IT’S THE LITTLE THINGS: AN EXPLORATION OF SMALL RNAs AND SELFISH GENETIC ELEMENTS OF THE HUMAN BACTERIAL PATHOGENS COXIELLA BURNETII AND BARTONELLA BACILLIFORMIS

Shaun Gregory Wachter

Follow this and additional works at: https://scholarworks.umt.edu/etd

Let us know how access to this document benefits you.

Recommended Citation


This Dissertation is brought to you for free and open access by the Graduate School at ScholarWorks at University of Montana. It has been accepted for inclusion in Graduate Student Theses, Dissertations, & Professional Papers by an authorized administrator of ScholarWorks at University of Montana. For more information, please contact scholarworks@mso.umt.edu.
IT’S THE LITTLE THINGS: AN EXPLORATION OF SMALL RNAs AND SELFISH GENETIC ELEMENTS OF THE HUMAN BACTERIAL PATHOGENS \textit{COXIELLA BURNETII AND BARTONELLA BACILLIFORMIS}

By

SHAUN GREGORY WACHTER

BS, University of Illinois at Urbana-Champaign, Champaign, IL, 2010
MS, Northern Illinois University, Dekalb, IL, 2013

Dissertation

presented in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in Cellular, Molecular, and Microbial Biology, Microbiology

The University of Montana
Missoula, MT

August 2020

Approved by:

Scott Whittenburg, Graduate School Dean

Michael Minnick, Chair
Division of Biological Sciences

Scott Samuels
Division of Biological Sciences

Steve Lodmell
Division of Biological Sciences

Dan Drecktrah
Division of Biological Sciences

Kent Sugden
Department of Chemistry and Biochemistry
It’s the little things: An exploration of small RNAs and selfish genetic elements of the human bacterial pathogens *Coxiella burnetii* and *Bartonella bacilliformis*

Chairperson: Dr. Michael Minnick

*Coxiella burnetii* is a Gram-negative gammaproteobacterium and zoonotic agent of Q fever in humans. Previous work in our lab has demonstrated that *C. burnetii* codes for several small RNAs (sRNAs) that are differentially expressed between *in vivo* and *in vitro* growth conditions. sRNAs serve as post-transcriptional regulatory effectors involved in the control of nearly all biological processes. We demonstrated that several of the identified sRNAs, namely *Coxiella burnetii* small RNA 3 (CbsR3), Cbsr13, and CbsR16, represent members of two novel families of miniature inverted-repeat transposable elements (MITEs), termed QMITE1 and QMITE2. Furthermore, we have characterized a highly expressed, infection-specific sRNA, CbsR12, and have determined that it is necessary for expansion of the *C. burnetii* intracellular niche in a human monocyte-derived alveolar macrophage cell line. We have determined that CbsR12 may participate in broad gene regulation by acting as an “RNA sponge” for the global regulatory RNA-binding protein CsrA. Additionally, CbsR12 is a *trans*-acting sRNA that targets transcripts of the *carA*, *metK*, and *cvpD* genes *in vitro* and *in vivo*.

*Bartonella bacilliformis* is a Gram-negative alphaproteobacterium and the etiological agent of Carrión’s disease in humans. *B. bacilliformis* is spread between humans through the bite of female phlebotomine sand flies. As a result, the pathogen encounters significant environmental shifts during its life cycle, including changes in pH and temperature. Bacterial sRNAs can serve as a means of rapid regulation under shifting environmental conditions. We therefore performed total RNA-sequencing analyses on *B. bacilliformis* grown *in vitro* then shifted to one of ten distinct conditions that simulate various environments encountered by the pathogen during its life cycle. From this, we identified 160 sRNAs significantly expressed under at least one of the conditions tested. Northern blot analysis was used to confirm the expression of eight novel sRNAs. We also characterized a *Bartonella bacilliformis* group I intron (Bbgpl) that disrupts an un-annotated tRNA<sub>CCU</sub><sup>Arg</sup> gene and determined that the intron splices *in vivo* and self-splices *in vitro*. Furthermore, we verified the predicted molecular targeting of a sand fly-specific sRNA, *Bartonella bacilliformis* small RNA 9 (BbsR9), to transcripts of the *ftsH*, *nuoF*, and *gcvT* genes, *in vitro*.  

Wachter, Shaun, Doctor of Philosophy, Summer 2020

Microbiology
Acknowledgments

I would like to first and foremost thank my amazing wife, Jenny Wachter, for her incredible support throughout the last five years of our lives. We fell in love over gonorrhea so many years ago, and our passion for microbiology continues to bring us together.

For my two beautiful children and native Montanans: Jesse and Elise.

I’d like to thank all of my colleagues who have made this research possible. First and foremost, Mike Minnick, my wonderful PI and mentor. The intellectual freedom you have given me, combined with little pushes when I started slacking, have helped me grow into the scientist I am today. Thank you for the funding throughout my PhD, the countless revisions on my manuscripts and grant/fellowship applications, and for maintaining a constant flow of coffee every single day.

Thank you to Linda Hicks for being the best lab technician and sympathetic ear a guy could have.

Thank you to Indu Warrier, Jim Battisti, Sarah Bidwell, and the Secor and Samuels lab for various support throughout the years, whether through research advice, instrument usage, or TA support.

Thank you to Jill Burke, Zooey Zephyr, and other members of the DBS office for various support and patience throughout this grad school experience.

Thank you to frequent collaborators of the Minnick lab, Dr. Rahul Raghavan and Dr. Matteo Bonazzi. Much of the success in my research is due to them.

Thank you to my Mom and Dad, my Stepmom Penny, Stepdad TJ, my cousins Vince, RJ, and Nate, my brother Bill, and good friend Dan, who have served as constant supports and sources of entertainment throughout this entire Montana adventure.

Thank you to the Simpson family for being the background noise to almost every study / writing session since 2006.

Big thanks to my car for surviving 100 mile round trips 5 days a week for 4 years, and thank you Stephen King, Brandon Sanderson, George R. R. Martin, Patrick Rothfuss, and Robert Jordan for the entertainment on all those trips.

Thank you to the Montana Academy of Sciences for providing the opportunities that led to receipt of a graduate student grant.

Finally, thank you to the University of Montana and the Department of Biological Sciences for providing the vehicle and setting for this amazing PhD experience. Thank you especially for the various scholarship/fellowship/grant opportunities that have
provided me with funding over the years, including the Linda Philips Knoblock Fellowship, the Toelle-Bekken Memorial Fund Fellowship, the Dr. Mitsuru J. Nakamura Graduate Scholarship, and the Robert J. and Carol Anthony Seim Scholarship.
Table of Contents

Title Page i
Abstract iii
Acknowledgements iv
Table of Contents vi
List of Figures viii
List of Tables x
List of Abbreviations xi

Chapter 1: Introduction
Small RNAs as a means of rapid gene regulation 1
Overview of sRNA functions 1
The Hfq chaperone 3
CsrA and RsmY/Z 6
sRNA-dependent regulation of virulence 9
Means of identifying sRNA targets 11
Bacterial selfish genetic elements 12
Overview of selfish genetic elements 12
Transposons, insertion sequences, MITEs, and Group I introns 13
Coxiella burnetii is a zoonotic, obligate intracellular human pathogen 16
Overview of pathogenicity and virulence factors 16
C. burnetii small RNAs 18
C. burnetii selfish genetic elements 18
Bartonella bacilliformis is a vector-borne, facultative intracellular human pathogen 19
Overview of pathogenicity 19
Infection cycle and virulence factors 21

Chapter 2: Identification of novel MITEs (miniature inverted-repeat transposable elements) in Coxiella burnetii: implications for protein and small RNA evolution

Abstract 23
Introduction 24
Materials and Methods 25
Results 27
CbsR3 and CbsR13 are members of a novel MITE family 27
QMITE1 copies encode basic peptides and overlap with annotated genes 34
The CbsR16 locus is a member of a second novel MITE family 34
QMITE2 loci are hot-spots for IS1111 insertion 40
QMITE2 is not specific to C. burnetii 41
Full-length QMITE2 displays inter-strain linkage and sequence conservation 42
The quantity of QMITE1 and QMITE2 insertions differs between Coxiella strains 45
QMITE copies affect sRNA genes 49
Discussion 49
Supplementary Material 57

Chapter 3: A CsrA-binding, trans-acting sRNA of Coxiella burnetii is necessary for optimal
intracellular growth and vacuole formation during early infection of host cells

Abstract
Importance
Introduction
Results
CbsR12 is a principal non-rRNA/tRNA/tmRNA transcript during C. burnetii infection of Vero and THP-1 cells
CbsR12 is processed by ribonuclease III in vitro
CbsR12 binds to C. burnetii recombinant CsrA-2, but not CsrA-1, in vitro
A cbsR12 mutant shows prolonged lag phase in axenic media
CbsR12 impacts intracellular replication of C. burnetii
CCV size correlates with CbsR12 production in THP-1 infection
CbsR12 binds to carA, metK, and cvpD transcripts in vitro
CbsR12 binds to metK, carA, cvpD, and ahecY transcripts in C. burnetii cells
CbsR12 negatively affects the quantity of cvpD transcripts and regulates synthesis of CarA and MetK

Discussion
Materials and Methods
Acknowledgments
Supplementary Material

Chapter 4: Novel small RNAs expressed by Bartonella bacilliformis under multiple conditions
reveal potential mechanisms for persistence in the sand fly vector and human host

Abstract
Introduction
Materials and Methods
Results
Identification of B. bacilliformis sRNAs
Verification of select B. bacilliformis sRNAs
Condition-specific sRNAs target mRNAs enriched in specific pathways
BbgpI is a group I intron that splices in vivo and self-splices in vitro
Bbs9 is a sand fly-specific sRNA
Bbs9 targets transcripts of fisH, nuoF, and gcvT in vitro

Discussion
Acknowledgments
Supplementary Material

Chapter 5: Conclusions and Future Directions

QMITEs as a source for sRNAs and highly basic proteins
enhC and QMITEs as a timeline for C. burnetii strain divergence
CbsR12 is a trans-acting sRNA that also binds CsrA
Determining the role of the methionine cycle in C. burnetii
BbsRs as a means of rapid regulation in rapidly changing environments
Bbs9 is a Bartonella-specific sRNA uniquely expressed in the arthropod vector

References
# List of Figures

## Chapter 1

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Generalized mechanism of sRNA regulation</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>Mechanisms of Hfq binding and gene regulation</td>
<td>5</td>
</tr>
<tr>
<td>1.3</td>
<td>Generic mechanism of CsrA negative regulation</td>
<td>7</td>
</tr>
<tr>
<td>1.4</td>
<td>The <em>L. pneumophila</em> CsrA regulatory cascade</td>
<td>9</td>
</tr>
<tr>
<td>1.5</td>
<td>Mechanisms and examples of sRNAs affecting virulence</td>
<td>10</td>
</tr>
<tr>
<td>1.6</td>
<td>Developmental cycle of <em>C. burnetii</em></td>
<td>17</td>
</tr>
<tr>
<td>1.7</td>
<td>Model of an acute <em>B. bacilliformis</em> infection</td>
<td>20</td>
</tr>
</tbody>
</table>

## Chapter 2

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Ambiguous and unambiguous reads map to the CbsR13 locus</td>
<td>29</td>
</tr>
<tr>
<td>2.2</td>
<td>CbsR13 loci contain a canonical IHF-binding site</td>
<td>30</td>
</tr>
<tr>
<td>2.3</td>
<td>CbsR13 loci represent a novel MITE, called QMITE1</td>
<td>34</td>
</tr>
<tr>
<td>2.4</td>
<td>CbsR16 is lowly transcribed, with some ambiguous reads mapping to it</td>
<td>36</td>
</tr>
<tr>
<td>2.5</td>
<td>CbsR16 loci have full-size and small versions</td>
<td>38</td>
</tr>
<tr>
<td>2.6</td>
<td>CbsR16 loci comprise another novel MITE family, termed QMITE2</td>
<td>39</td>
</tr>
<tr>
<td>2.7</td>
<td>QMITE2 is not unique to <em>C. burnetii</em></td>
<td>42</td>
</tr>
<tr>
<td>2.8</td>
<td>Locations of QMITE1 and QMITE2 insertions in the <em>C. burnetii</em> RSA 493 genome</td>
<td>46</td>
</tr>
<tr>
<td>2.9</td>
<td>Locations of QMITE1 and QMITE2 insertions in the <em>C. burnetii</em> Dugway</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>5J108-111 genome</td>
<td></td>
</tr>
<tr>
<td>S2.1-S2.11</td>
<td>Figures S2.1-11</td>
<td>57-64</td>
</tr>
</tbody>
</table>

## Chapter 3

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>CbsR12 binds to CsrA-2, but not CsrA-1, protein <em>in vitro</em></td>
<td>71</td>
</tr>
<tr>
<td>3.2</td>
<td>CbsR12 production and growth effects on <em>C. burnetii</em> grown in ACCM-2</td>
<td>73</td>
</tr>
<tr>
<td>3.3</td>
<td>CbsR12 production and growth effects on <em>C. burnetii</em> infecting THP-1 cells</td>
<td>75</td>
</tr>
<tr>
<td>3.4</td>
<td>CbsR12 affects CCV expansion in infected THP-1 cells</td>
<td>77</td>
</tr>
<tr>
<td>3.5</td>
<td>CbsR12 targets <em>carA</em>, <em>metK</em>, and <em>cvpD</em> transcripts <em>in vitro</em></td>
<td>79</td>
</tr>
<tr>
<td>3.6</td>
<td>CbsR12 targets several <em>C. burnetii</em> transcripts, including those of <em>metK</em>, <em>carA</em> and <em>cvpD</em></td>
<td>81</td>
</tr>
<tr>
<td>3.7</td>
<td>CbsR12 targets and upregulates translation of a <em>carA</em>-luciferase fusion construct</td>
<td>85</td>
</tr>
<tr>
<td>3.8</td>
<td>CbsR12 targets and downregulates translation of a <em>metK</em>-luciferase fusion construct</td>
<td>87</td>
</tr>
<tr>
<td>3.9</td>
<td>CarA and MetK proteins are differentially synthesized in MB-WT, MB-cbsR12, and MB-cbsR12-Comp strains</td>
<td>89</td>
</tr>
<tr>
<td>S3.1-S3.10</td>
<td>Figures S3.1-10</td>
<td>109-119</td>
</tr>
</tbody>
</table>

## Chapter 4

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Most <em>B. bacilliformis</em> sRNAs are expressed under specific conditions</td>
<td>132</td>
</tr>
<tr>
<td>4.2</td>
<td>Northern blot analyses confirm expression of eight putative <em>B. bacilliformis</em> sRNAs</td>
<td>133</td>
</tr>
<tr>
<td>4.3</td>
<td>Condition-specific sRNA targets are enriched in several GO terms and KEGG pathways</td>
<td>136</td>
</tr>
</tbody>
</table>
Figure 4.4: BbgpI self-splices *in vitro* and is spliced *in vivo* 139
Figure 4.5: BbsR9 targets transcripts of *ftsH*, *nuoF* and *gcvT* *in vitro* 142
Figure 4.6: BbsR9 binds its targets through several GC-rich predicted seed regions 144
Figure 4.7: BbsR9 binds to *ftsH*, *nuoF*, and *gcvT* transcripts via specific GC-rich seed regions 145
Figures S4.1-5 156-160

Chapter 5  

Figure 5.1: What is known of the CsrA regulatory cascade/regulon in *C. burnetii* 170
Figure 5.2: CBU_0636 may be a SAM transporter homolog 171
List of Tables

Chapter 2

Table 2.1: Full-size QMITE2 copies exhibit inter-strain linkage conservation 43
Table 2.2: QMITE effects on functional gene products 54

Chapter 3

Table 3.1: Top ten expressed genes across various C. burnetii growth conditions 69
Table 3.2: CbsR12 target prediction using various algorithms 78
Table 3.3: Predicted CsrA motifs in CbsR12 targets 82

Chapter 4

Table 4.1: Conditions used to prepare B. bacilliformis cultures for RNA-Seq experiments 129
Table 4.2: DESeq2 comparisons made 134
Table 4.3: mRNA targets for BbsR9, as predicted by the indicated algorithms 141
Tables S4.1-6 161-166
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>sRNA</td>
<td>small RNA</td>
</tr>
<tr>
<td>IGR</td>
<td>intergenic region</td>
</tr>
<tr>
<td>TSS</td>
<td>transcription start site</td>
</tr>
<tr>
<td>RACE</td>
<td>rapid amplification of cDNA ends</td>
</tr>
<tr>
<td>SD</td>
<td>Shine-Dalgarno</td>
</tr>
<tr>
<td>RBS</td>
<td>ribosome-binding site</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>TCS</td>
<td>two-component system</td>
</tr>
<tr>
<td>CCV</td>
<td><em>Coxiella</em>-containing vacuole</td>
</tr>
<tr>
<td>T4BSS</td>
<td>type IV-B secretion system</td>
</tr>
<tr>
<td>LPS</td>
<td>lipopolysaccharide</td>
</tr>
<tr>
<td>LCV</td>
<td>large-cell variant</td>
</tr>
<tr>
<td>SCV</td>
<td>small-cell variant</td>
</tr>
<tr>
<td>SAM</td>
<td>S-adenosyl methionine</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>quantitative reverse transcription polymerase chain reaction</td>
</tr>
<tr>
<td>RNase III</td>
<td>ribonuclease III</td>
</tr>
<tr>
<td>IHF</td>
<td>integration host fact</td>
</tr>
<tr>
<td>HGT</td>
<td>horizontal gene transfer</td>
</tr>
<tr>
<td>TPM</td>
<td>transcripts per million</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>OF</td>
<td>Oroya fever</td>
</tr>
<tr>
<td>VP</td>
<td>verruga peruana</td>
</tr>
<tr>
<td>HIBB</td>
<td>Bacto heart infusion blood agar</td>
</tr>
<tr>
<td>EMSA</td>
<td>electrophoretic mobility shift assay</td>
</tr>
<tr>
<td>TE</td>
<td>transposable element</td>
</tr>
<tr>
<td>Tn</td>
<td>transposon</td>
</tr>
<tr>
<td>IS</td>
<td>insertion sequence</td>
</tr>
<tr>
<td>MITE</td>
<td>miniature inverted-repeat transposable element</td>
</tr>
<tr>
<td>IVS</td>
<td>intervening sequence</td>
</tr>
<tr>
<td>REP</td>
<td>repetitive extragenic palindrome</td>
</tr>
<tr>
<td>TIR</td>
<td>terminal inverted-repeat</td>
</tr>
<tr>
<td>DR</td>
<td>direct repeat</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

As published, in parts, in:


Bacterial small RNAs as a means of rapid gene regulation

Overview of small RNA functions

Bacterial small RNAs (sRNAs) are small (<500 nts) transcripts that usually do not code for functional proteins. Instead, they serve as cis- and/or trans-acting regulators through a variety of mechanisms (reviewed in [1]). For example, cis-acting sRNAs are often coded antisense to a functional gene target. Upon transcription, the sRNA binds to the mRNA with perfect complementarity, usually culminating in ribonuclease degradation of the target. This effectively limits the free mRNA molecules available for translation (reviewed in [2]). Alternatively, trans-acting sRNAs are often coded in distant intergenic regions (IGRs) and bind to a variety of mRNAs through a more limited base-pairing mechanism involving a seed region of around ~7 - 12 nts (Figure 1.1). Many trans-acting sRNAs have been discovered in bacteria since Escherichia coli MicF was first described in 1984 [3]. These regulatory RNAs have been implicated in a variety of processes, including virulence [4], global regulation of transcription [5], iron homeostasis [6], protein degradation [7], and stress response [8, 9].
**Figure 1.1: Generalized mechanism of sRNA regulation.** sRNAs can positively or negatively regulate translation of the mRNAs to which they bind by freeing or creating a ribosome-binding site occlusion, respectively. Adapted from: Vanderpool *et al.* 2011.

Some sRNAs are widely conserved in bacteria. These sRNAs often serve an essential function, the loss of which may result in cell death. For example, 6S RNA is widely found in all bacterial phyla except for Deinococcus, Thermus, Thermotogae, Tenericutes, Elusimicrobia, and Fibrobacteres, although even in these phyla, highly diverged copies of 6S RNA may be found [10]. 6S RNA functions by binding and sequestering RNA polymerase complexed with sigma 70, leading to global inhibition of transcription under certain conditions [11]. 6S RNA was first discovered in *E. coli* in 1967 [12], but it wasn’t until more than 30 years later that it was found to regulate RNA polymerase activity during stationary phase [5]. Indeed, an *E. coli* strain lacking 6S RNA was found to be at a survival disadvantage during stationary phase [13].
Another example of a highly conserved sRNA is the tmRNA, a bifunctional sRNA that acts as both a tRNA and mRNA. tmRNA primarily functions by binding to and rescuing stalled ribosomes. Unlike an actual tRNA, though, tmRNA lacks an anticodon so instead functions as an mRNA by coding for a short peptide that is added to the C-terminal end of the nascent polypeptide chain [14]. This peptide targets the incomplete nascent protein for degradation. Like 6S RNA, tmRNA was first discovered and described in *E. coli* [15], although unlike 6S RNA, a mutant strain of *E. coli* unable to express tmRNA does not have a significant growth phenotype [7]. The *in vivo* stability of tmRNA is much like that of a tRNA, leading to a generally high abundance of the sRNA in bacterial cells [14]. High expression or conservation of a sRNA, then, is not necessarily linked to necessity. This stresses the need for functional characterization of sRNAs based on factors other than expression and conservation. Indeed, an important additional characteristic of a sRNA may be its reliance on RNA chaperones for function.

*The Hfq chaperone*

Typically, trans-acting sRNAs require assistance in “finding” their respective mRNA targets. In most bacteria, this is accomplished by the RNA chaperone Hfq, which binds to both sRNAs and mRNAs and plays the role of a molecular matchmaker (reviewed in [16]). Hfq is essential for the function of many trans-acting sRNAs that rely on limited base pairing to regulate their target mRNAs [17]. Hfq was first identified more than 50 years ago as a host factor required for replication of the Qβ bacteriophage in *E. coli* (reviewed in [18]). Since then, *hfq* homologs have been discovered in many bacteria, and mutants of these *hfq* genes expectedly cause pleiotropic phenotypes specific to the
repertoire of sRNAs they are associated with (reviewed in [19]). Hfq binds to its target sRNAs and mRNAs via its proximal face, distal face, rim, and C-terminal regions, which are solvent-exposed motifs with unique architectures able to bind RNA molecules [19]. For example, the proximal face of Hfq binds to poly-U stretches immediately following a hairpin loop structure, such as those found in Rho-independent terminators [20]. All known Hfq-binding sRNAs have been found to bind Hfq via this structure [20]. Meanwhile, the distal face of Hfq binds to both sRNAs and mRNAs via a sequence of ribonucleotides that varies between bacteria. For example, the E. coli Hfq distal face binds to (A-A-N)$_n$ repeat motifs, while Staphylococcus aureus Hfq binds to (A-L)$_n$ repeat motifs, where L is a linker ribonucleotide [21]. The rim region of E. coli Hfq is a secondary binding site for UA-rich regions in sRNAs and mRNAs [22], while the C-terminal disordered region of Hfq seems to function as a stabilizing force in the binding of some sRNAs [23]. Generally speaking, Hfq-binding sRNAs can be divided into two classes: Class I sRNAs bind to the proximal and rim domains of Hfq and target mRNAs that bind the distal face, and Class II sRNAs bind the proximal and distal domains of Hfq and target mRNAs with rim domain binding sites (Figure 1.2) [16].
Figure 1.2: Mechanisms of Hfq binding and gene regulation. Negative and positive regulation via sRNA-Hfq complexes is shown in (a) and (b), respectively. Hfq-binding sRNAs can lead to stabilization (c) or degradation (d) of the sRNA. Hfq-sRNA
complexes targeting mRNAs can also lead to mRNA degradation (e). Adapted from: Vogel J et al. 2011.

Hfq is not obligatory, however. For example, S. aureus has several sRNAs but does not require Hfq protein for their activities [24]. Similarly, Coxiella burnetii does not have a readily apparent hfq gene. However, this doesn't rule out the possibility of an atypical Hfq or some other novel RNA chaperone in these bacteria. Bartonella bacilliformis strain KC583 encodes a single hfq gene, although its function in the bacterium has yet to be elucidated.

CsrA and RsmY/Z
Some sRNAs act by binding to and titrating RNA-binding proteins, effectively sequestering them away from regulatory activities. For instance, C. burnetii codes for two homologs (CsrA-1, CsrA-2) of the RNA-binding protein CsrA (also referred to as repressor of stationary phase metabolites, RsmA), which has been shown to regulate metabolism, biofilm formation, and Type IV secretion in other bacteria [25-27]. CsrA functions as a homo-dimer where each monomer binds to a Shine-Dalgarno (SD)-like motif (AGGA or ANGGA), leading to inhibition and, in some cases, upregulation, of translation (Figure 1.3) [28]. CsrA is regulated by CsrA-binding sRNAs, termed CsrB/C (also called RsmY/Z). Classical CsrB/C sRNAs consist of a series of stem-loops containing exposed AGGA or ANGGA motifs that bind and sequester CsrA, effectively limiting its mRNA regulatory capabilities [29]. Some RsmY/Z sRNAs, however, differ in the number of stem-loop regions containing CsrA-binding sites, and can harbor far fewer
motifs than the classical CsrB/C *E. coli* counterparts [30, 31]. The CsrA regulatory cascade has not been studied in *C. burnetii*, in large part due to the absence of readily-discriminable RsmY/Z sRNAs, although the CsrA regulon in *Legionella pneumophila*, a close relative of *C. burnetii*, has been extensively studied [32, 33].

![Figure 1.3: Generic mechanism of CsrA-mediated negative regulation.](image)

Figure 1.3: Generic mechanism of CsrA-mediated negative regulation. CsrA targets SD-like motifs of stem-loop structures. CsrA first binds to a high affinity target upstream of a transcript’s ribosome-binding site (a). Then, the proximity of CsrA to the ribosome-binding site (b) leads to SD sequestration (c), preventing ribosome binding and inhibiting translation. Adapted from: Mercante J *et al.* 2009.
The regulatory cascade leading to CsrA production varies between bacteria. This is due in some part to CsrA being involved in adaptation to host infection conditions, which differs between pathogenic bacteria depending on the niche they inhabit. For example, *L. pneumophila* occupies an intracellular niche in which it is essential for the bacterium to avoid lysosomal degradation. CsrA is a repressor of *L. pneumophila* transmission phenotypes and an activator of intracellular replication [32]. Thus, a successful infection relies on production of CsrA within the intracellular niche, although tight regulation is required for the activation of transmission phenotypes during the later stages of infection.

In *L. pneumophila* and many other bacteria, CsrA is regulated by a two-component system (TCS) referred to as LetA/S, where LetS represents a sensor histidine kinase that senses some environmental stimuli that marks the necessity for transmission phenotypes. Concurrently, the RpoS sigma factor is also produced during *L. pneumophila* stationary phase. RpoS, along with the aid of the LetA response regulator, transcribes the RsmY/Z sRNAs, which bind to and sequester CsrA, allowing for the activation of transmission phenotypes and the production of effectors necessary for survival within the “next” intracellular niche ([Figure 1.4](#)) [27]. Like Hfq, though, CsrA is not ubiquitous among pathogens. *B. bacilliformis*, for example, does not code for any known CsrA homologs.
Figure 1.4: The *L. pneumophila* CsrA regulatory cascade. LetA/S and the RpoS sigma factor are indirect regulators of CsrA via the RsmY/Z sRNAs. Adapted from: Rasis M *et al.* 2009.

*sRNA-dependent regulation of virulence*

sRNAs have been implicated in the regulation of most bacterial processes, from transcription (6S RNA) and housekeeping (tmRNA), to the translation of specific gene subsets (RsmY/Z). Importantly, these sRNAs (6S RNA, tmRNA, and RsmY/Z) all perform their regulatory functions indirectly via protein binding. Most sRNAs function by directly binding mRNAs in *cis* or in *trans*, with or without the assistance of Hfq (Figure 1.5) [34]. For example, one of the most extensively studied *trans*-acting sRNAs is RNAIII of *S. aureus*. Firstly, while it is very uncommon for *trans*-acting sRNAs to
contain an open reading frame (ORF), RNAIII encodes the δ-hemolysin at its 5′ end that aids in virulence [35]. Furthermore, RNAIII targets multiple mRNAs *in trans*, leading to their up- or down-regulation. Among the down-regulated virulence factors are coagulase (*coa* gene) [36] and peptidoglycan hydrolase (*lytM* gene) [37], while the α-hemolysin (*hla* gene) is up-regulated [38]. This regulatory scheme facilitates *S. aureus* dissemination. Another example is the PapR sRNA of uropathogenic *E. coli* strains, which binds to the coding sequence of *papI* mRNA, thereby causing translational repression [39]. The product of *papI* is itself involved in the activation of P-fimbriae biosynthesis. P-fimbriae is an essential virulence factor involved in the attachment of uropathogenic *E. coli* to renal tissue [40]. Regulation by the PapR sRNA, then, prevents activation of P-fimbriae synthesis, when required, during infection.
Figure 1.5: Mechanisms and examples of sRNAs affecting virulence. Numerous examples are given of cis-encoded (A) and trans-encoded (B) sRNAs that affect the virulence of pathogenic bacteria. Adapted from: Chakravarty and Massé. 2019.

Means of identifying sRNA targets

Perhaps the most challenging step in determining the roles of sRNAs in virulence is identifying the mRNA targets to which a particular sRNA binds. As an initial step, it is common to employ algorithms that can scan a given genome for mRNAs that may contain seed regions able to be bound by a particular sRNA. Although several of these algorithms exist, each prioritizes a certain metric, and thus, some tend to perform better than others when comparing the predictions to known in vivo sRNA targets [41]. For example, the TargetRNA2 algorithm prioritizes conservation of the sRNA, followed by the accessibility of the sRNA seed region, the accessibility of the mRNA target region, and the energy of hybridization [42]. As a result, the algorithm is best suited for sRNAs with known homology to other sRNAs. Meanwhile, the IntaRNA 2.0 algorithm prioritizes the energy of hybridization, followed by accessibility of the sRNA and mRNA as determined by RNA secondary structure prediction. Conservation of the sRNA/mRNA, though, is not taken into consideration [43]. Meanwhile, the CopraRNA algorithm rigorously prioritizes conservation of the sRNA and mRNA, while its methods for computing free energy and sRNA/mRNA accessibility are more outdated [44]. It is helpful, then, to have some a priori knowledge of the sRNA of interest when considering the results of these algorithms. For example, the expression pattern of the sRNA in various stages of growth, in axenic culture vs. in vivo conditions, etc., are all valuable in
determining which \textit{in vitro} targets “make sense”. That said, these \textit{in silico} predictions need to be confirmed through other methods such as \textit{in vitro} binding assays and/or \textit{in vivo} cross-linking experiments.

RNA-RNA electrophoretic mobility shift assays (EMSAs) are valuable for the determination of specific sRNA-mRNA interactions. When combined with \textit{in vitro} mutagenesis of predicted seed regions, one can even identify the specific RNA bases mediating these interactions [45, 46]. There are also several methods for the \textit{in vivo} cross-linking of sRNAs to their targets. These sRNA-mRNA complexes are usually captured and sequenced in order to determine the biologically relevant repertoire of mRNA targets for any given sRNA. For example, cross-linking, ligation, and sequencing of hybrids (CLASH) is predicated on the affinity purification of UV cross-linked RNA-protein complexes [47]. Some of these RNA-protein complexes will contain multiple RNAs that are then ligated to each other before RNA-sequencing [47]. For example, CLASH has been used to determine the repertoire of \textit{E. coli} sRNA-mRNA interactions mediated by Hfq during nutrient-limiting conditions [48].

\textbf{Bacterial selfish genetic elements}

\textit{Overview of selfish genetic elements}

A selfish genetic element is a DNA segment that enhances its own transmission and maintenance at the expense of other genes in a given genome. This is regardless of whether such an element would enhance the fitness of an organism or not [49]. The concept of a selfish genetic element was first noted in 1928, when it was found that the female X chromosome in the fruit fly, \textit{Drosophila obscura}, seemed to be over-abundant
in fruit fly populations, indicating an uneven female:male sex ratio that could lead to extinction of the species [50]. In 1950, Barbara McClintock discovered the existence of transposable elements (TEs) in maize [51]. This represented the first characterization of a selfish genetic element. Despite this, it wasn’t until 1980 that Leslie Orgel and Francis Crick popularized the concept of selfish DNA, showing that these elements spread in a population regardless of their effect on fitness [52].

In the context of bacteria, selfish genetic elements constitute anywhere from 0-21% of a genome and seems to vary with ecology rather than phylogeny [53]. Notably, bacteria undergoing reductive evolution harbor larger amounts of selfish genetic elements when compared to other bacteria [54]. The impetus for reductive evolution seems to be the transition of an organism from free-living to obligate parasite [54]. As this transition occurs and some genes become unnecessary due to the nutrient replete environment of the host, large quantities of DNA may be lost. This leads to an increase in the number of pseudogenes (pseudogenization) and the multiplication and spread of selfish genetic elements, in part due to successive bottleneck events [55]. As an obligate intracellular pathogen, *C. burnetii* represents a bacterium undergoing reductive evolution. As such, it contains a high number of pseudogenes (10.1% of all ORFs) and a variety of different selfish genetic elements [56].

*Transposons, insertion sequences, MITEs, and Group I introns*

Bacterial TEs include transposons (Tns), insertion sequences (IS), miniature inverted-repeat transposable elements (MITEs), and group I introns, among others [57]. TEs are considered intracellular mobile genetic elements in that they rely on other mobile genetic elements for inter-cellular spread [57]. While Tns, IS, MITEs, and group I introns all
have similarities in function, there are some differences that distinguish one from another. Tns represent a broad term for TEss and can be divided into two major categories: Class I (retrotransposons) and Class II (DNA transposons). Retrotransposons are most often found in eukaryotes and so won’t be discussed. Class II Tns can be found in both eukaryotes and prokaryotes [57]. Furthermore, each class can be divided into autonomous Tns, which are able to catalyze their own transfer, and non-autonomous Tns, which require some other element for transposition [57]. Broadly, Tns represent large DNA segments that contain a gene coding for a transposase enzyme. That gene is flanked by terminal inverted-repeat (TIR) sequences that aid in targeting the Tn to specific DNA sequences in a cut-and-paste mechanism that is facilitated by the transposase enzyme [57].

Bacterial Tns usually encode other genes that may provide some adaptive benefit for an organism, such as an antibiotic resistance gene [58]. When no gene other than the transposase gene is present, a Tn is referred to as an IS. As a result, IS elements are smaller than Tns. IS elements are widely distributed in bacteria and may have large effects on genome evolution [57]. For example, the phenomenon of IS expansion is seen in bacteria that have recently adopted a host-restricted lifestyle [59]. IS expansion is thought to be an early consequence of host adaptation brought on by the sudden enhancement of genetic drift due to successive population bottlenecks in a host nutrient replete environment [60]. Ultimately, as a bacterium becomes adapted to the host, deletion of IS elements and adjacent DNA leads to genome reduction. As IS elements become non-autonomous due to loss of transposase function, these elements are eventually lost as the bacterial genome becomes more streamlined [57].
MITEs are non-autonomous class II TEs with defective or missing transposase genes. As such, they can only be mobilized in trans by transposases expressed from related Tns [61]. Most bacterial MITEs consist of 4-30 bp TIRs with a TA dinucleotide at their termini. MITEs are typically small (100-400 bp) and do not encode proteins; rather, their transcripts generate highly stable stem-loop structures [62]. MITE insertions have been implicated in virulence by fostering a plastic genome that enhances acquisition of virulence traits [63] and through physical insertions that alter ORFs and directly lead to virulence phenotypes [64]. Promoter regions and ORFs are common features of bacterial MITEs [65-68]. Moreover, integration host factor (IHF)-binding sites and methyltransferase binding domains have been reported [66, 69]. While most MITEs integrate into IGRs, some have been reported: a) in structural RNA genes [70], b) in protein-encoding genes to create in-frame protein fusions [71], and c) proximal to genes whose transcripts are regulated by the corresponding MITE RNA [72, 73]. Thus, MITEs can potentially interact at DNA, RNA or protein levels in a host bacterium, depending upon their structure and genomic sites of integration.

Group I introns are considered to be autonomous class II TEs that insert into tRNAs, rRNAs, and protein-coding genes. Generally speaking, group I introns are ribozymes that, upon transcription, catalyze their own splicing. Although this RNA splicing is auto-catalytic, they sometimes require protein co-factors for self-splicing in vitro, and it is presumed that all group I introns require protein co-factors to some extent for splicing in vivo [74]. Furthermore, some group I introns harbor a gene coding for a homing endonuclease, which is thought to further facilitate their transposition [75].
**Coxiella burnetii is a zoonotic, obligate intracellular human pathogen**

**Overview of pathogenicity**

*C. burnetii* is a Gram-negative, obligate intracellular bacterium and etiological agent of Q (query) fever in humans. Q fever most often manifests as an acute, flu-like illness, which in rare cases progresses to potentially life-threatening endocarditis [76]. *C. burnetii* undergoes a biphasic life cycle in which it alternates between a metabolically-active, replicative large-cell variant (LCV) and a non-replicative, spore-like small-cell variant (SCV) [77]. Upon aerosol transmission of SCVs to a mammalian host, *C. burnetii* is primarily endocytosed by alveolar macrophages, after which it survives acidification of the host phagolysosome and metamorphoses to LCVs. *C. burnetii* then utilizes the fusion of its *Coxiella*-containing vacuole (CCV) with lysosomes and autophagosomes in order to expand the intracellular niche (Figure 1.6) [78, 79]. CCV expansion is dependent on *C. burnetii* protein synthesis, but independent of replication, so expansion of the CCV is facilitated by a repertoire of Dot/Icm effector proteins secreted by a Type IV-B secretion system (T4BSS) [80, 81]. Many Dot/Icm substrates have been identified in recent years [82] and shown to modulate the host inflammasome [83], influence autophagosomal/lysosomal fusion with the CCV by various mechanisms [84-88], and regulate the host transcriptome after localizing to the nucleus [89, 90]. Little is known about regulation of *C. burnetii*’s T4BSS, although the PmrA response regulator has been shown to enhance synthesis of the T4BSS apparatus as well as certain Dot/Icm substrates [91].
Figure 1.6: Developmental cycle of *C. burnetii*. The developmental cycle of *C. burnetii* in alveolar macrophages is shown. The LCV (light blue) and SCV (dark blue) morphotypes are indicated as the infection proceeds. Adapted from: Minnick and Raghavan, 2011.

The only confirmed virulence factor in *C. burnetii* is lipopolysaccharide (LPS). *C. burnetii* LPS undergoes a process called phase variation where its length and molecular properties may change, leading to altered virulence [92]. For example, expression of phase I LPS (full-length LPS) leads to virulence in mammal hosts, while expression of phase II LPS (deep rough LPS) renders the LPS immunogenic, leading to clearance [93]. It has also been shown that repeated passages of *C. burnetii* in embryonated hen’s eggs
with phase I LPS (~10 passages) leads to a phase II LPS phenotype [94]. The noted
difference in immunogenicity is due to the steric hindrance of antibody binding to phase I
LPS compared to phase II LPS [95]. Recently, a series of C. burnetii genes was
implicated in successful synthesis of phase I LPS, and it was shown that an accumulation
of mutations in several genes leads to truncation and formation of phase II LPS in vivo
[96].

C. burnetii small RNAs
A previous study revealed 15 novel C. burnetii sRNAs that were differentially transcribed
either in LCVs vs. SCVs, or in host cell infections vs. growth in ACCM-2 medium [97,
98]. Of special interest were Coxiella burnetii small RNA 1 (CbsR1) and CbsR12, which,
along with being upregulated during infection compared to growth in axenic medium,
also have predicted Rho-independent terminators and strong predicted promoter elements
[97]. We hypothesize that these highly expressed, “infection-specific” sRNAs play
important roles in C. burnetii infection of mammalian cell lines.

C. burnetii selfish genetic elements
C. burnetii’s genome suggests that it is a relatively recent obligate intracellular pathogen,
based upon the high number of pseudogenes and selfish genetic elements [99]. Among
these elements are an intein [100], two group I introns [101], an intervening sequence
(IVS) [102], and TEs, including multiple copies of IS1111 [103]. The IS1111 transposon
has been studied extensively and found to preferentially insert into a palindromic DNA
sequence that is widely distributed throughout the C. burnetii genome [104]. This
palindromic DNA sequence has been described as a GTAG repetitive extragenic
palindrome (REP), although the nature and distribution of the REP has not been
described [105]. There is little information on other families of TEs in the *C. burnetii* genome. **We hypothesize that non-autonomous IS elements such as MITEs also exist within the *C. burnetii* genome.**

*Bartonella bacilliformis* is a vector-borne, facultative intracellular human pathogen

*Overview of pathogenicity and disease*

*B. bacilliformis* is a Gram-negative, facultative intracellular bacterium and the etiological agent of Carrión’s disease in humans. Carrión’s disease often manifests as a biphasic illness characterized by acute hemolytic anemia followed by eruptions of blood-filled hemangiomas of the skin [106]. Timely antibiotic administration restricts the fatality rate of Carrión’s disease to ~10%, although if left untreated, the rate has been reported to be as high as 88% [107, 108]. *B. bacilliformis* is transmitted between humans through the bite of female phlebotomine sand flies, specifically *Lutzomyia* spp. [109, 110]. The endemic region of Carrión’s disease has historically been limited to arid, high-altitude valleys (600 – 3200m) in the Andes Mountains of Peru, Colombia, and Ecuador, reflecting the habitat of the sand fly vector [111, 112].

The initial, acute stage of Carrión’s disease is referred to as Oroya fever (OF), and it is characterized by colonization of the entire circulatory system, leading to infection of ~61% of all circulating erythrocytes [112, 113]. This bacterial burden typically leads to severe anemia, fever, jaundice, and hepatomegaly, among other symptoms [114]. Weeks or months following OF, *B. bacilliformis* seemingly invades endothelial cells, where it triggers cell proliferation and angiogenesis. This event leads to formation of hemangiomas of the skin, referred to as verruga peruana (VP) (**Figure 1.7**). The VP stage
is chronic and lasts about one month to a year [106, 111]. Although Carrión’s disease can present as a severe illness, there are many documented cases with relatively milder symptoms and/or the onset of VP without having presented with OF symptoms [115]. In consideration of reports involving less virulent \textit{B. bacilliformis} strains and the possibility that other \textit{Bartonella} spp. can cause mild symptoms resembling Carrión’s disease, the incidence of the disease is likely underreported [116-118].

\textbf{Figure 1.7: Model of an acute \textit{B. bacilliformis} infection.} Upon transmission by a phlebotomine sand fly (not depicted), \textit{B. bacilliformis} may first invade endothelial cells, followed by dissemination to the blood stream and infection of erythrocytes, leading to cell death and hemolytic anemia. Subsequently, endothelial cells may provide a secondary niche during the chronic VP stage of infection. Adapted from: Gomes and Ruiz, 2017.
Infection cycle and virulence factors

The *B. bacilliformis* infection cycle is strikingly under-studied compared to other vector-borne pathogens. It is clear that the bacterium is transmitted by female *L. verrucarum* sand flies, although artificial feeding experiments showed that *L. longipalpis* can also “vector” the pathogen in the laboratory [110]. These studies also revealed that *B. bacilliformis* colonized and persisted in the lumen of the abdominal midgut of *L. verrucarum* but were digested along with the blood meal in *L. longipalpis* [110]. Despite this, viable bacteria were retrieved from both insects following a 7-d colonization period [110]. Other *Lutzomyia* spp. have been found to contain *B. bacilliformis* DNA, but colonization experiments have not been performed [119]. It has also been suggested that other mammals may serve as reservoir hosts for *B. bacilliformis*. However, serosurveys of animals that came into contact with infected humans were negative for *B. bacilliformis* DNA [120]. Interestingly, in various attempts to establish an animal model of *B. bacilliformis* infection, the bacterium was only able to infect rhesus macaques [121] and owl monkeys [122]. These results suggest that other primates could conceivably serve as natural reservoir hosts for *B. bacilliformis*, although there is a paucity of non-human primate species in *L. verrucarum*’s geographic range. Regardless, the lack of a small animal model severely limits the prospects of laboratory studies examining *B. bacilliformis* infections *in vivo*.

A number of virulence attributes are involved in *B. bacilliformis* pathogenesis, including erythrocyte attachment [123], invasion [124-126] and hemolysis [127]. Similarly, several factors have been implicated in endothelial cell invasion [128] and proliferation [129-131]. However, regulatory mechanisms that facilitate the pathogen’s
virulence, colonization, and persistence in the sand fly have not been explored, to date. The disparate environments encountered by *B. bacilliformis* during transmission from sand fly vector to human host, and back again, suggest that genetic regulatory mechanisms are used to rapidly adapt to prevailing conditions. For example, the temperature of the sand fly vector would be comparable to ambient temperatures in the geographical range of the insect. The competent vector, *L. verrucarum*, is endemic to high-elevation ranges of the Occidental and Inter-Andean valleys of Peru, Colombia, and Ecuador [132], where temperatures range from 17\(^\circ\) C - 22\(^\circ\) C; fairly consistent with laboratory “room temperature” [133]. Upon transmission to the human host, the bacterium would need to adjust to a human body temperature of ~37\(^\circ\) C. Similarly, human blood has a pH of ~7.4, while the pH of the sand fly (*L. longipalpis*) abdominal midgut after a blood meal is ~8.2, lowers to ~7.7 as the blood meal is digested, and decreases to ~6.0 after digestion [134, 135]. In contrast, the thoracic midgut is maintained at pH ~6.0, regardless of digestion status [134]. A rapid means of regulating virulence and stress-related factors to counteract sudden shifts in temperature and pH would be clearly adaptive for *B. bacilliformis*. **We hypothesize that sRNAs play a role in the rapid gene regulation necessary for *B. bacilliformis* adaptation to different aspects of its infection cycle.**
Chapter 2

Identification of novel MITEs (miniature inverted-repeat transposable elements) in Coxiella burnetii: implications for protein and small RNA evolution


Abstract

Coxiella burnetii is a Gram-negative gammaproteobacterium and zoonotic agent of Q fever. C. burnetii's genome contains an abundance of pseudogenes and numerous selfish genetic elements. MITEs are non-autonomous Tns that occur in all domains of life and are thought to be ISs that have lost their transposase function. Like most TEs, MITEs are thought to play an active role in evolution by altering gene function and expression through insertion and deletion activities. However, information regarding bacterial MITEs is limited. Here, we describe two MITE families discovered during research on small non-coding RNAs (sRNAs) of C. burnetii. Two sRNAs, Cbsr3 and Cbsr13, were found to originate from a novel MITE family, termed QMITE1. Another sRNA, CbsR16, was found to originate from a separate and novel MITE family, termed QMITE2. Members of each family occur ~50 times within the strains evaluated. QMITE1 is a typical MITE of 300-400 bp with short (2-3nt) direct repeats (DRs) of variable sequence and is often found overlapping annotated ORFs. Additionally, QMITE1 elements possess sigma-70 promoters and are transcriptionally active at several loci, potentially influencing expression of nearby genes. QMITE2 is smaller (150-190 bps), but has longer
(7-11nt) DRs of variable sequences and is mainly found in the 3' untranslated region (UTR) of annotated ORFs and IGRs. QMITE2 contains a GTAG REP that serves as a target for IS1111 TE insertion. Both QMITE1 and QMITE2 display inter-strain linkage and sequence conservation, suggesting that they are adaptive and existed before divergence of *C. burnetii* strains.

**Introduction**

*C. burnetii* is a Gram-negative, obligate intracellular gammaproteobacterium and the etiologic agent of Q fever in humans. Q fever is an acute, flu-like illness that can present with pneumonitis, hepatitis and malaise. In less than 5% of cases, chronic infection can develop with potentially life-threatening endocarditis as the most common manifestation [76]. *C. burnetii* undergoes a biphasic life cycle in which it alternates between a metabolically-active, replicative large-cell variant (LCV) and a dormant, spore-like small-cell variant (SCV) [136]. Upon inhalation of SCV’s by a mammalian host, alveolar macrophages internalize the bacteria and trap them within a highly acidic (pH ~4.5) parasitophorous vacuole that has features of a mature phagolysosome [78]. *C. burnetii* has adapted to survive in this acidic environment, where it forms a replicative niche. Dot/Icm effectors are translocated to the host cell in a T4BSS-dependent manner in order to establish and maintain the vacuole [81]. LPS is another critical virulence determinant in *C. burnetii* [137], although it has been found to be truncated (rough) in some strains, including the Nine Mile phase II laboratory strain, RSA 439 [138]. Interestingly, the Dugway 5J108-111 strain has a full-length LPS, but is avirulent [139]. Dugway is considered to be the most primitive of the sequenced *C. burnetii* strains based on a larger
genome with apparently less reductive evolution than virulent strains, such as RSA 493 [56]. It is hypothesized that Dugway either contains a gene(s) that impedes infection in humans, or that the virulent RSA 493 strain has some altered virulence gene(s) rendering it infective [140].

*C. burnetii* was recently shown to produce at least 15 sRNAs [97]. In this report, we show that *Coxiella burnetii* sRNA 3 (Cbsr3), CbsR13, and a newly defined sRNA, CbsR16, arose from two novel MITE families of the pathogen. Furthermore, we demonstrate how these novel MITE families can serve as a timeline for IS1111 transposition based upon their linkage and sequence conservation between strains. Finally, we show that although MITE copies show linkage and sequence conservation, an indel in a potential virulence-associated gene (*enhC*) affected by QMITE2 has created a truncated version of the gene in the virulent RSA 493 strain as compared to the avirulent Dugway strain.

**Materials and Methods**

Discontiguous MegaBLAST ([https://blast.ncbi.nlm.nih.gov/Blast.cgi](https://blast.ncbi.nlm.nih.gov/Blast.cgi)) was used as a local alignment program using default parameters to identify regions of homology to CbsR13 and CbsR16 in the *C. burnetii* genome (strains RSA 493; GenBank accession number AE016828.3 and Dugway 5J108-111; GenBank accession number CP000733.1). In order to compare the various QMITE loci in the RSA 493 genome, multiple sequence alignments of QMITE copies were performed using MUSCLE alignments via Geneious version 11.0.2 software with the default settings [141] ([https://www.geneious.com/download/](https://www.geneious.com/download/)). Phylogenetic analyses of various groups of
QMITE insertions were carried out by first trimming the MUSCLE alignments utilizing Gblocks version 0.91b software [142] (http://molevol.cmima.csic.es/castresana/Gblocks.html). This served to eliminate poorly aligned and highly divergent regions in the various alignments. The default parameters are exceptionally stringent and are catered towards longer input sequences. Thus, the minimum block length was reduced to four, and gap positions were allowed for half of the input sequences at each aligned position in order to accommodate the relatively shorter input sequences. Phylogenetic trees of these trimmed alignments were then constructed using FastTree version 2.1 [143] (http://www.microbesonline.org/fasttree/#FAQ). The generalized time-reversible model of nucleotide evolution was used and phylogeny was inferred using maximum likelihood. The resulting Newick tree file was visualized using FigTree version 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/). To support our designations of QMITEs as MITEs, supplemental MITE predictions of the C. burnetii RSA 493 genome were performed using MUSTv2 software [144] (http://www.healthinformaticslab.org/supp/resources.php). Predicted RNA secondary structures used to confirm the presence of TIRs were generated using mfold [145] (http://unafold.rna.albany.edu/?q=mfold). In order to demonstrate the potential for transcription of QMITE inserts, prediction of sigma-70 consensus promoter elements and Rho factor-independent terminators in QMITE inserts was performed using BPROM (http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb) and ARNold (http://rna.igmors.u-psud.fr/toolbox/arnold/), respectively. CIRCOS software [146] (http://circos.ca/software/download/circos/) was
used to visualize and depict positions of QMITEs on the *C. burnetii* chromosome. RNA-Seq data [Sequence Read Archive (SRA) database under accession number SRP041556] were analyzed using a custom pipeline, although various nesoni version 0.128 applications for processing high-throughput sequence data were also used (http://www.vicbioinformatics.com/software.nesoni.shtml). Transcripts per million (TPM) were calculated using custom perl and python scripts that can be accessed through GitHub (https://github.com/shawachter/TPM_Scripts). The Artemis genome browser was used to visualize alignment files generated from ambiguous and unambiguous read data (http://www.sanger.ac.uk/science/tools/artemis) [147]. Other figures were created using Powerpoint 2010 software (Microsoft, Redmond, WA).

**Results**

**CbsR3 and CbsR13 loci are members of a novel MITE family**

CbsR13 was originally identified as a *C. burnetii* sRNA by RNA-Seq analysis of the transcriptome [97]. It is often helpful to analyze both ambiguous and unambiguous reads associated with any RNA-Seq data. Ambiguous reads refer to those reads that can’t be aligned to one specific area of the genome because multiple copies of that sequence exist in the genome. Unambiguous reads refer to those that could only be mapped to one region of the genome. Upon visualization of ambiguous and unambiguous reads that map to the CbsR13 locus, we discovered that there were many ambiguous reads associated with it (Figure 2.1A). We also found that CbsR13 RNA produced a stable predicted secondary structure resembling a very long palindromic sequence (Figure 2.1B). Although a megaBLAST search produced several hits of high homology, the divergent
nature of the CbsR13 sequences necessitated use of a discontiguous megaBLAST search, which identified dozens of sequences with significant homology to CbsR13 in the genome. Specifically, the search identified 44 ranges, with E values of 8e-11 to 3e-123. Of these hits, 21 were at least 75% of the length of CbsR13 (>232 bp). It was noted upon alignment of the regions flanking these sequences that some of the ranges contained truncated 5′ ends and elongated 3′ ends. An artificial sequence combining the native CbsR13 sequence and the 3’ extension (see Figure S2.1) was thus used as an input for another discontiguous megaBLAST search. This search revealed 45 ranges, with E values from 9e-10 to 5e-123. Twenty-three of these hits were at least 75% of the input sequence length (>350 bp). A multiple alignment and phylogenetic analysis of these 23 sequences is shown in Figure 2.2A and Figure S2.2, respectively. The remaining 22 elements ranged in size from 39-321 bp (not shown), possibly representing degenerate forms of the original nucleotide sequences. One megaBLAST hit for the extended-CbsR13 corresponded to a large portion of the CbsR3 gene sequence (i.e., nt 481609-481806) (see Figure 2.2A, range 2) [97]. This result suggests that the two sRNAs share a common ancestor, although unambiguous TPM values from RNA-Seq show that CbsR13 is expressed at a markedly higher level relative to CbsR3 (Figure S2.3). Confirming what is seen in Figure 2.1A, the ambiguous TPMs associated with CbsR3 and CbsR13 are much higher than the unambiguous TPMs, indicating that additional CbsR13 loci are transcriptionally active (Figure S2.3). Indeed, a sigma-70 promoter search using BPROM predicts a promoter in the forward strand and two promoters in the reverse strand of the input sequence (Figure S2.1).
Figure 2.1: Ambiguous and unambiguous reads map to the CbsR13 locus. (A)
Artemis view of reads mapping to the CbsR13 locus (RSA 439 genome). The x-axis indicates the location (bp) on the chromosome and the y-axis indicates coverage of reads mapping to that location. Reads above the y-axis indicate antisense reads, whereas reads
below the y-axis indicate sense reads mapping to that genomic location. Blue lines signify ambiguous reads mapping to this locus, while the red lines denote unambiguous reads. (B) mfold secondary structure prediction of the CbsR13 sRNA. Red, blue, and green lines forming stem structures indicate G-C, A-U, and G-U base-pairing, respectively (predicted ΔG = -128.5 kcal/mol).

Figure 2.2: CbsR13 loci contain a canonical IHF-binding site. (A) MUSCLE sequence alignment of discontiguous megaBLAST hits (>75% of input sequence) associated with the extended-CbsR13 input sequence. Conserved bases appear as gray blocks, while unaligned bases appear as green, yellow, blue, and red bands, corresponding to T, G, C, and A bases, respectively. An identity indicator is shown above the sequence alignment,
where height signifies conservation of bases at that position, with a color indicator for overall identity between aligned ranges (green: 100%, yellow: 20-99%, red: 0-19%). The consensus sequence is shown above the identity indicator as colored bands indicating bases as described above. (B) The same alignment as shown in (A), focusing on the potential IHF-binding site. The sequence above the red line indicates the consensus IHF-binding site utilizing nucleotide notation, and above the alignment is a sequence logo where the height of the displayed bases indicates the relative identity of the aligned base at that position.

A common motif associated with bacterial TEs is an IHF-binding site [67]. IHF is a bacterial DNA-binding protein that binds to a specific DNA motif and facilitates bending of the DNA. It is thought that this bending aids in transposition of the locus [148]. The consensus IHF-binding nucleotide sequence is WATCAANNNNTTR [149]. Although IHF-binding sites are common in bacterial TEs, they are not always present in MITEs [67]. A manual search through the aligned ranges in Figure 2.2A, though, led to the discovery of a well-conserved IHF-binding site (Figure 2.2B). We chose Range 5 (Figure 2.2A) as a representative for this repeated sequence due to its completeness, and utilized mfold to visualize where this IHF-binding site was located and to see if the sequence had a TIR that could aid in the element’s categorization as a MITE. As shown in Figure 2.3A, it is clear that the element has a TIR of 21 bp in length. Based on the length of the element (~400 bp), the TIR, and the multiple loci scattered throughout the C. burnetii RSA 493 genome, we conclude that this element is a bona fide MITE. Moreover, no similar MITEs have been previously described, and BLASTn searches
found no orthologues in other genomes. Thus, we can conclude that this is a novel MITE that we designate as QMITE1. Other ranges in Figure 2.2A generated similar predicted secondary structures, with corresponding TIRs ranging from 21-28 nts (not shown). MUSTv2 software was also employed to confirm QMITE1 as a MITE (Figure S2.4) [144]. Using stringent parameters, MUSTv2 identified eight of the top ten most homologous ranges to the extended-CbsR13 input sequence and also identified 2-4 bp DRs of nucleotide compositions WW, SS, or GAAG. From this information, a model of QMITE1 was generated and is shown in Figure 2.3B.
Figure 2.3: CbsR13 loci represent a novel MITE, called QMITE1. (A) mfold secondary structure prediction of a selected QMITE1 (range 5; predicted $\Delta G = -192.72$ kcal/mol). Red bars bracket the TIRs and the blue line indicates the location of the potential IHF-binding site. (B) Model of QMITE1 depicting DRs as red arrow heads and the TIRs as hatched arrow heads. Length ranges for these features are also shown.

**QMITE1 copies encode basic peptides and overlap with annotated genes**

Along with being transcriptionally active, 19 QMITE1 copies fully contain short, annotated ORFs that encode predicted peptides with an average isoelectric point (pI) of 12.4. These basic peptides can be divided into three major groups based on sequence similarity (**Figure S2.5**), and they constitute the entire DUF1658 family of small, uncharacterized *C. burnetii* proteins in the Pfam database [150].

Other annotated genes that are affected by QMITE1 insertions mainly encode hypothetical proteins of unknown function. However, QMITE1 copies also overlap with several functional genes, including: *ubiB* C-terminal 2-bp overlap, *pntAA* C-terminal 42-bp overlap, *mutT* C-terminal 26-bp overlap, CBU_2058 proline/betaine transporter C-terminal 49-bp overlap, *nagZ* C-terminal 50-bp overlap, and CBU_2020 glutamate transporter C-terminal 3-bp overlap. The effect of these QMITE1 insertions in the 3′ UTRs of these genes could not be determined, although other MITE insertions in 3′ UTRs have been observed to translationally repress the affected genes [151].

**The CbsR16 locus is a member of a second novel MITE family**

We recently identified a new sRNA termed CbsR16 while analyzing CbsR12; a sRNA that is significantly upregulated during *C. burnetii’s* intracellular infection of host cells
[97]. The cbsR16 gene is located immediately downstream of the cbsR12 gene, which shares a bi-directional Rho-independent terminator with cbsR16 (data not shown). When viewing the CbsR16 locus with the Artemis genome browser, it was clear that there was minor differential expression of the locus when taking ambiguous reads into consideration (Figure 2.4A). Additionally, when we analyzed CbsR16 using mfold, the predicted secondary structure was highly stable (Figure 2.4B). Moreover, although QMITE1 is significantly transcribed at more than one location in the C. burnetii genome, CbsR16 is transcribed at a considerably lower level (Figure S2.3), with very minor TPM differences between mapped unambiguous and ambiguous transcripts. This indicates that although other sequences homologous to CbsR16 may exist in the RSA 493 genome, only the locus adjacent to CbsR12 is transcribed to any significant level. The strong secondary structure and minor presence of ambiguously mapped reads of CbsR16, though, warranted a genome-wide search for similar sequences.
**Figure 2.4: CbsR16 is lowly transcribed, with some ambiguous reads mapping to it.**

(A) Artemis view of reads mapping to the CbsR16 locus (RSA 439 genome). The x-axis shows the location (bp) on the chromosome and the y-axis indicates coverage of reads mapping to that location. Reads above the y-axis indicate antisense reads, whereas reads
below the y-axis indicate sense reads mapping to that genomic location. Blue lines signify ambiguous reads mapping to this locus, while red lines signify unambiguous reads. (B) mFold prediction of the CbsR16 sRNA secondary structure ($\Delta G = -85.24 \text{kcal/mol}$). Red, blue, and green lines forming stem structures indicate G-C, A-U, and G-U base-pairing, respectively.

A discontiguous megaBLAST search with CbsR16 resulted in 78 hits with E-values ranging from 1e-07 to 4e-33. We initially divided these 78 hits into two groups: full-size sequences and smaller sequences. From these pools, we aligned those that covered at least 75% of the input CbsR16 sequence. The full-size versions (Figure 2.5A) have a 5’ sequence of ~40 nts that is apparently missing in smaller versions of the element (Figure 2.5B). Phylogenetic trees for these full-size and smaller versions were constructed and can be seen in Figure S2.6 and Figure S2.7, respectively. As with QMITE1, we generated representative predicted secondary structures for the full-size (Range 7, Figure 2.6A) and small ranges (Range 9, Figure 2.6B). Although there are no IHF-binding sites in the CbsR16-like sequences, the full-size ranges have TIRs and are flanked by unique DRs of 7-9 bp, while the smaller ranges are essentially REP elements. Interestingly, these REP elements were previously reported in C. burnetii, although their status as a truncated MITE was not recognized [105]. Taken as a whole, the size (~190 bp), presence of TIRs and DRs, and their distribution across the RSA 493 genome suggest that the CbsR16-like loci are indeed MITEs. We therefore propose to designate this family of elements as QMITE2. A model of QMITE2 is shown in Figure 2.6C. The
smaller QMITE2 copies strongly resemble a REP element; i.e., they do not contain TIRs nor do they have discernible DRs in flanking genomic regions.

Figure 2.5: CbsR16 loci have full-size and small versions. (A) MUSCLE sequence alignment of discontiguous megaBLAST hits that returned full-size versions of the CbsR16 locus. Conserved bases appear as gray blocks, while unaligned bases appear as green, yellow, blue, and red bands, corresponding to T, G, C, and A bases, respectively. An identity indicator is shown above the sequence alignment, where the height signifies conservation of bases at that position, with a color indicator for overall identity between aligned ranges (green: 100%, yellow: 20-99%, red: 0-19%). Above this identity indicator is the consensus sequence, appearing as colored bands indicating bases as described above. (B) As in (A), except the MUSCLE alignment displays the top discontiguous megaBLAST hits (>75% of input sequence) associated with the CbsR16 locus, excluding all full-size hits. An asterisk indicates equivalent positions in the full-size and small versions of QMITE2.
Figure 2.6: CbsR16 loci comprise another novel MITE family, termed QMITE2. (A) mFold prediction of the RNA secondary structure of a full-size version of the CbsR16 repeated locus (range 7; predicted ΔG = -113.09 kcal/mol). Red, blue, and green lines forming stem structures indicate G-C, A-U, and G-U base-pairing, respectively. Red lines bracket the identified TIR. (B) As in (A), but depicting the secondary structure prediction of a small version of the CbsR16 repeated locus (range 9; predicted ΔG = -67.7 kcal/mol).
kcal/mol). (C) Model of QMITE2 depicting DRs as red arrow heads and TIRs as hatched arrow heads. Length ranges for these features are also shown.

As observed with QMITE1, QMITE2 copies may also affect certain annotated ORFs. Although they do not encode annotated genes like some QMITE1 copies, there is some overlap with neighboring functional genes, including a C-terminal 1-bp overlap with \( kdgK \), a C-terminal 8-bp overlap with \( ogt \), a C-terminal 7-bp overlap with \( recN \), a C-terminal 10-bp overlap with CBU_2078 Fic-Family protein, and a C-terminal 6-bp overlap with \( ruvB \). Additionally, although MUSTv2 identified QMITE1 in the RSA 493 genome, it was unable to find QMITE2 under stringent parameters. However, a full-size QMITE2 copy was identified using less stringent parameters (data not shown). The inability for MUSTv2 to identify QMITE2 most likely reflects the filtering parameters of the program itself. Namely, the program searches for copies of the MITE with similar DR’s. If a copy with a similar DR is not found, it will filter it out. QMITE2 has unique DR’s for each copy, making it difficult to detect.

**QMITE2 loci are hot-spots for IS1111 insertion**

While parsing various QMITE2 ranges, we found that 20 of the 21 annotated IS1111 TEs in the RSA 493 genome possessed a small QMITE2 located ~400 bp downstream of their stop codons. These small QMITE2 ranges were aligned and shown in Figure S2.8. A phylogenetic tree of these transposon-associated QMITE2 insertions was created and is shown in Figure S2.9. These ranges are nearly identical to the other small QMITE2 ranges (Figure 2.5B), except that they are missing 10-20 bp at the 5’ end. Upon closer inspection, these “missing” bases are actually located 5’ of the IS1111, indicating that the
transposon inserted into this region of QMITE2. Indeed, this has been described before, although the insertion site was not previously recognized as a MITE [104]. It is worth noting that these QMITE2 copies are more divergent than their IS1111-free counterparts, implying neutral selection while they are associated with IS1111. Interestingly, of the twenty IS1111 insertions in QMITE2 copies, only one clearly inserted into a full-length QMITE2 locus, as the TIR is still discernible up- and down-stream of the Tn insertion. The other QMITE2 loci may also have been full-length once, but their flanking sequences presumably diverged rapidly after insertion.

**QMITE2 is not specific to C. burnetii**

Unlike QMITE1, QMITE2 is apparently not unique to *C. burnetii*. A discontiguous megaBLAST search using the CbsR16 sequence yielded hits in multiple alphaproteobacteria, including *Bradyrhizobium* spp. and *Rhodobacter* spp. These hits had sizes of 83-100 nucleotides in length with E-values ranging from 1E-04 to 8E-07. QMITE2 also appeared in one location in *Lacinicrobium alkaliphilum*, a gammaproteobacteria. These sequences were aligned to the small version of QMITE2 (Figure 2.7) and a phylogenetic tree was constructed (Figure S2.10). The alignment indicates that although the majority of the sequence corresponding to the predicted stem structure of the small QMITE2 is conserved, the palindromic “tip” (see Figure 2.6B, bases 31-53) is more divergent among the alphaproteobacteria shown in the alignment. These results suggest that a majority of the palindromic stem structure may serve some function in *Bradyrhizobium* and *Rhodopseudomonas* spp., while the entirety of this stem is under purifying selection in *C. burnetii*. It’s also worth noting that the 3’ portion of QMITE2 is missing from the alphaproteobacterial MITEs. The 3’ end of QMITE2
comprises half of the TIR formed in the full-length QMITE2 suggesting that full-length QMITE2 never existed in the alphaproteobacterial species or was present further back in their evolutionary histories.

*Figure 2.7: QMITE2 is not unique to *C. burnetii*. MUSCLE sequence alignment of discontiguous megaBLAST hits that returned QMITE2 sequences in other organisms.*

Conserved bases appear as gray blocks, while unaligned bases appear as green, yellow, blue, and red bands, corresponding to T, G, C, and A bases, respectively. An identity indicator is shown above the sequence alignment, where the height signifies conservation of bases at that position with a color indicator for overall identity between aligned ranges (green: 100%, yellow: 20-99%, red: 0-19%). Above this identity indicator is the consensus sequence, appearing as colored bands indicating bases as described above.

Ranges 12, 13, 10, 9, 14 in the sequence alignment refer to small QMITE2 ranges included in the BLAST as shown in *Figure 2.5B*.

**Full-length QMITE2 displays inter-strain linkage and sequence conservation in *Coxiella***
Due to the unique DRs produced by individual full-length QMITE2 insertions, we were interested to see if these DRs displayed inter-strain linkage conservation. To accomplish this, full-length QMITE2 ranges were found in the *C. burnetii* Dugway strain and the DRs produced by these inserts were compared to those produced by QMITE2 inserts in the RSA 493 strain. If there were two DRs that were identical in sequence between strains, we determined if the associated QMITE2 copies were linked by observing syntenic genome blocks that were produced via genome rearrangements as the strains diverged [139]. We discovered that the Dugway strain contains 12 full-length QMITE2 copies versus 10 in RSA 493 (Table 2.1). Furthermore, seven of the nine discernible DRs in the RSA 493 strain had perfect homologs in the Dugway strain and displayed perfect linkage and sequence conservation. The single unique DR in RSA 493 without a counterpart in Dugway resulted from an IS1111 insertion in the corresponding position in Dugway’s genome. Likewise, 11 of the 12 full-length QMITE2 copies in Dugway had unique DRs associated with them and seven of these had perfect homologs in RSA 493, two had a IS1111 inserted into the corresponding position in RSA 493, one position belonged to a genomic segment unique to Dugway, and the final position displayed a QMITE2 inversion in RSA 493, leaving no discernible DR (Table 2.1). In summary, most DRs are conserved in both strains with a few lost via deletion, IS1111 insertion, or genome inversion events.

**Table 2.1**: Full-size QMITE2 copies exhibit inter-strain linkage conservation.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Range</th>
<th>TIR Length</th>
<th>DR Length</th>
<th>DR Sequence</th>
<th>Homolog?</th>
</tr>
</thead>
</table>

43
<table>
<thead>
<tr>
<th>RSA 493</th>
<th>c1006608-1006428</th>
<th>25</th>
<th>*</th>
<th>*</th>
<th>No**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1066751-1066922</td>
<td>29</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1380514-1380685</td>
<td>26</td>
<td>7</td>
<td>TCAGRGG</td>
<td>No***</td>
</tr>
<tr>
<td></td>
<td>c1168547-1168380</td>
<td>24</td>
<td>9</td>
<td>CCGTCAATA</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>c1360856-1360689</td>
<td>23</td>
<td>9</td>
<td>CACATCGAT</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>1988089-1988258</td>
<td>23</td>
<td>7</td>
<td>CAACATTW</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>1586332-1586502</td>
<td>23</td>
<td>9</td>
<td>GTTGGCAGCG</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>220015-220188</td>
<td>25</td>
<td>8</td>
<td>GGGTGTGGT</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>c970302-970140</td>
<td>24</td>
<td>7</td>
<td>GCTACTT</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>1252325-1252500</td>
<td>24</td>
<td>9</td>
<td>TTCTGTTTA</td>
<td>Yes</td>
</tr>
<tr>
<td>Dugway</td>
<td>c334562-334393</td>
<td>23</td>
<td>9</td>
<td>GTTGGCAGCG</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>c1836762-1836594</td>
<td>25</td>
<td>8</td>
<td>GGGTGTGGT</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>2151397-2151569</td>
<td>23</td>
<td>8</td>
<td>CAACATTW</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>117745-117908</td>
<td>22</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>c1299129-1298960</td>
<td>23</td>
<td>9</td>
<td>CCGTCAATA</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>c374053-373882</td>
<td>31</td>
<td>9</td>
<td>AATTTTAAC</td>
<td>No**</td>
</tr>
<tr>
<td></td>
<td>1295396-1295566</td>
<td>26</td>
<td>9</td>
<td>GTATCRTC</td>
<td>No***</td>
</tr>
<tr>
<td></td>
<td>1561569-1561721</td>
<td>21</td>
<td>13</td>
<td>CCTCTCTTCTTSA</td>
<td>No****</td>
</tr>
<tr>
<td></td>
<td>1384775-1384900</td>
<td>23</td>
<td>9</td>
<td>TTCTGTTTA</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>1261463-1261626</td>
<td>17</td>
<td>9</td>
<td>GGGCTTTCA</td>
<td>No***</td>
</tr>
<tr>
<td></td>
<td>c565652-565819</td>
<td>25</td>
<td>9</td>
<td>CACATCGAT</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>c1003105-1002901</td>
<td>24</td>
<td>7</td>
<td>GCTACTT</td>
<td>Yes</td>
</tr>
</tbody>
</table>
QMITE1 and QMITE2 copies in the RSA 493 and Dugway genomes

QMITE1 and QMITE2 (full-size and small) copies were mapped against the RSA 493 genome using Circos software (Figure 2.8) [146]. We identified 45 copies of QMITE1 and 78 copies of QMITE2 in the RSA 493 genome that in total affect 60 annotated ORFs, with 19 of these ORFs being completely contained within QMITE1 copies and encoding the DUF1658 family of proteins (see Figure S2.5). When combined, QMITE1 and QMITE2 copies make up 0.93% of the RSA 493 genome. Interestingly, our analysis revealed that there were generally higher concentrations of QMITE insertions in the second “half” of the genome (~1,000,000 – 1,995,488 bp), with small QMITE “deserts”. Accordingly, the first half of the genome was found to contain lower concentrations of QMITEs, with larger deserts (e.g., 570,000 – 690,000 bp) bearing no QMITE inserts.
Figure 2.8: Locations of QMITE1 and QMITE2 insertions in the *C. burnetii* RSA 493 genome. The outer ring depicts the RSA 493 chromosome in 100,000 bp increments. The next ring depicts locations of forward strand ORFs in green, IS1111 locations in blue, and non-IS1111 TEs in black, followed by reverse strand ORFs in red on the next ring also featuring IS1111 in blue and non-IS1111 TEs in black. The next ring depicts all chromosomal QMITE1 locations. Green ticks indicate QMITE1 insertions oriented in the forward, while red ticks indicate QMITE1 insertions in the reverse orientation. The next
ring depicts QMITE2 insertions, with green ticks indicating forward insertions, red ticks indicating reverse insertions, and blue ticks indicating IS1111-associated QMITE2 inserts. The following ring labels all of the locus tags for ORFs that have some overlap with either QMITE1 or QMITE2 insertions. ORFs labeled in blue are those that are encoded by QMITE1 insertions and represent the DUF1658 family of proteins. Finally, the colored links between blue-labeled ORFs are indicative of groupings of the proteins coded by these genes (see Figure S2.5).

The distribution of QMITE1 and QMITE2 in the Dugway genome is displayed in Figure 2.9. Due to linkage conservation of QMITE1 and QMITE2 copies between strains, the genomic locations of the QMITE copies are generally the same as RSA 493, although due to divergence between strains, there are some differences in the number of QMITE copies. Specifically, there are 53 copies of QMITE1 and 62 copies of QMITE2 that together comprise 0.91% of the Dugway genome. There are also 56 ORFs affected by MITEs in the Dugway strain. All of the functional annotated genes affected are the same in the two strains, except for the enhC gene, which shows a 3’ extension due to an indel linking the gene to a QMITE2 copy. Interestingly, a C-terminally extended EnhC protein has been previously described for the Dugway strain [139].
Figure 2.9: Locations of QMITE1 and QMITE2 insertions in the C. burnetii Dugway 5J108-111 genome. The outer ring depicts the RSA 493 chromosome in 100,000 bp increments. The next ring depicts locations of forward strand ORFs in green, IS1111 locations in blue, and non-IS1111 TEs in black, followed by reverse strand ORFs in red on the next ring also featuring IS1111 in blue and non-IS1111 TEs in black. The next ring depicts all chromosomal QMITE1 locations. Green ticks indicate QMITE1 insertions oriented in the forward, while red ticks indicate QMITE1 insertions in the
reverse orientation. The next ring depicts QMITE2 insertions, with green ticks indicating forward insertions, red ticks indicating reverse insertions, and blue ticks indicating IS1111-associated QMITE2 inserts. The following ring labels all of the locus tags for ORFs that have some overlap with either QMITE1 or QMITE2 insertions. ORFs labeled in blue are those that are encoded by QMITE1 insertions and represent the DUF1658 family of proteins. Colored links are omitted because the DUF1658 protein products remain the same between strains and largely depend on how the genome was annotated.

**QMITE copies affect sRNA genes**

New bacterial sRNAs can arise from degraded bacteriophage genes [152]. Similarly, we show that three sRNAs of *C. burnetii* are derived from MITEs. These results suggest that, as shown in eukaryotes [153], genomic parasitic elements can serve as a source for the generation of novel non-coding RNAs of bacteria. For instance, QMITE1 copies have inserted directly downstream of promoter elements for CbsR3 and CbsR13. Moreover, a QMITE2 copy has apparently provided the -10 promoter element for CbsR16, while the -35 promoter element is located directly upstream of the QMITE2 insert (Figure S2.11). All of these sRNAs show varying levels of expression (see Figure S2.3), indicating that they are being actively transcribed. Furthermore, previously published Northern blots have confirmed that CbsR3 and CbsR13 are transcribed and produce sRNA molecules of the expected size [102].

**Discussion**
We have described two novel MITE elements in *C. burnetii*, termed QMITE1 and QMITE2. Although their structures and distribution are clear, the nature of their transposition and origin remains indeterminate. Several lines of evidence suggest that QMITE copies are ancient and likely lost the ability to transpose before divergence of present-day *C. burnetii* strains. First, *C. burnetii* RSA 493 contains a plasmid called QpH1 that encodes type 4 secretion system substrates involved in virulence [154]. We could not detect QMITE copies in QpH1, or other *C. burnetii* plasmid types, indicating that either *Coxiella* gained the plasmid after the QMITEs lost the ability to transpose or that the plasmid is too gene-rich to contain stable QMITE copies. Second, the fact that QMITE copies show linkage conservation between strains suggests that they were present before the rearrangement of chromosomes that occurred during divergence of strains. Finally, the presence of QMITE deserts in *C. burnetii* chromosomes (see Figures 2.8 and 2.9), especially between CBU_0664 and CBU_0715, which code for non-IS1111 TEs, implies that horizontal gene transfer (HGT) was involved in forming these regions. Indeed, a recent report has shown that this region (608,000 – 660,000 bp; Figure 2.8), is rich in genes that were acquired via HGT, including some LPS biosynthesis genes that are essential to *C. burnetii*'s virulence [155]. The lack of QMITEs in this region indicates that it was acquired after QMITE1 and QMITE2 lost the ability to transpose, but before divergence of strains, since this region displays inter-strain linkage conservation. It is also worth noting that the chromosomal region harboring the *icm/dot* genes involved in type IV secretion display a paucity of QMITE inserts and is flanked by IS1111 TEs that have inserted into QMITE2 copies (see 1,540,000 to 1,580,000 bp in Figure 2.8). This
suggests that QMITE2 copies indirectly affected the evolution of *C. burnetii* from a free-living to an obligate parasite by fostering genome plasticity.

Interestingly, QMITE insertions can also be used as a marker for the transposition of certain IS1111 TEs. For example, it is likely that the IS1111 transposons at CBU_1217a and CBU_1186 in the RSA 493 strain inserted into these positions after divergence from the Dugway strain, because in Dugway there are full-size QMITE2 copies with discernible DRs that have no IS1111 TEs in these positions. Similarly, the CBUD_0567a IS1111 of Dugway inserted into that position after divergence, since there is a full-size QMITE2 copy at this position in the RSA 493 genome.

The uniqueness of the QMITE1 insert sequence suggests that it may have utility as a molecular signature for detecting *C. burnetii* in clinical or environmental samples. A current detection protocol utilizes PCR to amplify the so-called *htpAB*-repetitive element, which is part of the IS1111 TE [156]. Recent reports, however, have expressed concerns regarding this method due to the existence of IS1111 TEs in *Coxiella*-like endosymbionts, which may confound results obtained from environmental samples [157]. The QMITE1 sequence has variable ends, although it maintains a conserved core across insertions in the *C. burnetii* genome (see Figure 2.2A) that could easily serve as a sizeable DNA template for PCR amplification. Also, the abundance of insertion sites in the *C. burnetii* genome should ensure sensitivity of the assay.

Although results suggest that QMITE1 is unique to *C. burnetii*, a relative of QMITE2 was observed in very distantly-related alphaproteobacteria. However, these QMITE2 copies are not full-length and strongly resemble transposon-associated QMITE2 copies (see Figure S2.8). These alphaproteobacteria, including several *Bradyrhizobium*
and *Rhodopseudomonas* spp., are root nodule-associated, free-living bacteria that also encode several copies of the IS1111 TE in their genomes. There are several possible scenarios that could help explain the occurrence of QMITE2 between these distantly-related organisms. First, *C. burnetii* may have acquired QMITE2 from root nodule-associated bacteria via HGT (or vice versa) during its free-living past. Indeed, *C. burnetii*’s genome contains relics of competence, including an almost-complete type IV pilus system that could have facilitated uptake of foreign DNA [99]. Second, QMITE2 may be ancient, existing long before divergence of alpha- and gamma-proteobacteria. Finally, it is entirely possible that these alphaproteobacteria acquired QMITE2 copies via cut-and-paste transposition of IS1111 following HGT, which in turn left relics of QMITE2 scattered across their respective genomes. This latter explanation is certainly possible since full-size QMITE2 copies are absent in these species and the shorter QMITE2 copies they harbor are highly divergent.

All functional annotated genes affected by QMITE contain insertions located at their 3’ ends. The reason for this preference is unknown but may reflect the general tractability of the C-terminus of proteins to a change in amino acid composition. Indeed, when comparing these protein products to counterparts in *L. pneumophila*, there is no significant difference in the overall masses of the proteins, indicating that QMITE insertions neither extend nor truncate the proteins to a significant degree, although the amino acid composition is altered. These alterations are summarized in Table 2.2 below. In general, QMITE insertions into these genes increase the hypothetical pI of the encoded protein relative to predicted products lacking the QMITE insert. Such a chimera could have conceivably provided a subtle, adaptive advantage to *C. burnetii* as it transitioned
from a free-living bacterium to an obligate intracellular pathogen, as high pI proteins could potentially serve as proton sinks in an acidic host cell phagolysosome. In fact, many *C. burnetii* proteins have been described as having a very high pI, comparable to those found in the human stomach pathogen, *Helicobacter pylori* [99]. This may have been adequate to confer a selective advantage, but the alternative possibility is that QMITE insertions are simply under neutral selection with little to no effect on the fitness of the gene in question. Unfortunately, among the genes listed in Table 2.2, only orthologues for *recN* and *ruvB* are found in *H. pylori*. Similar to *C. burnetii*, these *H. pylori* (strain 26695) proteins have a theoretical pI of 5.84 and 5.86, respectively. This suggests that maintenance of an acidic pI was necessary and the minor change caused by the QMITE2 insert in these genes had little effect on fitness. It is also worth noting that there seems to be a preference for QMITE insertions at the 3′ end of DNA-binding genes and genes involved in DNA repair, such as *ogt*, *recN*, *mutT*, and *ruvB*. It is possible that these insertions are simply due to their proximity to these genomic locations during transposon-induced DNA repair. In fact, it has been found that transposition of TEs is increased upon genotoxic stress in bacteria [158]. There also does not appear to be any QMITE elements that affect the 5′ end of genes with known functions. This is most likely due to the necessity for a promoter element upstream; a feature that may not be provided by the QMITE insertion. Alternatively, there may be a transcription factor binding site(s) upstream of the ORF that is necessary for regulation of that gene. In contrast, there seems to be no locational preference for QMITE insertion into annotated hypothetical proteins, wherein QMITE insertions sometimes appear in-frame in the middle of the ORF (e.g., QMITE2 insertions in CBU_0752a and CBU_1269a).
An intriguing aspect of QMITE inserts is the influence they can have on sRNAs, depending on where they insert into the genome. It has been suggested that a class of MITEs in *Neisseria* spp., termed the Correia repeats, may insert near sRNA genes and alter their functions [66, 159]. This is similar to what is observed with QMITE1, especially those inserts that give rise to CbsR3 and CbsR13, two confirmed sRNAs.
harboring their own promoters upstream of the QMITE1 inserts and terminating within the confines of the insert itself (see Figure S2.11) [97]. When taking the unambiguous reads associated with all QMITE1 loci into account, the TPMs associated with QMITE1 loci reach approximately 9,342, or 0.93% of all transcripts expressed by C. burnetii. The fact that these promoter elements still exist after divergence of C. burnetii into separate strains speaks to the potential utility of the transcripts they produce, whether they: a) act in trans on mRNA target(s), b) affect expression of neighboring genes, or c) are actively translated to produce the high pI proteins listed in Figure S2.5.

The truly unique aspect of QMITE inserts is the sRNAs they may produce wherever they insert into the genome. It has been shown that the Correia repeats of N. meningitidis give rise to transcripts that are produced at varying levels depending on the specific repeat in question [160]. Here, we confirm this notion by showing that a QMITE2 insert in the coding region of the lowly transcribed sRNA CbsR16 provides the -10 promoter element for the sRNA (see Figure S2.11). Additionally, this seems to be one, if not the only, QMITE2 insert that is transcribed with near-equivalence of the ambiguous and unambiguous TPM data (see Figure S2.3). Additionally, although sRNAs arising from internal QMITE1 promoters have not been established, it is likely that transcripts are being produced by these inserts since many more ambiguous transcripts are associated with these loci than unambiguous transcripts (see Figure S2.3). As seen in Figure S2.1, these QMITE1 insertions also have identifiable promoters on both strands of DNA.

In general, ORFs that are affected by QMITE insertion events were the same between the two strains analyzed. One exception occurs in the Dugway strain’s enhC
gene, which codes for a protein that is thought to inhibit release of peptide fragments during infection by *L. pneumophila*, *C. burnetii*’s closest pathogenic relative [161,162]. The function of EnhC in *C. burnetii*’s pathogenicity has not yet been established, although recent studies have speculated that it may play a similar role to the *L. pneumophila* counterpart [163]. In the Dugway strain, *enhC* is extended due to an in-frame QMITE2 insertion at the 3’ end of the gene. Thus, the C-terminal 33 amino acids are presumably provided by the QMITE2 insertion, and the stop codon occurs immediately downstream. This same QMITE2 insert also exists in RSA 493, although an indel has resulted in a stop codon immediately preceding the element. It is unclear whether the C-terminal extension in Dugway affects EnhC function when compared to the altered protein product expressed by RSA 493. Conceivably, as the Dugway EnhC mRNA is transcribed, the highly stable stem structure conferred by QMITE2 could serve as a substrate for ribonuclease III (RNase III) processing. This could create an mRNA lacking a stop codon, which would, in turn, lead to ribosome stalling and eventual targeting of the nascent polypeptide for degradation [7]. Whether this process occurs as hypothesized is currently under investigation.

A variety of TEs have been previously described in *C. burnetii*. Here, we have characterized two novel MITE families that exist as multiple copies in all annotated strains of *C. burnetii*. QMITE1 is of importance because its promoter elements could influence expression of nearby genes. QMITE2 is noteworthy due to unique DRs that could allow for identification of syntenic blocks and visualization of chromosomal rearrangements that have occurred between *C. burnetii* strains as they diverged. QMITE loci could also be used to identify chromosomal regions derived through HGT after the
QMITE copies became inactive but prior to divergence of strains. The linkage conservation between QMITE1 and QMITE2 elements has helped us establish a timeline that suggests that these elements helped influence the evolution of *C. burnetii* on its path towards becoming an obligate pathogen by serving as sites for IS1111 transposition and inserting into and influencing annotated ORFs and sRNA genes. Finally, we have described the influence that QMITE insertions have had on CbsR3, CbsR13, and CbsR16 sRNA’s, the latter of which is produced from a promoter element within a QMITE2 insert.

**Supplementary Material**

Figure S2.1. Extended-QMITE1 sequence for discontiguous megaBLAST searches.

Predicted sigma-70 promoter elements for: **Forward -10** (red), **Forward -35** (red); **Reverse -10** (blue), **Reverse -35** (blue).
**Figure S2.2.** Maximum likelihood phylogenetic tree of QMITE1 inserts. Node labels are indicated at the corresponding locations, and a branch length legend is shown at the bottom of the figure.
Figure S2.3. QMITE-associated TPMs obtained by RNA-Seq from *C. burnetii* LCVs grown in infected Vero cells (n=2 biological replicates).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Ambiguity</th>
<th>Average TPM (n=2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CbsR3</td>
<td>Unambiguous</td>
<td>242</td>
</tr>
<tr>
<td></td>
<td>Ambiguous</td>
<td>6586</td>
</tr>
<tr>
<td>CbsR13</td>
<td>Unambiguous</td>
<td>373</td>
</tr>
<tr>
<td></td>
<td>Ambiguous</td>
<td>5875</td>
</tr>
<tr>
<td>CbsR16</td>
<td>Unambiguous</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Ambiguous</td>
<td>95</td>
</tr>
</tbody>
</table>

Figure S2.4. MUSTv2 search results indicating identified QMITE1 elements in the *C. burnetii* RSA 493 genome. Attributes of individual MITES are shown.

Figure S2.5. MUSCLE alignment of RSA 493 DUF1658 proteins.
Figure S2.6. Maximum likelihood phylogenetic tree of full-size QMITE2 inserts. Node labels are indicated at the corresponding locations, and a branch length legend is shown at the bottom of the figure.
Figure S2.7. Maximum likelihood phylogenetic tree of small QMITE2 inserts. Node labels are indicated at the corresponding locations, and a branch length legend is shown at the bottom of the figure.

Figure S2.8. MUSCLE alignment of transposon-associated QMITE2 inserts.
Figure S2.9. Maximum likelihood phylogenetic tree of transposon-associated QMITE2 inserts. Node labels are indicated at the corresponding locations, and a branch length legend is shown at the bottom of the figure.
Figure S2.10. Maximum likelihood phylogenetic tree of QMITE2 inserts found in alphaproteobacteria. Node labels are indicated at the corresponding locations, and a branch length legend is shown at the bottom of the figure.
CbsR13 Coding Sequence (181703-181765):

```
GGATGTGATCGAGGATGTGATGGAACCGGGTATAATTCTTTTTATATTCTTCCGGCTAGGGAGGGAATCATATCTAAACCACAGTTAGTGAAGACACTAAAAAAGTCCGTTAGGACCGTGGCGGAGGGGAGGAAATGAGGATTTAAAACAAACCTTCCAAGAAAGGGCAGTGAAGTGGCTGCCCTTGTCCGTTCGCTCGGTACGCTGCATGAGTTTCTCTCCTAATACTCAGACCCGGTCCCCTGGAGGAGCATTTGACCCCAACCTTCTACCCCTCAGGGATGAGGATGAC
```

CbsR3 Coding Sequence (c481204-481907):

```
TCTGAAACTCAATAGTAAGGTTACTATAATACAAATTGTTCTAAAATCTATCAATAGGGGAGGAGGACAGGAATCCTGGAATGACCTCAGAACTGGAGGTCGCCGCTACCCGGCAGGGGAAATAGGAGGATTTAAAACAAACCTTCCAAGAAAGGGCAGTGAAGTGGCTGCCCTTGTCCGTTCGCTCGGTACGCTGCATGAGTTTCTCTCCTAATACTCAGACCCGGTCCCCTGGAGGAGCATTTGACCCCAACCTTCTACCCCTCAGGGATGAGGATGAC
```

CbsR16 Coding Sequence (1403050-1403350):

```
TGCATGATGACATCTCTTCCATGTGGGAAATCTGGGCGCATTGCTAGAAATCAAAGATACAACAAAAAGGCTTGCACCCCATCTTGCGGAGGATAAGTGAAGTGGCTGCCCTTGTCCGTTCGCTCGGTACGCTGCATGAGTTTCTCTCCTAATACTCAGACCCGGTCCCCTGGAGGAGCATTTGACCCCAACCTTCTACCCCTCAGGGATGAGGATGAC
```

-10 promoter element
-35 promoter element

QMITE Insertion

Rho-independent terminator

Figure S2.11. QMITE insertions in functional sRNAs of *C. burnetii*. 

64
A CsrA-binding, trans-acting sRNA of Coxiella burnetii is necessary for optimal intracellular growth and vacuole formation during early infection of host cells


Abstract

Coxiella burnetii is an obligate intracellular gammaproteobacterium and zoonotic agent of Q fever. We previously identified 15 small non-coding RNAs (sRNAs) of C. burnetii. One of them, CbsR12 (Coxiella burnetii small RNA 12), is highly transcribed during axenic growth and becomes more prominent during infection of cultured mammalian cells. Secondary structure predictions of CbsR12 revealed four putative CsrA-binding sites in stem loops with consensus AGGA/ANGGA motifs. We subsequently determined that CbsR12 binds to recombinant C. burnetii CsrA-2, but not CsrA-1, proteins in vitro. Moreover, through a combination of in vitro and cell culture assays, we identified several in-trans mRNA targets of CbsR12. Of these, we determined that CbsR12 binds and upregulates translation of carA transcripts coding for carbamoyl phosphate synthetase A; an enzyme that catalyzes the first step of pyrimidine biosynthesis. In addition, CbsR12 binds and downregulates translation of metK transcripts coding for S-adenosyl methionine (SAM) synthetase, a component of the methionine cycle. Furthermore, we
found that CbsR12 binds to and downregulates the quantity of *cvpD* transcripts, coding for a type IVB effector protein, in mammalian cell culture. Finally, we found that CbsR12 is necessary for expansion of CCVs and affects growth rates in a dose-dependent manner in the early phase of infecting THP-1 cells. This is the first characterization of a *trans*-acting sRNA of *C. burnetii* and first example of a bacterial sRNA that regulates both CarA and MetK synthesis. CbsR12 is one of only a few identified *trans*-acting sRNAs that interacts with CsrA.

**Importance**

Regulation of metabolism and virulence in *C. burnetii* is not well understood. Here, we show that *C. burnetii* small RNA 12 (CbsR12) is highly transcribed in the metabolically active LCV compared to the non-replicative SCV. We show that CbsR12 directly regulates several genes involved in metabolism, along with a type IV effector gene, in *trans*. Additionally, we demonstrate that CbsR12 binds to CsrA-2 *in vitro* and induces autoaggregation and biofilm formation when transcribed ectopically in *E. coli*, consistent with other CsrA-sequestering sRNAs. These results implicate CbsR12 in the indirect regulation of a number of genes via CsrA-mediated regulatory activities. The results also support CbsR12 as a crucial regulatory component early on in a mammalian cell infection.

**Introduction**
A previous study by our group revealed 15 novel *C. burnetii* sRNAs that were differentially transcribed either in LCVs vs. SCVs, or in host cell infections vs. growth in ACCM-2 medium [97, 98]. Among these, *Coxiella burnetii* small RNA 12 (CbsR12) was found to be markedly upregulated in the intracellular niche as compared to ACCM-2. Northern blots also showed that CbsR12 was upregulated in SCVs vs. LCVs in ACCM-2, and revealed two distinct sizes of the sRNA, suggesting that either an alternative transcription start site (TSS) or ribonuclease processing of the sRNA was responsible. In a subsequent study, CbsR3 and CbsR13 were found to originate from transcribed loci of a selfish genetic element, termed QMITE1 [164]. However, despite the identification and verification of several CbsRs, none has been functionally characterized, to date.

In this study, we describe activities of a highly transcribed, infection-specific sRNA of *C. burnetii*, termed CbsR12. Our analyses show that CbsR12 binds to CsrA-2, but not CsrA-1 *in vitro*. We also establish that CbsR12 binds and up-regulates *carA* (CBU_1282) and down-regulates *metK* (CBU_2030) transcripts, *in trans*. The bacterial *carA* gene codes for carbamoyl-phosphate synthetase (small) subunit A (CarA), which forms a heterodimer with carbamoyl-phosphate synthetase (large) subunit B (CarB). The CarAB complex catalyzes the first step in pyrimidine biosynthesis and is involved in arginine biosynthesis in some bacteria [165]. The bacterial *metK* gene codes for SAM synthetase, an enzyme responsible for catalyzing production of SAM, the major donor of methyl groups during metabolism in prokaryotic cells. As a methyl donor, SAM affects DNA methylation and thus global transcription [166]). It has also been implicated in virulence, being necessary for the production of N-acyl homoserine lactones involved in bacterial quorum sensing (reviewed in [167]). We also implicate CbsR12 in expansion of
the CCV, as its size is directly correlated with levels of the sRNA. Furthermore, we find that CbsR12 binds \textit{ahcY} (CBU_2031) and \textit{cvpD} (CBU_1818) transcripts, which are components of the methionine cycle and a T4BSS effector protein, respectively [84]. Overall, this study highlights CbsR12 as a crucial component in early stages of a \textit{Coxiella} infection.

**Results**

\textbf{CbsR12 is a principal non-rRNA/tRNA/tmRNA transcript during \textit{C. burnetii} infection of Vero and THP-1 cells}

CbsR12 was first described as a highly-transcribed, infection-specific sRNA that was upregulated in SCVs compared to LCVs when analyzed by Northern blots [97]. The impetus for our study came when we analyzed previous RNA-Seq data (SRP041556) [97] by converting raw read data into TPM, a normalized measure of gene expression [168]. These results showed that CbsR12 was the most highly transcribed non-tmRNA transcript in both LCVs and SCVs during \textit{C. burnetii} infection of African green monkey kidney epithelial cells (Veros). Additional data from LCVs obtained from a \textit{C. burnetii} infection of monocytic THP-1 cells corroborates the observation that CbsR12 is a principal transcript during infection (Rahul Raghavan, unpublished data). Moreover, we were surprised to find that CbsR12 was more abundant in LCVs, not SCVs (Table 3.1).

<table>
<thead>
<tr>
<th>Rank</th>
<th>ACCM-2^b LCV (3dpi)</th>
<th>ACCM-2^b SCV (21dpi)</th>
<th>Vero LCV (3dpi)</th>
<th>Vero SCV (21dpi)</th>
<th>THP-1 LCV (3dpi)</th>
</tr>
</thead>
</table>

\textbf{Table 3.1. Top ten expressed genes\textsuperscript{a} across various \textit{C. burnetii} growth conditions.}
<table>
<thead>
<tr>
<th></th>
<th>Gene</th>
<th>Length</th>
<th>Gene</th>
<th>Length</th>
<th>Gene</th>
<th>Length</th>
<th>Gene</th>
<th>Length</th>
<th>Gene</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>tmRNA</td>
<td>(104,383)</td>
<td>tmRNA</td>
<td>(4,430)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(104,383)</td>
<td></td>
<td>tmRNA</td>
<td>(153,422)</td>
<td>tmRNA</td>
<td>(22,500)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CBU_1183</td>
<td>(32,842)</td>
<td>Intergenic</td>
<td>(2,632)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CbsR12</td>
<td>(79,870)</td>
<td>CbsR12</td>
<td>(8,521)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RNase P RNA</td>
<td>(63,637)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>CBU_0307</td>
<td>(13,902)</td>
<td>CBU_1538</td>
<td>(2,133)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RNase P RNA</td>
<td>(25,715)</td>
<td>CBU_0089a</td>
<td>(8,012)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CbsR12</td>
<td>(33,331)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>CBU_1224a</td>
<td>(11,023)</td>
<td>Intergenic</td>
<td>(1,977)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CBU_0089a</td>
<td>(13,882)</td>
<td>CBU_2034</td>
<td>(4,375)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6S RNA</td>
<td>(22,540)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>RNase P RNA</td>
<td>(10,3923)</td>
<td>tRNA</td>
<td>(1,902)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CbsR1</td>
<td>(11,445)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CBU_1170</td>
<td>(3,704)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CBU_0456</td>
<td>(17,603)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CBU_0311</td>
<td>(7,544)</td>
<td>CBU_0183</td>
<td>(1,593)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CBU_0718</td>
<td>(10,370)</td>
<td>CBU_0718</td>
<td>(3,703)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CBU_1183</td>
<td>(13,349)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>CBU_0474</td>
<td>(7,120)</td>
<td>Intergenic</td>
<td>(1,543)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CBU_1932</td>
<td>(9,734)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CBU_1280a</td>
<td>(3,522)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Intergenic</td>
<td>(6,928)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CBU_0157</td>
<td>(1,390)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CBU_1272</td>
<td>(9,011)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RNase P RNA</td>
<td>(3,429)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>CBU_0306</td>
<td>(6,666)</td>
<td>Intergenic</td>
<td>(1,340)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CBU_0711</td>
<td>(5,530)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intergenic</td>
<td>(3,340)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>CbsR12</td>
<td>(3,522)</td>
<td>Intergenic</td>
<td>(1,203)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CBU_1170</td>
<td>(4,633)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CBU_1272</td>
<td>(3,134)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Genes as annotated for RSA493 (NCBI Ref Seq: NC_002971.4) with corresponding TPM values in parentheses. ACCM-2 is an axenic growth medium (37). CbsR12 is indicated in boldface. N = 2 biological replicates per condition.

**CbsR12 is processed by ribonuclease III in vitro**

Previous Northern blot analyses showed that CbsR12 produced two bands of approximately 170 and 50 nucleotides, regardless of *C. burnetii* growth conditions or developmental stage [97]. We therefore wished to determine whether these RNAs arose from alternative TSSs for the *cbs12* gene or from RNase III processing of the full-length CbsR12 transcript. We first utilized 5' rapid amplification of cDNA ends (RACE) with total RNA derived from *C. burnetii* LCVs infecting Veros (3dpi) in order to determine
the TSS of CbsR12. This experiment revealed the full-size CbsR12 (~200 nucleotides) as expected, but also indicated that two potential alternative TSSs existed ~110 nucleotides upstream of the \textit{cbsR12} gene’s Rho-independent terminator (\textbf{Figure S3.1A}). To determine whether the TSSs were generated by RNase-mediated cleavage, we treated \textit{in vitro}-transcribed CbsR12 with recombinant \textit{C. burnetii} RNase III [102] and commercially-available \textit{E. coli} RNase III (New England BioLabs). Results showed that CbsR12 was similarly processed by both kinds of RNase III into two RNA fragments with sizes that closely resembled those in the previously-reported Northern blot analysis (\textbf{Figure S3.1B}) [97]. These results strongly suggest that the two sites are not alternative TSSs but are instead generated by RNase III processing.

\textbf{CbsR12 binds to \textit{C. burnetii} recombinant CsrA-2, but not CsrA-1, \textit{in vitro}}

The predicted secondary structure of CbsR12 also revealed four conserved single-strand sequence motifs among the various stem-loop structures (\textbf{Figure S3.1A}). This motif, AGGA/ANGGA, corresponds exactly to the conserved CsrA-binding motif of many bacteria [27, 30]. \textit{C. burnetii} contains two annotated and distinct types of CsrA (termed CsrA-1 [CBU_0024] and CsrA-2 [CBU_1050]) that share 65% primary sequence identity. To test the functionality of these domains, we employed an \textit{in vitro} binding assay and a RNA-protein EMSA to determine if CbsR12 binds to natively-purified recombinant CsrA-1 (rCsrA1) and/or CsrA-2 (rCsrA-2). EMSA results clearly showed that CbsR12 binds to rCsrA-2, but not rCsrA-1, \textit{in vitro} (\textbf{Figure 3.1}). In addition, a CbsR10 negative-control sRNA did not bind either rCsrA. Furthermore, the $K_D$ for rCsrA-2 was determined to be 130 nM, consistent with published values for CsrA-binding sRNAs [169].
Figure 3.1: CbsR12 binds to CsrA-2, but not CsrA-1, protein in vitro. EMSAs showing RNA-protein interactions between biotin-labeled, in vitro-transcribed CbsR12 (0 or 1 nM) and increasing concentrations of purified, rCsrA-1 (A) or rCsrA-2 (B). CbsR10 (at 0 or 1 nM; right of vertical black line) is included as a negative control, as the sRNA contains only a single discernible CsrA-binding motif.

A cbsR12 mutant shows prolonged lag phase in axenic media

A cbsR12 mutant of C. burnetii (strain RSA439 with Tn 327, hereafter referred to as MB-cbsR12), as well as an otherwise isogenic Tn insertional control strain (strain RSA439 with Tn 1832, hereafter referred to as MB-WT), were previously generated using a
Himar1-based Tn system [86]. The location of the Tn insertion of strain MB-cbsR12 is shown in Figure S3.2A. We also constructed a transposon-directed complement of strain MB-cbsR12 (hereafter referred to as MB-cbsR12-Comp) containing the wild-type cbsR12 gene plus ~100 bp of 5' and 3' flanking sequences, to include any potential transcriptional regulator element(s) that could influence cbsR12 expression. PCR was used to confirm the Tn insertions in the MB-cbsR12 and MB-cbsR12-Comp (Figure S3.2B).

Furthermore, we confirmed that the cbsR12 cassette inserted into an IGR between CBU_1788 and CBU_1789 (RSA493: accession number NC_002971.4), and we utilized copy-number quantitative PCR (qPCR) to confirm the single insertional event in MB-cbsR12-Comp (Figure S3.2C).

Next, we conducted growth curve analyses of MB-WT, MB-cbsR12, and MB-cbsR12-Comp strains grown axenically in ACCM-2 (Figures 3.2A, S3.3A, S3.3B) and assayed production of CbsR12 at incremental time points from the LCV stage (~ 1-6 d) through the SCV stage (> 7 d) by quantitative real-time PCR (qRT-PCR) (Figure 3.2B). Growth curve results showed that MB-cbsR12 displayed a prolonged lag phase from 1-3 d post-inoculation that was not observed in MB-WT or MB-cbsR12-Comp strains (Figures 3.2A, S3.3A, S3.3B). Following lag phase, MB-cbsR12 grew at a slightly increased rate relative to the other strains (6-9 d post-infection), but failed to reach cell numbers seen in the other two strains throughout the assay. The “wild-type” (MB-WT) and complemented (MB-cbsR12-Comp) strains produced essentially indistinguishable growth curves. The qRT-PCR results showed that the Tn insertion in MB-cbsR12 completely abrogated CbsR12 production (Figure 3.2B). The results also confirmed
cbsR12’s increased expression in LCVs compared to SCVs as copies of CbsR12 per C. burnetii genome were highest at 3 d post-inoculation.

Figure 3.2: CbsR12 production and growth effects on C. burnetii grown in ACCM-2. (A). Growth curves for MB-WT, MB-cbsR12, and MB-cbsR12-Comp in ACCM-2 as determined by qPCR. The 0dpi time point refers to the inoculum. Values represent the means ± standard error of means (SEM) of three technical replicates. Data are representative of three independent experiments with consistent and indistinguishable results. (B). CbsR12 production over time for MB-WT, MB-cbsR12, and MB-cbsR12-
Comp grown in ACCM-2 as determined by qRT-PCR. Values represent the means ± standard error of means (SEM) of three independent determinations.

**CbsR12 impacts intracellular replication of *C. burnetii***

*C. burnetii* typically infects alveolar macrophages during human infection. We therefore infected a differentiated human monocyte cell line (THP-1) with MB-WT, MB-*cbsR12*, and MB-*cbsR12*-Comp strains. Comparative growth curves showed that MB-*cbsR12* has a slower growth rate in exponential phase (1-3dpi) as compared to the two other strains, and never attained the bacterial cell numbers seen in infections with MB-WT or MB-*cbsR12*-Comp (**Figures 3.3A, S3.3C, S3.3D**). Furthermore, CbsR12 production in THP-1s correlated with replication efficiency of the individual strain. For example, production of CbsR12 in MB-WT and MB-*cbsR12*-Comp strains increased between 1dpi and 3dpi, and CbsR12 levels directly correlated to growth rates of the respective strains between these two time points. However, we observed a dysregulation of CbsR12 in MB-*cbsR12*-Comp infecting THP-1 cells that was strikingly different than what was seen during axenic growth. Specifically, we observed a maintenance of CbsR12 production throughout infection of THP-1s (**Figure 3.3B**), whereas in axenic growth there was a progressive drop-off in synthesis after 3dpi (see **Figure 3.2B**). These results suggest that *cbsR12* expression differs in this host-cell type and that a transcriptional regulatory motif may exist outside the bounds of the complementation insertion, resulting in dysregulation due to the genomic context of *cbsR12* in MB-*cbsR12*-Comp compared to MB-WT in a THP-1 infection.
Figure 3.3: CbsR12 production and growth effects on *C. burnetii* infecting THP-1 cells. (A). Growth curves for MB-WT, MB-\textit{cbsR12}, and MB-\textit{cbsR12-Comp} in THP-1 cells as determined by qPCR. The 0dpi time point refers to the inoculum. Values represent means ± standard error of means (SEM) of three technical replicates. Data are representative of three independent experiments with consistent and indistinguishable results. (B). CbsR12 production over time for MB-WT, MB-\textit{cbsR12}, and MB-\textit{cbsR12-Comp} infecting THP-1 cells, as determined by qRT-PCR. Values represent means ± standard error of means (SEM) of three independent determinations.
**CCV size correlates with CbsR12 production in THP-1 infection**

To more closely examine bacterial-host cell interactions, we employed immunofluorescence assays (IFAs) of *C. burnetii* infecting THP-1 cells. *C. burnetii* colonies and CCV boundaries were visualized using anti-Com1 (anti-Com1 [170]) and anti-LAMP1 antibodies at both 3dpi (late LCVs) and 7dpi (SCVs). Here, we define a *C. burnetii* colony as multiple *C. burnetii* inhabiting a LAMP1-decorated intracellular vacuole. LAMP1 is a host cell protein recruited to lysosomes and found on CCVs after lysosome fusion [171]. We observed a robust infection at 3dpi for MB-WT and MB-cbsR12-Comp strains, whereas the MB-cbsR12 strain only produced a few, small CCVs with relatively unclear boundaries (Figure 3.4A). In contrast, the MB-cbsR12-Comp strain produced CCVs that were similar in size to those generated by the MB-WT strain, reflecting the trend observed in their respective growth curves (see Figure 3.3A).

Quantitatively, the differences in CCV sizes between MB-cbsR12 and the other two strains were significant at 3dpi (Figure 3.4B). However, by 7dpi the CCVs were of similar size in MB-cbsR12 and MB-WT infections, indicating some compensatory or redundant mechanisms allowing for CCV expansion as the infection proceeded, even though MB-cbsR12 genome counts never reached wild-type levels (see Figure 3.3A). Interestingly, the MB-cbsR12-Comp strain formed consistently larger CCVs at 3dpi (significantly greater than MB-WT and MB-cbsR12) and 7dpi (significantly greater than MB-cbsR12), which meshes well with the sustained expression of cbsR12 evidenced throughout the course of infection (see Figure 3.3B). Taken as a whole, these results suggest that CbsR12 is important for optimum growth and establishment of CCVs early.
in the course of infection of THP-1 cells, and CbsR12 can influence CCV expansion throughout a THP-1 infection.

**Figure 3.4: CbsR12 affects CCV expansion in infected THP-1 cells.** (A).

Representative IFAs of MB-WT, MB-cbsR12 and MB-cbsR12-Comp infecting THP-1 cells at 3dpi and 7dpi. *C. burnetii* was probed with anti-Com1 antibodies coupled to Alexa Fluor 488 (green), CCV boundaries were labeled with anti-LAMP1 antibodies coupled to rhodamine (red), and host cell nuclei were labeled with DAPI (blue). (B).

Sizes of individual CCVs in log<sub>10</sub>(pixels) for MB-WT, MB-cbsR12, and MB-cbsR12-Comp. Measurements were taken from 46 individual images of random fields of view spanning three different experiments for each *C. burnetii* strain. Crossbars represent means ± standard error of means (SEM) (** = P < 0.01, one-way ANOVA). Scale bars = 20 µm.

**CbsR12 binds to carA, metK, and cvpD transcripts in vitro**

Although CbsR12 was identified as a CsrA-binding sRNA, nothing was known about the CsrA regulon in *C. burnetii*, making it difficult to ascribe intracellular phenotypes to regulation by CsrA. Therefore, we wanted to determine if CbsR12 could act by regulating
mRNAs in trans. To identify potential mRNA targets of CbsR12, we first employed three in silico sRNA target discovery algorithms. Each algorithm takes into consideration the extent of sRNA-mRNA hybridization, conservation of the sRNA, and the accessibility of both the sRNA and its target, although TargetRNA2 and IntaRNAv2 prioritize accessibility while CopraRNA prioritizes comparative interaction predictions among different strains of the indicated bacterium. From these search results, we omitted genes annotated as hypothetical and chose cvpD, metK, carA, purH, rpsA, and dnaA as potential targets based on conserved predictions (Table 3.2). To get a sense of CbsR12’s ability to bind to these potential mRNA targets, we next performed a RNA-RNA hybridization followed by EMSAs. The results clearly showed that CbsR12 bound to carA, metK, and cvpD transcripts in vitro, but did not interact with rpsA, purH, or dnaA transcripts (Figure 3.5). We further tested CbsR12’s specificity for these transcripts by performing dose-dependent and unlabeled-chase experiments. Results of the EMSA analyses showed that CbsR12 specifically bound carA, metK, and cvpD transcripts in a dose-dependent manner (Figure S3.4).

Table 3.2. CbsR12 targeta prediction using various algorithms

<table>
<thead>
<tr>
<th>Rank</th>
<th>TargetRNA2</th>
<th>IntaRNA</th>
<th>CopraRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CBU_1041 (0.001)</td>
<td>cvpD (0.000017)</td>
<td>cvpD (0.00001)</td>
</tr>
<tr>
<td>2</td>
<td>prlC (0.001)</td>
<td>CBU_0537 (0.00018)</td>
<td>CBU_0537 (0.000039)</td>
</tr>
<tr>
<td>3</td>
<td>suhB (0.001)</td>
<td>CBU_1161 (0.00018)</td>
<td>CBU_0922 (0.00053)</td>
</tr>
<tr>
<td>4</td>
<td>mutS (0.002)</td>
<td>CBU_0922 (0.00091)</td>
<td>CBU_0103 (0.00074)</td>
</tr>
<tr>
<td>5</td>
<td>bioD (0.003)</td>
<td>CBU_2028 (0.0012)</td>
<td>CBU_2028 (0.00078)</td>
</tr>
<tr>
<td>6</td>
<td>trmD (0.003)</td>
<td>CBU_0103 (0.0013)</td>
<td>rpsA (0.0013)</td>
</tr>
<tr>
<td>7</td>
<td>dnaA (0.005)</td>
<td>rpsA (0.0022)</td>
<td>metK (0.0014)</td>
</tr>
<tr>
<td>8</td>
<td>yciL (0.007)</td>
<td>metK (0.0023)</td>
<td>purH (0.0019)</td>
</tr>
<tr>
<td>9</td>
<td>metK (0.008)</td>
<td>purH (0.003)</td>
<td>CBU_1741 (0.0025)</td>
</tr>
<tr>
<td>10</td>
<td>CBU_0558 (0.01)</td>
<td>carA (0.015)</td>
<td>carA (0.0067)</td>
</tr>
</tbody>
</table>
Target genes identified through algorithms, ranked. Calculated p-values for each target are indicated in parentheses. Genes involved in subsequent experiments are indicated by bold-face type.

Figure 3.5: CbsR12 targets carA, metK, and cvpD transcripts in vitro. RNA-RNA EMSA showing hybridization reactions with 10 nM biotin-labeled CbsR12 and 5 nM in vitro-transcribed segments of carA, metK, cvpD, purH, rpsA or dnaA. Anti-CbsR12 represents a 10 nM positive control consisting of a transcript equal in size, but antisense, to the CbsR12 transcript. Arrows indicate un-bound bio-CbsR12 (blue) and bio-CbsR12 bound to RNA targets (red).
CbsR12 binds to *metK*, *carA*, *cvpD*, and *ahcY* transcripts in *C. burnetii* cells

To determine the CbsR12 targetome within *C. burnetii* cells, we employed a Crosslink-Seq technique previously used to detect intracellular mRNA targets of *E. coli* sRNAs [152]. For this procedure, we used *C. burnetii* LCVs grown in ACCM-2 to produce sufficient volumes to capture CbsR12 target RNAs for cDNA library preparation and RNA-Seq analysis. Hybridized RNAs in lysates from both MB-WT and MB-*cbsR12* strains were chemically cross-linked, captured by anti-CbsR12 probes, and analyzed by RNA-Seq in order to identify RNAs enriched in MB-WT compared to MB-*cbsR12*. Crosslink-Seq results confirmed that CbsR12 targeted *carA*, *metK*, and *cvpD* transcripts in *C. burnetii* cells (Figure 3.6), as demonstrated *in vitro* (see Figure 3.5). We also discovered an additional mRNA target, *ahcY*, coding for adenosylhomocysteinase, another component of the methionine cycle. Interestingly, *ahcY* was also predicted as an mRNA target by IntaRNA, although the p-value was not significant (p = 0.13). Additionally, *ahcY* is in an operon with and downstream of *metK*. To address whether *ahcY* is actually a target of CbsR12, or if it is a result of CbsR12’s binding to a polycistronic mRNA, we used the Artemis genome browser to observe Crosslink-Seq reads aligned to the *C. burnetii* RSA439 genome. This analysis showed distinct segments of these genes to which the captured reads mapped, suggesting they are separate binding events (Figure S3.5B). We also confirmed the other identified targets by the same method (Figures S3.5A, S3.5C).
Figure 3.6: CbsR12 targets several *C. burnetii* transcripts, including those of *metK*, *carA* and *cvpD*. Volcano plot highlighting mRNAs that occur at different levels between Crosslink-Seq experiments with strains MB-WT and MB-*cbsR12*. Labeled transcripts are indicated by a green dot and are significantly enriched in MB-WT versus MB-*cbsR12*, identifying them as targets of CbsR12. Black dots represent transcripts not significantly different between the strains tested. Orange dots represent transcripts having a log$_2$-fold change > 1, but a false discovery rate (FDR) > 0.05. There were no transcripts indicated by red dots, which would represent mRNAs with a FDR < 0.05 but a log$_2$-fold change in levels < 1. Data shown are representative of two biological replicates each of MB-WT.
and MB-cbsR12 Crosslink-Seq experiments. The potential CbsR12-binding site in the coding region of the ahcY transcript is inset.

Next, we wanted to determine whether CbsR12-mediated regulation of predicted in-trans mRNA targets would occur independently of CsrA. To this end, we searched for potential CsrA-binding (AGGA/ANGGA) motifs within the 100 bases up and downstream of the start codons of carA, metK, cvpD, and ahcY. This search showed that carA, metK and cvpD contained single potential CsrA-binding sites whereas cvpD had none. Moreover, the motifs of carA and metK did not occur in predicted ribosome-binding sites (RBS), suggesting that CsrA is unlikely to regulate the corresponding transcripts [33] (Table 3.3). Thus, CbsR12 regulation of carA, metK, and cvpD transcripts likely occurs as a direct result of in-trans binding by CbsR12 and independently of CsrA. In contrast, the ahcY sequence has two potential regulatory ANGGA CsrA-binding sites, so we cannot exclude the possibility of indirect regulatory effects caused by CbsR12-mediated sequestering of CsrA. As such, we did not further explore CbsR12-mediated regulation of ahcY.

### Table 3.3. Predicted CsrA motifs in CbsR12 targets

<table>
<thead>
<tr>
<th>CbsR12 Target Gene</th>
<th>Locus Tag&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Base Range&lt;sup&gt;b&lt;/sup&gt;</th>
<th>CsrA Motifs&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Position of Motif&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>carA</td>
<td>CBU_1282</td>
<td>c(1234872-1235073)</td>
<td>1</td>
<td>(71) – (75)</td>
</tr>
<tr>
<td>metK</td>
<td>CBU_2030</td>
<td>1936983-1937183</td>
<td>1</td>
<td>(-12) – (-8)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Locus Tag is the identifier for the gene in the CBU genome database.

<sup>b</sup> Base range indicates the position of the predicted CsrA motif within the gene.

<sup>c</sup> CsrA Motifs indicate the presence of potential CsrA-binding sites.

<sup>d</sup> Position of Motif refers to the specific position of the CsrA motif within the gene.
CbpR12 negatively affects the quantity of *cvpD* transcripts and regulates synthesis of *CarA* and *MetK*

Next, we set out to determine if CbpR12 regulates *carA*, *metK*, and *cvpD* transcripts in *C. burnetii*. First, we performed 5' RACE on the three transcripts in total RNA extracted from LCVs infecting mammalian cells (3dpi). 5' RACE results for the MB-WT *cvpD* gene indicated three apparent TSSs, including a TSS for the full-length transcript, several questionable “TSSs” within the CbpR12-binding site, and an alternative TSS for a short transcript downstream of the CbpR12-binding site and with its own predicted promoter element (Figure S3.6A). Interestingly, putative RBSs and start codons exist downstream of TSSs for both the full-length and short transcripts. Moreover, the two start codons are in-frame with each other and the existence of putative RBSs supports the possibility that translation occurs from both elements. The questionable “TSSs” within the CbpR12-binding region likely result from CbpR12-mediated RNase III degradation of *cvpD* mRNA, because 5' RACE results for MB-*cbsR12*, a strain that lacks CbpR12, did not produce TSSs in this region. We predict that CbpR12 down-regulates production of full-
length CvpD since the CbsR12-binding site occurs in the coding region. However, CbsR12 would predictably have no effect on production of the putative truncated CvpD, as the CbsR12-binding site occurs upstream of the alternative TSS (Figures S3.6A, S3.6B).

To determine if CbsR12 binds to and causes degradation of full-length cvpD transcripts, we performed qRT-PCR on MB-WT, MB-cbsR12, and MB-cbsR12-Comp LCVs obtained from infected THP-1 cells. These results clearly showed that the absence of CbsR12 in strain MB-cbsR12 led to a significant increase in full-length cvpD transcripts in LCVs (3dpi) (Figure S3.6C). At 7dpi, MB-WT and MB-cbsR12 levels were not significantly different, presumably due to reduced CbsR12 production in MB-WT SCVs (see Figure 3.3B). However, cvpD expression was significantly lower in MB-cbsR12-Comp, most likely due to the maintained production of CbsR12 in the strain’s SCVs (see Figure 3.3B). Whether or not two different forms of CvpD are produced from cvpD is unknown, although it appears that CbsR12 may negatively regulate the full-length cvpD transcript.

In order to determine if CbsR12 binds to and regulates carA and metK in a cellular environment, we devised a reporter assay in E. coli that measures the effects of CbsR12 production on translation of carA-luc or metK-luc fusion constructs. 5' RACE results for both MB-WT and MB-cbsR12 strains revealed that carA has two potential TSSs, a finding that is consistent with transcription of E. coli carA [165]. Based on the position of the TSSs, CbsR12 could only regulate the full-length carA transcript and not the shorter mRNA, whose transcription starts immediately upstream of the RBS and downstream of the CbsR12-binding site (Figure 3.7A). From these results, we hypothesized that CbsR12
binds to the 5' UTR of carA and upregulates translation by relieving the secondary structure that occludes the predicted RBS (Figures 3.7A, 3.7B). Results of the E. coli reporter assay confirmed our hypothesis, because translation of luciferase enzyme from a carA5'UTR-luc fusion was significantly upregulated in the presence of CbsR12 relative to a strain lacking the sRNA (Figure 3.7C).

Figure 3.7: CbsR12 targets and upregulates translation of a carA-luciferase fusion construct. (A). Secondary structure of the carA 5' UTR as predicted by mFold. Red asterisks indicate TSSs for the shorter transcripts as determined by 5' RACE. (Nucleotide 1 was determined to be the TSS for the full-length transcript by 5' RACE). Colored lines represent the start codon (green), predicted RBS (red), and determined CbsR12-binding sites (blue). (B). Representation of CbsR12 binding to the carA transcript as determined by IntaRNA, with respective base numbers indicated. The top strand in the model represents the carA sequence, while the bottom strand represents the complementary CbsR12 sequence. (C). carA-luc reporter assay indicating relative luminescence units produced by pBESTluc constructs with: 1) no luciferase production (Frameshifted
Luciferase), 2) pBESTluc vector (pBEST), 3) pBESTluc with the carA 5' UTR upstream of luc but lacking cbsR12 (pBEST + carA 5' UTR), and 4) carA 5' UTR upstream of luc plus the cbsR12 gene driven by a Ptac promoter (pBEST + carA 5' UTR + cbsR12).

Values represent means ± standard error of means (SEM) of three independent determinations (* = P < 0.05, student’s t test).

In contrast, CbsR12 was predicted to downregulate MetK translation by binding to the coding region of the transcript, immediately downstream of its start codon (Figure 3.8A). As is often the case with this type of sRNA-mediated regulation, RNase III would likely be recruited and the metK transcript cleaved, resulting in downregulation of the encoded protein product. Unexpectedly, 5' RACE analyses of metK mRNA also identified apparent alternative “TSSs” within the CbsR12-binding region, suggesting that the truncated mRNAs resulted from CbsR12-mediated RNase III processing (Figures 3.8A, 3.8B). Indeed, 5' RACE analysis of RNA from strain MB-cbsR12 infecting THP-1 cells did not detect the “TSSs”, suggesting they are a product of RNase III processing. Results of the reporter assays in E. coli confirmed our hypothesis, as the presence of CbsR12 significantly down-regulated translation of luciferase from the metK-luc fusion construct compared to a strain lacking the sRNA (Figure 3.8C).
Figure 3.8: CbsR12 targets and downregulates translation of a *metK*-luciferase fusion construct. (A). Secondary structure of the *metK* 5' UTR and initial coding sequence as predicted by mFold. Red asterisks indicate apparent alternative “TSSs” determined by 5' RACE. (Nucleotide 1 was determined to be the TSS for the full-length transcript by 5' RACE). Colored lines represent the start codon (green), a predicted RBS (red), and the determined CbsR12-binding site (blue). (B). Representation of CbsR12 binding to the *metK* transcript as determined by IntaRNA with base numbers indicated. The top strand in the model represents the *metK* sequence, while the bottom strand represents the complementary CbsR12 sequence. (C). *metK*-luc reporter assay indicating relative luminescence units produced by pBESTluc constructs with: 1) no luciferase production (Frameshifted Luciferase), 2) pBESTluc vector (pBEST), 3) pBESTluc with the CbsR12 binding site cloned in-frame into *luc* but lacking the *cbsR12* gene (pBEST + *metK*) and 4) pBESTluc with the CbsR12 binding site cloned in-frame into *luc* plus the *cbsR12* gene driven by a Ptac promoter (pBEST + *metK* + *cbsR12*). Values represent
mean ± standard error of means (SEM) of three independent determinations (* = P < 0.05, student's t test).

Although we determined that CbsR12 targets carA and metK transcripts in vitro and in E. coli cells, we were curious whether the absence of CbsR12 would also result in differential amounts of CarA and MetK proteins in C. burnetii. To this end, we performed Western blots with polyclonal antibody generated against recombinant C. burnetii CarA and MetK. As predicted, when proteins from whole-cell lysates of MB-WT, MB-cbsR12, and MB-cbsR12-Comp strains were compared, we found that CarA was synthesized in MB-WT and MB-cbsR12-Comp strains at comparable levels, but was undetectable in protein profiles of strain MB-cbsR12 (Figures 3.9A, S3.7A). In sharp contrast, MetK was highly synthesized in strain MB-cbsR12 but was produced at relatively lower and comparable levels in the MB-WT and MB-cbsR12-Comp strains (Figures 3.9B, S3.7B).
Figure 3.9: CarA and MetK proteins are differentially synthesized in MB-WT, MB-cbsR12, and MB-cbsR12-Comp strains. (A). Proteins (30 µg total) from MB-WT, MB-cbsR12, and MB-cbsR12-Comp LCVs (mid-log phase; 96h for MB-WT and MB-cbsR12-Comp and 144h for MB-cbsR12) grown in ACCM-2 were resolved on a 10-20% acrylamide gradient SDS-PAGE gel, blotted, probed with rabbit anti-CarA antibodies, and detected with chemiluminescence. The black arrow indicates CarA. (B). Proteins (60 µg total) from MB-WT, MB-cbsR12, and MB-cbsR12-Comp LCVs (mid-log phase; 96h for MB-WT and MB-cbsR12-Comp and 144h for MB-cbsR12) grown in ACCM-2 were resolved on a 10-20% acrylamide gradient SDS-PAGE gels, blotted, probed with rabbit anti-MetK antibodies, and detected with chemiluminescence. The black arrow indicates MetK.
Discussion

In this report, we show that CbsR12, a multifunctional sRNA that binds to mRNAs and to the regulatory protein CsrA-2, is important for proper Coxiella replication and CCV expansion during infection of human macrophage-like THP-1 cells.

Induction of cbsR12 expression in mammalian cell culture vs. in vitro conditions (see Table 3.1) led us to label CbsR12 as “infection-specific” and as a result we hypothesized that it played an important regulatory role in infection of host cells. The cbsR12 sequence is conserved among all C. burnetii strains sequenced to date, underscoring the potential for an important regulatory role, but from an evolutionary viewpoint. Interestingly, though, the cbsR12 gene is missing or degenerate in Coxiella-like endosymbionts [172], suggesting that the sRNA is important for a mammalian infection but is dispensable in endosymbionts that reside in arthropods.

Interestingly, CbsR12 binds rCsrA-2, but not rCsrA-1, in a dose-dependent manner, in vitro. Why CsrA-1 does not bind to consensus motifs present in CbsR12 is unclear, especially since both CsrA-1 and CsrA-2 maintain the critical L4 and R44 RNA-binding residues, although CsrA-1 has these residues at L4 and R46 [173]. There are several examples of pathogens harboring multiple copies of CsrA [174, 175]. For example, RsmF of P. aeruginosa is a homolog of RsmA (CsrA) and functions by binding a subset of mRNAs that RsmA also binds [175]. However, an rsmF mutant did not display a phenotype during infection [175]. Similarly, In C. burnetii, a transposon-mediated csrA-1 mutant was shown to have no intracellular phenotype [87]. It is also conceivable that CsrA-1 diverged during C. burnetii’s adaptation to an intracellular lifestyle and is no longer functional. It is also possible that CsrA-1 binds to a non-
canonical motif not present in CbsR12, although such CsrA homologs have not been described, to our knowledge. Regardless, it is necessary to examine the CsrA-1 and CsrA-2 regulons in order to determine their respective roles during infection.

In *L. pneumophila*, a close relative of *C. burnetii*, successful infection depends on a LetAS-RsmYZ-CsrA regulatory cascade. LetAS is a TCS that regulates production of two sRNAs, RsmY and RsmZ, which in turn act as RNA “sponges” that soak up CsrA and modulate its activity [33]. We determined that CbsR12 possesses only four CsrA-binding sites, similar to the RsmY/Z sRNAs of *L. pneumophila*. Interestingly, *L. pneumophila* RsmY/Z was implicated in the formation of cell aggregates and biofilms when the sRNAs were ectopically overproduced in *E. coli*, mimicking the effects of *E. coli*’s own CsrA-binding sRNAs [31]. Likewise, when we overproduced CbsR12 in *E. coli* reporter assays (Figures 3.7, 3.8), we observed a similar autoaggregative phenotype (Figure S3.8A). Moreover, CbsR12 induced biofilm production in *E. coli*, reflecting a CsrA-depleting phenotype (Figure S3.8B). Together, these results suggest that CbsR12 is a CsrA-sequestering, RsmY/Z-like sRNA, although further research is necessary to determine the exact influence of CbsR12 on the regulatory roles of CsrA-2.

*C. burnetii* has several potential transcription factors that are known to upregulate bacterial expression, including IHF [149], response regulator PmrA [176], and transcription factor DksA [177]. However, only PmrA has been well-studied, to date [91]. It is interesting to note that the *cbsR12* gene contains a close approximation to a PmrA-binding site (consensus sequence with less conserved nucleotides in lowercase: cTTAA-N2-TT-N2-cTTAA) [178] immediately upstream of its predicted -10 promoter element (*cbsR12* sequence: gTTTA-N2-TT-N1-gTTAA). However, the presence of this sequence
does not explain the prolonged CbsR12 production observed during a THP-1 infection by MB-cbsR12-Comp since the putative PmrA-binding sequence is present in the cbsR12 cassette that was inserted. We predict that expression of cbsR12 is regulated by an unidentified TCS in a fashion similar to the L. pneumophila LetAS TCS regulation of RsmYZ sRNAs [33]. In fact, the LetAS TCS is not unlike the GacAS TCS involved in RsmYZ-CsrA cascades of other bacteria [179]. C. burnetii codes for four different GacA response regulators that could bind upstream elements of cbsR12 and regulate its expression [99]. This upstream regulator may, in turn, help to explain the dysregulation of expression seen in MB-cbsR12-Comp during infection of THP-1 cells (see Figure 3.3B) that is not apparent during axenic growth (see Figure 3.2B). Alternatively, expression of cbsR12 could be upregulated by PmrA, and some other regulator may be involved in its down-regulation, in conjunction with RNase III-mediated decay (see Figure S3.1). Together, these would aid in suppression of CbsR12 as the LCV-to-SCV transition occurs, effectively freeing sequestered CsrA-2 to regulate the fate of target transcripts.

It is worth noting that we have identified a second sRNA, C. burnetii small RNA 1 (CbsR1), that possesses 5 putative CsrA-binding sites with an ANGGA motif (Figure S3.9B) [97]. Similar to CbsR12, CbsR1 is also produced at high levels in LCVs infecting Vero cells (see Table 3.1). Furthermore, cbsR1 harbors a putative LetA-binding site similar to that of L. pneumophila RsmY (Figure S3.9A) [31]. Together, these observations suggest that CbsR1 and CbsR12 may represent orthologs of RsmYZ, although further exploration of CbsR1 and its cooperativity with CbsR12 is required. If CbsR1 does indeed serve as a CsrA-binding sRNA, its potential, redundant regulatory
role may help to explain why MB-\textit{cbsR12} CCVs expanded to wild-type sizes as the infection progressed (see Figure 3.4B).

We also found that CbsR12 binds \textit{carA} transcripts (see Figures 3.5, 3.7) and upregulates production of \textit{C. burnetii} CarA (see Figure 3.9A). Pyrimidine metabolism in \textit{C. burnetii} presumably requires CarAB to catalyze the conversion of L-glutamine into carbamoyl phosphate and glutamate, since it is unable to shunt this process through the arginine dihydrolase pathway; \textit{C. burnetii} apparently lacks the necessary enzymes [180]. CbsR12-mediated upregulation of CarA in LCVs would result in increased production of pyrimidines that the pathogen requires for robust intracellular growth.

In \textit{E. coli}, \textit{carA} expression is tightly controlled by a series of transcriptional regulators and the two distinct promoters that are regulated by feedback from arginine and pyrimidines [165]. 5' RACE analysis showed two distinct TSSs for \textit{C. burnetii} carA mRNA, with the full-length transcript containing two CbsR12-binding sites and a shorter putative transcript lacking the site (see Figure 3.7A). We do not believe that the shorter, alternative TSS is due to RNase III-mediated degradation resulting from CbsR12 binding because this alternative TSS remained in 5' RACE analysis of the MB-\textit{cbsR12} strain. We do not know conditions under which the shorter transcript is produced, but it may involve feedback from arginine/pyrimidine in accordance with \textit{carA} regulation in \textit{E. coli}.

We also showed that CbsR12 binds to \textit{metK} transcripts and downregulates production of \textit{C. burnetii} MetK protein (see Figures 3.5, 3.8, 3.9B). MetK is a key component of the methionine cycle, which converts methionine to SAM via MetK, SAM to \textit{S}-adenosylhomocysteine (SAH) via various methylases, SAH to homocysteine via AhcY, and homocysteine to methionine via MetH/MetE. Cells produce homocysteine as
an input molecule through a series of reactions involving activated homoserines (reviewed in [181]). *C. burnetii* is a semi-auxotroph for methionine, since it can potentially grow without methionine in axenic media, albeit at a slower growth rate [182]. Interestingly, *C. burnetii* lacks several components of the methionine synthesis pathway, most notably the ability to produce activated homoserines. Most bacteria activate homoserine through addition of an O-succinyl group catalyzed by MetA or an O-acetyl group catalyzed by MetX (reviewed in [181]). *C. burnetii* apparently lacks genes coding for these enzymes. An ABC methionine transporter has been hypothesized [182] but not verified in *Coxiella*. If this is indeed a functional transporter, CbsR12’s negative regulation of *metK* transcripts makes sense in the context of the sRNA’s high level in LCVs, because any amount of scavenged methionine would be critical to growth. Shifting the equilibrium from SAM synthesis to methionine retention would be necessary as *C. burnetii* rapidly produces proteins to expand its intracellular niche.

SAM is a major methyl donor, is necessary for regulation of numerous enzymes, and has been implicated as a major contributor to virulence [183, 184]. Some bacteria lack *metK* and instead transport SAM directly [185]. There are many uncharacterized transporters encoded in the *C. burnetii* genome, so it is conceivable that a SAM transporter is present [99]. This would allow for SAM scavenging even when MetK production is downregulated by CbsR12. Furthermore, if SAM transport occurs, *C. burnetii* could synthesize methionine without having to scavenge it, since the amino acid can be synthesized from SAM without activated homoserine via the methionine cycle.

We also determined that CbsR12 actively targets *ahcY* transcripts (see Figure 3.6). AhcY is a component of the methionine cycle and catalyzes conversion of SAH into
homocysteine and adenosine. Based on the location of the CbsR12-binding site in the coding region of the ahcY transcript (Figure 3.6 inset), we predict that CbsR12 negatively regulates AhcY translation. The underlying reason for this negative regulation is unknown, although it could help to suppress adenosine and/or homocysteine accumulation in LCVs.

cvpD mRNA was also identified as a CbsR12 target through Crosslink-Seq (see Figure 3.6), and this was confirmed by RNA-RNA hybridization / EMSA and qRT-PCR analyses (see Figures 3.5, S3.6C). In this study, we found that CbsR12 was necessary for CCV expansion in early stages of a THP-1 infection. The mechanism for this is unclear, although it may involve regulation of cvpD, which is required for C. burnetii's intracellular replication and CCV expansion in infected THP-1 and HeLa cells [84]. CbsR12 is predicted to target the coding region of the cvpD transcript and would negatively regulate translation. However, in the context of cbsR12’s expression pattern, this is an unclear association, as one would expect upregulation of CvpD synthesis at a time when CbsR12 is highly produced in LCVs. However, 5' RACE analysis of cvpD transcripts in MB-WT and MB-cbsR12 provides a potential explanation, as an alternative cvpD promoter downstream of the CbsR12-binding site occurs that also possesses a putative RBS and start codon (see Figure S3.6A). From these results, we hypothesize that there are two gene isoforms of cvpD that are transcribed and differentially expressed depending on the C. burnetii morphotype. Due to high expression of cbsR12 in LCVs, the longer cvpD transcript isoform would be downregulated by RNase III. As expression of cbsR12 decreases as the infection proceeds, the longer transcript isoform would accumulate. qRT-PCR data support this explanation, since a lack of CbsR12 in MB-
cbsR12 significantly increased the quantity of long cvpD isoform transcripts (see Figure S3.6C). This hypothesis could be confirmed if the two putative CvpD products could be identified and distinguished.

The 2007-2010 Dutch outbreak involving C. burnetii yielded several newly annotated genomes specific to that epidemic [186]. Curiously, 7 of 13 strains analyzed contained a frameshift deletion in cvpD, leading to premature stop codons [186]. Among these, strains 18430 (NZ_CP014557.1), 14160-001 (NZ_CP014551.1), 701CbB1 (NZ_CP014553.1), and 2574 (NZ_CP014555.1) had single-base deletions that only affected the long cvpD isoform. These strains were isolated from aborted placentas of ruminants and cattle in the Netherlands and France [186]. Additional related strains include the Heizberg (NZ_CP014561.1), Henzerling (NZ_CP014559.1), and RSA 331 (NC_010117.1). These strains, which were isolated from patients with acute Q fever in northern Italy and Greece in the mid 1900’s, harbored 4-bp frameshift deletions near the middle of the cvpD coding region, affecting both long and short cvpD isoforms and introducing premature stop codons [186, 187]. Apparently, CvpD was dispensable for virulence in the latter strains, whereas the Dutch outbreak strains harbored intact cvpD genes, or cvpD genes with a 1-bp frameshift deletion only affecting the longer gene isoform. Thus, in these Dutch isolate strains it appears that CbsR12 regulation of cvpD is dispensable. Granted, there are many genotypic differences between RSA439 and the Dutch isolates [186], and some compensatory mechanism(s) may exist for the absence of cvpD. Alternatively, cvpD may be necessary during infection of human cell lines and dispensable in host-animal infections. Regardless, the role and regulation of the CvpD effector requires further investigation.
cbsR12’s high level of expression during infection likely facilitates and regulates the many functions we have described. CbsR12 not bound to CsrA-2 presumably acts in trans to facilitate efficient replication through translational up-regulation of CarA and down-regulation of MetK, and perhaps potentiates expansion of the CCV by means of cvpD transcript regulation. Furthermore, regulation is most likely independent of CsrA, because these genes lack multiple CsrA-binding sites necessary for regulation (see Table 3.3). It is worth noting that the CsrA-binding sites of CbsR12 do not overlap with metK and cvpD binding sites. Hence, it is feasible that CbsR12 may still regulate metK and cvpD in trans while bound to CsrA-2; in fact, a chaperone-like function such as this has recently been ascribed to CsrA [188].

CbsR12 is unique in that is one of only a few identified trans-acting sRNAs that also binds CsrA (reviewed in [189]). We hypothesize that CbsR12’s role in regulating C. burnetii replication and CCV expansion is due to a combination of in-trans mRNA (metK, carA, cvpD) targeting and regulation of CsrA-2 function. Our lab is currently working to elucidate the CsrA-1/CsrA-2 regulons, along with regulation of the putative CbsR12-CsrA-2 cascade of C. burnetii and the nature of CsrA-CbsR1 binding to clarify the interplay between CbsR12’s roles as a trans-acting and CsrA-sequestering sRNA.

Materials and Methods

Bacterial strains, cell lines and growth conditions

Strains, primers, and plasmids used in this study are listed in Figure S3.10. E. coli was grown in lysogeny broth (LB) supplemented with ampicillin (100 µg/mL) or kanamycin (50 µg/mL), as needed. When necessary, overnight cultures were expanded to 100 mL
LB, grown for 2 h, then supplemented with 1 mM IPTG for induction. *C. burnetii* Nine Mile phase II (strain RSA439, clone 4), MB-WT, MB-*cbsR12*, and MB-*cbsR12-Comp* were grown in ACCM-2 medium [98] supplemented with ampicillin (5 µg/mL) or kanamycin (350 µg/mL) at 5% CO₂, 2.5% O₂, 92.5% N₂ 37°C, and 100% humidity with continuous shaking at 75 RPM [98]. SCVs collected from Vero cells were used for all *C. burnetii* infections and growth curve experiments. Briefly, *C. burnetii* was used to infect Vero cell monolayers for 7d at 5% CO₂ and 37°C, after which the cultures were removed to room temperature and the flask lids tightened and covered for two additional weeks [190]. Following this, SCVs were harvested with digitonin, as previously described [100].

African green monkey kidney (Vero) epithelia (CCL-81; American Type Culture Collection; ATCC) and human monocytic leukemia (THP-1) cells (TIB-202; ATCC) cell lines were maintained in RPMI medium (Gibco) supplemented with 10% fetal bovine serum (FBS, RMBIO) in a humidified atmosphere with 5% CO₂ at 37°C. THP-1 cells were differentiated to macrophages by supplementing the growth medium with 200 nM phorbol myristate acetate (PMA, Sigma) overnight.

**Plasmid construction**

pBESTluc was used as a backbone for all reporter assay constructs and was included in the Luciferase Assay System kit (Promega). pBEST + *metK* was created by inserting nucleotides corresponding to the first 10 codons of *C. burnetii metK* immediately downstream of the *luc* start codon using a Q5 Site-Directed Mutagenesis Kit, as instructed (New England Biolabs). *cbsR12* was cloned into pBESTluc using primers containing XhoI and AfeI restriction sites on the forward and reverse primers,
respectively. The forward primer also encoded a \(Ptac\) promoter and \(lac\) operator. The PCR product was cloned into pBEST + \(metK\) using unique XhoI and AfeI restriction sites in an irrelevant IGR. A frameshifted \(luc\) construct was created as a byproduct of the \(metK\) Q5 mutagenesis of pBESTluc and contained a 1-bp frameshift deletion in the 5' end of \(luc\). pBEST + \(carA\) 5' UTR was created using primers specific to the 5' UTR of \(carA\) with HindIII and BamHI restriction sites on the forward and reverse primers, respectively. The forward primer also encoded the \(Ptac\) promoter. Nucleotides corresponding to the \(lac\) operator were inserted using a Q5 Site-Directed Mutagenesis Kit to create the final pBEST + \(carA\) 5' UTR construct. The \(cbsR12\) gene was inserted into this construct in the same fashion as for pBEST + \(metK\) + \(cbsR12\).

Recombinant CarA and MetK were generated by PCR amplification of \(carA\) and \(metK\) using forward and reverse primers containing BamHI and HindIII restriction sites, respectively. The resulting amplicons were cloned into compatible restriction sites of pQE30 (Qiagen) by standard protocol.

**Axenic growth of \(C.\ burnetii\)**

For growth curves and Crosslink-Seq experiments, \(3.33 \times 10^4\) genomic equivalents (GE)/mL of MB-WT, MB-\(cbsR12\), or MB-\(cbsR12\)-Comp were inoculated into 300 mL ACCM-2 in a 1-liter flask with either chloramphenicol (5 \(\mu\)g/mL), kanamycin (350 \(\mu\)g/mL) or both. GE/mL was initially determined from frozen cell stocks by qPCR as previously described [190], although different \(dotA\) primers were used (Figure S3.10). Cell viability was determined using a BacLight Bacterial Viability kit as instructed (Thermo Scientific).

**\(C.\ burnetii\) infection of differentiated THP-1 cells**
THP-1 cells were seeded onto 4-well chambered glass slides (Labtek) or T-75 flasks. After 2d of growth, 200 nM PMA was added along with fresh medium and cells were allowed to differentiate overnight. The PMA-supplemented medium was removed and fresh medium restored, after which differentiated THP-1 cells were allowed to recover for 4 h prior to *C. burnetii* infection at an MOI of 10. Initial infections were rocked for 2 h at room temperature before returning cells to 5% CO₂ and 37°C. At 1dpi the supernatant was removed, extracellular *C. burnetii* was washed away with warmed 1X PBS, and fresh medium was added.

**Total RNA and genomic DNA extraction and purification**

*C. burnetii* grown in ACCM-2 was centrifuged at 15,000 x g at 4°C for 15 min, after which pellets were resuspended in 1 mL TRI Reagent (Ambion). The suspension was incubated for 1 h at room temperature, frozen for 2 h at -80°C, thawed for 30 min at room temperature, then pipetted vigorously until homogenized. 100 µL BCP (Acros Organics) was added, the solution vortexed for 30 sec, incubated for 5 min at room temperature, and centrifuged at 12,000 x g at 4°C for 10 min. The aqueous phase was then collected and 300 µL 100% ethanol added. The mixture was immediately vortexed for 10 sec, and a RiboPure RNA Purification kit (Ambion) was used to collect, concentrate, and wash the resulting RNA. RNA was collected in nuclease-free H₂O and treated with DNase I for 1 h at 37°C. After RNA precipitation in 100% ethanol, the purified RNA was run on a NanoDrop spectrophotometer (Thermo Scientific) to determine concentration and purity.

In order to purify total RNA from *C. burnetii* grown in THP-1 cell lines, growth medium was first removed and replaced with 1 mL TRI Reagent. Flasks containing TRI Reagent were rocked for 1 h at room temperature, after which cells were mechanically
scrapped and collected into a 15-mL conical tube. The mixture was frozen overnight at -80°C, thawed to room temperature for 30 min, and the RNA purification procedure continued as described above.

Genomic DNA was purified from TRI Reagent mixtures according to manufacturer protocols (Ambion). The resulting DNA was purified using a Nucleotide Removal kit as instructed (Qiagen).

**Quantitative PCR (qPCR) and quantitative real-time PCR (qRT-PCR)**

qPCR and qRT-PCR experiments were performed as previously described [100] using 300 nM of primers specific to cbsR12 and a volume of iQ SYBR Green Supermix (Bio-Rad). The resulting reactions were cycled on a MyiQ Single-Color Real Time PCR Detection System (version 1.0 software) (Bio-Rad). CbsR12 cDNA copy number was normalized against dotA copy number derived from C. burnetii genomic DNA to obtain copy numbers / GE values. Growth curve GE/mL and GE/flask values were obtained from genomic DNA purified from the same cells from which total RNA was purified.

For ACCM-2 growth curves, 30 mL aliquots of a 300-mL culture were removed at each time point. gDNA was extracted and resuspended in 30 μL nuclease-free H₂O. 1 μL gDNA was used in subsequent qPCR reactions producing GE/mL values. For THP-1 growth curves, 8 separate T-75 flasks for each C. burnetii strain tested were inoculated simultaneously. At the specified time point, one flask was taken and 30 μL gDNA was again extracted. 1 μL gDNA was used in subsequent qPCR reactions and the GE/flask values were calculated. Each growth curve and qPCR reaction were performed in triplicate.

**RNase III assay**
RNase III assays were performed as previously described [102] using 200 nM CbsR12 substrate and the *C. burnetii* IVS RNA as a positive control [102]. Resulting reactions were electrophoresed on a 7% denaturing polyacrylamide gels and stained with 2 μg/mL acridine orange to visualize bands.

**Identification of transcription start sites**

5' RACE analysis of *carA, metK, cvpD*, and *cbsR12* transcripts was performed on MB-WT and MB-*cbsR12* RNA extracted from infected Vero (*cbsR12, carA*, and *metK*) or THP-1 (*cvpD*) cells at 3dpi using a 5' RACE System kit (Invitrogen) according to manufacturer protocols and with gene-specific primers (see Figure S3.10). Resulting PCR products were cloned into pCR2.1-TOPO as instructed (Invitrogen) and then sequenced with M13 universal primers by Sanger automated sequencing. In general, three biological replicate total RNA samples were obtained from MB-WT and MB-*cbsR12* infections of Vero and THP-1 cells. From these pools, 5' RACE was performed as above and 4 clones from each replicate were sequenced, producing 12 total clones analyzed for each total RNA pool for each gene analyzed.

**In silico and bioinformatics analyses**

RNA target predictions were carried out using TargetRNA2 [42], IntaRNA [43], and CopraRNA [44] algorithms with default settings. RNA was folded using mFold [145] and visualized with Visualization Applet for RNA [191]. Analyses of RNA-Seq data were carried out as previously described [164]. Briefly, raw fastq files were concatenated, quality-filtered with the FASTX toolkit ([http://hannonlab.cshl.edu/fastx_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)) then clipped, aligned, and filtered with Nesoni version 0.128 tools ([http://www.vicbioinformatics.com/software.nesoni.shtml](http://www.vicbioinformatics.com/software.nesoni.shtml)). TPM were calculated using
custom perl and python scripts that can be accessed through GitHub (https://github.com/shawachter/TPM_Scripts). Crosslink-Seq enrichment was accomplished by processing .bam files using featureCounts [192], followed by use of the DESeq2 package in R version 3.4.4 to obtain differentially expressed genes [193]. The Artemis genome browser was used to visualize generated alignment files (http://www.sanger.ac.uk/science/tools/artemis) [147].

All IFA images were processed and analyzed using Fiji [194] and Cell Profiler [195], respectively. Figures were created using R version 3.4.4, Inkscape (https://inkscape.org/release/inkscape-0.92.4/) and GIMP (https://www.gimp.org/downloads/).

**RNA-RNA hybridization and EMSA**

Regions of target genes were first selected for PCR amplification. The following regions were chosen based on inclusion of predicted 5’ UTRs and CbsR12-binding sites (+1 nucleotide designation is the first nucleotide of the annotated start codon): \(\text{carA} \, (-143) – (-1), \text{metK} \, (-26) – (+110), \text{cvpD} \, (-41) – (+101), \text{purH} \, (-100) – (+44), \text{dnaA} \, (-97) – (+61),\) and \(\text{rpsA} \, (-65) – (+77).\) PCR products (1µg) of desired templates were *in vitro*-transcribed overnight at 37°C with 2.5 mM Ribonucleotide Solution Mix (New England Biolabs) and when needed, 0.5 mM Bio-16-UTP (Invitrogen) using a MAXIscript T7 Transcription kit (Invitrogen). Resulting reactions were incubated for 1 h at 37°C with 1 µL TURBO DNase (Invitrogen), heated for 4 min at 85°C then immediately plunged in ice, and electrophoresed on a 7% polyacrylamide gel for 75 min at 100V. Gels were stained with a 2 µg/mL acridine orange solution and visualized bands were excised and eluted overnight into probe elution buffer (0.5M AmAc, 1mM EDTA, 0.2% SDS) at
37°C. The resulting solution was precipitated with ethanol overnight at -20°C, washed with 70% ethanol, and resuspended in nuclease-free H2O. RNA concentrations were determined using a NanoDrop spectrophotometer. Following this, 10 nM biotin-labeled CbsR12 and, unless otherwise noted, 5 nM target RNA were combined and heated for 5 min at 85°C. A high-salt TMN buffer (100 mM NaCl, 50 mM MgCl2, 100 mM Tris-Cl, 0.05% Tween-20) was then added and reactions were immediately plunged on ice for 30 sec and then incubated for 30 min at 37°C. A non-denaturing loading dye (0.25% bromophenol blue) was added and the resulting RNA mixtures were resolved on a 7% polyacrylamide gel for 1 h 20 min at 100V. RNA was transferred to a BrightStar-Plus Positively Charged Nylon Membrane (Ambion) using an electro-blot transfer system (Bio-Rad) and cross-linked with short-wave UV light in a GS Gene Linker UV Chamber (Bio-Rad). A North2South Chemiluminescent Hybridization and Detection Kit (Thermo Scientific) was used to detect resulting bands. The blot was imaged on a LAS-3000 imaging system (Fujifilm).

**RNA-protein electrophoretic mobility shift assay**

CbsR12-CsrA1/2 EMSAs were performed as previously described for CsrA-binding RNAs [169]. Biotin-labeled CbsR12 was synthesized *in vitro* as above for RNA-RNA EMSAs. *C. burnetii csrA-1* and *csrA-2* genes were cloned into pQE30, expressed, and resulting proteins natively purified as previously described [102]. 1 nM biotin-labeled CbsR12 diluted in TE buffer (10mM Tris-HCL, 1mM EDTA) was heated at 75°C for 3 min and equilibrated to room temperature for 10 min. Purified CsrA1/2 diluted in CsrA-binding buffer (1 µl in 10mM Tris-HCl, 10mM MgCl2, 100mM KCl, 10mM DTT, 10% glycerol, and 10 U RNASin [Promega]) was added, and reactions were incubated for 30
min at 37°C. Samples were immediately resolved on 10% non-denaturing polyacrylamide gels for 3h. Membrane transfer and imaging were performed as in RNA-RNA EMSAs described above. The K_D for CsrA-2 was determined as previously described [169].

**Reporter assay**

A Luciferase Assay System kit (Promega) was used. All pBESTluc constructs were transformed into *E. coli* Top 10 F'. Resulting *E. coli* strains were grown overnight at 30°C in 10 mL LB containing ampicillin (100 µg/mL) and 1% glucose in order to mitigate the autoaggregative effects of CbsR12. An aliquot (4.5 mL) of the overnight culture was inoculated into 40.5 mL LB with 100 µg/mL ampicillin and grown for 1.5 h at 30°C. IPTG was added (to 1 mM) and culture aliquots (100 µl) were removed at 0, 1, and 2 h time points. 80 µL LB and 20 µL CCLR lysis solution (1X CCLR, 25 mg BSA, 12.5 mg lysozyme, 7.5 mL water) were added to the aliquots and gently inverted until the solution clarified. 50 µL of the resulting lysate was aliquoted to a 96-well plate, 100 µL of luciferase assay substrate was added, and luminescence was immediately read with a SpectraMax M5 Plate Reader (Molecular Devices).

**Crosslink-Seq analysis**

RNA-RNA crosslinking was performed essentially as previously described [152, 196], except TRI Reagent was utilized for total RNA extraction as described above and 10 nmol of two distinct biotinylated *in vitro*-transcribed anti-CbsR12 RNAs were used as probes. The resulting captured RNA was sent to the Yale Center for Genomic Analysis for RNA-Seq analysis.

**Immunofluorescence assay**
IFAs on infected THP-1 cells were performed as previously described with modifications [86]. Briefly, 4-well chambered glass slides were coated for 30 min with a 0.2% solution of Sigmacote (Sigma). THP-1 cells were inoculated into chambered slides and incubated overnight or until 60% confluence was reached and then differentiated with 200 nM PMA. Confluent cells were then infected with MB-WT, MB-cbsR12, or MB-cbsR12-Comp strains at a MOI of 10. At 1dpi, infections were stopped by washing cells three times for 5 min in pre-warmed 1X PBS, after which fresh medium was added. At 3 or 7 dpi, the growth medium was removed and cells were fixed with ice-cold 100% methanol for 5 min at room temperature. Cells were washed three times for 5 min each with ice-cold 1X PBS, blocked for 1 h at room temperature with a 2% BSA solution in 1X PBS, and then incubated with anti-Com1 (1:1000) and anti-LAMP1 (1:50, H4A3 was deposited to the Developmental Studies Hybridoma Bank by August, J.T / Hildreth, J.E.K) antibodies for 2 h. Cells were washed and incubated with Alexa Fluor 488 (1:200, Thermo Scientific) and goat anti-mouse rhodamine antibodies (1:200, Thermo Scientific) along with DAPI (300 nM, Thermo Scientific) for 1 h. Cells were then washed three times for 5 min each in ice-cold 1X PBS and immediately imaged. Images were processed with Fiji [194]. Cell Profiler was used to measure CCV areas, as previously described [197]. Measurements were taken from 46 individual images of random fields of view spanning three different experiments for each C. burnetii strain.

**Protein synthesis, purification and antibody production**

Recombinant *Coxiella* RNase III was synthesized from a previously-generated pQE30 construct and purified as before [102]. *C. burnetii carA* and *metK* genes were cloned in-frame into pQE30 (Qiagen) and the resulting N-terminal His<sub>6</sub>-tagged proteins synthesized
and purified as previously described for *C. burnetii* RNA helicase [101]. Purified recombinant CarA and MetK proteins were submitted to General Bioscience, Inc., for rabbit polyclonal antibody production.

**Western blot**

Western blots were performed as previously described [97] with minor modifications. MB-WT, MB-*cbsR12*, and MB-*cbsR12*-Comp strains were grown to mid-log phase (4 d for MB-WT and MB-*cbsR12*-Comp, 6 d for MB-*cbsR12*) in ACCM-2. Proteins at 30 µg (CarA blot) or 60 µg (MetK blot) were resolved on 10% - 20% acrylamide gradient Tris-Glycine SDS-PAGE gels. Duplicate gels were run in parallel for Coomassie brilliant blue staining (CBB; 0.1% (w/v) in 50% methanol, 7% (v/v) acetic acid) in order to present a loading control. Blots were incubated with primary antibody solution (1X PBS pH7.4, 0.3% (v/v) Tween-20, 1:500 CarA / 1:5000 MetK primary antibody) for 2 h with rocking at room temperature. Blots were washed five times for 5 min in 1X PBS, then incubated for 1 h in secondary antibody solution (1X PBS + 1:2000 goat anti-rabbit::HRP). Blots were again washed five times for 5 min in 1X PBS and immediately developed using SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific) according to the manufacturer's protocol. Imaging was performed on a ChemiDoc XRS+ System (Bio-Rad).

**E. coli biofilm induction assay**

*E. coli* biofilm induction assays were performed as previously described [31]. 96-well plates were inoculated with an overnight culture of either pBEST or pBEST + *carA* 5’ UTR + *cbsR12* *E. coli* strains (see Figure S3.10). Cultures were allowed to grow for 3 h until induction with 1mM IPTG, after which cultures were allowed to grow an additional
21 h before subsequent staining with crystal violet. The average OD$_{570}$ readings of 10 wells were obtained by spectrophotometry.

**Generation of a CbsR12-complemented strain**

MB-$cbsR12$ was complemented as previously described, with modifications [198]. Briefly, wild-type $cbsR12$ along with 100 bp of flanking sequences were PCR-amplified using primers containing EcoRI and BamHI restriction sites. The amplicon was cloned into compatible restriction sites of pMini-Tn7-KAN by standard protocol [199]. The resulting plasmid was transformed into electrocompetent *E. coli* PIR1 cells for propagation. The pMini-Tn7-CbsR12-KAN plasmid (20 µg), along with 10 µg of a second plasmid containing the transposase, pMini-TnS2-ABCD, were transformed into MB-$cbsR12$ in a single electroporation reaction (25 kV, 500 ohms, 25 µF). Electroporated cells were allowed to recover 5 d in ACCM-2 supplemented with 1% FBS, then dilutions were plated onto ACCM-2 agar containing kanamycin (375 µg/mL). Isolated colonies were picked and re-cultured on ACCM-2 agar plates for several rounds. Colony-PCR was used to screen for the MB-$cbsR12$-Comp strain, and the location of the $cbsR12$ cassette was determined by PCR and Sanger automated sequencing. qPCR of MB-$cbsR12$-Comp genomic DNA utilizing primers specific to $cbsR12$ was used to ensure that a single Tn insertion event occurred.

**Data availability**

The sequencing reads from the Crosslink-Seq experiments are available at the NCBI sequencing read archive (accession number: PRJNA522455).

**Acknowledgments**
We thank Paul Beare for his generous donation of *E. coli* PIR1 cells, pMini-TnS2-ABCD plasmid, and the pMini-Tn7-KAN plasmid. We would thank Jenny Wachter for her contribution of the TPM calculator and Linda D. Hicks for her excellent technical assistance. The authors also wish to acknowledge the following grant support: RR was supported by NIH grants AI123464, AI126385 and AI133023. MFM was supported by NIH grants AI128575, AI123293 and AI119798. SW was supported by a research grant from the Montana Academy of Sciences.

**Supplementary Material**

**Figure S3.1: CbsR12 is processed by RNase III.** (A). CbsR12 secondary structure as predicted by mFold. Nucleotide 1 was determined to be the TSS for the full-size transcript by 5' RACE. Red asterisks indicate apparent alternative TSSs by 5' RACE. Dotted line indicates the putative RNase III processing area. Blue solid lines indicate consensus CsrA-binding sites. (B). RNase III assay of in vitro-transcribed CbsR12 with the *C. burnetii* IVS RNA as a positive control. Results from treatment with *E. coli* (Ec) RNase III (New England BioLabs), recombinant *C. burnetii* (Cb) RNase III or no-
enzyme controls are shown. Arrows indicate RNase III-processed (blue) and unprocessed (red) CbsR12 RNA.

Figure S3.2: Location of the MB-cbsR12 transposon insertion to inactivate cbsR12. (A). The cbsR12 gene and promoter elements are highlighted by the indicated colors, while the location of the Himar Tn insertion producing the MB-cbsR12 strain is marked by a black arrow. Red arrows denote primer-binding sites for PCR confirmation of the lesion (forward and reverse primers above and below their annealing sequences, respectively). (B). PCR products confirming Tn insertion in cbsR12 of MB-cbsR12 (red...
arrow) by loss of the ~250 bp amplicon and reintroduction of \textit{cbsR12} in MB-\textit{cbsR12}-Comp (blue arrow). (C). Copy number qPCR analysis confirming a single additional insertion of \textit{cbsR12} in the MB-\textit{cbsR12}-Comp strain. Values represent the means ± standard error of means (SEM) of three independent determinations.

**Figure S3.3: Additional biological replicates for ACCM-2 and THP-1 growth curves.** Growth curves for MB-WT, MB-\textit{cbsR12}, and MB-\textit{cbsR12}-Comp strains in ACCM-2 (A, B) or THP-1 cells (C, D) as determined by qPCR. The 0dpi time point refers to the inoculum. Values represent means ± standard error of means (SEM) of three technical replicates.
Figure S3.4: CbsR12 competitively binds *carA*, *metK*, and *cvpD* transcripts in a dose-dependent manner. RNA-RNA EMSAs showing hybridization reactions between biotin-labeled CbsR12 (Bio-CbsR12) and an *in vitro*-transcribed segment of *carA* (A), *metK* (B), or *cvpD* (C). Anti-CbsR12 represents a positive control consisting of a transcript equal in size but antisense to the CbsR12 transcript. A cold-chase sample containing Bio-CbsR12 + un-labeled CbsR12 + CarA/MetK/CvpD shows competitive (specific) binding relative to Bio-CbsR12 plus target alone, while increasing the dose of *carA/metK/cvpD* transcript (from 2 nM to 10 nM) increases the amount of retarded sample signal on the blot. Arrows indicate un-bound Bio-CbsR12 (blue) and Bio-CbsR12 bound to its RNA targets (red).
Figure S3.5: Artemis views of CbsR12 binding to carA, metK, ahcY, and cvpD transcripts. Artemis representation of Crosslink-Seq results for MB-WT. Red and blue lines represent the two biological replicates. (A). CbsR12 crosslinking with carA reads
(blue arrow). (B). CbsR12 crosslinking with *metK* reads (blue arrow) and *ahcY* reads (red arrow). (C). CbsR12 crosslinking with *cvpD* reads (blue arrow).
Figure S3.6: CbsR12 downregulates the quantity of transcripts arising from the 5' end of cvpD in LCVs from infected THP-1 cells. (A). cvpD gene sequence from the 5' TSS to the predicted downstream alternative start codon. Indicated colors highlight the TSSs, the CbsR12-binding site, the putative downstream promoter, putative RBSs, and start codons. Red arrows show primer annealing regions for qRT-PCR (forward and reverse primers above and below their respective annealing sequences). (B). Representation of CbsR12 binding to the cvpD transcript as determined by IntaRNA with base numbers indicated. The top strand in the model represents the metK transcript sequence, while the bottom strand represents the complementary CbsR12 sequence. (C). qRT-PCR of the 5' end of cvpD from MB-WT, MB-cbsR12, and MB-cbsR12-Comp LCVs (3dpi) and SCVs (7dpi) infecting THP-1 cells. Values represent means ± standard error of means (SEM) of three independent determinations (** = P < 0.01, one-way ANOVA, *** = P < 0.001, one-way ANOVA).
Figure S3.7: Coomassie blue-stained SDS-PAGE gels corresponding to Figure 3.9 to demonstrate loading consistency. (A). Proteins (30 µg total) from MB-WT, MB-cbsR12, and MB-cbsR12-Comp LCVs (mid-log phase; 96h for MB-WT and MB-cbsR12-Comp and 144h for MB-cbsR12) grown in ACCM-2 were resolved on a 10-20% acrylamide gradient SDS-PAGE gel and stained with Coomassie brilliant blue R. (B). Proteins (60 µg total) from MB-WT, MB-cbsR12, and MB-cbsR12-Comp LCVs (mid-log phase; 96h for MB-WT and MB-cbsR12-Comp and 144h for MB-cbsR12) grown in ACCM-2 were resolved on a 10-20% acrylamide gradient SDS-PAGE gel and stained with Coomassie brilliant blue R.
Figure S3.8: CbsR12 in *E. coli* leads to an autoaggregative phenotype and biofilm formation. (A). Overnight cultures of *E. coli* Top10 F’ harboring pBEST + carA5’UTR or pBEST + carA5’UTR + cbsR12 were inoculated into 3 mL LB supplemented with ampicillin (100 µg/mL), grown for 2 h at 37⁰ C with shaking, then induced with 1 mM IPTG for 3 h before photography. The red arrow indicates autoaggregation of *E. coli* in the presence of CbsR12. (B). *in vitro* biofilm formation assay of *E. coli* Top10 F’ harboring pBEST or pBEST + carA5’UTR + cbsR12. Crystal violet staining is indicative of adherence due to biofilm induction. Values represent the average OD570 readings of 10 wells ± standard error of means (SEM) of three independent determinations (* = P < 0.05, student’s t test, ** = P < 0.01, student’s t test).
**Figure S3.9: CbsR1 is an additional *C. burnetii* sRNA with RsmY/Z-like characteristics.** (A). The *cbsR1* gene, predicted promoter elements, and putative LetA-binding site are highlighted by the indicated colors. (B). CbsR1 secondary structure as predicted by mFold. Nucleotide 1 was predicted to be the TSS for the full-size transcript by analysis of RNA-Seq datasets. Blue solid lines indicate putative CsrA-binding sites.
Figure S3.10: Strains, plasmids, and primers used in the study.
Chapter 4

Novel small RNAs expressed by *Bartonella bacilliformis* under multiple conditions reveal potential mechanisms for persistence in the sand fly vector and human host

As published in:


Abstract

*Bartonella bacilliformis*, the etiological agent of Carrión’s disease, is a Gram-negative, facultative intracellular alphaproteobacterium. Carrión’s disease is an emerging but neglected tropical illness endemic to Peru, Colombia, and Ecuador. *B. bacilliformis* is spread between humans through the bite of female phlebotomine sand flies. As a result, the pathogen encounters significant and repeated environmental shifts during its life cycle, including changes in pH and temperature. In most bacteria, small non-coding RNAs (sRNAs) serve as effectors that may post-transcriptionally regulate the stress response to such changes. However, sRNAs have not been characterized in *B. bacilliformis*, to date. We therefore performed total RNA-sequencing analyses on *B. bacilliformis* grown *in vitro* then shifted to one of ten distinct conditions that simulate various environments encountered by the pathogen during its life cycle. From this, we identified 160 sRNAs significantly expressed under at least one of the conditions tested. sRNAs included the highly-conserved tmRNA, 6S RNA, RNase P RNA component, SRP
RNA component, \textit{ffH} leader RNA, and the alphaproteobacterial sRNAs \textit{ar45} and \textit{speF} leader RNA. In addition, 153 other potential sRNAs of unknown function were discovered. Northern blot analysis was used to confirm the expression of eight novel sRNAs. We also characterized a \textit{Bartonella bacilliformis} group I intron (BbgpI) that disrupts an un-annotated tRNA\textsubscript{CCU}^{\text{Arg}} gene and determined that the intron splices \textit{in vivo} and self-splices \textit{in vitro}. Furthermore, we demonstrated the molecular targeting of \textit{Bartonella bacilliformis} small RNA 9 (BbsR9) to transcripts of the \textit{ftsH}, \textit{nuoF}, and \textit{gcvT} genes, \textit{in vitro}.

\textbf{Introduction}

Bacteria often utilize sRNAs to rapidly and efficiently regulate gene products involved in multiple biological processes. sRNAs are small (< 500 nts) non-coding transcripts that typically serve to up- or down-regulate translation of proteins by binding to the respective mRNA in a \textit{cis} or \textit{trans} fashion [reviewed in [1]]. This fine-tuning of protein production can enhance tolerance to stressors, including temperature [8] and pH [200]. To our knowledge, sRNA research in \textit{Bartonella} is represented by a single report on \textit{B. henselae} [201]. We therefore utilized total RNA-Sequencing (RNA-Seq) to interrogate \textit{B. bacilliformis} transcriptomes to identify sRNAs expressed under a variety of conditions, including temperatures and pH levels consistent with the sand fly vector and human host. In doing so, we discovered 153 novel sRNAs expressed under at least one of the conditions tested. Furthermore, we characterized two of the sRNAs. The first RNA is a group I intron related to similar elements found in other alphaproteobacteria, while the
other is a novel *Bartonella*-specific sRNA expressed only under conditions that simulate the sand fly vector.

**Materials and Methods**

**Bacterial culturing**

Bacterial strains, primers, and plasmids utilized in this study are described in Table S4.1. *B. bacilliformis* strain KC583 (passages #4-7) was cultivated on HIBB plates, comprised of Bacto heart infusion agar (Becton Dickinson; Franklin Lakes, NJ) containing 4% defibrinated sheep blood and 2% sheep serum (Quad Five, Ryegate, MT), by volume, as previously described [131]. Following 4 d of growth, 4 confluent *B. bacilliformis* plates per biological replicate were either shifted to different temperatures for 2 h, harvested and shifted to different pH levels in an HIBB liquid medium for 2 h, harvested and shifted to a human blood sample for 2 h, or harvested and used to infect cultured human umbilical vein endothelial cells (HUVECs; PCS-100-013; American Type Culture Collection; Manassas, VA) for 24 h. *Escherichia coli* (TOP10) was grown for 16 h at 37°C with shaking in lysogeny broth (LB), or on LB plates, supplemented with kanamycin (25 μg/ml) and ampicillin (100 μg/ml), when required.

**HUVEC culturing and infection**

HUVECs were cultured and maintained as previously described [131]. *B. bacilliformis* infections were carried out for 24 h after which the medium was removed and cells were treated with gentamicin (10 μg/ml) for 1 h. Remaining viable extracellular *B. bacilliformis* cells were removed by washing 5 times for 10 min with phosphate-buffered
saline (PBS; pH 7.4) solution. Finally, cells were harvested into TRI Reagent (Ambion; Austin, TX), as previously described for infected Vero cells [202].

**Human blood infection**

Blood was drawn into vials containing sodium citrate to prevent coagulation. 1-ml aliquots were dispensed into fresh tubes, after which the lids were replaced with gas-permeable membranes. Blood samples were equilibrated at 37⁰ C (HB37 samples) or 37⁰ C in a blood-gas atmosphere (HBBG samples) for 1 h. Four HIBB plates of confluent *B. bacilliformis* for each equilibrated blood vial were harvested into PBS, pelleted at 16,000 x g for 5 min at 4⁰ C and washed twice in PBS with identical centrifugation steps. Cell pellets were resuspended into 300 μl equilibrated blood, then dispensed back into the corresponding tube. The tubes were incubated at the appropriate condition for 2 h, then 1 ml RNAlater solution (Thermo Fisher; Waltham, MA) was immediately added. Total RNA extraction was done as described below.

**Total RNA/genomic DNA isolation and preparation for RNA-Seq**

Upon shifting *B. bacilliformis* for the designated time periods, cells were either harvested directly into one volume of RNAlater solution (Thermo Fisher) or centrifuged at 10,000 x g at room temperature for 2 min, after which the pellet was resuspended in a volume of RNAlater. The cells were incubated at room temperature for 1 h then frozen at -80⁰ C for ≥ 2 h. The cells were thawed, centrifuged at 10,000 x g at 4⁰ C for 10 min, and resuspended in 1 ml of TRI Reagent (Sigma-Aldrich; St. Louis, MO). The cells were incubated at room temperature for 1 h then frozen at -80⁰ C for ≥ 2 h. Finally, cells were thawed and total RNA and genomic DNA isolation were done as previously described [202]. Total RNA pools from human blood infections were globin-depleted using a
GLOBINclear kit (Ambion) according to manufacturer’s specifications. HUVE, HB37, and HBBG samples were enriched for bacterial RNA using a MICROBEnrich kit (Ambion). RNA (1 μg) from three independent biological replicates of each condition was sent to the Yale Center for Genomic Analysis (PI25, PI30, PI37, pH06, pH07, and pH08 samples) or GENEWIZ (PIBG, HUVE, HB37, and HBBG samples) for bacterial rRNA depletion, stranded-library preparation, and HiSeq2500 (Illumina; San Diego, CA) 2x150 bp sequencing.

**Data analysis**

Raw reads were quality filtered and aligned as previously described [202]. Briefly, raw fastq files were concatenated, quality filtered with the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), and then clipped, aligned, and filtered with Neson version 0.128 tools (http://www.vicbioinformatics.com/software.neson.shtml). TPM were calculated using a custom Python script that can be accessed through GitHub (https://github.com/shawachter/TPM_Scripts). Stranded alignments were separated using SAMtools [203] and visualized using the Artemis genome browser [147].

sRNA identification was performed using the Artemis genome browser. RNA peaks were manually curated from IGRs and protein-coding gene regions. A read threshold for sRNA expression was devised for each condition tested based on reads that aligned to the *rpoD* gene (RpoD sigma factor), since TPM data suggest this gene is consistently expressed across all conditions. Using this method, putative sRNAs were identified, base ranges approximated, and sRNAs further characterized with putative promoters by manual searches using the conserved alphaproteobacterial sigma-70 promoter element, CTTGAC-N$_{17}$-CTATAT [204]. Rho-independent terminators were
identified using ARNold terminator prediction software (http://rna.igmors.unipud.fr/toolbox/arnold/).

Since TPM calculations were done in the context of the total transcriptome and were not strand-specific, it was necessary to further refine TPM values for cis-antisense sRNAs. This was done by considering the TPM value and proportion of the protein-coding gene to which the sRNA was antisense and subtracting the gene’s approximate TPM contribution with the following formula:

\[
sRNA \text{ Adjusted TPM} = sRNA \text{ Total TPM} - \left[ \frac{\text{Gene Total TPM}}{\text{Gene Length}} \right] \left( \frac{\text{Gene sRNA overlap Length}}{\text{Gene Length}} \right)
\]

This was accomplished with a custom python script located in a GitHub repository (https://github.com/shawachter/TPM_Scripts). After this calculation, sRNAs whose TPMs were < 300 were considered not expressed for the purposes of the UpSet plot. This TPM value was chosen because it most accurately reflected the read threshold used in the initial manual sRNA search. All TPM values were included in the heatmap and determination of condition-specific sRNAs. Differentially-expressed sRNAs were determined by featureCounts [192] and the DESeq2 package of R version 3.4.4 [193]. For DESeq2 analysis, the p-value distribution of the significantly differentially expressed genes (DEGs) was re-sampled using fdrtools in order to more accurately achieve the desired null distribution [205]. This effectively made the analysis more stringent by providing fewer significant DEGs.

Infection-specific and sand fly-specific sRNAs were identified based on expression patterns as described in Results. The IntaRNA 2.0 sRNA target prediction algorithm was used to determine potential genes regulated by the sRNAs [43]. Only mRNA targets with
a predicted IntaRNA 2.0 p-value < 0.01 were included in the potential targets list. Further targets with FDR values of < 0.05 were given special indications since these predicted bindings were considered especially strong. GO enrichment was performed utilizing the biobam Blast2GO program in the OmicsBox program suite (https://www.blast2go.com/blast2go-pro/download-b2g) using functional annotation of the B. bacilliformis KC583 genome as the background [206]. KEGG enrichment was performed using DAVID Bioinformatics Resources [207].

Figures were made using R version 3.4.4 and various Bioconductor packages including UpSetR [208], gplots (https://cran.r-project.org/web/packages/gplots/index.html), ggplot2 [209], and fdrtool [205]. Raw PNG images were modified into figures using Inkscape (https://inkscape.org/release/inkscape-0.92.4/) and Gimp (https://www.gimp.org/downloads/).

Identification of transcription start sites

5' RACE analyses of BbgpI and BbsR9 were performed using total RNA from B. bacilliformis shifted to liquid medium at pH 7 using a 5' RACE System kit (Invitrogen; Carlsbad, CA) according to manufacturer’s protocols and with gene-specific primers (Table S4.1). Resulting PCR products were cloned into pCR2.1-TOPO as instructed (Invitrogen), after which six arbitrary clones were sequenced with M13 universal primers by Sanger automated sequencing.

Northern blots

Northern blot analyses were carried out using total RNA extracted from B. bacilliformis under the noted conditions. Northern blot probes were synthesized in vitro by engineering probe-specific PCR primers to contain a T7 promoter then utilizing a MAXIscript T7
Transcription kit (Invitrogen) supplemented with 0.5 mM Bio-16-UTP (Invitrogen). *B. bacilliformis* total RNA (2 μg) was resolved on a 1% denaturing agarose gel for 130 min at 57 V in 1X MOPS running buffer (Quality Biological; Gaithersburg, MD). The gel was washed in nuclease-free H2O for 10 min, followed by another wash in 20X SSC buffer (3M NaCl, 0.3M sodium citrate, pH 7.0) for 15 min. RNA was transferred overnight to a BrightStar-Plus nylon membrane (Ambion) in 20X SSC via upward capillary transfer. RNA was crosslinked to the membrane using a GS Gene Linker UV chamber (Bio-Rad; Hercules, CA) at 150 mJ. Membrane pre-hybridization and probe hybridization were done with a North2South Chemiluminescent Hybridization and Detection Kit (Thermo Fisher) according to manufacturer’s protocol. 50 ng of the appropriate *in vitro*-transcribed biotin-labeled probe was hybridized to the membrane at 67°C overnight. Membranes were washed 3 times for 15 min at 67°C in 1X Hybridization Stringency Wash Buffer (Thermo Fisher), developed, and imaged with a ChemiDoc XRS+ system (Bio-Rad).

**qRT-PCR**

qRT-PCR was done on cDNA synthesized from 16 ng *B. bacilliformis* total RNA (for each 25 μl reaction) collected from various conditions using the Luna Universal One-Step RT-qPCR kit (New England BioLabs; Ipswich, MA) according to the manufacturer. *B. bacilliformis* total RNA was serially diluted and used as a standard curve, while primers targeting the *rpoD* housekeeping gene were used for normalization of gene expression between conditions. qRT-PCR was performed on a CFX Connect Real-Time System (Bio-Rad). cDNA from sRNAs of interest was analyzed for copy number, then divided by the copy number from the *rpoD* gene to achieve the sRNA transcripts / *rpoD*
transcript values.

**Mutagenesis and RNA-RNA EMSAs**

Mutagenesis of gcvT, nuoF, and ftsH target sequences was carried out *in vitro* using a Q5 mutagenesis kit (New England BioLabs) with specified primers (Table S4.1). Primers engineered with a T7 promoter sequence were used to amplify the gcvT (-100 to +100), nuoF (-86 to +100), ftsH (-100 to +105), RS02100 (-76 to +100), trmD (-50 to +100), and hflK (-70 to +100) target sequences, where nucleotide +1 represents the first nucleotide of the protein-coding sequence. PCR products were cloned into pCR2.1-TOPO as instructed (Invitrogen). Resulting plasmid DNA was used as the template for Q5 mutagenesis. Q5 clones were sequenced, re-amplified with T7-engineered primers, and *in vitro* transcribed using the MAXIscript T7 Transcription kit (Invitrogen) with or without 0.5 mM Bio-16-UTP (Invitrogen), as required. Dose-dependent RNA-RNA EMSAs were performed as previously described [202] using 2 nM biotin-labeled BbsR9 and varying concentrations of *in vitro*-transcribed, unlabeled target RNA.

**Data availability**

Aligned sequencing reads (BAM files) from all RNA-Seq experiments are available at the Sequencing Read Archive database (accession number PRJNA647605).

**Ethics statement**

The Institutional Biosafety Committee at the University of Montana granted approval for the experimental use of human blood (IBC 2019-05). Formal consent was obtained in verbal form from the blood donor (co-author MFM).

**Results**

128
Identification of *B. bacilliformis* sRNAs

We first analyzed the total transcriptomic profiles of *B. bacilliformis* following a timed shift from normal culture conditions (4-d incubation on HIBB plates at 30⁰ C) to various *in vitro* conditions that mimic the sand fly vector and human host (*Table 4.1*).

Specifically, we controlled for several environmental variables, including temperature, pH, solid vs. liquid media, and the presence of a human blood-gas atmosphere. Following quality control analysis of the resulting RNA-Seq datasets and correlation of variation analysis (*Table S4.2*), we discarded replicates that did not correlate well with others from the same condition. A principle component analysis (PCA) plot of the remaining RNA-Seq datasets confirmed statistical clustering of biological replicates (*Figure S4.1*).

**Table 4.1:** Conditions used to prepare *B. bacilliformis* cultures for RNA-Seq experiments.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Medium</th>
<th>Designation</th>
<th>Shift Time</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH 7.4, 25⁰ C</td>
<td>HIBB plates</td>
<td>Pl25</td>
<td>2 hours</td>
<td>Sand fly ambient temperature</td>
</tr>
<tr>
<td>pH 7.4, 30⁰ C</td>
<td>HIBB plates</td>
<td>Pl30</td>
<td>2 hours</td>
<td>Sand fly ambient temperature</td>
</tr>
<tr>
<td>pH 7.4, 37⁰ C</td>
<td>HIBB plates</td>
<td>Pl37</td>
<td>2 hours</td>
<td>Human host</td>
</tr>
<tr>
<td>pH 7.4, 37⁰ C with blood gas</td>
<td>HIBB plates</td>
<td>PlBG</td>
<td>2 hours</td>
<td>Human host</td>
</tr>
<tr>
<td>pH 6.0, 30⁰ C</td>
<td>HIBB liquid</td>
<td>pH06</td>
<td>2 hours</td>
<td>Sand fly post-blood meal</td>
</tr>
<tr>
<td>pH 7.4, 30⁰ C</td>
<td>HIBB liquid</td>
<td>pH07</td>
<td>2 hours</td>
<td>Human host / sand fly blood meal mid-digestion</td>
</tr>
<tr>
<td>pH 8.2, 30⁰ C</td>
<td>HIBB liquid</td>
<td>pH08</td>
<td>2 hours</td>
<td>Sand fly initial blood meal</td>
</tr>
<tr>
<td>pH 7.4, 37⁰ C</td>
<td>HUVECs in</td>
<td>HUVE</td>
<td>24 hours</td>
<td>Human</td>
</tr>
</tbody>
</table>
HIBB, Bacto heart infusion blood agar containing 4% defibrinated sheep blood and 2% sheep serum (vol/vol); HUVECs, human umbilical vein endothelial cells; EGM-Plus (Lonza), endothelial cell growth medium containing 2% fetal bovine serum and bovine brain extract.

* Blood gas is comprised of 5% CO$_2$, 2.5% O$_2$, and 92.5% N$_2$ at 100% humidity to simulate human blood.

Next, we visualized alignments for each RNA-Seq dataset and manually curated transcript peaks that could correspond to novel sRNAs. Peaks were found in IGRs, antisense to annotated genes (cis-anti) or as leader RNAs in 5' UTRs of annotated genes. The peaks were further refined based on proximity to neighboring peaks and a threshold read coverage based on expression of a housekeeping gene, rpoD (encoding sigma factor RpoD). We discovered 160 potential sRNAs, including seven highly-conserved sRNAs of other bacteria/alphaproteobacteria (**Table S4.3**). Of the 153 other potential sRNAs, 81 were located antisense to annotated genes, 57 were encoded in IGRs, and the remaining 15 were potential leader RNAs. Leader RNAs were included in the study because further analysis would be needed to determine if they are true leader RNAs, stand-alone sRNAs, or perhaps both. We also identified putative promoter elements for each identified sRNA based on approximated TSS’s. Next, we constructed an UpSet plot to visualize the
numbers of significantly expressed sRNAs shared between various combinations of conditions (Figure 4.1) [210]. Results of this analysis suggested that, while 19 of the 160 identified sRNAs were expressed regardless of circumstance, the majority of sRNAs were expressed under specific conditions. Following this, we calculated TPM for each sRNA under all ten conditions in the context of the total transcriptomes (Table S4.4). TPM is a normalized measure of gene expression, and although it is not always appropriate to compare TPM values across different RNA-Seq experiments [168], we constructed a heatmap to get a broad sense of sRNA expression patterns (Figure S4.2). These results revealed three distinct clusters of conditions with similar sRNA expression patterns and allowed us to identify interesting sRNAs for further characterization.
Figure 4.1: Most *B. bacilliformis* sRNAs are expressed under specific conditions. An UpSet plot is shown and displays the number of sRNAs shared among various combinations of conditions tested. The bar graph to the left indicates the quantity of sRNAs with a TPM >300 under the conditions shown. The connected nodes indicate shared conditions giving rise to the number of sRNAs expressed as indicated by the bar graph at the top.

Verification of select *B. bacilliformis* sRNAs
Of the 160 putative sRNAs discovered in *B. bacilliformis*, we assigned name designations to 22 of them based on appraisal of general relevance, including six widely-conserved sRNAs, such as BbtmRNA (*B. bacilliformis* tmRNA), and 15 novel *Bartonella*-specific sRNAs. We also gave a name designation to the *B. bacilliformis* group I intron (BbgpI), a group I intron with related elements previously identified, but not characterized, in other alphaproteobacteria [211]. To corroborate RNA-Seq expression results and verify sRNA expression, we chose eight novel sRNAs at random and conducted Northern blot analyses. Results of the Northern blots confirmed the expression of all eight sRNAs (Figure 4.2).

**Figure 4.2: Northern blot analyses confirm expression of eight putative *B. bacilliformis* sRNAs.**

*bacilliformis* sRNAs. Eight separate Northern blots were run under identical experimental conditions (see Chapter 4: Materials and Methods). RNA ladders from the respective blots were aligned with each other for presentation of the resolved total RNA. sRNA designations are shown above each blot. Exposure times (minutes, m; seconds, s) and origin of the RNA are indicated below each blot.
We analyzed the differential expression of *B. bacilliformis* genes across the ten tested conditions by performing relevant pairwise comparisons (**Table 4.2**) using the DESeq2 package in R version 3.4.4 [193]. For this analysis, transcriptomes from all ten conditions were compared simultaneously, while specific, relevant pairwise comparisons were made (**Table 4.2**). Results showed the greatest number of significant differentially expressed sRNAs by comparing solid-to-liquid media and host cell types. Other sRNA candidates were also found to be differentially regulated by these comparisons (see **Table S4.3**). We then utilized quantitative reverse transcription PCR (qRT-PCR) to validate the DESeq2 results. In doing so, we confirmed eleven sRNAs to be significantly differentially expressed under the relevant conditions (**Figure S4.3**).

**Table 4.2:** DESeq2 comparisons made.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Controlled Conditions</th>
<th>No. sRNA DEGs</th>
</tr>
</thead>
<tbody>
<tr>
<td>P130&lt;sup&gt;a&lt;/sup&gt; vs. PI25</td>
<td>Temperature</td>
<td>0</td>
</tr>
<tr>
<td>P130&lt;sup&gt;a&lt;/sup&gt; vs. PI37</td>
<td>Temperature</td>
<td>2</td>
</tr>
<tr>
<td>PI25&lt;sup&gt;a&lt;/sup&gt; vs. PI37</td>
<td>Temperature</td>
<td>0</td>
</tr>
<tr>
<td>P130&lt;sup&gt;a&lt;/sup&gt; vs. pH07</td>
<td>Solid/liquid media</td>
<td>12</td>
</tr>
<tr>
<td>pH07&lt;sup&gt;a&lt;/sup&gt; vs. pH08</td>
<td>pH level</td>
<td>0</td>
</tr>
<tr>
<td>pH07&lt;sup&gt;a&lt;/sup&gt; vs. pH06</td>
<td>pH level</td>
<td>0</td>
</tr>
<tr>
<td>pH06&lt;sup&gt;a&lt;/sup&gt; vs. pH08</td>
<td>pH level</td>
<td>7</td>
</tr>
<tr>
<td>PI37&lt;sup&gt;a&lt;/sup&gt; vs. PIBG</td>
<td>Blood gas</td>
<td>1</td>
</tr>
<tr>
<td>PIBG&lt;sup&gt;a&lt;/sup&gt; vs. HBBG</td>
<td>Solid/liquid media, human/sheep blood</td>
<td>6</td>
</tr>
<tr>
<td>PIBG&lt;sup&gt;a&lt;/sup&gt; vs. HUVEC</td>
<td>Solid/liquid media, cell type</td>
<td>18</td>
</tr>
<tr>
<td>pH07&lt;sup&gt;a&lt;/sup&gt; vs. HUVEC</td>
<td>Temperature, cell type</td>
<td>6</td>
</tr>
<tr>
<td>pH07&lt;sup&gt;a&lt;/sup&gt; vs. HB37</td>
<td>Temperature, human/sheep blood</td>
<td>3</td>
</tr>
<tr>
<td>HBBG&lt;sup&gt;a&lt;/sup&gt; vs. HUVEC</td>
<td>Cell type</td>
<td>17</td>
</tr>
</tbody>
</table>

DEGs, significant differentially expressed genes.

<sup>a</sup>Reference dataset
**Condition-specific sRNAs target mRNAs enriched in specific pathways**

We grouped several sRNAs based on their expression patterns across the ten conditions tested by using heatmap comparisons combined with data from the DESeq2 analysis. For example, multiple sRNAs were significantly and strictly expressed under conditions used to simulate or actually infect human cells (i.e., PlBG, HUVEC, HB37, and HBBG; see Table 4.1). These sRNAs were classified as human “infection-specific” based on their restricted upregulation (defined by TPM greater than the mean TPM plus one standard deviation) in at least two of these four conditions. Based on this definition, we identified 24 infection-specific sRNAs (see Table S4.4).

We were also curious to determine if the predicted mRNA targets of the infection-specific sRNAs significantly corresponded to particular gene classifications to provide clues regarding their upregulation under infection-specific conditions. Using gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses and the IntaRNA 2.0 sRNA target prediction program [43], we determined that the predicted mRNA targets of these sRNAs (Table S4.5) were enriched for several GO terms, including protein/amide transport and nucleotidyltransferase activity (Figure 4.3A). The pool of mRNA targets was also enriched for the glycerophospholipid/glycerolipid metabolism and nucleotide excision repair KEGG pathways (Figure 4.3B).
Figure 4.3: Condition-specific sRNA targets are enriched in several GO terms and KEGG pathways. A) Faceted bar graph of GO enrichment terms for infection-specific and sand fly-specific sRNA targets. Height of the bars indicates the number of sRNA targets containing that GO term, while the color displays the significance of enrichment. B) Faceted dot plot of KEGG enrichment terms for infection-specific and sand fly-specific sRNA targets. The enrichment score refers to the ratio of the number of gene targets corresponding to a particular pathway to the total number of genes in that pathway. Dot colors represent significance (p-value) of enrichment for that particular KEGG pathway.

We also determined that eight sRNAs were only expressed under conditions simulating the sand fly vector (i.e., Pl25, pH06, pH07, and pH08; see Table 4.1). Despite the fact that Pl25 clustered separately from pH06, pH07, and pH08 on the heatmap (Figure S4.2), we included it in the sand fly-specific conditions due to upregulation of several sRNAs under both pH06/pH08 and Pl25 conditions (BB026-1, BB103-2, BB103-3, and BB124; Table S4.4). These and other sRNAs were classified as “sand fly-specific sRNAs” based on their restricted upregulation in at least two of these four conditions. The predicted mRNA targets of the sand fly-specific sRNAs (Table S4.6) were enriched for the flavin adenine dinucleotide (FAD) binding GO term (Figure 4.3A) and the amino acid biosynthesis KEGG pathway (Figure 4.3B).

**BbgpI is a group I intron that splices in vivo and self-splices in vitro**

BB009 was initially identified as a sRNA of interest based on its high expression across multiple conditions (Table S4.4). A BLAST search of the BB009 gene sequence
(hereafter referred to as BbgpI) showed that it was highly homologous to a group I intron conserved in several other alphaproteobacteria and encoded in host tRNA_{CCU}^{Arg} genes [211]. Since *B. bacilliformis* has no annotated tRNA_{CCU}^{Arg} gene, we initially assumed that BbgpI was encoded in an IGR. However, such a location would be novel for group I introns, which are selfish genetic elements found in tRNA, rRNA, and rarely, protein-coding genes (reviewed in [74]).

To address this discrepancy, 5' RACE was utilized to determine the 5' end of the putative spliced-out RNA segment. From these results, we determined that BbgpI was flanked by CCT DRs, identical to those produced by the tRNA_{CCU}^{Arg} alphaproteobacterial group I intron (Figure S4.4A) [211]. Also, the predicted secondary structure of BbgpI possessed conserved group I intron stem structures (Figure S4.4B). Finally, we scanned the locus and flanking sequences with the tRNAscan-SE 2.0 web server and identified a tRNA_{CCU}^{Arg} gene (Figure S4.4C), but only when a sequence with BbgpI spliced out was used [212]. Taken together, these results suggest that BbgpI is a member of a conserved, alphaproteobacterial group I intron family and disrupts an unannotated tRNA_{CCU}^{Arg} gene (locus: c42404-42711) of *B. bacilliformis*.

Although homology and structural results suggested that BbgpI was a group I intron, it was unclear whether BbgpI was able to self-splice or whether a protein cofactor was required [211]. To address this question, we examined BbgpI’s ribozyme activity in *vitro*. Following *in vitro* transcription, cDNA synthesis, and PCR analysis with the primers shown in Figure S4.4A, we determined that BbgpI self-splices *in vitro* and is spliced *in vivo* (Figure 4.4), in keeping with other group I introns.
Figure 4.4: BbgpI self-splices in vitro and is spliced in vivo.

A) PCR analysis of an in vitro-transcribed (IVT) region of B. bacilliformis (Bb) genomic DNA (gDNA) containing BbgpI. Ethidium bromide-stained agarose (1%) gels are shown. PCR on the resulting cDNA using “Nested Primers” (see Figure S4.4A) produced unspliced DNA (450-bp band), a partially-spliced product of ~390 bp, plus a 218-bp band corresponding to the BbgpI flanking region where BbgpI self-spliced out (indicated by the red arrow). PCR on a Bb IVT RNA negative control did not produce product. B) As in A) but utilizing cDNA synthesized from B. bacilliformis total RNA using “Splice
Flank Primers” (see Figure S4.4A). A 308-bp amplicon was produced from gDNA, whereas a 76-bp band, corresponding to the BbgpI flanking region with BbgpI spliced out, was produced from cDNA generated from total RNA (indicated by the red arrow). PCR on a Bb total RNA negative control did not produce product.

BbsR9 is a sand fly-specific sRNA

BB092 (hereafter referred to as BbsR9) was initially identified as a sRNA of interest due to its restricted high-level expression under conditions that simulated the sand fly vector (Table 4.1, Table S4.4). In addition, BbsR9 was found to have well-defined, predicted sigma-70 promoter and Rho-independent terminator regions, and it was conserved among several other Bartonella spp. (Table S4.3). To elucidate the BbsR9 gene, the TSS was determined by 5' RACE. These results showed two possible sites with equal representation among the six clones sequenced (Figure S4.5A). We therefore wished to determine if two distinct transcripts of BbsR9 were made or if there was a single, dominant transcript for the sRNA. In addition, we wanted to confirm BbsR9 expression across the conditions examined. To this end, we set out to determine if BbsR9 would remain highly expressed in sheep blood shifted to 37⁰ C or if we would see a downregulation of the sRNA, as in B. bacilliformis shifted to human blood at 37⁰ C (HB37/HBBG; see Table S4.4). Northern blot analyses showed BbsR9 transcript and, together with the intensity of the signal, indicated a single, dominant TSS (bolded underlined in Figure S4.5A). Interestingly, we saw a distinct downregulation of BbsR9 when B. bacilliformis was shifted to sheep blood at 37⁰ C compared to 30⁰ C (Figure S4.5B). This decrease in RNA suggests that BbsR9 is primarily expressed under
conditions that simulate the sand fly vector and not the human host. Taken as a whole, results of the Northern blots and RNA-Seq suggest that both a liquid medium (PI30 vs. pH07; see Table S4.4) and a temperature below 37°C upregulate BbsR9.

**BbsR9 targets transcripts of ftsH, nuoF, and gcvT in vitro**

Since BbsR9 expression was restricted to sand fly-like conditions, we were interested in characterizing its mRNA targets to shed light on the sRNA’s role in regulation. To that end, we first utilized the TargetRNA2 [42], IntaRNA [43] and CopraRNA [44] algorithms to determine potential mRNA targets (Table 4.3). From these results, we selected transcripts of *ftsH*, *nuoF*, *gcvT*, *trmD*, *hflK*, and a predicted DNA response regulator (RS02100) as potential targets for characterization based on shared predictions between algorithms and the strength of predicted binding events.

**Table 4.3:** mRNA targets for BbsR9, as predicted by the indicated algorithms.

<table>
<thead>
<tr>
<th>Rank</th>
<th>TargetRNA2</th>
<th>IntaRNA</th>
<th>CopraRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>nuoF</em> (0.0001)</td>
<td><em>gcvT</em> (0.0033)</td>
<td>RS06660 (0.0013)</td>
</tr>
<tr>
<td>2</td>
<td>RS01360 (0.012)</td>
<td>RS01025 (0.0043)</td>
<td><em>nuoF</em> (0.0103)</td>
</tr>
<tr>
<td>3</td>
<td><em>czrB</em> (0.004)</td>
<td><em>trmD</em> (0.0060)</td>
<td>RS02895 (0.0113)</td>
</tr>
<tr>
<td>4</td>
<td><em>ftsE</em> (0.010)</td>
<td><em>ftsH</em> (0.0071)</td>
<td><em>gcvT</em> (0.0153)</td>
</tr>
<tr>
<td>5</td>
<td>RS05725 (0.012)</td>
<td><strong>RS02100</strong> (0.0090)</td>
<td>RS02955 (0.0232)</td>
</tr>
<tr>
<td>6</td>
<td><em>rplX</em> (0.033)</td>
<td>DUF475 (0.0125)</td>
<td><em>efp</em> (0.0286)</td>
</tr>
<tr>
<td>7</td>
<td><em>flgC</em> (0.043)</td>
<td>Pseudogene (0.0132)</td>
<td><em>ftsH</em> (0.0288)</td>
</tr>
<tr>
<td>8</td>
<td><em>aroP</em> (0.044)</td>
<td><em>tonB</em> (0.0136)</td>
<td><em>trmD</em> (0.0477)</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td><em>hflK</em> (0.0148)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td><em>nuoF</em> (0.0276)</td>
<td></td>
</tr>
</tbody>
</table>

p-values < 0.05 are indicated in parentheses; bolded gene targets were chosen for further study.
To demonstrate physical interactions between BbsR9 and the chosen mRNA candidates, RNA-RNA EMSAs were done using *in vitro*-transcribed BbsR9 and segments of the target mRNAs of interest with their predicted sRNA target regions. Results of the EMSAs showed that BbsR9 bound mRNAs of *ftsH*, *nuoF*, and *gcvT* *in vitro*, as judged by the novel hybrid RNA species showing markedly slower migration during gel electrophoresis (Figure 4.5). Hybrid RNAs were not observed for the other three candidate mRNAs, suggesting that sRNA binding did not occur.

Figure 4.5: BbsR9 targets transcripts of *ftsH*, *nuoF* and *gcvT* *in vitro.*
RNA-RNA EMSA of biotin-labeled \textit{in vitro}-transcribed BbsR9 binding to \textit{in vitro}-transcribed mRNA segments of the \textit{ftsH}, \textit{nuoF}, \textit{gcvT}, \textit{trmD}, BARBAKC583_RS02100 and \textit{hflK} genes. Red and blue arrows indicate bands corresponding to BbsR9 bound and unbound to target RNAs, respectively. Base values of the RNA size standard (ladder) are shown on the left.

We further characterized BbsR9-mRNA interactions by mutagenizing the predicted sRNA-binding regions of \textit{ftsH}, \textit{nuoF}, and \textit{gcvT} (Figure 4.6). The predicted \textit{ftsH}-binding region was extensive, so we created two distinct mutants for this target as well as a double-mutant (Figure 4.6). RNA-RNA EMSAs conducted with the mutagenized target mRNAs showed complete elimination of BbsR9 binding to all three targets \textit{in vitro} regardless of increasing target quantity present in the hybridization reaction (Figure 4.7). As expected, wild-type targets showed dose-dependent hybridization and signal intensity. Interestingly, abrogation of \textit{ftsH} transcript binding by BbsR9 was only observed with mutation 1 (Mut1) alone or in combination with mutation 2 (Mut2), whereas Mut2 alone did not prevent BbsR9 binding to the RNA (Figure 4.7). In consideration of the RNA secondary structure predictions and the EMSA results, we conclude that BbsR9 primarily targets mRNA transcripts via multiple GC-rich regions of a large, predicted stem-loop structure (see Figure 4.6) [145].
Figure 4.6: BbsR9 binds its targets through several GC-rich predicted seed regions.

A) Mfold secondary structure prediction of BbsR9 ($\Delta G = -46.9$ J mol$^{-1}$) with predicted seed regions for $ftsH$, $nuoF$, and $gcvT$ transcript binding indicated by red, blue, and green lines, respectively. B) Predicted IntaRNA BbsR9 target seed regions of the indicated transcripts. For the mRNA targets, nucleotide position +1 represents the first nucleotide of the respective start codon. Mutagenized bases of each mRNA are indicated in red.
Figure 4.7: BbsR9 binds to *ftsH*, *nuoF*, and *gcvT* transcripts via specific GC-rich seed regions.
RNA-RNA EMSAs showing dose-dependency of biotin-labeled BbsR9 binding to wild-type but not mutated, \textit{in vitro}-transcribed segments of \textbf{A-C) ftsH, D) nuoF} and \textbf{E) gcvT}. Mutated regions correspond to those shown in \textbf{Figure 4.6B}, and “Dbl” specifies the double \textit{ftsH} mutant. Red and blue arrows indicate bands corresponding to BbsR9 bound and unbound to target RNAs, respectively. All lanes contained 2 nM biotin-labeled BbsR9 in addition to increasing amounts of the indicated target (2 nM, 4 nM, 8 nM, and 16 nM).

\section*{Discussion}

In this study, we performed an extensive transcriptomic analysis of \textit{B. bacilliformis} grown \textit{in vitro} then shifted to one of 10 distinct conditions that mimic environments encountered by the bacterium during its natural life cycle. We chose these conditions in order to control for a variety of environmental factors that may directly influence expression of certain sRNAs. For example, temperature (25\textdegree C, 30\textdegree C, 37\textdegree C), pH levels (pH 6, pH 7.4, pH 8.2), solid/liquid substrates, and presence of a blood-gas atmosphere (5\% \text{CO}_2, 2.5\% \text{O}_2, \text{and} 92.5\% \text{N}_2 at 100\% \text{humidity}) were all examined. In addition, we included RNA-Seq experiments from experimental infections of low-passage human vascular endothelial cells (HUVE) and fresh human blood samples (HB37 and HBBG). From these experiments, we discovered 160 sRNAs expressed by \textit{B. bacilliformis} in at least one of the conditions tested.

Although we initially approached sRNA discovery using an automated approach, some clear-cut sRNAs were missed during the process. This issue led us to manually curate the 10 stranded RNA-Seq alignments, scanning each annotated gene, leader
region, and IGR for aligned reads forming peaks that could represent novel sRNAs. These peaks were required to surpass a pre-determined read coverage threshold determined independently for each condition based on reads aligned to the \textit{rpoD} (locus tag: BARBAKC583_RS04670) gene, which was consistently expressed across all 10 conditions (TPM \textasciitilde 300; \textbf{Table S4.4}). We remained consistent by using \textit{rpoD} as the housekeeping gene in qRT-PCR analyses (\textbf{Figure S4.3}) and employing a 300 TPM threshold for the purpose of the UpSet plot (\textbf{Figure 4.1}).

The putative sRNAs identified were organized into three categories (IGR, \textit{cis}-anti, or leader) depending on location of the corresponding sRNA locus. Each sRNA category has implications for its potential function. For example, IGR sRNAs are likely \textit{trans}-acting with small seed regions that often bind multiple mRNAs. \textit{Cis}-anti sRNAs most likely target the gene to which they are antisense, so target identification via algorithms such as IntaRNA would not be useful. Putative leader sRNAs are peaks that were identified sense to and in the 5' UTRs of protein-coding genes. Although the identified peaks appeared distinct from those within the actual coding sequence, the possibility remains that these peaks are not \textit{trans}-acting sRNAs. More likely, these leader RNAs may serve as \textit{cis}-acting regulatory components, like riboswitches, which are co-transcribed with the downstream protein-coding gene and harbor regulatory stem-loops that influence translation of the respective transcript [213]. Determining whether the identified leader sRNAs are \textit{cis}- and/or \textit{trans}-acting elements would require further experiments such as Northern blots and 5' RACE experiments to see if there is read-through into the downstream gene.
We performed Northern blot analyses on the putative leader sRNAs, BbspeF and BbsR7, and found that these are likely cis-acting leader RNAs, since the RNA sizes suggest read-through into the downstream gene (Figure 4.2). Northern blot analysis also verified the existence of six other sRNAs, although some of the results raise additional questions. For example, the presence of BbsR2 was detected, but the apparent band of ~450 bases is considerably larger than its predicted 284-base band (see Figure 4.2, Table S4.3). Although we identified a putative promoter element for BbsR2, we did not identify a Rho-independent terminator, so it is possible that the sRNA extends further downstream than predicted. It was also unclear whether BbsR3-1 / BbsR3-2 represented two distinct sRNAs. However, Northern blot analysis utilizing a probe against BbsR3-1 confirmed that there was a single transcript produced whose length (~600 bp) was equal to the sum of the predicted sizes of BbsR3-1 and BbsR3-2, indicating that this locus probably produces a single sRNA species (Figure 4.2). The BbsR7 blot also requires explanation. Here, several bands were identified, including smaller bands of ~200 bases and a larger band of ~600 bases. Since BbsR7 is predicted to be a leader sRNA, it is possible that the smaller bands represent the sRNA being independently expressed, while the larger band may represent BbsR7 being co-transcribed with the downstream gene (BARBAKC583_RS01695), which is 225 bp long (Figure 4.2). We also probed in the BbsR11-1 / BbsR11-2 region to determine if the two corresponding RNA-Seq peaks represented two distinct sRNAs. In this case, the Northern blot showed a single band that corresponded only to the predicted size of BbsR11-1 (Figure 4.2), suggesting that the locus harbors two distinct sRNAs. Northern blots for BbgpI and BbsR9 produced single
bands (Figure 4.2) that corresponded well to the estimated sizes of their respective peaks by RNA-Seq.

Among the sRNAs analyzed by Northern blot, Bbar45 and BbspeF are intriguing, non-coding RNA elements worthy of further characterization. Bbar45 belongs to the ar45 sRNA family first described in Sinorhizobium meliloti, but it is widely conserved in other Rhizobiales [214]. Functional characterization of sRNAs in the ar45 family has not been performed, although the S. meliloti ar45 can be co-immunoprecipitated with Hfq [215]. Since Hfq is an RNA chaperone that facilitates sRNA-mRNA interactions, we hypothesize that the S. meliloti ar45 sRNA may be trans-acting [16]. Here, we confirmed that Bbar45 is independently expressed from BbspeF, which lies immediately downstream (Figure 4.2). While this observation was previously observed in S. meliloti, it was unclear whether it was the case for other alphaproteobacteria [214]. Based on Northern blot results showing a transcript >700 bases (Figure 4.2), BbspeF is likely a leader RNA that is not independently expressed from its downstream gene. The speF leader RNA was initially discovered during a search for alphaproteobacterial riboswitches and was named for its upstream location relative to the Bacillus subtilis speF ortholog, which codes for an ornithine decarboxylase protein involved in polyamine biosynthesis [216]. However, metabolites of the polyamine biosynthesis pathway of B. subtilis were not shown to bind to the speF leader in vitro, leaving the element’s function unclear [216]. More experiments are needed to determine the regulatory role of the BbspeF leader as well as the function of the Bbar45 sRNA in B. bacilliformis.

An RNA secondary structure prediction of BbsR14 showed two stem-loops with nearly identical sequences of TTCCTCCTAA. Remarkably, these are anti-SD motifs
most often found in 16S rRNA, where they function in translational initiation. The presence of SD sequences outside of a RBS is rare, as they are selected against in the context of mRNAs, since they can cause ribosome stalling due to hybridization with 16S rRNA [217, 218]. One way in which sRNAs regulate translation is to bind directly to the RBS to occlude the ribosome and inhibit translational initiation [1]. In most cases, this is accomplished via a seed region that overlaps the SD sequence and extends up and/or downstream [6]. The predicted BbsR14 secondary structure displays unique potential seed regions solely comprised of anti-SD sequences. We speculate that this arrangement could provide opportunities for indiscriminate translational repression by the BbsR14 sRNA.

We also analyzed each of the identified B. bacilliformis sRNAs and discovered that BB019, BB113, and BB125-2 possessed a single anti-SD sequence (CCTCCT). Interestingly, of the four sRNAs that contain anti-SD sequences, BbsR14 and BB113 were significantly upregulated at pH08 relative to pH06 (see Table S4.3). Conditions of pH08 and pH06 were designed to simulate the initial and late stages of the sand fly after feeding, respectively. Thus, downregulation of translation may be advantageous for bacterial survival during initial stages within the sand fly’s midgut. As B. bacilliformis persists in the sand fly and infection proceeds, “gearing up” for a subsequent mammalian infection may occur as the insect prepares for another blood meal. Supporting this notion, we also identified 6S RNA as a sand fly-specific sRNA that was upregulated at pH08 vs. pH06, although not significantly (see Table S4.4). 6S RNAs function by binding to and sequestering the RNA polymerase holoenzyme [11]. The resulting global repression of transcription during the initial stages of sand fly infection and, to a lesser extent,
throughout a sand fly infection, could conceivably promote persistence of *B. bacilliformis* in the insect.

The mRNA target enrichment analyses for potential sand fly and infection-specific sRNAs provided insight into the regulation of pathways necessary for bacterial survival in these disparate environments. For example, targets of sand fly-specific sRNAs were significantly enriched for genes involved in the FAD-binding GO term and the biosynthesis of amino acids KEGG pathway (Figure 4.3). FAD-binding proteins include a wide array of proteins that participate in numerous biological processes. Enrichment of these genes may reflect a relatively low availability of FAD during residence in the sand fly. *B. bacilliformis* encodes a bifunctional riboflavin kinase/FAD synthetase (BARBAKC583_RS05700), and although this gene is relatively lowly expressed in all conditions tested, there is a downregulation of its expression under sand fly-like conditions (pH07, average TPM = 49.04) compared to human blood infections (HBBG, average TPM = 84.94). Enrichment of genes involved in the biosynthesis of amino acids is possibly explained by the likely downregulation of transcription and translation under sand fly-like conditions, where *B. bacilliformis* enters into a stationary phase that may promote persistence.

The human infection-specific sRNA targets were enriched in multiple GO terms associated with transferase activities, transporters, and the phospholipid biosynthetic process and KEGG pathways associated with glycerolipid/glycerophospholipid metabolism and nucleotide excision repair (Figure 4.3). Among these, there is a clear regulation of cell wall constituents during human infection conditions that would presumably be associated with morphological changes to the bacterium in the human host.
or perhaps as a means of expressing outer membrane proteins/transporters that aid in bacterial growth and replication during infection. This may very well also be in response to stressors encountered under these conditions, since nucleotide excision repair also seems to be significantly regulated by infection-specific sRNAs.

When analyzing mRNA targets of the infection-specific sRNAs, it was clear that numerous sRNAs were predicted to target the same mRNA in several cases (see Table S4.5). For example, of the 19 presumed trans-acting, infection-specific sRNAs, three independently target BARBAKC583_RS04310 transcripts, coding for lysylphosphatidylglycerol synthetase; an enzyme previously shown to augment a pathogen’s defense against host cationic antimicrobial immune peptides [219]. Additionally, four of the 19 predicted trans-acting, infection-specific sRNAs target BARBAKC583_RS00395 transcripts, coding for cobaltochelatase subunit CobT, which is involved in the synthesis of cobalamin (vitamin B₁₂), an essential coenzyme for many biological reactions [220]. It is difficult to ascribe roles to these mechanisms without knowing whether the sRNA-mediated regulation is positive or negative, although it is worth noting that redundant targeting is not a result of sRNA duplication, and each predicted binding site on these transcripts is unique. We hypothesize that redundant regulation of particular mRNAs may serve to “hyper-regulate” protein production in response to subtle differences in environmental cues. This kind of redundant regulation of mRNAs from multiple “sibling sRNAs” has been described in other pathogens, so further research into the function of sibling sRNAs of B. bacilliformis could be fruitful [221].

We found conservation of some sRNAs among other alphaproteobacteria species using discontinuous megaBLAST analysis (Table S4.3). Unfortunately, we were only
able to analyze IGR and leader sRNAs, since *cis*-anti sRNAs showed broad sequence conservation due to their close linkage to protein-coding genes. The majority of analyzed sRNAs was unique to *B. bacilliformis*, while the BB036 sRNA group was unique to the KC583 strain of *B. bacilliformis*. Five other sRNAs were widespread in *Bartonella* spp., including BbsR9 which was characterized in this study. Conservation of BbsR9 in other *Bartonella* spp. further highlights its potential importance. Since *Bartonella* spp. are typically transmitted to mammals by various arthropods (ticks, sand flies, fleas, lice, etc.), it is possible that BbsR9 plays a role in persistence in many vectors. Seven more sRNAs were found in additional alphaproteobacteria, including ubiquitous sRNAs like 6S RNA and tmRNA, conserved alphaproteobacteria sRNAs like Bbar45, and the tRNA*<sub>Arg</sub>*<sup>CCU</sup> group I intron.

BbgpI is a member of a tRNA*<sub>CCU</sub>*<sup>Arg</sup> group I intron family first identified in *Agrobacterium tumefaciens* and later found in other alphaproteobacteria [222, 211]. Group I introns are selfish genetic elements that insert into tRNAs, rRNAs, and protein-coding genes. Although group I introns are ribozymes and RNA splicing is auto-catalytic, they sometimes require protein co-factors for self-splicing *in vitro*, and it is presumed that all group I introns require protein co-factors to some extent for splicing *in vivo* [74]. Here, we have demonstrated that BbgpI self-splices *in vitro* and is spliced *in vivo*.

Furthermore, we have shown that BbgpI is not located in an IGR as presumed, but rather within an unannotated tRNA*<sub>CCU</sub>*<sup>Arg</sup> gene. Since the flanking tRNA*<sub>CCU</sub>*<sup>Arg</sup> gene retains all necessary tRNA domains (see Figure S4.4C), we predict that the tRNA is functional following intron splicing. This novel tRNA gene might have implications for future analyses of *B. bacilliformis* involving codon bias, conservation of tRNA genes, amino
acid scavenging, etc. Furthermore, this discovery suggests further optimization may be required for current tRNA scanning algorithms.

We also characterized the targeting and molecular interactions of BbsR9, as the sRNA was only appreciably expressed under pH06, pH07, and pH08 conditions (Table S4.4). For reference, these conditions reflect a liquid blood / serum environment at 30⁰C (Table 4.1) and simulate the sand fly’s midgut following a blood meal. It is interesting to note that the PI30 condition is identical to pH07 except that PI30 represents a solid medium. Furthermore, Northern blot analyses indicated that, in addition to the liquid medium requirements, BbsR9 expression was restricted to temperatures < 37⁰C (Figure S4.5B). The regulatory mechanisms that facilitate such an expression pattern warrant further investigation.

We verified several mRNA targets of BbsR9 using RNA-RNA EMSAs. Among the targets were transcripts of the *ftsH*, *nuoF*, and *gcvT* genes. First, *ftsH* codes for the FtsH zinc metalloprotease; a membrane-anchored, universal protease with various functions. FtsH has been extensively studied in *E. coli*, where it is the only protease essential for survival [reviewed in 223]. FtsH has also been described as required for regulation of optimal ratios of phospholipids and LPS in the outer membrane [223]. Whether BbsR9 regulation of *ftsH* transcripts is involved in bacterial protein turnover and/or modulation of membrane architecture in the context of the sand fly is unknown, but would be interesting to investigate. Second, the *nuoF* gene codes for the NADH-quinone oxidoreductase subunit F, a component of the type I NADH dehydrogenase enzyme and the initial step in the electron transport chain. NuoF is a component of the peripheral fragment of the NADH dehydrogenase complex and plays a role in oxidation
of NADH to generate a proton motive force [reviewed in 224]. Regulation of \textit{nuoF} transcripts could conceivably play a role in helping to establish the stationary phase as \textit{B. bacilliformis} persists in the sand fly. Finally, we determined that BbsR9 targets transcripts of \textit{gcvT}, which codes for the glycine cleavage system aminomethyltransferase, GcvT. The glycine cleavage system responds to high concentrations of glycine, breaking the amino acid down to CO$_2$, ammonia, and NADH [225]. In addition to redox reactions, NADH can be used to produce energy through cellular respiration. Of note is the potential interplay between sRNA targeting of \textit{nuoF} and \textit{gcvT} transcripts in this regard.

Interestingly, the glycine cleavage system has been implicated in contributing to bacterial persistence in animal and plant hosts [226]. In fact, \textit{gcvT} is essential for persistence of a closely-related pathogen, \textit{Brucella abortus}, in its animal host [227]. However, to our knowledge, the role of a glycine cleavage system in pathogen persistence in its arthropod vector has not been explored, to date. It is conceivable that \textit{B. bacilliformis} utilizes regulation of \textit{nuoF} and \textit{gcvT} to fine-tune levels of NAD+/NADH, thereby contributing to regulation of metabolism and persistence of the bacterium in the sand fly.

This study has provided further insight into the regulation of numerous processes by \textit{B. bacilliformis} in response to conditions encountered in the context of its sand fly vector and human host. We believe the results provide a strong foundation for future studies examining sRNA-mediated regulation in \textit{B. bacilliformis} and the regulatory mechanisms required for vector-host transmission.

\textbf{Acknowledgments}
The authors wish to thank Patty Langasek and Auguste Dutcher for technical assistance. This work was supported by NIH grant R21AI128575 (to MFM). RR was supported by NIH grants AI133023 and DE028409.

**Supplementary Material**

![B. bacilliformis RNA-Seq PCA Plot](image)

**Figure S4.1: B. bacilliformis RNA-Seq PCA plot.** Axes indicate the percentage of total variance that can be accounted for by two principle components. Colored dots indicate the retained biological replicates of the RNA-Seq analyses, and their distance apart is representative of overall relatedness in gene expression profiles. Experimental conditions are shown on the right.
Figure S4.2: *B. bacilliformis* sRNAs group into specific expression patterns. Heatmap of *B. bacilliformis* sRNA TPMs across the tested conditions (shown at the bottom). sRNAs group vertically based on similarity in expression patterns. Conditions group horizontally based on similarity in overall expression patterns. The $\log_{10}$ of the TPM value for each sRNA is indicated by a color gradient.

Figure S4.3: qRT-PCR confirmation of differential expression of several identified sRNAs. Faceted bar graph displaying the number of sRNA transcripts / rpoD transcript for select, differentially-expressed sRNAs and BB024, which was not shown to be differentially expressed. The condition / source of the total RNA is noted on the x-axis.
Significance was determined by students t-test (N = 9; * = p<0.05, ** = p<0.01, *** = p<0.001).

Figure S4.4: BbgpI is a group I intron inserted into an unannotated tRNA<sub>CCU<sup>Arg</sup></sub> gene of <i>B. bacilliformis</i>. A) Nucleotide sequence of BbgpI (bolded and underlined) and flanking chromosomal regions. Primer binding sites used for <i>in vitro</i> transcription (IVT) and PCR assays designed to show splicing of BbgpI <i>in vitro</i> and <i>in vivo</i> are indicated. B) Sequence of BbgpI outlining the conserved, characteristic stem structures (P1 to P9) with putative base pairings highlighted in green and yellow. Nucleotides predicted to
participate in base pairing are bolded and underlined. C) Sequence coding for the tRNA_{CCU}^{Arg} immediately flanking BbgpI. The two bolded underlined nucleotides represent the ends of the spliced out BbgpI. Conserved tRNA features are also outlined.

**Figure S4.5: BbsR9 is a sand fly-specific sRNA.** A) Nucleotide sequence of the bbsR9 gene with predicted promoter elements and Rho-independent terminator plus experimentally-determined TSS’s, highlighted in various colors or underlined,
respectively. An asterisk indicates the alternative TSS found by 5' RACE analysis.

**Northern blot analysis of BbsR9 expression under the indicated conditions.** The RNA ladder (2 min exposure) and resolved total RNA samples (30s exposure) were from the same blot but imaged using different exposure times.

<table>
<thead>
<tr>
<th>Table S4.1: Bacterial strains, primers, and plasmids used in the study.</th>
</tr>
</thead>
</table>

**Northern blot**

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Name</th>
<th>Sequence</th>
<th>Tm</th>
<th>Length</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGAATTGGGAACC</td>
<td>59.2</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGGAATAACGGTATTG</td>
<td>58.7</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCCCCAACAACTCTTG</td>
<td>63</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>aRT-PCR</td>
<td>Bb_gpI_Probe_F</td>
<td>TAATACGACTCACTATAGGGGGACTGTCTCTTC</td>
<td>57.5</td>
<td>33</td>
<td>This Study</td>
</tr>
<tr>
<td>aRT-PCR</td>
<td>Bb_gpI_Probe_R</td>
<td>TGTCACGGAATAGCTGATAAAGATC</td>
<td>55.8</td>
<td>22</td>
<td>This Study</td>
</tr>
<tr>
<td>aRT-PCR</td>
<td>Bb_gpI_Probe_F</td>
<td>TAATACGACTCACTATAGGCAAGAAAGAATCCGGGAATTGGGAACCCAG</td>
<td>59.6</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>aRT-PCR</td>
<td>Bb_gpI_Probe_R</td>
<td>TAATACGACTCACTATAGGGGCAATTGAGATTTGATATGGATCCTG</td>
<td>62.5</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>aRT-PCR</td>
<td>Bb_gpI_Splice_Sensitive</td>
<td>TAATACGACTCACTATAGGGGAAATGATTAAGAAACCGC</td>
<td>60.6</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>aRT-PCR</td>
<td>Bb_gpI_Splice_Sensitive</td>
<td>TAATACGACTCACTATAGGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>aRT-PCR</td>
<td>Bb_gpI_Splice_Sensitive</td>
<td>TAATACGACTCACTATAGGGCCCCTATGAGATTTGATATGGATCCTG</td>
<td>62.5</td>
<td>30</td>
<td>This Study</td>
</tr>
</tbody>
</table>

**Group I intron Splicing**

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Name</th>
<th>Sequence</th>
<th>Tm</th>
<th>Length</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Splice_NEW_F</td>
<td>GGTTACGATCACTATAGGGGAATTGGGAACC</td>
<td>59.6</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Splice_NEW_R</td>
<td>TAATACGACTCACTATAGGGGCAATTGAGATTTGATATGGATCCTG</td>
<td>62.5</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_F</td>
<td>GGTTACGATCACTATAGGGGAAATGATTAAGAAACCGC</td>
<td>60.6</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_R</td>
<td>TAATACGACTCACTATAGGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_F</td>
<td>GGTTACGATCACTATAGGGGCAATTGAGATTTGATATGGATCCTG</td>
<td>62.5</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_R</td>
<td>TAATACGACTCACTATAGGGGAAATGATTAAGAAACCGC</td>
<td>60.6</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_F</td>
<td>GGTTACGATCACTATAGGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_R</td>
<td>TAATACGACTCACTATAGGGGCAATTGAGATTTGATATGGATCCTG</td>
<td>62.5</td>
<td>30</td>
<td>This Study</td>
</tr>
</tbody>
</table>

**qRT-PCR**

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Name</th>
<th>Sequence</th>
<th>Tm</th>
<th>Length</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGAATTGGGAACC</td>
<td>59.2</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGGAATAACGGTATTG</td>
<td>58.7</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCCCCAACAACTCTTG</td>
<td>63</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGGACTGTCTCTTC</td>
<td>57.5</td>
<td>33</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGGAATAACGGTATTG</td>
<td>58.7</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCCCCAACAACTCTTG</td>
<td>63</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGGACTGTCTCTTC</td>
<td>57.5</td>
<td>33</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGGAATAACGGTATTG</td>
<td>58.7</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCCCCAACAACTCTTG</td>
<td>63</td>
<td>30</td>
<td>This Study</td>
</tr>
</tbody>
</table>

**Group I intron Splicing**

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Name</th>
<th>Sequence</th>
<th>Tm</th>
<th>Length</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Splice_NEW_F</td>
<td>GGTTACGATCACTATAGGGGAATTGGGAACC</td>
<td>59.6</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Splice_NEW_R</td>
<td>TAATACGACTCACTATAGGGGCAATTGAGATTTGATATGGATCCTG</td>
<td>62.5</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_F</td>
<td>GGTTACGATCACTATAGGGGAAATGATTAAGAAACCGC</td>
<td>60.6</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_R</td>
<td>TAATACGACTCACTATAGGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_F</td>
<td>GGTTACGATCACTATAGGGGCAATTGAGATTTGATATGGATCCTG</td>
<td>62.5</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_R</td>
<td>TAATACGACTCACTATAGGGGAAATGATTAAGAAACCGC</td>
<td>60.6</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_F</td>
<td>GGTTACGATCACTATAGGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_R</td>
<td>TAATACGACTCACTATAGGGGAAATGATTAAGAAACCGC</td>
<td>60.6</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_F</td>
<td>GGTTACGATCACTATAGGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>Group I intron Splicing</td>
<td>Bb_Intron_Spike_NEW_R</td>
<td>TAATACGACTCACTATAGGGGAAATGATTAAGAAACCGC</td>
<td>60.6</td>
<td>30</td>
<td>This Study</td>
</tr>
</tbody>
</table>

**qRT-PCR**

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Name</th>
<th>Sequence</th>
<th>Tm</th>
<th>Length</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGAATTGGGAACC</td>
<td>59.2</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGGAATAACGGTATTG</td>
<td>58.7</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCCCCAACAACTCTTG</td>
<td>63</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGGACTGTCTCTTC</td>
<td>57.5</td>
<td>33</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGGAATAACGGTATTG</td>
<td>58.7</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCCCCAACAACTCTTG</td>
<td>63</td>
<td>30</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGGACTGTCTCTTC</td>
<td>57.5</td>
<td>33</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGGGAATAACGGTATTG</td>
<td>58.7</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCTATAGCTGAAAGAC</td>
<td>60.3</td>
<td>36</td>
<td>This Study</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Bb_sscG_CPR</td>
<td>TAATACGACTCACTATAGGGCCCCAACAACTCTTG</td>
<td>63</td>
<td>30</td>
<td>This Study</td>
</tr>
</tbody>
</table>
### Table S4.2: Quality control results for *B. bacilliformis* RNA-Seq analyses.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Raw reads</th>
<th>Reads after Filter, Trim, and Pairing</th>
<th>% Reads Discarded</th>
<th>Reads with Bowtie Alignments</th>
<th>Bowtie Ambiguous Reads</th>
<th>Bowtie Unambiguous Reads</th>
<th>featureCounts Assigned Reads</th>
<th>% Ambiguous Reads</th>
<th>% Reads with alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pl25-1</td>
<td>27352690</td>
<td>20391632</td>
<td>23.6093369</td>
<td>1529263 1548384</td>
<td>2668779 756917</td>
<td>9827569 99.9492894</td>
<td>3525377 98.2013844</td>
<td>52.4498463</td>
<td>98.2013844</td>
</tr>
<tr>
<td>Pl25-2</td>
<td>25075128</td>
<td>19569572</td>
<td>23.7940137</td>
<td>1529263 1548384</td>
<td>2668779 756917</td>
<td>9827569 99.9492894</td>
<td>3525377 98.2013844</td>
<td>52.4498463</td>
<td>98.2013844</td>
</tr>
<tr>
<td>Pl25-3</td>
<td>29684466</td>
<td>22182380</td>
<td>24.8000677</td>
<td>1529263 1548384</td>
<td>2668779 756917</td>
<td>9827569 99.9492894</td>
<td>3525377 98.2013844</td>
<td>52.4498463</td>
<td>98.2013844</td>
</tr>
<tr>
<td>Pl30-1</td>
<td>28000000</td>
<td>21467982</td>
<td>24.9838016</td>
<td>1529263 1548384</td>
<td>2668779 756917</td>
<td>9827569 99.9492894</td>
<td>3525377 98.2013844</td>
<td>52.4498463</td>
<td>98.2013844</td>
</tr>
<tr>
<td>Pl30-2</td>
<td>26751499</td>
<td>20410930</td>
<td>24.6362136</td>
<td>1529263 1548384</td>
<td>2668779 756917</td>
<td>9827569 99.9492894</td>
<td>3525377 98.2013844</td>
<td>52.4498463</td>
<td>98.2013844</td>
</tr>
<tr>
<td>Pl30-3</td>
<td>30623621</td>
<td>23509074</td>
<td>24.9413588</td>
<td>1529263 1548384</td>
<td>2668779 756917</td>
<td>9827569 99.9492894</td>
<td>3525377 98.2013844</td>
<td>52.4498463</td>
<td>98.2013844</td>
</tr>
</tbody>
</table>

Replicates removed from global analysis
Table S1:3: Pufative sRNAs identified in B. bacilliformis by RNA-Seq analyses.
Table S4.4: Average TPMs of identified B. bacilliformis sRNAs.

<table>
<thead>
<tr>
<th>Name</th>
<th>PlBG</th>
<th>HUVE</th>
<th>HB37</th>
<th>HBBG</th>
<th>AVG</th>
<th>St Dev</th>
<th>Cut-Off Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB001</td>
<td>496.281673</td>
<td>355.0882</td>
<td>261.7784</td>
<td>...</td>
<td>HBBG (1 STDEV higher than mean)</td>
<td>S: Upregulated in ONLY two of four of Pl25, pH06, pH07, pH08 (1STDEV higher than mean)</td>
<td></td>
</tr>
</tbody>
</table>
Table S4.5: Predicted IntaRNA targets of *B. bacilliformis* infection-specific sRNAs.

An “X” indicates transcripts of the indicated gene to which the sRNA is predicted to bind (p < 0.01). Targets with a FDR < 0.05 are indicated with a red “X”.

| Gene Name | Product | BB003 | BB016 | BB020 | BB022 | BB024 | BB025 | BB035 | BB039 | BB048 | BB060 | BB075 | BB103-1 | BB103-4 | BB104-2 | BB105-1... |
|-----------|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|---------|---------|---------|---------|
|...|...|...|...|...|...|...|...|...|...|...|...|...|...|...|...|...|
Table S4.6: Predicted IntaRNA targets of *B. bacilliformis* sand fly-specific sRNAs.

An “X” indicates transcripts of the indicated gene to which the sRNA is predicted to bind (p < 0.01). Targets with a FDR < 0.05 are indicated with a red “X”.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Product</th>
<th>BB060</th>
<th>BB003</th>
<th>BB009</th>
<th>BB109-2</th>
<th>BB109-3</th>
<th>BB124</th>
</tr>
</thead>
<tbody>
<tr>
<td>BARBAKC583_RS04075</td>
<td>cold shock protein</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS01990</td>
<td>site-specific DNA-methyltransferase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS05060</td>
<td>metalloprotease</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS00527</td>
<td>transcriptional repressor</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS06065</td>
<td>HlyC/ConC family transporter</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS02225</td>
<td>bifunctional (ppGpp synthetase/guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050190</td>
<td>exoribonuclease</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050865</td>
<td>tetraacyl/disaccharide 4'-kinase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS020830</td>
<td>outer membrane protein assembly factor BamA</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050345</td>
<td>signal recognition particle protein</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS040685</td>
<td>NaPi cotransporter family protein</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS020620</td>
<td>Trk system potassium transporter TrkA</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS01875</td>
<td>FAD-binding oxidoreductase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050690</td>
<td>phosphoglycerate kinase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS000660</td>
<td>DUF1036 domain-containing protein</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS040120</td>
<td>BofA family transcriptional regulator</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS000665</td>
<td>tRNA uridine-5'-carboxymethylaminomethyl(34) MnmG</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS060555</td>
<td>2-polypropenylphosphol 6-hydroxylase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS020125</td>
<td>hydroxymethylpyrimidine/photophosphomethylpyrimidine kinase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS000655</td>
<td>LPS export ABC transporter ATP-binding protein</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050200</td>
<td>shikimate kinase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050325</td>
<td>glycine cleavage system aminomethyltransferase GcvT</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS000430</td>
<td>tRNA methyltransferase TrmD</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050100</td>
<td>DNA-binding response regulator</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050225</td>
<td>energy transducer TonB</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS030820</td>
<td>NADH-quinone oxidoreductase subunit NuqD</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050835</td>
<td>ATP-dependent zinc metalloprotease FltH</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050490</td>
<td>tRNA 2-thiouridine(34) synthase MnmA</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050220</td>
<td>TonB-dependent hemoglobin-transferin/lactoferrin receptor protein</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050045</td>
<td>succinate dehydrogenase hydrophobic membrane anchor protein</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050560</td>
<td>cell division protein FtsA</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050135</td>
<td>glutamate-5-semialdehyde dehydrogenase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050565</td>
<td>16S rRNA (cytidine(1402)-2'-O)-methyltransferase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050490</td>
<td>preprotein translocase subunit SecY</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050160</td>
<td>metal ABC transporter substrate-binding protein</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050755</td>
<td>peroxiredoxin</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050070</td>
<td>ribosomal RNA small subunit methyltransferase G</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050130</td>
<td>phospholipase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050440</td>
<td>23S-tetrahydrodiynidine-26,carboxylate N-succinyltransferase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050150</td>
<td>phosphomethylpyrimidine synthase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050180</td>
<td>SURF1 family protein</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050445</td>
<td>protease modulator HflC</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050815</td>
<td>amino acid permease</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050435</td>
<td>DNA repair protein FlexN</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050420</td>
<td>cold shock protein</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050380</td>
<td>305-ribosomal protein S10</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050220</td>
<td>305-ribosomal protein S10</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050620</td>
<td>acetylornithine transaminase</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARBAKC583_RS050210</td>
<td>DUF2059 domain-containing protein</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
QMITEs as a source for sRNAs and highly basic proteins

We have shown that QMITE1 and QMITE2 copies can serve as sources for novel sRNAs. Specifically, the sRNAs CbsR3 and CbsR13 were derived from QMITE1 loci, whereas CbsR16 was derived from a QMITE2 locus. Although QMITE1 and QMITE2 exist in multiple copies throughout C. burnetii genomes, RNA-Seq analysis has shown that certain loci have unambiguous reads mapping to them, indicating that these loci are transcriptionally active. Indeed, we have shown that predicted promoter elements exist within the confines of QMITE copies (see Figure S2.11). When transcribed, CbsR13 and CbsR16 produce large, stable stem-loop structures (see Figures 2.1B, 2.4B) that may serve as substrates for RNase III degradation. The roles for these sRNAs remain unclear, although their high ambiguous expression (see Figure S2.3) suggests that they may serve some adaptive role.

In addition to being the source for sRNAs, some QMITE1 copies contain a short ORF coding for highly basic proteins (average pI ~12.4). These uncharacterized proteins may confer some adaptive advantage for C. burnetii within the host, where it must survive low pH conditions within acidified phagolysosomes. Indeed, while the extracellular pH in this niche is ~ 4.5-5, C. burnetii maintains an intracellular pH of 5.1 to 6.95 [228]. Determining the roles of these proteins within the intracellular niche could provide further insights into how C. burnetii survives such an extreme environment.

enhC and QMITEs as a timeline for C. burnetii strain divergence
We determined that QMITE copies show inter-strain sequence and linkage conservation (see Table 2.1). While this is useful for establishing a timeline for IS1111 insertion and determining recent horizontal gene acquisition, it also allows for tracing individual QMITE insertions in order to gauge their effect on strain phenotype. For example, it is very interesting to note a QMITE2 insertion in the 3’ end of the *C. burnetii* Dugway strain’s *enhC* gene. This insertion effectively provides a C-terminal extension for the resulting EnhC protein, but it is very possible that the strong secondary structure of the QMITE2 insertion may cause ribosome stalling, leading to degradation of the transcript via tmRNA [229, 14]. Since EnhC has been implicated in the virulence of *L. pneumophila*, a close pathogenic relative of *C. burnetii*, it is conceivable that this QMITE2 insertion in the Dugway strain *enhC* gene renders it avirulent [161, 162, 139]. The avirulent nature of the Dugway strain remains a mystery, so this could provide valuable information on the factors necessary for successful host infection.

**CbsR12 is a trans-acting sRNA that also binds CsrA**

In this study, we carried out a comprehensive characterization of a highly expressed, infection-specific sRNA of *C. burnetii*, named CbsR12. This sRNA was found to be necessary for CCV expansion during early infection of a human monocyte-derived alveolar macrophage cell line (THP-1s) (see Figure 3.4). Growth rate also correlated to CbsR12 expression, both *in vitro* and during THP-1 infection (see Figures 3.2, 3.3). We also determined that the regulation of CbsR12 was dependent on the genomic context, as a transposon-based genetic complement of a *cbsR12* mutant targeted towards a different genomic locus resulted in the dysregulation of CbsR12 expression (see Figures 3.2, 3.3). As a result, we conclude that there may be an unknown transcriptional regulator of
CbsR12 expression. Indeed, there is a predicted PmrA binding site within the \textit{cbsR12} promoter, although this element was also present in the \textit{cbsR12} complementation cassette. As a result, it does not fully explain the noted dysregulation of CbsR12 expression. Determining the mechanisms for CbsR12 regulation would help in understanding its expression pattern during infection.

We determined that CbsR12 is a \textit{trans}-acting sRNA that engages in the post-transcriptional regulation of \textit{carA} and \textit{metK}. We also determined that CbsR12 binds to transcripts of \textit{cvpD} and \textit{ahcY} \textit{in vivo}. Additionally, CbsR12 binds to CsrA-2, but not CsrA-1, \textit{in vitro}. These data, in addition to the \textit{E. coli} biofilm induction data, indicate that in addition to its \textit{in-trans} activities, CbsR12 also serves as an RsmY/Z sRNA of \textit{C. burnetii}. These sRNAs act by binding CsrA, sequestering it away from its regulatory activities [29]. We have also determined that CbsR1, another highly expressed infection-specific sRNA, also harbors multiple CsrA-binding sites. Furthermore, CbsR1 seems to contain a classical GacA/LetA-binding site that is common amongst RsmY/Z sRNAs [27]. Since the CsrA regulon of \textit{C. burnetii} is wholly unknown, we hope that these conclusions will lead to more research into this important regulatory network. For example, determining the regulatory mechanisms of CbsR1 and CbsR12 expression would be important, as would determining the repertoire of mRNAs to which CsrA-1 and CsrA-2 bind. The regulation of CsrA-1 and CsrA-2 production would also be a fruitful area of research. Despite all that is unknown, we have developed a model of what we know about the \textit{C. burnetii} CsrA regulon, to date (\textbf{Figure 5.1}).
Figure 5.1: What is known of the CsrA regulatory cascade/regulon in *C. burnetii*. An uncharacterized host signal triggers a putative sensor kinase(s), leading to regulation of CbsR1/CbsR12 by an unknown response regulator. In turn, CbsR12 may promote CCV expansion via regulation of the CvpD T4BSS effector. CbsR12 also promotes replication through in-*trans* regulation of CarA and MetK production. CbsR1/CbsR12 may also bind CsrA-1/CsrA-2, leading to sequestration and indirect regulation of the CsrA-1/CsrA-2 targetome.

**Determining the role of the methionine cycle in *C. burnetii***
Our research into CbsR12 has lead to many additional questions regarding the methionine cycle in *C. burnetii*. Since *C. burnetii* is a semi- auxotroph for methionine, it is presumed that the bacterium can scavenge it from the host [182]. Indeed, an ABC transporter for methionine has been predicted in *C. burnetii* [182]. We determined that CbsR12 negatively regulates MetK production. MetK is responsible for converting methionine to SAM, which is an essential methyl group donor in bacterial cells [166]. Although the regulation of MetK production in the context of retaining scavenged methionine is intuitive, *C. burnetii* would seemingly also require a mechanism for producing or scavenging SAM. Indeed, SAM scavenging has been reported in another obligate intracellular pathogen, *Rickettsia prowazekii* [185]. Furthermore, we have determined, through position-specific iterated BLAST (PSI-BLAST) homology and *in silico* protein folding analysis [230], that CBU_0636 may be a SAM transporter homolog (Figure 5.2). If so, it may provide a compensatory mechanism for the down-regulation of MetK production by CbsR12. Also, determining the nature of *ahcY* regulation by CbsR12 may also provide insights into the role of the methionine cycle during infection.
**Figure 5.2: CBU_0636 may be a SAM transporter homolog.** I-TASSER protein structure comparisons of the *R. prowazekii* Madrid E SAM transporter (*sam* gene) (A) and the *C. burnetii* RSA493 CBU_0636 gene product (B).

**BbsRs as a means of rapid regulation in rapidly changing environments**

We performed total RNA-Seq analysis on *B. bacilliformis* grown *in vitro* then shifted to one of ten conditions designed to mimic the various environments encountered by the pathogen during its life cycle. In doing so, we discovered 160 novel sRNAs, some of which were found to be differentially expressed under certain conditions (see Table S4.3). We hope that the discovery of these sRNAs leads to many future characterization studies. Additionally, since we performed total RNA-Seq, global gene expression analyses can be carried out. Alternatively, the results may be used as reference datasets in future studies.

Determining the *B. bacilliformis* Hfq targetome would be useful in further characterizing the identified sRNAs. For example, by performing an *in vivo* crosslinking analysis such as CLASH [48], one could simultaneously determine the repertoire of Hfq-binding sRNAs along with the mRNAs they target. It is worth noting, though, that the majority (149 out of 160) of the sRNAs we identified do not have a predicted Rho-independent terminator, which is thought to be essential for binding Hfq (see Table S4.3) [20]. However, this doesn’t rule out the possibility of some unknown sRNA chaperone being involved.

Among the *B. bacilliformis* sRNAs we identified, the BB103 grouping (BB103-1, BB103-2, BB103-3, and BB103-4; see Table S4.3) is ideal for future characterization.
While each sRNA shows high expression under certain conditions (see Table S4.4), we determined that BB103-1 and BB103-4 are infection-specific, while BB103-2 and BB103-3 are sand fly-specific. Furthermore, BB103-1 and BB103-4 are both predicted to target transcripts of the RS04310 gene, coding for lysylphosphatidylglycerol synthetase, which is involved in the defense against host antimicrobial peptides by other pathogens (see Table S4.5) [219]. Determining the targetomes of these sRNAs and how they are regulated may help in understanding how B. bacilliformis adapts to the mammalian host and sand fly vector.

**BbsR9 is a Bartonella-specific sRNA uniquely expressed in the arthropod vector**

We determined that BbsR9 is a highly expressed, sand fly-specific sRNA that targets transcripts of the *ftsH, nuoF*, and *gcvT* genes, *in vitro* (see Figures 4.6, 4.7). We hypothesize that the regulation of these transcripts aids B. bacilliformis in persistence in the sandfly vector. Additionally, BbsR9 has a predicted Rho-independent terminator and is also conserved in some other Bartonella spp. Since pathogenic Bartonella spp. are vector-borne, we predict that BbsR9 is involved in persistence in a wide array of arthropod vectors. Experimental infections of *L. verrucarum* with a *bbsR9* mutant strain could be done to help determine its role in persistence. Additionally, since BbsR9 may be bound by Hfq, CLASH [48] could be useful in determining its repertoire of mRNA targets. Finally, since BbsR9 is appreciably expressed under very limited conditions (pH06, pH07, and pH08; see Tables 4.1, S4.4), it would be prudent to identify the specific regulators involved, as they may also be generically involved in the transition of *B. bacilliformis* from vector to host and back again.
References


69. Chen SL, Shapiro L. Identification of long intergenic repeat sequences associated 
with DNA methylation sites in *Caulobacter crescentus* and other alpha-

70. Ogata H, Audic S, Abergel C, Fournier PE, Claverie JM. Protein coding palindromes 
are a unique but recurrent feature in *Rickettsia*. Genome Res. 2002;12(5):808–816.


Whole-genome organization and functional properties of miniature DNA insertion 

73. De Gregorio E1, Silvestro G, Petrillo M, Carlomagno MS, Di Nocera PP. 
Enterobacterial repetitive intergenic consensus sequence repeats in yersinia: 

Mob DNA. 2014;5:8.

75. Wilson GW, Edgell DR. Phage T4 *mobE* promotes trans homing of the defunct 


77. Minnick MF, Raghavan R. Genetics of *Coxiella burnetii*: on the path of 

78. Romano PS, Gutierrez MG, Berón W, Rabinovitch M, Colombo MI. The autophagic 
pathway is actively modulated by phase II *Coxiella burnetii* to efficiently replicate in 


93. Hackstadt T, Peacock MG, Hitchcock PJ, Cole RL. Lipopolysaccharide variation in 
*Coxiella burnetii*: intrastrain heterogeneity in structure and antigenicity. Infect 

94. Fiset P. Phase variation of *Rickettsia* (*Coxiella*) *burneti*; a study of the antibody 

95. Hackstadt T. Steric hinderance of antibody binding to surface proteins of *Coxiella 

96. Beare PA, Jeffrey BM, Long CM, Martens CM, Heinzen RA. Genetic mechanisms of 

97. Warrier I, Hicks LD, Battisti JM, Raghavan R, Minnick MF. Identification of novel 

98. Omsland A, Beare PA, Hill J, Cockrell DC, Howe D, Hansen B, Samuel JE, Heinzen 
RA. Isolation from animal tissue and genetic transformation of *Coxiella burnetii* are 

genome sequence of the Q-fever pathogen *Coxiella burnetii*. Proc Natl Acad Sci U S 
A. 2003;100(9):5455-60.

100. Raghavan R, Hicks LD, Minnick MF. Toxic introns and parasitic intein in 
*Coxiella burnetii*: Legacies of a promiscuous past. J. Bacteriol. 2008;190(17):5934- 
5943.


134. Santos VC, Araujo RN, Machado LA, Pereira MH, Gontijo NF. The physiology of the midgut of *Lutzomyia longipalpis* (Lutz and Neiva 1912): pH in different
physiological conditions and mechanisms involved in its control. 2008;211(Pt 17):2792-8.


204. MacLellan SR, MacLean AM, Finan TM. Promoter prediction in the rhizobia. Microbiology. 2006;152(Pt 6):1751-63.


