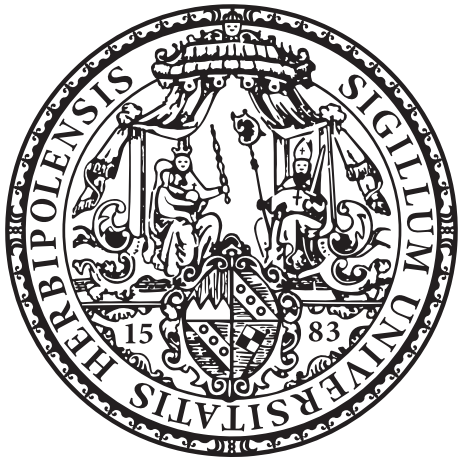


DATA MINING AND SOFTWARE DEVELOPMENT FOR RNA-SEQ-BASED APPROACHES IN BACTERIA

Data-Mining und Softwareentwicklung für RNA-seq-basierte Methoden bei Bakterien

Dissertation zur Erlangung des naturwissenschaftlichen Doktorgrades
der Graduate School of Life Sciences,
Julius-Maximilians-Universität Würzburg,
Klasse Infektion und Immunität

Vorgelegt von

THORSTEN DAVID BISCHLER

aus

Oberndorf am Neckar

Würzburg 2018

Eingereicht am: .....
Bürostempel

Mitglieder des Promotionskomitees:

Vorsitzender: Prof. Dr. Jörg Schultz

1. Betreuer: Prof. Dr. Cynthia M. Sharma

2. Betreuer: Prof. Dr. Thomas Dandekar

3. Betreuer: Prof. Dr. Jörg Vogel

4. Betreuer: Jun.-Prof. Dr. Björn Voß

Tag des Promotionskolloquiums: .....

Doktorurkunden ausgehändigt am: .....

DECLARATION

AFFIDAVIT

I hereby confirm that my thesis entitled **"Data mining and software development for RNA-seq-based approaches in bacteria"** 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, January 8, 2018

Thorsten David Bischler

EIDESSTATTLICHE ERKLÄRUNG

Hiermit erkläre ich an Eides statt, die Dissertation **"Data-Mining und Softwareentwicklung für RNA-seq-basierte Methoden bei Bakterien"** 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, 8. Januar 2018

Thorsten David Bischler

SUMMARY

RNA sequencing (RNA-seq) has in recent years become the preferred method for gene expression analysis and whole transcriptome annotation. While initial RNA-seq experiments focused on eukaryotic messenger RNAs (mRNAs), which can be purified from the cellular ribonucleic acid (RNA) pool with relative ease, more advanced protocols had to be developed for sequencing of microbial transcriptomes. The resulting RNA-seq data revealed an unexpected complexity of bacterial transcriptomes and the requirement for specific analysis methods, which in many cases is not covered by tools developed for processing of eukaryotic data.

The aim of this thesis was the development and application of specific data analysis methods for different RNA-seq-based approaches used to gain insights into transcription and gene regulatory processes in prokaryotes.

The differential RNA sequencing (dRNA-seq) approach allows for transcriptional start site (TSS) annotation by differentiating between primary transcripts with a 5'-triphosphate (5'-PPP) and processed transcripts with a 5'-monophosphate (5'-P). This method was applied in combination with an automated TSS annotation tool to generate global trancriptome maps for *Escherichia coli* (*E. coli*) and *Helicobacter pylori* (*H. pylori*).

In the *E. coli* study we conducted different downstream analyses to gain a deeper understanding of the nature and properties of transcripts in our TSS map. Here, we focused especially on putative antisense RNAs (asRNAs), an RNA class transcribed from the opposite strand of known protein-coding genes with the potential to regulate corresponding sense transcripts. Besides providing a set of putative asRNAs and experimental validation of candidates via Northern analysis, we analyzed and discussed different sources of variation in RNA-seq data.

The aim of the *H. pylori* study was to provide a detailed description of the dRNA-seq approach and its application to a bacterial model organism. It includes information on experimental protocols and requirements for data analysis to generate a genome-wide TSS map. We show how the included TSS can be used to identify and analyze transcriptome and regulatory features and discuss challenges in terms of library preparation protocols, sequencing platforms, and data analysis including manual and automated TSS annotation.

The TSS maps and associated transcriptome data from both *H. pylori* and *E. coli* were made available for visualization in an easily accessible online browser.

Furthermore, a modified version of dRNA-seq was used to identify transcriptome targets of the RNA pyrophosphohydrolase (RppH) in *H. pylori*. RppH initiates 5′-end-dependent degradation of transcripts by converting the 5′-PPP of primary transcripts to a 5′-P. I developed an analysis method, which uses data from complementary DNA (cDNA) libraries specific for transcripts carrying a 5′-PPP, 5′-P or both, to specifically identify transcripts modified by RppH. For this, the method assessed the 5′-phosphorylation state and cellular concentration of transcripts in *rppH* deletion in comparison to strains with the intact gene. Several of the identified potential RppH targets were further validated via half-life measurements and quantification of their 5′-phosphorylation state in wild-type and mutant cells. Our findings suggest an important role for RppH in post-transcriptional gene regulation in *H. pylori* and related organisms.

In addition, we applied two RNA-seq-based approaches, RNA immunoprecipitation followed by sequencing (RIP-seq) and cross-linking immunoprecipitation followed by sequencing (CLIP-seq), to identify transcripts bound by Hfq and CsrA, two RNA-binding proteins (RBPs) with an important role in post-transcriptional regulation.

For RIP-seq-based identification of CsrA binding regions in *Campylobacter jejuni* (*C. jejuni*), we used annotation-based analysis and, in addition, a self-developed peak calling method based on a sliding window approach. Both methods revealed *flaA* mRNA, encoding the major flagellin, as the main target and functional analysis of identified targets showed a significant enrichment of genes involved in flagella biosynthesis. Further experimental analysis revealed the role of *flaA* mRNA in post-transcriptional regulation.

In comparison to RIP-seq, CLIP-seq allows mapping of RBP binding sites with a higher resolution. To identify these sites an approach called "block-based peak calling" was developed and resulting peaks were used to identify sequence and structural constraints required for interaction of Hfq and CsrA with *Salmonella* transcripts.

Overall, the different RNA-seq-based approaches described in this thesis together with their associated analyis pipelines extended our knowledge on the transcriptional repertoire and modes of post-transcriptional regulation in bacteria. The global TSS maps, including further characterized asRNA candidates, putative RppH targets, and identified RBP interactomes will likely trigger similar global studies in the same or different organisms or will be used as a resource for closer examination of these features.

ZUSAMMENFASSUNG

RNA-Sequenzierung (RNA-seq) entwickelte sich in den letzten Jahren zur bevorzugten Methode für Genexpressionsanalysen und die Annotation ganzer Transkriptome. Nachdem sich erste RNA-seq-Experimente hauptsächlich mit eukaryotischen Boten-RNAs (mRNAs) beschäftigt hatten, da diese sich relativ einfach aus dem zellulären RNA-Gemisch aufreinigen lassen, war die Entwicklung von fortschrittlicheren Methoden nötig, um mikrobielle Transkriptome zu sequenzieren. Die sich daraus ergebenden RNA-seq-Daten enthüllten eine unerwartete Komplexität bakterieller Transkriptome und die Notwendigkeit der Anwendung spezifischer Analyseverfahren, welche von Tools zur Prozessierung eukaryotischer Daten häufig nicht zur Verfügung gestellt werden.

Das Ziel dieser Doktorarbeit war die Entwicklung und Anwendung spezifischer Verfahren zur Datenanalyse für verschiedene RNA-seq-basierte Methoden, um Erkenntnisse bezüglich Transkription und genregulatorischer Vorgänge bei Prokaryoten zu erlangen.

Die Differentielle-RNA-Sequenzierungsmethode (dRNA-seq) ermöglicht die Annotation von Transkriptionsstartpunkten (TSS), indem sie Primärtranskripte mit einem 5'-Triphosphat (5'-PPP) von prozessierten Transkripten mit einem 5'-Monophosphat (5'-P) unterscheidet. Diese Methode wurde in Kombination mit einem automatisierten TSS-Annotationstool zur Erstellung globaler Transkriptomkarten für *Escherichia coli (E. coli)* and *Helicobacter pylori (H. pylori)* verwendet.

In der *E. coli*-Studie haben wir verschiedene Folgeanalysen durchgeführt, um ein tieferes Verständnis für die Natur und Eigenschaften der in unserer Transkriptomkarte enthaltenen Transkripte zu erlangen. Das Hauptaugenmerk lag dabei auf mutmaßlichen Antisense-RNAs (asRNAs). Diese stellen eine RNA-Klasse dar, welche vom entgegengesetzten Strang von bekannten proteinkodierenden Genen transkribiert wird, und die das Potenzial hat, entsprechende Sense-Transkripte zu regulieren. Wir stellen nicht nur eine Liste mutmaßlicher asRNAs zur Verfügung, von der einige Kandidaten durch Northern Blots validiert wurden, sondern diskutierten auch von uns untersuchte Gründe für auftretende Variation bei RNA-seq-Daten.

Das Ziel der *H. pylori*-Studie war es, eine detaillierte Beschreibung der dRNA-seq-Methode und deren Anwendung auf einen bakteriellen Modellorganismus zur Verfügung zu stellen. Sie enthält Informationen bezüglich experimenteller Proto-

kolle und für die Datenanalyse notwendige Schritte, zur Erstellung einer genomweiten TSS-Karte. Wir zeigen, wie die enthaltenen TSS verwendet werden können, um verschiedene Transkriptomelemente, einschließlich solcher mit regulatorischen Eigenschaften, zu identifizieren und zu analysieren. Zusätzlich diskutieren wir Probleme, welche bei der Erstellung von Sequenzierlibraries, der Verwendung von Sequenzierplattformen und bei der Datenanalyse, einschließlich manueller und automatisierter TSS-Annotation, auftreten können.

Die TSS-Karten für *H. pylori* und *E. coli*, einschließlich der damit verbundenen Transkriptomdaten, haben wir in Form eines leicht zugänglichen Online-Browsers verfügbar gemacht.

Desweiteren wurde eine modifizierte Version der dRNA-seq-Methode verwendet, um Transkripte zu identifizieren, welche von der RNA Pyrophosphohydrolase (RppH) in *H. pylori* gespalten werden. RppH initiiert den vom 5'-Ende abhängigen RNA-Abbau, indem sie das 5'-PPP von Primärtranskripten in ein 5'-P umwandelt. Ich habe eine Analysemethode entwickelt, welche Daten basierend auf unterschiedlichen Komplementär-DNA (cDNA)-Libraries verwendet, welche entweder spezifisch für Transkripte mit einem 5'-PPP oder einem 5'-P sind, oder beides enthalten, um spezifisch Transkripte zu indentifizieren, die durch RppH modifiziert werden. Um dies zu erreichen wurden der 5'-Phosphorylierungsstatus und die zelluläre Konzentration der Transkripte zwischen einer *rppH*-Deletionsmutante und Stämmen mit intaktem Gen verglichen. Weiterhin wurden mehrere der identifizierten, von RppH gespaltenen Transkripte durch Messung ihrer Halbwertszeit und Quantifizierung ihres 5'-Phosphorylierungsstatus bei Wildtyp- und mutierten Zellen validiert. Unsere Ergebnisse lassen auf eine wichtige Rolle von RppH bei der Genregulation in *H. pylori* und verwandten Organismen schließen.

Zusätzlich haben wir zwei weitere RNA-seq-basierte Methoden namens RNA-Immunpräzipitation gefolgt von RNA-Sequenzierung (RIP-seq) und Quervernetzung und Immunpräzipitation gefolgt von RNA-Sequenzierung (CLIP-seq) verwendet, um Transkripte zu identifizieren, welche von Hfq und CsrA gebunden werden, zwei RNA-Bindeproteinen (RBPs), die eine wichtige Rolle bei posttranskriptionaler Regulation spielen.

Zur RIP-seq-basierten Identifikation von CsrA-Binderegionen bei *Campylobacter jejuni (C. jejuni)* haben wir eine annotationsbasierte Analyse und zusätzlich eine eigens entwickelte Peak-Bestimmungsmethode verwendet. Beide Methoden haben die *flaA* mRNA, welche das Hauptflagellin kodiert, als stärksten Bindepartner identifiziert. Die Funktionale-Anreicherungsanalyse hat außerdem eine Anreicherung von Genen ergeben, welche für die Flagellenbiosynthese von Bedeutung sind.

Im Vergleich zu RIP-seq ermöglicht CLIP-seq eine höhere Auflösung bei der Kartografierung von Bindestellen. Um diese Stellen zu identifizieren wurde eine Methode mit der Bezeichnung “block-based peak calling” entwickelt, und die daraus resultierenden Peaks wurden verwendet, um sequenz- und strukturabhängige Bedingungen zu bestimmen, die bei *Salmonella* für die Interaktion von Transkripten mit Hfq und CsrA notwendig sind.

Insgesamt betrachtet haben die verschiedenen RNA-seq-basierten Methoden, welche in dieser Doktorarbeit beschrieben wurden, in Kombination mit den damit verbundenen Analysepipelines, unser Verständnis des transkriptionellen Repertoires und der Art und Weise, wie posttranskriptionelle Regulation bei Bakterien abläuft, erweitert. Die globalen TSS-Karten, einschließlich der charakterisierten asRNA-Kandidaten, die mutmaßlich von RppH gespaltenen Transkripte und die identifizierten RBP-Interaktome werden höchstwahrscheinlich zur Durchführung ähnlicher Studien bei den gleichen oder anderen Organismen führen, oder können als Grundlage für eine detailliertere Untersuchung dieser Elemente verwendet werden.

CONTENTS

LIST OF FIGURES

ACRONYMS

5′-OH	5′-hydroxyl
5′-P	5′-monophosphate
5′-PPP	5′-triphosphate
asRNA	antisense RNA
asTSS	antisense TSS
ATP	adenosine triphosphate
B. subtilis	*Bacillus subtilis*
C. jejuni	*Campylobacter jejuni*
cDNA	complementary DNA
CDS	coding DNA sequence
ChIP-seq	chromatin immunoprecipitation followed by sequencing
CLIP-seq	cross-linking immunoprecipitation followed by sequencing
coIP	co-immunoprecipitation
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide
dRNA-seq	differential RNA sequencing

E. coli	*Escherichia coli*
eCLIP	enhanced CLIP
FPKM	Fragments per kilobase of exon model per million mapped reads
H. pylori	*Helicobacter pylori*
HITS-CLIP	high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation
irCLIP	infrared-CLIP
iTSS	internal TSS
methyl-seq	DNA methylation sequencing
miRNA	microRNA
mRNA	messenger RNA
ncRNA	non-coding RNA
NGS	next generation sequencing
nt	nucleotide
oTSS	orphan TSS
PAP I	*E. coli* poly(A) polymerase
PAR-CLIP	photoactivatable-ribonucleoside-enhanced-CLIP
piRNA	Piwi-interacting RNA
PNK	polynucleotide kinase
PNPase	polynucleotide phosphorylase
PSS	processed start sites
pTSS	primary TSS
RACE	rapid amplification of cDNA ends
RBP	RNA-binding protein
RIP-seq	RNA immunoprecipitation followed by sequencing
RNA-seq	RNA sequencing

RNA	ribonucleic acid
RNase	ribonuclease
RNase E	ribonuclease E
RNase II	ribonuclease II
RNase J	ribonuclease J
RNase R	ribonuclease R
RNase Y	ribonuclease Y
RPKM	Reads per kilobase of exon model per million mapped reads
RppH	RNA pyrophosphohydrolase
rRNA	ribosomal RNA
SBS	sequencing-by-synthesis
siRNA	short interfering RNA
SNP	single-nucleotide polymorphism
sRNA	small regulatory RNA
sTSS	secondary TSS
TAP	tobacco acid pyrophosphatase
TEX	terminator 5′-phosphate-dependent exonuclease
TPM	transcripts per million
tRNA	transfer RNA
TSS	transcriptional start site
UTR	untranslated region
UV	ultraviolet

1

AIM AND ORGANIZATION OF THE THESIS

The aim of this thesis was the development of appropriate biocomputational methods for the analysis of data derived from different RNA sequencing (RNA-seq)-based approaches conducted in several bacterial species. For this purpose, I integrated self-developed software together with existing tools to generate specific analysis pipelines.

The thesis is organized as follows. Chapter 2 provides general background on high-throughput sequencing, with focus on the RNA-seq-based approaches applied in this thesis and the associated requirements for data analysis.

Chapter 3 starts with a short summary of the findings of each publication included in this thesis followed by the original publications including supplementary materials. The publications are arranged in the following order:

Section 3.2: Maureen K. Thomason et al. "Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in Escherichia coli." en. In: *Journal of Bacteriology* 197.1 (Jan. 2015), pp. 18–28. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/JB.02096-14. URL: http://jb.asm.org/content/197/1/18 (visited on 01/07/2015)

Section 3.3: Thorsten Bischler et al. "Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in Helicobacter pylori." In: *Methods*. Bacterial and Archaeal Transcription 86 (Sept. 2015), pp. 89–101. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2015.06.012. URL: http://www.sciencedirect.com/science/article/pii/S1046202315002546 (visited on 02/29/2016)

Section 3.4: Thorsten Bischler et al. "Identification of the RNA Pyrophosphohydrolase RppH of Helicobacter pylori and Global Analysis of Its RNA Targets." en. In: *Journal of Biological Chemistry* 292.5 (Feb. 2017), pp. 1934–1950. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.M116.761171. URL: http://www.jbc.org/content/292/5/1934 (visited on 02/20/2017)

Section 3.5: Gaurav Dugar et al. "The CsrA-FliW network controls polar localization of the dual-function flagellin mRNA in Campylobacter jejuni." en. In: *Nature Communications* 7 (May 2016), p. 11667. DOI: 10.1038/ncomms11667. URL: http:

//www.nature.com/ncomms/2016/160527/ncomms11667/abs/ncomms11667.html (visited on 07/13/2016)

Section 3.6: Erik Holmqvist et al. "Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo." en. In: *The EMBO Journal* (Apr. 2016), e201593360. ISSN: 0261-4189, 1460-2075. URL: http://emboj.embopress.org/content/early/2016/04/04/embj.201593360 (visited on 04/05/2016)

The publications in sections 3.2 and 3.3 apply the standard differential RNA sequencing (dRNA-seq) approach for transcriptome-wide transcriptional start site (TSS) annotation.

In the publication in section 3.2 we aimed to gain further insights into the transcriptional repertoire of the widely-used model organsim *E. coli* strain K-12 with focus on antisense transcription.

In the publication in section 3.3 our aim was to provide a detailed description of the dRNA-seq approach using *Helicobacter pylori* (*H. pylori*) 26695 as an example including all steps required for data analysis and the generation of a TSS map. Furthermore, we wanted to analyze the effect of library preparation and higher read coverage based on Illumina sequencing by comparing our results to the findings of the original study [156].

The publication in section 3.4 uses a modified version of dRNA-seq to exploit its capability for processing site detection in order to globally identify targets of the RNA pyrophosphohydrolase (RppH) in *H. pylori*.

In the publication in section 3.5, we applied RNA immunoprecipitation followed by sequencing (RIP-seq) to identify ribonucleic acid (RNA) binding partners of CsrA in *Campylobacter jejuni* (*C. jejuni*), while cross-linking immunoprecipitation followed by sequencing (CLIP-seq) was used in the publication in section 3.6 to identify targets and binding sites of the RNA-binding proteins (RBPs) Hfq and CsrA in *Salmonella*.

Chapter 5 contains a common discussion of the findings of the five publications followed by conclusions and perspectives in chapter 6.

Contributions by others are listed in section a.1 in the Appendix.

2

INTRODUCTION

2.1 RNA SEQUENCING

The advent of high-throughput sequencing technologies has revolutionized genomic research in recent years. We are now able to collect an unprecedented amount of information on nucleic acid sequences with single-base resolution and at much lower costs compared to traditional Sanger sequencing platforms [64].

RNA-seq is a high-throughput method that can be used to qualitatively and quantitatively analyze the entire transcriptome of a cell or a collection of cells [127, 186]. The transcriptome consists of all transcripts expressed at a given time point and under specific physiological conditions and a thorough understanding of it is essential to interpret the functional constituents of a genome. Unlike formerly used hybridization-based approaches as microarrays or tiling arrays, which apply previously-designed probes that cover specific parts of or even a whole genome, RNA-seq does not require prior knowledge of sequence or structure of expressed genomic elements but can also be used for *de novo* sequencing and assembly of transcripts. In addition, RNA-seq has a higher dynamic range and needs less input material than array-based approaches without suffering from cross-hybridization.

RNA-seq can be used for detection of transcripts as well as quantitative profiling of transcript expression under different biological conditions. It is possible to compare expression between different genetic backgrounds, growth conditions or different tissues and cell types, and to detect changes upon exposure to chemical signals or environmental stresses. Qualitatively, RNA-seq is used to annotate all kinds of transcripts in pro- and eukaryotes, as messenger RNAs (mRNAs) or non-coding RNAs (ncRNAs) including small regulatory RNAs (sRNAs), and to elucidate the transcriptional structure of genes and operons with their 5'- and 3'-ends, alternatively spliced isoforms and post-transcriptional modifications.

2.1.1 *High-throughput sequencing technologies*

The term high-throughput sequencing describes methods to determine the sequences of a large number of molecules of either of the two principal nucleic acids: deoxyribonucleic acid (DNA) or RNA. The sequence describes the exact or-

der of nucleotides carrying the four bases adenine (A), guanine (G), cytosine (C) and thymine (T) in a single strand of DNA with uridine (U) instead of thymine (T) for an RNA strand. A contiguous sequence of bases derived from a single template molecule via a certain sequencing method is called a "read". Importantly, different sequencing methods can vary tremendously in terms of quality and maximum length of these reads.

To determine the sequence of an RNA molecule most sequencing methods require reverse transcription of RNA into complementary DNA (cDNA), which means that the read-out is not the actual RNA sequence but the DNA sequence of the genomic region from which it is transcribed.

2.1.1.1 *Short-read technologies*

The completion of the Human Genome Project in 2003 [33] greatly stimulated development of novel high-throughput sequencing approaches in the following years [115]. These technologies, previously called next generation sequencing (NGS) and currently second-generation sequencing, provided much higher throughput at greatly reduced costs in comparison to the Sanger sequencing method [150]. In exchange, they require extensive template amplification, which might introduce bias, and suffer from shorter read lengths and higher error rates [64, 108]. Due to the shorter read lengths in comparison to Sanger sequencing, they are also referred to as short-read sequencing techniques.

The first second-generation sequencer launched in 2005 was the 454 pyrosequencing machine [116]. Recently, the production of sequencers using this technology was stopped, but other second generation sequencing platforms encompassing Illumina, SOLiD [180] and Ion Torrent [148], to name some important examples, are still available. Therefore, there is not only one sequencing technology that works best for all purposes, and different options depending on the application should be considered.

Illumina sequencing is currently the most widely used technology on the market and supports a variety of different applications, such as, for example whole-genome or exome sequencing, epigenomics applications such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) [132] or DNA methylation sequencing (methyl-seq) [21] and, most important for the work described in this thesis, RNA-seq. In the following I will describe the Illumina sequencing method, which has been used for all RNA-seq-based approaches presented in this thesis.

Illumina sequencing is based on a sequencing-by-synthesis (SBS) approach, which makes use of a polymerase to add new nucleotides to an elongating strand where incorporation of each new base is detected via a fluorophore [64].

As is the case for most second-generation sequencing approaches, Illumina requires clonal amplification of the template sequence prior to sequencing. This is achieved by solid-phase bridge amplification [52] where two kinds of oligos, each complementary to one of the terminal adapter sequences, are bound to a glass slide (the flow cell). First, template strands hybridize with oligos on the flow cell. Next, the DNA strands bend and unbound terminal sequences bind to near complementary oligos for priming followed by synthesis of the complementary strand. The formed double-strand is denatured, the template is washed away and the process is repeated until dense clusters of identical DNA sequences consisting of forward and reverse strands are formed. Afterwards, all reverse strands are removed so that only forward strands are left as template for sequencing ([64, 115]).

For sequencing, Illumina applies 3'-blocked nucleotides. These deoxynucleotides (dNTPs) prevent strand elongation so that only one nucleotide can be incorporated at each elongating complementary strand per cycle. After hybridization of a sequencing primer, a mixture of all four dNTPs is added. Each dNTP is individually labeled with a fluorophore using either two- or four-color chemistry. Following incorporation of the complementary nucleotide, remaining dNTPs are washed away and the flow cell is imaged to identify the base that was added to each cluster. Afterwards, the fluorophore and blocking group is removed and the next sequencing cycle can begin. Imaging is conducted via total internal reflection fluorescence (TIRF) microscopy. Except for the NextSeq and MiniSeq sequencers, which apply the two-color chemistry, all other platforms use a separate laser channel for each of the four nucleotides.

In most cases, multiple cDNA libraries are sequenced in a single sequencing run. In order to discriminate which of the resulting sequencing reads belong to each library, a library-specific index sequence is integrated as part of the adapters and sequenced in one or two (dual-indexing) separate index reads. These index reads are used in the so-called demultiplexing step to assign reads from each cluster to the respective library. In general, demultiplexing is conducted by the sequencing facility using software provided by the manufacturer.

Most second-generation sequencing platforms, including Illumina sequencers, are capable of conducting paired-end sequencing. Here, in contrast to single-end sequencing, the DNA template is not only sequenced in one direction, but from both ends. Depending on the size of the fragment, forward and reverse reads can overlap or not. Paired-end sequencing has the advantage that, by taking into account fragment size, reads can be placed more accurately during assembly or alignment to a reference sequence. A special application of paired-end sequencing is

the sequencing of mate pair libraries that make it possible to sequence the ends of much larger fragments using a specific library preparation protocol [115].

2.1.1.2 *Long read technologies*

Short read sequencing technologies have problems resolving more complex parts of a genome, such as long repetitive regions and copy number or structural variations. Long read technologies represent a new third generation of sequencing technologies, which are able to generate reads with a length of several kilobases. Such long reads can span these complex regions and facilitate unambiguous read placement and determination of the size of genomic elements. Furthermore, they can be used for RNA-seq to allow precise analysis of isoforms or operon structures. Current long read technologies consist of single-molecule long read approaches and synthetic long read approaches. Synthetic long read approaches apply existing short read technologies in order to assemble long reads *in silico*. Platforms that are able to generate natural long reads consist of Pacific Biosciences single-molecule real time sequencing (SMRT) platforms [49] and nanopore sequencers [31] from Oxford Nanopore Technologies. Both technologies conduct sequencing of single molecules without the requirement for clonal amplification of a sequencing library. In addition, they do not require chemical cycling to add single dNTPs [64].

2.1.2 *Experimental design*

A crucial step for every RNA-seq-based experiment is the selection of an appropriate experimental design. This involves choosing adequate protocols for RNA extraction and library preparation, as well as considering the required number of replicates, sequencing depth, read length, and if sequencing should be conducted in single- or paired-end mode. Besides data acquisition, the steps and tools used for the analysis of the data play an important role.

In a typical RNA-seq experiment, a population of RNA, either total RNA or a certain fraction thereof, is extracted and converted into a cDNA library with adapters specific for the applied sequencing platform (Figure 2.1). In most cases, this process involves a PCR amplification step to increase the amount of cDNA fragments. For sequencing of short RNA species such as eukaryotic microRNAs (miRNAs), short interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs) or bacterial sRNAs, library preparation can be conducted directly. For longer RNAs, due to read-length limitations of most sequencing platforms, additional fragmentation of RNA or re-

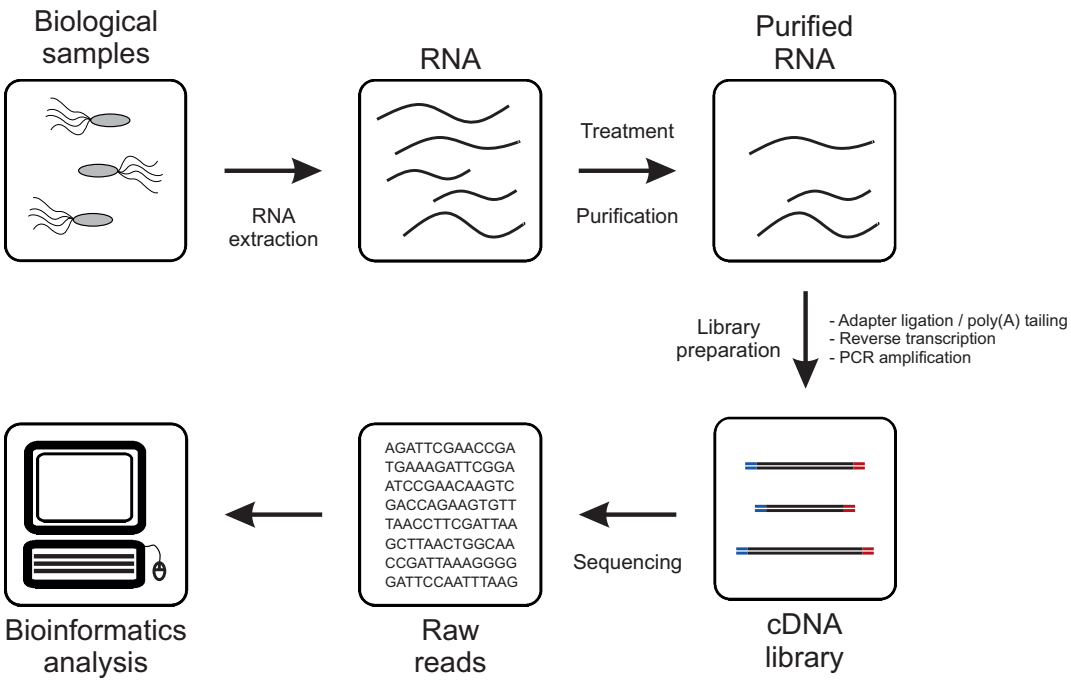


Figure 2.1: Steps of a typical RNA-seq experiment.

verse transcribed cDNA is required in advance to restrict the size of molecules to be sequenced.

A major challenge for RNA-seq experiments, especially with bacteria, has been the presence of abundant RNA species such as ribosomal RNAs (rRNAs) or transfer RNAs (tRNAs), which if sequenced would occupy almost the entire pool of sequencing reads. rRNA is especially problematic, since it typically constitutes more than 90% of the RNA in a cell. While this issue can be overcome for sequencing of mature eukaryotic mRNAs, which carry a poly(A) tail at their 3'-end, by conducting poly(A) selection via poly(T) oligomers or oligo(dT) amplification of cDNA, other methods were required for sequencing of ncRNA species and non-polyadenylated bacterial mRNAs. Early RNomics approaches for eukaryotic sRNA identification used Sanger sequencing of specialized cDNA libraries constructed by reverse transcription and vector cloning of size-selected RNA fractions [79]. While such approaches have also been used to identify prokaryotic sRNAs [184], they are not applicable to sequence whole bacterial transcriptomes. Methods for rRNA depletion in RNA-seq experiments utilize oligonucleotide-based removal of rRNAs via magnetic beads or size fractionation using gel electrophoresis [38, 162]. Importantly, even if such methods enrich the amount of non-rRNA reads, they can also cause problems. For example, rRNA depletion was recently found to introduce coverage bias [94].

In our publication "Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in *Helicobacter pylori*" [19], presented in section 3.3, we describe a library preparation protocol with inherent rRNA depletion via *E.*

coli poly(A) polymerase (PAP I), which has a preference for polyadenylating mRNAs over rRNAs [58]. This protocol was used in all studies presented in sections 3.2 to 3.5. In the CLIP-seq study presented in section 3.6, the experimental protocol includes co-immunoprecipitation (coIP) of a subset of cellular RNAs and ribonuclease (RNase) treatment, resulting in sufficient exclusion of rRNAs. Furthermore, steadily decreasing sequencing costs and the option for very deep sequencing counter the necessity for depletion of abundant RNA species.

Another important aspect of experimental design is the question of whether sequencing should be conducted in a strand-specific manner. Early RNA-seq studies used random hexamer priming to initiate reverse transcription [126]. The drawback of this method is that it does not retain strand information, which is important to identify and analyse overlapping transcripts or antisense RNAs (asRNAs), especially in complex bacterial transcriptomes. Different protocols have been developed to enable strand-specific sequencing of RNA pools [101]. We applied strand-specific library preparation protocols in all publications presented in this thesis, either via 5′ adapter ligation and poly(A)-tailing (sections 3.2 to 3.5) or by using a commercial strand-specific kit (section 3.6).

Illumina sequencers provide read lengths of up to 300 nucleotides (nts) in single-end mode (Illumina MiSeq v3, 2 x 300 nts paired-end) [64]. The single-end sequencing data in our studies consisted of read lengths ranging from ~100 nts based on HiSeq 2000/2500 machines (sections 3.2 to 3.5) to 120 nts using the older Genome Analyzer IIx (section 3.2). The paired-end sequencing data obtained for the publication in section 3.6 encompassed read pairs with a length of 2 x 75 nts. In addition, we used previously-generated 454 sequencing data (section 3.3) with a maximum mapped read length of ~350 nts [156]. In general, paired-end sequencing is used for applications such as *de novo* transcript assembly or analysis of isoform expression in eukaryotes where mapping both ends of long transcripts facilitates analysis [61, 86]. In addition, sequencing of longer reads results in improved mappability and transcript identification [61, 93].

Sequencing depth or library size describes the number of sequenced reads for a single cDNA library. Deeper sequencing results in increased transcript detection and improved quantification [126]. Nevertheless, optimal sequencing depth depends on the organism under study and the experimental approach used to answer a specific biological question. A target sequencing depth of 5 million reads has been established for primary transcriptome profiling in bacteria with a typical genome size of 5 megabases [155].

2.1.3 *Data analysis*

After sequencing on one of the available sequencing platforms (see section 2.1.1), the resulting output consists of the base sequence for each sequencing read and an associated quality score for each base call. Depending on the sequencing platform, this information is represented in a specific file format. For example, FASTQ is a common format that includes sequence and quality information [32]. Next, I will describe a general workflow for processing of Illumina sequencing data and give examples of common tools used for the different steps.

2.1.3.1 *Quality control and preprocessing*

First, quality of the reads assigned to each library is checked via a tool like FastQC [7], which among other metrics provides information on the quality distribution of base calls, sequence length distribution, GC content distribution, presence of duplicated or overrepresented sequences, per-base N content, per base sequence content, and k-mer content. The accuracy of base calling is measured by the Phred quality score (Q score), which represents the most common metric used to assess the accuracy of a sequencing platform. The score indicates the probability (P) that a base is called incorrectly by the sequencer and is calculated according to equation 2.1 [51].

$$Q = -10\log_{10}(P) \tag{2.1}$$

For example, a Phred score of 20 (Q20) represents a base call accuracy of 99%, meaning that an incorrect base is called with a probability of 1 in 100. For Illumina sequencing, quality commonly decreases towards the 3'-end of sequencing reads. In order to facilitate downstream processing, low quality bases are commonly trimmed from the 3'-end via tools like FASTX quality trimmer [69] or cutadapt [117].

Afterwards, additional preprocessing steps are required depending on the applied library preparation protocol. In cases where cDNA inserts are shorter than the sequenced read length, typical steps involve adapter or poly(A) clipping to ensure that resulting reads only consist of real transcriptome sequences. These and additional functions are also implemented in the FASTX toolkit [69] or cutadapt [117].

2.1.3.2 *Transcriptome profiling*

After a set of high-quality reads is generated for each library, different workflows can then be applied, depending on the presence of a reference genome or transcriptome for the used organism [34]. In cases where no reference exists, it is possible to conduct *de novo* transcriptome assembly via tools like SOAPdenovo-Trans [192], Oases [151], Trans-ABySS [65], or Trinity [67]. After *de novo* transcriptome assembly, reads are commonly mapped back to this newly generated reference using an ungapped mapper like Bowtie [96] and quantified via tools like Htseq-count [5] or RSEM [102]. Otherwise, if a transcriptome or genome assembly is available for the respective organism, the more common option is to align reads to this reference. For use with an existing reference transcriptome, reads can be aligned directly via Bowtie followed by transcript identification and read counting via, for example, RSEM. Another option is to conduct alignment-free quantification with tools such as kallisto [20] or Sailfish [134] that rely on k-mer counting in reads. For mapping to a reference genome, a splicing-aware mapper like TopHat [87, 176] or STAR [43] is used for eukaryotic organisms followed by application of, for example, Cufflinks [145] for annotation-based transcript identification or *de novo* transcript discovery and quantification. For these analysis workflows, and partly also for the preprocessing, one can select among a plethora of different analysis tools and pipelines. Because there is no optimal pipeline that covers all aspects of a specific RNA-seq analysis, sometimes different tools have to be combined to generate the desired analysis result. Figure 2.2 depicts a workflow for a typical RNA-seq experiment with different analysis paths, such as was applied in the publications presented in this thesis with slight modifications.

Next, I will give an overview of the problems that need to be addressed for calculation and comparison of expression levels between different samples based on RNA-seq data.

2.1.3.3 *Quantification and differential expression analysis*

Besides annotation of genomic features such as different kinds of transcripts, the most common application of RNA-seq is estimation of gene or transcript expression. This is primarily achieved by counting the number of reads that map to each gene or transcript or alternatively via alignment-free approaches as described above. Gene-level-based quantification can be conducted most easily via e.g Htseq-count [5], based on genomic locations of genes and exons provided in a gene transfer format (GTF) file. Importantly, it is not possible to compare expression levels among genes or samples using raw read counts, as these are affected by different

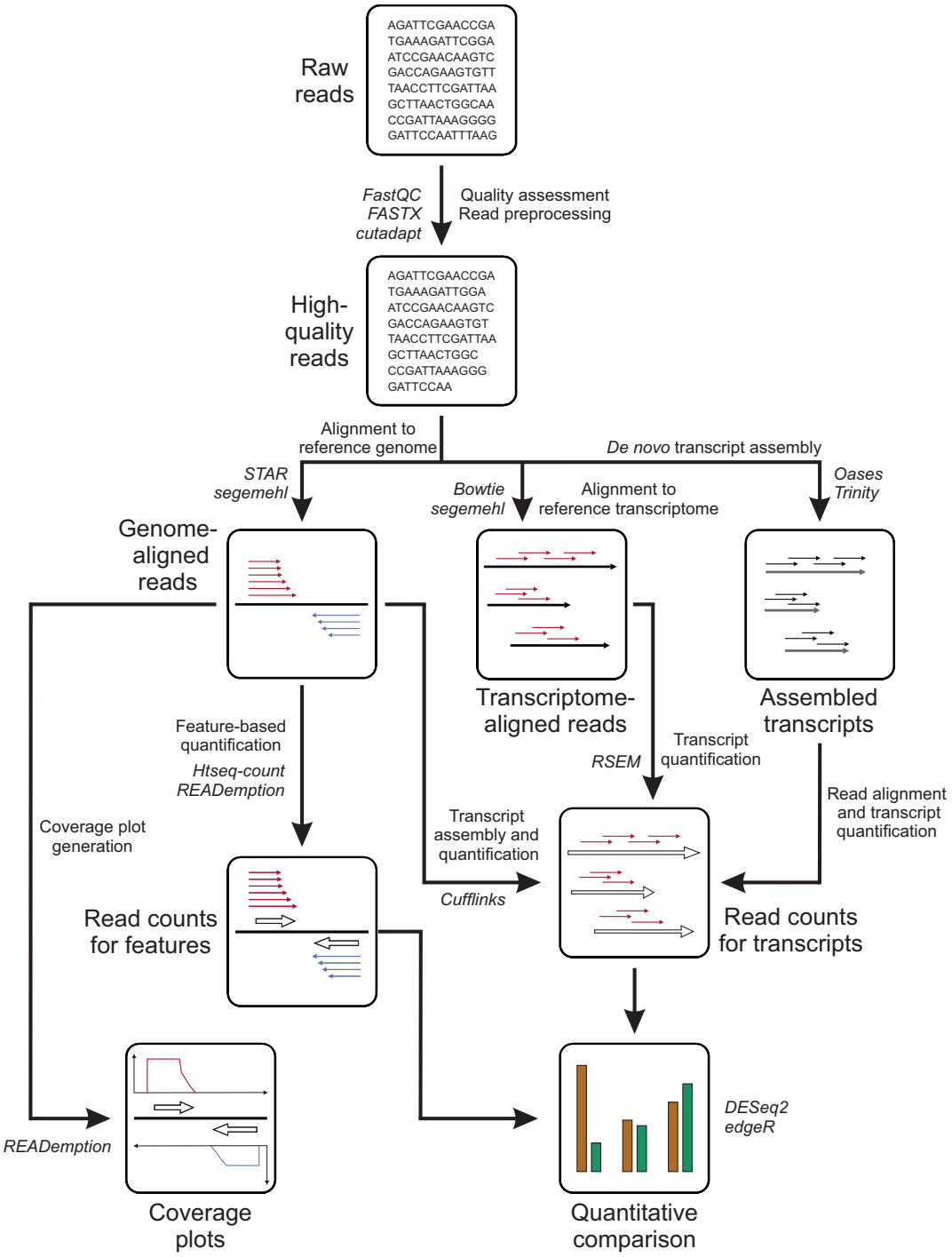


Figure 2.2: RNA-seq analysis workflow. Different analysis steps and outcomes are depicted and examples for tools (FastQC [7], FASTX [69], cutadapt [117], STAR [43], segemehl [75], Bowtie [96], Oases [151], Trinity [67], Htseq-count [5], READemption [56], RSEM [102], Cufflinks [145], edgeR [146] and DESeq2 [111]) are provided in italics. The step "Read alignment and transcript quantification" consists of the steps "Alignment to reference transcriptome" and "Transcript quantification", and can be conducted via the tools listed for these steps.

factors such as total number of reads, transcript length, and sequencing biases. Reads per kilobase of exon model per million mapped reads (RPKM) [126] is a measure to normalize read counts based on feature-length and library size. Fragments per kilobase of exon model per million mapped reads (FPKM) is the equivalent for paired-end sequencing data, where read pairs can map to the same transcript and are therefore counted only once. Together with transcripts per million (TPM), where only the order of operations in the normalization process differs and values always add up to 1,000,000, these measures are frequently used to report RNA-seq gene expression values.

When expression levels of the same feature are compared between samples, in contrast to between features in a single sample, a length normalization is not necessary. In the latter case, normalization is required to account for the fact that more reads are derived from a longer feature with the same expression level than a shorter one. Tools like Cufflinks [145] that estimate feature lengths from the data instead of using fixed annotations are likely to find significant differences in length for the same feature in different samples, which must be considered for the comparison. In general, TPM values are regarded as being more comparable among samples with approximately the same number of transcripts, since the sum over all TPM values is identical for each sample. Nevertheless, further biases might exist in the data, which have to be addressed by additional normalization techniques such as TMM (trimmed-mean of M-values) [146].

For comparison of expression levels among samples, a differential gene expression analysis is conducted. For this purpose, RPKM-, FPKM-, or TPM-normalized expression values should be avoided as they do not account for the fact that different samples might express very different RNA repertoires and can thus be heavily influenced by the presence of a small set of highly and differentially expressed features [22, 42]. Normalization methods that address this problem by ignoring such outliers include TMM [146], DESeq [4], PoissonSeq [104], or UpperQuartile [22].

These normalization methods work well for count data based on identical genomic features with a similar positional read distribution, but are not applicable for comparison on the level of eukaryotic transcripts where changes in transcript length or coverage along the trancript can occur together with additional biases. For this, more sophisticated statistical models as implemented in Cufflinks [145] or RSEM [102] are required to estimate expression levels of transcripts. One exception is, for example, DEXseq [6], which detects differential exon usage based on exonic read counts and applies the DESeq normalization [4].

Popular tools for differential expression analysis encompass methods that apply a negative binomial model as edgeR [146], DESeq2 [111], and baySeq [70], non-

parametric approaches as NOISeq [168] and SAMSeq [103], as well as methods for transcript-level-based quantification that also report differential expression on the gene-level like EBSeq [100] and Cuffdiff 2 [175]. An approach that applies a transformation of read counts to allow for linear modeling of the data is voom [98], which is used in combination with the limma package previously developed for the analysis of microarray data [144]. All tools have certain strenghts and weaknesses and no tool works best for all kinds of data. However, no matter which tool is applied, a very important aspect is the number of replicates used in an RNA-seq experiment [34]. Even if lower numbers (for example, three replicates) are common, a recent study suggests the use of at least six replicates for the design of an experiment with differential expression analysis [152].

2.1.3.4 *Bacterial RNA-seq analysis with READemption*

For the bacterial RNA-seq data presented in this thesis, I used the RNA-seq pipeline READemption [56], which was specifically developed in our lab to analyze RNA-seq data based on our library peparation method [19]. The pipeline includes size filtering of preprocessed reads and poly(A) clipping from the 3'-end. Alignment to a reference genome is conducted via the mapper segemehl [75], followed by generation of positional read coverage files in wiggle (WIG) format for visualization in a genome browser like the Integrated Genome Browser (IGB) [57] or the Integrative Genomics Viewer (IGV) [173]. Gene expression quantification is conducted based on provided gene annotations in GFF3 (gene feature format version 3) files and the resulting count data can be used for differential gene expression analysis via DESeq2 [111].

2.2 BACTERIAL TRANSCRIPTOME ANALYSIS

Bacterial transcriptome landscapes were found to be much more complex than originally thought [164]. New global approaches uncovered dense patterns of transcriptional activity along bacterial genomes that surpass the view of a simple mono- or polycistronic expression of mRNA, tRNA, and rRNA genes [38, 182].

Mapping of bacterial transcript boundaries via RNA-seq facilitates elucidation of operon structures, annotation of untranslated regions (UTRs), and discovery of novel transcripts such as sRNAs. However, due to the presence of a large amount of degradation fragments in the RNA pool, it is in most cases not possible to determine the precise genomic position of a TSS. Established methods for mapping of 5'-ends of single transcripts such as primer extension [172] or 5' rapid amplification of cDNA ends (RACE) [9, 16, 184] are time-consuming and thus cannot be

used on a global scale. RNomics approaches based on Sanger sequencing of cloned cDNAs have been used to identify bacterial sRNAs [184], but are also not applicable to whole transcriptomes. To overcome this problem, several RNA-seq-based protocols for sequencing of transcript 5′-ends have been developed [29, 30, 88, 138, 158, 190]. One of these methods is the dRNA-seq approach, which was used in several publications described in this thesis.

2.2.1 *Differential RNA sequencing*

dRNA-seq is an RNA-seq-based method that was specifically developed to annotate TSS in bacteria. First used to annotate the primary transcriptome of the human pathogen *H. pylori* [156], it was subsequently applied to a multitude of different organisms, including mainly bacteria but also archaea and eukaryotic organelles [155].

The method takes into account specific features of the RNA pool of a bacterial cell, which consists of primary transcripts that carry a 5′-triphosphate (5′-PPP) and processed transcripts with a 5′-monophosphate (5′-P) or to a lower extent a 5′-hydroxyl (5′-OH) group. dRNA-seq aims to selectively sequence primary transcripts to annotate the TSS of all transcripts in a bacterial cell. The protocol for construction of dRNA-seq libraries is explained in detail in the publication "Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in *Helicobacter pylori*" [19] presented in section 3.3. In brief, each RNA sample containing total RNA is split in two for the preparation of a matching pair of cDNA libraries. One half is treated with the enzyme terminator 5′-phosphate-dependent exonuclease (TEX) to generate a +TEX library, while the other half is left untreated to generate a −TEX library. The TEX enzyme selectively digests processed transcripts carrying a 5′-P, which results in an enrichment of primary transcripts in the +TEX sample. Apart from TEX treatment, the cDNA libraries are prepared in exactly the same way. The strand-specific protocol involves poly(A) tailing at the 3′-end of each RNA molecule followed by differential TEX treatment. Afterwards, treatment with tobacco acid pyrophosphatase (TAP) is conducted to convert 5′-PPP into 5′-P. This is required to allow ligation of a 5′ RNA linker, which cannot be ligated to a 5′-PPP or 5′-OH. Consequently, fragments with a 5′-OH group are not captured in the dRNA-seq libraries. First-strand cDNA synthesis is conducted via an oligo(dT) adapter primer and index sequences for multiplexing are incorporated during PCR amplification. As mentioned above, rRNA depletion is not required due to the inherent depletion via PAP I. Instead of applying poly(A) tailing, ligation of a 3′ linker together with a matching adapter primer would also be possible. However, this

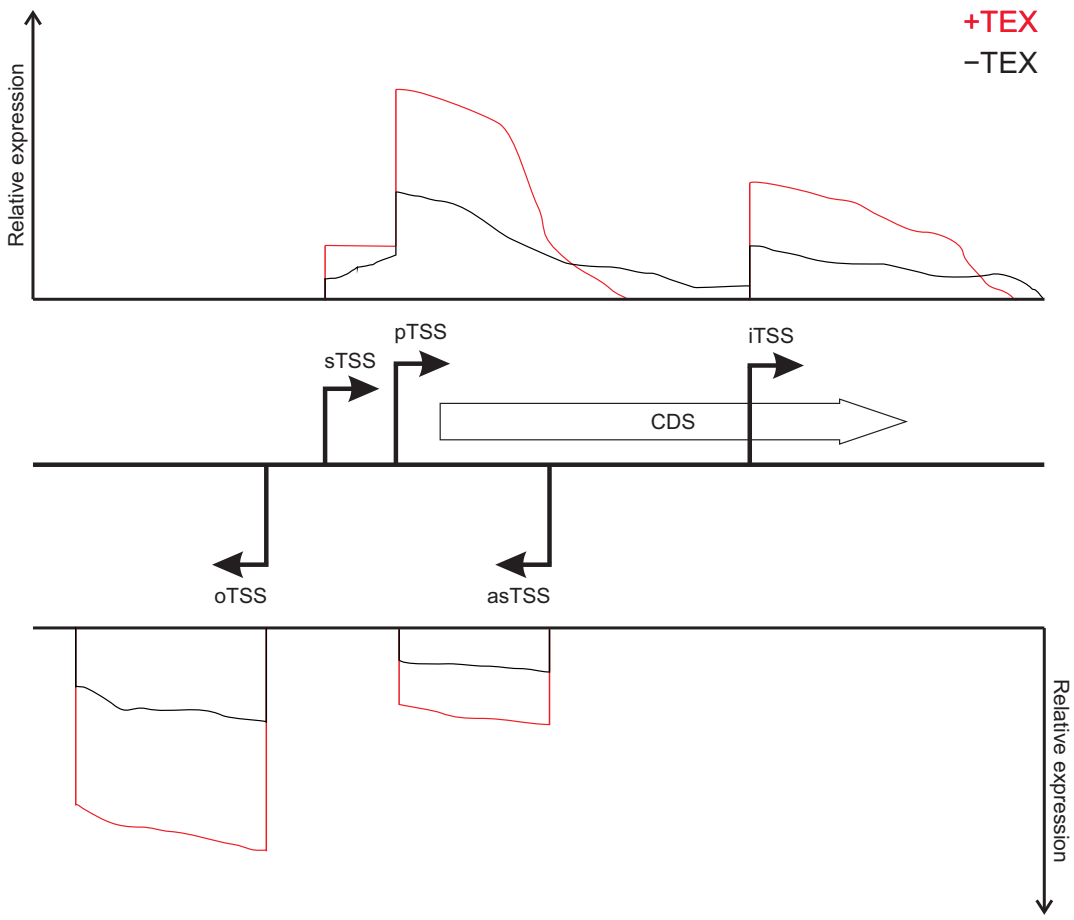


Figure 2.3: dRNA-seq enrichment and TSS classification.

again would result in higher rRNA concentrations and might require additional depletion or deeper sequencing.

Computational analysis of dRNA-seq data is conducted similar to regular RNA-seq data. Reads are quality-checked, preprocessed, and aligned to a reference genome. Expression of features is calculated based on the number of reads mapping to their genomic locations and positional coverage plots are generated for each library and strand for visualization in a genome browser or automated annotation of TSS (Figure 2.2). In these coverage plots, a charateristic enrichment of the +TEX compared to the −TEX library is observed at TSS positions, which is utilized for their annotation (Figure 2.3).

While in initial dRNA-seq studies TSS annotation was conducted via manual inspection of coverage plots in a genome browser [156], in subsequent studies, several software tools have been developed for automated TSS [3, 84] or transcript [18] annotation based on dRNA-seq data. Manual TSS annotation is laborious and in general, not reproducable and should therefore be avoided. In the dRNA-seq studies presented in this thesis [19, 170], we applied the tool TSSpredator, which was originally developed for TSS annotation of several *C. jejuni* strains [46]. The tool uses

positional coverage files as input and returns TSS positions with a classification according to their relative localization with respect to user-provided gene annotations. TSS located upstream of an annotated coding DNA sequence (CDS) are classified as primary TSS (pTSS) or secondary TSS (sTSS), with the pTSS having the highest expression level. In contrast, internal TSS (iTSS) are located inside a CDS, whereas antisense TSS (asTSS) are located on the opposite strand and within a certain distance to a CDS. TSS can be assigned to multiple of these classes and if not assigned to any class, are annotated as orphan TSS (oTSS) (Figure 2.3). Furthermore, TSSpredator is able to generate a comparative TSS map using either data from different biological conditions for a single strain or data from different closely related strains by mapping the transcriptome data to a common coordinate system, the so-called SuperGenome. In both cases, the use of replicates is supported. Despite this flexibility, application of TSSpredator is not trivial as TSS prediction strongly depends on selected values for a set of parameters, with the most important ones describing required expression and enrichment at the TSS position as well as support by a certain number of replicates. Besides TSS identification, other parameters are used to set the maximum allowed distance to annotated CDSs for classification of start sites as pTSS and sTSS on the sense strand or asTSS on the antisense strand. Choosing appropriate parameter thresholds for a specific organism and data set is a major challenge for the automated generation of a TSS map.

After a TSS map is generated, another difficulty is finding appropriate ways to make this data available to other researchers. Providing the TSS positions together with additional information in a table is useful for different kinds of downstream analysis but does not allow for visual inspection of read distribution and genomic context. A solution for this, which we used in our publications, is including this data in an easily accessible online browser like GenomeView [1].

2.2.2 *Analysis of antisense RNAs in E. coli*

One class of transcripts that can be studied via dRNA-seq are RNAs that are transcribed from the genomic strand opposite of annotated coding regions. These asRNAs have the potential to interact with their sense transcripts via complementary base-pairing or affect their expression via transcriptional interference [63, 171]. The initial dRNA-seq study identified asTSS for almost half of the annotated *H. pylori* genes [156], while subsequent studies in other bacteria reported amounts of genes with asRNAs between 2 and 30% [91, 92, 124, 130, 189, 191]. This variation could either be caused by differences in the extent of antisense transcription in the bacterial species or result from different experimental setups and data analysis methods.

Escherichia coli (*E. coli*) is a Gram-negative, facultatively anaerobic, rod-shaped Gammaproteobacterium that inhabits the gut of humans and warm-blooded animals. *E. coli* is one of the best-studied model organisms in research and different strains can be either commensals or significant human pathogens [169]. Nevertheless, the numbers of asRNAs reported by different transcriptome studies in *E. coli* [29, 30, 35, 44, 66, 88, 105, 122, 135, 136, 139, 149, 157] show a wide degree of variation, ranging from hundreds to thousands supporting the assumption that differences in the amount of reported asRNAs are likely of technical rather than biological origin.

Considering the number of reported asRNAs, even based only on the most conservative estimates, it is still surprising that only few functional members of this RNA class have been identified. Some asRNAs have been shown to affect transcription, stability or translation of their corresponding sense transcripts [63, 171]. Furthermore, asRNAs have been reported to play a role in global RNA processing in Gram-positive but not Gram-negative bacteria where they form duplexes with their overlapping sense transcripts and thereby enable digestion via the endoribonuclease RNase III [97]. Another function of asRNAs has been highlighted in a paradigm called the "excludon" where an unusually long asRNA with one or more included CDSs is transcribed opposite to divergent genes or operons with related or opposing functions. Being both an mRNA and an antisense regulator, these asRNAs can act as fine-tuning regulatory switches in bacteria [153]. Despite these findings, other studies conclude that most asRNAs result only from pervasive transcription [107, 139], inefficient transcription termination [130, 135, 136], collisions between replication and transcription machinery [133], or contamination with genomic DNA [66], and therefore do not have a biological function.

In the included study of *E. coli* (section 3.2), we used TSSpredator for the first time to generate a TSS map based on several growth conditions for a well-studied bacterial model organism. Besides selecting appropriate prediction parameters, it was also challenging to find appropriate computational methods to characterize the identified antisense transcripts and compare them to previous findings from other studies.

2.2.3 *Transcriptome mapping in Helicobacter pylori*

H. pylori is a Gram-negative, microaerophilic Epsilonproteobacterium that is present in about half of the human population. It resides in the acidic environment of the human stomach and represents a major human pathogen that can cause gastritis, peptic ulcers and gastric cancer [37, 165].

The original dRNA-seq study on *H. pylori* strain 26695, which has a genome size of ~1.6 megabases and ~1,600 annotated genes, identified more than 1,900 TSS via 454 sequencing of dRNA-seq libraries from five different biological conditions [156]. By taking into account genomic context, the TSS could be assigned to different genomic features including 5'UTRs and leaderless mRNAs, asRNAs for about half of the annotated CDSs, as well as more than 60 sRNAs. Furthermore, this study helped to elucidate operon structures, which were found to harbor a multitude of alternative suboperons.

By taking into account the existing data from the above study, we sought to conduct a comparison between manual TSS annotation based on 454 sequencing data and automated annotation via TSSpredator using the increased covarage of Illumina sequencing. Besides selection of parameters for TSS prediction, this required the development of different approaches for data comparison.

2.2.4 *Global identification of RppH targets in Helicobacter pylori*

While dRNA-seq was mainly used for TSS annotation, it can also be exploited for global annotation of processing sites. RNA degradation is an important mechanism for gene expression control in all organisms. In *E. coli* and other Gammaproteobacteria, mRNA decay is mediated by a set of RNases. It involves endonucleolytic cleavage by ribonuclease E (RNase E) and exonucleolytic cleavage from the 3'-end via polynucleotide phosphorylase (PNPase), ribonuclease II (RNase II), and ribonuclease R (RNase R) [78].

While Epsilonproteobacteria like *H. pylori* have homologs for 3'-exonucleases, they lack RNase E [85, 174]. In contrast, they contain two RNases important for RNA decay in Gram-positive bacteria like e.g. *Bacillus subtilis* (*B. subtilis*), the 5'-exoribonuclease ribonuclease J (RNase J) and the endoribonuclease ribonuclease Y (RNase Y) [118, 140, 154].

RNase E and RNase J have been shown to prefer RNA substrates with a 5'-P [113, 142]. However, as explained before, bacterial primary transcripts typically carry a 5'-PPP. Consequently, the generation of 5'-monophosphorylated substrates is an important step for RNA degradation by these enzymes. This can be achieved via two distinct mechanisms. Either monophosphorylated substrates are generated by RNase cleavage [163] or the 5'-PPP is converted to a 5'-P by the enzymatic activity of an RppH homolog [41, 142].

Since specific analysis tools were only available for the purpose of TSS annotation, a novel computational approach had to be developed which can utilize dRNA-seq data to examine processing of transcript 5' ends by RppH.

2.3 ANALYSIS OF BACTERIAL RNA-BINDING PROTEIN (RBP) INTERACTOMES

Interactions between RNA and proteins play an important role in various post-transcriptional processes. Besides RNA stability, which is affected by proteins such as, for example, RppH and various RNases, other RBPs can influence RNA structure, splicing, translation, localization, and export. Recent studies in eukaryotes identified a plethora of previously unknown RBPs and their binding sites [10, 13, 23, 90]. In contrast, only few bacterial RBPs have been characterized due to a lack of system-wide studies [14].

Hfq and CsrA are two bacterial RBPs with an important role in post-transcriptional regulation. Hfq is a key player in sRNA-mediated regulation. It serves as an RNA chaperone that promotes binding of many sRNAs to their respective target mRNAs. Hfq is conserved in about half of all bacterial species including Gammaproteobacteria like *E. coli* and *Salmonella*. Interestingly, despite the large number of sRNAs detected in Epsilonproteobacteria like *H. pylori* [156] and *C. jejuni* [46] no Hfq homolog could be identified in this bacterial class [27, 185].

CsrA, also referred to as RsmA or RsmE, is the central RBP of the widespread Csr (carbon storage regulator)/Rsm (repressor of stationary-phase metabolites) regulatory systems [147]. It primarily acts as a repressor of mRNA translation by binding to 5'UTRs [11]. In Gammaproteobacteria, the CsrB/C and RsmX/Y/Z sRNA families antagonize the function of CsrA [11, 147]. These sRNAs form structures representing several high-affinity CsrA binding sites that can titrate away the protein from its other targets [48].

C. jejuni is a Gram-negative, microaerophilic Epsilonproteobacterium that is the leading cause of bacterial food-borne disease in the industrial world [40, 195]. In this organism CsrA was shown to affect motility, biofilm formation, oxidative stress response, and infection [55], however, no global information on direct binding partners was yet available. In addition, both *C. jejuni* and *H. pylori* lack homologs of the antagonizing sRNAs [46, 156].

Salmonella enterica serovar Typhimurium is a Gram-negative Gammaproteobacterium that, as a food-borne pathogen, invades and replicates inside many eukaryotic host cells. As a bacterial model organism it has been widely used to study post-transcriptional regulation by sRNAs and the respective role of the RBPs Hfq and CsrA [72, 183, 187].

Studies using transcriptome and coIP approaches have suggested global roles for Hfq and CsrA in regulation of *Salmonella* virulence genes [8, 99, 159] but left open questions in terms of precise binding locations and mechanisms *in vivo*. While a more recent coIP approach predicted interactions of Hfq with hundreds of

sRNAs and more than thousand mRNAs [26], no such data is available for CsrA in *Salmonella*.

In order to identify binding partners of CsrA in *C. jejuni* and targets and binding sites of Hfq and CsrA in *Salmonella*, we applied two global RNA-seq-based approaches coined RIP-seq and CLIP-seq, respectively. In both cases, a major challenge was the selection of appropriate analysis software. In the following sections, I will give an overview of these experimental approaches including options for data analysis.

2.3.1 *RNA immunoprecipitation followed by sequencing (RIP-seq)*

RIP-seq is a method to identify binding sites of a specific RBP in the transcriptome to get insights into its biological function. The approach includes a coIP step where protein and bound RNA are purified from a lysed cell or tissue sample. For this, either an antibody specific for the protein of interest is used [71, 188] or, alternatively, the protein is modified with an epitope tag and an antibody against this tag is applied for the pull-down [71, 143]. Subsequently, an RNA-seq library is prepared from the extracted RNA.

In order to discriminate between real targets and unspecific RNA the use of appropriate control libraries is important. These can be based on total input RNA [188] or a pull-down via a control antibody unspecific for the protein of interest [71]. For epitope-tagged proteins, a coIP with the anti-tag antibody can be conducted on a wild-type sample where the protein of interest is untagged [143]. Figure 2.4 shows a typical enrichment in read coverage between experiment (signal) and control libraries observed for a transcript interacting with an RBP.

After sequencing and initial data processing as conducted for RNA-seq, RBP targets can be identified by calculating an enrichment score between experiment and control libraries. Here, enrichment of whole genomic features can be assessed via the same tools used for differential expression analysis based on normal RNA-seq data (see section 2.1.3). Additionally, since enrichment is not always observed spanning whole annotations or, in the case of unknown features, is located in intergenic regions, specific peak calling tools that do not rely on existing annotations have been developed. For example, Piranha [179], RIPseeker [106] and JAMM [80] are generic peak callers that can be used for RIP-seq data but also other approaches like CLIP-seq (see section 2.3.2).

The above-mentioned tools for peak detection have different drawbacks. Piranha [179] divides the genome into non-overlapping bins of a fixed size and calculates the number of read starts for each bin. Counts for control libraries can be supplied

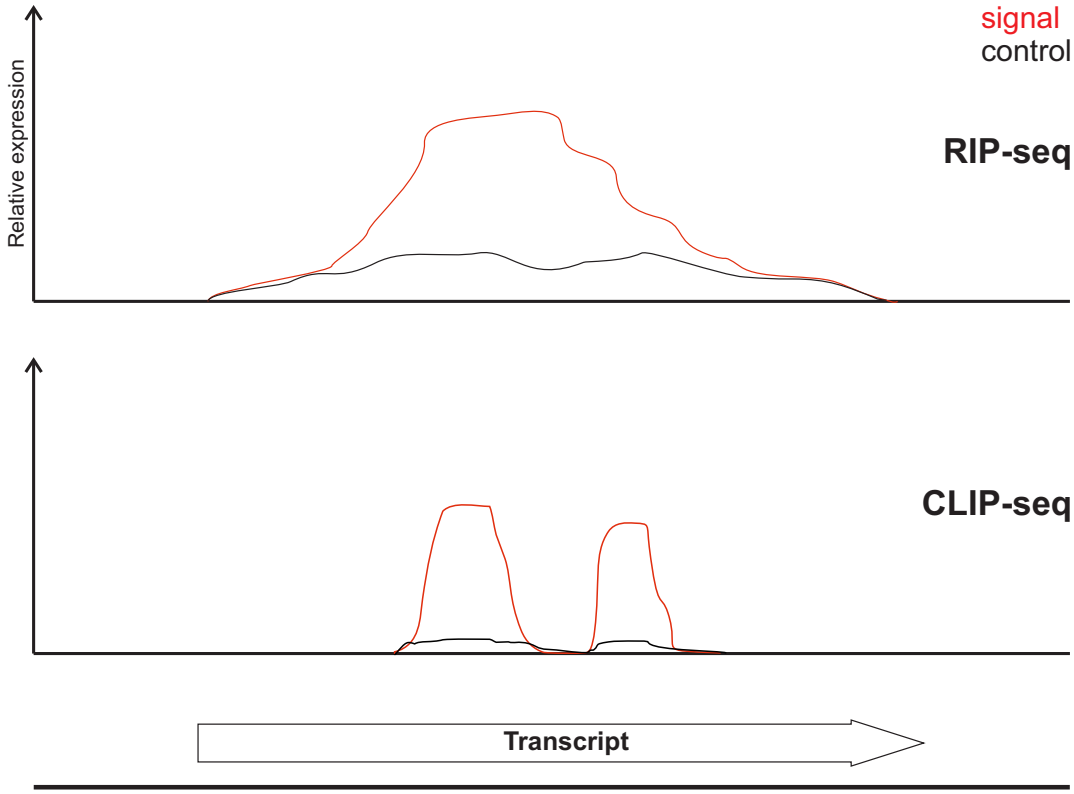


Figure 2.4: Identification of RBP binding sites.

as covariates to correct for differences in transcript abundance or protocol biases. Piranha assumes that most regions with read coverage are background and fits a distribution to calculate p-values corresponding to the probability of a bin being background. Unfortunately, the approach does not support replicates. RIPseeker [106] integrates replicate information by conducting peak detection for each replicate separately followed by subsequent merging of the predicted peaks into a consensus set. RIPseeker cannot call strand-specific peaks from the entire set of input reads but must be run separately for reads mapping to each strand. This problem also persists in JAMM [80], which in contrast is able to natively call consensus peaks based on several replicates.

2.3.2 *Cross-linking immunoprecipitation followed by sequencing (CLIP-seq)*

CLIP-seq, also known as high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP), is an extension of the RIP-seq approach that applies *in vivo* crosslinking by ultraviolet (UV) light to introduce covalent bonds between protein and bound RNA. This method has several advantages: (1) more stringent purification protocols can be used to remove unspecific RNA, (2) crosslinking allows trimming of the unprotected RNA parts to greatly increase binding-site

resolution, and (3) protein digestion often leaves a crosslinked peptide attached to purified RNA fragments, which results in mutations during reverse transcription that can be used to precisely map binding site positions [198]. In contrast to RIP-seq, control libraries can also be based on non-crosslinked versions of the CLIP samples. Figure 2.4 depicts typical read coverage patterns of RIP-seq and CLIP-seq experiment and control libraries for a transcript interacting with an RBP and exemplifies how the higher resolution of CLIP-seq allows identification of multiple binding sites while only one enriched region is detected by RIP-seq.

Several refinements of the original CLIP-seq approach have been developed. The photoactivatable-ribonucleoside-enhanced-CLIP (PAR-CLIP) method [68] uses photoreactive ribonucleoside analogs (e.g., 4-thiouridine) to enhance cross-linking efficiency and allows mapping of crosslink sites by T to C conversions in the cDNA sequence. The iCLIP method [89] takes advantage of the observation that reverse transcription frequently stops at crosslink sites, which results in cDNA fragments missed by the normal CLIP-seq approach. Library preparation for iCLIP includes these fragments to precisely map crosslink-nucleotides. Further protocols with specific enhancements encompass enhanced CLIP (eCLIP) [181] and infrared-CLIP (irCLIP) [197].

In general, peak calling for CLIP-seq data can be conducted via the same generic tools used for RIP-seq (see section 2.3.1). In addition, specific software has been developed for binding site identification based on data from CLIP-seq or related approaches. For example, PARalyzer calls crosslink sites from PAR-CLIP data by examining patterns of T to C conversions [36] and CLIPper [110] is used in the eCLIP pipeline [181] to identify peaks inside user-provided annotations based on read profiles.

Nevertheless, not all tools that are, in theory, applicable to a certain approach work equally well for each data set and most existing tools have been developed with a focus on eukaryotic data. We found our bacterial RIP-seq and CLIP-seq data to be quite complex with particularly high levels of background expression, which might interfere with appropriate detection of RBP binding. Since trials with existing tools did not yield satisfying results, development of novel peak calling approaches tailored to our data sets was required.

3

RESULTS

In the Results section, I present five manuscripts where different deep sequencing-based methods have been applied to gain novel biological insights into transcriptomes and gene regulation in different bacterial species. In these publications, I was mainly responsible for the whole or parts of the Bioinformatics analysis while wet lab experiments were conducted by other authors. Please see section a.1 in the Appendix for a listing of individual contributions.

3.1 SUMMARY OF RESULTS

3.1.1 *Analysis of antisense RNAs in E. coli*

In our publication "Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in *Escherichia coli*" presented in section 3.2, we applied the dRNA-seq approach to generate a genome-wide TSS map of *E. coli* strain K-12 substr. MG1655 based on three representative growth conditions. The strain has a genome size of ~4.6 megabases with ~4,500 annotated genes. Using different biological and technical library replicates, which were sequenced on three sequencing runs with two distinct Illumina sequencers, I examined different sources of variation in the dRNA-seq data using correlation analysis. Using TSSpredator, I predicted 14,868 potential TSS, of which 6,297 were detected under all three conditions. Using computational methods, I compared genes with a pTSS to operon annotations in the DOOR database [114] and our pTSS and sTSS to data from RegulonDB [149]. Furthermore, I examined the localization of iTSSs within genes and identified 212 divergently transcribed gene pairs with overlapping 5'UTRs.

For characterization of asRNA candidates, I compared expression levels of asTSS to those of other TSS and known asRNAs and found that most of them were expressed at lower levels compared to the other transcripts. Additionally, I calculated the overlap between our asRNA candidates to annotations from other RNA-seq-based studies [35, 44, 122, 139, 149, 157] and found large variations in numbers and only a limited amount of matching positions. Furthermore, we conducted differential expression analysis between the different conditions and analyzed promoter motifs

upstream of TSS. Finally, we experimentally verified 14 asRNA candidates via Northern analysis and found nine to be differentially affected by nucleases reported to be involved in asRNA processing. The complete *E. coli* TSS map is available via an easily accessible online browser at http://cbmp.nichd.nih.gov/segr/ecoli/.

3.1.2 *Transcriptome mapping in Helicobacter pylori*

In our publication "Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in *Helicobacter pylori*" [19] presented in section 3.3, we give a detailed description of the dRNA-seq approach using *H. pylori* 26695 as an example including all steps required for data analysis. In addition, we discuss different options for library preparation and sequencing platforms. Instead of examining different growth conditions, analyzed samples consist of several biological replicates for mid-log growth, which include the respective 454 libraries from the previous study [156] as well as three new replicates sequenced on two different Illumina sequencers. First, I conducted TSS prediction with TSSpredator using only the three Illumina replicates and compared the resulting TSS positions to the previous manual annotations [156]. Based on this analysis, we highlight differences observed between the replicates. For generation of the final TSS map and further examination of overlap to manual annotations, I also included the 454 data for TSS prediction. Based on these final annotations, we explain how TSS positions can be used to identify promoter motifs and detect regulatory elements such as riboswitches in 5'UTRs. We give examples for different genomic features that can be found, including intergenic sRNAs, *cis*-encoded asRNAs, and overlapping 5'UTRs that can result in antisense-mediated regulation. Finally, we provide the global TSS maps and cDNA coverage plots of the previous and newly generated *H. pylori* 26695 dRNA-seq data in an easily accessible online browser (http://www.imib-wuerzburg.de/research/hpylori/).

3.1.3 *Global identification of RppH targets in Helicobacter pylori*

In our study "Identification of the RNA Pyrophosphohydrolase RppH of *Helicobacter pylori* and Global Analysis of Its RNA Targets" [18] presented in section 3.4, we examined two potential RppH homologs in *H. pylori* and identified one of them as the real enzyme. We conducted *in vitro* characterization of its substrate specificity and applied a variant of dRNA-seq to globally identify transcriptome targets of RppH in *H. pylori*.

Besides TSS annotation, dRNA-seq can be used to analyze the amount of transcripts with a distinct 5′ status, i.e. monophosphorylated vs. triphosphorylated transcripts. In order to identify RppH targets, we compared relative levels and 5′-phosphorylation state of transcripts among isogenic *H. pylori* 26695 strains containing or lacking the *rppH* gene. Instead of the usual two dRNA-seq libraries, we constructed three distinct libraries from each sample by differential treatment with the enzymes TEX and TAP. TEX selectively digests processed transcripts carrying a 5′-P, while TAP converts 5′-PPPs into 5′-Ps. The library treated with both enzymes (+TEX/+TAP) is enriched for transcripts with a 5′-PPP, while the library treated only with TAP (−TEX/+TAP) captures both transcripts with a 5′-PPP and a 5′-P. The third library treated with neither TEX nor TAP (−TEX/−TAP), is specific for transcripts with a 5′-P due to the inability to ligate RNA 5′ adapters to a 5′-PPP. Library preparation, sequencing and data processing was conducted similar to the other dRNA-seq publications [19, 171]. Instead of using the resulting data for TSS prediction, I developed a computational approach for RppH target identification based on previously annotated TSS of mRNAs and ncRNAs as well as annotations for sRNAs [156]. The method takes into account changes in transcript expression based on (−TEX/+TAP) libraries and significant differences in 5′-phosphorylation based on (+TEX/+TAP) and (−TEX/−TAP) libraries between both wild type and *rppH* complementation versus the *rppH* deletion strain. Using this approach, I identified an overlapping set of 63 transcripts (53 mRNAs and 10 sRNAs) that were affected by RppH. Furthermore, we validated several of these potential RppH targets via half-life measurements and PABLO (phosphorylation assay by ligation of oligonucleotides) analysis [24, 25].

3.1.4 *CsrA target identification in Campylobacter jejuni*

In our publication "The CsrA-FliW network controls polar localization of the dual-function flagellin mRNA in *Campylobacter jejuni*" [47] presented in section 3.5, we applied a RIP-seq approach [143, 159] to globally identify RNAs that interact with CsrA in *C. jejuni* strains NCTC11168 and 81-176. For this we used chromosomally 3xFLAG-tagged and, as controls, their respective untagged wild-type strains. The coIP of protein and bound RNA was performed with an anti-FLAG antibody, and subsequent library preparation and sequencing was conducted similar to the −TEX libraries in the dRNA-seq publications [18, 19, 171]. Similar to normal RNA-seq experiments, the resulting reads were mapped to the respective reference genomes, followed by quantification and generation of positional coverage plots. Differential expression analysis of the CsrA-3xFLAG- versus control-coIP samples via Gfold

[54] identified *flaA* mRNA encoding the major flagellin as the main CsrA target with more than 300-fold enrichment, and functional analysis based on genes with >5-fold enrichment revealed an overrepresentation of the class "Surface structures", which includes a selection of flagellar genes. Visual inspection of cDNA coverage in the genome browser showed CsrA binding sites in form of enriched peaks in diverse regions of mRNA transcripts such as in 5'UTRs or between genes in polycistronic operons. To systematically identify peaks in the whole transcriptome, I developed a peak-detection algorithm based on a sliding-window approach, which uses normalized coverage files of the CsrA-3xFLAG and control coIP libraries as input to determine sites showing a continuous enrichment in the CsrA-3xFLAG-tagged library compared with the control. The approach predicted 328 potential CsrA binding sites based on a >5-fold enrichment in the NCTC11168 coIP. Motif analysis based on these peaks via MEME [12] and CMfinder [194] identified a (C/A)A(A/T)GGA sequence motif and a structural motif with AAGGA in the loop of a hairpin-structure, respectively. These findings agree with binding sites identified for other CsrA homologs [45]. Similar results were obtained for the 81-176 coIP.

Follow-up experiments revealed that indeed the *flaA* mRNA is translationally repressed by CsrA but that it also titrates CsrA activity and together with the FliW protein, which antagonizes CsrA, controls post-translational regulation of flagellar genes.

3.1.5 *Identification of RNA recognition patterns of Hfq and CsrA in Salmonella Typhimurium*

In our publication "Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking *in vivo*" [76] presented in section 3.6, we applied CLIP-seq to identify binding sites of the RBPs Hfq and CsrA in the transcriptome of *Salmonella enterica* serovar Typhimurium strain SL1344. For this study, we used *Salmonella* strains where either Hfq or CsrA was chromosomally FLAG-tagged and purification of protein-bound RNA was conducted with an anti-FLAG antibody. In order to identify binding sites with high confidence, we used three biological replicates of both crosslinked and non-crosslinked control samples for each protein. In contrast to the other studies described above, cDNA libraries were prepared using a commercial kit (NEBNext Multiplex Small RNA Library Prep Set for Illumina, #E7300, New England Biolabs) and sequencing was conducted on an Illumina NextSeq 500 in 2 x 75 cycle paired-end mode. The reason that paired-end sequencing was applied was that due to very short fragment sizes,

most fragments are sequenced completely from both directions. This yields no additional information on the genomic localization, but allows an additional interrogation of the sequence that facilitates differentiation between sequencing errors and crosslink-mutations.

I conducted data processing for this study in a similar fashion to the other publications, with the main difference that in order to remove putative PCR duplicates, which could obscure actual transcript abundance, I used the tool FastUniq [193] to collapse identical reads. Furthermore, to enhance peak resolution I conducted size filtering retaining only reads with a length between 12 and 25 nts. In addition, I addressed the presence of crosslink mutations by decreasing the required accuracy for read mapping via READemption [56] and segemehl [75] to 80%. Only uniquely aligned reads were used for further analysis.

Since none of the existing tools for peak detection yielded satisfying results, we investigated two different algorithmic options for peak calling. First, I developed an extended version of the sliding window approach used for RIP-seq-based annotation of binding regions, as described above, but with support for several replicates and using a more advanced statistical model based on repeated G–tests of goodness-of-fit [121]. In addition, our collaboration partners from the Backofen group developed a peak calling approach termed "block-based peak calling" [178]. In brief, the signal sequencing data from the crosslinked libraries is used to define clusters of blocks of overlapping reads via the blockbuster algorithm [95]. These blocks are subsequently joined to define peak boundaries using heuristics, which take into account peak shape. Finally, final binding sites are called by testing the resulting initial peaks for enrichment in comparison to the control libraries via DESeq2 [111]. Since block-based peak calling yielded slightly better results, this approach was used to identify the final set of binding sites (see section 5.3 for a discussion of both approaches).

Using peaks and identified crosslink-mutations as a basis, we conducted a thorough analysis of Hfq and CsrA interactions with their target RNAs including an investigation of the respective binding motifs. We confirmed the role of Hfq as a mediator of sRNA-target mRNA binding and explored ways to improve prediction of sRNA targets. In addition, our examination of CsrA binding sites revealed its function in direct regulation of *Salmonella* virulence genes.

3.2 GLOBAL TRANSCRIPTIONAL START SITE MAPPING USING DIFFERENTIAL RNA SEQUENCING REVEALS NOVEL ANTISENSE RNAS IN ESCHERICHIA COLI

JB Journals.ASM.org

Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in *Escherichia coli*

Maureen K. Thomason,[a*] Thorsten Bischler,[b] Sara K. Eisenbart,[a,b] Konrad U. Förstner,[b] Aixia Zhang,[a] Alexander Herbig,[c] Kay Nieselt,[c] Cynthia M. Sharma,[a,b] Gisela Storz[a]

Cell Biology and Metabolism Program, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, USA[a]; Research Center for Infectious Diseases (ZINF), University of Würzburg, Würzburg, Germany[b]; Center for Bioinformatics Tübingen (ZBIT), University of Tübingen, Tübingen, Germany[c]

While the model organism *Escherichia coli* has been the subject of intense study for decades, the full complement of its RNAs is only now being examined. Here we describe a survey of the *E. coli* transcriptome carried out using a differential RNA sequencing (dRNA-seq) approach, which can distinguish between primary and processed transcripts, and an automated prediction algorithm for transcriptional start sites (TSS). With the criterion of expression under at least one of three growth conditions examined, we predicted 14,868 TSS candidates, including 5,574 internal to annotated genes (iTSS) and 5,495 TSS corresponding to potential antisense RNAs (asRNAs). We examined expression of 14 candidate asRNAs by Northern analysis using RNA from wild-type *E. coli* and from strains defective for RNases III and E, two RNases reported to be involved in asRNA processing. Interestingly, nine asRNAs detected as distinct bands by Northern analysis were differentially affected by the *rnc* and *rne* mutations. We also compared our asRNA candidates with previously published asRNA annotations from RNA-seq data and discuss the challenges associated with these cross-comparisons. Our global transcriptional start site map represents a valuable resource for identification of transcription start sites, promoters, and novel transcripts in *E. coli* and is easily accessible, together with the cDNA coverage plots, in an online genome browser.

After many years of study, we are only now beginning to understand and appreciate the complexity of bacterial transcriptomes. With the recent advances in deep-sequencing technology, transcriptome sequencing (RNA-seq) now allows for the detection of transcripts that are present at low levels or were previously missed by other methods of detection, the generation of global transcript maps, and improved genome annotation (reviewed in references 1 and 2). While these studies provide vast amounts of information about bacterial transcriptomes and regulatory elements, they also raise challenges regarding comparisons between studies and functions of the newly identified transcripts.

One group of underappreciated transcripts being uncovered by these genome-wide analyses are RNAs that map opposite annotated coding regions, termed antisense RNAs (asRNAs). The abundance of pervasive antisense transcription start sites (asTSS) was first highlighted in an RNA-seq survey of the human pathogen *Helicobacter pylori*, where asTSS were identified opposite ~46% of the genes (3). Subsequent RNA-seq studies in cyanobacteria (4) and Gram-negative (5, 6) and Gram-positive (7–9) bacteria identified asRNAs expressed opposite 2 to 30% of annotated genes. This wide range in numbers of asRNAs reported may reflect differences in bacterial lifestyle or differences in the experimental setup or analyses of the RNA-seq data sets.

Even for the transcriptome analyses of the well-studied model organism *Escherichia coli* (10–22), the numbers of asRNAs reported range from hundreds to thousands. This significant variation is due, in part, to differences in cDNA library preparation, sequencing technology, and coverage as well as the criteria for what is considered an asRNA. For example, three different RNA-seq studies identified asRNAs opposite ~2.6% (13), ~23% (14), and ~80% (15) of genes. In another study, the number of asRNAs found opposite coding regions ranged from ~2% to ~28%, depending on the detection threshold (16).

Despite the hundreds of asRNAs reported, even using the most conservative estimates, it is surprising how few functions have been elucidated for these RNAs. A limited number of asRNAs have been shown to modulate transcription, stability, or translation of the corresponding sense transcripts (reviewed in references 23 and 24). Other recent genome-wide studies have proposed more general functions for asRNAs. These include asRNA-directed digestion of sense transcripts by RNase III in Gram-positive but not Gram-negative organisms (25) and reciprocal effects on the expression of sense RNAs in a so-called "excludon" model (reviewed in reference 26). Still other studies conclude most asRNAs lack function and result from pervasive transcription (16, 27), collisions between replication and transcription machinery (28), or inefficient transcription termination, particularly in the absence

Received 22 July 2014 **Accepted** 23 September 2014

Accepted manuscript posted online 29 September 2014

Citation Thomason MK, Bischler T, Eisenbart SK, Förstner KU, Zhang A, Herbig A, Nieselt K, Sharma CM, Storz G. 2015. Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli*. J Bacteriol 197:18–28. doi:10.1128/JB.02096-14.

Editor: R. L. Gourse

Address correspondence to Cynthia M. Sharma, cynthia.sharma@uni-wuerzburg.de, or Gisela Storz, storzg@mail.nih.gov.

* Present address: Maureen K. Thomason, Department of Microbiology, University of Washington, Seattle, Washington, USA.

M.K.T. and T.B. are joint first authors.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/JB.02096-14.

doi:10.1128/JB.02096-14

of the Rho protein (9, 17, 18), or correspond to contaminating genomic DNA (22).

To further explore the *Escherichia coli* transcriptome on a genome-wide scale, particularly the subset of asRNAs, we carried out differential RNA sequencing (dRNA-seq) analysis (reviewed in reference 29), which we analyzed by an automated TSS prediction algorithm (30). This approach led us to identify, across three growth conditions, >5,500 potential TSS within genes, 212 divergently transcribed gene pairs with overlapping 5′ untranscribed regions (UTRs), and >5,400 potential asRNA loci. We examined expression of 14 candidate asRNAs by Northern analysis and found 9 to be differentially degraded by RNase III and RNase E, two RNases implicated in asRNA-based regulation. Our global TSS map is one of the best and most sensitive data sets for promoter and transcript identification in the widely used model organism *E. coli* and is easily accessible at RegulonDB (21) and via an online browser at http://cbmp.nichd.nih.gov/segr/ecoli/.

MATERIALS AND METHODS

Strain construction. The strains and oligonucleotides used for this study are listed in Tables S1 and S2, respectively, in the supplemental material. The asRNA deletion control strains were constructed using λ Red-mediated recombination (31) to replace the region encompassing the asRNA signal along with 300 nucleotides (nt) on either side with a kanamycin cassette. Deletion constructs were confirmed by sequencing and moved into new wild-type or mutant backgrounds by P1 transduction.

Growth conditions. Cells were grown at 37°C in LB (10 g of tryptone, 5 g of yeast extract, 10 g of NaCl per liter) or M63 minimal glucose medium (supplemented with final concentrations of 0.001% vitamin B_1 and 0.2% glucose) to an optical density at 600 nm (OD_{600}) of ~0.4 and 2.0 for LB and an OD_{600} of ~0.4 for M63. At the indicated OD_{600}, 25 ml of cells (OD_{600} of 0.4) or 5 ml of cells (OD_{600} of 2.0) was combined in a 5:1 ratio of cells to stop solution (95% ethanol, 5% acid phenol [pH 4.5]), vortexed, incubated on ice for 10 min, and collected by centrifugation. Cell pellets were snap-frozen in an ethanol-dry ice slurry and stored at −80°C.

Deep-sequencing sample preparation. Details for sample preparation for deep sequencing can be found in Materials and Methods in the supplemental material. Briefly, RNA extraction for RNA-seq analysis was performed as described previously using hot-acid phenol chloroform (3, 32). RNA samples were treated with DNase I to remove contaminating genomic DNA. RNA samples free of genomic DNA were treated with terminator 5′-phosphate-dependent exonuclease (TEX) (Epicentre) followed by tobacco acid pyrophosphatase (TAP) treatment (Invitrogen) as described previously (3). Control reactions lacking terminator exonuclease were run in parallel for each sample. Unfractionated total RNA was used to construct cDNA libraries for sequencing on GAIIx and HiSeq 2000 machines.

Analysis of deep-sequencing data. For a detailed description of the read mapping, expression graph construction, normalization of expression graphs, correlation analysis, TSS prediction, comparison to other data sets, and other computational analyses, see Materials and Methods in the supplemental material.

(i) Read mapping. Between 1.8 and 9.8 million reads for each of the cDNA libraries were mapped to the *E. coli* MG1655 genome (NCBI accession no. NC_000913.2 [24 June 2004]) using our RNA-seq pipeline READemption (33) and *segemehl*, with an accuracy cutoff of 95% (34).

(ii) Correlation analysis. Nucleotide- and gene-wise Spearman and Pearson correlation coefficients were calculated based on concatenated values of forward and reverse strand position-wise coverage files and visualized using the R package *corrplot*. Gene-wise correlation values utilized read overlap counts based on NCBI annotations (accession no. NC_000913.2).

(iii) TSS prediction. Transcriptional start site (TSS) prediction was performed using the program TSSpredator (http://it.inf.uni-tuebingen.de/TSSpredator) (30). TSS were classified as primary TSS (pTSS), secondary TSS (sTSS), asTSS, internal TSS (iTSS), or orphan TSS (oTSS) based on the location relative to gene annotations. pTSS and sTSS are within 300 nucleotides upstream of a gene, with pTSS having the highest expression values. All other TSS associated with the gene are considered secondary. iTSS are internal to a gene on the sense strand, while asTSS are internal or within 100 nucleotides of a gene on the opposite strand of the annotation. oTSS do not meet any of the above requirements.

(iv) Comparison to DOOR. A table containing all operon annotations (1,526 single-gene operons and 851 operons consisting of multiple genes) was downloaded from the Database of prOkaryotic OpeRons (DOOR) 2.0 website (35) and compared to a final set of 2,441 TSS.

(v) Comparison of pTSS and sTSS to RegulonDB promoters. We extracted 6,406 TSS annotated based on the "strong evidence" classification (21) from the RegulonDB promoter table (version RegulonDB 8.5, 11-28-2013) and classified them according to our classification scheme, resulting in a set of 3,987 pTSS and sTSS. We conducted a pairwise comparison of the positions to our data (4,261 pTSS and sTSS) based on a maximum allowed distance of 3 nt.

(vi) Expression analysis and binning. Expression values for predicted TSS classified as exclusively antisense or exclusively primary or secondary were calculated based on overlap counts for a 50-nt window downstream of the respective TSS position from which reads per kilobase per million mapped reads (RPKM) values were calculated (36). The TSS were grouped into six bins according to their RPKM values.

(vii) Comparison of asRNAs detected in our and previous studies. asTSS annotations were retrieved from the Materials and Methods sections in the supplemental material from published studies (13, 14, 16, 19) or were downloaded from RegulonDB (data set version 3.0 [21] and data set version 2.0 [20]). We excluded the study by Li et al. (15), which revealed >82,000 asTSS, as this number is very high compared to previous studies and our study and thus would bias the comparative analyses. We compared the asTSS from each data set, including our 6,379 predicted asTSS, to the asTSS of all other data sets in a pairwise manner, requiring either a precise match of the annotated positions or allowing a variation of 1, 2, 3, or 10 nt.

Northern analysis. RNA extraction for Northern analysis was performed using TRIzol reagent (Invitrogen). Northern analysis of 10 μg of total RNA was performed on denaturing 8% acrylamide–7 M urea gels as described previously (37), with minor changes for detection using riboprobes (for details and oligonucleotides used to create the riboprobes, see Materials and Methods in the supplemental material).

RNA-seq data accession number. Raw sequence reads were uploaded to the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo) under accession no. GSE55199.

RESULTS

dRNA-seq reveals the primary transcriptome of *E. coli* MG1655. To detect the transcripts expressed by *E. coli*, we collected two independent biological replicates (B1 and B2 samples) from MG1655 wild-type cells grown to the exponential phase (OD_{600} of ~0.4) or stationary phase (OD_{600} of ~2.0) in LB medium (samples LB 0.4 and LB 2.0, respectively) or grown to the exponential phase (OD_{600} of ~0.4) in M63 minimal glucose medium (sample M63 0.4) (Fig. 1; see also Table S3 in the supplemental material). For all six biological samples, total RNA was extracted and subjected to dRNA-seq library preparation for primary transcriptome analysis as described previously (3). Specifically, prior to cDNA library construction, half of each RNA sample was treated with 5′ terminator exonuclease (+TEX samples), which degrades RNAs containing a 5′-monophosphate (5′-P), thereby enriching for primary transcripts containing 5′-triphosphates (5′-PPP). The other half of each sample was left untreated (−TEX samples) and thus contains both primary transcripts (5′-PPP) and processed RNAs

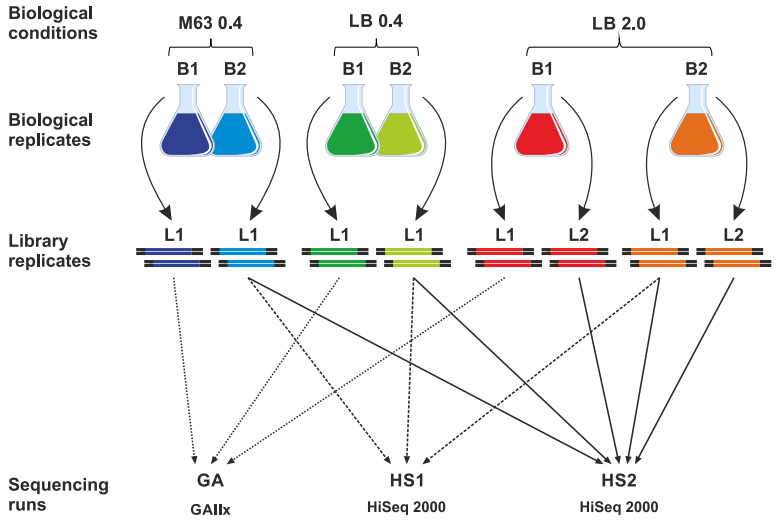


FIG 1 Summary of the biological, library, and Illumina sequencing replicates that were subjected to dRNA-seq analysis in this study.

(5′-P). Subsequently, the 5′-PPP ends in both samples were converted to 5′-P ends for cDNA library preparation.

The cDNA libraries of the first biological replicates (B1-L1) were sequenced on an Illumina Genome analyzer IIx (GA samples), while the second biological replicates (B2-L1) were sequenced on a HiSeq 2000 sequencer (HS1 samples). To examine variation between sequencing runs, the B2-L1 libraries were resequenced using the HiSeq 2000 (HS2 samples). To identify variation introduced during library preparation, technical replicates of the LB 2.0 libraries (B1-L2 and B2-L2 samples) were also generated and sequenced using the HiSeq 2000 (Fig. 1; see also Table S3 in the supplemental material).

Strand-specific sequencing resulted in a total number of ~1.8 to 3.6 million reads per sample for the GA set and ~5.3 to 9.8 million reads per sample for the HS sets after quality trimming (see Table S4 in the supplemental material). For all of the libraries, >70% of the reads could be mapped to the *E. coli* genome (NCBI accession no. NC_000913.2) indicating that the sequencing runs consisted of numerous high-quality reads. Read mapping analysis showed that for all three growth conditions, 65 to 80% of reads mapped to annotated regions of the genome while 2 to 6% mapped antisense to published annotations. The remainder of the reads mapped to unannotated intergenic regions, which also include UTRs (see Table S5 in the supplemental material). These data indicate the majority of transcripts correspond to the sense strand of genes; however, a small percentage of antisense transcription occurs, particularly opposite mRNAs.

Correlation analysis reveals variation associated with library preparation and sequencing platform. To assess the similarity between replicates, we calculated Spearman and Pearson correlation coefficients for nucleotide-wise expression values for both strands of all the −TEX and +TEX libraries (see Fig. S1 in the supplemental material). For each biological condition and both types of analysis, we noted the highest correlation among sequencing replicates (B2-L1-HS1 and B2-L1-HS2). The lowest correlation was between libraries sequenced on the GAIIx and HiSeq 2000, likely due to differences in sequence coverage and cDNA library preparation protocols for the two platforms. Since the nucleotide-wise correlations are sensitive to slight fluctuations in cDNA read counts, we also assessed the correlation coefficients for gene-wise expression values, defined as the number of mapped reads within genes annotated by NCBI, among the −TEX and +TEX libraries. Overall the correlation increased but had a pattern similar to that seen for the nucleotide-wise comparisons.

Despite the high correlation between replicates and overall similar cDNA coverage patterns, a few regions showed variable expression or enrichment in the +TEX libraries across samples, likely due to the number of reads produced by the different sequencing instruments combined with differences in library preparation. However, as we had high correlation between replicates, similar read distributions across replicates, and agreement on the positions of transcript ends, we proceeded with automated genome-wide TSS annotation.

The automated TSSpredator pipeline predicts previously unannotated TSS. Several RNA-seq-based studies have reported genome-wide annotations of 5′ ends of *E. coli* genes, but most cannot distinguish between primary and processed transcripts, limiting the potential to identify these distinct types of transcripts (12, 20). Our dRNA-seq approach allows for the precise annotation of TSS based on a characteristic enrichment pattern in the +TEX libraries relative to the −TEX libraries, which facilitates the differentiation between primary (5′-PPP) and processed (5′-P) transcripts (see Fig. S2A in the supplemental material) (3). In previous dRNA-seq studies, global (TSS) annotations were carried out by laborious manual inspection of enrichment patterns (3, 5, 6). To automate this annotation step, we utilized the TSSpredator pipeline recently developed to annotate TSS among multiple strains of *Campylobacter jejuni* (30). The TSSpredator prediction algorithm employs the dRNA-seq data to determine the location of a TSS based on identifying positions with sharp increases in expression in the +TEX library relative to the untreated −TEX control (see Fig. S2A and Materials and Methods in the supplemental material).

Using TSSpredator, TSS can be annotated in a comparative

manner among libraries through the integration of replicate information. If a strong enrichment is observed in one replicate, less strict parameters can be applied to the same position in other replicates to ensure identification of TSS despite differences in library or sequencing preparations while still maintaining stringent criteria for detection. To perform such an analysis for our replicates of the three biological conditions (see Materials and Methods in the supplemental material), we adjusted the "matching replicates" parameter, which defines the minimum number of replicates in which a TSS must be detected for a particular biological condition. For the M63 0.4 and LB 0.4 conditions, where only three replicates were available, we required a TSS to be detected in at least two replicates, while for the LB 2.0 condition, we required detection in at least three of the five replicates. All other parameters were set to default values as established previously (30).

We predicted a total of 14,868 potential TSS mapping throughout the *E. coli* genome (see Data set S1 in the supplemental material). Of these, 6,297 were detected under all three conditions, 1,151 were detected only in cells growing exponentially in M63 minimal medium, 470 TSS were found in cells growing exponentially in LB, and 1,947 were found in stationary-phase cells growing in LB (Fig. 2A; see also Fig. S3A in the supplemental material for examples of TSS detected under only one condition). The higher number of TSS identified for the LB stationary-phase cells might be a result of changes in transcriptional programs required to survive in the stationary phase (38).

TSSpredator automatically assigns TSS to five different classes: primary TSS (pTSS; main transcription start of a gene or operon), secondary TSS (sTSS; alternative start with lower expression), internal TSS (iTSS; start within a gene), antisense TSS (asTSS; transcript start antisense to a gene ±100 nt), and orphan TSS (oTSS; not associated with annotation) based on the location relative to existing gene annotation (see Fig. S2B in the supplemental material). A TSS can fall into more than one category, depending on its location relative to the surrounding gene annotations. For example, in the case of overlapping 5′ UTRs, a particular TSS can be both a pTSS and an asTSS. For downstream genes within operons, a pTSS can also be internal to the upstream genes. Among the 14,868 predicted TSS, we identified 2,672 pTSS (1,707 classified solely as pTSS), 1,589 sTSS (850 classified solely as sTSS), 5,574 iTSS (4,466 classified solely as iTSS), and 6,379 asTSS (5,495 classified solely as asTSS) (Fig. 2B).

To assess the coverage of our TSS predictions, we compared the number of TSS classified as pTSS only or pTSS and asTSS (2,057) and the number classified as pTSS and iTSS or pTSS, iTSS, and asTSS (615) with the number of genes classified as single-standing genes (1,526) or first genes within operons (851) in the Database of prOkaryotic OpeRons (DOOR) (35). In total, after excluding all TSS assigned to genes not annotated in DOOR (see Materials and Methods in the supplemental material), we used 2,441 of our TSS classified as pTSS. In agreement with the assumption that a pTSS must precede genes annotated as single genes or first genes in DOOR, we detected a pTSS for ~78% of the single-standing or first genes in operons (1,847/2,377) (see *iclR* in Fig. S3B in the supplemental material). The ~22% of single or first genes of operons for which no pTSS was predicted by our data (530/2,377) (see *ybeT* in Fig. S3B) generally were missed due to low read coverage. For several of the genes without detected TSS, we found a processing site upstream, as indicated by an enrichment in the −TEX compared to the +TEX libraries,

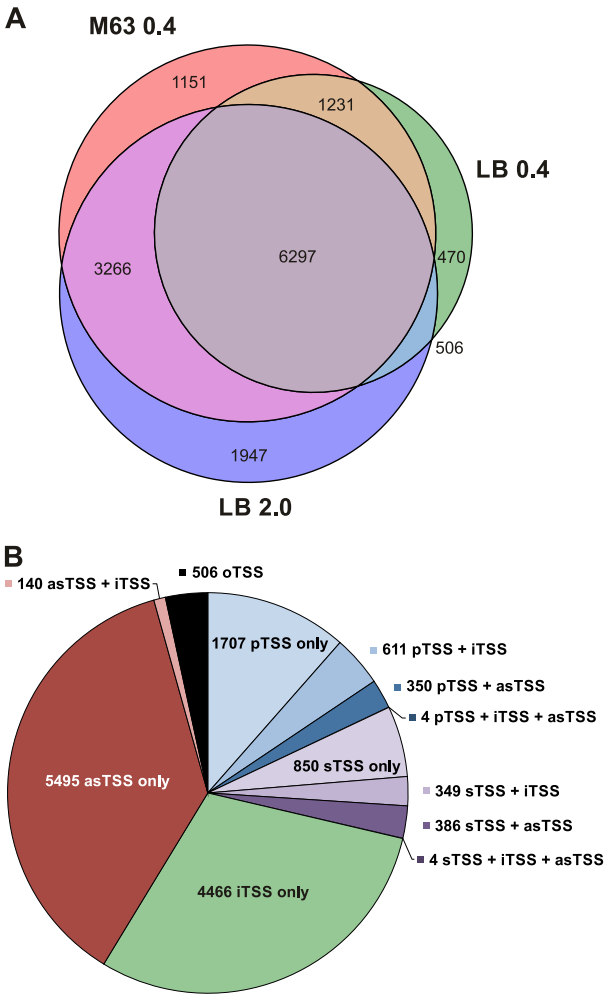


FIG 2 Automated TSS prediction across three different growth conditions using TSSpredator. (A) Distribution of predicted TSS across the biological conditions M63 0.4, LB 0.4, and LB 2.0. (B) Distribution of predicted TSS in the primary, secondary, internal, orphan, and antisense TSS classes (pTSS, sTSS, iTSS, oTSS, and asTSS, respectively).

indicating that they could be cotranscribed with upstream genes (see *fbaA* in Fig. S3B).

Approximately 24% (594/2,441) of genes for which we detected a pTSS were not classified in DOOR as single or first genes in an operon. The majority of these TSS likely correspond to real promoters that are located internal to upstream genes within an operon defined by DOOR (see *thrA* in Fig. S3B in the supplemental material). These TSS could drive transcription of unannotated alternative suboperons and thereby uncouple expression of the subset of genes from the longer operon. Some of these TSS are also found upstream of genes previously predicted to be in operons but are likely single genes (see *pheM* in Fig. S3B). Overall, these comparisons indicate that despite previous global transcriptome studies, the full complexity of the *E. coli* transcriptome is not yet known.

A comparison of our TSS predictions with TSS annotated in

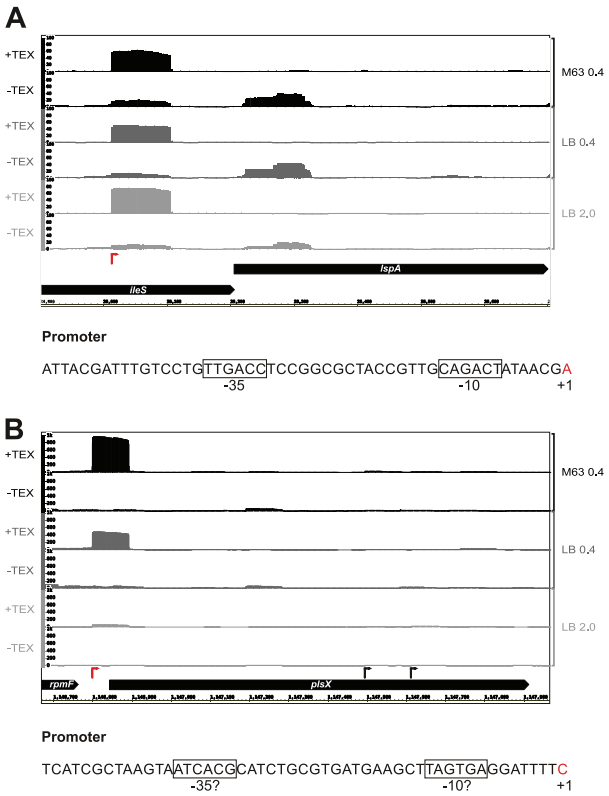


FIG 3 Examples of genes with newly detected pTSS. Screenshots showing the relative cDNA coverage plots for representative −TEX or +TEX libraries for the M63 0.4, LB 0.4, and 2.0 growth conditions across the genomic regions encompassing the *lspA* (A) and *plsX* (B) genes. The *x* axis depicts the genomic coordinates, while the *y* axis indicates the relative cDNA scores (normalized number of mapped cDNA reads). Red arrows indicate the previously unannotated TSS detected by our analysis. Promoter sequences for the new TSS, including the −10 and −35 sequences (boxed) and bases corresponding to TSS (red) are depicted below each plot.

RegulonDB (21), using a maximum distance of 3 nt, revealed that ~34% of our pTSS and sTSS overlap those annotated in RegulonDB (see Data set S1 in the supplemental material), while ~41% of the TSS from RegulonDB, classified as pTSS or sTSS, overlap our predictions. A TSS detected in our data but previously not annotated in RegulonDB is the pTSS for *lspA*, encoding a prolipoprotein signal peptidase, located internal to the upstream *ileS* gene (Fig. 3A). A promoter corresponding to the TSS based on direct experimental evidence was previously reported (39). Figure 3B shows a clear exponential-phase-specific pTSS for *plsX*, encoding a putative phosphate acyltransferase, although no evidence was present in RegulonDB, and the sequence does not carry an obvious promoter consensus sequence. These discrepancies illustrate that, even in a well-studied model organism like *E. coli*, TSS annotation is still incomplete. We next carried out further characterization of the noncanonical iTSS and asTSS.

iTSS are abundant and frequently located at the 3′ ends of genes. We identified 5,574 iTSS internal to annotated genes (Fig. 2B). It was recently reported that the majority of iTSS identified in the Gram-negative bacterium *Shewanella oneidensis* are present near the 5′ or 3′ ends of the genes (40). For a comparison, we examined the location of the 4,466 iTSS classified as iTSS only as one group and the 968 iTSS that are also annotated as pTSS or sTSS as a second group. Each annotated gene in which an iTSS was detected was divided into 10 equal sections, and the number of iTSS located in each section was counted for all genes. Those classified as iTSS only showed a broad distribution with similar numbers across the entire gene (see Fig. S4A in the supplemental material). In contrast, for the group of iTSS also classified as pTSS or sTSS, the majority (~86%) were located in the last 30% of the gene (see Fig. S4B). These 86% are likely TSS for downstream genes, driving alternative expression of suboperons (for an example, see *thrA* in Fig. S3B in the supplemental material) or the synthesis of small regulatory RNAs corresponding to the 3′ ends of mRNAs as was observed for the MicL RNA, whose promoter is within the *cutC* gene (41). Whether any of the iTSS in other categories result from spurious transcription or are generating functional alternative mRNAs or regulatory RNAs will require further characterization.

pTSS and sTSS from divergently transcribed gene pairs could also serve as asRNA regulators. In addition to the 5,495 TSS classified as asTSS only, we identified 350 pTSS and 386 sTSS that are also classified as asTSS. Examination of the regions encompassing these TSS revealed 212 divergently transcribed gene pairs with possible overlapping 5′ UTRs (see Data set S2A in the supplemental material), which could result in asRNA-mediated regulation of these genes (reviewed in reference 26) or could influence promoter occupancy (42). The set includes several gene pairs that encode proteins of opposing function, such as *entS* and *fepD*, encoding an enterobactin efflux system and a ferric enterobactin ABC transporter, respectively, and *pspF* and *pspA*, encoding the transcription factor PspF (phage shock protein F) and its antagonizing regulatory protein, PspA (see Fig. S5 in the supplemental material). Further characterization of these gene pairs will be required to determine if asRNA-mediated regulation occurs via the overlapping 5′ UTRs.

Some asTSS show high or differential levels of expression. Given that several asRNAs with characterized functions are expressed at high levels (reviewed in reference 43), we compared the relative expression levels for the 5,495 asTSS only (see Data set S3 in the supplemental material) to all pTSS only and sTSS only (see Data set S1 in the supplemental material) and TSS corresponding to known annotated asRNAs (see Table S6 in the supplemental material). We calculated reads per kilobase per million mapped reads (RPKM) values for all libraries utilizing a 50-nt window downstream of the predicted asTSS. The TSS were subsequently grouped into <10, 10 to 10^2, 10^2 to 10^3, 10^3 to 10^4, 10^4 to 10^5, and $>10^5$ bins according to their RPKM values. Using the highest expression value for a TSS among all conditions, we plotted the distribution of exclusively asTSS, exclusively pTSS or sTSS, and known annotated asRNAs for all conditions (Fig. 4A) and for each library individually (data not shown). For both the combined and individual sets, the distributions for pTSS or sTSS differ from the distributions for asTSS. Most pTSS or sTSS group within the 10^2 to 10^3 and 10^3 to 10^4 expression bins, while most asTSS are in the 10 to 10^2 and 10^2 to 10^3 expression bins (Fig. 4A). In contrast, most previously annotated asRNAs fall into the 10^3 to 10^4 and 10^4 to 10^5 bins. These distributions suggest that while a subset of asRNA candidates could be present at a high level, the majority of our predicted asRNA candidates might only exist as a few copies

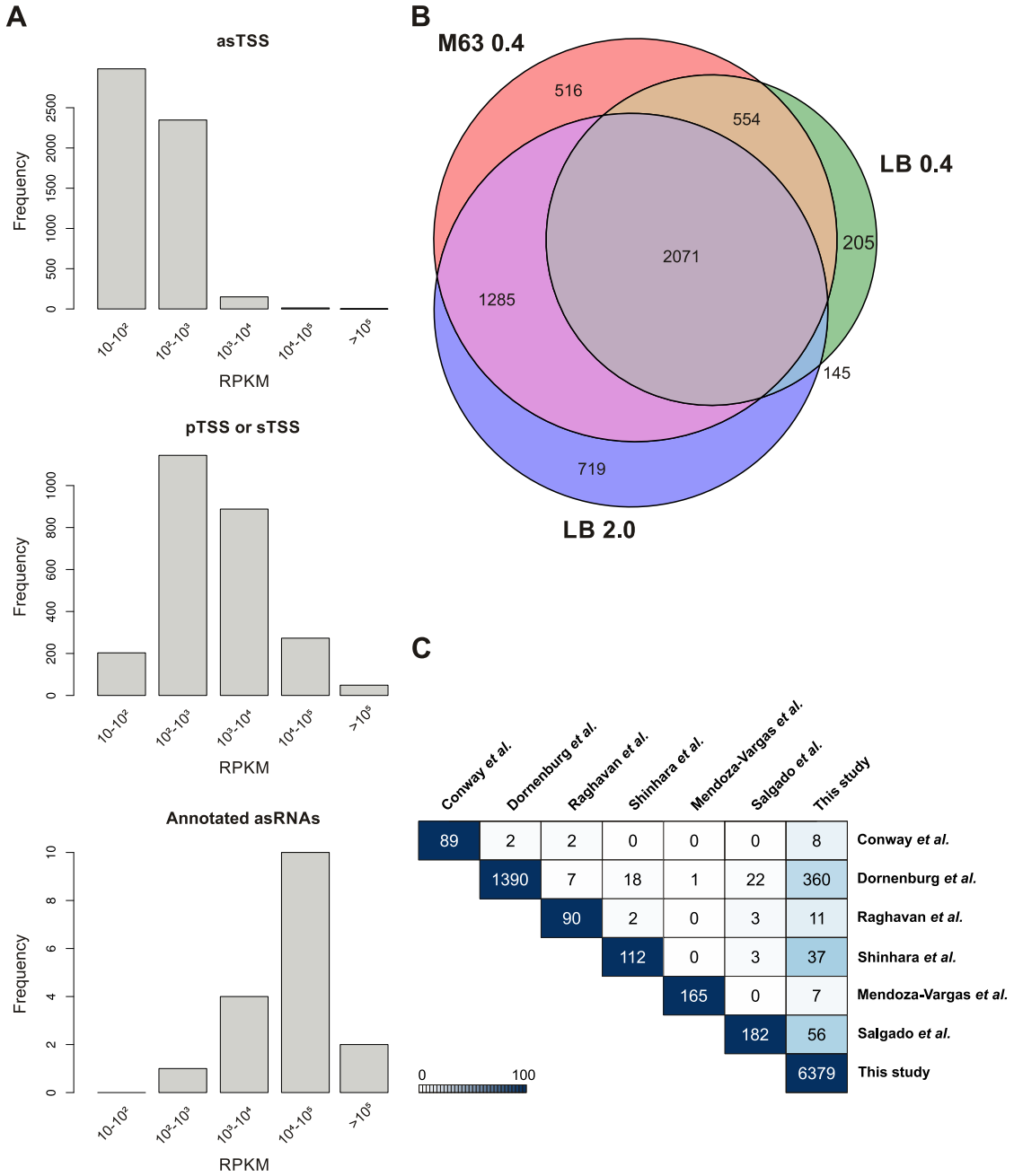


FIG 4 Comparison of asTSS. (A) Distribution of only asTSS, only pTSS or sTSS, and NCBI-annotated asRNAs in RPKM expression bins. The RPKM expression values were calculated based on cDNA read counts within 50-nt windows starting at the TSS. (B) Distribution of TSS classified exclusively as asTSS across the three biological conditions M63 0.4, LB 0.4, and LB 2.0. (C) Pairwise comparison of asTSS identified by our study and in previously published studies by Conway et al. (19), Dornenburg et al. (14), Raghavan et al. (16), Shinhara et al. (13), Mendoza-Vargas et al. (20), and Salgado et al. (21). The total numbers of annotated asTSS are shown on the main diagonal of the matrix. asTSS from the studies in the rows are compared to the studies in columns, and the number of TSS with exact matches is reported in the matrix entries. The background color depicts the percentage of overlapping asTSS relative to the total number of asTSS from the study in the particular row.

per cell or might be unstable transcripts that are rapidly degraded during RNA isolation and library preparation.

Since several functional asRNAs are expressed under specific conditions (44, 45), we also examined the distribution of the predicted asTSS across the different growth conditions (Fig. 4B). A total of 2,071 of the 5,495 asTSS were detected under all conditions. In general, candidate asRNAs in the $>10^5$ expression bin, showed a high signal for all growth conditions and library repli-

cates (see Data set S3 in the supplemental material). Like the overall TSS distribution, most condition-specific asTSS were detected in LB 2.0 (719), many of which are found in the 10^4 to 10^5 expression bin, followed by M63 0.4-specific asTSS (516), and LB 0.4-specific asTSS (205). There was significant overlap (1,285) between asTSS detected in exponential growth in M63 minimal glucose and stationary-phase LB medium, but limited overlap (145) between asTSS detected in the exponential- and stationary-phase LB samples. Again, these distributions mirror the ratios in the overall transcription profiles.

The majority of pTSS, iTSS, and asTSS are preceded by σ^{70} promoter elements. To detect potential differences between the promoters corresponding to the pTSS, iTSS, and asTSS, we compared the difference in expression for the pTSS only, iTSS only, and asTSS only detected in LB 0.4 with those detected in M63 0.4 (see Fig. S6 in the supplemental material). Overall, there were proportionally more pTSS showing differential expression than iTSS and asTSS. This suggests that the pTSS generally are more highly regulated.

We also examined the sequences upstream of the 1,707 pTSS only, 4,466 iTSS only, and 5,495 asTSS using the MEME software (46). With a window of −50 to +1 relative to the TSS, the promoter motifs derived for the three classes of TSS overall were very similar (see Fig. S7 in the supplemental material). All had a potential −10 element resembling the TATAAT consensus for the housekeeping σ^{70} transcription factor (reviewed in reference 47). The enrichment for two T residues comprising a potential σ^{70} −35 element was significantly less than what was observed for the −10 element; however, both pTSS and iTSS logos showed some enrichment for a G at position −14, characteristic of an extended −10 sequence associated with σ^{70} promoters with weak −35 elements. A window of −50 to +5 relative to the TSS revealed that a subset of pTSS, iTSS, and asTSS show enrichment for a purine at +1 and a pyrimidine at −1, features of *E. coli* σ^{70} promoters reported previously (10). Overall, despite differences in the dRNA-seq signal, most of the pTSS, iTSS, and asTSS are likely transcribed by the σ^{70} holoenzyme.

Comparison of asTSS prediction with published data sets reveals limited overlap in candidate asRNAs. A number of transcriptome data sets have recently been published for *E. coli* with different extents of antisense transcription reported (13–16, 19–21). Given the discrepancy in numbers of annotated asRNAs, we were interested in the extent of overlap between our asRNA predictions and those of the other studies. For our cross-study comparison, we only included studies where detailed asRNA annotations were provided. We compared our asTSS only (see Data set S3 in the supplemental material) to the asRNA candidates reported by each group rather than to the primary data, given the differences in data generation, analysis, quality, and quantity of reads mapping to the *E. coli* genome (see Table S7 in the supplemental material). We first required the TSS positions between two studies to match precisely (Fig. 4C). This resulted in very limited overlap across the studies. The largest overlap occurred between our data set and that of Shinhara et al. (13), with 33% of their asRNAs overlapping our predictions. In some cases, increasing the window size within which an asTSS could match, to 1, 2, 3, or 10 nt, increased the overlap between studies (see Fig. S8 in the supplemental material). For example, with the 1-nt window, 79% (71/90) of the asRNAs detected by Raghavan et al. (16) corresponded to an asTSS in our data compared to ~12% (11/90) when an exact match was required. In other cases, the increase in window size did not make much difference. There was no overlap between the asRNAs predicted by Mendoza-Vargas et al. (20) compared to Raghavan et al. (16), Shinhara et al. (13), and Salgado et al. (21), regardless of the window size. The discrepancies between the asTSS reported likely result from combinations of differences in the quality of the sequencing reads, analysis pipelines, expression cutoffs, and definitions of what constitutes an asRNA.

We also compared our asTSS map to a recent study by Lybecker et al. examining the double-stranded transcriptome of *E. coli* (48). The premise of this study was that RNAs under asRNA-mediated control would be present in double-stranded RNA duplexes and thus should be identified by coimmunoprecipitation (co-IP) with a double-stranded RNA (dsRNA)-specific antibody followed by RNA-seq. We compared these reported IP dsRNAs to our asTSS set and considered them to match if an asTSS is found within the region 10 nucleotides upstream of an IP dsRNA 5′ end to 10 nucleotides upstream of the corresponding 3′ end on at least one strand (see Data set S2B in the supplemental material). We excluded the class of overlapping 3′ UTRs identified by Lybecker et al. from our analysis as they are not covered by our dRNA-seq, which sequences from the 5′ end of transcripts. This comparison yielded matching asTSS for 63% of the IP dsRNAs (193/308).

Candidate asRNAs are detected as distinct bands by Northern analysis. As independent verification of the predicted asRNAs, we selected a panel of 14 candidate asRNAs for Northern analysis (Fig. 5; see also Fig. S9 and Table S8 in the supplemental material). While we primarily selected candidates from the two highest-expression bins (see Data set S3 in the supplemental material), we also randomly selected a few candidates, which showed differences in expression among growth conditions or were not detected by others, from the third expression bin. We employed riboprobes covering the region of the dRNA-seq signal and importantly also probed total RNA from control strains where the region of mapped signal was deleted from the *E. coli* chromosome. In addition, we included total RNA isolated from strains defective for ribonucleases reported to be involved in asRNA processing and degradation; an *rnc* mutant lacking RNase III, an endonuclease that cleaves double-stranded RNAs, and an *rne-131* mutant with defective RNase E, an essential endonuclease that associates with the RNA degradosome and cleaves single-stranded RNA. The C terminus of RNase E is deleted in the *rne-131* mutant, such that the enzyme can no longer associate with the degradosome, thus giving rise to reduced RNA turnover (49, 50).

We detected clear specific bands for RNA isolated from wild-type cells for six of the candidate asRNAs (as-*gsiB*, as-*argR*, as-*ymfL*, as-*eutB*, as-*speA*, and as-*yliF*) (Fig. 5; see also Fig. S9 in the supplemental material). Specific bands for five other candidates (as-*qorA*, as-*holE*, as-*serU*, as-*thrW*, and as-*ytfJ*) were most evident in one or both of the RNase mutant strains, while three candidates (as-*yeaJ*, as-*gmr*, and as-*yggN*) were only detected as smears. For 10 of the probes, we detected nonspecific bands present in all lanes serving as a loading control and emphasizing the importance of including samples from control deletion strains.

asRNAs show differential sensitivity to degradation by RNase E and RNase III. We were surprised to find that the RNase mutants had varied impacts on the levels of our asRNA candidates. First, counter to expectations, the levels of some asRNAs, such as as-*ymfL* and as-*speA* were decreased in both RNase mutant

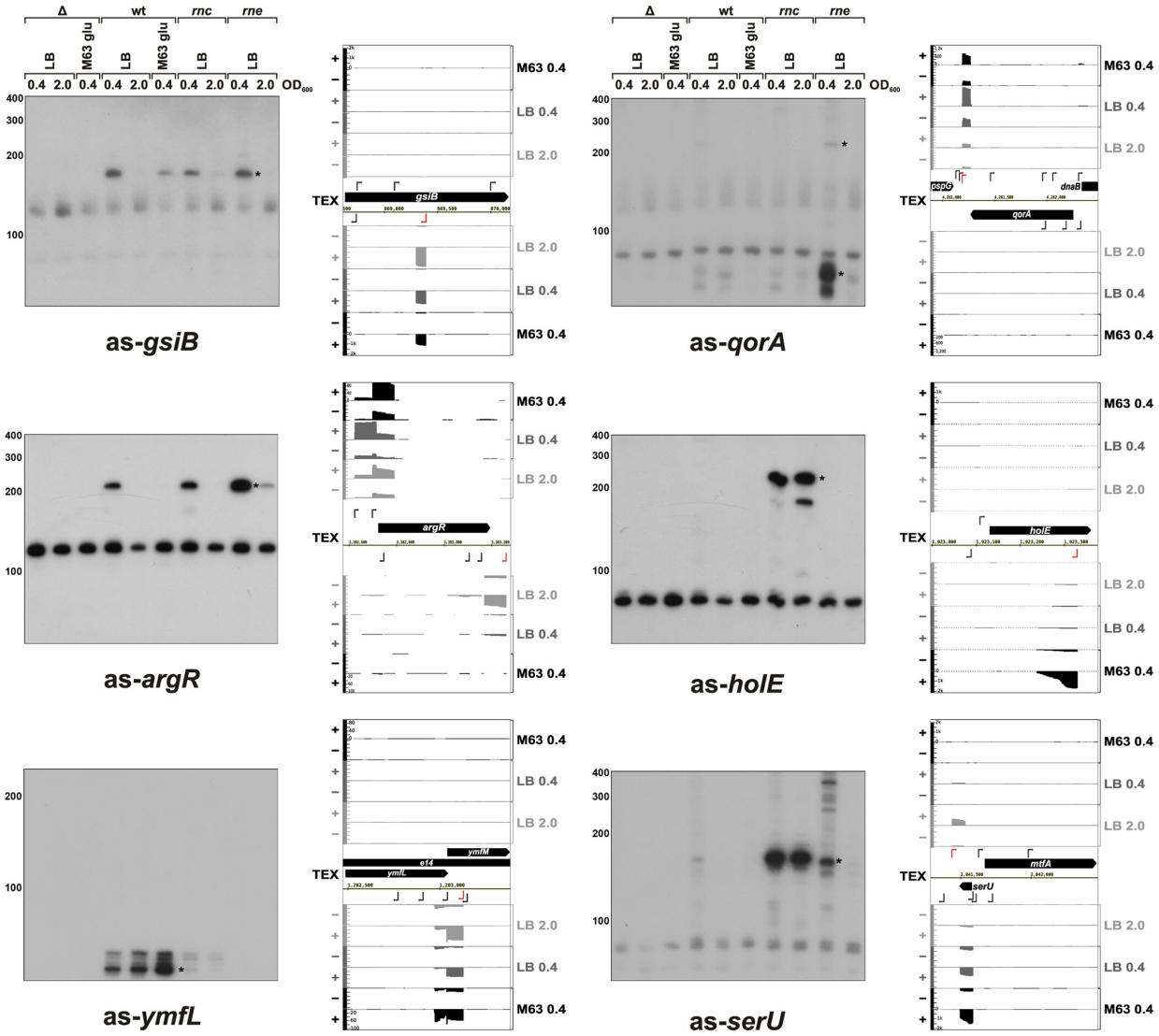


FIG 5 Northern blot detection and cDNA coverage plots of selected candidate asRNAs from the top three expression bins. In all cases, wild-type *E. coli* strain MG1655, the corresponding deletion strain for the particular asRNA as well as an *rnc* deletion strain, and an RNase E (*rne-131*) mutant strain were grown in LB or M63 supplemented with glucose until they reached the indicated OD_{600}. Samples were processed for Northern analysis and probed with a riboprobe specific for the asRNAs. The bands corresponding to the asRNAs are indicated with black stars. Schematics of cDNA coverage plots and genomic locations encoding the respective candidate asRNAs are shown on the right, with the position and direction of the asTSS indicated by red arrows. *y* axes indicating relative cDNA coverage have the same scale for the forward and reverse strands.

strains (Fig. 5; see also Fig. S9 in the supplemental material). Possibly these asRNAs are destabilized by interactions with RNAs that are normally degraded by RNase III and RNase E, or alternatively, processing is required for stabilization of these transcripts (51). The levels of three asRNAs (as-*argR*, as-*qorA*, and as-*eutB*) were elevated in the *rne* mutant, while the levels of four others (as-*holE*, as-*serU*, as-*ytfJ*, and as-*thrW*) were greatly elevated in the *rnc* mutant relative to the wild-type strain. Northern analysis carried out with RNA isolated from *rnc* mutants lacking the four chromosomal regions confirmed that the signal was specific (see Fig. S10 in the supplemental material; data not shown for as-*thrW*). Overall, these observations show that our detected asRNAs are substrates for different RNases and that regulation of asRNA levels by RNases may be more complex than previously thought.

DISCUSSION

In this study, we applied dRNA-seq and automated TSS prediction to the *E. coli* K-12 strain MG1655 grown under three different conditions to reveal >14,000 candidate TSS, of which >5,500 correspond to potential iTSS and >5,400 correspond to potential asRNAs. In contrast to previous *E. coli* transcriptome studies, dRNA-seq allowed us to globally map TSS since the approach specifically captures primary 5′ ends and thus allows discrimination between processed and primary transcripts. Our global TSS

map and coverage plots are integrated into RegulonDB and are easily accessible in an online browser at http://cbmp.nichd.nih.gov/segr/ecoli/, which allows researchers to readily identify candidate TSS and examine relative expression for their genes of interest. Our data represent a useful resource for the further characterization of promoters and novel RNAs in *E. coli*.

Automated TSS prediction has advantages and disadvantages. While the dRNA-seq analysis combined with automated TSS prediction used here provides a wealth of information, some reflection on the advantages and disadvantages is warranted. An automated approach for TSS annotation avoids potential bias introduced by manual annotation given that it follows defined rules and parameters. Automated annotation also facilitates rapid repetition of the analysis with different parameters or with additional data sets, a refinement that is impractical for manual annotation, especially for larger genomes or multiple strains or multiple conditions. However, choosing the right parameters for automated annotation, with the appropriate balance between sensitivity and specificity, can be difficult. Increasing the stringency for detection, for example, by filtering for those TSS whose step height is greater than 10 (see Data set S1 in the supplemental material) would reduce the number of TSS to ~4,400 (data not shown). Additionally, we analyzed the data for two of our LB 2.0 samples using another automated annotation program, TSSAR, with default parameters (52). This program predicted almost twice as many TSS as the TSSpredator program (data not shown) but is unable to integrate information from replicate samples. Therefore, for the TSS map presented in this study, we chose to use the parameters established on the basis of manual annotation of *Helicobacter pylori* dRNA-seq data and used for TSS annotation in *Campylobacter jejuni* (3, 30), which predicted TSS that were most consistent with manual annotation of selected regions of our *E. coli* data.

The automated TSSpredator program employed here led to the prediction of many more TSS candidates in *E. coli* than for other manually annotated data sets. To understand this difference between manual and automated annotations, we carried out automated TSS prediction using our *E. coli* parameters with the *Salmonella* dRNA-seq data sets from Kröger et al. (6). For the *Salmonella* dRNA-seq sets, we predicted ~22,000 potential candidate TSS, of which ~9,700 were found under all conditions (data not shown). These numbers are 4-fold higher than the TSS predicted by manual annotation. During manual TSS annotation, TSS corresponding to poorly expressed RNAs may not be annotated, resulting in underdetection of potential promoters transcribed at low levels. On the other hand, a higher discovery rate associated with automated TSS prediction may result in the false annotation of some promoters. It is also likely that our global map is still not saturated and that we have missed TSS that are not expressed under the limited growth conditions examined, as has been found for studies of *Salmonella* grown under a wide range of conditions (6).

Comparison of deep-sequencing data sets reveals sources of variation. When we compared our replicate deep-sequencing data sets, we found variation between different library preparations and sequencing platforms. The comparison of biological and technical replicates revealed that library preparation itself can lead to larger variation than found among biological replicates for which cDNA libraries were generated in parallel. 't Hoen et al. similarly found that library preparation is a major source of variation for human samples (53).

Our comparisons of asRNAs predicted by different published RNA-seq data sets further highlighted discrepancies and led us to consider additional sources of variation. Differences in RNA isolation protocols might limit the ability to capture unstable transcripts or RNAs of certain sizes. For example, small RNA fractions are often lost in column-based purification methods, and rRNA depletion kits can lead to unintended removal of non-rRNA transcripts. The use of terminator exonuclease (TEX) treatment to enrich for primary transcripts may miss the TSS of RNAs that are monophosphorylated due to the pyrophosphate removal by the enzyme RppH (54). However, we identified TSS for the majority of validated RppH targets, suggesting this is not a significant limitation in our data set (data not shown). Other inherent properties of the RNA molecules also can be a source for bias as it has been reported that RNAs with high GC content are less readily amplified and that linker ligation is more efficient when certain nucleotides are at the 3′ and 5′ ends (55).

Additionally, differences in data analysis, including differences in read quality filtering, mapping protocols (using all or only uniquely mapped reads), and especially different methods and thresholds for assembling and annotating transcripts, can lead to significantly different results. Despite a rapid increase in data generation, the availability of standardized RNA-seq analysis pipelines is still limited (56), particularly for bacterial transcriptomes. Nevertheless, RNA-seq has been an invaluable resource and has revolutionized bacterial, archaeal, and eukaryotic transcriptome analyses. Hopefully, as the field of deep sequencing continues to mature, standards for sample preparation, depth of sequencing, number of replicates sequenced, and data analysis as well as simple platforms for shared data visualization can be developed that will facilitate the comparisons of data generated by different groups.

Independent documentation of asRNAs is advised before functional analysis. Our dRNA-seq approach revealed more than 5,400 asTSS. We do not know how many of these predicted asTSS correspond to spurious transcripts rather than functional RNAs, although some show differential expression under the growth conditions examined (see Data set S3 in the supplemental material). The above-mentioned RNA-seq study of *Salmonella*, which analyzed RNAs from 22 different growth conditions, reported <500 asRNAs (5, 6). These authors found that ~1.75% of their reads mapped antisense to annotations (5) which is similar to what we observed (2 to 4%) (see Table S5 in the supplemental material) and to what has been reported for another *E. coli* RNA-seq study (~2%) (16). Thus, the high number of asTSS we detect probably is not due to large differences in general transcriptome coverage but rather is due to differences in data analyses and annotation. Moreover, we specifically enriched for the 5′ ends of transcripts, which might be more stable than internal degradation fragments, and did not include fragmentation steps that could result in the lower numbers of sequenced 5′ ends of transcripts.

As our comparison among different *E. coli* studies showed, there is extensive variation in asRNA annotation. Nevertheless, we found that several asRNAs were detected in multiple RNA-seq studies (Fig. 4C; see also Fig. S8 and Data set S3 in the supplemental material). Given the laborious process of functional investigation, however, we propose that further validation of asRNAs with appropriate controls is critical for defining candidates for further study. We independently validated expression of 14 candidate as-

RNAs by Northern analysis. For several of the asRNA candidates tested, nonspecific bands were detected in all lanes, emphasizing the importance of including samples from the control deletion strains. Expression was tested by Northern or quantitative PCR (qPCR) analysis for a subset of previously predicted asRNAs (13, 16, 48), although none of these studies included control deletion strains.

Overall, with the exponential increase in deep-sequencing studies and rapidly improving sequencing performance and coverage, more and more asRNA candidates will be reported in all organisms. To answer the questions of how many asRNAs identified in these analyses function as base-pairing RNA regulators, are used on a global scale for driving RNA processing, or are abortive transcripts resulting from degenerate promoters or RNA polymerase collisions, will require further experimental validation and characterization. Automated prediction of candidate asRNAs as reported here, combined with detection by multiple approaches, by multiple studies, or under specific growth conditions, will help identify those candidates most promising for future examination of phenotypes associated with the lack of the asRNA as well as mechanisms of asRNA action.

ACKNOWLEDGMENTS

We thank R. Reinhardt (Max Planck Genome Center, Cologne, Germany) for help with deep sequencing and *vertis* Biotechnologie AG for cDNA library preparation, E. P. Greenberg for work conducted by M.K.T. in his lab, and T. Conway for sharing data prior to publication. We also thank J. Hinton for discussions and J. Chen, S. Gottesman, and participants in the EMBO Lecture Course on the Biology of Bacterial Noncoding RNAs for comments on the manuscript.

Work in the Sharma laboratory is supported by the ZINF Young Investigator program of the Research Center for Infectious Diseases (ZINF) at the University of Würzburg, the Bavarian BioSysNet Program, and the Deutsche Forschungsgemeinshaft Project Sh580/1-1. Work in the Storz laboratory is supported by the Intramural Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development. For research stays in the Storz laboratory, S.K.E. was supported by a PROMOS travel scholarship (DAAD) of the University of Würzburg and C.M.S. was supported by a Boehringer Ingelheim Fonds travel allowance.

REFERENCES

1. **Croucher NJ, Thomson NR.** 2010. Studying bacterial transcriptomes using RNA-seq. Curr Opin Microbiol **13:**619–624. http://dx.doi.org/10.1016/j.mib.2010.09.009.
2. **van Vliet AH.** 2010. Next generation sequencing of microbial transcriptomes: challenges and opportunities. FEMS Microbiol Lett **302:**1–7. http://dx.doi.org/10.1111/j.1574-6968.2009.01767.x.
3. **Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, Stadler PF, Vogel J.** 2010. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature **464:**250–255. http://dx.doi.org/10.1038/nature08756.
4. **Mitschke J, Georg J, Scholz I, Sharma CM, Dienst D, Bantscheff J, Voss B, Steglich C, Wilde A, Vogel J, Hess WR.** 2011. An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. Proc Natl Acad Sci U S A **108:**2124–2129. http://dx.doi.org/10.1073/pnas.1015154108.
5. **Kröger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hebrard M, Händler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J, Hinton JC.** 2012. The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. Proc Natl Acad Sci U S A **109:**E1277–1286. http://dx.doi.org/10.1073/pnas.1201061109.
6. **Kröger C, Colgan A, Srikumar S, Händler K, Sivasankaran SK, Hammarlöf DL, Canals R, Grissom JE, Conway T, Hokamp K, Hinton JC.** 2013. An infection-relevant transcriptomic compendium for *Salmonella enterica* serovar Typhimurium. Cell Host Microbe **14:**683–695. http://dx.doi.org/10.1016/j.chom.2013.11.010.
7. **Wiegand S, Dietrich S, Hertel R, Bongaerts J, Evers S, Volland S, Daniel R, Liesegang H.** 2013. RNA-Seq of *Bacillus licheniformis:* active regulatory RNA features expressed within a productive fermentation. BMC Genomics **14:**667. http://dx.doi.org/10.1186/1471-2164-14-667.
8. **Wurtzel O, Sesto N, Mellin JR, Karunker I, Edelheit S, Becavin C, Archambaud C, Cossart P, Sorek R.** 2012. Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species. Mol Syst Biol **8:**583. http://dx.doi.org/10.1038/msb.2012.11.
9. **Nicolas P, Mader U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S, Becher D, Bisicchia P, Botella E, Delumeau O, Doherty G, Denham EL, Fogg MJ, Fromion V, Goelzer A, Hansen A, Hartig E, Harwood CR, Homuth G, Jarmer H, Jules M, Klipp E, Le Chat L, Lecointe F, Lewis P, Liebermeister W, March A, Mars RA, Nannapaneni P, Noone D, Pohl S, Rinn B, Rugheimer F, Sappa PK, Samson F, Schaffer M, Schwikowski B, Steil L, Stulke J, Wiegert T, Devine KM, Wilkinson AJ, van Dijl JM, Hecker M, Volker U, Bessieres P, Noirot P.** 2012. Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. Science **335:**1103–1106. http://dx.doi.org/10.1126/science.1206848.
10. **Kim D, Hong JS, Qiu Y, Nagarajan H, Seo J H, Cho BK, Tsai SF, Palsson BO.** 2012. Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling. PLoS Genet **8:**e1002867. http://dx.doi.org/10.1371/journal.pgen.1002867.
11. **Cho BK, Kim D, Knight EM, Zengler K, Palsson BO.** 2014. Genome-scale reconstruction of the sigma factor network in *Escherichia coli*: topology and functional states. BMC Biol **12:**4. http://dx.doi.org/10.1186/1741-7007-12-4.
12. **Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BØ.** 2009. The transcription unit architecture of the *Escherichia coli* genome. Nat Biotech **27:**1043–1049. http://dx.doi.org/10.1038/nbt.1582.
13. **Shinhara A, Matsui M, Hiraoka K, Nomura W, Hirano R, Nakahigashi K, Tomita M, Mori H, Kanai A.** 2011. Deep sequencing reveals as-yet-undiscovered small RNAs in *Escherichia coli*. BMC Genomics **12:**428. http://dx.doi.org/10.1186/1471-2164-12-428.
14. **Dornenburg JE, Devita AM, Palumbo MJ, Wade JT.** 2010. Widespread antisense transcription in *Escherichia coli*. mBio **1**(1):e00024-10. http://dx.doi.org/10.1128/mBio.00024-10.
15. **Li S, Dong X, Su Z.** 2013. Directional RNA-seq reveals highly complex condition-dependent transcriptomes in *E. coli* K12 through accurate full-length transcripts assembling. BMC Genomics **14:**520. http://dx.doi.org/10.1186/1471-2164-14-520.
16. **Raghavan R, Sloan DB, Ochman H.** 2012. Antisense transcription is pervasive but rarely conserved in enteric bacteria. mBio **3**(4):e00156-12. http://dx.doi.org/10.1128/mBio.00156-12.
17. **Peters JM, Mooney RA, Grass JA, Jessen ED, Tran F, Landick R.** 2012. Rho and NusG suppress pervasive antisense transcription in *Escherichia coli*. Genes Dev **26:**2621–2633. http://dx.doi.org/10.1101/gad.196741.112.
18. **Peters JM, Mooney RA, Kuan PF, Rowland JL, Keles S, Landick R.** 2009. Rho directs widespread termination of intragenic and stable RNA transcription. Proc Natl Acad Sci U S A **106:**15406–15411. http://dx.doi.org/10.1073/pnas.0903846106.
19. **Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J, San Miguel P, Shimada T, Ishihama A, Mori H, Wanner BL.** 2014. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. mBio **5**(4):e01442-14. http://dx.doi.org/10.1128/mBio.01442-14.
20. **Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juarez K, Contreras-Moreira B, Huerta AM, Collado-Vides J, Morett E.** 2009. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. PLoS One **4:**e7526. http://dx.doi.org/10.1371/journal.pone.0007526.
21. **Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, Weiss V, Solano-Lira H, Martínez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernández S, Alquicira-Hernández K, López-Fuentes A, Porrón-Sotelo L, Huerta AM, Bonavides-Martínez C, Balderas-Martínez YI, Pannier L, Olvera M, Labastida A, Jiménez-Jacinto V, Vega-Alvarado L, Del Moral-Chávez V, Hernández-Alvarez A, Morett E, Collado-Vides J.** 2013.

RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res **41:** D203–D213. http://dx.doi.org/10.1093/nar/gks1201.

22. **Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J.** 2012. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics **13:**734. http://dx.doi.org/10.1186/1471-2164-13-734.
23. **Thomason MK, Storz G.** 2010. Bacterial antisense RNAs: how many are there, and what are they doing? Annu Rev Genet **44:**167–188. http://dx.doi.org/10.1146/annurev-genet-102209-163523.
24. **Georg J, Hess WR.** 2011. cis-antisense RNA, another level of gene regulation in bacteria. Microbiol Mol Biol Rev **75:**286–300. http://dx.doi.org/10.1128/MMBR.00032-10.
25. **Lasa I, Toledo-Arana A, Dobin A, Villanueva M, de los Mozos IR,Vergara-Irigaray M, Segura V, Fagegaltier D, Penades JR, Valle J, Solano C, Gingeras TR.** 2011. Genome-wide antisense transcription drives mRNA processing in bacteria. Proc Natl Acad Sci U S A **108:**20172–20177. http://dx.doi.org/10.1073/pnas.1113521108.
26. **Sesto N, Wurtzel O, Archambaud C, Sorek R, Cossart P.** 2013. The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. Nat Rev Microbiol **11:**75–82. http://dx.doi.org/10.1038/nrmicro2934.
27. **Lin Y, Alvarez J R, Guan S, Mamanova L, McDowall KJ.** 2013. A combination of improved differential and global RNA-seq reveals pervasive transcription initiation and events in all stages of the life-cycle of functional RNAs in *Propionibacterium acnes*, a major contributor to widespread human disease. BMC Genomics **14:**620. http://dx.doi.org/10.1186/1471-2164-14-620.
28. **Passalacqua KD, Varadarajan A, Weist C, Ondov BD, Byrd B, Read TD, Bergman NH.** 2012. Strand-specific RNA-seq reveals ordered patterns of sense and antisense transcription in *Bacillus anthracis*. PLoS One **7:**e43350. http://dx.doi.org/10.1371/journal.pone.0043350.
29. **Sharma CM, Vogel J.** 2014. Differential RNA-seq: the approach behind and the biological insight gained. Curr Opin Microbiol **19:**97–105. http://dx.doi.org/10.1016/j.mib.2014.06.010.
30. **Dugar G, Herbig A, Forstner KU, Heidrich N, Reinhardt R, Nieselt K, Sharma CM.** 2013. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. PLoS Genet **9:**e1003495. http://dx.doi.org/10.1371/journal.pgen.1003495.
31. **Yu D, Ellis HM, Lee EC, Jenkins NA, Copeland NG, Court DL.** 2000. An efficient recombination system for chromosome engineering in *Escherichia coli*. Proc Natl Acad Sci U S A **97:**5978–5983. http://dx.doi.org/10.1073/pnas.100127597.
32. **Blomberg P, Wagner EGH, Nordström K.** 1990. Control of replication of plasmid R1: the duplex between the antisense RNA, CopA, and its target, CopT, is processed specifically in vivo and in vitro by RNase III. EMBO J **9:**2331–2340.
33. **Förstner KU, Vogel J, Sharma CM.** 13 August 2014. READemption—a tool for the computational analysis of deep-sequencing-based transcriptome data. Bioinformatics http://dx.doi.org/10.1093/bioinformatics/btu533.
34. **Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermuller J.** 2009. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol **5:**e1000502. http://dx.doi.org/10.1371/journal.pcbi.1000502.
35. **Mao X, Ma Q, Zhou C, Chen X, Zhang H, Yang J, Mao F, Lai W, Xu Y.** 2014. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. Nucleic Acids Res **42:**D654–D659. http://dx.doi.org/10.1093/nar/gkt1048.
36. **Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B.** 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods **5:**621–628. http://dx.doi.org/10.1038/nmeth.1226.
37. **Thomason MK, Fontaine F, De Lay N, Storz G.** 2012. A small RNA that regulates motility and biofilm formation in response to changes in nutrient availability in *Escherichia coli*. Mol Microbiol **84:**17–35. http://dx.doi.org/10.1111/j.1365-2958.2012.07965.x.
38. **Battesti A, Majdalani N, Gottesman S.** 2011. The RpoS-mediated general stress response in *Escherichia coli*. Annu Rev Microbiol **65:**189–213. http://dx.doi.org/10.1146/annurev-micro-090110-102946.
39. **Miller KW, Wu HC.** 1987. Cotranscription of the *Escherichia coli* isoleucyl-tRNA synthetase (*ileS*) and prolipoprotein signal peptidase (*lsp*) genes. Fine-structure mapping of the *lsp* internal promoter. J Biol Chem **262:** 389–393.
40. **Shao W, Price MN, Deutschbauer AM, Romine MF, Arkin AP.** 2014. Conservation of transcription start sites within genes across a bacterial genus. mBio **5**(4):e01398-14. http://dx.doi.org/10.1128/mBio.01398-14.
41. **Guo MS, Updegrove TB, Gogol EB, Shabalina SA, Gross CA, Storz G.** 2014. MicL, a new σ^E-dependent sRNA, combats envelope stress by repressing synthesis of Lpp, the major outer membrane lipoprotein. Genes Dev **28:**1620–1634. http://dx.doi.org/10.1101/gad.243485.114.
42. **Bendtsen KM, Erdossy J, Csiszovszki Z, Svenningsen SL, Sneppen K, Krishna S, Semsey S.** 2011. Direct and indirect effects in the regulation of overlapping promoters. Nucleic Acids Res **39:**6879–6885. http://dx.doi.org/10.1093/nar/gkr390.
43. **Fozo EM, Hemm MR, Storz G.** 2008. Small toxic proteins and the antisense RNAs that repress them. Microbiol Mol Biol. Rev **72:**579–589. http://dx.doi.org/10.1128/MMBR.00025-08.
44. **Andre G, Even S, Putzer H, Burguiere P, Croux C, Danchin A, Martin-Verstraete I, Soutourina O.** 2008. S-box and T-box riboswitches and antisense RNA control a sulfur metabolic operon of *Clostridium acetobutylicum*. Nucleic Acids Res **36:**5955–5969. http://dx.doi.org/10.1093/nar/gkn601.
45. **Mellin JR, Tiensuu T, Becavin C, Gouin E, Johansson J, Cossart P.** 2013. A riboswitch-regulated antisense RNA in *Listeria monocytogenes*. Proc Natl Acad Sci U S A **110:**13132–13137. http://dx.doi.org/10.1073/pnas.1304795110.
46. **Bailey TL, Elkan C.** 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol **2:**28–36.
47. **Feklistov A, Sharon BD, Darst SA, Gross CA.** 2014. Bacterial sigma factors: a historical, structural, and genomic perspective. Annu Rev Microbiol **68:**357–376. http://dx.doi.org/10.1146/annurev-micro-092412-155737.
48. **Lybecker M, Zimmermann B, Bilusic I, Tukhtubaeva N, Schroeder R.** 2014. The double-stranded transcriptome of *Escherichia coli*. Proc Natl Acad Sci U S A **111:**3134–3139. http://dx.doi.org/10.1073/pnas.1315974111.
49. **Massé E, Escorcia FE, Gottesman S.** 2003. Coupled degradation of a small regulatory RNA and its mRNA targets in *Escherichia coli*. Genes Dev **17:**2374–2383. http://dx.doi.org/10.1101/gad.1127103.
50. **Kido M, Yamanaka K, Mitani T, Niki H, Ogura T, Hiraga S.** 1996. RNase E polypeptides lacking a carboxyl-terminal half suppress a *mukB* mutation in *Escherichia coli*. J Bacteriol **178:**3917–3925.
51. **Opdyke JA, Fozo EM, Hemm MR, Storz G.** 2011. RNase III participates in GadY-dependent cleavage of the *gadX-gadW* mRNA. J Mol Biol **406:** 29–43. http://dx.doi.org/10.1016/j.jmb.2010.12.009.
52. **Amman F, Wolfinger MT, Lorenz R, Hofacker IL, Stadler PF, Findeiß S.** 2014. TSSAR: TSS annotation regime for dRNA-seq data. BMC Bioinformatics **15:**89. http://dx.doi.org/10.1186/1471-2105-15-89.
53. **'t Hoen PA, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, Laros JF, Buermans HP, Karlberg O, Brännvall M, GEUVADIS Consortium, den Dunnen JT, van Ommen GJ, Gut IG, Guigó R, Estivill X, Syvänen AC, Dermitzakis ET, Lappalainen T.** 2013. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol **31:**1015–1022. http://dx.doi.org/10.1038/nbt.2702.
54. **Deana A, Celesnik H, Belasco JG.** 2008. The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal. Nature **451:**355–358. http://dx.doi.org/10.1038/nature06475.
55. **Raabe CA, Tang TH, Brosius J, Rozhdestvensky TS.** 2014. Biases in small RNA deep sequencing data. Nucleic Acids Res **42:**1414–1426. http://dx.doi.org/10.1093/nar/gkt1021.
56. **McClure R, Balasubramanian D, Sun Y, Bobrovskyy M, Sumby P, Genco CA, Vanderpool CK, Tjaden B.** 2013. Computational analysis of bacterial RNA-Seq data. Nucleic Acids Res **41:**e140. http://dx.doi.org/10.1093/nar/gkt444.

SUPPLEMENTAL INFORMATION

Global transcriptional start site mapping using dRNA-seq reveals novel antisense RNAs in *Escherichia coli*

Maureen K. Thomason,[a,‡,*], Thorsten Bischler,[b,*] Sara K. Eisenbart,[a,b] Konrad U. Förstner,[b] Aixia Zhang,[a] Alexander Herbig,[c] Kay Nieselt,[c] Cynthia M. Sharma,[a,b,#] Gisela Storz[a,#]

Cell Biology and Metabolism Program, Eunice Kennedy Shriver National Institutes of Health, Bethesda, Maryland, USA[a]; Research Centre for Infectious Diseases (ZINF), University of Würzburg, Würzburg, Germany[b]; Integrative Transcriptomics, ZBIT (Center for Bioinformatics Tübingen), University of Tübingen, Tübingen, Germany[c]

[‡]Present address: Department of Microbiology, University of Washington, Seattle, Washington

[*]Joint First Authors.

[#]Address correspondence to Gisela Storz, storzg@mail.nih.gov or Cynthia M. Sharma, cynthia.sharma@uni-wuerzburg.de.

SUPPLEMENTAL MATERIALS AND METHODS

Deep sequencing sample preparation.

RNA extraction. Frozen cell pellets were thawed on ice and resuspended in 880 µl of lysis buffer (0.5 mg/ml lysozyme dissolved in TE pH 8.0, 1% SDS), mixed by inversion and incubated at 65˚C for 2 min or until the samples cleared. The samples were cooled and 88 µl of 1M sodium acetate, pH 5.2 was added along with 1 ml of acid phenol:chloroform (Ambion). Samples were incubated at 65˚C for 6 min with mixing and spun 10 min at 13,000 rpm, 4˚C. The aqueous layer was extracted a second time with chloroform using Phase Lock Gel 2.0 tubes (5Prime) after which the aqueous layer was ethanol precipitated, washed and resuspended in 100 µl of DEPC-H_2O. RNA concentration was measured by reading the absorbance at OD_{260} and the integrity was checked by running ~2 µg aliquots of each sample on a denaturing 1% agarose 1X TBE gel followed by ethidium bromide staining.

DNase I treatment. Total RNA (40 µg) was denatured at 65˚C for 5 min. The RNA was then combined with 1X DNase I buffer + $MgCl_2$ (Fermentas), 20 U of RNase Inhibitor (Invitrogen), and 10 U of DNase I (Fermentas) in a final volume of 100 µl. The mixture was incubated for 45 min at 37˚C and then extracted with phenol:chloroform:isoamylalcohol (Invitrogen) in 2 ml Phase Lock Gel Heavy tubes (5Prime). Samples were precipitated, washed, and resuspended in 40 µl of DEPC-H_2O. RNA concentration and integrity of ~100 ng aliquots were checked as above, and the absence of genomic DNA contamination was confirmed by PCR using primer MK0095 and MK0096.

Terminator exonuclease (TEX) and tobacco acid pyrophosphatase (TAP) treatment. TEX treatment was performed as described previously (1). Briefly, 7 µg of DNase I-treated RNA was denatured for 2 min at 90˚C, cooled on ice for 5 min and combined with 10 U RNase Inhibitor (Invitrogen), 1X Terminator Exonuclease Buffer A (Epicentre), and 7 U of Terminator Exonuclease (Epicentre) in a final reaction volume of 50 µl. Control reactions lacking terminator exonuclease were run in parallel for each sample. Reactions were incubated at 30˚C for 1 h and stopped by the addition of 0.5 µl of 0.5 M EDTA, 50 µl DEPC-H_2O and extraction with phenol:chloroform:isoamylalcohol with Phase Lock Gel 2.0 Tubes. The supernatant was precipitated, washed and resuspended in 11 µl of DEPC-H_2O. RNA concentration was determined by reading the absorbance at OD_{260}. TAP treatment was performed by incubating TEX-treated and untreated control samples with tobacco acid pyrophosphatase (TAP) for 1 h at

37˚C with 1X TAP Buffer (Invitrogen), 10 U RNase Inhibitor and 5 U Tobacco Acid Pyrophosphatase (Invitrogen) in a final reaction volume of 20 µl. The samples were extracted with phenol:chloroform:isoamylalcohol (Invitrogen), precipitated, washed, and resuspended in 20 µl of DEPC-H_2O. RNA concentration was determined by reading the absorbance at OD_{260}, and RNA integrity was checked on a denaturing 4% acrylamide-7M urea gel in 1X TBE and visualized with Stains-all nucleic acid stain (Sigma).

cDNA library construction. cDNA libraries for Illumina sequencing were constructed by vertis Biotechnology AG, Germany (http://www.vertis-biotech.com/) in a strand specific manner as described previously for eukaryotic microRNA (2) but omitting the RNA size-fractionation step prior to cDNA synthesis. In brief, equal amounts of RNA samples were poly(A)-tailed using poly(A) polymerase. Then, the 5'-PPP structures were removed using tobacco acid pyrophosphatase (TAP). Afterwards, an RNA adapter was ligated to the 5´-phosphate of the TAP-treated, poly(A)-tailed RNA. In the case of the GAIIx-libraries, the 5'linker contained the barcode sequence at its 3'end. For HiSeq 2000 libraries, the barcode was introduced in a later step during PCR-amplification of the cDNA library. First-strand cDNA was synthesized by using an oligo(dT)-adapter primer and the M-MLV reverse transcriptase. In a PCR-based amplification step using a high fidelity DNA polymerase the cDNA concentration was increased to 20-30 ng/µl. For all libraries the Agencourt AMPure XP kit (Beckman Coulter Genomics) was used to purify the DNA, which was subsequently analyzed by capillary electrophoresis.

For the GAIIx-libraries, PCR products for sequencing were generated using the following primers designed for amplicon sequencing according to the instructions of Illumina/Solexa:
5'-end_primer
5'-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGACGATCNNNN-3'
3'-end_primer
5'-CAAGCAGAAGACGGCATACGATTTTTTTTTTTTTTTTTTTTTTT-3'
The samples were run on an Illumina GAIIx instrument with 120 cycles in single-read mode.

For the HiSeq2000 libraries, a library-specific barcode for multiplex sequencing was included as part of a 3'-sequencing adapter. The following adapter sequences flank the cDNA inserts:
TrueSeq_Sense_primer

5'-
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC
T-3'
TrueSeq_Antisense_NNNNNN_primer (NNNNNN = 6n barcode for multiplexing)
5'-CAAGCAGAAGACGGCATACGAGAT-NNNNNN-
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC(dT25)-3'
The samples were run on an Illumina HiSeq 2000 instrument with 100 cycles in single-read mode.

Analysis of deep sequencing data.

Read mapping and coverage plot construction. To assure high sequence quality, the Illumina reads in FASTQ format were trimmed with a cutoff phred score of 20 by the program fastq_quality_trimmer from FASTX toolkit version 0.0.13. After trimming, poly(A)-tail sequences were removed and a size filtering step was applied in which sequences shorter than 12 nt were eliminated. The collections of remaining reads were mapped to the *E. coli* MG1655 genome (NCBI Acc.-No: NC_000913.2; Jun 24, 2004) using our RNA-seq pipeline READemption (3) and *segemehl* (4) with an accuracy cutoff of 95%. Coverage plots representing the numbers of mapped reads per nucleotide were generated. Reads that mapped to multiple locations contributed a fraction to the coverage value. For example, reads mapping to three positions contributed only 1/3 to the coverage values. Each graph was normalized to the number of reads that could be mapped from the respective library. To restore the original data range, each graph was then multiplied by the minimum number of mapped reads calculated over all libraries.

Normalization of expression graphs. Prior to the comparative analysis, the expression graphs with the cDNA coverages that resulted from the read mapping were further normalized. A percentile normalization step was applied to normalize the +TEX graphs. To this end, the 90^{th} percentile of all data values was calculated for each +TEX graph. This value was then used to normalize the +TEX graph as well as the respective −TEX graph. Thus, the relative differences between each +TEX and −TEX graph were not changed in this normalization step. Again, all graphs were multiplied with the overall lowest value to restore the original data range. To account for different enrichment rates, a third normalization step was applied. During this step, prediction of TSS candidates was performed for each replicate of each strain. These candidates

were then used to determine the median enrichment factor for each +/−TEX library pair. Using these medians all −TEX libraries were then normalized against the library with the strongest enrichment. Besides annotation of transcriptional start sites, the resulting graphs were also used for visualization in the Integrated Genome Browser (5).

Correlation analysis. To assess similarity between different libraries, nucleotide and gene-wise Spearman and Pearson correlation coefficients were calculated based on concatenated values of forward and reverse strand position-wise coverage files and visualized in a correlation matrix for both the +TEX and –TEX libraries using the R package *corrplot*. Gene-wise correlation values utilized read overlap counts based on NCBI annotations (Acc.-No: NC_000913.2). Each read with a minimum overlap of 10 nt was counted with a value based on the number of locations where the read was mapped. If the read overlapped more than one annotation, the value was divided by the number of regions and counted separately for each region (e.g. 1/3 for a read mapped to 3 locations).

Transcriptional start site (TSS) annotation. Based on the normalized expression graphs we conducted automated TSS prediction in a similar manner as described in Dugar *et al.* 2013 (6) utilizing TSSpredator (http://it.inf.uni-tuebingen.de/TSSpredator). In brief, for each position (i) in the expression graph corresponding to the TEX treated libraries, the algorithm calculates an expression height, e(i), and compares that expression height to the preceding position by calculating e(i) – e(i-1), which is termed the flank height. Additionally, the algorithm calculates a factor of height change e(i)/e(i-1). To determine if a TSS is a primary TSS and not a processed transcript end an enrichment factor is calculated as $e_{+TEX}(i)/e_{-TEX}(i)$, where $e_{+TEX}(i)$ is the expression height for the terminator exonuclease treated sample and $e_{-TEX}(i)$ is the expression height for the untreated sample. For all positions where these parameters exceed the predefined thresholds a TSS is annotated.

We set the thresholds for the “minimum flank height” and the “minimum factor of height change” which are used to determine if a TSS is “detected” to 0.3 and 2.0, respectively. Here, the value for the “minimum flank height” is a factor to the minimum 90^{th} percentile over all libraries resulting in an absolute value of 1.62. If the TSS candidate reaches these thresholds in at least one replicate of one condition, the thresholds are decreased for the other replicates to 0.1 (0.54 absolute) and 1.5, respectively. Furthermore, we set the “matching replicates” parameter which determines the number of replicates in which a TSS must exceed these thresholds in order to be

marked as “detected” within a condition to 2 for M63 0.4 and LB 0.4 and to 3 for LB 2.0. If a TSS was detected in a certain condition, the lowered thresholds also apply to all remaining libraries of the other conditions. Furthermore, we consider a TSS candidate to be enriched in a condition if the respective enrichment factor for at least one replicate is not less than 2.0. A TSS candidate has to be enriched in at least one condition and is discarded otherwise. If a TSS candidate is not enriched in a condition but still reaches the other thresholds it is only indicated as “detected”. However, a TSS candidate can only be labeled as detected in a condition if its enrichment factor is above 0.66. Otherwise we consider it to be a processing site. In order to take into account slight variations between TSS positions the respective parameters for clustering between replicates and conditions were set to a value of 1. In doing so a consensus TSS position in a three nucleotide window is determined based on the maximum “flank height” among the respective libraries. The same parameters were recently used in our comparative dRNA-seq analysis of multiple *Campylobacter jejuni* strains (6).

Comparison to Database of prOkaryotic OpeRons (DOOR). A table containing all operon annotations (1,526 single gene operons and 851 operons consisting of multiple genes) was downloaded from the DOOR 2.0 website (7) (on Sep 18, 2013). Two genes (b0816 and b1470) not included in the NCBI annotations used for the TSS prediction were discarded. We divided our genes with predicted primary TSS (2,672) into two groups. The first consisted of genes (2,057) for which the TSS was classified only as primary (1,707) or primary and antisense (350). The second consisted of genes (615) with a primary and internal TSS (611) or primary, internal and antisense (4). Furthermore, 231 genes from both groups lacking DOOR locus tags and information regarding operon structure were excluded from the analysis. This resulted in a final set of 2,441 TSS that were considered for the DOOR overlap: group one contained 1,562 primary and 566 primary and secondary and group two contained 309 primary and antisense and 4 primary, internal and antisense. The overlap was calculated by comparing the primary TSS-associated genes of either group to the set of single-standing genes and first genes in an operon from DOOR.

Analysis of iTSS localization. The 5,574 iTSS were split into two groups, iTSS that were also annotated as pTSS or sTSS (968) and the remaining iTSS (4,606) where iTSS in both groups can also be annotated as asTSS. Each gene in which an iTSS was detected was split into 10 equal-sized sections ordered from 5´ to 3´ end (the first section covers the first 10% of the

gene, the second section the second 10%, etc.) and the number of iTSS localizing to each section was counted over all genes. Histogram plots were generated to visualize the distribution of iTSS in each group.

Comparison of expression under different growth conditions. Expression values for TSS (based on a 50 nt window downstream of the TSS position) were calculated as described above for the gene-wise correlation analysis but using only libraries from the M63 0.4 and LB 0.4 conditions. Differential expression between these two conditions was assessed for pTSS only, iTSS only and asTSS only based on all replicates using DESeq2 (8). TSS with an adjusted p-value ≤ 0.05 were defined as differentially expressed. A Fisher exact test based on all TSS in the three classes was performed to determine if there is over- or underrepresentation of any class.

Detection of promoter motifs. Sequences from −50 to +1 as well as from −50 to +5 relative to the TSS were extracted for pTSS only, iTSS only and asTSS only and analyzed using MEME (9). For the −50 to +1 sequences a motif of length 48 nt was predicted to allow a distance of 0 to 3 nt to the TSS position. For the −50 to +5 sequences a motif length of 56 nt was applied.

Identification of overlapping 5' UTRs. Based on all primary and secondary TSS (see Data Set S1) all possible pairs of overlapping 5' UTRs with the first UTR on the forward and second one on the reverse strand were computed with the restriction that the overlapping region must have a minimum length of 10 nucleotides. The data for this is depicted in Data Set S2A.

Comparison of asRNAs detected in our and previous studies. To compare the annotations of previously detected antisense RNAs with our predictions, asRNA annotations were retrieved from the Supplemental Materials and Methods sections from previously published studies (10-12) or were downloaded from RegulonDB (13, 14). For the Raghavan *et al.* data, 90 asTSS were identified antisense to distinct genes which, combined with the gene names, was used to infer strand information for the asTSS (11). For the Shinhara *et al.* data, 229 novel candidate sRNAs were assembled from the mapped sequencing reads (12). From these reads we extracted the TSS positions of 112 candidate sRNAs labeled as "cis-antisense". For the Dornenburg *et al.* data, we used all 1,005 putative TSS located antisense to genes and 385 putative TSS located antisense to predicted untranslated regions (10). The TSS data obtained from RegulonDB consists of one data set with 1,490 TSS from which we extracted 165 TSS described as antisense to a gene (13) (Data set version 2.0), and a second set which includes 5,197 single TSSs and TSS clusters of varying length (14) (Data set version 3.0). We selected 182 single TSS and TSS clusters defined as either

located antisense to a gene or to be "convergent", i.e. the TSS is located sense to one gene and antisense to another. In these cases, we used the specific nucleotide position for a single TSSs and the respective region for a TSS cluster. Additionally, we included set of 89 asRNA predicted by Conway *et al*. (15). We compared the asTSS from each data set, including our 6,379 predicted asTSS, to the asTSS of all other data sets in a pair-wise manner, requiring either a precise match of the annotated positions or allowing a variation of 1, 2, 3 or 10 nt.

Comparison of asTSS to IP-dsRNAs. IP-dsRNAs were extracted from Lybecker *et al.* (16). Afterwards, all dsRNAs assigned to the category "Convergent" were excluded since overlapping 3'UTRs might not be covered in our dRNA-seq analysis as our protocol sequences from the 5' ends of genes. For the comparison we used all TSS annotated as asTSS including the ones also assigned to other classes. An overlap is reported if an asTSS is annotated in a region between 10 nucleotides upstream of the IP-dsRNA 5' end and 10 nucleotides upstream of the dsRNA 3' end on at least one strand.

Northern analysis. The oligonucleotides used to create the riboprobes are listed in Table S2. To synthesize the riboprobe, 3.5 µg of gene-specific PCR product was added to a reaction mix containing 5 µl of α-^{32}P-UTP, 1X T7 RNA polymerase buffer, 2000 U/ml T7 RNA polymerase (NEB), 20 U RNasin, 4 mM DTT (Invitrogen), 0.16 mg/ml BSA (NEB), 0.4 mM GTP, CTP, and ATP, and 0.01 mM UTP (Invitrogen) in a total volume of 25 µl. After 1 h incubation at 37˚C, 2000 U/ml of T7 RNA polymerase was added, and the samples were incubated an additional 1 h at 37˚C. After a final incubation with 1 U/µl of DNase I (Fermentas) at 37˚C for 15 min, probes were purified using G50 columns (GE Healthcare). RNA samples were separated on denaturing 8% acrylamide-7M urea gels, transferred and UV crosslinked to Zeta Probe GT membranes (Bio-Rad) as before (17). The membranes were subsequently incubated for 2 h at 50˚C in 20 ml hybridization buffer (50% formamide, 1.5X SSPE, 1% SDS, 0.5% dry milk), after which the hybridization buffer was exchanged for 20 ml of fresh buffer. Riboprobes were denatured with 100 µl of 10 mg/ml yeast RNA at 95˚C for 4 min, added to the membrane and left to hybridize overnight at 50˚C. The membranes were subsequently rinsed once with 2X SSC + 0.1% SDS, washed once for 20 min at 50˚C with 2X SSC + 0.1% SDS, and twice for 20 min at 50˚C with 0.1X SSC + 0.1% SDS. The membranes were rinsed two more times with 0.1X SSC + 0.1% SDS, allowed to dry, and exposed to Hyperfilm (Amersham) at −80˚C.

SUPPLEMENTAL FIGURES

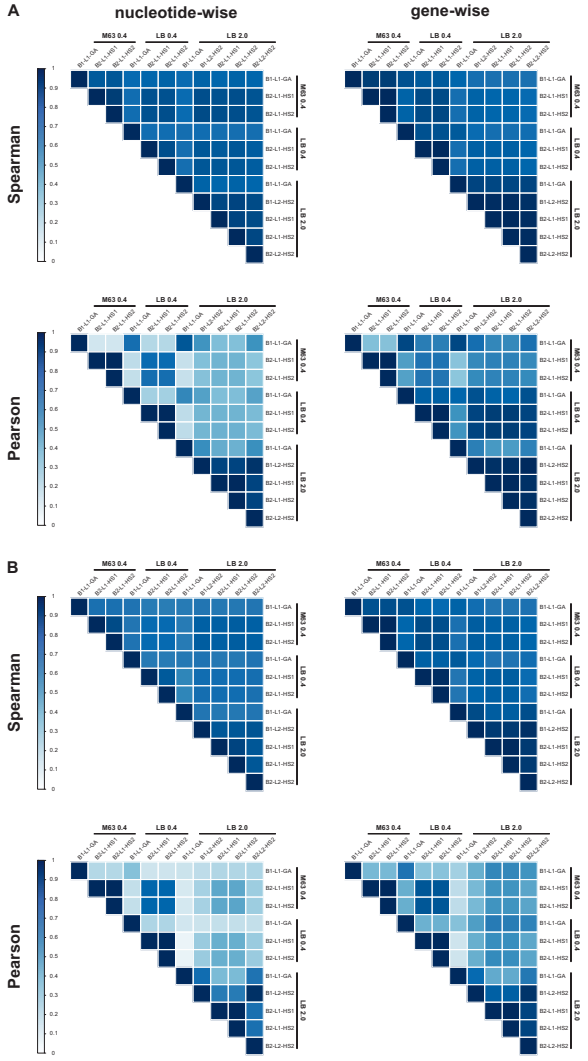


FIG. S1. Reproducibility of cDNA coverage among –TEX (A) and +TEX (B) cDNA libraries. Nucleotide-wise Spearman and Pearson correlation values for all possible combinations of library pairs from all growth conditions were based on expression values from both strands. Gene-wise Spearman and Pearson correlation values for all possible combinations of library pairs were based on expression values for genes as annotated by NCBI. A legend for the extent of correlation is given on the left

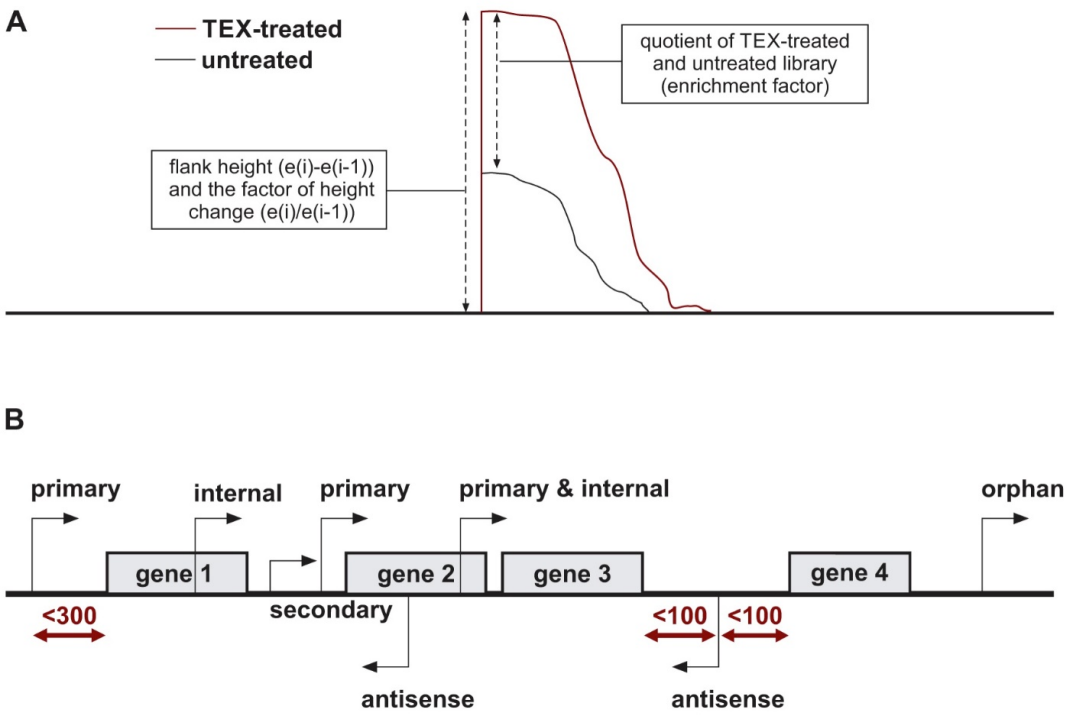


FIG. S2. TSS prediction parameters and classification. (A) Schematic representation of the criteria used for TSS prediction. (B) The different TSS classes (primary, secondary, internal, antisense, and orphan) are depicted according to their location relative to annotated genes. The height of the black arrows depicts differences in expression strength while the distance cutoffs for flanking genes are shown in red. The figures are adapted from (6).

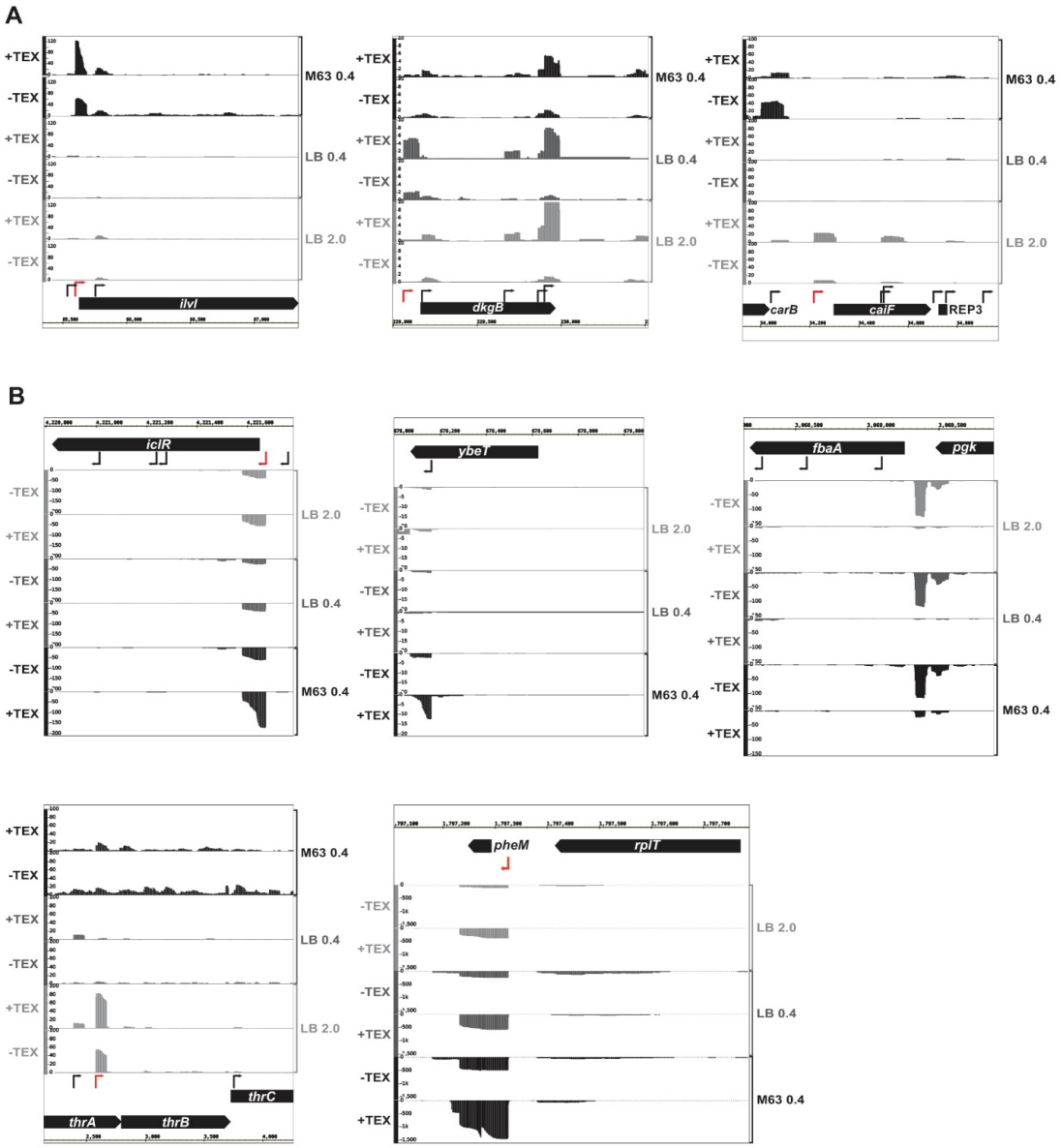


FIG. S3. cDNA coverage plots for examples of TSS detection. (A) Examples of TSS that were detected in only one of three conditions. Each panel shows a primary TSS (indicated by red arrow) that was detected only in one of the three examined growth conditions. (B) Examples of primary TSS (indicated by red arrows) detected in our study that agree or disagree with annotation from the DOOR database (7). The primary TSS detected for the *iclR* gene is in agreement with its annotation as a single gene transcription unit. We did not detect a primary

TSS for *ybeT* and *fbaA*. However, enrichment in the –TEX libraries for *fbaA* indicates the presence of an upstream processing site. The *thrB* gene, located within an operon, was assumed to not have a primary TSS. However, we detect an internal TSS within the upstream *thrA* gene that serves as a primary TSS for *thrB*. A primary TSS upstream of *pheM*, in the *rplT-pheM* operon, indicates *pheM* could be independently transcribed. Screenshots were taken for the B2 L1 HS1 libraries.

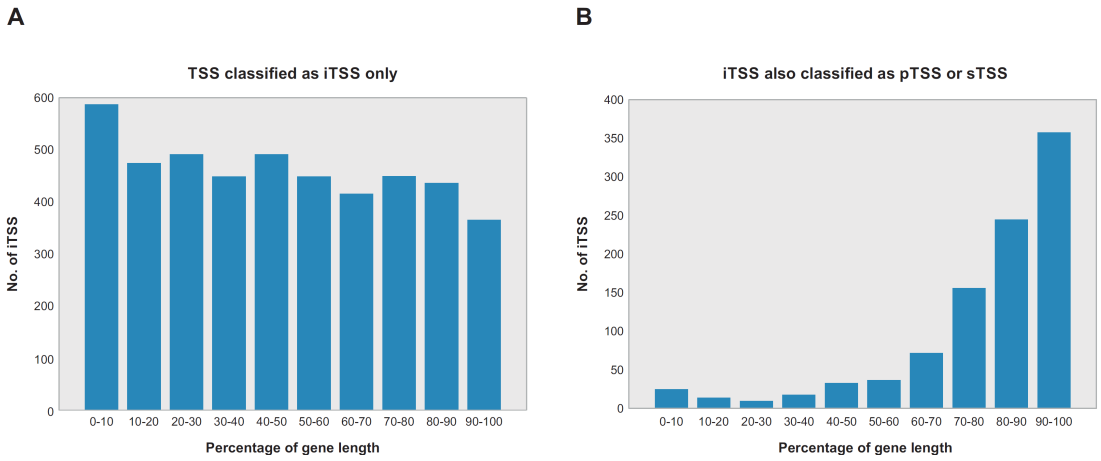


FIG. S4. Location of internal TSS (iTSS) relative to gene annotations. The relative location was determined for all TSS that are only classified as iTSS (A) or for iTSS that are also pTSS or sTSS (B). Each gene in which an iTSS was detected was divided into 10 equal sections and the number of iTSS located in each section was counted over all genes. Eight iTSS in (A) are internal to two distinct but overlapping gene annotations.

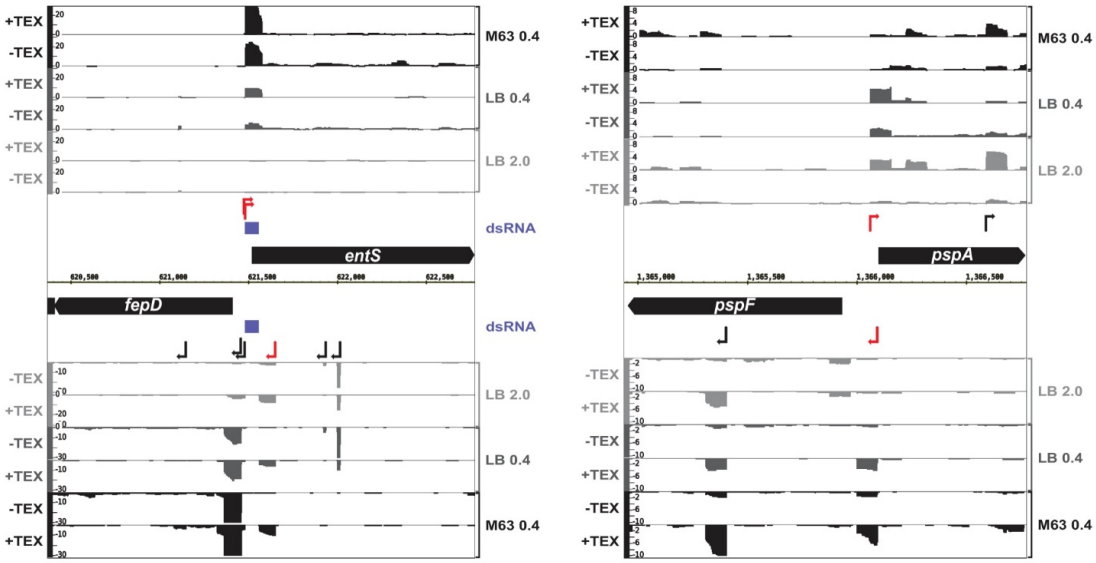


FIG. S5. cDNA coverage plots for examples of overlapping 5' UTRs. The TSS for overlapping 5' UTRs of divergently transcribed gene pairs *entS/fepD* and *pspA/pspF* are indicated in red. The locations of the annotated IP-dsRNAs reported by Lybecker *et al.* (16) are shown in blue where present. Screenshots were taken for the B2 L1 HS1 libraries.

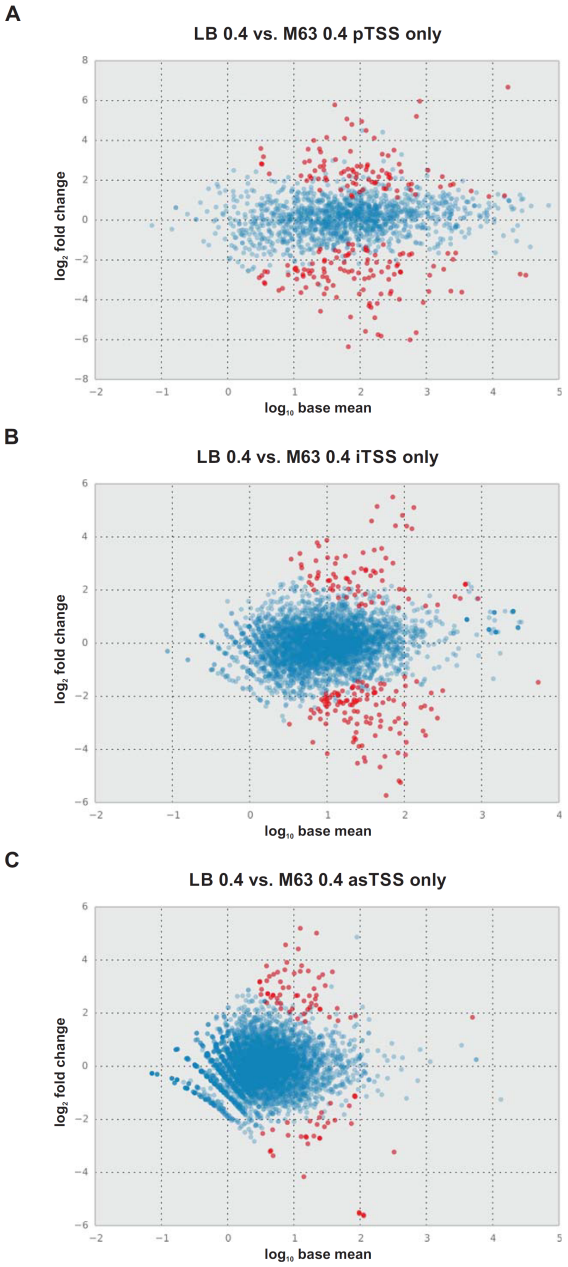


FIG S6. Comparison of mean expression levels and differential expression of TSS classified as pTSS only, iTSS only and asTSS only in LB 0.4 and M63 0.4. There is an overrepresentation of pTSS that show regulation (odd ratio 3.06, p-value: 3.75×10^{-32}) and underrepresentation of asTSS that show regulation (odd ratio 0.41, p-value: 2.98×10^{-18}). The iTSS show neither of these tendencies (odd ratio: 1.02, p-value: 0.79).

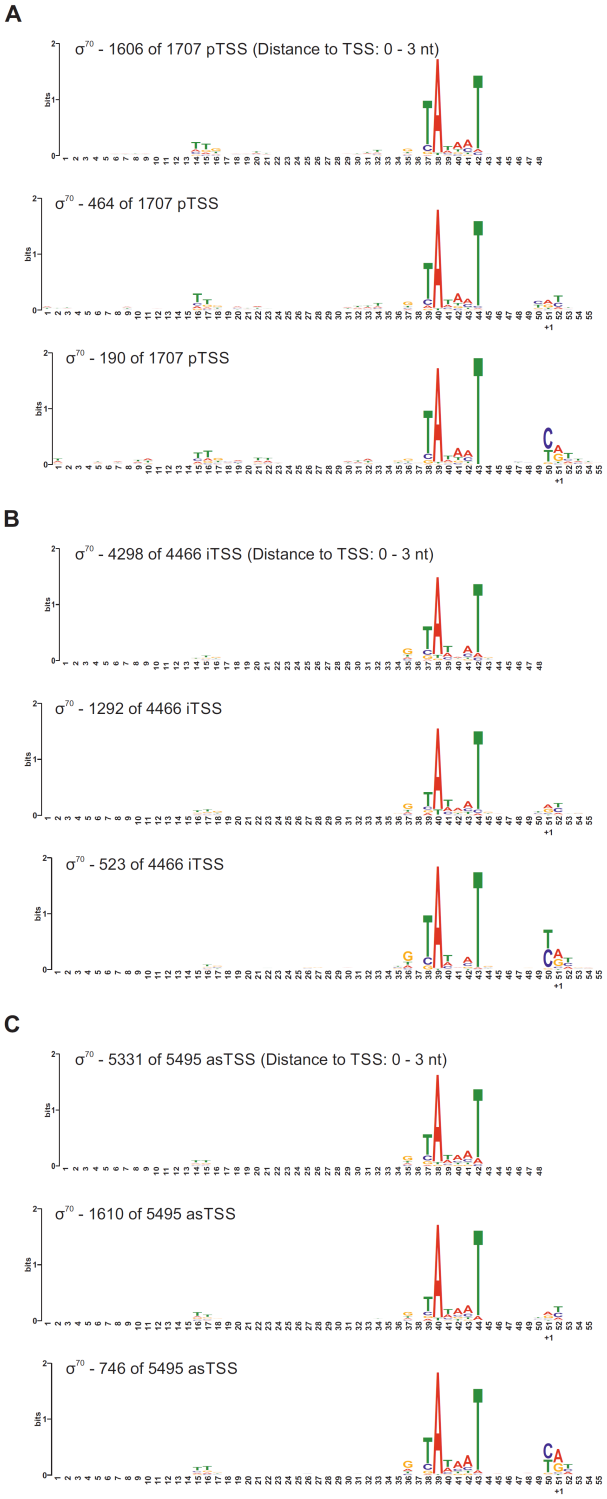

A
σ70 - 1606 of 1707 pTSS (Distance to TSS: 0 - 3 nt)
σ70 - 464 of 1707 pTSS
σ70 - 190 of 1707 pTSS
B
σ70 - 4298 of 4466 iTSS (Distance to TSS: 0 - 3 nt)
σ70 - 1292 of 4466 iTSS
σ70 - 523 of 4466 iTSS
C
σ70 - 5331 of 5495 asTSS (Distance to TSS: 0 - 3 nt)
σ70 - 1610 of 5495 asTSS
σ70 - 746 of 5495 asTSS
bits
+1

FIG S7. Promoter motifs for pTSS, iTSS and asTSS. The motif search was conducted in sequences extracted based on the 1,707 pTSS only (A), 4,466 iTSS only (B) and 5,495 asTSS only (see Supplemental Materials and Methods). The first motif in A, B and C was predicted for sequences ranging from position −50 to +1 relative to the TSS while the second and third represent the two top-scoring motifs based on sequences ranging from position −50 to +5. The first motif in A,B and C shows a canonical σ^{70} -10 sequence found in almost all sequences of the respective TSS class while the second and third motifs reveal a subset of TSS showing a slight enrichment for a pyrimidine at position −1, a purine at position +1 and a pyrimidine at position +2 as described previously (18).

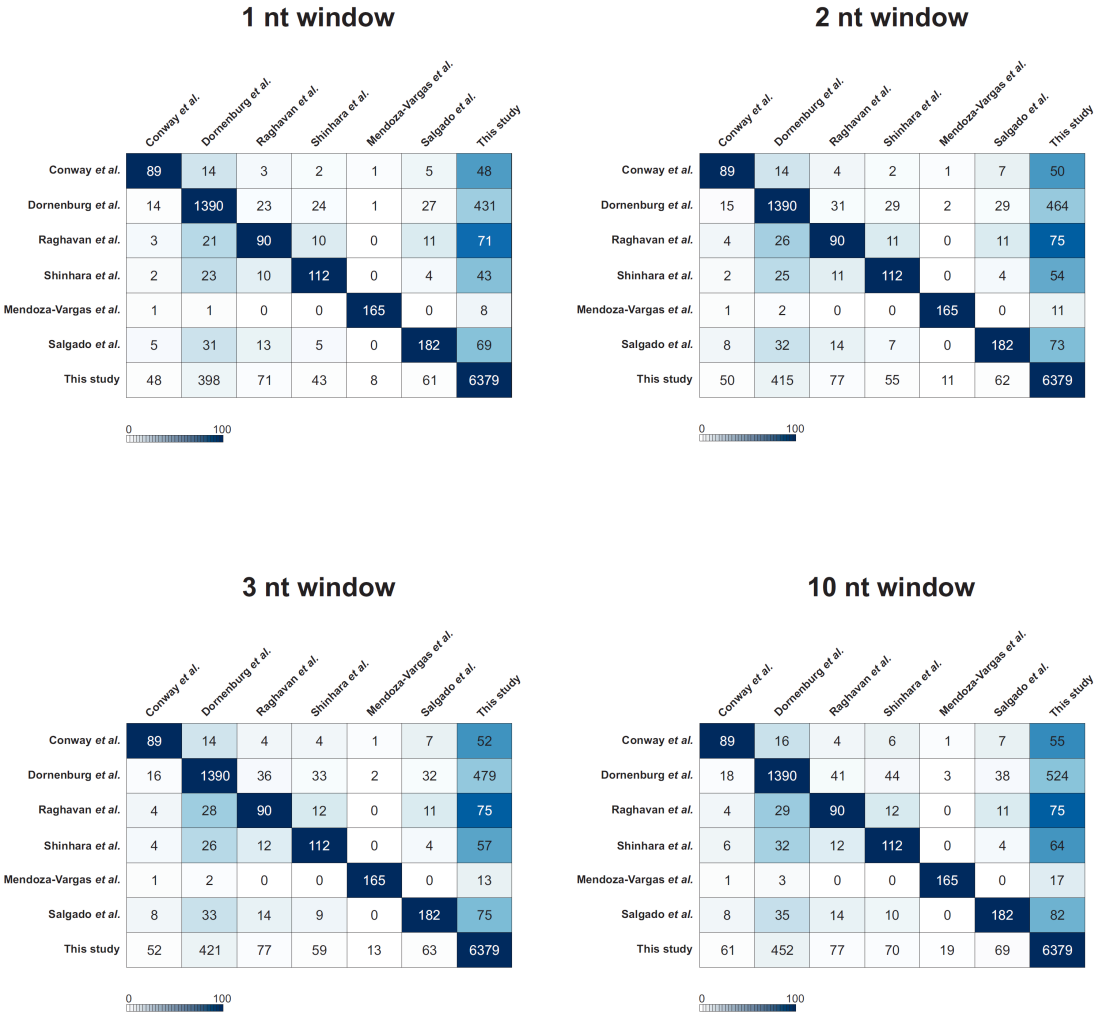


FIG. S8. Pair-wise comparisons of asTSS from this and other studies. The total numbers of annotated asTSS are shown on the main diagonal of the matrix. asTSS positions from the studies in the rows are compared to asTSS positions from the studies in the columns and the number of overlapping asTSS positions within a maximum distance of 1, 2, 3 and 10 nt is listed in the respective matrix entries. Differences in the number of matching TSS for a given pair of studies,

if either one or the other study is used as the basis for comparison, can be explained by cases where unequal numbers of TSS are matched (for example, a single TSS in study I matches several TSS from study II located in close proximity). Background color depicts the percentage of overlapping asTSS relative to the total number of asTSS from the respective study in the row.

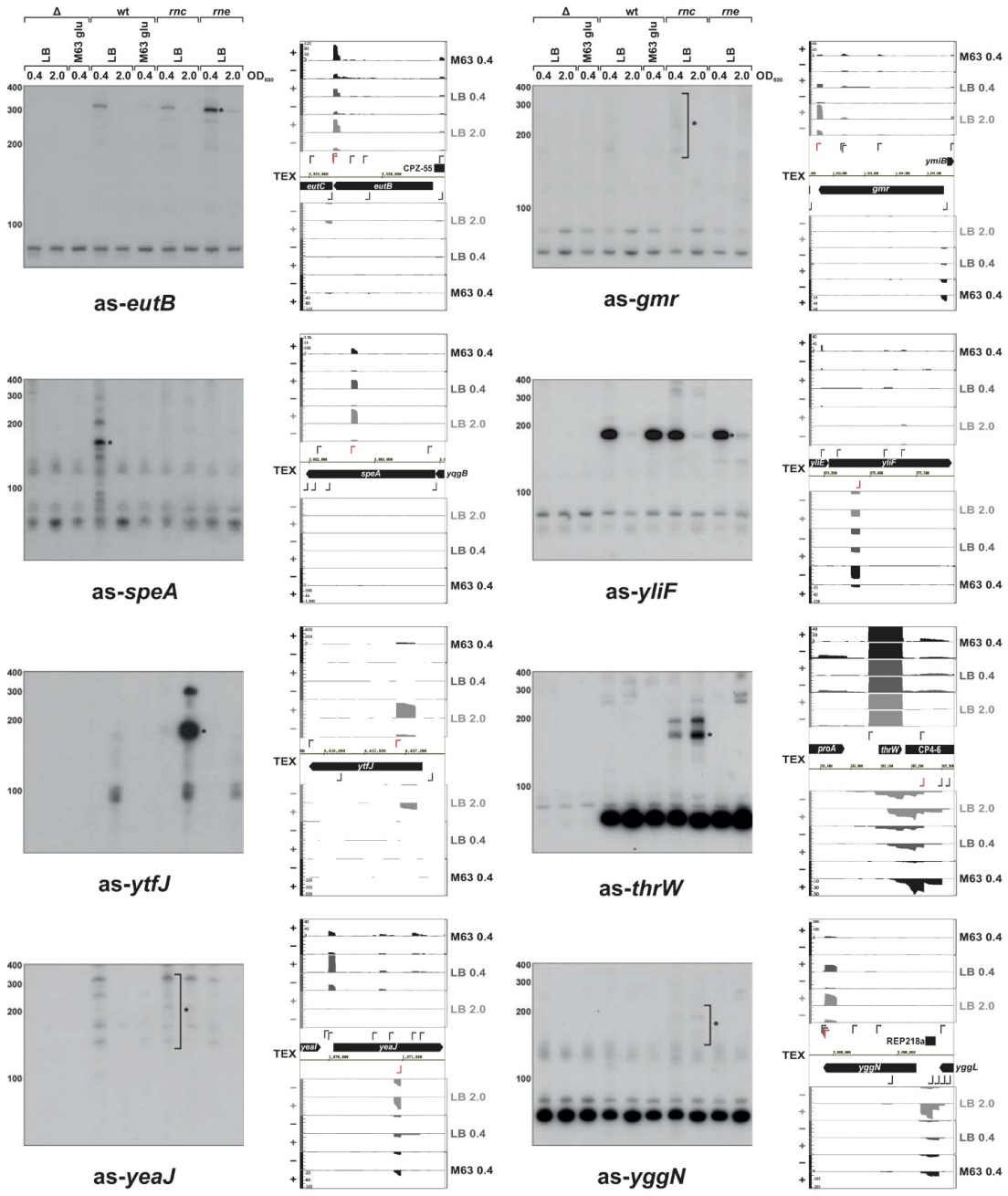


FIG. S9. Northern analysis for asRNA candidates. Northern blots and corresponding cDNA coverage plots of the –/+ TEX libraries for the three growth conditions are shown for the tested

asRNA candidates: as-*eutB*, as-*speA*, as-*ytfJ,* as-*yeaJ*, as-*gmr*, as-*yliF*, as-*thrW,* and as-*yggN*. Black stars indicate the primary bands detected for each asRNA candidate. Red arrows in the coverage plots indicate the positions of the asTSS relative to the corresponding sense gene.

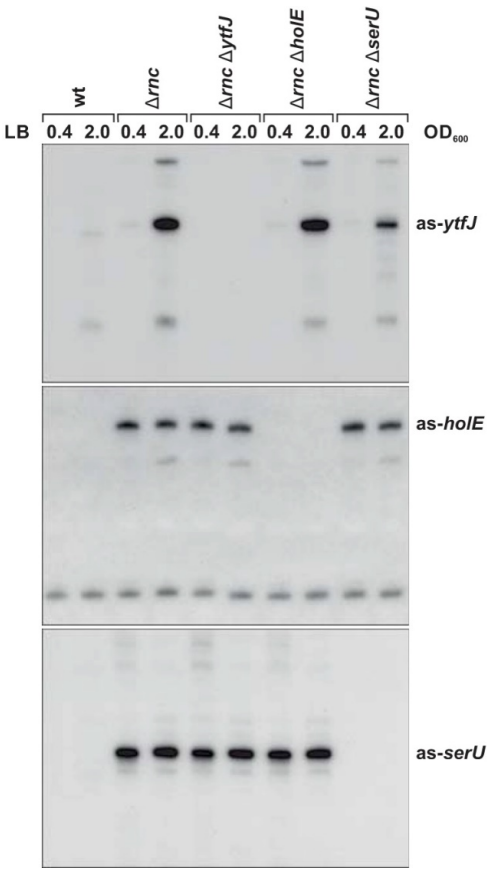


FIG. S10. Northern analysis for control *rnc* mutant strains. Strains deleted for both the *rnc* and the as-*ytfJ,* as-*holE* or as-*serU* loci were grown in LB to the indicated OD_{600} and northern analysis was performed as in Fig. 5 and Supplemental Fig. S9.

SUPPLEMENTAL TABLES

TABLE S1. Strains used in this study.

Name	MPK number	Genotype	Source
MG1655		*E. coli* F- λ- *ilvG- rfb-50 rph-1*	lab stock
NM500		MG1655 *mini-λ:tet*	N. Majdalani
NB478	MPK0331	MG1655 Δ*rnc::cat*	(19)
EM1377	MPK0330	MG1655 Δ*lac X174, rne131 zce-726::Tn10*	(20)
GSO659	MPK0310	MG1655 Δ*as-eutB::kan*	This study
GSO660	MPK0313	MG1655 Δ*as-ymfL::kan*	This study
GSO661	MPK0298	MG1655 Δ*as-qorA::kan*	This study
GSO662		MG1655 Δ*as-argR::kan*	This study
GSO663		MG1655 Δ*as-gmr::kan*	This study
GSO664		MG1655 Δ*as-holE::kan*	This study
GSO665		MG1655 Δ*as-yggN::kan*	This study
GSO666		MG1655 Δ*as-yliF::kan*	This study
GSO667		MG1655 Δ*as-speA::kan*	This study
GSO668		MG1655 Δ*as-gsiB::kan*	This study
GSO669		MG1655 Δ*as-yeaJ::kan*	This study
GSO670		MG1655 Δ*as-serU::kan*	This study
GSO671		MG1655 Δ*as-thrW::kan*	This study
GSO672		MG1655 Δ*as-ytfJ::kan*	This study
GSO718		MG1655 Δ*rnc::cat,* Δ*as-holE::kan*	This study
GSO719		MG1655 Δ*rnc::cat,* Δ*as-ytfJ::kan*	This study
GSO673		MG1655 Δ*rnc::cat,* Δ*as-serU::kan*	This study
GSO674		MG1655 Δ*rnc::cat,* Δ*as-thrW::kan*	This study

TABLE S2. Oligonucleotides used in this study.

Number	Sequence	Use
MK0095	CTTCCATGGCGTCAAGAAACAGC	Forward primer for gDNA contamination PCR check; use with MK0096
MK0096	GTTCTCAGCCTGTATCAGTCT	Reverse primer for gDNA contamination PCR check; use with MK0095
Oligos used for antisense deletion construction		
MK0383	GAACGAGAGTGATCGGCCAGGAAACGCAG CAGCGCCTGCGGTGTAGGCTGGAGCTGCT TC	Forward primer for deletion of antisense *eutB* strain construction; use with MK0399
MK0399	AAGGCCAATACCACCATCGGTATTCCGGGC ACCTTTAGCGATTCCGGGGATCCGTCGACC	Reverse primer for deletion of antisense *eutB* strain construction; use with MK0383
MK0385	CACACTGCGCTTTCAGCGCTTCAACAG	Forward primer for antisense *eutB* PCR check; use with MK0400
MK0400	AGCATCAGCGAACTGCGTGAG	Reverse primer for antisense *eutB* PCR check; use with MK0385
MK0415	CGGCTCCCAACGAATGACTCTGACGGGCA CTCCGTAGTGAGTGTAGGCTGGAGCTGCTT C	Forward primer for deletion of antisense *ymfL* strain construction; use with MK0416
MK0416	TTGCCGAGTGGTTACGCTGAAGCGGCTGA CTGGCTCGATGATTCCGGGGATCCGTCGA CC	Reverse primer for deletion of antisense *ymfL* strain construction; use with MK0415
MK0417	GCCAGCACACACCTTCGTCATTACT	Forward primer for antisense *ymfL* PCR check; use with MK0418
MK0418	GACTTAAGACCGGATATCTATC	Reverse primer for antisense *ymfL* PCR check; use with MK0417
MK0387	GCCGTGGTTACTACTGGCGATTGCGGTGGT GTGGGTTATTGTGTAGGCTGGAGCTGCTTC	Forward primer for deletion of antisense *qor* strain construction; use with MK0388
MK0388	ACCACGCGGGAGGAATTAACCGAGGCCAG TAATGAACTGATTCCGGGGATCCGTCGACC	Reverse primer for deletion of antisense *qor* strain construction; use with MK0387
MK0389	GTCATGCTGATGGTCACCGGCG	Forward primer for antisense *qor* PCR check; use with MK0389
MK0390	CTGCAACGCCGCGGCTTAATGGTCAG	Reverse primer for antisense *qor* PCR check; use with MK0390
AZ1312	GTACGCTTAATGCAGCAACAGTGGGTGTAG GCTGGAGCTGCTTC	Forward primer for deletion of antisense *argR* strain construction; use with AZ1313
AZ1313	GAAATGGTTTACTGCCTGCCAGCTATTCCG GGGATCCGTCGACC	Reverse primer for deletion of antisense *argR* strain construction; use with AZ1312
AZ1314	CGTTGCGCCTTTCTCTTCCGC	Forward primer for antisense *argR* PCR check; use with AZ1315
AZ1315	GCTCGGCTAAGCAAGAAGAACTAG	Reverse primer for antisense *argR* PCR check; use with AZ1314
AZ1282	GCGCCGGGCATCCTCAAATAGGTGTAGGC	Forward primer for deletion of antisense *gmr* strain

	TGGAGCTGCTTC	construction; use with AZ1283
AZ1283	CAACAATTTAGCCAACTAGGTGCGCATTCC GGGGATCCGTCGACC	Reverse primer for deletion of antisense *gmr* strain construction; use with AZ1282
AZ1284	GAAACTCCAGTGGCTTTTGCCAG	Forward primer for antisense *gmr* PCR check; use with AZ1285
AZ1285	CTGCACGTCAGCTCGCCG	Reverse primer for antisense *gmr* PCR check; use with AZ1284
AZ1300	GAGAGATCGGGTGGGGCA GGTGTAGGCTGGAGCTGCTTC	Forward primer for deletion of antisense *holE* strain construction; use with AZ1301
AZ1301	CGTAGCGAAGGGAGCGTGCATTCCGGGGA TCCGTCGACC	Reverse primer for deletion of antisense *holE* strain construction; use with AZ1300
AZ1302	GACATGCACCATGACTCTGATGG	Forward primer for antisense *holE* PCR check; use with AZ1303
AZ1303	CACCACTGAATCCTGTTTCAACACC	Reverse primer for antisense *holE* PCR check; use with AZ1302
AZ1306	CGCCAACTTTACCCACTGTGT AGGTGTAGGCTGGAGCTGCTTC	Forward primer for deletion of antisense *yggN* strain construction; use with AZ1307
AZ1307	GCAGCAAAATGCGCAGCCGTCATTCCGGG GATCCGTCGACC	Reverse primer for deletion of antisense *yggN* strain construction; use with AZ1306
AZ1308	GAAGCCGTCGGTCTTATCGCAG	Forward primer for antisense *yggN* PCR check; use with AZ1309
AZ1309	CGGTAAGCAATATTCCCTGAATGCC	Reverse primer for antisense *yggN* PCR check; use with AZ1308
AZ1288	CGCAAATACATACAATCCGGTCGG CGTGTAGGCTGGAGCTGCTTC	Forward primer for deletion of antisense *yliF* strain construction; use with AZ1289
AZ1289	CATTCTCCGCCTGGGATAAAAGTGGATTCC GGGGATCCGTCGACC	Reverse primer for deletion of antisense *yliF* strain construction; use with AZ1288
AZ1290	GCCCGCCATCCAGTACGCG	Forward primer for antisense *yliF* PCR check; use with AZ1291
AZ1291	CGGTACTGCGGGAAAAATTGTGCG	Reverse primer for antisense *yliF* PCR check; use with AZ1290
AZ1261	GTGGTCGATAGCACCGTCAGAGTGTGTAG GCTGGAGCTGCTTC	Forward primer for deletion of antisense *speA* strain construction; use with AZ1262
AZ1262	CAGTGCTTCGACGTCGGCGGATTCCGGGG ATCCGTCGACC	Reverse primer for deletion of antisense *speA* strain construction; use with AZ1261
AZ1263	CGAACACGTCAACCGCTTCGGTA	Forward primer for antisense *speA* PCR check; use with AZ1264
AZ1264	GTTGAAACCCTGCGTGAAGCCG	Reverse primer for antisense *speA* PCR check; use with AZ1263
AZ1255	CTGCGCGGTGCTGTGGTTATGGTGTAGGCT GGAGCTGCTTC	Forward primer for deletion of antisense *gsiB* strain construction; use with AZ1256
AZ1256	CGGAAGCGATCGATCCGACAACATTCCGG GGATCCGTCGACC	Reverse primer for deletion of antisense *gsiB* strain construction; use with AZ1255

AZ1257	GGTTGAAGCCGACCAGCCAG	Forward primer for antisense *gsiB* PCR check; use with AZ1258
AZ1258	GTGCTGGCGGAGAGTTATACCG	Reverse primer for antisense *gsiB* PCR check; use with AZ1257
AZ1273	CGAATCGACTGTTTAATCGCCTGAGGTGTA GGCTGGAGCTGCTTC	Forward primer for deletion of antisense *yeaJ* strain construction; use with AZ1274
AZ1274	GTCCTTCCTGCAGTCAGGAAGTAATTCCGG GGATCCGTCGACC	Reverse primer for deletion of antisense *yeaJ* strain construction; use with AZ1273
AZ1275	CCGGAAGAGAAGCCGATCTCTTTC	Forward primer for antisense *yeaJ* PCR check; use with AZ1276
AZ1276	GCTTTTGATCAGGCAGTGGAAGGC	Reverse primer for antisense *yeaJ* PCR check; use with AZ1275
AZ1249	GATACAAAGGCTTTCAAAAAAGCTGCGGTG TAGGCTGGAGCTGCTTC	Forward primer for deletion of antisense *serU* strain construction; use with AZ1250
AZ1250	TGTAAAAAACGTTCGGCAAGAGTGACATTC CGGGGATCCGTCGACC	Reverse primer for deletion of antisense *serU* strain construction; use with AZ1249
AZ1251	CAACGCGTCATAAATGTTTACGCAAGTG	Forward primer for antisense *serU* PCR check; use with AZ1252
AZ1252	CCACAAATGGCGCAGGATAAATTAAGAC	Reverse primer for antisense *serU* PCR check; use with AZ1251
AZ1267	CACCACTACAGCGGAACTTTCTTCAGTGTA GGCTGGAGCTGCTTC	Forward primer for deletion of antisense *ytfJ* strain construction; use with AZ1268
AZ1268	CTTAGTGACTATAGACTATCCGGGCATTCC GGGGATCCGTCGACC	Reverse primer for deletion of antisense *ytfJ* strain construction; use with AZ1267
AZ1269	CCCCTATCCCATAGATAACGATAGG	Forward primer for antisense *ytfJ* PCR check; use with AZ1270
AZ1270	GTATAACCGTCCACGGAACAGGATC	Reverse primer for antisense *ytfJ* PCR check; use with AZ1269
AZ1294	CGTAACAACGTAGTACGATGAACATTGCGT GTAGGCTGGAGCTGCTTC	Forward primer for deletion of antisense *thrW* strain construction; use with AZ1295
AZ1295	CGGAAGTGGCGGTAAGCACACATTCCGGG GATCCGTCGACC	Reverse primer for deletion of antisense *thrW* strain construction; use with AZ1294
AZ1296	CAC CCT AGC CGA TGC CGT G	Forward primer for antisense *thrW* PCR check; use with AZ1297
AZ1297	CGATGCCATCGCCCATATTCGTG	Reverse primer for antisense *thrW* PCR check; use with AZ1296
Oligos used for riboprobe construction		
MK0372	CTTGTCCCATTGACGCCATCACGCT	Forward primer for riboprobe construction for antisense *eutB*; use with MK0373
MK0373	CAGAGATGCATAATACGACTCACTATAGGG GGTGATGACATCATGCTCAACTAC	Reverse primer for riboprobe construction for antisense *eutB*; use with MK0372
MK0374	GTTATGACCGCTGGCGTTACTAAGG	Forward primer for riboprobe construction for antisense *qor*; use with MK0375
MK0375	CAGAGATGCATAATACGACTCACTATAGGG	Reverse primer for riboprobe construction for antisense

	CTGTTCTCTTTGATTGCCAGCGGTGTGA	*qor*; use with MK0374
MK0413	CGGTACAGGAAGCGCAATCAGTTGCGAG	Forward primer for riboprobe construction for antisense *ymfL*; use with MK0414
MK0414	CAGAGATGCATAATACGACTCACTATAGGG GAAGACGGTGTAGTGGAACCGCATGA	Reverse primer for riboprobe construction for antisense *ymfL*; use with MK0413
AZ1310	GTATTCATTGTGTGAATGACATGTCGC	Forward primer for riboprobe construction for antisense *argR* ; use with AZ1311
AZ1311	CAGAGATGCATAATACGACTCACTATAGGG ACCACCCCTGCTAACGGTTTC	Reverse primer for riboprobe construction for antisense *argR*; use with AZ1310
AZ1280	GCAAAAGGGGGAAAATGAATAATGC	Forward primer for riboprobe construction for antisense *gmr*; use with AZ1281
AZ1281	CAGAGATGCATAATACGACTCACTATAGGG CCAAGAACGGGATCAATGAGC	Reverse primer for riboprobe construction for antisense *gmr*; use with AZ1280
AZ1298	GCAGGCGTTATGTAAGAAAGTGTAACTC	Forward primer for riboprobe construction for antisense *holE*; use with AZ1299
AZ1299	CAGAGATGCATAATACGACTCACTATAGGG GGCATTTAAAGAACGCTACAATATGCCG	Reverse primer for riboprobe construction for antisense *holE*; use with AZ1298
AZ1304	GGAGATCTACAAAGTTAGAGGCAGG	Forward primer for riboprobe construction for antisense *yggN*; use with AZ1305
AZ1305	CAGAGATGCATAATACGACTCACTATAGGG TGGGGGGGCTGCAATCCTC	Reverse primer for riboprobe construction for antisense *yggN*; use with AZ1304
AZ1286	GGTTTTACCGTCAAAAGAGATAAACCCTG	Forward primer for riboprobe construction for antisense *yliF*; use with AZ1287
AZ1287	CAGAGATGCATAATACGACTCACTATAGGG CAGCTGTTGGTTGGTTTGCGCAC	Reverse primer for riboprobe construction for antisense *yliF*; use with AZ1286
AZ1259	CATGCTCAAATAAAGCTGCTCAGCC	Forward primer for riboprobe construction for antisense *speA*; use with AZ1260
AZ1260	CAGAGATGCATAATACGACTCACTATAGGG CTGCAGAAGATGCGCCGCG	Reverse primer for riboprobe construction for antisense *speA*; use with AZ1259
AZ1253	GCGTCACGTTCATACTGATATAACGC	Forward primer for riboprobe construction for antisense *gsiB*; use with AZ1254
AZ1254	CAGAGATGCATAATACGACTCACTATAGGG GCCCAAACTGGACAGCATAACCTG	Reverse primer for riboprobe construction for antisense *gsiB*; use with AZ1253
AZ1271	GCACCATATTCAGCAAAATTAACGCCG	Forward primer for riboprobe construction for antisense *yeaJ*; use with AZ1272
AZ1272	CAGAGATGCATAATACGACTCACTATAGGG CTAACCGATACCGATTCGGGGC	Reverse primer for riboprobe construction for antisense *yeaJ*; use with AZ1271
AZ1247	GAATCAAGTGCTGAATGTCACAGTATCG	Forward primer for riboprobe construction for antisense *serU*; use with AZ1248
AZ1248	CAGAGATGCATAATACGACTCACTATAGGG GGACCGGTCTCGAAAACCGGAG	Reverse primer for riboprobe construction for antisense *serU*; use with AZ1247
AZ1265	CAGCAATATGTTGCAGTACTCGCAC	Forward primer for riboprobe construction for antisense *ytfJ*; use with AZ1266

AZ1266	CAGAGATGCATAATACGACTCACTATAGGG CTACGCAAGATTCTGGCACTCAC	Reverse primer for riboprobe construction for antisense *ytfJ*; use with AZ1265
AZ1292	GTG AGC GAA GCC CTA TCA GGC	Forward primer for riboprobe construction for antisense *thrW*; use with AZ1293
AZ1293	CAGAGATGCATAATACGACTCACTATAGGG GCGCATTCGTAATGCGAAGGTCG	Reverse primer for riboprobe construction for antisense *thrW*; use with AZ1292

TABLE S3. Summary of *E. coli* K-12 MG1655 dRNA-seq libraries analyzed in this study.

Biological replicate	Library replicate	Sequencing run	Biological condition			Sequencing technique
			M63 0.4	LB 0.4	LB 2.0	
B1	L1	GA	x	x	x	GAIIx
B2	L1	HS1	x	x	x	HiSeq 2000
B2	L1	HS2	x	x	x	HiSeq 2000
B1	L2	HS2			x	HiSeq 2000
B2	L2	HS2			x	HiSeq 2000

TABLE S4. Read mapping statistics for the *E. coli* dRNA-seq libraries. This table contains the total number of reads after quality trimming, the number of mapped and uniquely mapped reads for each growth condition, library replicate and sequencing replicate (see Supplemental Materials and Methods). Percentage values are relative to the total number of reads after quality trimming.

	Library	Total number of reads after quality trimming	Mapped reads	% mapped reads	Uniquely mapped reads	% uniquely mapped reads
L1 GA	M63 0.4 B1 +TEX	2,226,993	2,179,278	97.86	1,353,514	60.78
	M63 0.4 B1 –TEX	3,618,430	3,587,012	99.13	1,641,287	45.36
	LB 0.4 B1 +TEX	3,386,772	3,319,935	98.03	2,450,714	72.36
	LB 0.4 B1 –TEX	2,348,269	2,313,292	98.51	1,448,100	61.67
	LB 2.0 B1 +TEX	1,812,576	1,662,957	91.75	1,255,341	69.26
	LB 2.0 B1 –TEX	2,950,320	2,833,237	96.03	1,369,732	46.43
L1 HS1	M63 0.4 B2 +TEX	8,676,235	8,320,773	95.90	5,616,514	64.73
	M63 0.4 B2 –TEX	8,180,585	7,878,531	96.31	4,840,039	59.16
	LB 0.4 B2 +TEX	6,173,388	6,034,843	97.76	4,521,459	73.24
	LB 0.4 B2 –TEX	7,901,590	7,050,172	89.22	3,178,987	40.23
	LB 2.0 B2 +TEX	7,039,151	6,785,582	96.40	4,724,268	67.11
	LB 2.0 B2 –TEX	5,486,064	4,662,787	84.99	2,771,432	50.52
L1 HS2	M63 0.4 B2 +TEX	8,785,626	8,433,738	95.99	5,688,313	64.75
	M63 0.4 B2 –TEX	9,814,115	9,463,702	96.43	5,799,209	59.09
	LB 0.4 B2 +TEX	5,878,169	5,753,542	97.88	4,301,601	73.18
	LB 0.4 B2 –TEX	8,321,493	7,434,035	89.34	3,356,765	40.34
	LB 2.0 B2 +TEX	7,792,864	7,520,559	96.51	5,221,142	67.00
	LB 2.0 B2 –TEX	6,648,330	5,654,792	85.06	3,352,443	50.43
L2 HS2	LB 2.0 B1 +TEX	6,336,509	5,383,288	84.96	3,970,765	62.66
	LB 2.0 B1 –TEX	7,782,920	5,842,996	75.07	3,399,792	43.68
	LB 2.0 B2 +TEX	6,543,088	5,961,511	91.11	3,984,821	60.90
	LB 2.0 B2 –TEX	5,327,191	3,753,567	70.46	2,217,871	41.63

TABLE S5. Mapping statistics based on strand and RNA class. This table indicates the number of reads mapped to the different RNA classes (mRNA, ncRNA, rRNA, tRNA, and tmRNA) for each strand across all libraries and biological conditions. The numbers for both the mapped and uniquely mapped reads per RNA class are shown with percentage values calculated from the total number of mapped reads regardless of mapped location (taken from Table S4) for the respective biological conditions.

		Biological condition											
		M63 0.4				LB 0.4				LB 2.0			
		Mapped reads		Uniquely mapped reads		Mapped reads		Uniquely mapped reads		Mapped reads		Uniquely mapped reads	
	Total*	39,863,034		24,938,876		31,905,819		19,257,626		50,061,276		32,267,607	
sense	mRNA	7,542,584	(19%)	7,459,916	(30%)	4,754,092	(15%)	4,679,204	(24%)	12,621,458	(25%)	12,449,865	(39%)
	ncRNA	2,144,882	(5%)	2,134,932	(9%)	1,974,888	(6%)	1,972,296	(10%)	6,865,538	(14%)	6,845,341	(21%)
	rRNA	9,337,211	(23%)	1,281,358	(5%)	8,409,540	(26%)	843,630	(4%)	16,392,800	(33%)	1,723,029	(5%)
	tRNA	8,608,679	(22%)	5,226,638	(21%)	5,500,223	(17%)	4,821,845	(25%)	3,695,560	(7%)	1,600,498	(5%)
	tmRNA	283,482	(1%)	283,431	(1%)	182,898	(1%)	182,884	(1%)	510,495	(1%)	510,408	(2%)
antisense	mRNA	976,736	(2%)	945,150	(4%)	566,133	(2%)	535,981	(3%)	1,900,343	(4%)	1,834,877	(6%)
	ncRNA	544,883	(1%)	541,313	(2%)	131,872	(0%)	131,022	(1%)	301,004	(1%)	300,376	(1%)
	rRNA	1,820	(0%)	70	(0%)	2,498	(0%)	30	(0%)	3,385	(0%)	435	(0%)
	tRNA	2,471	(0%)	2,022	(0%)	2,542	(0%)	2,152	(0%)	30,965	(0%)	29,200	(0%)
	tmRNA	8	(0%)	7	(0%)	12	(0%)	12	(0%)	26	(0%)	26	(0%)

*Total reads were calculated by summing all the reads mapped to the *E. coli* genome from all libraries (+TEX and –TEX) from a particular condition including reads that mapped to locations other than the listed RNA classes. The numbers were generated from the data in Supplemental Table S4.

TABLE S6. Known annotated asRNAs. This table contains the previously known annotated asRNAs taken from NCBI annotation.

Gene	Locus tag	Strand	Start	End	Description
arrS	b4704	-	3656009	3656077	asRNA ArrS, GadE-regulated, function unknown
gadY	b4452	+	3662887	3662991	asRNA regulator of transcriptional activator GadX mRNA
ohsC	b4608	+	2698542	2698618	asRNA regulator of *shoB* toxin
rdlA	b4420	+	1268546	1268612	asRNA RdlA affects LdrA translation; proposed addiction module in LDR-A repeat, with toxic peptide LdrA
rdlB	b4422	+	1269081	1269146	asRNA RdlB affects LdrB translation; proposed addiction module in LDR-B repeat, with toxic peptide LdrB
rdlC	b4424	+	1269616	1269683	asRNA RdlC affects LdrC translation; proposed addiction module in LDR-C repeat, with toxic peptide LdrC
rdlD	b4454	+	3698159	3698224	asRNA RdlD affects LdrD translation; proposed addiction module in LDR-D repeat, with toxic peptide LdrD
sibA	b4436	+	2151333	2151475	asRNA regulator of toxic IbsA protein; in SIBa repeat
sibB	b4437	+	2151668	2151803	asRNA regulator of toxic IbsB protein; in SIBb repeat
sibC	b4446	+	3054871	3055010	asRNA regulator of toxic IbsC protein; in SIBc repeat
sibD	b4447	-	3192745	3192887	asRNA regulator of toxic IbsD protein; in SIBd repeat
sibE	b4611	-	3193121	3193262	asRNA regulator of toxic IbsE protein; in SIBe repeat
sokB	b4429	+	1490143	1490198	asRNA blocking *mokB*, and hence *hokB*, translation
sokC	b4413	+	16952	17006	asRNA blocking *mokC*, and hence *hokC*, translation
sokE	b4700	+	606957	607015	asRNA at remnant *mokE/hokE* locus
sokX	b4701	+	2885376	2885431	asRNA, function unknown
symR	b4625	+	4577858	4577934	asRNA destabilizing divergent and overlapping *symE* mRNA

TABLE S7. Published *E. coli* transcriptome studies reporting asRNAs used for comparisons.

Reference	Strain	Growth conditions	Special treatment	Sequencing	Total number of reads	Annotated asTSS
Conway *et al.* (15)	K12 BW38028/ BW39452 Δ*rpoS*	Fermentor MOPS + 0.2% glucose OD_{600} ~0.1-stationary phase	Terminator Exonuclease (TEX) treatment	ABI SOLiD	72,147,745	89
Dornenburg *et al.* (10)	K12 MG1655	LB OD_{600} 0.7	rRNA depletion	Illumina (44 cycles)	8,967,903	1,390
Raghavan *et al.* (11)	K12 MG1655	LB OD_{600} ~0.5	rRNA depletion	Illumina GA II (35 cycles)	30,206,434	90
Shinhara *et al.* (12)	K12 BW25113	M63 glucose OD_{600} 0.76 at 37°C	Extraction of low-molecular weight RNAs	Illumina 1G (35 cycles)	12,473,172	112[#]
Mendoza-Vargas *et al.* (13)	K12 MG1655	LB and M63 glucose at 37°C and 30°C	rRNA depletion	Roche 454 GS20 (reads ~100 nt long)	~350,000	165
Salgado *et al.* *(14)*	K12 MG1655 / MG1655 Δ*rppH*	LB or MOPS media with 0.2% Glucose or 0.2% acetate at 37°C	rRNA depletion, enrichment for 5'-P or 5'-PPP	Illumina GAIIx (36 cycles)	77,628,858[*]	182

*Number of mapped non-rRNA sequences

[#]TSS represent extracted 5' end positions of reported antisense transcripts

TABLE S8. Candidate asRNAs tested by northern analysis.

Strand	Sense gene	Start-end	TSS expression bin	Size	Expression	Detected by others
+	*qorA*	4,261,162-4,261,303	>100,000	70, 200	*rne* LB 0.4	No
-	*gsiB*	869,344-869-393	10,000-100,000	170	LB 0.4; M63	(10)
-	*holE*	1,923,282-1,923,331	10,000-100,000	230	*rnc* LB 0.4 + 2.0	(10, 11, 14)
+	*serU*	2,041,439-2,041,488	10,000-100,000	170	*rnc* LB 0.4 +2.0; *rne*	(10)
+	*eutB*	2,555,334-2,555,448	10,000-100,000	310-320	LB 0.4; *rnc* 0.4; *rne* 0.4	(10, 12, 15)
+	*speA*	3,082,647-3,082,696	10,000-100,000	160	LB 0.4	No
+	*ytfJ*	4,437,157-4,437,206	10,000-100,000	180-190	*rnc* 0.4	(10)
-	*yeaJ*	1,870,929-1,870,978	1,000-10,000	multiple bands	LB 0.4; *rnc* 0.4+ 2.0; *rne* 0.4	(10, 11)
+	*gmr*	1,342,751-1,342,800	1,000-10,000	multiple bands	*rne* 0.4	(10)
-	*yliF*	874,841-874,890	1,000-10,000	185	LB 0.4; M63 0.4; *rnc* 0.4; *rne* 0.4	(10-12)
-	*thrW*	262,193-262,242	1,000-10,000	180	*rnc* LB 0.4 +2.0	No
+	*yggN*	3,098,913-3,098,986	1,000-10,000	multiple bands	*rnc* LB 0.4 +2.0	No
-	*argR*	3,383,214-3,383,263	1,000-10,000	210	LB 0.4; *rnc* 0.4; *rne* 0.4 + 2.0	No
-	*ymfL*	1,202,757-1,203,131	1,000-10,000	>50	LB 0.4 + 2.0; M63; *rnc* 0.4 + 2.0	(12, 15)

SUPPLEMENTAL DATA SETS

Data Set S1. TSS map. This table contains information on positions and assigned classes of all annotated TSS. It lists all TSS that were detected in at least one of the three conditions (column "detected" = 1). In case a TSS was not detected in a certain condition the value of detected is "0". Also, if the TSS is assigned to more than one class, there is one row for each class assignment and each associated gene.

Data Set S2. Overlapping 5' UTRs (Tab S2A) and comparison to IP-dsRNAs (Tab S2B). Tab S2A contains all pairs of overlapping 5' UTRs based on primary and secondary TSS, which can also be classified as internal and/or antisense, with a minimum overlap of 10 nt. Tab S2B contains all IP-dsRNAs described by Lybecker *et al.* (16) for which we found at least one matching asTSS.

Data Set S3. Exclusively asTSS bin expression table. This data set contains information on the expression of exclusively asTSS as well as the overlap to asRNAs from previous studies. The TSS are separated into bins according to the maximum RPKM value over all libraries. Each worksheet contains the TSS assigned to a specific bin.

SUPPLEMENTAL REFERENCES

1. **Sharma, C. M., S. Hoffmann, F. Darfeuille, J. Reignier, S. Findeiss, A. Sittka, S. Chabas, K. Reiche, J. Hackermuller, R. Reinhardt, P. F. Stadler, and J. Vogel.** 2010. The primary transcriptome of the major human pathogen *Helicobacter pylori.* Nature **464:**250-255.
2. **Berezikov, E., F. Thuemmler, L. W. van Laake, I. Kondova, R. Bontrop, E. Cuppen, and R. H. Plasterk.** 2006. Diversity of microRNAs in human and chimpanzee brain. Nat Genet **38:**1375-1377.
3. **Förstner, K. U., J. Vogel, and C. M. Sharma.** 2014. READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data. Bioinformatics **pii:**btu533.
4. **Hoffmann, S., C. Otto, S. Kurtz, C. M. Sharma, P. Khaitovich, J. Vogel, P. F. Stadler, and J. Hackermuller.** 2009. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol **5:**e1000502.
5. **Nicol, J. W., G. A. Helt, S. G. Blanchard, Jr., A. Raja, and A. E. Loraine.** 2009. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics **25:**2730-2731.
6. **Dugar, G., A. Herbig, K. U. Forstner, N. Heidrich, R. Reinhardt, K. Nieselt, and C. M. Sharma.** 2013. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. PLoS Genet **9:**e1003495.
7. **Mao, X., Q. Ma, C. Zhou, X. Chen, H. Zhang, J. Yang, F. Mao, W. Lai, and Y. Xu.** 2014. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. Nucleic Acids Res. **42:**D654-D659.
8. **Love, M. I., W. Huber, and S. Anders.** 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. bioRxiv **In press**.
9. **Bailey, T. L., and C. Elkan.** 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. **2:**28-36.
10. **Dornenburg, J. E., A. M. Devita, M. J. Palumbo, and J. T. Wade.** 2010. Widespread antisense transcription in *Escherichia coli*. MBio **1:**e00024-00010.
11. **Raghavan, R., D. B. Sloan, and H. Ochman.** 2012. Antisense transcription is pervasive but rarely conserved in enteric bacteria. MBio **3:**e00156-00112.

12. **Shinhara, A., M. Matsui, K. Hiraoka, W. Nomura, R. Hirano, K. Nakahigashi, M. Tomita, H. Mori, and A. Kanai.** 2011. Deep sequencing reveals as-yet-undiscovered small RNAs in *Escherichia coli*. BMC Genomics **12:**428.

13. **Mendoza-Vargas, A., L. Olvera, M. Olvera, R. Grande, L. Vega-Alvarado, B. Taboada, V. Jimenez-Jacinto, H. Salgado, K. Juarez, B. Contreras-Moreira, A. M. Huerta, J. Collado-Vides, and E. Morett.** 2009. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. PLoS One **4:**e7526.

14. **Salgado, H., 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.** 2013. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res. **41:**D203-D213.

15. **Conway, T., J. P. Creecy, S. M. Maddox, J. E. Grissom, T. L. Conkle, T. M. Shadid, J. Teramoto, P. San Miguel, T. Shimada, A. Ishihama, H. Mori, and B. L. Wanner.** 2014. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. mBio **5:**e01442-01414.

16. **Lybecker, M., B. Zimmermann, I. Bilusic, N. Tukhtubaeva, and R. Schroeder.** 2014. The double-stranded transcriptome of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **111:**3134-3139.

17. **Thomason, M. K., F. Fontaine, N. De Lay, and G. Storz.** 2012. A small RNA that regulates motility and biofilm formation in response to changes in nutrient availability in *Escherichia coli*. Mol Microbiol **84:**17-35.

18. **Kim, D., J. S. Hong, Y. Qiu, H. Nagarajan, J. H. Seo, B. K. Cho, S. F. Tsai, and B. O. Palsson.** 2012. Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling. PLoS Genet **8:**e1002867.

19. **Opdyke, J. A., E. M. Fozo, M. R. Hemm, and G. Storz.** 2011. RNase III participates in GadY-dependent cleavage of the *gadX-gadW* mRNA. J. Mol. Biol. **406:**29-43.

20. **Massé, E., F. E. Escorcia, and S. Gottesman.** 2003. Coupled degradation of a small regulatory RNA and its mRNA targets in *Escherichia coli*. Genes Dev. **17:**2374-2383.

3.3 DIFFERENTIAL RNA-SEQ (DRNA-SEQ) FOR ANNOTATION OF TRANSCRIPTIONAL START SITES AND SMALL RNAS IN HELICOBACTER PYLORI

Methods 86 (2015) 89–101

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in *Helicobacter pylori*

Thorsten Bischler [a], Hock Siew Tan [a], Kay Nieselt [b], Cynthia M. Sharma [a,*]

[a] Research Center for Infectious Diseases (ZINF), University of Würzburg, Josef-Schneider-Str. 2/Bau D15, 97080 Würzburg, Germany
[b] Integrative Transcriptomics, ZBIT (Center for Bioinformatics Tübingen), University of Tübingen, Sand 14, D-72076 Tübingen, Germany

ARTICLE INFO

Article history:
Received 8 April 2015
Received in revised form 7 June 2015
Accepted 9 June 2015
Available online 16 June 2015

Keywords:
Differential RNA-seq
Transcriptional start sites
Comparative transcriptomics
Small RNAs
Promoter motifs
Gene regulation
5′UTR

ABSTRACT

The global mapping of transcription boundaries is a key step in the elucidation of the full complement of transcriptional features of an organism. It facilitates the annotation of operons and untranslated regions as well as novel transcripts, including *cis*- and *trans*-encoded small RNAs (sRNAs). So called RNA sequencing (RNA-seq) based on deep sequencing of cDNAs has greatly facilitated transcript mapping with single nucleotide resolution. However, conventional RNA-seq approaches typically cannot distinguish between primary and processed transcripts. Here we describe the recently developed differential RNA-seq (dRNA-seq) approach, which facilitates the annotation of transcriptional start sites (TSS) based on deep sequencing of two differentially treated cDNA library pairs, with one library being enriched for primary transcripts. Using the human pathogen *Helicobacter pylori* as a model organism, we describe the application of dRNA-seq together with an automated TSS annotation approach for generation of a genome-wide TSS map in bacteria. Besides a description of transcriptome and regulatory features that can be identified by this approach, we discuss the impact of different library preparation protocols and sequencing platforms as well as manual and automated TSS annotation. Moreover, we have set up an easily accessible online browser for visualization of the *H. pylori* transcriptome data from this and our previous *H. pylori* dRNA-seq study.

1. Introduction

RNA-sequencing (RNA-seq) based on deep sequencing of cDNA libraries has been increasingly used as the method of choice for gene expression analysis and annotation of whole transcriptomes [1]. In comparison to hybridization-based techniques, such as microarrays or tiling arrays, RNA-seq has a higher dynamic range and requires less input material. Instead of applying previously designed probes that are prone to suffer from cross-hybridization issues, RNA-seq directly records the amount and boundaries of each transcript with single nucleotide (nt) resolution. The prior knowledge of the genomic sequence can facilitate the analysis and mapping of the sequenced cDNA reads, but is not necessarily required to detect and quantify a transcript. RNA-seq has greatly facilitated the annotation of transcript boundaries and the identification of novel transcripts in both pro- and eukaryotes [2–4]. While a major challenge for early bacterial RNA-seq experiments was the presence of highly abundant RNA species like rRNAs and tRNAs, which make up more than 95% of the RNA pool in a bacterial cell, this issue was overcome in eukaryotes by solely reverse-transcribing poly(A)-tailed mRNAs via oligo-d(T) priming during cDNA library preparation [4]. Since poly(A)-tails represent a degradation signal in bacteria, several strategies for rRNA removal including oligonucleotide-based removal of rRNAs with magnetic beads or size fractionation using gel electrophoresis (reviewed in [2,5]) were employed. The steadily dropping sequencing costs, bundled with a major increase in sequencing depth, nowadays provide sufficient coverage for the mRNA and non-abundant sRNA fractions without the necessity for additional depletion steps which were recently shown to introduce coverage bias [6].

In a typical RNA-seq experiment total RNA or a fraction thereof is first converted into cDNA in a reverse-transcription reaction, followed by PCR-based amplification of the library. Different library protocols are available, which are highly specific for the applied sequencing technique but can be subdivided into strand-specific and non-strand-specific protocols. Non-strand-specific protocols, for example, based on random hexamer priming and ligation of adapters to double-stranded cDNA have the drawback that they

* Corresponding author. Fax: +49 931 3182578.
E-mail address: cynthia.sharma@uni-wuerzburg.de (C.M. Sharma).

http://dx.doi.org/10.1016/j.ymeth.2015.06.012
1046-2023/

lose the information whether sequencing reads originate from the sense or the antisense strand. To overcome this problem, strand-specific protocols have been developed including direct sequencing of first strand cDNA [7], template switching PCR [8], RNA C to U conversion using bisulfite [9] or second strand synthesis with dUTP followed by degradation after adapter ligation [10]. Our below listed protocol combines 5′ end RNA linker ligation with poly(A)-tailing using *Escherichia coli* poly(A) polymerase [11–13]. After cDNA library construction and different quality checks, the samples are sequenced on one of the available deep sequencing platforms, resulting in millions of cDNA reads. The most commonly used techniques are the Illumina (Solexa), the 454 Life Sciences system, and ABI SOLiD sequencing. More recently developed single-molecule sequencing technologies comprise SMRT sequencing (Pacific Biosciences) or nanopore sequencing (Oxford Nanopore Technologies). Depending on the applied protocol and sequencing method, the reads are subjected to different pre-processing steps such as quality filtering or adapter or poly(A)-tail trimming. Afterwards, cDNA reads are commonly aligned to a genomic sequence and can then be used for gene expression profiling based on existing annotations, the generation of nucleotide-wise coverage plots for visualization in a genome browser and the annotation of novel transcripts.

RNA-seq-based mapping of bacterial transcript boundaries enables a global elucidation of operon structures and facilitates annotation of untranslated regions (UTRs) of protein coding genes, which potentially contain gene regulatory elements. Additionally, it allows for detection of novel transcripts such as small regulatory RNAs (sRNAs) and facilitates the discovery of previously non- or misannotated ORFs. Primer extension [14] or 5′ RACE (rapid amplification of cDNA ends) [15–17] are established methods for the determination of transcript 5′ ends of single genes, but they are time-consuming and impractical for global analysis. Therefore, several RNA-seq-based protocols for sequencing of 5′ ends of RNAs including a modified 5′ RACE approach have been developed, but many of them cannot clearly distinguish transcriptional start sites (TSS) from processing sites [18–23].

Here we give a detailed description of the differential RNA-seq (dRNA-seq) method, which allows for global annotation of all expressed TSS under the examined growth condition in an organism of interest in one sequencing experiment [24]. While it was originally developed to study the primary transcriptome of the major human pathogen *Helicobacter pylori* [12] it has since been successfully applied for determination of TSS in a wide range of pro- and eukaryotic organisms [24]. With >1900 unique TSS and at least one antisense TSS to 50% of all genes, the dRNA-seq approach revealed a very complex and compact transcriptional output from the small *H. pylori* genome and an unexpected number of >60 sRNAs [12]. While our previous *H. pylori* dRNA-seq approach was based on 454 sequencing of dRNA-seq libraries from *H. pylori* strain 26695 grown under different growth conditions, we here exemplify the use of Illumina-based dRNA-seq for annotation of TSS under a representative growth condition. We compare the results from the different sequencing platforms and among different replicates. Furthermore, we perform an automated TSS annotation using TSSpredator (http://it.inf.uni-tuebingen.de/TSSpredator), which we had initially applied for a comparative TSS annotation in multiple *Campylobacter jejuni* strains [25] and the generation of a global TSS map of *Escherichia coli* K12 MG1655 [26], and compare the automated TSS annotation with manual TSS annotations from the previous *H. pylori* dRNA-seq study [12]. We provide the global TSS maps and cDNA coverage plots of the previous and newly generated *H. pylori* 26695 dRNA-seq data in an easily accessible online browser (http://hpylori-tss.imib-zinf.net/).

2. Materials and methods

2.1. Helicobacter pylori growth conditions

Helicobacter pylori wild type strain 26695 (CSS-0065, kindly provided by D. Scott Merrell, Bethesda, MD) was grown on GC-agar (Oxoid) plates supplemented with 10% (V/V) donor horse serum (Biochrom AG), 1% (V/V) vitamin mix, 10 μg/ml vancomycin, 5 μg/ml trimethoprim, and 1 μg/ml nystatin as described previously [44]. For liquid cultures, 15 or 50 ml Brain Heart Infusion medium (BHI, Becton, Dickinson and Company) supplemented with 10% (V/V) FBS (Biochrom AG) and 10 μg/ml vancomycin, 5 μg/ml trimethoprim, and 1 μg/ml nystatin was inoculated with *H. pylori* grown on plates to a final OD_{600} of 0.02–0.05 and grown under agitation at 140 rpm in 25 cm^3 or 75 cm^3 cell culture flasks (Corning). Bacteria were grown at 37 °C in a HERAcell 150i incubator (Thermo scientific) in a microaerophilic environment (10% CO_2, 5% O_2, and 85% N_2). When the cultures reached mid-log phase (OD_{600} ~0.6), culture volumes of cells corresponding to a total amount of 4 OD_{600} were mixed with 0.2 volumes of stop-mix (95% EtOH and 5% phenol, V/V), frozen in liquid N_2 and stored at −80 °C until RNA extraction. In total, three biological replicates of bacteria grown to mid-log phase (ML) were harvested: B1, which was grown separately from B2 and B3, which were grown on the same day.

2.2. RNA extraction and DNase I treatment

Frozen cell pellets were thawed on ice and resuspended in lysis solution containing 600 μl of 0.5 mg/ml lysozyme in TE buffer (pH 8.0) and 60 μl 10% SDS. Bacterial cells were lysed by incubating the samples for 1–2 min at 65 °C. Afterwards, total RNA was extracted using the hot-phenol method as described previously [12,27]. DNase I (Fermentas) treatment was performed on total RNA according to manufacturer's instruction. Removal of residual genomic DNA was subsequently verified by control PCR using the oligos CSO-0790: GTTTTTTCTAGACGTTTAAAACAAGCCTGGT and CSO-0791: GTTTTTGAATTCCATGATGACTCCTTTAATTGAAA which amplify a ~594 nt long product of the HP1432 gene.

2.3. dRNA-seq library preparation and sequencing

Terminator exonuclease (TEX) treatment of RNA samples was performed as previously described [12]. cDNA libraries for Illumina sequencing were constructed by Vertis Biotechnology AG, Germany (http://www.vertis-biotech.com/) in a strand-specific manner as previously described for eukaryotic microRNA [28] but omitting the RNA size-fractionation step prior to cDNA synthesis. In brief, ~200 ng of RNA sample were poly(A)-tailed using 2.5 U *E. coli* poly(A) polymerase (NEB) for 5 min at 37 °C. TEX treatment (+TEX) and mock treatment without the enzyme (−TEX) were carried out after poly(A)-tailing. To this end, poly(A)-tailed RNA was denatured for 2 min at 90 °C, cooled on ice for 5 min and treated with 1.5 U of Terminator Exonuclease (Epicentre) for 30 min at 30 °C. Then, the 5′-PPP structures were removed using tobacco acid pyrophosphatase (TAP). TAP treatment was performed by incubating +TEX and −TEX samples with 5 U TAP for 15 min at 37 °C.

Afterwards, an RNA adapter (5′ Illumina sequencing adapter, 5′-UUUCCCUACACGACGCUCUUCCGAUCU-3′) was ligated to the 5′-P of the TAP-treated, poly(A)-tailed RNA for 30 min at 25 °C. First strand cDNA was synthesized by using an oligo(dT)-adapter primer (see below) and the M-MLV reverse transcriptase (AffinityScript, Agilent) by incubation at 42 °C for 20 min, ramp to 55 °C followed by 55 °C for 5 min. In a PCR-based amplification

step using a high fidelity DNA polymerase (Herculase II Fusion DNA Polymerases, Agilent) the cDNA concentration was increased to 10–20 ng/μl (initial denaturation at 95 °C for 2 min, 16–18 cycles 95 °C for 20 s and 68 °C for 2 min). A library-specific barcode for multiplex sequencing was included as part of a 3′ sequencing adapter. The TruSeq index primers for PCR amplification were used according to the instructions of Illumina. For all libraries the Agencourt AMPure XP kit (Beckman Coulter Genomics) was used to purify the DNA (1.8 × sample volume), and cDNA sizes were examined by capillary electrophoresis on a MultiNA microchip electrophoresis system (Shimadzu).

The following adapter sequences flank the cDNA inserts:

TruSeq_Sense_primer: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′, TruSeq_Antisense_NNNNNN_primer (NNNNNN = 6n barcode for multiplexing): 5′-CAAGCAGAAGACGGCATACGAGAT–NNNNNN-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC(dT25)-3′

The first biological replicate (B1) was sequenced on an Illumina HiSeq 2000 machine with 97 cycles while the second and third replicate (B2/B3) were sequenced on a HiSeq 2500 with 100 cycles. All sequencing was conducted in single-read mode.

2.4. Analysis of deep sequencing data

2.4.1. Read mapping and generation of coverage plots

To assure high sequence quality, the Illumina reads in FASTQ format were trimmed with a cutoff phred score of 20 by the program fastq_quality_trimmer from FASTX toolkit version 0.0.13. After trimming, poly(A)-tail sequences were removed and a size filtering step was applied in which sequences shorter than 12 nt were eliminated. The collections of remaining reads were mapped to the *H. pylori* 26695 (NCBI Acc.-No: NC_000915.1) genome using the RNA-seq pipeline READemption [29] and *segemehl* [30] with an accuracy cutoff of 95%. Coverage plots representing the numbers of mapped reads per nucleotide were generated. Reads that mapped to multiple locations with an equal score contributed a fraction to the coverage value. For example, reads mapping to three positions contributed only 1/3 to the coverage values. We chose this approach of including reads that map to multiple locations with relative scores rather than solely using uniquely mapped reads. It represents a tradeoff between introducing some uncertainty regarding the true origin of reads that map to multiple locations and not excluding all transcripts with true multiple copies in the genome like rRNAs, tRNAs and some of the sRNAs. Since only read mappings with equal scores were considered, most of the non-uniquely mapped reads likely corresponded to such duplicated or repetitive genes, rather than representing unspecifically mapped reads. Each resulting cDNA coverage graph was normalized by the number of reads that could be mapped from the respective library (typically several million reads when using Illumina sequencing) and afterwards multiplied by 1,000,000.

2.4.2. Coverage plot normalization by TSSpredator

Prior to the comparative analysis, the expression graphs with the cDNA coverages that resulted from the read mapping were further normalized using TSSpredator (http://it.inf.uni-tuebingen.de/TSSpredator). A percentile normalization step was applied to normalize the +TEX graphs. To this end, the 90th percentile of all data values was calculated for each +TEX graph. This value was then used to normalize the +TEX graph as well as the respective −TEX graph. Thus, the relative differences between each +TEX and −TEX graph were not changed in this normalization step. Again, afterwards all graphs were multiplied with the overall lowest value to restore the original data range. To account for different enrichment rates, a third normalization step was applied. During this step, prediction of TSS candidates was performed for each replicate. These candidates were then used to determine the median enrichment factor for each ±TEX library pair. Using these medians all −TEX libraries were then normalized against the library with the strongest enrichment. Besides annotation of TSS, the resulting graphs were also used for visualization in the Integrated Genome Browser [31].

2.4.3. Automated TSS annotation using TSSpredator

Based on the normalized expression graphs automated TSS prediction was performed similar to Thomason and Bischler et al. [26] and Dugar et al. [25] using TSSpredator. In brief, for each position *(i)* in the expression graph corresponding to the +TEX libraries, the algorithm calculates an expression height, $e(i)$, and compares that expression height to the preceding position by calculating $e(i)-e(i-1)$, which is termed the flank height. Additionally, the algorithm calculates a factor of height change $e(i)/e(i-1)$. To determine if a TSS is a real TSS and not a processed transcript end an enrichment factor is calculated as $e_{+\mathrm{TEX}}(i)/e_{-\mathrm{TEX}}(i)$, where $e_{+\mathrm{TEX}}(i)$ is the expression height for the TEX-treated sample and $e_{-\mathrm{TEX}}(i)$ is the expression height for the untreated sample. For all positions where these parameters (flank height, factor of height change, and enrichment) exceed the predefined thresholds a TSS is annotated.

We set the thresholds for the *minimum flank height* and the *minimum factor of height change*, which are used to determine if a TSS is detected to 0.3 and 2.0, respectively. Here, the value for the *minimum flank height* is a factor to the minimum 90th percentile over all libraries resulting in an absolute value of 2.94 (for predictions based on 4 replicates). If the TSS candidate reaches these thresholds in at least one replicate, the thresholds are decreased for the other replicates to 0.1 (0.98 absolute) and 1.5, respectively. Furthermore, we set the *matching replicates* parameter, which determines the number of replicates in which a TSS must exceed these thresholds in order to be marked as detected to 3. A TSS candidate is considered to be enriched, if the enrichment factor at the respective nucleotide position for at least one replicate is $\geqslant 2.0$. In order to take into account slight variations between TSS positions the respective parameter for clustering between replicates was set to a value of 1. In doing so, a consensus TSS position in a 3 nt window is determined based on the maximum flank height among the respective libraries.

Predicted TSS were assigned to five different classes based on their location with respect to predefined annotations: primary TSS (pTSS, main TSS within 300 nt upstream of a gene or operon), secondary TSS (sTSS, alternative TSS with lower flank height), internal TSS (iTSS, TSS within a gene), antisense TSS (asTSS, TSS antisense to a gene in a distance $\leqslant$100 nt), and orphan TSS (oTSS, TSS not associated with annotation). Please note that compared to our previous manually annotated TSS used in [12], we reduced the maximal window for pTSS and sTSS classification from 500 nt to 300 nt to have a more strict TSS classification. This might affect some of the classifications of previously annotated TSS, i.e. TSS $\leqslant$500 and >300 nt upstream of annotated genes. For example, some of the TSS that are also classified as iTSS or asTSS might have lost the primary or secondary classification whereas TSS solely classified as pTSS or sTSS would be annotated as oTSS. Moreover, our automated TSS prediction and classification employs an updated annotation file, which now also contains the annotations for validated sRNAs from *H. pylori*. Thus, these are now also listed with their primary TSS in Table S1.

2.4.4. Availability of sequencing data

Raw sequencing reads in FASTQ format and coverage files normalized by TSSpredator in wiggle (WIG) format are available via Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo) under accession number GSE67564.

Additionally, we used previous 454 sequencing data of a TEX-treated (+TEX) and an untreated library (−TEX) based on a sample collected at mid-log growth (B0) from the previous *H. pylori* dRNA-seq study [12] for which raw data were previously uploaded to the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra) under accession number SRA010186.

3. Results and discussion

3.1. The dRNA-seq approach for global mapping of TSS

The dRNA-seq approach allows for the precise mapping of TSS on a genome-wide scale via selective sequencing of primary transcripts [24]. For each biological sample, a cDNA library pair consisting of one library (+TEX) generated from RNA treated with terminator 5′ phosphate dependent exonuclease (TEX) and a second library (−TEX) generated from untreated total RNA is sequenced. TEX selectively digests processed transcripts with a 5′-P which results in an enrichment for primary transcripts that still carry a 5′-PPP in the +TEX library [12]. Fig. 1A depicts how TEX treatment of total RNA eliminates most of the processed RNAs including the abundant 16S and 23S rRNA. Another method that relies on initial TEX-treatment for depletion of processed transcripts employs a modified 5′ RACE approach [21,22]. However, compared to the dRNA-seq approach this approach does not include a direct comparison to an untreated library based on the same sample, which facilitates the discrimination of primary and processed transcripts.

An alternative strategy to identify TSS on a global scale is based on treatment of RNA with tobacco acid pyrophosphatase (TAP) and has been used for global identification of sRNAs and their TSS in *Clostridium difficile* [32] or the generation of a transcriptome map and analysis of pervasive transcription in *Propionibacterium acnes*

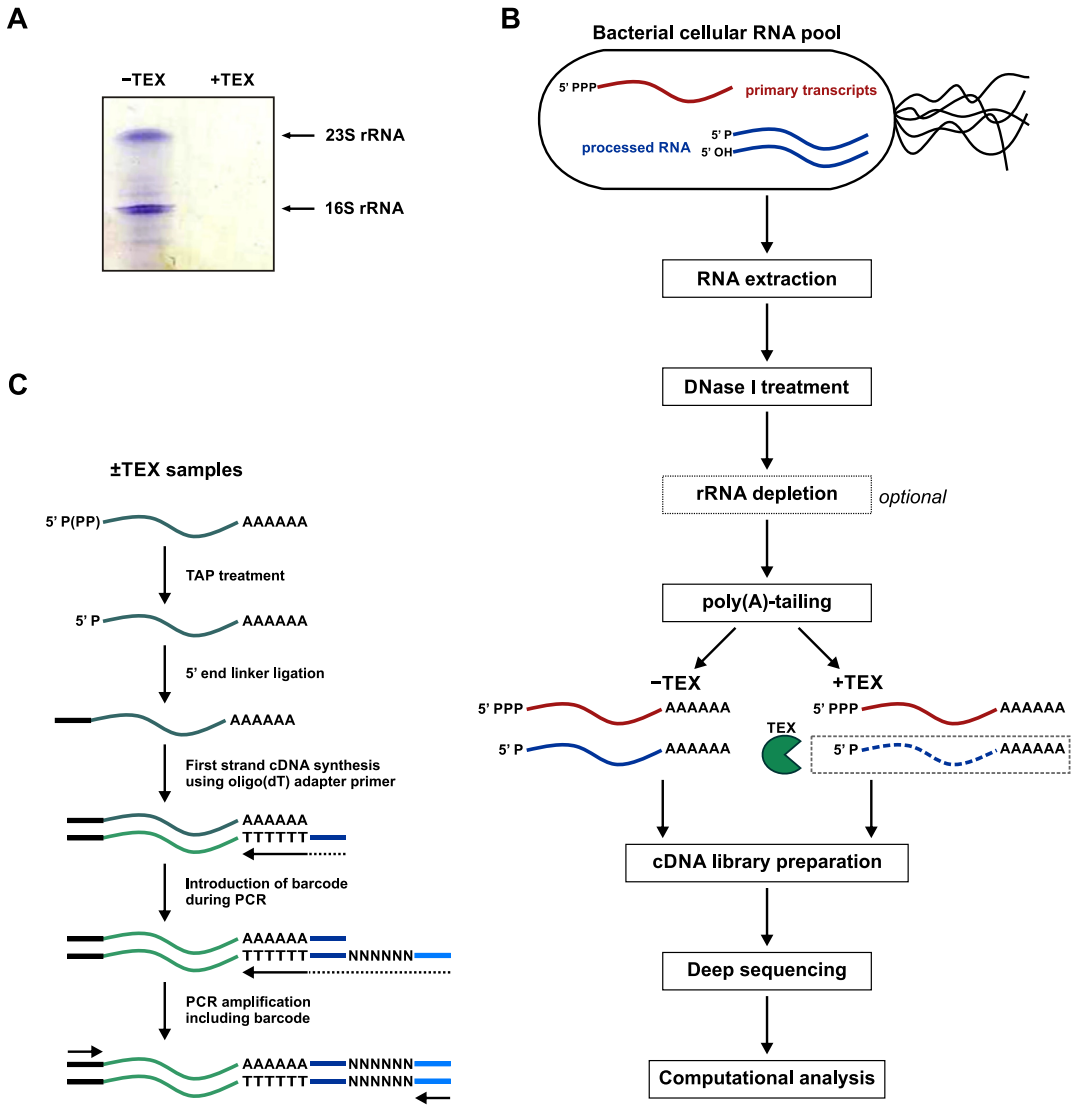


Fig. 1. Workflow for dRNA-seq-based primary transcriptome analysis. (A) *H. pylori* 26695 total RNA harvested at OD_{600} 0.6 with (+) and without (−) TEX treatment was separated on a 4% 7 M Urea polyacrylamide gel and stained with Stains-All (Sigma–Aldrich). Positions of bands for 16S and 23S rRNA are indicated on the right. (B) Representative workflow of a dRNA-seq experiment. (C) Illumina sequencing-specific cDNA library preparation protocol applied to both, +TEX and −TEX samples.

[33] as well as *Streptomyces coelicolor* and *Escherichia coli* [47]. TAP removes pyrophosphates from the 5′-PPP group of primary transcripts leaving a 5′-P end and making them accessible for 5′ end linker ligation. Comparison of library pairs generated from RNA with (+TAP) and without (−TAP) TAP treatment enables determination of TSS based on enrichment of primary transcripts in the +TAP versus the −TAP library. However, in contrast to the dRNA-seq approach this approach does specifically enrich for primary transcripts and does not deplete the abundant rRNAs and tRNAs so that deeper sequencing coverage might be required. A similar strategy was applied for TSS mapping in *E. coli* using 5′ polyphosphatase instead of TAP [23].

3.2. dRNA-seq of Helicobacter pylori strain 26695

Here, we describe the application of dRNA-seq using the gastric pathogen *H. pylori* 26695 as an example bacterium. *H. pylori* thrives in the acidic environment of the human stomach where it can cause gastritis, ulcers, gastric cancer and lead to lifelong, persistent infections [34,35]. *H. pylori* has a relatively small genome of 1.67 Mbp and encodes only for a small number of transcriptional regulators. The strain 26695 was originally isolated from a gastritis patient in the United Kingdom and was one of the first bacteria with a sequenced genome [36]. Strain 26695 is one of the most widely used strains in *H. pylori* research and a genome-wide map of TSS and operons based on dRNA-seq data was previously generated for this strain grown under five different biological conditions [12]. The conditions comprised bacteria grown to mid-logarithmic phase (ML), which represented the reference growth condition, or under acid stress (AS), grown in contact with responsive gastric epithelial AGS cells (AG) or non-responsive liver cells (HU), or in cell culture medium alone (PL). For each of these conditions, a single library was constructed and between 200,000 and 500,000 cDNA reads for each sample were generated by 454 pyrosequencing.

For every RNA-seq experiment, one important decision to make is the selection of an appropriate sequencing technology. Several platforms with differences in read length, reads per run, accuracy, price and time per run [37] are available on the market. Here, we will focus on protocols and data analyses that apply to the Illumina sequencing technology, which is currently the most widely-used platform for RNA-seq. To illustrate the necessary steps for generation of a TSS map and to assess the effects of the deeper Illumina sequencing coverage and the use of several replicates in comparison to the previous TSS annotations, we collected three biological replicate RNA samples (B1–B3) from *H. pylori* 26695 wild-type cells grown to mid-logarithmic phase in rich BHI medium +10% FCS. The protocol used to generate dRNA-seq data from these samples is shown in Fig. 1B and details are listed in the Materials and Methods section. After collection of cell samples, total RNA was isolated using the hot-phenol extraction [12,25,27] (see Section 2.2). It is crucial to obtain high-quality RNA in this step to avoid extensive sequencing of rRNA degradation fragments. Thus, an RNA quality check on agarose gels or using Bioanalyzer chips is recommended after removal of residual genomic DNA via DNase I treatment (see Section 2.2). An additional rRNA depletion is optional and is not necessary in most cases since sequencing coverage is no longer limiting. Due to the removal of processed RNAs, TEX treatment also decreases the fraction of rRNAs and tRNAs. Thus, together with the additional depletory effects due to lower preference of poly(A) addition by *E. coli* poly(A) polymerase (PAP I) described below and lower efficiency in reverse transcription for structured rRNAs during library construction, no additional rRNA depletion steps are required in a typical dRNA-seq experiment.

For the preparation of dRNA-seq libraries, either the ±TEX treatment can be the first step or each RNA sample can be first polyadenlyated using PAP I, followed by differential TEX treatment. Here we describe the latter order (Fig. 1B and C), which has the advantage that it ensures equal poly(A)-tailing for the corresponding ±TEX library pairs. The cDNA libraries for Illumina sequencing were generated in the same way for +TEX and −TEX samples and experimental details are given in Section 2.3. Strand-specificity of the sequencing is crucial to distinguish sense from antisense transcripts. In our method this is achieved by attaching a 5′ RNA adapter and a poly(A)-tail to each fragment prior to cDNA synthesis. First a poly(A)-tail was attached to the RNA molecules. It was shown that PAP I has a preference of polyadenylating mRNAs over rRNAs resulting in an inherent rRNA depletion in the resulting cDNA library [38]. Afterwards, each of the poly(A)-tailed biological replicates B1-B3 was split into two halves which were then differentially treated with TEX, resulting in −TEX samples covering RNAs with a 5′-P and a 5′-PPP and +TEX samples that are enriched for 5′-PPP RNAs. Next, the ±TEX samples were treated with TAP to cleave the 5′-PPP groups of primary transcripts leaving a 5′-P. This step is necessary to enable subsequent ligation of the 5′ end linkers that cannot be ligated to a 5′-PPP end. Please note that processed transcripts with a 5′-OH are not covered in the final cDNA libraries, although they are resistant to TEX removal, since they are not accessible for 5′ end RNA linker ligation. In case one is interested in capturing this class of transcripts, an additional treatment with polynucleotide kinase and ATP is required to generate 5′-P ends (for a protocol see [39]). After TAP treatment, an RNA linker was ligated to the transcripts in the ±TEX samples. Next, first strand cDNA was generated using an oligo(dT)-adapter primer and library-specific barcodes were introduced during PCR amplification of each library. All libraries were sequenced on either an Illumina 2000 (B1-HS1) or 2500 machine (B2-HS2 and B3-HS2). In total we sequenced between 4.1 and 8.1 Mio cDNA reads per library (Table 1). This represents a more than 10-fold higher coverage compared to the previous 454 libraries [12].

It was shown that the construction of cDNA libraries can be a major source of variation among RNA-seq experiments based on the same organism in both pro- and eukaryotes [26,40]. Especially, additional bias might be introduced by distinct library preparation protocols for different sequencing platforms due to differences in ligation efficiency and RNA structure or G/C-content-dependent differences in reverse transcription or PCR amplification efficacy [41]. The resulting variation in amplification of certain transcripts could be an explanation for observed differences among distinct studies of the same organism [40]. When comparing biological and technical replicates in a dRNA-seq analysis of *E. coli*, we observed larger variation for distinct library preparations from the same biological sample than among biological replicates for which the libraries were generated in parallel [26]. We therefore recommend, if possible, conducting cDNA library preparation for all samples simultaneously. This is even more important for quantitative gene-expression profiling experiments compared to qualitative transcriptome annotation approaches such as dRNA-seq.

3.3. dRNA-seq data analysis

After Illumina sequencing of the six B1-B3 ±TEX libraries, we assessed read quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). This software provides for each library a summary, which includes different quality metrics as, for example, sequence length distribution, GC content distribution, presence of duplicated or overrepresented sequences, per-base N content and most importantly base call quality scores. Read quality for Illumina sequencing typically decreases at the 3′ end of longer

Table 1
Mapping statistics for the *H. pylori* 26695 Illumina dRNA-seq libraries. This table summarizes the total number of sequenced cDNA reads after quality trimming, as well as the number of mapped and uniquely mapped reads for each library. Percentage values are relative to the number of cDNA reads that are >11 nt after poly(A) trimming.

Library	Total number of reads after quality trimming	Number of reads long enough after poly(A) trimming (>11 nt)	Mapped reads	% Mapped reads	Uniquely mapped reads	% Uniquely mapped reads
ML B1-HS1 + TEX	4,105,444	2,904,136	2,855,756	98.3	1,776,256	61.2
ML B1-HS1 – TEX	4,709,180	4,393,218	4,303,527	98.0	2,191,371	49.9
ML B2-HS2 + TEX	6,979,343	6,747,193	6,694,900	99.2	4,121,103	61.1
ML B2-HS2 – TEX	8,128,096	8,051,963	8,008,388	99.5	4,011,524	49.8
ML B3-HS2 + TEX	6,700,169	6,402,059	6,351,658	99.2	3,952,168	61.7
ML B3-HS2 – TEX	7,435,053	7,344,125	7,302,057	99.4	4,160,876	56.7

reads. Therefore, preprocessing of reads is important to facilitate alignment to the reference genome. In our pipeline (Fig. 2A) we conducted quality trimming from the 3′ end, poly(A) trimming, and size filtering in order to generate a set of high quality reads which were afterwards mapped to the *H. pylori* 26695 reference genome (NC_000915.1). For all steps starting from poly(A) trimming until coverage plot generation we used the RNA-seq analysis tool READemption [29].

To examine the percentage of reads mapped to individual RNA classes, we calculated the number of reads that overlapped for at least 10 nt in either sense or antisense direction, annotations for 5′UTRs, mRNAs, sRNAs, rRNAs, tRNAs and housekeeping RNAs (RNase P RNA, SRP RNA, tmRNA and 6S RNA) based on the previously generated *H. pylori* transcriptome map (Table 2) [12]. The amount of reads mapping to rRNA ranged between 48 and 55% for the –TEX libraries and between 35 and 38% for the +TEX libraries, indicating that TEX depletes these processed transcripts. Moreover, even in the –TEX libraries the observed rRNA fraction is lower than the expected 90–95% for abundant rRNAs which might be caused by multiple factors: (i) the poly(A)-tailing with lower preference for rRNAs during library construction mentioned above, (ii) the fact that no RNA fragmentation was conducted prior to cDNA synthesis, which would result in a large amount of rRNA fragments and (iii) the lower efficiency of reverse transcription of structured RNA. This further shows that an additional rRNA depletion step is not necessarily required in our protocol. Moreover, a clear enrichment (at least 2-fold) of the fractions of reads mapping to sRNAs as well as 5′UTRs is observed in the +TEX libraries compared to the respective –TEX libraries, showing a successful enrichment of primary transcripts and the 5′ ends of transcripts. Please note that sequencing initiates from the 5′ adapter, which further enriches for the 5′ ends of transcripts.

Based on the read mappings we computed per-strand coverage plots for each library (for details see Section 2.4) that indicate the number of mapped reads per nucleotide. In case a read mapped with the same score to multiple regions in the genome, only a corresponding fraction, e.g. a score of 0.5 reads in case of two equal mappings, was counted for the respective positions. The resulting cDNA coverage plots allow examination of the transcriptome in a genome browser, e.g., the Integrated Genome Browser (IGB) [31], with single nucleotide resolution. The visualized RNA-seq data can then be used for annotation of transcript boundaries or novel transcripts such as sRNAs.

3.4. Identification of TSS based on dRNA-seq

The differential RNA-seq approach leads to a characteristic cDNA coverage pattern of dRNA-seq library pairs at TSS. The

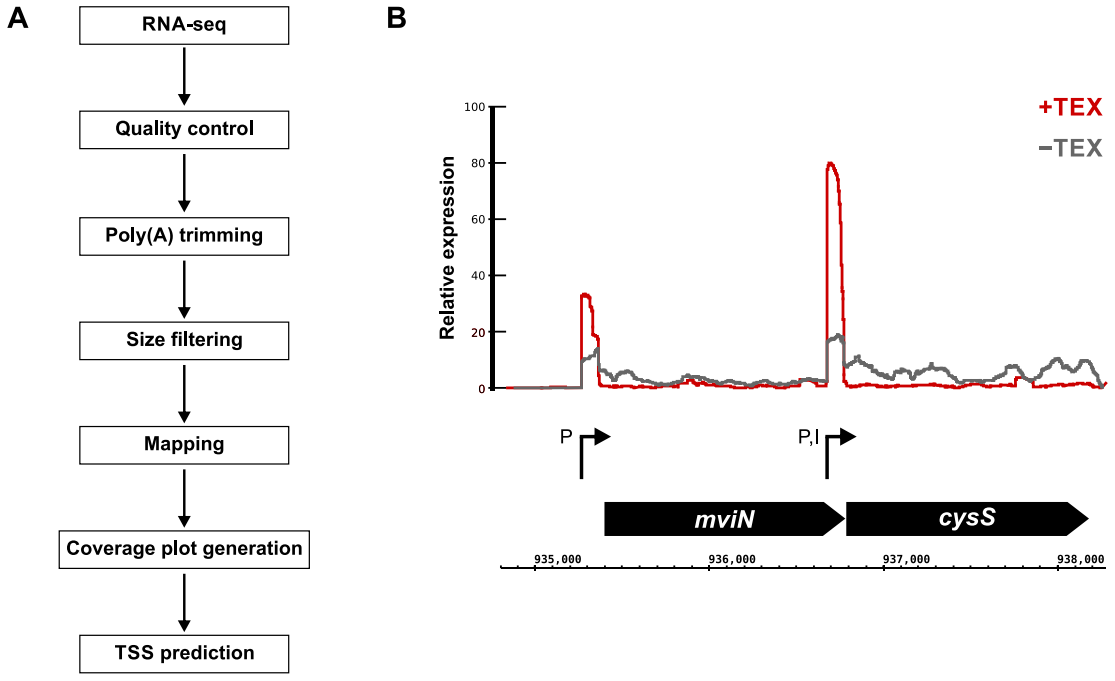


Fig. 2. Computational dRNA-seq analysis and TSS enrichment. (A) Workflow of the dRNA-seq data analysis pipeline. (B) Illustration of a representative cDNA enrichment pattern in the +TEX versus –TEX library at a TSS located upstream of the *mviN* gene and at a TSS internal to *mviN* and upstream of the *cysS* gene.

Table 2
Mapping statistics of cDNA reads based on strand and type of RNA class. This table indicates the number of cDNA reads that were mapped to the different RNA classes (5′UTR, mRNA, sRNA, rRNA, tRNA and housekeeping RNA) for each library. The numbers for the mapped reads per RNA class are shown with percentage values calculated from the total number of mapped reads regardless of mapped location (taken from Table 1) for the respective library. Housekeeping RNAs are RNase P RNA, SRP RNA, tmRNA and 6S RNA.

		Illumina library											
		ML B1 HS1 + TEX		ML B1 HS1 – TEX		ML B2 HS2 + TEX		ML B2-HS2 – TEX		ML B3 HS2 + TEX		ML B3 HS2 – TEX	
	Total*	2,855,756		4,303,527		6,694,900		8,008,388		6,351,658		7,302,057	
Sense	5′UTR	163,535	(6%)	119,712	(3%)	843,181	(13%)	336,795	(4%)	863,048	(14%)	349,806	(5%)
	mRNA	297,647	(10%)	895,532	(21%)	1,109,869	(17%)	2,467,397	(31%)	1,072,325	(17%)	2,641,220	(36%)
	sRNA	596,772	(21%)	255,731	(6%)	1,386,141	(21%)	210,665	(3%)	1,293,252	(20%)	196,253	(3%)
	rRNA	1,087,360	(38%)	2,205,744	(51%)	2,426,360	(36%)	4,382,063	(55%)	2,246,962	(35%)	3,497,798	(48%)
	tRNA	351,622	(12%)	428,298	(10%)	155,451	(2%)	76,466	(1%)	144,869	(2%)	75,009	(1%)
	Housekeeping RNA	113,695	(4%)	125,315	(3%)	201,105	(3%)	94,001	(1%)	183,972	(3%)	93,288	(1%)
Antisense	5′UTR	6,553	(0%)	4,650	(0%)	27,975	(0%)	4,052	(0%)	24,387	(0%)	4,551	(0%)
	mRNA	120,994	(4%)	125,865	(3%)	246,790	(4%)	119,946	(1%)	243,755	(4%)	130,290	(2%)
	sRNA	81,927	(3%)	39,202	(1%)	169,463	(3%)	49,897	(1%)	152,405	(2%)	43,984	(1%)
	rRNA	181	(0%)	204	(0%)	312	(0%)	54	(0%)	399	(0%)	57	(0%)
	tRNA	202	(0%)	401	(0%)	376	(0%)	693	(0%)	432	(0%)	765	(0%)
	Housekeeping RNA	114	(0%)	191	(0%)	200	(0%)	80	(0%)	308	(0%)	103	(0%)

* Total reads for each library also include reads that mapped to locations other than the listed RNA classes.

specific enrichment pattern (Fig. 2B) of the +TEX library compared to the corresponding −TEX library [12] indicates the start of a primary transcript and can thus be used to annotate TSS. Based on global examination of these enrichment patterns it is possible either to conduct manual TSS annotation based on visual inspection of the coverage plots in a genome browser, such as the IGB, or to use a tool for automated TSS annotation based on dRNA-seq data. In the previous dRNA-seq analysis *of H. pylori* 26695, we manually annotated 1907 unique TSS based on visual inspection of the enrichment patterns among the five examined growth conditions. Manual TSS annotation is laborious and time-consuming, especially when applied to large genomes or the comparative analysis of multiple strains or conditions including biological replicates. In comparison to automated TSS prediction, which follows defined rules, it is also likely to introduce human bias, based on individual perception of the data and thus might lead to different results for unclear cases. However, manual inspection can be useful in single cases to either confirm automated predictions or check for TSS patterns that were misinterpreted based on predefined parameter thresholds.

Multiple groups have in the meanwhile developed diverse TSS prediction tools, which make use of the information provided by dRNA-seq [42,43]. Here, we computationally annotated TSS in the three B1-B3 dRNA-seq data sets of *H. pylori* 26695 grown to mid-log phase utilizing the software TSSpredator (http://it.inf.uni-tuebingen.de/TSSpredator). We had originally applied this tool for a comparative TSS annotation in a dRNA-seq analysis of multiple *C. jejuni* strains [25], and also successfully applied it for TSS predictions among different growth conditions in *E. coli* K12 MG1655 [26]. To our knowledge, it is the most flexible of all currently available programs, with implemented support for comparative analysis and varying numbers of replicates. While other tools incorporate elaborate statistics to decide which genomic positions represent a TSS, TSSpredator applies specific heuristics to imitate manual TSS annotation with a set of tunable parameters. To ensure comparability between replicates, TSSpredator conducts additional normalization steps on the expression graphs of both, +TEX and −TEX libraries. Afterwards, each genomic position is checked for the presence of a potential TSS by assessing flank height and factor of height change in the +TEX libraries as well as enrichment between +TEX and −TEX libraries. When run comparatively, a TSS is annotated if it is detected and enriched in at least one strain or condition and in case multiple replicates are available the *matching replicates* parameter can be adjusted to determine the number of replicates in which a TSS must be detected but enrichment is only required in one of them. Here, we used the default settings of TSSpredator, which were established based on our manual annotation in *H. pylori* 26695 [12] and already applied in our previous studies [25,26].

In order to compare the TSS prediction based solely on the new Illumina mid-log dRNA-seq libraries to the manual annotations (1907 TSS) based on 454 data from the initial study [12], we ran TSSpredator using the three Illumina data sets as replicates. Requiring detection of a TSS in all replicates (*matching replicates* = 3), we predicted 1949 TSS. A comparison of these TSS positions with the 1907 manual TSS annotations requiring a precise match (cutoff 0 nt) resulted in 971 matching positions. The same comparison allowing for a maximum distance of three nt revealed an overlap of 1208 positions. This difference might be due to slight fluctuations in the actual TSS position for some promoters where transcription initiation is wobbly and the coverage shows a staircase-like pattern. In these cases, annotation of the major TSS is not always straightforward and slight variations in the libraries can lead to the annotation of neighboring positions. For this reason, we decided to tolerate such slight variations for this as well as subsequent comparisons reported below. The 1208 matching positions represent ~62% of our current predictions and ~63% of the previously annotated TSS (Fig. 3A). The additional 741 TSS predicted based on our current data are in most cases a result of the deeper coverage gained by Illumina sequencing and the support by several replicates for the mid-log growth condition. Previous TSS positions that are not detected in the Illumina dRNA-seq libraries are mainly caused by absence of or very low expression in at least one of the three Illumina replicates. In these cases, the respective TSS commonly shows a signal in one or more of the four other conditions assessed in the previous 454 study and was thus annotated. For example, in Fig. 3B the two TSS upstream of the HP0531 gene were annotated with matching positions in both, the previous manual annotation and the current TSSpredator prediction. The TSS within the HP0531 gene was only annotated by TSSpredator because there was no clear enrichment in the 454 data. In contrast, the TSS internal to HP0532 was only annotated in the 454 data as it was mainly expressed in the AS and HU conditions, but only very lowly expressed in the ML condition.

Furthermore, we noted some overall cDNA coverage variations, even within the same growth conditions, as observed for example between the B1-HS1 replicate and the B2/B3-HS2 replicates in Fig. 3B, which might be due to variations during library preparation. While such variations could be problematic for monitoring gene expression, especially of lowly expressed genes, when using

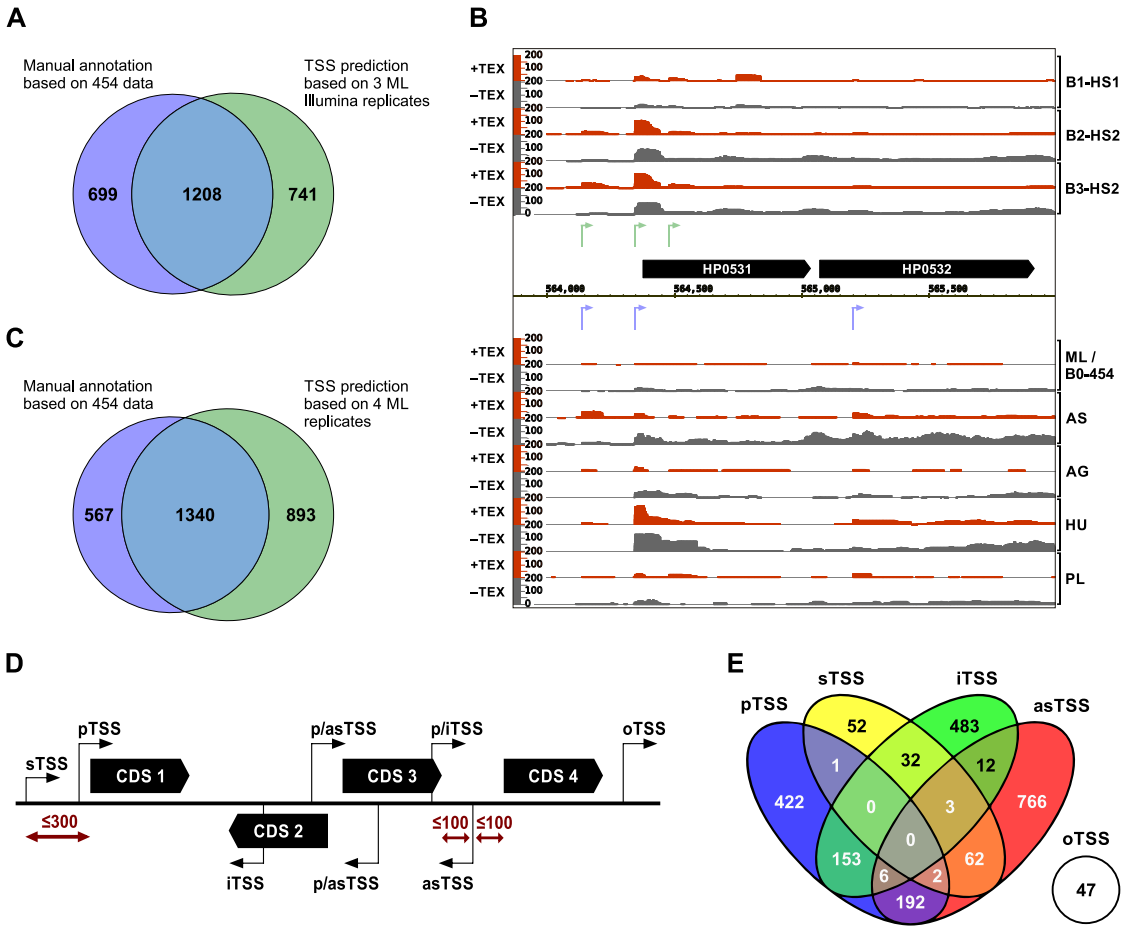


Fig. 3. TSS predictions in *H. pylori* 26695. (A) Comparison of manual TSS annotations from a previous study based on 454 sequencing of five different growth conditions to current TSSpredator predictions based on three biological replicates for the ML condition. (B) Example region encompassing the HP0531 and HP0532 genes, encoding the *cag* pathogenicity island proteins Cag11 and Cag12, respectively. cDNA coverage plots for the three Illumina ML replicates of our current data set are shown at the top with predicted TSS colored in green, while coverage plots for the five growth conditions from the previous 454 study [12] are shown at the bottom with manually annotated TSS colored in blue. Conditions or replicate identifiers are shown on the right (ML: mid-log growth; AS: acid stress; AG: *H. pylori* grown in the presence of AGS gastric cells; HU: *H. pylori* grown in the presence of Huh7 liver cells; PL: *H. pylori* in cell culture medium), while presence or absence of TEX treatment is indicated on the left. The *x*-axis reflects genomic positions while the *y*-axis indicates relative expression based on normalized read coverage. (C) Comparison of manual TSS annotations from the previous study based on 454 sequencing of five different growth conditions to current TSSpredator predictions based on four biological replicates (454 and Illumina) for the ML condition. (D) The location relative to annotated genes is depicted for the five different TSS classes (primary, secondary, internal, antisense, and orphan). The height of the black arrows indicates differences in expression strength while the distance cutoffs for flanking genes are shown in red. (E) The distribution according to TSS classes is depicted for the 2233 TSS predicted based on four ML replicates.

these data sets as replicates, this variability should not impede a qualitative dRNA-seq-based 5′ end mapping based on one or more conditions, since it does not affect the position of a TSS to be annotated. For measuring gene expression changes between different conditions, we recommend an approach that includes RNA fragmentation to cover full-length transcripts in combination with sample and library preparation in one experiment to reduce biological and technical variation among samples.

3.5. Comparison of *H. pylori* 454 and Illumina dRNA-seq data

To further examine the overlap between the old and new mid-log libraries and to investigate potential variation due to different sequencing platforms and library preparations, we complemented our three newly sequenced replicates with coverage plots for the ML condition based on 454 sequencing from [12] as an additional replicate (B0) and performed comparative TSSpredator predictions treating the four ML replicates as conditions. Using this setup, the resulting set of TSS encompassed 3240 distinct positions that were detected and enriched in at least one replicate and 1122 TSS which were found in all four (Fig. S1). A very good overlap (2211 TSS) with very few unique TSS positions for each library was observed between the B2-HS2 and B3-HS2 replicates, which were grown on the same day and for which library preparation was performed together, indicating that a careful and similar sample treatment is important to minimize variations. The second best overlap was observed between these two and the B1-HS1 replicate (1767 TSS), suggesting that slight differences in cultivation and potential biases introduced during library preparation can lead to differences in TSS expression and detection. The B1-HS1 and B0-454 replicates both introduced a similar amount of uniquely detected positions (392 and 329, respectively). This indicates again that not only differences in applied sequencing technologies and depth can play a role but also other experimental or technical

variation like differences in treatment and library preparation. In order to generate a comprehensive and reliable TSS map for the ML condition, we repeated the TSS prediction, now treating the 4 ML libraries as replicates. As a tradeoff between reliability and tolerance of data variation, we set the *matching replicates* parameter to a value of 3 (for details see Section 2.4). In total, we predicted 2233 TSS that were detected in at least three and enriched in at least one of the 4 ML replicates (Table S1). Out of these 2233 TSS found in ML growth, 1340 TSS were found in the set of our 1907 manually annotated TSS [12] (Fig. 3C).

TSSpredator automatically assigns TSS to five different classes according to their location in relation to annotated genes: primary TSS (pTSS), secondary TSS (sTSS), internal TSS (iTSS), antisense TSS (asTSS), and orphan TSS (oTSS) (for details see Section 2.4 and Fig. 3D). Notably, one TSS can independently be assigned to more than one category as, for example, in the presence of alternative suboperons the pTSS of the downstream gene can also be internal to the upstream gene (Figs. 2B and 3D). Similarly, in the case of overlapping 5′UTRs, the associated TSS can be both, a pTSS and an asTSS. Among the 2233 mid-log TSS, we identified 776 pTSS (422 classified only as pTSS), 152 sTSS (52 classified only as sTSS), 689 iTSS (483 classified only as iTSS), 1043 asTSS (766 classified only as asTSS) and 47 oTSS (Fig. 3E). This classification is based on the same gene annotations for the 26695 strain from NCBI (NC_000915.1) which we already used in our previous study [12], supplemented with annotations for validated sRNAs that we discovered at this time.

3.6. Detection of regulatory elements

Knowledge of genome-wide TSS positions facilitates the discovery of diverse transcriptome features including regulatory elements. Global inference of promoter motifs upstream of TSS can help to understand which sequence elements are important for transcription initiation and elucidate gene regulation pathways. In the previous *H. pylori* dRNA-seq study, an extended −10 box downstream of periodic AT-rich stretches was identified as the canonical promoter motif for the housekeeping sigma factor σ^{80} [12]. The same motif was later confirmed in our comparative dRNA-seq analysis as the consensus for the housekeeping σ^{08} in *Campylobacter jejuni* [25] reinforcing that this is a common feature of ε-proteobacterial promoters.

Annotated pTSS and sTSS of mRNA genes can be used to generate transcriptome-wide 5′UTR maps that can subsequently be utilized to search for *cis*-regulatory elements such as riboswitches and RNA thermometers. Additionally, they can contain sRNA binding sites, for example, the 5′UTR of the chemotaxis receptor TlpB which contains a poly(G) stretch far upstream of the start codon which is targeted by the sRNA RepG [44]. Our TSS map includes 925 pTSS and sTSS of which 790 are associated with mRNA genes. 20 of these TSS give rise to leaderless transcripts while the remaining 770 5′UTRs show an average and median length of ∼81 and 45 nt, respectively, and a clear peak in the distribution in a range between 20 and 40 nt (data not shown). This is consistent with earlier findings [12] and while leaderless mRNAs used to be considered rare in prokaryotes, unexpectedly high numbers have also been discovered in other bacteria [45–47]. On the other hand, in archaea, where leaderless mRNAs seem to represent the standard translational template, a dRNA-seq-based study in *Methanosarcina mazei* revealed that most mRNAs carry long 5′UTRs [48]. These findings underline the importance of 5′UTRs for translational control and the usefulness of dRNA-seq for their annotation.

3.6.1. Identification of cis- and trans-encoded sRNAs

The TSS map does not only provide information on transcription starts and regulatory mechanisms associated with already annotated genes or operons but also facilitates the discovery of novel regulatory elements, including sRNAs expressed from intergenic regions or antisense to ORFs. In the previous 454 dRNA-seq study we identified >60 sRNAs in *H. pylori* and an extensive antisense transcriptome. Fig. 4A shows an oTSS located in the intergenic region between the HP1399 and HP1400 genes annotated as arginase and iron(III) dicitrate transport protein FecA, respectively. This TSS was annotated as pTSS for HP1400 in our previous study as the 454 data did not provide any evidence for the existence of the newly predicted pTSS 294 nucleotides further downstream. The downstream TSS is clearly visible in the Illumina data and was also already mapped before by Ernst et al. [49] 2 nt further upstream via primer extension. Such oTSS could either belong to separate standing sRNA genes (e.g. an unannotated sRNA of ∼220 nt in the case of the TSS upstream of HP1400) or represent alternative promoters leading to transcription of longer 5′UTRs. The example of the TSS upstream of HP1400 indicates that even more transcriptome features can still be discovered when sequencing at higher coverage or more conditions are included.

Another prominent class of transcripts that is getting more and more attention are antisense RNAs (asRNAs) [50,51]. In the 454 data, >900 asTSS were detected and at least one asTSS expressed opposite of >50% of all genes. Based on our new mid-log data we detected 766 TSS solely classified as asTSS, indicating again a large set of asRNA candidates. 52% of these asTSS overlap with the 684 TSS solely classified as asTSS in the 454 datasets. Fig. 4B shows an example for an asTSS located internal and antisense to the *ispDF* gene annotated as bifunctional 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase/2-C-methyl-D-erythritol 2,4-cyclodiphosphate. This TSS was also previously annotated based on the 454 data. We do not know how many of our predicted asTSS represent functional RNAs or the amount corresponding only to spurious transcripts, as the number of reported antisense RNAs strongly varies and the function of most of these transcripts is still unclear [52]. However, we think that a global TSS map is the optimal starting point to find an answer to this question by conducting additional experiments like, for example, detection on Northern blots and discovery of associated phenotypes. Moreover, regulation of the asTSS expression under different growth or stress conditions as well as conservation in multiple strains could be further indications that they indeed have regulatory functions [26].

3.6.2. Overlapping 5′UTRs

Association of a TSS to more than one class, as mentioned above, can be used to select for regulatory elements. Divergently transcribed gene pairs with overlapping regions in the 5′UTR or even coding sequence (CDS) can result in asRNA-mediated gene regulation (reviewed in [53]) or affect promoter occupancy [54]. We found 200 pTSS and 67 sTSS that were additionally classified as asTSS. Requiring a minimum overlap of 10 nt and considering only TSS for mRNA genes, we identified 40 distinct overlapping 5′UTRs associated with 28 divergently transcribed gene pairs (Table S2). One example is shown in Fig. 4C, which depicts two hypothetical proteins (HP1162 and HP1163) with their associated pTSS. The 5′UTR of HP1162 almost completely overlaps CDS and 5′UTR of HP1163, possibly resulting in an asRNA-mediated regulation.

3.7. Accessibility of the H. pylori 26695 TSS map in an online browser

In the previous 454 dRNA-seq study, we provided the TSS map in a table that indicated the TSS positions. While such a table format is very useful for downstream analysis such as promoter motif predictions or 5′UTR calculations, sometimes it is also helpful to look at the cDNA coverage plots to see the overall read distribution for a gene of interest. Thus, we here used GenomeView [55] to set up an easily accessible online browser that directly includes the

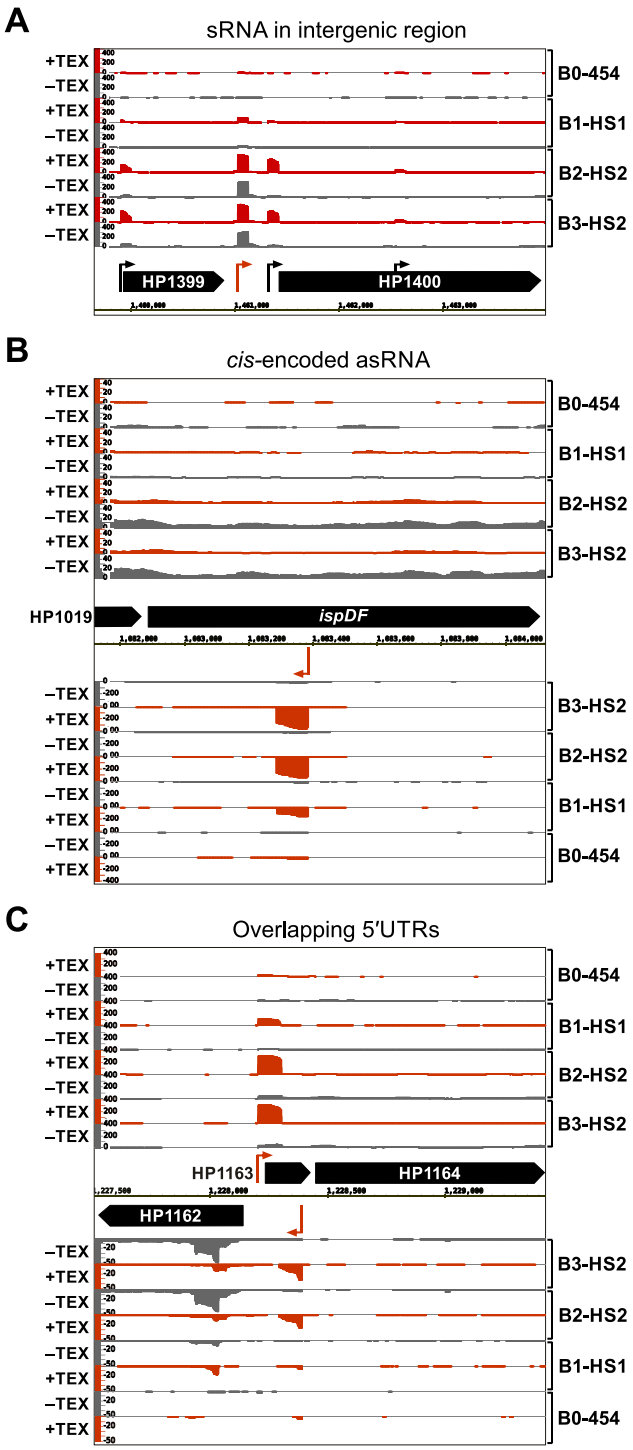


Fig. 4. Examples for transcripts and regulatory elements. Screenshots from IGB showing the relative cDNA coverage plots for ±TEX libraries of the four ML replicates. Red arrows indicate the genomic position of (A) a putative sRNA in the intergenic region between the HP1399 and HP1400 genes, (B) a putative asRNA transcribed from the opposite strand of the *ispDF* gene, and (C) two predicted p/asTSS for the divergently transcribed HP1162 and HP1163 genes, which indicate the presence of overlapping 5′UTRs.

complete set of predicted TSS from this study together with the respective coverage plots and gene annotations that were used for the prediction (Fig. 5). For comparison, we also added the previous manual TSS annotations from [12] for which coverage plots of all five biological conditions are loaded but not displayed by default. The browser allows for manual inspection of the data

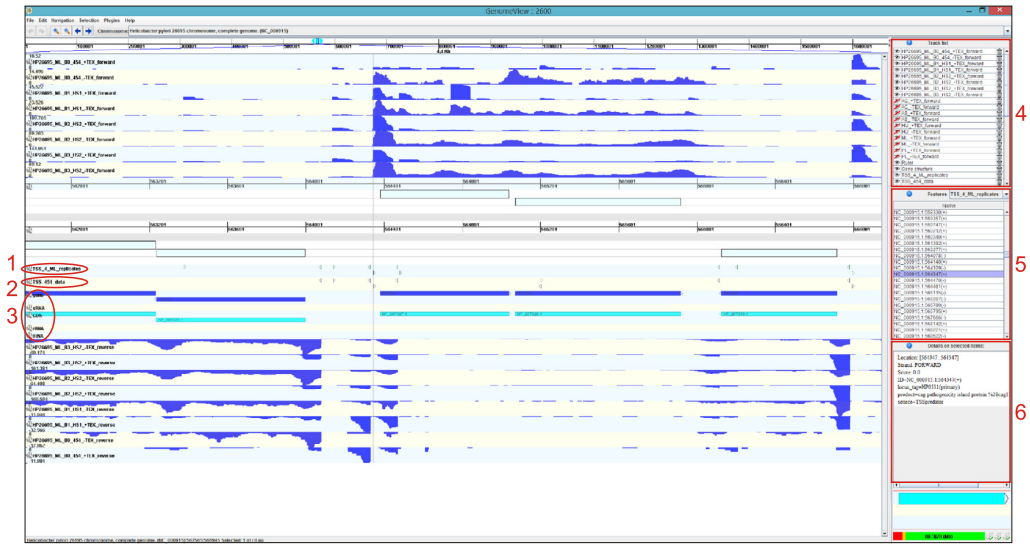


Fig. 5. Example screenshot of the online browser. cDNA coverage plots for the forward and reverse strand are displayed above and below the genomic axis, respectively. The browser depicts our new TSS annotations based on all 4 ML replicates (1), TSS annotations from the previous study based on 454 sequencing [12] (2), and annotations for genes, coding sequences (CDS), sRNAs, rRNAs and tRNAs (3). On the right the display and order of tracks can be altered in the track list (4), specific features can be selected (5) and details for selected items are displayed (6).

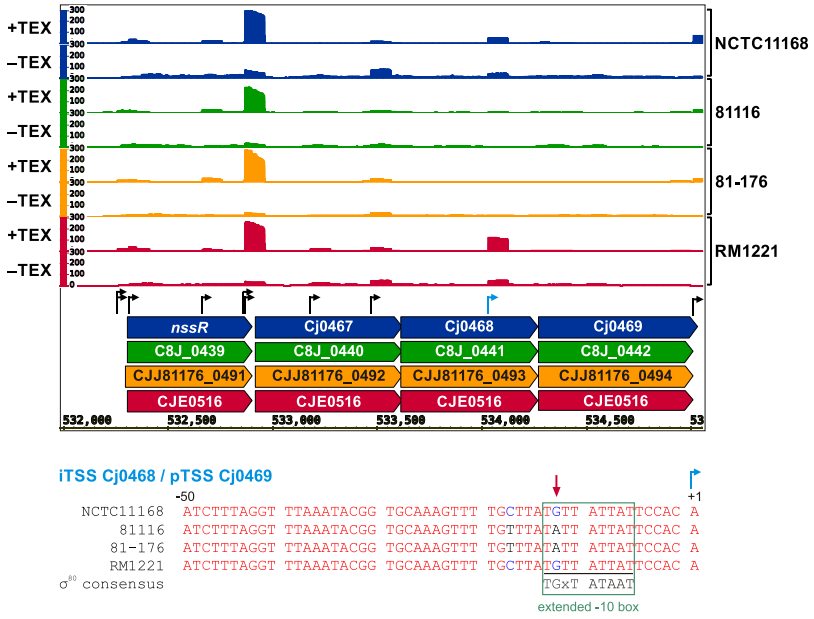


Fig. 6. Comparative TSS annotation in *C. jejuni* strains reveals strain-specific promoter usage. (Top) cDNA coverage for an example region of the SuperGenome of four *C. jejuni* strains which encompasses the *nssR*, Cj0467, Cj0468, Cj0469 operon. Black arrows indicate annotated TSS and the blue arrow a p/iTSS internal to Cj0468 and upstream of Cj0469 which is only detected in two strains (NCTC11168 and RM1221) and shows no expression in the other strains (81116 and 81-176). (Bottom) Multiple alignment of the promoter region −50 to +1 upstream of the blue p/iTSS based on the four *C. jejuni* strains. Differential expression of this TSS is likely caused by a G to A single nucleotide polymorphism (SNP) (red arrow) in the extended −10 box of strains 81116 and 81-176

including expression and enrichment at annotated TSS. Please note that there is no option for a consistent scaling of all coverage plots, which makes it necessary to compare the numbers representing relative expression values at the left end of each track. This online browser, which is available under http://www.imib-wuerzburg.de/research/hpylori/, greatly facilitates the data accessibility and allows researchers to examine the cDNA coverage plots and TSS for their genes of interest.

4. Conclusions

Genome-wide annotation of transcriptional features is crucial to understand the full complement of transcriptional regulation in an organism. Knowledge of precise positions of transcript 5′ ends gained by dRNA-seq is fundamental for a variety of downstream analyses like global prediction of promoter motifs or automated annotation of *cis*-regulatory features in 5′UTRs. In addition,

it provides a basis for the annotation of a plethora of novel transcripts including sRNAs and asRNAs as well as specific regulatory features like antisense-mediated regulation via overlapping 5′UTRs. Here, we provided a detailed description of the application of the dRNA-seq method to generate a transcriptome-wide TSS map using *H. pylori* 26695 as an example organism. We utilized an automated TSS prediction approach implemented in the tool TSSpredator, which greatly facilitates TSS annotation on a global scale.

We compared the predicted TSS positions based on four replicates of the ML condition to previous manual annotations from our initial study [12] and detected 1340 matching positions but in addition 893 novel TSS, which were previously missed due to low coverage or insufficient support by several growth conditions. Other TSS positions might have been missed as they were not expressed in the ML condition or due to a lack of enrichment in the +TEX library. This could for example be caused by processing of primary transcripts by the RNA pyrophosphohydrolase RppH which was shown to initiate degradation via cleavage of the 5′-PPP [56]. Moreover, some of the differences in TSS could also be due to slight differences in growth conditions or mutations in the 26695 clones upon sequential passages in different labs.

The TSSpredator tool is also capable of comparative analysis based on different bacterial strains or biological conditions. In a previous study, we used dRNA-seq together with TSSpredator to annotate TSS in four *Campylobacter jejuni* strains [25] in a comparative manner. Using a whole-genome alignment of multiple strains calculated by Mauve [57], TSSpredator computes a common coordinate system for all strains referred to as SuperGenome and TSS are then annotated by directly applying the above-mentioned detection and enrichment criteria to corresponding genomic positions. An example for a TSS that is only present in two of the four strains is shown in Fig. 6. The difference is likely caused by a single base mutation in the extended −10 box of the promoter region for the p/iTSS displayed in blue. The G at the second position of the consensus motif (TGxTATAAT) is replaced by an A in strains 81116 and 81-176 abolishing transcription in these strains. In strains NCTC11168 and RM1221 the TSS within Cj0468 uncouples transcription of the Cj0469 gene encoding an amino-acid ABC transporter ATP-binding protein from the *nssR*, Cj0467, Cj0468, Cj0469 operon. This indicates that while most comparative genomics studies consider SNPs in open reading frames that can lead to frameshift mutations or change protein function, also SNPs in non-coding parts can contribute to strain-specific gene expression and regulation and thereby add yet another layer of complexity. Such a comparative transcriptome analysis of multiple isolates might also help to examine the conservation and potential functions of the increasing number of *cis*-encoded antisense RNAs and helps to reveal conserved and strain-specific or species-specific sRNAs [25,58].

Overall, a comparative TSS analysis of multiple *H. pylori* strains and or *H. pylori* grown under different stress or growth conditions will provide further insight into conserved and strain-specific transcriptional features of this widespread human pathogen, which might underlie phenotypic differences among closely related strains. Together with variable host factors, these might contribute to the different clinical outcomes observed for *H. pylori* infections and to establish life-long persistent infections and adaptation to changing conditions in the human stomach.

Acknowledgements

We thank Anika Lins for technical assistance, Richard Reinhardt for help with deep sequencing, Alexander Herbig for discussions on TSSpredator, and Jörg Vogel for sharing previous 454 *H. pylori* RNA-seq data. The Sharma lab received financial support from the ZINF Young Investigator program at the Research Center for Infectious Diseases (ZINF) in Würzburg, Germany, the Bavarian Research Network for Molecular Biosystems (BioSysNet), DFG project Sh580/1-1, the Daimler-Benz-Foundation, and the Young Academy program of the Bavarian Academy of Sciences.

Appendix A. Supplementary data

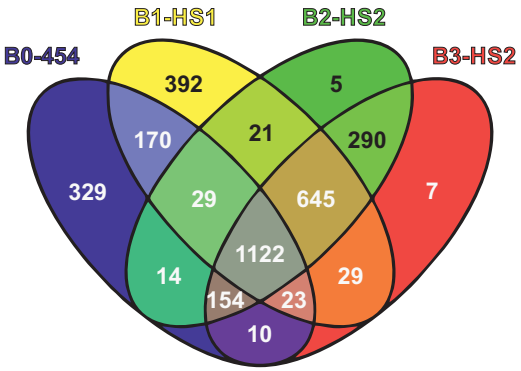
Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2015.06.012.

References

[1] K.O. Mutz, A. Heilkenbrinker, M. Lonne, J.G. Walter, F. Stahl, Curr. Opin. Biotechnol. 24 (2013) 22–30.
[2] N.J. Croucher, N.R. Thomson, Curr. Opin. Microbiol. 13 (2010) 619–624.
[3] A.H. van Vliet, FEMS Microbiol. Lett. 302 (2010) 1–7.
[4] Z. Wang, M. Gerstein, M. Snyder, Nat. Rev. Genet. 10 (2009) 57–63.
[5] R. Sorek, P. Cossart, Nat. Rev. Genet. 11 (2010) 9–16.
[6] N.F. Lahens, I.H. Kavakli, R. Zhang, K. Hayer, M.B. Black, H. Dueck, A. Pizarro, J. Kim, R. Irizarry, R.S. Thomas, G.R. Grant, J.B. Hogenesch, Genome Biol. 15 (2014) R86.
[7] N.J. Croucher, M.C. Fookes, T.T. Perkins, D.J. Turner, S.B. Marguerat, T. Keane, M.A. Quail, M. He, S. Assefa, J. Bahler, R.A. Kingsley, J. Parkhill, S.D. Bentley, G. Dougan, N.R. Thomson, Nucl. Acids Res. 37 (2009) e148.
[8] N. Cloonan, A.R. Forrest, G. Kolle, B.B. Gardiner, G.J. Faulkner, M.K. Brown, D.F. Taylor, A.L. Steptoe, S. Wani, G. Bethel, A.J. Robertson, A.C. Perkins, S.J. Bruce, C.C. Lee, S.S. Ranade, H.E. Peckham, J.M. Manning, K.J. McKernan, S.M. Grimmond, Nat. Methods 5 (2008) 613–619.
[9] Y. He, B. Vogelstein, V.E. Velculescu, N. Papadopoulos, K.W. Kinzler, Science 322 (2008) 1855–1857.
[10] D. Parkhomchuk, T. Borodina, V. Amstislavskiy, M. Banaru, L. Hallen, S. Krobitsch, H. Lehrach, A. Soldatov, Nucl. Acids Res. 37 (2009) e123.
[11] A. Sittka, S. Lucchini, K. Papenfort, C.M. Sharma, K. Rolle, T.T. Binnewies, J.C. Hinton, J. Vogel, PLoS Genet. 4 (2008) e1000163.
[12] C.M. Sharma, S. Hoffmann, F. Darfeuille, J. Reignier, S. Findeiß, A. Sittka, S. Chabas, K. Reiche, J. Hackermüller, R. Reinhardt, P.F. Stadler, J. Vogel, Nature 464 (2010) 250–255.
[13] P.T. McGrath, H. Lee, L. Zhang, A.A. Iniesta, A.K. Hottes, M.H. Tan, N.J. Hillson, P. Hu, L. Shapiro, H.H. McAdams, Nat. Biotechnol. 25 (2007) 584–592.
[14] J.A. Thompson, M.F. Radonovich, N.P. Salzman, J. Virol. 31 (1979) 437–446.
[15] L. Argaman, R. Hershberg, J. Vogel, G. Bejerano, E.G. Wagner, H. Margalit, S. Altuvia, Curr. Biol. 11 (2001) 941–950.
[16] J. Vogel, V. Bartels, T.H. Tang, G. Churakov, J.G. Slagter-Jager, A. Huttenhofer, E.G. Wagner, Nucl. Acids Res. 31 (2003) 6435–6443.
[17] B.A. Bensing, B.J. Meyer, G.M. Dunny, Proc. Natl. Acad. Sci. USA 93 (1996) 7794–7799.
[18] O. Wurtzel, R. Sapra, F. Chen, Y. Zhu, B.A. Simmons, R. Sorek, Genome Res. 20 (2010) 133–141.
[19] A. Mendoza-Vargas, L. Olvera, M. Olvera, R. Grande, L. Vega-Alvarado, B. Taboada, V. Jimenez-Jacinto, H. Salgado, K. Juarez, B. Contreras-Moreira, A.M. Huerta, J. Collado-Vides, E. Morett, PLoS One 4 (2009) e7526.
[20] B.K. Cho, K. Zengler, Y. Qiu, Y.S. Park, E.M. Knight, C.L. Barrett, Y. Gao, B.O. Palsson, Nat. Biotechnol. 27 (2009) 1043–1049.
[21] B.K. Cho, D. Kim, E.M. Knight, K. Zengler, B.O. Palsson, BMC Biol. 12 (2014) 4.
[22] D. Kim, J.S. Hong, Y. Qiu, H. Nagarajan, J.H. Seo, B.K. Cho, S.F. Tsai, B.O. Palsson, PLoS Genet. 8 (2012) e1002867.
[23] N. Singh, J.T. Wade, Methods Mol. Biol. 1103 (2014) 1–10.
[24] C.M. Sharma, J. Vogel, Curr. Opin. Microbiol. 19C (2014) 97–105.
[25] G. Dugar, A. Herbig, K.U. Förstner, N. Heidrich, R. Reinhardt, K. Nieselt, C.M. Sharma, PLoS Genet. 9 (2013) e1003495.
[26] M.K. Thomason, T. Bischler, S.K. Eisenbart, K.U. Förstner, A. Zhang, A. Herbig, K. Nieselt, C.M. Sharma, G. Storz, J. Bacteriol. 197 (2015) 18–28.
[27] P. Blomberg, E.G. Wagner, K. Nordstrom, EMBO J. 9 (1990) 2331–2340.
[28] E. Berezikov, F. Thuemmler, L.W. van Laake, I. Kondova, R. Bontrop, E. Cuppen, R.H. Plasterk, Nat. Genet. 38 (2006) 1375–1377.
[29] K.U. Förstner, J. Vogel, C.M. Sharma, Bioinformatics 30 (2014) 3421–3423.
[30] S. Hoffmann, C. Otto, S. Kurtz, C.M. Sharma, P. Khaitovich, J. Vogel, P.F. Stadler, J. Hackermüller, PLoS Comput. Biol. 5 (2009) e1000502.
[31] J.W. Nicol, G.A. Helt, S.G. Blanchard Jr., A. Raja, A.E. Loraine, Bioinformatics 25 (2009) 2730–2731.
[32] O.A. Soutourina, M. Monot, P. Boudry, L. Saujet, C. Pichon, O. Sismeiro, E. Semenova, K. Severinov, C. Le Bouguenec, J.Y. Coppee, B. Dupuy, I. Martin-Verstraete, PLoS Genet. 9 (2013) e1003493.
[33] Y.F. Lin, D.A. Romero, S. Guan, L. Mamanova, K.J. McDowall, BMC Genomics 14 (2013) 620.
[34] T.L. Cover, M.J. Blaser, Gastroenterology 136 (2009) 1863–1873.
[35] S. Suerbaum, P. Michetti, N. Engl. J. Med. 347 (2002) 1175–1186.
[36] J.F. Tomb, O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, B.A. Dougherty, K. Nelson, J. Quackenbush, L. Zhou, E.F. Kirkness, S. Peterson, B. Loftus, D. Richardson, R. Dodson, H.G.

Khalak, A. Glodek, K. McKenney, L.M. Fitzegerald, N. Lee, M.D. Adams, E.K. Hickey, D.E. Berg, J.D. Gocayne, T.R. Utterback, J.D. Peterson, J.M. Kelley, M.D. Cotton, J.M. Weidman, C. Fujii, C. Bowman, L. Watthey, E. Wallin, W.S. Hayes, M. Borodovsky, P.D. Karp, H.O. Smith, C.M. Fraser, J.C. Venter, Nature 388 (1997) 539–547.
[37] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, M. Law, J. Biomed. Biotechnol. 2012 (2012) 251364.
[38] J. Frias-Lopez, Y. Shi, G.W. Tyson, M.L. Coleman, S.C. Schuster, S.W. Chisholm, E.F. Delong, Proc. Natl. Acad. Sci. USA 105 (2008) 3805–3810.
[39] N. Heidrich, G. Dugar, J. Vogel, C.M. Sharma, Methods Mol. Biol. 1311 (2015) 1–21.
[40] P.A. t Hoen, M.R. Friedlander, J. Almlof, M. Sammeth, I. Pulyakhina, S.Y. Anvar, J.F. Laros, H.P. Buermans, O. Karlberg, M. Brannvall, J.T. den Dunnen, G.J. van Ommen, I.G. Gut, R. Guigo, X. Estivill, A.C. Syvanen, E.T. Dermitzakis, T. Lappalainen, Nat. Biotechnol. 31 (2013) 1015–1022.
[41] C.A. Raabe, T.H. Tang, J. Brosius, T.S. Rozhdestvensky, Nucl. Acids Res. 42 (2014) 1414–1426.
[42] F. Amman, M.T. Wolfinger, R. Lorenz, I.L. Hofacker, P.F. Stadler, S. Findeiß, BMC Bioinf. 15 (2014) 89.
[43] H. Jorjani, M. Zavolan, Bioinformatics 30 (2014) 971–974.
[44] S.R. Pernitzsch, S.M. Tirier, D. Beier, C.M. Sharma, Proc. Natl. Acad. Sci. USA 111 (2014) E501–E510.
[45] T. Cortes, O.T. Schubert, G. Rose, K.B. Arnvig, I. Comas, R. Aebersold, D.B. Young, Cell Rep. 5 (2013) 1121–1131.
[46] A. de Groot, D. Roche, B. Fernandez, M. Ludanyi, S. Cruveiller, D. Pignol, D. Vallenet, J. Armengaud, L. Blanchard, Genome Biol. Evol. 6 (2014) 932–948.
[47] D.A. Romero, A.H. Hasan, Y.F. Lin, L. Kime, O. Ruiz-Larrabeiti, M. Urem, G. Bucca, L. Mamanova, E.E. Laing, G.P. van Wezel, C.P. Smith, V.R. Kaberdin, K.J. McDowall, Mol. Microbiol. 94 (5) (2014) 963–987.
[48] D. Jager, C.M. Sharma, J. Thomsen, C. Ehlers, J. Vogel, R.A. Schmitz, Proc. Natl. Acad. Sci. USA 106 (2009) (1882) 21878–21882.
[49] F.D. Ernst, J. Stoof, W.M. Horrevoets, E.J. Kuipers, J.G. Kusters, A.H. van Vliet, Infection Immunity 74 (2006) 6821–6828.
[50] J. Georg, W.R. Hess, Microbiol. Mol. Biol. Rev. 75 (2011) 286–300.
[51] M.K. Thomason, G. Storz, Annu. Rev. Genet. 44 (2010) 167–188.
[52] J.T. Wade, D.C. Grainger, Nat. Rev. Microbiol. 12 (2014) 647–653.
[53] N. Sesto, O. Wurtzel, C. Archambaud, R. Sorek, P. Cossart, Nat. Rev. Microbiol. 11 (2013) 75–82.
[54] K.M. Bendtsen, J. Erdossy, Z. Csiszovszki, S.L. Svenningsen, K. Sneppen, S. Krishna, S. Semsey, Nucl. Acids Res. 39 (2011) 6879–6885.
[55] T. Abeel, T. Van Parys, Y. Saeys, J. Galagan, Y. Van de Peer, Nucl. Acids Res. 40 (2012) e12.
[56] A. Deana, H. Celesnik, J.G. Belasco, Nature 451 (2008) 355–358.
[57] A.C. Darling, B. Mau, F.R. Blattner, N.T. Perna, Genome Res. 14 (2004) 1394–1403.
[58] O. Wurtzel, N. Sesto, J.R. Mellin, I. Karunker, S. Edelheit, C. Becavin, C. Archambaud, P. Cossart, R. Sorek, Mol. Syst. Biol. 8 (2012) 583.

Figure S1


B1-HS1
B2-HS2
B0-454
B3-HS2
392
5
170
21
290
329
29
645
7
1122
14
29
154
23
10

3.4 IDENTIFICATION OF THE RNA PYROPHOSPHOHYDROLASE RPPH OF HELICOBACTER PYLORI AND GLOBAL ANALYSIS OF ITS RNA TARGETS

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 292, NO. 5, pp. 1934–1950, February 3, 2017

Identification of the RNA Pyrophosphohydrolase RppH of *Helicobacter pylori* and Global Analysis of Its RNA Targets*[S]

Received for publication, September 29, 2016, and in revised form, December 2, 2016 Published, JBC Papers in Press, December 14, 2016, DOI 10.1074/jbc.M116.761171

Thorsten Bischler[‡§1], Ping-kun Hsieh[¶1], Marcus Resch[‡§1], Quansheng Liu[¶], Hock Siew Tan[§], Patricia L. Foley[¶], Anika Hartleib[‡§], Cynthia M. Sharma[‡§2], and Joel G. Belasco[¶3]

From the [‡]Research Center for Infectious Diseases and the [§]Institute of Molecular Infection Biology, University of Würzburg, Josef-Schneider-Strasse 2/D15, 97080 Würzburg, Germany and the [¶]Kimmel Center for Biology and Medicine at the Skirball Institute and the Department of Microbiology, New York University School of Medicine, New York, New York 10016

Edited by Patrick Sung

RNA degradation is crucial for regulating gene expression in all organisms. Like the decapping of eukaryotic mRNAs, the conversion of the 5′-terminal triphosphate of bacterial transcripts to a monophosphate can trigger RNA decay by exposing the transcript to attack by 5′-monophosphate-dependent ribonucleases. In both biological realms, this deprotection step is catalyzed by members of the Nudix hydrolase family. The genome of the gastric pathogen *Helicobacter pylori*, a Gram-negative epsilonproteobacterium, encodes two proteins resembling Nudix enzymes. Here we present evidence that one of them, HP1228 (renamed HpRppH), is an RNA pyrophosphohydrolase that triggers RNA degradation in *H. pylori*, whereas the other, HP0507, lacks such activity. *In vitro*, HpRppH converts RNA 5′-triphosphates and diphosphates to monophosphates. It requires at least two unpaired nucleotides at the 5′ end of its substrates and prefers three or more but has only modest sequence preferences. The influence of HpRppH on RNA degradation *in vivo* was examined by using RNA-seq to search the *H. pylori* transcriptome for RNAs whose 5′-phosphorylation state and cellular concentration are governed by this enzyme. Analysis of cDNA libraries specific for transcripts bearing a 5′-triphosphate and/or monophosphate revealed at least 63 potential HpRppH targets. These included mRNAs and sRNAs, several of which were validated individually by half-life measurements and quantification of their 5′-terminal phosphorylation state in wild-type and mutant cells. These findings demonstrate an important role for RppH in post-transcriptional gene regulation in pathogenic Epsilonproteobacteria and suggest a possible basis for the phenotypes of *H. pylori* mutants lacking this enzyme.

Helicobacter pylori is a Gram-negative, microaerophilic epsilonproteobacterium that colonizes the stomachs of more than 50% of the world's population (1). Infection by this microorganism is associated with the development of gastritis, peptic ulcers, and adenocarcinoma (2). A variety of *H. pylori* proteins important for colonization and pathogenesis have been identified, but little is yet understood about how the biosynthesis of these factors is controlled, especially at the post-transcriptional level. For example, although RNA degradation is among the principal post-transcriptional mechanisms that control gene expression in all organisms, little is known about this process in Epsilonproteobacteria.

Much of what is understood about bacterial mRNA decay has come from studies of *Escherichia coli*. Most mRNAs in *E. coli* and other Gammaproteobacteria are degraded by a combination of endonucleolytic cleavage by ribonuclease E (RNase E) and 3′-exonucleolytic digestion by polynucleotide phosphorylase, RNase II, and RNase R (3). Although Epsilonproteobacteria contain homologs of the principal 3′-exonucleases present in *E. coli*, they lack RNase E (4, 5). Instead, to degrade mRNA, they rely on two ribonucleases absent from Gammaproteobacteria but present in Gram-positive bacteria: the endonuclease RNase Y and the 5′-exonuclease RNase J (6–9).

When initially synthesized, the 5′ ends of bacterial transcripts typically are triphosphosphorylated. However, RNase J and RNase E favor RNA substrates that have only one 5′-terminal phosphate (10, 11). This property has two important consequences. First, it enables these enzymes to rapidly degrade monophosphosphorylated intermediates generated by prior ribonuclease cleavage (12). Furthermore, it can assist them in attacking full-length transcripts whose 5′-triphosphate has been converted to a monophosphate by an RNA pyrophosphohydrolase (11, 13).

Every bacterial RNA pyrophosphohydrolase that has so far been identified is a member of the Nudix hydrolase family of proteins, as are most eukaryotic RNA decapping enzymes (14–

* Research in the Sharma laboratory was supported by the Young Investigator Program at the Research Center for Infectious Diseases in Würzburg, Germany; the Bavarian Research Network for Molecular Biosystems (BioSysNet); and DFG project Sh580/1-1 of the German Research Association (DFG). Research in the Belasco laboratory was supported by National Institutes of Health Grants 5R01GM035769 and 5R01GM112940 (to J. G. B.) and National Institutes of Health Fellowship T32AI007180 (to P. L. F.). The authors declare that they have no conflicts of interest with the contents of this article. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

[S] This article contains supplemental Tables S1–S3.

The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE86943).

[1] These authors contributed equally to this work.

[2] To whom correspondence may be addressed: Institute for Molecular Infection Biology, University of Würzburg, Josef-Schneider-Str. 2/D15, 97080 Würzburg, Germany. Tel.: 49-931/31-82560; E-mail: cynthia.sharma@uni-wuerzburg.de.

[3] To whom correspondence may be addressed: Skirball Institute, New York University School of Medicine, 540 First Ave., New York, NY 10016. Tel.: 212-263-5409; Fax: 212-263-2150; E-mail: joel.belasco@med.nyu.edu.

16). Nudix enzymes are present in all domains of life and have a variety of biochemical functions, most of which appear to involve the hydrolysis of substrates that contain a nucleoside diphosphate moiety (17). Besides their role in initiating RNA degradation (11, 13, 15, 19, 20), these enzymes have been implicated in a variety of metabolic pathways, such as those governing the synthesis or breakdown of folic acid (21), coenzyme A (22), ADP-ribose (23, 24), UDP-glucose (25), and mutagenic nucleotides such as 8-oxo-dGTP (26, 27).

The genomes of most species encode multiple Nudix enzymes, which can be identified by a characteristic sequence motif (the Nudix motif) (27) that usually is well conserved (17). Protein domains containing this motif typically fold so as to form a central four-stranded mixed β sheet (β strands 1, 3, 4, and 5) and an antiparallel β sheet (β strands 2 and 6) sandwiched between three α helices (α1, α2, and α3) (27). Those that act as RNA pyrophosphohydrolases (known by the genetic acronym RppH) are widespread in bacteria. However, their evolutionary divergence has made many of them difficult to identify on the basis of sequence alone. So far, two distinct families of RppH enzymes with recognizable sequence characteristics have been defined: those found in Alpha-, Beta-, Gamma-, and Epsilonproteobacteria and in flowering plants (*E. coli* RppH homologs) and those found in Bacillales but not in other Firmicutes (*Bacillus subtilis* RppH homologs) (16). These two families differ in their substrate specificity due to sequence differences external to the Nudix motif (16, 19, 28).

In addition to homologs of RNase J and RNase Y, the small genome of *H. pylori* (5) encodes two potential Nudix hydrolases, HP1228 and HP0507. HP1228 is able to catalyze the hydrolysis of the dinucleoside tetraphosphate Ap_4A *in vitro* (29), and it appears from its sequence to be a homolog of *E. coli* RppH. However, its ability to function as an RNA pyrophosphohydrolase has never been examined, either *in vitro* or *in vivo*, and no *H. pylori* RNAs whose longevity is HP1228-dependent have ever been identified. Here we report the identification and characterization of HP1228 as an RNA pyrophosphohydrolase in *H. pylori* (HpRppH). Our studies demonstrate the ability of the purified protein to convert 5′-terminal triphosphates to monophosphates and define its substrate specificity. By employing RNA-seq methods selective for either triphosphorylated or monophosphorylated 5′ ends, we have identified mRNAs and sRNAs targeted by this enzyme in *H. pylori*. By contrast, HP0507 appears to lack RNA pyrophosphohydrolase activity.

Results

The H. pylori Genome Encodes a Potential RppH Homolog—In *E. coli*, 5′-end-dependent RNA degradation is triggered by the RNA pyrophosphohydrolase RppH, a member of the Nudix hydrolase family (13). Like other members of this protein family, *E. coli* RppH contains a Nudix motif ($GX_5EX_7REUXEEXGU$, where U is a bulky aliphatic residue and *X* is any amino acid) (27), a telltale signature of Nudix domains (17). Examination of the genome of *H. pylori* strain 26695 (5) for encoded proteins that bear a Nudix motif revealed two candidates, HP1228 and HP0507 (29, 30). HP1228 contains a region that matches this motif at eight of nine positions ($GX_5EX_7REUXEEXG\underline{T}$; mismatch underlined), whereas HP0507 matches the motif at only four positions ($\underline{L}X_5\underline{K}X_7\underline{E}\underline{E}\underline{A}XEEXG\underline{Y}$; mismatches underlined). The sequence of HP1228, which is well conserved in other Epsilonproteobacteria (see the Kyoto Encyclopedia of Genes and Genomes website), is 34% identical to that of *E. coli* RppH (EcRppH) and contains each of the 23 amino acid residues that are strictly conserved in virtually all proteobacterial orthologs of EcRppH (Fig. 1*A*) (16). These sequence characteristics suggest that HP1228, like EcRppH, is an RNA pyrophosphohydrolase. We modeled the three-dimensional structure of HP1228 by using the X-ray crystal structure of EcRppH (31) as a template (Fig. 1, *B* and *C*). Most of the residues that are identical in these two proteins are clustered around a cavity that functions as the substrate-binding site and catalytic center of EcRppH. These residues include four glutamates that coordinate Mg^{2+} ions as well as other amino acids implicated in substrate recognition (16, 31). By contrast, the 19 residues that comprise the carboxyl terminus of EcRppH are entirely absent in HP1228 and many other EcRppH orthologs (16).

HpRppH Functions in Vitro as an RNA Pyrophosphohydrolase—Cellular phenotypes such as decreased resistance to hydrogen peroxide exposure (29) and a diminished ability to invade gastric epithelial cells (32) have been reported for *H. pylori* mutants unable to produce HP1228. However, the molecular function of this protein has remained unclear. To address this question, we tested HP1228 *in vitro* for RNA pyrophosphohydrolase activity. A 0.44-kb triphosphorylated *rpsT* RNA substrate (13) bearing a 5′-terminal γ-^{32}P label and an internal fluorescein label was treated with purified HP1228, and reaction samples were quenched at time intervals. The reaction products were then split into two portions and examined by gel electrophoresis and thin layer chromatography. HP1228 removed the radiolabel from the 5′ end of the transcript (Fig. 2*A*, *top*), yielding a mixture of radioactive pyrophosphate and orthophosphate (Fig. 2*B*). No such activity was observed for an HP1228 mutant in which an essential active site residue had been replaced (E57Q). γ-Phosphate removal by purified HP1228 was not accompanied by degradation of the transcript, whose fluorescence intensity was invariant (Fig. 2*A*, *bottom*).

To determine whether HP1228 generates monophosphorylated RNA as the other reaction product, we prepared another RNA substrate, $GA(CU)_{13}$, bearing a monophosphate, diphosphate, or triphosphate at the 5′ terminus and a single ^{32}P label between the first and second nucleotide. After treatment with HP1228, the RNA reaction product was subjected to alkaline hydrolysis, and the 5′-terminal nucleotide was examined by thin layer chromatography and autoradiography (Fig. 2*C*). HP1228-catalyzed hydrolysis of both triphosphorylated and diphosphorylated $GA(CU)_{13}$ generated monophosphorylated $GA(CU)_{13}$, which was detected as radiolabeled pGp after alkaline hydrolysis, whereas the corresponding monophosphorylated substrate was not affected by this enzyme. As expected, none of the substrates reacted with catalytically inactive HP1228 bearing an E57Q substitution. We conclude that HP1228 functions *in vitro* as an RNA pyrophosphohydrolase that is able to convert triphosphorylated and diphosphorylated substrates to

Functional Characterization of H. pylori RppH

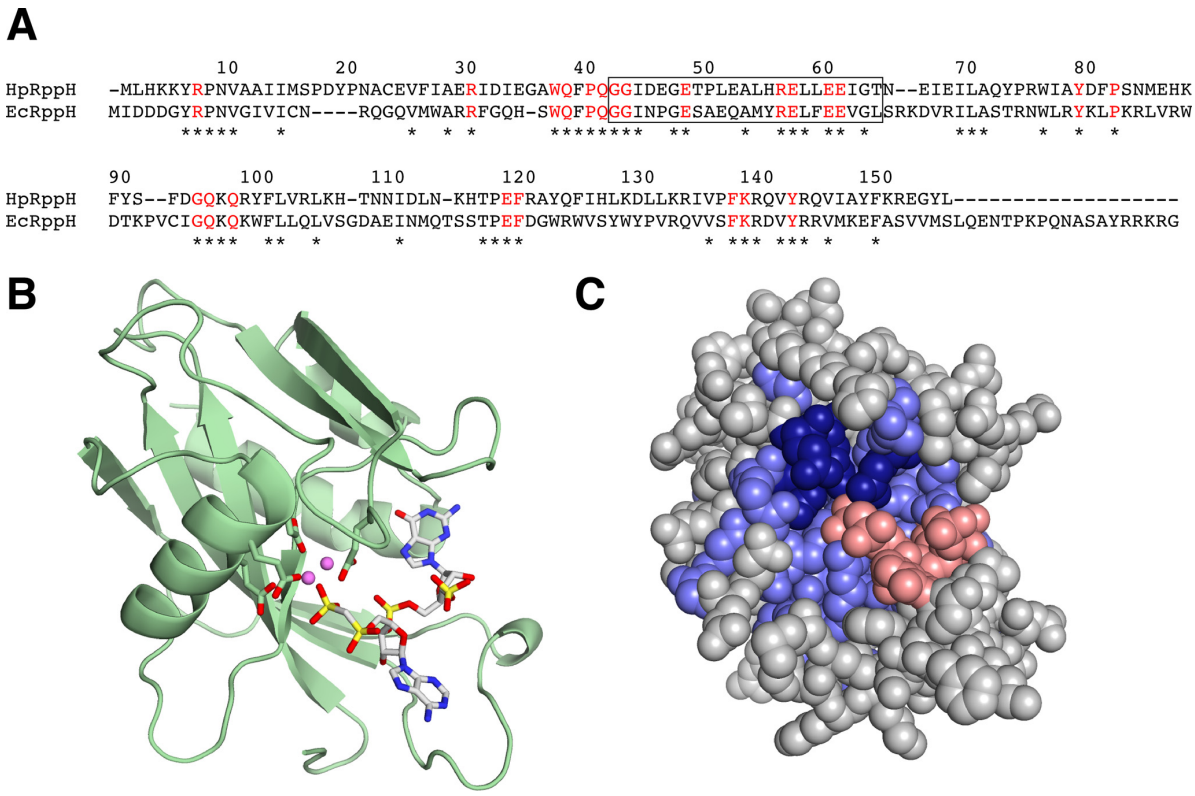


FIGURE 1. **RppH alignment and structure.** *A*, alignment of HpRppH (HP1228) and EcRppH. The sequences were aligned by analysis with ClustalW (18). *Asterisks* mark amino acid residues that are identical in the two sequences. Residues that are conserved in virtually all bacterial orthologs of EcRppH (16) are depicted as *red letters*. The region containing the Nudix motif is enclosed in a *rectangle*. *Numbers* correspond to the sequence of HpRppH. *B* and *C*, structural model of HpRppH bound to an RNA ligand. The structure of HpRppH was modeled by homology to the X-ray crystal structure of EcRppH bound to an oligonucleotide ligand and two Mg^{2+} ions (Protein Data Bank code 4S2X) (31) by using SWISS-MODEL on the ExPASy bioinformatics website (50). *B*, *ribbon model*. *Green ribbon*, HpRppH backbone. The four glutamate side chains (Glu-57, Glu-60, Glu-61, and Glu-118; *sticks*) that coordinate Mg^{2+} ions (*violet spheres*) are also shown. The diphosphorylated RNA ligand is depicted in a *stick representation*. *C*, *space-fill model*. *Blue*, HpRppH residues that are identical in EcRppH, which include the four glutamate residues (*dark blue*) that coordinate Mg^{2+} (not shown). *Gray*, HpRppH residues that differ from EcRppH. *Red*, diphosphorylated RNA ligand.

monophosphorylated products. These findings and the homology of HP1228 to EcRppH prompted us to rename it *H. pylori* RppH (HpRppH).

Requirement for Unpaired Nucleotides at the 5′ Terminus—To determine the minimum number of unpaired 5′-terminal nucleotides required for the reaction of RNA with HpRppH, we compared the reactivity of a set of structurally unambiguous substrates previously used to examine the specificity of EcRppH and *B. subtilis* RppH (BsRppH) (Fig. 3) (16, 19). A8, the prototype of these RNA substrates, comprised an 8-nucleotide single-stranded segment followed by two stem-loop structures, the first of which contained the only uracil base in the entire molecule. Synthesized by *in vitro* transcription in the presence of [γ-^{32}P]ATP and fluorescein-12-UTP, A8 contained a γ radiolabel within the 5′-terminal triphosphate and a single fluorescein label at the top of the first stem-loop. For use as an internal standard, we also prepared doubly labeled A8XL RNA, which differed from A8 only in having an additional stem-loop at the 3′ end.

Conversion of these triphosphorylated RNAs to monophosphorylated products was monitored by combining equal amounts of each with HpRppH, quenching reaction samples periodically, and separating the reaction products by gel electrophoresis (Fig. 4*A*). The extent of reaction at each time point was then determined for both A8 and A8XL by comparing the radioactivity of the corresponding gel band with its fluorescence intensity. As anticipated, the reaction rates of these two substrates were very similar.

The single-stranded segment at the 5′ end of A8 was then shortened from 8 to 4, 3, 2, or 1 nucleotide by removing nucleotides from its 3′ boundary to create A4, A3, A2, and A1 (Fig. 3), and the reactivity of these RNAs toward HpRppH was compared in the presence of A8XL. A4 and A3 were almost as reactive as A8, whereas A2 was significantly less reactive, and A1 was completely unreactive (Fig. 4, *A* and *B*). The addition of three unpaired nucleotides to the 3′ end of A1 (A1+3) (Fig. 3) did not improve its reactivity (Fig. 4*B*), providing evidence that its resistance to pyrophosphate removal by HpRppH resulted from an insufficient number of unpaired nucleotides at the 5′ end and not merely from its shorter overall length. The effect of the number of unpaired 5′-terminal nucleotides was similar for a related set of RNA substrates in which the first nucleotide was changed from A to G (Fig. 4*C*). These findings demonstrate that HpRppH, like EcRppH and BsRppH (16, 19), requires at least two unpaired nucleotides at the 5′ end of its substrates and prefers three or more.

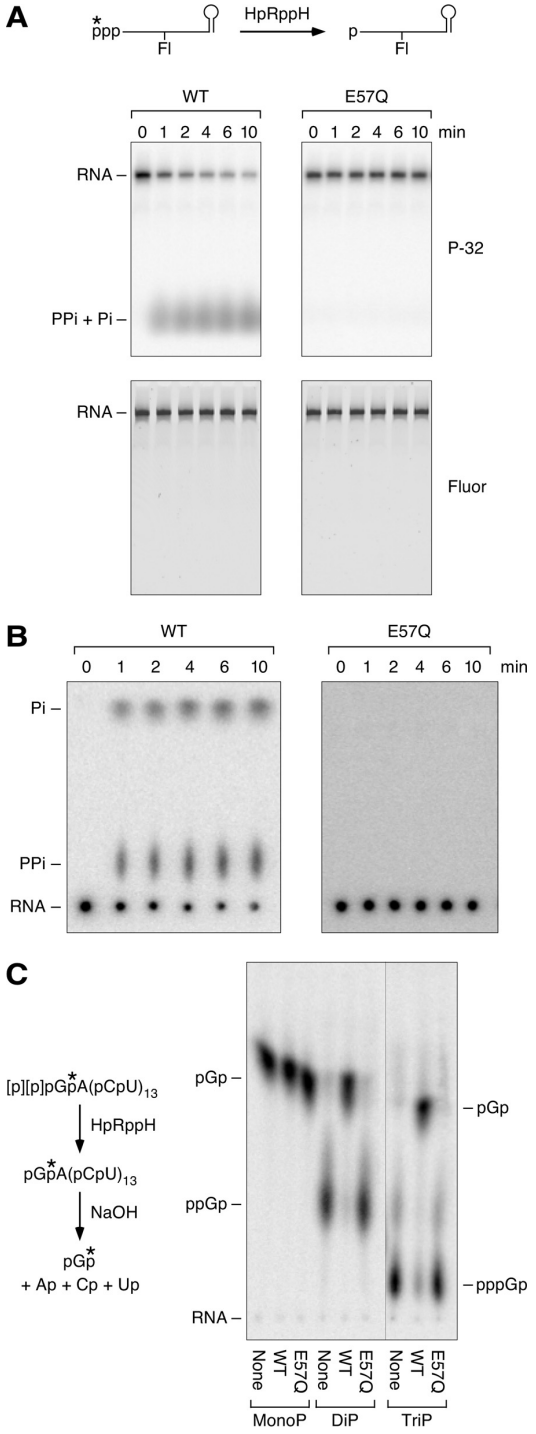


FIGURE 2. **RNA pyrophosphohydrolase activity of purified HpRppH.** *A* and *B*, release of pyrophosphate and orthophosphate from the 5′ end of triphosphorylated RNA by HpRppH. Triphosphorylated *rpsT* P1 RNA (13) bearing a 5′-terminal γ-^{32}P label (*) and an internal fluorescein label (*Fl*) (*A*, *top*) was treated with purified HpRppH or HpRppH-E57Q (75 nM), and reaction samples isolated at time intervals were analyzed by gel electrophoresis (with subsequent detection of radioactivity (*P-32*) and fluorescence (*Fluor*)) (*A*) or thin layer chromatography (with subsequent detection of radioactivity) (*B*). *PPi*, pyrophosphate; *Pi*, orthophosphate. *C*, conversion of triphosphorylated and diphosphorylated RNA to monophosphorylated RNA by HpRppH. Triphosphorylated (*TriP*), diphosphorylated (*DiP*), and monophosphorylated (*MonoP*) GA(CU)$_{13}$ bearing a single ^{32}P label (*) between the first and second nucleotides were treated with purified HpRppH or HpRppH-E57Q (75 nM), and the radiolabeled starting materials and reaction products were subjected to alkaline hydrolysis and analyzed by thin layer chromatography.

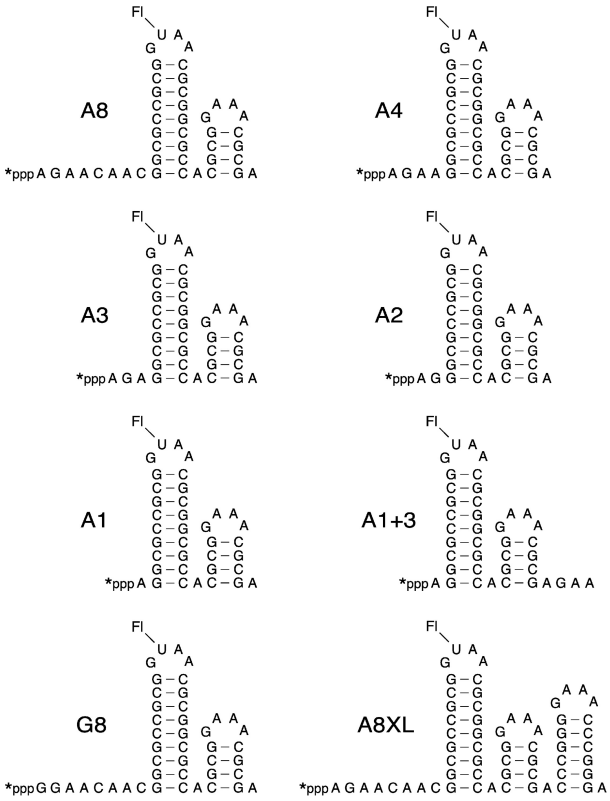


FIGURE 3. **HpRppH substrates.** The sequence and expected secondary structure of A8, A4, A3, A2, A1, A1+3, G8, and A8XL RNA are shown. Each bore a 5′-terminal triphosphate (*ppp*), a γ-^{32}P radiolabel (*) at the 5′ end, and a fluorescein label (*Fl*) at the top of the first stem-loop. In each RNA name, the letter indicates the identity of the 5′-terminal nucleotide, and the numeral indicates the number of unpaired nucleotides at the 5′ end. Truncated derivatives of A8 (A4, A3, A2, and A1) lacked 4–7 nucleotides from the 3′ boundary of the 5′-terminal single-stranded segment. G8, G4, G3, G2, G1, and G0 were identical to their A-series counterparts except for the presence of G instead of A at the 5′ end. A1+3 was the same as A1 except for three additional nucleotides at the 3′ end.

Effect of 5′-Terminal RNA Sequence—The requirement for unpaired nucleotides at the 5′ end of HpRppH substrates raised the possibility that this enzyme might also be affected by the identity of the nucleotides there. To determine whether HpRppH prefers substrates bearing certain 5′-terminal sequences, we replaced individual nucleotides in A4 (hereafter referred to as A4$_{AGAA}$ to reveal both the identity of the 5′-terminal nucleotide and the sequence of unpaired nucleotides at the 5′ end) and examined the effect of these substitutions on reactivity. A substitution mutant (G4$_{GGAA}$) in which the first nucleotide was changed from A to G (a majority of primary transcripts in bacteria begin with either of these two nucleotides (33)) was only slightly less reactive than A4$_{AGAA}$ (Fig. 5*A*). By contrast, pyrimidine substitutions at the second position

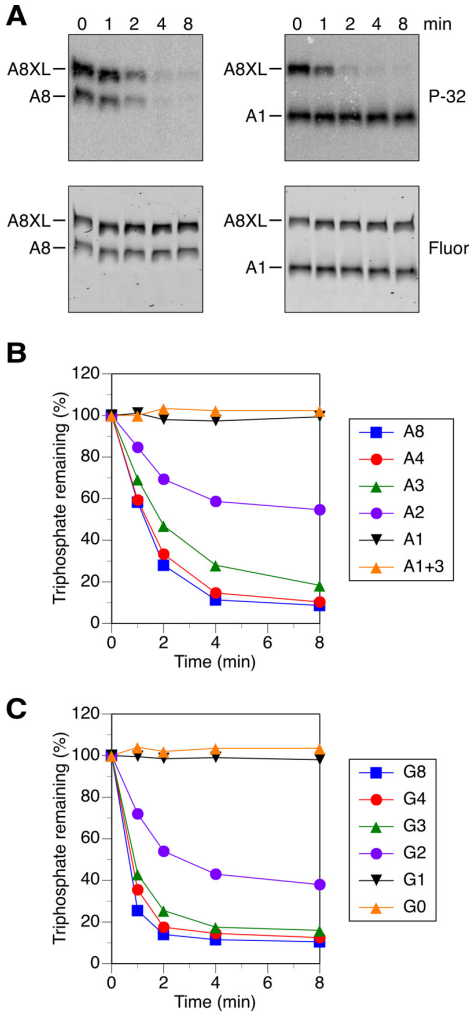


FIGURE 4. **Effect of the length of the 5′-terminal single-stranded segment on reactivity with HpRppH *in vitro*.** *A*, representative gel images. *In vitro* transcribed A8 and A1 bearing a γ-^{32}P radiolabel and an internal fluorescein label were mixed with labeled A8XL and treated with purified HpRppH (16 nM), and the radioactivity (*P-32*) and fluorescence (*Fluor*) of each RNA were monitored as a function of time by gel electrophoresis. *B* and *C*, graphs. HpRppH-catalyzed phosphate removal from A8, A4, A3, A2, A1, and A1+3 or from G8, G4, G3, G2, G1, and G0 was monitored as in *A* and quantified by normalizing the radioactivity remaining in each RNA to the corresponding fluorescence intensity. Each time point is the average of two or more independent measurements. Error bars have been omitted to improve the legibility of the graph; instead, the S.D. of each measurement is reported in supplemental Table S1.

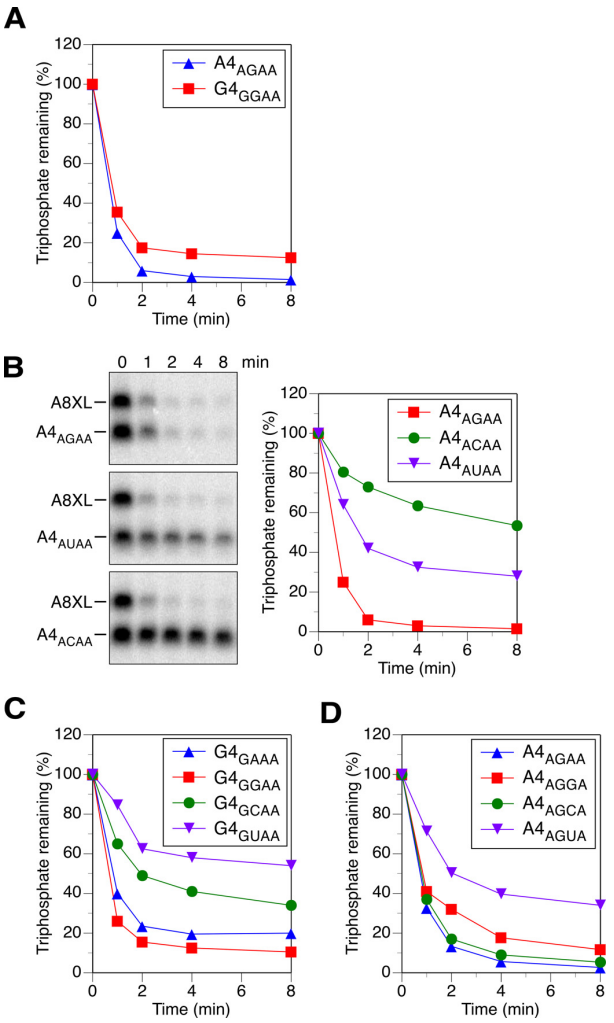


FIGURE 5. **Effect of the sequence of the first three RNA nucleotides on reactivity with HpRppH *in vitro*.** *A*, position 1. The reactivity of $A4_{AGAA}$ and $G4_{GGAA}$ was compared as in Fig. 4. The subscript in each RNA name indicates the sequence of the four unpaired nucleotides at the 5′ end. Consequently, $A4_{AGAA}$ was equivalent to A4. *B* and *C*, position 2. The reactivity of $A4_{AGAA}$, $A4_{ACAA}$, and $A4_{AUAA}$ and of $G4_{GGAA}$, $G4_{GAAA}$, $G4_{GCAA}$, and $G4_{GUAA}$ was compared. Although both radioactivity and fluorescence were measured, only the former is shown in the gel images. To avoid modifying the second nucleotide, $A4_{AUAA}$ and $G4_{GUAA}$ were not labeled with fluorescein; instead, the fluorescence of fluorescein-labeled A8XL was used to normalize the data from each time point. The synthesis of $A4_{AAAA}$ was not successful. *D*, position 3. The reactivity of $A4_{AGAA}$, $A4_{AGGA}$, $A4_{AGCA}$, and $A4_{AGUA}$ was compared. To avoid modifying the third nucleotide, $A4_{AGUA}$ was not labeled with fluorescein. The S.D. of each measurement is reported in supplemental Table S1.

significantly impaired reactivity. In particular, replacing the G at position 2 of either $A4_{AGAA}$ or $G4_{GGAA}$ with C or U (to create $A4_{ACAA}$, $A4_{AUAA}$, $G4_{GCAA}$, or $G4_{GUAA}$) slowed the reaction considerably but did not block it, whereas substituting A at that position in $G4_{GGAA}$ (to create $G4_{GAAA}$) had only a modest inhibitory effect (Fig. 5, *B* and *C*; synthesis of $A4_{AAAA}$ was not successful). Altering the third nucleotide had a substantial impact only when U was introduced there, as $A4_{AGGA}$ and $A4_{AGCA}$ were as reactive as $A4_{AGAA}$, whereas $A4_{AGUA}$ was less reactive (Fig. 5*D*). Overall, the 5′-terminal sequence specificity of HpRppH closely resembles that of its ortholog EcRppH in that both enzymes are rather promiscuous but prefer a purine at position 2, unlike BsRppH, which strictly requires G at position 2 (16, 19).

Inactivity of HP0507 as an RNA Pyrophosphohydrolase—In addition to HpRppH (HP1228), which contains an almost perfect Nudix motif (GX_5EX_7REU*X*EE*X*GT; mismatch underlined), the genome of *H. pylori* encodes another protein, HP0507, that contains a partial Nudix motif ($LX_5KX_7EEAXEEXGY$; mismatches underlined). HP0507 is 11% identical in overall sequence to EcRppH and has been

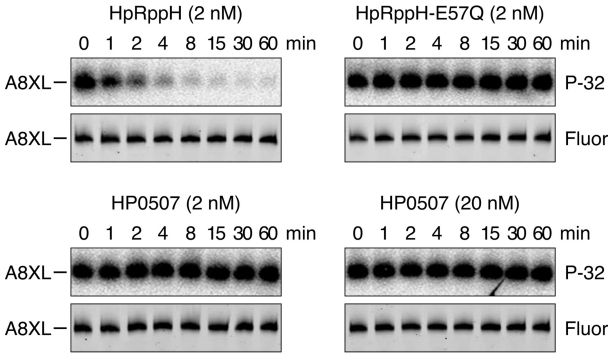


FIGURE 6. **Test of the putative Nudix hydrolase HP0507 for RNA pyrophosphohydrolase activity.** *In vitro* transcribed A8XL RNA radiolabeled at the 5′-terminal γ-phosphate and internally labeled with fluorescein (see Fig. 3) was treated with purified HpRppH (2 nM final concentration), catalytically inactive HpRppH-E57Q (2 nM), or HP0507 (2 or 20 nM), and reaction samples quenched at time intervals were subjected to gel electrophoresis. Hydrolytic release of the 5′-terminal radiolabel was detected by autoradiography (*P-32*), and the integrity of the remainder of the RNA molecule was monitored by fluorescence (*Fluor*).

implicated in virulence (30). To determine whether HP0507 has RNA pyrophosphohydrolase activity, we tested whether it can remove a γ radiolabel from triphosphorylated A8XL. Whereas 2 nM HpRppH released almost 90% of the radiolabel from this substrate within 4 min, no reactivity was observed for HP0507, even when 10-fold more enzyme (20 nM) was added and the reaction was monitored for 60 min (Fig. 6). Assuming the structural integrity of the recombinant protein, these findings indicate that HP0507 either is not an RNA pyrophosphohydrolase or has a strict RNA substrate specificity that prevents it from acting on A8XL.

Test for 8-Oxo-dGTPase Activity—Most bacterial species contain multiple Nudix hydrolases, each of which has a distinct function (17). Because HpRppH is the only *H. pylori* protein with a *bona fide* Nudix motif, we wondered whether it might have more than one function. Therefore, we tested whether it possesses another well known Nudix hydrolase activity: the ability of MutT-like proteins to protect cells from incorporating the mutagenic nucleotide 8-oxo-dGTP during DNA replication by selectively converting it to 8-oxo-dGMP (34). 8-Oxo-dGTP or dGTP was mixed with purified *E. coli* MutT (positive control), HpRppH, HP0507, EcRppH, or BsRppH. After 60 min, the starting material and products were separated by thin layer chromatography on fluorescent PEI-cellulose plates. As expected, MutT exhibited substantial 8-oxo-dGTPase activity at an enzyme concentration of just 1 nM and completely hydrolyzed the substrate at a concentration of 10 nM; only at a much higher enzyme concentration (100 nM) was it able to hydrolyze dGTP (Fig. 7). By contrast, neither HpRppH nor HP0507 detectably hydrolyzed 8-oxo-dGTP below an enzyme concentration of 100 nM, and neither had a preference for that substrate over dGTP. EcRppH and BsRppH were completely unable to hydrolyze either substrate. These results suggest that neither HpRppH nor HP0507 functions as a selective 8-oxo-dGTPase in *H. pylori*. This conclusion is consistent with a previous report that the frequency of spontaneous mutation is the same in wild-type and Δ*rppH* strains of *H. pylori* (29).

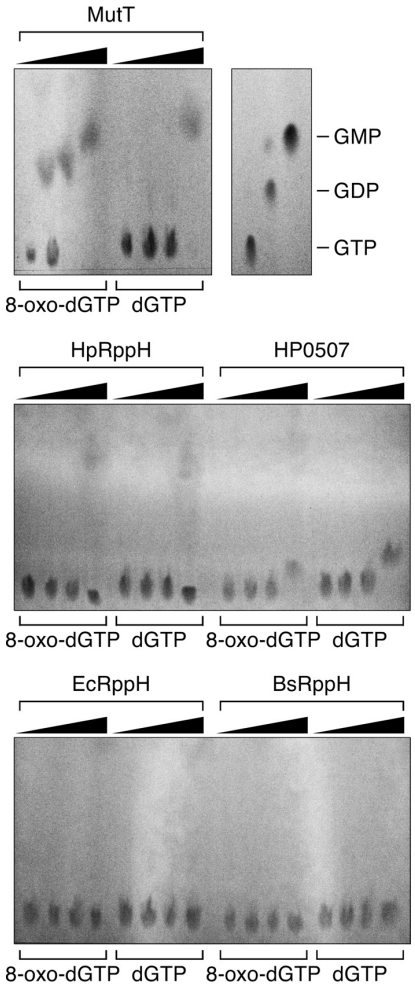


FIGURE 7. **Test of HpRppH and HP0507 for selective 8-oxo-dGTPase activity.** 8-Oxo-dGTP or dGTP (50 μM) was treated for 60 min with various concentrations of purified *E. coli* MutT, HpRppH, HP0507, EcRppH, or BsRppH (0, 1, 10, or 100 nM), and the reaction products were examined by thin layer chromatography on PEI-cellulose. GTP, GDP, and GMP served as mobility standards. Whereas MutT hydrolyzed 8-oxo-dGTP much faster than dGTP, the other enzymes either did not hydrolyze 8-oxo-dGTP detectably (EcRppH, BsRppH) or did so slowly and no faster than they hydrolyzed dGTP (HpRppH, HP0507).

Global Identification of RppH Targets by Differential RNA-seq[4]—To investigate the global role of HpRppH in converting 5′-triphosphates to monophosphates in *H. pylori*, we used a variant of differential RNA-seq (dRNA-seq) (35, 36) to compare the concentration and 5′-phosphorylation state of transcripts in isogenic *H. pylori* strains containing or lacking the *rppH* gene. For this purpose, we constructed two derivatives of the wild-type *H. pylori* strain 26695: an *rppH* deletion mutant (Δ*rppH*) and an *rppH* complementation strain (C*rppH*) bearing an ectopic copy of the *rppH* gene. The Δ*rppH* strain was gen-

[4] The abbreviations used are: RNA-seq, high-throughput RNA sequencing; dRNA-seq, differential RNA-seq; TEX, Terminator 5′-phosphate-dependent exonuclease; TAP, tobacco acid pyrophosphatase; 5′-P and 5′-PPP, 5′-monophosphorylated and 5′-triphosphorylated, respectively; TSS, transcription start site; nt, nucleotide(s); PABLO, phosphorylation assay by ligation of oligonucleotides.

erated by a non-polar chromosomal substitution in which the *rppH* gene of wild-type (WT) cells was replaced with a kanamycin resistance cassette (37). The C*rppH* strain was then constructed by complementing this deletion with an ectopic copy of the *H. pylori rppH* gene under the control of its own promoter (35), which was introduced at an unrelated locus (*rdxA*) previously used as a site for integrating genes into the *H. pylori* chromosome (38–41).

These isogenic *H. pylori* strains were grown to log phase, and total RNA isolated from each was used to generate three libraries specific for transcripts bearing 1) a 5′-triphosphate, 2) a 5′-monophosphate, or 3) either a 5′-triphosphate or a 5′-monophosphate (Fig. 8*A*). This was accomplished by differential treatment of total cellular RNA with Terminator 5′-phosphate-dependent exonuclease (TEX) and tobacco acid pyrophosphatase (TAP) (35, 36, 42). The 5′-exonuclease activity of TEX digests 5′-monophosphorylated (5′-P) RNAs but leaves triphosphorylated (5′-PPP) transcripts intact. Subsequent treatment of the latter set of transcripts with TAP generates monophosphorylated 5′ ends to which an RNA oligonucleotide can be ligated, thereby enabling cDNA synthesis. By contrast, treatment with TAP alone enables cDNA synthesis from both triphosphorylated and monophosphorylated RNAs, whereas treatment with neither enzyme allows cDNA synthesis only from cellular RNAs that are already monophosphorylated. Therefore, to identify RNAs in each category, we generated cDNA libraries specific for transcripts with a 5′-triphosphate (+TEX/+TAP), a 5′-monophosphate (−TEX/−TAP), or both (−TEX/+TAP) from all three strains (Fig. 8*A*) and subjected them to Illumina sequencing. In total, between 4.1 and 5.8 million reads were sequenced for each of the cDNA libraries, of which between 96.8 and 98.5% could be mapped to the *H. pylori* 26695 genome (Table 1).

Because RppH triggers the degradation of its targets by converting 5′-terminal triphosphates to monophosphates, both the cellular concentration of those transcripts and the percentage of each that is 5′-triphosphorylated (rather than monophosphorylated) are expected to be higher in Δ*rppH* cells than in WT and C*rppH* cells. Hence, we screened for *H. pylori* transcripts that fulfill both of these criteria to identify RNAs that are

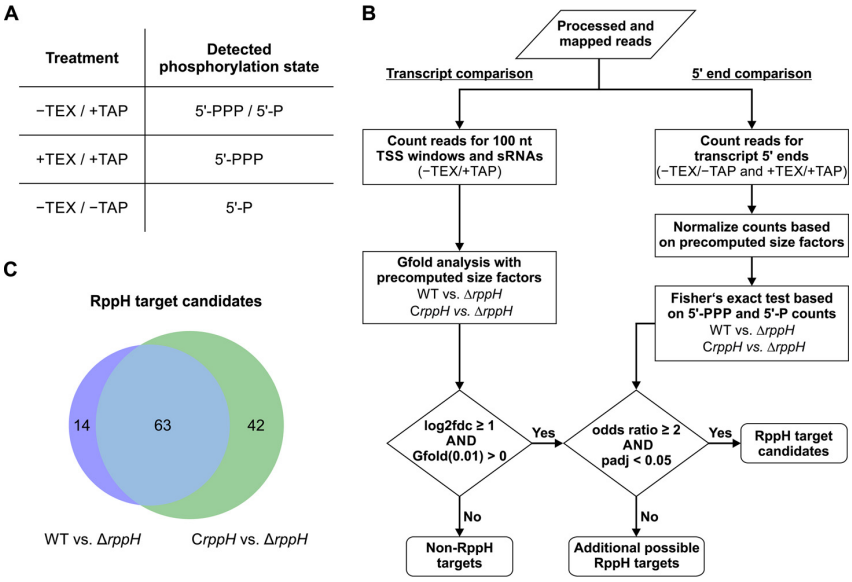


FIGURE 8. **Differential RNA-seq analysis of RNA 5′ ends in *H. pylori* cells containing or lacking HpRppH.** *A*, combinations of TEX/TAP treatments used to enrich for 5′-PPP transcripts, 5′-P transcripts, or both (5′-PPP/5′-P). *B*, computational pipeline used to identify RppH target candidates. To pass muster, a ≥2-fold increase in both the RNA concentration (log_2fdc ≥ 1) and the ratio of 5′-PPP to 5′-P ends (odds ratio ≥ 2) was required in Δ*rppH* cells *versus* WT and C*rppH* cells. Precomputed size factors were based on the number of mapped reads for each library. *C*, Venn diagram of RppH target candidates identified in Δ*rppH* cells *versus* WT or C*rppH* cells.

TABLE 1

Mapping statistics for the *H. pylori* 26695 Illumina libraries

This table summarizes the total number of sequenced cDNA reads after quality trimming, as well as the number of mapped and uniquely mapped reads for each sequencing library. Percentage values are relative to the number of reads that are >11 nt in length after poly(A) trimming.

Library	Total number of reads after quality trimming	Number of reads long enough after poly(A) trimming	Mapped reads	Percentage of mapped reads	Uniquely mapped reads	Percentage of uniquely mapped reads
HP26695_WT_+TEX_+TAP	4,105,444	2,904,136	2,855,756	98.3	1,776,256	61.2
HP26695_WT_−TEX_+TAP	4,709,180	4,393,218	4,303,527	98.0	2,191,371	49.9
HP26695_WT_−TEX_−TAP	4,541,183	3,735,492	3,637,516	97.4	1,801,667	48.2
HP26695_drppH_+TEX_+TAP	4,322,165	3,691,261	3,637,538	98.5	2,685,123	72.7
HP26695_drppH_−TEX_+TAP	5,687,933	5,367,180	5,285,689	98.5	2,913,153	54.3
HP26695_drppH_−TEX_−TAP	5,171,576	4,086,347	3,994,970	97.8	2,034,955	49.8
HP26695_CrppH_+TEX_+TAP	5,260,512	4,380,242	4,309,747	98.4	2,396,496	54.7
HP26695_CrppH_−TEX_+TAP	4,676,932	4,358,626	4,266,490	97.9	2,004,928	46.0
HP26695_CrppH_−TEX_−TAP	5,813,459	4,490,487	4,345,087	96.8	1,997,827	44.5

directly and productively targeted by HpRppH. To detect changes in RNA concentration, the relative numbers of transcripts in the −TEX/+TAP libraries (5′-PPP and 5′-P) were calculated on the basis of cDNA counts for windows of up to 100 nt encompassing previously annotated transcription start sites (TSSs) of mRNAs and non-coding RNAs (42) as well as full-length annotations for sRNAs (35) and then compared among the three strains by using Gfold (43). In addition, to detect changes in 5′-phosphorylation, transcript levels in the +TEX/+TAP (5′-PPP) and −TEX/−TAP (5′-P) libraries were calculated for a region from 5 nt upstream to 4 nt downstream of each TSS and then compared for WT *versus* Δ*rppH* as well as C*rppH versus* Δ*rppH* by a one-sided Fisher's exact test. In total, 63 of 925 transcripts (53 mRNAs and 10 sRNAs) were found to be at least 2-fold more abundant ($\log_2$fdc ≥1 and Gfold (0.01) > 0) in Δ*rppH* cells *versus* both WT and C*rppH* cells and additionally to be enriched at least 2-fold for monophosphorylated *versus* triphosphorylated 5′ ends (5′-P/5′-PPP ratio) in WT and C*rppH* cells compared with the Δ*rppH* mutant (one-sided Fisher's exact test; odds ratio ≥2 and Benjamini-Hochberg adjusted *p* value <0.05) (Fig. 8, *B* and *C*), evidence that they may be RppH targets. These 63 transcripts are summarized in the first sheet of supplemental Table S2. The 53 up-regulated mRNAs included 52 primary TSSs and one secondary TSS associated with 52 distinct genes. An additional 119 possible targets whose concentration increased ≥2-fold in Δ*rppH* cells without a corresponding reduction in the percentage of monophosphorylated 5′ ends are listed in supplemental Table S3.

sRNAs Targeted by RppH—Among the apparent HpRppH targets that we detected is the sRNA IsoA1 (HPnc6350) (supplemental Table S2). As judged from the RNA-seq data, the concentration of triphosphorylated IsoA1 and its abundance relative to its monophosphorylated counterpart were substantially higher in Δ*rppH* cells than in WT and C*rppH* cells (Fig. 9*A* and supplemental Table S2). IsoA1 belongs to a group of six structurally related *H. pylori* sRNAs, IsoA1–6 (RNA inhibitor of small ORF family A), that are each ~80 nt in length (35). They are transcribed antisense to the small ORFs *aapA1–6* (antisense RNA-associated peptide family A), which encode homologous peptides 22–30 amino acids in length. *In vitro*, IsoA1 has been shown to strongly and selectively inhibit the translation of *aapA1* mRNA (35). One other IsoA sRNA, IsoA3 (HPnc7630), as well as several additional sRNA candidates (including HPnc1980, HPnc3560, and HPnc7830) and potential *cis*-encoded antisense RNAs also appear to be targeted by HpRppH (supplemental Table S2). In contrast, a number of other sRNAs, such as the RNA polymerase inhibitor 6S RNA (HPnc6561, Fig. 9*A*) and HPnc2450 (supplemental Table S2), do not appear to be affected by HpRppH, indicating that this pyrophosphohydrolase targets sRNAs selectively.

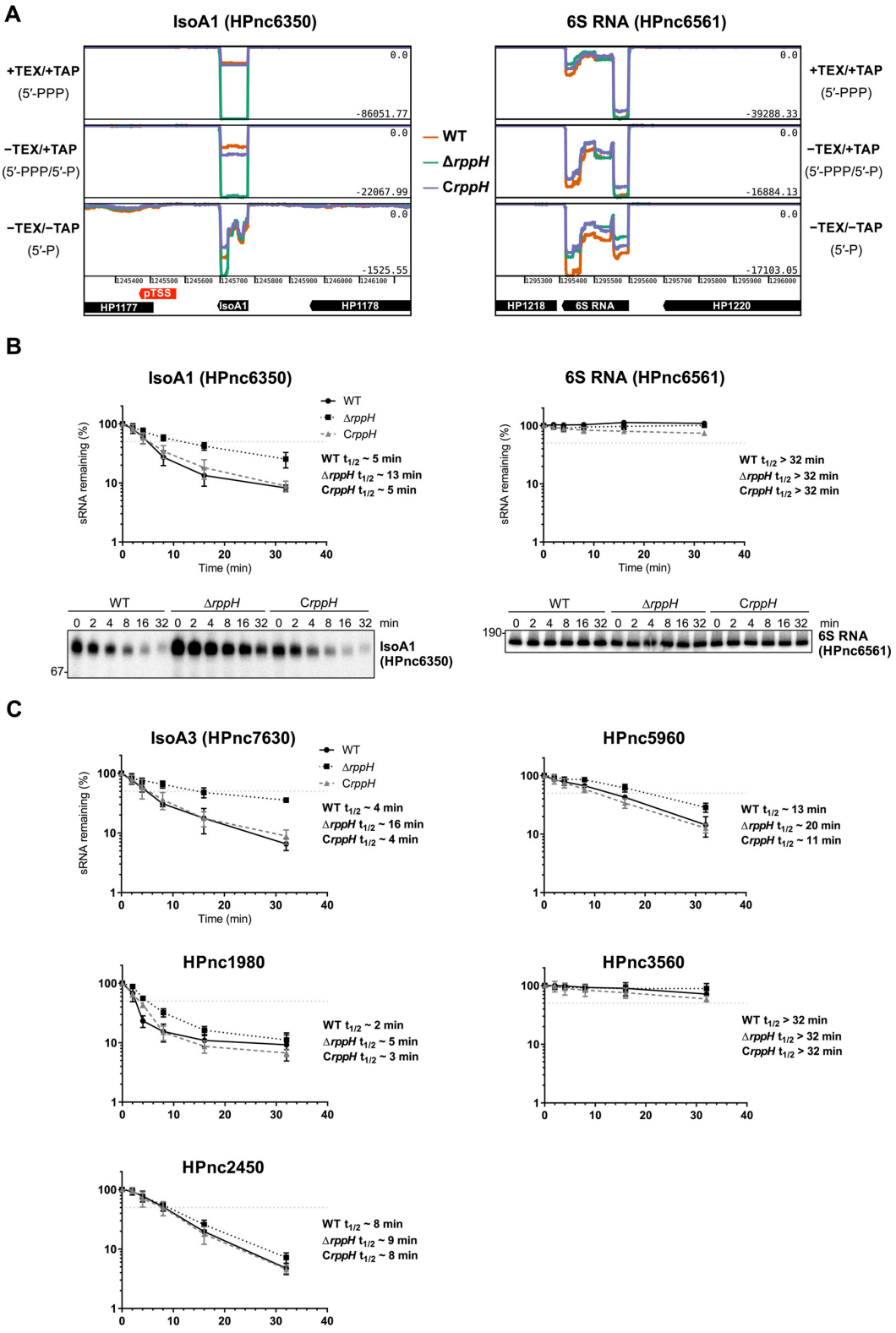
To independently validate these findings, we examined the effect of HpRppH on the degradation rates of several of its putative sRNA targets. This was achieved by treating log-phase cultures of isogenic WT, Δ*rppH*, and C*rppH* strains of *H. pylori* with rifampicin to arrest transcription and unmask degradation. Total RNA was then extracted from the cells at time intervals, and equal amounts were analyzed by Northern blotting. The half-life of IsoA1 sRNA increased from ~5 min in WT cells to ~13 min in Δ*rppH* cells (Fig. 9*B*, *left*). Complementation of the Δ*rppH* mutation with an ectopic copy of the gene (C*rppH*) restored the original 5-min half-life. Several other sRNAs judged by dRNA-seq to be candidate RppH targets, such as IsoA3 (HPnc7630), HPnc1980, and HPnc5960, were also significantly stabilized (1.5–4-fold) in the Δ*rppH* strain, whereas the stability of the long-lived HPnc3560 transcript did not increase noticeably (Fig. 9*C*). No change in lifetime was observed for 6S sRNA (HPnc6561) (Fig. 9*B*, *right*) or HPnc2450 (Fig. 9*C*), which served as negative controls.

mRNAs Targeted by RppH—In addition to potential sRNA targets, we identified 52 potential mRNA targets of HpRppH by dRNA-seq. For example, the *fldA* (HP1161) and *mda66* (HP0630) transcripts, encoding flavodoxin I (FldA) and an NADPH quinone reductase (MdaB), respectively, were more abundant and had a lower ratio of monophosphorylated to triphosphorylated 5′ ends in the Δ*rppH* mutant than in the WT and complemented strains (Fig. 10*A*). Other mRNAs that appeared to be targeted by HpRppH included those encoding cytochrome c_{553} (HP1227, encoded directly adjacent to HpRppH), cell binding factor 2 (HP0175), and outer membrane protein OMP18 (HP1125) (supplemental Table S2). Sensitivity to RppH was not significantly correlated with protein function, as defined by the PyloriGene database (44) (one-sided Fisher's exact test, calculated Benjamini-Hochberg adjusted *p* value >0.10 for every functional category; data not shown).

To corroborate the influence of HpRppH on two of its mRNA targets, we examined its effect on the lifetime and 5′-phosphorylation state of the *fldA* and *mda66* transcripts. First, we compared the half-lives of these mRNAs in cells containing or lacking RppH by using Northern blot analysis to monitor their disappearance after transcription inhibition with rifampicin. The half-lives of these transcripts increased from 7 min (*flaA*) or 10 min (*mda66*) in WT cells to >32 min in Δ*rppH* cells and returned to their original values in C*rppH* cells (Fig. 10*B*).

Next, we investigated the effect of RppH on the 5′-terminal phosphorylation state of these mRNAs by PABLO (phosphorylation assay by ligation of oligonucleotides), a splinted ligation assay specific for monophosphorylated 5′ ends (45, 46). This method is based on the ability of T4 DNA ligase to join a DNA oligonucleotide to a monophosphorylated RNA, but not its triphosphorylated counterpart, when their ends are juxtaposed by annealing them to a bridging oligonucleotide complementary to both. The percentage of the transcript that is monophosphorylated can then be determined by using denaturing gel electrophoresis and blotting to resolve the ligation product from its unligated counterpart and comparing the ligation yield with that of a fully monophosphorylated control (47). In this manner, we determined that a significant fraction of both *fldA* mRNA (27%) and *mda66* mRNA (16%) is monophosphorylated at steady state in WT cells and that this percentage declines to only 3–5% in Δ*rppH* cells (Fig. 10*C*). The percentage of these transcripts that was monophosphorylated was restored to normal by complementation of the genetic defect. Together, these findings confirm that *fldA* and *mda66* mRNA are direct targets of RppH and are degraded in *H. pylori* by an RppH-dependent mechanism.

Functional Characterization of H. pylori RppH



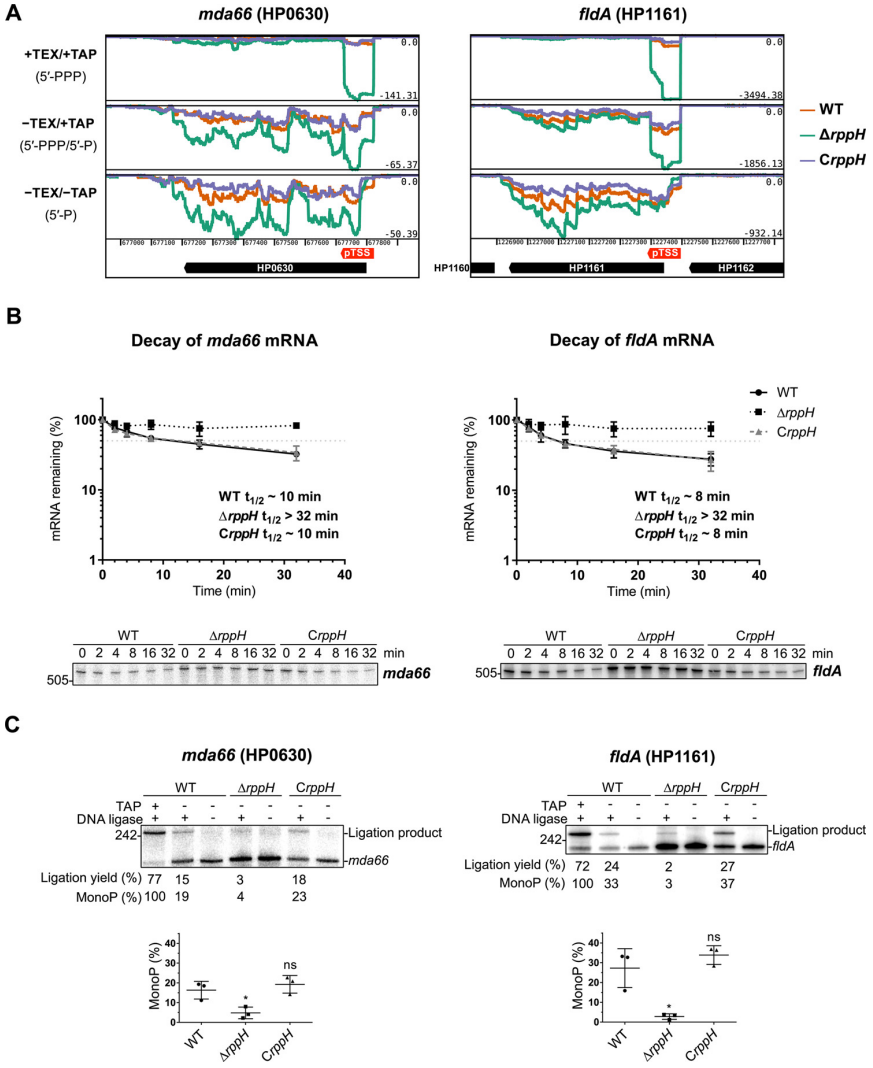


FIGURE 10. **mRNA targets of HpRppH.** *A*, screen shots of RNA-seq data for the HpRppH targets *mda66* mRNA (HP0630) and *fldA* mRNA (HP1161) in WT, Δ*rppH*, and C*rppH* cells, as visualized by using Artemis (56). *B*, half-lives of *mda66* mRNA (~621 nt long) and *fldA* mRNA (~548 nt long) in *H. pylori*. RNA degradation was monitored by Northern blotting analysis of equal amounts of total RNA extracted from WT, Δ*rppH*, and C*rppH* cells at various times after the addition of rifampicin to log-phase cultures. Data from three biological replicates of each of the three strains were averaged, and half-lives ($t_{1/2}$) were determined from the time at which 50% of the mRNA remained (*light gray dotted lines*). *C*, phosphorylation state of *mda66* and *fldA* mRNA in *H. pylori*. Total RNA extracted from WT, Δ*rppH*, and C*rppH* cells was examined by PABLO analysis to determine the 5′-phosphorylation state of the transcripts *in vivo*. *Top*, representative PABLO assays. RNA samples that had first been treated *in vitro* with TAP were analyzed in parallel so that the ligation yields of fully monophosphorylated transcripts could be used as correction factors for calculating the percentage of *mda66* and *fldA* that was monophosphorylated. *Bottom*, scatter plots showing the average of three independent PABLO experiments. *Error bars*, S.D. Student's *t* test was used for statistical comparison of the Δ*rppH* and C*rppH* data with the WT data. *, statistically significant difference ($p \leq 0.05$); *ns*, not significant ($p > 0.05$).

Discussion

In bacteria, RNA degradation typically commences by either of two mechanisms: 1) direct access of a ribonuclease to cleavage sites within transcripts or 2) 5′-end-dependent access in which RNA cleavage by a ribonuclease is facilitated by prior conversion of the 5′-terminal triphosphate to a monophosphate by an RNA pyrophosphohydrolase (3). Here we have identified the Nudix protein HP1228 as an RNA pyrophosphohydrolase important for RNA degradation in *H. pylori*, characterized its biochemical activity and substrate specificity *in vitro*, and identified several of its mRNA and sRNA targets *in vivo* by employing a global strategy based on high-throughput

FIGURE 9. **sRNA targets of HpRppH.** *A*, screen shots of RNA-seq data for the HpRppH target IsoA1 sRNA (HPnc6350) and the non-target 6S RNA (HPnc6561) in WT, Δ*rppH*, and C*rppH* cells, as visualized by using Artemis (56). *B*, half-lives of IsoA1 sRNA (~80 nt long) and 6S RNA (~180 nt long) in *H. pylori*. RNA degradation was monitored by Northern blotting analysis of equal amounts of total RNA extracted from WT, Δ*rppH*, and C*rppH* cells at various times after the addition of rifampicin to log-phase cultures. Data from four biological replicates of each of the three strains were averaged, and half-lives ($t_{1/2}$) were determined from the time at which 50% of the RNA remained (*light gray dotted lines*). *Error bars*, S.D. *C*, half-lives of additional sRNAs (HPnc7630, HPnc1980, HPnc5960, HPnc3560, and HPnc2450) in *H. pylori*, based on three biological replicates each.

sequencing. In view of these properties and the homology of HP1228 to *E. coli* RppH (EcRppH), we have renamed it HpRppH. Our findings suggest an important role for RppH in governing gene expression not only in *H. pylori* but also in other pathogenic Epsilonproteobacteria, where orthologs of this enzyme are ubiquitous.

Using *in vitro* assays, we have demonstrated that HpRppH converts triphosphorylated RNA 5′ ends to monophosphorylated ends while yielding a mixture of pyrophosphate and orthophosphate as by-products. The same two by-products are generated by EcRppH, albeit in a ratio that is more biased toward pyrophosphate (13), whereas BsRppH produces only orthophosphate (11), presumably by removing the γ- and β-phosphates consecutively. One other *H. pylori* protein, HP0507, may have a fold resembling a Nudix domain, as it contains a partial Nudix motif with matches at 4 of 9 positions. This protein has been implicated in *H. pylori* virulence (30), and orthologs appear to be present in other Epsilonproteobacteria and in *E. coli*. However, even at a high concentration, HP0507 exhibited no detectable RNA pyrophosphohydrolase activity when purified and assayed *in vitro*.

Like EcRppH (16) and BsRppH (19), HpRppH requires at least two unpaired nucleotides at the 5′ end of its substrates and prefers three or more. The purified enzyme is rather promiscuous with respect to the identity of those 5′-terminal nucleotides, although it has a slight preference for A over G at the first position and for a purine over a pyrimidine at the second position, properties shared by EcRppH (16) but not BsRppH (19), which strictly requires a G at the second position. The difference in specificity between the proteobacterial enzymes and BsRppH is explained by dissimilarities in the amino acid residues that line the pocket where the second nucleotide binds to each of these proteins (16, 28, 31), residues that are almost identical in HpRppH (Arg-30, Ala-36, Val-135, Phe-137, Lys-138) and EcRppH (Arg-27, Ser-32, Val-137, Phe-139, Lys-140) but very different in BsRppH (Asp-6, Tyr-86, Val-88, Ile-95, Lys-97, Phe-137, Ile-138, and Asp-141). Among these amino acids, the sole difference between the two proteobacterial enzymes is a residue (Ala-36 in HpRppH, Ser-32 in EcRppH) that contacts the Watson-Crick edge of the second nucleobase of the RNA ligand in X-ray crystal structures of EcRppH and contributes to the promiscuity of that ortholog (16, 31). The similarity of the substrate preferences of HpRppH and EcRppH despite their overall sequence divergence (34% identity) suggests that the many other proteobacterial and plant orthologs of these two enzymes are likely to share these properties.

To identify transcripts targeted by HpRppH in *H. pylori*, we employed a global dRNA-seq strategy in which three distinct enzymatic treatments were used to selectively enrich RNAs bearing a 5′-triphosphate and/or a 5′-monophosphate. By examining the effect of an *rppH* deletion on the number of 5′ ends that were triphosphorylated or monophosphorylated in *H. pylori*, we identified 53 mRNAs and 10 sRNAs whose degradation appears to be triggered by this enzyme (supplemental Table S2). Several of them were further validated by half-life measurements and PABLO analysis. To be classified as candidate RppH targets, transcripts had to fulfill two criteria in Δ*rppH* cells *versus* WT and C*rppH* cells: 1) a ≥2-fold increase in their cellular concentration and 2) a ≥50% decline in the ratio of monophosphorylated to triphosphorylated 5′ ends. These strict selection criteria were chosen to maximize the likelihood that only transcripts directly and productively targeted by HpRppH would be identified. Nevertheless, because of statistical uncertainty, the ≥2-fold effect used as a threshold, and the fact that only one growth condition was tested, it seems probable that HpRppH triggers the degradation of many additional *H. pylori* transcripts besides those identified here. Potential RppH targets whose concentration increased ≥2-fold in Δ*rppH* cells but whose phosphorylation state did not change sufficiently to satisfy the other requirement are listed in supplemental Table S3. For many of these 119 additional RNAs, the number of monophosphorylated 5′ ends detected in the −TEX/−TAP libraries may have been too low to be accurately quantified due to the susceptibility of such intermediates to rapid degradation.

HpRppH seems to target only a subset of *H. pylori* transcripts, as not all of the 925 5′ ends that were examined (second sheet of supplemental Table S2) satisfied the screening criteria. Therefore, although it is theoretically possible that this bacterial species contains a second, non-redundant RNA pyrophosphohydrolase, as has been proposed for *B. subtilis* and *Staphylococcus aureus* (19, 48), it is likely that a large number of *H. pylori* RNAs undergo rapid degradation by pathways that do not require prior conversion of the 5′-triphosphate to a monophosphate. Consistent with the existence of RppH-independent RNA decay pathways is the fact that *rppH* is not an essential gene in *H. pylori*, although its deletion reduces the growth rate of *H. pylori* 26695 by about one-third (data not shown).

The preference of purified HpRppH for a purine at the second position of its substrates is not reflected in the sequences at the 5′ end of the 63 candidate HpRppH targets identified *in vivo*, where there is a modest bias in favor of U at the expense of A and C at the second position (A:G:C:U (targeted transcripts/all transcripts) = 0.13/0.24 : 0.05/0.07 : 0.13/0.19 : 0.70/0.50 at position 2). For example, among the targets that were validated individually, IsoA1 and IsoA3 both have a purine (A) at position 2, whereas *mda66*, *fldA*, HPnc1980, and HPnc5960 each have a pyrimidine there (U, C, U, or U, respectively). This finding suggests that *H. pylori* transcripts degraded by a 5′-end-dependent mechanism have evolved not to maximize the RppH reaction rate but rather to allow sequence-dependent variations in that rate to contribute to differences in RNA lifetimes.

The fate of the monophosphorylated decay intermediates generated by RppH depends on the organism in which they are produced, as different bacterial species often have distinct ribonucleolytic arsenals (3). For example, *E. coli* and *B. subtilis* not only contain dissimilar RNA pyrophosphohydrolases but also utilize different sets of ribonucleases to degrade RNA. In *E. coli*, monophosphorylated decay intermediates are rapidly degraded by RNase E, a 5′-monophosphate-assisted endonuclease, whereas in *B. subtilis* they are degraded by RNase J, a 5′-monophosphate-dependent 5′-exonuclease (10, 11, 13, 49). *H. pylori* represents an interesting amalgam of those two species. Like *E. coli*, it is a proteobacterium, and it therefore contains an ortholog of EcRppH. However, as an epsilonproteobacterium, other aspects of RNA turnover in *H. pylori* more closely

resemble *B. subtilis*, as it lacks RNase E and instead is thought to utilize two other ribonucleases, RNase J and the endonuclease RNase Y, to degrade RNA (5, 8, 9). As a result, it is likely that the monophosphorylated decay intermediates generated by HpRppH are degraded exonucleolytically by RNase J, probably with help from RhpA, a DE*X*D-box RNA helicase with which RNase J forms a complex in *H. pylori* (8). Indeed, >80% of the likely and possible RppH targets that were previously examined for RNase J sensitivity (5, 8, 9) appear to be degraded by an RNase J-dependent mechanism (supplemental Tables S2 and S3). RNase J is also capable of functioning as an endonuclease (8), but this activity is not dependent on the 5′-phosphorylation state of RNA (11) and therefore is unlikely to contribute significantly to the degradation of transcripts productively targeted by RppH.

Previous studies have reported that HpRppH is constitutively expressed in *H. pylori* at various stages of growth and during stress (29) and that *H. pylori* Δ*rppH* mutants have a diminished capacity to invade gastric epithelial adenocarcinoma cells (32) and to survive hydrogen peroxide exposure (29). The latter two phenotypes probably are consequences of altered patterns of gene expression resulting from the increased stability of RNAs ordinarily targeted by RppH, and they illustrate the physiological importance of 5′-end deprotection by this enzyme. The fact that HpRppH is the only known *H. pylori* protein with a *bona fide* Nudix motif suggests that, of all of the metabolic functions of bacterial Nudix hydrolases (17), this may well be the most important.

Experimental Procedures

Protein Structure Prediction—A detailed structural model of HpRppH was generated on the basis of sequence homology to EcRppH by using a high-resolution X-ray crystal structure of EcRppH bound to an oligonucleotide ligand and two Mg^{2+} ions (Protein Data Bank code 4S2X) (31) as a template. The calculations were performed with SWISS-MODEL software (50) on the ExPASy bioinformatics website. PyMOL (51) was utilized to prepare figures from the resulting atomic coordinates.

In Vitro Assays of RNA Pyrophosphohydrolase Activity and Specificity—HpRppH (HP1228), HpRppH-E57Q, and HP0507, each bearing an amino-terminal hexahistidine tag, were produced in *E. coli*, purified by affinity chromatography on TALON beads (Clontech), and assayed for RNA pyrophosphohydrolase activity as described previously (13). Triphosphorylated *rpsT* P1 RNA bearing a 5′-terminal γ-^{32}P label and an internal fluorescein label and triphosphorylated, diphosphorylated, and monophosphorylated $GA(CU)_{13}$ bearing a single ^{32}P label between the first and second nucleotide were synthesized by *in vitro* transcription (13) and used as substrates in these assays. The specificity of HpRppH was examined as described previously with doubly labeled substrates (γ-^{32}P and fluorescein) prepared by *in vitro* transcription, except that the assays of substrate reactivity were performed in solutions containing 1 mM $MgCl_2$ and 16 nM HpRppH (19). Oligonucleotides and plasmids used to generate the DNA templates used for *in vitro* transcription have been described previously (13, 19, 45).

In Vitro Assays of 8-Oxo-dGTPase Activity—8-Oxo-dGTP or dGTP (50 μM) was combined with various concentrations of purified hexahistidine-tagged HpRppH, HP0507, *E. coli* MutT, *E. coli* RppH, or *B. subtilis* RppH (0, 1, 10, or 100 nM) in 500 μl of a buffer containing 5 mM Tris-HCl, pH 7.4, 1 mM $MgCl_2$, and 1 mM dithiothreitol. After 60 min at 37 °C, the reactions were quenched with EDTA (2 mM final concentration) and then concentrated to 5 μl by evaporation. The reaction products were separated by thin layer chromatography on fluorescent PEI-cellulose plates and visualized by irradiating the plates with ultraviolet light.

H. pylori Growth Conditions—*H. pylori* strains were grown on GC-agar (Oxoid) plates supplemented with 10% (v/v) donor horse serum (Biochrom AG), 1% (v/v) vitamin mix, 10 μg/ml vancomycin, 5 μg/ml trimethoprim, and 1 μg/ml nystatin. For transformant selection and growth of mutant strains, 20 μg/ml kanamycin or 16 μg/ml chloramphenicol were added. For liquid cultures, 10 or 50 ml of brain heart infusion (BHI) medium (BD Biosciences) supplemented with 10% (v/v) FBS (Biochrom AG) and 10 μg/ml vancomycin, 5 μg/ml trimethoprim, and 1 μg/ml nystatin were inoculated with *H. pylori* from a plate to a final A_{600} of 0.02–0.05 and grown under agitation at 140 rpm in 25- or 75-cm^3 cell culture flasks (PAA). Bacteria were grown at 37 °C in a HERAcell 150i incubator (Thermo Scientific) in a microaerophilic environment (10% CO_2, 5% O_2, and 85% N_2). *E. coli* strains were grown in Luria-Bertani (LB) medium supplemented with 100 μg/ml ampicillin, 20 μg/ml chloramphenicol, and/or 20 μg/ml kanamycin if applicable. Details about the generation of *H. pylori* mutant strains are provided below.

Construction of H. pylori Mutant Strains—All mutant strains were generated by natural transformation and homologous recombination of PCR-amplified constructs carrying either the *aphA*-3 kanamycin (37) or the *catGC* chloramphenicol resistance cassette (52) flanked by ~500-bp regions of homology upstream and downstream of the respective genomic locus, as described previously. Briefly, *H. pylori*, grown from frozen stocks until passage two, was streaked in small circles on a fresh plate and grown for 6–8 h at 37 °C under microaerophilic conditions. For transformation, 0.5–1.0 μg of purified PCR product was added to the cells. After incubation for 14–16 h at 37 °C, cells were restreaked on selective plates containing the indicated antibiotics. The genotypes of mutants were verified by PCR amplification and sequencing of genomic DNA isolated using the NucleoSpin plasmid kit (Macherey-Nagel, Bethlehem, PA). Table 2 lists all oligonucleotides used for cloning.

Construction of H. pylori rppH Deletion and Complementation Strains—To construct the *rppH* deletion strain, *H. pylori* 26695 ΔHP1228::*Kan*^{R} (CSS-0091, Δ*rppH* from 26695), overlap extension PCR was used to assemble a DNA fragment containing a non-polar *Kan*^{R} (*aphA*-3) cassette (37) flanked on one side by the first three codons of HP1228 (*rppH*) and ~500 additional upstream base pairs and on the other side by the last three codons of HP1228 and ~500 additional downstream base pairs. First, ~500 bp upstream of HP1228 codon 4 were amplified from genomic DNA of wild-type *H. pylori* 26695 (CSS-0065, kindly provided by D. Scott Merrell) using primers CSO-0121/-0122, and ~500 bp downstream of HP1228 codon 152 (the fourth to last codon) were amplified using primers CSO-0123/-0124. The *Kan*^{R} cassette was amplified using primers HPK1 and HPK2. The purified PCR products, corresponding to regions

TABLE 2
DNA oligonucleotides used in this study

Name	DNA sequences (5′–3′)	Description
CSO-0017	GTTTTTTCTAGAGATCAGCCTGCCTTTAGG	Cloning of *H. pylori* 26695 *rppH* complementation
CSO-0018	GTTTTTCTCGAGCTTAGCGCTTAATGAAACGC	Cloning of *H. pylori* 26695 *rppH* complementation
CSO-0033	GCATTTGAGCAAAAGAGGG	Verification of *H. pylori* 26695 *rppH* complementation
CSO-0034	GGCAAATCTTTAACCCTTTTG	Verification of *H. pylori* 26695 *rppH* complementation
CSO-0121	ACTTGTAATTGTATCATTTTAAGATCATT	Deletion of *H. pylori* 26695 *rppH*
CSO-0122	CTCCTAGTTAGTCACCCGGGTACATGTAGCATAGGTCTTTATTTTAGCT	Deletion of *H. pylori* 26695 *rppH*
CSO-0123	TGTTTTAGTACCTGGAGGGAATATATTTATAGGGTGTTAATCGTTCAA	Deletion of *H. pylori* 26695 *rppH*
CSO-0124	CCGTATAGATTTCGCACAAAT	Deletion of *H. pylori* 26695 *rppH*
CSO-0125	GGGATATGAATGTATAAAATCATATTTATT	Verification of *H. pylori* 26695 *rppH* deletion
CSO-0146	GTTTTTATCGATGTATGCTCTTTAAGACCCAGC	Cloning of *H. pylori* 26695 *rppH* complementation
CSO-0147	GTTTTTCATATGCTCGAATTCAGATCCACGTT	Cloning of *H. pylori* 26695 *rppH* complementation
CSO-0148	GTTTTTATCGATCATCAAAGCTTTAGCCAAATACAT	Cloning of *H. pylori* 26695 *rppH* complementation
CSO-0149	GTTTTTCATATGCCGTATTTTTGAACGATTAACAC	Cloning of *H. pylori* 26695 *rppH* complementation
CSO-0505	GTTCATAGCCTTTATCCACGA	Northern blotting probe for HP0630 (*mda66*) mRNA
CSO-1038	GTTCCCGCTGTCTGTCCC	Northern blotting probe for HP1161 (flavodoxin) mRNA
CSO-2298	CCGCTTTTAGCGAATGCTTGTCAAGTTATCATTCATATTGTTC	Y oligonucleotide for HP1161 for PABLO assay
CSO-2299	AAAAAAAAAAGAACAATATGAATGATAACTTG	X_{32} oligonucleotide for PABLO assay
CSO-2300	CAATCTGTTTGGGCTAGCTACAACGAAAATCACCCG	10–23 DNAzyme for PABLO assay of HP1161 mRNA
CSO-2301	AAATCGTCGCAGGCTAGCTACAACGACAGCGCTAAA	10–23 DNAzyme for PABLO assay of HP0630 mRNA
CSO-2302	TTCCTTTTCTAATAAAATAGCAAGTTATCATTCATATTGTTC	Y oligonucleotide for HP0630 for PABLO assay
HPK1	GTACCCGGGTGACTAACTAGG	Amplification of *aphA-3* cassette
HPK2	TATTCCCTCCAGGTACTAAAACA	Amplification of *aphA-3* cassette
JVO-0231	GAGTTTGTCATGGCTACCAA	Northern blotting probe for IsoA1
JVO-0514	CATGCCATGAAACACAAAAG	Northern blotting probe for IsoA3
JVO-2136	AACACGAATCATCTAGGCGAT	Northern blotting probe for 6S rRNA
JVO-2635	CGAGAAATACCTCCACACAAT	Northern blotting probe for HPnc2450
JVO-2715	ATCATATCTTATAAAGGCGTAACTTT	Northern blotting probe for HPnc1980, HPnc1990
JVO-3928	CTAATCATTTCTAAATCATGCTCG	Northern blotting probe for HPnc5960
JVO-3938	TCCTTATGGCTCAATTACAAGG	Northern blotting probe for HPnc3560
JVO-5257	TATAGGTTTTCATTTTCTCCCAC	Verification of *H. pylori* 26695 *rppH* deletion
pZE-A	GTGCCACCTGACGTCTAAGA	Colony PCR on pZE12-derived plasmids

upstream and downstream of HP1228 as well as the Kan^R cassette, were mixed at an equimolar ratio and subjected to overlap extension PCR using primers CSO-0121/-0124. The resulting deletion construct was gel-purified and substituted into the chromosome of CSS-0065 by transformation (natural competence) and recombination, yielding CSS-0091 (ΔHP1228::Kan^R). Positive clones from CSS-0091 were verified by PCR with primers CSO-0125 and JVO-5257.

To generate an *rppH* complementation strain, the *rppH* gene and ~200 additional base pairs on each side of it were amplified from genomic DNA of *H. pylori* 26695 (CSS-0065) using oligonucleotides CSO-0148/-0149. The PCR product was digested with NdeI (New England Biolabs, catalog no. R0111L) and ClaI (New England Biolabs, catalog no. R0197L). At the same time, plasmid pSP39-3 (41) was amplified using oligonucleotides CSO-0146/-0147 and, after digestion with DpnI, analogously digested with NdeI and ClaI and subsequently dephosphorylated with calf intestinal phosphatase (New England Biolabs, catalog no. M0290L). The PCR products of the plasmid backbone and of the *rppH* gene were purified, ligated, and transformed into *E. coli* Top 10 cells (CSS-0296, Invitrogen), yielding plasmid pSS4-2. Positive clones were selected on plates containing 100 μg/ml ampicillin and confirmed by colony PCR using oligonucleotides pZE-A/CSO-0017. Plasmid pSS4-2 contains both the *rppH* gene under the control of its own promoter and the *catGC* resistance cassette (52), flanked by the 5′ and 3′ parts of the *rdxA* locus, respectively. A PCR product amplified from pSS4-2 with oligonucleotides CSO-0017/-0018 was used for complementation of *H. pylori* 26695 ΔHP1228::Kan^R (CSS-0091), resulting in strain CSS-0148 (ΔHP1228::Kan^R; Δ*rdxA*::HP1228-*catGC*^{R}), which contains the *rppH* gene in an antisense orientation relative to the *catGC* cassette and the *rdxA* gene. Positive clones from CSS-0148 were verified by PCR with primers CSO-0034/-0148 and sequencing with CSO-0033.

RNA Isolation—Unless stated otherwise, *H. pylori* was grown in liquid culture to logarithmic phase (A_{600} ~1), and cells corresponding to an A_{600} of 4 were harvested, mixed with 0.2 volumes of stop mix (95% (v/v) EtOH, 5% (v/v) phenol), and immediately shock-frozen in liquid nitrogen. Frozen cell pellets were thawed on ice, centrifuged for 10 min at 3,250 × *g* at 4 °C, and resuspended in TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0) containing 0.5 mg/ml lysozyme and 1% (w/v) SDS. RNA was extracted using the hot phenol method as described and treated with DNase I (New England Biolabs) according to the manufacturer's instructions (35).

Examination of RNA Stability and Northern Blotting Analysis—To determine the stability of mRNAs and sRNAs in the various *H. pylori* strains, cells were grown to an A_{600} of ~1 and treated with rifampicin (final concentration, 500 μg/ml). Equal volumes of cells (5 ml) were withdrawn 0, 2, 4, 8, 16, and 32 min after the addition of rifampicin and immediately mixed with 0.2 volumes of stop solution (5% water-saturated phenol, 95% ethanol). The cells were promptly frozen in liquid nitrogen and stored at −80 °C until use. Total cellular RNA was isolated by the hot phenol method. For Northern blot analysis, 10 μg of total RNA were subjected to gel electrophoresis on 6% (v/v) polyacrylamide gels containing 7 M urea. RNA was subsequently transferred to a Hybond-XL membrane (GE Healthcare) by electroblotting and then UV-crosslinked to the membrane. Transcripts were detected by probing with 5′-end-labeled (γ-^{32}P) oligodeoxynucleotide probes complementary to specific RNAs of interest, as described (35). Radioactive bands were visualized with a Fuji FLA-3000 imager, and the band

intensities were quantified by using AIDA Image Analyzer software version 4.27 (Raytest).

PABLO—Total cellular RNA was extracted from various *H. pylori* mutant strains by the hot phenol procedure (35). As a control for the PABLO assay, a sample of total RNA from WT cells was treated with TAP to create fully monophosphorylated RNA, as described (46). Briefly, 50 μg of total WT RNA was combined in 44 μl of water with 5 μl of 10× TAP reaction buffer (Epicenter, catalog no. T19500), 1 μl of RNase inhibitor (Molox), and 0.5 μl of TAP (Epicenter, catalog no. T19500). This mixture was incubated at 37 °C for 2 h. Subsequently, 150 μl of autoclaved water was added to facilitate phenol extraction. The products were phenol-extracted once with water-equilibrated phenol and ethanol-precipitated. The pellets were washed with 75% ethanol and air-dried. The RNA was then resuspended in 25 μl of autoclaved water. After that, PABLO analysis was performed, using a portion of the TAP-treated RNA sample as a positive control, as described (46). For the assay, 15 μg of DNase I-treated total cellular RNA per reaction was combined with 2 μl of 10 μM oligonucleotide X_{32} (CSO-2299) and 4 μl of 1 μM oligonucleotide Y (CSO-2298 for HP1161, CSO-2302 for HP0630). To improve electrophoretic resolution of the ligation product, 4 μl of a 100 μM solution of a site-specific 10–23 DNAzyme oligonucleotide were included as well (CSO-2300 for HP1161, CSO-2301 for HP0630) (46). Water was added to bring the final volume to 45 μl. The samples were heated at 75 °C for 5 min and then cooled gradually to 30 °C before being placed on ice. A premixture (35 μl) containing the following components was added to each sample of RNA complexed with oligonucleotides X_{32} and Y: 10 μl of T4 DNA ligase (catalog no. M0202, New England Biolabs), 1 μl of RNase inhibitor (Molox), 8 μl of 10× ligation buffer (catalog no. M0202, New England Biolabs), 1.6 μl of 10 mM ATP, and 14.4 μl of H_2O. The resulting mixtures were incubated at 37 °C for 4 h and subsequently placed on ice. The ligation reactions were quenched by adding 120 μl of 10 mM EDTA, and the products were phenol-extracted and ethanol-precipitated. The pellets were washed with 75% ethanol and air-dried. The pellets containing the ligation products were dissolved in 5 μl of water, combined with 15 μl of RNA loading buffer (95% (v/v) formamide, 20 mM EDTA (pH 8.0), 0.025% (w/v) bromphenol blue, 0.025% (w/v) xylene cyanol), and heated at 95 °C for 5 min. Electrophoresis was performed on a 6% polyacrylamide gel containing 7 M urea. The gel was electroblotted onto a Hybond-XL membrane (GE Healthcare), and after UV cross-linking, the membrane was probed with radiolabeled DNA complementary to the transcript of interest. Radioactive bands corresponding to ligated and unligated RNA were visualized with a Fuji FLA-3000 imager, and ligation yields were calculated from the measured band intensities (yield = ligated/(unligated + ligated)) using AIDA software (Raytest, Germany).

cDNA Library Preparation and Deep Sequencing—RNA-seq libraries were constructed from total RNA samples harvested in logarithmic growth phase (WT A_{600} 0.7; Δ*rppH* A_{600} 0.5; C*rppH* A_{600} 0.7) in BHI medium. Residual genomic DNA was removed from the isolated total RNA by DNase I treatment. cDNA library preparation was performed by Vertis Biotechnology AG in a strand-specific manner as described previously for eukaryotic microRNA (53) but omitting the RNA size fractionation step before cDNA synthesis. In brief, the three RNA samples were each split into three portions. One portion was treated with TEX before the standard library preparation procedure described below to generate the +TEX/+TAP libraries. To this end, RNA was denatured for 2 min at 90 °C, cooled on ice for 5 min, and treated with 1.5 units of TEX (Epicenter) for 30 min at 30 °C. For the second portion, the TAP treatment (see below) was omitted to generate the −TEX/−TAP libraries. The standard procedure without modifications was used to generate the −TEX/+TAP libraries from the third portion. Here, ~200 ng of RNA sample were poly(A)-tailed using 2.5 units of *E. coli* poly(A) polymerase (New England Biolabs) for 5 min at 37 °C. The 5′-triphosphates were then converted to monophosphates with TAP. TAP treatment was performed by incubating the samples with 5 units of TAP for 15 min at 37 °C. Afterward, an RNA adapter (5′ Illumina sequencing adapter, 5′-UUUCCCUACACGACGCUCUUCCGAUCU-3′) was ligated to the 5′-P of the TAP-treated, poly(A)-tailed RNA for 30 min at 25 °C. First-strand cDNA was synthesized by using an oligo(dT)-adapter primer (see below) and Moloney murine leukemia virus reverse transcriptase (AffinityScript, Agilent) by incubation at 42 °C for 20 min, ramping to 55 °C, and further incubation at 55 °C for 5 min. In a PCR-based amplification step using a high-fidelity DNA polymerase (Herculase II Fusion DNA polymerases, Agilent), the cDNA concentration was increased to 20–30 ng/μl (initial denaturation at 95 °C for 2 min, followed by 14–16 cycles at 95 °C for 20 s and 68 °C for 2 min). A library-specific barcode for multiplex sequencing was included as part of a 3′-sequencing adapter. The TruSeq index primers for PCR amplification were used according to the instructions of Illumina. For all libraries, the Agencourt AMPure XP kit (Beckman Coulter Genomics) was used to purify the DNA (1.8× sample volume), and cDNA sizes were examined by capillary electrophoresis on a MultiNA microchip electrophoresis system (Shimadzu).

The following adapter sequences flanked the cDNA inserts: TrueSeq_Sense_primer, 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′; TrueSeq_Antisense_NNNNNN_primer (where NNNNNN represents the 6n barcode for multiplexing), 5′-CAAGCAGAAGACGGCATACGAGAT-NNNNNN-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC$(dT)_{25}$-3′. All libraries were sequenced using an Illumina HiSeq 2000 machine with 97 cycles in single-read mode.

Data Processing and Availability—To ensure high sequence quality, the Illumina reads in FASTQ format were trimmed with a cutoff phred score of 20 by the program *fastq_quality_trimmer* from FASTX toolkit version 0.0.13. Subsequent processing steps were conducted using the RNA-seq analysis pipeline READemption version 0.4.2 (54). These consisted of poly(A) tail removal followed by size filtering to keep only reads with a minimum length of 12 nt. Remaining reads from all libraries were mapped to the *H. pylori* 26695 reference genome (NC_000915.1) using *segemehl* version 0.2.0-418 (55). Read mapping statistics are summarized in Table 1.

Coverage plots representing the numbers of mapped reads per nucleotide were generated. Reads that mapped to multiple

(n) locations with an equal score contributed fractionally ($1/n$) to the coverage value. Each resulting coverage graph was normalized by the number of reads that could be mapped from the respective library (typically several million reads when using Illumina sequencing) and then multiplied by the minimum number of mapped reads calculated over all libraries. Coverage plots were visualized using Artemis (56).

Expression analysis for TSS windows as well as sRNA and housekeeping RNA annotations was also conducted using READemption. Here, read overlap counts for −TEX/+TAP libraries were calculated based on 100-nt windows encompassing previously annotated primary and secondary TSSs for mRNAs, tRNAs, and rRNAs (42) together with their downstream regions and using full-length annotations for sRNAs and housekeeping RNAs (35). Each read with a minimum overlap of 10 nt was counted with a value based on the number of locations where the read was mapped. If the read overlapped more than one annotation, the value was divided by the number of annotations and counted separately for each of them (*e.g.* ⅓ for a read mapped to three locations). For +TEX/+TAP and −TEX/−TAP libraries, read 5′ ends (first base only) matching to a region from 5 nt upstream to 4 nt downstream of each TSS were counted with a value based on the number of locations where the read was mapped but without considering overlap with more than one annotation. Read counts for +TEX/+TAP and −TEX/−TAP libraries were normalized as described above for the coverage plots. Size factors corresponding to this normalization were used for the pairwise Gfold comparison of −TEX/+TAP counts from WT and Δ*rppH* as well as C*rppH* and Δ*rppH* but were rescaled by the software, resulting in slightly different values for each comparison.

Raw sequencing reads in FASTQ format and normalized coverage files in wiggle (WIG) format are available via the Gene Expression Omnibus under accession number GSE86943. Two of the RNA-seq libraries have already been published in a previous study, where the TSS data used in the current analysis was also generated (42). These were the +TEX/+TAP and −TEX/+TAP libraries from the WT sample, which were used as a replicate for the differential RNA-seq approach described in the former publication.

Comparison between RppH and RNase J Targets—To analyze the overlap between HpRppH and RNase J targets, we extracted sequences for all *H. pylori* 26695 genes (protein-coding regions, tRNAs, and rRNAs) that were used to define TSSs in previous studies (35, 42) and for the sRNAs/housekeeping RNAs discovered at that time (35). Sequences for all *H. pylori* B8 genes used to identify RNase J targets (9) were downloaded from the MicroScope platform (57) in FASTA format. Orthologous genes in the two strains were identified by using Ortholuge software (58) while taking care to analyze sRNAs/housekeeping RNAs separately from other RNAs to avoid erroneous mappings between different RNA classes. Next, the reciprocal best BLAST matches in the *in1in2.out* files were combined and used to map identified B8 homologs to the *H. pylori* 26695 transcripts assessed in this study. As described previously (9), B8 annotations for which RNase J depletion resulted in a ≥2-fold increase in transcript concentration with an adjusted p value ≤0.05 were considered RNase J targets. Overlapping and non-overlapping target genes are identified in supplemental Tables S2 and S3.

Author Contributions—J. G. B. and C. M. S. designed the research; M. R., P.-K. H., Q. L., H. S. T., P. L. F., and A. H. performed the experiments; T. B. analyzed the data; and J. G. B., C. M. S., M. R., and T. B. wrote the manuscript.

Acknowledgments—We thank Konrad U. Förstner (SysMed Core Unit, University of Würzburg, Germany) for help with deep sequencing analysis, Richard Reinhardt (Max Planck Genome Centre, Cologne, Germany) for help with deep sequencing, and Stephanie Stahl (University of Würzburg) for help with the cloning of H. pylori mutants.

References

1. Cover, T. L., and Blaser, M. J. (2009) *Helicobacter pylori* in health and disease. *Gastroenterology* **136,** 1863–1873
2. van Amsterdam, K., van Vliet, A. H., Kusters, J. G., and van der Ende, A. (2006) Of microbe and man: determinants of *Helicobacter pylori*-related diseases. *FEMS Microbiol. Rev.* **30,** 131–156
3. Hui, M. P., Foley, P. L., and Belasco, J. G. (2014) Messenger RNA degradation in bacterial cells. *Annu. Rev. Genet.* **48,** 537–559
4. Kaberdin, V. R., Singh, D., and Lin-Chao, S. (2011) Composition and conservation of the mRNA-degrading machinery in bacteria. *J. Biomed. Sci.* **18,** 23
5. Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., *et al.* (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388,** 539–547
6. Mathy, N., Hébert, A., Mervelet, P., Bénard, L., Dorléans, A., Li de la Sierra-Gallay, I., Noirot, P., Putzer, H., and Condon, C. (2010) *Bacillus subtilis* ribonucleases J1 and J2 form a complex with altered enzyme behaviour. *Mol. Microbiol.* **75,** 489–498
7. Shahbabian, K., Jamalli, A., Zig, L., and Putzer, H. (2009) RNase Y, a novel endoribonuclease, initiates riboswitch turnover in *Bacillus subtilis*. *EMBO J.* **28,** 3523–3533
8. Redko, Y., Aubert, S., Stachowicz, A., Lenormand, P., Namane, A., Darfeuille, F., Thibonnier, M., and De Reuse, H. (2013) A minimal bacterial RNase J-based degradosome is associated with translating ribosomes. *Nucleic Acids Res.* **41,** 288–301
9. Redko, Y., Galtier, E., Arnion, H., Darfeuille, F., Sismeiro, O., Coppée, J. Y., Médigue, C., Weiman, M., Cruveiller, S., and De Reuse, H. (2016) RNase J depletion leads to massive changes in mRNA abundance in *Helicobacter pylori*. *RNA Biol.* **13,** 243–253
10. Mackie, G. A. (1998) Ribonuclease E is a 5′-end-dependent endonuclease. *Nature* **395,** 720–723
11. Richards, J., Liu, Q., Pellegrini, O., Celesnik, H., Yao, S., Bechhofer, D. H., Condon, C., and Belasco, J. G. (2011) An RNA pyrophosphohydrolase triggers 5′-exonucleolytic degradation of mRNA in *Bacillus subtilis*. *Mol. Cell* **43,** 940–949
12. Spickler, C., Stronge, V., and Mackie, G. A. (2001) Preferential cleavage of degradative intermediates of *rpsT* mRNA by the *Escherichia coli* RNA degradosome. *J. Bacteriol.* **183,** 1106–1109
13. Deana, A., Celesnik, H., and Belasco, J. G. (2008) The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal. *Nature* **451,** 355–358
14. Belasco, J. G. (2010) All things must pass: contrasts and commonalities in eukaryotic and bacterial mRNA decay. *Nat. Rev. Mol. Cell Biol.* **11,** 467–478
15. Messing, S. A., Gabelli, S. B., Liu, Q., Celesnik, H., Belasco, J. G., Piñeiro, S. A., and Amzel, L. M. (2009) Structure and biological function of the RNA pyrophosphohydrolase BdRppH from *Bdellovibrio bacteriovorus*. *Structure* **17,** 472–481

16. Foley, P. L., Hsieh, P. K., Luciano, D. J., and Belasco, J. G. (2015) Specificity and evolutionary conservation of the *Escherichia coli* RNA pyrophosphohydrolase RppH. *J. Biol. Chem.* **290,** 9478–9486
17. McLennan, A. G. (2006) The Nudix hydrolase superfamily. *Cell Mol. Life Sci.* **63,** 123–143
18. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680
19. Hsieh, P. K., Richards, J., Liu, Q., and Belasco, J. G. (2013) Specificity of RppH-dependent RNA degradation in *Bacillus subtilis*. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 8864–8869
20. She, M., Decker, C. J., Svergun, D. I., Round, A., Chen, N., Muhlrad, D., Parker, R., and Song, H. (2008) Structural basis of Dcp2 recognition and activation by Dcp1. *Mol. Cell* **29,** 337–349
21. Gabelli, S. B., Bianchet, M. A., Xu, W., Dunn, C. A., Niu, Z. D., Amzel, L. M., and Bessman, M. J. (2007) Structure and function of the *E. coli* dihydroneopterin triphosphate pyrophosphatase: a Nudix enzyme involved in folate biosynthesis. *Structure* **15,** 1014–1022
22. Cartwright, J. L., Gasmi, L., Spiller, D. G., and McLennan, A. G. (2000) The *Saccharomyces cerevisiae* PCD1 gene encodes a peroxisomal nudix hydrolase active toward coenzyme A and its derivatives. *J. Biol. Chem.* **275,** 32925–32930
23. Gabelli, S. B., Bianchet, M. A., Ohnishi, Y., Ichikawa, Y., Bessman, M. J., and Amzel, L. M. (2002) Mechanism of the *Escherichia coli* ADP-ribose pyrophosphatase, a Nudix hydrolase. *Biochemistry* **41,** 9279–9285
24. Kang, L. W., Gabelli, S. B., Cunningham, J. E., O'Handley, S. F., and Amzel, L. M. (2003) Structure and mechanism of MT-ADPRase, a nudix hydrolase from *Mycobacterium tuberculosis*. *Structure* **11,** 1015–1023
25. Yagi, T., Baroja-Fernández, E., Yamamoto, R., Muñoz, F. J., Akazawa, T., Hong, K. S., and Pozueta-Romero, J. (2003) Cloning, expression and characterization of a mammalian Nudix hydrolase-like enzyme that cleaves the pyrophosphate bond of UDP-glucose. *Biochem. J.* **370,** 409–415
26. Bessman, M. J., Frick, D. N., and O'Handley, S. F. (1996) The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes. *J. Biol. Chem.* **271,** 25059–25062
27. Mildvan, A. S., Xia, Z., Azurmendi, H. F., Saraswat, V., Legler, P. M., Massiah, M. A., Gabelli, S. B., Bianchet, M. A., Kang, L. W., and Amzel, L. M. (2005) Structures and mechanisms of Nudix hydrolases. *Arch. Biochem. Biophys.* **433,** 129–143
28. Piton, J., Larue, V., Thillier, Y., Dorléans, A., Pellegrini, O., Li de la Sierra-Gallay, I., Vasseur, J. J., Debart, F., Tisné, C., and Condon, C. (2013) *Bacillus subtilis* RNA deprotection enzyme RppH recognizes guanosine in the second position of its substrates. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 8858–8863
29. Lundin, A., Nilsson, C., Gerhard, M., Andersson, D. I., Krabbe, M., and Engstrand, L. (2003) The NudA protein in the gastric pathogen *Helicobacter pylori* is an ubiquitous and constitutively expressed dinucleoside polyphosphate hydrolase. *J. Biol. Chem.* **278,** 12574–12578
30. Guo, B. P., and Mekalanos, J. J. (2002) Rapid genetic analysis of *Helicobacter pylori* gastric mucosal colonization in suckling mice. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 8354–8359
31. Vasilyev, N., and Serganov, A. (2015) Structures of RNA complexes with the *Escherichia coli* RNA pyrophosphohydrolase RppH unveil the basis for specific 5′-end-dependent mRNA decay. *J. Biol. Chem.* **290,** 9487–9499
32. Liu, H., Semino-Mora, C., and Dubois, A. (2012) Mechanism of *H. pylori* intracellular entry: an *in vitro* study. *Front. Cell Infect. Microbiol.* **2,** 13
33. Kim, D., Hong, J. S., Qiu, Y., Nagarajan, H., Seo, J. H., Cho, B. K., Tsai, S. F., and Palsson, B. Ø. (2012) Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling. *PLoS Genet.* **8,** e1002867
34. Maki, H., and Sekiguchi, M. (1992) MutT protein specifically hydrolyses a potent mutagenic substrate for DNA synthesis. *Nature* **355,** 273–275
35. Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F., and Vogel, J. (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464,** 250–255
36. Sharma, C. M., and Vogel, J. (2014) Differential RNA-seq: the approach behind and the biological insight gained. *Curr. Opin. Microbiol.* **19,** 97–105
37. Skouloubris, S., Thiberge, J. M., Labigne, A., and De Reuse, H. (1998) The *Helicobacter pylori* UreI protein is not involved in urease activity but is essential for bacterial survival *in vivo*. *Infect. Immun.* **66,** 4517–4521
38. Croxen, M. A., Sisson, G., Melano, R., and Hoffman, P. S. (2006) The *Helicobacter pylori* chemotaxis receptor TlpB (HP0103) is required for pH taxis and for colonization of the gastric mucosa. *J. Bacteriol.* **188,** 2656–2665
39. Goodwin, A., Kersulyte, D., Sisson, G., Veldhuyzen van Zanten, S. J., Berg, D. E., and Hoffman, P. S. (1998) Metronidazole resistance in *Helicobacter pylori* is due to null mutations in a gene (*rdxA*) that encodes an oxygen-insensitive NADPH nitroreductase. *Mol. Microbiol.* **28,** 383–393
40. Liechti, G., and Goldberg, J. B. (2012) *Helicobacter pylori* relies primarily on the purine salvage pathway for purine nucleotide biosynthesis. *J. Bacteriol.* **194,** 839–854
41. Pernitzsch, S. R., Tirier, S. M., Beier, D., and Sharma, C. M. (2014) A variable homopolymeric G-repeat defines small RNA-mediated posttranscriptional regulation of a chemotaxis receptor in *Helicobacter pylori*. *Proc. Natl. Acad. Sci. U.S.A.* **111,** E501–E510
42. Bischler, T., Tan, H. S., Nieselt, K., and Sharma, C. M. (2015) Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in *Helicobacter pylori*. *Methods* **86,** 89–101
43. Feng, J., Meyer, C. A., Wang, Q., Liu, J. S., Shirley Liu, X., and Zhang, Y. (2012) GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics* **28,** 2782–2788
44. Boneca, I. G., de Reuse, H., Epinat, J. C., Pupin, M., Labigne, A., and Moszer, I. (2003) A revised annotation and comparative analysis of *Helicobacter pylori* genomes. *Nucleic Acids Res.* **31,** 1704–1714
45. Celesnik, H., Deana, A., and Belasco, J. G. (2007) Initiation of RNA decay in *Escherichia coli* by 5′ pyrophosphate removal. *Mol. Cell* **27,** 79–90
46. Celesnik, H., Deana, A., and Belasco, J. G. (2008) PABLO analysis of RNA: 5′-phosphorylation state and 5′-end mapping. *Methods Enzymol.* **447,** 83–98
47. Luciano, D. J., Hui, M. P., Deana, A., Foley, P. L., Belasco, K. J., and Belasco, J. G. (2012) Differential control of the rate of 5′-end-dependent mRNA degradation in *Escherichia coli*. *J. Bacteriol.* **194,** 6233–6239
48. Bonnin, R. A., and Bouloc, P. (2015) RNA Degradation in *Staphylococcus aureus*: diversity of ribonucleases and their impact. *Int. J. Genomics* **2015,** 395753
49. Mathy, N., Bénard, L., Pellegrini, O., Daou, R., Wen, T., and Condon, C. (2007) 5′-to-3′ exoribonuclease activity in bacteria: role of RNase J1 in rRNA maturation and 5′ stability of mRNA. *Cell* **129,** 681–692
50. Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., and Schwede, T. (2009) Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Protoc.* **4,** 1–13
51. The PyMOL Molecular Graphics System, Version 1.8, Schrödinger, LLC.
52. Boneca, I. G., Ecobichon, C., Chaput, C., Mathieu, A., Guadagnini, S., Prévost, M. C., Colland, F., Labigne, A., and de Reuse, H. (2008) Development of inducible systems to engineer conditional mutants of essential genes of *Helicobacter pylori*. *Appl. Environ. Microbiol.* **74,** 2095–2102
53. Berezikov, E., Thuemmler, F., van Laake, L. W., Kondova, I., Bontrop, R., Cuppen, E., and Plasterk, R. H. (2006) Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* **38,** 1375–1377
54. Förstner, K. U., Vogel, J., and Sharma, C. M. (2014) READemption: a tool for the computational analysis of deep-sequencing-based transcriptome data. *Bioinformatics* **30,** 3421–3423
55. Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F., and Hackermüller, J. (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.* **5,** e1000502

Functional Characterization of H. pylori RppH

56. Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., and Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16,** 944–945
57. Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., Le Fèvre, F., Longin, C., Mornico, D., Roche, D., Rouy, Z., Salvignol, G., Scarpelli, C., Thil Smith, A. A., Weiman, M., and Médigue, C. (2013) MicroScope: an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res.* **41,** D636–D647
58. Fulton, D. L., Li, Y. Y., Laird, M. R., Horsman, B. G., Roche, F. M., and Brinkman, F. S. (2006) Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics* **7,** 270

Identification of the RNA Pyrophosphohydrolase RppH of *Helicobacter pylori* and Global Analysis of Its RNA Targets

Thorsten Bischler, Ping-kun Hsieh, Marcus Resch, Quansheng Liu, Hock Siew Tan, Patricia L. Foley, Anika Hartleib, Cynthia M. Sharma and Joel G. Belasco

J. Biol. Chem. 2017, 292:1934-1950.
doi: 10.1074/jbc.M116.761171 originally published online December 14, 2016

3.5 THE CSRA-FLIW NETWORK CONTROLS POLAR LOCALIZATION OF THE DUAL-FUNCTION FLAGELLIN MRNA IN CAMPYLOBACTER JEJUNI

nature COMMUNICATIONS

ARTICLE

Received 12 Dec 2015 | Accepted 18 Apr 2016 | Published 27 May 2016

DOI: 10.1038/ncomms11667 OPEN

The CsrA-FliW network controls polar localization of the dual-function flagellin mRNA in *Campylobacter jejuni*

Gaurav Dugar[1], Sarah L. Svensson[1], Thorsten Bischler[1], Sina Wäldchen[2], Richard Reinhardt[3], Markus Sauer[2] & Cynthia M. Sharma[1]

The widespread CsrA/RsmA protein regulators repress translation by binding GGA motifs in bacterial mRNAs. CsrA activity is primarily controlled through sequestration by multiple small regulatory RNAs. Here we investigate CsrA activity control in the absence of antagonizing small RNAs by examining the CsrA regulon in the human pathogen *Campylobacter jejuni*. We use genome-wide co-immunoprecipitation combined with RNA sequencing to show that CsrA primarily binds flagellar mRNAs and identify the major flagellin mRNA (*flaA*) as the main CsrA target. The *flaA* mRNA is translationally repressed by CsrA, but it can also titrate CsrA activity. Together with the main *C. jejuni* CsrA antagonist, the FliW protein, *flaA* mRNA controls CsrA-mediated post-transcriptional regulation of other flagellar genes. RNA-FISH reveals that *flaA* mRNA is expressed and localized at the poles of elongating cells. Polar *flaA* mRNA localization is translation dependent and is post-transcriptionally regulated by the CsrA-FliW network. Overall, our results suggest a role for CsrA-FliW in spatiotemporal control of flagella assembly and localization of a dual-function mRNA.

[1] Research Centre for Infectious Diseases (ZINF), University of Würzburg, Josef-Schneider-Str. 2/D15, Würzburg D-97080, Germany. [2] Department of Biotechnology and Biophysics, University of Würzburg, Am Hubland, Würzburg D-97074, Germany. [3] Max Planck Genome Centre Cologne, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, Cologne D-50829, Germany. Correspondence and requests for materials should be addressed to C.M.S. (email: cynthia.sharma@uni-wuerzburg.de).

Post-transcriptional control involves a complex interplay between mRNAs, small regulatory RNAs (sRNAs) and protein regulators. Although regulatory functions have typically been attributed to proteins or sRNAs, mRNAs have canonically been considered as targets of this regulation. However, regulatory functions have recently also been described for mRNAs that either encode sRNAs in their untranslated regions (UTRs) or act as sponges that sequester other regulatory factors[1–4].

The widespread bacterial Csr/Rsm (Carbon storage regulator/Regulator of secondary metabolism) regulatory network[5] is an ideal model system to study the complex post-transcriptional cross-talk between mRNAs, sRNAs and protein regulators. About 75% of all sequenced bacterial genomes encode a homologue of the central RNA-binding protein (RBP) of this system, CsrA (RsmA/E). CsrA is a pleiotropic regulator of global physiological phenomena in Gammaproteobacteria[5] and considered the most conserved post-transcriptional virulence regulator[6]. CsrA mainly acts by repression of translation initiation via binding to 5′ regions of mRNAs[7]. The homodimeric CsrA binds GGA-rich motifs that are often located in hairpin loops and/or overlap the Shine-Dalgarno (SD) sequence[5]. In Gammaproteobacteria, CsrA activity is regulated through the CsrB/C and RsmX/Y/Z families of sRNAs[5,7]. These antagonizing sRNAs are often induced by environmental signals[6] and harbour multiple stem-loops with high-affinity GGA motifs that sequester CsrA/RsmA[8]. Despite the presence of CsrA, many bacteria lack homologues of these antagonizing sRNAs. Also, the global CsrA regulon and its general biological function outside the Gammaproteobacteria are unclear. In the Gram-positive *Bacillus subtilis*, the flagellar assembly protein FliW antagonizes CsrA via direct binding[9]. Although FliW homologues are relatively widespread[9], protein-mediated regulation of CsrA has not yet been shown outside *B. subtilis*. Whether FliW can cooperate with RNA-mediated regulation of CsrA is also unknown.

In the Gram-negative Epsilonproteobacterium *Campylobacter jejuni*, currently the leading cause of bacterial gastroenteritis in humans, CsrA affects motility, biofilm formation, oxidative stress response and infection[10]. Despite several phenotypic analyses of *csrA* deletion strains[10–12], direct CsrA targets in Epsilonproteobacteria are largely unknown. Global transcriptome studies indicated that both *C. jejuni* and the related pathogen *Helicobacter pylori*[13–16], which both carry potential FliW homologues, lack the CsrA-antagonizing sRNAs.

Here we use co-immunoprecipitation (coIP) combined with RNA sequencing[17,18] (RIP-seq) to globally determine the direct RNA-binding partners of *C. jejuni* CsrA and investigate whether RNA-based regulation of CsrA occurs in the absence of canonical antagonizing sRNAs. Our genome-wide approach reveals many mRNAs of flagellar genes as potential CsrA targets and we demonstrate that *flaA* mRNA, encoding the major flagellin, has dual (coding and regulatory) function. As the most abundantly co-purified transcript, *flaA* mRNA is the main target of CsrA translational repression. In addition, the *flaA* leader can act as an mRNA-derived RNA antagonist of CsrA. Together with the main CsrA antagonist, the FliW protein, *flaA* mRNA titrates CsrA to regulate expression of other flagellar genes.

In addition, using confocal and super-resolution microscopy imaging, we show that *flaA* mRNA is expressed in elongating cells and localizes to the cell poles of the amphitrichous *C. jejuni*. In contrast to eukaryotes[19], RNA localization is so far only poorly understood in prokaryotes. Bacterial mRNAs can remain localized close to their genomic site of transcription[20] or can migrate to places in the cell where their encoded products are required in a translation-independent manner involving *cis*-acting signals in the RNA itself[21]. Besides the mechanisms of bacterial RNA localization, even less is known about how this process may be regulated and which, if any, RBPs are involved. Here we show, based on a variety of *C. jejuni* mutants that disrupt or maintain *flaA* translation, that polar *flaA* mRNA localization requires its translation. Furthermore, we demonstrate that FliW facilitates polar flagellin mRNA localization by antagonizing CsrA-mediated translational repression of *flaA*. The unexpected role of the CsrA-FliW system in spatial control of flagellin mRNA expression provides new insight into the role of RBPs in bacterial mRNA localization, a process only recently described in prokaryotes.

Results

Global RIP-seq reveals direct CsrA targets in *C. jejuni*. To globally identify *C. jejuni* CsrA targets and any RNA regulators of CsrA activity, we applied a RIP-seq approach[17,18]. The *csrA* (Cj1103) gene was chromosomally 3xFLAG-tagged at its C-terminus in strains NCTC11168 and 81-176. CsrA-3xFLAG is constitutively expressed during growth in rich medium, and neither introduction of the FLAG-tag nor deletion of *csrA* affects *C. jejuni* growth under the examined conditions (Supplementary Fig. 1). We performed coIPs on mid-exponential-phase lysates of *csrA*-3xFLAG strains and, as control, their respective untagged wild-type (WT) strains (Fig. 1a and Supplementary Fig. 2a). After conversion of co-purified RNAs into cDNA and deep sequencing, 93.2–95.8% of the 4.6–6.2 million sequenced reads for the individual libraries were mapped to the respective genomes (Supplementary Table 1). Most of the NCTC11168 control-coIP library reads mapped to presumably non-specifically pulled-down abundant classes of RNA (rRNA, tRNA and housekeeping RNAs; Fig. 1b and Supplementary Table 2). In contrast, a ~36-fold and ~5-fold enrichment for reads mapped to 5′UTRs or open reading frames (ORFs) of mRNAs, respectively, was observed in the CsrA-3xFLAG coIP library (Fig. 1b). No specific sRNA enrichment was detected. As the coIP of strain 81-176 showed similar enrichment patterns (Supplementary Fig. 2b), we focused on strain NCTC11168.

***C. jejuni* CsrA primarily binds flagellar mRNAs.** Functional enrichment analysis of the 154 top CsrA targets with >5-fold enrichment in the CsrA-3xFLAG- versus control-coIP (Supplementary Data 1) revealed an overrepresentation of mRNAs from the class 'Surface Structures', including flagellar genes (Supplementary Fig. 3a,b). In fact, 90% of the reads mapping to the >5-fold-enriched CsrA targets belonged to flagella- or motility-related genes (Fig. 1c). The alternative sigma factors RpoN (σ^{54}) and FliA (σ^{28}) hierarchically control flagellar expression in *Campylobacter*[22]. Early genes are expressed from RpoD/σ^{70}-dependent promoters, whereas class 2 (middle) and class 3 (late) genes are RpoN- and FliA-dependent, respectively[22]. Most of the enriched transcripts belonged to either class 2 or class 3 (Table 1 and Supplementary Fig. 3c). The most abundantly co-purified transcript, with more than 300-fold enrichment, was *flaA* mRNA, encoding the major flagellin (Fig. 1c).

cDNA peaks reveal CsrA binds in diverse mRNA regions. Visual inspection of the cDNA read-patterns showed that numerous flagellar mRNAs, including *flaA*, *flaG* and *flgI* (encoding the major flagellin, a gene involved in flagellum formation, and a P-ring component, respectively) showed strong enrichment in their 5′UTRs (Fig. 1d and Supplementary Fig. 4a). CsrA binding was also observed between two genes in polycistronic mRNAs, such as the Cj0310c-Cj0309c and Cj0805-*dapA* operons. Analysis of the potential CsrA-binding sites in an *Escherichia coli* green fluorescent protein (GFP) reporter-system,

NATURE COMMUNICATIONS | DOI: 10.1038/ncomms11667 ARTICLE

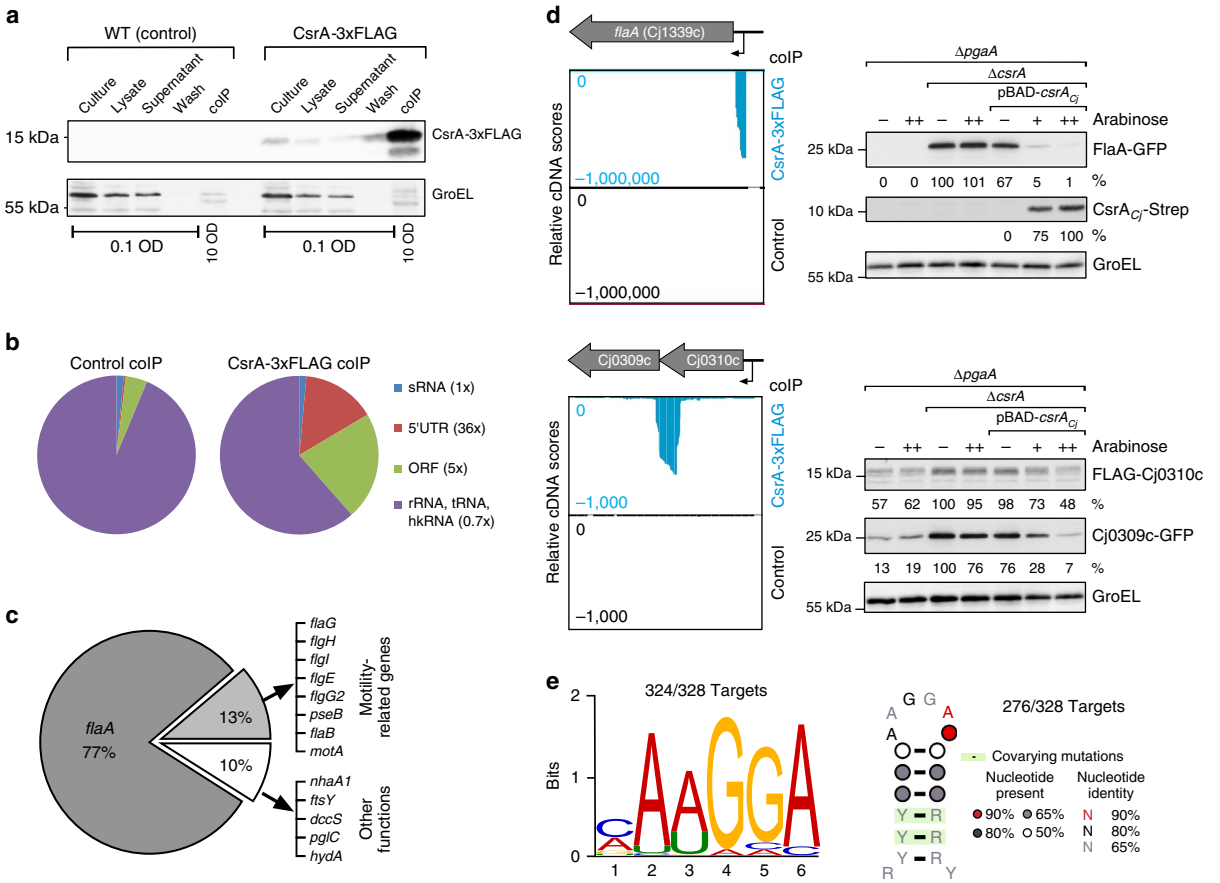


Figure 1 | RIP-seq analysis of *C. jejuni* CsrA. (**a**) Western blot analysis of coIP samples of *C. jejuni* NCTC11168 WT and *csrA*-3xFLAG strains using anti-FLAG antibody confirms a successful CsrA-3xFLAG pulldown in the tagged strain. The amount of samples loaded (OD_{600} of bacteria) is indicated. GroEL served as loading control. (**b**) Pie charts showing relative proportions of mapped cDNA reads of different RNA classes in the coIP libraries (hkRNA: housekeeping RNAs). Numbers in brackets indicate the relative enrichment of the respective RNA class in the CsrA-3xFLAG versus control coIPs. (**c**) Pie chart showing the percentages and enriched genes of mapped reads for all >5-fold enriched CsrA target genes. (**d**) (Left) Mapped RNA-seq reads for the control (black) and CsrA-3xFLAG coIP (blue) in strain NCTC11168. Grey arrows: ORFs; black arrows: transcriptional start sites (TSS). Examples of enrichment patterns in 5′ UTRs (*flaA*) and between genes in a polycistron (Cj0310c-Cj0309c operon; encoding two paralogous efflux proteins). (Right) Western blot analysis using anti-FLAG and anti-GFP antibodies of reporter fusions to potential *C. jejuni* CsrA target genes in *E. coli* Δ*pgaA*, Δ*pgaA*/Δ*csrA* and Δ*pgaA*/Δ*csrA* + pBAD-$csrA_{Cj}$ (complementation with *C. jejuni* CsrA-Strep under control of an arabinose-inducible pBAD promoter) strains. Putative CsrA targets from *C. jejuni* were fused in-frame (for example, 33 aa for *flaA*) to GFP or a FLAG-*lacZ*′ tag (Supplementary Fig. 4b). As deletion of *csrA* dramatically enhanced biofilm formation and led to poor growth in liquid culture in our *E. coli* strain, reporter experiments were performed in a Δ*pgaA* background. GroEL served as loading control. Protein samples corresponding to 0.1 OD_{600} were loaded. Quantifications of reporter expression are given below the blots. (**e**) (Left) CsrA-binding motif predicted by MEME[24] (*E*-value = 2.1E-11). (Right) Consensus secondary structure motif of *C. jejuni* CsrA-binding sites predicted by CMfinder[62].

originally developed to study sRNA-mediated regulation[23], revealed all of the tested 5′UTR targets (*flaA*, *flaG*, *flgI*, *flaB*, *pseB* and Cj1249) were highly upregulated (>10-fold) in the absence of *E. coli csrA* as measured by western blot and FACS analyses (Fig. 1d and Supplementary Figs 4 and 5). Reduced reporter fusion expression was restored by complementation of Δ*csrA* with *C. jejuni* CsrA. Using an operon reporter, where the C-terminal part of the upstream gene is fused to FLAG-*lacZ*′ and the N-terminal part of the downstream gene to GFP, we observed that both *E. coli* and *C. jejuni* CsrA can repress the downstream genes in polycistrons (Cj0310c-Cj0309c and Cj0805-*dapA*). Expression of the upstream genes was only slightly affected and they do not contain any strong internal transcriptional start sites that could lead to uncoupled transcription of the downstream genes[14]. As we observed that potential SD sequences right at the 3′ end of the upstream genes are covered by CsrA target sites, CsrA probably interferes with ribosome binding and translation of the downstream genes and thereby might mediate discoordinate operon regulation.

Automated peak-detection reveals a CsrA-binding motif. To automatically identify CsrA-binding regions and a binding motif from coIP cDNA enrichment patterns, we developed a peak-detection algorithm based on a sliding window approach (see the Methods for details). This approach predicted 328 potential CsrA-binding sites with >5-fold enrichment in the NCTC11168 coIP (Supplementary Data 2). As a control, peak

Table 1 | Enrichment of genes involved in flagellar biosynthesis in the CsrA coIP data.

	C. jejuni NCTC11168		*C. jejuni* 81-176	
Enrichment (reads)	**5′UTR**	**ORF**	**5′UTR**	**ORF**
Regulation of expression (class 1)				
***rpoN* (Cj0670)**	1.5 × (26)	**25 × (1,747)**	–	**8 × (634)**
fliA (Cj0061c)	–	1.1 × (121)	–	0.7 × (79)
flgS (Cj0793)	–	1.7 × (43)	1.3 × (1)	1.7 × (15)
flgR (Cj1024)	1.2 × (4)	1.1 × (137)	1.3 × (4)	1.2 × (101)
Flagellar protein secretion (class 1)				
***flgM* (Cj1464)**	–	**5 × (1429)**	–	**2.3 × (380)**
***fliF* (Cj0318)**	–	**6.1 × (1,005)**	–	**7.8 × (1,105)**
flhA (Cj0882c)	–	1.2 × (82)	–	0.9 × (53)
flhB (Cj0335)	0.6 × (3)	1.2 × (99)	0.7 × (1)	1.0 × (67)
fliO (Cj0352)	–	1.4 × (41)	–	0.7 × (1)
fliP (Cj0820c)	–	1.2 × (29)	–	0.5 × (15)
fliQ (Cj1675)	–	1.3 × (45)	–	0.9 × (29)
fliR (Cj1179c)	–	1.2 × (6)	–	0.5 × (5)
fliH (Cj0320)	–	4.8 × (202)	–	1.3 × (100)
fliI (Cj0195)	–	3.6 × (442)	–	0.6 × (86)
Basal body components (classes 1 and 2)				
fliE (Cj0526c)	–	2.4 × (268)	–	1.4 × (120)
flgC (Cj0527c)	–	1.2 × (397)	–	0.9 × (217)
flgB (Cj0528c)	1.6 × (7)	1.6 × (181)	0.7 × (2)	0.9 × (61)
***flgG2* (Cj0697)**	–	**43.9 × (8,133)**	–	**77.4 × (9,670)**
***flgG* (Cj0698)**	1.2 × (1)	1.4 × (253)	**5.3 × (4)**	1.3 × (165)
flgJ (Cj1463)	–	4.4 × (180)	–	1.0 × (26)
***flgI* (Cj1462)**	**170.5 × (5,750)**	**52.7 × (12,087)**	**157.1 × (1,666)**	**61.5 × (5,401)**
***flgA* (Cj0769c)**	0.8 × (2)	**15.8 × (410)**	0.9 × (4)	**3.7 × (104)**
***flgH* (Cj0687c)**	**200.9 × (1,911)**	**20.1 × (3,288)**	**110.1 × (917)**	**27.6 × (2,487)**
Flagellar hook components (class 2)				
***flgE* (Cj1729c)**	–	**68.2 × (104,324)**	–	**7.1 × (3,967)**
flgD (Cj0042)	–	3.6 × (1015)	–	2.0 × (284)
flgE2 (Cj0043)	–	2.5 × (1045)	–	1.1 × (250)
fliK (Cj0041)	–	4.1 × (613)	–	1.2 × (89)
Cj0040*	**356.2 × (3,389)**	**110.6 × (9,277)**	**38 × (230)**	**20.4 × (727)**
flgK (Cj1466)	–	0.7 × (4)	–	1.0 × (141)
flgL (Cj0887c)	–	2.0 × (484)	2.3 × (74)	0.9 × (193)
Flagellar filament components (classes 2 and 3)				
***flaA* (Cj1339c)**	**304.5 × (693,471)**	**111 × (473,588)**	**324.7 × (158,590)**	**45.3 × (138,159)**
***flaB* (Cj1338c)**	**58.8 × (915)**	**14.1 × (17,880)**	**59.4 × (1,170)**	**14.9 × (29,530)**
***fliD* (Cj0548)**	–	**6.8 × (4,348)**	–	**5.4 × (3,929)**
fliS (Cj0549)		1.6 × (149)		1.3 × (165)
flaC (Cj0720)	1.2x (344)	1.2 × (1,298)	1.3 × (239)	1.2 × (1,237)
Other enriched genes (>5x) involved in flagella formation				
***pseB* (Cj1293)**	**119.7 × (2,298)**	**9.5 × (2,280)**	**34.5 × (470)**	4.1 × (759)
***pseI* (Cj1317)**	1.7 × (22)	**7.2 × (864)**	2.3 × (14)	0.8 × (111)
***flaG* (Cj0547)**	**346.1 × (11,077)**	**72.4 × (18,150)**	**168.5 × (3,701)**	**84.2 × (16,012)**
***motA* (Cj0337c)**	**10.3 × (89)**	1.8 × (660)	1.4 × (16)	0.8 × (271)
Cj0951c	–	**15.2 × (79)**	2 × (3)	1.3 × (194)
Cj0248	**5.5 × (120)**	1.8 × (387)	1.1 × (38)	0.9 × (257)
***flhX* (Cj0848c)**	–	**7.5 × (13)**	–	1.5 × (7)

CoIP, co-immunoprecipitation; UTR, untranslated region; ORF, open reading frame.
Classification of flagellar genes is based on ref. 75. Transcripts with >5-fold enrichment in cDNA read counts in the CsrA-3xFLAG versus control coIP libraries are highlighted in bold. Numbers in brackets indicate the absolute cDNA read counts in the CsrA-3xFLAG coIP libraries.
*Cj0040 (unknown function) is the first gene of the hook gene operon.

detection was performed in reverse manner by scanning for enriched regions in the control- versus CsrA-3xFLAG-coIP. This analysis revealed only five peaks, without a common motif, indicating a high specificity of the peaks detected in the CsrA-3xFLAG-coIP. MEME[24] analysis of the 328 enriched sequences revealed a (C/A)A(A/T)GGA motif in 324/328 input sequences (Fig. 1e). Analysis of the 81-176 coIP led to a similar motif (Supplementary Fig. 2c). To check if a similar motif can be found in non-enriched regions, we conducted the peak-detection in reverse manner using a cutoff of only >1-fold enrichment in the control- versus CsrA-3xFLAG-coIP. This revealed 448 'enriched' sites in the control library. Subsequent motif prediction did not yield any significant motifs, further supporting high specificity of the coIP approach. Consensus-structure motif screening of the enriched CsrA-coIP sequences revealed an AAGGA motif in a hairpin-structure loop in 276/328 input sequences (Fig. 1e). These *C. jejuni* sequence/structural motifs agree with binding sites of other CsrA homologues[25].

NATURE COMMUNICATIONS | DOI: 10.1038/ncomms11667 ARTICLE

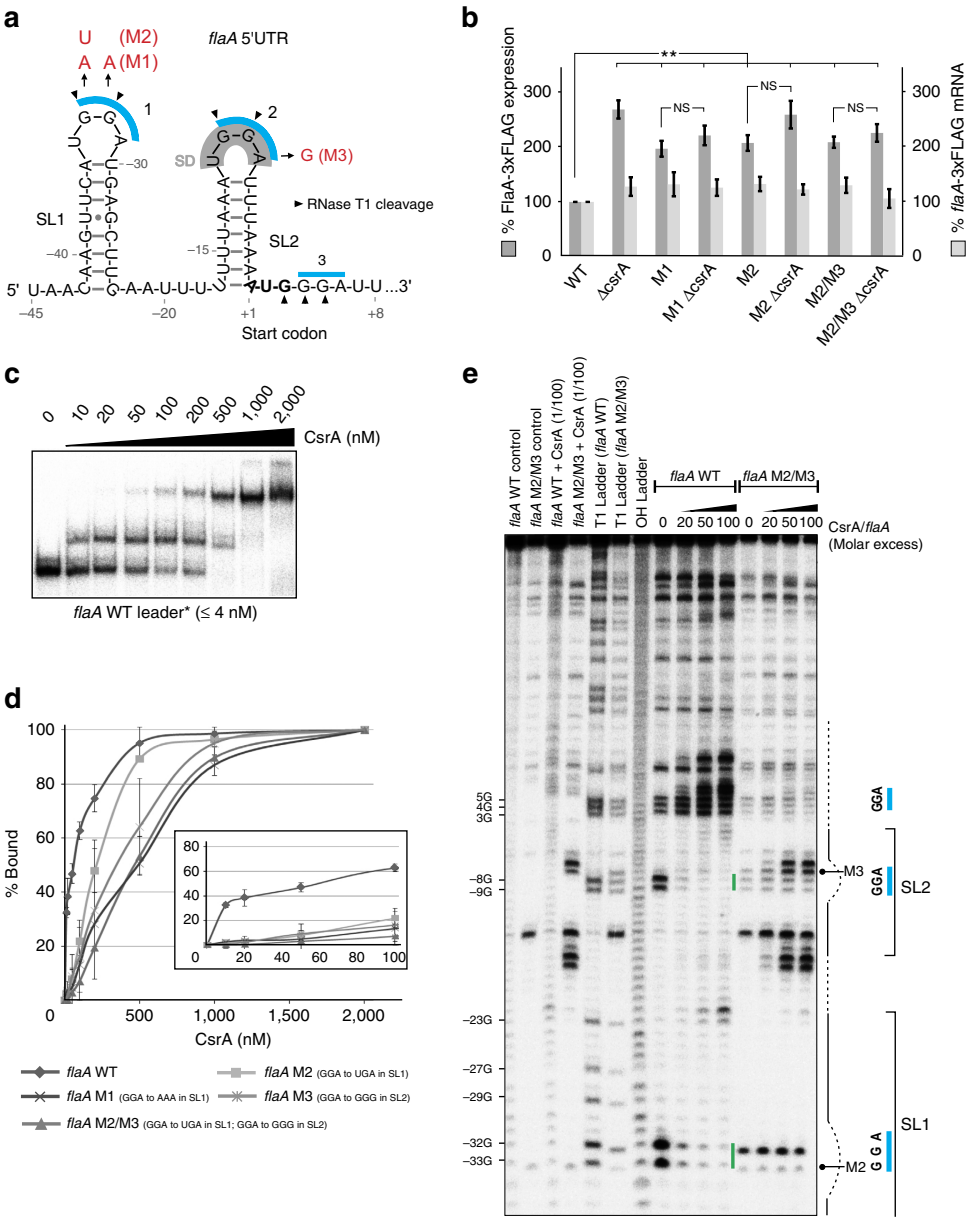


Figure 2 | CsrA represses *flaA* translation by binding to its 5′UTR. (**a**) Predicted secondary structure of the *flaA* leader using Mfold[74]. Blue bars indicate GGA motifs; grey: SD sequence. Black triangles indicate RNase T1 cleavages from the structure probing in **c**. (**b**) Western blot quantification ($n=5$ biological replicates) of FlaA with a C-terminal 3xFLAG epitope tag integrated at its native locus (FlaA-3xFLAG) and northern blot analysis of *flaA* mRNA ($n=3$ biological replicates) in Δ*csrA* and various *flaA* 5′UTR mutant strains. Shown is the mean ± s.e.m (**$P<0.01$ using Student's *t*-test, NS: not significant). Mutations are depicted in red in **a**. (**c**) Gel-shift assays using ~0.04 pmol *in vitro*-transcribed and 5′ end-labelled *flaA* leader (−45 to +99 relative to the start codon) with increasing concentrations of CsrA. (**d**) Affinity binding curves determined by gel-shift assays for ^{32}P-labelled *flaA* WT and mutant leaders (≤4 nM) based on three replicates. The inset represents an enlargement of the binding curves for low CsrA concentrations. Shown is the mean ± s.d. (**e**) Footprinting assays of ~0.2 pmol ^{32}P-labelled *flaA* WT and *flaA* M2/M3 mutant leaders in the absence or presence of increasing CsrA concentrations (molar excess of 0, 20, 50 and 100 CsrA) using RNase T1. Untreated *flaA* leader alone or incubated with 100-fold excess of CsrA served as controls and RNase T1- or alkali (OH)-digested *flaA* leader as ladders, respectively. Blue lines: GGA motifs; green lines: protection from RNA cleavage upon addition of CsrA. The secondary structure of the *flaA* leader according to **a** is depicted on the right.

***flaA* mRNA is translationally repressed by CsrA**. The flagellar filament, consisting mainly of the FlaA flagellin, is among the last components produced during flagellum assembly. In our coIP, 77% of the reads from >5-fold enriched genes mapped to *flaA*, indicating it as the main CsrA target (Fig. 1c). Secondary-structure predictions revealed that the 45-nt-long *flaA* 5′UTR can fold into two stem-loops (SL1 and SL2), both of which harbour an ANGGA motif in their loops (Fig. 2a). The second ANGGA motif

covers the ribosome-binding site and a third GGA is present as the second codon. The *flaA* 5′UTR secondary structure is conserved and supported by compensatory base-pair changes in other *Campylobacter* species (Supplementary Figs 6 and 7, and Supplementary Methods). A chromosomally 3xFLAG-tagged FlaA was ~3-fold upregulated in a Δ*csrA* strain compared with WT on western blots (Fig. 2b and Supplementary Fig. 8a, lanes 1 and 2). To show that CsrA affected translation by binding to the *flaA* leader, we introduced chromosomal point-mutations into the two putative GGA CsrA-binding motifs (M1: $SL1_{GGA\rightarrow AAA}$, M2: $SL1_{GGA\rightarrow UGA}$, and M3: $SL2_{GGA\rightarrow GGG}$; Fig. 2a,b and Supplementary Fig. 8a, lanes 3–8). Like deletion of *csrA*, mutation of the GGA motifs resulted in two- to threefold elevated FlaA-3xFLAG protein expression. FlaA-3xFLAG levels were not affected by deletion of *csrA* in the *flaA* leader mutants, indicating CsrA binding was abolished in these strains. Northern blot analysis showed *flaA*-3xFLAG mRNA levels are only mildly affected in the different mutant strains, further indicating post-transcriptional regulation of *flaA* by CsrA (Fig. 2b and Supplementary Fig. 8a).

In vitro gel-shift assays using recombinant *C. jejuni* CsrA-Strep and T7-transcribed, 5′-end radiolabelled *flaA* WT leader showed strong CsrA binding ($K_d = \sim 50$ nM) with two defined shifts, indicating at least two CsrA-binding sites (Fig. 2c). In contrast, *flaA* leaders with GGA point-mutations in either SL1 (M1 and M2), SL2 (M3) or both SL1 and SL2 (M2/M3) showed four- to tenfold higher K_d values (200–500 nM), confirming that the mutations reduced CsrA binding (Fig. 2d and Supplementary Fig. 9a). To map CsrA-binding sites on the *flaA* leader, we performed *in-vitro* footprinting assays with labelled *flaA* leader in the absence or presence of CsrA using enzymatic and chemical cleavage (RNase T1; single stranded G-residues and lead(II) acetate; single-stranded RNA). Cleavage patterns without CsrA confirmed the predicted *flaA* leader structure (Fig. 2e and Supplementary Fig. 8b). A clear protection was observed at the SL1 and SL2 GGA motifs of the WT leader upon addition of increasing CsrA amounts, but not for a *flaA* M2/M3 mutant with disrupted binding motifs. The third GGA downstream of the start codon was not protected. Overall, our data suggest *C. jejuni* CsrA represses *flaA* translation by high-affinity binding to the two GGA-containing stem-loops SL1 and SL2 in the *flaA* leader.

The flagellar assembly factor FliW binds CsrA in *C. jejuni*. The constitutive expression of CsrA during routine culture (Supplementary Fig. 1) suggested modulation of its activity rather than its expression. Because homologues of the CsrB/C sRNAs are absent in *C. jejuni*, we hypothesized that other RNAs, or even proteins, might control CsrA activity in *Campylobacter*. One candidate (Cj1075, 129 aa) is a potential homologue of the flagellar assembly factor, FliW, which has a role in motility[26,27] but is otherwise uncharacterized. In *B. subtilis*, FliW binds CsrA and antagonizes CsrA-mediated translational repression of *hag* mRNA, encoding the major flagellin[9]. FliW can also bind Hag, which accumulates in the cytoplasm before flagellar hook completion. Hag thus sequesters FliW from CsrA, allowing CsrA to repress Hag synthesis. Upon completion of the hook, Hag is secreted, FliW is released and CsrA repression of *flaA* translation is relieved. Thus, this Hag-FliW-CsrA partner-switch mechanism ensures appropriate temporal flagellin synthesis. In Epsilonproteobacteria, *fliW* homologues are present, but, unlike *Bacillus,* are not encoded adjacent to *csrA* (Fig. 3a). To investigate whether FliW can interact with CsrA and FlaA in *C. jejuni*, we performed protein–protein coIP experiments using chromosomal C-terminal 3xFLAG-tag fusions as bait. The anticipated interaction partners were tagged with mCherry at their C-terminus to allow detection by western blotting. In a FliW-3xFLAG-coIP, CsrA-mCherry was successfully co-purified, indicating the two proteins can interact (Supplementary Fig. 10). Similarly, FliW-mCherry was co-purified in a FlaA-3xFLAG-coIP, indicating conserved interactions between all three proteins. As control, none of the proteins was co-purified in coIPs with strains that carry the mCherry-fusion proteins but not the FLAG-tagged proteins.

FliW antagonizes CsrA-mediated translational repression. To determine whether the FliW–CsrA interaction could antagonize CsrA function in Epsilonproteobacteria, we used FlaA protein levels as a read-out for CsrA activity (Fig. 3b). Whereas FlaA-3xFLAG was ~3-fold upregulated in Δ*csrA*, deletion of *fliW* led to ~6-fold downregulation, consistent with further repression of *flaA* translation by additional CsrA released upon deletion of its protein antagonist (Fig. 3b). A Δ*csrA*/Δ*fliW* double deletion confirmed that the observed downregulation was indeed mediated through CsrA, as FlaA-3xFLAG levels increased back to those in the Δ*csrA* mutant. Despite strong reduction of FlaA-3xFLAG protein levels, a ~2-fold higher *flaA* mRNA level was observed upon deletion of *fliW*, indicating additional effects of FliW on *flaA* expression (Supplementary Fig. 11a). Thus, we constructed a transcriptional reporter composed of the unrelated Cj1321 5′UTR and its early coding region (Cj1321_mini) under the control of the *flaA* promoter. This reporter was, like the endogenous *flaA* mRNA, ~2-fold upregulated in the Δ*fliW* mutant (Supplementary Fig. 11b). As Cj1321 is independent of CsrA-mediated control, FliW seems to have a negative effect (direct or indirect) on *flaA* transcription.

To uncouple transcriptional control of *flaA* from its translational regulation, we replaced the σ^{28}-dependent *flaA* promoter in the FlaA-3xFLAG strain with a constitutive σ^{70}-dependent *metK* promoter. Upon deletion of *csrA* in this strain, a ~3-fold increase in FlaA-3xFLAG level was observed, further confirming post-transcriptional regulation of FlaA-3xFLAG protein expression by CsrA (Fig. 3b). Like for the strain expressing FlaA-3xFLAG from its native promoter, FlaA-3xFLAG expressed from the *metK* promoter was strongly downregulated upon deletion of *fliW* and was restored to Δ*csrA* levels in the Δ*csrA*/Δ*fliW* double mutant. This further indicates FliW antagonizes CsrA-mediated translational repression of *flaA* in a promoter-independent manner. In addition, decreased *flaA* mRNA stability was observed upon *fliW* deletion in rifampicin stability assays. This is consistent with increased translational repression of *flaA* in the absence of *fliW*, despite overall higher steady-state *flaA* mRNA levels because of FliW-dependent increased transcription (Supplementary Fig. 11c).

In line with strong downregulation of the FlaA protein upon *fliW* deletion, transmission electron microscopy revealed shorter flagella on Δ*fliW* bacteria compared with those of the WT strain (Fig. 3c,d). In fact, the flagella of Δ*fliW* appeared similar to those of a Δ*flaA* mutant strain and of bacteria lacking σ^{28} (Δ*fliA*), required for *flaA* transcription. In contrast, the Δ*csrA* and Δ*csrA*/Δ*fliW* strains expressed normal flagellar filaments. The short flagella of the Δ*fliW* strain are probably composed mainly of the minor flagellin FlaB, which is transcribed from an RpoN (σ^{54})-dependent promoter. Upon deletion of both flagellin genes (Δ*flaA*/Δ*flaB*), the bacteria no longer had filaments but the hook structure was visible at the poles (black arrowheads, Fig. 3c). Furthermore, a Δ*rpoN* mutant strain had neither flagella nor hooks. Motility assays revealed that the Δ*csrA* or Δ*fliW* strains showed a halo-radius reduction to 78% and 72% of WT, respectively (Fig. 3d). Likely due to its shorter flagella, Δ*fliW* also showed slower autoagglutination than WT, but greater than

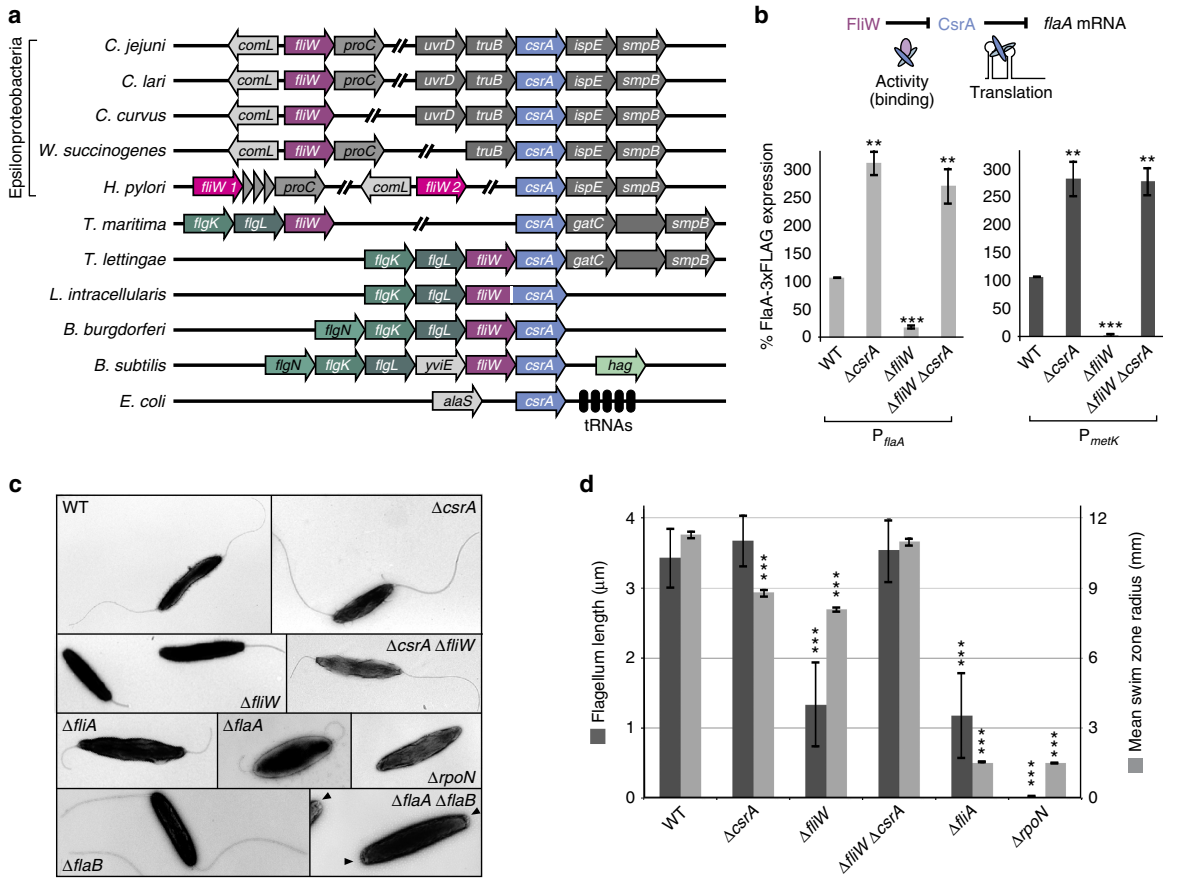


Figure 3 | The flagellar assembly factor FliW binds and antagonizes CsrA. (**a**) Genomic context of *csrA* and *fliW* homologues in diverse bacterial species (*Campylobacter* spp: *C. jejuni, C. lari, C. curvus*; *Wolinella succinogenes*; *Helicobacter pylori*; Thermotogales: *T. maritima, T. lettingae*; *Lawsonia intracellularis*; *Borrellia burgdorferi*; *Bacillus subtilis*; *Escherichia coli*). Blue: *csrA* homologs; dark or light red: *fliW* homologues; shades of green: flagellar genes. (**b**) (Top) Scheme of the antagonizing effect of FliW on CsrA-mediated translational repression of *flaA* mRNA by direct binding of FliW to CsrA. (Bottom, left) Quantification of FlaA-3xFLAG using western blot in *C. jejuni* WT, Δ*csrA*, Δ*fliW* and Δ*csrA*/Δ*fliW* strains in mid-log phase ($n=3$ biological replicates). Plotted is the mean ± s.e.m (**$P<0.01$, ***$P<0.001$ using Student's *t*-test). (Bottom, right) Quantification of FlaA-3xFLAG using western blot in WT, Δ*csrA*, Δ*fliW* and Δ*csrA*/Δ*fliW* strain backgrounds where the *flaA* promoter has been exchanged with the constitutive *metK* promoter. Please note that FlaA-3xFLAG levels expressed from the P_{metK} *promoter* represent ~70% compared with the expression from its native P_{flaA} promoter. (**c**) Transmission electron micrographs of indicated strains harvested from MH agar. Black triangles indicate hook structures. (**d**) Average flagella length (dark grey bars) of indicated strains from transmission electron micrographs using ImageJ ($n>25$ measurements). Plotted is the mean ± s.d. (***$P<0.001$ versus WT using Student's *t*-test). Motility was measured as average swimming distance (light grey bars) in soft agar. Bars show the mean ± s.e.m (***$P<0.001$ versus WT using Student's *t*-test).

the non-motile Δ*fliA* and Δ*rpoN* mutants (Supplementary Fig. 12). Overall, these data suggest that, besides a mild effect on *flaA* transcription, FliW affects post-transcriptional control of FlaA, and therefore filament assembly and motility, in a CsrA-dependent manner.

Expression of flagellar mRNAs is not affected in Δ*csrA*. Besides *flaA* mRNA, many other flagellar targets, such as the 5′UTRs of *flaG*, *flaB* and *flgI*, were strongly enriched in the CsrA-3xFLAG-coIP (>346-, >58- and >170-fold, respectively; Table 1). The *flaG*, *flaB* and *flgI* leaders also have one or more GGA-containing motifs near their SD (Fig. 4a). *In vitro* gel-shift assays of *in vitro* transcribed *flaG*, *flaB* and *flgI* leaders, and several other co-purified flagellar mRNAs (Cj0040, *flgA* and *flgM*), confirmed CsrA binding (Fig. 4b and Supplementary Fig. 9b). The non-enriched Cj1324 mRNA, encoding a gene involved in flagellin modification, or an unrelated mRNA fragment from *H. pylori* did not shift with CsrA, confirming specific binding of CsrA to coIP-enriched transcripts (Supplementary Fig. 9c). However, CsrA affinity for *flaG*, *flaB* and *flgI* leaders was lower (K_d = >350 nM) than for the *flaA* WT leader (K_d = ~50 nM, Fig. 4b). Although FlaA-3xFLAG was upregulated upon *csrA* deletion (Fig. 2b), chromosomally tagged FlaG-3xFLAG, FlaB-3xFLAG and FlgI-3xFLAG levels did not change substantially (Fig. 4a).

FliW and *flaA* mRNA titrate CsrA-mediated repression. The observed strong CsrA-mediated regulation of *flaG*, *flaB* and *flgI* in the *E. coli* reporter system (Supplementary Figs 4 and 5) indicates that CsrA can, in principle, regulate these targets. Thus, we hypothesized that FliW, or even abundant mRNAs, might sequester CsrA under the examined routine growth conditions, obscuring any regulatory effect on these low-affinity targets. Because *flaA* mRNA is highly abundant[14] and expressed at the end of the flagellar cascade, we reasoned *flaA* mRNA might itself

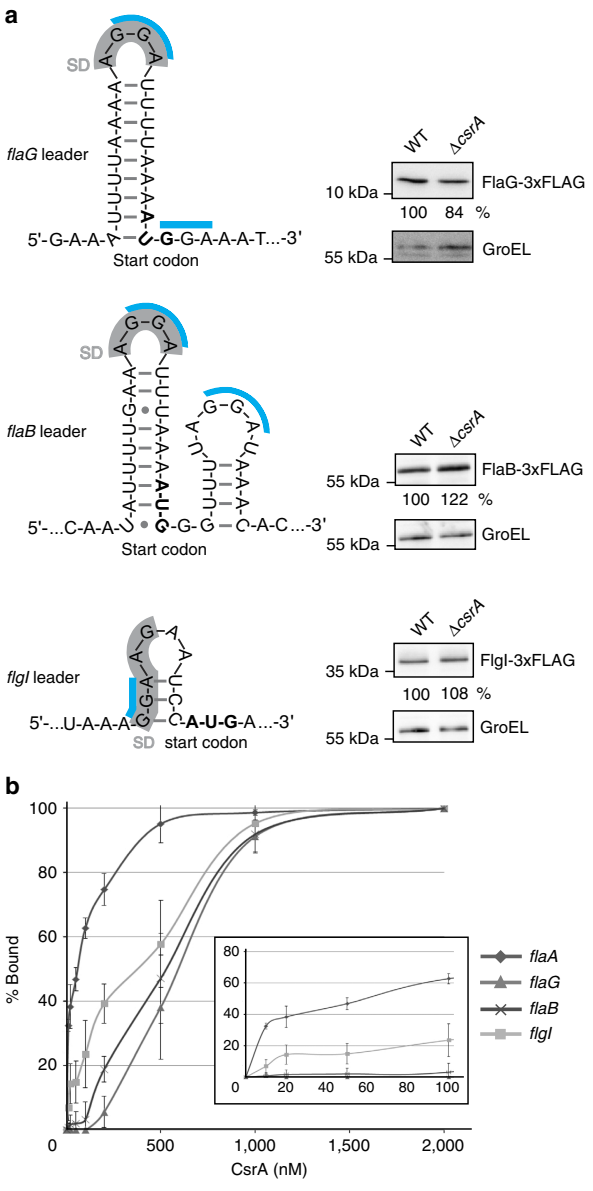


Figure 4 | CsrA binds to other flagellar target mRNAs but *csrA* deletion does not affect their translation. (**a**) (Left) Predicted secondary structures of *flaG*, *flaB* and *flgI* leaders using Mfold[74] with putative GGA binding-sites of CsrA (blue) and SD sequences (grey). (Right) Western blot analyses of FlaG-3xFLAG, FlaB-3xFLAG and FlgI-3xFLAG in *C. jejuni* WT or Δ*csrA* strains. (**b**) CsrA-binding affinities of flagella mRNA leaders (≤4 nM) determined by *in vitro* gel-shift assays. The inset represents an enlargement of the binding curves for low CsrA concentrations. Shown is the mean ± s.d.

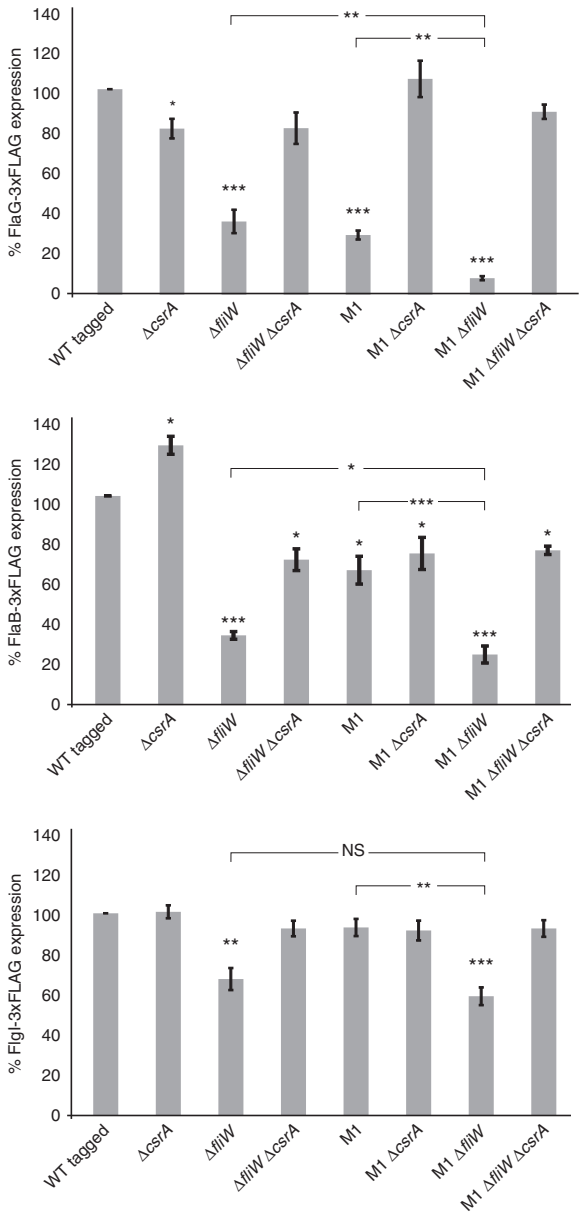


Figure 5 | The *flaA* 5′UTR and FliW inhibit CsrA-mediated regulation of flagella genes. Quantification of FlaG-3xFLAG, FlaB-3xFLAG and FlgI-3xFLAG levels using western blot of the indicated *C. jejuni* NCTC11168 strains grown to mid-log phase (M1: GGA→AAA in SL1 of *flaA* 5′UTR). Values were calculated based on at least three biological replicates. Shown is the mean ± s.e.m (*$P<0.05$, **$P<0.01$, ***$P<0.001$, using Student's *t*-test). NS, not significant.

titrate CsrA activity. To investigate the role of FliW and the *flaA* mRNA as CsrA antagonists, we analysed FlaG-3xFLAG, FlaB-3xFLAG and FlgI-3xFLAG protein expression in loss-of-function strains of both antagonists. In line with FliW acting as a general CsrA antagonist that limits CsrA activity, deletion of *fliW* led to a ~3-fold decrease in FlaG-3xFLAG level, which was restored to WT level in a Δ*csrA*/Δ*fliW* double mutant (Fig. 5 and Supplementary Fig. 13a).

Because *flaG* and *flaA* are primarily transcribed from σ^{28}-dependent promoters[28] and are thus expressed at the same time, monitoring FlaG-3xFLAG might reveal the potential role of *flaA* 5′UTR as a CsrA antagonist. The chromosomal M1 *flaA* leader mutation (GGA → AAA in SL1, Fig. 2a), which leaves the coding region intact but abolishes CsrA binding (Fig. 2d), decreased FlaG-3xFLAG levels ~3-fold (Fig. 5 and Supplementary Fig. 13a). Upon introduction of Δ*csrA*, FlaG-

3xFLAG expression was restored to WT levels, indicating decreased FlaG expression in the *flaA*-M1 mutant is dependent on CsrA, and suggesting that the *flaA* leader can also titrate CsrA. Combining both Δ*fliW* and *flaA*-M1 led to a tenfold reduction in FlaG-3xFLAG levels, showing their cumulative effect in antagonizing CsrA. In line with this, the M1/Δ*fliW*/Δ*csrA* triple mutant restored FlaG-3xFLAG levels back to WT levels (Fig. 5 and Supplementary Fig. 13a). Growth curves showed that there was no major impact on growth of the individual mutations under the examined conditions (Supplementary Fig. 13b). Although the Δ*fliW* and M1/Δ*fliW* mutants showed a slightly increased growth rate compared with WT, this increase was less than a non-motile Δ*fliA* strain.

To further confirm the role of the *flaA* 5′UTR as a CsrA antagonist, a ~250-nt long *flaA_mini* transcript comprising the *flaA* leader and first 17 codons followed by a stable ribosomal *rrnB* terminator was ectopically expressed from the native *flaA* promoter (Supplementary Fig. 14a). Expression of the *flaA_mini* transcript in a Δ*fliW* mutant, which has strong CsrA-mediated *flaA* translational repression, increased FlaA-3xFLAG levels around 2.6-fold (Supplementary Fig. 14b). This indicates *flaA_mini* can bind and antagonize CsrA and partially relieve CsrA-mediated repression of *flaA* translation. A smaller, yet significant, complementation of the effect of a *fliW* deletion was also observed for FlaG-3xFLAG levels.

Next, the effect of the two antagonists on CsrA-mediated regulation of the RpoN-dependent genes *flaB* and *flgI* was evaluated. A similar, yet less pronounced effect compared with FlaG-3xFLAG, was observed for FlaB-3xFLAG upon single or double mutations of *fliW* and M1. In contrast, FlgI-3xFLAG levels were only significantly reduced upon *fliW* deletion (Fig. 5 and Supplementary Fig. 13a). Overall, this reveals FliW as the major CsrA antagonist under the examined growth conditions that titrates, along with the *flaA* mRNA antagonist, CsrA from lower affinity flagellar targets such as *flaG*.

***flaA* mRNA localizes to the poles of elongating cells**. As *flaA* mRNA can titrate CsrA activity, we wondered when *flaA* mRNA levels change to modulate CsrA activity. Expression of *flaA* mRNA appeared constitutive during growth (Supplementary Fig. 13c). However, in the amphitrichously flagellated *C. jejuni*, after every cell division, a new flagellum has to be synthesized at the new pole of each daughter cell. As bacteria in batch culture are not synchronized in cell cycle, differences in *flaA* mRNA expression might be obscured because of the population-based northern analysis. To monitor *flaA* mRNA expression in single bacteria, we performed RNA-FISH (fluorescence *in situ* hybridization) in fixed *C. jejuni* cells from exponential phase. Although the control RNA, 16S rRNA (Fig. 6a, green), was visible in all

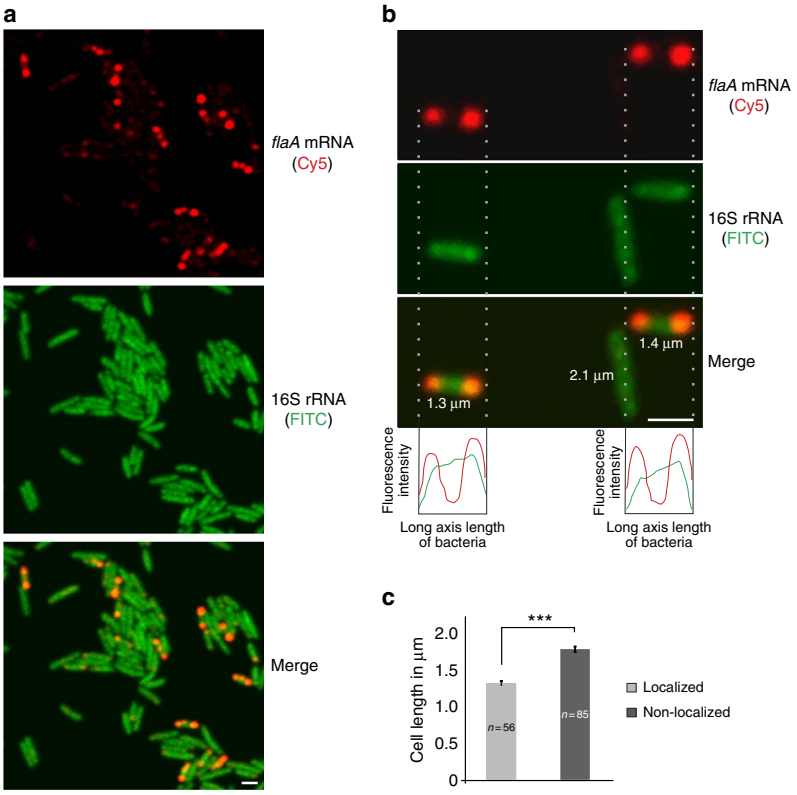


Figure 6 | *flaA* mRNA localizes to the poles of shorter cells. (**a**) RNA-FISH analysis of 16S rRNA (FITC-labelled DNA oligonucleotide probe, green) and *flaA* mRNA (14 Cy5-labelled single-stranded DNA oligonucleotide probes, red) in *C. jejuni* WT cells in mid-log phase using confocal microscopy (scale bar, 1 µm). (**b**) A magnified RNA-FISH image showing the distribution of fluorescence signals. *flaA* mRNA (Cy5) and 16S rRNA (FITC) signals were quantified along the long axis length of bacteria using ImageJ software and were subsequently merged as shown at the bottom of the panel (scale bar, 1 µm). The length of individual cells was also quantified using ImageJ. Statistical analysis for average *flaA* mRNA and 16S rRNA signals over the cell length is provided in Supplementary Fig. 15. (**c**) Average *C. jejuni* WT cell lengths in bacteria where *flaA* mRNA is localized (56 cells) or non-localized (85 cells), $***P<10^{-15}$ using Student's *t*-test.

cells, *flaA* mRNA (Fig. 6a, red) was detected in only some of the cells. As a negative control, we also performed *flaA* mRNA FISH on a Δ*fliA* mutant strain (Fig. 7a), which showed no expression of *flaA* (Supplementary Fig. 11a). Whereas 16S rRNA was equally distributed throughout the cell, *flaA* mRNA was specifically detected at the cell poles in ~20% of WT cells (Fig. 6a,b). Quantification of cell length across the population showed that cells with localized *flaA* mRNA were significantly shorter than cells without *flaA* expression (Fig. 6b,c). Live-cell imaging of a non-motile *C. jejuni* strain (Δ*fliA*) over two or three division cycles showed regular patterns of an increase in cell length until cells divide at mid-cell, resulting in short daughter cells (Supplementary Fig. 16). This indicates shorter cells likely correspond to cells that have divided and are elongating. Together, these data suggest differential expression of *flaA* mRNA during the cell cycle and accumulation in elongating cells at the required site of its encoded protein.

FliW impacts *flaA* mRNA localization via CsrA. To investigate whether CsrA-FliW impacts *flaA* mRNA localization, we next performed RNA-FISH in Δ*fliW*, Δ*csrA* and Δ*fliW*/Δ*csrA* mutant strains. Although *csrA* deletion had no effect on *flaA* localization, it was completely abolished in a Δ*fliW* mutant (Fig. 7a). Instead of a polar localization, *flaA* mRNA was now dispersed throughout the cell. The loss of *flaA* mRNA localization upon *fliW* deletion was not due to lower transcript abundance as its mRNA level is increased despite strong repression at the protein level (Supplementary Fig. 11a). Strikingly, *flaA* mRNA localization was restored to the cell poles in the Δ*fliW*/Δ*csrA* double mutant, showing CsrA affects localization of *flaA* mRNA. As a further confirmation of *flaA* mRNA localization, we performed super-resolution imaging of *flaA* mRNA FISH in WT and mutant strains using *direct* stochastic optical reconstruction microscopy (*d*STORM)[29], which has only recently been applied for bacterial RNA localization[30]. *d*STORM analysis fully supported and complemented the observations from confocal microscopy analysis (Fig. 7b and Supplementary Fig. 17). Overall, this suggests a model where *flaA* translation is required for polar localization: upon deletion of *fliW*, CsrA is released and in turn strongly represses *flaA* mRNA translation to impede its localization to the poles.

Polar *flaA* mRNA localization requires its translation. To support the translation-dependent model of *flaA* localization, we constructed several point mutants in the native *flaA* gene that either maintain or disrupt *flaA* translation (Fig. 8a). Mutation of the start codon of *flaA* (AUG → AAG (X1) or AUU (X2)) to abolish translation initiation resulted in dispersed *flaA* mRNA (Fig. 8b). In contrast, when the start codon was changed to an alternative start codon (AUG → GUG (X3)), *flaA* mRNA still localized to the cell poles, indicating translation of *flaA* mRNA is indeed required for polar localization. Mutation of the third *flaA* codon to a stop codon (UUU → UAG (X4)) also resulted in a completely dispersed *flaA* mRNA signal (Fig. 8b and Supplementary Fig. 17). In contrast, *flaA* mRNA with a synonymous silent mutation (UUU → UUC (X5); both encoding Phe)

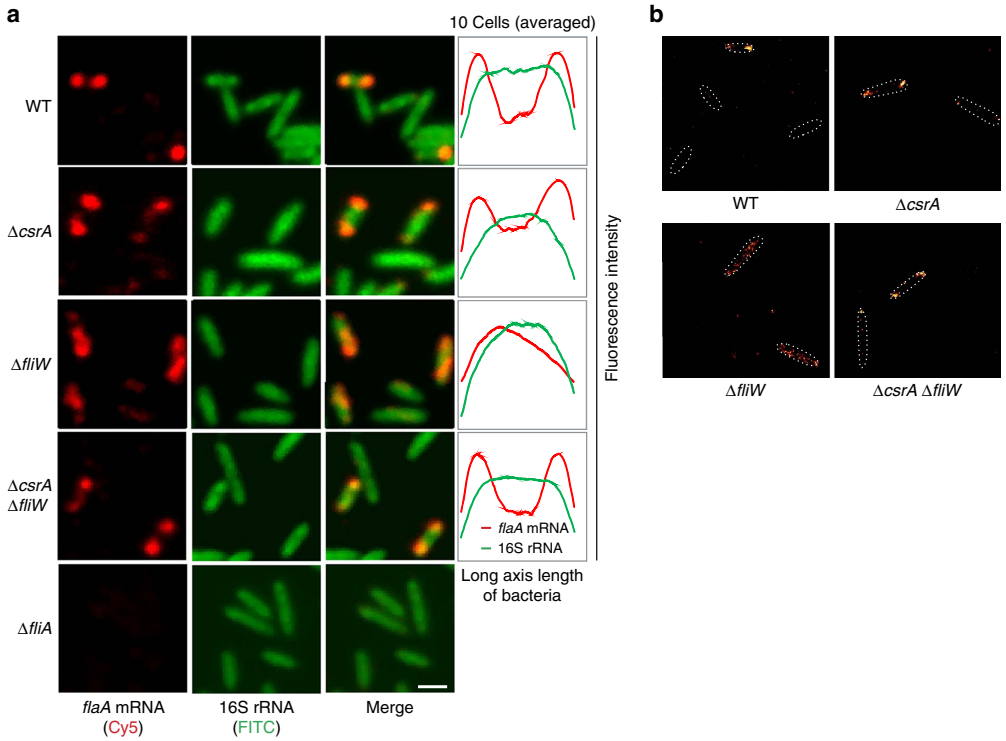


Figure 7 | CsrA and FliW influence *flaA* mRNA localization to the poles. (**a**) RNA-FISH analysis (Left: confocal microscopy images; Right: averaged fluorescence intensity along the long axis based on 10 cells) of 16S rRNA (green) and *flaA* mRNA (red) in *C. jejuni* NCTC11168 WT, Δ*csrA*, Δ*fliW*, Δ*csrA*/Δ*fliW* and Δ*fliA* strains in mid-log phase. FITC and Cy5 channels were merged in the microcopy images in the third lanes (scale bar, 1 μm). (**b**) Super-resolution microscopy imaging of *flaA* mRNA RNA-FISH (14 Cy5-labelled oligos) in the indicated *C. jejuni* strains using *d*STORM imaging. Cell boundaries from bright-field images are depicted by white dotted lines.

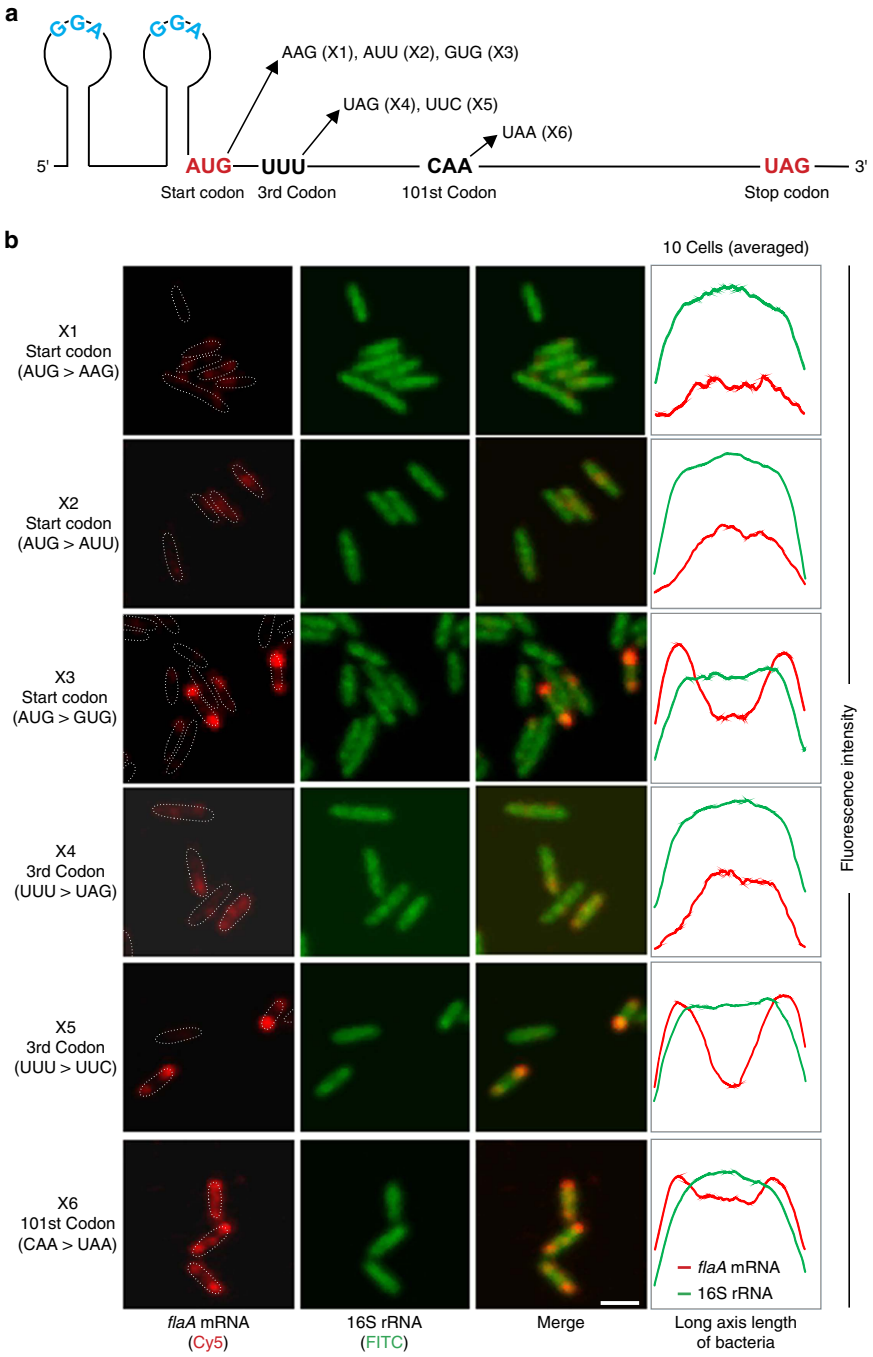


Figure 8 | Translation is required for *flaA* mRNA localization to the cell poles. (**a**) Point mutations in *flaA* mRNA that were introduced at the native *flaA* locus. Mutations X1, X2, X4 and X6 abolish or prematurely stop *flaA* translation, whereas X3 and X5 represent silent mutations. (**b**) RNA-FISH analysis (Left: confocal microscopy images; Right: averaged fluorescence intensity along the long axis based on 10 cells) of *C. jejuni* point mutant strains depicted in **a**. FITC and Cy5 channels were merged in the third rows of the microscopy images (scale bar, 1 μm).

at the third codon localized similarly to the WT mRNA. Some of the mutations that abolish translation (X1, X2) lead to reduced (50–80% of WT) *flaA* mRNA levels (Supplementary Fig. 18). Nonetheless, as the strain expressing the *flaA* mRNA with a stop mutation at the third codon (X4), which also showed abolished polar mRNA localization, had even higher (~170%) *flaA* expression levels than WT, it is unlikely that reduced (or increased) *flaA* mRNA levels lead to loss of localization. To determine the effect of terminating translation at a downstream position, we introduced a stop codon at the 101st codon of *flaA* (CAA → UAA (X6)). This mutant showed partial polar *flaA* mRNA localization, suggesting the N-terminal peptide might be

required for recruiting *flaA* mRNA to the cell poles. Overall, these data support a role of the FliW/CsrA post-transcriptional network in controlling translation-dependent polar *flaA* mRNA localization in *C. jejuni*.

Discussion

Using genome-wide RIP-seq, we have identified direct RNA targets of the translational regulator CsrA in a bacterium that lacks the canonical antagonizing sRNAs. Our study revealed the major flagellin mRNA is both the main CsrA target and a dual-function mRNA, which can titrate CsrA activity together with the FliW protein, the main CsrA antagonist (Fig. 9). Compared with microarray-based transcriptome analyses of *csrA* loss-of-function strains[31,32], which might reveal indirect effects or miss targets because of a lack of changes in target mRNA levels despite translational repression, a coIP approach facilitates the identification of direct targets and binding sites. Sanger sequencing of cDNAs from an RsmA-coIP identified six target mRNAs in *P. aeruginosa*[32]. RNA-seq of a CsrA-coIP in *E. coli* revealed 721 co-purified transcripts[33], and *in vivo* ultraviolet crosslinking combined with RNA-seq (CLIP-seq) revealed 467 potential CsrA-binding sites in *Salmonella typhimurium*, including binding sites in many virulence mRNAs[34]. In our RIP-seq approach, we used untagged WT strains as a negative control to allow for elimination of non-specifically bound transcripts. Our peak-detection tool confirmed the high specificity of this approach, as it detected an 'ANGGA' sequence in 324/328 targets, which resembles the CsrA consensus-motif determined by *in-vitro* selection[25]. Besides canonical binding to 5′UTRs or early codons[5,35], our coIP also revealed CsrA binding within coding regions or between genes in polycistrons to mediate discoordinate operon regulation.

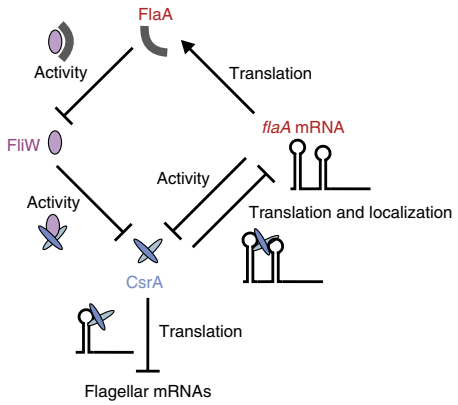


Figure 9 | Model depicting the *C. jejuni* CsrA-FliW regulatory network. Schematic representation of the regulatory circuit and the putative roles of CsrA, FliW and FlaA proteins along with *flaA* mRNA in the CsrA-FliW regulon of *C. jejuni*. The post-transcriptional regulatory protein CsrA represses translation of multiple flagellar mRNAs including *flaA* mRNA, encoding the major flagellin, by direct binding to the mRNAs. The FliW protein can directly bind and titrate CsrA activity and in-turn affects CsrA-mediated post-transcriptional regulation of flagellar genes. FliW can also bind to the FlaA protein, which releases FliW-mediated sequestration of CsrA. The abundant *flaA* mRNA is the main target of CsrA translational repression but can also act as a regulatory sponge and titrate CsrA activity together with the main CsrA antagonist FliW. Furthermore, *flaA* mRNA localizes to the cell poles of elongating cells. Polar localization of *flaA* mRNA itself is dependent on its translation, which is controlled by the CsrA-FliW regulatory network.

Our coIP approach revealed many mRNAs of flagellar genes as direct CsrA targets. The motility defect of Δ*csrA* suggests that tight regulation of flagellar genes by CsrA, and especially of the major flagellin FlaA, is required for proper motility. Balancing CsrA activity through the antagonizing protein FliW also appears crucial for flagellar assembly, as we observed that a *C. jejuni* NCTC11168 Δ*fliW* mutant expresses short flagella, as also reported in other strains[27,36], and is defective for autoagglutination and motility in both *B. subtilis* and *C. jejuni*[9,26]. Although CsrA impacts motility by directly controlling flagellin expression in *C. jejuni*, *B. subtilis* and *Borrelia*, the strong motility defect of an *E. coli csrA* mutant[37] is due to a requirement of CsrA for stabilization of the mRNA encoding the master regulator FlhDC[38]. The flagellum also plays an essential, multi-factorial role in *C. jejuni* colonization and pathogenesis, including secretion of Cia/Fed effectors[28,39], and is required for proper cell division[40]. Future studies might reveal CsrA-affected phenotypes beyond motility.

Instead of CsrA-activity control by antagonizing sRNAs[5], we demonstrated that the *flaA* mRNA itself can titrate CsrA. This represents a new mode of CsrA activity control by a target mRNA-derived antagonist. The *flaA* leader has higher affinity for CsrA compared to other flagellar targets. It has two GGA motifs in adjacent hexaloops, resembling high-affinity CANGGANG-containing apical hexaloop structures targeted by CsrA/RsmE[25,41]. The 21-nt spacing between the *flaA* GGA motifs is close to the 18-nt optimal intersite distance for binding of a CsrA dimer[42]. Whereas *flaA* mRNA probably only binds one CsrA dimer, multiple RsmE dimers are cooperatively assembled on RsmZ sRNA[8,41]. CsrA titration by a 5′UTR has recently been shown to mediate hierarchical control of fimbriae expression in *Salmonella typhimurium*[43]. The *fimAICDHF* mRNA leader, which in contrast to *flaA* mRNA is not itself a CsrA target, cooperates with the CsrB/C sRNAs to antagonize CsrA-mediated activation of plasmid-encoded fimbriae. Small RNAs other than CsrB/C can also sequester CsrA in addition to functioning as antisense RNAs[44]. Global approaches such as RIP-seq are ideally suited to identify additional antagonizing sRNAs or members of the emerging class of dual-function, cross-regulating mRNAs[2,3].

Analysis of *flaA* mRNA expression in single bacteria using RNA-FISH showed that this transcript localizes to the poles of shorter, and presumably elongating, cells. As a new flagellum is synthesized after each cell division at the new pole of the amphitrichous *C. jejuni*, polar *flaA* mRNA localization might facilitate this process. This temporal and spatial modulation of *flaA* mRNA expression might also affect CsrA-mediated regulation of other flagellar genes through mediating varying levels of this CsrA RNA antagonist. Mutations that either abolish or maintain translation showed *flaA* translation is required for its polar localization. Bacterial mRNA localization has only recently been described and unlike eukaryotes the underlying mechanisms and regulation of this process are poorly understood[45,46]. Besides co-translational targeting of mRNAs to the required sites of their encoded products, translation-independent mechanisms of RNA localization have also been described[20,21], including spatial expression according to chromosome organization. We observed that a *flaA* mRNA variant with a premature stop-codon mutation at the 101st codon partially localizes, suggesting a role of the N-terminus in directing the nascent peptide along with the mRNA to the secretion apparatus. Little is known how flagellar substrates are selected for secretion, as they do not share a secretion-signal sequence or cleavable signal peptide. N-terminal domains are required for secretion of flagellar proteins in diverse bacteria, including *C. jejuni*[36], and both 5′UTR and N-terminal peptide secretion signals have been shown to contribute to secretion efficiency[47]. In addition, flagellar

chaperones play a role in regulating the coupling of translation to secretion of flagellar substrates[48]. In *Yersinia*, *cis*-encoded RNA-localization elements in the early coding region are required for secretion of effector proteins by type III secretion systems[49]. Future studies will identify and clarify the role of elements, either in the protein N-terminus or the mRNA 5′UTR, as well as potential interaction partners that are crucial for directing the peptide and/or mRNA to the cell poles and secretion apparatus. Besides the requirement of *flaA* translation for localization, other factors such as the *flaA* genomic location or the transcriptional complex might also contribute to polar *flaA* mRNA localization.

Our study revealed an unexpected function for the CsrA-FliW network in spatial and temporal gene-expression control, and specifically FliW affects translation-dependent polar localization of the flagellin mRNA by antagonizing CsrA-mediated translational repression. The limited CsrA activity in WT cells under standard growth conditions, because of sequestration by the FliW protein antagonist, probably allows sufficient translation of *flaA* mRNA for its polar localization. Strong CsrA-mediated translation repression of *flaA* upon *fliW* deletion is probably responsible for the diffuse *flaA* localization in the Δ*fliW* mutant. CsrA binding might mediate storage of translationally inactive *flaA* mRNA until synthesis of FlaA is required or proper localization is achieved, similar to mRNP granules in eukaryotes[50]. Future studies will show whether other flagellar mRNAs also polarly localize and if the CsrA-FliW regulatory network also impacts their localization. CsrA-mediated regulation of mRNA localization might also occur in *B. subtilis* and *B. burgdorferi*, where CsrA overexpression represses the major flagellin[51–53]. An analogous system might have also evolved in the Alphaproteobacterium *Caulobacter crescentus*, which encodes two proteins with opposing activities on flagellin regulation, FlaF and FlbT, whereby FlbT post-transcriptionally regulates flagellin expression[54].

Our identification of *C. jejuni* CsrA titration by FliW indicates that CsrA-activity control by a protein antagonist, a mechanism first identified in the Gram-positive *B. subtilis*[9], is more widespread than previously appreciated. Besides the post-transcriptional effect of FliW on *flaA* and other flagellar genes by antagonizing CsrA, deletion of *fliW* directly or indirectly increases *flaA* transcription. Transcription of *hag* is also twofold upregulated in *B. subtilis* upon *fliW* deletion[9,55]. Although FliW appears to be the main CsrA antagonist, its synergistic interplay with the *flaA* mRNA antagonist affects other flagellar genes showed that RNA-based regulation can also impact CsrA activity in this type of Csr network. Gammaproteobacterial genomes encode CsrA[56] as well as the antagonizing sRNAs[5] and an anti-correlation between the presence of the CsrB/C sRNAs and FliW has been observed[57]. As the *csrA* gene is located next to a tRNA cluster in *E. coli*, this strongly suggests the pleiotropic function of CsrA in Gammaproteobacteria might have been horizontally acquired, followed by evolution of the antagonizing sRNAs. Thus, the conserved or possibly more ancient function of the CsrA-FliW system might be to mediate temporal and spatial control of proper flagellum assembly. During our conservation analysis we observed that certain non-flagellated *Campylobacter* species, such as *C. hominis*, *C. gracilis* and *C. ureolyticus*, lack *csrA* and *fliW* homologues, further supporting their conserved function in flagellar regulation. Further studies are required to unravel the full complexity of the CsrA-FliW regulatory network and its impact on RNA localization.

Methods

Bacterial strains, oligonucleotides and plasmids. All *C. jejuni* and *E. coli* strains used in this study are listed in Supplementary Table 3 and DNA oligonucleotides in Supplementary Table 4, respectively. Plasmids are summarized in Supplementary Table 5.

Bacterial growth conditions. *C. jejuni* strains were routinely grown on Müller-Hinton agar plates or with shaking in Brucella broth (BB), both supplemented with 10 μg ml^{-1} vancomycin, at 37 °C under microaerobic (10% CO_2, 5% O_2) conditions as described previously[14]. The agar was further supplemented with marker-selective antibiotics (20 μg ml^{-1} chloramphenicol, 50 μg ml^{-1} kanamycin , 20 μg ml^{-1} gentamicin or 250 μg ml^{-1} hygromycin B) where appropriate. *E. coli* strains were grown aerobically at 37 °C in Luria-Bertani (LB) medium supplemented with appropriate antibiotics. For induction of arabinose-inducible pBAD promoter, 0.001% (+) or 0.003% (++) L-arabinose was added to LB media.

Construction of bacterial mutant strains. All *C. jejuni* mutant strains (deletion, chromosomal 3xFLAG-tagging, chromosomal point mutations) were constructed using double-crossover homologous recombination. Cloning strategies and the generation of constructs are described in detail in the Methods and Supplementary Methods. Oligonucleotides used to amplify regions of upstream/downstream homology and resistance cassettes for homologous recombination, as well as recipient strains and oligonucleotides for validation of mutant strains by colony PCR, are listed in Supplementary Table 6 for each generated strain. Introduction of PCR products with 500 bp homologous ends or genomic DNA with mutant constructs into *C. jejuni* was performed by electroporation or natural transformation, respectively, as described previously[14].

Construction of 3xFLAG epitope-tagged proteins in *C. jejuni*. *C. jejuni* genes were chromosomally tagged at their C-terminus either by cloning of constructs for C-terminal epitope tagging on plasmids or by construction of 3xFLAG constructs by overlap PCR.

Tagging of proteins using PCR products amplified from plasmid constructs. The CsrA, FlaA, FlgI and FlaB proteins were fused to a 3xFLAG epitope at their C-termini by cloning regions encoding ~500 bp of their C-terminal coding region (C-term) and ~500 bp downstream of the stop codon (DN) into plasmid pGG1 to flank a 3xFLAG tag and *aphA-3* Kan^R cassette. Afterwards, the 3xFLAG-tag constructs were amplified by PCR and introduced into the chromosome of *C. jejuni* strains by electroporation and double-crossover homologous recombination. An example of this plasmid cloning strategy is described for *csrA*. Approximately 500 bp of the region downstream of *csrA* was amplified from genomic DNA (gDNA) with primers CSO-0173/-0174. These primers included *Xba*I and *Eco*RI sites, respectively. Following cleanup, the PCR product was digested with *Eco*RI and *Xba*I and ligated into a similarly digested pGG1 backbone, generated by inverse PCR with primers CSO-0074/-0075, to create pGD2-1. The plasmid was verified by colony PCR with primers JVO-0054/CSO-0173 and the sequence was verified using JVO-0054. Next, the backbone of this plasmid, including the *csrA* 'DN' region, was amplified by PCR with primers CSO-0073 (*Xho*I) and JVO-5142 (blunt). The C-terminal coding region of *csrA* (~500 bp) without the stop codon was amplified with primers CSO-0171/-0172 from NCTC11168 WT gDNA. The sense primer (CSO-0172) included an *Xho*I site, whereas the antisense primer (CSO-0171) contained a 5′-phosphate. Both the plasmid backbone with the 'DN' insert and the C-term insert were digested with *Xho*I and ligated to create plasmid pGD4-1. Integration of the PCR product was confirmed by colony PCR using primers CSO-0172/-0023 and the plasmid was validated by sequencing using CSO-0023. The entire integration cassette was then amplified with Phusion High-Fidelity DNA polymerase (NEB) using primers CSO-0172/-0173 and electroporated into *C. jejuni* and selected on kanamycin plates. Mutants were confirmed by colony PCR with primers CSO-0196/-0023 and western blot analysis with an anti-FLAG antibody.

3xFLAG tagging of proteins by overlap PCR. Construction of a C-terminal 3xFLAG translational fusion at its native locus was performed by overlap PCR for *flaG* as described in Supplementary Methods for gene deletions, but with the following modifications. The final overlap PCR product contained ~500 bp of the C-terminal coding region of *flaG* minus the stop codon (C-term) and ~500 bp downstream of *flaG* (DN) for homologous recombination. These regions flanked an in-frame 3xFLAG tag and stop codon followed by an *aphA-3* Kan^R cassette. For example, for tagging *flaG*, the 3xFLAG tag and Kan^R cassette was amplified from plasmid pGG1 with primers JVO-5142 and HPK2. The 'C-term' region of *flaG* was amplified using primers CSO-1002/-1098, where CSO-1098 is antisense and contains region of complementarity at its 5′ end to the 3xFLAG tag/JVO-5142, from NCTC11168 gDNA. The 'DN' region was amplified using primers CSO-1099/-1003, where CSO-1099 is sense to *flaG* DN and contains a region of complementarity to the 3′ end of the Kan^R cassette/primer HPK2. If the coding region of the target gene contained sequences required for expression of a downstream ORF (that is, SD sequence or codons), these sequences were included in the 'DN' amplicon. The three PCR products were then used for overlap PCR with primers CSO-1002/-1003, and the resulting amplicon was electroporated into *C. jejuni*, followed by selection of positive clones on kanamycin plates. Mutants

were checked by colony PCR with primers CSO-1005/HPK2 and western blot analysis with an anti-FLAG antibody.

Introducing chromosomal point mutations into the *flaA* leader. To introduce point mutations into the 5′UTR of *flaA* at the native locus, a 1,100-bp region around the *flaA* promoter was amplified using oligos CSO-0752/-0753. These primers introduced *Xho*I and *Xba*I sites, respectively, into the resulting PCR product. After *Xho*I and *Xba*I digestion, the product was then ligated into a similarly digested plasmid pJV752-1, resulting in plasmid pGD70-5. Plasmid pGD70-5 was checked by colony PCR using primers pZE-A/CSO-0753 and sequencing with pZE-A. Next, plasmid pGD70-5 was amplified by inverse PCR using primers CSO-0754/-0755, thereby introducing *Nde*I and *Bam*HI restriction sites 40 nt upstream of the *flaA* transcriptional start site (TSS). An *aac*(3)-IV gentamicin resistance cassette with its own promoter and terminator was amplified using CSO-0483/-0576 and introduced into PCR-amplified pGD70-5 in the reverse orientation to *flaA*, just upstream of its promoter, using the *Nde*I/*Bam*HI restriction sites, resulting in plasmid pGD76-1. Plasmid pGD76-1 was checked by colony PCR using primers CSO-0576/-0753 and sequencing with CSO-0753.

Point mutations were then introduced into the *flaA* 5′UTR by inverse PCR on pGD76-1 using complementary oligos harbouring the desired mutation, followed by *Dpn*I digestion and transformation of the resulting purified PCR product into *E. coli* TOP10. For introduction of the *flaA* M1 mutation (GGA > AAA in stem-loop SL1 of the *flaA* leader), oligonucleotides CSO-1114/-1115 were used for PCR on pGD76-1. The mutation was confirmed in the resulting plasmid pGD92-1 by sequencing with CSO-0753. Similarly, the *flaA* M2 (GGA > UGA in stem-loop SL1 of the *flaA* leader), M3 (GGA > GGG in stem-loop SL2 of the *flaA* leader), X1 start codon (AUG > AAG), X2 start codon (AUG > AUU), X3 start codon (AUG > GUG), X4 3rd codon (UUU > UAG), X5 3rd codon (UUU > UUC) and X6 101st codon (CAA > UAA) mutations were introduced using primer pairs CSO-0757/-0758, CSO-1116/-1117, CSO-2019/-2020, CSO-2827/-2828, CSO-2825/-2826, CSO-2829/-2830, CSO-2831/-2832 and CSO-2833/-2834, respectively, resulting in plasmids pGD77-1, pGD93-1, pGD114-2, pGD205-1, pGD204-1, pGD206-1, pGD207-1 and pGD208-1, respectively. For combination of the *flaA* M2 and M3 mutations, a similar mutagenesis approach was performed based on PCR amplification of the M2 plasmid pGD77-1 using oligonucleotides CSO-1116/-1117, resulting in pGD95-1 harbouring both the mutations. To introduce the *flaA* 5′UTR mutations into *C. jejuni*, a PCR product covering the homologous ends and the gentamicin resistance cassette was amplified from the respective WT (pGD76-1) or mutant plasmids using CSO-0752/-0850 and electroporated into *C. jejuni* as described above. To confirm introduction of point mutation in *C. jejuni,* colony PCR was performed using CSO-0576/-0753 and sequencing with CSO-0850.

Construction of *E. coli* mutants. The *E. coli* Δ*pgaA* and Δ*pgaA* Δ*csrA* deletion strains were constructed in the TOP10 background using the λ Red protocol[58]. Briefly, a kanamycin resistance gene, amplified from plasmid pKD4 using primers CSO-0652/-0653, was used to replace the entire *pgaA* ORF excluding the start and stop codon. The mutant strain was verified by colony PCR using the primer pairs CSO-0654/-0653 and CSO-0652/-0655. After verification, helper plasmid pCP20 containing FLP recombinase was introduced to remove the kanamycin resistance marker[58]. The helper plasmid, which is temperature-sensitive and carries an ampicillin resistance marker, was then cured by recovering colonies at 37 °C and confirming ampicillin sensitivity, resulting in strain CSS-0556. Similarly, the ORF of the *csrA* gene excluding the start and stop codon was then replaced by the kanamycin resistance marker (amplified using CSO-0611/-0612) in the Δ*pgaA* strain resulting in strain CSS-0557, harbouring both *pgaA* and *csrA* deletions. The *csrA* deletion was verified by colony PCR using primer pairs CSO-0639/-0612 and CSO-0611/-0640.

RIP-seq of *C. jejuni* CsrA-3xFLAG. coIP combined with RNA-seq (RIP-seq) to identify direct RNA-binding partners of CsrA-3xFLAG in *C. jejuni* was performed as previously described[18,59] with minor modifications.

CoIP of RNA with CsrA-3xFLAG. CoIP of chromosomally epitope-tagged *C. jejuni* CsrA with an anti-FLAG antibody and Protein A-Sepharose beads was performed from lysates of *C. jejuni* NCTC11168 and 81-176 WT (control) and isogenic *csrA*-3xFLAG strains grown in 100 ml (50 ml × 2 flasks) BB containing 10 μg ml^{-1} vancomycin to mid-exponential phase (OD_{600} = 0.6) at 37 °C as described previously for *H. pylori*[18]. Cells were harvested by centrifugation at 6,000*g* for 15 min at 4 °C. Afterwards, cell pellets were resuspended in 1 ml Buffer A (20 mM Tris-HCl, pH 8.0, 150 mM KCl, 1 mM $MgCl_2$, 1 mM dithiothreitol (DTT)) and subsequently centrifuged (3 min, 11,000*g*, 4 °C). The pellets were shock-frozen in liquid nitrogen and stored at − 80 °C. Frozen pellets were thawed on ice and resuspended in 0.8 ml Buffer A. An equal volume of glass beads was then added to the cell suspension. Cells were then lysed using a Retsch MM40 ball mill (30 s^{-1}, 10 min) in pre-cooled blocks (4 °C) and centrifuged for 2 min at 15,200*g*, 4 °C. The supernatant was transferred to a new tube, and an additional 0.4 ml of Buffer A was added to the remaining un-lysed cells with beads. Lysis of the remaining cells was achieved by a second round of lysis at 30 s^{-1} for 5 min. Centrifugation was repeated and this second supernatant was combined with the first one. The combined supernatant was centrifuged again for 30 min at 15,200*g*, 4 °C for clarification and the resulting supernatant (lysate fraction) was transferred to a new tube. The lysate was incubated with 35 μl anti-FLAG antibody (Monoclonal ANTI-FLAG M2, Sigma, #F1804) for 30 min at 4 °C on a rocker. Next, 75 μl of Protein A-Sepharose (Sigma, #P6649), prewashed with Buffer A, was added and the mixture was rocked for another 30 min at 4 °C. After centrifugation at 15,200*g* for 1 min, the supernatant was removed. Pelleted beads were washed five times with 0.5 ml Buffer A. Finally, 500 μl Buffer A was added to the beads and RNA and proteins were separated by phenol-chloroform-isoamyl alcohol extraction and precipitated as described previously[18]. From each coIP, 700–1,000 ng of RNA was recovered. 100 μl of 1 × protein loading buffer (62.5 mM Tris-HCl, pH 6.8, 100 mM DTT, 10% (v/v) glycerol, 2% (w/v) SDS, 0.01% (w/v) bromophenol blue) was added to the final protein sample precipitated along with beads. This sample was termed the coIP sample. For verification of a successful coIP, protein samples equivalent to 1.0 OD_{600} of cells were obtained during different stages of the coIP (culture, lysate, supernatant, wash and coIP (beads)) for further western blot analysis. One hundred microlitres of 1 × protein loading buffer was added to the protein samples and boiled for 8 min. Protein sample corresponding to an OD_{600} of 0.1 or 0.15 (culture, lysate, supernatant and wash fraction) and 10 or 5 (for proteins precipitated from beads) were used for western blot analysis.

RIP-Seq cDNA library preparation. Residual gDNA was removed from the coIP RNA samples isolated from the control (WT) and CsrA-3xFLAG coIPs of the two strains *C. jejuni* NCTC11168 and 81-176 using DNase I treatment. cDNA libraries for Illumina sequencing were constructed by vertis Biotechnologie AG (http://www.vertis-biotech.com) in a strand-specific manner as described previously[14]. In brief, equal amounts of RNA samples were poly(A)-tailed using poly(A) polymerase. Then, 5′-triphosphates were removed using tobacco acid pyrophosphatase, and an RNA adapter was then ligated to the resulting 5′-monophosphate. First-strand cDNA was synthesized with an oligo(dT)-adapter primer using M-MLV reverse transcriptase. In a PCR-based amplification step, using a high-fidelity DNA polymerase, the cDNA concentration was increased to 20–30 ng μl^{-1}. For all libraries, the Agencourt AMPure XP kit (Beckman Coulter Genomics) was used to purify the DNA, which was subsequently analysed by capillary electrophoresis.

A library-specific barcode for multiplex sequencing was included as part of a 3′-sequencing adapter. The following adapter sequences flank the cDNA inserts:

TrueSeq_Sense_primer

5′-AATGATACGGCGACCACCGAGATCTACACTC TTTCCCTACACGAC GCTCTTCCGATCT-3′

TrueSeq_Antisense_NNNNNN_primer (NNNNNN = 6nt barcode for multiplexing)

5′-CAAGCAGAAGACGGCATACGAGAT-NNNNNN-GTGACTGGAG TTCAGACGTGTGCTCTTCCGATC(dT25)-3′.

The samples were sequenced on an Illumina HiSeq instrument with 100 cycles in single-read mode. The resulting read numbers are listed in Supplementary Table 1.

Analysis of deep sequencing data. To assure high sequence quality, the Illumina reads in FASTQ format were trimmed with a cutoff phred score of 20 by the programme fastq_quality_trimmer from FASTX toolkit version 0.0.13. After trimming, poly(A)-tail sequences were removed and a size filtering step was applied in which sequences shorter than 12 nt were eliminated. The collections of remaining reads were mapped to the *C. jejuni* NCTC11168 (NCBI Acc.-No: NC_002163.1) and 81-176 (NCBI Acc.-No: NC_008770.1, NC_008787.1, NC_008790.1) genomes using *segemehl*[60] with an accuracy cutoff of 95%. Mapping statistics are listed in Supplementary Table 1. Coverage plots representing the numbers of mapped reads per nucleotide were generated. Reads that mapped to multiple locations contributed a fraction to the coverage value. For example, reads mapping to three positions contributed only one-third to the coverage values. Each graph was normalized to the number of reads that could be mapped from the respective library. To restore the original data range, each graph was then multiplied by the minimum number of mapped reads calculated over all libraries.

The overlap of sequenced cDNA reads to annotations was assessed for each library by counting all reads overlapping selected annotations on the sense strand. These annotations consist of strain-specific NCBI gene annotations complemented with annotations of previously determined 5′UTRs and small RNAs[14]. Each read with a minimum overlap of 10 nt was counted with a value based on the number of locations where the read was mapped. If the read overlapped more than one annotation, the value was divided by the number of regions and counted separately for each region (for example, one-third for a read mapped to three locations).

Enrichment analysis of CsrA targets. Enrichment of transcripts in the CsrA-3xFLAG coIP versus control coIP libraries was determined based on mapped cDNA read counts for annotations provided in NC_002163.gff (NCBI) for NCTC11168 using GFOLD version 1.0.9 (ref. 61) but with manually defined normalization constants based on the number of reads that could be mapped to the respective libraries. For determination of genes enriched in the CsrA-3xFLAG-tagged library, log2 fold changes (FCs) rather than GFOLD values were used.

Similar analysis was done for strains 81-176 using annotations provided in NC_008787.gff (chromosome), NC_008770.gff (pVir plasmid) and NC_008790.gff (pTet plasmid).

Peak detection and CsrA-binding motif analyses. To automatically define CsrA-bound RNA regions or peaks from the CsrA-3xFLAG coIP data sets, an in-house tool 'sliding_window_peak_calling_script' was developed based on a sliding window approach. A detailed description of the tool will be described elsewhere. The script has been deposited at Zenodo (https://zenodo.org/record/49292) under DOI 10.5281/zenodo.49292 (http://dx.doi.org/10.5281/zenodo.49292). The script is written in Python 3 and requires installation of the Python 3 packages *numpy* and *scipy* for execution.

In brief, the 'sliding_window_peak_calling_script' software uses normalized wiggle files of the CsrA-3xFLAG and control coIP libraries as input to determine sites showing a continuous enrichment of the CsrA-3xFLAG-tagged library compared with the control. The identification of enriched regions is based on four parameters: a minimum required fold change (FC) for the enrichment, a factor multiplied by the 90th percentile of the wiggle graph, which reflects the minimum required expression (MRE) in the tagged library, a window size in nt (WS), for which the previous two values are calculated in a sliding window approach, and a nucleotide step size (SS), which defines the steps in which the window is moved along the genomic axis. All consecutive windows that fulfill the enrichment requirements are assembled into a single peak region. The peak detection is performed separately for the forward and reverse strand of each replicon. For the CsrA-3xFLAG coIP data set, the following parameters were used: FC = 5, MRE = 3, WS = 25 and SS = 5.

For the prediction of consensus motifs based on the peak sequences, MEME[24] and CMfinder 0.2.1 (ref. 62) were used. For MEME[24] predictions, the following settings were applied: Search 0 or 1 motif of length 4–7 bp per sequence in the given strand only. To search for the presence of a structural motif, CMfinder 0.2.1 (ref. 62) was run on the enriched peak sequences with default parameters except for allowing a minimum single stem loop candidate length of 20 nt. The top-ranked motif incorporated 276 of the 328 sequences and was visualized by R2R[63].

Functional classes enrichment analysis. To check for overrepresentation of functional classes of CsrA-bound genes, we considered genes with at least fivefold enrichment in their 5′UTR and/or coding sequence in the CsrA-3xFLAG coIP library (versus control) as CsrA-bound and the remaining genes as unbound. We applied an existing functional classification[64] of genes from strain NCTC11168 to determine statistically enriched functional classes. Because a similar classification was not available for strain 81-176, a table with orthologue mappings between the two strains was downloaded from OrthologueDB[65] and used to assign the NCTC11168 functional classes to their respective 81-176 counterparts. Genes in our annotation lists without an existing functional classification in NCTC11168 or without an orthologue match were assigned to class 5.I, defined as 'Unknown', in the original classification scheme. Genes encoded on the pVir and pTet plasmids of strain 81-176 were assigned to new pVir and pTet classes, respectively. Functional overrepresentation was analysed for each functional class via a two-sided Fisher's exact test followed by multiple-testing correction using the Benjamini–Hochberg method. An adjusted *P*-value of 0.05 was selected as significance threshold for functional overrepresentation.

Protein–protein coIP. The FliW and CsrA protein–protein coIP was performed exactly as described for the RIP-seq coIP protocol (see above) until the step where beads were washed five times with Buffer A. After washing, the beads were suspended in 200 µl of 1 × protein loading buffer (62.5 mM Tris-HCl, pH 6.8, 100 mM DTT, 10% (v/v) glycerol, 2% (w/v) SDS, 0.01% (w/v) bromophenol blue) and boiled for 8 min. Lysate samples corresponding to an OD_{600} of 0.05 and 2 (for proteins precipitated from beads) were used for western blot analysis.

SDS-PAGE and immunoblotting. Protein analyses were performed on cells collected from *C. jejuni* in mid-exponential phase (OD_{600} 0.5–0.6) or *E. coli* cultures in late-exponential phase (OD_{600} 1.0–1.5). Cells were collected by centrifugation at 11,000*g* for 3 min. Cell pellets were resuspended in 100 µl of 1 × protein loading buffer (62.5 mM Tris-HCl, pH 6.8, 100 mM DTT, 10% (v/v) glycerol, 2% (w/v) SDS, 0.01% (w/v) bromophenol blue) and boiled for 8 min. For western blot analysis, samples corresponding to an OD_{600} of 0.02 to 0.1 were separated by 12, 15 or 18% (v/v) SDS-polyacrylamide (PAA) gels and transferred to a nitrocellulose membrane by semidry blotting. Membranes were blocked for 1 h with 10% (w/v) milk powder/TBS-T (Tris-buffered saline-Tween-20) and incubated overnight with primary antibody at 4 °C. Membranes were then washed with TBS-T, followed by 1 h incubation with secondary antibody. After washing, the blot was developed using enhanced chemiluminescence-reagent. GFP-, FLAG- and Strep-tagged proteins of interest were detected with monoclonal anti-GFP (1:1,000 in 3% BSA/TBS-T; Roche, #11814460001), monoclonal anti-FLAG (1:1,000 in 3% BSA/TBS-T; Sigma-Aldrich, #F1804-1MG) or monoclonal anti-Strep (1:10,000 in 3% BSA/TBS-T; IBA GmbH, #2-1507-001) primary antibodies and anti-mouse IgG (1:10,000 in 3% BSA/TBS-T; GE-Healthcare, #RPN4201) secondary antibody. mCherry-tagged proteins were detected using a polyclonal anti-mCherry (1:4,000 in 3% BSA/TBS-T; Acris, #AB0040-20) primary antibody and an anti-goat (1:10,000 in 3% BSA/TBS-T; Santa Cruz Biotechnology, #sc2020) secondary antibody. A monoclonal antibody specific for GroEL (1:10,000 in 3% BSA/TBS-T; Sigma-Aldrich, # G6532-5ML) and an anti-rabbit IgG (1:10,000 in 3% BSA/TBS-T; GE-Healthcare, #RPN4301) secondary antibody were used as a loading control. Images of full blots that were cropped in main Figures are shown in Supplementary Fig. 19.

Validation of CsrA targets with a GFP reporter system. Validation of CsrA targets was performed using a heterologous *E. coli* system previously developed for validation of sRNA–mRNA interactions[23]. Selected candidate *C. jejuni* CsrA target sequences from the coIP were cloned as translational fusions to GFP or FLAG in plasmids pXG-10 or pXG-30 as listed in Supplementary Tables 5 and 7. Levels of FLAG or GFP translational fusions were then determined by western blotting or FACS in *E. coli* Δ*pgaA*, Δ*pgaA* Δ*csrA* and a Δ*pgaA* Δ*csrA* strain harbouring plasmid pGD72-3 with *C. jejuni* CsrA-Strep under the control of an arabinose-inducible promoter.

Flow cytometric analysis. For FACS analysis of GFP reporter fluorescence in *E. coli*, cells corresponding to 1 OD_{600} were collected from LB cultures in log phase and resuspended in 0.25 ml PBS. Cells were then fixed for 10 min with 0.25 ml of 4% paraformaldehyde, collected by centrifugation and washed twice with 0.5 ml PBS before final resuspension in 0.5 ml PBS. A 1/100 dilution of the fixed sample in PBS was used for measurement. Measurements (50 000 counts per sample) were performed on a BD FACSCalibur machine and analysed using FlowJo (V10).

Purification of *C. jejuni* CsrA. Recombinant, C-terminal Strep-tagged *C. jejuni* CsrA (Cj1103) was overexpressed and purified from *E. coli* TOP10 Δ*pgaA*/Δ*csrA* using Strep-Tactin Sepharose (IBA GmbH, #2-1202-001). Primers and plasmids used for cloning are listed in Supplementary Tables 4 and 5 The *csrA* gene, including its SD sequence, was fused to a C-terminal Strep-tag in the arabinose-inducible plasmid pBAD/Myc-His A (Invitrogen) for overexpression and affinity purification. The *csrA*-coding region and SD were amplified from *C. jejuni* NCTC11168 genomic DNA using primers CSO-0746/-0747, and the pBAD/Myc-His A plasmid was amplified by inverse PCR with JVO-0900/-0901 as previously described[66]. CSO-0747 and JVO-0901 introduce an *Xba*I site to the insert and vector, respectively, whereas CSO-0746 has a 5′-phosphate to facilitate blunt-end ligation. *Xba*I-digested insert and vector were then ligated, resulting in pGD68-1. Plasmid pGD68-1 was checked by colony PCR using primers pBAD-FW/CSO-0747 and sequencing with pBAD-FW. A Strep-tag (WSHPQFEK) was then added at the C-terminus of *csrA* by inverse PCR using oligonucleotides CSO-0852/-0853, resulting in plasmid pGD72-3. Plasmid pGD72-3 was checked by sequencing with pBAD-FW. Plasmid pGD72-3 was then introduced into an *E. coli* TOP10 Δ*pgaA*/Δ*csrA* deletion strain resulting in strain CSS-0931. CSS-0931 was grown in 500 ml LB broth with 100 µg ml^{-1} of ampicillin at 37 °C and shaking at 220 r.p.m. to an OD_{600} of 0.3, at which time L-arabinose was added to a final concentration of 0.01%. The culture was then incubated for an additional 8 h at 18 °C. Cells were harvested by centrifugation at 7,000*g* for 30 min at 4 °C. The pellet was resuspended in 5 ml of Buffer W (IBA GmbH, #2-1003-100). The rest of the protocol was followed as per the manufacturer's instructions using 1 ml Gravity flow Strep-Tactin Sepharose. After washing steps, the CsrA-Strep protein was finally eluted using Buffer E (IBA GmbH, #2-1000-025) in three successive steps (E1: 0.8 ml, E2: 1.4 ml and E3: 0.8 ml). The majority of CsrA-Strep was concentrated in the E2 fraction. Concentration was quantified using Roti-Quant (Carl ROTH, #K015.3), and the protein was stored at − 20 °C in 50 µl aliquots.

RNA isolation. Bacteria were grown to the indicated growth phase and culture volume corresponding to a total amount of 4 OD_{600} was harvested and mixed with 0.2 volumes of stop-mix (95% ethanol and 5% phenol, vol/vol). The samples were snap-frozen in liquid nitrogen and stored at − 80 °C until RNA extraction. Frozen samples were thawed on ice and centrifuged at 4 °C to collect cell pellets. Cell pellets were lysed by resuspension in 600 µl of a solution containing 0.5 mg ml^{-1} lysozyme in TE buffer (pH 8.0) and 60 µl of 10% SDS. The samples were incubated for 1–2 min at 65 °C to ensure lysis. Afterwards, total RNA was extracted using the hot-phenol method as described previously[13,14].

Northern blot analysis. For northern blot analysis, 5–10 µg RNA sample was loaded per lane. After separation on 6% PAA gels containing 7 M urea, RNA was transferred to Hybond-XL membranes (GE-Healthcare) by electroblotting. After blotting, the RNA was ultraviolet cross-linked to the membrane and hybridized with γ32P-ATP end-labelled DNA oligonucleotides (Supplementary Table 4).

Rifampicin RNA stability assays. To determine the stability of *flaA* mRNA in *C. jejuni* NCTC11168 WT, Δ*csrA*, Δ*fliW* and Δ*csrA* Δ*fliW* strains, cells were grown to an OD_{600} of 0.45 (mid-log phase) and treated with rifampicin to a final concentration 500 µg ml^{-1}. Samples were harvested for RNA isolation at indicated time points following rifampicin addition (0, 4, 8, 16 and 32 min) as described

above. After RNA isolation, 10 μg of each RNA sample was used for northern blot analysis as detailed above.

***In-vitro* T7 transcription and RNA labelling.** DNA templates containing the T7 promoter sequence were generated by PCR using oligos and DNA templates listed in Supplementary Table 8. T7 *in-vitro* transcription of RNAs was carried out using the MEGAscript T7 kit (Ambion) and sequences of the resulting T7 transcripts are listed in Supplementary Table 8. *In vitro* transcribed RNAs were quality checked and 5′ end-labelled (γ^{32}P) as previously described[66,67].

Gel mobility shift assays. Gel-shift assays were performed using ~0.04 pmol 5′-labelled RNA (4 nM final concentration) with increasing amounts of purified *C. jejuni* CsrA in 10 μl reactions. In brief, 5′-radiolabelled RNA (^{32}P, 0.04 pmol in 6 μl) was denatured (1 min, 95 °C) and cooled for 5 min on ice. Yeast tRNA (1 μg) and 1 μl of 10 × RNA Structure Buffer (Ambion: 10 mM Tris, pH 7, 100 mM KCl, 10 mM $MgCl_2$) was then added to the labelled RNA. CsrA protein (2 μl diluted in 1 × Structure Buffer) was added to the desired final concentrations (0 mM, 10 nM, 20 nM, 50 nM, 100 nM, 200 nM, 500 nM, 1 μM or 2 μM CsrA). Binding reactions were incubated at 37 °C for 15 min. Before loading on a pre-cooled native 6% PAA, 0.5 × TBE gel, samples were mixed with 3 μl native loading buffer (50% (v/v) glycerol, 0.5 × TBE, 0.2% (w/v) bromophenol blue). Gels were run in 0.5 × TBE buffer at 300 V at 4 °C for 3 h. Gels were dried and analysed using a PhosphoImager (FLA-3000 Series, Fuji).

***In vitro* structure probing assays.** *In vitro* structure probing of *flaA* WT and *flaA* M1/M2 leaders with RNase T1 and lead(II) acetate was performed as previously described[68]. For each reaction, 0.1 pmol of a labelled *flaA* leader variant was denatured for 1 min at 95 °C and chilled on ice for 5 min. One microgram yeast tRNA as competitor and 10 × RNA Structure Buffer was added (provided together with RNase T1, Ambion). Unlabelled recombinant *C. jejuni* CsrA protein was then added at 0-, 20-, 50- or 100-fold molar excess. After incubation for 15 min at 37 °C, 2 μl RNase T1 (0.01 U μl^{-1}) or 2 μl freshly prepared lead(II)-acetate solution (25 mM) were added and reactions were incubated for 3 min or 90 s, respectively. As a control, ~0.1 pmol labelled RNA with 100-fold excess CsrA was also prepared without nuclease/lead(II) treatment. The reactions were stopped by addition of 12 μl Gel loading buffer II (#AM8546G, Ambion). For RNase T1 ladders, ~0.1 pmol labelled RNA was denatured in 1 × Structure Buffer for 1 min at 95 °C and afterwards incubated with 0.1 U μl^{-1} RNase T1 for 5 min. The OH ladder was generated by incubation of ~0.1 pmol labelled *flaA* WT leader RNA in 1 × alkaline hydrolysis buffer (Ambion) for 5 min at 95 °C. Ladders and samples were then separated on 10% (v/v) PAA/7M urea gels in 1 × TBE buffer. Gels were dried, exposed to a screen and analysed using a PhosphorImager (FLA-3000 Series, Fuji).

Transmission electron microscopy. *C. jejuni* WT and mutant strains were grown for 14 h on MH plates supplemented with vancomycin (10 μg ml^{-1}). Cells were resuspended gently in PBS using a cotton swab and centrifuged at 5,000*g* for 5 min. The cell pellet was resuspended in 2% glutaraldehyde in 0.1 M cacodylate and incubated at 4 °C overnight. The next day, samples were stained with 2% uranyl acetate and imaged using a Zeiss EM10 transmission electron microscope.

Motility assays. *C. jejuni* strains were inoculated from the appropriate selective MH agar plates into 20 ml BB containing 10 μg ml^{-1} vancomycin and grown microaerobically with shaking at 37 °C to an OD_{600} of ~0.5. Cells were harvested by centrifugation at 6,500*g* for 5 min and resuspended at an OD_{600} of 0.5 in BB. For each strain, 0.5 μl of bacterial suspension was inoculated into motility soft-agar plates (MH broth + 0.4% agar) poured the day before. Plates were incubated right-side-up for ~24 h microaerobically at 37 °C. Three measurements of each motility halo were made for each inoculation, which were averaged to give the mean swim distance for each strain on a plate. All strains were inoculated together on six replicate plates and the mean swim distance ± standard error on these plates was used to the compare motility of each strain.

Autoagglutination assay. Autoagglutination was determined as described previously[26]. Briefly, strains grown in liquid cultures for motility assays were resuspended in PBS, pH 7.4, to an OD_{600} of 1.0. Two millilitres were placed into three replicate tubes and the OD_{600} was measured. Tubes were incubated at 37 °C microaerobically without shaking, and at indicated time points, 100 μl was carefully removed from the top of the suspension, diluted tenfold in PBS, and the OD_{600} was measured. Measurements were normalized to the optical density of each strain at the zero time point.

Time-lapse microscopy to monitor cell division. *C. jejuni* Δ*fliA* mutant cells corresponding to an OD_{600} of 0.5 were collected from BB culture in log phase by centrifugation and resuspended in 0.5 ml BB. The cells were further serially diluted 100- and 1,000-fold in BB. Five microlitres of the diluted samples were spotted on a BB-agarose (1%) plate. The plate was incubated under microaerophilic conditions at 37 °C for 10 min. The agarose patch was excised and inverted onto a Petri dish with a glass bottom. Single cells were then monitored over time using several bright-field images in a fluorescence microscope (Leica DMI6000 B) maintained at 37 °C under aerobic conditions.

RNA FISH. RNA-FISH was performed as previously described[69] with some modifications. A total amount of cells corresponding to two OD_{600} was collected from BB cultures in mid-log phase (OD_{600} = 0.4) and resuspended in 0.5 ml PBS. Cells were then fixed for 3 h with 0.5 ml 4% paraformaldehyde at room temperature, collected by centrifugation and washed twice with 0.5 ml PBS before final resuspension in 0.5 ml 70% ethanol. After 10 min, cells were collected by centrifugation and resuspended in 95% ethanol and incubated at room temperature for 1 h. Cells were again collected by centrifugation, completely dried in a laminar flow hood and then washed once with 2 × SSC before final resuspension in 0.5 ml of 2 × SSC containing 10% formamide. Fluorescently labelled DNA oligos (14 Cy5-labelled oligos to detect *flaA* mRNA and one FITC-labelled oligo specific for 16 S rRNA, Sigma, Supplementary Table 4) were then added at a concentration of 10 ng μl^{-1} and incubated at 37 °C overnight. The next day, cells were collected by centrifugation and washed three times for 1 h at 37 °C with 0.5 ml of 2 × SSC containing 10% formamide before final resuspension in 2 × SSC (50–250 μl). Cells were then imaged in a Leica Confocal TCS SP5 II microscope using sequential scanning mode.

***d*STORM.** For super-resolution imaging, *C. jejuni* cells were grown, fixed and labelled using the above-described RNA-FISH protocol (14 Cy5-labelled DNA oligonucleotides to detect *flaA* mRNA and a FITC oligo to label 16S rRNA, Sigma, Supplementary Table 4). Labelled cells were immobilized on poly-D-lysine (Sigma-Aldrich)-coated eight-well chambered cover glasses (Sarstedt). For fluorophore photo switching, a buffer with a pH of 8.3–8.5 was used[30,70] containing 50 mM Tris-HCl (pH 8), 10% glucose, 1% 2-mercaptoethanol (Carl Roth), 3 U ml^{-1} pyranose oxidase (Sigma-Aldrich) and 90 U ml^{-1} catalase (Sigma-Aldrich) in 2 × SSC.

*d*STORM was performed on a wide-field setup for localization microscopy[71]. An optically pumped semiconductor laser (Genesis MX STM-Series, Coherent) with a wavelength of 639 nm (maximum power of 1 W) was used for excitation of Cy5 and a diode laser (iBeam smart Family, TOPTICA Photonics) with a wavelength of 405 nm (maximum power of 120 mW) was used for reactivation of Cy5. Laser beams were cleaned-up by bandpass filters (Semrock/Chroma) and combined by appropriate dichroic mirrors (LaserMUX filters, Semrock). Afterwards they were focused onto the back focal plane of the high numerical oil-immersion objective (Olympus APON 60XO TIRF, numerical aperture 1.49), which is part of an inverted fluorescence microscope (Olympus IX71). To separate the excitation light from the fluorescence light, suitable dichroic beam splitters (Semrock) were placed into the light path before the laser beams enter the objective. Fluorescence light collected by the objective was filtered by appropriate detection filters (Semrock/Chroma) and was detected by an EMCCD camera with 512 × 512 pixels (iXon Ultra 897, Andor Technology). The pixel size in the image was 129 nm px^{-1}. Cy5 was excited with the 639-nm laser at a maximum intensity of 4.19 kW cm^{-2}. During imaging, the 405-nm laser was switched on to keep up a suitable switching ratio. Its laser power was increased successively to a maximum intensity of 0.04 kW cm^{-2}. For every image, 5,000–25,000 frames were taken with an integration time of 15 ms per frame. For every imaged area, additionally a bright-field image was taken to identify single bacteria. Data analysis was performed using rapid*d*STORM open source software[72].

Statistical analysis. All data for western, northern blot or FISH analysis are presented as mean ± s.e.m. Statistical analysis was carried out using Student's *t*-test. For statistical comparison of two groups, a two-tailed paired Student's *t*-test was used. A value of $P < 0.05$ was considered significant and marked with an asterisk (*) as explained in the legends. For FISH analysis, fluorescence data curves from 10 cells from a single image were merged as a single averaged curve after cell length normalization. The data were acquired and normalized over cell length using ImageJ and subsequently the merged average curve was generated using Microsoft Excel.

Code availability. The 'sliding_window_peak_calling_script' for identification of CsrA-binding sites based on RIP-seq data has been deposited at Zenodo (https://zenodo.org/record/49292) under DOI: 10.5281/zenodo.49292 (http://dx.doi.org/10.5281/zenodo.49292).

Data availability. The raw, de-multiplexed reads as well as coverage files of the RIP-seq libraries have been deposited in the NCBI Gene Expression Omnibus[73] under the accession number GSE58419. The authors declare that all other data supporting the findings of this study are available within the article and its supplementary information files, or from the corresponding author upon request.

References

1. Kartha, R. V. & Subramanian, S. Competing endogenous RNAs (ceRNAs): new entrants to the intricacies of gene regulation. *Front. Genet.* **5,** 8 (2014).
2. Figueroa-Bossi, N., Valentini, M., Malleret, L., Fiorini, F. & Bossi, L. Caught at its own game: regulatory small RNA inactivated by an inducible transcript mimicking its target. *Genes Dev.* **23,** 2004–2015 (2009).
3. Miyakoshi, M., Chao, Y. & Vogel, J. Cross talk between ABC transporter mRNAs via a target mRNA-derived sponge of the GcvB small RNA. *EMBO J.* **34,** 1478–1492 (2015).
4. Miyakoshi, M., Chao, Y. & Vogel, J. Regulatory small RNAs from the 3' regions of bacterial mRNAs. *Curr. Opin. Microbiol.* **24C,** 132–139 (2015).
5. Romeo, T., Vakulskas, C. A. & Babitzke, P. Post-transcriptional regulation on a global scale: form and function of Csr/Rsm systems. *Environ. Microbiol.* **15,** 313–324 (2013).
6. Heroven, A. K., Bohme, K. & Dersch, P. The Csr/Rsm system of *Yersinia* and related pathogens: a post-transcriptional strategy for managing virulence. *RNA Biol.* **9,** 379–391 (2012).
7. Babitzke, P. & Romeo, T. CsrB sRNA family: sequestration of RNA-binding regulatory proteins. *Curr. Opin. Microbiol.* **10,** 156–163 (2007).
8. Duss, O. *et al.* Structural basis of the non-coding RNA RsmZ acting as a protein sponge. *Nature* **509,** 588–592 (2014).
9. Mukherjee, S. *et al.* CsrA-FliW interaction governs flagellin homeostasis and a checkpoint on flagellar morphogenesis in *Bacillus subtilis*. *Mol. Microbiol.* **82,** 447–461 (2011).
10. Fields, J. A. & Thompson, S. A. *Campylobacter jejuni* CsrA mediates oxidative stress responses, biofilm formation, and host cell invasion. *J. Bacteriol.* **190,** 3411–3416 (2008).
11. Barnard, F. M. *et al.* Global regulation of virulence and the stress response by CsrA in the highly adapted human gastric pathogen *Helicobacter pylori*. *Mol. Microbiol.* **51,** 15–32 (2004).
12. Kao, C. Y., Sheu, B. S. & Wu, J. J. CsrA regulates *Helicobacter pylori* J99 motility and adhesion by controlling flagella formation. *Helicobacter* **19,** 443–454 (2014).
13. Sharma, C. M. *et al.* The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464,** 250–255 (2010).
14. Dugar, G. *et al.* High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. *PLoS Genet.* **9,** e1003495 (2013).
15. Porcelli, I., Reuter, M., Pearson, B. M., Wilhelm, T. & van Vliet, A. H. Parallel evolution of genome structure and transcriptional landscape in the Epsilonproteobacteria. *BMC Genomics* **14,** 616 (2013).
16. Taveirne, M. E., Theriot, C. M., Livny, J. & DiRita, V. J. The complete *Campylobacter jejuni* transcriptome during colonization of a natural host determined by RNAseq. *PLoS One* **8,** e73586 (2013).
17. Sittka, A. *et al.* Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet.* **4,** e1000163 (2008).
18. Rieder, R., Reinhardt, R., Sharma, C. M. & Vogel, J. Experimental tools to identify RNA-protein interactions in *Helicobacter pylori*. *RNA Biol.* **9,** 520–531 (2012).
19. Martin, K. C. & Ephrussi, A. mRNA localization: gene expression in the spatial dimension. *Cell* **136,** 719–730 (2009).
20. Montero Llopis, P. *et al.* Spatial organization of the flow of genetic information in bacteria. *Nature* **466,** 77–81 (2010).
21. Nevo-Dinur, K., Nussbaum-Shochat, A., Ben-Yehuda, S. & Amster-Choder, O. Translation-independent localization of mRNA in *E. coli*. *Science* **331,** 1081–1084 (2011).
22. Lertsethtakarn, P., Ottemann, K. M. & Hendrixson, D. R. Motility and chemotaxis in *Campylobacter* and *Helicobacter*. *Annu. Rev. Microbiol.* **65,** 389–410 (2011).
23. Urban, J. H. & Vogel, J. Translational control and target recognition by *Escherichia coli* small RNAs in vivo. *Nucleic Acids Res.* **35,** 1018–1037 (2007).
24. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**(Web Server issue): W202–W208 (2009).
25. Dubey, A. K., Baker, C. S., Romeo, T. & Babitzke, P. RNA sequence and secondary structure participate in high-affinity CsrA-RNA interaction. *RNA* **11,** 1579–1587 (2005).
26. Golden, N. J. & Acheson, D. W. Identification of motility and autoagglutination *Campylobacter jejuni* mutants by random transposon mutagenesis. *Infect. Immun.* **70,** 1761–1771 (2002).
27. de Vries, S. P. *et al.* Motility defects in *Campylobacter jejuni* defined gene deletion mutants caused by second-site mutations. *Microbiology* **161,** 2316–2327 (2015).
28. Barrero-Tobon, A. M. & Hendrixson, D. R. Identification and analysis of flagellar coexpressed determinants (Feds) of *Campylobacter jejuni* involved in colonization. *Mol. Microbiol.* **84,** 352–369 (2012).
29. Heilemann, M. *et al.* Subdiffraction-resolution fluorescence imaging with conventional fluorescent probes. *Angew. Chem.* **47,** 6172–6176 (2008).
30. Fei, J. *et al.* RNA biochemistry. Determination of in vivo target search kinetics of regulatory noncoding RNA. *Science* **347,** 1371–1374 (2015).
31. Lawhon, S. D. *et al.* Global regulation by CsrA in *Salmonella typhimurium*. *Mol. Microbiol.* **48,** 1633–1645 (2003).
32. Brencic, A. & Lory, S. Determination of the regulon and identification of novel mRNA targets of *Pseudomonas aeruginosa* RsmA. *Mol. Microbiol.* **72,** 612–632 (2009).
33. Edwards, A. N. *et al.* Circuitry linking the Csr and stringent response global regulatory systems. *Mol. Microbiol.* **80,** 1561–1580 (2011).
34. Holmqvist, E. *et al.* Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. *EMBO J.* **35,** 991–1011 (2016).
35. Yakhnin, H. *et al.* CsrA represses translation of sdiA, which encodes the N-acylhomoserine-L-lactone receptor of *Escherichia coli*, by binding exclusively within the coding region of *sdiA* mRNA. *J. Bacteriol.* **193,** 6162–6170 (2011).
36. Barrero Tobon, A. M. & Hendrixson, D. R. Flagellar biosynthesis exerts temporal regulation of secretion of specific *Campylobacter jejuni* colonization and virulence determinants. *Mol. Microbiol.* **93,** 957–974 (2014).
37. Wei, B. L. *et al.* Positive regulation of motility and *flhDC* expression by the RNA-binding protein CsrA of *Escherichia coli*. *Mol. Microbiol.* **40,** 245–256 (2001).
38. Yakhnin, A. V. *et al.* CsrA activates *flhDC* expression by protecting *flhDC* mRNA from RNase E-mediated cleavage. *Mol. Microbiol* **87,** 851–866 (2013).
39. Konkel, M. E. *et al.* Secretion of virulence proteins from *Campylobacter jejuni* is dependent on a functional flagellar export apparatus. *J. Bacteriol.* **186,** 3296–3303 (2004).
40. Balaban, M. & Hendrixson, D. R. Polar flagellar biosynthesis and a regulator of flagellar number influence spatial parameters of cell division in *Campylobacter jejuni*. *PLoS Pathog.* **7,** e1002420 (2011).
41. Duss, O., Michel, E., Diarra Dit Konte, N., Schubert, M. & Allain, F. H. Molecular basis for the wide range of affinity found in Csr/Rsm protein-RNA recognition. *Nucleic Acids Res.* **42,** 5332–5346 (2014).
42. Mercante, J., Edwards, A. N., Dubey, A. K., Babitzke, P. & Romeo, T. Molecular geometry of CsrA (RsmA) binding to RNA and its implications for regulated expression. *J. Mol. Biol.* **392,** 511–528 (2009).
43. Sterzenbach, T. *et al.* A novel CsrA titration mechanism regulates fimbrial gene expression in *Salmonella typhimurium*. *EMBO J.* **32,** 2872–2883 (2013).
44. Jorgensen, M. G., Thomason, M. K., Havelund, J., Valentin-Hansen, P. & Storz, G. Dual function of the McaS small RNA in controlling biofilm formation. *Genes Dev.* **27,** 1132–1145 (2013).
45. Keiler, K. C. RNA localization in bacteria. *Curr. Opin. Microbiol.* **14,** 155–159 (2011).
46. Buskila, A. A., Kannaiah, S. & Amster-Choder, O. RNA localization in bacteria. *RNA Biol.* **11,** 1051–1060 (2014).
47. Singer, H. M., Erhardt, M. & Hughes, K. T. Comparative analysis of the secretion capability of early and late flagellar type III secretion substrates. *Mol. Microbiol.* **93,** 505–520 (2014).
48. Karlinsey, J. E., Lonner, J., Brown, K. L. & Hughes, K. T. Translation/secretion coupling by type III secretion systems. *Cell* **102,** 487–497 (2000).
49. Sorg, J. A., Miller, N. C. & Schneewind, O. Substrate recognition of type III secretion machines--testing the RNA signal hypothesis. *Cell Microbiol.* **7,** 1217–1225 (2005).
50. Buchan, J. R. mRNP granules: Assembly, function, and connections with disease. *RNA Biol.* **11,** 1019–1030 (2014).
51. Yakhnin, H. *et al.* CsrA of *Bacillus subtilis* regulates translation initiation of the gene encoding the flagellin protein (*hag*) by blocking ribosome binding. *Mol. Microbiol.* **64,** 1605–1620 (2007).
52. Sze, C. W. *et al.* Carbon storage regulator A (CsrA(Bb)) is a repressor of *Borrelia burgdorferi* flagellin protein FlaB. *Mol. Microbiol.* **82,** 851–864 (2011).
53. Ouyang, Z., Zhou, J. & Norgard, M. V. CsrA (BB0184) is not involved in activation of the RpoN-RpoS regulatory pathway in *Borrelia burgdorferi*. *Infect. Immun.* **82,** 1511–1522 (2014).
54. Anderson, P. E. & Gober, J. W. FlbT, the post-transcriptional regulator of flagellin synthesis in *Caulobacter crescentus*, interacts with the 5' untranslated region of flagellin mRNA. *Mol. Microbiol.* **38,** 41–52 (2000).
55. Mukherjee, S., Babitzke, P. & Kearns, D. B. FliW and FliS function independently to control cytoplasmic flagellin levels in *Bacillus subtilis*. *J. Bacteriol.* **195,** 297–306 (2013).
56. Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **28,** 3442–3444 (2000).
57. Zere, T. R. *et al.* Genomic targets and features of BarA-UvrY (-SirA) signal transduction systems. *PLoS One* **10,** e0145035 (2015).

58. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97,** 6640–6645 (2000).
59. Sittka, A., Sharma, C. M., Rolle, K. & Vogel, J. Deep sequencing of *Salmonella* RNA associated with heterologous Hfq proteins in vivo reveals small RNAs as a major target class and identifies RNA processing phenotypes. *RNA Biol.* **6,** 266–275 (2009).
60. Hoffmann, S. *et al.* Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.* **5,** e1000502 (2009).
61. Feng, J. *et al.* GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics* **28,** 2782–2788 (2012).
62. Yao, Z., Weinberg, Z. & Ruzzo, W. L. CMfinder--a covariance model based RNA motif finding algorithm. *Bioinformatics* **22,** 445–452 (2006).
63. Weinberg, Z. & Breaker, R. R. R2R--software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics* **12,** 3 (2011).
64. Gundogdu, O. *et al.* Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. *BMC Genomics* **8,** 162 (2007).
65. Whiteside, M. D., Winsor, G. L., Laird, M. R. & Brinkman, F. S. OrthologeDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis. *Nucleic Acids Res.* **41**(Database issue): D366–D376 (2013).
66. Papenfort, K. *et al.* SigmaE-dependent small RNAs of *Salmonella* respond to membrane stress by accelerating global *omp* mRNA decay. *Mol. Microbiol.* **62,** 1674–1688 (2006).
67. Sittka, A., Pfeiffer, V., Tedin, K. & Vogel, J. The RNA chaperone Hfq is essential for the virulence of *Salmonella typhimurium*. *Mol. Microbiol.* **63,** 193–217 (2007).
68. Sharma, C. M., Darfeuille, F., Plantinga, T. H. & Vogel, J. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes Dev.* **21,** 2804–2817 (2007).
69. Russell, J. H. & Keiler, K. C. Subcellular localization of a bacterial regulatory RNA. *Proc. Natl Acad. Sci. USA* **106,** 16405–16409 (2009).
70. Swoboda, M. *et al.* Enzymatic oxygen scavenging for photostability without pH drop in single-molecule experiments. *ACS Nano* **6,** 6364–6369 (2012).
71. van de Linde, S. *et al.* Direct stochastic optical reconstruction microscopy with standard fluorescent probes. *Nat. Protoc.* **6,** 991–1009 (2011).
72. Wolter, S. *et al.* rapidSTORM: accurate, fast open-source software for localization microscopy. *Nat. Methods* **9,** 1040–1041 (2012).
73. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30,** 207–210 (2002).
74. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406–3415 (2003).
75. Hendrixson, D. R. *Regulation of Flagellar Gene Expression and Assembly. Campylobacter* 3rd edn (ASM, 2008).

Acknowledgements

We thank Sandy R. Pernitzsch, Erik Holmqvist, Jörg Vogel, Gisela Storz and Stan Gorski for critical comments on our manuscript and/or fruitful discussions. We thank Konrad U. Förstner for help with RNA-seq analysis, Lars Barquist for help with CMfinder motif analysis, Hilde Merkert for help with electron microscopy and Belinda Aul for technical assistance. G.D. is supported by the Graduate School for Life Sciences (GSLS), Würzburg. Research in the Sharma's laboratory is supported by the Young Investigator program at the Research Center for Infectious Diseases in Würzburg, Germany, the Bavarian Research Network for Molecular Biosystems (BioSysNet), the Bavarian Academy of Science and Humanities and the Daimler and Benz foundation.

Author contributions

G.D., S.L.S. and C.M.S. designed the study. G.D., S.L.S., T.B., S.W., M.S. and C.M.S. performed the experiments and analysed the data. R.R. provided reagents and deep sequencing and T.B. performed the RNA-seq data analysis. G.D., S.L.S. and C.M.S. wrote the manuscript with input from all the authors.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

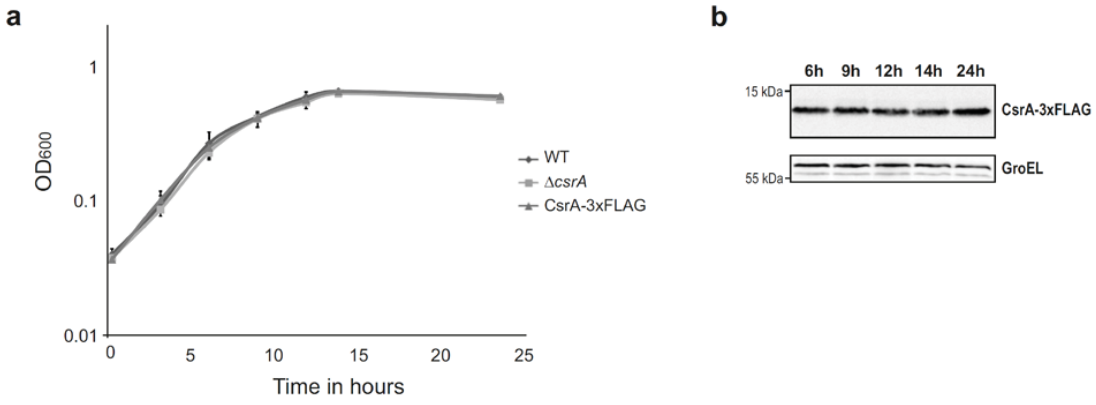
Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Dugar, G. *et al.* The CsrA-FliW network controls polar localization of the dual-function flagellin mRNA in *Campylobacter jejuni*. *Nat. Commun.* 7:11667 doi: 10.1038/ncomms11667 (2016).

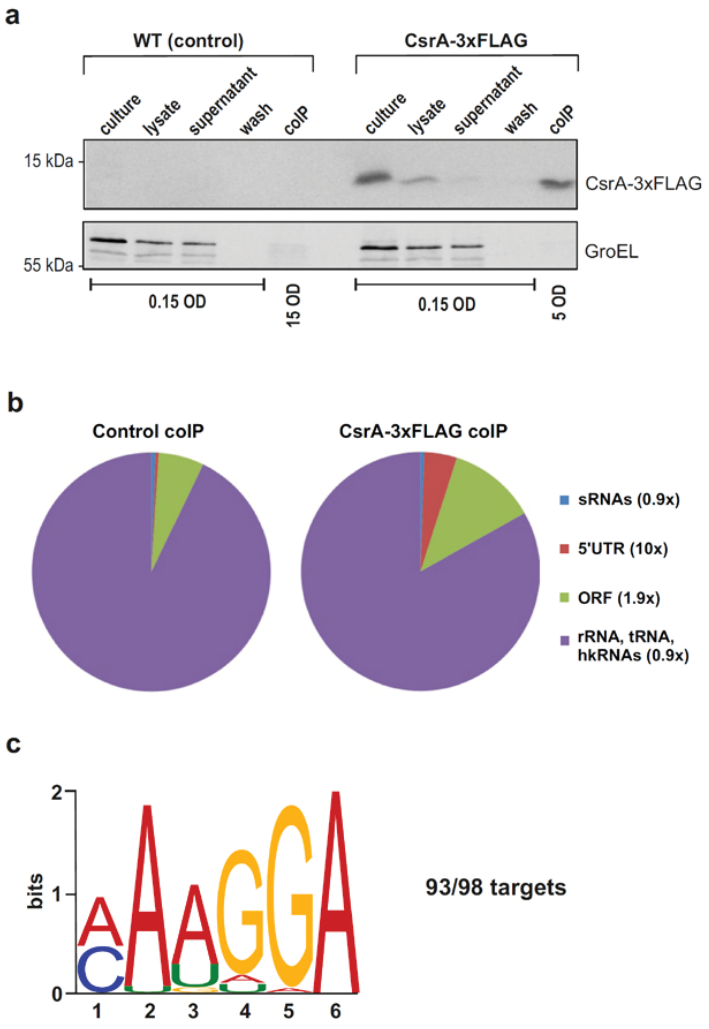
Supplementary Figures

Supplementary Figure 1



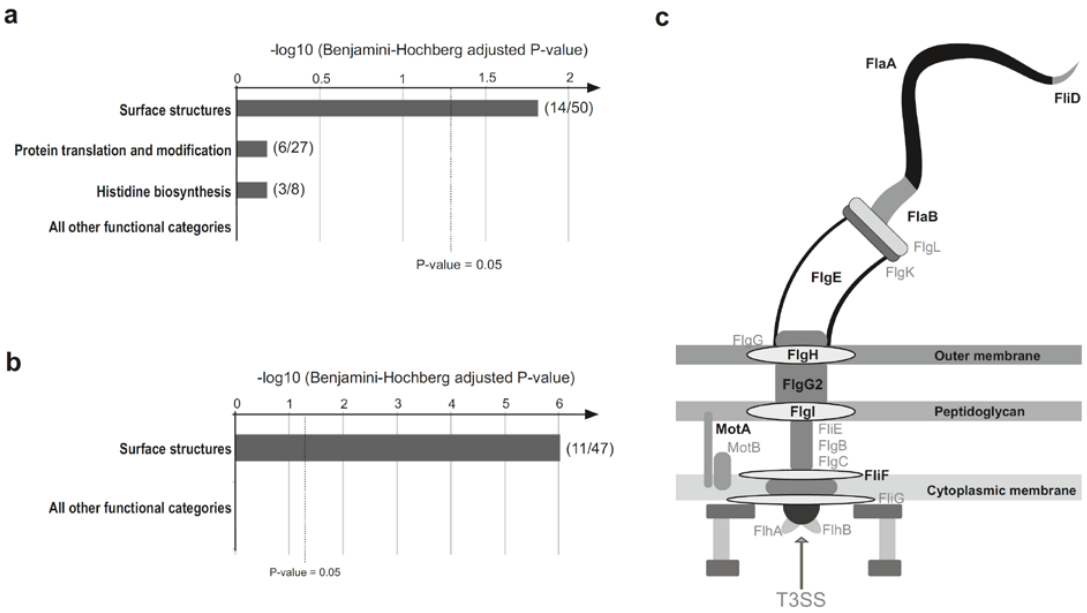
Supplementary Figure 1. Growth curves and expression of *C. jejuni* CsrA in strain NCTC11168. (a) Semi-log growth curves over 24 h for *C. jejuni* NCTC11168 wild-type (WT), Δ*csrA* and CsrA-3xFLAG tagged strains grown in Brucella broth in duplicate. Error bars are mean ± s.e.m. **(b)** Western blot analysis of CsrA-3xFLAG expression during growth in liquid culture in *C. jejuni* strain NCTC11168. Total protein samples corresponding to 0.05 OD_{600} were loaded for different time points. GroEL was probed as sample processing control on a separate blot.

Supplementary Figure 2



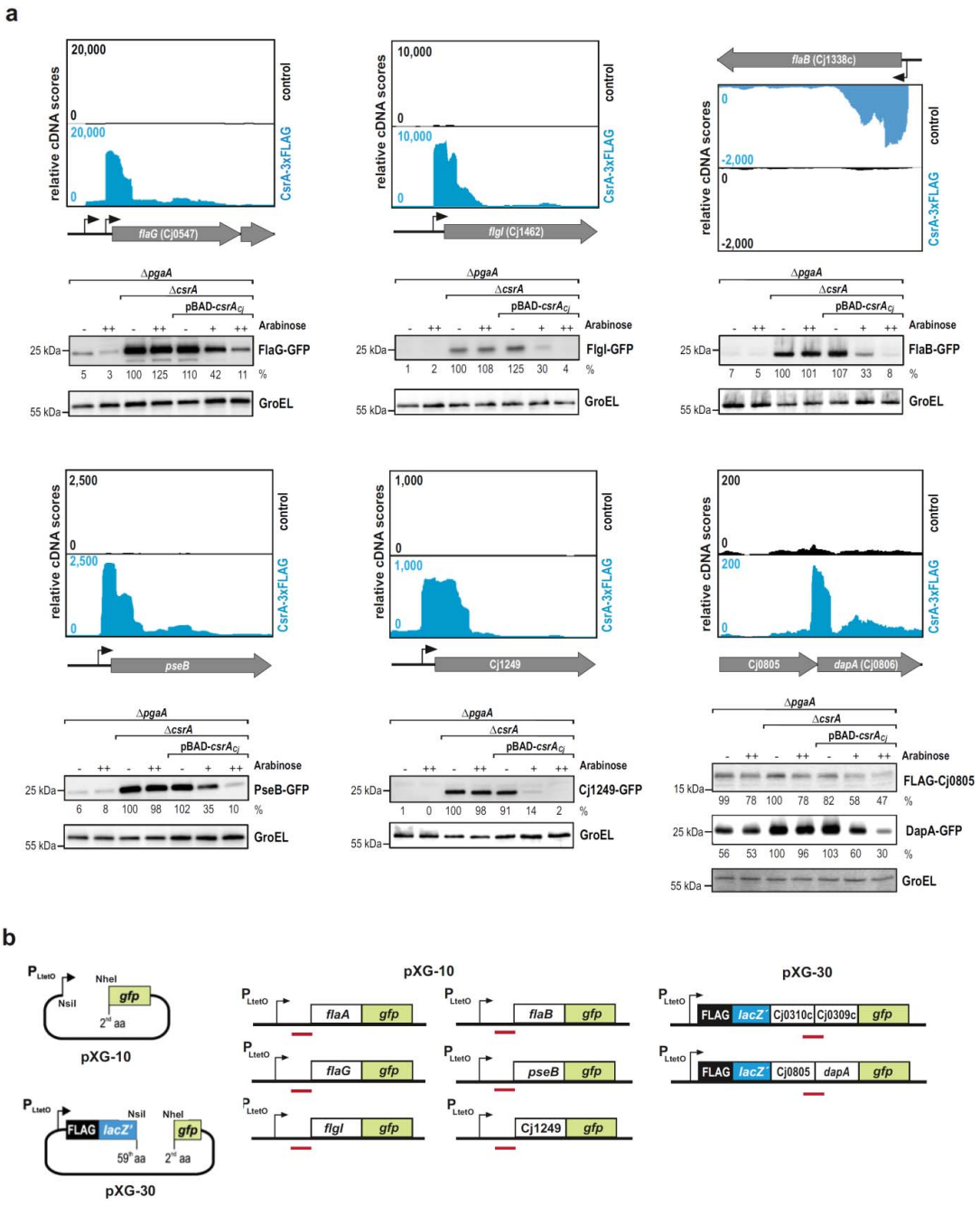
Supplementary Figure 2. RIP-seq of CsrA-3xFLAG in strain 81-176. (a) Western blot of protein samples from CsrA-3xFLAG and control coIPs from strain 81-176. The protein amount loaded in each lane corresponds to the OD_{600} of cells as indicated below. GroEL was probed as sample processing control on a separate blot. **(b)** Pie charts showing the relative proportions of mapped cDNAs of different RNA classes in the control and CsrA-3xFLAG coIP libraries from strain 81-176. Values in brackets for each RNA class denote its relative enrichment in the CsrA-3xFLAG vs. control coIP. **(c)** Consensus motif for CsrA determined by MEME using peak sequences enriched more than 5-fold in the CsrA-3xFLAG coIP from *C. jejuni* strain 81-176.

Supplementary Figure 3



Supplementary Figure 3. Statistical analysis for overrepresentation of functional categories among potential CsrA target genes. (a), (b) Statistical analysis was performed on all genes with more than 5-fold enrichment in their 5'UTR and/or coding sequence in the CsrA-coIP in comparison to non-enriched genes. P-values, adjusted using the Benjamini-Hochberg method, were calculated for overrepresentation of enriched genes in each functional category in strain NCTC11168 **(a)** and strain 81-176 **(b)**. Values in brackets denote the number of enriched genes in the particular functional category (only functional categories with non-zero $\log_{10}$ P-values are shown for clarity). Functional categories are based on reannotation of the NCTC11168 genome[1]. **(c)** Schematic representation of structural components of the *Campylobacter* flagellum. Proteins encoded by mRNAs which showed >5-fold enrichment in the CsrA coIP are marked in black and bold. T3SS: Type III secretion system.

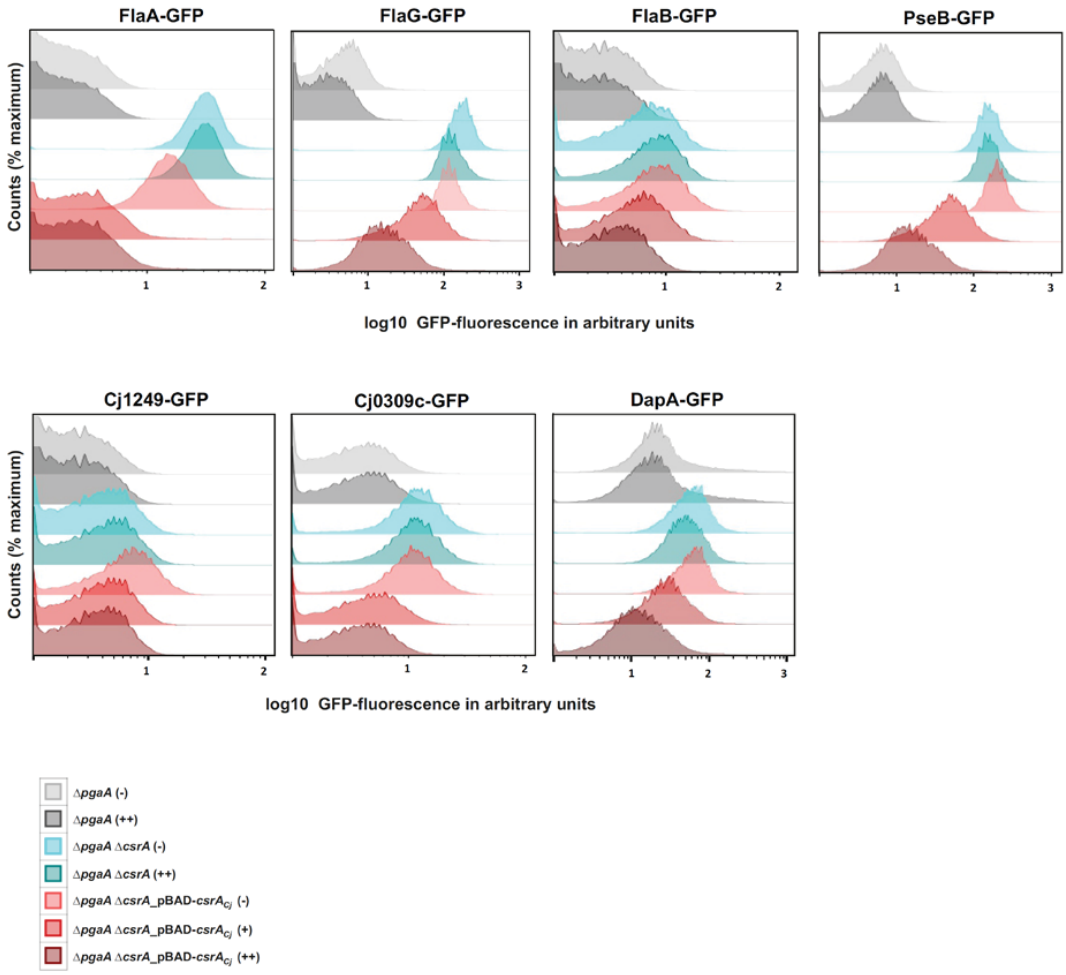
Supplementary Figure 4



Supplementary Figure 4. Validation of potential CsrA targets using a GFP reporter system in *E. coli*. **(a)** *(Top panels)* Examples of enrichment patterns indicating potential CsrA binding sites in 5'UTRs (*flgG*, *flgI*, *flaB*, *pseB*, and Cj1249 mRNAs) and between genes in polycistronic transcripts (Cj0805-*dapA* operon; encoding a zinc protease and dihydrodipicolinate synthase) that were tested in the *E. coli* system. Mapped cDNA reads are shown for the control (black) and

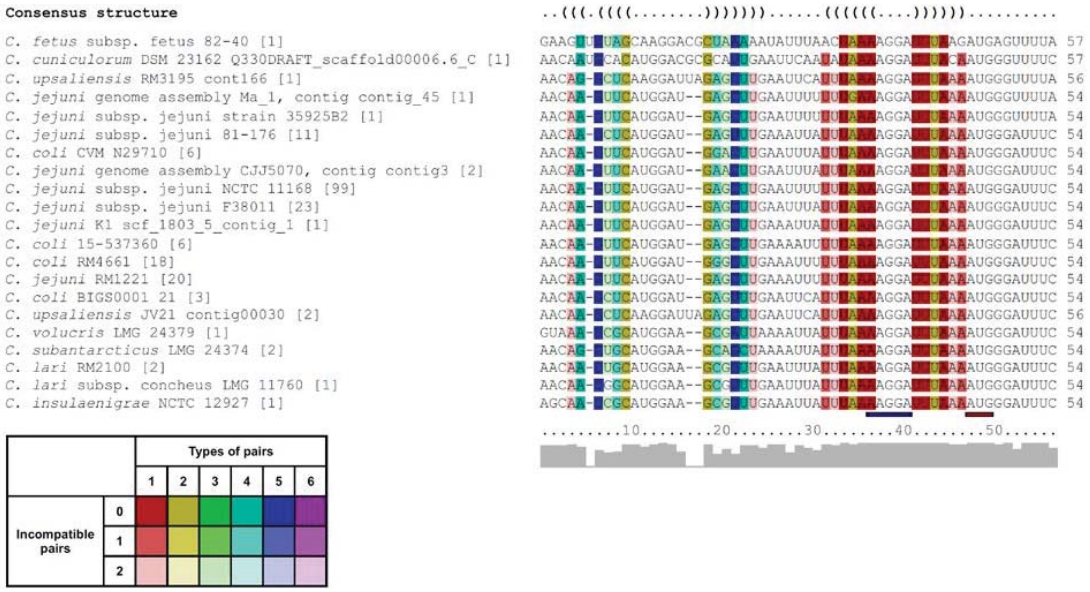
CsrA-3xFLAG coIP (blue) libraries in strain NCTC11168. ORFs are indicated by grey arrows and TSS - based on dRNA-seq[2] - by black arrows, respectively. *(Lower panels)* Western blot analysis using anti-FLAG and anti-GFP antibodies of reporter translational fusions to potential *C. jejuni* CsrA target genes in *E. coli* Δ*pgaA*, Δ*pgaA/*Δ*csrA,* and Δ*pgaA/*Δ*csrA* complemented with plasmid pGD72-3 carrying arabinose-inducible *csrA*-Strep from *C. jejuni.* Strains were grown to late log phase in LB medium or LB media supplemented with 0.001% (+) or 0.003% (++) L-arabinose. **(b)** Overview of GFP reporter plasmids pXG-10 and pXG-30[3] used to validate 5'UTR and intergenic/ORF targets, respectively. For putative targets in 5'UTRs, the entire leader, as well as the first few codons, were cloned as a translational fusion to *gfp* in low-copy vector pXG-10 (pSC101* origin). Fusions to *gfp* were made downstream of the +1 (TSS) site of the P_L promoter and were transcribed from a constitutive λ PLtetO-1 promoter (PL derivative). To examine CsrA binding between genes in polycistrons, or in downstream regions of ORFs, the operon plasmid pXG-30 was used. The C-terminus of the upstream ORF was fused in-frame after a short artificial reading frame composed of a FLAG epitope and truncated *lacZ* gene, whereas the N-terminus of the downstream gene was fused in frame to *gfp*, thus mimicking operon mRNA expression. Putative CsrA binding regions are marked by a red line.

Supplementary Figure 5



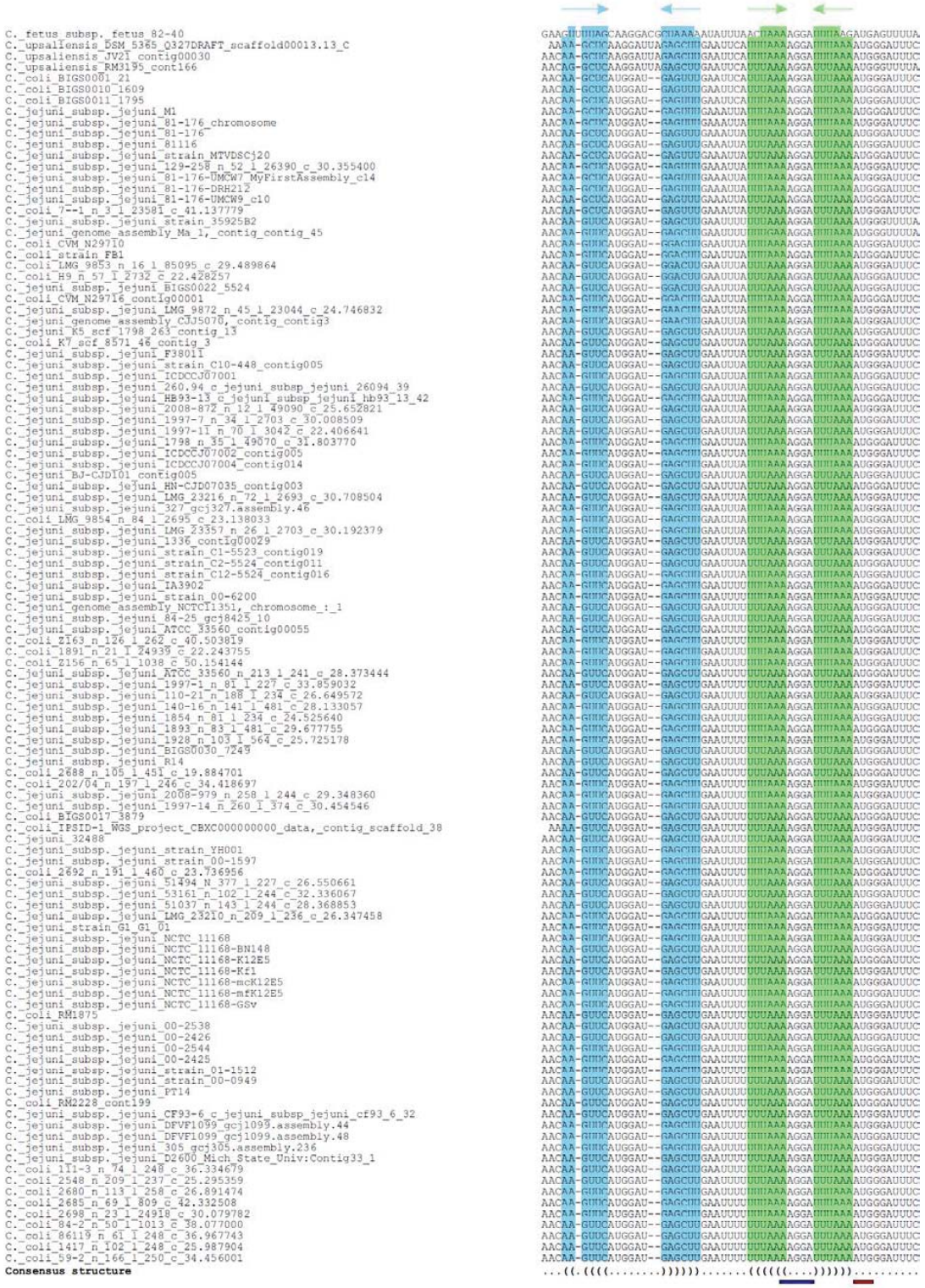
Supplementary Figure 5. FACS analysis of GFP reporter fusions. pXG-10- or pXG-30-based GFP reporter plasmids from Supplementary Fig. 3b were introduced into *E. coli* strains Δ*pgaA,* Δ*pgaA/*Δ*csrA,* or Δ*pgaA/*Δ*csrA* complemented with arabinose-inducible *C. jejuni csrA*-Strep. All strains were grown to late log phase in LB or LB supplemented with 0.001% (+) or 0.003% (++) L-arabinose, and GFP levels were measured by flow cytometry. Data acquired in each experiment is plotted in fluorescence histograms generated from all events measured (50,000 events). Cellular fluorescence is given in arbitrary units (GFP intensity). Regulation by CsrA is visible as a shift of the peak of fluorescence curves to the right (higher GFP intensity) in the Δ*pgaA/*Δ*csrA* background and a shift to the left upon complementation with CsrA-Strep. Please note that levels of FlgI-GFP were not detectable in FACS due to low expression or fluorescence of this fusion.

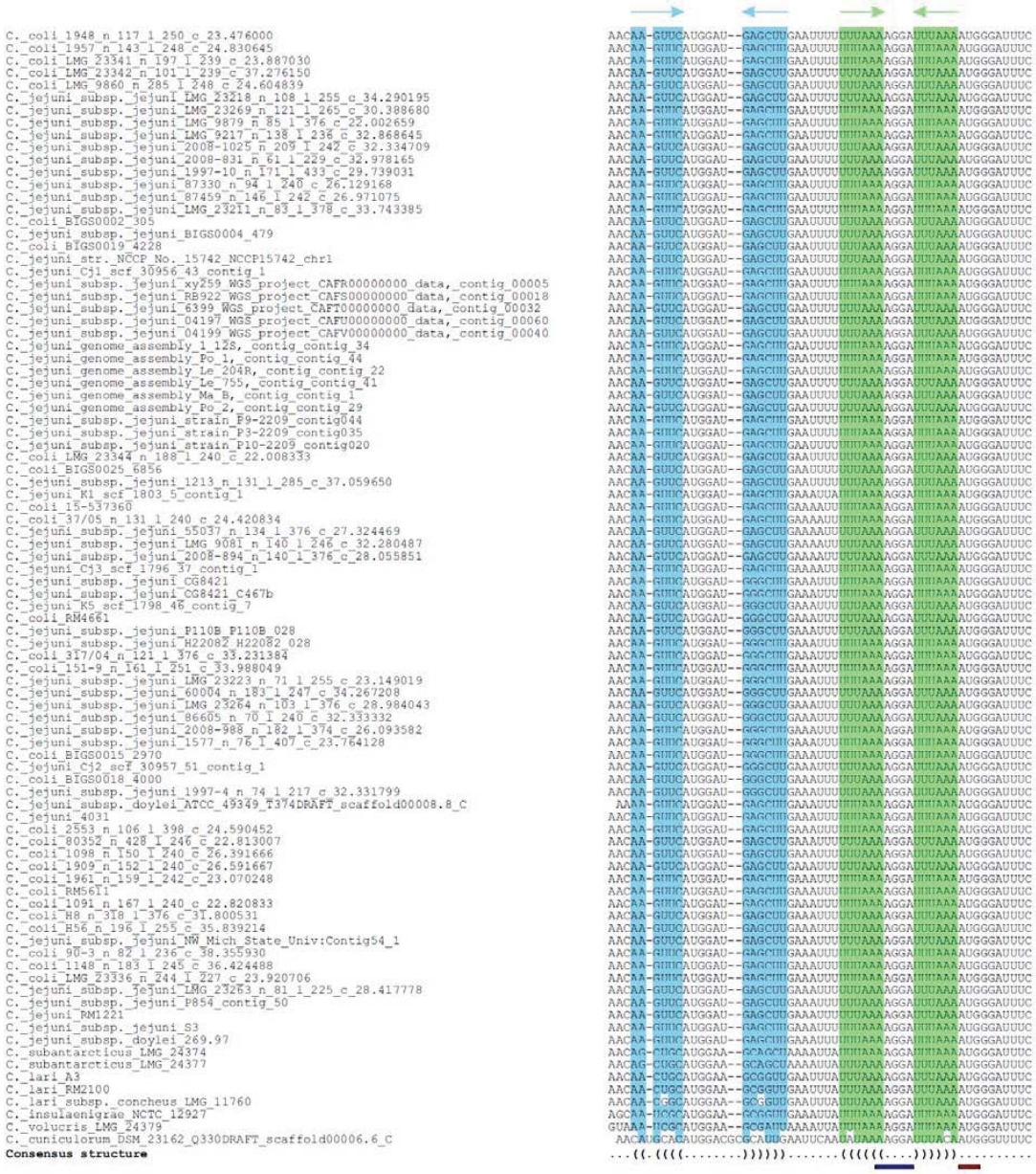
Supplementary Figure 6



Supplementary Figure 6. Structure annotated sequence alignment of *flaA* 5'UTR. Structure-annotated sequence alignment based on representative unique sequences of the *flaA* 5'UTR and the first 10 nt of the coding region from different *Campylobacter* species. The numbers in square brackets indicate how often an identical sequence occurred in the original complete alignment (see Supplementary Fig. 7). The consensus structure is shown in bracket-dot notation at the top and sequence conservation as a bar chart at the bottom. The color legend was adapted from the RNAalifold online help (http://rna.tbi.univie.ac.at/help.html) and reflects how many different kinds of nucleotide combinations support a certain base pair. Here, paler colors indicate that a base pair cannot be formed by all sequences in the alignment. The RBS and start codon of *flaA* mRNA are indicated by blue and red bars below the alignment, respectively.

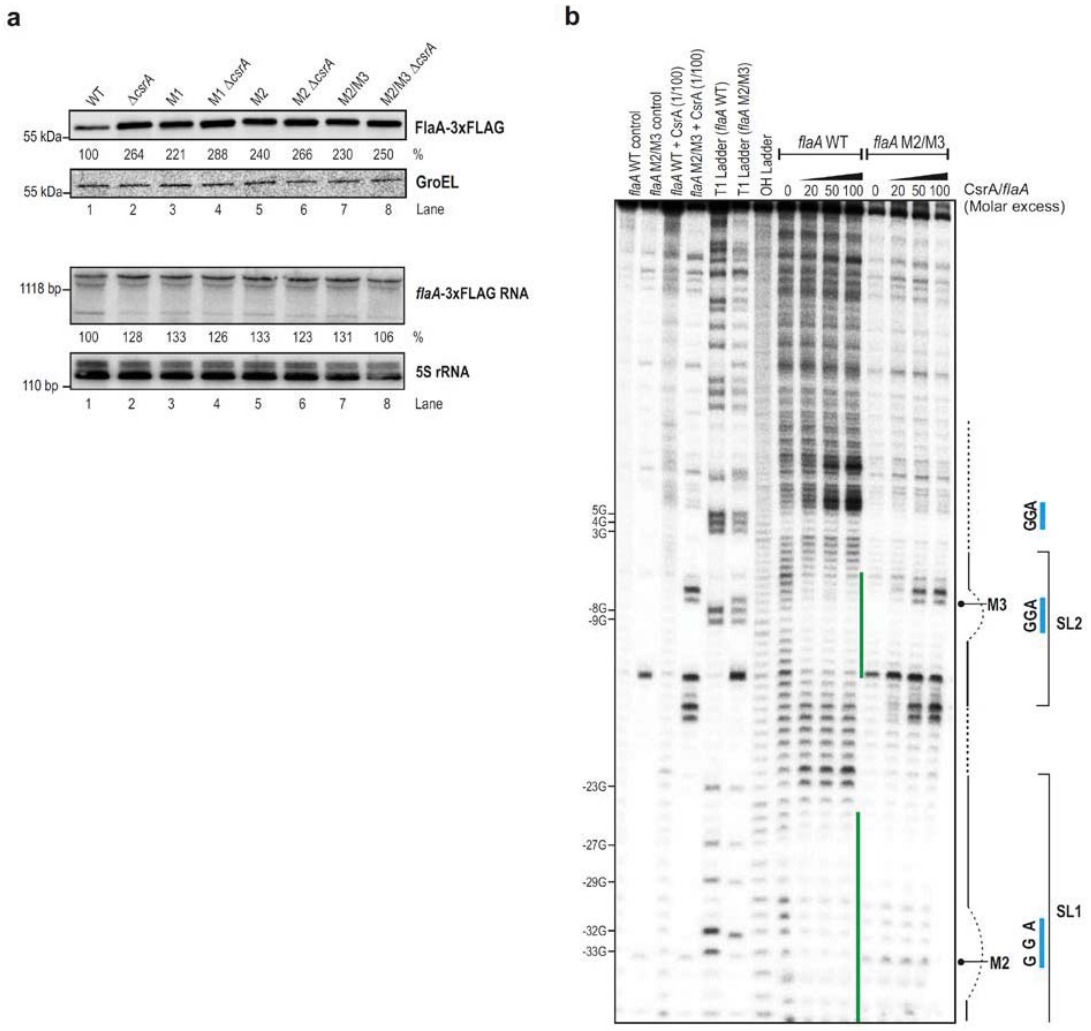
Supplementary Figure 7





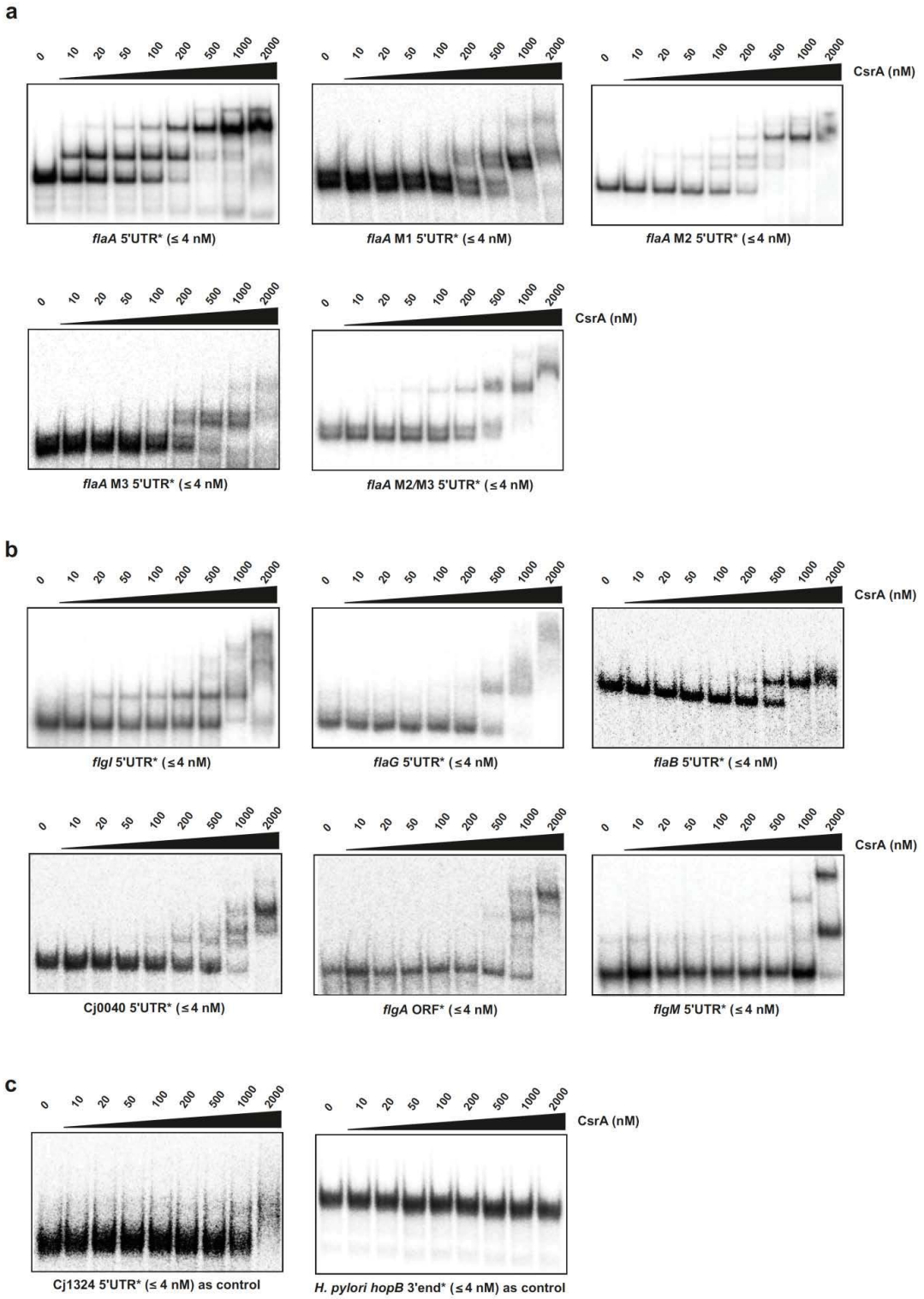
Supplementary Figure 7. Sequence alignment of *flaA* leaders from *Campylobacter* species. Sequence alignment of redundant sequences for the *flaA* 5'UTR and the first 10 nt of the coding region from different *Campylobacter* species. The consensus structure calculated based on the collapsed version of the alignment (see Supplementary Fig. 6) is shown at the bottom of each page. The blue and green arrows at the top mark stem loops in the consensus structure while pointing to the unpaired loop region, respectively. The nucleotides of each sequence below the stems are marked in the respective color if the base pair can be formed and have a white background otherwise. The RBS and start codon of *flaA* mRNA are indicated by blue and red bars below the alignment, respectively.

Supplementary Figure 8



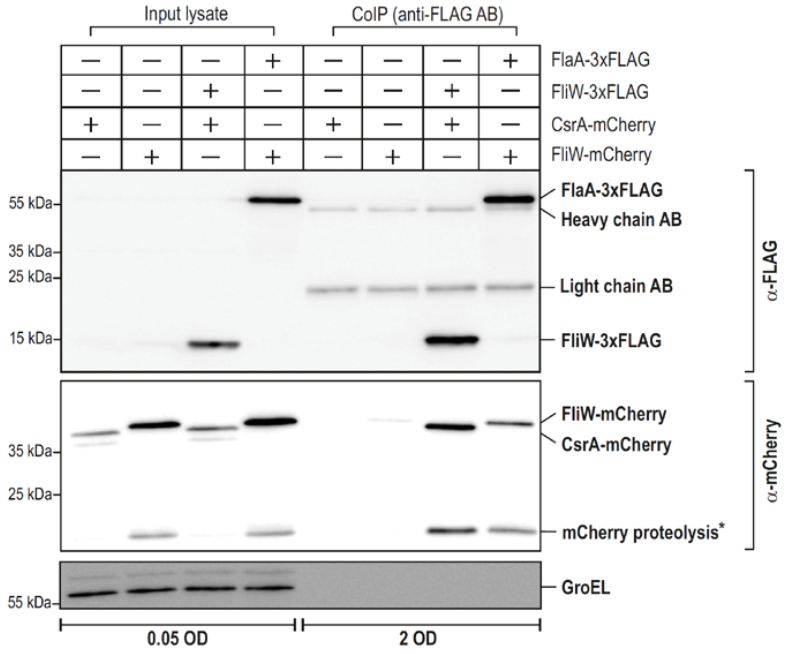
Supplementary Figure 8. Analysis of regulation and binding of CsrA to *flaA* mutant leaders. (a) Representative Western and Northern blot of FlaA-3xFLAG and its mRNA in various *flaA* 5'UTR mutant strains (from main Fig. 2b) from liquid cultures in log phase. Anti-FLAG antibody was used to detect FlaA-3xFLAG. See main Fig. 2b for M1, M2 and M3 mutations. GroEL was probed as sample processing control on a separate blot. **(b)** Footprinting assays of ~0.2 pmol P^{32} labeled *flaA* (WT or M2/M3 mutant) leaders in the absence or presence of increasing *C. jejuni* CsrA concentrations (CsrA/*flaA* molar ratio of 0, 20, 50 and 100) using lead(II) acetate. Untreated *flaA* leader alone, or *flaA* leader incubated with 100-fold excess CsrA, served as controls. Partially RNase T1- or alkali (OH)-digested *flaA* WT and M2/M3 leaders are included as ladders. Blue lines: three GGA motifs in the WT *flaA* leader; green lines: regions protected from cleavage upon increasing CsrA concentration. Bands representing the M2 or M3 mutations are marked next to the gel.

Supplementary Figure 9



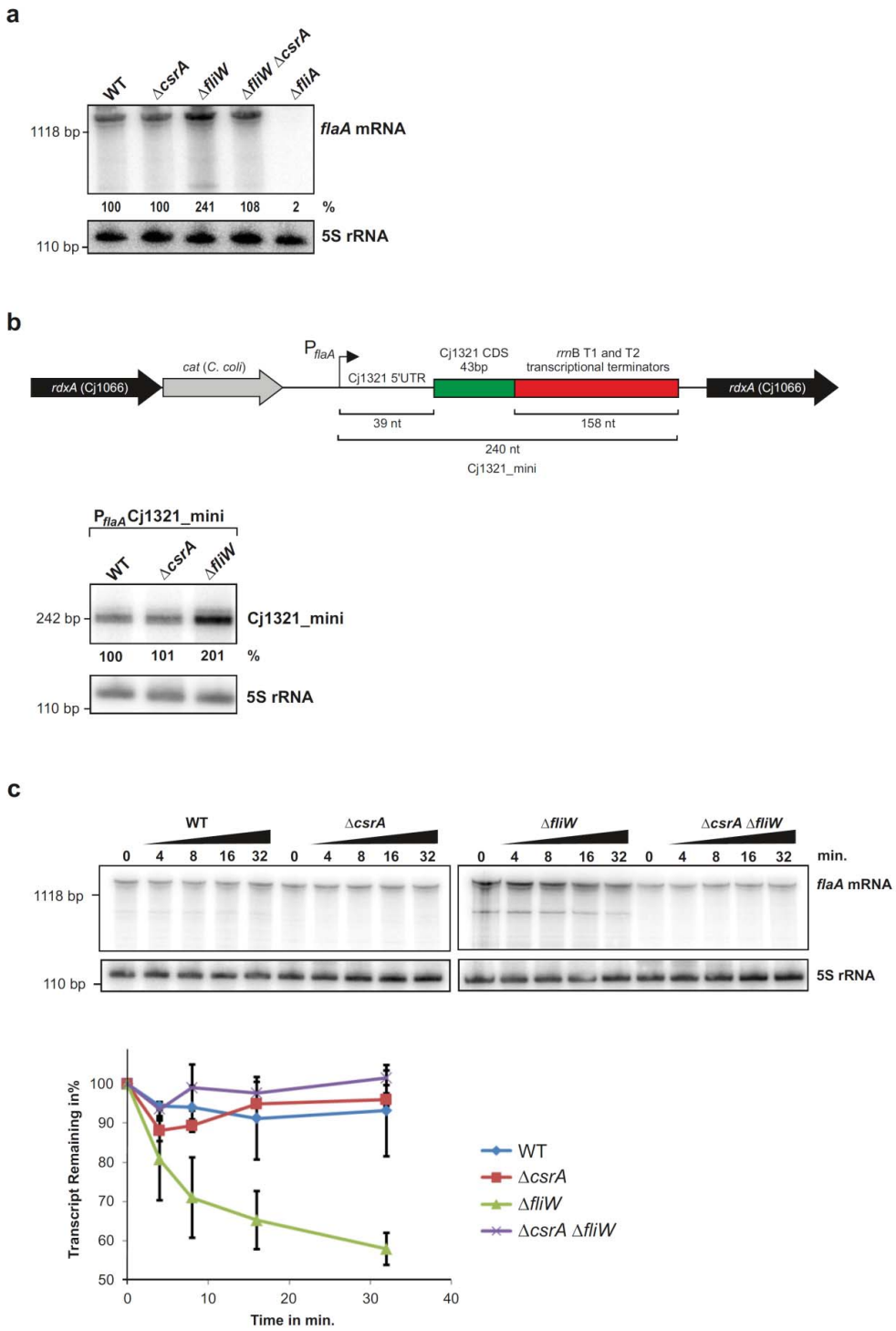
Supplementary Figure 9. *In vitro* gel-shift assays of 5'-labeled T7-transcripts and purified *C. jejuni* CsrA. (a) Gel-shift assays with 5'-labeled WT *flaA* leader and its mutant variants (M1, M2, M3, and M2/M3) with increasing concentrations of CsrA. M1: GGA→AAA in stem-loop 1 (SL1, see Fig. 2a); M2: GGA>UGA in SL1; M3: GGA>GGG in stem-loop 2 (SL2); M2/M3: combination of M2 and M3. **(b)** Gel-shift assays of T7-transcribed, 5'-labeled RNAs of flagellar targets with increasing concentrations of CsrA. **(c)** Gel-shift assays with negative controls using the leader of Cj1324 from *C. jejuni* (one GGA, not enriched in either coIP) and a fragment of the unrelated *hopB* mRNA (no GGAs, from *Helicobacter pylori* G27).

Supplementary Figure 10



Supplementary Figure 10. Protein-protein coIP confirms direct interactions of FliW with CsrA and FlaA. CsrA-mCherry and FliW-mCherry were specifically co-purified in a coIP of FliW-3xFLAG and FlaA-3xFLAG, respectively, using an anti-FLAG antibody. In a negative control reaction with non-tagged FliW and FlaA, CsrA-mCherry and FliW-mCherry were not pulled down. Western Blots were performed for the input lysates and coIP protein samples (FLAG) using anti-FLAG and anti-mCherry antibodies. GroEL served as loading control for the input lysate samples on the mCherry blot and was not detected in the coIP fraction. (*Please note that the lower band represents partially hydrolysed mCherry resulting from sample preparation[4].)

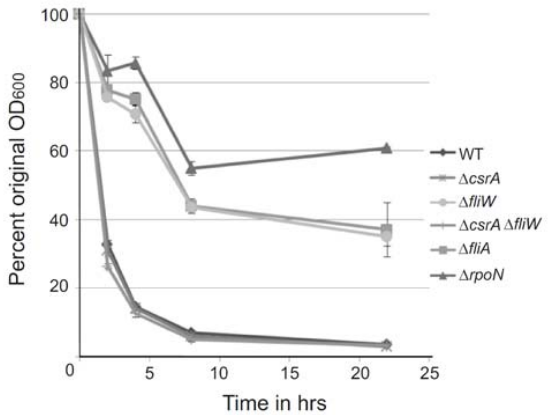
Supplementary Figure 11



Supplementary Figure 11. Examination of effects of FliW on *flaA* transcription and translation. (a) Northern blot analysis of *flaA* mRNA using oligonucleotide probe CSO-0486, as

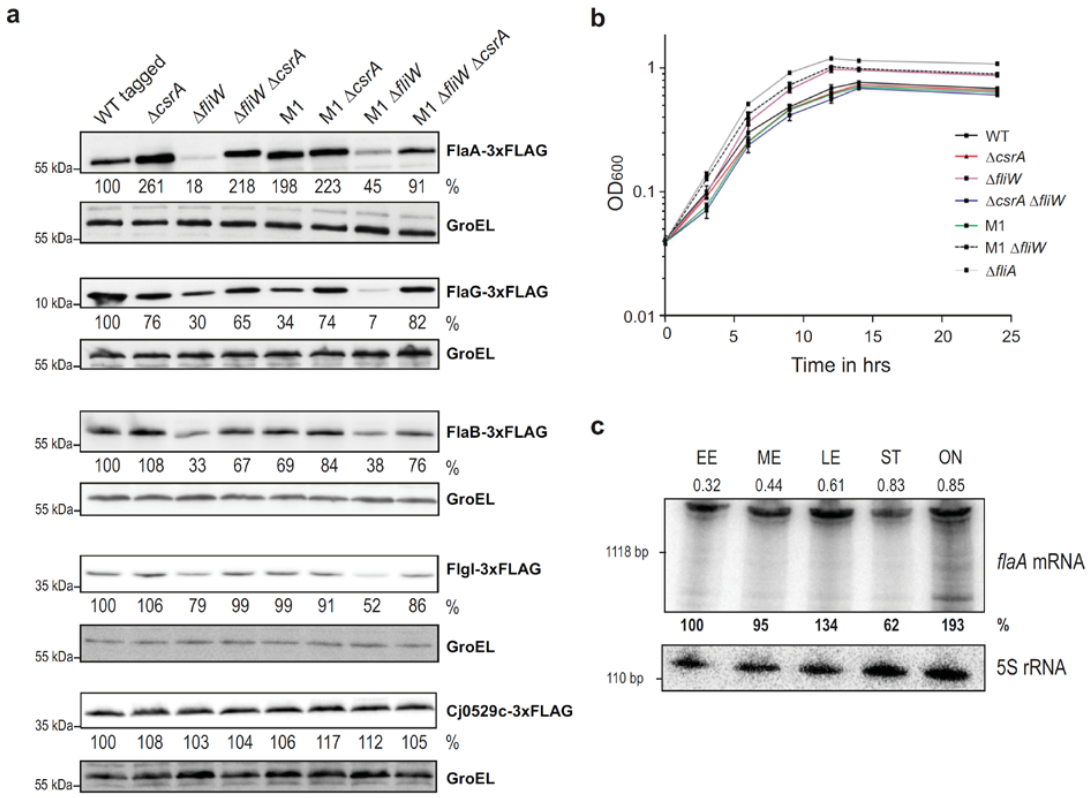
well as 5S rRNA (loading control, CSO-0192) in *C. jejuni* NCTC11168 wild-type and the indicated mutant strains grown to mid-log phase. Relative expression levels are quantified below the blot. **(b)** (*top*) Schematic representation of a Cj1321_mini gene construct that was expressed from the P_{flaA} promoter and was integrated into the unrelated *rdxA* locus of *C. jejuni*. The Cj1321 gene is not a CsrA target according to the coIP results and thus should be independent of CsrA. The Cj1321 5'UTR and the first 42 bp of the coding sequence were fused to a stable *rrnB* terminator to express as stable mini transcript under control of the P_{flaA} promoter. Northern blotting was used to monitor expression of the stable Cj1321_mini transcript, which was used as a transcriptional reporter for the *flaA* promoter. The ~240-nt long Cj1321_mini transcript was detected by Northern blot analysis of total RNA from *C. jejuni* cells expressing Cj1321_mini either in the wildtype, Δ*csrA*, or Δ*fliW* background using oligonucleotide probe CSO-2746. Probing for 5S rRNA (CSO-0192) served as a loading control. **(c)** *flaA* mRNA rifampicin stability assay. Northern blot probed for *flaA* mRNA in *C. jejuni* wildtype (WT) and the mutant strains over a time course after rifampicin addition (0-32 min) using oligonucleotide probe CSO-2835. Averaged quantification of *flaA* mRNA transcript levels over time from two independent rifampicin stability assays is shown. Error bars indicate mean ± s.e.m.

Supplementary Figure 12



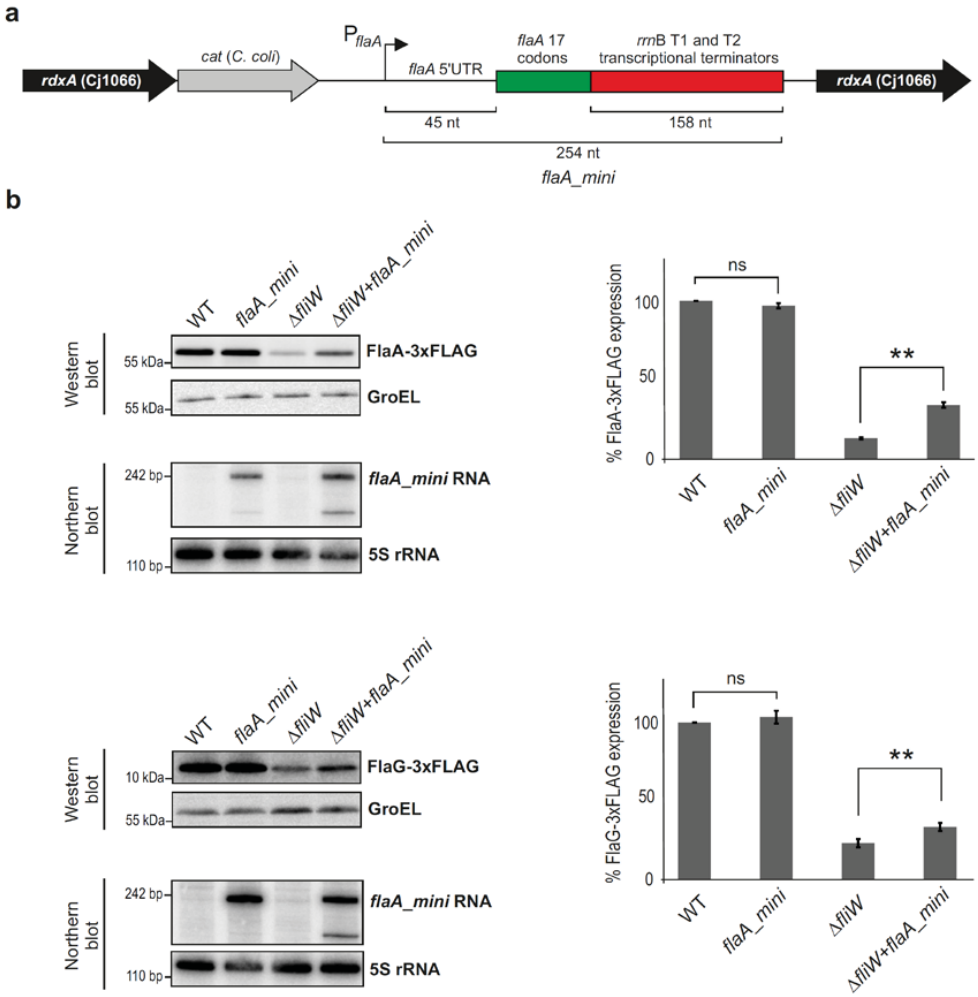
Supplementary Figure 12. Influence of *csrA* and *fliW* deletion on autoagglutination. Autoagglutination assay of *C. jejuni* WT and mutant strains (OD_{600} of supernatants of 1.0 OD bacterial suspensions grown in Brucella broth at the indicated time points) in PBS. Error bars indicate the s.e.m.

Supplementary Figure 13



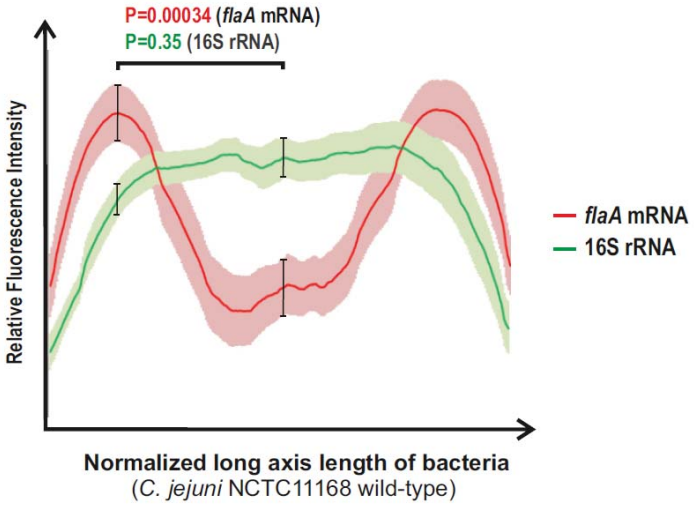
Supplementary Figure 13. Representative Western blots of 3xFLAG-tagged CsrA targets in various mutant strains and Northern blot of *flaA* mRNA over growth in the WT strain. **(a)** FlaA-3xFLAG, FlaG-3xFLAG, FlaB-3xFLAG, and FlgI-3xFLAG levels, as well as Cj0529c-3xFLAG levels as a negative control, were examined by Western blot in *C. jejuni* NCTC11168 wildtype (WT), Δ*csrA*, Δ*fliW*, Δ*csrA*/Δ*fliW,* M1, M1/Δ*csrA*, M1/Δ*fliW,* and M1/Δ*fliW*/Δ*csrA* strains. Cells were grown to mid-log phase in liquid culture, and protein samples (amounts corresponding to an OD_{600} of cells of 0.02 for FlaA-3xFLAG, 0.075 for FlaG-3xFLAG, or 0.05 for FlaB-3xFLAG, FlgI-3xFLAG and Cj0529c-3xFLAG) were analyzed by Western blotting with an anti-FLAG antibody. GroEL levels served as loading control for FlgI-3xFLAG and Cj0529c-3XFLAG and was probed as sample processing control on separate blots for FlaA-, FlaG- and FlaB-3xFLAG. **(b)** Semi-log growth curves in Brucella broth over 24 h for the untagged strains from **(a)** based on 2 biological replicates. Error bars indicate mean ± s.e.m. **(c)** Northern blot analysis of *flaA* mRNA using RNA extracted from *C. jejuni* samples collected at different growth phases (EE-Early Exponential, ME-Mid Exponential, LE-Late Exponential, ST-Stationary and ON-Overnight). The OD_{600} of the culture is also indicated below each phase. 5S rRNA was probed as a loading control.

Supplementary Figure 14



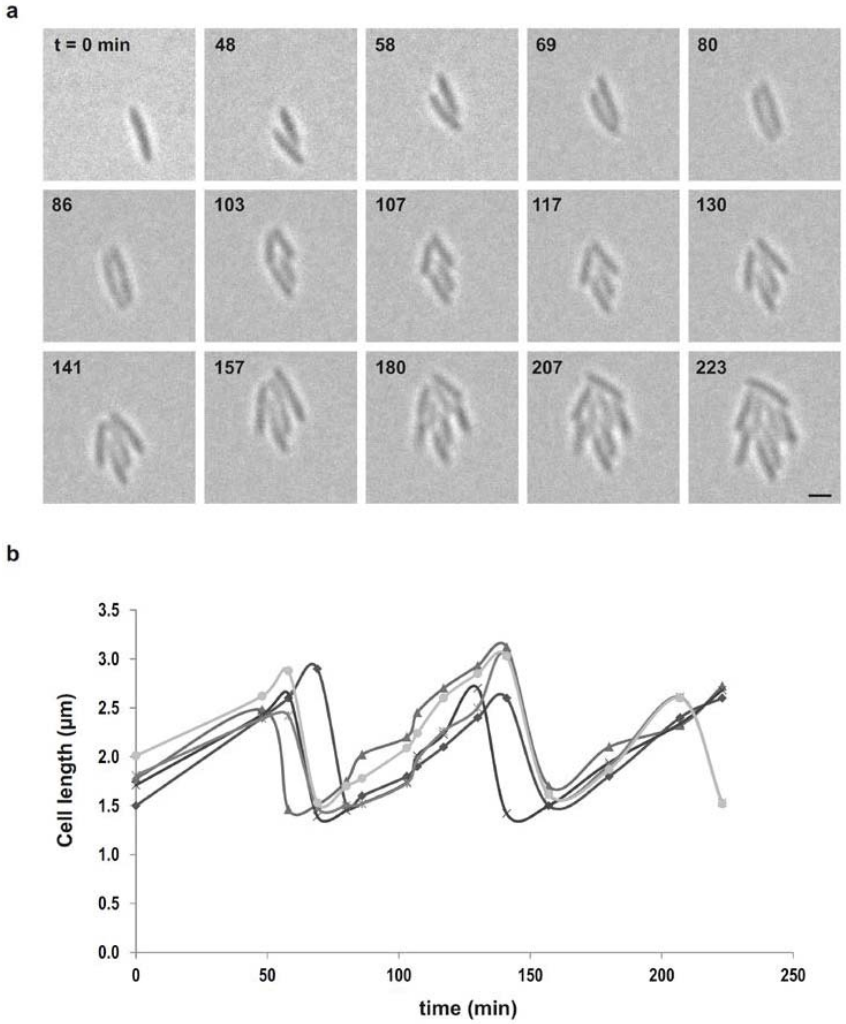
Supplementary Figure 14. An ectopically-expressed *flaA* mini-gene can partially complement CsrA-mediated effects on FlaA and FlaG translation upon *fliW* deletion. (a) Schematic representation of a *flaA_mini* gene construct expressed from the *rdxA* locus of *C. jejuni*. **(b)** *(Left)* Representative Western blots of FlaA-3xFLAG and FlaG-3xFLAG used for the quantifications on the right and Northern blots of *flaA_mini* RNA expression from liquid cultures in log phase. Anti-FLAG antibody was used to detect the tagged proteins. GroEL was probed as sample processing control on separate blots. *(Right)* Quantification of FlaA-3xFLAG (top panel) and FlaG-3xFLAG (bottom panel) determined by Western blot in the indicated strains ($n \geq 4$). Error bars indicate mean ± s.e.m (**P<0.01). Deletion of *fliW* leads to strong CsrA-mediated translational repression of *flaA*-3xFLAG and *flaG*-3xFLAG due to release of CsrA repression in the absence of the FliW protein antagonist. Expression of the stable *flaA_mini* transcript partially relieves CsrA-mediated translational repression of *flaA*-3xFLAG and *flaG*-3xFLAG upon *fliW* deletion, indicating that it can sequester CsrA and act as an antagonist of CsrA activity.

Supplementary Figure 15



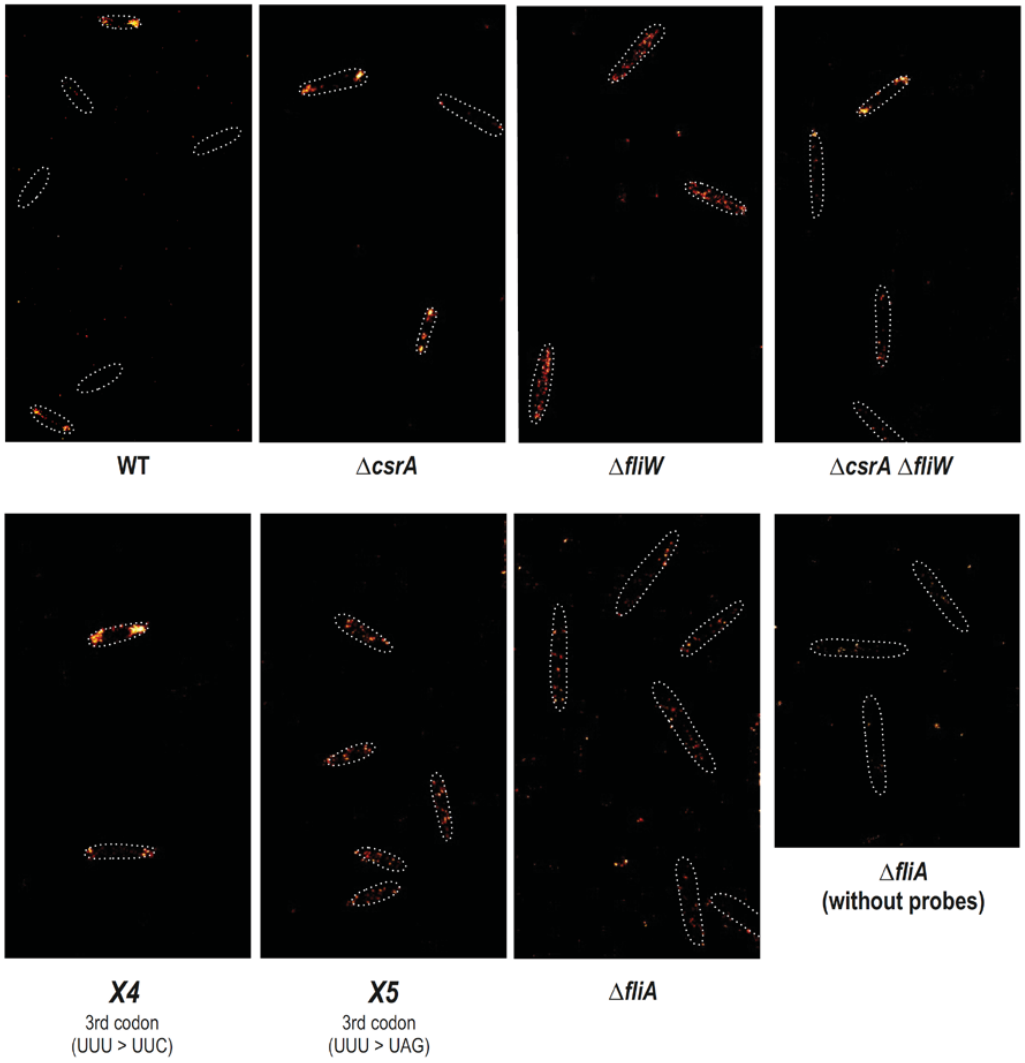
Supplementary Figure 15. Averaged fluorescence intensity curves for *flaA* mRNA and 16S rRNA FISH signals. Averaged fluorescence intensities (based on 10 cells) from RNA-FISH analysis of 16S rRNA (green) and *flaA* mRNA (red) plotted along the long cell axis for *C. jejuni* NCTC11168 WT. The shaded regions along the curves mark the boundary of errors bars (±s.e.m.) of 320 points along the long axis. The points taken for statistical analysis are highlighted by black error bars (Student's *t*-test). These points correspond to the pole and the mid-cell of the bacterium.

Supplementary Figure 16



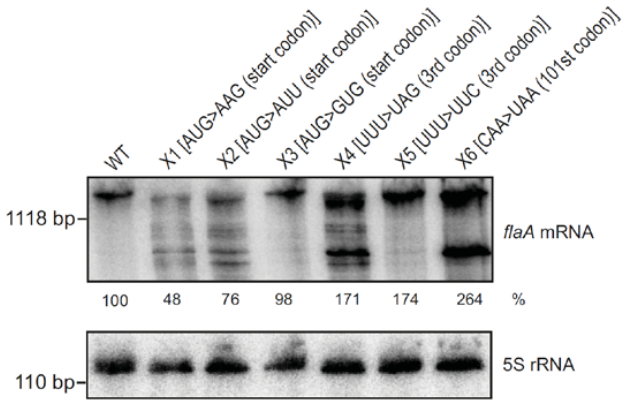
Supplementary Figure 16. Live-cell imaging of growth of non-motile *C. jejuni* over 2-3 generations. **(a)** A dividing non-motile (Δ*fliA*) *C. jejuni* cell (collected from Brucella broth culture in log phase) was imaged under a fluorescence microscope in bright field mode. Fifteen images were taken over a period of 223 min and 3 cell divisions. The black bar (lower right) represents 1 µm in length. **(b)** Cell lengths were measured over 2-3 divisions for five representative cells using ImageJ and were plotted along the time frame (each curve represents one cell). Please note that the cell lengths were longer (up to ~3 µm) compared to those of the fixed WT cells used in the FISH analysis (Figure 6, up to ~2 µm). This length difference could be due to slightly different morphology of the non-motile strain or different growth conditions (aerobic) used during microscopy.

Supplementary Figure 17



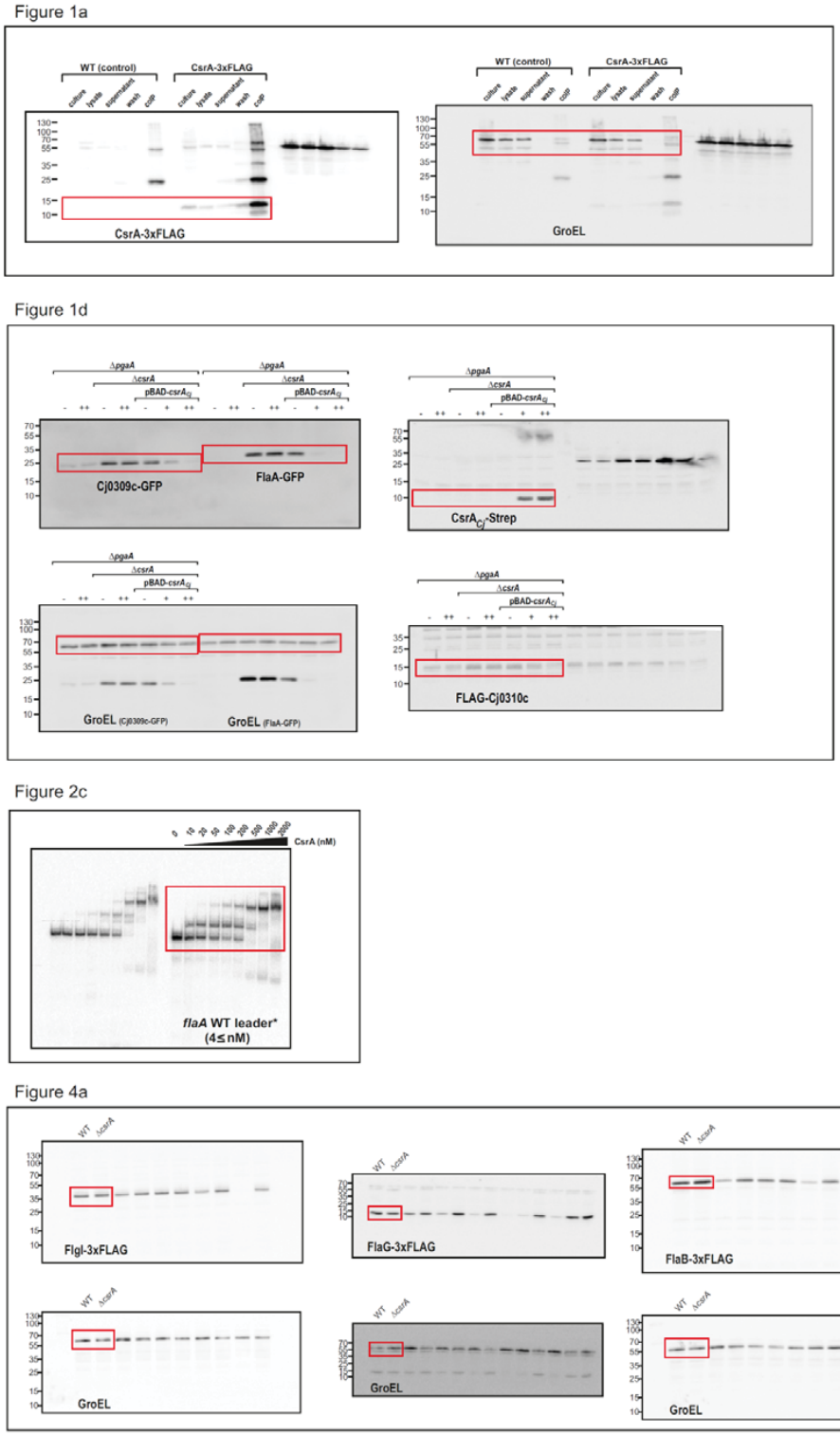
Supplementary Figure 17. Super-resolution imaging of *flaA* mRNA. RNA-FISH analysis of *flaA* mRNA (14 Cy5-labeled oligos) in the indicated *C. jejuni* strains using *d*STORM imaging. Cell boundaries from bright field images are depicted by white dotted lines. As a negative control, the *C. jejuni fliA* deletion strain was analyzed with and without probes to check for background signals.

Supplementary Figure 18



Supplementary Figure 18. Northern blot analysis of *flaA* mutant mRNAs. Northern blot analysis of *flaA* mRNA using oligonucleotide probe CSO-2835, as well as of 5S rRNA (loading control, probe CSO-0192) in *C. jejuni* NCTC11168 wild-type and the indicated mutant strains with point mutations in the *flaA* coding region. RNA was extracted from cells grown to mid-log phase.

Supplementary Figure 19



Supplementary Figure 19. Uncropped images of all Western blots and gels shown in the main Figures. The cropped parts that are shown in the main Figures are marked by red boxes.

Supplementary Tables

Supplementary Table 1. Mapping statistics of *C. jejuni* CsrA coIP RNA-seq libraries. The table indicates the total number of sequenced cDNA reads considered in the analysis, the number of reads that were removed due to insufficient length (<12 nt) after poly(A)-tail clipping (before read mapping), the number of reads that were successfully mapped to the reference genomes (or the pVir and pTet plasmids of strain 81-176) using *segemehl* (see Methods), the number of mappings (i.e. some reads map to different locations with the same score), and the number of uniquely-mapped reads. For the number of mapped reads and number of uniquely mapped reads, the percentage values (relative to the total number of reads) are also listed.

	***C. jejuni* NCTC11168 Control coIP**	***C. jejuni* NCTC11168 CsrA-3xFLAG coIP**
Total number of reads	6,214,261	5,389,919
Failed size filter after clipping	144,133	113,458
Total number of mapped reads	5,933,127	5,164,774
Total number of mappings (NC_002163)	13,207,712	8,897,385
Uniquely mapped reads	2,079,985	3,103,694
% mappable reads	95.48	95.82
% of uniquely mapped reads	33.47	57.58

	***C. jejuni* 81-176 Control coIP**	***C. jejuni* 81-176 CsrA-3xFLAG coIP**
Total number of reads	6,053,715	4,605,355
Failed size filter after clipping	295,445	223,949
Total number of mapped reads	5,641,676	4,299,931
Total number of mappings	13,439,403	9,866,887
Uniquely mapped reads	1,545,585	1,385,653
% mappable reads	93.19	93.37
% of uniquely mapped reads	25.53	30.09
Mapped reads in plasmid pVir (NC_008770)	16,383	9,753
Mapped reads in chromosome (NC_008787)	5,601,153	4,268,583
Mapped reads in plasmid pTet (NC_008790)	24,140	21,595
Mappings in plasmid pVir (NC_008770)	17,832	10,780
Mappings in chromosome (NC_008787)	13,396,129	9,833,383
Mappings in plasmid pTet (NC_008790)	25,442	22,724

Supplementary Table 2. Distribution of mapped reads to annotations. The table indicates the total number of mapped reads overlapping annotations for different RNA classes (sRNAs, 5'UTRs, ORFs, rRNAs, tRNAs, housekeeping RNAs, and pseudogenes) with a minimum overlap size of 10 nt (for details see Methods). Absolute read numbers and percentage values based on the total number of reads overlapping all annotations are shown for all libraries.

	***C. jejuni* NCTC11168 Control coIP**		***C. jejuni* NCTC11168 CsrA-3xFLAG coIP**	
sRNAs	28,720.50	0.50%	25,507.28	0.51%
5'UTRs	23,806.78	0.41%	756,447.88	15.08%
ORFs	249,394.84	4.30%	1,092,168.37	21.77%
rRNAs	3,285,700.41	56.66%	1,593,895.71	31.78%
tRNAs	1,509,545.38	26.03%	1,191,322.49	23.75%
housekeeping RNAs	699,295.33	12.06%	351,895.93	7.02%
pseudogenes	2,893.90	0.05%	4,752.24	0.09%
total	5,799,357.14	100.00%	5,015,989.91	100.00%

	***C. jejuni* 81-176 Control CoIP**		***C. jejuni* 81-176 CsrA-3xFLAG CoIP**	
sRNAs	32,282.87	0.58%	22,503.67	0.53%
5'UTRs	23,432.45	0.42%	187,578.08	4.40%
ORFs	344,388.16	6.16%	509,091.48	11.94%
rRNAs	3,641,665.55	65.09%	2,630,055.93	61.69%
tRNAs	1,267,744.45	22.66%	698,004.19	16.37%
housekeeping RNAs	284,913.59	5.09%	215,899.56	5.06%
total	5,594,427.06	100.00%	4,263,132.91	100.00%

Supplementary Table 3. Bacterial strains. List of all *C. jejuni* and *E. coli* strains used in this study. All strains were generated in this study unless otherwise stated. All *C. jejuni* strains correspond to NCTC11168 background unless otherwise stated.

Name	Description	Strain number	Resistance
***C. jejuni* strains**	All *C. jejuni* strains have NCTC11168 background unless otherwise stated		
NCTC11168	WT strain; (Kindly provided by Arnoud van Vliet, Institute of Food Research, Norwich, UK)	CSS-0032	-
81-176	WT strain; (Patricia Guerry, Naval Medical Research Center, Silver Spring, MD, USA)	CSS-0063	Tet^R
CsrA-3xFLAG NCTC11168	*csrA*-3xFLAG::*aphA-3* C-terminal 3xFLAG tag at native locus (Cj1103) in NCTC11168 background	CSS-0625	Kan^R
CsrA-3xFLAG 81-176	81-176, C-terminal 3xFLAG tag in 81-176 background	CSS-0604	Tet^R Kan^R
Δ*csrA*	*csrA*::*cat,* Deletion of *csrA* (Cj1103)	CSS-0643	Cm^R
Δ*fliW*	*fliW*::*aac*(3)-IV Deletion of *fliW* (Cj1075)	CSS-0820	Gm^R
Δ*csrA* Δ*fliW*	*csrA::cat; fliW::aac(3)-IV* Deletion of *csrA* and *fliW*	CSS-1134	Cm^R Gm^R
Δ*rpoN*	*rpoN*::*aac*(3)-IV Deletion of *rpoN* (Cj0670)	CSS-1141	Gm^R
Δ*fliA*	*fliA*::*aac*(3)-IV Deletion of *fliA* (Cj0061c)	CSS-1133	Gm^R
Δ*flaA*	*flaA*::*aphA-3* Deletion of *flaA* (Cj1339c)	CSS1512	Kan^R
Δ*flaB*	*flaB*::*aphA-3* Deletion of *flaB* (Cj1338c)	CSS-2892	Kan^R
Δ*flaA* Δ*flaB*	*flaAB*::*aphA-3* Deletion of *flaA* (Cj1339c) and *flaB* (Cj1338c)	CSS-2891	Kan^R
***flaA*-3xFLAG**	*flaA*-3xFLAG::*aphA-3* C-terminal 3xFLAG tag of *flaA* at native locus (Cj1339c)	CSS-0640	Kan^R
flaA*-3xFLAG Δ*csrA	*flaA*-3xFLAG::*aphA-3; csrA*::*cat* Deletion of *csrA* in *flaA*-3xFLAG background	CSS-0644	Kan^R Cm^R
flaA*-3xFLAG Δ*fliW	*flaA*-3xFLAG::*aphA-3; fliW*::*aac* Deletion of *fliW* in *flaA*-3xFLAG background	CSS-1100	Gm^R Kan^R
flaA*-3xFLAG Δ*csrA* Δ*fliW	*flaA*-3xFLAG::*aphA-3; csrA*::*cat; fliW*::*aac*(3)-IV Deletion of *csrA* and *fliW* in *flaA*-3xFLAG background	CSS-1107	Kan^R Cm^R Gm^R
FlaA-3xFLAG M1	*flaA*[5'UTR $SL1^{GGA\rightarrow AAA}$::*aac*(3)-IV]-3xFLAG::*aphA-3* $SL1^{flaA}$ 5'UTR point mutant in *flaA*-3xFLAG background	CSS-0991	Kan^R Gm^R
FlaA-3xFLAG M1 Δ*csrA*	*flaA*[5'UTR $SL1^{GGA\rightarrow AAA}$::*aac*(3)-IV]-3xFLAG::*aphA-3; csrA*::*cat* $SL1^{flaA}$ point mutant and *csrA* deletion in *flaA*-3xFLAG background	CSS-1410	Kan^R Cm^R Gm^R
FlaA-3xFLAG M1 Δ*fliW*	*flaA*[5'UTR $SL1^{GGA\rightarrow AAA}$::*cat*]-3xFLAG::*aphA-3; fliW*::*aac*(3)-IV Deletion of *fliW* in *flaA* 5'UTR point mutant/3xFLAG-tag	CSS-1418	Kan^R Cm^R Gm^R
FlaA-3xFLAG M1 Δ*fliW* Δ*csrA*	*flaA*-3xFLAG::*aphA-3; flaA*[5'UTR $SL1^{GGA\rightarrow AAA}$::*cat*]; *fliW*::*aac*(3)-IV *csrA*::*aph(7'')* $SL1^{flaA}$ point mutant, *fliW* deletion and *csrA* deletion in *flaA*-3xFLAG background	CSS-1554	Kan^R Cm^R Gm^R Hyg^R
FlaA-3xFLAG M2	*flaA*[5'UTR $SL1^{GGA\rightarrow UGA}$::*aac*(3)-IV]-3xFLAG::*aphA-3* $SL1^{flaA}$ point mutant in *flaA*-3xFLAG background	CSS-0955	Kan^R Gm^R
FlaA-3xFLAG M2 Δ*csrA*	*flaA*[5'UTR $SL1^{GGA\rightarrow UGA}$::*aac*(3)-IV]-3xFLAG::*aphA-3 csrA*::*cat* $SL1^{flaA}$ point mutant and *csrA* deletion in *flaA*-3xFLAG background	CSS-1105	Kan^R Cm^R Gm^R
FlaA-3xFLAG M2/M3	*flaA*[5'UTR $SL1^{GGA\rightarrow UGA}$ $SL2^{GGA\rightarrow GGG}$::*aac*(3)-IV]-3xFLAG::*aphA-3* $SL1^{flaA}$ and $SL2^{flaA}$ double point mutant in *flaA*-3xFLAG background	CSS-1095	Kan^R Gm^R
FlaA-3xFLAG M2/M3 Δ*csrA*	*flaA*[5'UTR $SL1^{GGA\rightarrow UGA}$ $SL2^{GGA\rightarrow GGG}$::*aac*(3)-IV]-3xFLAG::*aphA-3 csrA*::*cat* $SL1^{flaA}$ and $SL2^{flaA}$ double point mutant and *csrA* deletion in *flaA*-3xFLAG background	CSS-1106	Kan^R Cm^R Gm^R
FlaG-3xFLAG	*flaG*-3xFLAG::*aphA-3*; C-terminal 3xFLAG-tag of *flaG* at native locus (Cj0547)	CSS-0968	Kan^R

FlaG-3xFLAG Δ*csrA*	*flaG*-3xFLAG::*aphA-3*; *csrA*::*cat* Deletion of *csrA* in *flaG*-3xFLAG background	CSS-0983	Kan^R Cm^R
FlaG-3xFLAG Δ*fliW*	*flaG*-3xFLAG::*aphA-3*; *fliW*::*aac*(3)-IV Deletion of *fliW* in FlaG-3xFLAG background	CSS-1112	Kan^R Gm^R
FlaG-3xFLAG Δ*csrA* Δ*fliW*	*flaG*-3xFLAG::*aphA-3; csrA*::*cat; fliW*::*aac*(3)-IV Deletion of *csrA* and *fliW* in FlaG-3xFLAG background	CSS-1204	Kan^R Cm^R Gm^R
FlaG-3xFLAG M1	*flaG*::3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*aac*(3)-IV] SL1^{flaA} point mutant in *flaG*-3xFLAG background	CSS-1399	Kan^R Gm^R
FlaG-3xFLAG M1 Δ*csrA*	*flaG*-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*aac*(3)-IV] *csrA*::*cat* SL1^{flaA} point mutant and *csrA* deletion in *flaG*-3xFLAG background	CSS-1411	Kan^R Cm^R Gm^R
FlaG-3xFLAG M1 Δ*fliW*	*flaG*-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*cat*] *fliW*::*aac*(3)-IV SL1^{flaA} point mutant and *fliW* deletion in *flaG*-3xFLAG background	CSS-1421	Kan^R Cm^R Gm^R
FlaG-3xFLAG M1 Δ*fliW* Δ*csrA*	*flaG*-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*cat*]*; fliW*::*aac*(3)-IV *csrA*::*aph(7")*; SL1^{flaA} point mutant, *fliW* deletion and *csrA* deletion in *flaG*-3xFLAG background	CSS-1435	Kan^R Cm^R Gm^R Hyg^R
FlgI-3xFLAG	*flgI*-3xFLAG::*aphA-3* C-terminal 3xFLAG tag of *flgI* at native locus (Cj1462)	CSS-0967	Kan^R
FlgI-3xFLAG Δ*csrA*	*flgI*-3xFLAG::*aphA-3; csrA*::*cat* Deletion of *csrA* in *flgI*-3xFLAG background	CSS-0982	Kan^R Cm^R
FlgI-3xFLAG Δ*fliW*	*flgI*-3xFLAG::*aphA-3; fliW*::*aac*(3)-IV Deletion of *fliW* in *flgI*-3xFLAG background	CSS-1114	Kan^R Gm^R
FlgI-3xFLAG Δ*csrA* Δ*fliW*	*flgI*-3xFLAG::*aphA-3; csrA*::*cat; fliW*::*aac*(3)-IV Deletion of *csrA* and *fliW* in *flgI*-3xFLAG background	CSS-1203	Kan^R Cm^R Gm^R
FlgI-3xFLAG M1	*flgI*-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*aac*(3)-IV] SL1^{flaA} point mutant in *flgI*-3xFLAG background	CSS-1426	Kan^R Gm^R
FlgI-3xFLAG M1 Δ*csrA*	*flgI*-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*aac*(3)-IV] *csrA*::*cat* SL1^{flaA} point mutant and *csrA* deletion in *flgI*-3xFLAG background	CSS-1542	Kan^R Cm^R Gm^R
FlgI-3xFLAG M1 Δ*fliW*	*flgI*-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*cat*] *fliW*::*aac*(3)-IV SL1^{flaA} point mutant and *fliW* deletion in *flgI*-3xFLAG background	CSS-1420	Kan^R Cm^R Gm^R
FlgI-3xFLAG M1 Δ*fliW* Δ*csrA*	*flgI*-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*cat*]*; fliW*::*aac*(3)-IV *csrA*::*aph(7")* SL1^{flaA} point mutant, *fliW* deletion and *csrA* deletion in *flgI*-3xFLAG background	CSS-1436	Kan^R Cm^R Gm^R Hyg^R
FlaB-3xFLAG	*flaB*-3xFLAG::*aphA-3;* C-terminal 3xFLAG tag of *flaB* at native locus (Cj1338c)	CSS-0641	Kan^R
FlaB-3xFLAG Δ*csrA*	*flaB*-3xFLAG::*aphA-3; csrA*::*cat* Deletion of *csrA* in *flaB*-3xFLAG background	CSS-0645	Kan^R Cm^R
FlaB-3xFLAG Δ*fliW*	*flaB*-3xFLAG::*aphA-3; fliW*::*aac*(3)-IV Deletion of *fliW* in *flaB*-3xFLAG background	CSS-1201	Kan^R Gm^R
FlaB-3xFLAG Δ*csrA* Δ*fliW*	*flaB*-3xFLAG::*aphA-3; csrA*::*cat; fliW*::*aac*(3)-IV Deletion of *csrA* and *fliW* in *flaB*-3xFLAG background	CSS-1202	Kan^R Cm^R Gm^R
FlaB-3xFLAG M1	*flaB*-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*aac*(3)-IV] SL1^{flaA} point mutant in *flaB*-3xFLAG background	CSS-1405	Kan^R Gm^R
FlaB-3xFLAG M1 Δ*csrA*	*flaB*-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*aac*(3)-IV] SL1^{flaA} point mutant and *csrA* deletion in *flaB*-3xFLAG background	CSS-1543	Kan^R Cm^R Gm^R
FlaB-3xFLAG M1 Δ*fliW*	*flaB*-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*cat*] *fliW*::*aac*(3)-IV SL1^{flaA} point mutant and *fliW* deletion in *flaB*-3xFLAG background	CSS-1419	Kan^R Cm^R Gm^R
FlaB-3xFLAG M1 Δ*fliW* Δ*csrA*	*flaB*-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*cat*]*; fliW*::*aac*(3)-IV *csrA*::*aph(7")* SL1^{flaA} point mutant, *fliW* deletion and *csrA* deletion in *flaB*-3xFLAG background	CSS-1438	Kan^R Cm^R Gm^R Hyg^R
Cj0529-3xFLAG	Cj0529-3xFLAG::*aphA-3* C-terminal 3xFLAG tag of Cj0529 at native locus	CSS-1541	Kan^R
Cj0529-3xFLAG Δ*csrA*	Cj0529-3xFLAG::*aphA-3; csrA*::*cat* Deletion of *csrA* in Cj0529-3xFLAG background	CSS-1431	Kan^R Cm^R
Cj0529-3xFLAG Δ*fliW*	Cj0529-3xFLAG::*aphA-3; fliW*::*aac*(3)-IV Deletion of *fliW* in Cj0529-3xFLAG background	CSS-1430	Kan^R Gm^R
Cj0529-3xFLAG Δ*csrA* Δ*fliW*	Cj0529-3xFLAG::*aphA-3; csrA*::*cat; fliW*::*aac*(3)-IV Deletion of *csrA* and *fliW* in Cj0529-3xFLAG background	CSS-1437	Kan^R Cm^R Gm^R
Cj0529-3xFLAG M1	Cj0529-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*aac*(3)-IV] SL1^{flaA} point mutant in Cj0529-3xFLAG background	CSS-1432	Kan^R Gm^R
Cj0529-3xFLAG M1 Δ*csrA*	Cj0529-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*aac*(3)-IV]*; csrA*::*cat* SL1^{flaA} point mutant and *csrA* deletion in Cj0529-3xFLAG background	CSS-1576	Kan^R Cm^R Gm^R

Cj0529-3xFLAG M1 Δ*fliW*	Cj0529-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*cat]; fliW*::*aac*(3)-IV SL1^{flaA} point mutant and *fliW* deletion in Cj0529-3xFLAG background	CSS-1575	Kan^R Cm^R Gm^R
Cj0529-3xFLAG M1 Δ*fliW* Δ*csrA*	Cj0529-3xFLAG::*aphA-3; flaA*[5'UTR SL1$^{GGA\rightarrow AAA}$::*cat]; fliW*::*aac*(3)-IV *csrA*:: *aph(7")* SL1^{flaA} point mutant, *fliW* deletion and *csrA* deletion in Cj0529-3xFLAG background	CSS-1582	Kan^R Cm^R Gm^R Hyg^R
flaA X1	*flaA*[Start Codon$^{AUG\rightarrow AAG}$::*aac*(3)-IV]	CSS-1586	Gm^R
flaA X2	*flaA*[Start Codon$^{AUG\rightarrow AUU}$::*aac*(3)-IV]	CSS-3089	Gm^R
flaA X3	*flaA*[Start Codon$^{AUG\rightarrow GUG}$::*aac*(3)-IV]	CSS-3087	Gm^R
flaA X4	*flaA*[3rd Codon$^{UUU\rightarrow UAG}$::*aac*(3)-IV]	CSS-3091	Gm^R
flaA X5	*flaA*[3rd Codon$^{UUU\rightarrow UUC}$::*aac*(3)-IV]	CSS-3093	Gm^R
flaA X6	*flaA*[101st Codon$^{CAA\rightarrow UAA}$::*aac*(3)-IV]	CSS-3095	Gm^R
P$_{MetK}$-*flaA*	Exchange of *flaA* native promoter with *metK* promoter	CSS-3096	Gm^R
P$_{MetK}$-*flaA*-3xFLAG	*flaA*-3xFLAG::*aphA-3*; P$_{MetK}$-*flaA*:: *aac*(3)-IV C-terminal 3xFLAG tag of *flaA* in *flaA* promoter exchanged strain	CSS-3098	Gm^R Kan^R
P$_{MetK}$-*flaA*-3xFLAG Δ*csrA*	*flaA*-3xFLAG::*aphA-3*; P$_{MetK}$-*flaA*:: *aac*(3)-IV*; csrA*::*cat* C-terminal 3xFLAG tag of *flaA* in *flaA* promoter exchanged strain	CSS-3102	Gm^R Kan^R Cm^R
P$_{MetK}$-*flaA*-3xFLAG Δ*fliW*	*flaA*-3xFLAG::*aphA-3*; P$_{MetK}$-*flaA*:: *aac*(3)-IV*; fliW*::*hyg* C-terminal 3xFLAG tag of *flaA* in *flaA* promoter exchanged strain	CSS-3104	Gm^R Kan^R Hyg^R
P$_{MetK}$-*flaA*-3xFLAG Δ*csrA* Δ*fliW*	*flaA*-3xFLAG::*aphA-3*; P$_{MetK}$-*flaA*:: *aac*(3)-IV; *csrA*::*cat; fliW*::*aph(7")* C-terminal 3xFLAG tag of *flaA* in *flaA* promoter-exchanged strain	CSS-3124	Gm^R Kan^R Cm^R Hyg^R
FliW-3xFLAG	*fliW*-3xFLAG::*aac*(3)-IV C-terminal 3xFLAG tag of *fliW* at native locus (Cj1075)	CSS-0962	Gm^R
FliW-mCherry	*fliW*-mCherry::*aac*(3)-IV C-terminal mCherry tag of *fliW* at native locus (Cj1075)	CSS-3073	Gm^R
CsrA-mCherry	*csrA*-mCherry::*aphA-3* C-terminal mCherry tag of *csrA* at native locus (Cj1103)	CSS-3071	Kan^R
FliW-3xFLAG CsrA-mCherry	*fliW*-3xFLAG::*aac*(3)-IV *csrA*-mCherry::*aphA-3*	CSS-3126	Gm^R Kan^R
FlaA-3xFLAG FliW-mCherry	*flaA*-3xFLAG::*aphA-3* *fliW*-mCherry::*aac*(3)-IV	CSS-3128	Gm^R Kan^R
***flaA*_mini**	*flaA*_mini::*rdxA* Introduction of *flaA_mini* in *rdxA* (Cj1066) complementation locus	CSS-3075	Cm^R
FlaA-3xFLAG *flaA*_mini	*flaA*-3xFLAG::*aphA-3*; *flaA_mini*::*rdxA* Introduction of *flaA*_mini in *rdxA* (Cj1066) in *flaA*-3xFLAG background	CSS-3076	Cm^R Kan^R
FlaA-3xFLAG Δ*fliW* *flaA*_mini	*flaA*-3xFLAG::*aphA-3*; *fliW*::*aac*(3)-IV*; flaA_mini*::*rdxA* Introduction of *flaA_mini* in *rdxA* (Cj1066) in *flaA*-3xFLAG background with *fliW* deletion	CSS-3106	Cm^R Kan^R Gm^R
FlaG-3xFLAG *flaA*_mini	*flaG*-3xFLAG::*aphA-3*; *flaA_mini*::*rdxA* Introduction of *flaA_mini* in *rdxA* (Cj1066) in *flaG*-3xFLAG background	CSS-3080	Cm^R Kan^R
FlaG-3xFLAG Δ*fliW* *flaA*_mini	*flaG*-3xFLAG::*aphA-3*; *fliW*::*aac*(3)-IV*; flaA_mini*::*rdxA* Introduction of *flaA_mini* in *rdxA* (Cj1066) in *flaG*-3xFLAG background with *fliW* deletion	CSS-3112	Cm^R Kan^R Gm^R
Cj1321_mini	Cj1321_mini::*rdxA* Introduction of *flaA_mini* in *rdxA* (Cj1066) complementation locus	CSS-3130	Cm^R
Cj1321_mini Δ*csrA*	Cj1321_mini::*rdxA csrA*::*aph(7")* Introduction of *flaA_mini* in *rdxA* (Cj1066) complementation locus	CSS-3131	Cm^R Hyg^R
Cj1321_mini Δ*fliW*	Cj1321_mini::*rdxA fliW*::*aac*(3)-IV Introduction of *flaA_mini* in *rdxA* (Cj1066) complementation locus	CSS-3132	Cm^R Gm^R
***E. coli* strains**			
TOP10	*mcrA* Δ(*mrr-hsdRMS-mcrBC*) Φ80*lacZ* ΔM15 Δ*lacX74 deoR recA1 araD139* Δ(*ara-leu*)7697 *galU galK rpsL* endA1 nupG (from Invitrogen)	CSS-0070	Str^R
Δ*pgaA*	*pgaA* deletion in TOP10 background	CSS-0556	Str^R
Δ*pgaA* Δ*csrA*	*pgaA* and *csrA* deletion in TOP10 background	CSS-0557	Str^R Kan^R

Supplementary Table 4. DNA oligonucleotides. List of all DNA oligonucleotides used in this study for PCR amplification, Northern blot hybridization, and FISH assays. DNA sequences are given in 5' to 3' direction; P- denotes a 5' monophosphate.

Name	Sequence (5' → 3')	Description
CSO-0023	CCACCAGCTTATATACCTTAGCA	Antisense to *aphA-3* for verification
CSO-0073	CTAACAAGCTTTCATCTACGCA	3xFLAG Tagging using pGG1
CSO-0074	GTTTTTGAATTCTATTCCCTCCAGGTACTAAAACA	3xFLAG Tagging using pGG1
CSO-0075	TCCTTCACAAAGAAGGGG	3xFLAG Tagging using pGG1
CSO-0171	P-TTTGATTAGTTTTTTGCTTAAGTCAT	Cloning of *csrA*-3xFLAG in pGG1
CSO-0172	GTTTTTCTCGAGCTCTTTAGAGCGCATTAAAGAA	Cloning of *csrA* -3xFLAG in pGG1
CSO-0173	GTTTTTTCTAGACAAGATATTTGTGGAAAAGTCC	Cloning of *csrA*-3xFLAG in pGG1
CSO-0174	GTTTTTGAATTCATCAAATGAAAGCTTACGCTAA	Cloning of *csrA*-3xFLAG in pGG1
CSO-0196	GTATTTGATTGCAAGATCTTAAGC	Verification of *csrA*-3xFLAG in *C. jejuni*
CSO-0392	TACTCCTTAAGTCTTGATGATCAA	Verification of *csrA* deletion in *C. jejuni*
CSO-0393	TCCTAGTTAGTCACCCGGGTACCTTGATAATATTAACAT TTTTCAACCT	Deletion of *csrA* using *hyg* in *C. jejuni*
CSO-0394	TGCAAGGAATTATCTCCTATACAC	Deletion of *csrA* using *hyg/cat* in *C. jejuni*
CSO-0395	ATCATAAACAGCTTTAGTTTGGC	Deletion of *csrA* using *hyg/cat* in *C. jejuni*
CSO-0396	ATTGTTTTAGTACCTGGAGGGAATAGCAAAAAACTAATC AAATGAAAG	Deletion of *csrA* using *hyg* in *C. jejuni*
CSO-0483	GTTTTTGGATCCTTTTATGGATAATTTTTAAAATCATTTG	Cloning of *aac*(3)-IV upstream of *flaA* 5'UTR
CSO-0486	GTGTTAATACGAAATCCCATTTTAAATC	NB detection *flaA* mRNA (Binds 5'UTR)
CSO-0553	P-CTGTAGTAATCTTAAAACATTTTGTTGA	Cloning of *flaA*-3xFLAG in pGG1
CSO-0554	GTTTTTCTCGAGTGGTTATTCTTCTGTTAGTGCCT	Cloning of *flaA*-3xFLAG in pGG1
CSO-0555	GTTTTTTCTAGAGCGATATTGTCAAGTTCTTCC	Cloning of *flaA*-3xFLAG in pGG1
CSO-0556	GTTTTTGAATTCTTTACAAAAGCTGCAATATATACAAA	Cloning of *flaA*-3xFLAG in pGG1
CSO-0557	CTCTCAAGCTTCTGTTTCTTTAAG	Verification of *flaA*-3xFLAG in *C. jejuni*
CSO-0558	P-TTGAAGAAGTTTTAAAACATTTTGC	Cloning of *flaB*-3xFLAG in pGG1
CSO-0559	GTTTTTCTCGAGTTAGTGCCTATATGAGTAGCGC	Cloning of *flaB*-3xFLAG in pGG1
CSO-0560	GTTTTTTCTAGAGTGCTAGGATAGAAAGCGCT	Cloning of *flaB*-3xFLAG in pGG1 / Overlap PCR construction of *flaB* deletion with *aphA-3*
CSO-0561	GTTTTTGAATTCTTTCTTAGATGCTTTTATGCATCT	Cloning of *flaB*-3xFLAG in pGG1
CSO-0562	GATGCTAATATCGCTGATGC	Verification of *flaB*-3xFLAG
CSO-0575	CAATACGAATGGCGAAAAG	*aac*(3)-IV cloning in pGG1
CSO-0576	GTTTTTCATATGAAACACCCCCATAAGTGCAATTATGGG GATAAATCATCTCGTTCTCCGCTC	Cloning of *aac*(3)-IV upstream of *flaA* 5'UTR
CSO-0577	P-CATTTATTCCTCCTAGTTAGTCACC	*aac*(3)-IV cloning in pGG1
CSO-0606	GTTTTTATGCATTTTATTCAAGAAAATTCAACTACGG	Cj0805-Cj0806 (*dapA*) cloning in pXG-30
CSO-0607	GTTTTTGCTAGCTTGCTCATCAACTTTTCCAT	Cj0805-Cj0806 (*dapA*) cloning in pXG-30
CSO-0608	GTTTTTATGCATGCAATTTTACTTTTAAGTATTATAGCCC	Cj0310c-Cj0309c cloning in pXG-30
CSO-0609	GTTTTTGCTAGCAAGTTCTTTCATGATCACCACG	Cj0310c-Cj0309c cloning in pXG-30
CSO-0611	TACAGAGAGACCCGACTCTTTTAATCTTTCAAGGAGCAA AGAATGGTGTAGGCTGGAGCTGCTTC	Deletion of *csrA* in *E. coli* (using λ-red system)
CSO-0612	TTTGAGGGTGCGTCTCACCGATAAAGATGAGACGCGGA AAGATTAGGTCCATATGAATATCCTCCTTAG	Deletion of *csrA* in *E. coli* (using λ-red system)
CSO-0613	AACAAATCGGAATTTACGGA	Amplification of *C. coli cat* cassette
CSO-0614	GGCACCAATAACTGCCTTAA	Amplification of *C. coli cat* cassette

CSO-0615	CTCCGTAAATTCCGATTTGTTCTTGATAATATTAACATTTTTCAACCT	Deletion of *csrA* using *cat* in *C. jejuni*
CSO-0616	TTTTAAGGCAGTTATTGGTGCCGCAAAAAACTAATCAAATGAAAG	Deletion of *csrA* using *cat* in *C. jejuni*
CSO-0621	GTTTTTATGCATTAACAAGTTCATGGATGAGCTT	*flaA* cloning in pXG-10
CSO-0622	GTTTTTGCTAGCACTAAGTCTGCTTAAAGAAGCATC	*flaA* cloning in pXG-10
CSO-0639	GATGTAATGTGTTTGTCATTGCT	Verification of *csrA* deletion in *E. coli*
CSO-0640	GAGACTTAAGTTGAATGAACGG	Verification of *csrA* deletion in *E. coli*
CSO-0652	AGATACAGAGAGAGATTTTGGCAATACATGGAGTAATACAGGATGGTGTAGGCTGGAGCTGCTTC	Deletion of *pgaA* in *E.coli* (using λ-red system)
CSO-0653	GCATCAGGAGATATTTATTTCCATTACGTAACATATTTATCCTTAGGTCCATATGAATATCCTCCTTAG	Deletion of *pgaA* in *E.coli* (using λ-red system)
CSO-0654	TCTCTCTTCCGCGTTTAATAAC	Verification of *pgaA* deletion in *E.coli*
CSO-0655	CTGTGGCGGTATAAATGATG	Verification of *pgaA* deletion in *E.coli*
CSO-0694	GTTTTTATGCATACAATAGATTAAAGGAAGAATCCAT	*flgI* cloning in pXG-10
CSO-0695	GTTTTTGCTAGCACCTATAAGTTGGTTATCTCTTACACC	*flgI* cloning in pXG-10
CSO-0701	GTTTTTTTTAATACGACTCACTATAGGAAAGCTGGTGGCGCTG	*in vitro* transcription of *hopB 3'*end, carries T7 promoter
CSO-0702	GTTAAATCAAAGCCTATAAAAGGCC	*in vitro* transcription of *hopB 3'*end
CSO-0709	GTTTTTTTTAATACGACTCACTATAGGTAACAAGTTCATGGATGAGCTT	*in vitro* transcription of *flaA* leader, carries T7 promoter
CSO-0710	ACTAAGTCTGCTTAAAGAAGCATCT	*in vitro* transcription of *flaA* leader
CSO-0713	GTTTTTTTTAATACGACTCACTATAGGACAATAGATTAAAGGAAGAATCCAT	*in vitro* transcription of *flgI* leader, carries T7 promoter
CSO-0714	ACCTATAAGTTGGTTATCTCTTACACC	*in vitro* transcription of *flgI* leader
CSO-0748	P-TAACAAGTTCATGGATGAGCTT	*flaA* leader cloning in pBAD plasmid
CSO-0749	GTTTTTTCTAGAGTTTGCTTTTGCATTTAAAGCT	*flaA* leader cloning in pBAD plasmid
CSO-0752	GTTTTTCTCGAGAAGGTGGAGCAAGGATTAA	*flaA* cloning in pJV752.1 / Overlap PCR construction of *flaA* deletion with *aphA-3*
CSO-0753	GTTTTTTCTAGATCTTAGAAGATTGAGTTGCTCC	*flaA* cloning in pJV752.1
CSO-0754	GTTTTTCATATGCAATAAAATTTCATACTTTTGACA	Cloning of *aac*(3)-IV upstream of *flaA* 5'UTR
CSO-0755	GTTTTTGGATCCTAAAGTATAAAATATTTTTTTGATTGCA	Cloning of *aac*(3)-IV upstream of *flaA* 5'UTR
CSO-0756	TATGCAGGCAAAGGTGAAG	Verification of *flaA* deletion in *C. jejuni*
CSO-0757	TAACAAGTTCATTGATGAGCTTGAATTTTTTTAAAAG	Introduction of SL1$^{GGA>UGA}$ (M2) mutation into *flaA* 5' UTR
CSO-0758	TTCAAGCTCATCAATGAACTTGTTAAATGCTATATCGT	Introduction of SL1$^{GGA>UGA}$ (M2) mutation into *flaA* 5' UTR
CSO-0831	GTTTTTCATATGTGTTTTAGTACCTGGAGGGAATA	*aac(3)-IV* cloning in pGG1
CSO-0832	GTTTTTCATATGTCATCTCGTTCTCCGCTC	*aac(3)-IV* cloning in pGG1
CSO-0852	CGGGTGGCTCCATTTGATTAGTTTTTTGCTTAAGT	Addition of Strep-tag to *csrA* in pBAD plasmid
CSO-0853	P-CAGTTCGAAAAATGAAAGCTTACGCTCTAGA	Addition of Strep-tag to *csrA* in pBAD plasmid
CSO-0997	GATAACGAATATAATCAGCATTGC	Deletion of *fliW* using *aac*(3)-IV in *C. jejuni*
CSO-0998	TCCTAGTTAGTCACCCGGGTACGCATTTCAGCTAGGGTCATG	Deletion of *fliW* using *aac*(3)-IV in *C. jejuni*
CSO-0999	TGTGTTTTAGTACCTGGAGGGAATACCGACTTTTTTCAAGCTGATC	Deletion of *fliW* using *aac*(3)-IV in *C. jejuni*
CSO-1000	GACAAACCTTCATAAACTCCAG	Deletion of *fliW* using *aac*(3)-IV in *C. jejuni*
CSO-1002	GTTTTTCTCGAGAAATTTTGGCACAGTTTTTGCTTA	Overlap PCR construction of *flaG*-3xFLAG with *aphA-3* cassette
CSO-1003	GTTTTTTCTAGACCTGTGTTTACAATCTTAGCAAC	Overlap PCR construction of *flaG*-3xFLAG with *aphA -3* cassette
CSO-1005	GTGATAGAAGATTTGATCTTGC	Verification of *flaG*-3xFLAG in *C. jejuni*
CSO-1011	P-GATGATCTCCAAATCCGCGT	Cloning of *flgI*-3xFLAG in pGG1
CSO-1012	GTTTTTCTCGAGAAATTCCACAAAATTTTAGCC	Cloning of *flgI*-3xFLAG in pGG1

CSO-1013	GTTTTTTCTAGATGCGATTTTACTCGCTTTATCA	Cloning of *flgI*-3xFLAG in pGG1
CSO-1014	GTTTTTGAATTCACGCGGATTTGGAGATCATC	Cloning of *flgI*-3xFLAG in pGG1
CSO-1015	AACTGTAATGGGCGGAGCTA	Verification of *flgI*-3xFLAG in *C. jejuni*
CSO-1072	TAAAGCCTGATTACGATTTGGC	Verification of *fliW* deletion in *C. jejuni*
CSO-1081	GTTTTTTTTAATACGACTCACTATAGGTAACAAGTTCATTGATGAGCTTG	*in vitro* transcription of *flaA* M1/M2 variant, carries T7 promoter
CSO-1082	GTTTTTTTTAATACGACTCACTATAGGAAATTAAAATTTTAAAAAGGAAGTTAAA	*in vitro* transcription of Cj0040, carries T7 promoter
CSO-1083	TCTAAAACTTGAAGCAAAACTTC	*in vitro* transcription of Cj0040
CSO-1084	GTTTTTTTTAATACGACTCACTATAGGACTAGCAATAGGAAATTTTAAAAAG	*in vitro* transcription of *flaG*, carries T7 promoter
CSO-1085	TGTGTCTCACTTGTTCTTTGG	*in vitro* transcription of *flaG*
CSO-1088	GTTTTTTTTAATACGACTCACTATAGGCAAATGTTGATGTTTTAATCGAA	*in vitro* transcription of *flgA*, carries T7 promoter
CSO-1089	TTTACCCACTACGATACCTTG	*in vitro transcription of flgA*
CSO-1092	GTTTTTTTTAATACGACTCACTATAGGATTATAACTAAGATCAAGGAG	*in vitro* transcription of *flgM*, carries T7 promoter
CSO-1093	CTTTATCTATTCTATTTGTATTTAATG	*in vitro* transcription of *flgM*
CSO-1098	TCACCGTCATGGTCTTTGTAGTCACTCTCCTTATCAAATATCATTCC	Overlap PCR construction of *flaG*-3xFLAG with *aphA -3* cassette
CSO-1099	ATTGTTTTAGTACCTGGAGGGGGAATGATATTTGATAAGGAGAGT	Overlap PCR construction of *flaG*-3xFLAG with *aphA -3* cassette
CSO-1114	TAACAAGTTCATAAATGAGCTTGAATTTTTTTAAAAGG	Introduction of SL1$^{GGA>AAA}$ (M1) mutation into *flaA* 5'UTR
CSO-1115	ATTCAAGCTCATTTATGAACTTGTTAAATGCTATATCG	Introduction of SL1$^{GGA>AAA}$ (M1) mutation into *flaA* 5'UTR
CSO-1116	TTTTTTAAAAGGGTTTAAAATGGGATTTCGTATTAACA	Introduction of SL2$^{GGA>GGG}$ (M3) mutation into *flaA* 5'UTR
CSO-1117	TCCCATTTTAAACCCTTTTAAAAAAATTCAAGCTCAT	Introduction of SL2$^{GGA>GGG}$ (M3) mutation into *flaA* 5'UTR
CSO-1138	GTTTTTCATATGTATAAAATATTTTTTTGATTGCACGATATAGCATTTAACAAGTTCATGGATGAGCTT	Cloning of *flaA_mini* in *Campylobacter rdxA* complementation plasmid
CSO-1139	GTTTTTATCGATAAGGCCCAGTCTTTCGACT	Cloning of *flaA_mini* in *Campylobacter rdxA* complementation plasmid
CSO-1144	AGTGGAAAAGTTCTTTTAGACGG	Overlap PCR construction of *rpoN* deletion with *aac*(3)-IV
CSO-1145	TCCTAGTTAGTCACCCGGGTACCTTGGGTGATTTTTTGCTTTAACA	Overlap PCR construction of *rpoN* deletion with *aac*(3)-IV
CSO-1146	TGTGTTTTAGTACCTGGAGGGAATATCTATCAATCTATCAAACCCATTAC	Overlap PCR construction of *rpoN* deletion with *aac*(3)-IV
CSO-1147	CATTGGACGCTCAGGACG	Overlap PCR construction of *rpoN* deletion with *aac*(3)-IV
CSO-1148	AACAACTTTTTATATGATATGTGGAC	Verification of *rpoN* deletion in *C. jejuni*
CSO-1149	GAATTCTTAGGTCATTTAAGCGC	Overlap PCR construction of *fliA* deletion with *aac*(3)-IV
CSO-1150	TCCTAGTTAGTCACCCGGGTACTTTCTTTAGCATTTGTGCATAAGC	Overlap PCR construction of *fliA* deletion with *aac*(3)-IV
CSO-1151	TGTGTTTTAGTACCTGGAGGGAATATAAAAAACTTAGAGAAAGGCTAGTG	Overlap PCR construction of *fliA* deletion with *aac*(3)-IV
CSO-1152	GATAACAATCTCATTTTGAGATACG	Overlap PCR construction of *fliA* deletion with *aac*(3)-IV
CSO-1153	TGCAGATGCAAACATTAAAAATCC	Verification of *fliA* deletion in *C. jejuni*
CSO-1407	CAGCCAAACAACTTGGACTT	Verification of Cj0529c-3xFLAG in *C. jejuni*
CSO-1408	TCACCGTCATGGTCTTTGTAGTCTTTTCCTTTTTGTAAATTTATGGCTT	Overlap PCR construction of Cj0529-3xFLAG with *aphA -3* cassette
CSO-1409	GCTCCTTATGATGAAGGAGT	Overlap PCR construction of Cj0529-3xFLAG with *aphA -3* cassette
CSO-1410	CTTTAACTTAATTTAGAGCTTGC	Overlap PCR construction of Cj0529-3xFLAG with *aphA -3* cassette
CSO-1411	ATTGTTTTAGTACCTGGAGGGAATATTTTTGATATTTTTATACAAAATAGTTAA	Overlap PCR construction of Cj0529-3xFLAG with *aphA -3* cassette
CSO-1471	GTTTTTTTTAATACGACTCACTATAGGTAACAAGTTCATAAATGAGCTTGA	*in vitro* transcription of *flaA* 5'UTR M1 variant, carries T7 promoter
CSO-1548	TCCTAGTTAGTCACCCGGGTATTTAAATCCTTTTAAAAAAATTCAAGCT	Overlap PCR construction of *flaA* deletion with *aphA-3*

CSO-1549	ATTGTTTTAGTACCTGGAGGGAATATTTACAAAAGCTGC AATATATACAAAT	Overlap PCR construction of *flaA* deletion with *aphA-3*
CSO-1550	ATAGCTTGACCTAAAGTGGCT	Overlap PCR construction of *flaA* deletion with *aphA-3*
CSO-1665	GTTTTTTTTAATACGACTCACTATAGGATTTTTATTAAATT GAAGGGGTGGG	*in vitro* transcription of Cj1324, carries T7 promoter
CSO-1666	TACCTTCTTTATCTTTTGTAAAATTAATAC	*in vitro* transcription of Cj1324
CSO-1678	GTACCCGGGTGACTAACTAGGGTGACTAACTAGGAGGA ATAAATG	Amplification of Hyg^R cassette
CSO-1679	TATTCCCTCCAGGTACTAAAACAGTCATATTCCCTCCAG GTATCA	Amplification of Hyg^R cassette
CSO-1815	GTTTTTATGCATCGATGCAATATTTTGAAAGGATT	*flaB* cloning in pXG-10
CSO-1816	GTTTTTGCTAGCACCTGAACTAAGTCTGCTTAAA	*flaB* cloning in pXG-10
CSO-1817	GTTTTTTTTAATACGACTCACTATAGGCGATGCAATATTT TGAAAGGATT	*in vitro* transcription of *flaB* leader, carries T7 promoter
CSO-1818	ACCTGAACTAAGTCTGCTTAAA	*in vitro* transcription of *flaB* leader
CSO-1819	TAGAAATTTCAAGGAAGAAATATGCATGGAAAAATAGCT ATTTATATGGATTCTACAGGACGTGGAACCG	Cj1249 cloning in pXG-10
CSO-1820	CTAGCGGTTCCACGTCCTGTAGAATCCATATAAATAGCT ATTTTTCCATGCATATTTCTTCCTTGAAATTTCTATGCA	Cj1249 cloning in pXG-10
CSO-1823	GTTTTTATGCATACTAGCAATAGGAAATTTTAAAAAG	*flaG* cloning in pXG-10
CSO-1824	GTTTTTGCTAGCCTGTGTCTCACTTGTTCTTTG	*flaG* cloning in pXG-10
CSO-1825	GTTTTTATGCATAAAAAACTTAAGCAAAGGAAGGC	*pseB* cloning in pXG-10
CSO-1826	GTTTTTGCTAGCTTCTAGCAAAACCTTAGTATAAGTT	*pseB* cloning in pXG-10
CSO-1895	[CY5] ATTGGTGTTAATACGAAATCCCATTT	FISH oligo 1 to detect *flaA* mRNA (Cy5-labeled 5' end)
CSO-1896	[CY5] AATTCAAGCTCATCCATGAACTTGT	FISH oligo 2 to detect *flaA* mRNA (Cy5-labeled 5' end)
CSO-1963	[CY5] CGTTTGCTTTTGCATTTAAAGCTG	FISH oligo 3 to detect *flaA* mRNA (Cy5-labeled 5' end)
CSO-1964	[CY5] CTTAAAGAAGCATCTAAACTTTTACTAT	FISH oligo 4 to detect *flaA* mRNA (Cy5-labeled 5' end)
CSO-2006	[FITC] GCTGCCTCCCGTAGGAGT	Universal FISH oligo to detect 16s rRNA (FITC- labeled 5' end)
CSO-2019	AAGGATTTAAAAAGGGATTTCGTATTAACACCAAT	Introduction of start $codon^{AUG>AAG}$ mutation into *flaA* 5'UTR
CSO-2020	ATACGAAATCCCTTTTTAAATCCTTTTAAAAAAATTCAAG C	Introduction of start $codon^{AUG>AAG}$ mutation into *flaA* 5'UTR
CSO-2023	[CY5] GAGCTAAGATATTTGCTTTAGAGTAG	FISH oligo 5 to detect *flaA* mRNA (Cy5-labeled 5' end)
CSO-2024	[CY5] TGCCATGGCATAAGAGCCGCT	FISH oligo 6 to detect *flaA* mRNA (Cy- labeled 5' end)
CSO-2150	GTGAGCAAGGGCGAGGA	Amplification of *mCherry* (2^{nd} codon to stop)
CSO-2151	TTACTTGTACAGCTCGTCCAT	Amplification of *mCherry* (2^{nd} codon to stop)
CSO-2155	CCTCCTCGCCCTTGCTCACTTTTTTAATATAATTAGCAAT TTGATCA	Overlap PCR construction of *fliW-mCherry* with *aac*(3)-IV cassette
CSO-2156	CCTCCTCGCCCTTGCTCACTTTGATTAGTTTTTTGCTTAA GTCAT	Overlap PCR construction of *csrA-mCherry* with *aphA -3* cassette
CSO-2746	GTAAAGCCACCCGCTCCTATG	NB oligo to detect Cj1321_mini
CSO-2809	[Cy5] CACGGATTTGCGATTCTGCTG	FISH oligo 7 to detect *flaA* mRNA (Cy5-labeled 5' end)
CSO-2810	[Cy5] GTGATGTTGTTTATAGTTGATGTAAC	FISH oligo 8 to detect *flaA* mRNA (Cy5-labeled 5' end)
CSO-2811	[Cy5] AACTAAGGCTCCATTAGCATCAC	FISH oligo 9 to detect *flaA* mRNA (Cy5 -labeled 5' end)
CSO-2812	[Cy5] GTAATCTACTTTACCGATTTTTACCC	FISH oligo 10 to detect *flaA* mRNA (Cy5-labeled 5' end)
CSO-2813	[Cy5] AGATCTTAAACTATCTGCTATCGC	FISH oligo 11 to detect *flaA* mRNA (Cy5-labeled 5' end)
CSO-2814	[Cy5] TAGATATAGCTTGACCTAAAGTATTAG	FISH oligo 12 to detect *flaA* mRNA (Cy5-labeled 5' end)
CSO-2815	[Cy5] TATTAGCATCAATTTGTCCTTTTGAC	FISH oligo 13 to detect *flaA* mRNA (Cy5 -labeled 5' end)
CSO-2816	[Cy5] GTAATCTTAAAACATTTTGTTGAACAGAA	FISH oligo 14 to detect *flaA* mRNA (Cy5-labeled 5' end)
CSO-2821	P-TTTGATAAGTTTATTTGGATACAATTGTGGTAACAAGT TCATGGATGAGCTT	Exchange of *flaA* promoter with *metK* promoter
CSO-2841	CTTCAGGTTCAGGTTATTCTG	Verification of *flaB* deletion in *C. jejuni*

CSO-2842	GGCTCAGGTTTTTCAAGTGG	Overlap PCR construction of *flaB* deletion with *aphA-3*
CSO-2843	TCCTAGTTAGTCACCCGGGTAATCCTAAAACCCATTTTA AATCCTT	Overlap PCR construction of *flaB* deletion with *aphA-3*
CSO-2844	ATTGTTTTAGTACCTGGAGGGAATACAGCAAAATGTTTT AAAACTTCTTC	Overlap PCR construction of *flaB* deletion with *aphA-3*
CSO-2853	GCTTGATAAATTAAAAATTTACTAAAATTAGGATCCTTTT ATGGATAATTTTTAAA	Exchange of *flaA* promoter with *metK* promoter
CSO-2825	AAAGGATTTAAAGTGGGATTTCGTATTAACACCAA	Introduction of start codon$^{AUG>GUG}$ mutation into *flaA* 5' UTR
CSO-2826	TACGAAATCCCACTTTAAATCCTTTTAAAAAAATTCAAGC	Introduction of start codon$^{AUG>GUG}$ mutation into *flaA* 5' UTR
CSO-2827	AGGATTTAAAATTGGATTTCGTATTAACACCAATG	Introduction of start codon$^{AUG>AUU}$ mutation into *flaA* 5' UTR
CSO-2828	AATACGAAATCCAATTTTAAATCCTTTTAAAAAAATTCAA GC	Introduction of start codon$^{AUG>AUU}$ mutation into *flaA* 5' UTR
CSO-2829	TTAAAATGGGATAGCGTATTAACACCAATGTTGCAG	Introduction of 3rd codon$^{UUU>UAG}$ mutation into *flaA* 5' UTR
CSO-2830	GGTGTTAATACGCTATCCCATTTTAAATCCTTTTAAAAAA AT	Introduction of 3rd codon$^{UUU>UAG}$ mutation into *flaA* 5' UTR
CSO-2831	TTAAAATGGGATTCCGTATTAACACCAATGTTGCAG	Introduction of 3rd codon$^{UUU>UUC}$ mutation into *flaA* 5' UTR
CSO-2832	GGTGTTAATACGGAATCCCATTTTAAATCCTTTTAAAAAA AT	Introduction of 3rd codon$^{UUU>UUC}$ mutation into *flaA* 5' UTR
CSO-2833	ACTCAAGCGGCTTAAGATGGACAAAGTTTAAAAACAAG	Introduction of 101st codon$^{CAA>UAA}$ mutation into *flaA* 5' UTR
CSO-2834	TTTGTCCATCTTAAGCCGCTTGAGTTGCCTTA	Introduction of 101st codon$^{CAA>UAA}$ mutation into *flaA* 5' UTR
CSO-2835	GCTGCAACATTGGTGTTAATACG	NB oligo to detect *flaA* mRNA in all 5'UTR point mutants and *flaA_mini*
HPK1	GTACCCGGGTGACTAACTAGG	Amplification of *aphA-3* cassette
HPK2	TATTCCCTCCAGGTACTAAAACA	Amplification of *aphA-3* cassette
JVO-0054	GGGATCAAGCCTGATTG	Sense to *aphA-3* for verification
JVO-0900	GGAGAAACAGTAGAGAGTTGC	Antisense oligo for inverse PCR on pBAD/*Myc*-His A
JVO-0901	TTTTTTCTAGATTAAATCAGAACGCAGA	Sense oligo for inverse PCR on pBAD /*Myc*-His A
JVO-5142	GACTACAAAGACCATGACGG	Sense oligo to 3xFLAG tag
pBAD-FW	ATGCCATAGCATTTTTATCC	Verification of insert in pBAD/*Myc*-His A plasmid
pZE-A	GTGCCACCTGACGTCTAAGA	Verification of insert in pJV752.1

Supplementary Table 5. Plasmids. List of all plasmids used in this study.

Name	Description/Generation	Origin/ Marker	Reference
pJV752.1	Cloning vector, pZE12-*luc* with modified p15A origin	p15A mod/ Amp^R	[5]
pUC1813 *apra*	Carries *aac*(*3*)-IV gene	pBR322/ Gm^R	[6]
pAC1H	Carries *aph*(7") gene	ColE1/pBR 322/Hyg^R	[7]
pBAD/Myc-His A	pBAD expression plasmid	pBR322/ Amp^R	Invitrogen
pZE12-luc	General expression plasmid	ColE1/ Amp^R	[8]
pXG-10	Standard plasmid for directional cloning of a target mRNA as N-terminal translational fusion to GFP	pSC101*/ Cm^R	[3]
pXG-30	Plasmid for cloning operon fusions with the N-terminus of downstream gene fused to GFP and the C-terminus of upstream gene fused to a short artificial reading frame composed of a FLAG epitope and truncated *lacZ* gene	pSC101*/ Cm^R	[3]
pGD68-1	pBAD::$CsrA_{Cj}$ based on pBAD/Myc-His A	pBR322/ Amp^R	This study
pGD72-3	pBAD::$CsrA_{Cj}$-Strep, based on pGD68-1	pBR322/ Amp^R	This study
pGG1	Plasmid (based on pZE12-luc) harbouring 3xFLAG and non-polar *aphA-3* cassette. Used for introduction of 'UP' and 'DN' regions of a gene of interest to be FLAG-tagged	ColE1/ Kan^R Amp^R	Sharma lab
pGD78-1	*aphA -3* ORF in pGG1 replaced by *aac*(*3*)*-IV* ORF	ColE1/ Gm^R Amp^R	This study
pGD4-1	Plasmid harbouring *csrA*-3xFLAG C-terminal translational fusion, *csrA* upstream and downstream regions, and *aphA-3* cassette in pGG1 for chromosomal epitope tagging at native locus	ColE1/ Kan^R Amp^R	This study
pMW5-2	Plasmid harbouring *flaA*-3xFLAG C-terminal translational fusion, *flaA* upstream and downstream regions, and *aphA-3* cassette in pGG1 for chromosomal epitope tagging at native locus	ColE1/ Kan^R Amp^R	This study
pMW6-1	Plasmid harbouring *flaB*-3xFLAG C-terminal translational fusion, *flaB* upstream and downstream regions, and *aphA-3* cassette in pGG1 for chromosomal epitope tagging at native locus	ColE1/ Kan^R Amp^R	This study
pSSv1-2	Plasmid harbouring *flgI*-3xFLAG C-terminal translational fusion, *flgI* upstream and downstream regions, and *aphA-3* cassette in pGG1 for chromosomal epitope tagging at native locus	ColE1/ Kan^R Amp^R	This study
pGD70-5	Plasmid harbouring 1,100 bp region around *flaA* promoter; based on pJV752.1	p15A mod/ Amp^R	This study
pGD76-1	*aac*(3)-IV gentamicin cassette introduced upstream of *flaA* promoter in pGD70-5 in reverse orientation to *flaA*	p15A mod/ Gm^R Amp^R	This study
pGD92-1	M1 (SL1 GGA>AAA) mutation in *flaA* 5'UTR in pGD76-1	p15A mod/ Gm^R Amp^R	This study
pGD77-1	M2 (SL1 GGA>UGA) mutation in *flaA* 5'UTR in pGD76-1	p15A mod/ Gm^R Amp^R	This study
pGD93-1	M3 (SL2 GGA>GGG) mutation in *flaA* 5'UTR in pGD76-1	p15A mod/ Gm^R Amp^R	This study
pGD95-1	M2 (SL1 GGA>UGA) /M3 (SL2 GGA>GGG) mutation in *flaA* 5'UTR in pGD76-1	p15A mod/ Gm^R Amp^R	This study
pGD114-2	Start codon mutation in *flaA* (AUG→AAG) in pGD76-1	p15A mod/ Gm^R Amp^R	This study
pGD205-1	Start codon mutation in *flaA* (AUG→AUU) in pGD76-1	p15A mod/ Gm^R Amp^R	This study
pGD204-1	Start codon mutation in *flaA* (AUG→GUG) in pGD76-1	p15A mod/ Gm^R Amp^R	This study
pGD206-1	3rd codon mutation in *flaA* (AUG→UAG) in pGD76-1	p15A mod/ Gm^R Amp^R	This study
pGD207-1	3rd codon mutation in *flaA* (AUG→UUC) in pGD76-1	p15A mod/ Gm^R Amp^R	This study
pGD208-1	101st codon mutation in *flaA* (CAA→UAA) in pGD76-1	p15A mod/ Gm^R Amp^R	This study
pGD209-1	*flaA* promoter replaced by *metK* promoter in pGD76-1	p15A mod/	This study

		Gm^R Amp^R	
pGD107-1	*aac(3)*-IV gentamicin cassette replaced by *cat* cassette in pGD92-1	p15A mod/ Cm^R Amp^R	This study
pGD31-1	5'UTR along with first 33 codons of *flaA* fused in-frame to *gfp* in pXG-10	pSC101*/ Cm^R	This study
pGD111-1	5'UTR along with first 25 codons of *flaG* fused in-frame to *gfp* in pXG-10	pSC101*/ Cm^R	This study
pGD38-1	5'UTR along with first 35 codons of *flgI* fused in-frame to *gfp* in pXG-10	pSC101*/ Cm^R	This study
pGD109-1	5'UTR along with first 35 codons of *flaB* fused in-frame to *gfp* in pXG-10	pSC101*/ Cm^R	This study
pGD112-1	5'UTR along with first 26 codons of *pseB* fused in-frame to *gfp* in pXG-10	pSC101*/ Cm^R	This study
pGD110-1	5'UTR along with first 16 codons of Cj1249 fused in-frame to *gfp* in pXG-10	pSC101*/ Cm^R	This study
pGD28-1	Last 17 codons of Cj0310c and first 23 codons of Cj0309c fused in-frame to FLAG-*lacZ* and *gfp,* respectively, in pXG-30	pSC101*/ Cm^R	This study
pGD27-1	Last 17 codons of Cj0805 and first 25 codons of *dapA* fused in-frame to FLAG-*lacZ* and *gfp,* respectively, in pXG-30	pSC101*/ Cm^R	This study

Supplementary Table 6. Construction of *C. jejuni* mutants.

Mutation	'UP' PCR primers	'DN' PCR primers	Cassette/ Primers	UP-cassette-DN amplification primers	Mutant validation primers
Single gene deletions (overlap PCR)					
Δ*csrA* (Cm^R)	CSO-0394 CSO-0615	CSO-0616 CSO-0395	*cat* (CSO-0613/-0614)	CSO-0394 CSO-0395	CSO-0392 CSO-0614
Δ*csrA* (Hyg^R)	CSO-0393 CSO-0394	CSO-0395 CSO-0396	*aph(7'')* (CSO-1678/-1679)	CSO-0394 CSO-0395	CSO-0392 HPK2
Δ*fliW* (Gm^R)	CSO-0997 CSO-0998	CSO-0999 CSO-1000	*aac*(3)-IV (HPK1/HPK2)	CSO-0997 CSO-1000	CSO-1072 HPK2
Δ*fliW* (Hyg^R)	CSO-0997 CSO-0998	CSO-0999 CSO-1000	*aph(7'')* (CSO-1678/-1679)	CSO-0997 CSO-1000	CSO-1072 HPK2
Δ*flaA* (Kan^R)	CSO-0752 CSO-1548	CSO-1549 CSO-1550	*aphA-3* (HPK1/HPK2)	CSO-0999 CSO-1000	CSO-0756 CSO-0023
Δ*flaB* (Kan^R)	CSO-2842 CSO-2843	CSO-2844 CSO-0560	*aphA-3* (HPK1/HPK2)	CSO-0999 CSO-1000	CSO-2841 CSO-0023
Δ*flaAB* (Kan^R)	CSO-0752 CSO-1548	CSO-1549 CSO-1550	*aphA-3* (HPK1/HPK2)	CSO-0999 CSO-1000	CSO-0756 CSO-0023
Δ*rpoN* (Gm^R)	CSO-1144 CSO-1145	CSO-1146 CSO-1147	*aac*(3)-IV (HPK1/HPK2)	CSO-1144 CSO-1147	CSO-1148 HPK2
Δ*fliA* (Gm^R)	CSO-1149 CSO-1150	CSO-1151 CSO-1152	*aac*(3)-IV (HPK1/HPK2)	CSO-1149 CSO-1152	CSO-1153 HPK2
3xFLAG tags (cloned in pGG1)					
***csrA*-3xFLAG** (NCTC11168 and 81-176) (pGD4-1)	CSO-0171 CSO-0172	CSO-0173 CSO-0174	*aphA-3* (from pGG1)	CSO-0172 CSO-0173	CSO-0023 CSO-0196
***flaA*-3xFLAG** (pMW5.2)	CSO-0553 CSO-0554	CSO-0555 CSO-0556	*aphA-3* (from pGG1)	CSO-0554 CSO-0555	CSO-0023 CSO-0557
***flgI*-3xFLAG** (pSSv1.2)	CSO-1011 CSO-1012	CSO-1013 CSO-1014	*aphA-3* (from pGG1)	CSO-1012 CSO-1013	CSO-1015 HPK2
***flaB*-3xFLAG** (pMW6.1)	CSO-0558 CSO-0559	CSO-0560 CSO-0561	*aphA-3* (from pGG1)	CSO-0559 CSO-0560	CSO-0562 HPK2
3xFLAG tags (overlap PCR)					
***flaG*-3xFLAG**	CSO-1002 CSO-1098	CSO-1099 CSO-1003	*aphA-3* (HPK1/HPK2)	CSO-1002 CSO-1003	CSO-1005 HPK2
Cj0529-3xFLAG	CSO-1408 CSO-1409	CSO-1410 CSO-1411	*aphA-3* (HPK1/HPK2)	CSO-1409 CSO-1410	CSO-1407 CSO-0023
For construction of *flaA* 5'UTR point mutations refer to Methods.					

Supplementary Table 7. GFP fusions for validating CsrA-target interactions in *E. coli*.

Plasmid	*C. jejuni* target gene(s)	Primers used for target amplification	Plasmid backbone	Colony PCR	Description
pGD31-1	*flaA*	CSO-0621 CSO-0622	pXG-10	CSO-0621 CSO-0155	5'UTR along with first 33 codons of *flaA* fused in-frame to *gfp*
pGD111-1	*flaG*	CSO-1823 CSO-1824	pXG-10	CSO-1823 CSO-0155	5'UTR along with first 25 codons of *flaG* fused in-frame to *gfp*
pGD38-1	*flgI*	CSO-0694 CSO-0695	pXG-10	CSO-0694 CSO-0155	5'UTR along with first 35 codons of *flgI* fused in-frame to *gfp*
pGD109-1	*flaB*	CSO-1815 CSO-1816	pXG-10	CSO-1815 CSO-0155	5'UTR along with first 35 codons of *flaB* fused in-frame to *gfp*
pGD112-1	*pseB*	CSO-1825 CSO-1826	pXG-10	CSO-1825 CSO-0155	5'UTR along with first 26 codons of *pseB* fused in-frame to *gfp*
pGD110-1	Cj1249	CSO-1819* CSO-1820*	pXG-10	CSO-1819 CSO-0155	5'UTR along with first 16 codons of Cj1249 fused in-frame to *gfp*
pGD28-1	Cj0310c-Cj0309c	CSO-0608 CSO-0609	pXG-30	CSO-0608 CSO-0155	Last 17 codons of Cj0310c and first 23 codons of Cj0309c were fused in-frame to FLAG-*lacZ* and *gfp*, respectively.
pGD27-1	Cj0805-*dapA*	CSO-0606 CSO-0607	pXG-30	CSO-0606 CSO-0155	Last 17 codons of Cj0805 and first 25 codons of *dapA* were fused in-frame to FLAG-*lacZ* and *gfp*, respectively

Sequencing on plasmids was performed using oligonucleotide CSO-0155 which binds antisense to *gfp*.

*CSO-1819/-1820 were annealed together to yield the insert (without PCR) for direct introduction into pXG-10.

Supplementary Table 8. Details of RNAs used for *in vitro* work. WT GGA motifs are marked in blue and introduced point mutations in the leader variants are marked in red. Start (ATG) and stop (TAA) codons are underlined.

Name	DNA template (plasmid or gDNA)	Primers	Size of T7-transcript [nt]	Sequence (5' → 3')
***flaA* WT leader**	gDNA NCTC11168	CSO-0709 CSO-0710	144	UAACAAGUUCAU**GGA**UGAGCUUGAAUUUUUUUAAAA**GGA**UUUAAA**A** **UG**GGAUUUCGUAUUAACACCAAUGUUGCAGCUUUAAAUGCAAAAGC AAACGCUGAUUUAAAUAGUAAAAGUUUAGAUGCUUCUUUAAGCAGAC UUAGU
***flaA* M1 leader**	pGD92-1	CSO-1656 CSO-0710	144	UAACAAGUUCAU**AAA**UGAGCUUGAAUUUUUUUAAAA**GGA**UUUAAA**A** **UG**GGAUUUCGUAUUAACACCAAUGUUGCAGCUUUAAAUGCAAAAGC AAACGCUGAUUUAAAUAGUAAAAGUUUAGAUGCUUCUUUAAGCAGAC UUAGU
***flaA* M2 leader**	pGD77-1	CSO-1081 CSO-0710	144	UAACAAGUUCAU**UGA**UGAGCUUGAAUUUUUUUAAAA**GGA**UUUAAA**A** **UG**GGAUUUCGUAUUAACACCAAUGUUGCAGCUUUAAAUGCAAAAGC AAACGCUGAUUUAAAUAGUAAAAGUUUAGAUGCUUCUUUAAGCAGAC UUAGU
***flaA* M3 leader**	pGD93-1	CSO-0709 CSO-0710	144	UAACAAGUUCAU**GGA**UGAGCUUGAAUUUUUUUAAAA**GGG**UUUAAA**A** **UG**GGAUUUCGUAUUAACACCAAUGUUGCAGCUUUAAAUGCAAAAGC AAACGCUGAUUUAAAUAGUAAAAGUUUAGAUGCUUCUUUAAGCAGAC UUAGU
***flaA* M2/M3 leader**	pGD95-1	CSO-1081 CSO-0710	144	UAACAAGUUCAU**UGA**UGAGCUUGAAUUUUUUUAAAA**GGG**UUUAAA**A** **UG**GGAUUUCGUAUUAACACCAAUGUUGCAGCUUUAAAUGCAAAAGC AAACGCUGAUUUAAAUAGUAAAAGUUUAGAUGCUUCUUUAAGCAGAC UUAGU
***flgI* leader**	gDNA NCTC11168	CSO-0713 CSO-0714	128	ACAAUAGAUUAAAGGAAGAAUCC**AUG**AGAGUUUUAACGAUAUUUUUA CUCUUUAUGACAAGCAUUUUUGCAGUGCAAAUCAAGGAUGUAGCAA AUACUGUAGGUGUAAGAGAUAACCAACUUAUAGGU
***flaG* leader**	gDNA NCTC11168	CSO-1084 CSO-1085	108	ACUAGCAAUAGGAAAUUUUAAAAAGGAUUUUAAA**AUG**GAAAUAUCGA AGGCAAAUGGGCAAAUGGAUACAGCUUUGGCAAACAUUAGCCAAAG AACAAGUGAGACACA
***flaB* leader**	gDNA NCTC11168	CSO-1817 CSO-1818	132	CGAUGCAAUAUUUUGAAAGGAUUUAAA**AUG**GGUUUUAGGAUAAACA CCAACAUCGGUGCAUUAAAUGCACAUGCAAAUUCAGUUGUUAAUGC UAGAGAACUGGAUAAGUCUUUAAGCAGACUUAGUUCAGGU
Cj0040 leader	gDNA NCTC11168	CSO-1082 CSO-1083	117	AAAUUAAAAUUUUAAAAAGGAAGUUAAA**AUG**UCAAAACCAUUAAAUG AAGAGAUUUUUGUUGAAUUUAAAAGUGAUCUAGCUGAAAGAAAAAAU GAAGUUUUGCUUCAAGUUUUAGA
***flgA* 3'end**	gDNA NCTC11168	CSO-1088 CSO-1089	113	CAAAUGUUGAUGUUUUAAUCGAACUUGUGGCUUUGCAAAGUGCAAA UAUGGGCGAAAGGAUUCGUGCAAAAAACAAAGAAGGUAAAGUUAUG CAAGGUAUCGUAGUGGGUAAA
***flgM* leader**	gDNA NCTC11168	CSO-1092 CSO-1093	97	AGGAUUAUAACUAAGAUCAAGGAGGCAGAA**AUG**AUCAAUCCUAUACA ACAAAGUUAUGUGGCAAAUACCGCAUUAAAUACAAAUAGAAUAGAUA AAG
Cj1324 leader	gDNA NCTC11168	CSO-1665 CSO-1666	100	AUUUUUAUUAAAUUGAAGGGGUGGGGA**AUG**AUUUAUUGUGAUCACU GCGUGAUGCCAAAUACUAGACCUGGUAUUAAUUUUACAAAAGAUAAA GAAGGUA
***hopB* 3'end with UTR**	gDNA *H. pylori* G27	CSO-0701 CSO-0702	107	AAAGCUGGUGGCGCUGAAGUGAAAUACUUCCGCCCUUAUAGCGUGU AUUGGGUCUAUGGCUACGCCUUC**UAA**AAAAGCUCAAGGCCUUUUAU AGGCUUUGAUUUAAC

Supplementary Methods

Transformation of *C. jejuni* for mutant construction. Transformation of *C. jejuni* was performed by electroporation or natural transformation as described previously[2, 9, 10, 11]. For electroporation, strains grown from frozen stocks until passage one or two on MH agar were harvested into cold electroporation buffer (272 mM sucrose, 15% v/v glycerol) and washed twice with the same buffer. Cells (50 µl) were mixed with 200-400 ng PCR product on ice and electroporated (Biorad MicroPulser) in a 1 mm gap cuvette (PEQLAB) at 2.5 kV. Cells were then transferred with Brucella broth to a non-selective MH plate and recovered overnight at 37 °C microaerobically before plating on the appropriate selective medium.

In some cases, *C. jejuni* double or triple mutants were constructed by natural transformation of the genomic DNA from the appropriate donor strain[10, 12]. Genomic DNA was extracted from the donor strain by phenol-chloroform extraction and ethanol precipitation. Specifically, bacteria were harvested from one-day-old selective MH plates into SET buffer (150 mM NaCl, 15 mM EDTA, 10 mM Tris-HCl; pH 8.0), collected by centrifugation, and resuspended in SET buffer. SDS and proteinase K were then added to final concentrations of 0.5% (w/v) and 100 µg/ml, respectively, and suspensions were incubated at 55 ˚C for 2h. Protein was then removed by extraction with an equal volume of phenol-chloroform-isoamyl alcohol (25:24:1), separation of phases by centrifugation at 13,000 rpm for 8 min, and re-extraction of the aqueous phase with an equal volume of chloroform with centrifugation at 13,000 rpm to separate phases. DNA was then precipitated from the final aqueous phase with 1/10 vol. 3M sodium acetate, pH 5.3 and 2 vol. absolute ethanol. After overnight incubation at -20 °C, precipitated DNA was collected by centrifugation at 13,000 rpm for 10 min and washed once with 75% cold ethanol. DNA pellets were resuspended in 100 µl water with shaking at 65 °C. For transformations, recipient strains were grown from frozen stocks, patched into small circles on a non-selective MH plate, and grown for 2-3 h at 37 °C under microaerobic conditions. One hundred ng of donor genomic DNA (gDNA) was then added to the patches and plates were incubated for an additional 4-5h. Patched cells were then harvested into 1 ml Brucella broth and 10 or 100 µl was plated on the appropriate selective MH agar. Colonies were re-streaked onto selective plates, and colony PCR was performed to confirm presence of desired mutations from both donor and recipient strain.

Construction of *C. jejuni* deletion strains by overlap PCR. All *C. jejuni* deletion mutant strains listed in Supplementary Table 3 were generated by double-crossover homologous recombination with PCR products of deletion cassettes that were constructed by overlap PCR (for details see Supplementary Table 6) and electroporated into bacteria as described above.

PCR products carried *aphA-3* kanamycin[13], *C. coli cat* chloramphenicol[14], *aac*(3)-IV gentamicin[6], or *aph*(7″) hygromycin[7] resistance cassettes flanked by ~500 bp of homologous sequence up- and downstream of the coding region of the target gene. Non-polar resistance cassettes were amplified from plasmids that carry the resistance markers using primers HPK1/HPK2 (Kan^R), CSO-1678/-1679 (Hyg^R), or CSO-0613/-0614 (Cm^R). The *aphA-3* (Kan^R) ORF was replaced by the *aac*(3)-IV (Gm^R) ORF (amplified using CSO-0575/-0832 and *Nde*I digested) in the plasmid pGG1 (amplified using CSO-0577/-0831 and *Nde*I digested) leaving the HPK1/HPK2 binding sites intact. The resulting plasmid pGD78-1 was used to amplify the *aac*(3)-IV (Gm^R) cassette using the same HPK1/HPK2 primers.

As an example, the construction of the chloramphenicol resistant *C. jejuni* NCTC11168 Δ*csrA*::*cat* deletion mutant is described. About 500 bp upstream of the *csrA* (Cj1103) start codon was amplified from genomic DNA (gDNA) of *C. jejuni* NCTC11168 WT using 'UP' primers (CSO-0394/-0615). Likewise, ~500bp downstream of the *csrA* stop codon was amplified using 'DN' primers (CSO-0616/-0395). The 5' ends of the antisense-UP primer and sense-DN primer contained ~25 bp of sequence homologous to the sense or antisense primer (CSO-0613/-0614), respectively, used to amplify the *cat* resistance cassette. PCR products were purified (Macherey-Nagel NucleoSpin PCR cleanup kit), and UP, DN, and resistance cassette amplicons were then added together in a ratio of 50:50:90 ng to a 100 µl Phusion polymerase PCR reaction with sense-UP and antisense-DN primers (CSO-0394/-0395) at a final concentration of 0.06 µM. Overlap PCR was performed with the following conditions: 1 cycle of [98 °C, 3 min; 61 °C, 1 min; 72 °C, 10 min; 98 °C, 1 min], 40 cycles of [98 °C, 15 s; 57 °C, 20 s; 72 °C, 1 min], followed by a 10 min final extension at 72 °C. Following verification of product size by agarose gel electrophoresis and purification (Macherey-Nagel NucleoSpin PCR cleanup kit), the resulting overlap PCR product was electroporated into the appropriate recipient *C. jejuni* strain. Deletion mutants for *flaA*::Kan^R, *flaB*::Kan^R, *flaAB*::Kan^R, *fliW*::Gm^R, *fliW*::Hyg^R *fliA*::Gm^R, *rpoN*::Gm^R, and *csrA*::Hyg^R were constructed similarly (see Supplementary Table 6).

Sequence and structure conservation of the *flaA* 5'UTR. In order to identify homologous *flaA* 5'UTR regions in different *Campylobacter* species and strains we ran nucleotide blast (blastn[15]) with parameters optimized for more dissimilar sequences (discontiguous megablast) using both the NCBI nucleotide collection (nr/nt) and the NCBI whole genome shotgun (wgs) contigs as databases. As query we used a 130 nt-long sequence encompassing the 100 nt upstream and the first 30 nt of the *C. jejuni* NCTC11168 *flaA* coding region. This includes the *flaA* promoter region, its 5'UTR, and the beginning of the coding sequence. Based on all hits (~200) we extracted the 130 nt from the target sequences. If BLAST hits with an optimal score were truncated they were extended on either side to obtain 130 nt. We excluded all sequences with undefined bases or

without proper species/strain association. Afterwards, sequences for *C. fetus* subsp. *fetus* 82-40 and *C. concisus* 13826, which were not found by BLAST, were added manually to the set. The sequences were aligned using MUSCLE[16] with default parameters and a conserved Sigma28 (FliA) -10 box (CGATAT) was observed in all of them. The alignment was trimmed to the region including only the *flaA* 5'UTR (based on the region according to the 5' UTR of *C. jejuni* NCTC11168) and the first 10 nt of the coding sequence. More dissimilar sequences disrupting the alignment (*C. peloridis* LMG 23910, *C. lari* NCTC11845, *C. lari* NCTC12892, *C. lari* RM16701, *C. lari* CCUG 22395, *C. lari* RM16712, *C. curvus* 525.92, *Campylobacter* sp. FOBRC14 ctg120009214739, *C. curvus* DSM 6644 C514DRAFT scaffold00004.4_C, and *C. concisus* 13826) were removed. Based on the resulting alignment all identical sequences were collapsed keeping only one representative sequence per cluster. Subsequently, a consensus structure was predicted using RNAalifold[17] with RIBOSUM scoring and default values for all other parameters. The resulting structure-annotated sequence alignment is shown in Supplementary Fig. 6 and the full alignment including additional strains is shown in Supplementary Fig. 7.

Supplementary References

1. Gundogdu O, Bentley SD, Holden MT, Parkhill J, Dorrell N, Wren BW. Re-annotation and re-analysis of the Campylobacter jejuni NCTC11168 genome sequence. *BMC Genomics* 2007, **8:** 162.

2. Dugar G, Herbig A, Forstner KU, Heidrich N, Reinhardt R, Nieselt K*, et al.* High-Resolution Transcriptome Maps Reveal Strain-Specific Regulatory Features of Multiple Campylobacter jejuni Isolates. *PLoS Genet* 2013, **9**(5)**:** e1003495.

3. Urban JH, Vogel J. Translational control and target recognition by Escherichia coli small RNAs in vivo. *Nucleic Acids Res* 2007, **35**(3)**:** 1018-1037.

4. Zhang B, Rapolu M, Liang Z, Han Z, Williams PG, Su WW. A dual-intein autoprocessing domain that directs synchronized protein co-expression in both prokaryotes and eukaryotes. *Scientific reports* 2015, **5:** 8541.

5. Sharma CM, Darfeuille F, Plantinga TH, Vogel J. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes Dev* 2007, **21**(21)**:** 2804-2817.

6. Bury-Mone S, Skouloubris S, Dauga C, Thiberge JM, Dailidiene D, Berg DE*, et al.* Presence of active aliphatic amidases in Helicobacter species able to colonize the stomach. *Infect Immun* 2003, **71**(10)**:** 5613-5622.

7. Cameron A, Gaynor EC. Hygromycin B and apramycin antibiotic resistance cassettes for use in Campylobacter jejuni. *PLoS One* 2014, **9**(4)**:** e95084.

8. Lutz R, Bujard H. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res* 1997, **25**(6)**:** 1203-1210.

9. Hansen CR, Khatiwara A, Ziprin R, Kwon YM. Rapid construction of Campylobacter jejuni deletion mutants. *Lett Appl Microbiol* 2007, **45**(6)**:** 599-603.

10. McLennan MK, Ringoir DD, Frirdich E, Svensson SL, Wells DH, Jarrell H*, et al.* Campylobacter jejuni biofilms up-regulated in the absence of the stringent response utilize a calcofluor white-reactive polysaccharide. *J Bacteriol* 2008, **190**(3)**:** 1097-1107.

11. Miller JF, Dower WJ, Tompkins LS. High-voltage electroporation of bacteria: genetic transformation of Campylobacter jejuni with plasmid DNA. *Proc Natl Acad Sci U S A* 1988, **85**(3)**:** 856-860.

12. Wassenaar TM, Fry BN, van der Zeijst BA. Genetic manipulation of Campylobacter: evaluation of natural transformation and electro-transformation. *Gene* 1993, **132**(1)**:** 131-135.

13. Skouloubris S, Thiberge JM, Labigne A, De Reuse H. The Helicobacter pylori UreI protein is not involved in urease activity but is essential for bacterial survival in vivo. *Infect Immun* 1998, **66**(9)**:** 4517-4521.

14. Boneca IG, Ecobichon C, Chaput C, Mathieu A, Guadagnini S, Prevost MC*, et al.* Development of inducible systems to engineer conditional mutants of essential genes of Helicobacter pylori. *Appl Environ Microbiol* 2008, **74**(7)**:** 2095-2102.

15. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990, **215**(3)**:** 403-410.

16. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004, **32**(5)**:** 1792-1797.

17. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* 2008, **9:** 474.

3.6 GLOBAL RNA RECOGNITION PATTERNS OF POST-TRANSCRIPTIONAL REGULATORS HFQ AND CSRA REVEALED BY UV CROSSLINKING IN VIVO

Published online: April 4, 2016

Resource

SOURCE DATA | TRANSPARENT PROCESS | OPEN ACCESS

THE EMBO JOURNAL

Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking *in vivo*

Erik Holmqvist[1], Patrick R Wright[2], Lei Li[1], Thorsten Bischler[1], Lars Barquist[1], Richard Reinhardt[3], Rolf Backofen[2,4,*] & Jörg Vogel[1,**]

Abstract

The molecular roles of many RNA-binding proteins in bacterial post-transcriptional gene regulation are not well understood. Approaches combining *in vivo* UV crosslinking with RNA deep sequencing (CLIP-seq) have begun to revolutionize the transcriptome-wide mapping of eukaryotic RNA-binding protein target sites. We have applied CLIP-seq to chart the target landscape of two major bacterial post-transcriptional regulators, Hfq and CsrA, in the model pathogen *Salmonella* Typhimurium. By detecting binding sites at single-nucleotide resolution, we identify RNA preferences and structural constraints of Hfq and CsrA during their interactions with hundreds of cellular transcripts. This reveals 3′-located Rho-independent terminators as a universal motif involved in Hfq–RNA interactions. Additionally, Hfq preferentially binds 5′ to sRNA-target sites in mRNAs, and 3′ to seed sequences in sRNAs, reflecting a simple logic in how Hfq facilitates sRNA–mRNA interactions. Importantly, global knowledge of Hfq sites significantly improves sRNA-target predictions. CsrA binds AUGGA sequences in apical loops and targets many *Salmonella* virulence mRNAs. Overall, our generic CLIP-seq approach will bring new insights into post-transcriptional gene regulation by RNA-binding proteins in diverse bacterial species.

Keywords CLIP; CsrA; Hfq; non-coding RNA; peak calling; post-transcriptional control; small RNA; terminator; translation

Subject Categories Methods & Resources; Microbiology, Virology & Host Pathogen Interaction; RNA Biology

DOI 10.15252/embj.201593360 | Received 9 November 2015 | Revised 25 February 2016 | Accepted 26 February 2016 | Published online 4 April 2016

The EMBO Journal (2016) 35: 991–1011

Introduction

The fate of RNA molecules in the cell is largely determined at the post-transcriptional level by RNA–protein interactions. RNA-binding proteins (RBPs) are responsible for essential traits such as RNA stability, structure, translatability, export, and localization. Recent screens in human cells have suggested that the number of proteins with RNA-binding properties may be vastly underestimated (Baltz *et al*, 2012; Castello *et al*, 2012; Kramer *et al*, 2014), prompting new systematic searches for RBPs in many eukaryotic systems (Ascano *et al*, 2013). By comparison, our knowledge of the scope and binding preferences of prokaryotic RBPs is lagging behind eukaryotic systems, and new approaches are needed to fully elucidate the roles of RBPs in post-transcriptional control in bacterial pathogens (Barquist & Vogel, 2015). That is, although the structural details of the interactions of many positively and negatively acting proteins with DNA have been established, the paucity of understanding regarding RBPs has been holding back the field of bacterial gene regulation.

Salmonella enterica serovar Typhimurium is a widely studied food-borne bacterial pathogen that invades and replicates in many different eukaryotic host cells. Over the past decade, *Salmonella* has become a bacterial model organism to study post-transcriptional regulation by small regulatory RNAs (sRNAs) and two associated RBPs, Hfq and CsrA (Vogel, 2009; Hébrard *et al*, 2012; Westermann *et al*, 2016). Transcriptomic and RNA co-immunoprecipitation (coIP) analyses have suggested that Hfq and CsrA play global roles in the regulation of *Salmonella* virulence genes (Lawhon *et al*, 2003; Sittka *et al*, 2008; Ansong *et al*, 2009), but precisely how and where these proteins bind cellular transcripts *in vivo* remains to be fully understood.

Hfq is a widely conserved bacterial RBP of the Sm family of proteins which have a ring-like multimeric quaternary structure (Wilusz & Wilusz, 2005). In the Gram-negative bacteria *Salmonella* and *Escherichia coli*, coIP studies have predicted interactions of Hfq

1 Institute for Molecular Infection Biology, University of Würzburg, Würzburg, Germany
2 Bioinformatics Group, Department of Computer Science, Albert Ludwig University Freiburg, Freiburg, Germany
3 Max Planck Genome Centre Cologne, Max Planck Institute for Plant Breeding Research, Cologne, Germany
4 BIOSS Centre for Biological Signaling Studies, University of Freiburg, Freiburg, Germany
*Corresponding author. Tel: +49 761 203 7460; E-mail: backofen@informatik.uni-freiburg.de
**Corresponding author. Tel: +49 931 318 2576; E-mail: joerg.vogel@uni-wuerzburg.de

Published online: April 4, 2016
The EMBO Journal

with hundreds of sRNAs and an excess of one thousand mRNAs (Chao *et al*, 2012; Zhang *et al*, 2013; Bilusic *et al*, 2014). By helping sRNAs to regulate target mRNAs, Hfq modulates a variety of physiological traits including phosphosugar detoxification (Rice *et al*, 2012; Papenfort *et al*, 2013), catabolite repression (Beisel *et al*, 2012), envelope stress (Figueroa-Bossi *et al*, 2006; Gogol *et al*, 2011; Guo *et al*, 2014; Chao & Vogel, 2016), metal homeostasis (Desnoyers & Masse, 2012; Coornaert *et al*, 2013), biofilm formation (Holmqvist *et al*, 2010; Jørgensen *et al*, 2012; Mika *et al*, 2012; Thomason *et al*, 2012), motility (De Lay & Gottesman, 2012), and virulence (Sittka *et al*, 2007; Koo *et al*, 2011; Westermann *et al*, 2016). In pathogenic *Vibrio* species, Hfq and sRNAs regulate similarly complex traits, for example, quorum sensing or biofilm formation (Feng *et al*, 2015; Papenfort *et al*, 2015).

Mechanistically, Hfq promotes sRNA–mRNA annealing by increasing the rate of duplex formation (Møller *et al*, 2002; Zhang *et al*, 2002; Lease & Woodson, 2004; Link *et al*, 2009; Fender *et al*, 2010), while at the same time protecting sRNAs from the activity of cellular ribonucleases (Vogel & Luisi, 2011). In addition, Hfq may recruit auxiliary protein factors such as RNase E to promote the decay of target mRNAs (Morita & Aiba, 2011; Bandyra *et al*, 2012).

Structural studies of *Salmonella* Hfq confirmed the homohexameric ring model (Sauer & Weichenrieder, 2011). The two faces of the ring, denoted proximal and distal, both bind RNA, but show affinity for different RNA sequences: the proximal face tends to target single-stranded U-rich sequences, whereas the distal face interacts with single-stranded A-rich sequences (Schumacher *et al*, 2002; Mikulecky *et al*, 2004; Link *et al*, 2009). More recently, the rim of the Hfq hexamer has emerged as a third RNA-binding surface which interacts with UA-rich RNA and promotes intermolecular RNA annealing (Updegrove & Wartell, 2011; Sauer *et al*, 2012; Panja *et al*, 2013; Dimastrogiovanni *et al*, 2014). Whereas most of these findings stem from studying Hfq interactions with selected model substrates *in vitro*, details of transcriptome-wide Hfq binding within RNA *in vivo* emerged only recently through a crosslinking-based study in pathogenic *E. coli* (Tree *et al*, 2014). However, while this study captured many known Hfq targets, it generally failed to observe Hfq binding to sRNA 3′ ends, thus contrasting with the emerging mechanistic model from recent biochemical and structural studies whereby Hfq is loaded onto sRNAs via their 3′ located poly(U) stretch (Otaka *et al*, 2011; Sauer & Weichenrieder, 2011; Ishikawa *et al*, 2012; Dimastrogiovanni *et al*, 2014).

CsrA, initially identified as a regulator of carbon storage and glycogen biosynthesis in *E. coli* (Romeo *et al*, 1993), belongs to the large CsrA/Rsm family of RBPs that influence physiology and virulence in numerous pathogenic and non-pathogenic bacteria (Lenz *et al*, 2005; Brencic & Lory, 2009; Heroven *et al*, 2012; Romeo *et al*, 2013; Vakulskas *et al*, 2015). CsrA/Rsm proteins primarily affect translation of mRNAs by binding to 5′ untranslated regions (UTRs). A wealth of genetic, biochemical, and structural data shows that these proteins generally recognize GGA motifs in apical loops of RNA secondary structures (Dubey *et al*, 2005; Duss *et al*, 2014a). Other reported mechanisms of CsrA activity in the cell include promotion of Rho-dependent transcription termination, or mRNA stabilization by masking of RNase E cleavage sites (Yakhnin *et al*, 2013; Figueroa-Bossi *et al*, 2014). CsrA may also govern a large post-transcriptional regulon, as inferred from transcriptomic and RNA co-purification data in *Salmonella* and *E. coli*, respectively (Lawhon *et al*, 2003; Edwards *et al*, 2011).

The CsrA/Rsm proteins are themselves regulated by sRNAs such as CsrB and RsmZ, which contain multiple GGA sites that titrate the protein away from mRNA targets (Liu *et al*, 1997; Weilbacher *et al*, 2003; Valverde *et al*, 2004). Structural studies of one CsrA-like protein revealed a sequential and cooperative assembly of the protein on antagonistic sRNAs (Duss *et al*, 2014b). Antagonists of CsrA activity also include the Hfq-dependent sRNA McaS in *E. coli* (Holmqvist & Vogel, 2013; Jørgensen *et al*, 2013) and a sponge-like mRNA in *Salmonella* (Sterzenbach *et al*, 2013). Again, despite the strong interest in these proteins, the global binding preferences of CsrA/Rsm *in vivo* remain unknown.

Approaches combining *in vivo* crosslinking and RNA deep sequencing have been increasingly used to globally map the cellular RNA ligands and binding sites of eukaryotic RBPs *in vivo* (Darnell, 2010; König *et al*, 2011; Ascano *et al*, 2012). Such methods are now widely used in cell culture, tissues, and even whole animals. The purification of RNA–protein complexes after *in vivo* crosslinking by ultraviolet (UV) light offers several advantages over traditional coIP. Firstly, the UV-induced covalent bonds between protein and RNA survive denaturing conditions, facilitating stringent purification protocols. Secondly, crosslinking enables trimming by ribonucleases to yield protein-protected RNA fragments, pinpointing binding regions with unprecedented resolution. Thirdly, the attachment of a crosslinked peptide to a purified RNA fragment often causes mutations during reverse transcription which identify direct RNA–protein contacts at single-nucleotide resolution (Zhang & Darnell, 2011).

Here, we have employed UV crosslinking of RNA–protein complexes in living bacterial cells, followed by stringent purification and sequencing of crosslinked RNA, to detect transcriptome-wide binding sites of Hfq and CsrA in *Salmonella*. As well as confirming known binding sites at nucleotide resolution, our study identifies a plethora of new sites that reveal the specificities of Hfq and CsrA interactions with their RNA ligands. Our contact maps for Hfq interacting sRNAs and their target mRNAs support a model for Hfq as a mediator of RNA duplex formation and provide new insight into improving sRNA-target prediction. The discovery of CsrA-binding sites in mRNAs shows that CsrA is a direct regulator of *Salmonella* virulence genes.

Results

Selective enrichment of crosslinked RNA ligands

To comprehensively analyze direct targets of RBPs *in vivo*, we established a CLIP-seq protocol for purification of crosslinked RNA–protein complexes from bacterial cells irradiated with UV light (Fig 1A). *Salmonella* strain SL1344 expressing chromosomally FLAG-tagged Hfq was cultured in LB medium to an OD_{600} of 2.0. One half of this culture was then irradiated with UV light while the other half was left untreated. This growth condition activates the invasion genes of *Salmonella*, that is it enabled us to also capture potential Hfq interactions with virulence-associated transcripts. Hfq–RNA complexes were immunoprecipitated in cell lysates with a monoclonal anti-FLAG antibody followed by several stringent washes.

Published online: April 4, 2016
Erik Holmqvist et al Hfq and CsrA CLIP The EMBO Journal

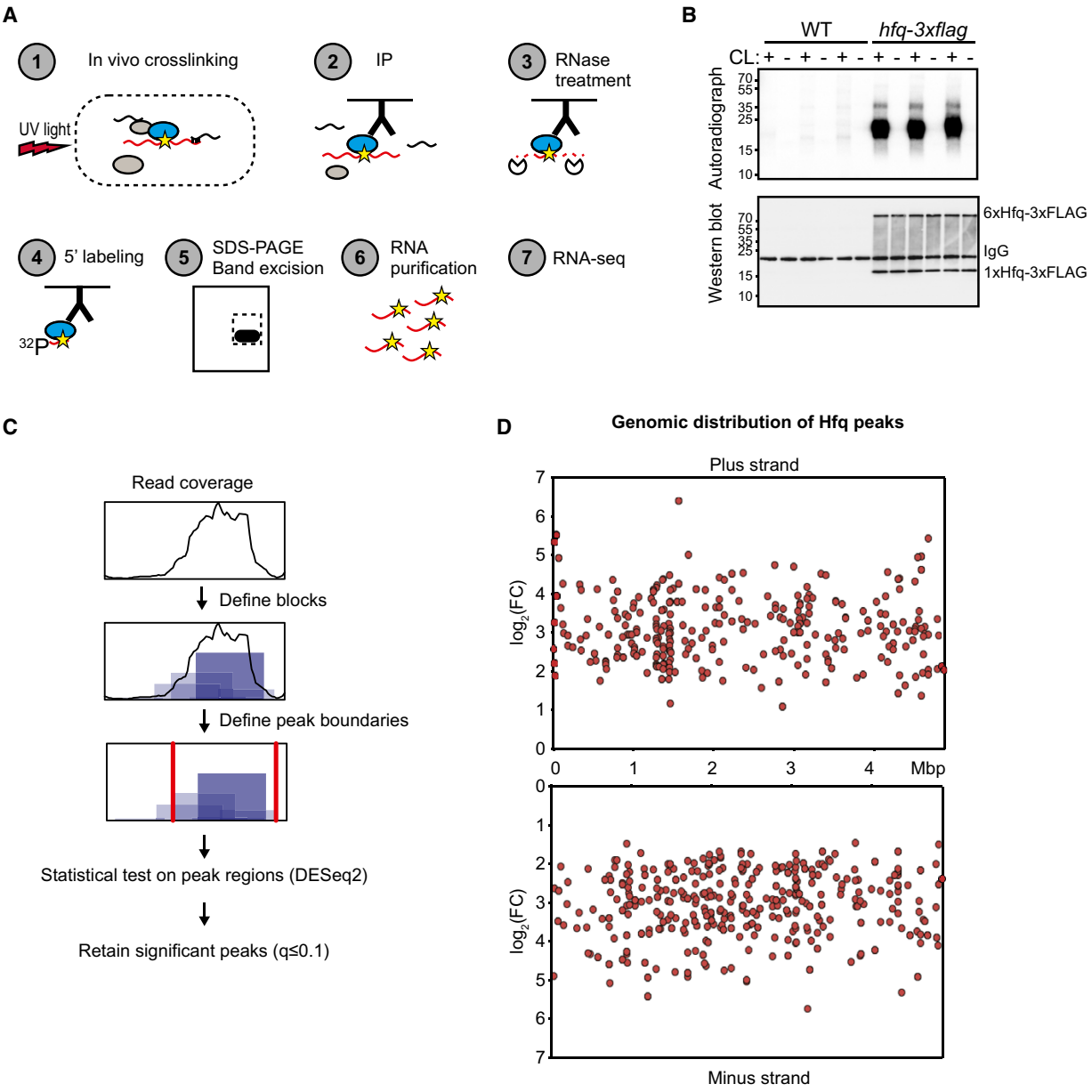


Figure 1. CLIP-seq of Hfq-3xFLAG in *Salmonella*.

A Schematic representation of the CLIP-seq protocol for bacterial RBPs that was established and used in this study. UV: ultraviolet.

B Detection of crosslinked, immunoprecipitated, and radioactively labeled RNA–protein complexes after separation on denaturing SDS–polyacrylamide gels and transfer to nitrocellulose membranes. Radioactive signals were detected by phosphorimaging (top). Detection of Hfq-3xFLAG proteins by Western blot using an anti-FLAG antibody served as a control for successful immunoprecipitation (bottom). CL: crosslinking.

C Schematic representation of binding site determination (peak calling).

D Fold change (*y*-axis) and genomic position (*x*-axis) of Hfq peaks. Mbp: mega basepair.

After on-bead RNase treatment, dephosphorylation, and radioactive labeling of RNA 5′ ends, the complexes were eluted, separated by denaturing SDS–PAGE, and transferred to a membrane. UV irradiation itself did not interfere with protein recovery (as judged by Western blot), but a strong radioactive signal corresponding to bound labeled RNA was detected only in tagged and crosslinked samples, indicating that unspecific RNA–protein interactions were successfully depleted (Fig 1B). RNA–protein complexes from

Published online: April 4, 2016
The EMBO Journal

crosslinked and control samples were extracted from the membrane and treated with proteinase to yield RNA ligands for analysis by Illumina sequencing. The number of sequencing reads obtained for each cDNA library is given in Appendix Fig S1. To avoid biases introduced during library amplification, reads originating from potential PCR duplicates were removed for all downstream analyses.

A very important step in the analysis of CLIP-seq data is peak calling, which is used to differentiate between specific und unspecific binding. Here, two major problems in standard CLIP-seq protocols may confound peak calling approaches. Firstly, in contrast to traditional RNA immunoprecipitation and sequencing (RIP-seq), where comparison to a non-tagged strain or the omission of the antibody serves to control for background noise, CLIP-seq approaches usually lack a standardized negative control. Secondly, in contrast to chromatin immunoprecipitation and DNA sequencing (ChIP-seq), transcript abundance impacts read coverage independent of the affinity of the RBP for a given target. Standard peak callers such as Piranha (Uren *et al*, 2012) assume the majority of sites to be noise, so the sum of all sites can be used to fit a background model. However, this assumption is problematic if the RBP is a ubiquitous binder and the genome size is rather small. Both criteria apply in our case. To overcome these problems, we developed a specific peak calling algorithm able to identify Hfq-binding sites throughout the *Salmonella* transcriptome. The algorithm first divides consecutive reads into blocks and then merges overlapping blocks into peaks (Fig 1C). Subsequently, based on three biological replicates and three control replicates, each peak was tested for significant enrichment in the crosslinked samples versus the non-crosslinked samples using DESeq2 (Love *et al*, 2014). This strategy identified 640 significant ($q \leq 0.1$) Hfq peaks (Table EV1) which are distributed across the *Salmonella* transcriptome (Fig 1D).

As a significant advantage of CLIP-seq over simple coIP, crosslinking-induced mutations narrow RNA–protein contacts down to individual nucleotides (Zhang & Darnell, 2011). Thus, we compared the nature of read mutations that (i) occurred in both mate pairs for each read (to discriminate from sequencing errors), (ii) were exclusively present in libraries from crosslinked cultures, and (iii) overlapped with Hfq peaks (Table EV2). T to C mutations were by far the most common crosslink-specific mutation (Fig 2A), and more than half of the Hfq peaks (347/640) contained at least one crosslink-specific mutation. To provide a better display of peak density, the *Salmonella* chromosome was divided into bins of 2×10^4 basepairs. Plotting peak numbers per bin identified certain chromosomal regions in which the density of Hfq peaks is unusually high (Fig 2B). Interestingly, transcripts from the two major pathogenicity islands, SPI-1 and SPI-2, attract the highest Hfq peak density, supporting the crucial role of Hfq in *Salmonella* virulence (Sittka *et al*, 2007). Dividing the Hfq peaks into different RNA classes shows that the majority map to sRNAs and mRNAs, the two RNA classes previously known to be targets of Hfq (Fig 2C). In summary, combining CLIP-seq with a new peak calling algorithm and identification of crosslinking-induced mutations provides the basis for a detailed investigation of Hfq–RNA interactions.

Hfq binding in mRNAs

To analyze the general distribution of the 551 Hfq-binding sites detected in mRNAs, we performed a meta-gene analysis of Hfq peaks with respect to mRNA start and stop codons (for polycistronic mRNAs, only the start codon of the first cistron and the stop codon of the last cistron was used). The greatest peak densities were found in 5′UTRs and 3′UTRs (Fig 2D) and confirmed—on the level of individual transcripts—previously predicted Hfq activity, for example, in the 5′UTR of *chiP* mRNA which is a target of ChiX sRNA (Figueroa-Bossi *et al*, 2009), or the 3′UTR of *hilD* mRNA encoding a virulence regulator (Lopez-Garrido *et al*, 2014) (Fig 2E and F).

To test whether Hfq recognizes disparate sequences in different parts of mRNAs, we divided the mRNA peaks into those that map to 5′UTRs, CDSs, or 3′UTRs. Using the MEME algorithm (Bailey *et al*, 2015), only the combined 3′UTRs yielded a significant consensus motif (Fig 2G). This motif strongly resembles Rho-independent transcription terminators present at the 3′ end of many bacterial transcripts, namely GC-rich hairpins followed by single-stranded uridine tails (Wilson & von Hippel, 1995). Indeed, we found a strong enrichment of Hfq 3′UTR peaks at predicted Rho-independent terminators that were specific to mRNAs (Fig 2H; all sRNA terminators were excluded from this analysis). Moreover, CMfinder analysis (Yao *et al*, 2006) on the Hfq 3′UTR peaks resulted in a motif comprising a hairpin structure followed by a U-rich sequence, strongly resembling a Rho-independent terminator (Fig EV1), suggesting that Hfq binds to mRNA 3′ ends.

Hfq binding in sRNAs

We next compared our crosslinking data to Hfq-binding sites in well-investigated sRNAs. For example, SgrS was proposed to contain an Hfq-binding module consisting of two distinct binding sites: the poly(U) sequence of the Rho-independent terminator at the very 3′ end of SgrS, and an internal hairpin preceded by a U-rich sequence (Ishikawa *et al*, 2012). In accordance with this, we detected two Hfq peaks within SgrS that mapped to the previously reported binding sites (Fig 3A and B). In addition, the only

Figure 2. Genomic distribution of Hfq-binding sites.

A Percentage of the occurrence of the indicated mutations among all crosslink-specific mutations found within Hfq peaks.

B Hfq peak distribution along the *Salmonella* chromosome divided in bins of 2×10^4 basepairs each. The genomic positions of the pathogenicity islands SPI-1 and SPI-2 are indicated. Mbp: mega basepair.

C Distribution of Hfq peaks among the indicated RNA classes. Numbers in parentheses give the number of called peaks that overlapped with annotations belonging to the respective RNA class.

D Global peak density distribution (meta-gene analysis) around start and stop codons. For this analysis, only those start and stop codons were used that are flanked by a 5′UTR or 3′UTR, respectively. Vertical dashed lines indicate the position of start and stop codons, respectively.

E, F Read coverage at the *chiP* (E) and *hilD* (F) loci in libraries from crosslinked and non-crosslinked samples. Exp: experiment, CL: crosslinking

G Consensus motif generated by MEME using sequences of Hfq peaks mapping to mRNA 3′UTRs.

H Meta-gene analysis of peak distribution around genomic positions of predicted Rho-independent terminators.

Published online: April 4, 2016
Erik Holmqvist et al Hfq and CsrA CLIP The EMBO Journal

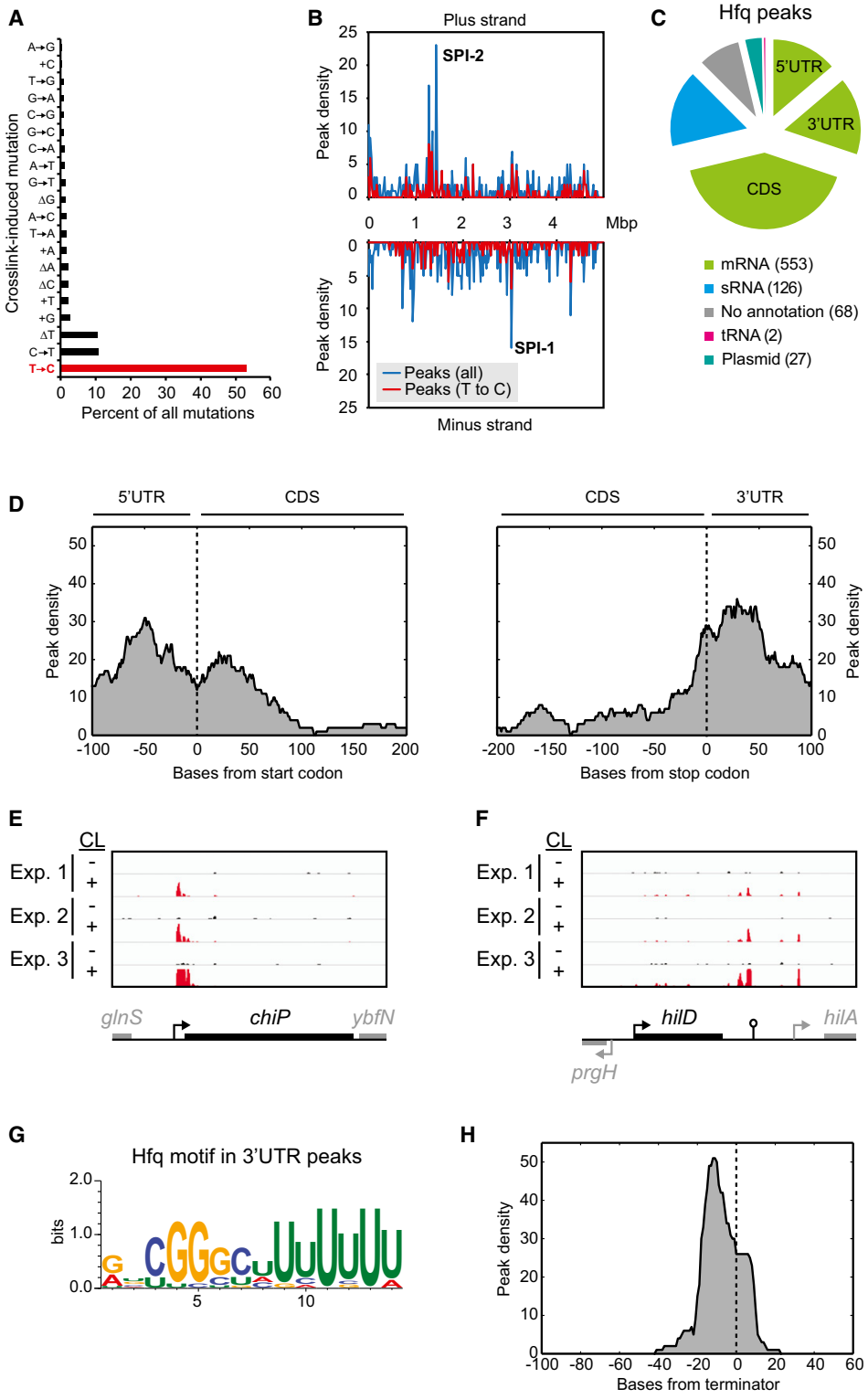


Figure 2.

Published online: April 4, 2016
The EMBO Journal

crosslink-induced mutations detected in SgrS occur within the above-described U-rich sequences (Fig 3B). Likewise, we compared our crosslinking data with the interactions observed in a co-crystal of *Salmonella* Hfq and the sRNA RydC (Dimastrogiovanni *et al*, 2014). The X-ray crystallization data suggest Hfq interacts with four regions on RydC: the proximal site of Hfq interacts with the U-rich 3′ end of RydC; the rim of Hfq interacts with U23/U24, U46/U47, and the RydC 5′ end (Dimastrogiovanni *et al*, 2014). Out of the eight positions in RydC with crosslinking-induced mutations, seven perfectly fit with the crystal structure (Fig 3D). Mutations were found in the 5′ end of RydC, at positions U23, U24, U46, U47, and in the RydC 3′ end (Fig 3D). Taken together, these examples demonstrate that our crosslinking experiments faithfully capture Hfq–RNA interactions at single-nucleotide resolution, in excellent agreement with published work.

The distribution of Hfq peaks over all sRNA sequences suggests that Hfq may interact with different regions in different sRNAs; however, there is a strong bias for Hfq binding toward sRNA 3′ ends (Fig 3E). As for the 3′UTR-binding motif (Fig 2G), the consensus motif found using MEME in peaks mapping to within sRNAs resembles the 3′ region of a Rho-independent terminator (Fig 3F). Following the demonstration of Hfq interactions with 3′ portions of a few sRNAs (Sauer & Weichenrieder, 2011; Ishikawa *et al*, 2012), our screen provides the first global analysis to suggest that Hfq interacts with the 3′ end of many sRNAs detected under the growth condition studied. Taken together, Rho-independent terminators constitute a general Hfq-binding motif shared by mRNAs and sRNAs.

Hfq binding in sRNA–mRNA pairs

A key function of Hfq is to facilitate sRNA–mRNA duplex formation (Møller *et al*, 2002; Zhang *et al*, 2002; Kawamoto *et al*, 2006; Fender *et al*, 2010); this activity seems to require Hfq binding in mRNAs proximal to the site of sRNA pairing, as suggested by studies of *rpoS* mRNA which is regulated by multiple sRNAs (Soper *et al*, 2011). The simultaneous binding of both the sRNA and cognate mRNA by an Hfq hexamer may then accelerate RNA duplex formation at the rim of the protein (Panja *et al*, 2013). To understand where Hfq needs to bind within its ligand to facilitate RNA duplex formation, we performed a meta-gene analysis of Hfq peaks that mapped close to seed pairing regions in known sRNA–mRNA target pairs. In mRNAs, Hfq peaks were significantly more likely to occur 5′ of the respective sRNA interaction site ($P < 0.05$, two-tailed sign test, $n = 17$) (Fig 4A). By contrast, Hfq peaks in sRNAs were found significantly more often 3′ of sRNA seed sequences ($P < 10^{-4}$, two-tailed sign test, $n = 24$) (Fig 4A). This result supports a model whereby Hfq is sandwiched between the mRNA and sRNA of a cognate pair prior to RNA duplex formation (Fig 4B).

The presence of an Hfq site close to an sRNA site in an mRNA improves target regulation (Beisel *et al*, 2012). Therefore, we asked whether our Hfq-binding data could increase the success of sRNA-target predictions. To this end, the top 20 mRNA targets predicted by the CopraRNA algorithm (Wright *et al*, 2013) for each of 17 selected sRNAs were intersected with the list of crosslinked mRNAs, giving 48 predicted mRNA targets with at least one Hfq peak (Fig 4C, Table EV3). Strikingly, inclusion of the Hfq peaks increased the fraction of true positives from 15% to 40% ($P < 10^{-5}$, Fisher's exact test) (Fig 4C).

For experimental validation, we selected the *mglB* mRNA as a new candidate target of Spot42 sRNA. Recognition would occur by a previously established seed sequence within Spot42 (Beisel & Storz, 2011) at a conserved site downstream of the Hfq peak in *mglB* (Figs 4D and EV2). Of note, the levels of MglB, a CRP-cAMP-activated galactose ABC transporter (Zheng *et al*, 2004), are increased in Hfq-deficient cells, predicting that Spot42 represses the *mglB* mRNA in an Hfq-dependent manner (Fig EV2; Sittka *et al*, 2007; Beisel & Storz, 2011). In agreement with this prediction, deletion of *spf* (encoding Spot42) resulted in elevated levels of the *mglB* mRNA (Fig 4E). Reciprocally, we observed a 10-fold repression of this target after pulse-expression of Spot42 (Fig 4F). Spot42 repressed a constitutively transcribed translational *mglB-gfp* fusion, but not a *lacZ-gfp* control, confirming that the regulation occurs at the post-transcriptional level (Fig 4G). To test whether the observed regulation indeed relies on the predicted basepairing, we introduced disruptive mutations in the *mglB-gfp* and Spot42 plasmids (Fig 4H). Deletion of *spf* on the chromosome leads to increased expression of wild-type *mglB-gfp* but not of the mutant *mglB*-gfp* construct (Fig 4H). Likewise, while wild-type Spot42 repressed *mglB-gfp* but not *mglB*-gfp*, the Spot42* mutant repressed *mglB*-gfp* but not *mglB-gfp* (Fig 4H), strongly indicating that the observed regulation indeed relies on basepairing between Spot42 and the *mglB* mRNA, as predicted. In conclusion, these results indicate that knowing which mRNAs are bound by Hfq can dramatically improve the prediction of sRNA targets.

Transcriptome-wide mapping of CsrA-binding sites

Following the successful identification of Hfq-binding sites, we applied our CLIP-seq protocol to CsrA, an RBP that recognizes transcripts very differently compared to Hfq. CsrA has affinity for GGA sequences present in loop regions of hairpins in mRNA 5′UTRs and in a few sRNAs (Vakulskas *et al*, 2015). A *Salmonella* strain carrying a chromosomal *csrA::3xflag* allele was subjected to the same crosslinking and immunoprecipitation strategy described above. As with Hfq, radioactively labeled CsrA-RNA complexes were detected only in crosslinked samples (Fig EV3). Plotting all CsrA peaks obtained from three biological replicates along the *Salmonella* transcriptome revealed a strong enrichment within CsrB and CsrC; almost 40% of reads from all peaks map to these sRNA antagonists of CsrA (Fig 5A and Table EV4), consistent with them being the major cellular ligands of CsrA (Romeo *et al*, 2013). The *glgC* mRNA, the first transcript shown to be directly regulated by CsrA in *E. coli* (Liu *et al*, 1995; Baker *et al*, 2002), was also highly recovered in our experiments (0.5% of reads, Fig 5A and Table EV4).

The CsrB RNA carries multiple hairpins with GGA sequences which serve as high-affinity-binding sites for CsrA. Intriguingly, the read distribution within CsrB is not uniform. Regions with high read densities are separated by low-read regions (Fig 5B). Aligning the CsrA reads on the predicted secondary structure of CsrB, we find that read coverage is highest in the hairpin structures, indicating that these are indeed preferentially bound by CsrA (Fig 5B). Some hairpins show higher coverage than others, perhaps reflecting a hierarchy in CsrA capture by CsrB similar to the proposed step-wise sequestration of the homologous RsmE protein by RsmZ RNA in *Pseudomonas* (Duss *et al*, 2014b). Regarding CsrA mRNA interactions, reads from the *glgC* transcript almost perfectly overlapped

Published online: April 4, 2016
Erik Holmqvist et al Hfq and CsrA CLIP The EMBO Journal

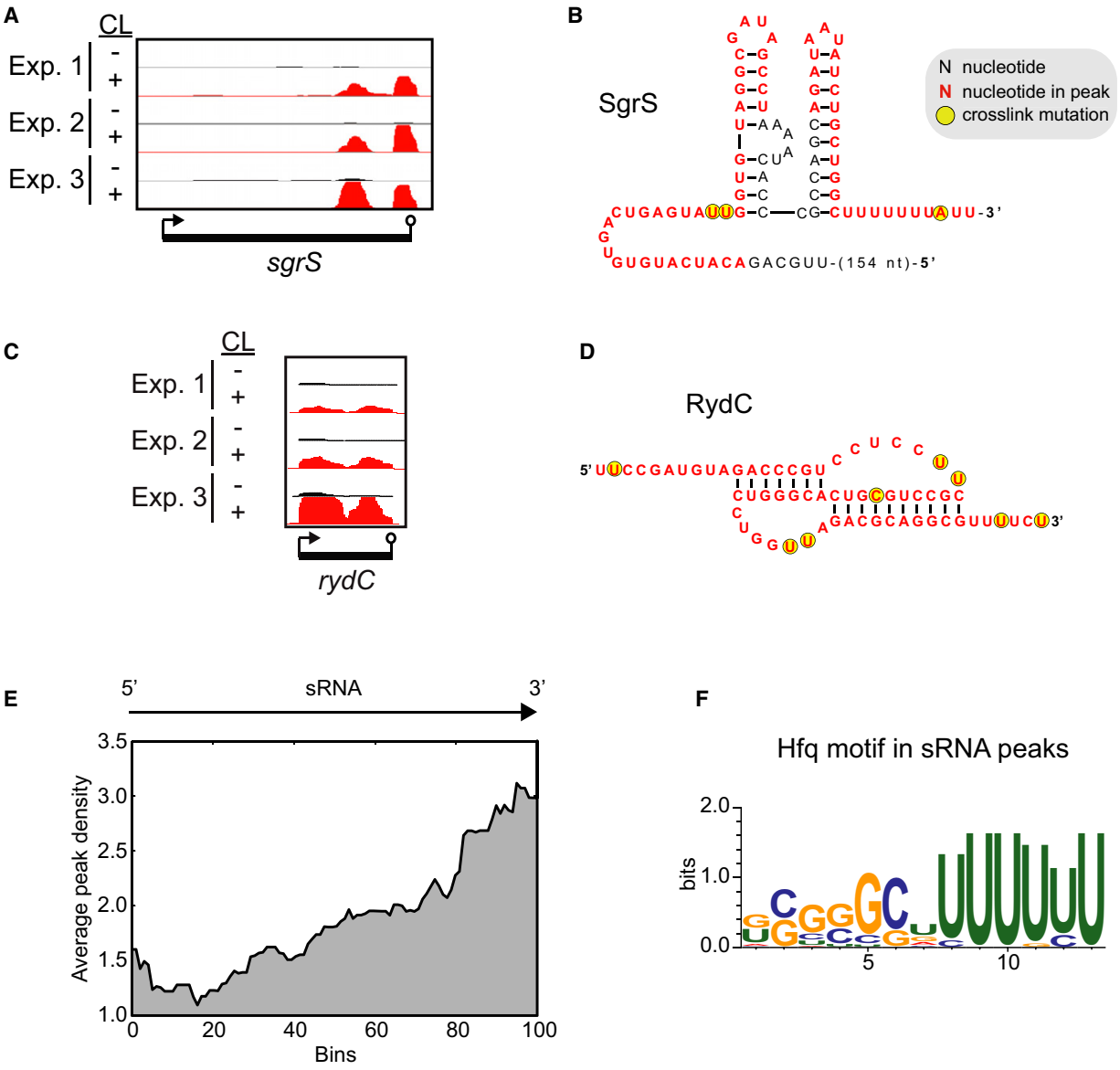


Figure 3. Hfq binding in *Salmonella* sRNAs.

A Read coverage in libraries from crosslinked and non-crosslinked samples at the *sgrS* locus. CL: crosslinking

B Predicted secondary structure of the sRNA SgrS. Nucleotides corresponding to a Hfq peak and positions of crosslink-induced mutations are color coded as highlighted in the legend.

C Read coverage in libraries from crosslinked and non-crosslinked samples at the *rydC* locus. CL: crosslinking.

D Predicted pseudoknot structure of the sRNA RydC. Nucleotides corresponding to an Hfq peak and positions of crosslink-induced mutations are color coded as highlighted in (B).

E Meta-gene analysis of the peak distribution along *Salmonella* sRNAs. Length normalization was achieved through proportional binning according to the different lengths of the sRNA sequences.

F Consensus motif generated by MEME using sequences of peaks mapping to sRNAs as input.

with a GGA-containing hairpin structure in the *glgC* leader (Fig 5C), which was previously defined as the element through which CsrA exercises translational repression in *E. coli* (Baker *et al*, 2002). The detection of CsrA peaks in these two well-documented targets of CsrA suggests that our method readily captures *bona fide* CsrA-binding sites (Fig 5A–C).

Published online: April 4, 2016
The EMBO Journal

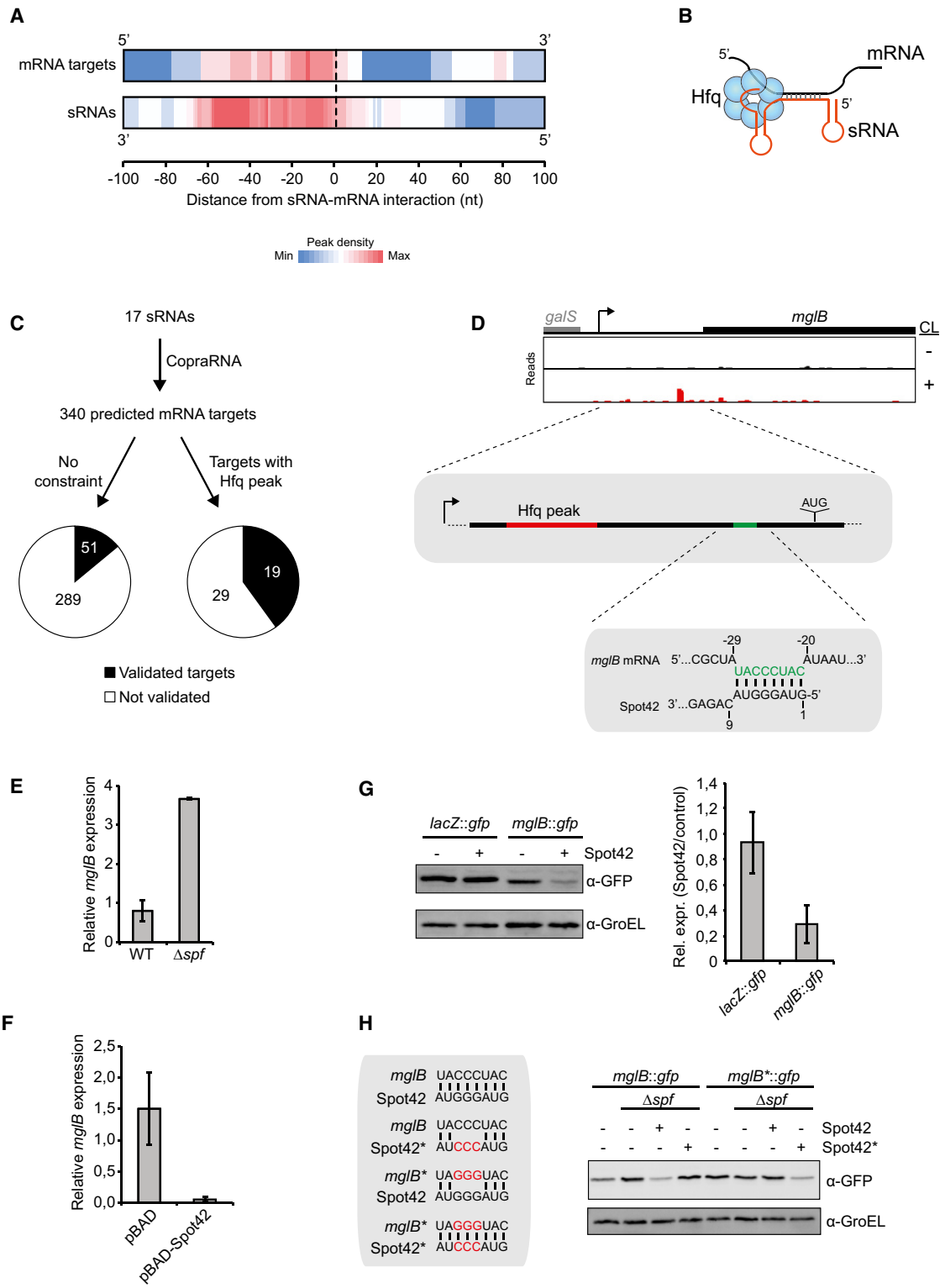


Figure 4.

Published online: April 4, 2016
Erik Holmqvist et al Hfq and CsrA CLIP The EMBO Journal

Figure 4. Hfq binding in validated sRNA–mRNA pairs.

A Distribution of Hfq peaks with respect to sRNA interaction sites in mRNA targets and seed sequences in sRNAs, respectively.
B Putative model of Hfq interaction with cognate sRNA–mRNA pairs.
C Workflow for the integration of Hfq peak information during sRNA-target prediction using CopraRNA. The pie charts show the number of previously validated targets among all predictions, or among predicted targets with Hfq peaks, respectively.
D Read coverage from Hfq CLIP-seq at the *mglB* locus (top), location of the detected Hfq peak (red) and the predicted Spot42 interaction site (green) in the *mglB* 5′UTR (middle), and the predicted basepair interaction between Spot42 and *mglB* (bottom). The Spot42 interaction site in *mglB* is highlighted in green.
E qRT–PCR analysis of *mglB* mRNA expression in wt *Salmonella* or in an isogenic Δ*spf* strain. Samples were collected from cells grown in LB medium to an optical density of 0.3 (OD_{600}). Means and error bars representing standard deviations are based on two biological replicates.
F qRT–PCR analysis of *mglB* mRNA expression in *Salmonella* Δ*spf* 10 min after induction of Spot42 overexpression from plasmid pBAD–Spot42. Plasmid pBAD was used as a control. Means and error bars representing standard deviations are based on two biological replicates.
G Western blot analysis of GFP expression from plasmid-expressed translational *lacZ-gfp* and *mglB-gfp* fusions in the presence or absence of Spot42 overexpression. Quantification of Western blot signals is shown on the right. Means and error bars representing standard deviations are based on three biological replicates. GFP fusion proteins were detected with an anti-GFP antibody, while an anti-GroEL antibody was used to determine the amount of protein loaded on the gel.
H Western blot analysis of GFP expression from the wild-type *mglB-gfp* or mutant *mglB*-gfp* fusions upon deletion and overexpression of wild-type Spot42 or the Spot42* mutant. The predicted interactions between Spot42 and *mglB*, as well as the introduced mutations, are shown.

Source data are available online for this figure.

CsrA consensus motif

We called a total of 467 CsrA peaks, most of which map to within mRNAs (Fig 6A and Table EV4). Meta-gene analysis showed an enrichment of peaks in 5′UTRs compared to CDSs and 3′UTRs, with the strongest enrichment of peaks close to start codons, consistent with CsrA being a regulator of translation initiation (Fig 6B).

High-affinity CsrA–RNA interactions are defined by both RNA sequence and structure (Romeo *et al*, 2013). Interrogation of the CsrA peaks showed that each contained at least one minimal GGA triplet and more than half of them an ANGGA sequence (Fig 6C). Searching all peak regions using the MEME algorithm, we established [A/C]UGGA as the CsrA recognition motif in *Salmonella* (Fig 6D).

Similar to Hfq, we observed that crosslinking of CsrA to RNA frequently causes mutations during reverse transcription. T to C transitions were most prominent (Fig 6E, Table EV5), and these were most often found immediately upstream of a GGA motif (Fig 6E). To analyze the structural context of CsrA-binding sites, we performed CMfinder analysis on all CsrA peaks (Yao *et al*, 2006). Two of the resulting motifs, the one with the highest rank score and the one detected in the most peak sequences (Fig 6F left and right, respectively), consist of stem-loops with a GGA sequence present in the loop regions. Thus, our CLIP analysis confirms the preference for CsrA to interact with AUGGA sequences present in apical loops of hairpin structures. These are the first global data to prove the previous biochemical and genetical studies of individual CsrA ligands, which increasingly suggested ANGGA as a general recognition motif in a variety of bacterial species (Valverde *et al*, 2004; Dubey *et al*, 2005; Majdalani *et al*, 2005; Mercante *et al*, 2006; Babitzke *et al*, 2009; Lapouge *et al*, 2013).

CsrA regulates *Salmonella* virulence genes

Binding of CsrA to target mRNAs typically results in reduced mRNA translation and/or stability (Romeo *et al*, 2013). Since the vast majority of the CsrA sites detected here were previously unknown, we wondered whether they were functional in terms of CsrA-mediated gene regulation. One primary genomic area of CsrA peak density was the invasion gene island SPI-1; likewise, a KEGG pathway analysis suggested enrichment of CsrA peaks in mRNAs encoding *Salmonella* virulence proteins (Fig 7A and B). Our crosslinking data (Table EV4) not only support the previously proposed direct regulation of *hilD* mRNA (encoding a SPI-1 transcription factor) by CsrA (Martinez *et al*, 2011), but also predict CsrA to target dozens of additional virulence-associated mRNAs from both *Salmonella*'s pathogenicity islands and the core genome (Appendix Fig S2).

To test whether the presence of CsrA peaks correlates with CsrA-mediated gene regulation, we constructed translational *gfp*-fusion reporters (Corcoran *et al*, 2012) to several virulence-associated ORFs from the core genome (*sopD2*) or the SPI-1 locus (*sic-sip* and *prg* operons). GFP fusion plasmids were transformed into Δ*csrB*Δ*csrC* cells harboring either a plasmid expressing CsrB, or an empty control plasmid, reasoning that CsrB-mediated titration of CsrA will translate into GFP reporter regulation. This strategy was chosen to circumvent the genetic instability observed in *csrA* deletion strains (Altier *et al*, 2000). While co-expression of CsrB had no effect on a *lacZ-gfp* control plasmid (pXG10-SF), it caused a strong derepression of a *glgC-gfp* fusion chosen as positive control (Fig EV4), arguing that this experimental setup faithfully monitors CsrA-mediated regulation.

SopD2 is an effector protein that promotes *Salmonella* replication inside macrophages (Figueira *et al*, 2013), and CLIP-seq data identified several CsrA peaks in the *sopD2* 5′UTR and CDS (Fig 7C). Western blot analysis showed that *sopD2-gfp* expression is repressed when CsrA activity is increased as a result of deletion of *csrB* and *csrC* (Fig 7D). This is reversed by complementing the double sRNA deletion strain with *csrB* on a plasmid (Fig 7D). A CsrA peak in the 5′UTR of *sopD2* overlaps with a predicted RNA hairpin structure with two GGA motifs in the loop (Fig 7E). A *sopD2-gfp* fusion in which both GGA motifs were each replaced by CCU totally abolished the regulation, strongly indicating that CsrA directly represses the production of SopD2 (Fig 7E). In further support of this, overexpression of CsrB upregulates the synthesis of endogenous SopD2 in wild-type *Salmonella* (Fig EV5).

The *prgHIJK-orgA* operon encodes components of the SPI-1 type III secretion system needed for host cell invasion, and CsrA peaks were detected in its four-first cistrons (Fig 7F). Western blot analysis with translational fusions encompassing cistron junctions with the downstream cistron being fused to *gfp* showed that translation of *prgI* and *prgJ* is activated upon CsrB overexpression, whereas

Published online: April 4, 2016
The EMBO Journal

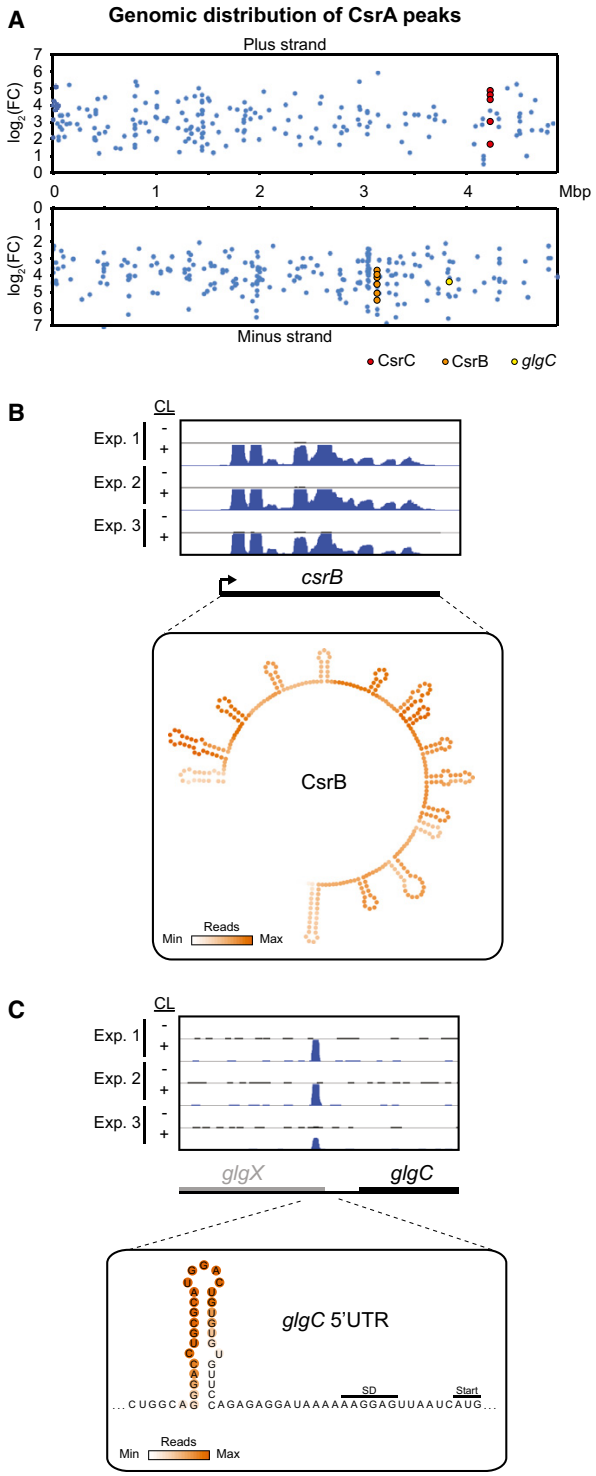


Figure 5. CLIP-seq of *Salmonella* CsrA-3xFLAG captures previously known CsrA-binding sites.

A Fold change (*y*-axis) and genomic position (*x*-axis) of CsrA peaks. Peaks mapping to the known CsrA ligands CsrB, CsrC, and *glgC* are indicated.

B Read coverage from CsrA CLIP-seq at the *csrB* locus (top). A heat map of the average read coverage at the *csrB* locus superimposed on the predicted secondary structure of *Salmonella* CsrB (bottom). The CsrB structure was predicted by MFOLD (Zuker, 2003).

C Read coverage from CsrA CLIP-seq at the *glgC* locus (top). A heat map of the average read coverage at the *glgC* locus superimposed on the predicted secondary structure of the 5′UTR of the *Salmonella glgC* mRNA (bottom).

prgK is not affected (Fig 7G). Of note, the major peaks are located in *prgI* and *prgJ* (Fig 7F). Similarly, the *sicA-sipBCDA-iacP* operon encodes a protein chaperone (SicA), four effector proteins (SipB, SipC, SipD, and SipA), and a putative acyl carrier protein (IacP), and CsrA peaks are distributed across this operon (Fig 7H). Of the four fusions cloned from this operon, three (*sicA, sipC, and sipA*) were clearly upregulated upon CsrB overexpression, indicating that expression from the respective cistrons is repressed by CsrA (Fig 7I). In conclusion, the results shown in Fig 7 strongly indicate that CsrA peaks indeed mark mRNAs that are under direct control of CsrA and suggest that direct regulation of virulence functions by CsrA includes many more mRNAs than previously known.

Discussion

Historically, molecular biologists have focused on the interactions between individual proteins with target nucleic acids *in vitro*, but this approach does not scale well and fails to account for the complexity observed in transcriptional networks. Post-genomic approaches can now potentially provide the global data required to understand post-transcriptional gene regulation in bacteria (Barquist & Vogel, 2015). Specifically, *in vivo* crosslinking methods can determine protein-binding sites within RNA at high resolution and permit stringent purification that diminishes non-specific contamination. Nevertheless, these CLIP-seq approaches have been associated with considerable background noise that, if left uncorrected, increased the identification of false positive interactions (Friedersdorf & Keene, 2014). Here, we have sequenced libraries prepared from both UV crosslinked and non-crosslinked bacterial cultures to control for background RNA, yielding a high-confidence transcriptome-wide map of the binding sites of the two global RNA-binding proteins Hfq and CsrA.

We have shown that Hfq selectively and primarily crosslinks to *Salmonella* mRNAs and sRNAs (Fig 2), in accordance with our previous Hfq coIP results (Sittka *et al*, 2008; Chao *et al*, 2012). More importantly, while relatively few Hfq–sRNA interactions have been studied in biochemical or structural detail, we can faithfully reproduce such results with single-nucleotide resolution in our crosslinking experiment, as shown in Fig 3 for the model sRNAs RydC and SgrS (Ishikawa *et al*, 2012; Dimastrogiovanni *et al*, 2014). Global analysis revealed that Hfq peaks in mRNAs are enriched in 5′UTRs and 3′UTRs as compared to CDS regions (Fig 2), consistent with a role for Hfq in both sRNA-dependent regulation at mRNA 5′ regions and 3′ end-dependent processes. Analysis of Hfq peak density over the *Salmonella* transcriptome revealed strong enrichment in transcripts expressed from the major pathogenicity islands SPI-1 and SPI-2 (Fig 2B). This may in part be explained by the higher content of A and U residues in these transcripts compared

Published online: April 4, 2016
Erik Holmqvist et al Hfq and CsrA CLIP The EMBO Journal

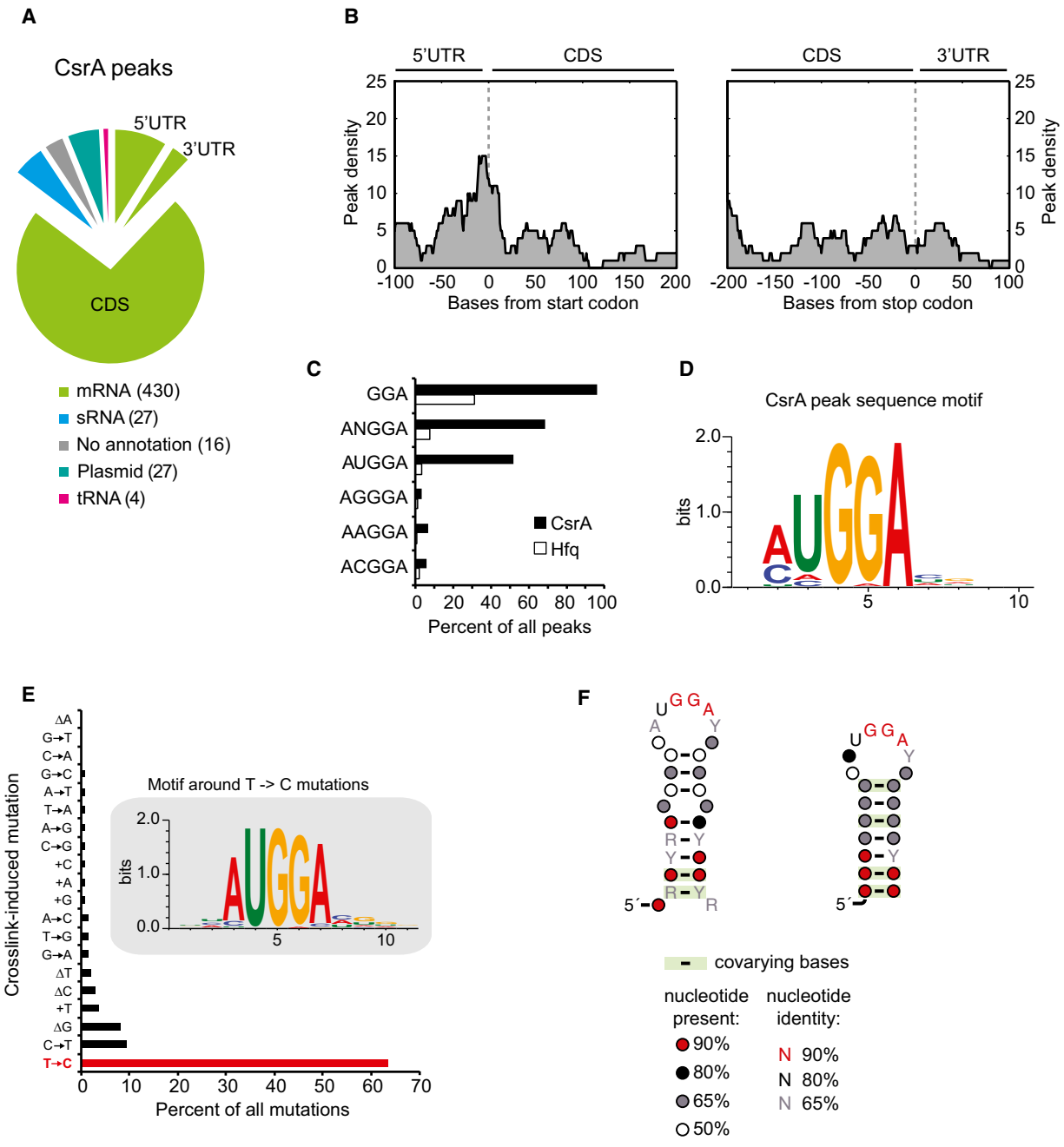


Figure 6. Sequence and structure analysis of CsrA-binding sites.

A Distribution of CsrA peaks among the indicated RNA classes. Numbers in parenthesis represent the number of called peaks that were mapped within annotations belonging to the respective RNA class.

B Meta-gene analysis of CsrA peaks around start and stop codons. For this analysis, only those start and stop codons were used that are flanked by a 5′UTR or 3′UTR, respectively.

C Percentage of peaks that contain the indicated sequences.

D Consensus motif generated by MEME based on all CsrA peak sequences.

E Percentage of the occurrence of the indicated mutations among all crosslink-specific mutations found within CsrA peaks. The inset shows the consensus motif generated with MEME using sequences flanking a crosslink-specific T to C mutation as input.

F Consensus motifs generated by CMfinder based on all CsrA peaks.

Published online: April 4, 2016
The EMBO Journal

Figure 7. CsrA plays a major role in the regulation of *Salmonella* virulence genes.

A CsrA peak density distribution along the *Salmonella* chromosome in bins of 2×10^4 basepairs. The genomic positions of *Salmonella* pathogenicity islands SPI-1 and SPI-2 are indicated.

B KEGG pathways that were found significantly enriched among gene annotations to which CsrA peaks were mapped. Pathways that are related to *Salmonella* pathogenicity are highlighted in red.

C Read coverage from CsrA CLIP-seq at the *sopD2* locus. Light blue bars represent called peaks.

D Western blot analysis of SopD2-GFP expression from a translational *sopD2-gfp* fusion on a plasmid in the indicated strain backgrounds. Plus sign indicates the presence of plasmid pCsrB. Minus sign indicates the presence of the control vector pJV300. SopD2-GFP signals were detected with an anti-GFP antibody. Expression of GroEL served as a loading control and was detected with an anti-GroEL antibody.

E Predicted secondary structure of the *sopD2* 5′UTR. Peak position, GGA motifs, and introduced mutations are indicated. GFP fluorescence measurements from the wild-type *sopD2-gfp* fusion or a 2xCCU mutant upon *csrBcsrC* deletion and CsrB complementation. Means and error bars representing standard deviations are based on three independent experiments.

F Read coverage at the *prgHIJK-orgAB* locus from a CsrA CLIP-seq experiment.

G Western blot analysis of the expression from the indicated plasmid-borne translational GFP fusions in the presence of plasmids pCsrB (plus signs) or pJV300 (minus signs).

H Read coverage at the *sicA-sipBCDA-iacP* locus from a CsrA CLIP-seq experiment.

I Western blot analysis of the expression from the indicated plasmid-borne translational GFP fusions in the presence of plasmids pCsrB (plus signs) or pJV300 (minus signs).

Source data are available online for this figure.

to those expressed from the core genome (Hensel, 2004). Comprehensive analysis of sRNA peaks revealed a strong enrichment of Hfq binding at 3′ ends (Fig 3). The highly enriched consensus motifs found in peak sequences from either mRNA 3′UTRs or sRNAs, respectively, both resemble the 3′ region of Rho-independent terminators (Figs 2, 3 and EV1) and were indeed found in 3′UTRs of mRNAs predicted to transcriptionally terminate in a Rho-independent manner (Fig 2).

The strong evidence for Hfq binding to 3′ ends in mRNAs and sRNAs presented here agrees with previous reports on individual Hfq ligands. Hfq protects RNA from 3′ to 5′ exonuclease activity by binding to, and stimulating the addition of, non-templated poly(A) sequences to RNA 3′ ends by poly(A) polymerase PAPI (Hajnsdorf & Regnier, 2000; Le Derout *et al*, 2003). The sRNA SgrS strongly depends on Hfq binding at its 3′ poly(U) tail for both stability and target regulation (Otaka *et al*, 2011), and the destabilization of SgrS in the absence of Hfq is dependent on the exonuclease PNPase (Andrade *et al*, 2012).

That Hfq binds so commonly to mRNA 3′ ends may be very relevant for sRNA evolution. Cloning or RNA-seq-based studies have identified many sRNAs derived from mRNA 3′UTRs (Vogel *et al*, 2003; Kawano *et al*, 2005; Sittka *et al*, 2008; Chao *et al*, 2012). Whether these sRNAs are produced from internal promoters or by endonucleolytic cleavage of the parental mRNA, they often possess a Rho-independent terminator shared with the mRNA expressed from the same locus (Miyakoshi *et al*, 2015b). Several 3′ UTR-derived sRNAs have been shown to be functional, for example DapZ (Chao *et al*, 2012), MicL (Guo *et al*, 2014), or SroC (Miyakoshi *et al*, 2015a), suggesting that mRNA 3′UTRs may serve as evolutionary birthplaces for sRNAs (Miyakoshi *et al*, 2015b; Updegrove *et al*, 2015). This extends to other types of regulatory transcripts such as recently discovered sRNA sponges that are made from the 3′ end of tRNA precursors (Lalaouna *et al*, 2015).

A key finding from our analysis of the crosslinking data is that we were able to locate Hfq-binding sites in relation to sRNA–mRNA interaction sites (Fig 4). Our observation of preferential binding of Hfq to 5′ of the sRNA interaction site in an mRNA target, and 3′ of the seed sequence in the recognizing sRNA, supports a model whereby Hfq brings the two RNAs together to facilitate RNA duplexing. We used this global information on Hfq binding to substantially improve sRNA-target predictions (Fig 4), illustrating how global RNA–protein interaction maps can foster a better understanding of post-transcriptional networks and discovering the *mglB* mRNA as a target for the sRNA Spot42 (Fig 4). MglB is a transporter of the non-preferred carbon source galactose, and its expression is activated by CRP–cAMP (Zheng *et al*, 2004). Thus, the regulation of *mglB* by Spot42 fits with a proposed model in which Spot42 and CRP form a feed-forward loop to reduce leaky expression of proteins during carbon foraging (Fig EV2; Beisel & Storz, 2011).

The fact that Hfq binds RNA on three distinct faces of the hexamer, each with a different sequence preference, produces a challenge for CLIP-seq methods in that ligation of sequencing adapters to RBP-bound RNA, as well as UV irradiation, may introduce biases in binding site detection. This may explain why our Hfq CLIP-seq data contrast with a recent crosslinking study of Hfq in *E. coli* (Tree *et al*, 2014). This latter study identified neither the 3′-located terminator-like consensus motif nor an enrichment of Hfq-binding sites in sRNA 3′ ends. Instead, the authors concluded that Hfq binding occurs in the seed sequences located in the middle or at the 5′ end of sRNAs. These differences can be explained by differences in the protocols: 3′ adapter ligation to RNA in complex with Hfq (Tree *et al*, 2014) versus adapter ligation after the RNA fragments are released from Hfq (this study). As RNA 3′ ends may not be accessible to ligation when bound to the proximal side of Hfq, adapter ligation to purified RNA as performed here may be the preferred strategy for CLIP approaches when studying proteins that target RNA 3′ ends.

In addition, Tree *et al* (2014) reported a general ARN motif in Hfq crosslink regions, which seemed consistent with structural data on the interaction between the distal face of Hfq and A-rich sequences (Link *et al*, 2009), and the involvement of mRNA located ARN sequences in sRNA-dependent regulation (Salim & Feig, 2010; Beisel *et al*, 2012; Salim *et al*, 2012; Peng *et al*, 2014). Reviewing our CLIP-seq data, on the one hand, almost all (38/39) Hfq peaks in mRNAs known to be targeted by sRNAs (including *rpoS, ompA, ompC, cfa,* and *mglB*) contain at least one ARN motif (Table EV1). On the other hand, we only detected Hfq peaks in 30% of the previously described sRNA targets (Table EV1) (Wright *et al*, 2013), and we did not observe a significant enrichment of ARN motifs among the mRNA peak sequences compared to randomly selected

Published online: April 4, 2016
Erik Holmqvist et al Hfq and CsrA CLIP The EMBO Journal

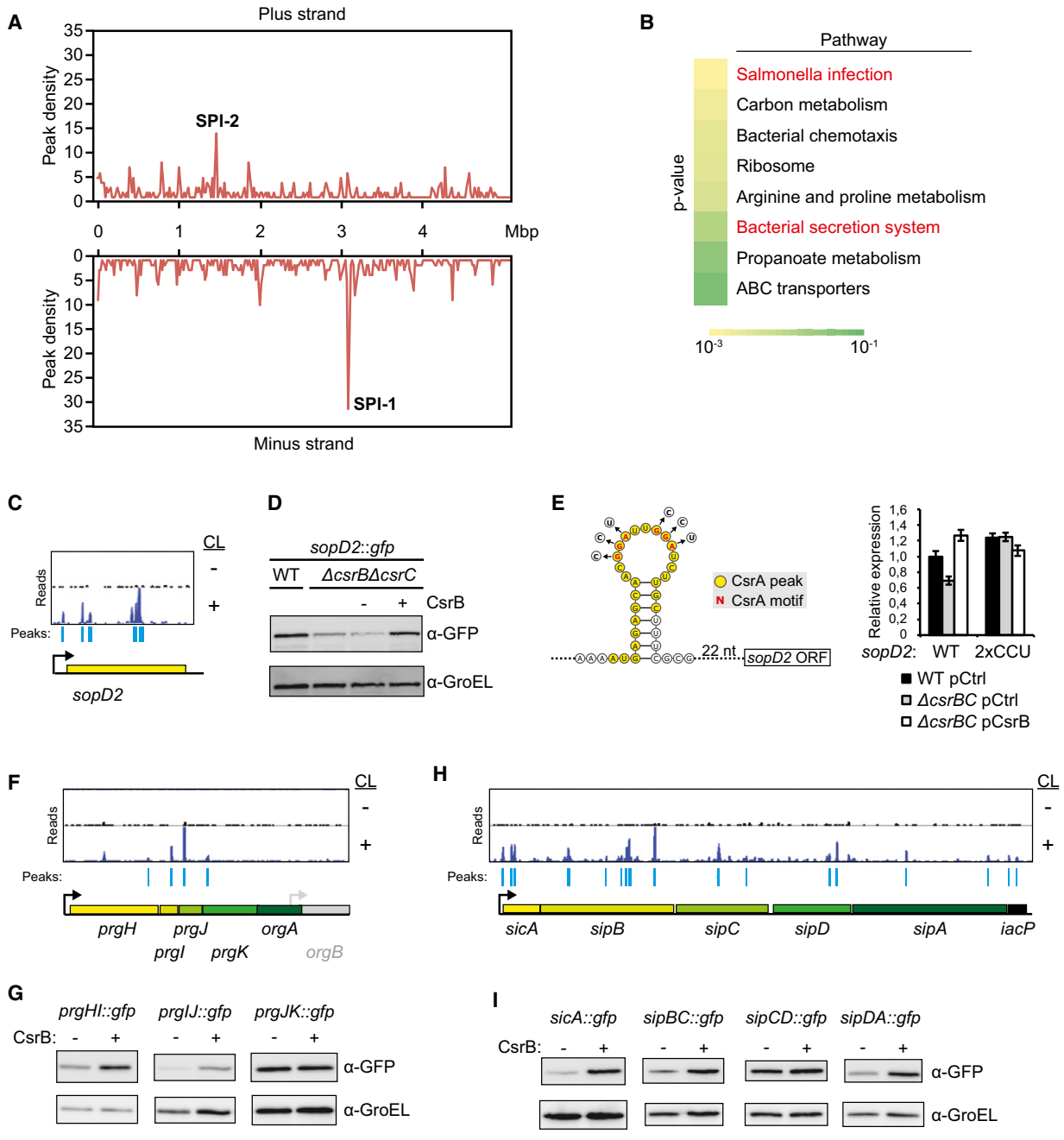


Figure 7.

sequences. One explanation for this discrepancy may be that uridines are more prone to crosslink than other nucleosides (Sugimoto *et al*, 2012); this bias together with the above-discussed adaptor ligation issues may explain why we preferentially detect binding of Hfq at 3′-located U-rich sequences, while the different adapter ligation strategy forced preferential detection of A-rich sequences in the previous *E. coli* study (Tree *et al*, 2014).

Moreover, the canonical view that sRNAs generally interact with the proximal side of Hfq and mRNA targets with the distal side has already been challenged: a recent study showed that some sRNAs use ARN sequences to interact with the distal side of Hfq, whereas their cognate targets harbor 5′UTR-located UA-rich rim-binding sequences (Schu *et al*, 2015). In support of this finding, we find crosslinking mutations in an ARN sequence in the sRNA ChiX and in a UA-rich

Published online: April 4, 2016
The EMBO Journal

sequence in the cognate target mRNA *chiP* (*ybfM*) (Table EV2). Taken together, we propose that mapping of the *in vivo* binding events at each of the three Hfq interaction faces, applying CLIP-seq to mutant Hfq proteins, should be undertaken to further test the current model of distinct "sRNA" and "mRNA" binding faces of Hfq.

These issues with Hfq notwithstanding, the successful application of our crosslinking protocol to CsrA, an RBP with very different targets and recognition mode to Hfq, strongly supports the general applicability of our crosslinking protocol. In contrast to Hfq-binding regions, the vast majority of the detected CsrA-binding sites contain the crucial GGA motif for CsrA–RNA interactions (Figs 5 and 6; Vakulskas *et al*, 2015). CsrA is known to regulate virulence gene expression in *Salmonella*, and a direct interaction between CsrA and *hilD* mRNA, encoding a transcriptional activator of SPI-1, has been described (Martinez *et al*, 2011). In addition to binding *hilD* mRNA, our crosslinking data suggests that CsrA binds to a plethora of virulence-associated mRNAs (Appendix Fig S2). The regulatory potential of newly discovered CsrA-binding sites in virulence-associated mRNAs was confirmed using GFP reporters (Fig 7), consistent with previous reports showing that the levels of some of these mRNAs depend on the intracellular CsrA concentration (Altier *et al*, 2000; Lawhon *et al*, 2003). Even though our validation of CsrA targets is far from comprehensive, it already expands the number of *Salmonella* virulence mRNAs that are post-transcriptionally regulated by CsrA sixfold. Based on our findings, it is likely that more virulence mRNAs are directly regulated by CsrA.

In *Escherichia coli*, the Hfq-dependent McaS sRNA was recently reported to titrate CsrA, suggesting that sRNAs other than CsrB and CsrC may be functional CsrA interaction partners (Jørgensen *et al*, 2013). Interestingly, we also detected binding sites for CsrA in sixteen sRNAs in addition to CsrB and CsrC (Fig 6 and Table EV4), although the read coverage of these additional sRNAs was far below that of CsrB and CsrC. The majority of these sRNAs (14 of 16) carry between one and six GGA motifs, and many of the corresponding peak sequences (12 of 16) fold into hairpins with GGA sequences in the loops (Appendix Fig S3), suggesting that they possess bona fide CsrA-binding sites. Apart from a few well-characterized Hfq-binding sRNAs, of which only one (SdsR) harbors GGA motifs, the majority of the sRNAs that crosslinked to CsrA are uncharacterized. Comparative expression analysis revealed that several of these sRNAs (STnc1890, STnc2080, STnc1210, STnc1480, PinT, and SdsR) are induced in late stationary phase, a growth condition in which CsrB and CsrC are repressed (Kröger *et al*, 2013). This suggests that these six sRNAs may compete with CsrB and CsrC under specific conditions. Future studies will be required to determine whether or not these sRNAs are functional CsrA antagonists, or perhaps are regulated by CsrA.

Bacteria express a plethora of regulatory RBPs for which no global binding site information is available. Examples of these include proteins with RNA-binding domains found in cold-shock proteins (the Csp family of proteins) and proteins such as ProQ that possess a FinO-like RNA-binding domain (Phadtare *et al*, 1999; Mark Glover *et al*, 2015). We believe that our procedure for global mapping of the Hfq and CsrA interactomes with cellular RNA will lay the foundations for future studies of other important bacterial RBPs and may also rapidly identify proteins with putative RNA-binding potential. Such studies should be a major future direction in the study of post-transcriptional phenomena in bacteria and will shed light on this shadowy area of gene regulation.

Materials and Methods

Oligodeoxyribonucleotides

DNA oligonucleotides are listed in Appendix Table S1.

Bacterial strains and plasmids

All experiments were performed with *Salmonella enterica* serovar Typhimurium strain SL1344 or derivatives thereof as listed in Appendix Table S2. All plasmids used in this study are listed in Appendix Table S3. Construction of strains and plasmids is described in Appendix Supplementary Methods. The addition of a FLAG-tag to Hfq or CsrA affected neither bacterial growth nor regulation of known Hfq or CsrA targets, indicating that the tag did not compromise protein function (Appendix Fig S4).

UV crosslinking, immunoprecipitation, and RNA purification

For each biological replicate, 200 ml bacterial culture was grown until an OD_{600} of 2.0. Half of the culture was directly placed in a 22 × 22 cm plastic tray and irradiated with UV-C light at 800 mJ/cm^2. Cells were pelleted in 50 ml fractions by centrifugation for 40 min at 6,000 *g* and 4°C, resuspended in 800 µl NP-T buffer (50 mM NaH_2PO_4, 300 mM NaCl, 0.05% Tween, pH 8.0) and mixed with 1 ml glass beads (0.1 mm radius). Cells were lysed by shaking at 30 Hz for 10 min and centrifuged for 15 min at 16,000 *g* and 4°C. Cell lysates were transferred to new tubes and centrifuged for 15 min at 16,000 *g* and 4°C. The cleared lysates were mixed with one volume of NP-T buffer with 8 M urea, incubated for 5 min at 65°C in a thermomixer with shaking at 900 rpm and diluted 10× in ice-cold NP-T buffer. Anti-FLAG magnetic beads (Sigma) were washed three times in NP-T buffer (30 µl 50% bead suspension was used for a lysate from 100 ml bacterial culture), added to the lysate, and the mixture was rotated for one hour at 4°C. Beads were collected by centrifugation at 800 *g*, resuspended in 1 ml NP-T buffer, transferred to new tubes, and washed 2× with high-salt buffer (50 mM NaH_2PO_4, 1 M NaCl, 0.05% Tween, pH 8.0) and 2× with NP-T buffer. Beads were resuspended in 100 µl NP-T buffer containing 1 mM $MgCl_2$ and 2.5 U benzonase nuclease (Sigma) and incubated for 10 min at 37°C in a thermomixer with shaking at 800 rpm, followed by a 2-min incubation on ice. After one wash with high-salt buffer and two washes with CIP buffer (100 mM NaCl, 50 mM Tris–HCl pH 7.4, 10 mM $MgCl_2$), the beads were resuspended in 100 µl CIP buffer with 10 units of calf intestinal alkaline phosphatase (NEB) and incubated for 30 min at 37°C in a thermomixer with shaking at 800 rpm. After one wash with high-salt buffer and two washes with PNK buffer (50 mM Tris–HCl pH 7.4, 10 mM $MgCl_2$, 0.1 mM spermidine), one-tenth of the beads was removed for subsequent Western blot analysis. The remaining beads were resuspended in 100 µl PNK buffer with 10 U of T4 polynucleotide kinase and 10 µCi γ-^{32}P-ATP and incubated for 30 min at 37°C. After three washes with NP-T buffer, the beads were resuspended in 20 µl Protein Loading buffer (0.3 M Tris–HCl pH 6.8, 0.05% bromophenol blue, 10% glycerol, 7% DTT) and incubated for 3 min at 95°C. The magnetic beads were collected on a magnetic separator, and the supernatant was loaded and separated on a 15% SDS–polyacrylamide gel. RNA–protein complexes were transferred

Published online: April 4, 2016
Erik Holmqvist et al Hfq and CsrA CLIP The EMBO Journal

to a nitrocellulose membrane, the protein marker was highlighted with a radioactively labeled marker pen and exposed to a phosphor screen for 30 min. The autoradiogram was used as a template to cut out the labeled RNA–protein complexes from the membrane. Each membrane piece was further cut into smaller pieces, which were incubated for 30 min in a thermomixer at 37°C with shaking at 1,000 rpm in 400 µl PK solution [50 mM Tris–HCl pH 7.4, 75 mM NaCl, 6 mM EDTA, 1% SDS, 10 U of SUPERaseIN (Life Technologies) and 1 mg/ml proteinase K (ThermoScientific)] whereafter 100 µl 9 M urea was added and the incubation was continued for additional 30 min. About 450 µl of the PK solution/urea was mixed with 450 µl phenol:chloroform:isoamyl alcohol in a phase-lock tube and incubated for 5 min in a thermomixer at 30°C with shaking at 1,000 rpm followed by centrifugation for 12 min at 16,000 *g* and 4°C. The aqueous phase was precipitated with 3 volumes of ice-cold ethanol, 1/10 volume of 3 M NaOAc pH 5.2, and 1 µl of GlycoBlue (Life Technologies) in LoBind tubes (Eppendorf). The precipitate was pelleted by centrifugation (30 min, 16,000 *g*, 4°C), washed with 80% ethanol, centrifuged again (15 min, 16,000 *g*, 4°C), dried 2 min at room temperature, and resuspended in 10 µl sterile water.

cDNA library preparation

To enable sequencing on Illumina instruments, libraries were prepared using the NEBNext Multiplex Small RNA Library Prep Set for Illumina (#E7300, New England Biolabs) according to the manufacturer's instructions. About 2.5 µl purified RNA (or sterile water as negative control) was mixed with 0.5 µl 3′ SR Adaptor (diluted 1:10) and 0.5 µl nuclease-free water, incubated for 2 min at 70°C and chilled on ice. After addition of 5 µl 3′ ligation reaction buffer and 1.5 µl 3′ ligation enzyme mix, the samples were incubated for 60 min at 25°C. About 0.25 µl SR RT primer and 2.5 µl nuclease-free water were added followed by incubation for 5 min at 75°C, 15 min at 37°C, and 15 min at 25°C. For ligation of the 5′ adaptor, the sample was mixed with 0.5 µl 5′ SR adaptor (denatured, diluted 1:10), 0.5 µl 10× ligation reaction buffer, and 1.24 µl ligation enzyme mix and incubated for 60 min at 25°C. cDNA synthesis was carried out by the addition of 4 µl first strand synthesis reaction buffer, 0.5 µl murine RNase inhibitor, and 0.5 µl Protoscript reverse transcriptase and incubation at 50°C for 60 min. The reverse transcription activity was inhibited by a 15-min incubation at 70°C. The cDNA was amplified by PCR by mixing 10 µl cDNA sample with 25 µl 2× LongAmp Taq PCR master mix, 1.25 µl SR primer and 17.5 µl nuclease-free water in a thermal cycler with the following program: 30 s at 94°C, 18 rounds of (15 s at 94°C, 30 s at 62°C, and 15 s at 70°C). The PCRs were purified on columns (QIAGEN), eluted in 10 µl sterile water, and loaded on 6% polyacrylamide gels with 7 M urea together with a 50 bp DNA size marker (ThermoScientific). Gels were stained with SYBRGold (Life Technologies), and fragments between 140 and 250 bp were excised from the gels. Elution of DNA fragments was performed in 500 µl DNA elution buffer (NEB) at 16°C overnight in a thermomixer at 1,000 rpm followed by EtOH precipitation. Pellets were resuspended in 10 µl sterile water. About 2 µl gel-purified DNA was mixed with 25 µl 2× LongAmp Taq PCR master mix, 2 µl each of primer JVO-11007 and JVO-11008 (10 µM), and 19 µl sterile water and amplified using the following program: 30 s at 94°C, 6 rounds of (15 s at 94°C, 30 s at 60°C, and 15 s at 65°C). PCRs were purified on columns (QIAGEN) and eluted in 15 µl sterile water.

Sequencing

High-throughput sequencing was performed at vertis Biotechnologie AG, Freising, Germany. Twelve cDNA libraries were pooled on an Illumina NextSeq 500 mid-output flow cell and sequenced in paired-end mode (2 × 75 cycles). Raw sequencing reads in FASTQ format and coverage files normalized by DESeq2 size factors are available via Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo) under accession number GSE74425.

Processing of sequence reads and mapping

To assure high sequence quality, read 1 (R1) and read 2 (R2) files containing the Illumina paired-end reads in FASTQ format were trimmed independently from each other with a Phred score cutoff of 20 by the program fastq_quality_trimmer from FASTX toolkit version 0.0.13. In the same step, after quality trimming NEB, R1 and R2 3′-adapters (R1: AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC, R2: GATCGTCGGACTGTAGAACTCTGAACGTGTAGATCTCGGTGGTCGCCGTATCATT) were trimmed using Cutadapt version 1.7.1 (Martin, 2011) and reads without any remaining bases were discarded. Afterward, reads without a mate in the complementary read file were excluded using cmpfastq (http://compbio.brc.iop.kcl.ac.uk/software/cmpfastq.php). In order to remove putative PCR duplicates, paired-end reads were collapsed using FastUniq (Xu *et al*, 2012). Subsequently, a size filtering step was applied in which read pairs with at least one read shorter than 12 nt or longer than 25 nt were eliminated. The collections of remaining reads were mapped to the *Salmonella* Typhimurium SL1344 chromosome (NCBI Acc.-No: NC_016810.1) and plasmid (NCBI Acc.-No: NC_017718.1, NC_017719.1, NC_017720.1) reference sequences using the RNA-seq pipeline READemption version 0.3.5 (Förstner *et al*, 2014) and segemehl version 0.2.0 (Hoffmann *et al*, 2014) with an accuracy cutoff of 80%. From the results, only reads mapping uniquely to one genomic position were considered for all subsequent analysis. Pearson correlations between all libraries were calculated on nucleotide read coverage (Appendix Fig S5).

Coverage plots representing the numbers of mapped reads per nt were generated for each replicon and strand to facilitate data visualization in a genome browser. Each resulting cDNA coverage graph was normalized using the DESeq2 (Love *et al*, 2014) size factors calculated during peak calling.

For all analyses related to annotated genomic features such as CDSs, tRNAs, and rRNAs, gene annotations from NCBI were used. We defined *ad hoc* transcriptional units (TUs) based on NCBI CDS annotations, transcription start site (TSS) annotations from Kröger *et al* (2013) and Rho-independent terminator predictions by RNIE (Gardner *et al*, 2011). Briefly, TUs were defined as starting on annotated primary TSSes and ending either with a predicted Rho-independent terminator or in the presence of an intergenic gap greater than 500 nt on the coding strand. In the absence of an upstream TSS, an arbitrary 100 nt 5′UTR was added upstream of the first CDS in the TU, and similarly in the absence of a terminator, an arbitrary 100 nt 3′UTR was added. In the event of a predicted primary TSS within an intergenic gap of less than 500 nt on the coding strand, the TU was ended 100 nt downstream of the preceding CDS, or at the end of the preceding CDS if the predicted primary TSS was less than 100 nt downstream. We defined 5′UTRs as the

Published online: April 4, 2016
The EMBO Journal

regions from the start of each predicted TU to the position upstream of the first CDS in the TU and 3′UTRs as the regions from one nt downstream of the last CDS in the TU to the end of the TU. sRNA annotations are based on Perkins *et al* (2009), Chinni *et al* (2010), Kröger *et al* (2013), and KU Förstner and J Vogel (unpublished data).

Peak calling

Peak calling was performed as a two-step process. In the first step, we defined peak regions using the blockbuster algorithm for defining discrete blocks of overlapping reads (Langenberger *et al*, 2009) across all crosslinked libraries for each RNA-binding protein investigated. Mapped and collapsed reads were filtered to only contain properly paired reads. The resulting BAM files were converted to BED format using BEDTools (v2.17.0) (Quinlan & Hall, 2010). These BED files were concatenated for all crosslinked libraries. Subsequently, each read pair in the concatenated BED file was merged into a single unit representing the sequenced RNA fragment. Only fragments ≤ 25 nt and ≥ 12 nt were retained for further analysis. The resulting BED file was reformatted to satisfy the blockbuster input specifications. Blockbuster uses a greedy approach based on a Gaussian smoothing of read profiles to identify clusters of overlapping read blocks. For this procedure, we required blocks to contain at least 10 reads (i.e., the minBlockHeight option was set to 10) and clusters had to be separated by at least one base (i.e., the distance parameter was set to 1). This procedure resulted in a large set of clusters consisting of overlapping blocks of reads. We then iteratively decomposed each cluster of overlapping blocks into peaks, taking into consideration the local frequency of read counts within the cluster. We first selected the block with the highest read count from the cluster under consideration. All blocks that overlapped with this block were removed from the cluster, and a peak was defined using these overlapping blocks. This procedure, of selecting the next largest block, was repeated in the reduced cluster until no more blocks were left that contained greater than 1% of the total cluster read count (see Appendix Supplementary Methods for a formalized description of this procedure).

In the second step of our peak calling analysis, we applied DESeq2 (v1.2.10) (Love *et al*, 2014) to test each peak for a reproducible relative read count enrichment in triplicate crosslinked libraries compared to non-crosslinked controls. Reads per peak were counted using HTSeq-count (v 0.6.1p1) (Anders *et al*, 2015) for all libraries with the mode option set to "union", the order option set to "name" and the stranded option set to "yes". DESeq2 was then run with default options in R. We considered peaks genuine if they had a normalized average expression of ≥ 10 in the crosslinked libraries and a statistically significant enrichment in crosslinked libraries compared to non-crosslinked controls, defined as a false discovery rate (FDR) corrected *P*-value of 0.1 or less.

CopraRNA–Hfq peaks overlap

CopraRNA (Wright *et al*, 2013, 2014) target predictions were performed for all sRNAs from the benchmark dataset of (Wright *et al*, 2013) that had an associated Hfq peak in our data (that is, all except RyhB). Two hundred nucleotides upstream and 100 nucleotides downstream of annotated start codons were specified as potential target regions. The top 20 CopraRNA predictions for each sRNA candidate were subsequently intersected with mRNA candidates that show an Hfq peak in our data. To test for enrichment of known targets in the intersected lists, the number of known targets in the unfiltered top 20 CopraRNA predictions and the number of known targets in the lists resulting from the intersection were compared. The benchmark dataset (Wright *et al*, 2013) was considered as a reference for verified targets and was extended with the interactions between Spot42-glpF (Beisel *et al*, 2012), OxyS-cspC (Tjaden *et al*, 2006), and RybB-STM1530 (Wright, 2012). The unfiltered list of top 20 predictions for 17 individual target predictions contains 51 verified targets in a total list of length 340. The filtered list has a length of 48 and contains 19 verified targets. The interaction between Spot42–mglB discovered in this study was not used for enrichment analysis. A one-sided Fisher's exact test was employed to test for enrichment of known targets in the filtered list relative to the unfiltered list. The test was performed in R statistics using the Fisher's exact test function with the "alternative" parameter set to "greater". For this, we considered that 19 candidates are Hfq bound and verified, 29 candidates are Hfq bound and not verified, 32 candidates are not Hfq bound and verified and 260 candidates are not Hfq bound and not verified. Based on these numbers, the test matrix is given as matrix(c(19,32,29,260), nrow = 2, ncol = 2) in R notation. For the sake of simplicity, we considered targets verified in *E. coli* also to be targets in *Salmonella*. Even though this may not hold true for every single target, this is unlikely to change the principle findings of this analysis.

Analysis of crosslink-specific mutations

For the detection of crosslinking-induced mutation sites from the CLIP-seq data, only uniquely mapped, paired-end reads were considered and used for mutation calling using samtools (v 0.1.19). To reduce bias caused by sequencing errors, we required the mutated sites to be present in both paired reads. A python script adapted from the PIPE-CLIP package (Chen *et al*, 2014) was applied to identify sites significantly enriched in mutations in each library. The number of mutations at each position was modeled as the result of a Bernoulli process with p equal to the observed mutation rate across all positions. Positions were counted as significantly enriched in mutations if the probability of a mutation count greater than or equal to that observed at the position was less than 0.01 under the implied binomial distribution. The final requirement for a site to be considered enriched for crosslinking-induced mutations was that it had to be present in at least two of the libraries from the crosslinked samples and absent in all of the libraries from non-crosslinked samples.

Global analysis of binding regions

The peak density was calculated by counting the number of peaks along the specified annotation features, which included start codons in single-cistron mRNAs and in the first cistron in multigene operons, stop codons in single-cistron mRNAs and in the last cistron in operons, sRNAs, and predicted Rho-independent terminators. These features were retrieved from the extended *Salmonella* Typhimurium SL1344 annotation described above.

Published online: April 4, 2016
Erik Holmqvist et al Hfq and CsrA CLIP The EMBO Journal

Analysis of sequence and structure motifs

The sequences of peaks or sequences 10 nucleotides upstream and downstream of crosslinking mutation sites were used for sequence motif identification using MEME (Bailey *et al*, 2015) with one base shift allowed while the remaining parameters were set at default values. To verify the specificity of the peak motifs found in Hfq peaks from 3′UTRs or sRNAs, the following analysis was performed: for each annotation feature with an Hfq peak, a sequence of the same length as the Hfq peak mapping to that feature but randomly positioned within the feature was extracted. This procedure was repeated ten times. The resulting sequences were used as input for MEME.

To search for the presence of a structural motif, CMfinder 0.2.1 (Yao *et al*, 2006) was run on sequences from peak regions extended by additional 10 nt upstream and downstream, using default parameters except for allowing a minimum single stem loop candidate length of 20 nt. The top-ranked motif incorporated 396 sequences while the motif detected most frequently was found in 416 of the 467 sequences. Both motifs were visualized using R2R (Weinberg & Breaker, 2011) and are depicted in Fig 6F.

Analysis of Hfq peaks in known sRNA–mRNA pairs

Distributions of Hfq peaks in sRNAs and mRNAs with validated basepair interaction sites (Wright *et al*, 2013) were calculated and visualized as a heat map using Excel. The interactions used were restricted to those mRNAs where an Hfq peak was detected within 100 nt on either side of a validated sRNA interaction site.

Pathway analysis

Pathway information was retrieved from the KEGG database (Kanehisa & Goto, 2000), the *Salmonella* SL1344 genome annotation (Kröger *et al*, 2012), and a selection of regulons curated from literature sources. Pathway enrichment analysis was performed using Fisher's exact test, and *P*-values were corrected for multiple testing using the Benjamini–Hochberg method.

Western blot

To analyze immunoprecipitated material in the CLIP experiments, one-tenth of the magnetic beads from each sample was resuspended in 10 µl protein loading buffer and heated 4 min at 95°C. The magnetic beads were collected on a magnetic separator, and the supernatant was loaded and separated on a 15% SDS–polyacrylamide gel followed by transfer of proteins to a nitrocellulose membrane. To detect FLAG-tagged proteins, the membrane was blocked in TBS-T with 5% milk powder, washed in TBS-T for 10 min, incubated for 1 h with anti-FLAG antibody (Sigma) diluted 1:1,000 dilution in TBS-T with 3% BSA, washed in TBS-T for 10 min, incubated for 1 h with anti-mouse-HRP antibody (ThermoScientific) diluted 1:10,000 dilution in TBS-T with 3% BSA, and finally washed in TBS-T two times for 10 min before adding the ECL substrate and taking captions with a CCD camera (ImageQuant, GE Healthcare).

To analyze the expression of GFP fusion proteins, bacterial cultures were harvested at an OD_{600} of 1.0, and cell pellets were boiled in protein loading buffer and separated on 12% SDS–polyacrylamide gels. Proteins were transferred to PVDF membranes and GFP signals were detected as described above but using an anti-GFP antibody (Roche) followed by HRP-coupled anti-mouse antibody (ThermoScientific).

qRT–PCR

Total RNA was extracted using hot phenol, and contaminating DNA was removed by DNase I treatment. qRT–PCRs were carried out using the RNA-to-Ct 1-step kit (ThermoFisher) with 50 ng of RNA per reaction. Relative gene expression was calculated using the $\Delta\Delta C_t$ method (Livak & Schmittgen, 2001) by normalization to the *rfaH* mRNA.

Expanded View for this article is available online.

Acknowledgements
We thank Stan Gorski, Tony Romeo, Gerhart Wagner, and Jay Hinton for comments on the manuscript and Victoria McParland for excellent technical assistance. We also thank Yanjie Chao, Kai Papenfort, Verena Pfeiffer, Dimitri Podkaminski, and Franziska Seifert for providing bacterial strains and plasmids. The Vogel group is supported by funds from DFG and the Bavarian BioSysNet program. Both the Vogel and Backofen group received funds from a joint BMBF eBio:RNASys grant. Erik Holmqvist acknowledges support by an EMBO long-term fellowship and by the Wenner-Gren Foundations. Lars Barquist is supported by a Research Fellowship from the Alexander von Humboldt Foundation.

Author contributions
JV and EH designed the experiments. EH performed the experiments. PRW, TB, and RB implemented the peak calling strategy. LL implemented the calling of crosslink mutations and performed the motif analysis together with TB. PRW, LL, TB, LB, and EH carried out additional bioinformatical analyses. RR carried out sequencing and data analysis. EH and JV wrote the manuscript, which was discussed, modified, and improved by all authors. JV and RB supervised the project.

Conflict of interest
The authors declare that they have no conflict of interest.

References

Altier C, Suyemoto M, Lawhon SD (2000) Regulation of Salmonella enterica serovar typhimurium invasion genes by csrA. *Infect Immun* 68: 6790 – 6797

Anders S, Pyl PT, Huber W (2015) HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31: 166 – 169

Andrade JM, Pobre V, Matos AM, Arraiano CM (2012) The crucial role of PNPase in the degradation of small RNAs that are not associated with Hfq. *RNA* 18: 844 – 855

Ansong C, Yoon H, Porwollik S, Mottaz-Brewer H, Petritis BO, Jaitly N, Adkins JN, McClelland M, Heffron F, Smith RD (2009) Global systems-level analysis of Hfq and SmpB deletion mutants in Salmonella: implications for virulence and global protein translation. *PLoS ONE* 4: e4809

Ascano M, Hafner M, Cekan P, Gerstberger S, Tuschl T (2012) Identification of RNA-protein interaction networks using PAR-CLIP. *Wiley Interdiscip Rev RNA* 3: 159 – 177

Ascano M, Gerstberger S, Tuschl T (2013) Multi-disciplinary methods to define RNA-protein interactions and regulatory networks. *Curr Opin Genet Dev* 23: 20 – 28

Published online: April 4, 2016
The EMBO Journal Hfq and CsrA CLIP *Erik Holmqvist et al*

Babitzke P, Baker CS, Romeo T (2009) Regulation of translation initiation by RNA binding proteins. *Annu Rev Microbiol* 63: 27 – 44

Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME Suite. *Nucleic Acids Res* 43: W39 – 49

Baker CS, Morozov I, Suzuki K, Romeo T, Babitzke P (2002) CsrA regulates glycogen biosynthesis by preventing translation of glgC in *Escherichia coli*. *Mol Microbiol* 44: 1599 – 1610

Baltz AG, Munschauer M, Schwanhausser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M, Wyler E, Bonneau R, Selbach M, Dieterich C, Landthaler M (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell* 46: 674 – 690

Bandyra KJ, Said N, Pfeiffer V, Gorna MW, Vogel J, Luisi BF (2012) The seed region of a small RNA drives the controlled destruction of the target mRNA by the endoribonuclease RNase E. *Mol Cell* 47: 943 – 953

Barquist L, Vogel J (2015) Accelerating discovery and functional analysis of small RNAs with new technologies. *Annu Rev Genet* 49: 367 – 394

Beisel CL, Storz G (2011) The base-pairing RNA spot 42 participates in a multioutput feedforward loop to help enact catabolite repression in *Escherichia coli*. *Mol Cell* 41: 286 – 297

Beisel CL, Updegrove TB, Janson BJ, Storz G (2012) Multiple factors dictate target selection by Hfq-binding small RNAs. *EMBO J* 31: 1961 – 1974

Bilusic I, Popitsch N, Rescheneder P, Schroeder R, Lybecker M (2014) Revisiting the coding potential of the *E. coli* genome through Hfq co-immunoprecipitation. *RNA Biol* 11: 641 – 654

Brencic A, Lory S (2009) Determination of the regulon and identification of novel mRNA targets of Pseudomonas aeruginosa RsmA. *Mol Microbiol* 72: 612 – 632

Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, Krijgsveld J, Hentze MW (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* 149: 1393 – 1406

Chao Y, Papenfort K, Reinhardt R, Sharma CM, Vogel J (2012) An atlas of Hfq-bound transcripts reveals 3′ UTRs as a genomic reservoir of regulatory small RNAs. *EMBO J* 31: 4005 – 4019

Chao Y, Vogel J (2016) A 3′ UTR-derived small RNA provides the regulatory noncoding arm of the inner membrane stress response. *Mol Cell* 61: 352 – 363

Chen B, Yun J, Kim MS, Mendell JT, Xie Y (2014) PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome Biol* 15: R18

Chinni SV, Raabe CA, Zakaria R, Randau G, Hoe CH, Zemann A, Brosius J, Tang TH, Rozhdestvensky TS (2010) Experimental identification and characterization of 97 novel npcRNA candidates in Salmonella enterica serovar Typhi. *Nucleic Acids Res* 38: 5893 – 5908

Coornaert A, Chiaruttini C, Springer M, Guillier M (2013) Post-transcriptional control of the *Escherichia coli* PhoQ-PhoP two-component system by multiple sRNAs involves a novel pairing region of GcvB. *PLoS Genet* 9: e1003156

Corcoran CP, Podkaminski D, Papenfort K, Urban JH, Hinton JC, Vogel J (2012) Superfolder GFP reporters validate diverse new mRNA targets of the classic porin regulator, MicF RNA. *Mol Microbiol* 84: 428 – 445

Darnell RB (2010) HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA* 1: 266 – 286

De Lay N, Gottesman S (2012) A complex network of small non-coding RNAs regulate motility in *Escherichia coli*. *Mol Microbiol* 86: 524 – 538

Desnoyers G, Masse E (2012) Noncanonical repression of translation initiation through small RNA recruitment of the RNA chaperone Hfq. *Genes Dev* 26: 726 – 739

Dimastrogiovanni D, Fröhlich KS, Bandyra KJ, Bruce HA, Hohensee S, Vogel J, Luisi BF (2014) Recognition of the small regulatory RNA RydC by the bacterial Hfq protein. *ELife* 3: e05375

Dubey AK, Baker CS, Romeo T, Babitzke P (2005) RNA sequence and secondary structure participate in high-affinity CsrA-RNA interaction. *RNA* 11: 1579 – 1587

Duss O, Michel E, Diarra Dit Konte N, Schubert M, Allain FH (2014a) Molecular basis for the wide range of affinity found in Csr/Rsm protein-RNA recognition. *Nucleic Acids Res* 42: 5332 – 5346

Duss O, Michel E, Yulikov M, Schubert M, Jeschke G, Allain FH (2014b) Structural basis of the non-coding RNA RsmZ acting as a protein sponge. *Nature* 509: 588 – 592

Edwards AN, Patterson-Fortin LM, Vakulskas CA, Mercante JW, Potrykus K, Vinella D, Camacho MI, Fields JA, Thompson SA, Georgellis D, Cashel M, Babitzke P, Romeo T (2011) Circuitry linking the Csr and stringent response global regulatory systems. *Mol Microbiol* 80: 1561 – 1580

Fender A, Elf J, Hampel K, Zimmermann B, Wagner EG (2010) RNAs actively cycle on the Sm-like protein Hfq. *Genes Dev* 24: 2621 – 2626

Feng L, Rutherford ST, Papenfort K, Bagert JD, van Kessel JC, Tirrell DA, Wingreen NS, Bassler BL (2015) A qrr noncoding RNA deploys four different regulatory mechanisms to optimize quorum-sensing dynamics. *Cell* 160: 228 – 240

Figueira R, Watson KG, Holden DW, Helaine S (2013) Identification of salmonella pathogenicity island-2 type III secretion system effectors involved in intramacrophage replication of S. enterica serovar typhimurium: implications for rational vaccine design. *MBio* 4: e00065

Figueroa-Bossi N, Lemire S, Maloriol D, Balbontin R, Casadesus J, Bossi L (2006) Loss of Hfq activates the sigmaE-dependent envelope stress response in Salmonella enterica. *Mol Microbiol* 62: 838 – 852

Figueroa-Bossi N, Valentini M, Malleret L, Fiorini F, Bossi L (2009) Caught at its own game: regulatory small RNA inactivated by an inducible transcript mimicking its target. *Genes Dev* 23: 2004 – 2015

Figueroa-Bossi N, Schwartz A, Guillemardet B, D'Heygere F, Bossi L, Boudvillain M (2014) RNA remodeling by bacterial global regulator CsrA promotes Rho-dependent transcription termination. *Genes Dev* 28: 1239 – 1251

Förstner KU, Vogel J, Sharma CM (2014) READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data. *Bioinformatics* 30: 3421 – 3423

Friedersdorf MB, Keene JD (2014) Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biol* 15: R2

Gardner PP, Barquist L, Bateman A, Nawrocki EP, Weinberg Z (2011) RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic Acids Res* 39: 5845 – 5852

Gogol EB, Rhodius VA, Papenfort K, Vogel J, Gross CA (2011) Small RNAs endow a transcriptional activator with essential repressor functions for single-tier control of a global stress regulon. *Proc Natl Acad Sci USA* 108: 12875 – 12880

Guo MS, Updegrove TB, Gogol EB, Shabalina SA, Gross CA, Storz G (2014) MicL, a new sigmaE-dependent sRNA, combats envelope stress by repressing synthesis of Lpp, the major outer membrane lipoprotein. *Genes Dev* 28: 1620 – 1634

Hajnsdorf E, Regnier P (2000) Host factor Hfq of *Escherichia coli* stimulates elongation of poly(A) tails by poly(A) polymerase I. *Proc Natl Acad Sci USA* 97: 1501 – 1505

Hébrard M, Kröger C, Srikumar S, Colgan A, Händler K, Hinton JC (2012) sRNAs and the virulence of Salmonella enterica serovar Typhimurium. *RNA Biol* 9: 437 – 445

Published online: April 4, 2016
Erik Holmqvist et al Hfq and CsrA CLIP **The EMBO Journal**

Hensel M (2004) Evolution of pathogenicity islands of Salmonella enterica. *Int J Med Microbiol* 294: 95 – 102

Heroven AK, Bohme K, Dersch P (2012) The Csr/Rsm system of Yersinia and related pathogens: a post-transcriptional strategy for managing virulence. *RNA Biol* 9: 379 – 391

Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, Kunz M, Holdt LM, Teupser D, Hackermuller J, Stadler PF (2014) A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol* 15: R34

Holmqvist E, Reimegård J, Sterk M, Grantcharova N, Römling U, Wagner EGH (2010) Two antisense RNAs target the transcriptional regulator CsgD to inhibit curli synthesis. *EMBO J* 29: 1840 – 1850

Holmqvist E, Vogel J (2013) A small RNA serving both the Hfq and CsrA regulons. *Genes Dev* 27: 1073 – 1078

Ishikawa H, Otaka H, Maki K, Morita T, Aiba H (2012) The functional Hfq-binding module of bacterial sRNAs consists of a double or single hairpin preceded by a U-rich sequence and followed by a 3′ poly(U) tail. *RNA* 18: 1062 – 1074

Jørgensen MG, Nielsen JS, Boysen A, Franch T, Møller-Jensen J, Valentin-Hansen P (2012) Small regulatory RNAs control the multi-cellular adhesive lifestyle of *Escherichia coli*. *Mol Microbiol* 84: 36 – 50

Jørgensen MG, Thomason MK, Havelund J, Valentin-Hansen P, Storz G (2013) Dual function of the McaS small RNA in controlling biofilm formation. *Genes Dev* 27: 1132 – 1145

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27 – 30

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30: 3059 – 3066

Kawamoto H, Koide Y, Morita T, Aiba H (2006) Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. *Mol Microbiol* 61: 1013 – 1022

Kawano M, Reynolds AA, Miranda-Rios J, Storz G (2005) Detection of 5′- and 3′-UTR-derived small RNAs and cis-encoded antisense RNAs in *Escherichia coli*. *Nucleic Acids Res* 33: 1040 – 1050

König J, Zarnack K, Luscombe NM, Ule J (2011) Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet* 13: 77 – 83

Koo JT, Alleyne TM, Schiano CA, Jafari N, Lathem WW (2011) Global discovery of small RNAs in Yersinia pseudotuberculosis identifies Yersinia-specific small, noncoding RNAs required for virulence. *Proc Natl Acad Sci USA* 108: E709 – E717

Kramer K, Sachsenberg T, Beckmann BM, Qamar S, Boon KL, Hentze MW, Kohlbacher O, Urlaub H (2014) Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat Methods* 11: 1064 – 1070

Kröger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hebrard M, Handler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J *et al* (2012) The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *Proc Natl Acad Sci USA* 109: E1277 – 1286

Kröger C, Colgan A, Srikumar S, Handler K, Sivasankaran SK, Hammarlof DL, Canals R, Grissom JE, Conway T, Hokamp K, Hinton JC (2013) An infection-relevant transcriptomic compendium for Salmonella enterica Serovar Typhimurium. *Cell Host Microbe* 14: 683 – 695

Lalaouna D, Carrier MC, Semsey S, Brouard JS, Wang J, Wade JT, Masse E (2015) A 3′ external transcribed spacer in a tRNA transcript acts as a sponge for small RNAs to prevent transcriptional noise. *Mol Cell* 58: 393 – 405

Langenberger D, Bermudez-Santana C, Hertel J, Hoffmann S, Khaitovich P, Stadler PF (2009) Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics* 25: 2298 – 2301

Lapouge K, Perozzo R, Iwaszkiewicz J, Bertelli C, Zoete V, Michielin O, Scapozza L, Haas D (2013) RNA pentaloop structures as effective targets of regulators belonging to the RsmA/CsrA protein family. *RNA Biol* 10: 1031 – 1041

Lawhon SD, Frye JG, Suyemoto M, Porwollik S, McClelland M, Altier C (2003) Global regulation by CsrA in Salmonella typhimurium. *Mol Microbiol* 48: 1633 – 1645

Le Derout J, Folichon M, Briani F, Deho G, Regnier P, Hajnsdorf E (2003) Hfq affects the length and the frequency of short oligo(A) tails at the 3′ end of *Escherichia coli* rpsO mRNAs. *Nucleic Acids Res* 31: 4017 – 4023

Lease RA, Woodson SA (2004) Cycling of the Sm-like protein Hfq on the DsrA small regulatory RNA. *J Mol Biol* 344: 1211 – 1223

Lenz DH, Miller MB, Zhu J, Kulkarni RV, Bassler BL (2005) CsrA and three redundant small RNAs regulate quorum sensing in Vibrio cholerae. *Mol Microbiol* 58: 1186 – 1202

Link TM, Valentin-Hansen P, Brennan RG (2009) Structure of *Escherichia coli* Hfq bound to polyriboadenylate RNA. *Proc Natl Acad Sci USA* 106: 19292 – 19297

Liu MY, Yang H, Romeo T (1995) The product of the pleiotropic *Escherichia coli* gene csrA modulates glycogen biosynthesis via effects on mRNA stability. *J Bacteriol* 177: 2663 – 2672

Liu MY, Gui G, Wei B, Preston JF III, Oakford L, Yuksel U, Giedroc DP, Romeo T (1997) The RNA molecule CsrB binds to the global regulatory protein CsrA and antagonizes its activity in *Escherichia coli*. *J Biol Chem* 272: 17502 – 17510

Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25: 402 – 408

Lopez-Garrido J, Puerta-Fernandez E, Casadesus J (2014) A eukaryotic-like 3′ untranslated region in Salmonella enterica hilD mRNA. *Nucleic Acids Res* 42: 5894 – 5906

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550

Majdalani N, Vanderpool CK, Gottesman S (2005) Bacterial small RNA regulators. *Crit Rev Biochem Mol Biol* 40: 93 – 113

Mark Glover JN, Chaulk SG, Edwards RA, Arthur D, Lu J, Frost LS (2015) The FinO family of bacterial RNA chaperones. *Plasmid* 78: 79 – 87

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17: 10 – 12

Martinez LC, Yakhnin H, Camacho MI, Georgellis D, Babitzke P, Puente JL, Bustamante VH (2011) Integration of a complex regulatory cascade involving the SirA/BarA and Csr global regulatory systems that controls expression of the Salmonella SPI-1 and SPI-2 virulence regulons through HilD. *Mol Microbiol* 80: 1637 – 1656

Mercante J, Suzuki K, Cheng X, Babitzke P, Romeo T (2006) Comprehensive alanine-scanning mutagenesis of *Escherichia coli* CsrA defines two subdomains of critical functional importance. *J Biol Chem* 281: 31832 – 31842

Mika F, Busse S, Possling A, Berkholz J, Tschowri N, Sommerfeldt N, Pruteanu M, Hengge R (2012) Targeting of csgD by the small regulatory RNA RprA links stationary phase, biofilm formation and cell envelope stress in *Escherichia coli*. *Mol Microbiol* 84: 51 – 65

Mikulecky PJ, Kaw MK, Brescia CC, Takach JC, Sledjeski DD, Feig AL (2004) *Escherichia coli* Hfq has distinct interaction surfaces for DsrA, rpoS and poly(A) RNAs. *Nat Struct Mol Biol* 11: 1206 – 1214

Published online: April 4, 2016
The EMBO Journal Hfq and CsrA CLIP *Erik Holmqvist et al*

Miyakoshi M, Chao Y, Vogel J (2015a) Cross talk between ABC transporter mRNAs via a target mRNA-derived sponge of the GcvB small RNA. *EMBO J* 34: 1478 – 1492

Miyakoshi M, Chao Y, Vogel J (2015b) Regulatory small RNAs from the 3′ regions of bacterial mRNAs. *Curr Opin Microbiol* 24: 132 – 139

Møller T, Franch T, Hojrup P, Keene DR, Bachinger HP, Brennan RG, Valentin-Hansen P (2002) Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Mol Cell* 9: 23 – 30

Morita T, Aiba H (2011) RNase E action at a distance: degradation of target mRNAs mediated by an Hfq-binding small RNA in bacteria. *Genes Dev* 25: 294 – 298

Otaka H, Ishikawa H, Morita T, Aiba H (2011) PolyU tail of rho-independent terminator of bacterial small RNAs is essential for Hfq action. *Proc Natl Acad Sci USA* 108: 13059 – 13064

Panja S, Schu DJ, Woodson SA (2013) Conserved arginines on the rim of Hfq catalyze base pair formation and exchange. *Nucleic Acids Res* 41: 7536 – 7546

Papenfort K, Sun Y, Miyakoshi M, Vanderpool CK, Vogel J (2013) Small RNA-mediated activation of sugar phosphatase mRNA regulates glucose homeostasis. *Cell* 153: 426 – 437

Papenfort K, Förstner KU, Cong JP, Sharma CM, Bassler BL (2015) Differential RNA-seq of Vibrio cholerae identifies the VqmR small RNA as a regulator of biofilm formation. *Proc Natl Acad Sci USA* 112: E766 – 775

Peng Y, Soper TJ, Woodson SA (2014) Positional effects of AAN motifs in rpoS regulation by sRNAs and Hfq. *J Mol Biol* 426: 275 – 285

Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, He M, Croucher NJ, Pickard DJ, Maskell DJ, Parkhill J, Choudhary J, Thomson NR, Dougan G (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. *PLoS Genet* 5: e1000569

Phadtare S, Alsina J, Inouye M (1999) Cold-shock response and cold-shock proteins. *Curr Opin Microbiol* 2: 175 – 180

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841 – 842

Rice JB, Balasubramanian D, Vanderpool CK (2012) Small RNA binding-site multiplicity involved in translational regulation of a polycistronic mRNA. *Proc Natl Acad Sci USA* 109: E2691 – 2698

Romeo T, Gong M, Liu MY, Brun-Zinkernagel AM (1993) Identification and molecular characterization of csrA, a pleiotropic gene from *Escherichia coli* that affects glycogen biosynthesis, gluconeogenesis, cell size, and surface properties. *J Bacteriol* 175: 4744 – 4755

Romeo T, Vakulskas CA, Babitzke P (2013) Post-transcriptional regulation on a global scale: form and function of Csr/Rsm systems. *Environ Microbiol* 15: 313 – 324

Salim NN, Feig AL (2010) An upstream Hfq binding site in the fhlA mRNA leader region facilitates the OxyS-fhlA interaction. *PLoS ONE* 5: e13028

Salim NN, Faner MA, Philip JA, Feig AL (2012) Requirement of upstream Hfq-binding (ARN)x elements in glmS and the Hfq C-terminal region for GlmS upregulation by sRNAs GlmZ and GlmY. *Nucleic Acids Res* 40: 8021 – 8032

Sauer E, Weichenrieder O (2011) Structural basis for RNA 3′-end recognition by Hfq. *Proc Natl Acad Sci USA* 108: 13065 – 13070

Sauer E, Schmidt S, Weichenrieder O (2012) Small RNA binding to the lateral surface of Hfq hexamers and structural rearrangements upon mRNA target recognition. *Proc Natl Acad Sci USA* 109: 9396 – 9401

Schu DJ, Zhang A, Gottesman S, Storz G (2015) Alternative Hfq-sRNA interaction modes dictate alternative mRNA recognition. *EMBO J* 34: 2557 – 2573

Schumacher MA, Pearson RF, Möller T, Valentin-Hansen P, Brennan RG (2002) Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein. *EMBO J* 21: 3546 – 3556

Sittka A, Pfeiffer V, Tedin K, Vogel J (2007) The RNA chaperone Hfq is essential for the virulence of Salmonella typhimurium. *Mol Microbiol* 63: 193 – 217

Sittka A, Lucchini S, Papenfort K, Sharma CM, Rolle K, Binnewies TT, Hinton JC, Vogel J (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet* 4: e1000163

Soper TJ, Doxzen K, Woodson SA (2011) Major role for mRNA binding and restructuring in sRNA recruitment by Hfq. *RNA* 17: 1544 – 1550

Sterzenbach T, Nguyen KT, Nuccio SP, Winter MG, Vakulskas CA, Clegg S, Romeo T, Bäumler AJ (2013) A novel CsrA titration mechanism regulates fimbrial gene expression in Salmonella typhimurium. *EMBO J* 32: 2872 – 2883

Sugimoto Y, König J, Hussain S, Zupan B, Curk T, Frye M, Ule J (2012) Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol* 13: R67

Thomason MK, Fontaine F, De Lay N, Storz G (2012) A small RNA that regulates motility and biofilm formation in response to changes in nutrient availability in *Escherichia coli*. *Mol Microbiol* 84: 17 – 35

Tjaden B, Goodwin SS, Opdyke JA, Guillier M, Fu DX, Gottesman S, Storz G (2006) Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res* 34: 2791 – 2802

Tree JJ, Granneman S, McAteer SP, Tollervey D, Gally DL (2014) Identification of bacteriophage-encoded anti-sRNAs in pathogenic *Escherichia coli*. *Mol Cell* 55: 199 – 213

Updegrove TB, Wartell RM (2011) The influence of *Escherichia coli* Hfq mutations on RNA binding and sRNA*mRNA duplex formation in rpoS riboregulation. *Biochim Biophys Acta* 1809: 532 – 540

Updegrove TB, Shabalina SA, Storz G (2015) How do base-pairing small RNAs evolve? *FEMS Microbiol Rev* 39: 379 – 391

Uren PJ, Bahrami-Samani E, Burns SC, Qiao M, Karginov FV, Hodges E, Hannon GJ, Sanford JR, Penalva LO, Smith AD (2012) Site identification in high-throughput RNA-protein interaction data. *Bioinformatics* 28: 3013 – 3020

Vakulskas CA, Potts AH, Babitzke P, Ahmer BM, Romeo T (2015) Regulation of bacterial virulence by Csr (Rsm) systems. *Microbiol Mol Biol Rev* 79: 193 – 224

Valverde C, Lindell M, Wagner EG, Haas D (2004) A repeated GGA motif is critical for the activity and stability of the riboregulator RsmY of Pseudomonas fluorescens. *J Biol Chem* 279: 25066 – 25074

Vogel J, Bartels V, Tang TH, Churakov G, Slagter-Jäger JG, Hüttenhofer A, Wagner EG (2003) RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res* 31: 6435 – 6443

Vogel J (2009) A rough guide to the non-coding RNA world of Salmonella. *Mol Microbiol* 71: 1 – 11

Vogel J, Luisi BF (2011) Hfq and its constellation of RNA. *Nat Rev Microbiol* 9: 578 – 589

Weilbacher T, Suzuki K, Dubey AK, Wang X, Gudapaty S, Morozov I, Baker CS, Georgellis D, Babitzke P, Romeo T (2003) A novel sRNA component of the carbon storage regulatory system of *Escherichia coli*. *Mol Microbiol* 48: 657 – 670

Weinberg Z, Breaker RR (2011) R2R–software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics* 12: 3

Westermann AJ, Förstner KU, Amman F, Barquist L, Chao Y, Schulte LN, Müller L, Reinhardt R, Stadler PF, Vogel J (2016) Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. *Nature* 529: 496 – 501

Wilson KS, von Hippel PH (1995) Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proc Natl Acad Sci USA* 92: 8793 – 8797

Published online: April 4, 2016
Erik Holmqvist et al Hfq and CsrA CLIP **The EMBO Journal**

Wilusz CJ, Wilusz J (2005) Eukaryotic Lsm proteins: lessons from bacteria. *Nat Struct Mol Biol* 12: 1031 – 1036

Wright PR (2012) hIntaRNA – Comparative prediction of sRNA targets in prokaryotes. Diploma, Albert Ludwig University Freiburg, Freiburg, Germany

Wright PR, Richter AS, Papenfort K, Mann M, Vogel J, Hess WR, Backofen R, Georg J (2013) Comparative genomics boosts target prediction for bacterial small RNAs. *Proc Natl Acad Sci USA* 110: E3487 – E3496

Wright PR, Georg J, Mann M, Sorescu DA, Richter AS, Lott S, Kleinkauf R, Hess WR, Backofen R (2014) CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains. *Nucleic Acids Res* 42: W119 – 123

Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S (2012) FastUniq: a fast *de novo* duplicates removal tool for paired short reads. *PLoS ONE* 7: e52249

Yakhnin AV, Baker CS, Vakulskas CA, Yakhnin H, Berezin I, Romeo T, Babitzke P (2013) CsrA activates flhDC expression by protecting flhDC mRNA from RNase E-mediated cleavage. *Mol Microbiol* 87: 851 – 866

Yao Z, Weinberg Z, Ruzzo WL (2006) CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics* 22: 445 – 452

Zhang A, Wassarman KM, Ortega J, Steven AC, Storz G (2002) The Sm-like Hfq protein increases OxyS RNA interaction with target mRNAs. *Mol Cell* 9: 11 – 22

Zhang C, Darnell RB (2011) Mapping *in vivo* protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol* 29: 607 – 614

Zhang A, Schu DJ, Tjaden BC, Storz G, Gottesman S (2013) Mutations in interaction surfaces differentially impact *E. coli* Hfq association with small RNAs and their mRNA targets. *J Mol Biol* 425: 3678 – 3697

Zheng D, Constantinidou C, Hobman JL, Minchin SD (2004) Identification of the CRP regulon using *in vitro* and *in vivo* transcriptional profiling. *Nucleic Acids Res* 32: 5874 – 5893

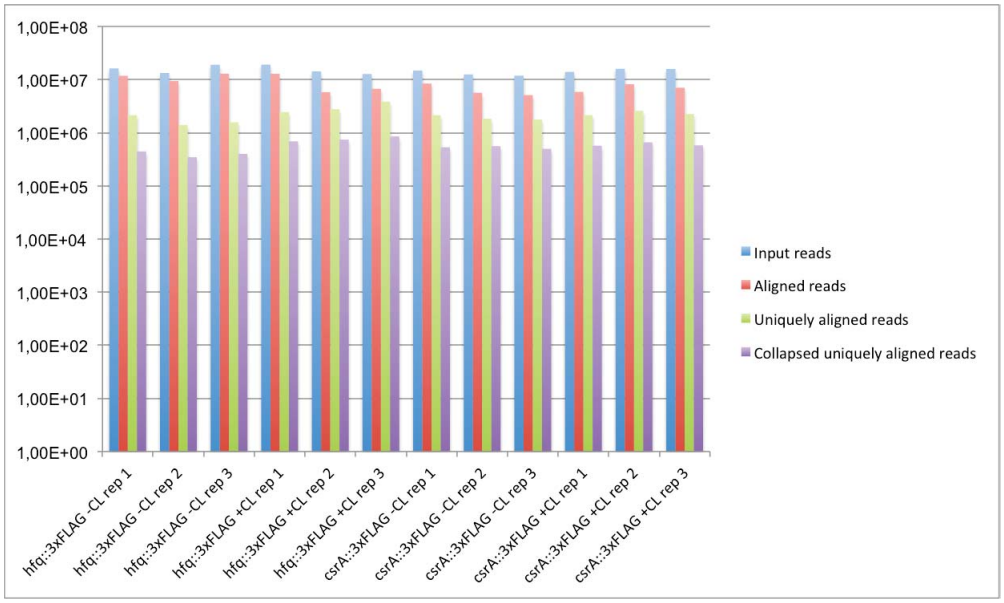
Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31: 3406 – 3415

Appendix

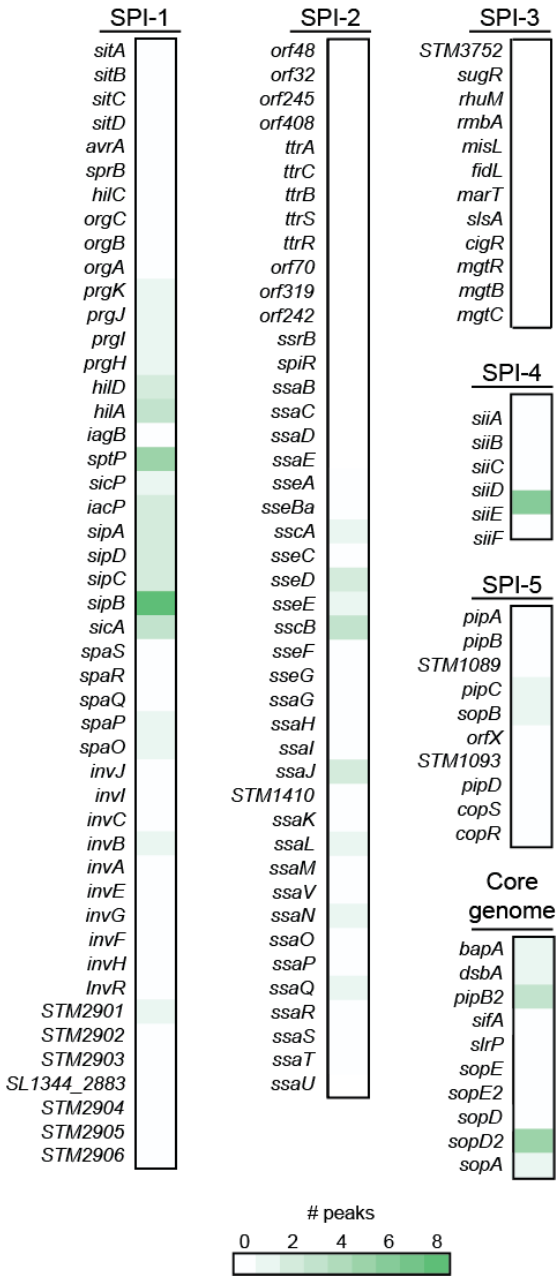
Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking *in vivo*

Erik Holmqvist, Patrick R. Wright, Lei Li, Thorsten Bischler, Lars Barquist, Richard Reinhardt, Rolf Backofen, Jörg Vogel

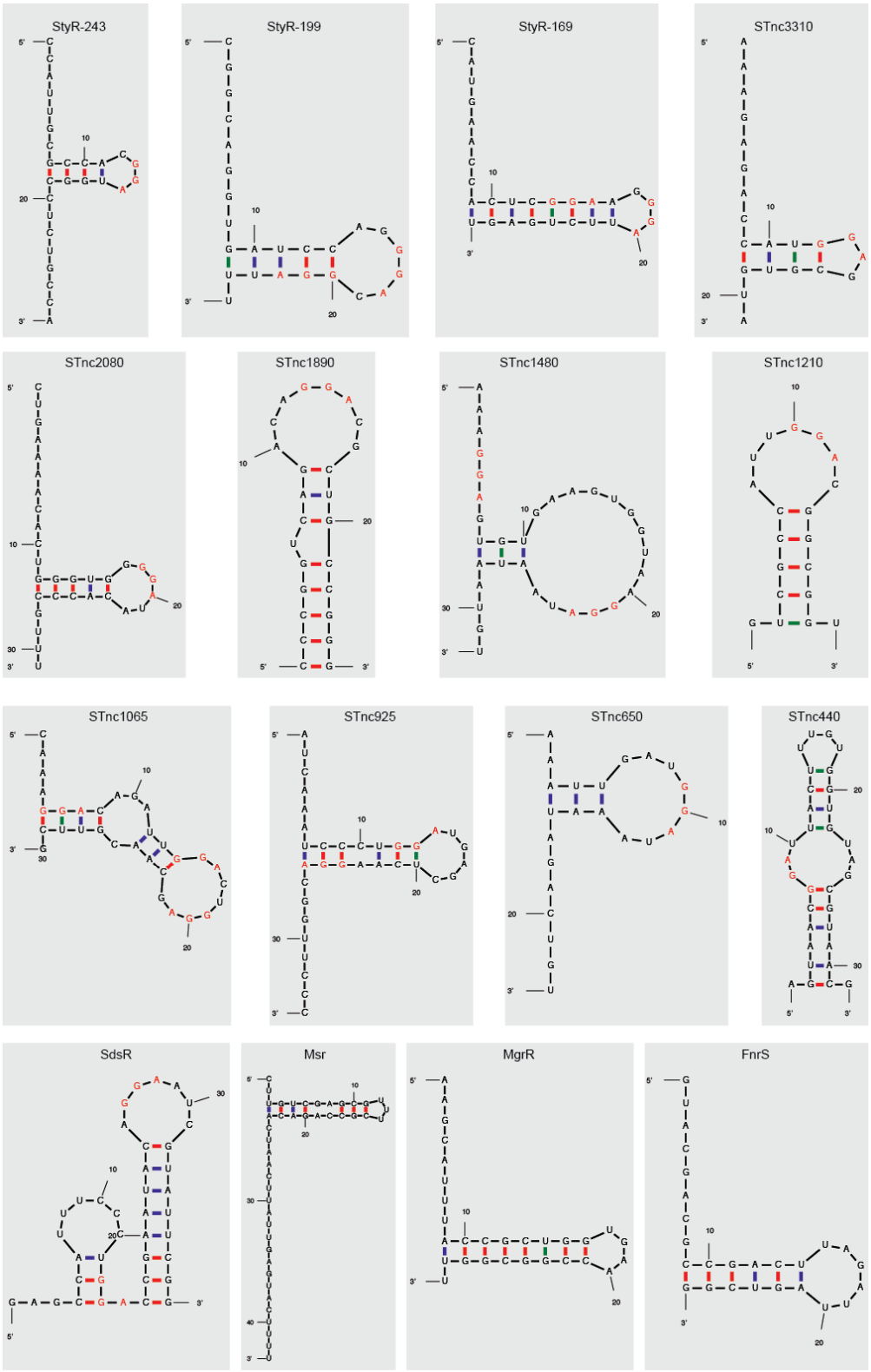
Table of contents:



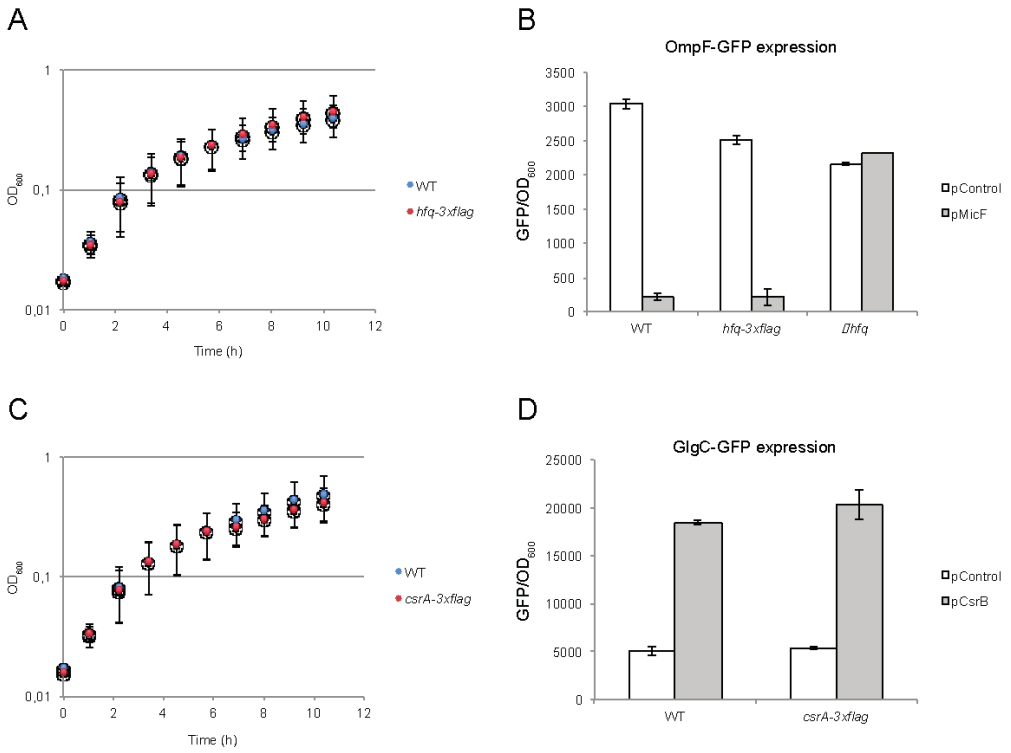
Appendix Fig. S1. Sequencing reads obtained for each library used in this study.



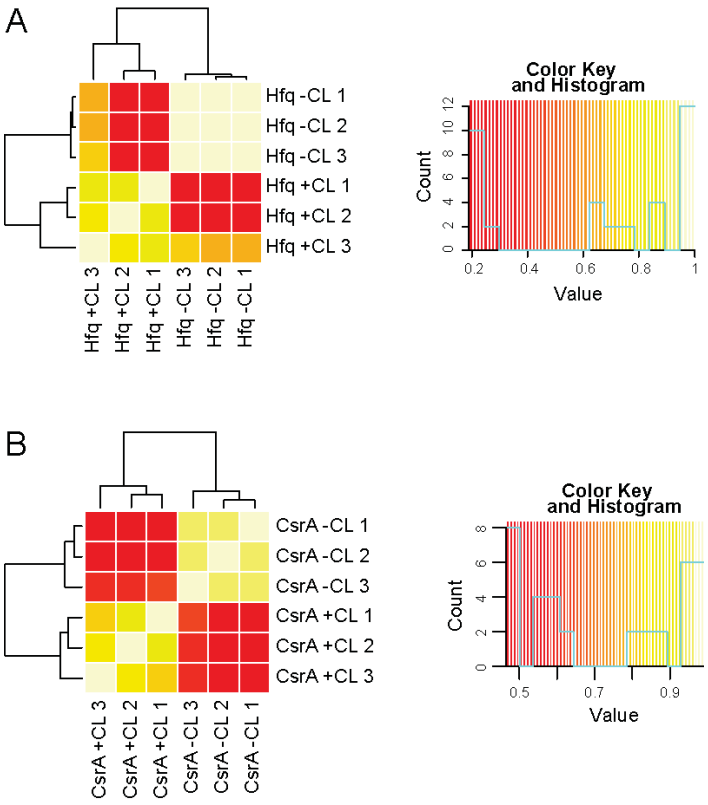
Appendix Fig. S2. CsrA peak distribution over Salmonella pathogenicity islands and virulence factors encoded on the core genome.



Appendix Fig. S3. Secondary structure predictions of CsrA peak sequences in sRNAs. GGA motifs are highlighted in red. Structure predictions were made with MFOLD.



Appendix Fig. S4. Addition of a FLAG-tag to Hfq or CsrA does not impair protein function. A) Growth of wild type and *hfq::3xflag* strains. B) OmpF-GFP expression in the presence or absence of MicF overexpression was monitored in wild type, *hfq::3xflag*, and Δ*hfq* strain backgrounds. C) Growth of wild type and *csrA::3xflag* strains. D) GlgC-GFP expression in the presence or absence of CsrB overexpression was monitored in wild type and *csrA::3xflag* strain backgrounds. In each panel, error bars show standard deviations determined from three independent experiments.



Appendix Fig. S5. Pearson correlations calculated on nucleotide read coverage from CLIP-seq experiments on Hfq (A) and CsrA (B).

Appendix Supplementary Methods

Construction of bacterial strains and plasmids

The *csrA::3xFLAG::* Km^R and Δ*spf::*Km^R alleles were constructed using the Lamdba Red system with PCR products amplified from pSUB11 or pKD4, respectively. To construct the Δ*csrB*Δ*csrC::*Km^R strain, the *csrB* and *csrC* alleles were first separately deleted using the Lambda Red system with PCR products amplified from pKD4. The resulting Δ*csrB::*Km^R strain was healed from the Kanamycin resistance with pCP20 followed by transduction of the *csrC::*Km^R allele using phage P22. The CsrB overexpression plasmid pCsrB was constructed by blunt/XbaI cloning of PCR products (JVO-10759/JVO-10760) into pZE12-luc as previously described (Urban and Vogel 2007). Plasmid pBAD-Spot42 was constructed through cloning of a PCR product (JVO-294/JVO-930) into a pBAD backbone as described previously (Papenfort et al. 2006). GFP-fusion plasmids where constructed as described previously (Corcoran et al. 2012) using primers listed in Table S1. Plasmids pEH728, pEH731 and pEH734 were generated through site-directed mutagenesis on plasmids pFS-102, pJV765-18 and pEH456, respectively, using primers listed in Appendix Table S1.

Fluorescence measurements

Bacterial growth and fluorescence measurements were essentially carried out as previously described (Holmqvist et al. 2013). Briefly, overnight cultures of bacterial strains were diluted 1:100 in fresh M9 medium, transferred to a 96-well plate (100 µl per well), and incubated at 37°C with shaking in an Infinite M-200 plate reader (Tecan) controlled by the i-control software (Tecan). The optical density was measured with the following parameters: wavelength 600 nm, bandwidth 9 nm. Fluorescence (GFP) was monitored with excitation wavelength 480 nm, excitation bandwidth 9 nm, emission wavelength 520 nm, and emission bandwidth 20 nm. Measurements were taken at 7 min intervals.

Iterative decomposition of blockbuster clusters into peaks

The output of blockbuster is a set of clusters C, each consisting of a set of read blocks. With $b(B)$, $e(B)$ and $l(B)$ we denote the left end, right end and the length of block B. The size of a block $S(B)$ is the number of reads assigned by blockbuster to this block. The size of a cluster is defined as the sum of block sizes, i.e., $S(C) = \sum_{B \in C} S(B)$. We decompose each cluster into peaks using the following iterative procedure. While a cluster C still contains non-processed blocks, we select the largest block B_m in C, i.e.,

$$B_m = \underset{\substack{B \in C \\ S(B) \geq 0.01 S(C)}}{\operatorname{argmax}} S(B)$$

The peak is defined by selecting all blocks in C that overlap with B_m. These blocks are then removed from C. However, for the final peak boundaries, we only consider those overlapping blocks that overlap with at least half of B_m. In more detail, we consider an overlapping block B only if $b(B) \leq b(B_m) + \left\lfloor \frac{1}{2} l(B_m) \right\rfloor$ and $e(B) \geq e(B_m) - \left\lfloor \frac{1}{2} l(B_m) \right\rfloor$. Furthermore, the block B has to satisfy a size restriction according to the size of B_m. Thus, we consider local frequency of reads and hence exclude putative candidates for noise. Specifically, we have chosen to consider only blocks that satisfy $S(B) \geq 0.1 S(B_m)$.

Appendix Table S1. Oligodeoxyribonucleotides used in the study.

Oligo name	Sequence (5' to 3')	Used for
JVO-03591	Ccagcgtatccaggctgaaaaatcccagcagtccagttacgactacaaagaccatgacgg	3xFLAG tagging of *csrA*
JVO-03592	Accatatcaacagtgaggttgaaaaaagtcatgaagggacccatatgaatatcctccttag	3xFLAG tagging of *csrA*
JVO-00112	Tgaaaatctggcgcgaagaataacaaaaaaaagggagcactgtatgtgtaggctggagctgcttc	Deletion of *csrB*
JVO-00113	Aaggtattgtctgtaagcgtcttgtaagacaaggtgaaacaggcggtccatatgaatatcctccttag	Deletion of *csrB*
JVO-00117-B	Aaaagggatttgccgtgtcggtatcttgtgagtttaccccaaaagagtgtaggctggagctgcttc	Deletion of *csrC*
JVO-00117-C	ggcggaatctagcagaaagcaagcaaagaaaaaaggcgacagaggtccatatgaatatcctccttag	Deletion of *csrC*
JVO-10759	Gtcgacagggagtcgtacaacg	Cloning of *csrB* in pZE12-luc, fwd
JVO-10760	gtttttctagaatgagtcgtcatgttaaaaacctcaatga	Cloning of *csrB* in pZE12-luc, rev
JVO-00291	ctatgtaaagaatgaaaaaaaattacgaaaaggtctttctgtgttgctcagtgtaggctggagctgcttc	Deletion of *spf*
JVO-00292	ggatggcgccgctgcgcgtcttatccggcctacggtgtgagcgaaactttggtccatatgaatatcctccttag	Deletion of *spf*
JVO-00294	gtttttttctagagcaccggtcgaagagat	Cloning of *spf* in pBAD, fwd
JVO-00930	5'p-gtagggtacagaggtaagatgttc	Cloning of *spf* in pBAD, rev
JVO-01476	Atcgtcgattccggtctttg	qRT-PCR *mglB*, fwd
JVO-01477	Ctgattgttggcgtcattcagt	qRT-PCR *mglB*, rev
JVO-01117	Tcagccattttgtgcgctt	qRT-PCR *rfaH*, fwd
JVO-01118	ttcaggatcgacaacgcctt	qRT-PCR *rfaH*, rev
JVO-02846	gtttttatgcatgttttgaacgccgctag	Cloning of *mglB* in pXG10-SF, fwd
JVO-02847	gtttttgctagcgaataacagacttgccatcac	Cloning of *mglB* in pXG10-SF, rev
JVO-10553	gtttttatgcattatagttttacccgaaggggaatgg	Cloning of *glgC* in pXG10-SF, fwd
JVO-10554	gtttttgctagctacacgatcgttcttctctaaactcac	Cloning of *glgC* in pXG10-SF, rev
JVO-10573	gtttttatgcataaatagagtgtggttttaatcaaaaaatgaga	Cloning of *sopD2* in pXG10-SF, fwd
JVO-10574	gtttttttctagatataagcatattgcgacaactcgac	Cloning of *sopD2* in pXG10-SF, rev
JVO-13333	gtttttatgcatatggacgatgttaccgcagca	Cloning of *prgHI* in pXG30-SF, fwd
JVO-13334	gtttttgctagcacggaagttctgaataatggcagca	Cloning of *prgHI* in pXG30-SF, rev
JVO-13335	gtttttatgcatgcaacaccttggtcaggctatc	Cloning of *prgIJ* in pXG30-SF, fwd
JVO-13336	gtttttgctagctgagcgtaatagcgtttcaacagcc	Cloning of *prgIJ* in pXG30-SF, rev
JVO-13337	gtttttatgcattcgattgcaactattgtccctgaga	Cloning of *prgJK* in pXG30-SF, fwd
JVO-13338	gtttttgctagcagctcgcggagacgatacc	Cloning of *prgJK* in pXG30-SF, rev
JVO-10557	gtttttatgcatatcagataaacgcagtcgttaagttctac	Cloning of *sicA* in pXG10-SF, fwd
JVO-13349	gtttttgctagcttccttttcttgttcactgtgctgc	Cloning of *sicA* in pXG10-SF, rev
JVO-13341	gtttttatgcatggattaggcgtcgataagaaaacgg	Cloning of *sipBC* in pXG30-SF, fwd
JVO-13342	gtttttgctagcagcgcgaatattgcctgcg	Cloning of *sipBC* in pXG30-SF, rev
JVO-13343	gtttttatgcatatggatatgacccgaatcgatgcg	Cloning of *sipCD* in pXG30-SF, fwd
JVO-13344	gtttttgctagctccttgcaggaagcttttggc	Cloning of *sipCD* in pXG30-SF, rev
JVO-13345	gtttttatgcatctgaaatcttatggatccggttatgtc	Cloning of *sipDA* in pXG30-SF, fwd
JVO-13346	gtttttgctagcctgtttgatcagcgcgggaaaa	Cloning of *sipDA* in pXG30-SF, rev
JVO-14260	gctatagggtacataataaaaccggag	Mutagenesis of plasmid pFS-102
JVO-14261	tatgtaccctatagcgtaaaaaaatgcc	Mutagenesis of plasmid pFS-102
JVO-14262	tccgtaccctacagaggtaagat	Mutagenesis of plasmid pJV765-18
JVO-14263	tctgtagggtacggagaaacag	Mutagenesis of plasmid pJV765-18
JVO-14264	agcaaccctttccttcttgctttcgcggtaaataat	Mutagenesis of plasmid pEH456
JVO-14265	gcaagaaggaaagggttgctctcatttttttgattaaaacca	Mutagenesis of plasmid pEH456
JVO-11007	aatgatacggcgaccaccg	cDNA library preparation
JVO-11008	caagcagaagacggcatacg	cDNA library preparation

Appendix Table S2. Bacterial strains used in the study.

Strain	Relevant markers/genotype	Reference/source
SL1344	*Salmonella typhimurium*, Str^R *hisG rpsL xyl*	(Stocker et al. 1983)
JVS-1338	SL1344 *hfq::3xFLAG*	(Pfeiffer et al. 2007)
JVS-0584	SL1344 Δ*hfq*	(Sittka et al. 2007)
JVS-4317	SL1344 *csrA::3xFLAG* Km^R	This study
JVS-0129	SL1344 Δ*csrB* Δ*csrC*::Km^R	This study
JVS-0118	SL1344 Δ*spf*::Km^R	This study
JVS-3902	SL1344 *sopD2::3xFLAG* Km^R	(Papenfort et al. 2012)

Appendix Table S3. Plasmids used in the study.

Plasmid trivial name	Plasmid stock name	Cloned fragment	Origin/marker	Reference/source
pBAD	pKP8-35		pBR322, AmpR	(Papenfort et al. 2006)
pBAD-Spot42	pJV765-18	*spf*	pBR322, AmpR	This study
	pJV300		ColE1, AmpR	(Sittka et al. 2007)
pCsrB	pEH476	*csrB*	ColE1, AmpR	This study
	pXG-0	*luc*	pSC101*, CmR	(Urban and Vogel 2007)
	pXG10-SF	*lacZ*	pSC101*, CmR	(Corcoran et al. 2012)
pMglB-GFP	pFS102-1	*mglB*	pSC101*, CmR	This study
pGlgC-GFP	pEH451	*glgC*	pSC101*, CmR	This study
pSopD2-GFP	pEH456	*sopD2*	pSC101*, CmR	This study
pSicA-GFP	pEH646	*sicA*	pSC101*, CmR	This study
pSipBC-GFP	pEH683	*sipB-sipC*	pSC101*, CmR	This study
pSipCD-GFP	pEH651	*sipC-sipD*	pSC101*, CmR	This study
pSipDA-GFP	pEH652	*sipD-sipA*	pSC101*, CmR	This study
pPrgHI-GFP	pEH648	*prgH-prgI*	pSC101*, CmR	This study
pPrgIJ-GFP	pEH649	*prgI-prgJ*	pSC101*, CmR	This study
pPrgJK-GFP	pEH650	*prgJ-prgK*	pSC101*, CmR	This study
pMicF	pDP31	*micF*	ColE1, AmpR	(Corcoran et al. 2012)
pOmpF-GFP	pDP23	*ompF*	pSC101*, CmR	(Corcoran et al. 2012)
pBAD-Spo42*	pEH731	*spf*	pBR322, AmpR	This study
pMglB*-GFP	pEH728	*mglB*	pSC101*, CmR	This study
pSopD2-2xCCU-GFP	pEH734	*sopD2*	pSC101*, CmR	This study

sRNA sequences used for CopraRNA target predictions:

ArcZ
>NC_000913
gugcggccu gaaaaacagu gcu gu gcc cuu guaacucau cauaauaauuu ac ggc gca gc caa gauuuccc u gguguu ggc gca guauuc gc gcacc
ccggucuagcc ggggucauuuuuu
>NC_016810
gugcggccu gaaaacaggacu gc gc cuuu gacaucau cauaauaagcac ggc gca gcca c gauuuccc u gguguu ggc gca guauu c gc gcacc cc g
gucaaaccggggucauuuuuu
>NC_009792
gugcggccu gaaaagca gagcu gc gc cu gu guaaaaaacaaucau aacuu ac ggc gca gc cac gauuuccc u gguguu ggc gca guauuc gc gcacc c
cggucaauccggggucauuuuuu
>NC_013716
gugcggccu gaaaau ga gc gcu gc gc guuuaaaau au gagaauaacuuac c gc gca gcuac gauuuccc u gguguu ggc gca guauu c gc gcacc cc
gguuaaauccggggucauuuuuu
>NC_009778
gagcgccuugaucacucaggcc gc gcca c gua guuaaaaagauc aacaucauu aaucaau ggc gaggcc ac gauuu cccu gguguu ggc gca gu auu
cgcgcacccc gguuuagcc ggggucauuuuuu
>NC_009436
gugcggccucauucuuaaggcagc gucc au gc gaaaacaacacaaagaaucaggc gc ggcca c gauuuccc u gguguu ggc gca guauuu c gc gcacc
ccgguuuauccggggucauuuuuu
>NC_011740
gugcggccu gaaaaacagu gcu gc gc cuu gguuacaaac gacaauaauuuac ggc gca gcca u aauuuccc u gguguu ggc gca guauu c gc gcacc
ccgguuaauccggggucauuuuuu
>NC_009648
gcgagcaauccacuucuu c gguu gc gc cac gu aacaacauca cucaaacaacacu ggcu caacca cca guucccu gguguu ggc gca gu auuc gc gca
ccccggucugucc ggggu cauuuuuu
>NC_012917
uuaagacacgaauauc c gca cuu gc ga guuu acaaaaccu gaaaucuaaau gca ggugcau guuuuccc u gguguu ggc gcauaaauu c gc gcacc cc
ggcuucggcc ggggu cauuuuuu
>NC_005126
ugauguacggagaauuc c ucauuu acuc gcc gcaac caaagauau auca gaua ggc gu gu aaaaa guuuuccc u gguguu ggc gca guauu c gc gca
ccccaaccucgguu ggggguuauuuuuu
>NC_010554
augauguauggaauagcuu cauc c uauuc gccuau gu aau gau aauc aaaaaa gc ga gua gaaaa guuuccc u gguguu ggc gca guauuc gc gcac
cccaaccuuc gguu ggggguuauuuuuu
>NC_003197
gugcggccu gaaaacaggacu gc gc cuuu gacaucau cauaauaagcac ggc gca gcca c gauuuccc u gguguu ggc gca guauu c gc gcacc cc g
gucaaaccggggucauuuuuu
>NC_009832
cuguaugauguuuaggaauucu cuac aac cu gcuc gacca gaacauuaaaacca auac gca gguucacaauuuccc u gguguu ggc gcaauauu c gc
gcaccccggccua gguc ggggu cauuuuuu
>NC_007606
gugcggccu gaaaaacagu gcu gu gcc cuu guaacucau cauaauaauuu ac ggc gca gc caa gauuuccc u gguguu ggc gca guauuc gc gcacc
ccggucuagcc ggggucauuuuuu
>NC_004337
gugcggccu gaaaaacagu gcu gu gcc cuu guaacucau cauaauaauuu ac ggc gca gc caa gauuuccc u gguguu ggc gca guauuc gc gcacc
ccggucuagcc ggggucauuuuuu
>NC_007384
gugcggccu gaaaaacagu gcu gu gcc cuu guaacucau cauaauaauuu ac ggc gca gc caa gauuuccc u gguguu ggc gca guauuc gc gcacc
ccggucuagcc ggggucauuuuuu
>NC_007712
uuuauucaugaaugcuggac gucuuuuc ccc gu cccu caacc gcc guuaa gu aau gac gac gauaua ca guuuccc u gguguu ggc gcaau cuuc gc
gcaccccggca gcau aa guc ggggu cauuuuuu
>NC_008800
gacuggg gu gaac gaa ggca gcc aac gca cau gc aacuu gaa guau gac gggu auu gca gguu aac gauuu cccu gguguu ggc gca gu auuc gc gc
accccggccuc ggucggggu uauuuuuu
>NC_003143
guaugauguaugaaagaauccu gacaaccu gc gaauu cacu c gaaauc gaaaua auac gca gguu aac guuuuccc u gguguu ggc gca gucuu c gc
gcaccccggccuc ggucggggu uauuuuuu
>NC_006155
guaugauguaugaaagaauccu gacaaccu gc gaauu cacu c gaaauc gaaaua auac gca gguu aac guuuuccc u gguguu gac gca gu cuuc gc
gcaccccggccuc ggucggggu uauuuuuu

ChiX
>NC_000913

acaccgucgcuuaaagugacggcauaauaauaaaaaauugaaauuccucuuugacgggccaauagcgauauuggccauuuuuuu
>NC_016810
gauccgaagcgaaagcgucgggauaauaauaacgaugaaauuccucuuugacgggccaauagcgauauuggccauuuuuuu
>NC_003197
gauccgaagcgaaagcgucgggauaauaauaacgaugaaauuccucuuugacgggccaauagcgauauuggccauuuuuuu
>NC_009436
caaccgagggucuccuucggcauaauaauaacgaugaaauuccucuuugacgggccaauagaaauauuggccauuuuuuu
>NC_009792
uaaccagggcgcuacgucccuggcauaaauaauaacgaugaaauucucucuuugacgggccaauagaaauauuggccauuuuuuu
>NC_013716
acaccgucgcuuaaagcggcggcauaacaauaaugaugaaauuccucuuugacgggccaauagcgauauuggccauuuuuuu
>NC_009778
aaccguccgcuaaggcgcacggcauaacgacaauaacgaaaaguuccucuuugacgggccaguagcgauacuggccuucuuuuu
>NC_011740
acaccgucgcuuaaagugacggcauaauaauaaaaaauugaaauuccucuuugacgggccaauagcgauauuggccauuuuuuu
>NC_009648
gauccgggaugcaaauccccgggauaauaauaaugaugaaaauuccucuuugacgggccaauagcaauauuggccauuuuuuu
>NC_009832
cguuaacagggguuacaaugcguauaacuacaauacaagaaaauuccucuuugacuggccaguagcgauauuggccacuuuuuu
>NC_007613
acaccgucgcuuaaagugacggcauaauaauaaaaaauugaaauuccucuuugacgggccaauagcgauauuggccauuuuuuu
>NC_007606
acaccgucgcuuaaagugacggcauaauaauaaaaaauugaaauuccucuuugacgggccaauagugauauuggccauuuuuuu
>NC_004337
acaccgucgcuuaaagugacggcauaauaauaaaaaauugaaauuccucuuugacgggccaauagcgauauuggccauuuuuuu
>NC_007384
acaccgucgcuuaaagugacggcauaauaauaaaaaauugaaauuccucuuugacgggccaauagcgauauuggccauuuuuuu

CyaR
>NC_000913
gcugaaaaacauaacccauaaaauugcuagcuguaccaggaaccacccuccuuagccugugguaaucucccuuacacgggcuuauuu
>NC_016810
gcugaaaaacauaacccauaaaaugcuagcuguaccaggaaccaccuccuuggccugcguaaucucccuuacgcaggcuuauuuuuu
>NC_009792
gcugaaaaacauaacccauaaaaugcuagcuguaccaggaaccaccuccuuagccugcguaaucucccuuacgcaggcuuauuuuuu
>NC_013716
gcugaaaaacauaacccauaaaaugcuagcuguaccaggaaccaccuccuuggccugcguaaucucccuuacgcaggcuuauuu
>NC_009778
gcugaaaaacauaauccauaaaugcccguuguaccaggaaccaccuccuucagccugcguaaucucccuuacgcaggcuuauuu
>NC_009436
gcugaaaaacauaacccauaaaaugcuagcuguaccaggaaccaccuccuuagccugcguaaucucccuuacgcaggcuuauuuuuu
>NC_011740
gcugaaaaacauaacccauaaaaugcuaagcuguaccaggaaccaccuccuuagccugcguaaucucccuuacgcaggcugauuu
>NC_009648
gcugaaaaacauaacccauaaaaugcuaguuguaccaggaaccaccuccuuagccugcguaaucucccuuacgcgggcuuauuu
>NC_012917
gcugagaaaacauagaacgaaaaauaagcguagugauacuacuaggaaccaccuccuuggccagcucaaucucccuugagcuggcuuuucu
>NC_005126
gguguaaaaauaaaaauuaaaaaaguuuuuacaguaagacuaggaaccaccuccuuggccggcucaaucucccuugagcuggccuuuucuuu
>NC_003197
gcugaaaaacauaacccauaaaaugcuagcuguaccaggaaccaccuccuuggccugcguaaucucccuuacgcaggcuuauuuuuu
>NC_009832
gcuuaaaacuaagaacgaaaaauauuugcauagcaauacuaggaaccaccuccuuagccggucccaaucucccuuggacuggcuuuuuc
>NC_007613
gcugaaaaacauaacccauaaaaugcuagcuguaccaggaaccaccuccuuagccugugguaaucucccuuacacgggcuuauuu
>NC_007606
gcugaaaaacauaacccauaaaaugcuagcuguacuaggaaccaccuccuuagccugugguaaucucccuuacacgggcuuauuu
>NC_004337
gcugaaaaacauaauccauaaaaugcuagcuguaccaggaaccaccuccuuagccugugguaaucucccuuacacgggcuuauuu
>NC_007384
gcugaaaaacauaacccauaaaaugcuagcuguaccaggaaccaccuccuuagccugugguaaucucccuuacacgggcuuauuu
>NC_008800
gaguacaaucaagaacuaaaaaaguuuuaaagcaagacuaggaaccaccuccuuggccaacccaaucucccuuggggcuggccuuuuucuuu
>NC_003143
aguacaaucaauaacuaaaaaaguguuaaguaaauacuaggaaccaccuccuuggcuagcccaaucucccuugggcuggccuuuuc
>NC_006155
aguacaaucaauaacuaaaaaaguguuaaguaaauacuaggaaccaccuccuuggcuagcccaaucucccuugggcuggccuuuuc

DsrA

>NC_000913
aacacaucagauuuccuggugu aacgaauuuuuuaagugcuucuugcuuaagcaaguuucaucccgacccccucagggucgggauuuuu
>NC_016810
cucacaucagauuuccuggugu aacgaauuuucaagugcuucuugcauaagcaaguuugaucccgacccguaagggccgggauuuuu
>NC_003197
cucacaucagauuuccuggugu aacgaauuuucaagugcuucuugcauaagcaaguuugaucccgacccguaagggccgggauuuuu
>NC_009792
cgcacaucagauuuucuggugu aacgaauuuucaagugcuucuugcauaagcaaguuugauccggcucugcgagccgggauuuuu
>NC_013716
cgcacaucagauuuccuggugu aaugaauuuucaagugcuucuugcaucagcaaguuuaaucccgacccgucagggucgggauuuuu
>NC_009778
ucgacauccguuucccuggugu aacgaauuuuaagugcuucuugcuuucgcaagcuuaucccggcucccccaggccgggauauuu
>NC_009436
cccacaucagauuuccuggugu aacgaauuuacaagugcuucuugcauaagcaaguucaucccggucaucccccaugggccgggauuuuu
>NC_011740
cgcacaucagauuuccuggugu aacgaauuuucaagugcuucuugcauaagcaaguuuccucccgauccuucacaggaucgggauuuuu
>NC_009648
aacgcaucggauuucccggugu aacgaauuuucaagugcuucuugcauuagcaaguuugauccgacuccugcgagucgggauuuuu
>NC_007613
aacacaucagauuuccuggugu aacgaauuuuuuaagugcuucuugcuuaagcaaguuucaucccgacccccucagggucgggauuuuu
>NC_007606
aacacaucagauuuccuggugu aacgaauuuuuuaagugcuucuugcuuaagcaaguuucaucccgacccccucagggucgggauuuuu
>NC_004337
aacacaucagauuuccuggugu aacgaauuuuuuaagugcuucuugcuuaagcaaguuucaucccgacacccucagggucgggauuuuu
>NC_007384
aacacaucagauuuccuggugu aacgaauuuuuuaagugcuucuugcuuaagcaaguuucaucccgacccccucagggucgggauuuuu

FnrS

>NC_000913
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuugaauuacugcauagcacaauugauucguacgacgcc
gacuuugaugagucggcuuuuuuuuu
>NC_016810
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuugaauuacugcauagcacaauugauucguacgacgcc
gacuuagauuagucggcuuuuuuuuu
>NC_003197
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuugaauuacugcauagcacaauugauucguacgacgcc
gacuuagauuagucggcuuuuuuuuu
>NC_009792
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuugaauuacugcauagcacaauugauucguacgacgcc
gacuuagauuagucggcuuuuuuuuu
>NC_013716
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuugaauuacugcauagcacaauugauucguacgacgcc
gacuuaaauaagucggcuuuuuuuuu
>NC_009778
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuugaauuacugcauagcacaauugauucguacgaugcc
gacuuguuuaagucggcuuuuuuuug
>NC_009436
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuugaauuacugcauagcacaauugauucguacaaugcc
gacuaaauuagucggcuuuuuuuuu
>NC_011740
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuugaauuacugcauagcacaauugauucguacgacgcc
gacuuagauaagucggcuuuuuuuuu
>NC_009648
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuugaauuacugcauagcacaauugauucguacgacgcc
ggcuuuguugagucggcuuuuuuuuu
>NC_012917
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuugaauuacugcauagcacaauugauucacaccaugcc
gacaguuuugucggcuuuuuuuuu
>NC_005126
gcaggugaaaacaacguuaagcgaugaacguuguucucccguaaauuguaguuuuuucucacauaagucuuuauacagaauaauugccccauucuugugcc
gauauuauuuuuauaauaucggcuuuuuuuua
>NC_010554

gcaggugaau acaacguuga gcgaugaacguugugcuccauaauuguaguuuuucucauauugaguucuuaaauacagaauaauugaccauuauuacaccgauguuaaaauaaacaucgguuuuuuuuu
>NC_009832
gcaggugaaugcaacgucaagcgaugggcguugugcuccauaauugucuuacuuucuuauauuuagaauuacugcauagcacauugauucauacgaugccgguuuaaaucaccggcauuuuuuuu
>NC_007606
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuggaauugcugcauagcacaauugauucguacgacgccgacuuugaugagucggcuuuuuuuu
>NC_004337
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuugaauuacugcauagcacaauugauucguacgacgccgacuuugaugagucggcuuuuuuuu
>NC_007384
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauugucuuacuuccuuuuuugaauuacugcauagcacaauugauucguacgacgccgacuuugaugagucggcuuuuuuuu
>NC_007712
gcaggugaaugcaacgucaagcgaugggcguauugucuuacuuucuuuuugaauuauugcauagcguauuuaauucagaugaugccgggugaauagcacggcauuuuuuuu
>NC_008800
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauuuguaauacuuuuauuuuugaauuacugcauagcacauaauuguuacgaugccgauguugucucaacaucggcauuuuuuuu
>NC_003143
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauuuguuauacuuuuauuuugaauuacugcauagcauuuaauuguuacggugccgauguugaaaaacaucggcauuuuuuuu
>NC_006155
gcaggugaaugcaacgucaagcgaugggcguugcgcuccauauuuguuauacuuuuauuuugaauuacugcauagcauuuaauuguuacggugccgauguugaaaaacaucggcauuuuuuuu

GcvB
>NC_000913
acuuccugagccggaacgaaaaguuuuaucggaaugcguguucuggugaacuuuuggcuuacgguugugauguugugugugugcaauugucugcgauucagaccaugguagcaaagcuaccuuuuuuucacuucuguacauuuacccugucugccauagugauuaauguagcaccgccuaauugcggugcuuu
>NC_016810
acuuccugagccggaacgaaaaguuuuaucggaaugcguguucugaugggcuuuuggcuuacgguugugauguugugugugcaauugucugcgauucagaccacgguagcgagacuacccuuuuucacuuccuguacauuuacccugucugcauagugauuaauguagcaccgccauauugcggugcuuu
>NC_009792
acuuccugagccggaacgaaaaguuuuaucggaaugcguguucugaugggcuuuuggcuuacgguugugauguugugugugcaauugucugcgauucagaccaugguagcgaagcuaccuuuuuucacuuccuguacauuuacccugucugccauagugauuaauguagcaccgccaauugcggugcuuu
>NC_013716
acuuccugagccggaacgaaaaguuuuaucggaaugcguguucuggugggcuuuuggcuuacgguugugauguugugugugcaauugucugcgauucagaccacgguagcgaagcuacccuuuuucacuuccuguacauuuacccugucugcauagugauuaauguagcaccgccuauugcggugcuuu
>NC_009778
acuuccugagccggaacgaaaagucuuuaagaaugagaguucuggagggcuuuuggcuuacgguugugauguugugugugcauuuggucugcgauucagaccaucguagcuaagcuacucuuucacuuccuguacauuuacccugucugccauagugauuaauguagcaccgcacuuugcggugcuuu
>NC_009436
acuuccugagccggaacgaaaagcuuuuuuuuggaaugcgguguucugaagggcuuuuggcuugcgguugugauguugugugugcaauuuggucugcuauucagaucaugguagcaaagcuacccuuuuuucacuuccuguacauuuacccugucugccauaguuaauguagcaccgccuauuugcggugcuuu
>NC_011740
acuuccugagccggaacgaaaaguuuuaucggaaugcguguucuggugaacuuuuggcuuacgguugugauguugugugugcaauugucugcgauucagaccacgguagcauagcuacccuuuuucacuuccuguacauuuacccugucugccauagugauuaauguagcaccgccuuauugcggugcuuu
>NC_009648
acuuccugagccggaacgagaagcuuuuuuuuggaaugcguguuccaucaagcuuuuggcuuacgguugugauguugugugugcaauuggucuguuuuugcagacccugguagcaaaagcuaccccuuuucacuuccuguacauuuacccugucugccauaguuuuaauguagcaccgcaacgcgcggugcuuu
>NC_012917
acuuccugggccggaacgaaaagugcggaugggugaccugaggugcuuuuggcuugugguugugauguugugugcuauuuguuugucugccuuuugcagauguggguagcgagucuacccuauucacuuccuguacauuuacccugucugccaaagugauuuuuuugugugagcaccgcaaauugcggugcuuu
>NC_005126

acuuccuaagccggaacgaaaagugaaucagcuuaacgcuaugaaacuuuuggcuuugugguugugauguuguguuugcaaguugucugggaaacc
ggacccuguagcucaagcuacuguuuuucacuuccuguacauuuacccugucuguccauagugauuuuuuaaugcaccgcaguugcggugcuuu
>NC_010554
acuuccuaagccggaacgaaaaguuguguaguagcaauacugcaaaacuuuuggcuuugugguugugauguuguguuugcaaguugucuggaaauu
ccagacuuuguagcugaugcuacuguuuuuuuucacuuccuguacauuuacccugucuguccauagugauuuuauaaugcaccgccaaggcggugc
uuu
>NC_003197
acuuccugagccggaacgaaaaguuuuaucggaaugcguguucugaugggcuuuuggcuuacgguugugauguuguguuguuguguuugcaauug
gucugcgauucagaccacgguagcgagacuacccuuuuucacuuccuguacauuuacccugucuguccauagugauuaauguagcaccgccauauug
cggugcuuu
>NC_009832
acuucccgggccggaacgaaaagggagugggugucgaggacgccgaugaacuuuuggcuugugguugugauguuguguuugcaaauugucuggcaa
uccagacgugguagcuaagcuacuguuuuucacuuccuguacauuuacccugucuguccauagugauuuuaugcagcaccgcaaucuuugcggugcu
uu
>NC_007606
acuuccugagccggaacgaaaaguuuuaucggaaugcguguucuggugaacuuuuggcuuacgguugugauguuguguuguuguguuugcaauug
gucugcgauucagaccaugguagcaaagcuaccuuuuuucacuuccuguacauuuacccugucuguccauagugauuaauguagcaccgccuaauug
cggugcuuu
>NC_004337
acuuccugagccggaacgaaaaguuuuaucggaaugcguauucuggugaacuuuuggcuuacgguugugauguuguguuguuguguuugcaauug
gucugcgauucagaccaugguagcaaagcuaccuuuuuucacuuccuguacauuuacccugucuguccauagugauuaauguagcaccgccuaauug
cggugcuuu
>NC_007384
acuuccugagccggaacgaaaaguuuuaucggaaugcguguucuggugaacuuuuggcuuacgguugugauguuguguuguuguguuugcaauug
gucugcgauucagaccaugguagcaaagcuaccuuuuuucacuuccuguacauuuacccugucuguccauagugauuaauguagcaccgccuaauug
cggugcuuu
>NC_007712
cauucauaagccggaacgaaaagaacuuaggaagauaaugauuucaaagcgcuuuuggcgagugaugugauguuguguuugcagauugggccgccau
ugcggacccccuuuuucuuaaauccguuuuggauauauuuuucugucuguccauagagaaaaguagcaccgcgccugcggugcuuu
>NC_008800
acuucccaagccggaacgaaaaguagguuuaguauccagguacugaauugcuuuuggcuuugugguugugauguuguguuugcaaauggucuggug
ugccagacauaguagcuaagcuacuguuuuuuuucacuuccuguacauuuacccugucuguccauagugauuuuaugcagcaccgcuauuuucgcgg
ugcuuu
>NC_003143
acuuccccuagccggaacgaaaaguagguuugguaucccagguacugaaaugcuuuuggcuuugugguggugauguuguguuugcaaauggucugg
uaugccagacauaguagcuaagcuacugcuuuuuuucacuuccuguacauuuacccugucuguccauagugauuuuaugcagcaccguuauuuuggc
ggugcuuu
>NC_006155
acuuccccuagccggaacgaaaaguagguuugguaucccagguacugaaaugcuuuuggcuuugugguggugauguuguguuugcaaauggucugg
uaugccagacauaguagcuaagcuacugcuuuuuuucacuuccuguacauuuacccugucuguccauagugauuuuaugcagcaccguuauuuuggc
ggugcuuu

GlmZ
>NC_000913
guagaugcucauuccaucucuuauguucgccuuagugccucauaaacuccggaaugacgcagagccguuuacggugcuuaucguccacugacagaug
ucgcuuaugccucaucagacaccauggacacaacguugagugaagcacccacuuguugucauacagaccuguuuuaacgccugcuccguaauaagagc
aggcguuuuuuu
>NC_016810
guagaugcucauuccaucucuuauguucgccuucuggccucauaaacucaggaaugaugcagagccguuuacggugcuuaucguccacugacagaug
ucgcuuacgccucaucaaacacccuggacacaacguugagugaagcacccuuuauguugucauacagaccuguuuugacgccugccccuuaaccggg
caggcguuuuuuu
>NC_009792
guagaugcucauuccaucucuuauguucgccuuagugccucauaaacucaggaaugaugcagagccguuuacggugcuuaucguccacagacagaug
ucgcuuacgccucaucaaacaccauggacauaacguugagugaagcacccauuuauguugucauacagaccuguuuugacgccugcucauuugagca
ggcguuuuuuu
>NC_013716
guagaugcucauuccaucucuuaugcucgccauagugccucauaaacucaggaaugaugcagagccguuuacggugcuuaucguccacugacagaug
ucgcuucggccucaucagacaccauggacauaacguugagugaagcacccauuuauguugucauacagaccuguuuugacgccugccccuuaccggg
caggcguuuuuuu
>NC_009778
guagaugcucauuccacuuauuauguuugcuccggcuucauaaacccaggaaugacgcagagccguuuacggugcuuaucguccacagacagauguc
gcuucggccucaucaaacaccauggacauaacguugagugaagcaccacauuuguugucaaacagaccuguuuugauaccugccuucgggcagguau
uuuuuu
>NC_009436

guagaugcucauuccaccu cuuau guuc gccuua gu gccuc auaa acuc a ggaau gac gc a ga gc cauuua c ggu gcuuau c gucc acc gaca gau g
ucgcuucggccucau caa aca ccau ggaca caa c guu ga gu gaa gca ccaa aauau guu gu c gaau a gaucu guuu aac gccu gcuuuuu a gca ggc g
uuuuuuu
>NC_011740
guagaugcucauuccaucucuu au guuc gc cuua gu gc cucau aaa cuca ggaau gac gca ga gcc guuu ac ggu gcuu auc guc cacu gaca gau g
ucgcuuaugccucaucaa aca ccau ggacau aac guu ga gu gaa gcac cca auu guu guca aac a gac cu guuuuaac gccu g cucc gu aauaa ga gc
aggcguuuuuuu
>NC_009648
guagaugcucauuccaucucuu au guuc gc cuuc gu gc cucau aaa cucc ggaau gau gca ga gcc guucuu ac ggu gcuu auc guc cacu gaca ga
ugucgcauuuaugccuc auca aac acc au ggac auuac guu ga gu gaa gcac cca auuu guu guca aac a gac cu guuuuaac gccu gcccu gauuu
cagcgcaggc guuuuuuu
>NC_012917
guagaugcucauuccaccuu cuau gcuu gcuuc ggcuu cauaauccu gggaau gau gca ga gcca auuu ga ggu gc cuauc gu cca acuc ca gau ga
agaugcaaaucuccau c gggcuuuc c ggac auaa c guu ga gu ga ggcac cauucu guu gu cu gaca gac cu gauauuuu caa c gcuua cc guuuau c
gguaagcguuuuuuu
>NC_005126
guagaugcucauccu auuucuuau gauu gcau au gcuuc auaa acc ca gggau gau gca ga gcc gauuauc ggu gccuauu gu ccau gu cac a gau g
aguaagauauaacccu cuuuucau c gcc gacc ggacuc aac guu ga gu ga ggc acuau cc gucu gu a gac cu gauu guauuu gu aca ccuau cuuua
uuuuuaagguaggu guuuuuuu
>NC_010554
guagaugcucauuccacuc cuuau gaca gc auau ga cuucau aaa ccua ggaau gau gca ga gcc gauu auc ggu gc cuac gu cca c guuauc gau g
aaacaccauucaucaccaa ca ggac gaa ac guu ga gu ga ggca ccau cc gucu gu a gac cu gauu guau auuuuau gc accu gu auuuuuuaau ac g
ggu guuuuuuu
>NC_003197
guagaugcucauuccaucucuu au guuc gc cuucu ggc cucau aaa cuca ggaau gau gca ga gcc guuua c ggu gcuuau c gucc acu gaca gau g
ucgcuuacgccucauca aac acc cu ggaca aca c guu ga gu gaa gc acc ccuuuau guu gu caua ca gac cu guuuu ga c gccu gccc cuuaa cc ggg
caggcguuuuuuu
>NC_009832
guagaugcucauuccaccu cuuau guuu gc cuua ggcuu caua aac ccu ggga au gac gca ga gcc gauuu aa ggu gc cuauu gc cca cca gaac ga
ugucagcguugcuu gca gcc gca gacau cac acuc c gggcau aac guu ga gu ga ggc acc gccc cu guu guccu a gac cu gauu gcuuuuuu auaca
cuugccaccgcggcaa gu guuuuuuu
>NC_007613
guagaugcucauuccaucucuu au guuc gc cuua gu gc cucau aaa cucc ggaau gac gca ga gcc guuu ac ggu gcuu auc guc cacu gaca gau g
ucgcuuaugccucauca gacac cau ggacac aac guu ga gu gaa gcac cca cuu guu gucau aca gac cu guuuu aac gccu gcu cc gua auaa ga gc
aggcguuuuuuu
>NC_007606
guagaugcucauuccaucuc cuau guuc gccuua gu gccuc auaa acuc c ggaau gac gca ga gcc guuuac ggu gcuuauc gu cca cu gaca gau g
ucgcuuaugccucauca gacac cau ggacac aac guu ga gu gaa gcac cca cuu guu gucau aca gac cu guuuu aac gccu gcu cc gua auaa ga gc
aggcguuuuuua
>NC_004337
guagaugcucauuccaucucuu au guuc gc cuua gu gc cucau aaa cucc ggaau gac gca ga gcc guuu ac ggu gcuu auc guc cacu gaca gau g
ucgcuuaugccucauca gacac cau ggacac aac guu ga gu gaa gcac cca cuu guu gucau aca gac cu guuuu aac gccu gcu cc gua auaa ga gc
aggcguuuuuuu
>NC_007384
guagaugcucauuccaucucuu au guuc gc cuua gu gc cucau aaa cucc ggaau gac gca ga gcc guuu ac ggu gcuu auc guc cacu gaca gau g
ucgcuuaugccucauca gacac cau ggacac aac guu ga gu gaa gcac cca cu guu gucau aca gac cu guuuu aac gccu gcu cc gua auaa ga gca
ggcguuuuuuu
>NC_008800
guagaugcucaucc cacuu auuau gau a gccu gguuuuu aua gc caauuu a gcuuc auaa aac ca gggau gac gc a ga gc c gauuuua gggu gccua
uuguccauguaaacgau guu gaau aucuuc auca caua cc gggc auaa c guu ga gu ga ggc acc gac auu guu gucu gua gac cu gaaa auuuca ga
cgcuugcccuuauc ggcaa gc guuuuuuu
>NC_003143
guagaugcucaucc cacuu auuau gaca gcuu ggc cauu aa ggcu aauuua gcuucau aaa ccc a gggau gac gca ga gcc gauuuua ga gu gccua
uuguccauguaaccgau guu ga gua au gucau cac auac c gggcau aac guu ga gu ga ggc acu gau auu guu gucuau c gac cu gaa aauuuua ga
cacuugcccuuuucggcaa gu guuuuuuu
>NC_006155
guagaugcucaucc cacuu auuau gaca gcuu ggc cauu aa ggcu aauuua gcuucau aaa ccc a gggau gac gca ga gcc gauuuua ga gu gccua
uuguccauguaaccgau guu ga gua au gucau cac auac c gggcau aac guu ga gu ga ggc acu gau auu guu gucuau c gac cu gaa aauuuua ga
cacuugcccuuuucggcaa gu guuuuuuu

MicA
>NC_000913
gaaagacgc gc auuu guuauc aucau cccu gaauuc a ga gau gaaauuuu ggcca cuca c ga gu ggc cuuuu
>NC_016810
gaaagacgc gc auuu guuauc aucau cccu guuuu ca gc gau gaa auuuu ggcc acuc c gu ga gu ggc cuuuu
>NC_009792

gaaagacgc gc auuu guuauc aucau cccu guuuu ca ga gau gaa auuuu ggcc acuc ac ga gu ggcc uuuu
>NC_013716
gaaagacgc gc auuu guuauc aucau cccu gauuuca ga gau ga aaauuuu ggc cacu ccc ga gu ggcc uuuu
>NC_009778
gaaagacgc gc auuu guuauc aucau cacu ga guuca ga gau ga caauuu ggcca ca gc gau gu ggcc uuuu
>NC_009436
gaaagacgc gc auuu guuauc aucau cccu gauuuca ga gau guuauuuu ggcca ca gc gau gu ggcc auuu
>NC_011740
gaaagacgc gc auuu guuauc aucau cccu gaauuc a ga gau gaaauuuu ggcca cuca c ga gu ggcc uuuu
>NC_009648
gaaagacgc gc auuuauuau caucaucau cccu gaauc a ga gau gaaa guuu ggcca ca gu gau gu ggcc uuuu
>NC_012917
gaaagacgc gc auuuauuau caucaucc cuauua ga gau guuuauuu ggcca ca guuucu gu ggcc uuuu
>NC_003197
gaaagacgc gc auuu guuauc aucau cccu guuuu ca gc gau gaa auuuu ggcc acuc c gu ga gu ggc cuuuu
>NC_009832
gaaagacgc gc auuu guuauc aucau cccu gu cauc a ga gau gcauauuu ggcca cauu gau gu ggcc uuuu
>NC_007613
gaaagacgc gc auuu guuauc aucau cccu gaauuc a ga gau gaaauuuu ggcca cuca c ga gu ggc cuuuu
>NC_007606
gaaagacgc gc auuu guuauc aucau cccu gaauuc a ga gau gaaauuuu ggcca cuca c ga gu ggc cuuuu
>NC_004337
gaaagacgc gc auuu guuauc aucau cccu gaauuc a ga gau gaaauuuu ggcca cucc c ga gu ggc cuuuu
>NC_007384
gaaagacgc gc auuu guuauc aucau cccu gaauuc a ga gau gaaauuuu ggcca cuca c ga gu ggc cuuuu
>NC_007712
gaaagaugc gcauuu guu aucau cauc ccu guua aca ggaau guua auuua gc cac a guuucu gu ggc cuuuu
>NC_008800
gaaagacgc gc auuu guuauc aucau cccu guu auca ga gau guua auuu ggcc aca gcaau gu ggc cuuuu
>NC_003143
gaaagacgc gc auuu guuauc aucau ccccuu cuauua ga gau guua auuu ggcc aca gu gau gu ggc cuuuu
>NC_006155
gaaagacgc gc auuu guuauc aucau ccccuu cuauua ga gau guua auuu ggcc aca gu gau gu ggc cuuuu

MicC
>NC_003197
guuauaugccuuuauuguca cauauu cauuuu guc gcu gggcc auu gc guua accuuu gcuuucc a gc guau aaauu gaca a gccc gaac ggau gu
ucgggcuuuuuuu
>NC_016810
guuauaugccuuuauuguca cauauu cauuuu guc gcu gggcc auu gc guua accuuu gcuuucc a gc guau aaauu gaca a gccc gaac ggau gu
ucgggcuuuuuuu
>NC_000913
guuauaugccuuuauuguca ca gauuuu auuuucu guu gggc cauu gc auu gcc acu gauuuu cca acau auaa aaa gaca a gccc gaaca guc gu c
cgggcuuuuuuu
>NC_009792
guuauaugccuuuauuguca cauuuu gcuuuuuu c guu gggcc auu gc gau aa gua cu gaucuu cca gc gaau gaauu gaca a gccc gaac caa ggu
ucgggcuuuuuuu
>NC_013716
guuauaugccuuuauuguca au guuu gcuuuuu guu ggc ccauu gc gaa gc guacu gauuu gcc aac aauc auaau gaca a gccc gaac gaau guu c
gggcuuuuuuu
>NC_009648
guuauaugccuuuauuguca u gcca auaauuu auu guu gcc gu cucauu cu gc gga au gau guu guuuau c gguaa aac gaca a gccc gaac guu g
uguucgggcuuuuuuu
>NC_011740
guuauaugccuuucuuguca cuauu gcuuuuu au gcu gggcu acc gca gauuu acuc auuuac ca gc aauau aauc gaca a gccc gaa caa au gucc
gggcuuuuuuu
>NC_007613
guuauaugccuuuauuguca ca gauuuu auuuucu guu gggc cauu gc auu gcc acu gauuuu cca acau auaa aaa gaca a gccc gaaca guc gu c
cgggcuuuuuuu
>NC_007606
guuauaugccuuuauuguca ca gauuuu auuuucu guu gggc cauu gc auu gcua cu gauuuuc caa cauau aaaa aa gaca a gccc gaaca guc guc
cgggcuuuuuuu
>NC_004337
guuauaugccuuuauuguca ca gauuuu auuuucu guu gggc cauu gc auu gcc acu gauuuu cca acau auaa aaa gaca a gccc gaaca guc gu c
cgggcuuuuuuu
>NC_007384

guuauaugccuuuauugucacagauuuuauuuucuguugggccauugcauugccacugauuuuccaacauauaaaaagacaagcccgaacagucguc
cgggcuuuuuuu

MicF
>NC_000913
gcuaucaucauuaacuuuauuuauuaccgucauucauuucugaaugucuguuuacccuauuucaaccggaugccucgcauucgguuuuuuuuu
>NC_009792
gcuaucaucauuaacuuuauuuauuaccgucauucauuucugaaugucuguuuacccuauuucaaccggaugccucgcauucgguuuuuuuuu
>NC_013716
gcuaucaucauuaacuuuauuuauuaccgucauucauuucugaaugucuguuuacccuauuucagccgaacguuuucacguucgguuuuuuuuu
>NC_009778
gcuaucaucauuaacuuuauuuauuaccgucauucagaucugaauguucguuuuacccuauuaccgccggaugcucgcauccggcauuuuuu
>NC_009436
gcuaucaucauuaacuuuauuuauuaccgucauucauuuuugaaugucuguuuaucccuaauugagccgagugcaaugcauucgguuuuuuuu
>NC_011740
gcuaucaucauuaacuuuauuuauuaccgucauucauauucugaaugucuguuuacccuauucaaccggaugcgaagcauccgguuuuuuuu
>NC_009648
gcuaucaucauuaacuuuauuuauuaccgucauucaguucugaaugucuguuuacccuauuucaaccggaugccucgcauccgguuuuuuuu
>NC_003197
gcuaucaucauuaacuuuauuuauuaccgucauucacuucugaaugucuguuuacccuauuucaaccggaugcuucgcauucgguuuuuuuu
>NC_016810
gcuaucaucauuaacuuuauuuauuaccgucauucacuucugaaugucuguuuacccuauuucaaccggaugcuucgcauucgguuuuuuuu
>NC_007613
gcuaucaucauuaacuuuauuuauuaccgucauucauuucugaaugucuguuuacccuauuucaaccggaugccucgcauucgguuuuuuuua
>NC_007606
gcuaucaccauuaacuuuauuuauuaccgucauucauuucugaaugucuguuuacccuacuucaaccggaugccucgcauccgguuuuuuuu
>NC_004337
gcuaucaucauuaacuuuauuuauuaccgucauucauuucugaaugucuguuuacccuauuucaaccggaugccucgcauucgguuuuuuua
>NC_007384
gcuaucaucauuaacuuuauuuauuaccgucauucauuucugaaugucuguuuacccuauuucaaccggaugccucgcauccgguuuuuuuu

OmrA
>NC_000913
cccagagguauugauuggugagauuauucgguacgcucuucguacccugucucuugcaccaaccugcgcggaugcgcagguuuuuuuu
>NC_011740
cccagagguauugauuggugagauuauucgguacgcgcuucguacccugucucuugcaccaaccugcgcggaugcgcagguuuuuuuu
>NC_003197
cccagagguauugauuggugagauuauucgguacgcucuucguaccugucucuugcaccaaccugcgcggaugcgcagguuuuuuuu
>NC_016810
cccagagguauugauuggugagauuauucgguacgcucuucguaccugucucuugcaccaaccugcgcggaugcgcagguuuuuuuu
>NC_007613
cccagagguauugauuggugagauuauucgguacgcucuucguacccugucucuugcaccaaccugcgcggaugcgcagguuuuuuuu
>NC_004337
cccagagguauugauuggugagauuauucgguacgcucuucguacccugucucuugcaccaaccugcgcggaugcgcagguuuuuuuu

OmrB
>NC_000913
cccagagguauugauaggugaagucaacuucgggguugagcacaugaauuacaccagccugcgcagaugcgcagguuuuuuuu
>NC_009792
cccagagguauugauaggugaaaucagcuuucgggguugaucacaagaauuacaccaaccugcgcaucaugugcagguuuuuuuu
>NC_013716
cccagagguauugauaggugacgucaacuuucgggguugaacacacgaauuacaccaaccugcgcagaugcgcagguuuuuuuu
>NC_009778
cccagagguauugauaggugaagucagcgacuuagcugaucacaacacuuacaccaaccugcgcggaugcgcagguuuuuuuu
>NC_009436
cccagagguauugauaggugaaaucagcuccgguugauuaacacgauuugcaccaaccugcguccauacgcagguuuuuuuu
>NC_011740
cccagagguauugauaggugguucaacuauaaaguugaccacuugaauuacaccaaccugcgcagaugcgcagguuuuuuuu
>NC_009648
cccagagguauugauaggugagucaacgucacguugaccacuuuacuuacaccagccugcgcagaugcgcagguuuuuuuu
>NC_012917
cccagagguauugauuggugauauaucgaugugcucuguacauugaaaccauuugauuacaccaaccuacgcggaugcguagguuuuuuuu
>NC_003197
cccagagguauugauaggugggaaucaacgucauuguugaucacacgaauuacaccaaccugcguagagaugcgcagguuuuuuuu

>NC_016810
cccagagguauu gau a ggu ggaau caa c gucauu guu gaucacac gaauua cac caa ccu gc gu a ga gau gc gca gguuuuuuuu
>NC_009832
cccagagguauu gauu ggu gaaucu ca gc aacuu guu guu gaac ccau auuaauuu gcac caa ccua c gca gau gc gua gguuuuuuuu
>NC_007613
cccagagguauu gauu ggu ga gauuauu c gguac gcucuuc gu acc cu gucucuu gcac caa ccu gc gc ggau gc gca gguuuuuuuu
>NC_007606
cccagagguauu gau a ggu gaa gu caa cuuc ggguu ga gcac au gaauu aca cca gccu gc gca gau gc gca gguuuuuuuu
>NC_004337
cccagagguauu gau a ggu gaa gu caa cuuc gu guu ga gcac au gaauu aca cca gccu gc gca gau gc gca gguuuuuuuu
>NC_007384
cccagagguauu gau a ggu gaa gu caa cuuc ggguu ga gcac au gaauu aca cca gccu gc gca gau gc gca gguuuuuuuu
>NC_008800
cccagagguauuaauu ggu ga gua auca acau ac gcu gu gu guu aaa gcca guuuuuu auuu gca cc ga ccua c gca gau gc gua gguuuuuuuu
>NC_003143
cccagagguauua auu ggu gaau aauc aac auuc gcu gu gu auca aa gau c guuuuuuauuu gc acc gaccu ac gc a gau gc gu a gguuuuuuuu
>NC_006155
cccagagguauua auu ggu gaau aauc aac auuc gcu gu gu auca aa gau c guuuuuuauuu gc acc gaccu ac gc a gau gc gu a gguuuuuuuu

OxyS
>NC_000913
gaaacgga gc ggc accu cuuuuaa cccuu gaa guc acu gc cc guuuc ga ga guuucu caa cuc gaauaa cuaaa gcc aac gu gaacuuuu gc ggaucu ccaggauccgc
>NC_009792
gaaacgga gc ggua ccucuuu aac ccuu gaa guca cc gc cc guuc aaa ga guuuuucuc aacu c gaaauaa cuaaa gcc aac gu gaacuuuu gc ggac ccuaugguccgc
>NC_013716
gaaacgga gc ggua c guuuaa cccuu gaa gc cac c gca c guuca ga ga guuucu cuca acc c gaau aacu aaa gcca ac gu ga acuuuu gc gga cc cc gu ggu ccgc
>NC_009778
gc gc gga gaa c auca cuccuu acc cuca cu ga gu gau aac cc gc aca ca ga gu cucucu guu a gcc gu auaa cuaa a gcc aac gu gaacu auuua gc ggacaguaa gucc gc
>NC_009436
uagacgaggc ggca c guauuu ga cccuu gac guc ccc gcc ga gu ca gac ga guuuau cccu aacu c gaa caa cuaa a gcc aac gu gaacuuuu gc gga ccccguggu ccgc
>NC_011740
gaaacgga gc ggc accu cuuuuaa cccuu gaa guc acu gc cc guuuc ga ga guuucu caa cuc gaauaa cuaaa gcc aac gu gaacuuuu gc ggaucu ccaggauccgc
>NC_009648
aauacgcccauaaa gac ggu cuac cu gu gaa aauc acu ga ccc gu cac acu guuucu cuac cc gaa caa cuaa a gcc aac gu gaacuuuu gc gga ccu ugcguccgc
>NC_003197
agaacgga gc gguuucu c guuuaa cccuu gaa ga cac c gcc c guuca ga ggguaucu cuc gaa ccc gaa auaa cuaa a gcc aac gu gaacuuuu gc gg accucuggu ccgc
>NC_016810
agaacgga gc gguuucu c guuuaa cccuu gaa ga cac c gcc c guuca ga ggguaucu cuc gaa ccc gaa auaa cuaa a gcc aac gu gaacuuuu gc gg accucuggu ccgc
>NC_007613
gaaacgga gc ggc accu cuuuuaa cccuu gaa guc acu gc cc guuuc ga ga guuucu caa cuc gaauaa cuaaa gcc aac gu gaacuuuu gc ggaucu ccaggauccgc
>NC_007606
gaaacgga gc ggc accu cuuuuaa cccuu gaa guc acu gc cc guuuc ga ga guuucu caa cuc gaauaa cuaaa gcc aac gu gaacuuuu gc ggaucu ccaggauccgc
>NC_004337
gaaacgga gc ggc accu cuuuuaa cccuu gaa guc acu gc cc guuuc ga ga guuucu caa cuc gaauaa cuaaa gcc aac gu gaacuuuu gc ggaucu ccaggauccgc
>NC_007384
gaaacgga gc ggc accu cuuuuaa cccuu gaa guc acu gc cc guuuc ga ga guuucu caa cuc gaauaa cuaaa gcc aac gu gaacuuuu gc ggaucu ccaggauccgc

RprA
>NC_000913
acgguuauaaaaucaac auauu gauuu auaa gcau gga aauc cccu ga gu gaa aca ac ga auu gcu gu gu gua gu cuuu gcc cauc ucc ac gau ggg cuuuuuuu
>NC_009792

acgguuauaaaaucaacgccuugauuuauaaagcauggaaaucccugagugaaacaacgaauugcugugguaguuuugcccaucucccacgauggg
ccuuuuuu
>NC_013716
aacguuauaaaaucaacacauugauuuauaagcauggaaaucccugagugaaacaacgaauugcugugguagucuuugcccaucucccacgaugggc
uuuuuuu
>NC_009778
acgguuauaaaaucaacgccuugauuuauaaaacauggaaaucccugagugaaacaacgaauugcugugguagucuuugcccaucucccacgauggg
cuuuuuuu
>NC_009436
acgguuauaaaaucaacaccuugauuuauaagcauggaaaucccugagugaaacaacgaauugcugugguagucuuugcccaucucccacgauggg
cuuuuuuu
>NC_011740
acgguuauaaaaucaacguauugauuuauaagcauggaaaucccugagugaaacaacgaauugcugugguagucuuugcccaucucccacgauggg
cuuuuuuu
>NC_009648
acgguuauaaaaucaacaccuugauuuauaagcauggaaaucccugagugaaacaacgaauugcugugguagucuuugcccaucucccacgauggg
cuuuuuuu
>NC_012917
gcuaguauuaaccuacuguaugucaauggagguuuccuccucgucagaguuaguaacgaauugcugugguagucuuugcccaucccugagauggg
ccuuuuuu
>NC_005126
ugggcuaugaacgagacguuuauuucuccuuuugaugguucuacuuguuagaaacgccuauauugcugugguagucccuugccuaucacgcccaug
auaggcuuuuuuu
>NC_010554
augguuuugccuauuuuauuaauagauuaaaaggcgagaccaaaguauauaaaauauugcugugguagucuuugccugucagcccaugauaggcu
uuuuuu
>NC_003197
acgguuauaaaaucaacacauugauuuauaagcauggaaaucccugagugaaacaacgaauugcugugguagucuuugcccgucuccuacgauggg
cuuuuuuu
>NC_016810
acgguuauaaaaucaacacauugauuuauaagcauggaaaucccugagugaaacaacgaauugcugugguagucuuugcccgucuccuacgauggg
cuuuuuuu
>NC_009832
ccugauuuaaagcaagauuuccaaacaauguaaugccauacaugugaauugcugugguagucuuugcccgucuccuaugaugggcuuuuuuu
>NC_007606
acgguuauaaaaucaacauauugauuuauaagcauggaaaucccugagugaaacaacgaauugcugugguagucuuugcccaucucccacgauggg
cuuuuuuu
>NC_004337
acgguuauaaaaucaacauauugauuuauaagcauggaaaucccugagugaaacaacgaauugcugugguagucuuuuccaucucccacgauggg
cuuuuuuu
>NC_007384
acgguuauaaaaucaacauauugauuuauaagcauggaaaucccugagugaaacaacgaauugcugugguagucuuugcccaucucccacgauggg
cuuuuuuu
>NC_007712
augcguguuauuagcucuaugauuuaguaaugauuuuuaaugcuacccggcaugacagaauugcugugguagucuuugcccaucuuuaagauggg
uuuuuuu
>NC_008800
cccgggauauuuaaauaaauuaucucauaaugacuuauuuaaauaucuguaaucgaauauaaguauugcugugguagucuuugccugucaccuaag
acgggcuuuuuuu
>NC_003143
aggcgcuuaaaauaaccauuuaaugauagcuuauuuaaaacgucugugaucuaguacauguauugcugugguagucuuugccugucaccuaagacag
gcuuuuuuu
>NC_006155
aggcgcuuaaaauaaccauuuaaugauagcuuauuuaaaacgucugugaucuaguacauguauugcugugguagucuuugccugucaccuaagacag
gcuuuuuuu

RybB
>NC_000913
gccacugcuuuucuuugauguccccauuuuguggagcccaucaaccccgccauuucgguucaagguugaugggguuuuuu
>NC_009792
gccacugcuuuucuuugauguccccaauuuguggagcccaucaaccccgccauuuuggguucaagguugaugggguuuuuu
>NC_013716
gccacugcuuuucuuugauguccccauuuuguggagcccaucaaccccgccguuucgguucaagguugaugggguuuuuu
>NC_009778
gccacugcuuuucuuugauguccccauuuuguggagcccaucaaccccgccaucuuggguucaagguugaugggguuuuuu
>NC_009436

gccacugcuuuucuuugau gu ccc ccauuuu gu gga gcc cauc aac ccc gcc auuuu gguu caa gguu gau gggu uuuuu
>NC_011740
gccacugcuuuucuuugau gu ccc caauuu gu gga gcc cauc aac ccc gcc guuuu gguu caa gguu gau gggu uuuuu
>NC_009648
gccacugcuuuucuuugau gu ccc cauuuu gu gga gccc auca acc cc gc cacuu c gguu caa gguu gau gggu uuuuu
>NC_012917
gccacugcuuuucuuugauu ccc cuuauu ga gga gcc cauc guu ccc gccuuu c a gguu caa gaa c gau gggc guuuu
>NC_005126
gccacugcuuuucuuugau aucc ccau auu ga gga acc c gaua gu caau ccauuuu gguu c ga ga cuauc gggu uuuuu
>NC_010554
gccacugcuuuucuuugau gu ccc caa auauu ga gga gcc c gaua guuu caa ccc cuuuu ggu gaa auaa cu guc gggu uuuuu
>NC_003197
gccacugcuuuucuuugau gu ccc cauuuu gu gga gccc auca acc cc gc cauuuc gguu caa gguu ggu gggu uuuuu
>NC_016810
gccacugcuuuucuuugau gu ccc cauuuu gu gga gccc auca acc cc gc cauuuc gguu caa gguu ggu gggu uuuuu
>NC_009832
gccacugcuuuucuuugau gu ccc cauuuu gu gga gccc guu a guccc gccu auuua gguu caa ga cua gc gggu uuuuu
>NC_007606
gccacugcuuuucuuugau gu ccc cauuuu gu gga gccc auca acc cc gc cauuuc gguu caa gguu ggu gguuuuuuu
>NC_004337
gccacugcuuuucuuugau gu ccc cauuuu gu gga gccc auca acc cc gc cauuuc gguu caa gguu gau gggu uuuuu
>NC_007384
gccacugcuuuucuuugau gu ccc cauuuu gu gga gccc auca acc cc gc cauuuc gguu caa gguu gau gggu uuuuu
>NC_007712
gccacugcuuuucuuugau gu ccc cauuuu ga gga gcc caa gc gucc c gccu auua gguu caa gac gcuu gguuuuuuu
>NC_008800
gccacugcuuuucuuugau gu ccc cauauu ga gga gc cc gau a gucc c gccuu cuua gguu caa ga cua guc gggu uuuu
>NC_003143
gccacugcuuuucuuugau gu ccc cauauu ga gga gc cc ggua gu ccc gccuucuu a gguu caa gacu aac c gggu uuuu
>NC_006155
gccacugcuuuucuuugau gu ccc cauauu ga gga gc cc ggua gu ccc gccuucuu a gguu caa gacu aac c gggu uuuu

SgrS
>NC_000913
gaugaagcaa ggggu gc ccc au gc guc a guuuuauc a gca cuauuuua cc gc gaca gc gaa guu gu gcu gguu gc guu gguua a gc guc cca caa c gauuaaccaugcuu gaa gga cu gau gc a gu gggau ga cc gc aauucu gaaa guu gacuu gc cu gcau cau gu gu ga cu ga guauu ggu gu aaa auca cccgccagca gauuau accu gcu gguuuuuuuu
>NC_009792
gaugaagcaauau gga ga ggucuu c gau gc gu ca guuuuauuu gaa gua cuuuac c gc gaca gaa ggguu guccu gguu ggcuu gucc ga gc gc gc cgcagcgcuuaaa gau gcuu gaa ga acu gau gca gu gggaggu gacu gc cu gac cuc ggauuu cc gaauu gca gacau cau gu gu ga cu ga guauu g guguuuaucgccgc gcc a gca gauaaauu ccu gcu ggcuuuuuuc
>NC_013716
gaugaagcaca au gaa ggggu gauca au gc gc ca guucu aucu gaaauauuuu acc gc gaca gaaa gauu guc cu gguu ggcuu gc cu ga gc gc gcc gcagcgcuuaaaaaaua cu ggaa gaacu gau gca gu gggaggu gaaaac cu gacu cc ggau a ggca gacau cau gu gu ga cu ga guauu ggu gca ggc uauagccucgauucacc gc gcc a gca ggu auuaucu gcu ggcuuuuuuc
>NC_009778
gcuuaugaaaacgu c gu ca acau gca gguuuuau ca gc a guauuuc a gc gc aac gcaac gucuuuc c ggcuc au ggcu ggccc gc cu gaac gguaaa acgccgcaac gaau gcu ggau ga gaucau gca gu gggau gu gacauu ccc ggu gac guuuaa gc gc c gcua a gc ggacauc ac gu gua acu ga gu au uggu gcuuuucaccu gcc a gca gu gu gcuacu gcu ggcuuuuuuc
>NC_009436
gaugaagcaa ggga gu gaac gau gaa ga ggu ca acc gcac gu ca guuuuau ca gc a guacuuuuu a gc gacaaaa gga gu gu ccu ggcu ggccc gcc agugcgccga gca gc ggcu ga aaau gcu ggaa gaauu gau gca gu ggga gguu ac gc aaac gacuu cu gaac gcu ga cuu gccu c gauc au gu gu ga cugaguauuggu gcuu auca ccc gcca gcaa aau guuuuu gcu ggcuuuuuuc
>NC_011740
gaugaagcga gga ggu ggga u gcau ca guuuuau ca gc ggu auuuuac c gca acu ga aaa guu gu ccu gguuau auu gguuaa gc gccu ccc gg cguuuagcaaugcu gga a gaa cu gau gc a gu ggga guu ga gu cucc gu aacu gaauu gc c gac aucau gu gu gacu ga gu auu ggu guaa aauc a cccgccagca gaua auac cu gcu ggcuuuuuuu
>NC_012917
gguggcuuau ggc gauuu ccc guuu gu acc gguu cuauc aaa ccua cuuauc gaccu gu aa ggc aaa gu gguu gc gau ggau gu c gac cca aca gc g cguugcauuauugca gc a ggca ac gc a gu ggca ccu gaa c gaa au gucu gau gaa ga guac c gucauu ggcucua ggu ggu gu gaa ga gau guauuc ugaguaaugguu gau guuu cac ca gc ca gc ga guuuucuc gcu ggcuuuuuuu
>NC_003197
gaugaagcaa ga ggaa ga ggu cacu au gc gc ca guucu gguu ga gauauuuu gcc gc gac ggaaaa aac gu ccu ggcu ggcuu gccu ga gc gc acc g cagcgcuuaaaaau gcu c gc gga acu gau gca gu ggga ggc gac c gauu gaa gcc aauu gc a gac aucau gu gu gacu ga gu auu ggu gua ggc gau agccuaaaaucaccc gc ca gc a gaua auaucu gcu ggcuuuuuuu
>NC_016810

gaugaagcaaggaaggaggucacuaugcgccaguucugguugagauauuuugccgcgacggaaaaaacgucculggcuggcuugccugagcgcaccg
cagcgcuuaaaaaaugcucgcggaacugaugcagugggaggcgaccgauugaagccaaauugcagacaucaugugugacugaguauugguguaggcgau
agccuaaaaaucacccgccagcagauaauaucugcuggcuuuuuuuu
>NC_009832
ggcgacggaugaaggguuuucuugucaaaacaguucuaucagcgcuauuucagcgcggugcgucgccagcgggccgauugguggggcuggugccug
aacaggcgcggcuggagaugcuggcccaucugacccagugggacauccacuccaugacggauaagcaauaccgugagcaucugugaggcucggguag
aaaaaguuucugaguaaugguguuguccaccagccagugggaucauaccacugguuuuuuua
>NC_007613
gaugaagcaaggaggugcccaugcgucaguuuuaucagcacuauuuuaccgcgacagcgaaguugugcugguugcguugguuaagcgucccacaac
gauuaaccaugcuugaaggacugaugcagugggaugaccgcaauucugaaaguugacuugccugcaucaugugugacugaguauuggugguaaaauca
cccgccagcagauuauaccugcugguuuuuuuuu
>NC_007606
gaugaagcaaggaggugcccaugcgucaguuuuaucagcacuauuuuaccgcgacagcgaaguugugcugguugcguugguuaagcgucccacaac
gauuaaccaugcuugaagaacugaugcagugggaggggagucauucugacuacugauuugcggacaucaugugugacugaguauuggugguuaaucg
ccacgccagcagugauuaucugcuggcuuuuuuuc
>NC_004337
gaugaagcaaggaggugcccaugcgucaguuuuaucagcacuauuuuaccgcgacagcgaaguugugcugguugcguugguuaagcgucccacaac
gauuaaccaugcuugaaggacugaugcagugggaugaccgcaauucugaaaguugacuugccugcaucaugugugacugaguauuggugguaaaauca
cccgccagcagauuauaccugcugguuuuuuuuu
>NC_007384
gaugaagcaaggaggugcccaugcgucaguuuuaucagcacuauuuuaccgcgacagcgaaguugugcugguugcguugguuaagcgucccacaac
gauuaaccaugcuugaaggacugaugcagugggaugaccgcaauucugaaaguugacuugccugcaucaugugugacugaguauuggugguaaaauca
cccgccagcagauuauaccugcugguuuuuuuuu

Spot42
>NC_000913
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuuagccgcccagucaguaaugacug
gggcguuuuuua
>NC_016810
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuuagccgcccagucaguuuaugac
ugggcguuuuuua
>NC_003197
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuuagccgcccagucaguuuaugac
ugggcguuuuuua
>NC_009792
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuuagccgcccagucaguaaugacug
gggcguuuuuua
>NC_013716
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuuagccgcccagucaguuucugac
ugggcguuuuuua
>NC_009778
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuuagccgcccagucagaaaugacug
gggcguuuuuua
>NC_009436
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuuagccgcccagucaguaaugacug
gggcguuuuuua
>NC_011740
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuuagccgcccagucaguaaugacug
gggcguuuuuua
>NC_009648
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuuagccgcccggucauuuaugacc
ggggcguuuuuua
>NC_012917
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuagccgcccgcucaguuuugagcgg
ggcguuuuuua
>NC_005126
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuagcugaauauuagcugcccagucaauuuugacugg
ggcauuuuuuu
>NC_010554
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuagcugaauauuagcugcccagucgacuuuauuucga
cuggggcauuuuuuu
>NC_009832
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuagccgcccagucaguauugacug
gggcguuuuuua
>NC_007606

guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuuagccgcccagucaguaaugacug
gggcguuuuuua
>NC_004337
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuuagccgcccagucaguaaugacug
gggcguuuuuua
>NC_007384
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcugaauauuuuagccgcccagucaguaaugacug
gggcguuuuuua
>NC_007712
guaggguacacagagguaagauguucuaucuuucaggccuuuuacuucacguaaucggacuuggcuaaguauuagcugcccagucauuuaaugacu
ggggcguuuuuua
>NC_008800
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcuguauauuagccgcccagucauuuauugacug
gggcguuuuuua
>NC_003143
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcuguauauuagccgcccagucauuuauugacug
gggcguuuuuug
>NC_006155
guaggguacagagguaagauguucuaucuuucagaccuuuuacuucacguaaucggauuuggcuguauauuagccgcccagucauuuauugacug
gggcguuuuuua

Supplementary references

Corcoran CP, Podkaminski D, Papenfort K, Urban JH, Hinton JC, Vogel J. 2012. Superfolder GFP reporters validate diverse new mRNA targets of the classic porin regulator, MicF RNA. *Molecular microbiology* **84**: 428-445.

Holmqvist E, Reimegard J, Wagner EG. 2013. Massive functional mapping of a 5'-UTR by saturation mutagenesis, phenotypic sorting and deep sequencing. *Nucleic acids research* **41**: e122.

Papenfort K, Pfeiffer V, Mika F, Lucchini S, Hinton JC, Vogel J. 2006. SigmaE-dependent small RNAs of Salmonella respond to membrane stress by accelerating global omp mRNA decay. *Molecular microbiology* **62**: 1674-1688.

Papenfort K, Podkaminski D, Hinton JC, Vogel J. 2012. The ancestral SgrS RNA discriminates horizontally acquired Salmonella mRNAs through a single G-U wobble pair. *Proceedings of the National Academy of Sciences of the United States of America* **109**: E757-764.

Pfeiffer V, Sittka A, Tomer R, Tedin K, Brinkmann V, Vogel J. 2007. A small non-coding RNA of the invasion gene island (SPI-1) represses outer membrane protein synthesis from the Salmonella core genome. *Molecular microbiology* **66**: 1174-1191.

Sittka A, Pfeiffer V, Tedin K, Vogel J. 2007. The RNA chaperone Hfq is essential for the virulence of Salmonella typhimurium. *Molecular microbiology* **63**: 193-217.

Stocker BA, Hoiseth SK, Smith BP. 1983. Aromatic-dependent "Salmonella sp." as live vaccine in mice and calves. *Developments in biological standardization* **53**: 47-54.

Urban JH, Vogel J. 2007. Translational control and target recognition by Escherichia coli small RNAs in vivo. *Nucleic acids research* **35**: 1018-1037.

The EMBO Journal Hfq and CsrA CLIP *Erik Holmqvist et al*

Expanded View Figures

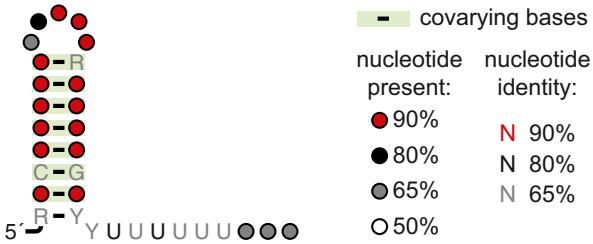
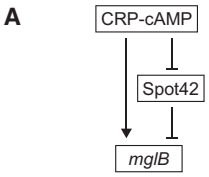


Figure EV1. RNA motif generated from Hfq 3′UTR peaks using the CMfinder algorithm.



B

```
ECO   gagcauuuaucucaagcacuacccugcauaa-gaaaaaccggagauac-----
CKO   gagcauuuuucucaaguacuacccugcauaa--uaaaaccggagauacc----
CRO   gggcauuucucugaagcacuacccugcauaa--uaaaaccggaguuacc----
STM   cgggcauuuuuuuacgcuauacccuacauaa--uaaaaccggagcuacc----
ENT   cguuauucacaagaagaacuacccugcauaaaaaaaaaaccggagaua------
SPR   uuggccuaacaccgagcaagacccuacauaa------accggagacacacaau
YEN   ugguguuaucacccuguacuacccuacauaa-----acccggagauaua-aau
            *        *    *****.*****       ******  *
```

Figure EV2. *mglB* mRNA is a putative target for the sRNA Spot42.

A Putative feed-forward loop between CRP-cAMP, Spot42, and *mglB*.

B Conservation of the predicted Spot42-binding site in *mglB* mRNA. Sequence alignment of RNA sequences upstream of the *mglB* start codon. Gray shading highlights the predicted Spot42-binding site. The alignment was made using MAFFT (Katoh *et al*, 2002). An asterisk indicates nucleotides that are identical in all sequences. ECO: *Escherichia coli* MG1655, CKO: *Citrobacter koseri*, CRO: *Citrobacter rodentium*, STM: *Salmonella* Typhimurium LT2, ENT: *Enterobacter* sp. 638, SPR: *Serratia proteamaculans*, YEN: *Yersinia enterocolitica* 8081.

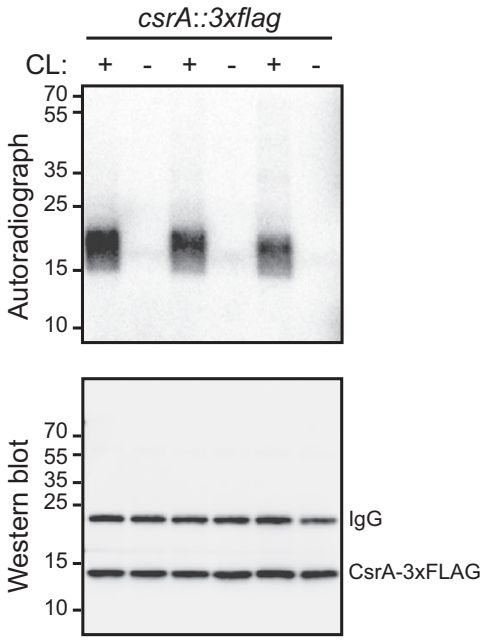


Figure EV3. UV-crosslinking selectively enriches CsrA–RNA complexes.
Autoradiograph showing radioactive RNA–protein complexes after separation on SDS–PAGE and transfer to a nitrocellulose membrane (top). The same samples were analyzed by Western blot using an anti-FLAG antibody to ensure that all samples contained the same amount of CsrA-3xFLAG protein (bottom).

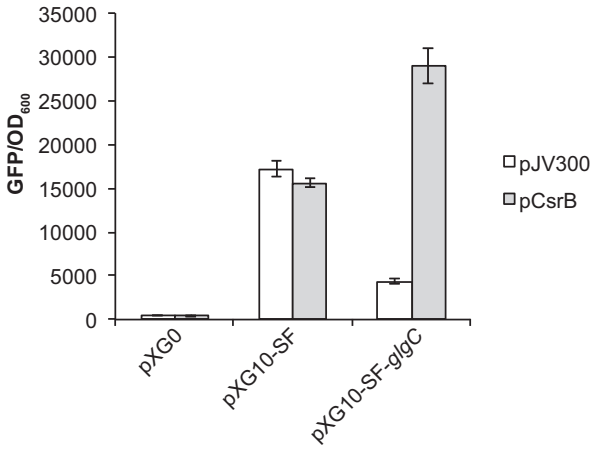


Figure EV4. Verification of CsrA-mediated regulation of a *glgC-gfp* translational fusion.
GFP fluorescence normalized to cell density from plasmids pXG0 (no GFP expression control), pXG10-SF (GFP-expressing *lacZ-gfp* control fusion), and pXG10-SF-*glgC* (GFP expression controlled by the *glgC* mRNA leader), in combination with a CsrB overexpression plasmid or the empty control plasmid pJV300. Means and standard deviations are based on three experiments.

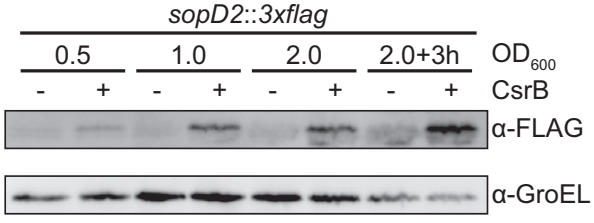


Figure EV5. Western blot analysis of endogenously expressed SopD2-3xFLAG protein in the presence or absence of CsrB overexpression.

4

MATERIAL AND METHODS

This chapter lists and indexes the materials and methods I personally applied for data analysis in the publications presented in section 3 of this thesis.

4.1 GLOBAL TRANSCRIPTIONAL START SITE MAPPING USING DIFFERENTIAL RNA SEQUENCING REVEALS NOVEL ANTISENSE RNAS IN ESCHERICHIA COLI

4.1.1 *Read mapping and coverage plot construction*

The description is located on pages 29 and 42.

4.1.2 *Normalization of expression graphs*

The description is located on page 42f.

4.1.3 *Correlation analysis*

The description is located on pages 29 and 43.

4.1.4 *Transcriptional start site (TSS) annotation*

The description is located on pages 29 and 43f.

4.1.5 *Comparison to Database of prOkaryotic OpeRons (DOOR)*

The description is located on pages 29 and 44.

4.1.6 *Comparison of pTSS and sTSS to RegulonDB promoters*

The description is located on page 29.

4.1.7 *Analysis of iTSS localization*

The description is located on page 44f.

4.1.8 *Expression analysis and binning*

The description is located on page 29.

4.1.9 *Comparison of expression under different growth conditions*

The description is located on page 45.

4.1.10 *Identification of overlapping 5′ UTRs*

The description is located on page 45.

4.1.11 *Comparison of asRNAs detected in our and previous studies*

The description is located on pages 29 and 45f.

4.1.12 *Comparison of asTSS to IP-dsRNAs*

The description is located on page 46.

4.2 DIFFERENTIAL RNA-SEQ (DRNA-SEQ) FOR ANNOTATION OF TRANSCRIPTIONAL START SITES AND SMALL RNAS IN HELICOBACTER PYLORI

4.2.1 *Read mapping and generation of coverage plots*

The description is located on page 80.

4.2.2 *Coverage plot normalization by TSSpredator*

The description is located on page 80.

4.2.3 *Automated TSS annotation using TSSpredator*

The description is located on page 80.

4.3 IDENTIFICATION OF THE RNA PYROPHOSPHOHYDROLASE RPPH OF HELICOBACTER PYLORI AND GLOBAL ANALYSIS OF ITS RNA TARGETS

4.3.1 *Data Processing and Availability*

The description is located on page 106f.

4.3.2 *Comparison between RppH and RNase J Targets*

The description is located on page 107.

4.4 THE CSRA-FLIW NETWORK CONTROLS POLAR LOCALIZATION OF THE DUAL-FUNCTION FLAGELLIN MRNA IN CAMPYLOBACTER JEJUNI

4.4.1 *Analysis of deep sequencing data*

The description is located on page 125.

4.4.2 *Enrichment analysis of CsrA targets*

The description is located on page 125f.

4.4.3 *Peak detection and CsrA-binding motif analyses*

The description is located on page 126.

The "sliding_window_peak_calling_script" I developed for identification of CsrA-binding sites based on RIP-seq data has been deposited at Zenodo (https://zenodo.org/record/49292) under DOI: 10.5281/zenodo.49292 (http://dx.doi.org/10.5281/zenodo.49292).

4.4.4 *Functional classes enrichment analysis*

The description is located on page 126.

4.4.5 *Sequence and structure conservation of the flaA 5'UTR*

The description is located on page 169f.

4.5 GLOBAL RNA RECOGNITION PATTERNS OF POST-TRANSCRIPTIONAL REGULATORS HFQ AND CSRA REVEALED BY UV CROSSLINKING IN VIVO

4.5.1 *Processing of sequence reads and mapping*

The description is located on page 188f.

4.5.2 *Analysis of structure motifs*

The description is located on page 190.

5

DISCUSSION

5.1 RNA-SEQ

In my thesis, I presented several publications that highlight how RNA-seq-based approaches can be used to answer different biological questions. The dRNA-seq method was applied for global annotation of TSS in two bacterial species, *E. coli* and *H. pylori*, as well as qualitative and quantitative analysis of associated transcripts including mRNAs, sRNAs and asRNAs (sections 3.2 and 3.3). In addition, we used a modified version of dRNA-seq to globally identify *in vivo* transcriptome targets of the RppH enzyme in *H. pylori* (section 3.4). Furthermore, we applied two related approaches to map transcriptome interaction sites of two bacterial RBPs. RIP-seq was used to identify targets of CsrA in *C. jejuni* (section 3.5) and CLIP-seq to precisely map binding sites of Hfq and CsrA in *Salmonella* (section 3.6).

5.1.1 *Sequencing*

A crucial step in experimental design for RNA-seq-based approaches is the selection of appropriate parameters for sequencing of cDNA libraries including required read numbers, read lengths, and if single- or paired-end sequencing should be applied.

Our dRNA-seq studies focused on the analysis of transcript 5'-ends and therefore did not require additional sequencing from the 3'-end. Furthermore, obtained read lengths were sufficient for unambiguous alignment of most reads except reads mapping to rRNAs, which exist in multiple genomic copies. For the CLIP-seq-based approach described in section 3.6, sequencing of shorter reads was recommended due to the overall short fragment sizes resulting from the experiment. In this case, we conducted paired-end sequencing to facilitate analysis of crosslink-mutations as identical mutations in both reads of a pair are unlikely to be the result of sequencing errors.

In our *H. pylori* and *C. jejuni* studies (sections 3.3 to 3.5) we sequenced between ~4.1 and 8.1 million reads, which should largely be sufficient due to their smaller genome sizes of ~1.6 megabases. In our *E. coli* study (section 3.2), read numbers for the HiSeq libraries were all above the 5 million threshold. Only the libraries sequenced on the Genome Analyzer IIx yielded numbers between ~1.8 and 3.6

million reads, which might explain part of the variation we observed in the data. In the CLIP-seq study conducted in *Salmonella* (section 3.6) we sequenced between ~17 and 33 million reads per library (data not shown) to sustain a sufficient amount of uniquely aligned reads (between ~340,000 and 850,000) for binding site detection after the extensive size filtering and collapsing steps.

5.1.2 *Data analysis*

As mentioned before, there is no optimal pipeline for all kinds of RNA-seq data, but in many cases an existing workflow can be applied to different experiments with only minor customization. For example, the READemption pipeline [56] has been applied in all studies presented in this thesis with slightly different parameter settings and in combination with different tools for preprocessing of the data. On the contrary, similar to sequencing technologies and experimental protocols, new software is being developed constantly and it is important to find a middle course between standardization and innovation.

For comparison, Rockhopper is a tool specifically designed for the analysis of bacterial RNA-seq data [120]. The workflow includes read alignment to a reference genome, transcript assembly and quantification, differential expression analysis, characterization of operon structures, and visualization in a genome browser. Rockhopper has the advantage that many important analysis steps are included in a single pipeline without requirement for external tools. However, this also makes it less flexible and complicates application of other tools for specific analyses. For example, transcript assembly is based solely on RNA-seq read coverage without the option to include additional information as for example provided by dRNA-seq.

5.1.3 *Reproducibility and sources of variation*

Reproducibility is a major concern for all published scientific results. Due to a variety of possible biases, RNA-seq-based experiments are particularly prone to variation resulting from technical rather than biological differences. In our dRNA-seq studies for mapping of TSS in *E. coli* and *H. pylori* (section 3.2 and 3.3) we identified library preparation, especially for different sequencing platforms, as a major source of variation. Similar findings have been described after comparison of RNA-seq experiments conducted in different laboratories based on human samples [167]. In our studies, biological replicates, for which library preparation and sequencing was conducted in parallel, showed much higher correlation than samples for which libraries were prepared on different days or even using different proto-

cols and sequencing platforms. Furthermore, sequencing of the same library on distinct sequencing runs yielded almost perfect correlation, suggesting very high reproducibility. Despite normalization to account for differences in sequencing coverage, large variation in read numbers between samples as observed between our Illumina and 454 replicates (section 3.3) can yield different results for quantitative but also qualitative analysis, such as TSS annotation. Further sources of variation include RNA isolation where unstable RNAs or fragments of specific sizes might not be captured, and rRNA depletion, which is prone to introduce coverage bias (see above) or might unintentionally remove non-rRNAs. During library preparation, different adjacent nucleotides can influence efficiency of adapter ligation, and RNA structure or modifications can lead to differences in reverse transcription. Furthermore, variations in G/C-content of transcripts impact PCR amplification efficacy [137].

Besides generation of RNA-seq data, variation can also be introduced during data analysis. In general, a consistent workflow is used for data processing and downstream analysis in a single RNA-seq experiment. A more complicated scenario represents the comparison of results from different experiments or between studies conducted in different laboratories. In the publication presented in section 3.2, we compared our asRNA candidates to annotations from other *E. coli* studies [35, 44, 122, 139, 149, 157] and observed large variation in numbers of reported asRNAs and only limited overlap among all studies. This variation could be caused by differences in the experimental setup but likely also by different analysis workflows including quality filtering, alignment, transcript annotation, and the approach used to classify transcripts as asRNAs.

There is no common standard for conducting an RNA-seq experiment, but it is important to consider sources of technical variation and try to minimize bias, especially within a single experiment. To achieve this, identical protocols for RNA extraction and library preparation should be used as much as possible with respect to the experimental setup. Based on our results, we recommend collecting biological samples for all replicates on the same day and prepare cDNA libraries in parallel. As there is no standard pipeline for RNA-seq data analysis, especially in bacteria, all steps in the analysis workflow should be considered carefully and executed via scripts that allow reproduction of the entire analysis. Furthermore, obtained results should be validated via independent experiments.

5.2 BACTERIAL TRANSCRIPTOME ANALYSIS

The dRNA-seq approach allows for TSS identification via comparison of a TEX-treated (+TEX) to an untreated library (−TEX) constructed from the same sample. The +TEX library mainly captures primary transcripts with a 5′-PPP while the −TEX library includes both primary transcripts and processed transcripts with a 5′-P. Currently, processed transcripts bearing a 5′-OH are ignored by the standard dRNA-seq approach. These transcripts are not degraded by TEX but are also excluded from both libraries since the 5′ RNA linker cannot be ligated to the 5′-OH group. However, minor adjustments to the protocol to include conversion of 5′-OH to 5′-P ends via polynucleotide kinase (PNK) and adenosine triphosphate (ATP) treatment (see [73] for a protocol) would enable sequencing of this class of transcripts.

A drawback of the dRNA-seq approach is its limitation to the annotation and quantification of RNA 5′-ends rather than providing insights on the extent of whole transcripts. For this it can be complemented with other RNA-seq-based methods. For example, an additional conventional RNA-seq library can be generated based on fragmentation of the −TEX sample to gain read coverage over complete transcripts. In the original dRNA-seq study, such data was used in combination with DOOR annotations [114] to elucidate operon structures [156]. Furthermore, a novel approach named term-seq for sequencing of exposed RNA 3′-ends was used to globally map transcript termini in *Bacillus subtilis*. Here, the majority of identified sites showed sequence and structural features of Rho-independent transcription terminators confirming their specificity [39]. In addition, Rho-independent terminators can also be predicted computationally [62]. A novel tool that integrates dRNA-seq and conventional RNA-seq data with computational terminator predictions to annotate all kinds of transcriptional features in bacterial and archaeal genomes is ANNOgesic [196].

Besides using fragmented conventional RNA-seq data based on short read sequencing to get whole-transcript coverage, another option is application of long read sequencing to sequence entire mono- and polycistronic transcripts. The PacBio IsoSeq™ protocol can directly sequence whole eukaryotic transcripts with a length of up to 10 kb [141]. Furthermore, it has been applied to insect mitochondrial transcriptome profiling [59] and an application to full-length sequencing of prokaryotic transcripts is under development [119].

Further recent methods for identification of primary and processed transcripts encompass tagRNA-seq [83] and Cappable-seq [50]. tagRNA-seq is based on labeling primary and processed transcripts with distinct sequence tags to allow differentiation between TSS and processed start sites (PSS), while Cappable-seq ap-

plies labeling of 5′-PPP ends with a biotin derivative to allow purification of primary transcripts via streptavidin beads. These methods claim to be superior to dRNA-seq in terms of specificity or the ability to annotate TSS based on a single sequencing library, but have so far been applied to a limited number of organisms while dRNA-seq has been successfully used on a multitude of bacterial and archaeal species [155].

In the studies presented in sections 3.2 and 3.3, we applied dRNA-seq in combination with automated TSS prediction via TSSpredator to generate global TSS maps for the bacterial model organims *E. coli* and *H. pylori*, respectively. For *E. coli*, we used several replicate dRNA-seq libraries generated from RNA samples harvested from bacteria growing under three different conditions to annotate >14,000 candidate TSS, while the *H. pylori* TSS map consist of >2,200 TSS based on four biological replicates from mid-log growth.

Although the number of annotated *E. coli* TSS seems quite high in comparison to previous publications, e.g. Kim et al. identified >3,700 TSS using a modified 5′ RACE approach [88], a recent study using Cappable-seq has reported the presence of >16,000 clustered TSS [50]. Calculating the overlap between Cappable-seq TSS and a composite dataset consisting of promoter annotations from RegulonDB, Kim TSS and our TSS detected in the M63 0.4 condition (16,855 TSS) yielded an overlap of 9,600 TSS [50]. Possibly, even more matching positions would have been identified when including our TSS detected under the LB 0.4 and LB 2.0 condition. The authors state additional TSS identified by their approach under similar growth conditions are to a certain extent the result of deeper sequencing. Together, these findings suggest that seemingly high numbers of identified TSS still do not represent the full complement of transcription activity in organims like *E. coli* and that deeper sequencing and analysis of additional biological conditions might further increase the amount of annotations. In addition, repeated identification of matching TSS positions results in an increased confidence in these sites.

Considering dRNA-seq data for more different biological conditions as conducted in previous studies using manual TSS annotation [92, 156] facilitates annotation of condition-specific TSS. Besides TSS prediction in a single organism, TSSpredator has also been applied to annotate TSS in four *C. jejuni* strains [46]. By mapping the genomes of the different strains to a common coordinate system, the so-called SuperGenome, this approach can give insights into strain- or species-specific differences in transcription or gene regulation caused, for example, by single-nucleotide polymorphisms (SNPs) in promoter regions or non-coding parts of a transcript. Such comparative analyses applied to multiple isolates of different bacterial species

might also help to identify specific or conserved sRNAs and give insights into the conservation of antisense transcription [46, 191].

For further discussion of numbers and classes of identified TSS and a comparison of the *H. pylori* TSS annotations to the original study [156] please refer to the respective publication in section 3.2 and 3.3.

Manual TSS annotation, as conducted in the initial *H. pylori* dRNA-seq study, is a laborious and time consuming process, which is especially impractical when data includes multiple strains or conditions with several replicates. Automated annotation approaches like TSSpredator [46] follow defined rules based on specific parameters and therefore avoid biases inherent to manual annotation. Furthermore, analysis can be easily repeated using different parameters or including additional data sets. However, choosing optimal parameters for a specific organism and data set can be difficult and manual inspection of a subset of predicted TSS in a genome browser is recommended. Another option is conducting parameter optimization based on an inital subset of manually annotated TSS. Such an approach for defining optimal TSSpredator parameters is implemented in the tool ANNOgesic [196].

Further TSS prediction tools which utilize dRNA-seq data include TSSAR [3] and TSSer [84]. TSSAR models read counts in transcriptionally active regions based on the Poisson distribution and applies the Skellam distibution to identify significantly enriched primary transcripts locally via a sliding window approach. TSSer calculates enrichment of putative TSS positions via a 'z-score' or, when replicates are available, using a Bayesian framework to quantify the probability that a genomic position is overrepresented across a number of TEX-treated samples. Furthermore, it requires a local enrichment of the putative TSS position compared to the neighboring genomic positions. In contrast to TSSpredator, which is based on a set of fixed cutoffs, both tools apply statistical models and require less parameters. However, only TSSer supports multiple replicates and neither of the two is able to integrate data from multiple strains or conditions. In addition, TSS prediction of both tools is influenced by selected size of locally analyzed regions.

In order to globally catalog TSS in an organism of interest, we suggest a strategy where dRNA-seq data is generated based on growth under different stress or growth conditions, possibly using multiple related strains, and TSS annotation is conducted via an automated tool like TSSpredator.

5.2.1 *Analysis of transcriptome features*

Global TSS maps facilitate identification and analysis of diverse transcriptome features, including promoter regions, 5'UTRs and leaderless mRNAs as well as *cis*- and *trans*-encoded sRNAs.

For *E. coli*, we identified a common housekeeping σ^{70} promoter motif [53] upstream of most TSS assigned to pTSS, iTSS and asTSS (section 3.2), suggesting that there is no preference of a specific transcript class for transcription via the σ^{70} holoenzyme. A similar motif has been described in another study based on a likewise high number of identified TSS [50].

Annotation of 5'UTRs of protein-coding genes based on pTSS and sTSS is the starting point for a number of downstream analyses. Their length can be correlated with translation rates. In addition, they can be searched for *cis*-regulatory elements like riboswitches and RNA thermometers or target sites of sRNAs [155] or RBPs like CsrA (see section 3.5). Furthermore, 5'UTRs of divergently transcribed genes that overlap with each other on opposite strands can be examined for their role in antisense-mediated regulation [153] or transcriptional interference [15, 63, 171].

In *E. coli* we identified 212 gene pairs with overlapping 5'UTRs based on pTSS and sTSS additionally classified as asTSS (section 3.2). The genes encoded by some of these antisense transcript pairs are annotated to have opposing functions and could serve as a starting point for deeper analysis to understand if and how the associated transcripts affect each other. The same goes for the 28 divergently transcribed gene pairs found in *H. pylori* (section 3.3).

Besides TSS that likely represent transcription starts of already annotated genes or operons, the TSS map also includes information regarding the presence of novel transcripts as *trans*-acting sRNAs or asRNAs. sRNAs are frequently found in intergenic regions but can also be derived from 3' regions of mRNAs, either by transcription from their own promoter or via processing of the parent transcript [26, 125]. The presence of an oTSS in an intergenic region could either indicate transcription of a so far unknown sRNA or represent a very long 5'UTR with a length above the applied detection threshold (see Fig. 4A in section 3.3 for an example). Likewise, a pTSS or sTSS associated with a certain gene could actually represent the TSS of an sRNA in close proximity to the respective gene. In case of a 3'UTR-derived sRNA, the transcription start could either be annotated as an oTSS or iTSS depending on its localization downstream of or at the 3'-end of a CDS, respectively. In any case, closer examination of oTSS but also distant pTSS or sTSS as well as iTSS at the 3'-end of coding regions is a good starting point for discovery of novel sRNAs. To examine if short transcripts in intergenic regions constitute untranslated sRNAs or might

represent novel small mRNAs with a short open reading frame (sORF), dRNA-seq can be combined with another RNA-seq-based approach termed ribosome profiling [81, 82, 131], which is able to define translated regions by sequencing mRNA bound by actively translating ribosomes.

Besides *trans*-acting sRNAs, asRNAs transcribed from the opposite strand of annotated coding regions represent an emerging class of transcripts with potential regulatory functions. In the study presented in section 3.2, we identified >5,400 asTSS that were not assigned to any other class. It is hard to estimate how many of these asRNA candidates have an actual function or just result from pervasive transcription. A recent study reported an exponential dependence of the number of asRNAs on genomic A/T content and that only asRNAs expressed over a certain threshold can function as regulators of their respective sense transcripts [109]. This supports the hypothesis that most asRNAs are the result of transcriptional noise from spurious promoters that arise more frequently in bacteria with higher A/T content. In accordance with this, we only found a limited number of our asRNA candidates to be present at high levels or differentially expressed among growth conditions. Furthermore, only a few of them were detected in multiple RNA-seq studies. While some candidates might only be expressed at functional levels under specific biological conditions, the low overlap between studies is likely caused by major differences in experimental and computational methods used to annotate asRNAs (see section 5.1). To gain additional confidence in our predictions we further evaluated 14 selected candidates by detection on Northern blots.

Overall, the rising number of RNA-seq-based studies and the development of new experimental approaches for transcript identification [50, 83] will result in growing numbers of reported putative asRNAs. To investigate how many of them function as specific antisense regulators, are involved in global processing of sense transcripts, or are just expressed as spurious transcripts will require further examination. Given the effort required for functional characterization and the elucidation of mechanisms of action a careful selection of appropriate transcripts is essential. Automated TSS annotation as conducted in our dRNA-seq studies (sections 3.2 and 3.3), together with assessment of expression levels, detection by multiple studies and independent experimental validation of single candidates can give important hints to identify the most promising candidates.

5.2.2 *RppH target identification*

Bacterial RNA degradation initiates either via direct internal cleavage of transcripts mediated by an RNase or a 5′-end-dependent mechanism where conversion of a terminal 5′-PPP to a 5′-P by RppH facilitates RNase processing [78].

We applied a modified version of the dRNA-seq approach [156] to globally identify transcript targets of RppH in *H. pylori* (section 3.4). For this, we complemented the two standard dRNA-seq libraries with a third library specific for transcripts with a 5′-P. We applied these libraries to analyze the influence of an *rppH* deletion on the 5′-phosphorylation state of trancripts and identified 53 mRNAs and 10 sRNAs whose degradation is potentially triggered by this enzyme. In addition, we further validated several of these transcripts by half-life measurements and PABLO analysis.

Since the analysis of RppH targets was conducted based on previous TSS annotations [156] it should be noted that TEX enrichment in the standard dRNA-seq approach might fail due to fast processing of primary transcripts via RppH. This could result in missing TSS annotations for transcripts that are heavily affected by RppH, which consequently were also not considered for RppH target indentification. However, since we found many putative target candidates, RppH activity in most cases is likely not strong enough to prevent TSS annotation. Additionally, we identified TSS for the majority of validated RppH targets in *E. coli* (see section 3.2) further supporting the assumption that this is not a general problem of the dRNA-seq approach.

For a more detailed discussion of RppH binding preferences, identification and validation of target candidates and the role of RppH in RNA degradation please refer to section 3.4.

In addition to the detection of RppH processing sites, RNA-seq-based methods can also be used to map RNase cleavage sites. For example, TIER-seq has been used to identify cleavage sites of the endoribonuclease RNase E in *Salmonella* [28]. As RNase E was shown to prefer substrates with a 5′-P [113], a combination of the two approaches could be used to analyze the interplay of RppH and RNase E during RNA processing.

Besides its biological role, RppH can also be used in practical applications. The TAP enzyme, which was essential for many experimental protocols as, for example, library preparation in the dRNA-seq approach, is not available on the market anymore. It was shown that *E. coli* RppH can be used as a replacement for TAP to generate monophosphorylated 5′-ends of RNA molecules [74, 128].

5.3 IDENTIFICATION OF RBP TARGETS

In two publications presented in this thesis, we applied RNA-seq-based approaches to globally identify targets of two distinct bacterial RBPs. RIP-seq was used to detect binding partners of CsrA in *C. jejuni* (section 3.5) and CLIP-seq was applied to precisely map binding sites of Hfq and CsrA in *Salmonella* (section 3.6). While the sequencing data from RIP-seq is based on whole RBP-bound transcripts or longer fragments thereof, CLIP-seq reads map more precisely to their respective binding sites as they result from only sequencing RNA regions that are protected from digestion through RBP binding.

Both RIP-seq and CLIP-seq have different strengths and weaknesses. In RIP-seq, RNA-protein interactions are not stabilized by crosslinking, which can result in loss of target RNA or inclusion of non-specifically bound RNAs if washing conditions are not carefully adjusted. Furthermore, low affinity targets are likely not recovered by the approach. In CLIP-seq, off-target effects are reduced via crosslinking and more stringent washes, and binding sites can be determined with a higher resolution due to RNase digestion of unprotected parts of bound RNA molecules. In addition, crosslink-induced mutations can be used to identify crosslinked nucleotides. However, crosslinking efficiency is rather low and prone to be biased by the presence of specific nucleotides and amino acid residues or RNA structures [188].

To address some of the shortcomings of each method, a recent digestion-optimized RIP-seq approach was developed to investigate protein-RNA interactions with binding site resolution. DO-RIP-seq [129] incorporates elements of both RIP-seq and CLIP-seq by conducting partial digestion of protein-bound RNA without covalent crosslinking. Using two distinct analysis workflows, it allows quantification of binding on whole transcript as well as binding site level.

We do not know to what extent our experiments might be affected by the issues that can arise for RIP-seq and CLIP-seq approaches. In the Hfq CLIP-seq experiment, we found our binding sites to preferentially map to U-rich sequences in Rho-independent terminators of sRNAs as well as mRNAs, while a recent Hfq crosslinking study in *E. coli* seemed to be biased towards detection of A-rich sequences [177]. This could be explained by a combination of a preference for crosslinking uridines [166] and differences in the applied protocols. In the study by Tree *et al.* [177], 3′ adapter ligation was performed with RNA in complex with Hfq, while in our study after RNA fragments were released from Hfq (see section 3.6). Hence, 3′-ends may not have been accessible for adapter ligation in the other study when still bound to the proximal side of Hfq that tends to target U-rich sequences. However, these differences are mainly due to the complexity introduced by Hfq, which binds RNA

on three distinct faces. Despite the underrepresentation of A-rich sequences, we annotated many known and potential Hfq binding sites in *Salmonella* mRNAs and sRNAs, identified Rho-independent terminators as a general binding motif, and found that in many cases Hfq binds 5′ to sRNA-binding sites in mRNA targets and 3′ to seed sequences in cognate sRNAs, which supports a model where Hfq facilitates duplex formation by bringing together the two RNAs (section 3.6).

Additionally, we found similar binding preferences in terms of sequence and structural motifs based on RIP-seq in *C. jejuni* (section 3.5) and CLIP-seq in *Salmonella* (section 3.6), which agree with binding sites of other CsrA homologs [45]. This further supports the validity of both approaches.

Please refer to sections 3.5 and 3.6 for a detailed discussion of targets and binding preferences of CsrA in *C. jejuni* and of Hfq and CsrA in *Salmonella*, respectively.

A widely used peak caller for RIP-seq and CLIP-seq data is Piranha [179]. This tool assumes that most regions with read coverage are noise and are therefore used to fit a background model. This causes problems especially when the RBP under study has many targets in a relatively small genome, which is the case for both Hfq and CsrA. In addition, Piranha does not support replicates.

Other tools like RIPseeker [106] or JAMM [80] have not been tested on our data but have technical limitations. Both tools are not able to conduct strand-specific peak calling and only JAMM has native support for several replicates while RIPseeker conducts subsequent merging of peaks called based on single replicates.

For RIP-seq-based CsrA target identification we applied Gfold [54] on the level of known genomic annotations and a self-developed peak calling method based on a sliding window approach for a more accurate detection of binding sites independent of annotation (section 3.5). In contrast, a different peak calling approach termed "block-based peak calling" [178] was used to identify binding sites of Hfq and CsrA in *Salmonella* (section 3.6). I implemented the block-based peak calling approach as well as a modified version of the sliding window approach with support for several replicates in the tool PEAKachu (`https://github.com/tbischler/PEAKachu`, Bischler and Wright *et al.*, manuscript in preparation). The sliding window approach is more suitable for RIP-seq data due to its flexibility in adjusting window and step size for detection of longer enriched regions, while the block-based peak calling method was found to yield better results for more narrow CLIP-seq peaks. Besides the manuscripts presented in this thesis, PEAKachu has also been applied to predict DHX9 binding sites in the human genome based on another UV-crosslinking method (FLASH: fast ligation of RNA after some sort of affinity purification for high-throughput sequencing) [2]. This shows that PEAKachu is a general purpose tool that can not only be used for peak calling based on standard

RIP-seq and CLIP-seq experiments, but also modified versions of these methods and potentially also other kinds of data with a similar nature.

6

CONCLUSION AND PERSPECTIVE

RNA-seq-based approaches can be applied to a multitude of biological questions. Conventional RNA-seq data can be used for gene expression analysis and annotation of transcriptional features. However, more advanced protocols are required for precise mapping of transcript boundaries, especially in complex bacterial transcriptomes. Annotation of transcriptional features is essential to elucidate the full complement of transcriptional regulation in an organism. The work presented in this thesis where the dRNA-seq approach was applied (sections 3.2 and 3.3) exemplifies how global annotation of transcript 5'-ends facilitates downstream analyses like prediction of promoter motifs and identification of 5'UTRs, which can subsequently be searched for *cis*-reglatory elements. Furthermore, it represents a good starting point to annotate and characterize a multitude of regulatory features including sRNAs, asRNAs, and specific antisense-mediated regulation via overlapping 5'UTRs.

Automated TSS prediction via TSSpredator as conducted for the *H. pylori* and *E. coli* dRNA-seq data greatly facilitates generation of a global TSS map. However, selecting optimal parameter values is a major challenge for each data set and using the same global cutoffs for the whole transcriptome might impede detection of lowly expressed transcripts or increase the amount of falsely annotated TSS in highly-transcribed regions. These issues could be addressed by the integration of a locally applied statistical approach similar to other tools like TSSAR [3] or TSSer [84] while keeping the advanced capabilities of TSSpredator to conduct comparative TSS prediction based on multiple strains or conditions.

There is still ongoing discussion on the extent of transcribed and functional asRNAs in bacteria. Using dRNA-seq, we identified a plethora of asRNA candidates in *E. coli* (section 3.2). Further investigation will be required to answer the question how many of these have regulatory functions or just result from pervasive transcription. However, computational methods including promoter analysis, classification according to expression, assessment of differential expression between conditions, comparison to existing data, and conservation analysis, if data from multiple strains are available, can give important hints to select candidates for further examination. We independently validated 14 candidates via Northern blot and highlighted the importance of including control deletion strains for this anal-

ysis. Overall, results from this study will help to identify appropriate candidates for future examination of phenotypes and regulatory mechanisms associated with asRNAs.

The global TSS maps for *E. coli* and *H. pylori* are both available in table format, which facilitates automated downstream analysis. In addition, we integrated TSS positions and coverage plots in easily accessible online genome browsers, which allow visual inspection of individual TSS positions in their genomic context. This enables researchers to identify candidate TSS and examine relative expression for their genes of interest.

Our findings from *in vitro* analysis of RppH activity and substrate specificity, as well as global identification of *in vivo* targets (section 3.4), suggest an important role for this enzyme in gene expression control of *H. pylori* and related organisms. Further investigation will be required to unveil the full complement of RppH-based regulation in combination with different RNases and its association with phenotypes under specific biological conditions.

Our global target identification strategy employing a modified dRNA-seq approach represents a useful tool, which can be applied to globally analyze RppH-based regulation in other bacterial species including *E. coli* and *B. subtilis*, where only a limited number of RppH targets has been identified [41, 77, 112, 142].

Bacteria express a multitude of regulatory RBPs for which global information on RNA binding partners or precise binding sites is either not available at all or only exists for specific bacterial species. In the last two publications presented in this thesis (sections 3.5 and 3.6), we applied RIP-seq and CLIP-seq to identify the interactomes of the two bacterial RBPs Hfq and CsrA. We developed specific analysis pipelines for both approaches with peak calling as the key step for binding site annotation. The integration of the two applied peak calling approaches in the tool PEAKachu will facilitate similar analysis in future studies and help to uncover the interactomes and binding preferences of other RBPs, such as ProQ [160] or cold-shock proteins like CspC and CspE [123].

Overall, the work presented in this thesis describes the application of different RNA-seq-based approaches and associated computational analysis methods to gain insights into transcription and post-transcriptional regulation in bacteria. Experimental approaches involved cDNA library preparation for sequencing on the Illumina platform, which was found to cause data variation when conducted at different times or for distinct sequencers. Recent developments in sequencing technologies, such as direct RNA sequencing using nanopore sequencers [60, 161], might reduce bias associated with library preparation and facilitate data quality and analysis.

BIBLIOGRAPHY

[1] Thomas Abeel, Thomas Van Parys, Yvan Saeys, James Galagan, and Yves Van de Peer. "GenomeView: a next-generation genome browser." In: *Nucleic Acids Research* 40.2 (Jan. 2012), e12–e12. ISSN: 0305-1048. DOI: 10.1093/nar/gkr995. URL: https://academic.oup.com/nar/article/40/2/e12/2409789 (visited on 01/05/2018).

[2] Tuğçe Aktaş, İbrahim Avşar Ilık, Daniel Maticzka, Vivek Bhardwaj, Cecilia Pessoa Rodrigues, Gerhard Mittler, Thomas Manke, Rolf Backofen, and Asifa Akhtar. "DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome." en. In: *Nature* 544.7648 (Apr. 2017), pp. 115–119. ISSN: 0028-0836. DOI: 10.1038/nature21715. URL: https://www.nature.com/nature/journal/v544/n7648/full/nature21715.html (visited on 10/31/2017).

[3] Fabian Amman, Michael T. Wolfinger, Ronny Lorenz, Ivo L. Hofacker, Peter F. Stadler, and Sven Findeiß. "TSSAR: TSS annotation regime for dRNA-seq data." In: *BMC Bioinformatics* 15 (Mar. 2014), p. 89. ISSN: 1471-2105. DOI: 10.1186/1471-2105-15-89. URL: https://doi.org/10.1186/1471-2105-15-89 (visited on 11/13/2017).

[4] Simon Anders and Wolfgang Huber. "Differential expression analysis for sequence count data." en. In: *Genome Biology* 11.10 (Oct. 2010), R106. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-10-r106. URL: http://genomebiology.com/2010/11/10/R106 (visited on 03/14/2012).

[5] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. "HTSeq—a Python framework to work with high-throughput sequencing data." In: *Bioinformatics* 31.2 (Jan. 2015), pp. 166–169. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu638. URL: https://academic.oup.com/bioinformatics/article/31/2/166/2366196/HTSeq-a-Python-%20framework-to-work-with-high (visited on 09/25/2017).

[6] Simon Anders, Alejandro Reyes, and Wolfgang Huber. "Detecting differential usage of exons from RNA-seq data." en. In: *Genome Research* 22.10 (Oct. 2012), pp. 2008–2017. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.133744.111. URL: http://genome.cshlp.org/content/22/10/2008 (visited on 09/28/2017).

[7] S Andrews. *FastQC A Quality Control tool for High Throughput Sequence Data*. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (visited on 09/22/2017).

[8] Charles Ansong, Hyunjin Yoon, Steffen Porwollik, Heather Mottaz-Brewer, Brianne O. Petritis, Navdeep Jaitly, Joshua N. Adkins, Michael McClelland, Fred Heffron, and Richard D. Smith. "Global Systems-Level Analysis of Hfq and SmpB Deletion Mutants in Salmonella: Implications for Virulence and Global Protein Translation." In: *PLOS ONE* 4.3 (Mar. 2009), e4809. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0004809. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004809 (visited on 11/12/2017).

[9] Liron Argaman, Ruth Hershberg, Jörg Vogel, Gill Bejerano, E. Gerhart H Wagner, Hanah Margalit, and Shoshy Altuvia. "Novel small RNA-encoding genes in the intergenic regions of Escherichia coli." In: *Current Biology* 11.12 (June 2001), pp. 941–950. ISSN: 0960-9822. DOI: 10.1016/S0960-9822(01)00270-6. URL: http://www.sciencedirect.com/science/article/pii/S0960982201002706 (visited on 10/05/2017).

[10] Manuel Ascano, Stefanie Gerstberger, and Thomas Tuschl. "Multi-disciplinary methods to define RNA–protein interactions and regulatory networks." In: *Current Opinion in Genetics & Development*. Cancer genomics 23.1 (Feb. 2013), pp. 20–28. ISSN: 0959-437X. DOI: 10.1016/j.gde.2013.01.003. URL: http://www.sciencedirect.com/science/article/pii/S0959437X13000117 (visited on 10/16/2017).

[11] Paul Babitzke and Tony Romeo. "CsrB sRNA family: sequestration of RNA-binding regulatory proteins." In: *Current Opinion in Microbiology*. Cell regulation (RNA special issue) 10.2 (Apr. 2007), pp. 156–163. ISSN: 1369-5274. DOI: 10.1016/j.mib.2007.03.007. URL: http://www.sciencedirect.com/science/article/pii/S1369527407000240 (visited on 10/16/2017).

[12] Timothy L. Bailey, Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. "MEME Suite: tools for motif discovery and searching." In: *Nucleic Acids Research* 37.suppl_2 (July 2009), W202–W208. ISSN: 0305-1048. DOI: 10.1093/nar/gkp335. URL: https://academic.oup.com/nar/article/37/suppl_2/W202/1135092/MEME-Suite-tools-fo%20r-motif-discovery-and-searching (visited on 10/13/2017).

[13] Alexander G. Baltz et al. "The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts." English. In: *Molecular Cell*

46.5 (June 2012), pp. 674–690. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2012.05.021. URL: http://www.cell.com/molecular-cell/abstract/S1097-2765(12)00437-6 (visited on 10/16/2017).

[14] Lars Barquist and Jörg Vogel. "Accelerating Discovery and Functional Analysis of Small RNAs with New Technologies." In: *Annual Review of Genetics* 49.1 (2015), pp. 367–394. DOI: 10.1146/annurev-genet-112414-054804. URL: https://doi.org/10.1146/annurev-genet-112414-054804 (visited on 10/15/2017).

[15] Kristian Moss Bendtsen, János Erdőssy, Zsolt Csiszovszki, Sine Lo Svenningsen, Kim Sneppen, Sandeep Krishna, and Szabolcs Semsey. "Direct and indirect effects in the regulation of overlapping promoters." In: *Nucleic Acids Research* 39.16 (Sept. 2011), pp. 6879–6885. ISSN: 0305-1048. DOI: 10.1093/nar/gkr390. URL: https://academic.oup.com/nar/article/39/16/6879/2411724/Direct-and-indirect-effe%20cts-in-the-regulation-of (visited on 10/26/2017).

[16] B. A. Bensing, B. J. Meyer, and G. M. Dunny. "Sensitive detection of bacterial transcription initiation sites and differentiation from RNA processing sites in the pheromone-induced plasmid transfer system of Enterococcus faecalis." en. In: *Proceedings of the National Academy of Sciences* 93.15 (July 1996), pp. 7794–7799. ISSN: 0027-8424, 1091-6490. URL: http://www.pnas.org/content/93/15/7794 (visited on 10/05/2017).

[17] Thorsten Bischler, Ping-kun Hsieh, Marcus Resch, Quansheng Liu, Hock Siew Tan, Patricia L. Foley, Anika Hartleib, Cynthia M. Sharma, and Joel G. Belasco. "Identification of the RNA Pyrophosphohydrolase RppH of Helicobacter pylori and Global Analysis of Its RNA Targets." en. In: *Journal of Biological Chemistry* 292.5 (Feb. 2017), pp. 1934–1950. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.M116.761171. URL: http://www.jbc.org/content/292/5/1934 (visited on 02/20/2017).

[18] Thorsten Bischler, Matthias Kopf, and Björn Voß. "Transcript mapping based on dRNA-seq data." en. In: *BMC Bioinformatics* 15.1 (Apr. 2014), p. 122. ISSN: 1471-2105. DOI: 10.1186/1471-2105-15-122. URL: http://www.biomedcentral.com/1471-2105/15/122/abstract (visited on 05/12/2014).

[19] Thorsten Bischler, Hock Siew Tan, Kay Nieselt, and Cynthia M. Sharma. "Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in Helicobacter pylori." In: *Methods*. Bacterial and Archaeal Transcription 86 (Sept. 2015), pp. 89–101. ISSN: 1046-2023. DOI:

10.1016/j.ymeth.2015.06.012. URL: http://www.sciencedirect.com/science/article/pii/S1046202315002546 (visited on 02/29/2016).

[20] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. "Near-optimal probabilistic RNA-seq quantification." en. In: *Nature Biotechnology* 34.5 (May 2016), pp. 525–527. ISSN: 1087-0156. DOI: 10.1038/nbt.3519. URL: http://www.nature.com/nbt/journal/v34/n5/full/nbt.3519.html?foxtrotcallback=true (visited on 09/26/2017).

[21] Alayne L. Brunner et al. "Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver." en. In: *Genome Research* 19.6 (June 2009), pp. 1044–1056. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.088773.108. URL: http://genome.cshlp.org/content/19/6/1044 (visited on 10/19/2017).

[22] James H. Bullard, Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments." In: *BMC Bioinformatics* 11 (Feb. 2010), p. 94. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-94. URL: https://doi.org/10.1186/1471-2105-11-94 (visited on 09/28/2017).

[23] Alfredo Castello et al. "Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins." In: *Cell* 149.6 (June 2012), pp. 1393–1406. ISSN: 0092-8674. DOI: 10.1016/j.cell.2012.04.031. URL: http://www.sciencedirect.com/science/article/pii/S0092867412005764 (visited on 06/18/2012).

[24] Helena Celesnik, Atilio Deana, and Joel G. Belasco. "Initiation of RNA Decay in Escherichia coli by 5′ Pyrophosphate Removal." In: *Molecular Cell* 27.1 (July 2007), pp. 79–90. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2007.05.038. URL: http://www.sciencedirect.com/science/article/pii/S1097276507003644 (visited on 10/11/2017).

[25] Helena Celesnik, Atilio Deana, and Joel G. Belasco. "Chapter 5 PABLO Analysis of RNA: 5′-Phosph orylation State and 5′-End Mapping." In: *Methods in Enzymology*. Vol. 447. RNA Turnover in Bacteria, Archaea and Organelles. DOI: 10.1016/S0076-6879(08)02205-2. Academic Press, Jan. 2008, pp. 83–98. URL: http://www.sciencedirect.com/science/article/pii/S0076687908022052 (visited on 10/11/2017).

[26] Yanjie Chao, Kai Papenfort, Richard Reinhardt, Cynthia M Sharma, and Jörg Vogel. "An atlas of Hfq-bound transcripts reveals 3′ UTRs as a genomic reservoir of regulatory small RNAs." en. In: *The EMBO Journal* 31.20 (Oct. 2012), pp. 4005–4019. ISSN: 1460-2075. DOI: 10.1038/emboj.2012.229.

URL: http://onlinelibrary.wiley.com/doi/10.1038/emboj.2012.229/abstract (visited on 10/27/2017).

[27] Yanjie Chao and Jörg Vogel. "The role of Hfq in bacterial pathogens." In: *Current Opinion in Microbiology*. Host-microbe interactions: bacteria 13.1 (Feb. 2010), pp. 24–33. ISSN: 1369-5274. DOI: 10.1016/j.mib.2010.01.001. URL: http://www.sciencedirect.com/science/article/pii/S1369527410000032 (visited on 10/16/2017).

[28] Yanjie Chao et al. "In Vivo Cleavage Map Illuminates the Central Role of RNase E in Coding and Non-coding RNA Pathways." In: *Molecular Cell* 65.1 (Jan. 2017), pp. 39–51. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2016.11.002. URL: http://www.sciencedirect.com/science/article/pii/S1097276516307109 (visited on 12/31/2017).

[29] Byung-Kwan Cho, Donghyuk Kim, Eric M. Knight, Karsten Zengler, and Bernhard O. Palsson. "Genome-scale reconstruction of the sigma factor network in Escherichia coli: topology and functional states." In: *BMC Biology* 12 (Jan. 2014), p. 4. ISSN: 1741-7007. DOI: 10.1186/1741-7007-12-4. URL: https://doi.org/10.1186/1741-7007-12-4 (visited on 10/05/2017).

[30] Byung-Kwan Cho, Karsten Zengler, Yu Qiu, Young Seoub Park, Eric M. Knight, Christian L. Barrett, Yuan Gao, and Bernhard Ø Palsson. "The transcription unit architecture of the Escherichia coli genome." en. In: *Nature Biotechnology* 27.11 (Nov. 2009), pp. 1043–1049. ISSN: 1087-0156. DOI: 10.1038/nbt.1582. URL: https://www.nature.com/nbt/journal/v27/n11/full/nbt.1582.html (visited on 10/05/2017).

[31] James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. "Continuous base identification for single-molecule nanopore DNA sequencing." en. In: *Nature Nanotechnology* 4.4 (Apr. 2009), pp. 265–270. ISSN: 1748-3387. DOI: 10.1038/nnano.2009.12. URL: http://www.nature.com/nnano/journal/v4/n4/abs/nnano.2009.12.html?foxtrotcallback%20=true (visited on 08/30/2017).

[32] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants." In: *Nucleic Acids Research* 38.6 (Apr. 2010), pp. 1767–1771. ISSN: 0305-1048. DOI: 10.1093/nar/gkp1137. URL: https://academic.oup.com/nar/article/38/6/1767/3112533/The-Sanger-FASTQ-file-for%20mat-for-sequences-with (visited on 09/27/2017).

[33] Francis S. Collins, Michael Morgan, and Aristides Patrinos. "The Human Genome Project: Lessons from Large-Scale Biology." en. In: *Science* 300.5617 (Apr. 2003), pp. 286–290. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1084564. URL: http://science.sciencemag.org/content/300/5617/286 (visited on 08/14/2017).

[34] Ana Conesa et al. "A survey of best practices for RNA-seq data analysis." In: *Genome Biology* 17 (Jan. 2016), p. 13. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0881-8. URL: https://doi.org/10.1186/s13059-016-0881-8 (visited on 09/24/2017).

[35] Tyrrell Conway et al. "Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing." en. In: *mBio* 5.4 (Aug. 2014), e01442–14. ISSN: , 2150-7511. DOI: 10.1128/mBio.01442-14. URL: http://mbio.asm.org/content/5/4/e01442-14 (visited on 11/03/2014).

[36] David L. Corcoran, Stoyan Georgiev, Neelanjan Mukherjee, Eva Gottwein, Rebecca L. Skalsky, Jack D. Keene, and Uwe Ohler. "PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data." In: *Genome Biology* 12 (Aug. 2011), R79. ISSN: 1474-760X. DOI: 10.1186/gb-2011-12-8-r79. URL: https://doi.org/10.1186/gb-2011-12-8-r79 (visited on 10/18/2017).

[37] Timothy L. Cover and Martin J. Blaser. "Helicobacter pylori in Health and Disease." In: *Gastroenterology*. Intestinal Microbes in Health and Disease 136.6 (May 2009), pp. 1863–1873. ISSN: 0016-5085. DOI: 10.1053/j.gastro.2009.01.073. URL: http://www.sciencedirect.com/science/article/pii/S0016508509003394 (visited on 10/09/2017).

[38] Nicholas J Croucher and Nicholas R Thomson. "Studying bacterial transcriptomes using RNA-seq." In: *Current Opinion in Microbiology*. Antimicrobials/Genomics 13.5 (Oct. 2010), pp. 619–624. ISSN: 1369-5274. DOI: 10.1016/j.mib.2010.09.009. URL: http://www.sciencedirect.com/science/article/pii/S1369527410001360 (visited on 09/19/2017).

[39] Daniel Dar, Maya Shamir, J. R. Mellin, Mikael Koutero, Noam Stern-Ginossar, Pascale Cossart, and Rotem Sorek. "Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria." en. In: *Science* 352.6282 (Apr. 2016), aad9822. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aad9822. URL: http://science.sciencemag.org/content/352/6282/aad9822 (visited on 10/25/2017).

[40] Javid I. Dasti, A. Malik Tareen, Raimond Lugert, Andreas E. Zautner, and Uwe Groß. "Campylobacter jejuni: A brief overview on pathogenicity-associated factors and disease-mediating mechanisms." In: *International Journal of Medical Microbiology* 300.4 (Apr. 2010), pp. 205–211. ISSN: 1438-4221. DOI: 10.1016/j.ijmm.2009.07.002. URL: http://www.sciencedirect.com/science/article/pii/S1438422109000575 (visited on 10/13/2017).

[41] Atilio Deana, Helena Celesnik, and Joel G. Belasco. "The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal." en. In: *Nature* 451.7176 (Jan. 2008), pp. 355–358. ISSN: 0028-0836. DOI: 10.1038/nature06475. URL: https://www.nature.com/nature/journal/v451/n7176/full/nature06475.html (visited on 10/11/2017).

[42] Marie-Agnès Dillies et al. "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis." In: *Briefings in Bioinformatics* 14.6 (Nov. 2013), pp. 671–683. ISSN: 1467-5463. DOI: 10.1093/bib/bbs046. URL: https://academic.oup.com/bib/article/14/6/671/189645/A-comprehensive-evaluation-%20of-normalization (visited on 09/28/2017).

[43] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. "STAR: ultrafast universal RNA-seq aligner." In: *Bioinformatics* 29.1 (Jan. 2013), pp. 15–21. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts635. URL: https://academic.oup.com/bioinformatics/article/29/1/15/272537/STAR-ultrafast-un%20iversal-RNA-seq-aligner (visited on 09/26/2017).

[44] James E. Dornenburg, Anne M. DeVita, Michael J. Palumbo, and Joseph T. Wade. "Widespread Antisense Transcription in Escherichia coli." en. In: *mBio* 1.1 (May 2010), e00024–10. ISSN: , 2150-7511. DOI: 10.1128/mBio.00024-10. URL: http://mbio.asm.org/content/1/1/e00024-10 (visited on 10/12/2017).

[45] Ashok K. Dubey, Carol S. Baker, Tony Romeo, and Paul Babitzke. "RNA sequence and secondary structure participate in high-affinity CsrA–RNA interaction." en. In: *RNA* 11.10 (Oct. 2005), pp. 1579–1587. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.2990205. URL: http://rnajournal.cshlp.org/content/11/10/1579 (visited on 10/13/2017).

[46] Gaurav Dugar, Alexander Herbig, Konrad U. Förstner, Nadja Heidrich, Richard Reinhardt, Kay Nieselt, and Cynthia M. Sharma. "High-Resolution Transcriptome Maps Reveal Strain-Specific Regulatory Features of Multi-

ple Campylobacter jejuni Isolates." In: *PLoS Genet* 9.5 (May 2013), e1003495. DOI: 10.1371/journal.pgen.1003495. URL: http://dx.doi.org/10.1371/journal.pgen.1003495 (visited on 07/09/2013).

[47] Gaurav Dugar, Sarah L. Svensson, Thorsten Bischler, Sina Wäldchen, Richard Reinhardt, Markus Sauer, and Cynthia M. Sharma. "The CsrA-FliW network controls polar localization of the dual-function flagellin mRNA in Campylobacter jejuni." en. In: *Nature Communications* 7 (May 2016), p. 11667. DOI: 10.1038/ncomms11667. URL: http://www.nature.com/ncomms/2016/160527/ncomms11667/abs/ncomms11667.html (visited on 07/13/2016).

[48] Olivier Duss, Erich Michel, Maxim Yulikov, Mario Schubert, Gunnar Jeschke, and Frédéric H.-T. Allain. "Structural basis of the non-coding RNA RsmZ acting as a protein sponge." en. In: *Nature* 509.7502 (May 2014), pp. 588–592. ISSN: 0028-0836. DOI: 10.1038/nature13271. URL: https://www.nature.com/nature/journal/v509/n7502/full/nature13271.html (visited on 10/16/2017).

[49] John Eid et al. "Real-Time DNA Sequencing from Single Polymerase Molecules." en. In: *Science* 323.5910 (Jan. 2009), pp. 133–138. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1162986. URL: http://science.sciencemag.org/content/323/5910/133 (visited on 08/30/2017).

[50] Laurence Ettwiller, John Buswell, Erbay Yigit, and Ira Schildkraut. "A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome." In: *BMC Genomics* 17 (2016), p. 199. ISSN: 1471-2164. DOI: 10.1186/s12864-016-2539-z. URL: http://dx.doi.org/10.1186/s12864-016-2539-z (visited on 03/07/2017).

[51] Brent Ewing and Phil Green. "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities." en. In: *Genome Research* 8.3 (Mar. 1998), pp. 186–194. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.8.3.186. URL: http://genome.cshlp.org/content/8/3/186 (visited on 09/22/2017).

[52] Milan Fedurco, Anthony Romieu, Scott Williams, Isabelle Lawrence, and Gerardo Turcatti. "BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies." In: *Nucleic Acids Research* 34.3 (Jan. 2006), e22–e22. ISSN: 0305-1048. DOI: 10.1093/nar/gnj023. URL: https://academic.oup.com/nar/article/34/3/e22/1048340/BTA-a-novel-reagent-for-DN%20A-attachment-on-glass (visited on 08/22/2017).

[53] Andrey Feklístov, Brian D. Sharon, Seth A. Darst, and Carol A. Gross. "Bacterial Sigma Factors: A Historical, Structural, and Genomic Perspective." In: *Annual Review of Microbiology* 68.1 (Sept. 2014), pp. 357–376. ISSN: 0066-4227. DOI: 10.1146/annurev-micro-092412-155737. URL: http://www.annualreviews.org/doi/10.1146/annurev-micro-092412-155737 (visited on 10/26/2017).

[54] Jianxing Feng, Clifford A. Meyer, Qian Wang, Jun S. Liu, X. Shirley Liu, and Yong Zhang. "GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data." en. In: *Bioinformatics* 28.21 (Nov. 2012), pp. 2782–2788. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bts515. URL: http://bioinformatics.oxfordjournals.org/content/28/21/2782 (visited on 11/12/2012).

[55] Joshua A. Fields and Stuart A. Thompson. "Campylobacter jejuni CsrA Mediates Oxidative Stress Responses, Biofilm Formation, and Host Cell Invasion." en. In: *Journal of Bacteriology* 190.9 (May 2008), pp. 3411–3416. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/JB.01928-07. URL: http://jb.asm.org/content/190/9/3411 (visited on 12/15/2017).

[56] Konrad U. Förstner, Jörg Vogel, and Cynthia M. Sharma. "READemption—a tool for the computational analysis of deep-sequencing–based transcriptome data." en. In: *Bioinformatics* 30.23 (Dec. 2014), pp. 3421–3423. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btu533. URL: http://bioinformatics.oxfordjournals.org/content/30/23/3421 (visited on 09/16/2015).

[57] Nowlan H. Freese, David C. Norris, and Ann E. Loraine. "Integrated genome browser: visual analytics platform for genomics." en. In: *Bioinformatics* 32.14 (July 2016), pp. 2089–2095. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btw069. URL: http://bioinformatics.oxfordjournals.org/content/32/14/2089 (visited on 12/21/2016).

[58] Jorge Frias-Lopez, Yanmei Shi, Gene W. Tyson, Maureen L. Coleman, Stephan C. Schuster, Sallie W. Chisholm, and Edward F. DeLong. "Microbial community gene expression in ocean surface waters." en. In: *Proceedings of the National Academy of Sciences* 105.10 (Mar. 2008), pp. 3805–3810. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0708897105. URL: http://www.pnas.org/content/105/10/3805 (visited on 10/23/2017).

[59] Shan Gao, Yipeng Ren, Yu Sun, Zhenfeng Wu, Jishou Ruan, Bingjun He, Tao Zhang, Xin Yu, Xiaoxuan Tian, and Wenjun Bu. "PacBio full-length transcriptome profiling of insect mitochondrial gene expression." In: *RNA Biol-*

ogy 13.9 (Sept. 2016), pp. 820–825. ISSN: 1547-6286. DOI: 10.1080/15476286.2016.1197481. URL: http://dx.doi.org/10.1080/15476286.2016.1197481 (visited on 10/26/2017).

[60] Daniel R. Garalde et al. "Highly parallel direct RNA sequencing on an array of nanopores." en. In: *bioRxiv* (Aug. 2016), p. 068809. DOI: 10.1101/068809. URL: https://www.biorxiv.org/content/early/2016/08/12/068809 (visited on 11/16/2017).

[61] Manuel Garber, Manfred G. Grabherr, Mitchell Guttman, and Cole Trapnell. "Computational methods for transcriptome annotation and quantification using RNA-seq." en. In: *Nature Methods* 8.6 (June 2011), pp. 469–477. ISSN: 1548-7091. DOI: 10.1038/nmeth.1613. URL: https://www.nature.com/nmeth/journal/v8/n6/full/nmeth.1613.html (visited on 10/24/2017).

[62] Paul P. Gardner, Lars Barquist, Alex Bateman, Eric P. Nawrocki, and Zasha Weinberg. "RNIE: genome-wide prediction of bacterial intrinsic terminators." In: *Nucleic Acids Research* 39.14 (Aug. 2011), pp. 5845–5852. ISSN: 0305-1048. DOI: 10.1093/nar/gkr168. URL: https://academic.oup.com/nar/article/39/14/5845/1374243/RNIE-genome-wide-predict%20ion-of-bacterial-intrinsic (visited on 10/25/2017).

[63] Jens Georg and Wolfgang R. Hess. "cis-Antisense RNA, Another Level of Gene Regulation in Bacteria." en. In: *Microbiology and Molecular Biology Reviews* 75.2 (June 2011), pp. 286–300. ISSN: 1092-2172, 1098-5557. DOI: 10.1128/MMBR.00032-10. URL: http://mmbr.asm.org/content/75/2/286 (visited on 10/11/2017).

[64] Sara Goodwin, John D. McPherson, and W. Richard McCombie. "Coming of age: ten years of next-generation sequencing technologies." en. In: *Nature Reviews Genetics* 17.6 (June 2016), pp. 333–351. ISSN: 1471-0056. DOI: 10.1038/nrg.2016.49. URL: http://www.nature.com/nrg/journal/v17/n6/full/nrg.2016.49.html (visited on 07/17/2017).

[65] Manfred G. Grabherr et al. "Full-length transcriptome assembly from RNA-Seq data without a reference genome." en. In: *Nature Biotechnology* 29.7 (July 2011), pp. 644–652. ISSN: 1087-0156. DOI: 10.1038/nbt.1883. URL: http://www.nature.com/nbt/journal/v29/n7/full/nbt.1883.html?foxtrotcallback=true (visited on 09/24/2017).

[66] Brian J. Haas, Melissa Chin, Chad Nusbaum, Bruce W. Birren, and Jonathan Livny. "How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes?" In: *BMC Genomics* 13 (Dec. 2012), p. 734. ISSN: 1471-2164. DOI:

10.1186/1471-2164-13-734. URL: https://doi.org/10.1186/1471-2164-13-734 (visited on 10/12/2017).

[67] Brian J. Haas et al. "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." en. In: *Nature Protocols* 8.8 (Aug. 2013), pp. 1494–1512. ISSN: 1754-2189. DOI: 10.1038/nprot.2013.084. URL: https://www.nature.com/nprot/journal/v8/n8/full/nprot.2013.084.html (visited on 09/24/2017).

[68] Markus Hafner et al. "Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP." In: *Cell* 141.1 (Apr. 2010), pp. 129–141. ISSN: 0092-8674. DOI: 10.1016/j.cell.2010.03.009. URL: http://www.sciencedirect.com/science/article/pii/S009286741000245X (visited on 10/17/2017).

[69] Hannon Lab. *FASTX Toolkit*. URL: http://hannonlab.cshl.edu/fastx_toolkit/.

[70] Thomas J. Hardcastle and Krystyna A. Kelly. "baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data." In: *BMC Bioinformatics* 11.1 (Aug. 2010), p. 422. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-422. URL: https://doi.org/10.1186/1471-2105-11-422 (visited on 09/29/2017).

[71] Steven R. Head, H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, and Phillip Ordoukhanian. "Library construction for next-generation sequencing: Overviews and challenges." In: *BioTechniques* 56.2 (Feb. 2014). ISSN: 1940-9818, 0736-6205. DOI: 10.2144/000114133. URL: http://www.biotechniques.com/BiotechniquesJournal/2014/February/Library-construc%20tion-for-next-generation-sequencing-Overviews-and-challenges/biotechniques-34988%209.html (visited on 10/17/2017).

[72] Magali Hébrard, Carsten Kröger, Shabarinath Srikumar, Aoife Colgan, Kristian Händler, and Jay C. D. Hinton. "sRNAs and the virulence of Salmonella enterica serovar Typhimurium." In: *RNA Biology* 9.4 (Apr. 2012), pp. 437–445. ISSN: 1547-6286. DOI: 10.4161/rna.20480. URL: http://dx.doi.org/10.4161/rna.20480 (visited on 10/18/2017).

[73] Nadja Heidrich, Gaurav Dugar, Jörg Vogel, and Cynthia M. Sharma. "Investigating CRISPR RNA Biogenesis and Function Using RNA-seq." en. In: *CRISPR*. Methods in Molecular Biology. DOI: 10.1007/978-1-4939-2687-9_1. Humana Press, New York, NY, 2015, pp. 1–21. ISBN: 978-1-4939-2686-2 978-

1-4939-2687-9. URL: https://link.springer.com/protocol/10.1007/978-1-4939-2687-9_1 (visited on 10/25/2017).

[74] Jonathan Hetzel, Sascha H. Duttke, Christopher Benner, and Joanne Chory. "Nascent RNA sequencing reveals distinct features in plant transcription." en. In: *Proceedings of the National Academy of Sciences* 113.43 (Oct. 2016), pp. 12316–12321. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1603217113. URL: http://www.pnas.org/content/113/43/12316 (visited on 11/16/2017).

[75] Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M. Sharma, Philipp Khaitovich, Jörg Vogel, Peter F. Stadler, and Jörg Hackermüller. "Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures." In: *PLoS Comput Biol* 5.9 (Sept. 2009), e1000502. DOI: 10.1371/journal.pcbi.1000502. URL: http://dx.doi.org/10.1371/journal.pcbi.1000502 (visited on 11/27/2012).

[76] Erik Holmqvist, Patrick R. Wright, Lei Li, Thorsten Bischler, Lars Barquist, Richard Reinhardt, Rolf Backofen, and Jörg Vogel. "Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo." en. In: *The EMBO Journal* (Apr. 2016), e201593360. ISSN: 0261-4189, 1460-2075. URL: http://emboj.embopress.org/content/early/2016/04/04/embj.201593360 (visited on 04/05/2016).

[77] Ping-kun Hsieh, Jamie Richards, Quansheng Liu, and Joel G. Belasco. "Specificity of RppH-dependent RNA degradation in Bacillus subtilis." en. In: *Proceedings of the National Academy of Sciences* (Apr. 2013). ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1222670110. URL: http://www.pnas.org/content/early/2013/04/19/1222670110 (visited on 05/24/2013).

[78] Monica P. Hui, Patricia L. Foley, and Joel G. Belasco. "Messenger RNA Degradation in Bacterial Cells." In: *Annual Review of Genetics* 48.1 (2014), pp. 537–559. DOI: 10.1146/annurev-genet-120213-092340. URL: https://doi.org/10.1146/annurev-genet-120213-092340 (visited on 10/11/2017).

[79] A. Hüttenhofer. "RNomics: Identification and Function of Small Non-Protein-coding RNAs in Model Organisms." en. In: *Cold Spring Harbor Symposia on Quantitative Biology* 71 (Jan. 2006), pp. 135–140. ISSN: 0091-7451, 1943-4456. DOI: 10.1101/sqb.2006.71.007. URL: http://symposium.cshlp.org/content/71/135 (visited on 11/11/2017).

[80] Mahmoud M. Ibrahim, Scott A. Lacadie, and Uwe Ohler. "JAMM: a peak finder for joint analysis of NGS replicates." en. In: *Bioinformatics* 31.1 (Jan. 2015), pp. 48–55. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/

btu568. URL: http://bioinformatics.oxfordjournals.org/content/31/1/48 (visited on 04/19/2016).

[81] Nicholas T. Ingolia. "Ribosome profiling: new views of translation, from single codons to genome scale." En. In: *Nature Reviews Genetics* 15.3 (Mar. 2014), p. 205. ISSN: 1471-0064. DOI: 10.1038/nrg3645. URL: https://www.nature.com/articles/nrg3645 (visited on 12/14/2017).

[82] Nicholas T. Ingolia, Sina Ghaemmaghami, John R. S. Newman, and Jonathan S. Weissman. "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling." en. In: *Science* 324.5924 (Apr. 2009), pp. 218–223. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1168978. URL: http://www.sciencemag.org/content/324/5924/218 (visited on 12/08/2015).

[83] Nicolas Innocenti et al. "Whole-genome mapping of 5′ RNA ends in bacteria by tagged sequencing: a comprehensive view in Enterococcus faecalis." en. In: *RNA* 21.5 (May 2015), pp. 1018–1030. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.048470.114. URL: http://rnajournal.cshlp.org/content/21/5/1018 (visited on 05/26/2015).

[84] Hadi Jorjani and Mihaela Zavolan. "TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data." In: *Bioinformatics* 30.7 (Apr. 2014), pp. 971–974. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt752. URL: https://academic.oup.com/bioinformatics/article/30/7/971/236374/TSSer-an-automat%20ed-method-to-identify (visited on 09/29/2017).

[85] Vladimir R. Kaberdin, Dharam Singh, and Sue Lin-Chao. "Composition and conservation of the mRNA-degrading machinery in bacteria." In: *Journal of Biomedical Science* 18 (Mar. 2011), p. 23. ISSN: 1423-0127. DOI: 10.1186/1423-0127-18-23. URL: https://doi.org/10.1186/1423-0127-18-23 (visited on 10/11/2017).

[86] Yarden Katz, Eric T. Wang, Edoardo M. Airoldi, and Christopher B. Burge. "Analysis and design of RNA sequencing experiments for identifying isoform regulation." en. In: *Nature Methods* 7.12 (Dec. 2010), pp. 1009–1015. ISSN: 1548-7091. DOI: 10.1038/nmeth.1528. URL: https://www.nature.com/nmeth/journal/v7/n12/full/nmeth.1528.html (visited on 10/24/2017).

[87] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." In: *Genome Biology*

14 (Apr. 2013), R36. ISSN: 1474-760X. DOI: 10.1186/gb-2013-14-4-r36. URL: https://doi.org/10.1186/gb-2013-14-4-r36 (visited on 09/26/2017).

[88] Donghyuk Kim, Jay Sung-Joong Hong, Yu Qiu, Harish Nagarajan, Joo-Hyun Seo, Byung-Kwan Cho, Shih-Feng Tsai, and Bernhard Ø Palsson. "Comparative Analysis of Regulatory Elements between Escherichia coli and Klebsiella pneumoniae by Genome-Wide Transcription Start Site Profiling." In: *PLOS Genetics* 8.8 (Aug. 2012), e1002867. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002867. URL: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002867 (visited on 10/05/2017).

[89] Julian König, Kathi Zarnack, Gregor Rot, Tomaž Curk, Melis Kayikci, Blaž Zupan, Daniel J. Turner, Nicholas M. Luscombe, and Jernej Ule. "iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution." en. In: *Nature Structural & Molecular Biology* 17.7 (July 2010), pp. 909–915. ISSN: 1545-9993. DOI: 10.1038/nsmb.1838. URL: https://www.nature.com/nsmb/journal/v17/n7/full/nsmb.1838.html (visited on 10/17/2017).

[90] Katharina Kramer, Timo Sachsenberg, Benedikt M. Beckmann, Saadia Qamar, Kum-Loong Boon, Matthias W. Hentze, Oliver Kohlbacher, and Henning Urlaub. "Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins." en. In: *Nature Methods* 11.10 (Oct. 2014), pp. 1064–1070. ISSN: 1548-7091. DOI: 10.1038/nmeth.3092. URL: https://www.nature.com/nmeth/journal/v11/n10/full/nmeth.3092.html (visited on 10/16/2017).

[91] Carsten Kröger et al. "The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium." en. In: *Proceedings of the National Academy of Sciences* 109.20 (May 2012), E1277–E1286. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1201061109. URL: http://www.pnas.org/content/109/20/E1277 (visited on 10/12/2017).

[92] Carsten Kröger et al. "An Infection-Relevant Transcriptomic Compendium for Salmonella enterica Serovar Typhimurium." In: *Cell Host & Microbe* 14.6 (Dec. 2013), pp. 683–695. ISSN: 1931-3128. DOI: 10.1016/j.chom.2013.11.010. URL: http://www.sciencedirect.com/science/article/pii/S1931312813004113 (visited on 10/14/2014).

[93] Paweł P. Łabaj, Germán G. Leparc, Bryan E. Linggi, Lye Meng Markillie, H. Steven Wiley, and David P. Kreil. "Characterization and improvement of RNA-Seq precision in quantitative transcript expression profil-

ing." In: *Bioinformatics* 27.13 (July 2011), pp. i383–i391. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr247. URL: https://academic.oup.com/bioinformatics/article/27/13/i383/183214/Characterizati%20on-and-improvement-of-RNA-Seq (visited on 10/24/2017).

[94] Nicholas F. Lahens et al. "IVT-seq reveals extreme bias in RNA sequencing." In: *Genome Biology* 15 (June 2014), R86. ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-6-r86. URL: https://doi.org/10.1186/gb-2014-15-6-r86 (visited on 10/23/2017).

[95] David Langenberger, Clara Bermudez-Santana, Jana Hertel, Steve Hoffmann, Philipp Khaitovich, and Peter F. Stadler. "Evidence for human microRNA-offset RNAs in small RNA sequencing data." en. In: *Bioinformatics* 25.18 (Sept. 2009), pp. 2298–2301. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btp419. URL: http://bioinformatics.oxfordjournals.org/content/25/18/2298 (visited on 04/27/2016).

[96] Ben Langmead and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." en. In: *Nature Methods* 9.4 (Apr. 2012), pp. 357–359. ISSN: 1548-7091. DOI: 10.1038/nmeth.1923. URL: https://www.nature.com/nmeth/journal/v9/n4/full/nmeth.1923.html (visited on 09/25/2017).

[97] Iñigo Lasa et al. "Genome-wide antisense transcription drives mRNA processing in bacteria." en. In: *Proceedings of the National Academy of Sciences* 108.50 (Dec. 2011), pp. 20172–20177. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1113521108. URL: http://www.pnas.org/content/108/50/20172 (visited on 10/11/2017).

[98] Charity W. Law, Yunshun Chen, Wei Shi, and Gordon K. Smyth. "voom: precision weights unlock linear model analysis tools for RNA-seq read counts." In: *Genome Biology* 15.2 (Feb. 2014), R29. ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-2-r29. URL: https://doi.org/10.1186/gb-2014-15-2-r29 (visited on 09/29/2017).

[99] Sara D. Lawhon, Jonathan G. Frye, Mitsu Suyemoto, Steffen Porwollik, Michael McClelland, and Craig Altier. "Global regulation by CsrA in Salmonella typhimurium." en. In: *Molecular Microbiology* 48.6 (June 2003), pp. 1633–1645. ISSN: 1365-2958. DOI: 10.1046/j.1365-2958.2003.03535.x. URL: http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2958.2003.03535.x/abstract (visited on 11/12/2017).

[100] Ning Leng, John A. Dawson, James A. Thomson, Victor Ruotti, Anna I. Rissman, Bart M. G. Smits, Jill D. Haag, Michael N. Gould, Ron M. Stewart, and Christina Kendziorski. "EBSeq: an empirical Bayes hierarchical model

for inference in RNA-seq experiments." In: *Bioinformatics* 29.8 (Apr. 2013), pp. 1035–1043. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt087. URL: https://academic.oup.com/bioinformatics/article/29/8/1035/228913/EBSeq-an-empiri%20cal-Bayes-hierarchical-model-for (visited on 09/29/2017).

[101] Joshua Z. Levin, Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, Andreas Gnirke, and Aviv Regev. "Comprehensive comparative analysis of strand-specific RNA sequencing methods." en. In: *Nature Methods* 7.9 (Sept. 2010), pp. 709–715. ISSN: 1548-7091. DOI: 10.1038/nmeth.1491. URL: https://www.nature.com/nmeth/journal/v7/n9/full/nmeth.1491.html (visited on 09/25/2017).

[102] Bo Li and Colin N. Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." In: *BMC Bioinformatics* 12 (Aug. 2011), p. 323. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-323. URL: https://doi.org/10.1186/1471-2105-12-323 (visited on 09/26/2017).

[103] Jun Li and Robert Tibshirani. "Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data." en. In: *Statistical Methods in Medical Research* 22.5 (Oct. 2013), pp. 519–536. ISSN: 0962-2802. DOI: 10.1177/0962280211428386. URL: https://doi.org/10.1177/0962280211428386 (visited on 09/29/2017).

[104] Jun Li, Daniela M. Witten, Iain M. Johnstone, and Robert Tibshirani. "Normalization, testing, and false discovery rate estimation for RNA-sequencing data." In: *Biostatistics* 13.3 (July 2012), pp. 523–538. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxr031. URL: https://academic.oup.com/biostatistics/article/13/3/523/248016/Normalization-tes%20ting-and-false-discovery-rate (visited on 09/28/2017).

[105] Shan Li, Xia Dong, and Zhengchang Su. "Directional RNA-seq reveals highly complex condition-dependent transcriptomes in E. coli K12 through accurate full-length transcripts assembling." In: *BMC Genomics* 14 (July 2013), p. 520. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-520. URL: https://doi.org/10.1186/1471-2164-14-520 (visited on 10/12/2017).

[106] Yue Li, Dorothy Yanling Zhao, Jack F. Greenblatt, and Zhaolei Zhang. "RIPSeeker: a statistical package for identifying protein-associated transcripts from RIP-seq experiments." en. In: *Nucleic Acids Research* 41.8 (Apr. 2013), e94–e94. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkt142.

URL: http://nar.oxfordjournals.org/content/41/8/e94 (visited on 04/24/2013).

[107] Yu-fei Lin, David Romero A, Shuang Guan, Lira Mamanova, and Kenneth J. McDowall. "A combination of improved differential and global RNA-seq reveals pervasive transcription initiation and events in all stages of the life-cycle of functional RNAs in Propionibacterium acnes, a major contributor to wide-spread human disease." In: *BMC Genomics* 14 (Sept. 2013), p. 620. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-620. URL: https://doi.org/10.1186/1471-2164-14-620 (visited on 10/12/2017).

[108] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. *Comparison of Next-Generation Sequencing Systems*. en. Research article. DOI: 10.1155/2012/251364. 2012. URL: https://www.hindawi.com/journals/bmri/2012/251364/ (visited on 08/08/2017).

[109] Verónica Lloréns-Rico, Jaime Cano, Tjerko Kamminga, Rosario Gil, Amparo Latorre, Wei-Hua Chen, Peer Bork, John I. Glass, Luis Serrano, and Maria Lluch-Senar. "Bacterial antisense RNAs are mainly the product of transcriptional noise." en. In: *Science Advances* 2.3 (Mar. 2016), e1501363. ISSN: 2375-2548. DOI: 10.1126/sciadv.1501363. URL: http://advances.sciencemag.org/content/2/3/e1501363 (visited on 10/02/2017).

[110] Michael T. Lovci et al. "Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges." en. In: *Nature Structural & Molecular Biology* 20.12 (Dec. 2013), pp. 1434–1442. ISSN: 1545-9993. DOI: 10.1038/nsmb.2699. URL: https://www.nature.com/nsmb/journal/v20/n12/full/nsmb.2699.html (visited on 10/18/2017).

[111] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." In: *Genome Biology* 15 (2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. URL: http://dx.doi.org/10.1186/s13059-014-0550-8 (visited on 08/16/2016).

[112] Daniel J. Luciano, Monica P. Hui, Atilio Deana, Patricia L. Foley, Kevin J. Belasco, and Joel G. Belasco. "Differential Control of the Rate of 5'-End-Dependent mRNA Degradation in Escherichia coli." en. In: *Journal of Bacteriology* 194.22 (Nov. 2012), pp. 6233–6239. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/JB.01223-12. URL: http://jb.asm.org/content/194/22/6233 (visited on 02/07/2013).

[113] George A. Mackie. "Ribonuclease E is a 5'-end-dependent endonuclease." en. In: *Nature* 395.6703 (Oct. 1998), pp. 720–724. ISSN: 0028-0836. DOI: 10.1038/27246. URL: https://www.nature.com/nature/journal/v395/n6703/full/395720a0.html (visited on 10/11/2017).

[114] Fenglou Mao, Phuongan Dam, Jacky Chou, Victor Olman, and Ying Xu. "DOOR: a database for prokaryotic operons." In: *Nucleic Acids Research* 37.suppl_1 (Jan. 2009), pp. D459–D463. ISSN: 0305-1048. DOI: 10.1093/nar/gkn757. URL: https://academic.oup.com/nar/article/37/suppl_1/D459/1008910/DOOR-a-database-for%20-prokaryotic-operons (visited on 10/25/2017).

[115] Elaine R. Mardis. "Next-Generation Sequencing Platforms." In: *Annual Review of Analytical Chemistry* 6.1 (2013), pp. 287–303. DOI: 10.1146/annurev-anchem-062012-092628. URL: https://doi.org/10.1146/annurev-anchem-062012-092628 (visited on 08/14/2017).

[116] Marcel Margulies et al. "Genome sequencing in microfabricated high-density picolitre reactors." en. In: *Nature* 437.7057 (Sept. 2005), pp. 376–380. ISSN: 0028-0836. DOI: 10.1038/nature03959. URL: http://www.nature.com/nature/journal/v437/n7057/full/nature03959.html?foxtrotcal%20lback=true (visited on 08/16/2017).

[117] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads." en. In: *EMBnet.journal* 17.1 (May 2011), pp. 10–12. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200. URL: http://journal.embnet.org/index.php/embnetjournal/article/view/200 (visited on 09/16/2015).

[118] Nathalie Mathy, Agnès Hébert, Peggy Mervelet, Lionel Bénard, Audrey Dorléans, Inés Li de la Sierra-Gallay, Philippe Noirot, Harald Putzer, and Ciarán Condon. "Bacillus subtilis ribonucleases J1 and J2 form a complex with altered enzyme behaviour." en. In: *Molecular Microbiology* 75.2 (Jan. 2010), pp. 489–498. ISSN: 1365-2958. DOI: 10.1111/j.1365-2958.2009.07004.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2009.07004.x/abstract (visited on 10/11/2017).

[119] Matthew Boitano. *Full-length cDNA sequencing of prokaryotic transcriptome and metatranscriptome samples.* Conference Poster. 2017. URL: http://www.pacb.com/wp-content/uploads/Boitano-ASM-2017-Full-Length-cDNA-Sequenc%20ing-of-Prokaryotic-Transcriptome-and-Metatranscriptome-Samples.pdf (visited on 10/26/2017).

[120] Ryan McClure, Divya Balasubramanian, Yan Sun, Maksym Bobrovskyy, Paul Sumby, Caroline A. Genco, Carin K. Vanderpool, and Brian Tjaden. "Computational analysis of bacterial RNA-Seq data." In: *Nucleic Acids Research* 41.14 (Aug. 2013), e140–e140. ISSN: 0305-1048. DOI: 10.1093/nar/gkt444. URL: https://academic.oup.com/nar/article/41/14/e140/1078655 (visited on 11/12/2017).

[121] John H. McDonald. "Repeated G–tests of goodness-of-fit." In: *Handbook of Biological Statistics*. 3rd ed. Baltimore, Maryland: Sparky House Publishing, 2014, pp. 90–93.

[122] Alfredo Mendoza-Vargas et al. "Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in E. coli." In: *PLOS ONE* 4.10 (Oct. 2009), e7526. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0007526. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0007526 (visited on 10/12/2017).

[123] Charlotte Michaux, Erik Holmqvist, Erin Vasicek, Malvika Sharan, Lars Barquist, Alexander J. Westermann, John S. Gunn, and Jörg Vogel. "RNA target profiles direct the discovery of virulence functions for the cold-shock proteins CspC and CspE." en. In: *Proceedings of the National Academy of Sciences* 114.26 (June 2017), pp. 6824–6829. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1620772114. URL: http://www.pnas.org/content/114/26/6824 (visited on 11/16/2017).

[124] Jan Mitschke et al. "An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp. PCC6803." en. In: *Proceedings of the National Academy of Sciences* 108.5 (Feb. 2011), pp. 2124–2129. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1015154108. URL: http://www.pnas.org/content/108/5/2124 (visited on 10/12/2017).

[125] Masatoshi Miyakoshi, Yanjie Chao, and Jörg Vogel. "Regulatory small RNAs from the 3′ regions of bacterial mRNAs." In: *Current Opinion in Microbiology*. Cell regulation 24.Supplement C (Apr. 2015), pp. 132–139. ISSN: 1369-5274. DOI: 10.1016/j.mib.2015.01.013. URL: http://www.sciencedirect.com/science/article/pii/S1369527415000223 (visited on 10/27/2017).

[126] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." en. In: *Nature Methods* 5.7 (July 2008), pp. 621–628. ISSN: 1548-7091. DOI: 10.1038/nmeth.1226. URL: https://www.nature.com/nmeth/journal/v5/n7/full/nmeth.1226.html (visited on 09/27/2017).

[127] Kai-Oliver Mutz, Alexandra Heilkenbrinker, Maren Lönne, Johanna-Gabriela Walter, and Frank Stahl. "Transcriptome analysis using next-generation sequencing." In: *Current Opinion in Biotechnology*. Analytical biotechnology 24.1 (Feb. 2013), pp. 22–30. ISSN: 0958-1669. DOI: 10.1016/j.copbio.2012.09.004. URL: http://www.sciencedirect.com/science/article/pii/S0958166912001310 (visited on 01/12/2015).

[128] Francesco Neri, Stefania Rapelli, Anna Krepelova, Danny Incarnato, Caterina Parlato, Giulia Basile, Mara Maldotti, Francesca Anselmi, and Salvatore Oliviero. "Intragenic DNA methylation prevents spurious transcription initiation." en. In: *Nature* 543.7643 (Feb. 2017), pp. 72–77. ISSN: 0028-0836. DOI: 10.1038/nature21373. URL: https://www.nature.com/nature/journal/v543/n7643/full/nature21373.html (visited on 11/16/2017).

[129] Cindo O. Nicholson, Matthew B. Friedersdorf, Laura S. Bisogno, and Jack D. Keene. "DO-RIP-seq to quantify RNA binding sites transcriptome-wide." In: *Methods*. Protein-RNA: Structure Function and Recognition 118-119.Supplement C (Apr. 2017), pp. 16–23. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2016.11.004. URL: http://www.sciencedirect.com/science/article/pii/S1046202316304376 (visited on 11/01/2017).

[130] Pierre Nicolas et al. "Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus subtilis." en. In: *Science* 335.6072 (Mar. 2012), pp. 1103–1106. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1206848. URL: http://science.sciencemag.org/content/335/6072/1103 (visited on 10/12/2017).

[131] Eugene Oh et al. "Selective ribosome profiling reveals the co-translational chaperone action of trigger factor in vivo." In: *Cell* 147.6 (Dec. 2011), pp. 1295–1308. ISSN: 0092-8674. DOI: 10.1016/j.cell.2011.10.044. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3277850/ (visited on 10/15/2015).

[132] Peter J. Park. "ChIP–seq: advantages and challenges of a maturing technology." en. In: *Nature Reviews Genetics* 10.10 (Oct. 2009), pp. 669–680. ISSN: 1471-0056. DOI: 10.1038/nrg2641. URL: https://www.nature.com/nrg/journal/v10/n10/full/nrg2641.html (visited on 10/19/2017).

[133] Karla D. Passalacqua, Anjana Varadarajan, Charlotte Weist, Brian D. Ondov, Benjamin Byrd, Timothy D. Read, and Nicholas H. Bergman. "Strand-Specific RNA-Seq Reveals Ordered Patterns of Sense and Antisense Transcription in Bacillus anthracis." In: *PLOS ONE* 7.8 (Aug. 2012), e43350. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0043350. URL: http://journals.

plos.org/plosone/article?id=10.1371/journal.pone.0043350 (visited on 10/12/2017).

[134] Rob Patro, Stephen M. Mount, and Carl Kingsford. "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms." en. In: *Nature Biotechnology* 32.5 (May 2014), pp. 462–464. ISSN: 1087-0156. DOI: 10.1038/nbt.2862. URL: https://www.nature.com/nbt/journal/v32/n5/full/nbt.2862.html (visited on 09/26/2017).

[135] Jason M. Peters, Rachel A. Mooney, Jeffrey A. Grass, Erik D. Jessen, Frances Tran, and Robert Landick. "Rho and NusG suppress pervasive antisense transcription in Escherichia coli." en. In: *Genes & Development* 26.23 (Dec. 2012), pp. 2621–2633. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.196741.112. URL: http://genesdev.cshlp.org/content/26/23/2621 (visited on 10/12/2017).

[136] Jason M. Peters, Rachel A. Mooney, Pei Fen Kuan, Jennifer L. Rowland, Sündüz Keleş, and Robert Landick. "Rho directs widespread termination of intragenic and stable RNA transcription." en. In: *Proceedings of the National Academy of Sciences* 106.36 (Sept. 2009), pp. 15406–15411. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0903846106. URL: http://www.pnas.org/content/106/36/15406 (visited on 10/02/2013).

[137] Carsten A. Raabe, Thean-Hock Tang, Juergen Brosius, and Timofey S. Rozhdestvensky. "Biases in small RNA deep sequencing data." In: *Nucleic Acids Research* 42.3 (Feb. 2014), pp. 1414–1426. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1021. URL: https://academic.oup.com/nar/article/42/3/1414/1052322/Biases-in-small-RNA-deep-%20sequencing-data (visited on 10/23/2017).

[138] Rahul Raghavan, Alan Sage, and Howard Ochman. "Genome-Wide Identification of Transcription Start Sites Yields a Novel Thermosensing RNA and New Cyclic AMP Receptor Protein-Regulated Genes in Escherichia coli." en. In: *Journal of Bacteriology* 193.11 (June 2011), pp. 2871–2874. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/JB.00398-11. URL: http://jb.asm.org/content/193/11/2871 (visited on 10/05/2017).

[139] Rahul Raghavan, Daniel B. Sloan, and Howard Ochman. "Antisense Transcription Is Pervasive but Rarely Conserved in Enteric Bacteria." en. In: *mBio* 3.4 (Aug. 2012), e00156–12. ISSN: , 2150-7511. DOI: 10.1128/mBio.00156-12. URL: http://mbio.asm.org/content/3/4/e00156-12 (visited on 10/12/2017).

[140] Yulia Redko, Sylvie Aubert, Anna Stachowicz, Pascal Lenormand, Abdelkader Namane, Fabien Darfeuille, Marie Thibonnier, and Hilde De Reuse. "A minimal bacterial RNase J-based degradosome is associated with translating ribosomes." In: *Nucleic Acids Research* 41.1 (Jan. 2013), pp. 288–301. ISSN: 0305-1048. DOI: 10.1093/nar/gks945. URL: https://academic.oup.com/nar/article/41/1/288/1178313/A-minimal-bacterial-RNase-%20J-based-degradosome-is (visited on 10/11/2017).

[141] Anthony Rhoads and Kin Fai Au. "PacBio Sequencing and Its Applications." In: *Genomics, Proteomics & Bioinformatics*. SI: Metagenomics of Marine Environments 13.5 (Oct. 2015), pp. 278–289. ISSN: 1672-0229. DOI: 10.1016/j.gpb.2015.08.002. URL: http://www.sciencedirect.com/science/article/pii/S1672022915001345 (visited on 10/26/2017).

[142] Jamie Richards, Quansheng Liu, Olivier Pellegrini, Helena Celesnik, Shiyi Yao, David H. Bechhofer, Ciarán Condon, and Joel G. Belasco. "An RNA Pyrophosphohydrolase Triggers 5'-Exonucleolytic Degradation of mRNA in Bacillus subtilis." In: *Molecular Cell* 43.6 (Sept. 2011), pp. 940–949. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2011.07.023. URL: http://www.sciencedirect.com/science/article/pii/S1097276511005879 (visited on 10/11/2017).

[143] Renate Rieder, Richard Reinhardt, Cynthia Sharma, and Jörg Vogel. "Experimental tools to identify RNA-protein interactions in Helicobacter pylori." In: *RNA Biology* 9.4 (Apr. 2012), pp. 520–531. ISSN: 1547-6286. DOI: 10.4161/rna.20331. URL: http://dx.doi.org/10.4161/rna.20331 (visited on 07/14/2016).

[144] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. "limma powers differential expression analyses for RNA-sequencing and microarray studies." In: *Nucleic Acids Research* 43.7 (Apr. 2015), e47–e47. ISSN: 0305-1048. DOI: 10.1093/nar/gkv007. URL: https://academic.oup.com/nar/article/43/7/e47/2414268/limma-powers-differential-%20expression-analyses-for (visited on 09/29/2017).

[145] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. "Identification of novel transcripts in annotated genomes using RNA-Seq." In: *Bioinformatics* 27.17 (Sept. 2011), pp. 2325–2329. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr355. URL: https://academic.oup.com/bioinformatics/article/27/17/2325/223194/Identification%20-of-novel-transcripts-in-annotated (visited on 09/26/2017).

[146] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." en. In: *Bioinformatics* 26.1 (Jan. 2010), pp. 139–140. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btp616. URL: http://bioinformatics.oxfordjournals.org/content/26/1/139 (visited on 11/30/2012).

[147] Tony Romeo, Christopher A. Vakulskas, and Paul Babitzke. "Post-transcriptional regulation on a global scale: form and function of Csr/Rsm systems." en. In: *Environmental Microbiology* 15.2 (Feb. 2013), pp. 313–324. ISSN: 1462-2920. DOI: 10.1111/j.1462-2920.2012.02794.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1462-2920.2012.02794.x/abstract (visited on 10/16/2017).

[148] Jonathan M. Rothberg et al. "An integrated semiconductor device enabling non-optical genome sequencing." en. In: *Nature* 475.7356 (July 2011), pp. 348–352. ISSN: 0028-0836. DOI: 10.1038/nature10242. URL: http://www.nature.com/nature/journal/v475/n7356/full/nature10242.html?foxtrotcal%20lback=true (visited on 08/29/2017).

[149] Heladia Salgado et al. "RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more." In: *Nucleic Acids Research* 41.D1 (Jan. 2013), pp. D203–D213. ISSN: 0305-1048. DOI: 10.1093/nar/gks1201. URL: https://academic.oup.com/nar/article/41/D1/D203/1068643/RegulonDB-v8-0-omics-dat%20a-sets-evolutionary (visited on 10/12/2017).

[150] F. Sanger, S. Nicklen, and A. R. Coulson. "DNA sequencing with chain-terminating inhibitors." en. In: *Proceedings of the National Academy of Sciences* 74.12 (Dec. 1977), pp. 5463–5467. ISSN: 0027-8424, 1091-6490. URL: http://www.pnas.org/content/74/12/5463 (visited on 08/04/2017).

[151] Marcel H. Schulz, Daniel R. Zerbino, Martin Vingron, and Ewan Birney. "Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels." In: *Bioinformatics* 28.8 (Apr. 2012), pp. 1086–1092. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts094. URL: https://academic.oup.com/bioinformatics/article/28/8/1086/195757/Oases-robust-de%20-novo-RNA-seq-assembly-across-the (visited on 09/24/2017).

[152] Nicholas J. Schurch et al. "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?" en. In: *RNA* 22.6 (June 2016), pp. 839–851. ISSN: 1355-8382, 1469-9001.

DOI: 10.1261/rna.053959.115. URL: http://rnajournal.cshlp.org/content/22/6/839 (visited on 07/28/2016).

[153] Nina Sesto, Omri Wurtzel, Cristel Archambaud, Rotem Sorek, and Pascale Cossart. "The excludon: a new concept in bacterial antisense RNA-mediated gene regulation." en. In: *Nature Reviews Microbiology* 11.2 (Feb. 2013), pp. 75–82. ISSN: 1740-1526. DOI: 10.1038/nrmicro2934. URL: http://www.nature.com/nrmicro/journal/v11/n2/full/nrmicro2934.html (visited on 02/05/2014).

[154] Karen Shahbabian, Ailar Jamalli, Léna Zig, and Harald Putzer. "RNase Y, a novel endoribonuclease, initiates riboswitch turnover in Bacillus subtilis." en. In: *The EMBO Journal* 28.22 (Nov. 2009), pp. 3523–3533. ISSN: 0261-4189, 1460-2075. DOI: 10.1038/emboj.2009.283. URL: http://emboj.embopress.org/content/28/22/3523 (visited on 10/11/2017).

[155] Cynthia M Sharma and Jörg Vogel. "Differential RNA-seq: the approach behind and the biological insight gained." In: *Current Opinion in Microbiology*. Ecology and industrial microbiology • Special Section: Novel technologies in microbiology 19 (June 2014), pp. 97–105. ISSN: 1369-5274. DOI: 10.1016/j.mib.2014.06.010. URL: http://www.sciencedirect.com/science/article/pii/S1369527414000800 (visited on 12/11/2014).

[156] Cynthia Sharma et al. "The primary transcriptome of the major human pathogen Helicobacter pylori." In: *Nature* 464.7286 (2010), pp. 250–255. ISSN: 0028-0836. URL: http://dx.doi.org/10.1038/nature08756 (visited on 02/15/2012).

[157] Atsuko Shinhara, Motomu Matsui, Kiriko Hiraoka, Wataru Nomura, Reiko Hirano, Kenji Nakahigashi, Masaru Tomita, Hirotada Mori, and Akio Kanai. "Deep sequencing reveals as-yet-undiscovered small RNAs in Escherichia coli." In: *BMC Genomics* 12 (Aug. 2011), p. 428. ISSN: 1471-2164. DOI: 10.1186/1471-2164-12-428. URL: https://doi.org/10.1186/1471-2164-12-428 (visited on 10/12/2017).

[158] Navjot Singh and Joseph T. Wade. "Identification of Regulatory RNA in Bacterial Genomes by Genome-Scale Mapping of Transcription Start Sites." en. In: *Therapeutic Applications of Ribozymes and Riboswitches*. Methods in Molecular Biology. DOI: 10.1007/978-1-62703-730-3_1. Humana Press, Totowa, NJ, 2014, pp. 1–10. ISBN: 978-1-62703-729-7 978-1-62703-730-3. URL: https://link.springer.com/protocol/10.1007/978-1-62703-730-3_1 (visited on 10/05/2017).

[159] Alexandra Sittka, Sacha Lucchini, Kai Papenfort, Cynthia M. Sharma, Katarzyna Rolle, Tim T. Binnewies, Jay C. D. Hinton, and Jörg Vogel. "Deep Sequencing Analysis of Small Noncoding RNA and mRNA Targets of the Global Post-Transcriptional Regulator, Hfq." In: *PLOS Genetics* 4.8 (Aug. 2008), e1000163. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000163. URL: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000163 (visited on 10/12/2017).

[160] Alexandre Smirnov, Konrad U. Förstner, Erik Holmqvist, Andreas Otto, Regina Günster, Dörte Becher, Richard Reinhardt, and Jörg Vogel. "Grad-seq guides the discovery of ProQ as a major small RNA-binding protein." en. In: *Proceedings of the National Academy of Sciences* 113.41 (Oct. 2016), pp. 11591–11596. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1609981113. URL: http://www.pnas.org/content/113/41/11591 (visited on 11/16/2017).

[161] Andrew M. Smith, Miten Jain, Logan Mulroney, Daniel R. Garalde, and Mark Akeson. "Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing." en. In: *bioRxiv* (Apr. 2017), p. 132274. DOI: 10.1101/132274. URL: https://www.biorxiv.org/content/early/2017/04/29/132274 (visited on 11/16/2017).

[162] Rotem Sorek and Pascale Cossart. "Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity." en. In: *Nature Reviews Genetics* 11.1 (Jan. 2010), pp. 9–16. ISSN: 1471-0056. DOI: 10.1038/nrg2695. URL: https://www.nature.com/nrg/journal/v11/n1/full/nrg2695.html (visited on 09/25/2017).

[163] Catherine Spickler, Victoria Stronge, and George A. Mackie. "Preferential Cleavage of Degradative Intermediates of rpsT mRNA by the Escherichia coli RNA Degradosome." en. In: *Journal of Bacteriology* 183.3 (Feb. 2001), pp. 1106–1109. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/JB.183.3.1106-1109.2001. URL: http://jb.asm.org/content/183/3/1106 (visited on 10/11/2017).

[164] D. Stazic and B. Voß. "The complexity of bacterial transcriptomes." In: *Journal of Biotechnology*. Bioinformatics for Biotechnology and Biomedicine 232.Supplement C (Aug. 2016), pp. 69–78. ISSN: 0168-1656. DOI: 10.1016/j.jbiotec.2015.09.041. URL: http://www.sciencedirect.com/science/article/pii/S0168165615301474 (visited on 12/13/2017).

[165] Sebastian Suerbaum and Pierre Michetti. "Helicobacter pylori Infection." In: *New England Journal of Medicine* 347.15 (Oct. 2002), pp. 1175–1186. ISSN:

0028-4793. DOI: 10.1056/NEJMra020542. URL: http://dx.doi.org/10.1056/NEJMra020542 (visited on 10/09/2017).

[166] Yoichiro Sugimoto, Julian König, Shobbir Hussain, Blaž Zupan, Tomaž Curk, Michaela Frye, and Jernej Ule. "Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions." In: *Genome Biology* 13 (Aug. 2012), R67. ISSN: 1474-760X. DOI: 10.1186/gb-2012-13-8-r67. URL: https://doi.org/10.1186/gb-2012-13-8-r67 (visited on 11/01/2017).

[167] Peter A. C. 't Hoen et al. "Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories." en. In: *Nature Biotechnology* 31.11 (Nov. 2013), pp. 1015–1022. ISSN: 1087-0156. DOI: 10.1038/nbt.2702. URL: https://www.nature.com/nbt/journal/v31/n11/full/nbt.2702.html (visited on 10/23/2017).

[168] Sonia Tarazona, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. "Differential expression in RNA-seq: A matter of depth." en. In: *Genome Research* 21.12 (Dec. 2011), pp. 2213–2223. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.124321.111. URL: http://genome.cshlp.org/content/21/12/2213 (visited on 09/29/2017).

[169] Olivier Tenaillon, David Skurnik, Bertrand Picard, and Erick Denamur. "The population genetics of commensal Escherichia coli." en. In: *Nature Reviews Microbiology* 8.3 (Mar. 2010), pp. 207–217. ISSN: 1740-1526. DOI: 10.1038/nrmicro2298. URL: http://www.nature.com/nrmicro/journal/v8/n3/full/nrmicro2298.html?foxtrotcallbac%20k=true#B6 (visited on 10/05/2017).

[170] Maureen K. Thomason, Thorsten Bischler, Sara K. Eisenbart, Konrad U. Förstner, Aixia Zhang, Alexander Herbig, Kay Nieselt, Cynthia M. Sharma, and Gisela Storz. "Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in Escherichia coli." en. In: *Journal of Bacteriology* 197.1 (Jan. 2015), pp. 18–28. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/JB.02096-14. URL: http://jb.asm.org/content/197/1/18 (visited on 01/07/2015).

[171] Maureen Kiley Thomason and Gisela Storz. "Bacterial Antisense RNAs: How Many Are There, and What Are They Doing?" In: *Annual Review of Genetics* 44.1 (2010), pp. 167–188. DOI: 10.1146/annurev-genet-102209-163523. URL: https://doi.org/10.1146/annurev-genet-102209-163523 (visited on 10/11/2017).

[172] J. A. Thompson, M. F. Radonovich, and N. P. Salzman. "Characterization of the 5′-terminal structure of simian virus 40 early mRNA's." en. In: *Journal of Virology* 31.2 (Aug. 1979), pp. 437–446. ISSN: 0022-538X, 1098-5514. URL: http://jvi.asm.org/content/31/2/437 (visited on 10/05/2017).

[173] Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration." In: *Briefings in Bioinformatics* 14.2 (Mar. 2013), pp. 178–192. ISSN: 1467-5463. DOI: 10.1093/bib/bbs017. URL: https://academic.oup.com/bib/article/14/2/178/208453/Integrative-Genomics-Viewer%20-IGV-high-performance (visited on 09/26/2017).

[174] Jean-F. Tomb et al. "The complete genome sequence of the gastric pathogen Helicobacter pylori." en. In: *Nature* 388.6642 (Aug. 1997), pp. 539–547. ISSN: 0028-0836. DOI: 10.1038/41483. URL: https://www.nature.com/nature/journal/v388/n6642/full/388539a0.html (visited on 10/11/2017).

[175] Cole Trapnell, David G. Hendrickson, Martin Sauvageau, Loyal Goff, John L. Rinn, and Lior Pachter. "Differential analysis of gene regulation at transcript resolution with RNA-seq." en. In: *Nature Biotechnology* 31.1 (Jan. 2013), pp. 46–53. ISSN: 1087-0156. DOI: 10.1038/nbt.2450. URL: https://www.nature.com/nbt/journal/v31/n1/full/nbt.2450.html (visited on 09/29/2017).

[176] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. "TopHat: discovering splice junctions with RNA-Seq." In: *Bioinformatics* 25.9 (May 2009), pp. 1105–1111. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp120. URL: https://academic.oup.com/bioinformatics/article/25/9/1105/203994/TopHat-discover%20ing-splice-junctions-with-RNA-Seq (visited on 09/26/2017).

[177] Jai J. Tree, Sander Granneman, Sean P. McAteer, David Tollervey, and David L. Gally. "Identification of Bacteriophage-Encoded Anti-sRNAs in Pathogenic Escherichia coli." In: *Molecular Cell* 55.2 (July 2014), pp. 199–213. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2014.05.006. URL: https://www.sciencedirect.com/science/article/pii/S1097276514004006 (visited on 11/01/2017).

[178] Michael Uhl, Torsten Houwaart, Gianluca Corrado, Patrick R. Wright, and Rolf Backofen. "Computational analysis of CLIP-seq data." In: *Methods*. Protein-RNA: Structure Function and Recognition 118.Supplement C (Apr. 2017), pp. 60–72. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2017.02.

006. URL: http://www.sciencedirect.com/science/article/pii/S1046202317300828 (visited on 10/15/2017).

[179] Philip J. Uren, Emad Bahrami-Samani, Suzanne C. Burns, Mei Qiao, Fedor V. Karginov, Emily Hodges, Gregory J. Hannon, Jeremy R. Sanford, Luiz O. F. Penalva, and Andrew D. Smith. "Site identification in high-throughput RNA–protein interaction data." en. In: *Bioinformatics* 28.23 (Dec. 2012), pp. 3013–3020. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bts569. URL: http://bioinformatics.oxfordjournals.org/content/28/23/3013 (visited on 08/25/2015).

[180] Anton Valouev et al. "A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning." en. In: *Genome Research* 18.7 (July 2008), pp. 1051–1063. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.076463.108. URL: http://genome.cshlp.org/content/18/7/1051 (visited on 08/28/2017).

[181] Eric L. Van Nostrand et al. "Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)." en. In: *Nature Methods* 13.6 (June 2016), pp. 508–514. ISSN: 1548-7091. DOI: 10.1038/nmeth.3810. URL: http://www.nature.com/nmeth/journal/v13/n6/full/nmeth.3810.html (visited on 06/14/2016).

[182] Van Vliet and Arnoud H.m. "Next generation sequencing of microbial transcriptomes: challenges and opportunities." In: *FEMS Microbiology Letters* 302.1 (Jan. 2010), pp. 1–7. ISSN: 0378-1097. DOI: 10.1111/j.1574-6968.2009.01767.x. URL: https://academic.oup.com/femsle/article/302/1/1/522192/Next-generation-sequencin%20g-of-microbial (visited on 10/06/2017).

[183] Jörg Vogel. "A rough guide to the non-coding RNA world of Salmonella." en. In: *Molecular Microbiology* 71.1 (Jan. 2009), pp. 1–11. ISSN: 1365-2958. DOI: 10.1111/j.1365-2958.2008.06505.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2008.06505.x/abstract (visited on 10/18/2017).

[184] Jörg Vogel, Verena Bartels, Thean Hock Tang, Gennady Churakov, Jacoba G. Slagter-Jäger, Alexander Hüttenhofer, and E. Gerhart H. Wagner. "RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria." In: *Nucleic Acids Research* 31.22 (Nov. 2003), pp. 6435–6443. ISSN: 0305-1048. DOI: 10.1093/nar/gkg867. URL: https://academic.oup.com/nar/article/31/22/6435/2375976/RNomics-in-Escherichia-c%20oli-detects-new-sRNA (visited on 10/05/2017).

[185] Jörg Vogel and Ben F. Luisi. "Hfq and its constellation of RNA." en. In: *Nature Reviews Microbiology* 9.8 (Aug. 2011), pp. 578–589. ISSN: 1740-1526. DOI: 10.1038/nrmicro2615. URL: http://www.nature.com/nrmicro/journal/v9/n8/full/nrmicro2615.html (visited on 06/03/2012).

[186] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." en. In: *Nature Reviews Genetics* 10.1 (Jan. 2009), pp. 57–63. ISSN: 1471-0056. DOI: 10.1038/nrg2484. URL: http://www.nature.com/nrg/journal/v10/n1/full/nrg2484.html (visited on 07/10/2012).

[187] Alexander J. Westermann, Konrad U. Förstner, Fabian Amman, Lars Barquist, Yanjie Chao, Leon N. Schulte, Lydia Müller, Richard Reinhardt, Peter F. Stadler, and Jörg Vogel. "Dual RNA-seq unveils noncoding RNA functions in host–pathogen interactions." en. In: *Nature* 529.7587 (Jan. 2016), pp. 496–501. ISSN: 0028-0836. DOI: 10.1038/nature16547. URL: https://www.nature.com/nature/journal/v529/n7587/full/nature16547.html (visited on 10/18/2017).

[188] Emily C. Wheeler, Van Nostrand, Eric L, and Gene W. Yeo. "Advances and challenges in the detection of transcriptome-wide protein–RNA interactions." en. In: *Wiley Interdisciplinary Reviews: RNA* (Aug. 2017). ISSN: 1757-7012. DOI: 10.1002/wrna.1436. URL: http://onlinelibrary.wiley.com/doi/10.1002/wrna.1436/full (visited on 10/17/2017).

[189] Sandra Wiegand, Sascha Dietrich, Robert Hertel, Johannes Bongaerts, Stefan Evers, Sonja Volland, Rolf Daniel, and Heiko Liesegang. "RNA-Seq of Bacillus licheniformis: active regulatory RNA features expressed within a productive fermentation." In: *BMC Genomics* 14 (Oct. 2013), p. 667. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-667. URL: https://doi.org/10.1186/1471-2164-14-667 (visited on 10/12/2017).

[190] Omri Wurtzel, Rajat Sapra, Feng Chen, Yiwen Zhu, Blake A. Simmons, and Rotem Sorek. "A single-base resolution map of an archaeal transcriptome." en. In: *Genome Research* 20.1 (Jan. 2010), pp. 133–141. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.100396.109. URL: http://genome.cshlp.org/content/20/1/133 (visited on 10/05/2017).

[191] Omri Wurtzel, Nina Sesto, J. R. Mellin, Iris Karunker, Sarit Edelheit, Christophe Bécavin, Cristel Archambaud, Pascale Cossart, and Rotem Sorek. "Comparative transcriptomics of pathogenic and non-pathogenic Listeria species." en. In: *Molecular Systems Biology* 8.1 (Jan. 2012), p. 583. ISSN: 1744-4292, 1744-4292. DOI: 10.1038/msb.2012.11. URL: http://msb.embopress.org/content/8/1/583 (visited on 10/12/2017).

[192] Yinlong Xie et al. "SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads." In: *Bioinformatics* 30.12 (June 2014), pp. 1660–1666. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu077. URL: https://academic.oup.com/bioinformatics/article/30/12/1660/380938/SOAPdenovo-Tra%20ns-de-novo-transcriptome-assembly (visited on 09/24/2017).

[193] Haibin Xu, Xiang Luo, Jun Qian, Xiaohui Pang, Jingyuan Song, Guangrui Qian, Jinhui Chen, and Shilin Chen. "FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads." In: *PLOS ONE* 7.12 (Dec. 2012), e52249. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0052249. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0052249 (visited on 10/18/2017).

[194] Zizhen Yao, Zasha Weinberg, and Walter L. Ruzzo. "CMfinder—a covariance model based RNA motif finding algorithm." In: *Bioinformatics* 22.4 (Feb. 2006), pp. 445–452. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btk008. URL: https://academic.oup.com/bioinformatics/article/22/4/445/184450/CMfinder-a-covar%20iance-model-based-RNA-motif (visited on 10/13/2017).

[195] Kathryn T. Young, Lindsay M. Davis, and Victor J. DiRita. "Campylobacter jejuni: molecular biology and pathogenesis." en. In: *Nature Reviews Microbiology* 5.9 (Sept. 2007), pp. 665–679. ISSN: 1740-1526. DOI: 10.1038/nrmicro1718. URL: https://www.nature.com/nrmicro/journal/v5/n9/full/nrmicro1718.html (visited on 10/13/2017).

[196] Sung-Huan Yu, Jörg Vogel, and Konrad Ulrich Förstner. "ANNOgesic: A Pipeline To Translate Bacterial/Archaeal RNA-Seq Data Into High-Resolution Genome Annotations." en. In: *bioRxiv* (May 2017), p. 143081. DOI: 10.1101/143081. URL: https://www.biorxiv.org/content/early/2017/05/29/143081 (visited on 10/25/2017).

[197] Brian J. Zarnegar, Ryan A. Flynn, Ying Shen, Brian T. Do, Howard Y. Chang, and Paul A. Khavari. "irCLIP platform for efficient characterization of protein-RNA interactions." en. In: *Nature Methods* 13.6 (June 2016), pp. 489–492. ISSN: 1548-7091. DOI: 10.1038/nmeth.3840. URL: http://www.nature.com/nmeth/journal/v13/n6/full/nmeth.3840.html (visited on 10/17/2017).

[198] Chaolin Zhang and Robert B. Darnell. "Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data." en. In: *Nature Biotechnology* 29.7 (July 2011), pp. 607–614. ISSN: 1087-0156. DOI: 10.1038/

nbt.1873. URL: https://www.nature.com/nbt/journal/v29/n7/full/nbt.1873.html (visited on 10/13/2017).

a

APPENDIX

A.1 STATEMENT OF INDIVIDUAL AUTHOR CONTRIBUTIONS AND OF LEGAL SECOND PUBLICATION RIGHTS

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Graduate School
Life Sciences

"Dissertation Based on Several Published Manuscripts"

Statement of individual author contributions and of legal second publication rights

(If required please use more than one sheet)

Publication (complete reference):

Thomason, M.K.*, **Bischler, T.***, Eisenbart, S.K., Förstner, K.U., Zhang, A., Herbig, A., Nieselt, K., Sharma, C.M., Storz, G., 2015. Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in Escherichia coli. J. Bacteriol. 197, 18–28. doi:10.1128/JB.02096-14

*joined first authors

Participated in	**Author Initials,** Responsibility decreasing from left to right						
Study Design	GS	CMS	KN	MKT	TB	KUF	AH
Methods Development	MKT	TB	AH	KUF	SKE	AZ	
Data Collection	MKT	TB	SKE	AZ			
Data Analysis and Interpretation	TB	MKT	GS	CMS	KUF	SKE	
Manuscript Writing	GS	MKT	CMS	TB			
Writing of Introduction	GS	MKT	CMS	TB			
Writing of Materials & Methods	TB	MKT	GS	CMS	SKE		
Writing of Results	GS	CMS/TB/MKT					
Writing of Discussion	GS	MKT	CMS	TB			
Writing of First Draft	MKT	TB					

Explanations (if applicable):

No other shared first author will use this publication for submission of a PhD thesis based on several published manuscripts.

Publication (complete reference):

Bischler, T., Tan, H.S., Nieselt, K., Sharma, C.M., 2015. Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in Helicobacter pylori. Methods, Bacterial and Archaeal Transcription 86, 89–101. doi:10.1016/j.ymeth.2015.06.012

Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design	CMS	TB	HST	KN	
Methods Development	TB	CMS	KN		
Data Collection	TB	HST			
Data Analysis and Interpretation	TB	CMS	HST		
Manuscript Writing	TB/CMS	HST			
Writing of Introduction	TB	CMS			
Writing of Materials & Methods	TB	CMS	HST		
Writing of Results	TB	CMS			
Writing of Discussion	TB	CMS			
Writing of First Draft	TB				

Explanations (if applicable):

Publication (complete reference):

Bischler, T.*, Hsieh, P.-K.*, Resch, M.*, Liu, Q., Tan, H.S., Foley, P.L., Hartleib, A., Sharma, C.M., Belasco, J.G., 2016. Identification of the RNA Pyrophosphohydrolase RppH of Helicobacter pylori and Global Analysis of Its RNA Targets. J. Biol. Chem. jbc.M116.761171. doi:10.1074/jbc.M116.761171

*joined first authors

Participated in	**Author Initials,** Responsibility decreasing from left to right						
Study Design	JGB	CMS					
Methods Development	TB	PLF	QL				
Data Collection	PKH	HST	TB	MR	QL	PLF	AH
Data Analysis and Interpretation	JGB	TB	CMS	PKH	MR	QL	PLF
Manuscript Writing	JGB	CMS	MR	TB	PKH	HST	
Writing of Introduction	JGB	CMS	MR	TB			
Writing of Materials & Methods	MR	TB	JGB	HST			
Writing of Results	JGB	CMS	TB	MR			
Writing of Discussion	JGB	CMS	TB	MR	PKH		
Writing of First Draft	MR	TB					

Explanations (if applicable):

No other shared first author will use this publication for submission of a PhD thesis based on several published manuscripts.

Thorsten Bischler's main contribution to this publication consists of the analysis and interpretation of the high-throughput sequencing data as well as the selection of targets for experimental validation.

Publication (complete reference):

Holmqvist, E., Wright, P.R., Li, L., **Bischler, T.**, Barquist, L., Reinhardt, R., Backofen, R., Vogel, J., 2016. Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. The EMBO Journal 35, 991–1011. doi:10.15252/embj.201593360

Participated in	**Author Initials,** Responsibility decreasing from left to right					
Study Design	EH	JV				
Methods Development	EH	PRW/TB/LL	LB	RB		
Data Collection	EH					
Data Analysis and Interpretation	EH	PRW/TB/LL	LB			
Manuscript Writing	EH	JV	RB	PRW	TB	LL
Writing of Introduction	EH	JV	RB			
Writing of Materials & Methods	EH	PRW	TB	LL	LB	
Writing of Results	EH	JV	RB			
Writing of Discussion	EH	JV	RB			
Writing of First Draft	EH					

Explanations (if applicable):

Publication (complete reference):

Dugar, G., Svensson, S.L., **Bischler, T.**, Wäldchen, S., Reinhardt, R., Sauer, M., Sharma, C.M., 2016. The CsrA-FliW network controls polar localization of the dual-function flagellin mRNA in Campylobacter jejuni. Nat Commun 7, 11667. doi:10.1038/ncomms11667

Participated in	**Author Initials,** Responsibility decreasing from left to right					
Study Design	CMS	GD	SLS	MS		
Methods Development	GD/TB	SW				
Data Collection	GD	SLS/TB	SW	RR		
Data Analysis and Interpretation	GD	CMS	SLS/TB	SW	MS	
Manuscript Writing	CMS	GD	SLS			
Writing of Introduction	CMS	GD				
Writing of Materials & Methods	GD	SLS	TB	SW	CMS	
Writing of Results	CMS	GD	SLS			
Writing of Discussion	CMS	GD	SLS			
Writing of First Draft	GD	CMS	SLS			

Explanations (if applicable):

The doctoral researcher confirms that she/he has obtained permission from both the publishers and the co-authors for legal second publication.

The doctoral researcher and the primary supervisor confirm the correctness of the above mentioned assessment.

Thorsten Bischler

Doctoral Researcher's Name | Date | Place | Signature

Cynthia Sharma

Primary Supervisor's Name | Date | Place | Signature

"Dissertation Based on Several Published Manuscripts"

Statement of individual author contributions to figures/tables/chapters included in the manuscripts

(If required please use more than one sheet)

Publication (complete reference):

Thomason, M.K.*, **Bischler, T.***, Eisenbart, S.K., Förstner, K.U., Zhang, A., Herbig, A., Nieselt, K., Sharma, C.M., Storz, G., 2015. Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in Escherichia coli. J. Bacteriol. 197, 18–28. doi:10.1128/JB.02096-14

*joined first authors

Figure	**Author Initials,** Responsibility decreasing from left to right				
1	TB				
2	TB	KUF			
3	TB				
4	TB	MKT			
5	TB/MKT/AZ/SKE				
S1	TB				
S2	(TB)				
S3	TB				
S4	TB				
S5	TB				
S6	KUF/TB				
S7	KUF				
S8	TB	MKT			
S9	TB/MKT/AZ/SKE				
S10	AZ				
Table	**Author Initials, Responsibility decreasing from left to right**				
S1	MKT/SKE/AZ				
S2	MKT/SKE/AZ				
S3	TB				
S4	TB				
S5	TB				
S6	MKT				
S7	MKT	TB			
S8	MKT	SKE			
Data Sets	**Author Initials, Responsibility decreasing from left to right**				
S1	TB				
S2	TB				
S3	TB				

Explanations (if applicable):

Figure S2 was adapted from Dugar *et al.*, PLoS Genet, 2013.

Publication (complete reference):

Bischler, T., Tan, H.S., Nieselt, K., Sharma, C.M., 2015. Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in Helicobacter pylori. Methods, Bacterial and Archaeal Transcription 86, 89–101. doi:10.1016/j.ymeth.2015.06.012

Figure	**Author Initials,** Responsibility decreasing from left to right				
1	TB	HST			
2	TB				
3	TB				
4	TB				
5	TB				
6	TB				
S1	TB				
Table	**Author Initials,** Responsibility decreasing from left to right				
S1	TB				
S2	TB				

Explanations (if applicable):

Publication (complete reference):

Bischler, T.*, Hsieh, P.-K.*, Resch, M.*, Liu, Q., Tan, H.S., Foley, P.L., Hartleib, A., Sharma, C.M., Belasco, J.G., 2016. Identification of the RNA Pyrophosphohydrolase RppH of Helicobacter pylori and Global Analysis of Its RNA Targets. J. Biol. Chem. jbc.M116.761171. doi:10.1074/jbc.M116.761171

*joined first authors

Figure	**Author Initials,** Responsibility decreasing from left to right				
1	JGB				
2	QL				
3	JGB				
4	PKH				
5	PKH				
6	PLF				
7	PLF				
8	TB				
9	HST	TB	MR		
10	HST	TB	MR		
Table	**Author Initials,** Responsibility decreasing from left to right				
1	TB				
2	HST	MR	AH		
S1	PKH				
S2	TB				
S3	TB				

Explanations (if applicable):

The results of the high-throughput sequencing-based RppH target analysis, which was Thorsten Bischler's main contribution to the publication, are shown in Supplementary Tables S2 and S3.

Publication (complete reference):

Holmqvist, E., Wright, P.R., Li, L., **Bischler, T.**, Barquist, L., Reinhardt, R., Backofen, R., Vogel, J., 2016. Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. The EMBO Journal 35, 991–1011. doi:10.15252/embj.201593360

Figure	**Author Initials,** Responsibility decreasing from left to right				
1	EH	PRW	TB		
2	EH	LL	TB		
3	EH	LL	TB		
4	EH	PRW			
5	EH	TB			
6	EH	LL	TB		
7	EH	LL	TB		
S1	EH	TB			
S2	EH				
S3	EH				
S4	EH				
S5	LB				
EV1	TB	EH			
EV2	EH	PRW			
EV3	EH				
EV4	EH				
EV5	EH				
Table	**Author Initials,** Responsibility decreasing from left to right				
S1	EH				
S2	EH				
S3	EH				
EV1	PRW	EH	LB		
EV2	LL	EH	LB		
EV3	PRW	EH			
EV4	PRW	EH	LB		
EV5	LL	EH	LB		

Explanations (if applicable):

Publication (complete reference):

Dugar, G., Svensson, S.L., **Bischler, T.**, Wäldchen, S., Reinhardt, R., Sauer, M., Sharma, C.M., 2016. The CsrA-FliW network controls polar localization of the dual-function flagellin mRNA in *Campylobacter jejuni*. Nat Commun 7, 11667. doi:10.1038/ncomms11667

Figure	**Author Initials,** Responsibility decreasing from left to right				
1	GD	TB			
2	GD				
3	GD	SLS			
4	GD	SLS			
5	GD				

6	GD				
7	GD	SW			
8	GD				
9	GD				
S1	SLS				
S2	GD/TB				
S3	GD/TB				
S4	GD				
S5	GD				
S6	TB				
S7	TB				
S8	GD				
S9	GD				
S10	GD				
S11	GD				
S12	SLS				
S13	GD	SLS			
S14	GD				
S15	GD				
S16	GD				
S17	GD	SW			
S18	GD				
S19	GD				
Table	**Author Initials,** Responsibility decreasing from left to right				
1	TB	GD			
S1	TB				
S2	TB				
S3	GD	SLS			
S4	GD				
S5	GD				
S6	GD				
S7	GD				
S8	GD				
Data	**Author Initials,** Responsibility decreasing from left to right				
1	TB				
2	TB				

Explanations (if applicable):

I also confirm my primary supervisor's acceptance.

Thorsten Bischler

Doctoral Researcher's Name Date Place Signature

A.2 CURRICULUM VITAE

Thorsten Bischler

Curriculum vitae

Personal Information

Date of Birth	May, 7, 1983
Place of Birth	Oberndorf a.N.
Nationality	German
Marital status	single

Academic qualifications

since 02/2012 **Doctoral Study Program "Life Sciences" - Graduate School of Life Sciences (final degree Dr. rer. nat.)**, *University of Würzburg.*

10/2009 – 01/2012 **Master of Science in Bioinformatics and Systems Biology**, *Albert Ludwig University of Freiburg, (1.6).*

09/2002 – 04/2008 **Diplom(FH) degree in Bioinformatics**, *FH Bingen, University of Applied Sciences, (2.3).*

08/1993 – 07/2002 **Allgemeine Hochschulreife (general higher education entrance qualification)**, *Gymnasium am Rosenberg*, Oberndorf a.N., *(1.4).*
Advanced courses: Mathematics, Biology

Research/Work experience

since 02/2012 **Doctoral thesis**, *Research Center for Infectious Diseases / Institute of Molecular Infection Biology, University of Würzburg.*
Title: *Data mining and software development for RNA-seq-based approaches in bacteria*

01/2011 – 01/2012 **Major practical course / Master thesis**, *Genetics & Experimental Bioinformatics, Institute of Biology III, Albert Ludwig University of Freiburg.*
Title: *Algorithms and applications for comparative RNA studies*

04/2007 – 03/2008 **Diploma thesis**, *Experimental Ophthalmology, Dept. of Ophthalmology, Johannes Gutenberg University of Mainz.*
Title: *Development of a software and method for discovering disease correlated post-translational modifications*

10/2006 – 02/2007 **Internship**, *Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), Johannes Gutenberg University of Mainz.*
Title: *Identification of multivariate dependencies by means of Copula models*

St.-Rochus-Str. 43 – 97078 Würzburg, Germany
0931/90796803 • *thorsten.bischler@uni-wuerzburg.de*

Positions and Memberships

04/2014 – 03/2015 **Students' Speaker of the Graduate Programme "Infectious Diseases Research"**, *Graduate School of Life Sciences*, University of Würzburg.

06/2013 – 06/2014 **ISCB member**, *International Society for Computational Biology*.

10/2012 – 03/2014 **Substitute Students' Speaker of the Graduate Programme "Infectious Diseases Research"**, *Graduate School of Life Sciences*, University of Würzburg.

Publications

Thorsten Bischler, Ping-kun Hsieh, Marcus Resch, Quansheng Liu, Hock Siew Tan, Patricia L. Foley, Anika Hartleib, Cynthia M. Sharma, and Joel G. Belasco. Identification of the RNA Pyrophosphohydrolase RppH of Helicobacter pylori and Global Analysis of Its RNA Targets. *Journal of Biological Chemistry*, 292(5):1934–1950, February 2017.

Gaurav Dugar, Sarah L. Svensson, Thorsten Bischler, Sina Wäldchen, Richard Reinhardt, Markus Sauer, and Cynthia M. Sharma. The CsrA-FliW network controls polar localization of the dual-function flagellin mRNA in Campylobacter jejuni. *Nature Communications*, 7:11667, May 2016.

Erik Holmqvist, Patrick R. Wright, Lei Li, Thorsten Bischler, Lars Barquist, Richard Reinhardt, Rolf Backofen, and Jörg 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, May 2016.

Thorsten Bischler, Hock Siew Tan, Kay Nieselt, and Cynthia M. Sharma. Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in Helicobacter pylori. *Methods*, 86:89–101, September 2015.

Maureen K. Thomason, Thorsten Bischler, Sara K. Eisenbart, Konrad U. Förstner, Aixia Zhang, Alexander Herbig, Kay Nieselt, Cynthia M. Sharma, and Gisela Storz. Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in Escherichia coli. *Journal of Bacteriology*, 197(1):18–28, January 2015.

Thorsten Bischler, Matthias Kopf, and Björn Voß. Transcript mapping based on dRNA-seq data. *BMC Bioinformatics*, 15(1):122, April 2014.

Thorsten Bischler

St.-Rochus-Str. 43 – 97078 Würzburg, Germany
0931/90796803 • *thorsten.bischler@uni-wuerzburg.de*

ACKNOWLEDGMENTS

At the end of this thesis, I want to thank all people who have accompanied and helped me during my time as a doctoral researcher.

First of all, I want to thank my primary supervisor Prof. Dr. Cynthia Sharma for giving me the opportunity to write this thesis and for her support, fruitful discussions and encouragement during the whole time I spent in her group.

Next, I want to thank all members of my thesis committee, which alongside Prof. Sharma consisted of Prof. Dr. Thomas Dandekar, Prof. Dr. Jörg Vogel and Jun.-Prof. Dr. Björn Voß, for helpful scientific discussions during our annual meetings but also other occasions.

I want to especially thank Dr. Konrad Förstner, who acted as an additional supervisor and helped me a lot with all kinds of bioinformatical but also other scientific and non-scientific problems.

Special thanks also go to the co-authors of my publications including people from our institute but also external collaboration partners. Without their work finishing this thesis would not have been possible.

I want to thank Freya, Konrad, Kristina, Sarah and Björn for providing helpful comments on my thesis.

Furthermore, let me thank all members of the Sharma, Vogel and Bioinformatics groups and also all other people at the institute, who have always been a great support and provided a pleasureable atmosphere during but also outside work.

In addition, I want to thank Hilde Merkert and Josef Heger for technical assistance and Mrs. Meece and the other secretaries for their help with all kinds of administrative problems.

I also want to thank Dr. Gabriele Blum-Oehler and the GSLS team for their support during my time as a doctoral researcher.

Last but not least, I want to thank my parents, family and friends for their love and continuous support not only during my thesis but my whole life.