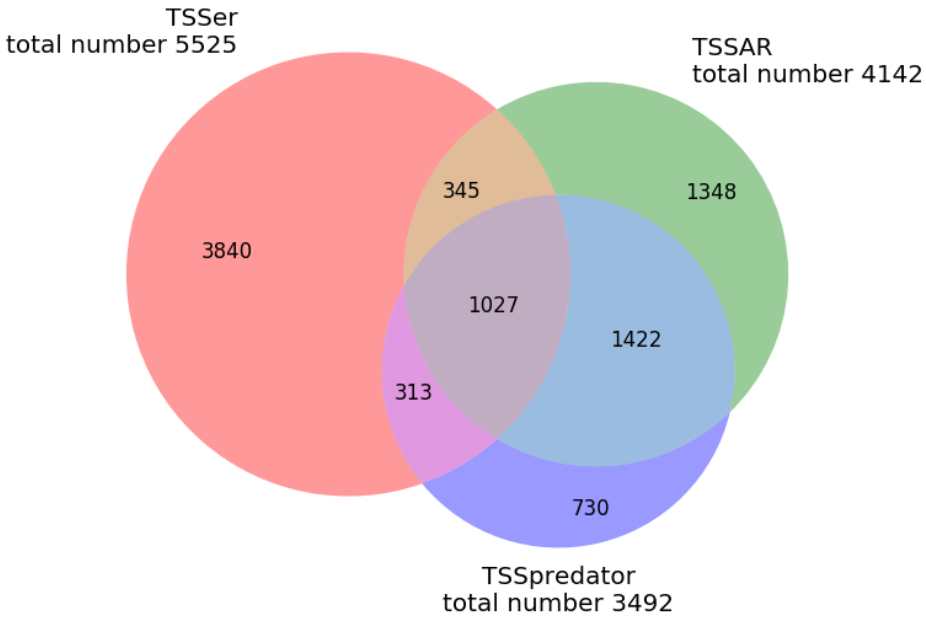


Figure 3.4: Venn diagram of comparing TSS prediction tools.

TSSpredator contains several parameters influencing the results of predictions significantly. In order to produce the precise annotations of TSSs, a parameter optimization method built on a subset of manually curated TSSs was created and integrated into ANNOgesic (details are described in chapter 2 - Methods and Materials). By using the optimized parameters to perform TSS prediction, a precise TSS sets were generated. For testing the optimization, three dRNA-Seq datasets from *S. aureus* HG003, *H. pylori* 26695 and *C. jejuni* 81116 were used. For *S. aureus* HG003. a manually curated TSS set of whole genome was available for the comparison. For the other two genomes, only small sets of manually curated TSSs (first 200 kb of genome sequence) were used. Moreover, the TSSs manually detected within first 200 to 400 kb were used as test sets. As displayed in Table 3.3, the optimization slightly improved the sensitivity for *H. pylori* 26695 (from 96.8% to 99.6%) and *S. aureus* HG003 (from 91.8% to 93.8%), while significantly raised the sensitivity for *C. jejuni* 81116 (from 67.1% to 98.7%) with similar specificity. Moreover, a comparison between TSSs and transcripts was performed. The amount of TSSs predicted by optimized parameters and located within transcripts was nearly the same as the TSSs detected by default parameters for *H. pylori* 26695 (83% for optimized parameters and 82% for default parameters), but slightly increased for *S. aureus* HG003 (99.6% for optimized parameters from 92.1% for default parameters) and even significantly raised for *C. jejuni* 81116 (96% for optimized parameters and 81% for default parameters).

Besides PSs represent the borders of transcripts, some transcripts undergo processing, which influences their biological activity. In addition, 3' UTR-derived sRNAs

Table 3.3: The comparison of the TSS predictions with optimized and default parameters

Strains	Parameters	Sensitivity (TP)	Specificity (FP)	Missing
S. aureus HG003	Default	91.8% (1,032)	99.91% (2,460)	92
	Optimization	93.8% (1,054)	99.98% (564)	70
H. pylori 26695	Default	96.8% (244)	99.98% (32)	8
	Optimization	99.6% (251)	99.98% (32)	1
C. jejuni 81116	Default	67.1% (104)	99.98% (31)	51
	Optimization	98.7% (153)	99.99% (7)	2

The percentages are of the sensitivity or specificity. The numbers in brackets indicate true positives or false positives. The optimization was tested for whole genome in *S. aureus* HG003, and for 200kb in *H. pylori* 26695 and *C. jejuni* 81116.

may be generated by internal processing [34, 35]. Actually, TSSpredator can not only be used for detecting TSSs, but also for identifying PSs by searching for the reverse enrichment pattern (relative enrichment in TEX- libraries). In order to improve the prediction, parameter optimization was performed as well. As done for the optimization of TSS prediction, the manually annotated PSs in the first 200 kb of the genomes were used as a training set, and the manually curated PSs from the first 200 to 400 kb were used as a test set. The performances of the predictions with default and optimized parameters were similar in *H. pylori* 26695, but had a significant improvement in *C. jejuni* 81116. In *S. aureus* HG003, around 100 false positives were removed via optimization (Table 3.4).

Table 3.4: The comparison of the optimized and default parameters of TSSpredator for PS prediction

Strains	Parameters	Sensitivity (TP)	Specificity (FP)	Missing
S. aureus HG003	Default	100% (82)	99.96% (143)	0
	Optimization	100% (82)	99.99% (11)	0
H. pylori 26695	Default	92.9% (26)	99.99% (7)	2
	Optimization	92.9% (26)	99.99% (7)	2
C. jejuni 81116	Default	61.3% (19)	99.99% (2)	12
	Optimization	93.5% (29)	99.99% (6)	2

The percentages are of the sensitivity or specificity. The numbers in brackets indicate true positives or false positives.

Based on TSS and PS predictions with optimized parameters, the candidates of TSSs and PSs may be annotated globally and precisely. For *S. aureus* HG003, 1,766 PSs and 1,688 TSSs consisting of 1,047 primary, 178 secondary, 338 internal, 232 antisense and 134 orphan TSSs were detected. Additionally, a Venn diagram of different TSS classes was generated by ANNOgesic automatically (Figure 3.5).

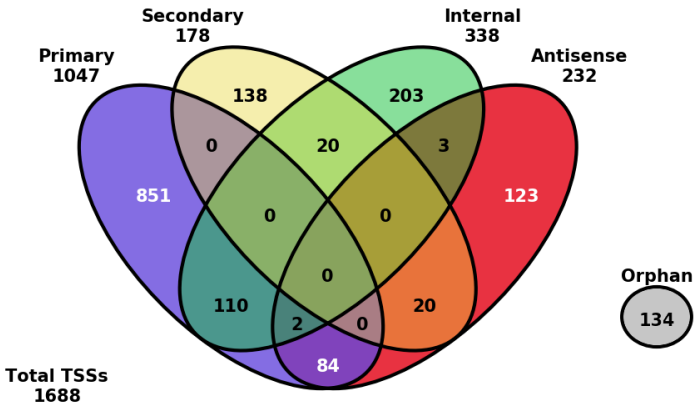


Figure 3.5: The distribution of TSS classes of *S. aureus* HG003.

Terminators

For the detection of transcript boundaries, TSS is an important feature for identifying the transcript border in the 5' end. However, the 3' end of a transcript is usually not so clear and sharp. Due to this, the data from RNA-Seq generated after transcript fragmentation and the information of terminators which are clear landmarks of the transcript borders in the 3' ends are required.

Terminators can be separated into two types based on the dependence of Rho factor involved in the termination of transcription. Rho factor is a hexameric-ring-shaped protein that binds to the pause site of the terminator (C-rich/G-poor region after ORF) to terminate the transcription [114]. In *Escherichia coli* strains, Rho factor is an essential protein to regulate the transcription. However, it is non-essential in certain bacteria, like the main target species of this study, *S. aureus* [115]. A Rho-independent terminator is normally composed of a stable CG-rich stem-loop (7-20 base pairs). The stem-loop can bind tightly to *NusA*, which is bound to an RNA polymerase to stall the transcription [116, 117]. Numerous Rho-independent terminator prediction tools are built based on the specific secondary structures [96, 118].

TransTermHP [96] and RNIE [118] are two representative tools for the prediction of Rho-independent terminators based on genome sequences. In order to integrate the best tool into ANNOgesic, the comparison between TransTermHP and RNIE was performed for the genome sequence of *S. aureus* HG003. As shown in Figure 3.6, TransTermHP detected more candidates than RNIE which only identified 137

Rho-independent terminators of which 80% were also found by TransTermHP. In order to put these number into perspective, one has to consider that in principle each operon should be associated with a terminator and 1,498 operons were found in *S. aureus* HG003 (see the section – Operon). This indicates that RNIE contains many false negatives. Thus, TransTermHP was integrated into ANNOgesic. However, the candidates of terminators generated from TransTermHP are not always supported by RNA-Seq data because several terminators may only function in specific conditions (Figure 3.7E and F). In order to improve the prediction, two further novel approaches based on RNA-Seq data and the given genome annotations were developed and integrated into ANNOgesic.

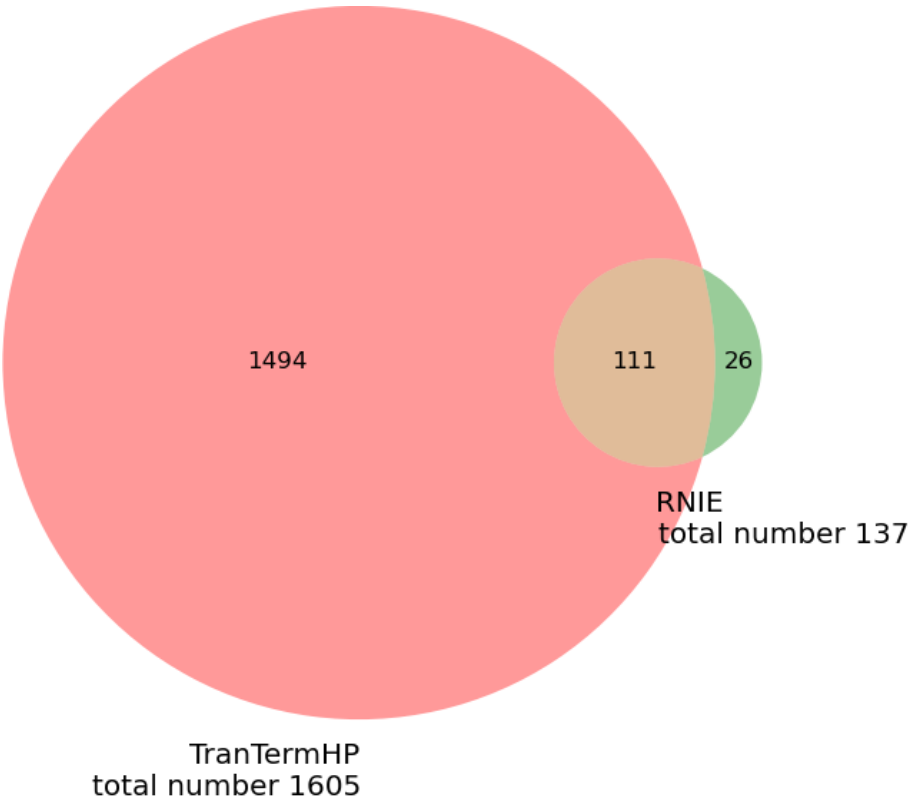


Figure 3.6: The comparison of Rho-independent terminator prediction tools by RNIE and TransTermHP based on *S. aureus* HG003 genome.

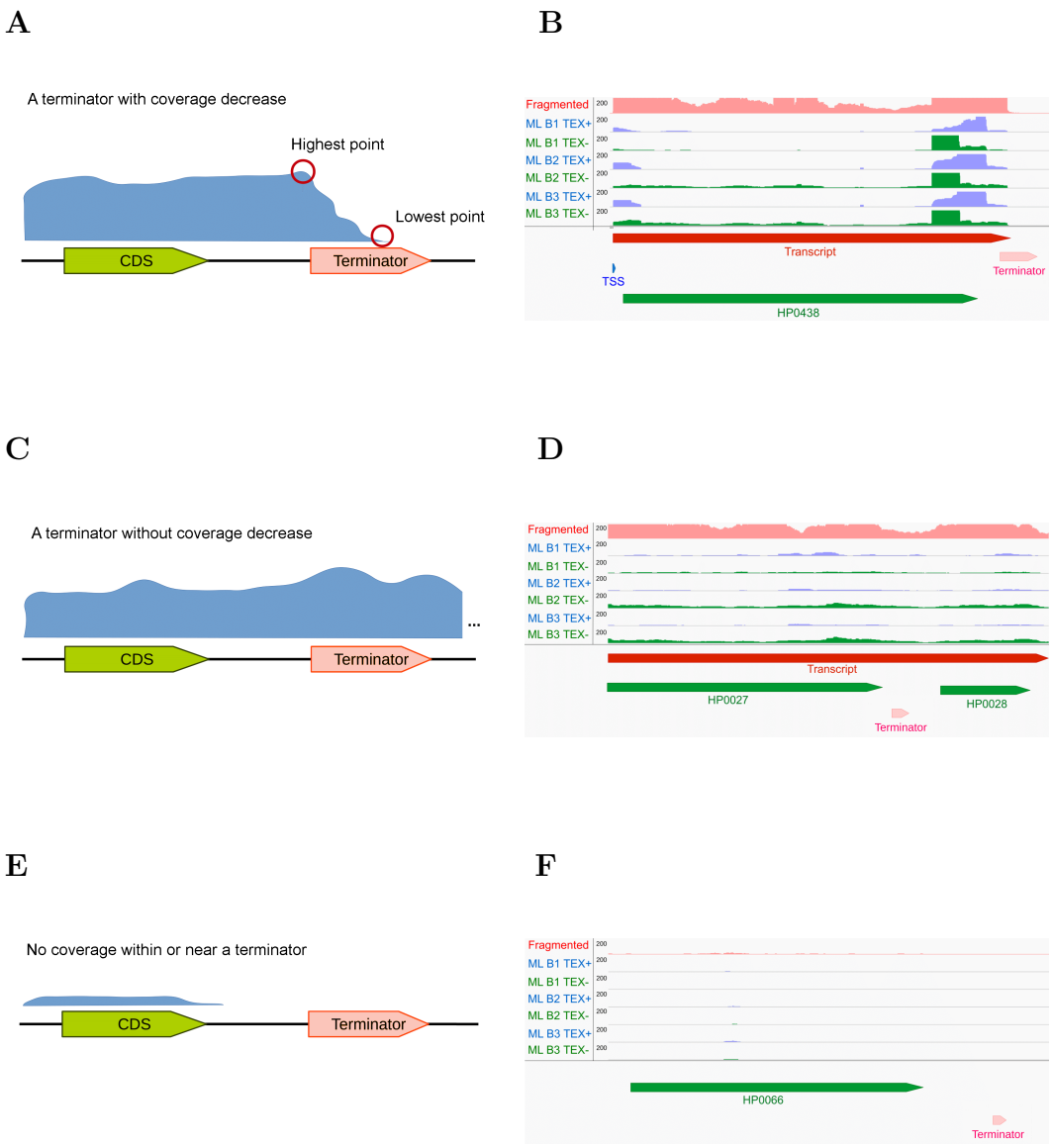


Figure 3.7: The method and an example for detecting coverage decrease of terminators. **(A)** and **(B)** represent a high-confidence terminator which shows a significant drop of coverage. **(C)** and **(D)** show a terminator which has no a significant decrease of coverage. A terminator without showing expression is shown in **(E)** and **(F)**. In **(B)**, **(D)**, **(F)**, the coverage of RNA-Seq generated after transcript fragmentation, TEX+ and TEX- of dRNA-Seq are represented as pink, blue and green coverages, respectively. In the annotation track, terminators, TSSs, CDSs and transcripts are showed as pink, blue, green, and red bars, respectively.

The first new approach is for increasing the sensitivity of terminator prediction. Since secondary structure of Rho-independent terminator is an important feature, RNAfold was applied to check the secondary structure of the intergenic region between the two converging genes in order to find the potential terminators [48, 52] (Figure 3.8). In case the region forms a stem-loop and the tails of stem-loop are A/T rich region, it is considered as a Rho-independent terminator. In the prediction default setting, the maximum nucleotides of the potential terminator region are 80, T rich tail of the 3' end contains more than 5 Thymines, the stem-loop contains 4 to 20 nucleotides (75% nucleotides is able to make pairs), and the length of the loop is between 3 to 10 nucleotides.

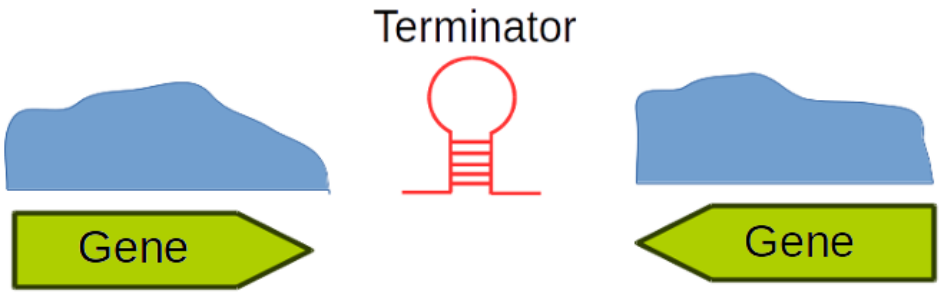


Figure 3.8: Detecting Rho-independent terminator based on convergent gene pairs. The blue curve-blocks, green arrows and red stem-loop represent the read coverages, two convergent genes, and a potential Rho-independent terminator, respectively.

A general observation was that regions of terminators show a sharp decrease of coverage. Based on this, a second approach for the detection of terminators was developed. For that location with a significant decrease of coverage in the 3'end of a feature is searched in order to find high-confidence candidates for terminators (Figure

3.7A and B). On the other hands, if the terminator candidates lack expression or express without the drop of coverage, they might be false positives or not functional terminators for the selected conditions (Figure 3.7C - F). By default setting, the sharp coverage drops are located within the region of the terminator candidate, or within 30 nucleotides upstream and downstream from the potential terminator. Moreover, the minimum ratio of the lowest and highest read coverage value must be 0.5 or more.

In *S. aureus* HG003, the number of Rho-independent terminators detected by TransTermHP is 1,525, and by the approach of checking secondary structures of the intergenic regions between convergent genes is 524. However only 1,031 (68%) terminators from TransTermHP and 421 (80%) terminators from convergent gene based approach contain a significant coverage decrease. 270 terminators were detected by using both methods, and 248 of them contain significant coverage drops. Overall, 1,779 Rho-independent terminators were identified in *S. aureus* HG003, and 1,181 of them are high-confidence terminators (with a significant coverage decrease).

UTRs

UTR is considered as an essential feature for understanding the RNA-RNA interaction and the regulation of genes since numerous important sequences are located in UTRs such as riboswitches, RNA thermometers and ribosome binding sites [119, 120]. Additionally, UTR-derived sRNAs are discovered recently [18, 34, 35]. Despite this high importance, the available tools for detecting UTRs are still few, and all of the current tools are only based on genome sequences [121].

Since transcript boundaries and CDSs can be identified by using ANNOgesic, a comparison of the positions of CDSs, TSSs, terminators, and transcripts was performed for detecting 5' UTRs and 3' UTRs. The region between a TSS and the following downstream CDS is a 5' UTR; in addition, the sequence between a terminator or the 3' end of a transcript and the last upstream CDS is a 3' UTR. ANNOgesic detected 1,041 5' UTRs and 869 3' UTR in *S. aureus* HG003. The distribution for UTRs was shown in Figure 3.9.

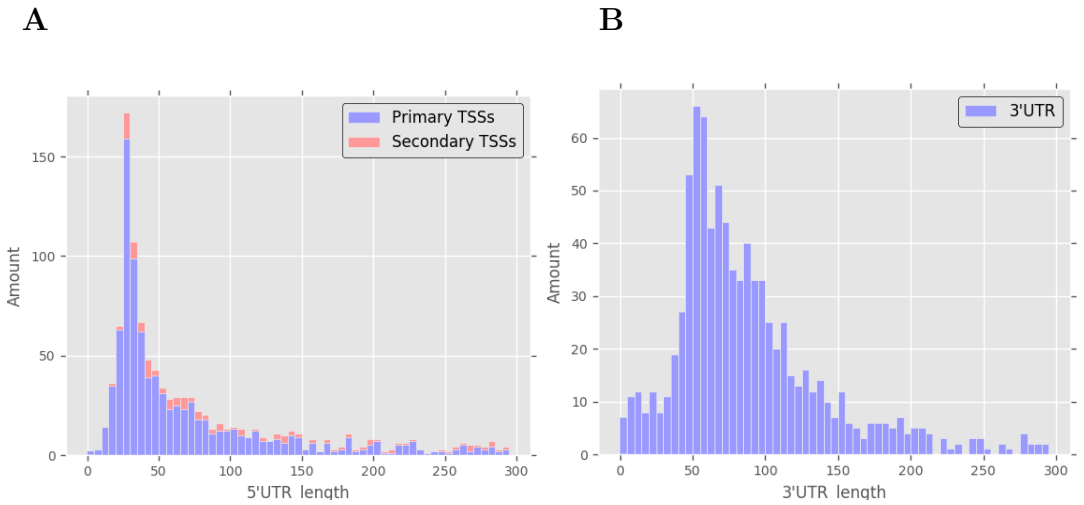


Figure 3.9: The distribution of UTR lengths for *S. aureus* HG003 was generated by ANNOgesic. **(A)** is 5' UTRs and **(B)** is 3' UTRs.

Promoters

Promoters are located upstream of genes and can be bound by transcription factors and RNA polymerases. In bacteria, the most common promoters are two short consensus sequences located around 10 (Pribnow Box) and 35 nucleotides upstream from TSSs (Figure 3.10). These promoters can specifically interact with RNA

polymerase via sigma factor (σ^{70}) which is a transcription initiation factor [122, 123]. Therefore, the detection of promoters is an important step to understand the mechanism of transcription factor interaction and the regulation of transcription.

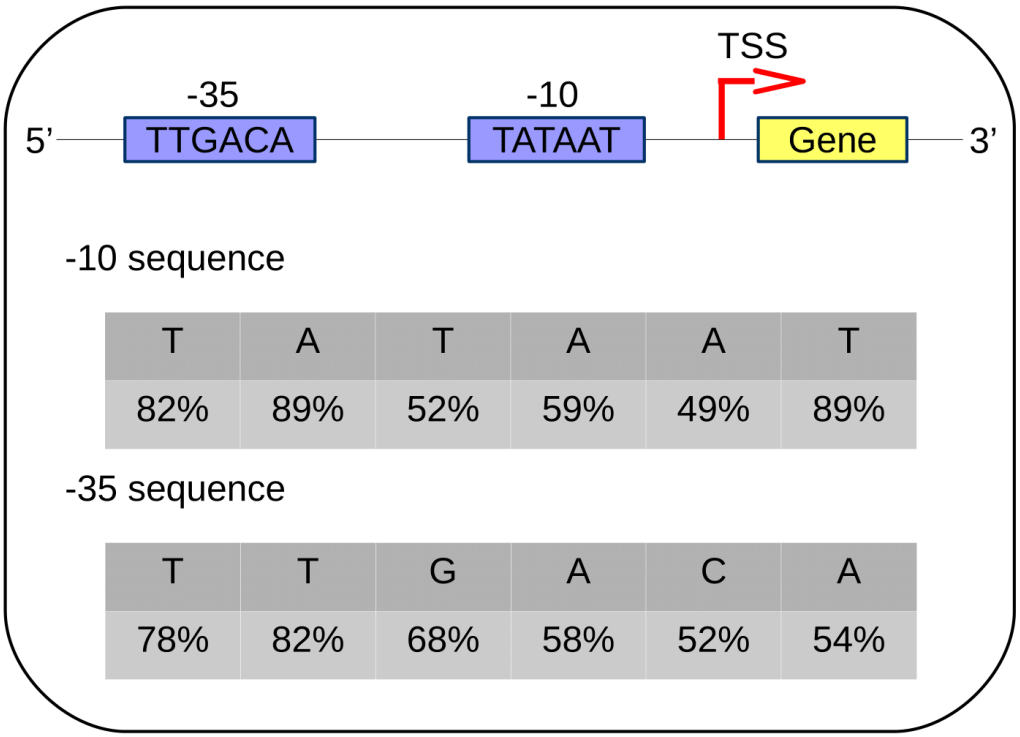


Figure 3.10: The probability for occurrence of nucleotides in promoter sequences in *E. coli*. 10 and 35 nucleotides upstream from TSS are two consensus promoter sequences [123].

For detecting promoter motifs, ANNOgesic integrates MEME [97] (which can identify ungapped motifs) and GLAM2 [98] (which is able to discover gapped motifs). These two tools not only detect the promoter candidates with the information of the corresponding sequences, but also generate the figures of the sequence motifs. In the default setting, 50 nucleotides upstream from TSS were used for searching promoter motifs, and the length of the promoter is set by 45, 50 and 2-10 nucleotides. In *S.*

aureus HG003, 1,547 Pribnow Boxes were found from 45 nucleotides upstream of TSSs (Figure 3.11).

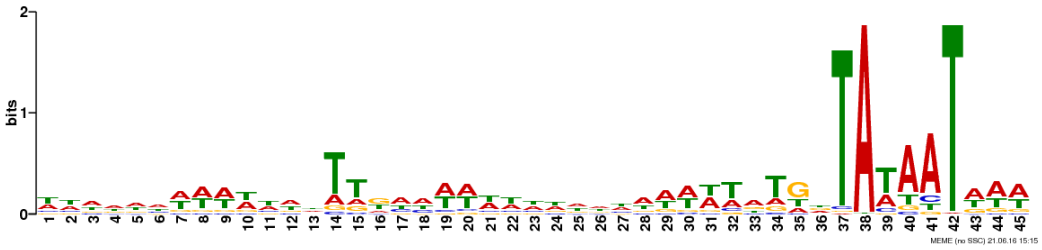


Figure 3.11: The Pribnow Box detected in the upstream sequences of 1,547 TSSs (92%) in *Staphylococcus aureus* HG003.

Operons

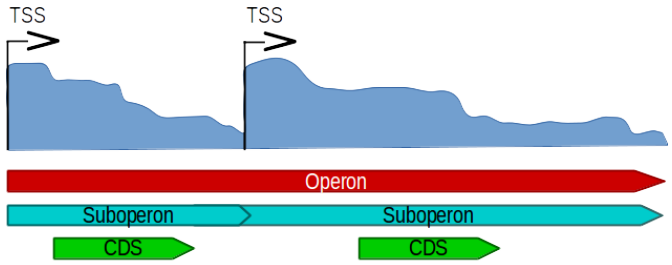
An operon is a functional unit containing the genes regulated by the same transcription factor and promoter. The cluster of genes are transcribed together and might have related functions. However, only few computational tools are available for detecting such feature. ProOpDB [124] is a representative tool that integrates data from KEGG [55], COG [125], Pfam [126], and STRING [127] to store and detect prokaryotic operons via machine learning approaches (neural network). An operon prediction tool based on RNA-Seq data was not existed so far.

Since all requirements underlying features – TSSs, CDSs, and transcripts can be predicted by ANNOgesic, operons as well as sub-operons associated with different TSSs in the same operon can be detected by it as well (Figure 3.12A). As part of that, ANNOgesic classified the operons to monocistronic operons (operons contain only single genes) and polycistronic operons (operons consist of multiple genes).

For *S. aureus* HG003, 2,659 transcripts composed of 1,027 monocistronic operons,

472 polycistronic operons, and 1,160 transcripts which are not associated with genes were detected. Additionally, within these operons, only 47 of them contain sub-operons (Figure 3.12B).

A



B

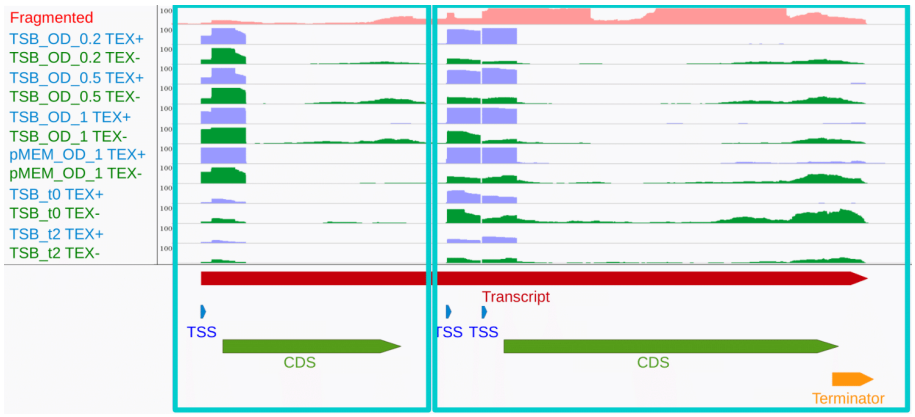


Figure 3.12: A schema and an example of operon and sub-operon detection. **(A)** Sub-operons were detected by searching for multiple TSSs located in the intergenic region of an operon. **(B)** An example of an operon with sub-operons in *S. aureus* HG003. The operon is from 1,874,426 to 1,876,261 at the forward strand. The pink, green, and blue coverage represent transcript fragmented library, TEX- and TEX+ libraries of dRNA-seq. In the annotation track, the blue spots, orange bar, pink bars and green bars represent TSSs, terminator, operon/transcript, and CDSs. The two CDSs are located in the same operon, but in different sub-operons (two hollow light blue squares).

sRNAs

Detection of sRNAs

In order to detect these sRNAs based on the genome sequence, numerous tools were developed. ANNOgesic offers a novel RNA-Seq-based method which is different from most of the available tools that use only genome sequence to detect and classify different types of sRNAs [36–42].

For detecting sRNAs, ANNOgesic extracts short expressed non-annotated transcripts (the default setting: 30 - 500 nucleotides long) as the potential sRNAs. If the length of a non-annotated transcript is longer than the length threshold (given by the users), the information of the read coverage is used for checking a significant drop of coverage in order to define the border of 3' end (similar to Figure 3.7A). If a potential sRNA does not overlap with any CDSs in both the forward and reverse strands, it is considered as an intergenic sRNA. However, if CDSs exist in the complementary strand of the potential sRNA, it is marked as an antisense sRNA. For the detection of UTR-derived sRNAs, a novel method based on the information of transcripts, TSSs, and PSs was developed. A 5' UTR-derived sRNA should start with a TSS or a PS as well as show a PS or a point containing significant coverage decrease in the 3' end. The detection sRNAs located in interCDS (the region between two consecutive CDSs within the same transcript) is based on searching a TSS or a PS in the 5' end and a sharp coverage drop or a PS in the 3' end. For 3' UTR- derived sRNAs, they must start either with a TSS or a PS and end with the transcript or at a PS. The

identification and classification of sRNAs are illustrated in Figure 3.13.

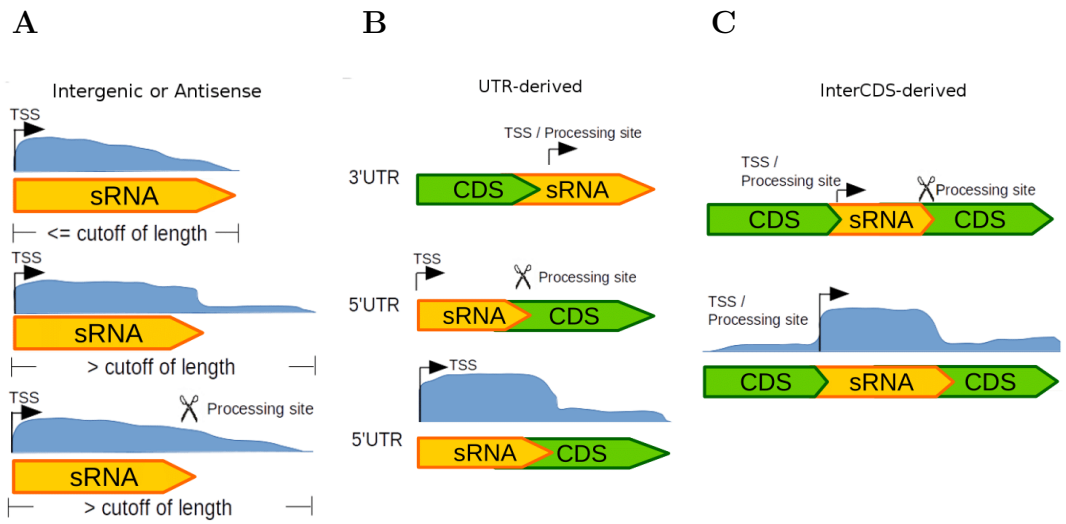


Figure 3.13: Detection of intergenic, antisense, and UTR-derived sRNAs. **(A)** Detection of intergenic and antisense sRNAs. The upper panel shows a normal case, a non-annotated transcript starts with a TSS, and is within the normal length of sRNA. The middle panel shows a TSS-associated transcript which is longer than the length threshold. In this case, the coverage (blue region) is used for searching the significant coverage drops. The bottom panel shows a similar case to the middle panel, but a PS is detected in the 3' end of the sRNA. **(B)** Identification of UTR-derived sRNAs. For 3' UTR-derived sRNA, if the transcript starts either with a TSS or a PS, it is marked as a 3' UTR-derived sRNA. For 5' UTR-derived sRNA, if the transcript starts with a TSS and shows a coverage which significant drops or a PS in the 3' end, it is considered as a 5' UTR-derived sRNA. **(C)** Detection of interCDS-derived sRNAs. It is similar to the detection of 5' UTR-derived sRNA, but the transcript can start with a PS as well.

In order to remove the false positives, several filters were applied and integrated into ANNOgesic. First of all, if homologous sequences of a sRNA candidate were found in sRNA databases based on a BLAST+ search [128], it is marked as a known sRNA. In *S. aureus* HG003, two sRNA databases were used - *i*) BSRD [129] which stores experimentally confirmed sRNAs of all bacterial species and *ii*) experimentally validated sRNAs in SRD [64] which only stores sRNAs of *S. aureus* from both experimental and computational identifications. If a sRNA candidate does not have homologous sequences in sRNA databases, it needs to pass the following filters, otherwise it is considered as a false positive. For excluding the potential protein-coding sequences, a BLAST+ [128] search in the NCBI non-redundant protein database was performed. If a potential sRNA got a hit, it is tagged as a potential protein-coding sequence, and removed from the list of sRNA candidates. After excluding potential protein-coding sequences, the remaining sRNA candidates which start with a TSS and form a stable secondary structure (folding free energy change normalized by length should be smaller than 0.05 $\vec{\Delta}$G/nt) are included in the final sRNA set. By using ANNOgesic and applying all the filters, 256 sRNAs which consist of 75 intergenic sRNAs, 25 antisense sRNAs, 54 5' UTR-derived sRNAs, 64 3' UTR-derived sRNAs and 38 interCDS-derived sRNAs were identified in *S. aureus* HG003 (Figure 3.14). Moreover, this set of sRNAs is composed of 62 known sRNAs and 194 novel sRNAs (Table 3.5).

Table 3.5: Previously published sRNAs which were detected in *S. aureus* HG003

sRNA name	Amount	sRNA name	Amount	sRNA name	Amount
RsaOT	1	SsrA	2	Teg45	1
SbrC/RsaC/RsaOW2	2	RsaOB	1	RsaA	2
SprA2/RsaJ	1	RsaOW2	2	Sau-6053	1
RsaOR/SprX	2	RsaG	1	RsaE	1
RsaOI/Sau-6477	1	SprG2	1	RsaOU	1
RsaD	1	RsaOM	1	RsaX25	1
RsaOL	1	Sau-5949	2	RsaOC	1
Sau-19	1	SprD	1	Sau-02/SprF2	1
RsaOQ	1	SprB	1	RNAIII	1
SbrC/RsaC	1	SprA2	4	SprC	1
SprA/SprA1	2	SprF2	1	SbrC	1
RNaseP-bact-a	1	SbrB	1	RsaK	1
Teg1	1	RsaOR	2	RsaH	2
SprF4	1	Teg70	1	Teg76	1
Teg35	1	Sau-63	1	RsaOG	1
Sau-5971	1	SprF3/SprG3	1	RsaOV	1
RsaOE	1				

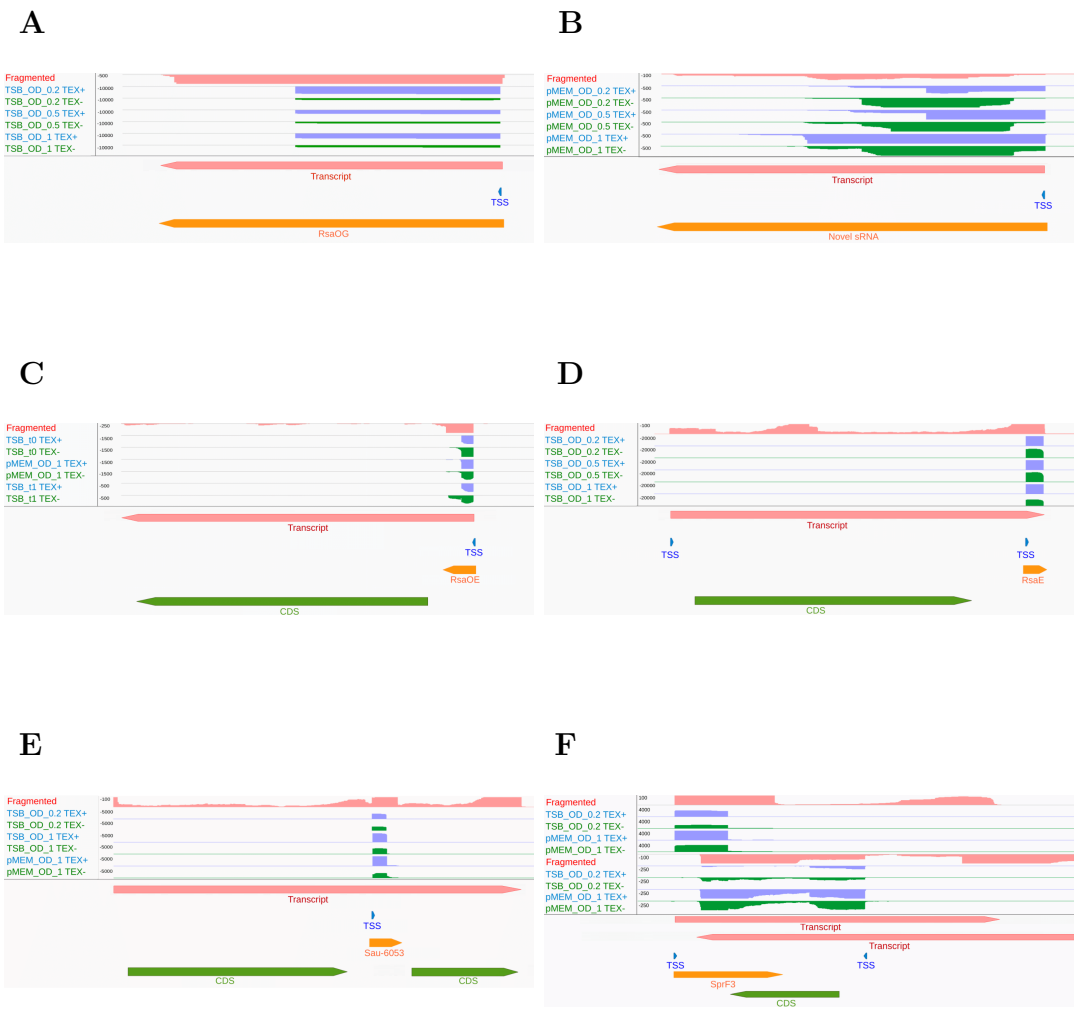


Figure 3.14: Examples of sRNAs *S. aureus* HG003. Red, blue, and green coverages represent the library of fragmented transcripts from RNA-Seq, TEX+ and TEX- libraries from dRNA-Seq, respectively. In the annotation tracks, red, blue, green, and orange bars represent transcripts, TSSs, CDSs and a sRNA, respectively. **(A)** and **(B)** are for intergenic sRNAs. **(A)** A known sRNA at the region between 2,377,278 to 2,377,456 at the reverse strand, and **(B)** a novel sRNA located from 1,801,971 to 1,802,267 at the reverse strand. **(C)** An example of 5' UTR-derived sRNA located from 1,791,238 to 1,791,456 at the reverse strand. **(D)** A 3' UTR-derived sRNA found at the location from 911,380 to 911,494 at the forward strand. **(E)** An interCDS-derived sRNA detected from 931,870 to 932,058 at the forward strand of the genome. An example of antisense sRNA shown in **(F)** was discovered between 2,211,957 to 2,212,213 at the forward strand.

ANNOgesic also provides numerous functions and visualizations for analyzing the predicted sRNAs, such as comparing sRNAs with terminators, sORFs, and promoters as well as generating secondary structural figures, dot plots, and mountain plots by using Vienna RNA package [48] (Figure 3.15).

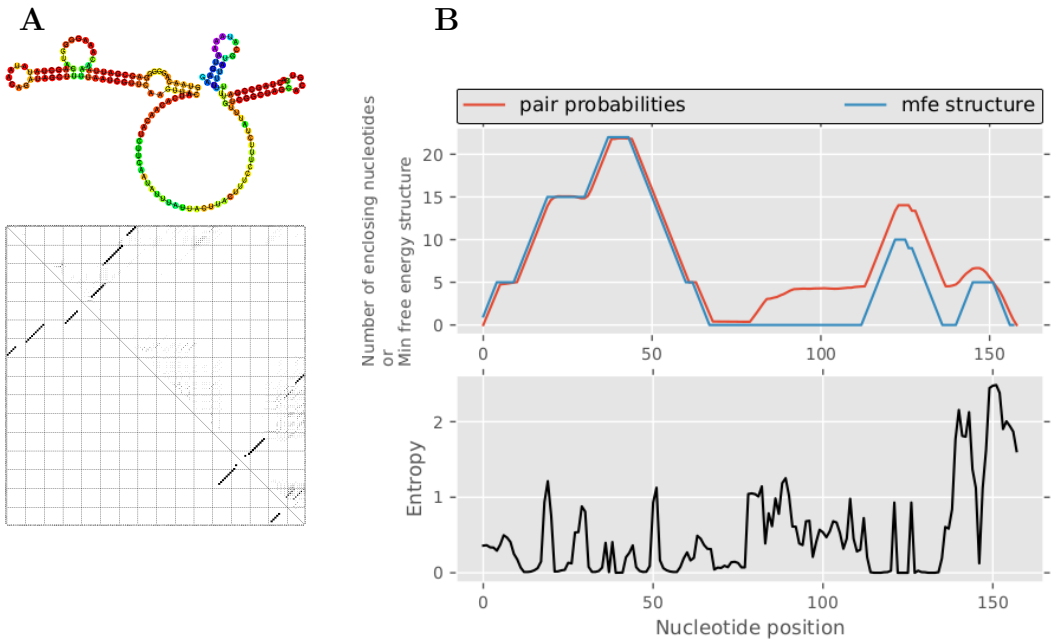


Figure 3.15: An example (RsaOG) of sRNA secondary structure analysis. **(A)** The potential secondary structure of RsaOG (upper panel), and the plot of the secondary structure (bottom panel). **(B)** The mountain plot (upper panel) and the entropy plot (bottom panel) of RsaOG. The folding stability and the possibility were revealed in the these figures generated by using RNAfold of Vienna RNA package [48].

Ranking of sRNA candidates

In order to analyze the performance of ANNOgesic capability for prediction of sRNA, it was applied to four bacterial genomes - *Helicobacter pylori* 26695, *Campylobacter jejuni* 81116, *Staphylococcus aureus* HG003, and *Escherichia coli* K-12. The previously described sRNA sets of these four genomes were selected as benchmarking sets.

Those sets were taken from two dRNA-Seq based publications for *Helicobacter pylori* 26695 and *Campylobacter jejuni* 81116 [24, 60], from a RefSeq annotation file for *Escherichia coli* K-12 in RefSeq, and from a microarray based study for *Staphylococcus aureus* HG001 which has only one gene difference to *S. aureus* HG003 [130]. Some of sRNAs of the benchmarking sets were removed since they overlap with CDSs or are not expressed in the chosen conditions. Based on the comparison of the predicted sRNA sets and the benchmarking sets, around 80% to 90% of previously reported sRNAs in these four bacterial genomes were detected by applying ANNOgesic (Table 3.6).

Table 3.6: The sensitivity of the sRNA detection in ANNOgesic

Strains	Sensitivity (TP)	Total sRNAs
H. pylori 26695*	90% (53)	59
C. jejuni 81116	84% (26)	31
S. aureus HG003**	80% (28)	35
E. coli K-12	86% (50)	58

*The RNA-Seq data of *H. pylori* 26695 did not include 454 Sequencing data which was used for several conditions in the publication from Sharma *et al.* [60]. Thus, some of the described sRNAs were not considered in this study.

**The benchmarking set of *S. aureus* HG003 is from *S. aureus* HG001

In order to check the resolution of the sRNA detection in ANNOgesic, the comparison between the locations of benchmarking sets and predicted sRNA sets was performed. As displayed in Figure 3.16, almost all of the differences of the positions in the 5' end of sRNAs are less than 10 nts because TSSs are precisely detected by using dRNA-Seq data. However, the resolution of the 3' end in the strains except *E. coli* K-12 is low since no RNA-Seq protocol was applied for specifically detecting

terminators in this study. Although the resolution of sRNA detection in the 3' end is not such high as the resolution in the 5' end, the majority of sRNA location differences are still less than 50 nts. The results of *S. aureus* HG003 are worst because the benchmarking sRNAs are from *S. aureus* HG001 which is not exactly the same as *S. aureus* HG003 and underlying data is from microarray which has lower resolution. For improving the resolution of sRNA detection, the data from a RNA-Seq protocol which can identify terminators in high-resolution like Term-Seq [13] is required.

Although ANNOgesic provides potential sRNAs, the selection of the reliable sRNA candidates for experimental validation is still an issue. In order to address this, a ranking system based on the average of the read coverages of sRNA candidates and the promoter information was developed. Equation 3.1 shows the scoring function of the ranking system. In case the assigned promoters (default is Pribnow box) are found upstream of the sRNA, the score of the sRNA is the average of the read coverages multiplied by 2 (default setting). Otherwise, the score is the average coverage value. In Figure 3.17, the previously published sRNAs show higher scores (ranking in the front). The p-values of the t-test between the previously published sRNAs and the rest predicted ones are 1.631e-09, 4.629e-04, 6.606e-07 and 2.528e-13 for *H. pylori* 26695, *C. jejuni* 81116, *S. aureus* HG003 and *E. coli* K-12, respectively (Figure 3.17). The results of these analyses showed that the ranking system is a good indicator for the reliability of the sRNA prediction.

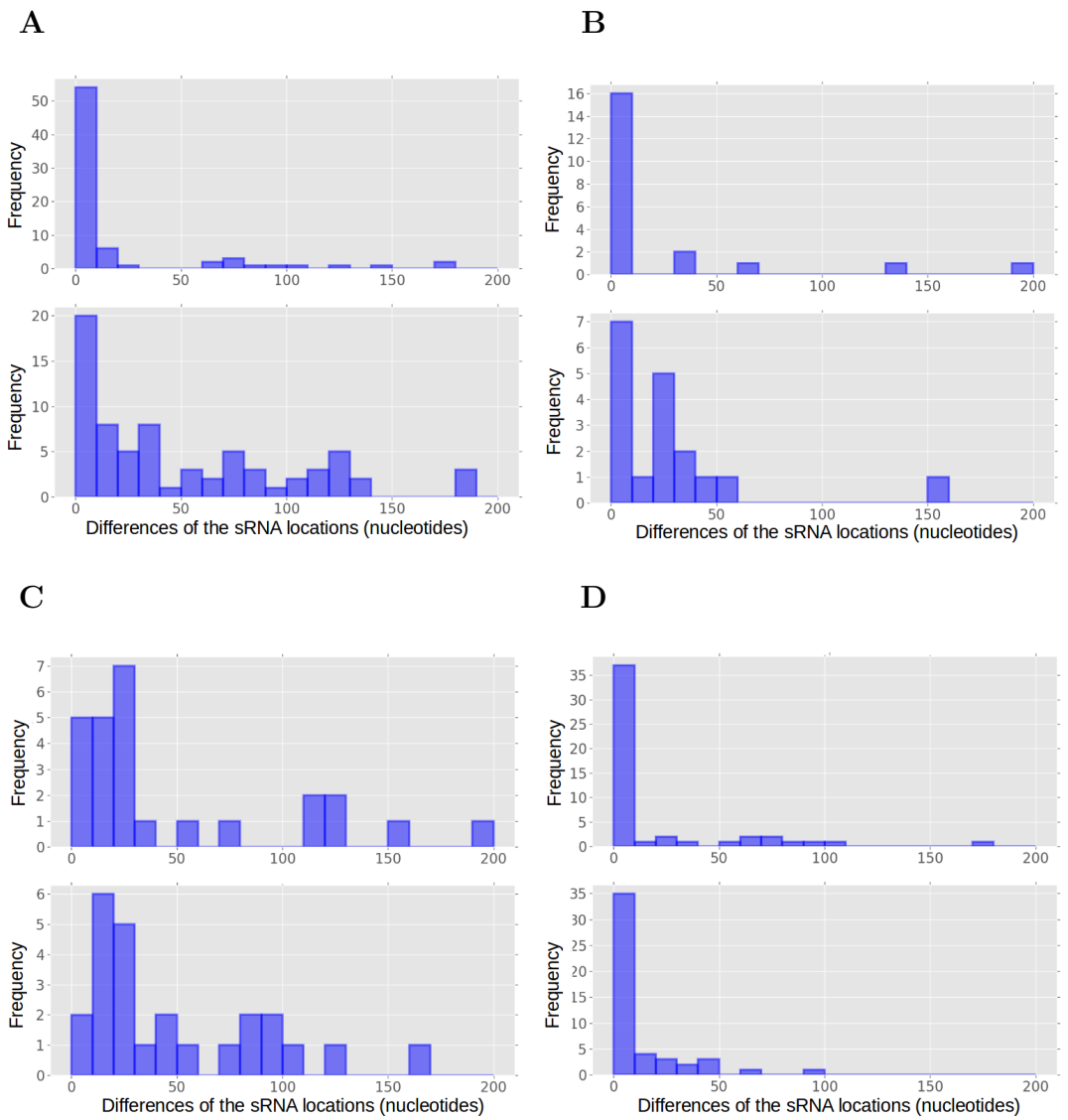


Figure 3.16: The resolution of sRNA detection in ANNOgesic. The figures show the results of the comparison between the positions of the benchmarking sets and the predicted set. The differences of the sRNA positions between the benchmarking sets in the 5' end is presented in each upper panel, and the comparison of the sRNA positions for the 3' end is shown in each bottom panel. **(A)**, **(B)**, **(C)** and **(D)** present the resolution of sRNA detection for *H. pylori* 26695, *C. jejuni* 81116, *S. aureus* HG003, and *E. coli* K-12, respectively. All of them show high-resolution in the 5' end, but only *E. coli* K-12 shows high resolution in the 3' end.

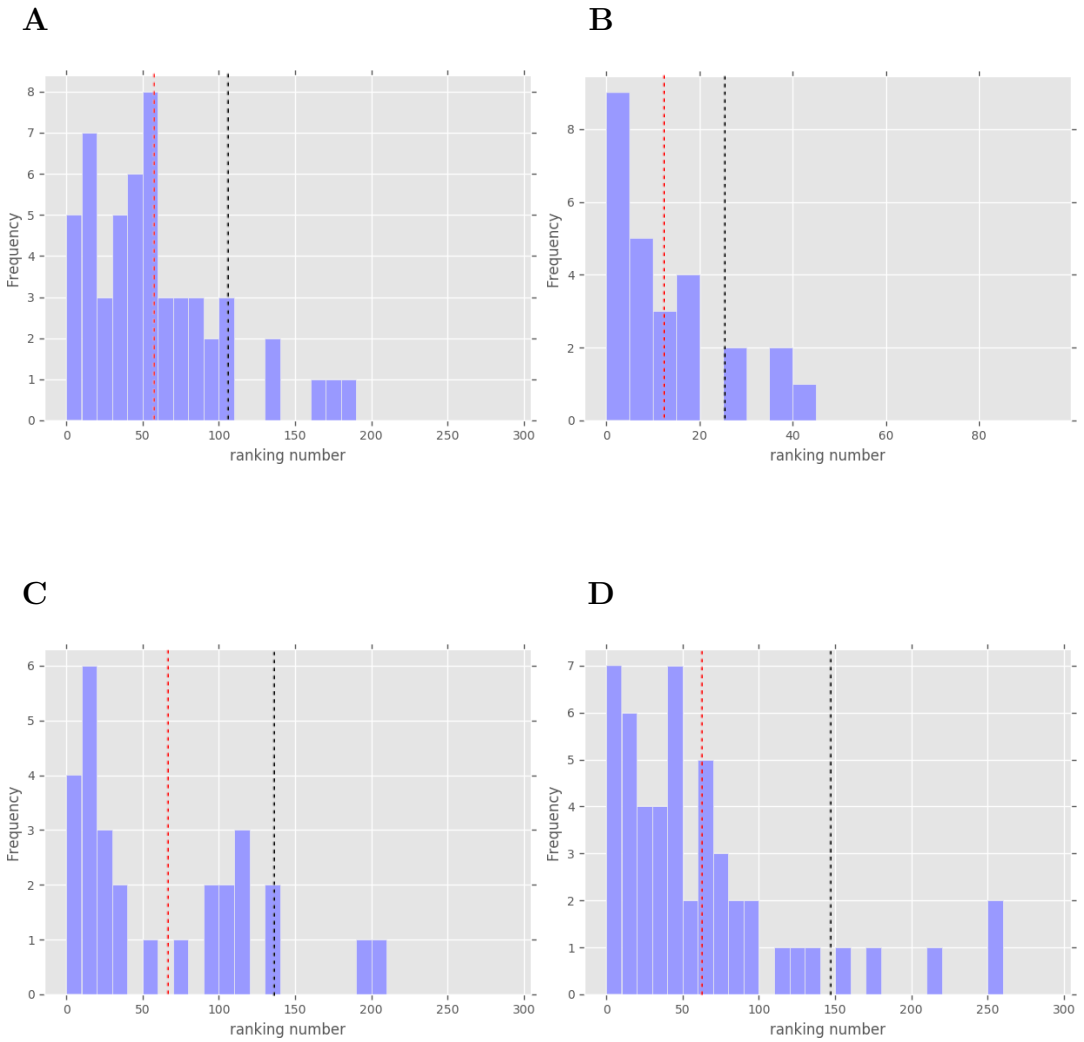


Figure 3.17: Distribution of the ranking of benchmarking sRNAs. The red dash lines show the average ranking numbers of previously reported sRNAs (57.25, 13.19, 63.32 and 61.76 for *H. pylori* 26695, *C. jejuni* 81116, *S. aureus* HG003, and *E. coli* K-12, respectively), and the black dash lines represent the average ranking numbers of the rest sRNA populations (106.17, 25.05, 136.25 and 147.41 for *H. pylori* 26695, *C. jejuni* 81116, *S. aureus* HG003, and *E. coli* K-12, respectively). The p-values of the t-test show that the ranking scores between the benchmarking sets and rest populations are significantly different. **(A)** is the histogram of *H. pylori* 26695, **(B)** presents the histogram of *C. jejuni* 81116, the histogram of *S. aureus* HG003 is shown in **(C)**, and the histogram presented in **(D)** is for *E. coli* K-12.

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

Equation. 3.1: S is the score for sRNA ranking. If no promoter is found upstream of the sRNA, S is the average coverage of the sRNA (presented by C). If the sRNA is associated with a promoter, S is assigned by P (defined by the users) times of the average coverage of the sRNA.

A sRNA candidate for regulation of fluid shear stress

ANNOgesic was also applied to *Pseudomonas aeruginosa* CF_PA39 in order to predict the sRNA candidates which may be regulated by the fluid shear stress. A newly detected sRNA candidate (sRNA10) which was significantly down-regulated under low fluid shear conditions compared to high fluid shear conditions was discovered by applying ANNOgesic [73] (Table 3.7). Thus, the sRNA detection of ANNOgesic not only identifies the known sRNAs but also provides the reliably potential sRNAs for experimental validation.

A long non-coding RNA - SRR42

Until recently long non-coding RNAs only be described in eukaryotic genomes. However, a highly expressed long non-coding RNA (SSR42) which can regulate virulence factors was found in *S. aureus* in 2012 by Morrison J. M. *et al.* [131]. It is also detected in *S. aureus* HG003 (1,249 nucleotides long) based on RNA-Seq data (Figure 3.18). Furthermore, the result of the multiple sequence alignment shows

Table 3.7: The sRNAs which are significantly down-regulated under low fluid shear conditions compared to high fluid shear conditions in *P. aeruginosa* CF_PA39

sRNA	Length (BP)	Position in the genome	Experimental validation	Fold change (RNA-Seq)
sRNA10	202	Intergenic region: PA3964-PA3965	No	-2.35
SPA0117	201	Intergenic region: PA3049 (rmf)-PA3050 (pyrD)*	Yes	-1.94
P8	78	Intergenic region: PA1030-PA1031	Yes	-1.85
SPA0003	137	Intergenic region: PA2729-PA2730	Yes	-1.58

The data is from the study of Dingemans *et al.* [73] , and only the sRNAs down-regulated ≥ 1.50-fold were included.

*The *SPA0117* sRNA overlaps the both genes.

that SRR42 widely exists in all strains of *S. aureus*, but does not exist in other *Staphylococcus* members.

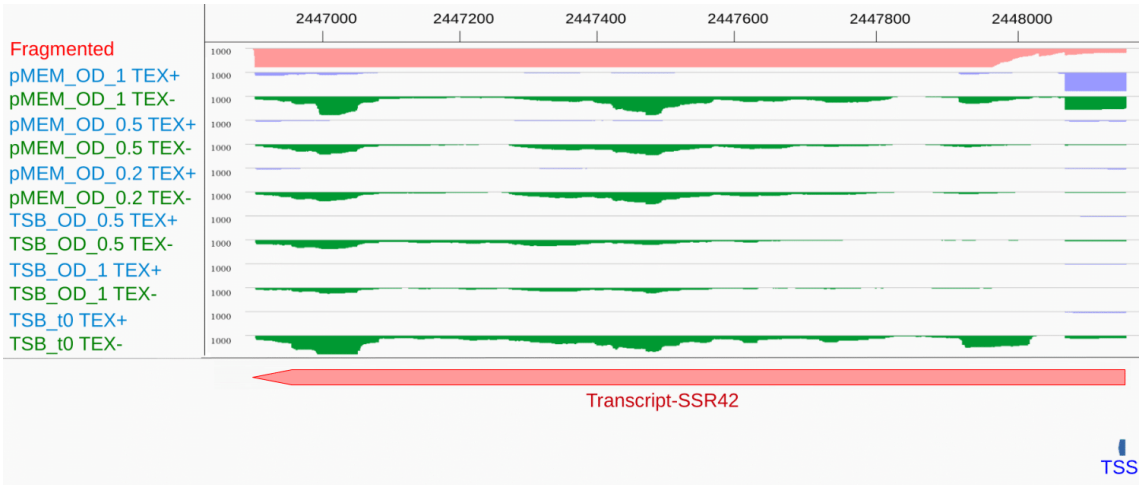


Figure 3.18: A SSR42 is located at the reverse strand from 2,446,903 to 2,448,151 in *S. aureus* HG003. The red coverage region, blue coverage regions and green coverage regions are for the library of fragmented transcripts from RNA-Seq, TEX+ and TEX- libraries from dRNA-Seq, respectively. The pink bar represents a SSR42 and the blue spot shows a TSS.

Targets of sRNAs

A module of sRNA target prediction was generated and integrated into ANNOgesic in order to understand the functions of sRNAs. Based on a resent review study of sRNA target prediction tools, CopraRNA [45], IntaRNA [46], RNAplex [47, 48], and RNAup [48, 49] are the four most outstanding tools [43, 44] for predicting the targets of sRNAs. ANNOgesic integrates RNAup, RNAplex and IntaRNA for the target prediction since CopraRNA needs manually selected homologs from different species of an sRNA and due to this, it cannot be used for constructing an automatic tool.

For numerous sRNAs, it has been proven that they can regulate the translation of bacterial mRNAs by masking the Shine-Dalgarno (SD) or the start codon in the 5' end of mRNA coding region [111]. Thus, 200 nucleotides upstream of CDSs and 100 nucleotides downstream of the start codons of CDSs were extracted as potential binding sequences of sRNAs to use IntaRNA [46], RNAplex [47, 48], and RNAup [48, 49]). ANNOgesic selects the mRNA targets predicted as top 20 interactions (default setting) in all of the three methods to provide the reliable candidates based on the information of binding energy. Moreover, some important information are provided as well, such as the interacting regions, the nucleotides of the base pairing, and the binding energy. The results of applying ANNOgesic to *S. aureus* HG003 shows that only 23.5% sRNA-mRNA interactions were detected in all of these three methods. Moreover, an interesting discovery reported in previously publications is also revealed in the result - most of the sRNAs can bind to multiple targets (average interacted targets is 4.9) for regulating various pathways, but the majority of the

targets tend to only interact with a specific sRNA (a target is regulated by 1.3 sRNAs) [132, 133] (Figure 3.19).

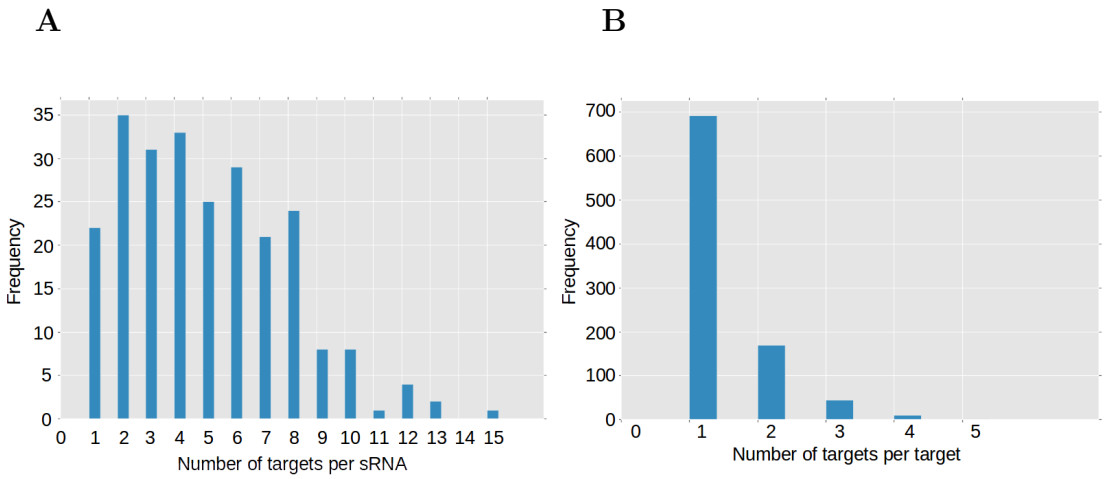


Figure 3.19: Number of interacting partners for sRNAs show that almost all of the sRNAs can bind to multiple mRNAs. However, most of the sRNA targets are only regulated by a specific sRNA. **(A)** The histogram of the number of interacting targets per sRNA. **(B)** The histogram of the number of interacting sRNAs per mRNA. Only the top 20 sRNA-mRNA interactions of RNAup, RNAplex, and IntaRNA predictions were considered.

Functions of sRNAs

Besides using sRNA target prediction tools, applying gene co-expression analysis was another option to detect the functions of sRNAs. For example, by appliying co-expression analysis for dual RNA-Seq data in *Salmonella*, a sRNA (PinT) and its functions were detected and validated by experiments [134]. Thus, in this study, gene co-expression analysis (using Spearman correlation coefficient) was also applied to allocate the functions of sRNAs for *S. aureus* HG003 based on the data of the 14 RNA-Seq samples (seven time points of bacterial populations growing in two different

conditions, a rich medium and a poor medium). Moreover, GO enrichment analysis was used for extracting the genes which contain enriched GO terms. Since the expression kinetics of the genes which are in the same group are similar, the functions of sRNAs can be detected by retrieving the functions of the genes co-expressed and inversely expressed with the sRNAs.

In the gene co-expression analysis of *S. aureus* HG003, many examples with previously published sRNAs show that their functions are closely related to the genes located in the same cluster. RNAIII, which is a widely studied sRNA, acts as the effector of the *agr* quorum-sensing system for regulating virulence genes [135, 136]. RNAIII regulates the expression of the repressor of toxins (rot), which is a global regulator of virulence gene expression in *S. aureus*, by occluding the Shine-Dalgarno sequence and blocking the translation [137, 138]. Some genes like *coa*, *lytM* and *spa* are also repressed by RNAIII [139, 140]. Moreover, RNAIII transcript also contains delta-haemolysin gene (*hld*) and activates translation of *hla* mRNA [136]. In our results of co-expression analysis, RNAIII (from 2,093,091 to 2,093,248 at the reverse strand in *S. aureus* HG003) was co-expressed with the genes of *agr* family, *hla*, and some virulence factors (Figure 3.20A, Appendix Table A.2). Furthermore, Dunman *et. al* provided a list of genes which are up-regulated and down-regulated with *agr* [140]. The list contains many genes of the *hut* family which are also shown in our co-expression results. Moreover, most of the inversely expressed genes of RNAIII also match to the list of *agr*-down-regulated genes which are provided by Dunman *et. al* like the *coa* and *lyt* family [140] (Figure 3.20B, Appendix Table A.2). The Spr sRNA family is another largely studied sRNA. Members of the family associated group

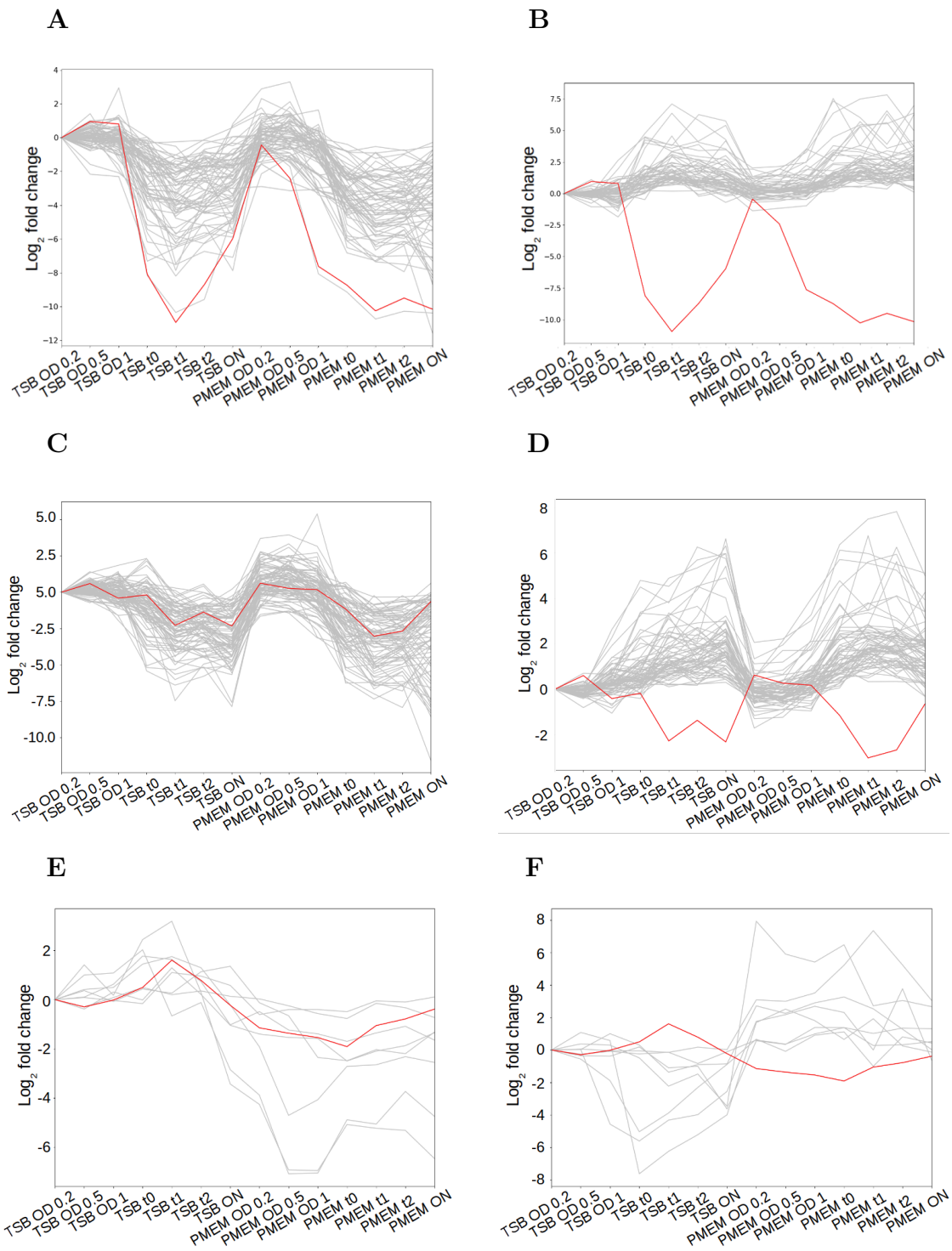
includes phage proteins and virulence proteins. The Spr sRNA family is expressed from pathogenicity islands which contain virulence and antibiotic resistance genes. Moreover, pathogenicity islands can be transferred through horizontal gene transfer like phages, plasmids and transposons [70, 141]. Thus, the Spr sRNA family is mainly co-expressed with phage proteins and virulence proteins like leukocidin and capsular polysaccharide biosynthesis protein. In addition, the Spr sRNA family is also inversely expressed with numerous tRNA-synthetases. It might be due to that the pathogenicity island usually uses tRNA loci for integration and recombination [142]. The kinetic curves of SprG4 [143] (from 942,430 to 942,482 at the forward strand) and its correlated proteins are shown in Figure 3.20C and D as an example (Appendix Table A.3).

Besides the functions of the known sRNAs, numerous potential functions of newly detected sRNAs can be characterized based on the application of gene co-expression analysis as well. For examples, a novel sRNA (from 90,947 to 91,092 at the forward strand) is mainly inversely expressed with the proteins involved in the conversion of lactate to pyruvate like L-lactate dehydrogenase, and co-expressed with some members of *bio* family which can regulate the biotin metabolism (Figure 3.20E and F, Appendix table A.4). Furthermore, biotin is a cofactor responsible for carbon dioxide transfer in pyruvate carboxylase used for converting pyruvate to oxaloacetate [144]. Thus, this sRNA may be able to decrease the storage level of pyruvates by repressing the conversion of lactate to pyruvate and enhancing the carboxylation of pyruvate. Moreover, a newly discovered sRNA (from 641,099 to 641,200 at the forward strand) is highly co-expressed with several iron transportation

related proteins such as ferrichrome transport permease, and inversely expressed with a ferritin, which is a storage of irons, as well as with two proteins down-regulated at iron-depleted conditions (*glpK* and *glmS*) [145, 146] (Figure 3.20G and H, Appendix table A. 5). Based on the analysis, this newly detected sRNA may be a regulator which can activate the iron transportation and inhibit storage of irons. Furthermore, a novel sRNA (from 2,485,411 to 2,485,628 at reverse strand) may be involved in the regulation of the purine metabolism because its expression values are highly correlated with the expression values of the *pur* family (Figure 3.20I, Appendix table A.6). In addition, two of the three genes inversely expressed with this sRNA are transporters (the last one is a phage integrase) (Figure 3.20J, Appendix table A.6). These predictions provide valuable information for the experimental validation.

sORFs

Small open reading frames (sORFs) are short sequences (normally $\leq$ 100 residues) with a start codon and a stop codon which form a potential protein-coding regions. The product of sORF is a short peptide (sPEP) which is usually lost in the process of protein extraction and purification. Therefore, sPEP may not be able to be detected in typical proteomic screens because of its small size and rapid degradation [147]. Due to these experimental difficulties, using computational approaches for detecting sORFs becomes an important method.


A
B
C
D
E
F
Log2 fold change
TSB OD 0.2
TSB OD 0.5
TSB OD 1
TSB t0
TSB t1
TSB t2
TSB ON
PMEM OD 0.2
PMEM OD 0.5
PMEM OD 1
PMEM t0
PMEM t1
PMEM t2
PMEM ON

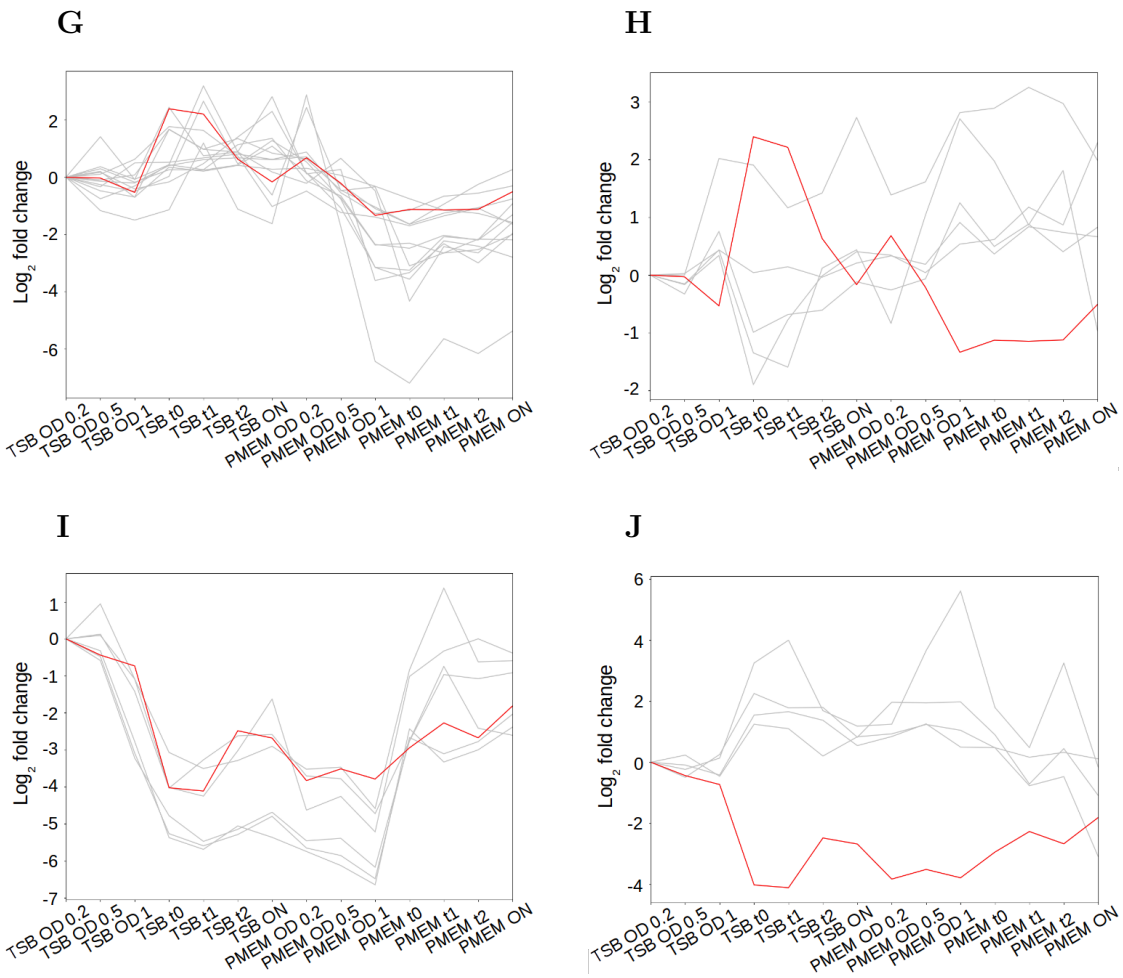


Figure 3.20: Examples of allocating sRNA functions by using gene co-expression analysis. X-axis represents the 14 conditions of RNA-Seq data. The rich media and poor media are presented as TSB and pMEM, respectively. For the time points, OD 0.2, OD 0.5, and OD 1 mean $OD_{600} = 0.2$, $OD_{600} = 0.5$, $OD_{600} = 1$, respectively. t0, t1, and t2 indicate 0 hour, 1 hour, and 2 hours after entering stationary phase, respectively. ON represents overnight. Y-axis shows $\log_2$ fold change of gene quantification values. The red line and the gray lines represent the queried sRNA and the other genes, respectively. **(A)**, **(C)**, **(E)**, **(G)** and **(I)** show the co-expressed genes with the queried sRNA, and the others present the genes inversely expressed with the queried sRNA. **(A)** and **(B)** are for a known sRNA - RNAIII, **(C)** and **(D)** show the cluster for another known sRNA - SprG4 the others are for newly discovered sRNAs.

However, the currently available tools and databases for detecting sORFs are

few. CodonW is one of the widely used tools [148]. It was developed based on the assumption that the codon usages of true ORFs are not random. All of the true sORFs should contain the optimal codon pattern. Therefore, CodonW can detect sORFs by searching the specific codon patterns. Another commonly used sORF detection tool is sORF finder which can detect sORFs by calculating nucleotide composition frequency and coding potential score [149]. However, sORF finder can only detect sORFs with less than 100 codons. Moreover, a review study reported that these two tools may not be the solution of sORF detection due to their low accuracy. For detecting sORFs with less than 100 residues, the sensitivities of these two tools are only 35% to 65% at 20% false positive rate, and 25% to 50% at 5% false positive rate. In addition, for identifying sORFs with 100-150 residues, the true positive rates are not higher than 70% either at 5% false positive rate or 20% false positive rate in both of the methods [150].

In order to develop an approach with higher precision for detecting sORFs, a new method based on RNA-Seq data was created and integrated into ANNOgesic. This approach searches for the short expressed non-annoated transcripts (default setting is within 30 to 150 base-pairs) containing start and stop codons which can form potential sORFs. Moreover, a ribosome binding site (RBS) must be detected between a TSS and 3 to 15 nucleotides upstream of the start codon (Figure 3.21). The sequence length of the sORF as well as the sequence patterns of the start codon, the stop codon, and the RBS can be assigned by the users to satisfy some special requests like using non-canonical start codons.

For *S. aureus* HG003, 181 sORFs which comprise 10 antisense sORFs, 34 intergenic

sORFs, 42 3' UTR-derived sORFs, 33 interCDS-derived sORFs, and 62 5' UTR-derived sORFs were detected by ANNOgesic.

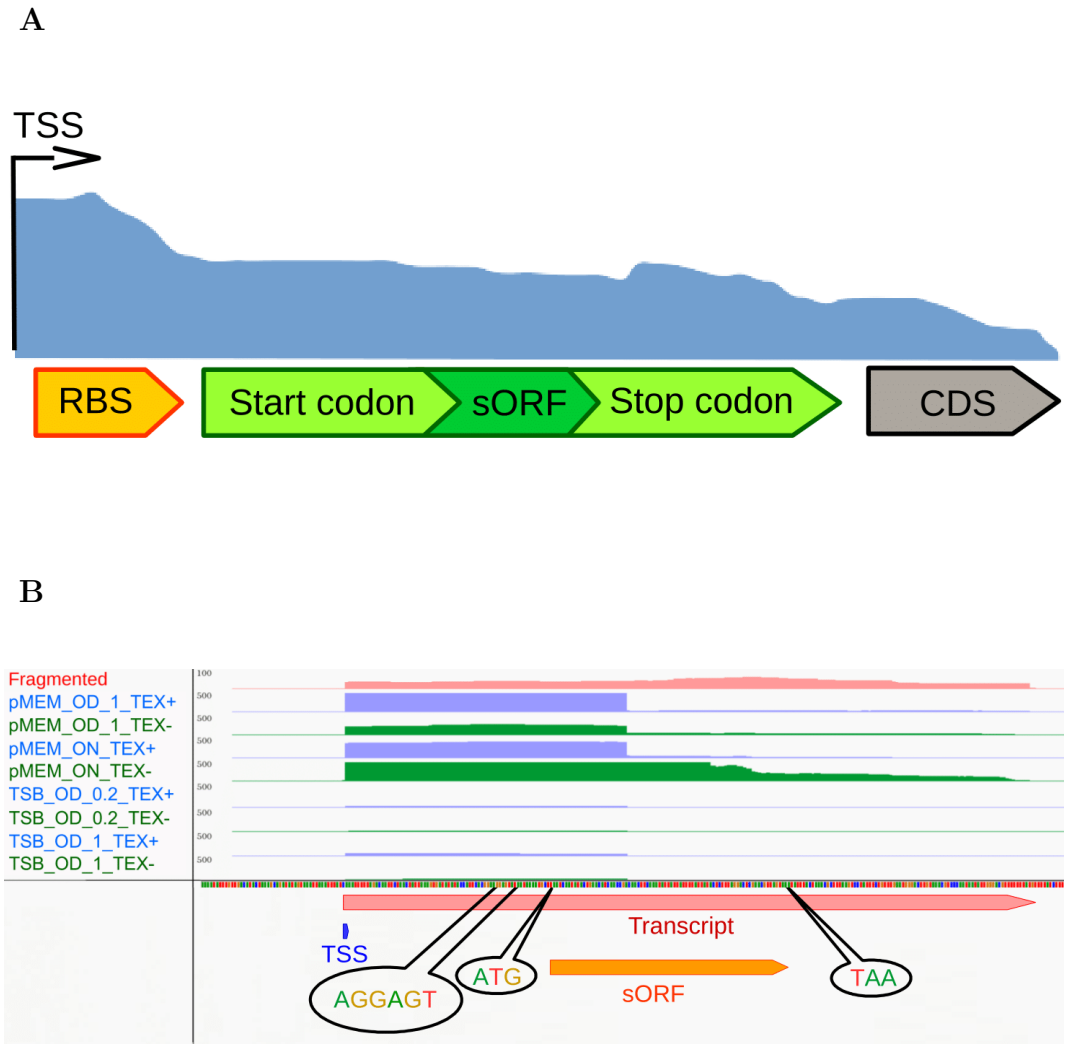


Figure 3.21: The method and an example of sORF detection. **(A)** The approach of sORF detection. A short non-annotated transcript (default 30 - 150 nts) containing start and stop codons which can form an ORF is considered as a sORF candidate. Moreover, a RBS must be discovered between a TSS and the start codon otherwise the candidates are excluded. **(B)** An example of a sORF from *S. aureus* HG003. The pink, blue and green coverages represent the library of RNA-Seq generated after transcript fragmented, TEX+ and TEX- libraries of dRNA-Seq, respectively. The TSS, transcript, and sORF are presented as blue, pink and green bars, respectively. A start codon, stop codon, and Shine-Dalgarno Sequence were detected. The location of this sORF is from 2,111,563 to 2,111,646 bp at the forward strand.

Circular RNAs

Circular RNAs (circRNAs) are recently discovered genomic features [151,152]. Unlike conventional linear RNAs, they are a special type of non- coding RNA and forms a closed continuous loop (the 3' and 5' ends are joined together). Due to the low expression of circRNAs, the detection of circRNAs is still a challenge for both experimental and computational approaches. Since the first genome-wide detection of circRNAs based on RNA-Seq data was published in 2012 [152], RNA-Seq has become a potent method for the detection of circRNAs. The RNA-Seq-based methods for identifying circRNAs mainly focus on searching for the splice sites located at the two terminals of a RNA-Seq read (Figure 3.22). Currently, almost all of the previously reported circRNAs were found in eukaryotic genomes because the splicing events rarely occur in bacterial and archaeal organisms. However, a study published in 2012 reported a transcriptome-wide discovery of circRNAs in archaea [153].

In 2014, the functions of Segemehl were extended in order to predict circRNAs by searching and classifying different types of splice sites. The recall and precision of Segemehl for detecting circRNAs are 85% and 98%, respectively [105]. Based on the outstanding performance, ANNOgesic integrates it for its detection of circRNAs. Moreover, ANNOgesic compares circRNA candidates with genome annotations in order to exclude the false positives which are marked as CDSs, tRNAs, or rRNAs. Furthermore, the circRNA candidates with low ratio between supporting reads and total reads were removed as well. For *S. aureus* HG003, no candidate can be detected after removing the false positives.

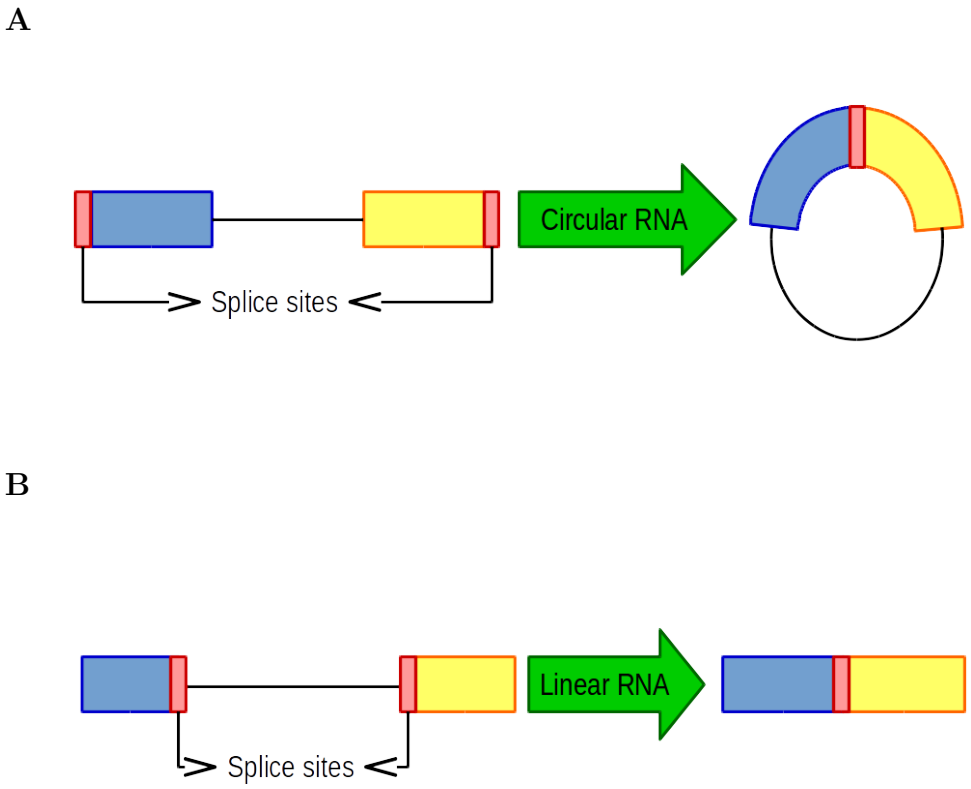


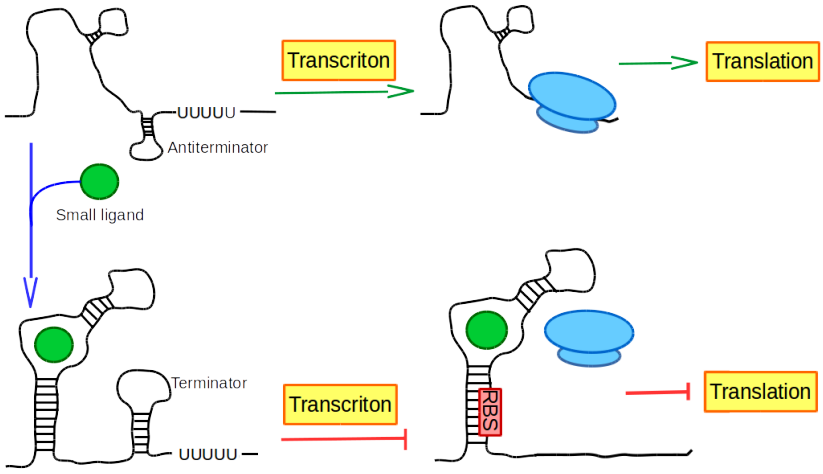
Figure 3.22: Detection of circRNA is based on searching the splice sites located at the two ends within a RNA-Seq read. **(A)** The splice sites of a circRNA. **(B)** The splice sites of a normal linear RNA.

Riboswitches and RNA thermometers

Riboswitches and RNA thermometers (RNATs) are two structured regulatory RNAs located in the 5' UTRs of bacterial genomes. Riboswitches can regulate downstream gene at the level of the transcription termination, translation initiation or RNA stability by interacting with small molecules (Figure 3.23A). The previous studies reported that RNATs can influence translation initiation based on the change of temperature (Figure 3.23B) [154]. For the prediction of these two important reg-

ulators, ANNOgesic extracts the potential sequences that are between TSSs (or the starting point of the transcript if no TSS is detected) and downstream CDSs, and associated with ribosome binding site to search for the homologs in the Rfam database by running Infernal [106, 107]. For *S. aureus* HG003, 22 riboswitches and 11 RNA thermometers were found (Table 3.8).

A



B

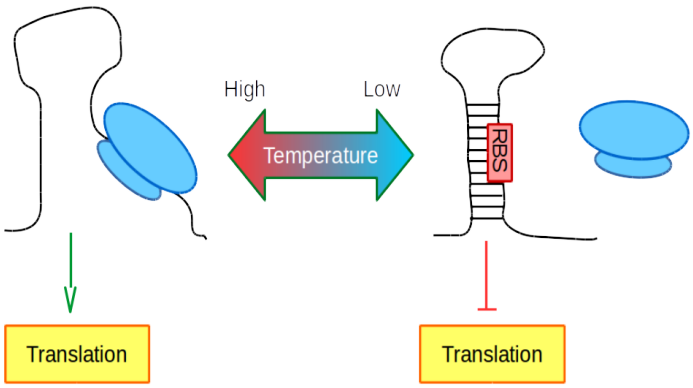


Figure 3.23: Mechanisms of riboswitches and RNA thermometers. **(A)** Riboswitch. **(B)** RNA thermometer.

Table 3.8: Riboswitches and RNA thermometers in *S. aureus* HG003

Riboswitch		RNA thermometer	
Name	Number	Name	Number
drz-agam-2-2	1	FourU	1
Purine	1	Phe_leader	5
FMN	2	ROSE_3	1
yybP-ykoY	1	PrfA	2
glmS	1	hsp17	2
SAM-SAH	2		
Glycine	1		
Lysine	1		
SAM_V	1		
PreQ1	3		
SAM	3		
speF	1		
TPP	2		
preQ1-II	1		

CRISPRs

Clustered regularly interspaced palindromic repeat (CRISPR) plays a vital role in a bacterial immunological system to resist phage invasion. It consists of several repeated sequences, and each repeated sequence is followed by a spacer from the foreign DNA, such as viruses or plasmids [155]. Based on the specific sequence pattern of CRISPRs, numerous CRISPR prediction tools were constructed. One of the widely used tools is CRISPR recognition tool (CRT) requiring only within 2 seconds to detect the CRISPR candidates of a 2 million bp genome with 90% recall and 100% precision [108]. Thus, CRT was integrated into ANNOgesic for the detection of CRISPRs. Moreover, ANNOgesic makes comparisons between CRISPR

candidates and genome annotations in order to remove the false positives which are reported as CDSs, tRNA, rRNA, etc.

In this study, the CRISPR detection of ANNOgesic was applied to *S. aureus* HG003, *H. pylori* 26695, and *C. jejuni* 81116. A CRISPR with 5 repeated sequences and a CRISPR with 8 repeated sequences were found in *S. aureus* HG003 and *C. jejuni* 81116, respectively (Figure 3.24). However, in *H. pylori* 26695, no CRIPSR was detected.

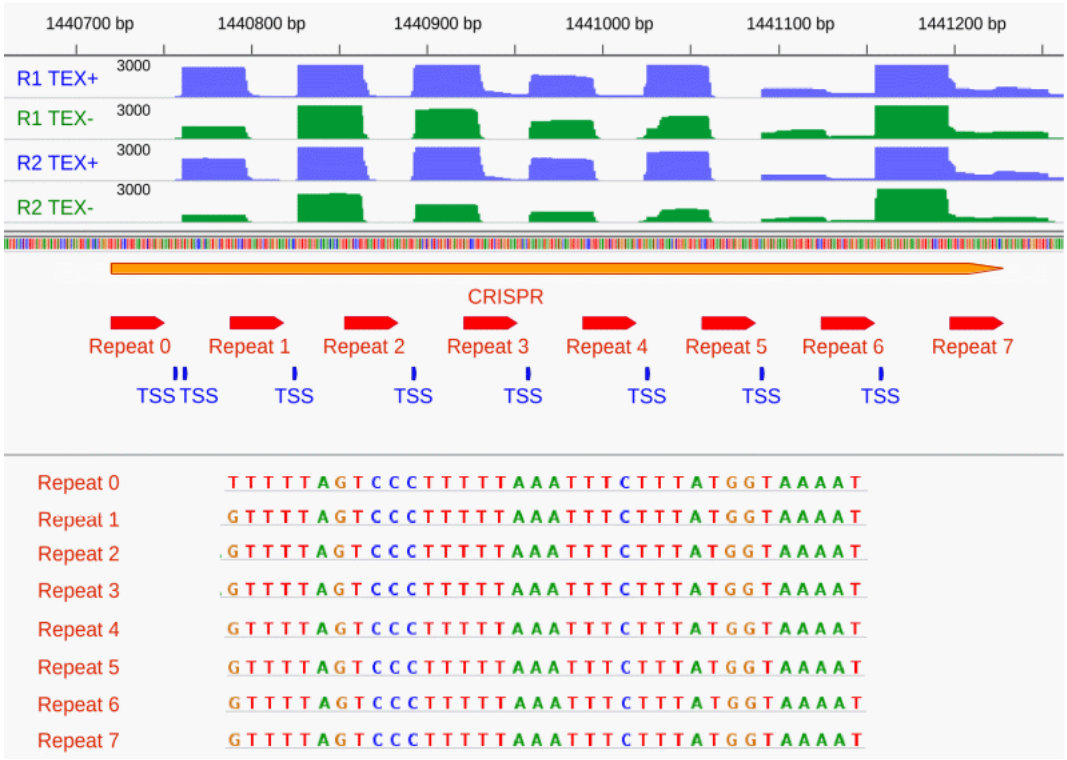


Figure 3.24: An example of CRISPRs in *C. jejuni* 81116. The transcript fragmented RNA-Seq library, TEX+ library and TEX- libraries of dRNA-Seq are presented as the pink, blue, and green coverages mean , respectively. The whole region of CRISPR, repeat units, and TSSs are represented by the orange bar, red bars, and blue spots, respectively. The CRISPR is from 1,440,718 to 1,441,215 bp.

Functional labeling system

GO terms

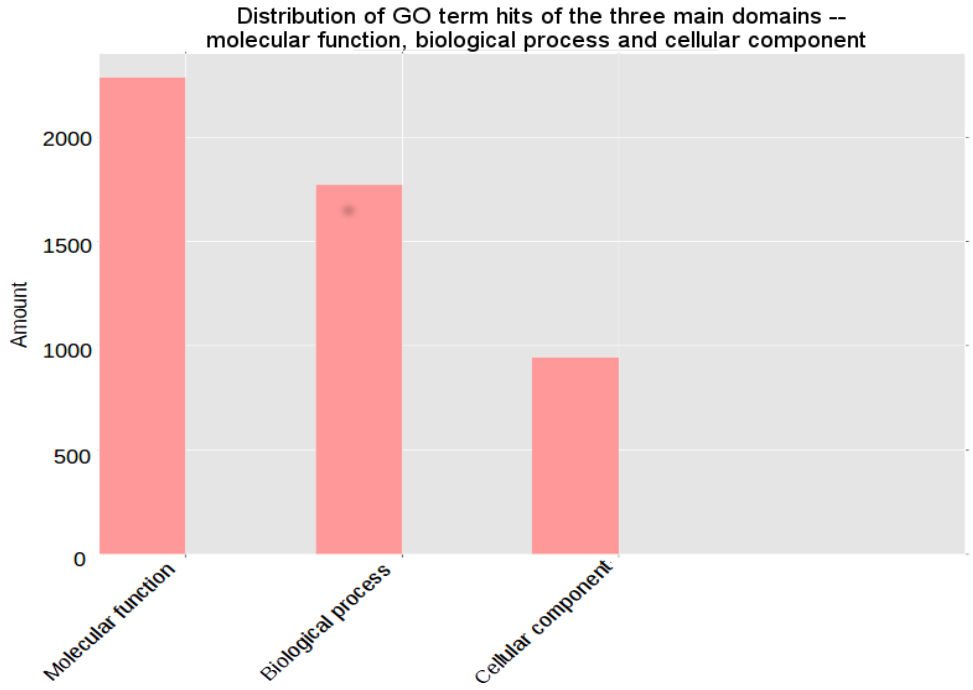
Since vast amount of genomic features were identified based on RNA-Seq data in recent years, understanding the metabolisms and regulations of these genomic features becomes an urgent need. Thus, a sophisticated method for addressing the functions of genes and classifying them is required. Gene Ontology (GO) is a major bioinformatic resource for annotating and cataloguing functions and locations of gene products. It provides numerous structured and controlled vocabularies for describing biological processes, molecular functions, and cellular components of gene products [56, 57]. Moreover, GO slim, which is a simplified version of the Gene Ontologies for generating a broad overview of the GO terms, was developed as well [57]. Since GO slim contains only a subs of GO terms without the details of the species-specific, fine-grained terms, it is a useful resource and widely applied for generating a summary of the GO terms of a genome or a gene population.

In order to identify the functions of genes, ANNOgesic allocates GO and GO slim terms to CDSs by searching for protein IDs in Universal Protein Resource (Uniprot) [99, 100]. Uniprot provides a ID mapping list used for converting different IDs to GO terms. Moreover, ANNOgesic can also retrieve the GO terms only for expressed CDSs based on the information of the transcripts. Therefore, the comparison between the GO terms of expressed CDSs and the GO terms of all CDSs can be performed. In addition, the variation of the GO terms of expressed CDSs

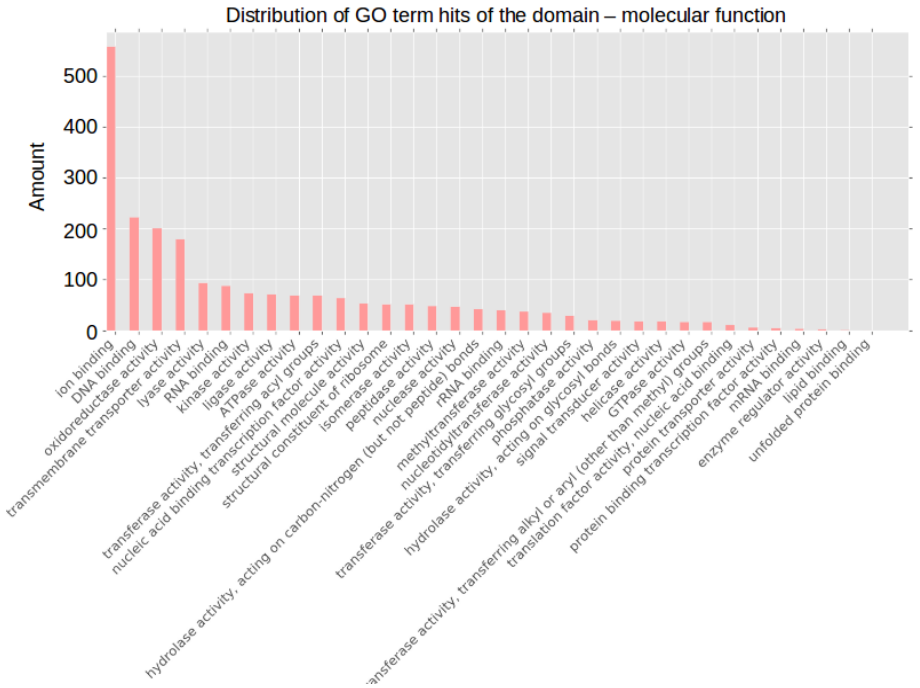
between different conditions can be detected as well.

In *S. aureus* HG003, the GO terms of 2,073 CDSs (75% of all CDSs) were found in Uniprot mapping list. Around 46% of the CDSs belong to molecular function domain, 35% of the CDSs are involved in biological process domain and the rest CDSs are located in cellular component domain (Figure 3.25A). In the domain of molecular function, the majority is ion binding (24%), following up by DNA binding (9.8%), oxidoreductase activity (8.8%), and transmembrane transporter activity (7.9%) (Figure 3.25B). In the domain of biological process, the proteins related to biosynthetic process (12.7%), transport (9.3%), and cellular nitrogen compound metabolic process (9.1%) are more than others (Figure 3.25C). Cytoplasm (35%), plasma membrane (23%) and cytosol (11%) occupy around 70% of cellular component domain (Figure 3.25D).

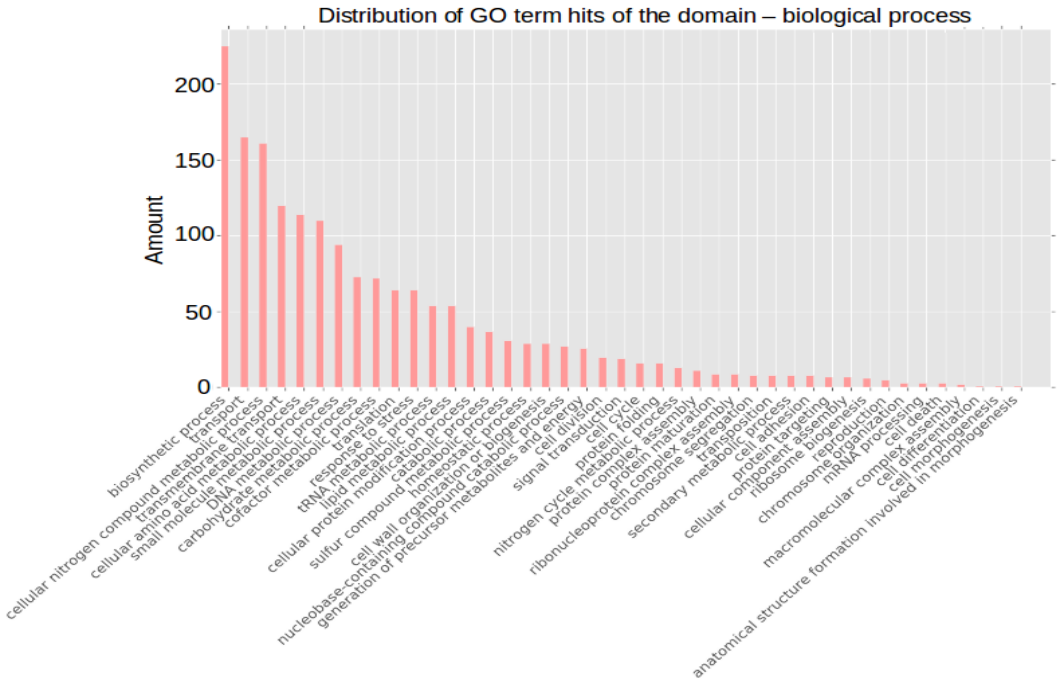
A



B


Distribution of GO term hits of the domain – molecular function
Amount
0
100
200
300
400
500
ion binding
DNA binding
oxidoreductase activity
transmembrane transporter activity
lyase activity
RNA binding
kinase activity
ligase activity
ATPase activity
transferring acyl groups
transferase activity, transferring acyl groups
nucleic acid binding transcription factor activity
structural molecule activity
structural constituent of ribosome
isomerase activity
peptidase activity
nuclease activity
rRNA binding
methyltransferase activity
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
nucleotidyltransferase activity
transferase activity, transferring glycosyl groups
hydrolase activity, acting on glycosyl bonds
phosphatase activity
signal transducer activity
helicase activity
GTPase activity
transferase activity, transferring alkyl or aryl (other than methyl) groups
translation factor activity, nucleic acid binding
protein transporter activity
protein binding transcription factor activity
mRNA binding
enzyme regulator activity
lipid binding
unfolded protein binding

C


Distribution of GO term hits of the domain – biological process
Amount
0
50
100
150
200
biosynthetic process
transport
cellular nitrogen compound metabolic process
transmembrane transport
cellular amino acid metabolic process
small molecule metabolic process
DNA metabolic process
carbohydrate metabolic process
cofactor metabolic process
translation
response to stress
tRNA metabolic process
lipid metabolic process
cellular protein modification process
carbohydrate metabolic process
sulfur compound metabolic process
homeostatic process
cell wall organization or biogenesis
nucleobase-containing compound catabolic process
generation of precursor metabolites and energy
cell division
signal transduction
cell cycle
nitrogen cycle metabolic process
protein folding
protein complex assembly
ribonucleoprotein complex assembly
protein maturation
chromosome segregation
secondary metabolic process
transposition
cellular component organization
cell adhesion
protein targeting
protein assembly
ribosome biogenesis
reproduction
chromosome organization
macromolecular complex assembly
mRNA processing
cell death
anatomical structure formation involved in morphogenesis
cell differentiation
cell morphogenesis

D

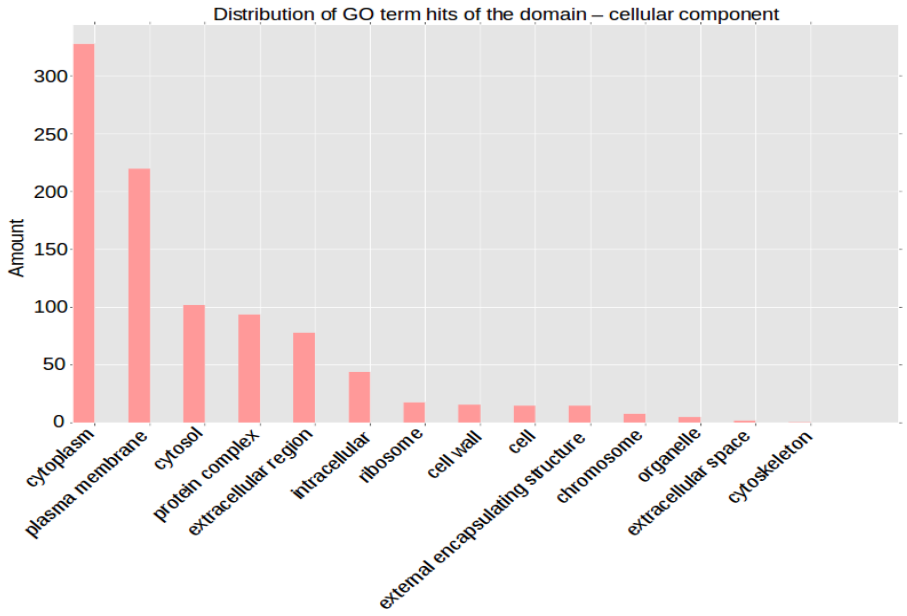


Figure 3.25: Distribution of GO terms in *S. aureus* HG003. **(A)**, **(B)**, **(C)**, and **(D)** are for the three root domains, molecular function, biological process, and cellular component, respectively.

Subcellular localizations

The bacterial proteins usually carry out their unique functions in specific locations of the cell. Therefore, the functions and the regulation networks may be related to the subcellular localizations of the proteins. For example, most of the proteins located at the cell membrane are involved in transportation, toxin secretion, or signal exportation such as ABC type transporters and the proteins of the Sec secretion system. Moreover, identification of subcellular localizations is also widely applied for searching drug and vaccine targets. For identifying the subcellular localizations, a lot of novel tools and studies were developed and published.

However, most of the subcellular localization prediction tools are built based on

eukaryotic genomes, and may not be able to apply to bacterial species. A study for comparing the available subcellular localization prediction tools for bacterial genomes was published [156]. Eight computational prediction tools (PSORT I [157], PSORTb [103, 104], Proteome Analyst [158], SubLoc [159], CELLO [160], PSLpred [161], LOCtree [162], and P-CLASSIFIER [163]) were included in this comparison. For general prediction of the subcellular localization, the precision scores of PSORTb (97%) and Proteome Analyst (95%) are the two highest ones, and the recalls of SubLoc (85%) and LOCtree (87%) are the best. However, some of these tools may be specialized in the predictions of several specific locations. Thus, an analysis of the feature-based predictions (exported proteins, cytoplasmic membrane proteins, and outer-membrane proteins) was done as well. The performance of PSORTb and Proteome Analyst are the best performing tools in these feature-based predictions. Moreover, the first study of comparing the computational and laboratory methods for the detection of subcellular localization in bacteria was published in 2005 [164]. The results of this study show that the precision of the computational approach (PSORTb) exceeds the precision from the high-throughput laboratory approach. The error rate of the high-throughput laboratory approach is 14.3% across 10 strains, but the error rate of PSORTb is only 0.7%. Therefore, the computational prediction is a crucial and non- ignorable step for detecting the subcellular localizations.

Based on the results of the previous assessments, PSORTb shows high accuracy to all the tests and can predict all locations. Hence, it was integrated into ANNOgesic in order to construct a module for identifying subcellular localization. For *S. aureus* HG003, the subcellular localizations of 2,211 CDSs (567 unknown) were detected.

Around 73% proteins are located at cytoplasm (47%) and cytoplasmic membrane (26%). The proteins that are located extracellular and in cell wall only occupy 4.2% and 1.5% of all proteins (Figure 3.26).

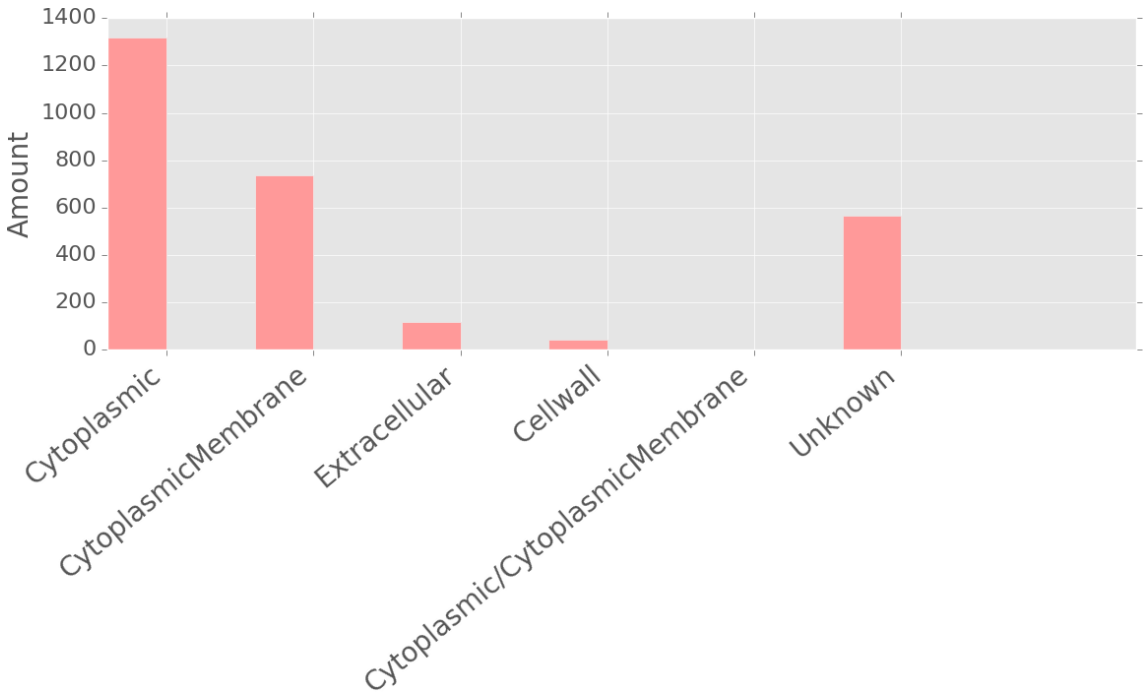


Figure 3.26: Distribution of subcellular localizations in *S. aureus* HG003. **(A)** excludes the proteins which can not be predicted by PSORTb, and **(B)** includes all proteins.

Protein-protein interactions

Detecting protein-protein interaction (PPI) is essential for understanding the protein functions and pathways. Thus, PPI is considered as an important key for drug design. However, the experimental detection of PPI is a difficult task because of the high false positives and false negative rate [165]. Hence, computational approaches have become a huge benefit for providing high-confidence candidates for experimental

validation. Due to the importance of the computational PPI predictions, numerous tools were created based on different methods and resources, such as machine learning approaches, homology searches, and sequence or structure based algorithms [166]. Furthermore, a lot of PPI databases storing, searching as well as exchanging the information of PPIs were constructed.

In the available PPI databases, STRING [101] is one of the most powerful databases updated regularly. It provides a crucial assessment for the reliability of the PPIs including physical and functional associations. In addition, GO term analysis and protein clustering were applied in STRING as well. Currently, STRING not only stores the information of PPIs for over 2,000 organisms, but also provides clear visualizations and variant types of interaction information (Figure 3.27). Therefore, STRING is a helpful resource for the selection of PPIs to perform experimental validation.

STRING integrated a text-mining system which can search publications in Pubmed for helping the selection of the reliable and interesting PPIs. However, the text-mining system of STRING searches the articles only based on the protein names, not based on the keywords like interaction, binding, activate, etc (Figure 3.27E). Hence, several false positive hits are still found in this database. In order to solve this issue, PIE, which is a text-mining tool for detecting PPIs based on the protein names and the keywords of PPIs in publications, was developed [102]. PIE also provides a score system for evaluating the precision of PPIs, and a Pubmed ID list for retrieving the articles. By the application of PIE, the reliability of PPI detection is improved significantly.

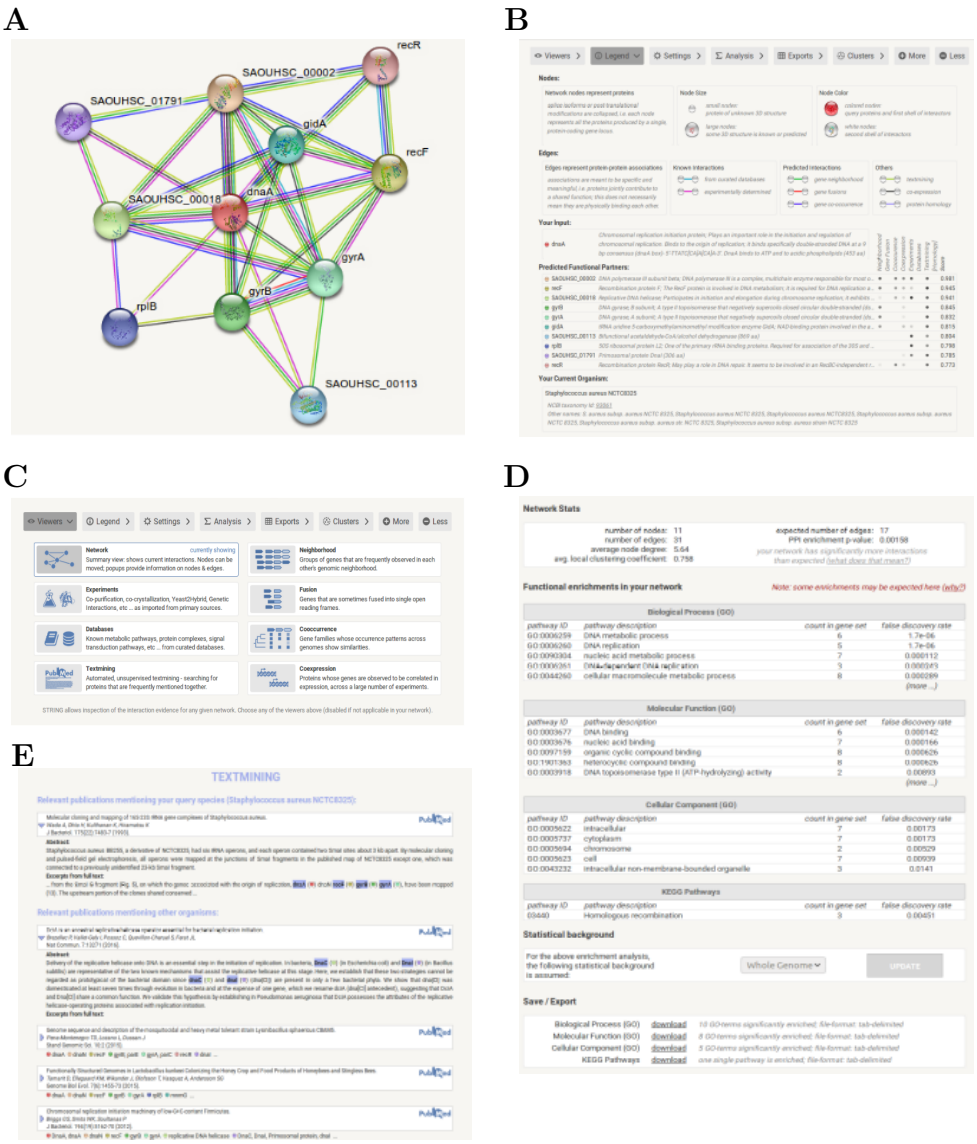


Figure 3.27: An example of a STRING search – searching dnaA of *S. aureus* NCTC8325. **(A)** The visualization of PPI network. **(B)** The legends of the PPI network. **(C)** The options for visualization. **(D)** The GO term analysis in STRING. **(E)** The text-mining for searching publications of PPI in Pubmed. The figures were retrieved from the version 10.5 of STRING website [101] (http://string-db.org).

For PPI detection, ANNOgesic retrieves the data from STRING and apply PIE for selecting high-confidence PPIs (high PIE scores). Moreover, all the results can be viewed by applying a clear visualization method. As displayed in Figure 3.27, the visualization of STRING cannot show the information of the supporting literature, and the text-mining method in STRING is also not an ideal one. On the contrary, ANNOgesic can generate clear figures of PPIs with the information of supporting literature and the PIE scores (Figure 3.28). Figure 3.28A shows that although the amount of publications which support the interaction between *gyrA* and *grlB* are more than the interaction between *dnaA* and *dnaD*, the reliability of the former is higher than the latter. In addition, several interesting points are revealed by comparing the Pubmed searches with or without assigning strain names. For example, the interaction between *dnaA* and *dnaI* has not beed reported in *S. aureus* (Figure 3.28B), but it was fully studied in other organisms (Figure 3.28A).

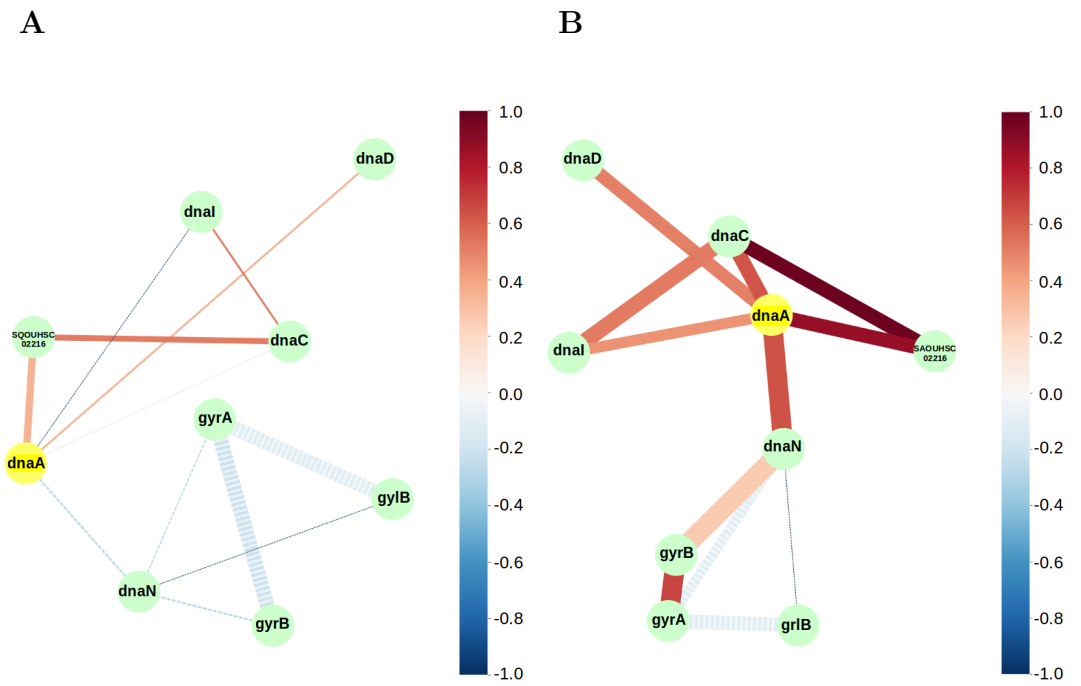


Figure 3.28: An example of new visualization of PPI by using ANNOgesic (dnaA of *S. aureus* NCTC8325). The yellow circles show the queried protein (dnaA) in *S. aureus* NCTC8325. The other proteins related to the queried one are presented as green circles. The dotted lines represent the interactions without supporting literature, the dashdot lines represent the interactions with supporting literature though their PIE scores are below 0, and the solid lines mean that the interactions are supported in the literature with high PIE scores (higher than 0). The thickness of the lines indicate the amount of the articles which support the interaction. Moreover, the color of the connections indicates the PIE scores. **(A)** The result of the Pubmed search with the specific word – "Staphyloccocus aureus". **(B)** The result of the Pubmed search without the strain name.

Assessment of ANNOgesic predictions

In order to assess the performance of ANNOgesic, the predictions of ANNOgesic based on dRNA-Seq and conventional RNA-Seq data sets of *Escherichia coli* K12 MG1655 by Thomason *et al.* (dRNA-Seq) [75] and McClure *et al.* [27] (conventional

RNA-Seq) are compared with the previously reported genomic features in several databases [76–81]. The results show that most of the predictions can achieve a high sensitivity of 80% (Table 3.9). However, TSS prediction represent an exception with low detection rate may mainly because the dRNA-Seq method may achieve higher sensitivity in detecting TSSs than the other protocols. In order to test our hypothesis and investigate the quality of the previously reported TSSs in RegulonDB, a comparison between three deposited TSS datasets (Salgado *et al.* generated with Illumina RNA-Seq [167], and Mendoza- Vargas*et al.* generated with Roche 454 high-throughput pyrosequencing [168], and Roche 5'RACE [168]) was used and an extremely low overlap was found (Figure 3.29). Moreover, the 50 nucleotides at the upstream TSSs were extracted for searching the common promoter motifs. Based on the results of using MEME [97], the promoter motifs were only found in 0% to 7% of the deposited TSSs while 80% of TSSs detected by ANNOgesic are associated with promoters (Table 3.10). Due to this result, the previously reported TSS sets (including the TSS information of promoter set in RegulonDB) may not be able to represent a benchmarking set for evaluating the accuracy of the TSS predictions of ANNOgesic.

Table 3.9: The comparison between ANNOgesic predictions and several databases

Feature	Database	Sensitivity of *E. coli* from dRNA-Seq [75]	Sensitivity of *E. coli* from conventional RNA-Seq [27]	Sensitivity of *H. pylori* [60]	Sensitivity of *C. Jejuni* [24]
Transcript	EcoCyc [76]	86%	90%	-[i]	-
Operon	DOOR[2] [77]	72%	70%	74%	80%
	RegulonDB [78] [a]	90%	89%	-	-
sRNA[b]	RefSeq [79]	90%	70%	-[j]	-
	RegulonDB	80%	55%	-	-
	Others	-	-	90% [k]	84% [l]
TSS[c]	RegulonDB (3 datasets)	∼6%	-	-	-
Terminator[d]	RegulonDB	72%	70%	-	-
	EcoCyc	86%	84%	-	-
UTR[e]	RegulonDB	5' UTR 86%	-	-	-
		3' UTR 63%[f]	-	-	-
Promoter[g]	RegulonDB	39%	-	-	-
sORF[h]	Hemm *et. al* [80]	74%	-	-	-
Riboswitch	EcoCyc	83%	-	-	-
CRISPR	CRISPRdb [81]	100%	100%	100%	100%

[a]The features marked as "weak evidence" confidence level by RegulonDB were excluded.

[b]The non expressed sRNAs in published datasets were removed.

[c]The overlapped TSSs of three datasets are few. Moreover, most of the published TSSs (< 8%) are not associated with promoters.

[d]The terminators which do not contain coverage significant drop were removed.

[e]The non expressed UTRs in published datasets were excluded.

[f]The information of 3' end is usually lost in dRNA-Seq data.

[g]Based on TSSs information in the promoter set, only 22% promoters can be detected [97].

[h]The non expressed sORFs in published datasets were removed.

[i]"-" represents the feature of the strain has no proper dataset from the database or can not be generated.

[j]The sRNA comparison for *H. pylori* and *C. Jejuni* are done by other literature which shown in manuscript. [k]sRNAs of *H. pylori* is from Sharma *et al.* [60].

[l]sRNAs of *C. Jejuni* is from Dugar *et al.* [24].

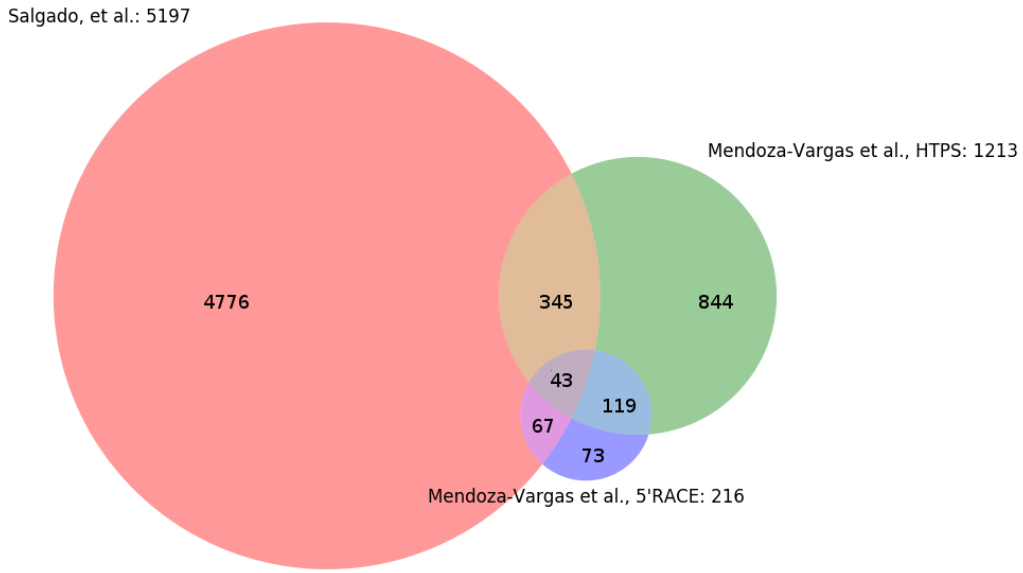


Figure 3.29: The overlap of three previously published TSS datasets in RegulonDB [167, 168].

Table 3.10: Number of TSSs and their associated promoter motifs in RegulonDB [78]

Dataset	Total TSSs	Number of promoters
Salgado *et al.* Illumina RNA-Seq [167]	5,197	374 (7%)
Mendoza-Vargas *et al.* Roche 454 high-throughput pyrosequencing [168]	1,213	23 (2%)
Mendoza-Vargas *et al.* Roche 5'RACE [168]	216	0 (0%)
Using the TSSs from the promoter set in RegulonDB for running MEME [97]	6,478	1,450 (22%)

Generation of coverage plots via the IGV API

ANNOgesic also contains several modules helping the users to review the annotations. In order to compare the different genomic features, ANNOgesic offers an user friendly module which can merge all the given genomic features to generate an annotation file in GFF3 format. Moreover, the parental transcripts can be detected and assigned to each genomic feature.

If the number of the libraries of dRNA-Seq is large, checking TSSs or PSs becomes a difficult task because the TEX+ and TEX- libraries of dRNA-Seq need to be distinguished laboriously. Because of this, A module of ANNOgesic was developed for generating screenshots by using IGV application programming interfaces (API) which is a set of commands, functions, protocols, and objects provided for developers to build applications easily [169]. Afterward, the tracks of screenshots can be colorized automatically (Figure 3.30). By using this approach, the users just need to check the screenshots without the manual manipulation of genome browser.

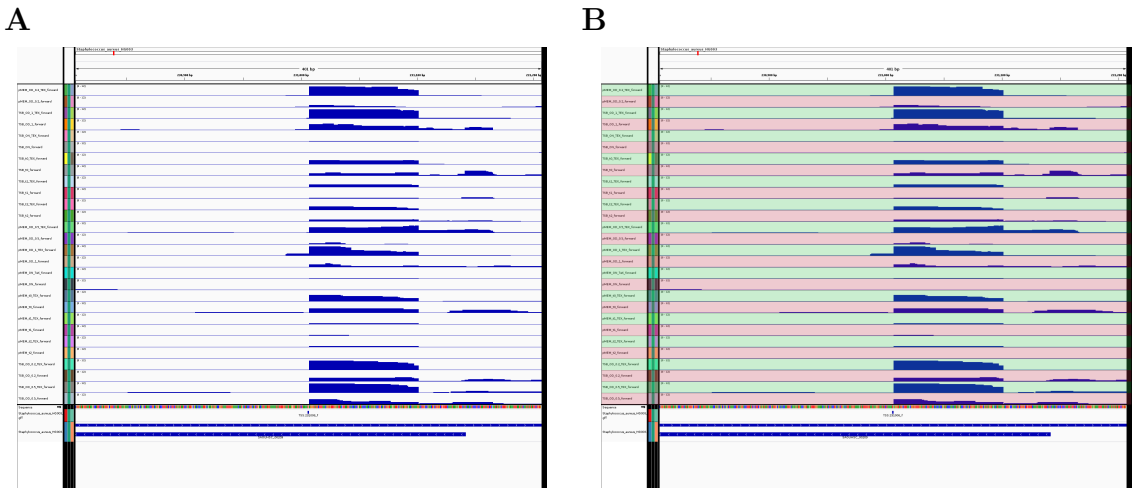


Figure 3.30: ANNOgesic can generate and colorize screenshots via IGV API automatically. This case has 28 libraries of *S. aureus* HG003. **(A)** The TEX+ and TEX- libraries of dRNA-Seq are distinguished with difficulty due to the vast number of libraries. **(B)** This figure is generated from ANNOgesic for providing an easy way to view the data.

An interactive interface for browsing and searching generated annotations and interactions

In order to search and analyze the data of annotations, an interactive table and figure were generated based on the Python libraries Bokeh [170] and Biocircos [171], respectively (Figure 3.31, 3.32). The interactive table provides a simple way for sorting, browsing, and comparing the data. It also links to several public databases for obtaining more information of the genomic features like Gene in NCBI, Rfam, and CRISPRdb [81,107,172]. Additionally, the associated transcripts, TSSs, and PSs can be found in the table as well. Moreover, the results of co-expression analysis for sRNAs, including the interactive plots and all the details of the genes co-expressed

and inversely expressed with the queried sRNA, can be provided in this interactive table. Besides the interactive table, an interactive figure was generated for an overview of all the annotations for *Staphylococcus aureus* HG003.

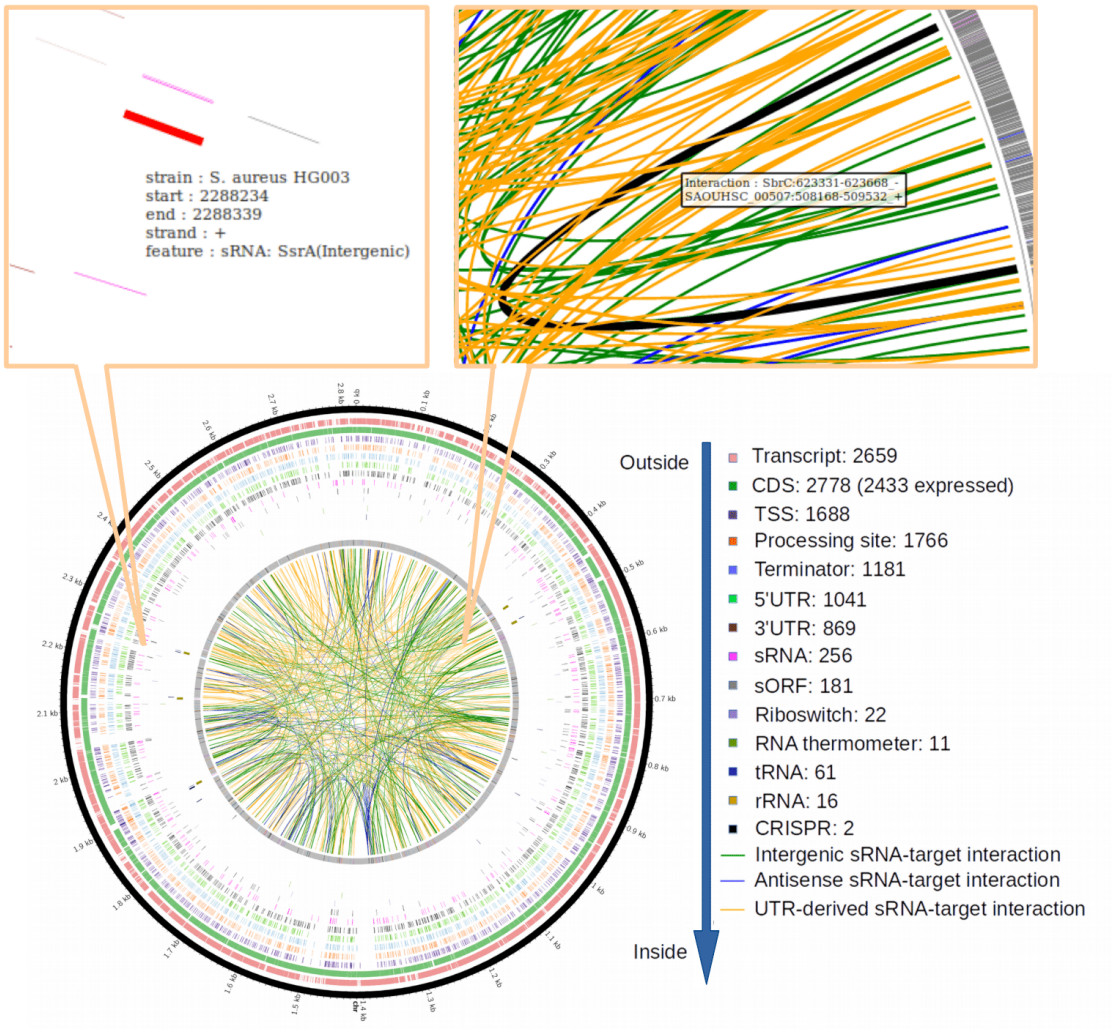


Figure 3.31: Screenshots of the interactive figure for an overview of the annotations of *Staphylococcus aureus* HG003. The interactive figure can show the detailed information of all functional genes and sRNA target interactions.

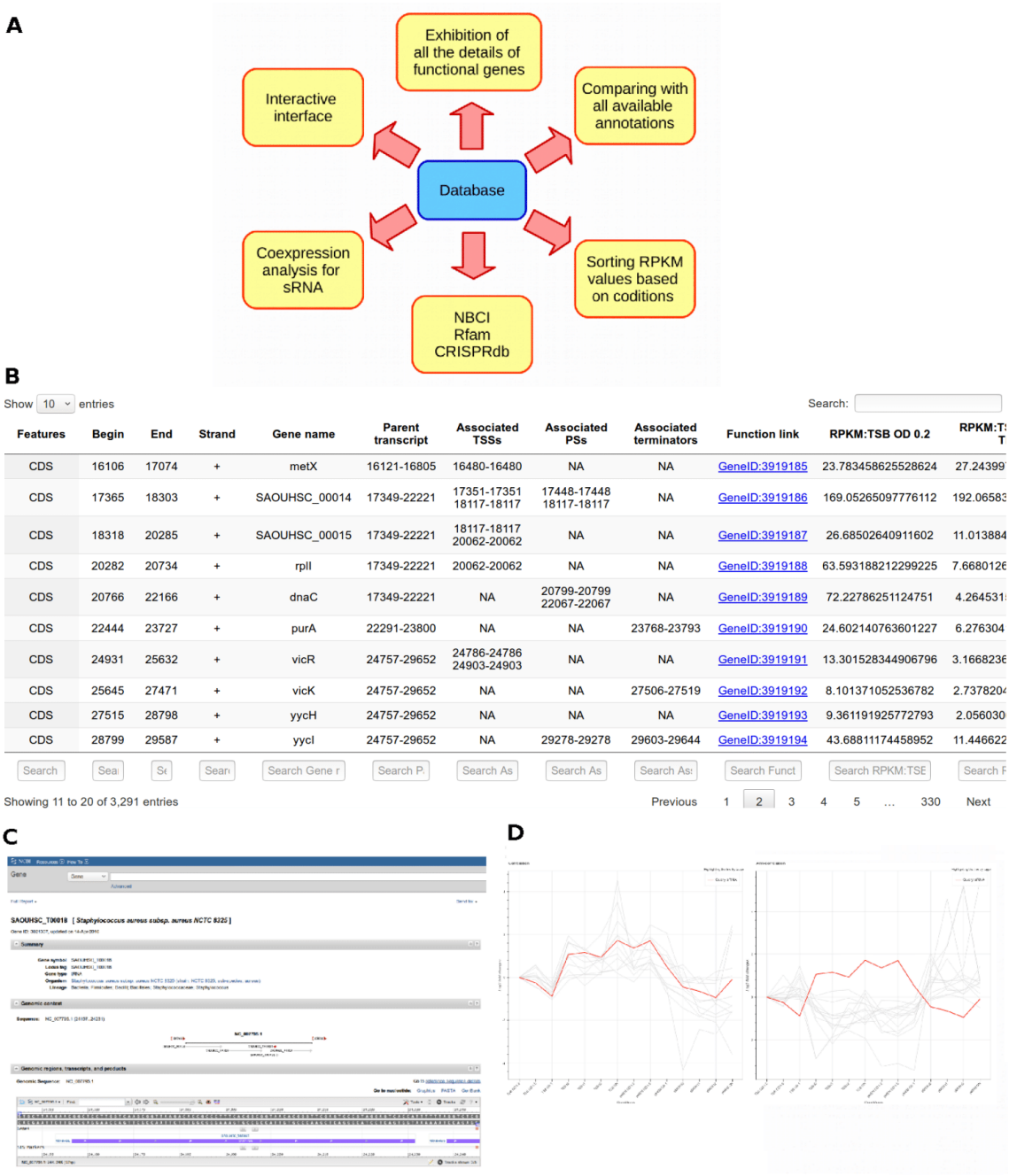


Figure 3.32: Screenshots of the interactive table of *S. aureus* HG003. (A) The interactive table allow one to browse the detailed information of functional genes, compare the annotations, sort the RPKM values, link to public databases, and show the results of co-expression analysis for sRNAs. (B) The RPKM values can be sorted by clicking the headers of RPKM values. The information of the functions of CDSs, riboswitches, RNA thermometers, and CRISPRs can be checked by connecting to public database (presented by (C)). Moreover, the results of co-expression analysis of sRNAs can be viewed like (D).

Chapter 4

Discussion

The achivements of ANNOgesic

ANNOgesic is the first tool developed for the detection of multiple bacterial and archaeal genomic features based on RNA-Seq data. It can predict all genomic features for a strain systematically. The biasses and shortages of identifying the genomic features by using different tools separately can be significantly reduced. It integrates a number of novel methods developed for detecting the genomic features which cannot be detected by previously available tools, and improved third-party tools by removing false positives and parameter optimization. ANNOgesic does not only generate precise genome annotations, but also provides numerous useful statistic analyses and visualizations. Furthermore, ANNOgesic is a flexible modular tool with a consistent and user friendly interface. It has been widely and successfully applied to many bacterial and archaeal genomes. Based on the application of ANNOgesic, numerous gene candidates as well as their potential functions can be found, and

many hypotheses can be made for experimental testing based on them.

sRNAs missed by using ANNOgesic

One of the core modules of ANNOgesic is the sRNA detection which has a high accuracy and sensitivity as shown in benchmarking with published sRNA sets. For examples, ANNOgesic can detect 80% to 90% of previously reported sRNAs in *S. aureus* HG003, *E. coli* K-12, *H. pylori* 26695, and *C. jejuni* 81116. Although the majority of the previously published sRNAs can be found by using ANNOgesic, several known sRNAs were still missed in the ANNOgesic analysis. The missing sRNAs can be classified into two classes. The first class is the set of the lowly expressed sRNAs (Figure 4.1A). Although these sRNAs can be detected by decreasing the cutoff of read coverage, the number of false positives would also be increased. Moreover, some of these low expressed published sRNAs are only detected by RNA-Seq but not by Northern blot or RT-PCR. The final class, the sRNAs are not associated with any TSSs (Figure 4.1B). Although ANNOgesic can detect sRNAs without using dRNA-Seq data, false positive rate would be also increased. These three classes reveal a trade between false-positive and false-negative rates. Without experimental validations, it is difficult to set proper thresholds. Thus, in order to provide a reliable sRNA set for the selection of experimental validations, ANNOgesic generates two lists of sRNA candidates - one list contains the sRNAs that passed all the filters like having a TSS associated with it. The other list covers all sRNAs without filtering. These two lists can be beneficial for the priority of sRNAs experimental validations.

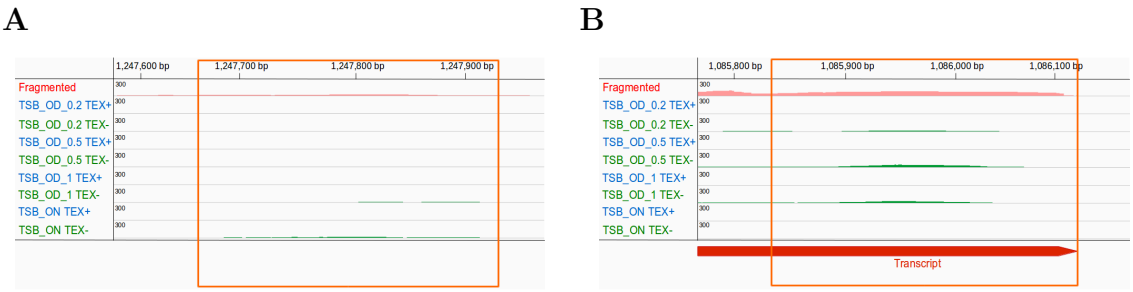


Figure 4.1: The published sRNAs missed by using ANNOgesic. The library of RNA-Seq generated after transcript fragmentation, TEX+ and TEX- libraries of dRNA-Seq are presented as the pink, blue and green coverages, respectively. The orange rectangles indicate the region of the previously reported sRNAs. **(A)** The previously reported sRNAs show low expression. **(B)** The published sRNAs are not associated with any TSSs. Although a transcript (red bar) can be detected, the sRNA can not be identified by ANNOgesic analysis. These cases are from *S. aureus* HG003.

Requirement for an automatic function detection in a gene co-expression analysis

Since the potential functions of sRNAs may be related to their co-expressed or inversely expressed genes detected by applying gene co-expression analysis, investigation of the functions of these co-expressed and inversely expressed genes is a fruitful approach. However, a time-consuming manual detection for characterizing gene functions still needs to be performed. Although GO terms of the genes in a gene clusters can be detected in an automatic way, numerous gene clusters possess diverse GO terms. For the cases without inconsistent GO terms allocation, ANNOgesic is an useful tool for automatically annotating the functions for gene clusters.

Comparison between sRNA target prediction and gene co-expression analysis

Since sRNA target prediction and gene co-expression analysis are both methods for characterizing and understanding the functions of sRNAs, a comparison between these two analyses was performed in this study. However, only 3% sRNA targets were detected by the both methods (Figure 4.2). It may be result from the low accuracy of sRNA target prediction tools since their recall is lower than 80% (some tools are even lower than 60%) [43]. Although numerous filters were applied to sRNA target prediction of ANNOgesic for removing false positives, their occurrences cannot be excluded completely. Furthermore, a biological pathway can be directly and indirectly controlled by multiple regulators. On the other hands, a regulator can regulate numerous interactions. Due to these reasons, the results between sRNA target prediction and gene co-expression analysis are inconsistent. Thus, applying and comparing these two methods is necessary to understanding and characterizing functions of sRNAs.

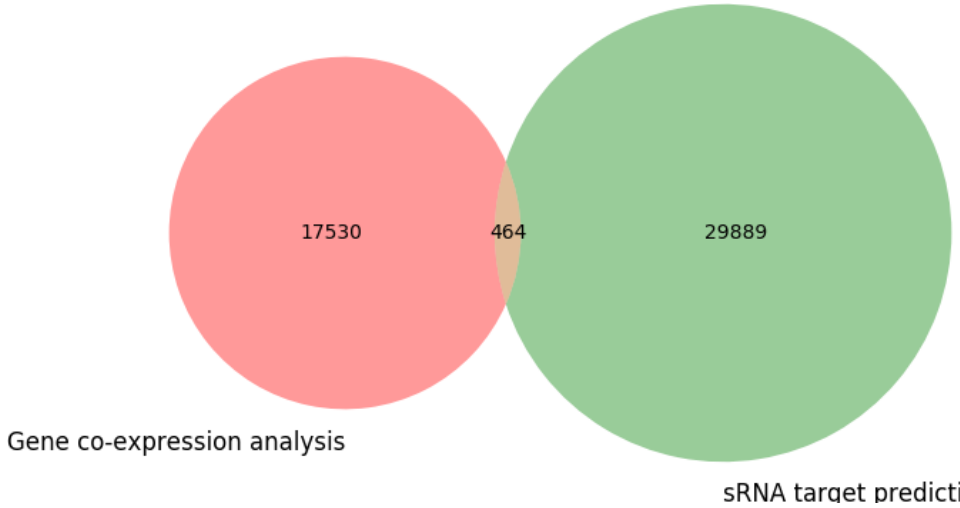
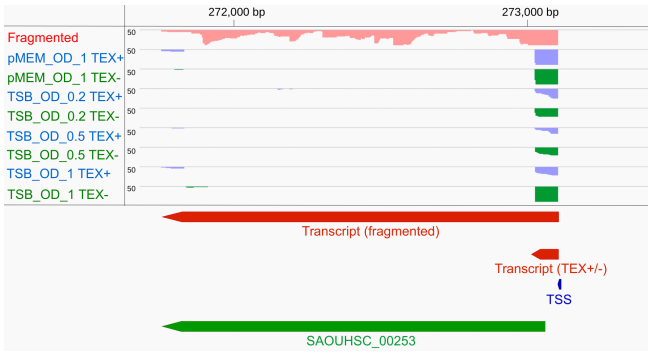


Figure 4.2: Overlapping of sRNA target prediction and gene co-expression analysis.

Advantages of using RNA-Seq data generated with multiple protocols

Although ANNOgesic can process RNA-Seq data from multiple protocols to generate precise annotations, for the majority of species, only data sets from a single RNA-Seq protocol are available. Due to this, ANNOgesic can also generate genome annotations with such limited single method. Since several genomic features can be detected much more precisely by applying the specific RNA-Seq protocols, applying ANNOgesic to the RNA-Seq data from a single protocol will negative influence the results. For examples, the 3' end of transcript boundary may not be identified precisely without the data from RNA- Seq generated after transcript fragmentation (Figure 4.3A). In addition, TSSs, especially internal TSSs, can not be predicted without dRNA-Seq data (Figure 4.3B).

A



B

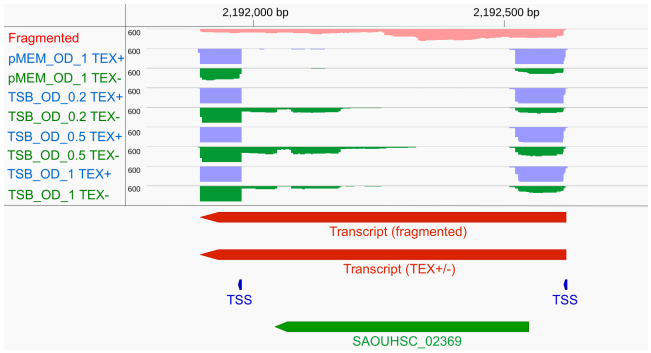


Figure 4.3: Examples of the comparison between the data from dRNA-Seq and RNA-Seq generated after transcript fragmentation in *S. aureus* HG003. The pink, blue and green coverages represent the library of RNA-Seq generated after transcript fragmentation, TEX+ libraries and TEX- libraries of dRNA-Seq, respectively. The red bars, green bars and blue spots represent transcripts, CDSs (SAOUHSC_00253: 271580 to 273103 at the reverse strand, SAOUHSC_02369: 2192012 to 2192542 at the reverse strand,) and TSSs **(A)** Fragmented libraries are a benefit for detecting the 3' end of the transcript. However, the length of transcript will be underestimated if only dRNA-Seq data was used. **(B)** dRNA-Seq data is used to identify TSSs with high resolution, especially internal TSSs which cannot be detected based on only fragmented libraries.

Choice of parameters

Setting proper cutoffs is an important step for detecting genomic features by using ANNOgesic such as transcript detection, sRNA detection, etc. On the one hand, using the cutoff makes ANNOgesic more flexible to meet users' requirements; on the other hand, an inappropriate setting may influence the predictions and generate misleading annotations. As displayed in Figure 4.4A, the annotations of transcripts depend on different read coverage cutoffs. Moreover, the result of gene co-expression analysis is significantly influenced by the cutoff of Spearman correlation coefficient. Using a loose cutoff may generate numerous false positives increasing the difficulty for characterizing the potential functions. However, applying a strict cutoff to gene co-expression analysis may give rise to misleading results or even hinder the discovery of the functions of the queried sRNAs due to a lack of associated genes. In principle, applying ROC curve (receiver operating characteristic curve) is the ideal way to set cutoffs by plotting true positive rate (TPR) against the false positive rate (FPR). However, plotting ROC curve can not be performed since no golden standard exists currently. Thus, the cutoffs still need to be adjusted for specific genomes or sRNAs by the user. For example, the default setting 0.77 (97.5 percentile) and -0.77 (2.5 percentile) as the cutoff of positive and negative correlation coefficients for gene co-expression analysis, respectively. Based on this setting, an iron-transportation related group (Figure 4.4B and C) contains a 30S ribosome protein S7 (*rpsG*) which is neither related to iron-transportation nor co-expressed with the members of the group perfectly. If the cutoff were set as 0.79 for positive correlation coefficient, *rpsG*

would be removed from the group (Figure 3.22H).

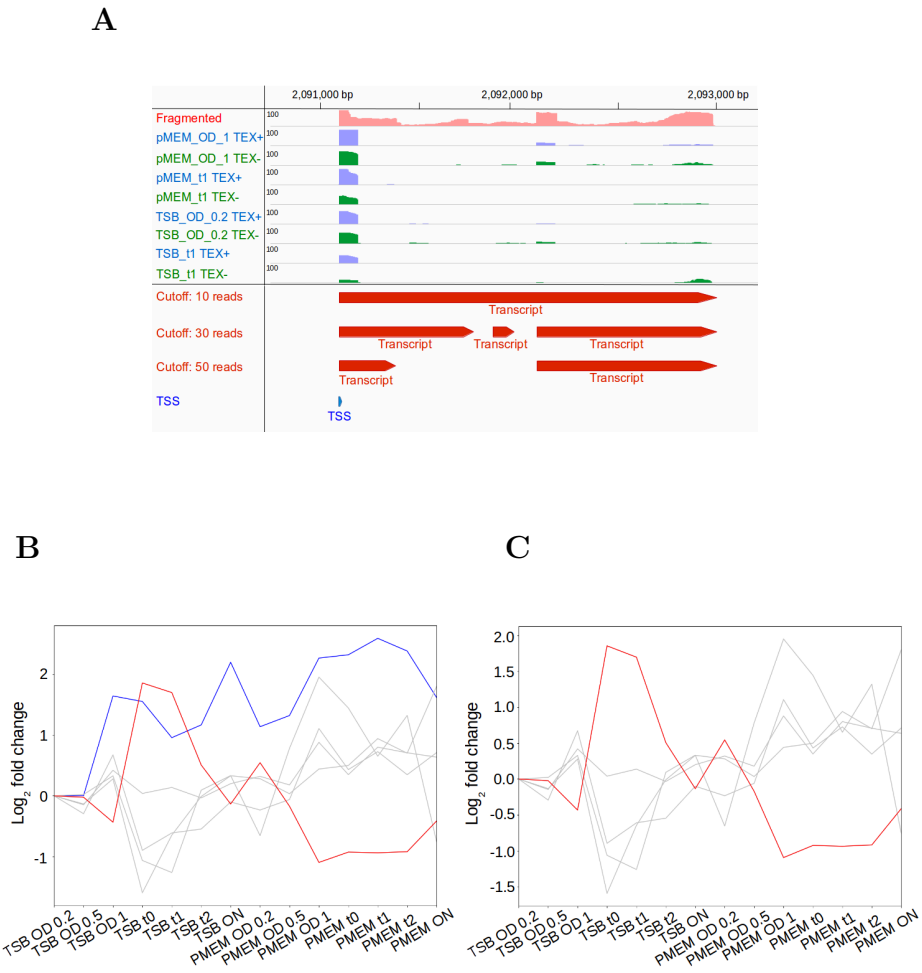


Figure 4.4: Examples for cutoff settings of ANNOgesic. **(A)** Based on the cutoffs of read coverages, the results of transcript detections can be different. The library of RNA-Seq generated after transcript fragmentation, TEX+ libraries and TEX- libraries of dRNA-Seq are presented as the pink, blue and green coverages, respectively. The transcripts from the upper track to the bottom track were detected by setting cutoff of minimum read coverages as 10, 30 and 50. The last track of annotation is for TSSs. **(B)** and **(C)** The kinetic curves of the genes anti-correlated with a novel sRNA which may regulate iron-transportation. The cutoffs of Spearman correlation coefficient of **(B)** and **(C)** are set as -0.77 and -0.79, respectively. The red, gray and blue lines represent the queried novel sRNA, the genes revealed anti-correlated expression with the queried sRNA, and 30S ribosome protein S7 *(*rpsG), respectively.

Pitfalls and limitations of ANNOgesic

Although ANNOgesic was successfully applied to many bacterial and achaeal genomes for generating high-quality genome annotations, a few pitfalls and limitations still exist. Until now, only few of the sRNAs newly predicted by ANNOgesic have been experimentally being validated. Therefore, the number of false postives may be underestimated. Moreover, some of the ANNOgesic's predictions are based on the genome annotations which can be retrieved from public database. Since the naming system of genome annotations is not well defined, diverse names of the same genomic feature and misannotations sometimes happen. The accuracy of the predictions may be influenced by the incorrect genome annotations.

An obvious shortage is that ANNOgesic integrates more than 20 third-party tools which need to be installed one after another. This large number of dependencies come with certain effort during the setup. For examples, the paths of the executive files, environment variable settings, and the versions of the tools need to be managed. In order to overcome this shortcoming, a Docker image [92] that contains all software dependencies is provided. By the application of Docker image, ANNOgesic can be installed and executed in any machine that supports Docker.

Although ANNOgesic can detect numerous genomic features in high resolution, the running time of several modules of ANNOgesic are relative long such as sRNA target prediction and PPI network detection. In fact, most of these modules spend a lot of time on running the third-party software like RNAup [48, 49] for sRNA target prediction. Excluding the time for running the third-party software, all genomic

features of *S. aureus* HG003 (2,821,354 base pairs) with 29 RNA-Seq libraries (around 5 million reads per library) can be detected within one day on a mid-sized server.

Since some genomic feature detections rely strongly on the information of other genomic features like sRNA detection which requires TSS or PS information, the accuracy of a genomic feature detection may be influenced by other features' predictions. Although ANNOgesic improved the performances of the previously available tools, several genomic features still cannot be detected precisely without applying some specific RNA-Seq protocols such as using Term-Seq [13] for terminator and riboswitch detections and ribosome profiling [14] for sORF prediction. Therefore, integrating more results from RNA-Seq based protocols into ANNOgesic may raise the accuracy of the specific genomic feature detections significantly.

Perspectives

In previous publications have shown that Term-Seq [13] and ribosome profiling [14] can be applied to detect several genomic features and improve genome annotations. Using Term-Seq, not only the annotations of Rho-independent terminators and riboswitches in high resolution, but also the novel ones that cannot be found by applying ANNOgesic, can be identified. Ribosome profiling, which can be applied for detecting the transcripts undergoing translation based on the short mRNA sequences bound to ribosomes, can be beneficial for improving the identification of sORFs. These two protocols can also improve a lot of detections which depend on the information of terminators and sORFs, such as operons, UTRs, and sRNAs. Therefore, these two protocol can be used to extend ANNOgesic in the future.

Based on the application of ANNOgesic, numerous novel sRNAs have been detected in this study. In order to validate these novel sRNAs, RT-PCR or Northern plot need to be done. Moreover, to understand the functions of sRNAs, gene co-expression analysis was used and many potential functions of sRNAs were characterized. Based on those predictions, knock-out experiments can be performed to validate the functions of sRNAs. This will be one of the most important follow-up tasks.

Third generation sequencing technologies can generate sequencing reads with a different approach from second generation platforms. It can produce sequencing reads in unprecedented lengths which can strongly increase the quality of genome assemblies. Since the importance of the applications of third generation sequencing platforms like Nanopore and PacBio raise quickly, adapting ANNOgesic to be able to handle long read data and may improve the quality of its predictions [173].

Conclusion

In my doctoral study, a tool of generating RNA-Seq-based annotations for bacterial and archaeal genomes, ANNOgesic, was developed. Numerous comparisons between the predictions of ANNOgesic and published datasets were done, and high performance of the tools was shown in this study. The genome sequence and an extensive annotations of *S. aureus* HG003, which is a potential model strain for studying both virulence and antibiotic resistance, were generated by applying ANNOgesic. Both ANNOgesic and the information of genomic features of *S. aureus* HG003 may help for the community of microbiology.

References

[1] Z. Wang, M. Gerstein, and M. Snyder. RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2017.

[2] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, 2009.

[3] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9:357–359, 2012.

[4] S. Hoffmann, Christian Otto, Stefan 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 Computational Biology*, 5(9):e1000502, 2009.

[5] 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 (Oxford, England)*, 29(1):15–21, 2013.

[6] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman, and A. Regev. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nature Protocols*, 8(8):1494–1512, 2013.

[7] B. Tjaden. De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biology*, 16(1):1, 2015.

[8] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1), 2016.

[9] K. U. Förstner, J. Vogel, and C. M. Sharma. READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data. *Bioinformatics (Oxford, England)*, 30(23):3421–3423, 2014.

[10] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 2014.

[11] C. M. Sharma and J. Vogel. Differential RNA-seq: the approach behind and the biological insight gained. *Current Opinion in Microbiology*, 19:97–105, 2014.

[12] 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, 2016.

[13] 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(6282):aad9822–aad9822, 2016.

[14] N. T. Ingolia. Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Reviews. Genetics*, 15(3):205–213, 2014.

[15] S. Melamed, A. Peer, R. Faigenbaum-Romm, Y. E. Gatt, N. Reiss, A. Bar, Y. Altuvia, L. Argaman, and H. Margalit. Global mapping of small rna-target interactions in bacteria. *Molecular Cell*, 63(5):884–897, 2016.

[16] R. B. Darnell. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley interdisciplinary reviews. RNA*, 1(2):266–286, 2010.

[17] J. Zhao, T. K. Ohsumi, J. T. Kung, Y. Ogawa, D. J. Grau, Kavitha Sarma, J. J. Song, R. E. Kingston, M. Borowsky, and J. T. Lee. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular Cell*, 40(6):939–953, 2010.

[18] 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. *Proceedings of the National Academy of Sciences*, 113(41):11591–11596, 2016.

[19] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23):4636–4641, 1999.

[20] 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 Research*, 33:W686–W689, 2005.

[21] 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 Research*, 35(9):3100–3108, 2007.

[22] T. Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, 30(14):2068–2069, 2014.

[23] T. Weinmaier, A. Platzer, J. Frank, H.-J. Hellinger, P. Tischler, and T. Rattei. ConsPred: a rule-based (re-)annotation framework for prokaryotic genomes: Table 1. *Bioinformatics*, page btw393, 2016.

[24] 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 genetics*, 9(5):e1003495, 2013.

[25] 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.

[26] H. Jorjani and M. Zavolan. TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data. *Bioinformatics (Oxford, England)*, 30(7):971–974, 2014.

[27] 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 Research*, 41(14):e140, 2013.

[28] X. Chen, W.-C. Chou, Q. Ma, and Y. Xu. SeqTU: A web server for identification of bacterial transcription units. *Scientific Reports*, 7:43925, 2017.

[29] S. C. Forster, A. M. Finkel, J. A. Gould, and P. J. Hertzog. RNA-eXpress annotates novel transcript features in RNA-seq data. *Bioinformatics (Oxford, England)*, 29(6):810–812, 2013.

[30] E. Sallet, J. Gouzy, and T. Schiex. EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes. *Bioinformatics (Oxford, England)*, 30(18):2659–2661, 2014.

[31] G. Storz, J. Vogel, and K. M. Wassarman. Regulation by small rnas in bacteria: Expanding frontiers. *Molecular Cell*, 43(6):880–891, 2011.

[32] L. S. Waters and G. Storz. Regulatory rnas in bacteria. *Cell*, 136(4):615–628, 2009.

[33] E. Holmqvist and E. G. H. Wagner. Impact of bacterial srnas in stress responses. *Biochem Soc Trans*, 45(6):880–891, 2017.

[34] Y. Chao and J. Vogel. A 3' utr-derived small rna provides the regulatory noncoding arm of the inner membrane stress response. *Molecular Cell*, 61(3):352–363, 2016.

[35] 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. *The EMBO journal*, 31(20):4005–4019, 2012.

[36] Elena Rivas, Robert J. Klein, Thomas A. Jones, and Sean R. Eddy. Computational identification of noncoding RNAs in e. coli by comparative genomics. *Current Biology*, 11(17):1369–1373, 2001.

[37] C. Pichon and B. Felden. Intergenic sequence inspector: searching and identifying bacterial RNAs. *Bioinformatics*, 19(13):1707–1709, 2003.

[38] Stefan Washietl, Ivo L. Hofacker, and Peter F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2454–2459, 2005.

[39] J. Livny. sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Research*, 33(13):4096–4105, 2005.

[40] Jonathan Livny, Hidayat Teonadi, Miron Livny, and Matthew K. Waldor. High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. *PLoS ONE*, 3(9):e3197, 2008.

[41] Y. Y. Leung, P. Ryvkin, L. H. Ungar, B. D. Gregory, and L.-S. Wang. CoRAL: predicting non-coding RNAs from small RNA-sequencing data. *Nucleic Acids Research*, 41(14):e137–e137, 2013.

[42] Jayavel Sridhar, Suryanarayanan Ramkumar Narmada, Radhakrishnan Sabarinathan, Hong-Yu Ou, Zixin Deng, Kanagaraj Sekar, Ziauddin Ahamed Rafi, and Kumar Rajakumar. sRNAscanner: A computational tool for intergenic small RNA detection in bacterial genomes. *PLoS ONE*, 5(8):e11970, 2010.

[43] A. Pain, A. Ott, H. Amine, T. Rochat, P. Bouloc, and D. Gautheret. An assessment of bacterial small RNA target prediction programs. *RNA biology*, 12(5):509–513, 2015.

[44] S. U. Umu and P. P. Gardner. A comprehensive benchmark of RNAâĂŞRNA interaction prediction tools for all domains of life. *Bioinformatics*, 33(7):btw728, 2016.

[45] P. R. Wright, J. Georg, M. Mann, D. A. Sorescu, A. S. Richter, S. Lott, R. Kleinkauf, W. R. Hess, and R. Backofen. CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains. *Nucleic Acids Research*, 42:W119–W123, 2014.

[46] 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.

[47] H. Tafer and I. L. Hofacker. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics (Oxford, England)*, 24(22):2657–2663, 2008.

[48] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA package 2.0. *Algorithms for molecular biology: AMB*, 6:26, 2011.

[49] U. Mücksteine, H. Tafer, S. H. Bernhart, M. Hernandez-Rosales, J. Vogel, P. F. Stadler, and I. L. Hofacker. Translational control by RNA-RNA interaction: Improved computation of RNA-RNA binding thermodynamics. *Bioinformatics Research and Development*, 13:114–127, 2008.

[50] A.K. Dubey, C.S. Baker, T. Romeo, and P. Babitzke. Rna sequence and secondary structure participate in high-affinity csra-rna interaction. *RNA*, 11(10):1579–1587, 2005.

[51] O. Duss, E. Michel, N. Diarra dit Konté, M. Schubert, and F.H. Allain. Molecular basis for the wide range of affinity found in csr/rsm protein-rna recognition. *Nucleic Acids Research*, 42(8):5332–5346, 2014.

[52] R. B. Denman. Using RNAFOLD to predict the activity of small catalytic RNAs. *BioTechniques*, 15(6):1090–1095, 1993.

[53] Z. Yao, Z. Weinberg, and W. L. Ruzzo. CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics (Oxford, England)*, 22(4):445–452, 2006.

[54] N. R. Markham and M. Zuker. UNAFold: software for nucleic acid folding and hybridization. *Methods in Molecular Biology (Clifton, N.J.)*, 453:3–31, 2008.

[55] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45:D353–D361, 2017.

[56] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[57] The Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research*, 2014.

[58] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, pages 418–429, 2000.

[59] P. Langfelder and S. Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

[60] 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(7286):250–255, 2010.

[61] 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 Research*, 38(19):6620–6636, 2010.

[62] O. Wurtzel, R. Sapra annd F. Chen, Y. Zhu, B. A. Simmons, and R. Sorek. A single-base resolution map of an archaeal transcriptome. *Genome Research*, 20(1):133–141, 2010.

[63] 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(9):1760–1774, 2012.

[64] M. Sassi, Y. Augagneur, T. Mauro, L. Ivain, S. Chabelskaya, M. Hallier, O. Sallou, and B. Felden. SRD: a staphylococcus regulatory RNA database. *RNA (New York, N.Y.)*, 2015.

[65] R. K. Carroll, A. Weiss, W. H. Broach, R. E. Wiemels, A. B. Mogen, K. C. Rice, and L. N. Shaw. Genome-wide annotation, identification, and global transcriptomic analysis of regulatory or small RNA gene expression in *Staphylococcus aureus*. *mBio*, 7(1):e01990–15, 2016.

[66] S. Herbert, A.-K. Ziebandt, K. Ohlsen, T. Schäfer, M. Hecker, D. Albrecht, R. Novick, and F. Götz. Repair of global regulators in staphylococcus aureus 8325 and comparative analysis with other clinical isolates. *Infection and Immunity*, 78(6):2877–2889, 2010.

[67] K. Plata, A. E. Rosato, and G. Wegrzyn. Staphylococcus aureus as an infectious agent: overview of biochemistry and molecular genetics of its pathogenicity. *Acta Biochimica Polonica*, 56(4):597–612, 2009.

[68] J. Hacker and J.B. Kaper. Pathogenicity islands and the evolution of microbes. *Annual Review of Microbiology*, 54:641âĂŞ679, 2000.

[69] C. Ubeda, P. Barry, J. R. Penades, and R. P. Novick. A pathogenicity island replicon in staphylococcus aureus replicates as an unstable plasmid. *Proceedings of the National Academy of Sciences*, 104(36):14182–14188, 2007.

[70] C. Pichon and B. Felden. Small RNA genes expressed from staphylococcus aureus genomic and pathogenicity islands with specific expression among pathogenic strains. *Proceedings of the National Academy of Sciences*, 102(40):14249–14254, 2005.

[71] P. van Helden. Data-driven hypotheses. *EMBO reports*, 14(2):104, 2013.

[72] F. Mazzocchi. Could big data be the end of theory in science? *EMBO reports*, 16(10):1250–1255, 2015.

[73] 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(4), 2016.

[74] B. Remes, T. Rische-Grahl, K.M.H. Müller, K.U. Fürstner, S.H. Yu, L. Weber, A. Jäger, V. Peuser, and G. Klug. An RpoHI-dependent response promotes outgrowth after extended stationary phase in the alphaproteobacterium *Rhodobacter sphaeroides*. *Journal of Bacteriology*, pages JB.00249–17, 2017.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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 Research*, 37(16):e106–e106, 2009.

[83] 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 Biology*, 11(5):R57, 2010.

[84] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11):1422–1423, 2009.

[85] S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[86] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[87] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science conference*, pages 11–15, 2008.

[88] Git repository of annogesic. `https://github.com/Sung-Huan/ANNOgesic`.

[89] Documentation of annogesic. `http://annogesic.readthedocs.io/en/latest/index.html`.

[90] pip3. `https://pip.pypa.io`.

[91] Docker image of annogesic. `https://hub.docker.com/r/silasysh/annogesic`.

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

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

[94] 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 (Oxford, England)*, 25(16):2078–2079, 2009.

[95] T. D. Otto, G. P. Dillon, W. S. Degrave, and M. Berriman. RATT: Rapid annotation transfer tool. *Nucleic Acids Research*, 39(9):e57, 2011.

[96] 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 Biology*, 8(2):R22, 2007.

[97] Timothy L. Bailey, Nadya Williams, Chris Misleh, and Wilfred W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34:W369–373, 2006.

[98] M. C. Frith, N. F. W. Saunders, B. Kobe, and T. L. Bailey. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Computational Biology*, 4(5):e1000071, 2008.

[99] R. P Huntley, T. Sawford, M. J. Martin, and C. O'Donovan. Understanding how and why the gene ontology and its annotations evolve: the GO within UniProt. *GigaScience*, 3(1):4, 2017-05-24.

[100] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45:D158–D169, 2017-05-24.

[101] 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.

[102] 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 Research*, 36:W411–415, 2008.

[103] J. L. Gardy. PSORT-b: improving protein subcellular localization prediction for gram-negative bacteria. *Nucleic Acids Research*, 31(13):3613–3617, 2003.

[104] 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 (Oxford, England)*, 26(13):1608–1615, 2010.

[105] 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 Biology*, 15(2):R34, 2014.

[106] E. P. Nawrocki and S. R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics (Oxford, England)*, 29(22):2933–2935, 2013.

[107] 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 Research*, 43:D130–137, 2014.

[108] 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.

[109] H. Tang, D. Klopfenstein, B. Pedersen, P. Flick, K. Sato, F. Ramirez, J. Yunes, and C. Mungall. GOATOOLS: Tools for gene ontology. *Zenodo*, 2015.

[110] M. Sassi, B. Felden, and Y. Augagneur. Draft genome sequence of staphylococcus aureus subsp. aureus strain HG003, an NCTC8325 derivative. *Genome Announcements*, 2(4), 2014.

[111] M. Bouvier, C. M. Sharma, F. Mika, K. H. Nierhaus, and J. Vogel. Small RNA binding to 5' mRNA coding region inhibits translational initiation. *Molecular Cell*, 32(6):827–837, 2008.

[112] 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*. *The EMBO Journal*, 35(9):991–1011, 2016.

[113] M. Miyakoshi, Y. Chao, and J. Vogel. Regulatory small RNAs from the 3âĂš regions of bacterial mRNAs. *Current Opinion in Microbiology*, 24:132–139, 2015.

[114] D. L. Kaplan and M. O'Donnell. Rho factor: Transcription termination in four steps. *Current Biology*, 13(18):R714–R716, 2003.

[115] R. S. Washburn, A. Marra, A. P. Bryant, M. Rosenberg, and D. R. Gentry. rho is not essential for viability or virulence in staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 45(4):1099–1103, 2001.

[116] A. Ray-Soni, M. J. Bellecourt, and R. Landick. Mechanisms of bacterial transcription termination: All good things must end. *Annual Review of Biochemistry*, 85:319–347, 2016.

[117] P. J. Farnham and T. Platt. Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription in vitro. *Nucleic Acids Research*, 9(3):563–577, 1981.

[118] P. P. Gardner, L. Barquist, A. Bateman, E. P. Nawrocki, and Z. Weinberg. RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic Acids Research*, 39(14):5845–5852, 2011.

[119] M. Mandal and R. R. Breaker. Gene regulation by riboswitches. *Nature Reviews Molecular Cell Biology*, 5(6):451–463, 2004.

[120] J. Kortmann and F. Narberhaus. Bacterial RNA thermometers: molecular zippers and switches. *Nature Reviews Microbiology*, 10(4):255–265, 2012.

[121] G. Grillo, A. Turi, F. Licciulli, F. Mignone, S. Liuni, S. Banfi, V. A. Gennarino, D. S. Horner, G. Pavesi, E. Picardi, and G. Pesole. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Research*, 38:D75–D80, 2010.

[122] D. F. Browning and S. J. W. Busby. The regulation of bacterial transcription initiation. *Nature Reviews Microbiology*, 2(1):57–65, 2017-05-04.

[123] C. B. Harley and R. P. Reynolds. Analysis of e. coli promoter sequences. *Nucleic Acids Research*, 15(5), 1987.

[124] B. Taboada, R. Ciria, C. E. Martinez-Guerrero, and E. Merino. ProOpDB: Prokaryotic operon DataBase. *Nucleic Acids Research*, 40:D627–D631, 2012.

[125] M. Y. Galperin, K. S. Makarova, Y. I. Wolf, and E. V. Koonin. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, 43:D261–269, 2015.

[126] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman. The pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44:D279–D285, 2016.

[127] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, and Lars J. Jensen. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41:D808–815, 2013.

[128] 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.

[129] 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 Research*, 41:D233–238, 2013.

[130] U. Mäder, P. Nicolas, M. Depke, J. Pané-Farré, M. Debarbouille, M. M. van der Kooi-Pol, C. Guérin, S. Dérozier, A. Hiron, H. Jarmer, A. Leduc, S. Michalik, E. Reilman, M. Schaffer, F. Schmidt, P. Bessières, P. Noirot, M. Hecker, T. Msadek, U. Völker, and J. M. van Dijl. Staphylococcus aureus transcriptome architecture: From laboratory to infection-mimicking conditions. *PLoS genetics*, 12(4):e1005962, 2016.

[131] J. M. Morrison, E. W. Miller, M. A. Benson, F. Alonzo, P. Yoong, V. J. Torres, S. H. Hinrichs, and P. M. Dunman. Characterization of SSR42, a novel virulence factor regulatory RNA that contributes to the pathogenesis of a staphylococcus aureus USA300 representative. *Journal of Bacteriology*, 194(11):2924–2938, 2012.

[132] L. Feng, S. T. Rutherford, K. Papenfort, J. D. Bagert, J. C. van Kessel, D. A. Tirrell, N. S. Wingreen, and B. L. Bassler. A qrr noncoding rna deploys four different regulatory mechanisms to optimize quorum-sensing dynamics. *Cell*, 160(1-2):228–240, 2015.

[133] Y. Sun and C. K. Vanderpool. Physiological consequences of multiple-target regulation by the small rna sgrs in *Escherichia coli*. *Journal of Bacteriology*, 195(21):4804–4815, 2013.

[134] A. J. Westermann, K. U. Förstner, F. Amman, L. Barquist, Y. Chao, L. N. Schulte, L. Müller, R. Reinhardt, P. F. Stadler, and J. Vogel. Dual rna-seq unveils noncoding rna functions in hostâĂŞpathogen interactions. *Nature*, 529(7587):496–501, 2016.

[135] E. Morfeldt, D. Taylor, A. von Gabain, and S. Arvidson. Activation of alpha-toxin translation in *Staphylococcus aureus* by the trans-encoded antisense rna, rnaiii. *EMBO Journal*, 14(18):4569–4577, 1995.

[136] Y. Benito, F.A. Kolb, P. Romby, G. Lina, J. Etienne, and F. Vandenesch. Probing the structure of rnaiii, the *Staphylococcus aureus agr* regulatory rna, and identification of the rna domain involved in repression of protein a expression. *RNA*, 6(5):668–679, 2000.

[137] S. Boisset, T. Geissmann, E. Huntzinger, P. Fechter, N. Bendridi, M. Possedko, C. Chevalier, A.C. Helfer, Y. Benito, A. Jacquier, C. Gaspin, F. Vandenesch, and P. Romby. *Staphylococcus aureus* rnaiii coordinately represses the synthesis of virulence factors and the transcription regulator rot by an antisense mechanism. *Genes Development*, 21(11):1353–1366, 2007.

[138] E. Geisinger, R.P. Adhikari, R. Jin, H.F. Ross, and R.P. Novick. Inhibition of rot translation by rnaiii, a key feature of agr function. *Molecular Microbiology*, 61(4):1038–1048, 2006.

[139] C. Chevalier, S. Boisset, C. Romilly, B. Masquida, P. Fechter, T. Geissmann, F. Vandenesch, and P. Romby. *Staphylococcus aureus* rnaiii binds to two distant

regions of coa mrna to arrest translation and promote mrna degradation. *PLOS Pathogens*, 6(3):e1000809, 2010.

[140] P. M. Dunman, E. Murphy, S. Haney, D. Palacios, G. Tucker-Kellogg, S. Wu, E. L. Brown, R. J. Zagursky, D. Shlaes, , and S. J. Projan. Transcription profiling-based identification of *Staphylococcus aureus* genes regulated by the agr and/or sara loci. *Journal of Bacteriology*, 183(24):7341–7353, 2001.

[141] H. Schmidt and M. Hensel. Pathogenicity islands in bacterial pathogenesis. *Clinical Microbiology Reviews*, 17(1):14–56, 2004.

[142] A. Ritter, G. Blum, L. Emödy, M. Kerenyi, A. Böck, B. Neuhierl, W. Rabsch, F. Scheutz, and J. Hacker. trna genes and pathogenicity islands: influence on virulence and metabolic properties of uropathogenic escherichia coli. *Molecular Microbiology*, 17(1):109–121, 1995.

[143] M.L. Pinel-Marie, R. Brielle, and B. Felden. Dual toxic-peptide-coding staphylococcus aureus RNA under antisense regulation targets host cells and bacterial rivals unequally. *Cell Reports*, 7(2):424–435, 2014.

[144] S. Jitrapakdee, M. St Maurice, I. Rayment, W. W. Cleland, J. C. Wallace, and P. V. Attwood. Structure, mechanism and regulation of pyruvate carboxylase. *Biochemical Journal*, 413(3):369–387, 2008.

[145] R. Ansell and L. Adler. The effect of iron limitation on glycerol production and expression of the isogenes for NAD $^{+}$ -dependent glycerol 3-phosphate dehydrogenase in *Saccharomyces cerevisiae*. *FEBS Letters*, 461(3):173–177, 1999.

[146] D. B. Friedman, D. L. Stauff, G. Pishchany, C. W. Whitwell, V. J. Torres, and E. P. Skaar. *Staphylococcus aureus* redirects central metabolism to increase iron availability. *PLoS Pathogens*, 2(8):e87, 2006.

[147] S. J. Andrews and J. A. Rothnagel. Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics*, 15:193–204, 2014.

[148] J. Peden. Analysis of codon usage. *PhD Thesis, University of Nottingham, UK.*, 1999.

[149] K. Hanada, K. Akiyama, T. Sakurai, T. Toyoda, K. Shinozaki, and S.H. Shiu. sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics*, 26(3):399–400, 2010.

[150] H. Cheng, W. S. Chan, Z. Li, D. Wang, S. Liu, and Y. Zhou. Small open reading frames: current prediction techniques and future prospect. *Current Protein & Peptide Science*, 12(6):503–507, 2011.

[151] J.E. Wilusz and P.A. Sharp. Molecular biology. a circuitous route to noncoding rna. *Science*, 340(6131):440–441, 2013.

[152] J. Salzman, C. Gawad, P. L. Wang, N. Lacayo, and P. O. Brown. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS ONE*, 7(2):e30733, 2012.

[153] M. Danan, S. Schwartz, S. Edelheit, and R. Sorek. Transcriptome-wide discovery of circular RNAs in archaea. *Nucleic Acids Research*, 40(7):3131–3142, 2012.

[154] J. Roßmanith and F. Narberhaus. Exploring the modular nature of riboswitches and RNA thermometers. *Nucleic Acids Research*, 44(11):5410–5423, 2016.

[155] J. D. Sander and J. K. Joung. CRISPR-cas systems for editing, regulating and targeting genomes. *Nature Biotechnology*, 32(4):347–355, 2014.

[156] J. L. Gardy and F. S. L. Brinkman. Methods for predicting bacterial protein subcellular localization. *Nature Reviews Microbiology*, 4(10):741–751, 2006.

[157] K. Nakai and M. Kanehisa. Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins: Structure, Function, and Genetics*, 11(2):95–110, 1991.

[158] Z. Lu, D. Szafron, R. Greiner, P. Lu, D.S. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, 20(4):547–556, 2004.

[159] S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, 2001.

[160] C. S. Yu, C. J. Lin, and J. K. Hwang. Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Science*, 13(5):1402–1406, 2004.

[161] M. Bhasin, A. Garg, and G. P. S. Raghava. PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics*, 21(10):2522–2524, 2005.

[162] R. Nair and B. Rost. Mimicking cellular sorting improves prediction of subcellular localization. *Journal of Molecular Biology*, 348(1):85–100, 2005.

[163] J. Wang, W. K. Sung, A. Krishnan, and K. B. Li. Protein subcellular localization prediction for gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. *BMC bioinformatics*, 6:174, 2005.

[164] S. Rey, J. L. Gardy, and F. S. L. Brinkman. Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. *BMC genomics*, 6:162, 2005.

[165] J. Zahiri, J. H. Bozorgmehr, and A. Masoudi-Nejad. Computational prediction of protein-protein interaction networks: Algo-rithms and resources. *Current Genomics*, 14(6):397–414, 2013.

[166] L. C. Xue, D. Dobbs, A. M.J.J. Bonvin, and V. Honavar. Computational prediction of protein interfaces: A review of data driven methods. *FEBS Letters*, 589(23):3516–3526, 2015.

[167] 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.

[168] 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.

[169] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, 2011.

[170] Bokeh Development Team. Bokeh: Python library for interactive visualization. *http://www.bokeh.pydata.org*, 2014.

[171] Y. Cui, X. Chen, H. Luo, Z. Fan, J. Luo, S. He, H. Yue, P. Zhang, and R. Chen. BioCircos.js: an interactive circos JavaScript library for biological data visualization on web applications. *Bioinformatics (Oxford, England)*, 32(11):1740–1742, 2016.

[172] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 45:D12–D17, 2017.

[173] M. Miyamoto, D Motooka, K Gotoh, T Imai, K Yoshitake, N Goto, T Iida, T Yasunaga, T Horii, K Arakawa, M Kasahara, and S Nakamura. Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics*, 21:699, 2014.

Appendix A

Appendix

Table A.1: Comparison between the published scaffolds of *Staphylococcus aureus* HG003 and the complete sequence generated by applying ANNOgesic

Scaffold IDs	Length of Scaffold	Start* (Scaffold)	End* (Scaffold)	Start (Complete)	End (Complete)	Identity
JPPU01000001	392186	97756	391821	1	294073	99.00%
JPPU01000001	392186	1	97755	2723598	2821354	83.00%
JPPU01000002	158759	1912	158759	292038	448884	96.00%
JPPU01000002	158759	1412	1903	289423	289914	95.00%
JPPU01000002	158759	888	1387	294081	294580	92.00%
JPPU01000002	158759	1	432	290492	290923	87.00%
JPPU01000002	158759	433	674	291453	291695	77.00%
JPPU01000003	109412	1	109412	2237297**	2127886**	98.00%
JPPU01000004	49454	1	49454	2122817	2073364	99.00%
JPPU01000005	166029	1	166029	2072430	1906402	98.00%
JPPU01000006	82030	1	82030	1897331	1815302	96.00%
JPPU01000007	107453	53470	107453	1760733	1706751	99.00%
JPPU01000007	107453	1	53908	1814362	1760455	97.00%
JPPU01000008	198693	1	198693	1705793	1507101	84.00%
JPPU01000009	711094	1	711094	1506304	795209	84.00%

Scaffold IDs	Length of Scaffold	Start* (Scaffold)	End* (Scaffold)	Start (Complete)	End (Complete)	Identity
JPPU01000010	243280	36464	170019	757478	623923	99.00%
JPPU01000010	243280	169563	239154	624246	554649	99.00%
JPPU01000010	243280	1	26834	795082	768251	99.00%
JPPU01000010	243280	26495	36467	768646	758674	99.00%
JPPU01000010	243280	238837	243280	554798	550349	84.00%
JPPU01000011	52090	1	52090	550092	498003	85.00%
JPPU01000012	39162	1	39162	493164	454003	97.00%
JPPU01000013	20707	1	20707	2244284	2264991	95.00%
JPPU01000014	455511	1	305732	2265943	2571675	99.00%
JPPU01000014	455511	304653	455510	2572516	2723373	96.00%
JPPU01000015	2945	1	2945	1901470	1898526	97.00%
JPPU01000016	1448	1	1448	1897205	1898652	100.00%
JPPU01000016	1448	1	1448	2237171	2238618	100.00%
JPPU01000016	1448	1	1448	1423793	2238618	97.00%
JPPU01000017	1206	1	1206	2266069	2264865	98.00%
JPPU01000017	1206	1	1206	1815428	1814236	98.00%
JPPU01000017	1206	1	1206	1705667	1706877	97.00%
JPPU01000018	1049	1	1049	2072443	2073490	99.00%
JPPU01000018	1049	1	1049	1506178	1507227	99.00%
JPPU01000019	5390	1	3122	453849	450729	100.00%
JPPU01000019	5390	3716	5390	1904854	1906528	100.00%
JPPU01000019	5390	1	3122	2122739	2125860	99.00%
JPPU01000019	5390	1	3122	498129	495008	99.00%
JPPU01000019	5390	1	3122	1901344	1904464	99.00%
JPPU01000019	5390	3716	5390	494727	493053	99.00%
JPPU01000019	5390	3716	5390	2126323	2127997	99.00%
JPPU01000019	5390	3716	5390	2242734	2244410	99.00%
JPPU01000019	5390	3716	5390	450448	448773	99.00%

*"Start" and "End" indicate the aligned region. "Scaffold" in the bracket means the position is for the published scaffolds, and "Complete" means the position is for the genome generated in this study.

**If the value of "Start" is larger than the value of "End", it means the scaffold was aligned on the reverse strand of the complete genome.

Table A.2: The co-expressed and inversely expressed genes of RNAIII

Co-expressed genes of RNAIII (from 2,093,091 to 2,093,248 at the reverse strand)				
Location	Strand[1]	Product[2]	Gene name [3]	C.C.[4]
10893-12407	+	histidine ammonia-lyase	hutH	0.89011
119492-120160	+	capsular polysaccharide biosynthesis protein	capA	0.77143
171258-172712	+	PTS system transporter	murP	0.82418
254123-254566	+	murein hydrolase regulator LrgA	lrgA	0.78022
259909-260823	-	ribokinase	rbsK	0.79780
314327-316399	+	lipase	geh	0.77582
369730-371253	-	alkyl hydroperoxide reductase subunit F	ahpF	0.93407
547752-550739	+	fibrinogen-binding protein SdrC	sdrC	0.77582
684182-685555	+	deoxyribodipyrimidine photolyase	phrB	0.80659
758680-759654	+	excinuclease ABC subunit B	-	0.85055
819904-821154	+	aminotransferase	nifS	0.78462
1022387-1023214	+	inositol monophosphatase family protein	suhB2	0.89451
1062563-1064344	+	excinuclease ABC subunit C	uvrC	0.80220
1071233-1071634	-	formyl peptide receptor-like 1 inhibitory protein	flr	0.80659
1076411-1077370	-	alpha-hemolysin	hla	0.84176
1275056-1275679	-	LexA repressor	lexA	0.82418
1287158-1289863	+	aconitate hydratase	citB	0.78022
1340719-1341438	+	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-acetyltransferase	dapD	0.77582
1355915-1357183	-	dihydrolipoamide succinyltransferase	odhB	0.78022
1357197-1359995	-	2-oxoglutarate dehydrogenase E1 component	sucA	0.85934
1553447-1554793	-	glycine dehydrogenase subunit 1	gcvPA	0.84176
1554813-1555904	-	glycine cleavage system aminomethyltransferase T	gcvT	0.85934
1556063-1556587	-	shikimate kinase	aroK	0.85055
1615283-1616035	-	LamB/YcsF family protein	lamB	0.88132
1617410-1617859	-	acetyl-CoA carboxylase biotin carboxyl carrier protein subunit	accB	0.82418
1691103-1692128	-	glyceraldehyde 3-phosphate dehydrogenase 2	gapB	0.80659
1819130-1820722	+	phosphoenolpyruvate carboxykinase	pckA	0.82418
1857741-1857884	-	gallidermin superfamily epiA protein	epiA	0.78462
1953330-1953689	-	phi PVL orf 50-like protein	-	0.77143

Co-expressed genes of RNAIII (from 2,093,091 to 2,093,248 at the reverse strand)				
Location	Strand[1]	Product[2]	Gene name[3]	C.C.[4]
2078217-2079440	+	succinyl-diaminopimelate desuccinylase	dapE	0.79341
2093500-2093634	-	delta-hemolysin	hld	0.98242
2094649-2095893	+	accessory gene regulator protein C	argC2	0.94725
2371786-2374740	-	formate dehydrogenase subunit alpha	fdhA	0.79341
2395160-2396398	-	imidazolonepropionase	hutI	0.92967
2396398-2398059	-	urocanate hydratase	hutU	0.85495
2400200-2401135	-	formimidoylglutamase	hutG	0.77143
2484745-2485431	-	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	gpmA	0.79341
2523000-2524226	-	amino acid ABC transporter ATP-binding protein	opuCA	0.82418
2584900-2586258	-	gluconate permease	gntP	0.79341
2779030-2781072	-	lipase	lip	0.82857
2796251-2797222	-	lactonase Drp35	drp35	0.81099
15795-16101	+	Teg1	Teg1	0.79780
384151-384256	-	Sau-63	Sau-63	0.78462
466471-466566	+	sRNA 00042	sRNA 00042	0.82857
623331-623668	-	SbrC	SbrC	0.89890
639706-639869	-	RsaD	RsaD	0.86374
774252-774423	-	RsaH	RsaH	0.78462
803886-803995	-	sRNA 00103	sRNA 00103	0.82418
812909-813098	+	RsaOM	RsaOM	0.79780
817532-817631	+	sRNA 00107	sRNA 00107	0.78901
1077604-1077702	-	sRNA 00132	sRNA 00132	0.82418
1194046-1194121	+	sRNA 00142	sRNA 00142	0.81978
1248023-1248136	-	sRNA 00151	sRNA 00151	0.91648
1355805-1355897	-	sRNA 00165	sRNA 00165	0.81538
1463875-1464375	-	RsaOR/SprX	RsaOR/SprX	0.84176
1731924-1732006	-	sRNA 00195	sRNA 00195	0.78022
1848996-1849113	-	SprB	SprB	0.89451
1863777-1863901	-	sRNA 00216	sRNA 00216	0.77582
1922182-1922252	-	sRNA 00228	sRNA 00228	0.85934
2111497-2111738	+	sRNA 00256	sRNA 00256	0.86813
2211957-2212213	+	SprF3/SprG3	SprF3/SprG3	0.83297
2377278-2377456	-	RsaOG	RsaOG	0.78022
2447792-2448151	-	sRNA 00302	sRNA 00302	0.93846
2505368-2505471	+	sRNA 00309	sRNA 00309	0.87253

Co-expressed genes of RNAIII (from 2,093,091 to 2,093,248 at the reverse strand)				
Location	Strand[1]	Product[2]	Gene name[3]	C.C.[4]
2555734-2555921	-	RsaOT	RsaOT	0.82418
2556328-2556416	+	Sau-19	Sau-19	0.84176
2622999-2623103	-	RsaOU	RsaOU	0.77582
2778594-2778873	-	SprA2	SprA2	0.81538
2795435-2795528	+	sRNA 00336	sRNA 00336	0.81978
Inversely expressed genes of RNAIII (from 2,093,091 to 2,093,248 at the reverse strand)				
Location	Strand[1]	Product[2]	Gene name[3]	C.C.[4]
211733-213643	+	staphylocoagulase	coa	-0.79341
243806-244975	+	teichoic acid biosynthesis protein F	tagF	-0.90769
465209-466102	+	dimethyladenosine transferase	ksgA	-0.78901
473154-473726	+	peptidyl-tRNA hydrolase	pth	-0.78901
482597-483136	+	hypoxanthine phosphoribosyltransferase	hpt	-0.81978
570077-570907	-	phosphomethylpyrimidine kinase	thiD1	-0.79780
634530-635825	-	penicillin-binding protein 4	pbp4	-0.79341
766954-767889	+	thioredoxin reductase	trxB	-0.94286
892098-893342	+	3-oxoacyl- synthase	fab	-0.80220
926875-928050	-	diacylglycerol glucosyltransferase	ypfP	-0.77582
1006693-1007244	-	peptide deformylase	def	-0.77582
1014431-1015525	+	ABC transporter	potA	-0.81538
1055512-1056450	-	ribonuclease HIII	rnhC	-0.81978
1068963-1069550	+	nucleoside-triphosphatase	rdgB	-0.80220
1107128-1109881	+	isoleucyl-tRNA synthetase	ileS	-0.77582
1126849-1127472	+	guanylate kinase	gmk	-0.78462
1175047-1176354	+	tRNA (uracil-5-)-methyltransferase Gid	gid	-0.81538
1187555-1189258	+	prolyl-tRNA synthetase	proS	-0.78022
1430639-1431610	-	bifunctional biotin operon repressor/biotin–[acetyl-CoA-carboxylase] synthetase BirA	birA	-0.86374
1431597-1432799	-	tRNA CCA-pyrophosphorylase	papS	-0.77582
1468673-1470583	-	SLT orf 636-like protein	-	-0.86374
1534633-1535907	-	2-oxoisovalerate dehydrogenase, E2 component, dihydrolipoamide acetyltransferase	bmfBB	-0.82418
1589638-1590390	-	16S rRNA (uracil(1498)-N(3))-methyltransferase	rsmE	-0.82857
1596246-1597370	-	coproporphyrinogen III oxidase	hemN	-0.81538
1607427-1608233	-	shikimate 5-dehydrogenase	aroE	-0.84615

Inversely expressed genes of RNAIII (from 2,093,091 to 2,093,248 at the reverse strand)				
Location	Strand[1]	Product[2]	Gene name[3]	C.C.[4]
1620425-1621048	-	uridine kinase	udk	-0.78022
1652611-1653750	-	queuine tRNA-ribosyltransferase	tgt	-0.85934
1677870-1679132	-	ATP-dependent protease ATP-binding subunit ClpX	clpX	-0.77582
1778125-1779786	-	polysaccharide biosynthesis protein	-	-0.79780
1866399-1867799	-	protoporphyrinogen oxidase	hemY	-0.80659
1886616-1887437	+	ribosomal large subunit pseudouridine synthase D	yhcT	-0.89011
2007644-2008720	+	nitric oxide synthase oxygenase subunit	nos	-0.89890
2029594-2031348	-	MHC class II analog protein	truncated mapW	-0.77582
2034686-2034982	-	peptidoglycan hydrolase	lytA	-0.84615
2336051-2337073	-	molybdenum cofactor biosynthesis protein A	moaA	-0.82418
2682290-2682661	-	aspartate alpha-decarboxylase	panD	-0.87692
2684410-2685270	+	2-dehydropantoate 2-reductase	panE	-0.80659
2820529-2820882	-	ribonuclease P	rnpA	-0.87692
298695-298804	-	sRNA 00018	sRNA 00018	-0.82857
357753-357861	+	sRNA 00023	sRNA 00023	-0.77143
483201-483370	+	sRNA 00043	sRNA 00043	-0.84176
686630-686743	-	sRNA 00077	sRNA 00077	-0.85495
993614-993798	-	sRNA 00125	sRNA 00125	-0.81099
1123759-1123970	-	sRNA 00138	sRNA 00138	-0.79341
1325685-1325800	-	sRNA 00160	sRNA 00160	-0.82418
1545833-1545942	+	sRNA 00179	sRNA 00179	-0.82857
1619706-1619848	+	Sau-5949	Sau-5949	-0.80220
1745968-1746232	-	sRNA 00196	sRNA 00196	-0.77143
1771226-1771318	+	sRNA 00198	sRNA 00198	-0.77582
1865076-1865185	-	sRNA 00217	sRNA 00217	-0.80220
2244481-2244535	-	sRNA 00273	sRNA 00273	-0.83736
2254308-2254399	-	sRNA 00276	sRNA 00276	-0.87692
2595206-2595316	+	sRNA 00320	sRNA 00320	-0.87253

[1] "+" in this column means the gene is at the forward, "-" meanns the gene is at the reverse strand.

[2] the sRNAs which were newly detected in this study are presented by "novel sRNA".

[3] If no gene name can be found, "-" would be shown in this column.

[4] C.C. means Spearman correlation coefficient

Table A.3: The co-expressed and inversely expressed genes of SprG4

Co-expressed genes of SprG4 (from 942,430 to 942,474 at the forward strand)				
Location	Strand[1]	Product[2]	Gene name [3]	C.C.[4]
10893-12407	+	histidine ammonia-lyase	hutH	0.79341
119492-120160	+	capsular polysaccharide biosynthesis protein	capA	0.84450
314327-316399	+	lipase	geh	0.77143
388295-388975	+	superantigen-like protein	ssl1	0.78901
758680-759654	+	excinuclease ABC subunit B	-	0.90330
759662-762508	+	excinuclease ABC subunit A	uvrA	0.77143
1064668-1065282	+	succinate dehydrogenase cytochrome b-558 subunit	sdhC	0.83736
1076411-1077370	-	alpha-hemolysin	hla	0.77582
1275056-1275679	-	LexA repressor	lexA	0.77582
1355915-1357183	-	dihydrolipoamide succinyltransferase	odhB	0.92967
1357197-1359995	-	2-oxoglutarate dehydrogenase E1 component	sucA	0.90330
1473305-1479505	-	phage tail tape meausure protein	-	0.87253
1857741-1857884	-	gallidermin superfamily epiA protein	epiA	0.77143
1931076-1932899	-	phiETA ORF57-like protein	-	0.78022
1932899-1934809	-	phi ETA orf 56-like protein		0.77582
1934824-1936725	-	phi ETA orf 55-like protein	-	0.79780
1941979-1942560	-	phage structural protein	-	0.78901
1942974-1943321	-	HK97 family phage protein	-	0.86813
1944270-1945244	-	phage head protein	-	0.79341
1947232-1948767	-	SPP1 family phage portal protein	-	0.77143
1950669-1951091	-	int gene activator RinA	-	0.78022
1951262-1951450	-	transcriptional activator rinb-like protein	-	0.80220
1952821-1953066	-	phi PVL orf 52-like protein	-	0.80659
1953081-1953329	-	phi PVL orf 51-like protein	-	0.86813
1953330-1953689	-	phi PVL orf 50-like protein	-	0.81978
1953960-1954145	-	PV83 orf 23-like protein	-	0.87692
1956596-1957402	-	phi PV83 orf 20-like protein	-	0.83736
1957374-1958066	-	phi PV83 orf 19-like protein	-	0.85495
1959992-1960312	-	phi PVL orf 39-like protein	-	0.84176
2078217-2079440	+	succinyl-diaminopimelate desuccinylase	dapE	0.77582
2266102-2266611	-	alkaline shock protein 23	asp23	0.80220
2400200-2401135	-	formimidoylglutamase	hutG	0.79780
2431146-2432624	-	malate:quinone oxidoreductase	mqo1	0.79780
2489222-2490151	+	gamma-hemolysin h-gamma-II subunit	hlgA	0.79341
2491668-2492645	+	leukocidin f subunit	hlgB	0.78022
2586375-2587928	-	gluconate kinase	gntK	0.86813

Co-expressed genes of SprG4 (from 942,430 to 942,474 at the forward strand)				
Location	Strand[1]	Product[2]	Gene name[3]	C.C.[4]
2650759-2652267	-	squalene synthase	crtN	0.81978
2652279-2653142	-	squalene desaturase	crtM	0.80659
2779030-2781072	-	lipase	lip	0.81099
201742-201995	+	RsaG	RsaG	0.79341
384153-384256	-	Sau-63	Sau-63	0.89890
386295-386383	+	sRNA 00030	sRNA 00030	0.81099
454093-454131	+	sRNA 00040	sRNA 00040	0.81978
454348-454384	+	sRNA 00041	sRNA 00041	0.81538
466471-466566	+	sRNA 00044	sRNA 00044	0.86374
624446-624539	+	SbrC/RsaC	SbrC/RsaC	0.78901
639706-639869	-	RsaD	RsaD	0.81538
721370-721484	-	sRNA 00086	sRNA 00086	0.79341
788253-788698	+	SsrA	SsrA	0.80659
801482-801578	-	RsaOL	RsaOL	0.84176
803886-803995	-	sRNA 00108	sRNA 00108	0.81978
833764-833849	+	sRNA 00113	sRNA 00113	0.83736
1194046-1194121	+	sRNA 00148	sRNA 00148	0.89011
1248030-1248136	-	sRNA 00158	sRNA 00158	0.79780
1349498-1349815	+	RsaOW2	RsaOW2	0.83297
1418742-1419051	-	RNaseP bact a	RNaseP bact a	0.92088
1462718-1462934	-	sRNA 00184	sRNA 00184	0.82418
1463876-1464375	-	RsaOR/SprX	RsaOR/SprX	0.81099
1638992-1639233	-	sRNA 00195	sRNA 00195	0.91648
1771659-1771725	+	Sau-5949	Sau-5949	0.89451
1818838-1818983	-	sRNA 00217	sRNA 00217	0.81099
1832869-1832985	-	SprA/SprA1	SprA/SprA1	0.93407
1848999-1849113	-	SprB	SprB	0.84615
1897234-1897324	-	SbrC/RsaC/RsaOW2	SbrC/RsaC/RsaOW2	0.88132
1922186-1922252	-	sRNA 00238	sRNA 00238	0.82418
1923575-1923871	-	sRNA 00239	sRNA 00239	0.89011
1962568-1962663	-	sRNA 00245	sRNA 00245	0.79780
2027313-2027386	+	sRNA 00254	sRNA 00254	0.87692
2237200-2237290	-	SbrC/RsaC/RsaOW2	SbrC/RsaC/RsaOW2	0.88132
2252765-2252898	+	sRNA 00290	sRNA 00290	0.85055
2363055-2363204	-	sRNA 00306	sRNA 00306	0.84615
2377298-2377456	-	RsaOG	RsaOG	0.82418
2498328-2498395	+	SprA2/RsaJ	SprA2/RsaJ	0.82418
2502526-2502622	+	SprA2	SprA2	0.96044
2505368-2505458	+	sRNA 00328	sRNA 00328	0.82857
2530723-2530817	+	sRNA 00331	sRNA 00331	0.78462

Co-expressed genes of SprG4 (from 942,430 to 942,474 at the forward strand)				
Location	Strand[1]	Product[2]	Gene name[3]	C.C.[4]
2551638-2551722	-	sRNA 00333	sRNA 00333	0.88132
2552154-2552329	-	sRNA 00334	sRNA 00334	0.77143
2778594-2778873	-	SprA2	SprA2	0.84615
Inversely expressed genes of SprG4 (from 942,430 to 942,474 at the forward strand)				
Location	Strand[1]	Product[2]	Gene name[3]	C.C.[4]
211733-213643	+	staphylocoagulase	coa	-0.80659
253270-254010	+	two-component response regulator	lytR	-0.87253
359013-360110	+	GTP-dependent nucleic acid-binding protein EngD	ychF	-0.88132
367374-367955	+	phosphoglycerate mutase family protein	-	-0.79780
381705-383246	+	GMP synthase	guaA	-0.78901
445285-446958	+	DNA polymerase III subunits gamma and tau	dnaX	-0.78901
465209-466102	+	dimethyladenosine transferase	ksgA	-0.80659
482597-483136	+	hypoxanthine phosphoribosyltransferase	hpt	-0.82418
517929-518477	+	transcription antitermination protein	nusG	-0.78901
519288-519980	+	50S ribosomal protein L1	rplA	-0.82857
634530-635825	-	penicillin-binding protein 4	pbp4	-0.78901
674809-675684	-	undecaprenyl pyrophosphate phosphatase	uppP	-0.87692
753021-753896	+	peptide chain release factor 2	prfB	-0.80220
785349-787721	+	ribonuclease R	rnr	-0.78022
809038-809358	-	thioredoxin	-	-0.79341
834275-835732	+	D-alanine–poly(phosphoribitol) ligase subunit 1	dltA	-0.84176
1023661-1025508	+	GTP-binding protein TypA	typA	-0.85934
1051881-1052939	+	phenylalanyl-tRNA synthetase subunit alpha	pheS	-0.86374
1107128-1109881	+	isoleucyl-tRNA synthetase	ileS	-0.83736
1124877-1126574	-	fibrinogen-binding protein A-like protein	fbpA	-0.96484
1151543-1151776	+	acyl carrier protein	acpP	-0.88132
1157931-1159298	+	signal recognition particle protein	ffh	-0.86813
1181767-1182648	+	elongation factor Ts	tsf	-0.87253
1182785-1183507	+	uridylate kinase	pyrH	-0.77582
1187555-1189258	+	prolyl-tRNA synthetase	proS	-0.78901
1198912-1199262	+	ribosome-binding factor A	rbfA	-0.89011

Inversely expressed genes of SprG4 (from 942,430 to 942,474 at the forward strand)				
Location	Strand[1]	Product[2]	Gene name [3]	C.C.[4]
1259354-1259887	+	thermonuclease	nucI	-0.89890
1318354-1319613	+	methicillin resistance factor	femB	-0.85934
1345915-1346115	-	cold shock protein	cspA	-0.85934
1431597-1432799	-	tRNA CCA-pyrophosphorylase	papS	-0.79780
1447060-1448370	-	GTP-binding protein EngA	engA	-0.78462
1450479-1451138	-	cytidylate kinase	cmk	-0.79341
1514320-1515057	-	ribosomal large subunit pseudouridine synthase B	rluB	-0.80659
1546679-1547143	-	acetyl-CoA carboxylase biotin carboxyl carrier protein subunit	accB	-0.88571
1597933-1599756	-	GTP-binding protein LepA	lepA	-0.79341
1600398-1601372	-	DNA polymerase III subunit delta	holA	-0.78022
1608247-1609347	-	GTP-binding protein YqeH	yqeH	-0.80659
1620425-1621048	-	uridine kinase	udk	-0.79780
1639304-1641070	-	aspartyl-tRNA synthetase	aspS	-0.89890
1641086-1642348	-	histidyl-tRNA synthetase	hisS	-0.84615
1646762-1647280	-	adenine phosphoribosyltransferase	apt	-0.87253
1652611-1653750	-	queuine tRNA-ribosyltransferase	tgt	-0.91209
1653773-1654798	-	S-adenosylmethionine:tRNA ribosyltransferase-isomerase	queA	-0.79780
1679283-1680584	-	trigger factor	tig	-0.89451
1711922-1712779	-	acetyl-CoA carboxylase carboxyltransferase subunit beta	accD	-0.80659
1729300-1730523	-	thiamine biosynthesis protein ThiI	thiI	-0.89011
1744663-1745925	-	tyrosyl-tRNA synthetase	tyrS	-0.87253
1994708-1996711	-	NAD-dependent DNA ligase	lig	-0.79780
2148174-2149694	-	DEAD-box ATP dependent DNA helicase	cshA	-0.87253
2180586-2181662	-	peptide chain release factor 1	prfA	-0.85934
2330439-2331704	-	peptidoglycan pentaglycine interpeptide biosynthesis protein FmhB	fmhB	-0.89451
2336051-2337073	-	molybdenum cofactor biosynthesis protein A	moaA	-0.80220
705004-705162	-	RsaOC	RsaOC	-0.80659

Inversely expressed genes of SprG4 (from 942,430 to 942,474 at the forward strand)				
Location	Strand[1]	Product[2]	Gene name[3]	C.C.[4]
993614-993798	-	sRNA 00130	sRNA 00130	-0.81538
1180702-1180838	+	sRNA 00145	sRNA 00145	-0.85934
1551723-1551821	+	sRNA 00190	sRNA 00190	-0.78462
1623931-1624057	-	sRNA 00193	sRNA 00193	-0.89451
1884625-1884760	+	sRNA 00229	sRNA 00229	-0.83297
1904523-1904649	-	sRNA 00236	sRNA 00236	-0.78901

[1] "+" in this column means the gene is at the forward, "-" meanns the gene is at the reverse strand.

[2] the sRNAs which were newly detected in this study are presented by "novel sRNA".

[3] If no gene name can be found, "-" would be shown in this column.

[4] C.C. means Spearman correlation coefficient

Table A.4: The co-expressed and inversely expressed genes of a novel sRNA – sRNA 00008

co-expressed genes of sRNA 00008 (from 90,947 to 91,086 at the forward strand)				
Location	Strand[1]	Product[2]	Gene name [3]	C.C.[4]
447372-447968	+	recombination protein RecR	recR	0.78462
729365-730336	+	ribonucleotide-diphosphate reductase subunit beta	nrdF	0.80220
2131925-2132695	-	RNA polymerase sigma factor SigB	sigB	0.77143
2493175-2493867	-	6-carboxyhexanoate–CoA ligase	bioW	0.91648
2495980-2497338	-	adenosylmethionine–8-amino-7-oxononanoate aminotransferase BioA	bioA	0.80220
2497316-2498002	-	dethiobiotin synthase	bioD	0.81978
Inversely expressed genes of sRNA 00008 (from 90,947 to 91,086 at the forward strand)				
Location	Strand[1]	Product[2]	Gene name [3]	C.C.[4]
205909-208158	+	formate acetyltransferase	pflB	-0.83297
228483-229292	+	L-lactate dehydrogenase	lctE	-0.80220
500139-501026	+	pyridoxal biosynthesis lyase PdxS	pdxS	-0.82418
691003-691923	+	1-phosphofructokinase	fruB	-0.89011
1234933-1236606	+	aerobic glycerol-3-phosphate dehydrogenase	glpD	-0.85934
1594510-1595136	-	heat shock protein GrpE	grpE	-0.79780
2276133-2276303	-	PTS system lactose-specific transporter subunit IIBC	lacE	-0.89011
2433005-2434603	-	L-lactate permease	lctP	-0.81099

[1] "+" in this column means the gene is at the forward, "-" meanns the gene is at the reverse strand.

[2] the sRNAs which were newly detected in this study are presented by "novel sRNA".

[3] If no gene name can be found, "-" would be shown in this column.

[4] C.C. means Spearman correlation coefficient

Table A.5: The co-expressed and inversely expressed genes of a novel sRNA – sRNA 00076

co-expressed genes of sRNA 00076 (from 641,099 to 641,200 at the forward strand)				
Location	Strand[1]	Product[2]	Gene name [3]	C.C.[4]
78527-79519	-	periplasmic binding protein	sirA	0.77143
80727-81737	+	2,3-diaminopropionate biosynthesis protein SbnB	sbnB	0.89890
264826-266202	-	drug transporter	-	0.83297
642034-643038	+	ferrichrome transport permease FhuB	fhuB	0.89011
643035-644051	+	ferrichrome ABC transporter permease	fhuG	0.84615
727082-729247	+	ribonucleotide-diphosphate reductase subunit alpha	nrdE	0.78901
729365-730336	+	ribonucleotide-diphosphate reductase subunit beta	nrdF	0.77143
1048274-1049050	+	iron compound ABC transporter permease	isdF	0.82857
1050053-1050376	+	heme-degrading monooxygenase IsdG	isdG	0.84176
2139879-2140439	-	potassium-transporting ATPase subunit C	kdpC	0.78901
2340202-2340696	+	molybdenum cofactor biosynthesis protein MoaC	moaC	0.79341
2541036-2541905	-	nickel ABC transporter permease	opp-1C	0.80659
Inversely expressed genes of sRNA 00076 (from 641,099 to 641,200 at the forward strand)				
Location	Strand[1]	Product[2]	Gene name [3]	C.C.[4]
530440-530910	+	30S ribosomal protein S7	rpsG	-0.78022
1233279-1234775	+	glycerol kinase	glpK	-0.91648
1981615-1982115	+	ferritin	ftn	-0.83297
2172989-2173201	-	F0F1 ATP synthase subunit C	atpE	-0.79341
2217068-2218873	-	glucosamine–fructose-6-phosphate aminotransferase	glmS	-0.79341

[1] "+" in this column means the gene is at the forward, "-" meanns the gene is at the reverse strand.

[2] the sRNAs which were newly detected in this study are presented by "novel sRNA".

[3] If no gene name can be found, "-" would be shown in this column.

[4] C.C. means Spearman correlation coefficient

Table A.6: The co-expressed and inversely expressed genes of a novel sRNA – sRNA 00324

co-expressed genes of sRNA 00324 (from 2,485,411 to 2,485,628 at the reverse strand)				
Location	Strand[1]	Product[2]	Gene name [3]	C.C.[4]
979384-979803	+	5-(carboxyamino)imidazole ribonucleotide mutase	purE	0.81538
982551-984740	+	phosphoribosylformylglycinamidine synthase II	purL	0.77582
984719-986203	+	amidophosphoribosyltransferase	purF	0.81978
986196-987224	+	phosphoribosylaminoimidazole synthetase	purM	0.83297
987808-989286	+	bifunctional phosphoribosylamino-imidazolecarboxamide formyltransferase/IMP cyclohydrolase	purH	0.78901
989308-990555	+	phosphoribosylamine–glycine ligase	purD	0.88571
Inversely expressed genes of sRNA 00324 (from 2,485,411 to 2,485,628 at the reverse strand)				
Location	Strand[1]	Product[2]	Gene name [3]	C.C.[4]
1965879-1966925	+	phage family integrase	int	-0.81978
2071481-2072251	+	repressor	-	-0.78901
2274590-2276128	-	PTS system lactose-specific transporter subunit IIBC	lacE	-0.81099
2714829-2716829	-	permease domain-containing protein	-	-0.85934

[1] "+" in this column means the gene is at the forward, "-" meanns the gene is at the reverse strand.

[2] the sRNAs which were newly detected in this study are presented by "novel sRNA".

[3] If no gene name can be found, "-" would be shown in this column.

[4] C.C. means Spearman correlation coefficient

Curriculum Vitae

Personal information

Birth: November 7th, 1983 in Chung-Hua, Taiwan
Nationality: Taiwan

Education

Sept. 2013 – Jan. 2018
PhD student in the group of Prof. Dr. Jörg Vogel and Dr. Konrad Förstner
Institute of Molecular Infection Biology (IMIB)
University of Würzburg, Würzburg, Germany

Sept. 2006 – June 2008
Master of Science – Bioinformatics
Institute of Bioinformatics and Systems Biology
National Chiao Tung University, Hsin-Chu, Taiwan
Thesis title: Prediction of functional sites of proteins from protein structures. Supervised by Prof. Dr. Jenn-Kang Hwang

Sept. 2002 – June 2006
Bachelor of Science – Life science
Department of Life Science
National Tsing Hua University, Hsin-Chu, Taiwan
Thesis title: The relationship between cell-cell adhesion and cell repair. Supervised by Prof. Dr. Jui-Chou Hsu

Experience

July 2011 – June 2013
Research Assistant – Bioinformatics
Institute of Bioinformatics and Systems Biology
National Chiao Tung University, Hsin-Chu, Taiwan
Project title: The relationship between sequence conservation and structure similarity. Supervised by Prof. Dr. Jenn-Kang Hwang

Sept. 2009 – June 2011
Trainee of Juridical Party of the Church in Taipei Assembly Hall
Bible Truth and Church Service Training Center
the Church in Taipei Assembly Hall, Taipei, Taiwan
Project title: Discuss the revelation and progress of grace from the view of GodâĂŹs eternal economy.

July 2008 – June 2009
Corporal of Government Soldier
Government affairs
Command center of Hengshan, Ministry of National Defense R.O.C, Taipei, Taiwan

_____ _____

Place and Date Sung-Huan Yu

Publication list

Publications during the PhD

S.H. Yu, J. Vogel and K.U. Förstner (2018) "ANNOgesic: A Tool to translate bacterial/archaeal RNA-Seq data into high-resolution genome annotations", *GigaScience*, **7**, giy096

S.H. Yu, P. Tanwer, M. Sharan, M. Sauer, C. Schuster, A. Smirnov, A, Herbig, R. Bertram, K. Nieselt, K.U. Förstner and J. Vogel (2017) "A high resolution genome annotation of Staphylococcus aureus HG003 and functional analysis of its sRNAs", manuscript preparation

L. Langhanki, P. Berger, J. Treffon, B. Bunk, **S.H. Yu**, K.U. Foerstner, J. Vogel, B. Kahl, A. Mellmann (2017) "Transcriptomic and epigenomic mechanisms play a role in the adaptation of of Staphylococcus aureus during long-term persistence in humans.", manuscript preparation

B. Remes, T. Rische-Grahl, T. Müller, K.U. Förstner, **S.H. Yu**, W. Lennart, A. Jäger, V. Peuser, and G. Klug (2017) "An RpoHI-dependent response promotes outgrowth after extended stationary phase in the alphaproteobacterium *Rhodobacter sphaeroides*", *J. Bacteriol.*, JB.00249-17

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

S.W. Yeh, J.W. Liu, **S.H. Yu**, C.H. Shih, J.K. Hwang, J. Echave (2014) "Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure", *Mol. Biol. Evol.*, **31**, p. 135-139

S.W. Yeh, T.T. Huang, J.W. Liu, **S.H. Yu**, C.H. Shih, J.K. Hwang, J. Echave (2014)

"Local packing density is the main structural determinant of the rate of protein sequence evolution at site level", *Biomed Res. Int.*

Previous publications

S.W. Huang, **S.H. Yu**, C.H. Shih, H.W. Guan, T.T. Huang, J.K. Hwang (2011) "On the relationship between catalytic residues and their protein contact number", *Curr. Protein Pept. Sci.*, **12**, pp. 574-579

Y.L. Lai, S.C. Yen, **S.H. Yu**, J.K. Hwang (2007) "pKNOT: the protein KNOT web server", *Nucleic Acids Res.*, **35**, p. W420-W424

C.H. Shih, S.W. Huang, S.C. Yen, Y.L. Lai, **S.H. Yu**, J.K. Hwang (2007) "A simple way to compute protein dynamics without a mechanical model", *Proteins: Struct., Funct., Bioinf.*, **68**, p. 34-38

Acknowledgements

I would like to thank the people who support and guide me during the years of my PhD. I am so grateful that I can meet and work with them. Without their help, the thesis will be impossible.

I express my gratitude the my supervisor Prof. Dr. Jörg Vogel for giving me a wonderful opportunity to do such exciting and charming project. His fully support and professional suggestions always brought numerous inspirations to me. Working with him is a tremendous help for my scientific development. Thanks my PhD committee – Prof. Dr. Thomas Dandekar and Prof. Dr. Cynthia Sharma for the fruitful discussions and the valuable feedbacks on my work.

I am thankful to Dr. Konrad Förstner for his for bringing me to this interesting field. He not only guided me patiently in my PhD research, but also share his rich research experience to me for enlarging my vision of science.

I would like to acknowledge my colleagues of IMIB and Core Unit Systemmedizin especially Elena Katzowitsch, Margarete Göbel, Dr. Malvika Sharan, Dr. Panagiota Arampatzi, Dr. Richa Bharti, Silvia Di Giorgio, Dr. Kristina Döring, and Thorsten Bischler. Because of their fully support, my PhD thesis can be accomplished such successfully. Working with them is my great honor.

I am extremely grateful to my family. Because of their kindly understanding, I can study abroad and fully concentrate to my work without any worries. They are my comfort and joy. Thanks to all of my friends for helping me to take care of my parents and my dear daughter.

Last but not least, I would like to give the glory to my Lord – Jesus Christ. He is the source of wisdom and grace for leading me to overcome all of the difficulties. His accompany make me never feel lonely in my research career.

Affidavit

I hereby confirm that my thesis entitled Development and application of computational tools for RNA-seq based transcriptome annotations is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Würzburg,
Place, Date Signature

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation Entwicklung und Anwendung bioinformatischer Werkzeuge für RNA-Seq-basierte Transkriptom-Annotationen eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Würzburg,
Ort, Datum Unterschrift