Uncovering *Coxiella burnetii*'s Pathogenicity by Elucidating its Metabolism and Host Interactions

Jess Annai Millar

*Portland State University*

Let us know how access to this document benefits you.

Follow this and additional works at: [https://pdxscholar.library.pdx.edu/open_access_etds](https://pdxscholar.library.pdx.edu/open_access_etds)

Part of the [Biology Commons](https://pdxscholar.library.pdx.edu/open_access_etds)

Recommended Citation

Millar, Jess Annai, "Uncovering *Coxiella burnetii*'s Pathogenicity by Elucidating its Metabolism and Host Interactions" (2017).

*Dissertations and Theses. Paper 3937.*

10.15760/etd.5821

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.
Uncovering *Coxiella burnetii*’s Pathogenicity by Elucidating its
Metabolism and Host Interactions

by

Jess Annai Millar

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science
in
Biology

Thesis Committee:
Rahul Raghavan
Susan E. Masta
Kenneth M. Stedman

Portland State University
2017
ABSTRACT

Coxiella burnetii, the etiologic agent of acute Q fever and chronic endocarditis, has a unique biphasic life cycle, which includes a metabolically active intracellular form that occupies a large lysosome-derived acidic vacuole. C. burnetii is the only bacterium known to thrive within such a hostile intracellular niche, and this ability is fundamental to its pathogenicity; however, very little is known about genes that facilitate Coxiella’s intracellular growth. This lack of knowledge of Coxiella’s basic biology and molecular pathogenesis is a critical barrier to developing more effective therapies.

In this study, we aimed to understand both bacterial and host factors that have important roles during C. burnetii infections. Using an evolutionary genomics approach, we identified metabolic pathways that are critical to C. burnetii’s ability to grow intracellularly. Among those found, the most promising are fatty acid, biotin, and heme biosyntheses pathways. Coxiella has horizontally acquired extra copies of genes that enhance these processes; when these genes were disrupted, Coxiella’s growth was significantly inhibited. Also, by analyzing the host transcriptome, we identified human genes, including microRNA (miRNA) genes that are important during C. burnetii infections. Coxiella induces the expression of multiple anti-apoptotic miRNAs, which likely have a role in inhibiting apoptosis in order to sustain the intracellular replication of the pathogen. The biosynthetic pathways and miRNAs identified in this study are ideal targets for developing more effective therapeutic strategies against Q fever and its chronic and often fatal complications.
ACKNOWLEDGEMENTS

I would like to thank my research advisor Rahul Raghavan of the Biology Department at Portland State University. He has provided continual support but also generous freedom that has allowed me to design aspects and conduct much of this project independently. I would also like to acknowledge Rahul and Tina Schroyer for their assistance in writing my thesis. Their feedback has helped me become a better writer. I want to thank Abraham Moses and Fenil Kacharia for their assistance with conducting all the physical experiments and other lab duties. I also want to thank Todd Smith for taking the time to supplement my learning in bioinformatics. I would also like to thank my thesis committee: Susan Masta and Kenneth Stedman. I appreciate their time and effort reviewing my paper and supporting my research and education.

This work was supported by startup funds from PSU to Rahul Raghavan, as well as the American Heart Association Beginning Grant-in-Aid, Collins Medical Trust Research Grant, Medical Research Foundation of Oregon New Investigator Grant, and NIH National Institute of Allergy and Infectious Diseases grants (R03 AI123464-01A1, R15 AI126385-01A1). Fellowship and partial funding was provided by the Portland State Laurels Graduate Award, Elsa Jorgenson Award, President's Equal Access Scholarship, and Forbes-Lea Research Award at PSU. Outside funding was also provided by the Sigma Xi Grants-in-Aid of Research Copernicus Fund, as well as travel funds from American Society for Microbiology, Pacific Northwest Women in Science, Society for Applied Microbiology, and Society for the Study of Evolution.
TABLE OF CONTENTS

| CHAPTERS |
|---------------------------|------------------|
| 1. Introduction | 1 |
| 2. Horizontally acquired biosynthesis genes boost *Coxiella burnetii*’s physiology | 5 |
| 2.1 Abstract | 5 |
| 2.2 Introduction | 7 |
| 2.3 Materials and Methods | 10 |
| 2.4 Results and Discussion | 13 |
| 3. *Coxiella burnetii* and *Leishmania mexicana* residing within similar parasitophorous vacuoles elicit disparate host responses | 30 |
| 3.1 Abstract | 30 |
| 3.2 Introduction | 32 |
| 3.3 Materials and Methods | 35 |
| 3.4 Results and Discussion | 39 |
| 4. Conclusion | 51 |
REFERENCES ...........................................................................................................................................................................53

APPENDICES

A. Additional Files........................................................................................................................................................................72
B. Genome rearrangements can make and break small RNA genes ..............75
C. Emergence of new sRNAs in enteric bacteria is associated with low expression and rapid evolution ...............................................107
D. Accumulation and expression of multiple antibiotic resistance genes in Arcobacter cryaerophilus that thrives in sewage.................................142
E. Whole-genome sequence of Coxiella burnetii Nine Mile RSA 439 (phase II, clone 4), a laboratory workhorse strain.........................................175
# LIST OF TABLES

## TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Expression of biotin and heme biosyntheses genes in <em>C. burnetii</em></td>
<td>24</td>
</tr>
<tr>
<td>3.1</td>
<td>KEGG pathways enriched in <em>Coxiella</em>-infected and <em>Leishmania</em>-infected THP-1 cells</td>
<td>42</td>
</tr>
<tr>
<td>3.2</td>
<td>MicroRNAs perturbed by <em>Coxiella</em> and <em>Leishmania</em> infections</td>
<td>49</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Biogenesis of <em>Coxiella</em>-containing vacuole (CCV)</td>
</tr>
<tr>
<td>2.1</td>
<td>Horizontally acquired genes in <em>C. burnetii</em></td>
</tr>
<tr>
<td>2.2</td>
<td>Maximum Likelihood tree of CBU_0678, gained via HGT</td>
</tr>
<tr>
<td>2.3</td>
<td>Bayesian tree of CBU_0678, gained via HGT</td>
</tr>
<tr>
<td>2.4</td>
<td>A horizontally acquired fatty acid biosynthesis operon in <em>C. burnetii</em></td>
</tr>
<tr>
<td>2.5</td>
<td>Bayesian tree of CBU_0038, gained via HGT</td>
</tr>
<tr>
<td>2.6</td>
<td>Heme and biotin syntheses effect on <em>C. burnetii</em>’s growth</td>
</tr>
<tr>
<td>2.7</td>
<td>tRNA(^{\text{Glu}})(^2) was horizontally acquired by <em>C. burnetii</em></td>
</tr>
<tr>
<td>3.1</td>
<td>Identification of differentially expressed genes</td>
</tr>
<tr>
<td>3.2</td>
<td>Protein-protein interaction analysis</td>
</tr>
<tr>
<td>3.3</td>
<td>Validation by qPCR of expression levels estimated by RNA-seq</td>
</tr>
<tr>
<td>3.4</td>
<td>Differential expression of mRNA isoforms</td>
</tr>
</tbody>
</table>
Coxiella burnetii is the etiological agent of acute Q fever and a chronic disease commonly manifested as endocarditis (Maurin and Raoult 1999). C. burnetii has a unique biphasic lifestyle. The pathogen persists in the environment as a metabolically quiescent small cell variant (SCV), which transforms into a metabolically active large cell variant (LCV) within an acidic (pH ~4.5) lysosome-derived Coxiella-containing vacuole (CCV) (Figure 1.1; Voth and Heinzen 2007). Coxiella is the only bacterium known to thrive in such a compartment, and this ability is considered critical to its pathogenicity.

CDC has classified Coxiella as a select agent due to its past use as a bioweapon, environmental stability, aerosol transmission, and extremely low infectious dose [ID50 of one to ten bacteria in a guinea pig model (Tigertt et al. 1961)]. Untreated chronic Q fever is associated with high mortality rate (>60%) (Karakousis et al. 2006) and is very difficult to treat because it requires prolonged antibiotic regimen: a combination of doxycycline and hydroxychloroquine (1.5 to 3 years), or doxycycline and fluoroquinolone (3 to 4 years) (Gould et al. 2012). However, strains resistant to doxycycline are prevalent (Rouli et al. 2012). C. burnetii is found worldwide and can cause epidemics, such as a recent epidemic in the Netherlands where thousands of people were infected (van der Hoek et al. 2012). It has also been detected among U.S. military personnel (Faix et al. 2008), and its high prevalence in US environmental samples (Kersh et al. 2010; Loftis et al. 2010) highlights the urgent need to understand its biology and
**Figure 1.1 – Biogenesis of *Coxiella*-containing vacuole (CCV).** The pathogen occurs as a ‘spore-like’ small cell variant (SCV) in the environment. Most human infections occur through inhalation of aerosols contaminated with animal material. Shortly after entering into a host cell (typically an alveolar macrophage), SCV transforms into a metabolically active form called the large cell variant (LCV) within a lysosome-derived, acidic (pH ~4.5) *Coxiella*-containing vacuole (CCV).
virulence mechanisms.

Current research into *Coxiella*’s pathogenicity is mainly focused on host cell entry and on manipulation of the host cell environment through its Type Four Secretion System (TFSS) (e.g., Weber et al. 2013; Newton et al. 2014; Larson et al. 2015, Graham et al. 2015; Seshadri et al. 2003; Beare et al. 2009, Chen et al. 2010). This bias has left a large void in our understanding of *Coxiella*’s intracellular metabolism and host response to *Coxiella* infection. This lack of knowledge of *Coxiella*’s basic biology and molecular pathogenesis is a critical barrier to developing more effective therapies.

In this study, we aimed to understand both bacterial and host factors that have important roles during *C. burnetii* infections. In Chapter 2, using an evolutionary genomics approach, we identified metabolic pathways that are critical to *C. burnetii*’s ability to grow intracellularly. Among those found, the most promising are fatty acid, biotin, and heme biosyntheses pathways. *Coxiella* has horizontally acquired extra copies of genes that enhance these processes; when these genes are disrupted *Coxiella*’s growth in the cell is significantly inhibited. In Chapter 3, by analyzing the host transcriptome, we identified host genes, including microRNA (miRNA) genes that are important during *C. burnetii* infections. *Coxiella* induces the expression of multiple anti-apoptotic miRNAs, which likely have a role in inhibiting apoptosis in order to sustain the intracellular replication of the pathogen. One well-studied example is miR-148a, which inhibits apoptosis by targeting BCL-2, thereby possibly promoting *Coxiella* infection. The biosynthetic pathways and miRNAs identified in this study are ideal targets for
developing more effective therapeutic strategies against Q fever and its chronic and often fatal complications.
CHAPTER 2
Horizontally acquired biosynthesis genes boost *Coxiella burnetii*’s physiology

2.1 ABSTRACT

*Coxiella burnetii*, the etiologic agent of acute Q fever and chronic endocarditis, has a unique biphasic life cycle, which includes a metabolically active intracellular form that occupies a large lysosome-derived acidic vacuole. *C. burnetii* is the only bacterium known to thrive within such an hostile intracellular niche, and this ability is fundamental to its pathogenicity; however, very little is known about genes that facilitate *Coxiella*’s intracellular growth. Recent studies indicate that *C. burnetii* evolved from a tick-associated ancestor and that the metabolic capabilities of *C. burnetii* are different from that of *Coxiella*-like bacteria found in ticks. Horizontally acquired genes that allow *C. burnetii* to infect and grow within mammalian cells likely facilitated the host shift; however, because of its obligate intracellular replication, *C. burnetii* would have lost most genes that have been rendered redundant due to the availability of metabolites within the host cell. Based on these observations, we reasoned that horizontally derived biosynthetic genes that have been retained in the reduced genome of *C. burnetii* are ideal candidates to begin to uncover its intracellular metabolic requirements. Our analyses identified a large number of putative genes of foreign-origin in *C. burnetii*, including tRNA\textsuperscript{Glu} that is potentially required for heme biosynthesis, and genes involved in the production of lipopolysaccharide — a virulence factor, and of critical metabolites such as
fatty acids and biotin. In comparison to wild-type *C. burnetii*, a strain that lacks tRNA$^{\text{Glu2}}$ exhibited reduced growth, indicating its importance to *Coxiella’s* physiology.

Additionally, by using chemical agents that block heme and biotin biosyntheses, we show that these pathways are promising targets for the development of new anti-*Coxiella* therapies.
2.2 INTRODUCTION

*Coxiella burnetii* is the etiological agent of acute Q fever and a chronic disease commonly manifested as endocarditis (Maurin and Raoult 1999). Most human infections occur through inhalation of aerosols originating from ruminants that shed *C. burnetii* during parturition and in milk. The pathogen persists in the environment as a metabolically quiescent small cell variant (SCV), which transforms into a metabolically active large cell variant (LCV) within a lysosome-derived, acidic (pH ~4.5), *Coxiella*-containing vacuole (CCV) (Voth and Heinzen 2007). The unique ability of *C. burnetii* to thrive in this inhospitable vacuole is fundamental to its physiology and pathogenicity; however, the metabolic processes that allow its intracellular growth are unknown.

The evolutionary origin of *C. burnetii* is also not clearly understood. The closest relatives of *C. burnetii* are tick-associated bacteria, indicating that *C. burnetii* evolved from a tick-associated ancestor (Smith et al. 2015; Duron et al. 2015; Gottlieb et al. 2015). Interestingly, *Coxiella*-like bacteria found in ticks cannot infect mammalian cells and are unable to grow in ACCM-2, a culture medium that supports robust growth of *C. burnetii* (Duron et al. 2015; Omsland et al. 2011). These observations suggest that despite their close evolutionary relationship, the human pathogen and the tick-associated strains have different virulence and metabolic capabilities. Concomitantly, when the genomes of *C. burnetii* were sequenced, it became clear that the pathogen has acquired several virulence and metabolic genes via horizontal gene transfer (HGT) (Raghavan et al. 2008). For example, *C. burnetii* contains a tryptophan biosynthesis operon of Chlamydial origin,
and a Type Four Secretion System (TFSS) and eukaryote-like ankyrin repeat sequence-containing effector proteins that are essential for CCV generation (Seshadri et al. 2003; Beare et al. 2009). Horizontal acquisition of foreign DNA occurs in bacteria through transformation, conjugation or transduction via mobile genetic elements such as plasmids, integrons, bacteriophages, transposons, retrotransposons etc (Ochman et al. 2000). The novel DNA is stably maintained and spreads through the recipient population if it offers selective advantage (e.g. antibiotic resistance), allowing the bacterium to adapt to the new environment (Eisen 2000; Frost et al. 2005; Thomas and Nielsen 2005).

Unlike eukaryotic genomes, which contain large fractions of nonfunctional DNA (e.g., >80% of human genome), bacterial genomes are tightly packed with functional genes (Moran 2002; Ochman and Davalos 2006). In bacteria there is a bias towards deletion over insertion, hence, DNA is retained in a bacterial genome only if selection is acting effectively to preserve it (Mira et al. 2001). For instance, when the tick-associated ancestor of C. burnetii evolved into a mammalian pathogen that replicates only within the CCV, several biosynthetic genes would be rendered redundant if corresponding metabolic intermediates are available within the host cell. In addition, although genes acquired via HGT might have been critical in facilitating the host shift, many of them could become expendable in the new lifestyle (e.g. Lo et al. 2015). These superfluous metabolic and HGT-origin genes would subsequently be deleted from the genome due to a lack of selection pressure to maintain them (Ochman and Moran 2001). Furthermore, the intracellular niche limits C. burnetii’s opportunity to gain new genes from the environment via HGT. Thus, intracellular pathogens such as C. burnetii tend to have
reduced genomes in comparison to related free-living bacteria (e.g., *C. burnetii*’s genome is ~2 million bp, whereas *E. coli*’s is ~5 million bp). Based on these observations, we reasoned that a significant number of HGT-origin genes that have been retained in *C. burnetii* would be critical to its intracellular fitness. In this study, we identified a large number of horizontally derived genes, including those for the synthesis of LPS, fatty acids, heme, and biotin that augment the physiological capability of *C. burnetii*. 
2.3 MATERIALS AND METHODS

**C. burnetii growth assay**

*C. burnetii* Nine Mile phase II RSA 439 was grown in ACCM-2 medium as described previously (Omsland et al. 2011; Warrier et al. 2014), and incubated at 37°C, 2.5% O₂, and 5.0% CO₂ using a Galaxy 170 R incubator (New Brunswick Scientific, NJ). Chloramphenicol (8 µg/ml) and/or Kanamycin (375 µg/ml) were added as necessary. Growth was measured using PicoGreen as described previously (Martinez et al. 2015). Briefly, 50µl of culture was mixed with 5µl of Triton X-100 Surfact-Amps 10% detergent solution (Thermo Scientific) in 96-well black-bottom Cliniplates (Thermo Scientific), and allowed to incubate at room temperature for 10 minutes with shaking. PicoGreen (Life Technologies) was diluted 1:200 in TE buffer and 55µL was added to the wells, and incubated at room temperature with shaking for 5 minutes. Wells were excited at 495 nm and emission was read at 519 nm using a Victor X5 2030 Multiplate Reader (Perkin Elmer). To determine whether *Coxiella* can utilize external heme as its sole iron source, ferrous sulfate was either omitted from ACCM-2 preparations or substituted with hemin (Alfa Aesar), and to assay the importance of heme and biotin biosyntheses, gabaculine (Enzo Life Sciences) or MAC13772 (Maybridge) was added to ACCM-2. Briefly, a 10mM stock of hemin was made in 1.5M NaOH, a 300mg/mL stock of MAC13772 was made in dimethyl sulfoxide (DMSO), and a 100 mM stock of gabaculine was made in distilled water. One µl of solution from these stocks was added to 1ml of ACCM-2 to attain final concentrations of 10µM hemin, 300µg/ml MAC13772,
or 100µM gabaculine. ACCM-2 containing the same amount of solvent was used as control, and growth was measured after 7 days using PicoGreen as described above.

**Generation of tRNA\textsuperscript{Glu}\textsubscript{2} deletion and complementation strains**

To delete tRNA\textsuperscript{Glu}\textsubscript{2}, ~1200 bp on each side of the gene along with Chloramphenicol acetyltransferase (CAT) gene was cloned into the vector pJC-Kan, and the mutant was generated as described previously (Beare et al. 2012). Insertion of the CAT gene in place of tRNA\textsuperscript{Glu}\textsubscript{2} was confirmed using PCR and DNA sequencing. To generate a complementation strain, we used pKM244, a low-copy plasmid that is maintained stably in *C. burnetii* (Chen et al. 2010). Because the tRNA\textsuperscript{Glu}\textsubscript{2}-deletion strain already contained a CAT gene, we amplified a kanamycin resistance gene under the control of the 1169 promoter from pJB-Kan (Beare et al. 2012) and cloned it into pKM244 using NheI and AatII to generate pAM100. We inserted tRNA\textsuperscript{Glu}\textsubscript{2} along with its flanking intergenic regions into pAM100 using BamHI. *C. burnetii* was transformed (400 ohms, 2.5 kV, 25 mF) with either empty pAM100 or pAM100 with cloned tRNA\textsuperscript{Glu}\textsubscript{2}, as described previously (Omsland et al. 2011; Beare et al. 2012). Growth in ACCM-2 of wild-type (with empty pAM100), deletion (with empty pAM100) and complementation strains were measured at day-7 using PicoGreen as described above.

**Detection of horizontal gene transfer**

Horizontally acquired genes were identified using HGTector (Zhu et al. 2014). *Coxiella* was set as self-group, and Legionellales was set as exclusion group. BLAST parameter
thresholds were set at 70% identity and an E-value of a maximum of $1 \times 10^{-5}$. To validate the HGT data, phylogenetic analyses were conducted for several genes projected to be the result of HGT. To this end, TBLASTN search was conducted against the NCBI nr database and the top 100 matches with at least 30% identity, 70% coverage, and E-value of less than $1 \times 10^{-10}$ were chosen. Sequence alignment was performed using Clustal Omega (Sievers et al. 2011), and ambiguously aligned regions were removed using Gblocks (Talavera and Castresana 2007). The evolution model GTR+I+G (General Time Reversible plus Invariant sites plus Gamma distribution) was selected using jModelTest2 (Darriba et al. 2012). Bayesian trees were constructed using MrBayes as implemented in Geneious (Huelsenbeck and Ronquist 2001; Kearse et al. 2012). A chain length of 1,000,000 was used with a burn-in fraction of 25% and sampling every 100 trees. Maximum Likelihood trees were constructed using RAxML (Stamatakis et al. 2008) as implemented in Geneious with 1,000 bootstrap replicates.
2.4 RESULTS AND DISCUSSION

Identification of horizontally acquired genes in *C. burnetii*

HGT is a major driver of evolution and adaptation in bacteria (Lerat et al. 2005; Price et al. 2008; Treangen and Rocha 2011). By examining differences in nucleotide composition (e.g. GC%) HGT-origin genes can be tentatively identified; however, because most transfers occur between closely related bacteria, and because genes from distant organisms will evolve over time to reflect the base composition of the recipient genome (Lawrence and Ochman 1997), this approach is not always effective. Another common approach is to perform reciprocal-BLAST analysis (e.g. Raghavan et al. 2012), where, if the top hit is from a distantly related organism, the gene is of putative HGT origin. While this approach provides rapid results, the best BLAST hit may not always be the closest phylogenetic relative of the query gene (Koski and Golding 2001). This is especially the case for organisms such as *C. burnetii* that have few close relatives represented in public NCBI databases. For instance, *C. burnetii* is the only defined species within its genus, and complete genomes are available for only three other bacteria within the order Legionellales — *Legionella, Rickettsiella*, and *Diplorickettsia*. In order to circumvent this limitation, we utilized a phylogeny-informed BLAST-based approach (HGTector, Zhu et al. 2014), but applied very strict criteria to categorize a gene as horizontally acquired. We set *Coxiella* as the self-group, and Legionellales as the exclusion group, which captured HGT events where only *Coxiella* has acquired a particular gene from outside of Legionellales and ignored any events where the genes
could also have been transferred elsewhere within the order. In addition, because the phylogenetic position of Legionellales within Gammaproteobacteria is not well resolved (Williams et al. 2010), we also ignored any gene that was potentially gained from another member of the class Gammaproteobacteria. Using this ultra-conservative approach, we were able to identify 172 ‘high-confidence’ horizontally acquired genes on the chromosome of *C. burnetii* RSA 493 (Figure 2.1, Additional_file_A1.csv), whereas, none of the genes located on *C. burnetii* plasmids (QpH1, QpRS, QpDG) were deemed to be of horizontal origin.

Orthologs of all 172 HGT-origin genes are present in *C. burnetii* Dugway 5J108-111 (NC_009727.1), *C. burnetii* 3262 (CP013667.1) and *C. burnetii* Z3055 (NZ_LK937696.1), however orthologs of CBU_1991 were not detected in *C. burnetii* RSA 331 (CP000890.1), CBU_2021 and CBU_0562 in *C. burnetii* G_Q212 (NC_011527.1), and CBU_0007a, CBU_0167, CBU_0168, CBU_0768 and CBU_0792 in *C. burnetii* K_Q154 (NC_011528.1). Of the 172 HGT genes, 18 are located close to transposase genes in *C. burnetii* RSA 493, suggesting a role for transposons or insertion sequences in acquiring foreign genes. In addition to the large number of transposases (>30), all *C. burnetii* strains contain several other selfish genetic elements that are horizontally exchanged between bacteria, including two Group I introns, an intein, and an intervening sequence (Seshadri et al. 2003; Beare et al. 2009; Raghavan et al. 2007; Raghavan et al. 2008; Warrier et al. 2016). Proliferation of mobile genetic elements and extensive genome rearrangements are hallmarks of bacteria that have recently shifted to host-associated lifestyles, indicating that the obligate intracellular growth of *C. burnetii*
Figure 2.1 – Horizontally acquired genes in *C. burnetii*. Two outer rings show ORFs (black bars) on forward and reverse strands, respectively. Third ring shows positions of 172 HGT-origin genes, and the inner ring contains biosynthetic genes examined in this study. Genes are colored according to their putative donors, as shown in the center. Number of genes acquired from each donor is also shown.
is of recent origin (Ochman and Moran 2001; McCutcheon and Moran 2012).

As expected, our analysis excluded several genes that were likely acquired via HGT because they are also present in other members of the order Legionellales e.g. TFSS genes, eukaryote-like sterol reductase genes, plasmid genes with eukaryotic domains etc. (Seshadri et al. 2003; Beare et al. 2009; Gilk et al. 2010; Voth et al. 2013). Without better resolution of *C. burnetii*’s phylogenetic position in the bacterial tree, it is difficult to discern whether these genes were acquired from a common ancestor or were gained independently by each bacterium as they adapted to their respective intracellular niches (Gottlieb et al. 2015). Intriguingly, *C. burnetii* encodes several intact or pseudogenized genes that encode the components of a Type IV pilus (Seshadri et al. 2013). A related Type IV pilus enables *Acinetobacter baumannii* to acquire DNA from the environment with high efficiency (Smith et al. 2007), indicating that during an earlier stage during *C. burnetii*’s evolution, a functional Type IV pilus endowed it with the ability to acquire foreign genes proficiently, thereby likely facilitating its transition into a mammalian pathogen from a tick-associated ancestor (Smith et al. 2015; Duron et al. 2015; Gerhart et al. 2016).

**HGT contributed to the lipopolysaccharide profile of *C. burnetii***

HGT-origin genes are strewn all across the *C. burnetii* genome, indicating, as shown previously (Raghavan et al. 2008), that there are no prototypical pathogenicity islands, but, based on their clustered genome locations, several genes appear to have been acquired en bloc (*Figure 2.1*). A group of genes of particular interest is CBU_0678 to
CBU_0683, which is part of an operon that encodes genes involved in the biosynthesis of LPS (Seshadri et al. 2003; Narasaki and Toman 2012). We conducted in-depth phylogenetic analyses of CBU_0678, the most upstream gene of this cluster in order to validate the HGTector results. Based on both Maximum Likelihood and Bayesian phylogenetic analyses (Figure 2.2, Figure 2.3), the closest orthologs of this gene are present in members of Alphaproteobacteria. Furthermore, O-polysaccharide biosynthesis genes tend to occur as an operon in most bacteria, and since several genes in this location (CBU_0673, CBU_0676, CBU_0678 to CBU_0682) were also likely acquired from Alphaproteobacteria (Figure 2.2), it is highly likely that the genes were acquired in a single event. Previous studies have shown that LPS genes are horizontally transferred among bacteria (Nelson and Selander 1994), probably due to their importance in host-pathogen interactions (Narasaki and Toman 2012). In fact, full-length LPS is the only C. burnetii virulence factor established in an immunocompetent animal model of infection. It protects the pathogen from innate immune response (Shannon et al. 2005), and avirulent Nine Mile phase II strain produces a severely truncated LPS due to the loss of 22 LPS biosynthesis genes, including CBU_0678 to CBU_0682 (Beare et al. 2006). The observation that important virulence factors in C. burnetii (e.g., LPS, TFSS, effector proteins) were assembled via HGT illustrates the significance of this process in the evolutionary history of this intracellular pathogen.

**HGT enhanced C. burnetii’s fatty acid metabolism**

Another set of genes in C. burnetii that is of putative foreign origin is CBU_0034 to
**Figure 2.2 – Maximum Likelihood tree of CBU_0678, gained via HGT.** A Maximum Likelihood phylogeny for CBU_0678, a LPS biosynthesis gene, is shown. Gammaproteobacteria are in red, Alphaproteobacteria in blue, Delta/Epsilonproteobacteria in green, and Planctomycetes in black. Bootstrap values of >50 are indicated at the nodes. Branch length measured in number of substitutions per site. The evolution model GTR+I+G (General Time Reversible plus Invariant sites plus Gamma distribution) was selected using jModelTest2 (Darriba et al. 2012). Maximum Likelihood tree was constructed using RAxML (Stamatakis et al. 2008) as implemented in Geneious with 1,000 bootstrap replicates.
Figure 2.3 - Bayesian tree of CBU_0678, gained via HGT. A Bayesian phylogenetic tree of CBU_0678. Posterior probability shown at nodes. *C. burnetii* highlighted in red. Number of taxa collapsed into each branch is shown within parentheses. Branch length measured in number of substitutions per site. The evolution model GTR+I+G (General Time Reversible plus Invariant sites plus Gamma distribution) was selected using jModelTest2 (Darriba et al. 2012). Bayesian tree was constructed using MrBayes as implemented in Geneious (Huelsenbeck and Ronquist 2001; Kearse et al. 2012). A chain length of 1,000,000 was used with a burn-in fraction of 25% and sampling every 100 trees.
CBU_0038 (Figure 2.4). This gene cluster appears to have originated in either a Deltaproteobacteria or a Spirochete. We explored the HGT origin of this presumptive operon through phylogenetic analysis of its first gene, CBU_0038 (Figure 2.4A, Figure 2.5). Furthermore, this set of genes has a similar arrangement in both *C. burnetii* and in *Spirochaeta africana* DSM 8902, its best BLAST hit (Additional_file_A1.csv), indicating that the genes were transferred en bloc. In addition, an IS1111A transposase (CBU_0040) is proximally located to the operon, illustrating a probable role for this mobile genetic element in HGT (Figure 2.4B). Based on homology, CBU_0034 to CBU_0038 encode ACP, FabB, FabZ, FabA and FabH, respectively, which are involved in the synthesis of unsaturated fatty acids (Zhang and Rock 2008; Feng and Cronan 2009). In contrast, vertically inherited genes (CBU_0493 to CBU_0497) are responsible for the synthesis of saturated fatty acids (Gilk 2012), denoting that *C. burnetii*’s ability to produce unsaturated fatty acids was enhanced through HGT. *C. burnetii* also contains a putative fatty acid desaturase (CBU_0920) that introduces double bonds into existing fatty acids (Gilk 2012). This gene also appears to be of HGT origin, but was not included in our analysis because it is present in other Legionellales. In addition to fatty acid biosynthesis genes, *C. burnetii* has also gained three pyruvate dehydrogenases (CBU_0686, CBU_0692, and CBU_0693) (Figure 2.1). Pyruvate dehydrogenase complex converts pyruvate into acetyl-CoA, which is a critical metabolite for both fatty acid biosynthesis and ATP generation (de Kok et al. 1998). Thus, HGT appears to have played a significant role in shaping *C. burnetii*’s fatty acid metabolism. Furthermore, a recent study showed that transposon-mediated disruption of CBU_0035 and CBU_0038
Figure 2.4 – A horizontally acquired fatty acid biosynthesis operon in *C. burnetii*.

(A) Maximum Likelihood phylogenetic tree for CBU_0038, a fatty acid biosynthesis gene. Gammaproteobacteria is colored red, Delta/Epsilonproteobacteria in green, Spirochaetes in brown, Bacteriodetes in purple, and Firmicutes in blue. Bootstrap values of > 50 are indicated at the nodes. Branch length measured in number of substitutions per site. The evolution model GTR+I+G (General Time Reversible plus Invariant sites plus Gamma distribution) was selected using jModelTest2 (Darriba et al. 2012). Maximum Likelihood tree was constructed using RAxML (Stamatakis et al. 2008) as implemented in Geneious with 1,000 bootstrap replicates. (B) Fatty acid biosynthesis genes have similar arrangement in *C. burnetii* and Spirochaeta africana, and an IS1111A transposase is located next to the operon in *C. burnetii*. 
Figure 2.5 – Bayesian tree of CBU_0038, gained via HGT. A Bayesian phylogenetic tree of CBU_0038. Posterior probability shown at nodes. *C. burnetii* highlighted in red. Number of taxa collapsed into each branch is shown within parentheses. Branch length measured in number of changes per site. The evolution model GTR+I+G (General Time Reversible plus Invariant sites plus Gamma distribution) was selected using jModelTest2 (Darriba et al. 2012). Bayesian tree was constructed using MrBayes as implemented in Geneious (Huelsenbeck and Ronquist 2001; Kearse et al. 2012). A chain length of 1,000,000 was used with a burn-in fraction of 25% and sampling every 100 trees.
resulted in reduced *C. burnetii* growth within Vero cells, suggesting that this process is a critical constituent of the pathogen’s physiology (Martinez et al. 2014).

**HGT enhanced *C. burnetii*’s biotin metabolism**

Fatty acid biosynthesis enzymes require biotin as a cofactor, and biotin in turn is synthesized by utilizing a portion of the fatty acid biosynthesis pathway (Lin et al. 2010). Unlike most other bacteria, *C. burnetii* contains two copies of the gene *bioC*, the first committed step in biotin production. The two *bioC* genes are only 45% similar at the nucleotide level, indicating that they were not formed by a recent duplication event in *C. burnetii*. Further, *bioC.2* (CBU_1004) has an 11 bp overlap with *bioH* and is part of the *bioA-bioBFHCD-birA* regulon (CBU_1008 to CBU_1002), whereas *bioC.1* (CBU_0467) is a single gene located at a different part of the genome, indicating that *bioC1* is of horizontal origin. To understand the functional relevance of these genes, we examined their expression levels in *C. burnetii* grown in ACCM-2 and within Vero cells (Warrier et al. 2014). The RNA-seq data revealed that all biotin biosynthesis genes (*bioA, bioB, bioC.1, bioC.2, bioD, bioF*, and *bioH*) were expressed under both conditions, but their expression was significantly higher in Vero cells than in ACCM-2 (*Table 2.1*). In contrast, the expression of *birA*, the transcriptional repressor of biotin operon showed the opposite pattern, suggesting that biotin is likely synthesized under both conditions, with possible upregulation within the host cell. Furthermore, a small molecule (MAC13772) that blocks biotin biosynthesis in *E. coli* (Zlitni et al. 2013) inhibited the growth of *C. burnetii* in ACCM-2, indicating that the production of biotin is a critical process in this
Table 2.1 - Expression of biotin and heme biosyntheses genes in *C. burnetii* grown in ACCM-2 or Vero cells for 72 h measured using RNA-seq.

<table>
<thead>
<tr>
<th>Gene</th>
<th>ACCM-2 (LCV)</th>
<th>Vero (LCV)</th>
<th>Fold Change</th>
<th>p-value (FDR corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biotin</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bioA</td>
<td>2,093</td>
<td>3,915</td>
<td>1.87</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>bioB</td>
<td>1,562</td>
<td>10,118</td>
<td>6.48</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>bioC.1</td>
<td>9,659</td>
<td>15,678</td>
<td>1.62</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>bioC.2</td>
<td>1,562</td>
<td>2,909</td>
<td>1.86</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>bioD</td>
<td>1,590</td>
<td>6,233</td>
<td>3.92</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>bioE</td>
<td>1,159</td>
<td>5,354</td>
<td>4.62</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>bioH</td>
<td>916</td>
<td>4,836</td>
<td>5.28</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>birA</td>
<td>5,803</td>
<td>3,961</td>
<td>0.68</td>
<td>0.0055</td>
</tr>
<tr>
<td><strong>Heme</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hemA</td>
<td>6,447</td>
<td>10,335</td>
<td>1.60</td>
<td>0.0007</td>
</tr>
<tr>
<td>hemB</td>
<td>49,632</td>
<td>5,778</td>
<td>0.12</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>hemC</td>
<td>6,652</td>
<td>4,783</td>
<td>0.72</td>
<td>0.0183</td>
</tr>
<tr>
<td>hemD</td>
<td>4,534</td>
<td>6,420</td>
<td>1.42</td>
<td>0.0110</td>
</tr>
<tr>
<td>hemE</td>
<td>29,442</td>
<td>16,762</td>
<td>0.57</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>hemF</td>
<td>9,104</td>
<td>12,200</td>
<td>1.34</td>
<td>0.0291</td>
</tr>
<tr>
<td>hemH</td>
<td>3,343</td>
<td>8,085</td>
<td>2.42</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>hemK</td>
<td>3,497</td>
<td>6,141</td>
<td>1.76</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
pathogen (Figure 2.6). Biotin synthesis has been shown to be critical to the virulence of other human pathogens such as *Francisella tularensis* and *Mycobacterium tuberculosis* (Woong Park et al. 2011; Feng et al. 2014). Thus, novel pharmaceutical agents that block this process hold promise as a potential broad-spectrum agent to treat intracellular pathogens.

**Heme biosynthesis is an essential metabolic process in *C. burnetii***

*C. burnetii*’s genome is small compared to that of free-living bacteria such as *E. coli* (~2 Mb and ~5 Mb, respectively). Correspondingly, it contains less than half the number of tRNAs than in *E. coli* (42 and 89, respectively). This reduction has occurred due to loss of redundant tRNAs; for instance, while *E. coli* has five copies of tRNA<sub>Ile</sub>, *C. burnetii* has only one copy. In contrast to other tRNAs, *C. burnetii* contains an additional tRNA<sub>Glu</sub> isoacceptor (tRNA<sub>Glu</sub>2, anticodon CUC) that is not found in any other Gammaproteobacteria, denoting that it was gained via HGT (Figure 2.7A). Additionally, a toxin-antitoxin system (CBU_0284, CBU0285) that was probably horizontally acquired is located adjacent to tRNA<sub>Glu</sub>2, further strengthening the evidence for its putative HGT origin (Figure 2.7B). Because tRNA<sub>Glu</sub>2 is retained in such a streamlined genome, it most likely has a critical function; in fact, a tRNA<sub>Glu</sub>2-deletion strain had significantly slower growth than the wild-type strain in ACCM-2, which was restored in a complementation strain, signifying the importance of the HGT-derived tRNA to optimum bacterial fitness (Figure 2.6). However, the major function of tRNA<sub>Glu</sub>2 is unlikely to be protein synthesis because tRNA<sub>Glu</sub>2 (anticodon CUC) cannot decode the major glutamate
Figure 2.6 – Heme and biotin syntheses effect on *C. burnetii*’s growth. Growth of *C. burnetii* in ACCM-2 after 7 days was measured using PicoGreen. Fluorescence of each strain relative to that of control (wild-type grown in ACCM-2; dashed line) is shown. MAC13772: ACCM-2 supplemented with 300 µg/ml MAC13772, a biotin biosynthesis inhibitor; tRNAGlu2: tRNAGlu2-deletion strain; Complement: tRNAGlu2-deletion complemented with intact tRNAGlu2 on pAM100; No Fe: ACCM-2 without FeSO4; Hemin: ACCM-2 with hemin in place of FeSO4; Gabaculine: ACCM-2 supplemented with 100 µM gabaculine, a heme biosynthesis inhibitor. Statistically significant differences in growth from control are indicated by (**) p < 0.001 and (*) p < 0.01 (unpaired t-test).
Figure 2.7 – \( tRNA^{\text{Glu}}_2 \) was horizontally acquired by \textit{C. burnetii}. (A) 16S rDNA Neighbor-Joining phylogenetic tree of Gammaproteobacteria showing the presence (+) or absence (−) of \( tRNA^{\text{Glu}} \) (anticodon CUC). (B) A putative toxin/antitoxin system is located adjacent to \( tRNA^{\text{Glu}}_2 \).
codon GAA (76%), whereas tRNA$^{\text{Glu}}_1$ (anticodon UUC) can decode both GAA and GAG codons efficiently. Additionally, Coxiella proteins are not enriched for glutamate in comparison to *E. coli* (~6% of all codons is for Glu in both bacteria). Taken together, these data suggest that protein biosynthesis is not tRNA$^{\text{Glu}}_2$’s major function in *C. burnetii*.

An alternate function for tRNA$^{\text{Glu}}_2$ is in heme biosynthesis. Heme is an iron-containing tetrapyrrole that serves multiple cellular functions, including in respiration, energy generation, oxidative reactions and signal transduction (Almiron et al. 2001). The universal precursor of tetrapyrrole biosynthesis is 5-aminolevulinic acid (ALA) (Anzaldi and Skaar 2010). There are two alternate pathways in nature for the synthesis of ALA: C5 pathway found in most bacteria, and Shemin pathway present in most eukaryotes, including humans (Frankenberg et al. 2003). The starting point of the C5 pathway is Glutamyl-tRNA$^{\text{Glu}}$, which is converted into ALA using two consecutive enzymes HemA and HemL. In Shemin pathway, ALA is synthesized from Succinyl-CoA and glycine by the enzyme ALAS (Frankenberg et al. 2003). The remaining steps are shared between the C5 and Shemin pathways. We examined the *C. burnetii* genome and discovered that it encodes an intact C5 heme biosynthesis pathway, and RNA-seq data confirmed that these genes are expressed during growth (Table 2.1). Additionally, heme biosynthesis pathway is conserved in all *C. burnetii* strains but is absent in non-pathogenic Coxiella present in ticks, and a *C. burnetii* strain with a transposon insertion in the hemL gene had reduced growth in Vero cells (Martinez et al. 2014), implying the importance of heme biosynthesis to the human pathogen.
Several bacteria augment heme biosynthesis with heme transported from the outside (Mike et al. 2013). *C. burnetii* encodes a transporter (*feoAB*) for ferrous iron, which is required for heme biosynthesis, but has no known transporters for ferric ions or heme. Consequently, we reasoned that the pathogen might not be able to utilize extracellular heme as its sole source of iron. To test this, we inoculated equal amounts of *Coxiella* into ACCM-2 medium with FeSO4 (standard recipe) or into ACCM-2 in which FeSO4 was replaced with equimolar amount of hemin. Growth was measured after 7 days using PicoGreen as described previously (Martinez et al. 2015). As shown in Figure 2.6, *Coxiella* growth in FeSO4-containing ACCM-2 was significantly higher than in hemin-containing ACCM-2 (p < 0.001, paired t-test), whereas the fluorescence measurements between the control (ACCM-2 without FeSO4) and the hemin-containing ACCM-2 samples were not significantly different (p > 0.05, paired t-test), showing that *C. burnetii* cannot utilize external heme. To test *C. burnetii*’s requirement for heme biosynthesis, we treated *C. burnetii* with increasing concentrations of gabaculine, an inhibitor of HemL (Wang et al. 1997), and found that 100 uM gabaculine significantly inhibited bacterial growth (p < 0.001, paired t-test) (Figure 2.6). These data show the probable of heme biosynthesis to *C. burnetii*’s normal physiology, and indicates that heme biosynthesis genes are potential targets for the development of new anti-*Coxiella* therapies.
CHAPTER 3

Coxiella burnetii and Leishmania mexicana residing within similar parasitophorous vacuoles elicit disparate host responses

3.1 ABSTRACT

Coxiella burnetii is a bacterium that thrives in an acidic parasitophorous vacuole (PV) derived from lysosomes. Leishmania mexicana, a eukaryote, has also independently evolved to live in a morphologically similar PV. As Coxiella and Leishmania are highly divergent organisms that cause different diseases, we reasoned that their respective infections would likely elicit distinct host responses despite producing phenotypically similar parasite-containing vacuoles. The objective of this study was to investigate, at the molecular level, the macrophage response to each pathogen. Infection of THP-1 (human monocyte/macrophage) cells with Coxiella and Leishmania elicited disparate host responses. At 5 days post-infection, 1057 genes were differentially expressed (746 genes up- and 311 genes down-regulated) in C. burnetii infected cells when compared to uninfected cells, whereas 698 genes (534 genes up- and 164 genes down-regulated) were differentially expressed in L. mexicana infected cells. Interestingly, of the 1755 differentially expressed genes identified in this study, only 126 genes (~7%) are common to both infections. We also discovered that 1090 genes produced mRNA isoforms at significantly different levels under the two infection conditions, suggesting that alternate proteins encoded by the same gene might have important roles in host response to each
infection. Additionally, we detected 257 micro RNAs (miRNAs) that were expressed in THP-1 cells and identified miRNAs that were specifically expressed during *Coxiella* or *Leishmania* infections. Collectively, this study identified host mRNAs and miRNAs that were influenced by *Coxiella* and/or *Leishmania* infections, and our data indicate that although their PVs are morphologically similar, *Coxiella* and *Leishmania* have evolved different strategies that perturb distinct host processes to create and thrive within their respective intracellular niches.
3.2 INTRODUCTION

Macrophages that phagocytize pathogens and recruit other immune cells are critical for the elimination of potential infections. Within macrophages, engulfed pathogens are transported inside phagosomes that later fuse with lysosomes to generate the phagolysosome. Most pathogens are degraded within the phagolysosome, which has a very harsh environment (low pH, high concentration of lysosomal hydrolases, presence of cationic peptides etc.) (Kinchen and Ravichandran 2008; Flannagan et al. 2009).

Several pathogens have evolved strategies to survive and replicate within macrophages: *Toxoplasma gondii* prevents the fusion of its vacuoles with the endosomal pathway; *Salmonella enterica* Typhimurium, *Mycobacterium tuberculosis* and *Legionella pneumophila* block maturation of phagosomes into phagolysosomes; *Shigella flexneri* and *Listeria monocytogenes* escape into cytoplasm from phagosomes before lysosomal fusion; *Trypanosoma cruzi* escapes from phagosomes after fusion with lysosomes (Flannagan et al., 2009; Swanson and Fernandez-Moreia, 2002).

Unlike most other pathogens, *Coxiella* (a bacterium) and *Leishmania* (a eukaryote) have independently evolved the ability to thrive in a parasitophorous vacuole (PV) that is derived from the fusion of phagosomes with lysosomes (Voth and Heinzen 2007; Alix et al. 2011). *Coxiella burnetii* (the only defined species within this genus) causes human Q fever and chronic endocarditis. The bacterium is shed in milk, urine, and birth products of animals, and can survive in the environment via a “spore-like” form called the small cell variant (SCV). *C. burnetii* is usually acquired via inhalation, and
initially infects alveolar macrophages but then spreads to mononuclear phagocytes of other tissues. Within the macrophage, SCV transforms into a metabolically active form called the large cell variant (LCV), and multiple *Coxiella*-containing vacuoles merge to form a single large vacuole that fuses with endolysosomal vesicles to give rise to the mature *Coxiella* PV (van Schaik et al. 2013).

*Leishmania* is a genus of trypanosomatid parasite that comprises several species of medical and veterinary importance that cause cutaneous, mucocutaneous, or visceral diseases. It has a dimorphic lifecycle that alternates between an extracellular promastigote form in insect vectors and an intracellular amastigote form in mammalian hosts (Herwaldt 1999). The primary host cells of *Leishmania* are macrophages, but it can also infect neutrophils, fibroblasts and dendritic cells (Contreras et al. 2014). Similar to the biogenesis of *Coxiella* PV, the *Leishmania*-containing vacuole also fuses with endolysosomal vesicles to give rise to the mature *Leishmania* PV. However, the morphology of PV varies among different *Leishmania* species. In several species, including *L. donovani*, *L. infantum*, and *L. major*, only one or two amastigotes reside within each PV, which segregates into new vacuoles after parasite replication. In contrast, as observed for *Coxiella* PVs, parasites of the *Leishmania mexicana* complex such as *L. mexicana* and *L. amazonensis* form communal PVs that continuously enlarge as the parasites replicate (Real et al. 2010). Interestingly, coinfection studies have shown that PVs formed by *L. amazonensis* amastigotes can fuse with *C. burnetii* PVs but not with PVs containing *L. major* amastigotes, suggesting that the intracellular niches generated by *Leishmania mexicana* complex parasites and *Coxiella* may be compositionally rather
similar (Real et al. 2010; Beare et al. 2011; Veras et al. 1995; Rabinovitch and Veras 1996; Newton and Roy 2011).

Both *Coxiella* and *Leishmania* actively participate in the creation of their respective PVs (Voth and Heinzen 2007; Waller and McConville 2002). To begin to understand how the two distantly related pathogens generate phenotypically similar PVs, we compared host gene expression in human macrophage cells (THP-1) infected with either *C. burnetii* or *L. mexicana*. Our data show that the bacterium and the eukaryote elicit distinct host messenger RNA (mRNA) and microRNA (miRNA) responses, indicating that despite their superficial similarity, generation and maintenance of the *Coxiella* PV and *Leishmania* PV involve distinct host processes.
3.3 MATERIALS AND METHODS

*C. burnetii* and *L. mexicana* infection of THP-1 cells, RNA extraction and RNA-seq

THP-1 human macrophage cells (TIB-202; ATCC) were maintained in RPMI 1640 medium (Gibco) supplemented with 10% fetal calf serum (Gibco) at 37°C in 5% CO₂. Cells were incubated in the presence of 200 nM phorbol 12-myristate 13-acetate (PMA; EMD Biosciences) for 24 h to induce differentiation into adherent, macrophage-like cells. Prior to infection, PMA-containing medium was replaced with fresh RPMI without PMA. Cells were infected with either *C. burnetii* (Nine Mile phase II, RSA 493) or promastigotes of *L. mexicana* (MNYZ/BZ/62/M379) at an approximate multiplicity of infection of 25 and incubated for five days. Growth medium was replaced every two days and formation of *Coxiella* and *Leishmania* PVs was monitored microscopically. At 5 d post-infection, growth medium was replaced with 1 ml of TRI reagent (Life Technologies) and total RNA was extracted, genomic DNA was removed by DNase (Life Technologies) treatment, as per instructions. RNA from two samples each of uninfected, *Coxiella*-infected, and *Leishmania*-infected THP-1 cells were used to prepare mRNA and small RNA Illumina sequencing libraries. To analyze gene expression, the six mRNA libraries were pooled into a single lane of an Illumina HiSeq 2000 (2x75 cycles). For miRNA identification, the six small RNA libraries were pooled into a single Illumina Miseq lane (1x50 cycles). All RNA-seq reads are available at National Center for Biotechnology Information Sequence Read Archive (Accession SRP045986).
Mapping sequencing reads and identification of differentially expressed genes

Adapters were removed from reads and were filtered by quality (>Q20) and length (>50bp) using Trimmomatic v0.30 (Bolger et al. 2014). *Homo sapiens* reads were filtered for possible contamination by mapping to *C. burnetii* genome (NC_002971.3) using BWA MEM v0.7.5 (Li and Durbin 2010) and *L. mexicana* genome (NZ_CADB00000000.1) using Tophat v2.0.11 (Kim et al. 2013) and were removed. After processing, reads were mapped to *Homo sapiens* Genome Reference Consortium Human Build 37 (GCF_000001405.13) using CLC Genomic Workbench v6.5. To identify differential gene expression, replicate data were pooled for pairwise comparisons and quantile normalized using CLC Genomic Workbench v6.5. Genes were filtered based on at least 10 raw reads mapping to each sample, and a log2 transformed fold change of one standard deviation above or below the mean. Differentially expressed genes were chosen based on significance (P < 0.05, FDR-corrected beta-binomial distribution test).

Raw read counts mapped to each mRNA isoform were exported from CLC into EBSeq (Leng et al. 2013) and differential expression of isoforms was determined based on significant EBSeq values (P < 0.05, FDR-corrected).

For quantitative PCR (qPCR) validation of gene expression, 1 µg of DNase-treated RNA and oligo-dT primers were used to prepare cDNA (Thermo Scientific). A subset of genes involved in host cell death (TGFB2, RIPK2, CYR61, CYP1B1, NFKBIA) was selected and qPCR was performed using SYBR green on an Agilent Mx3000P System. Fold difference value for each gene was calculated using the $2^{\Delta\Delta CT}$ method with GAPDH as the control (Livak and Schmittgen 2001). To assess the
correlation between expression estimates from RNA-seq and qPCR [As shown previously (Raghavan et al. 2012)], we calculated the Pearson correlation coefficient between fold difference values calculated by each method for *Coxiella* - and *Leishmania*-infected cells.

**Gene Ontology (GO) analysis and protein-protein interaction networks**

Gene Ontology (GO) terms were found using Database for Annotation, Visualization and Integrated Discovery (DAVID) and the GO FAT filter. GO-term enrichment tests were also performed with DAVID (Huang et al. 2009a, 2009b). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways over-represented among differentially expressed genes were chosen based on the level of statistical significance (P < 0.01). Protein-protein interaction networks were visualized using STRING 9.1 (Franceschini et al. 2013). Proteins unconnected to the main graph were removed. Markov Clustering was performed on STRING confidence scores using an inflation factor of 2 to visualize subgraphs of interacting protein processes (Brohée and van Helden 2006). GO-terms were overlaid onto the graphs using STRING to identify what processes were represented in the separate subgraphs.

**Identification of miRNAs**

Sequencing reads were cleaned by removing adapters and filtered by quality (>Q20) and length (>15bp) using Trimmomatic (Bolger et al. 2014). Replicate data was pooled and miRNAs were identified using CLC based on having an average of at least 10 reads
mapped to mature 5’ or 3’ miRNAs annotated in mirBase (Kozomara and Griffiths-Jones 2014).
3.4 RESULTS AND DISCUSSION

*C. burnetii* and *L. mexicana* infections induce robust but non-overlapping host responses

Human monocyte/macrophage cell line THP-1 was used to evaluate host responses against *C. burnetii* and *L. mexicana*. Previous studies have investigated host responses during early stages (6 h pi to 72 h pi) of infections by *C. burnetii* and by various *Leishmania* species (Ren et al. 2003; Mahapatra et al. 2010; Rabhi et al. 2013, 2012; de Muylder et al. 2011); however, because the transformation from the infective form (SCV and promastigote, respectively) to the replicative form (LCV and amastigote, respectively) occur at differing rates in the two pathogens, we analyzed a later point during infection (5d pi) when both pathogens have generated large PVs that fill most of the host cell volume. When compared to uninfected THP-1 cells, 1057 genes (746 up- and 311 down-regulated) were differentially expressed in *C. burnetii* infected THP-1 cells, whereas 698 genes (534 up- and 164 down-regulated) were differentially expressed in *L. mexicana* infected cells ([Figure 3.1, Additional_file_A2.csv](#), [Additional_file_A3.csv](#)). Interestingly, the sets of genes affected by the two pathogens are very different. Of the 1755 total genes identified in this study, only 126 genes (~7%) are differentially expressed under both conditions, and no metabolic pathways were significantly enriched within this common set of genes ([Figure 3.1, Additional_file_A4.csv](#)). A previous study that compared THP-1 cell response to infections by *Coxiella* and *Chlamydia trachomatis* (an intracellular bacterium), reported
Figure 3.1 – Identification of differentially expressed genes. Gene expression in (A) *Coxiella burnetii*-infected and (B) *Leishmania mexicana*-infected THP-1 cells in comparison to uninfected THP-1 cells are shown. Differentially expressed genes are highlighted in red and green. (C) Comparison of genes differentially expressed in *C. burnetii*-infected and *L. mexicana*-infected cells. Arrows indicate up-regulation or down-regulation of genes.
an overlap of ~25% of genes differentially expressed between the two infections (Ren et al. 2003). The low overlap between the host responses to *Coxiella* and *Leishmania*, and the higher magnitude of differentially expressed genes of host response to *C. burnetii* than that to *L. mexicana* possibly reflects the more distant evolutionary relationship between the bacteria and the eukaryotic parasite compared to the two bacterial pathogens previously studied. Apoptosis and host cell immune response pathways were the most significantly enriched KEGG pathways in *Coxiella* infected cells (Table 3.1), as observed in previous microarray-based studies (Ren et al. 2003; Mahapatra et al. 2010). Repression of host cell death by *Coxiella* has been reported previously (Voth et al. 2007; Lührmann and Roy 2007), and is thought to promote intracellular growth of *Coxiella* within large PVs; conversely, induction of Toll-like Receptor signaling pathways and production of cytokines and chemokines participate in the host response to *Coxiella* infection (Zamboni et al. 2004; Graham et al. 2013).

In *Leishmania*-infected cells, purine metabolism was the only KEGG pathway that was significantly perturbed (Table 3.1). *Leishmania* is dependent on host for its purine supply (McConville et al. 2007), and three genes (ADSSL1, RRM2, PRPS1) involved in purine biosynthesis or salvage pathways were significantly overexpressed in infected THP-1 cells. Intriguingly, a majority of “purine metabolism” genes listed in Table 3.1 regulate the levels of intracellular second messengers cAMP and cGMP. Adenylate cyclases (ADCY4, ADCY8, ADCY9, ADCY10) catalyze the formation of cAMP from ATP; guanylate cyclase (GUCY1A2) catalyzes the conversion of GTP to cGMP; phosphodiesterases (PDE4C, PDE6G, PDE2A) catalyze the hydrolysis of cAMP
<table>
<thead>
<tr>
<th>Sample</th>
<th>KEGG Term</th>
<th>Description</th>
<th>Genes</th>
<th>Fold Change</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coxiella</strong>-infected</td>
<td>hsa04210</td>
<td>Apoptosis</td>
<td>BID, IRAK2, TNF, XIAP, RELA, TP53, NFKBIA, ENDOD1, NFKB1, BIRC3,</td>
<td>3.51</td>
<td>2.1E-06</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TNFRSF10A, CASP10, PRKAR2B, IRAK3, TNFRSF10B, PPP3CC, IL1B, PIK3R5,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PIK3R3, IL1A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>hsa04621</td>
<td>NOD-like receptor signaling pathway</td>
<td>CXCL1, TNF, XIAP, IL8, RELA, CXCL2, NFKBIA, NFKB1, BIRC3, NOD2, RIPK2,</td>
<td>3.20</td>
<td>5.4E-04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IL1B, TNFAIP3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>hsa04060</td>
<td>Cytokine-cytokine receptor interaction</td>
<td>CXCL1, TNFRSF21, CCL3, TNF, CXCL5, CXCL3, CXCL2, TNFSF15, CXCL6,</td>
<td>1.81</td>
<td>1.6E-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IL7R, CCL4, TGFB2, LIF, CCL2, IL23A, CCL20, CCL3L1, IL4R, TNFRSF18,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IL15RA, IL1B, IL1A, BMP2, IL8, CD40, IL11RA, TNFRSF10A, INHBA, ACVR2B,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TNFRSF10B, VEGFA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>hsa04062</td>
<td>Chemokine signaling pathway</td>
<td>CXCL1, ADCY4, CCL3, LYN, CXCL5, IL8, HCK, CXCL3, RELA, CXCL2, NFKBIA,</td>
<td>1.96</td>
<td>2.2E-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ADRBK2, NFKB1, CXCL6, CCL4, CCL2, CCL20, CCL3L1, GNG10, SOS2, PIK3R5,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GNB4, PIK3R3, GNG7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>hsa05222</td>
<td>Small cell lung cancer</td>
<td>E2F1, TRAF1, XIAP, PTGS2, RELA, TP53, ITGA2, NFKBIA, NFKB1,</td>
<td>2.55</td>
<td>2.8E-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BIRC3, LAMB3, PIK3R5, PIK3R3, TRAF3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>hsa05200</td>
<td>Pathways in cancer</td>
<td>TRAF1, E2F1, BID, PTGS2, XIAP, STAT5A, MITF, NFKBIA, NFKB1,</td>
<td>1.63</td>
<td>4.2E-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NFKB2, TCF7L2, MMP1, TGFB2, LAMB3, SOS2, PIK3R5, CCNA1, PIK3R3, FGF2,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TRAF3, BMP2, IL8, VHL, RELA, TP53, ITGA2, BIRC5, BIRC3, FZD4, DAPK3,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CTNNB3, RAD51, SMO, ETS1, VEGFA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>hsa04620</td>
<td>Toll-like receptor signaling pathway</td>
<td>CCL3, TNF, IL8, RELA, NFKBIA, NFKB1, CD40, CCL4, CD86, MAP3K8, IL1B,</td>
<td>2.27</td>
<td>5.5E-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PIK3R5, PIK3R3, CD14, TRAF3</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Leishmania</strong>-infected</td>
<td>hsa00230</td>
<td>Purine metabolism</td>
<td>ADCY4, ADSSL1, ADCY8, POLA1, PDE4C, PDE6G, POLE4, PDE2A, ADCY9, RRM2,</td>
<td>2.40</td>
<td>5.1E-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PKLR, GUCY1A2, ADCY10, PRPS1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
and/or cGMP. Previous studies have shown that *Leishmania* resists host antimicrobial activities by modulating several host signaling pathways, including Ca\(^{2+}\)- and PKC-dependent pathways, JAK-STAT pathways, and MAP kinases (Olivier et al. 2005). Similarly, *Leishmania* could be subverting the host’s cAMP and cGMP signaling pathways in order to suppress immune responses and to promote its intracellular growth.

A protein-protein interaction network analysis using the STRING database (Franceschini et al. 2013) confirmed that *Coxiella* infection induced the expression of genes involved in negative regulation of cell death (**Figure 3.2A**). In contrast, this analysis identified that genes involved in positive regulation of cell death were upregulated in *Leishmania*-infected cells (**Figure 3.2B**). We confirmed this trend by analyzing the expression of a subset of cell death-related genes using qPCR (**Figure 3.3**). The induction of host cell death during later stages of infection probably aids in the cell-to-cell transfer of *Leishmania* amastigotes within membrane blebs, as shown recently (Real et al. 2014).

**Differential expression of mRNA isoforms in infected and uninfected cells**

In human cells, alternate splicing of pre-mRNA can give rise to several isoforms of the mature mRNA, and proteins derived from them may have distinct cellular roles (Lareau et al. 2004). In addition to expanding the proteome, cells utilize alternate splicing as a regulatory tool. For example, a short splice variant of human tryptophan-tRNA synthase, but not the full length protein, regulates angiogenesis (Wakasugi et al. 2002). Isoform generation may also have a role in host cell response against infections. Different
Figure 3.2 – Protein-protein interaction analysis. Protein-protein interaction networks of up-regulated genes in (A) *C. burnetii*-infected and (B) *L. mexicana*-infected THP1 cells visualized in STRING. Colors of circles are based on Markov Clustering with an inflation factor of 2. Highlighted clusters are labeled with their GO or KEGG categories.
Figure 3.3 - Validation by qPCR of expression levels estimated by RNA-seq. Fold difference values calculated by qPCR correlated well with fold difference values calculated by RNA-seq.
isoforms of p53 (encoded by TP53 gene) are involved in host defense against both bacterial (Helicobacter pylori) and viral (Influenza and Simian virus 40) infections (Terrier et al. 2013). Similarly, Hepatitis C virus activates the immunologic isoform of nitric oxide synthase (NOS) gene, which induces NO production (Machida et al. 2004). Transcriptome analysis (RNA-seq) is a powerful approach to identify differential isoform expression under different conditions at a genome-wide scale (Eswaran et al. 2013; Lo et al. 2014). We used RNA-seq to investigate whether infection by either Coxiella or Leishmania induced differential expression of human gene isoforms. We identified 689 isoforms from 626 genes that were differentially expressed in C. burnetii-infected cells, and 651 isoforms from 569 genes in Leishmania-infected cells, when compared to uninfected THP-1 cells (Figure 3.4, Additional_file_A5.csv, Additional_file_A6.csv). As observed for full-length mRNAs, there was minimal overlap between the sets of genes with differential expression of isoforms under each infection condition (only 105 common genes). Additionally, no KEGG pathways were significantly enriched in either gene set, indicating that differential isoform expression is a cell-wide phenomenon. Cumulatively, our data revealed that in addition to differences that are apparent at the gene level, the mostly unexplored realm of isoform variation could contribute to host responses to infections.

Coxiella and Leishmania infections perturb the expression of apoptosis-related miRNAs

Expression of various protein-coding genes in humans is regulated by miRNAs. These
Figure 3.4 – Differential expression of mRNA isoforms. Differential isoform expression of VPS8 in *C. burnetii* infected, *L. mexicana* infected, and uninfected THP1 cells are shown as a representation of isoform analysis. (A) The full-length VPS8 gene is depicted with colored bars representing exons. Isoforms 1 (B) and 7 (C) of VPS8 that have significantly different expression in the three samples are shown. Each gray line above an isoform represents 10 mapped reads.
small non-coding RNAs regulate the expression of target genes by base-pairing with mRNAs, thereby either blocking translation or causing target degradation or destabilization (Fabian et al. 2010). They are involved in many, if not all, biological processes, including metabolic pathways, cell proliferation and apoptosis. Recently, miRNAs have been shown to be an important part of host cell response to viral, bacterial and parasitic infections (Schulte et al. 2011; Schnitger et al. 2011; Lagos et al. 2010). In addition, some viruses, including Herpes viruses and Hepatitis C virus, have the ability to interfere with the host miRNA network to promote viral growth (Cullen 2011; Jopling et al. 2005). Recent studies also showed that eukaryotic intracellular pathogens such as Cryptosporidium parvum and Toxoplasma gondii promote intracellular replication by altering host cell miRNA networks (Hakimi and Ménard 2010; Zeiner et al. 2010). To identify miRNAs that are potentially perturbed by C. burnetii or L. mexicana infections, we sequenced and enumerated miRNAs expressed by uninfected, Coxiella-infected, and Leishmania-infected THP-1 cells. We identified 257 miRNAs that were expressed in THP-1 cells (Additional_file_A7.csv), which includes 50 of the 64 miRNAs reported by a recent study that examined miRNAs expressed in human macrophages in response to Leishmania major infection (Lemaire et al. 2013). Among the 257 miRNAs, seven were upregulated and one was down regulated in Coxiella-infected cells, and three were upregulated and two were down regulated in Leishmania-infected cells (Table 3.2). Intriguingly, several of the differentially expressed miRNAs have been shown in previous studies to regulate host cell death: miR-145 modulates the expression of KLF4 (Davis-Dusenbery et al. 2011), a transcription factor for TP53, which regulates apoptosis.
Table 3.2 – MicroRNAs (miRNAs) perturbed by *Coxiella* and *Leishmania* infections

<table>
<thead>
<tr>
<th>Sample</th>
<th>miRNA</th>
<th>Fold Change (log2)</th>
<th>P value</th>
<th>Regulation</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Coxiella</em>-infected</td>
<td>mir-148a-3p</td>
<td>-0.58</td>
<td>0.024</td>
<td>Down</td>
<td>Pro-apoptotic a</td>
</tr>
<tr>
<td></td>
<td>mir-181d-5p</td>
<td>0.78</td>
<td>&lt;0.001</td>
<td>Up</td>
<td>Anti-apoptotic b</td>
</tr>
<tr>
<td></td>
<td>mir-193a-5p</td>
<td>0.81</td>
<td>&lt;0.001</td>
<td>Up</td>
<td>Pro-apoptotic c</td>
</tr>
<tr>
<td></td>
<td>mir-362-5p</td>
<td>0.89</td>
<td>0.015</td>
<td>Up</td>
<td>Anti-apoptotic d</td>
</tr>
<tr>
<td></td>
<td>mir-361-5p</td>
<td>0.95</td>
<td>0.004</td>
<td>Up</td>
<td>Anti-apoptotic e</td>
</tr>
<tr>
<td></td>
<td>mir-194-2-5p</td>
<td>1.05</td>
<td>0.024</td>
<td>Up</td>
<td>Anti-apoptotic f</td>
</tr>
<tr>
<td></td>
<td>mir-28-3p</td>
<td>1.12</td>
<td>0.024</td>
<td>Up</td>
<td>Neither g</td>
</tr>
<tr>
<td></td>
<td>mir-28-5p</td>
<td>1.35</td>
<td>&lt;0.001</td>
<td>Up</td>
<td>Pro-apoptotic f</td>
</tr>
<tr>
<td><em>Leishmania</em>-infected</td>
<td>mir-145-5p</td>
<td>-1.00</td>
<td>0.002</td>
<td>Down</td>
<td>Pro-apoptotic h</td>
</tr>
<tr>
<td></td>
<td>mir-221-5p</td>
<td>-0.62</td>
<td>&lt;0.001</td>
<td>Down</td>
<td>Anti-apoptotic i</td>
</tr>
<tr>
<td></td>
<td>mir-15b-5p</td>
<td>0.56</td>
<td>0.035</td>
<td>Up</td>
<td>Pro-apoptotic j</td>
</tr>
<tr>
<td></td>
<td>mir-29b-1-3p</td>
<td>1.09</td>
<td>0.002</td>
<td>Up</td>
<td>Pro-apoptotic k</td>
</tr>
<tr>
<td></td>
<td>mir-29b-2-3p</td>
<td>1.19</td>
<td>&lt;0.001</td>
<td>Up</td>
<td>Pro-apoptotic k</td>
</tr>
</tbody>
</table>

a (Zhang et al., 2011), b (Wang et al., 2010), c (Nakano et al., 2013), d (Xia et al., 2014), e (Wu et al., 2013), f (Zhang et al., 2014), g (Almeida et al., 2012), h (Davis-Dusenbery et al., 2011), i (le Sage et al., 2007), j (Cimmino et al., 2005), k (Garzon et al., 2009)
(Rowland et al. 2005); miR-15b and miR-29b are known to be pro-apoptotic in leukemia cells (Cimmino et al. 2005; Garzon et al. 2009); miR-148a promotes apoptosis by targeting BCL2 in colorectal cancer cells (Zhang et al. 2011); miR-181d also targets BCL-2 and promotes apoptosis in glioma cells (Wang et al. 2012). These results complement gene expression data (Figure 3.2), and indicate that miRNAs may have important roles in inhibiting host cell death during *Coxiella* infection, and promoting host cell death during *Leishmania* infection.
Coxiella’s basic biology and molecular pathogenesis has remained elusive and generally understudied. There is an urgent need to expand research to fill this void and to develop more effective therapies. By examining several critical biosynthetic pathways, we show that horizontal gene transfer (HGT) has played an important role in shaping the intracellular metabolic capability of C. burnetii. We found that not only has Coxiella horizontally acquired extra copies of genes that enhance fatty acid, biotin and heme biosyntheses, but also the mechanism for increased heme production stems from a non-translational function of tRNA\textsuperscript{Glu}\textsubscript{2}. Combined with the finding that Coxiella’s growth is significantly reduced when HemL protein (part of the heme biosynthesis pathway) is inhibited, we showed that heme biosynthesis is critical to Coxiella’s intracellular growth, and importantly, because HemL is not found in humans, it could be a prime target for developing new therapeutics.

Due to the uncertainty in C. burnetii’s phylogenetic relationship with other Gammaproteobacteria, we focused on a subset of genes that were likely acquired from distantly related bacteria. However, horizontal exchange usually occurs at higher frequency between closely related bacteria (Ochman et al. 2000). Hence, it is likely that many more genes in C. burnetii are of HGT origin from more closely related bacteria. Availability of more genomes of bacteria in the phylogenetic neighborhood of C. burnetii along with more detailed evolutionary, genetic, and functional analyses are required to
identify all HGT-origin genes in *C. burnetii*. Then, we can begin to understand the full impact of HGT on the pathogen’s physiology and pathogenicity.

At the host level, the genome-wide gene, mRNA-isoform, and miRNA expression patterns were distinct between macrophages infected with either *C. burnetii* or *L. Mexicana*. This pathogen-specific response indicates that even though both pathogens have converged on a similar intracellular niche, they utilize distinct programs to generate and maintain their respective intracellular vacuoles. Of particular interest were several miRNAs known to inhibit apoptosis in human cells that are induced or repressed during *Coxiella* infection. One such miRNA is miR-148a, which inhibits apoptosis by targeting BCL-2. The expression of miR-148a was significantly reduced in *Coxiella*-infected cells when compared to uninfected cells; concordantly, the expression of Bcl-2 was higher in *Coxiella* infected cells. Based on this, our working hypothesis is that miR-148a regulates Bcl-2 mRNA levels, thereby promoting *Coxiella* infection. This current study focused mainly on RNA-seq analysis, thus further experiments need to be conducted to confirm that Bcl-2 is indeed the target of miR-148a in THP-1 cells. Additional work will be required to assess the role of the individual miRNAs identified here in promoting *Coxiella* infection, and how these miRNAs can be targeted to develop new, more effective therapies.
REFERENCES


Raghavan R, Miller SR, Hicks LD, Minnick MF. 2007. The unusual 23S rRNA gene of *Coxiella burnetii*: Two self-splicing group I introns flank a 34-base-pair exon, and


Wang LY, Brown L, Elliott M, Elliott T. 1997. Regulation of heme biosynthesis in *Salmonella typhimurium*: Activity of glutamyl-tRNA reductase (HemA) is greatly...


Upregulation of miR-194 contributes to tumor growth and progression in 
pancreatic ductal adenocarcinoma. *Oncol Rep* 31: 1157-1164. doi: 
10.3892/or.2013.2960.

genome-wide discovery of putative horizontal gene transfers. *BMC Genomics* 15: 

Zlitni S, Ferruccio LF, Brown ED. 2013. Metabolic suppression identifies new 
10.1038/nchembio.1361.
APPENDIX A

Additional_file_A1.csv – 172 ‘high-confidence’ horizontally acquired genes on the chromosome of *C. burnetii* RSA 493 and their putative donors.

Additional_file_A2.csv – Differentially expressed genes in *Coxiella*-infected vs. uninfected THP-1 cells.

Additional_file_A3.csv – Differentially expressed genes in *Leishmania*-infected vs. uninfected THP-1 cells.

Additional_file_A4.csv – Differentially expressed genes that overlap between *Coxiella*-infected vs. uninfected and *Leishmania*-infected vs. uninfected THP-1 cells.

Additional_file_A5.csv – Differentially expressed mRNA isoforms in *Coxiella*-infected vs. uninfected THP-1 cells.

Additional_file_A6.csv – Differentially expressed mRNA isoforms in *Leishmania*-infected vs. uninfected THP-1 cells.

Additional_file_A7.csv – All miRNAs detected in this study.
Additional_file_A8.pdf – Supplementary tables and figures for Appendix B: “Genome rearrangements can make and break small RNA genes”.

Additional_file_A9.pdf – Supplementary tables and figures for Appendix C: “Emergence of new sRNAs in enteric bacteria is associated with low expression and rapid evolution”.

Additional_file_A10.csv – Table of 81 E. coli sRNAs and their descriptive statistics used in “Emergence of new sRNAs in enteric bacteria is associated with low expression and rapid evolution”.

Additional_file_A11.csv – Table of 127 S. enterica sRNAs and their descriptive statistics used in “Emergence of new sRNAs in enteric bacteria is associated with low expression and rapid evolution”.

Additional_file_A12.csv – Table of 49 E. coli rfam sRNAs and their descriptive statistics used in “Emergence of new sRNAs in enteric bacteria is associated with low expression and rapid evolution”.

Additional_file_A13.csv – Table of sRNAs conserved between E. coli and S. enterica and their descriptive statistics used in “Emergence of new sRNAs in enteric bacteria is associated with low expression and rapid evolution”.
Additional_file_A14.pdf – Supplementary tables and figures for Appendix D:

“Accumulation and expression of multiple antibiotic resistance genes in *Arcobacter cryaerophilus* that thrives in sewage”.

Additional_file_A15.csv – Expression of ORFs mapped to antibiotic resistance categories

Additional_file_A16.csv – Comparison of overlapping genes and annotations in *Arcobacter*

Additional_file_A17.csv – Comparison of *Arcobacter* in RAST

Additional_file_A18.csv – Expression of ORFs mapped to virulence factors
APPENDIX B

Genome rearrangements can make and break small RNA genes

B.1 ABSTRACT

Small RNAs (sRNAs) are short, transcribed regulatory elements that are typically encoded in the intergenic regions (IGRs) of bacterial genomes. Several sRNAs, first recognized in *Escherichia coli*, are conserved among enteric bacteria, but because of the regulatory roles of sRNAs, differences in sRNA repertoires might be responsible for features that differentiate closely related species. We scanned the *E. coli* MG1655 and *Salmonella enterica* Typhimurium genomes for nonsyntenic IGRs as a potential source of uncharacterized, species-specific sRNAs and found that genome rearrangements have reconfigured several IGRs causing the disruption and formation of sRNAs. Within an IGR that is present in *E. coli* but was disrupted in *Salmonella* by a translocation event is an sRNA that is associated with the FNR/CRP global regulators and influences *E. coli* biofilm formation. A *Salmonella*-specific sRNA evolved de novo through point mutations that generated a $\sigma^{70}$ promoter sequence in an IGR that arose through genome rearrangement events. The differences in the sRNA pools among bacterial species have previously been ascribed to duplication, deletion, or horizontal acquisition. Here, we show that genomic rearrangements also contribute to this process by either disrupting sRNA-containing IGRs or creating IGRs in which novel sRNAs may evolve.
**B.2 INTRODUCTION**

RNAs that do not code for proteins are critical to gene regulation in all domains of life. In bacteria, small RNAs (sRNAs) are typically 50–200 nucleotides (nt) in length and are usually encoded in genomic regions between protein-coding genes (intergenic regions or IGRs). They can control gene expression by modulating transcription, translation, or mRNA stability (Storz et al. 2011). The application of technologies that interrogate entire transcriptomes has revealed unexpectedly large numbers of sRNAs in bacterial genomes (Raghavan, Groisman, et al. 2011; Kroger et al. 2012). But unlike protein-coding genes, the mechanisms by which new sRNA genes arise and the forces that shape the sRNAs contents of genomes are not well understood (Gottesman and Storz 2011). Some sRNAs, such as 6S RNA (Wassarman and Storz 2000), are broadly conserved among bacteria, whereas several others are species- or even strain-specific (Gottesman and Storz 2011; Skippington and Ragan 2012). The sRNA transcriptomes of the enterics *Escherichia coli* and *Salmonella enterica* show substantial overlap; however, some of the orthologous IGRs display different patterns of expression and several sRNAs are present in only one of the species (Raghavan et al. 2012).

Differences in sRNA gene contents among bacteria can arise from lineage-specific loss or from the emergence of new sRNAs through duplication (Lenz et al. 2004; Wilderman et al. 2004) or horizontal acquisition (Pichon and Felden 2005; Sittka et al. 2008). An examination of the distribution of sRNAs within the *E. coli/ Shigella* complex showed that the variation in the presence of known sRNAs was dominated by gene loss
through deletions (Skippington and Ragan 2012). However, because this study focused only on those sRNAs that were originally characterized in a single strain of *E. coli*, it was biased toward the recognition of deletion events as it could not detect unique sRNAs in the genomes of other strains. Applying a broader phylogenetic perspective, homologs of a dual-function sRNA, SgrS, have been detected in distantly related Gammaproteobacteria (Horler and Vanderpool 2009), and an exhaustive survey of sRNAs revealed that most *E. coli* sRNAs originated after *Enterobacteriales* split from other Gammaproteobacteria (Peer and Margalit 2014). This lineage-specific sRNA accumulation seems to be related to the evolution of the RNA-binding protein Hfq; however, the mechanisms by which new bacterial sRNAs emerge or are lost remain largely unknown.

In eukaryotes, there are cases where novel regulatory RNAs have evolved through gene duplication, by de novo origination from noncoding sequences, and from the degradation of protein-coding genes (Kaessmann 2010); but in bacteria, the mechanisms by which new regulatory RNAs arise are much less clear. Because new genes can form through the chimeric assembly of fragments from various sources—one well-known example of this is the jingwei gene of *Drosophila* (Long and Langley 1993)—we first adopted a structural genomics and RNA sequencing (RNA-seq)-based approach to identify new sRNA genes and then tested for sRNA functions. This combination of comparative and experimental analyses identified several previously unrecognized sRNAs and uncovered the sources of these differences in sRNA repertoires. We find that genome rearrangements have disrupted and formed IGRs containing functional sRNAs, thereby causing disparity in the sRNA contents of related bacterial species.
B.3 MATERIALS AND METHODS

RNA Sequencing

For sRNA discovery, *E. coli* K-12 MG1655 (GenBank NC_000913.2) and *S. enterica* subsp. *enterica* serovar Typhimurium str. 14028S (GenBank NC_016856.1) were grown in lysogeny broth (LB) to $\text{OD}_{600} \approx 0.5$ and then harvested by centrifugation. Total RNA was extracted from bacterial pellets using TRI reagent (Life Technologies), and cleaned on RNeasy columns (Qiagen) to remove spurious transcripts, transfer RNAs and 5S ribosomal RNA (rRNA). Genomic DNA was degraded by DNase treatment (Life Technologies) and 16S and 23S rRNAs were removed with a MICROBExpress kit (Life Technologies). Strand-specific RNA-seq libraries were synthesized (Raghavan et al. 2012), and each library was sequenced on the Illumina GA II platform (35 cycles) at the Yale Center for Genome Analysis.

Mapping Sequencing Reads

To identify sRNAs, sequencing reads were mapped onto the published *E. coli* (NC_000913.2) or *Salmonella* Typhimurium (NC_016856.1) genomes using MAQ (Li et al. 2008) and examined with Artemis (Rutherford et al. 2000), as described previously (Raghavan, Groisman, et al. 2011; Raghavan, Sage, et al. 2011; Raghavan et al. 2012). Those previously uncharacterized sRNAs identified in *E. coli* are numbered and given the prefix EcsR (*E. coli* sRNA) and those in *Salmonella*, SesR (*S. enterica* sRNA).
To characterize regions that are differentially expression, sequencing reads were mapped onto *E. coli* (NC_000913.2) using the CLC Genomics Workbench. Genes with at least one read per sample and at least 20 total reads across all samples were chosen based on raw gene read counts from CLC mapping. Differential expression analysis of genes was performed using the DESeq R package (Anders and Huber 2010). Genes were chosen for downstream analysis based on significance \( P < 0.05 \), FDR-corrected. Gene Ontology (GO) terms were found using Database for Annotation, Visualization and Integrated Discovery (DAVID) and the GO FAT filter (Huang et al. 2009). GO-term enrichment tests were also performed with DAVID. GO-terms overrepresented among differentially expressed genes were chosen based on the level of statistical significance \( P < 0.05 \), Benjamini-corrected.

**sRNA Target Identification**

To identify sRNA-regulated genes, EcsR1 was cloned into the *NheI* and *HindIII* sites behind the arabinose-inducible promoter on plasmid pBAD using the polymerase chain reaction (PCR) primers 5′-CCG CTA GCG TTT TAG TAT CCG CAT AAA GTG TAA C-3′ and 5′-CTA AGC TTT CCT GCC CGC TGT TAT GGC G-3′. *Escherichia coli* or *Salmonella*, transformed with either the empty pBAD vector (control) or pBAD+ EcsR1 (test), were grown in LB to OD\(_{600}\) ≈ 0.5 and induced with 0.2% arabinose for 15 min, as previously described (Durand and Storz 2010). RNA was extracted and processed for Illumina sequencing as above. Four Illumina mRNA-seq libraries (two control samples and two test samples) were prepared for each bacterium and multiplexed into a single
lane of an Illumina HiSeq 2000 (101 cycles) at the Genomic Sequencing and Analysis Facility at University of Texas at Austin.

**Measuring Hfq Stabilization of sRNAs**

An *Hfq*-deleted strain of *E. coli* (JW4130-1) and its isogenic parent strain (BW25113) (Baba et al. 2006) were obtained from Yale Coli Genetic Stock Center and grown to mid-log phase (OD$_{600}$ ≈ 0.5) in LB. Total RNA was DNase-treated, and 1 µg used as template for preparing cDNA. The abundances of EcsR1 in the wild-type and *Hfq*-deleted strains were determined by quantitative PCR (primers: 5′-TTT TTG TGT AAT GAC GGA GTT CA-3′, and 5′-GCG GGC TT TTC TGC TTA TT-3′), and calculated from Ct (threshold cycle) values.

**Identification of Unique IGRs**

Orthologous genes common to *E. coli* and *Salmonella* were identified using a reciprocal BLAST best-hit approach (Raghavan et al. 2012). Gene order of each orthologous gene-pair was determined with GeneOrder 4.0 (Mahadevan and Seto 2010), and in cases where the genomic locations were not syntenic in the two species, we searched for gene-pairs with adjacent novel IGRs using Artemis.

**Identification of -10 Promoter Elements and sRNA Homologs**

Transcriptional start sites (TSSs) for novel sRNAs were identified from RNA-seq data as described previously (Raghavan, Sage, et al. 2011; Raghavan et al. 2012). TSSs for
flanking genes were identified as above and were confirmed using published data (Kroger et al. 2012; Keseler et al. 2013). The $\sigma^{70}$-10 motif has a 6-bp consensus sequence TATAAT; however, promoters often have imperfect matches to the consensus and can be located anywhere in a window ranging from approximately 4 to 18 bp upstream of the TSS (Huerta and Collado-Vides et al. 2003). To identify potential -10 elements associated with sRNAs, we searched this 15-bp window for any hexamers that matched at least 4 of the 6 bp in the consensus sequence including the two most highly conserved positions, A2 and T6 (Huerta et al. 2006). Bacterial genomes were queried for homologs of sRNAs identified in this study by analyzing a combination of sequence identity, secondary structure conservation, and genomic location as described previously (Raghavan, Groisman, et al. 2011).

**Detection of sRNA 3′-Ends**

A modified Rapid Amplification of cDNA Ends (RACE) procedure (Raghavan, Groisman, et al. 2011) was used to determine the 3′-ends of sRNAs as follows: Total RNA, depleted of 16S and 23S rRNA using a MICROBExpress kit (Life Technologies), was dephosphorylated with alkaline phosphatase (NEB), and a short oligonucleotide adapter (5′-P-UCG UAU GCC GUC UUC UGC UUG UidT-3′) was ligated to 3′-ends using T4 RNA ligase (NEB). The 3′ adapter-ligated RNA was reverse-transcribed using a primer complementary to the adapter (5′-CAA GCA GAA GAC GGC ATA CGA-3′), and the resulting cDNA was used as template in PCR reactions using primers specific to sRNAs (EcsR1: 5′-AGA TGA CAC TTT TTG TGT AAT GAC G-3′; EcsR2: 5′-TAT
CGC GCT ACT TCA GGA TGA TGT A-3') along with the adapter-complementary primer. Amplicons were resolved on 3% low-range ultra agarose (Bio-Rad) gels to determine their lengths, and their nucleotide sequences were determined by Sanger sequencing.

**Biofilm Assay**

EcsR1-deletion strain of *E. coli* was constructed using λ Red-mediated recombination (Datsenko and Wanner 2000). *Escherichia coli* or *Salmonella* strains grown overnight at 37 °C in LB (or LB with 100 µg/ml ampicillin) were diluted 1:100 in fresh media and grown in 96-well microtiter plates for 48 h at 28 °C without shaking. Planktonic growth (OD$_{600}$) of *E. coli* and *Salmonella* strains measured on a Victor X5microplate reader (Perkin Elmer) did not significantly differ from each other. Supernatants containing nonadhered cells were discarded, and samples were washed twice with distilled water and the attached biofilm in each well was stained with 0.1% crystal violet for 30 min. Unbound stain was removed by washing with distilled water. To quantify biofilm production, the crystal violet associated with biofilms was dissolved in 100% ethanol and absorbance (A$_{600}$) was measured, and normalized to the OD$_{600}$ value of each strain, as described previously (Gualdi et al. 2008). Average intensity of biofilm formation for each strain was generated from at least four replicate experiments.
B.4 RESULTS

Genome Rearrangements Form Unique IGRs

To identify IGRs that are unique to either *E. coli* or *Salmonella*, we compared the genomic locations of all orthologous genes in the two genomes. Because *E. coli* and *Salmonella* genomes are largely syntenic, the majority of IGRs situated between orthologous gene-pairs in the two genomes are also syntenic. However, there are several instances where orthologous protein-coding genes are situated at different relative locations in each genome (apparent as data points that do not lie along the diagonal in supplementary fig. S1, found in Additional_file_A8.pdf). After examining each of these cases, we identified chimeric IGRs present in either *E. coli* or *Salmonella* that were generated through the rearrangement of 68 genes (supplementary table S1, found in Additional_file_A8.pdf).

Unique IGRs Contain Novel sRNAs

We performed a directional RNA-seq analysis on *E. coli* and *Salmonella* grown under identical conditions to determine whether any of the species-specific IGRs contained highly transcribed regions. After mapping sequencing reads onto the respective genomes, we detected “transcriptional peaks,” which usually indicate the presence of sRNAs, in four of the species-specific IGRs, two in *E. coli* and two in *Salmonella* (Figure B.1). Transcripts mapping to the corresponding locations in the *E. coli* and *Salmonella* genomes have been observed in previous studies (Tjaden et al. 2002; Dornenburg et al.
Figure B.1 – Expression profiles within nonsyntenic IGRs. Putative sRNAs were detected by RNA-seq analysis of transcript levels within IGRs in *E. coli* (A, B) and *Salmonella* (C, D). For uniformity, the number of sequencing reads mapped to IGRs is limited to 2,000 (dashed line). Arrows showing the orientation of ORFs and putative sRNAs are not drawn to scale.
2010; Kroger et al. 2013) further verifying their transcriptional status, and there were no potential open reading frames (ORFs) of substantial length within these transcripts indicating that they represent sRNAs.

TSSs and 3′-ends of the transcribed sequences detected in these IGRs were identified from RNA-seq data (Raghavan et al. 2012), and a modified 3′-RACE procedure (Raghavan, Groisman, et al. 2011) was used to confirm the sRNAs in E. coli (supplementary fig. S2, found in Additional_file_A8.pdf), yielding the following results: 1) The sRNA (EcsR1) within the IGR between uspF and ompN genes in E. coli is 126 nt (genomic location 1433654–1433779), 2) the sRNA (EcsR2) within the IGR between yagU and ykgJ genes in E. coli is 166 nt (genomic location 302905–303070), 3) the Srna (SesR1) within the IGR between STM14_1512 and STM14_1513 genes in Salmonella is 105 nt (genomic location 1347963–1348067), and 4) the sRNA (SesR2) within the IGR between STM14_1869 and STM14_1870 genes in Salmonella is 111 nt (genomic location 1636380–1636490).

Homologs of EcsR1 and EcsR2 are present in all 66 E. coli genomes available in the RefSeq database (supplementary table S2, found in Additional_file_A8.pdf). SesR1 homologs were detected in all 44 S. enterica and S. bongori genomes (supplementary table S3, found in Additional_file_A8.pdf), and a recent study reported an sRNA (STnc1990) at the homologous position in S. enterica Typhimurium SL1344 (Kroger et al. 2013). SesR2 is conserved in 20 S. enterica genomes and in the two sequenced S. bongori strains (supplementary table S3, found in Additional_file_A8.pdf). However, the STM14_1869–STM14_1870 IGR is not maintained in S. bongori due to the loss of
the STM14_1870 ortholog. Because SesR2 is absent from a few S. enterica serovars but present in the S. bongori outgroup, this sRNA is ancestral to Salmonella and was subsequently lost in some S. enterica lineages.

A *Salmonella*-specific IGR Formed Through HGT-mediated Genome Rearrangement

The STM14_1869–STM14_1870 IGR is present in Salmonella but not in E. coli. *Escherichia coli* possesses a gene, *yjgH*, that is orthologous to STM14_1869, but contained no ortholog for STM14_1870. Further analysis uncovered that STM14_1870 and its neighboring gene STM14_1871 constitute the toxin and antitoxin, respectively, of the StbED toxin–antitoxin (TA) system (Unterholzner et al. 2013) (Figure B.2).

Orthologs of stbED TA genes are present on several enterobacterial plasmids and prophages, and are horizontally transferred between bacteria (Anantharaman and Aravind 2003; Unterholzner et al. 2013). Additionally, the succeeding gene in the *Salmonella* genome, STM14_1872, is also homologous to a gene of bacteriophage origin (Figure B.2), further indicating that the IGR between STM14_1869 and STM14_18670 was created by the introduction of genes through horizontal gene transfer (HGT)-mediated events.

Evolution of a New sRNA in a *Salmonella* IGR

To determine whether SesR2 was introduced along with its horizontally acquired neighboring genes into Salmonella, we searched the IGRs downstream of the *stbE* gene
Figure B.2 – *Salmonella* IGR formed through an HGT-mediated genome rearrangement. Most homologs of STM14_1870 (*stbE*, blue arrow) and STM14_1871 (*stbD*, green arrow) are situated on bacterial plasmids, and STM14_1872 (purple arrow) is a prophage gene. Both the STM14_1869–STM14_1870 IGR and the sRNA (SesR2) are present only in *Salmonella*. In *Shigella*, an insertion sequence (IS1294) flanks the *stbE* gene.
in several enterobacterial genomes, but could not detect homologous sRNAs (Figure B.2). Because a promoter is required for the new sRNA to be transcribed, we compared the region that contains the sRNA’s TSS (5’-end of STM14_1869) with its homologous sequence in *E. coli* (5’-end of *yjgH*) and in other enterics. As shown in Figure B.3 (and supplementary figure S3, found in Additional_file_A8.pdf), a putative σ70 promoter (CATAAT, located -6 to -11 bp from sRNA’s TSS) is uniquely present in *Salmonella*, indicating that this sRNA originated de novo in the *Salmonella*-specific IGR. Additionally, the promoter and sRNA sequences are conserved in *S. bongori* and in those *S. enterica* serovars that maintain an intact STM14_1869–STM14_1870 IGR.

**Rearrangement-Induced Loss of a *Salmonella* IGR**

In *E. coli*, the *ompN* and *usf* genes are adjacent, separated by a 140-bp IGR that contains EcsR1; whereas in *Salmonella*, the *ompN* gene is in an alternate location, situated between the STM14_1775 and *rstA* genes. To determine the ancestry of these gene arrangements—specifically whether the *usf*-*ompN* IGR was gained by *E. coli* or lost by *Salmonella*—we analyzed the organization of the orthologous regions in the genomes of other enteric bacteria. The *usf*-*ompN* IGR is present and intact in other enteric species (*Klebsiella pneumoniae*, *Enterobacter aerogenes*, and *Citrobacter koseri*) establishing that this IGR predates the split between *E. coli* and *Salmonella* and was lost in *Salmonella* due to the relocation of *ompN* gene (Figure B.4, Figure B.5). As a consequence of this genome rearrangement in *Salmonella*, EcsR1 was split into two fragments located ≈200 kb apart, neither of which is transcribed (supplementary fig.
**Figure B.3 – Evolution of a new sRNA promoter.** Sequences immediately upstream of STM14_1869 (4471230–4471245) and its ortholog yigH in *E. coli* (1636369–1636384) are aligned. Numbers of RNA-seq reads mapping to this region are shown (black, *Salmonella*; blue, *E. coli*). The new *Salmonella* $\sigma^{70}$ promoter and sRNA (SesR2) transcription start site are boxed. Asterisks indicate point mutations that differentiate the *Salmonella* sequence from the corresponding region in *E. coli*. 
Figure B.4 – Distribution of the uspF–ompN IGR among enteric species. (A)
Alignment of genomic regions containing the uspF–ompN IGR in E. coli and three other enteric species. Note that in both Citrobacter koseri and Klebsiella pneumoniae, small ORFs (gray arrows situated between uspF [purple] and ompN [blue]) have been predicted to occur in this IGR. (B) Phylogenetic tree (modified from Petty et al. 2010) showing the presence or absence of the uspF–ompN IGR among species.
Figure B.5 – Loss of uspF–ompN IGR through genome rearrangement. The uspF–ompN IGR of *E. coli* was fragmented in *Salmonella* due to the translocation of *ompN* to a site adjacent to STM14_1775. The predicted FNR- and CRP-binding sites (yellow and blue, respectively; overlapping region in green) upstream of the sRNA (EcsR1) transcription start site (sRNA TSS) are shown. A predicted Rho-independent terminator (stem-loop structure) situated 3’ of the sRNA is also depicted.
S4, found in Additional_file_A8.pdf).

**EcsR1 Is Associated with Global Regulators in *E. coli***

The *uspF–ompN* IGR and EcsR1 are present in all strains of *E. coli*, which suggests that it maintains a regulatory function. To identify genes that are potentially under the control of EcsR1, we examined the effect of its overexpression on *E. coli* genes genome-wide, an approach that has been used previously to characterize the regulatory targets of sRNAs (Durand and Storz 2010; Beisel and Storz 2011). When analyzed by RNA-seq, the expression levels of 43 genes were significantly different (*P* < 0.05) in the EcsR1-overexpressing strain when compared with wild-type *E. coli* (supplementary table S4, found in Additional_file_A8.pdf). A GO analysis uncovered bacterial membrane (GO:0031090), carbohydrate catabolic process (GO:0016052), and nitrate metabolic process (GO:0042126) as processes that were significantly enriched (*P* < 0.05) in our data set. Eleven downregulated genes were associated with these GO terms, of which nine were regulated by CRP and/or FNR (supplementary table S5, found in Additional_file_A8.pdf) (Constantinidou et al. 2006; Keseler et al. 2013). In concert with these observations, a 22-nt palindromic sequence with features resembling the consensus CRP-binding site and a putative 15-nt FNR-binding sequence *(Figure B.5)* were identified upstream of EcsR1, indicative of the sRNA being part of the CRP and FNR regulons. Expression of another *E. coli* sRNA, FnrS, is known to be affected by both FNR and CRP (Durand and Storz 2010), showing that CRP and FNR regulons overlap and may control multiple sRNAs. It has been shown previously that the
transcriptional regulator CRP can control the expression of both an sRNA and the sRNA’s target genes, and this “feed-forward loop” is thought to aid in the efficient modulation of gene expression in *E. coli* (Beisel and Storz 2011).

Because many sRNAs in *E. coli* associate with and are stabilized by the RNA-binding protein Hfq (De Lay et al. 2013), we examined whether Hfq stabilizes EcsR1. We measured the abundance of EcsR1 in wild-type and Hfq-deficient strains of *E. coli*, and found it to be significantly more abundant in the wild-type strain (supplementary fig. S5, found in Additional_file_A8.pdf), reinforcing its identity as an sRNA.

**EcsR1 Impacts Biofilm Formation**

Because CRP and FNR control carbohydrate and nitrate metabolism during biofilm formation (Van Alst et al. 2007; Karatan and Watnick 2009), we constructed EcsR1-deletion and EcsR1-overexpression *E. coli* strains and measured the impact of this sRNA on biofilm formation. As shown in Figure B.6, biofilm production increased significantly (*P* < 0.0001) in the EcsR1-deleted strain when compared with wild-type *E. coli*.

Reintroduction of a plasmid-borne copy of EcsR1 into the deletion strain reduced biofilm formation to the same level as that of the wild-type strain (Figure B.6), indicating that biofilm-inhibition is an sRNA-specific phenotype.

**Expression of *E. coli* EcsR1 in *Salmonella* Activates Invasion-Associated Genes**

Biofilm production is important to virulence of enteric pathogens, so we tested the effects of EcsR1 on biofilm production in *Salmonella* by reintroducing the sRNA in an
Figure B.6 – Biofilm formation is influenced by EcsR1. *Escherichia coli* biofilms stained with crystal violet were measured (A<sub>600</sub>) after 48-h growth at 28 °C and normalized to OD<sub>600</sub> value. A wild-type strain, an EcsR1-deleted strain (ΔEcsR1), a ΔEcsR1 strain containing pBAD with cloned EcsR1 (ΔEcsR1-pBAD-EcsR1), and a ΔEcsR1 strain containing empty pBAD (ΔEcsR1-empty pBAD) were tested. Asterisks indicate a statistically significant difference between wild-type and ΔEcsR1 strains (P < 0.0001).
expression vector. There was no significant difference in biofilm production between the wild-type and EcsR1-overexpression strains; the overexpression of EcsR1 in Salmonella alters the expression of 128 genes genome-wide (supplementary table S6, Supplementary Material online). GO analysis revealed nine processes (representing 27 genes) that were significantly enriched within this gene set, with “pathogenesis” (GO:0009405) being the most highly significant (supplementary table S7, found in Additional_file_A8.pdf). Among genes regulated by this sRNA, 22 are known to promote Salmonella invasion of host cells, most of which are situated within SPI-1 pathogenicity island (Fàbrega and Vila 2013).
B.5 DISCUSSION

Our search for species-specific sRNAs was directed toward IGRs that were unique to either *E. coli* or *Salmonella* because most bacterial regulatory sRNAs are contained within these noncoding regions, although 3′-untranslated regions (UTRs) of mRNAs and promoters within mRNAs can also give rise to sRNAs (Chao et al. 2012; Guo et al. 2014). We found that genome rearrangements have altered IGRs and, in doing so, caused disparity in sRNA contents of these two species. A newly discovered sRNA (EcsR1), situated within the IGR between the *uspF–ompN* genes in *E. coli*, is absent from *Salmonella* due to the translocation of a genomic segment containing the *ompN* gene. This sRNA is associated with the FNR and CRP regulons, and its expression impacts *E. coli* biofilm formation. Additionally, we identified an sRNA (SesR2) unique to *Salmonella* that evolved de novo in an IGR that was formed through a phage-mediated genome rearrangement. Although disparities in genome architectures are common among related species, these are the first known cases where rearrangements have caused the generation and destruction of sRNAs.

The main source of rearrangement events in bacterial genomes is homologous recombination across identical sequences. *Escherichia coli* and *Salmonella* contain numerous classes of repeat elements that can serve as templates for exchange (Rocha 2004). In addition, recombination between bacteriophage sequences in a genome can result in altered genome architectures in related bacteria (Brüssow et al. 2004); large proportions of both *E. coli* and *Salmonella* genomes consist of prophage genes (Bobay et
al. 2013). Notwithstanding the large number of targets for homologous exchange, the
gene order of *E. coli* and *Salmonella* has been well conserved despite an estimated 100-
Myr divergence (Ochman et al. 1999). The major difference in their genome architectures
involves a large 600-kb inversion spanning the replication terminus and approximately 50
small-scale translocation events (supplementary fig. S1, found in

**Additional_file_A8.pdf**). This contrasts the situation in many bacteria, such as *Yersinia*
and *Portiera* (Parkhill et al. 2001; Sloan and Moran 2013), in which there have been
substantial changes in gene arrangement among closely related strains. The source of this
variation has been ascribed not only to the differences among species in their repertoires
of DNA recombination and repair enzymes (Tamas et al. 2002) but also to selection on
gene order and position (Suyama and Bork 2001; Ballouz et al. 2010; Treangen and
Rocha 2011). Our analyses show that some fraction of the rearrangements that shuffle
IGRs may affect organismal fitness by disrupting or generating regulatory elements.

Although the IGR between *yagU* and *ompN* was disrupted and split in
the *Salmonella* genome, some portions of it—approximately 70 nt of the 5′-end of EcsR1
and 20 nt of its 3′-end—are still recognizable adjacent to the *uspF* and *ompN* genes
in *Salmonella* (**Figure B.5**). It is likely that these sRNA segments are not transcribed and
are not functional in *Salmonella* (supplementary fig. S4, found in

**Additional_file_A8.pdf**, and Kroger et al. 2013) because nucleotide substitutions in the
putative CRP- and FNR-binding regions (supplementary fig. S6, found in

**Additional_file_A8.pdf**) have rendered them inactive. The reintroduction of EcsR1
into *Salmonella* did not affect biofilm production but instead triggered the increased
expression of several virulence genes, particularly those within the SPI-1 pathogenicity island (supplementary table S7, found in Additional_file_A8.pdf). Multiple factors, including the biofilm machinery, are known to regulate the expression of these invasion-associated genes (Fàbrega and Vila 2013), suggestive of links between the different phenotypes produced by this sRNA in *E. coli* and *Salmonella*. Additional experiments are necessary to understand how EcsR1 induces diverse phenotypes in the two species; nevertheless, our findings demonstrate the potential of sRNAs to influence bacterial adaptation and evolution.

In addition to losing the biofilm-reducing sRNA (EcsR1), *Salmonella* has gained, again by a rearrangement event, an IGR that contains an sRNA (SesR2) that is not present in other enteric species. Because none of the corresponding regions in *E. coli* displays any appreciable transcript production, the evolution of this new sRNA in *Salmonella* also required the de novo formation of a new promoter sequence (Figure B.3 and supplementary fig. S3, found in Additional_file_A8.pdf). In bacterial genomes, $\sigma^{70}$ promoter-like sequences are able to arise spontaneously through point mutations, especially in IGRs (Stone and Wray 2001; Huerta et al. 2006; Mendoza-Vargas et al. 2009), and transcription can originate from newly evolved $\sigma^{70}$ promoters (Mendoza-Vargas et al. 2009; Raghavan et al. 2012). Therefore, it is most likely that an incipient promoter in the newly formed STM14_1869–STM14_1870 IGR gave rise to the transcript that evolved into SesR2. An alternate possibility is that SesR2 was introduced into *Salmonella* with the HGT event that brought in the entire STM14_1870–STM14_1872 region, as has been proposed for other sRNAs located close to transposon
insertion sites in *Salmonella* (Sittka et al. 2008). However, no similar sRNA is detectable in the homologous regions found on various enteric plasmids and genomes, making this scenario less likely. Finally, because the first 55 nt of this sRNA is complementary to the 5′-UTR of STM14_1869 (TSS of STM14_1869 is located 63 bp upstream of coding region), it could be functioning as an antisense RNA to regulate STM14_1869 expression, as shown previously for other genes in *Salmonella* (Lee and Groisman 2010).

In bacteria, differences in protein-coding gene contents between closely related species are either due to new genes that arose by gene duplication or HGT (Lerat et al. 2005; Blount et al. 2012; Nasvall et al. 2012), or due to gene loss through pseudogenization and deletion (Mira et al. 2001; Kuo and Ochman 2010). Although the mechanisms that shape bacterial sRNA gene repertoires are not well understood, duplication, deletion, and HGT have also been attributed to this process (Gottesman and Storz 2011). In this report, we show that genome rearrangements that create and disrupt IGRs can result in the gain or loss of sRNA genes in bacteria. Because sRNAs regulate myriad metabolic processes, this disparity in sRNA repertoires between closely related bacteria might also contribute to niche adaptation and speciation events.


Horler RS, Vanderpool CK. Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence. Nucleic Acids Res. 2009;37:5465–5476.


APPENDIX C

Emergence of new sRNAs in enteric bacteria is associated with low expression and rapid evolution

C.1 ABSTRACT

Non-coding small RNAs (sRNAs) are critical to post-transcriptional gene regulation in bacteria. However, unlike for protein-coding genes, the evolutionary forces that shape sRNAs are not understood. We investigated sRNAs in enteric bacteria and discovered that recently emerged sRNAs evolve at significantly faster rates than older sRNAs. Concomitantly, younger sRNAs are expressed at significantly lower levels than older sRNAs. This process could potentially facilitate the integration of newly emerged sRNAs into bacterial regulatory networks. Furthermore, it has previously been difficult to trace the evolutionary histories of sRNAs because rapid evolution obscures their original sources. We overcame this challenge by identifying a recently evolved sRNA in *Escherichia coli*, which allowed us to determine that novel sRNAs could emerge from vestigial bacteriophage genes, the first known source for sRNA origination.
C.2 INTRODUCTION

Non-coding RNAs (ncRNAs) regulate gene expression in all domains of life. In Eukaryotes, post-transcriptional control of gene expression by ncRNA such as microRNA (miRNA) and small interfering RNA is now recognized as a fundamental layer of gene regulation (Wilson and Doudna 2013). While less studied, Archaea also contain a large repertoire of ncRNAs (Bernick et al. 2012). In Bacteria, a major class of ncRNA is small RNAs (sRNAs), which are around 50–200 nucleotides in length, and regulate gene expression by binding to messenger RNAs (mRNAs) (Gottesman and Storz 2011). Broadly, sRNAs are classified into two major types: (1) trans-acting sRNAs that are encoded in intergenic regions and regulate the expression of distantly located genes via imperfect complementarity, and (2) cis-acting sRNAs (also called antisense RNAs) that are transcribed from the opposite strand of adjacent target genes and regulate gene expression through perfectly complementary regions (Thomason and Storz 2010; Georg and Hess 2011). Some of the other major classes of bacterial ncRNA are riboswitches and RNA thermometers that are located in the untranslated regions of certain mRNAs (Breaker 2011; Kortmann and Narberhaus 2012), and intraRNAs that originate from within protein-coding genes (Miyakoshi et al. 2015).

Several advantages of sRNAs over proteins have led to their emergence as important gene regulators in bacteria. For instance, relatively lower energy is required to synthesize sRNAs; they act rapidly and their co-degradation along with target mRNAs allows precise control of regulatory circuits (Storz et al. 2011; Updegrove et al. 2015).
Each sRNA typically controls multiple genes belonging to interconnected processes, including metabolic pathways, quorum sensing, biofilm formation, and virulence (Michaux et al. 2014), and because there are hundreds of sRNAs in each bacterium, their contribution to bacterial adaptation and phenotypic diversity could be tantamount to that of protein-coding genes (Gottesman and Storz 2011; Raghavan et al. 2011; Kröger et al. 2013). Hence, differences in sRNA sequences and contents between closely related bacteria could have a substantial impact on bacterial physiology and pathogenicity.

However, we have minimal knowledge about the evolutionary forces that shape bacterial sRNAs. An analysis of the conservation of sRNAs in *Escherichia coli* showed that variation in sRNA contents between strains is mainly due to deletions (Skippington and Ragan 2012), and a broader examination of sRNA conservation across bacteria revealed that most *E. coli* sRNAs originated after enteric bacteria split from other Gammaproteobacteria (Peer and Margalit 2014). In addition to this cycle of birth and loss, sRNA genes evolve at faster rates than protein-coding genes, making it difficult to identify sRNA homologs in distantly related bacteria (Gardner et al. 2005; Hoeppner et al. 2012). One group of bacteria that was shown to be at optimum evolutionary distance for effective evolutionary analysis of sRNAs is the family Enterobacteriaceae (Lindgreen et al. 2014), which includes *E. coli* and *Salmonella enterica*, two model organisms in which sRNAs have been investigated thoroughly (e.g., Raghavan et al. 2011; Kröger et al. 2013).

In this study, we estimated the evolutionary ages of >200 sRNAs present in *E. coli* and *S. enterica*, and show that younger sRNAs are expressed at significantly lower
levels than older sRNAs, and that younger sRNAs have significantly higher rates of evolution than older sRNAs. The low expression of new sRNAs could mitigate the negative effects of nascent sRNA–mRNA interactions, whereas their rapid evolution could generate beneficial interactions that facilitate their integration into bacterial regulatory networks. We also show that most sRNAs are evolving under purifying selection, and discovered that a young sRNA in *E. coli* originated from a vestigial bacteriophage protein-coding gene, thereby revealing the first known source for sRNA origination in bacteria.
C.3 MATERIALS AND METHODS

Bacterial Strains and Plasmids

*Escherichia coli* K-12 MG1655 was used in all experiments. For EcsR2 expression vector construction, EcsR2 gene was amplified using the following primers:

\[
\text{5'}\text{ATGCTAGCGCAGATAGTCAGTGAGTATATC3'}, \text{5'}\text{GACGTCGCAGATAGTCAGTGAGTATATC3'}, \text{5'}\text{GACGTCGCAGATAGTCAGTGAGTATATC3'}, \text{5'}\text{GACGTCGCAGATAGTCAGTGAGTATATC3'},
\]

and cloned into the plasmid pBAD (Guzman et al. 1995) by digesting both the PCR product and pBAD with NheI and AatII restriction enzymes (restriction sites on primers are underlined). EcsR2-deletion strain was constructed using \(\lambda\) Red-mediated recombination (Datsenko and Wanner 2000).

RNA-Seq and Crosslink-Seq

Highest level of EcsR2 expression was observed during exponential phase growth (Fig. S4, found in Additional_file_A9.pdf). Hence, for the RNA-seq analysis, *E. coli* transformed with either empty pBAD (control) or pBAD with cloned EcsR2 (test) that were grown in Lysogeny Broth (LB) aerobically to OD\(_{600}\) of ~0.5. Cultures were supplemented with arabinose (0.2%) for 10 min to induce the expression of EcsR2, 0.2 volume stop solution (5% water-saturated phenol, 95% ethanol) was added, and the cells were harvested by centrifugation. Total RNA was extracted using TRI reagent, treated with DNase, and ribosomal RNAs were removed using MICROBExpress kit (Life Technologies). RNA-seq (Illumina HiSeq 2000, 100 cycles, single-end) was performed using two control and test samples at the Genomic Sequencing and Analysis Facility at
the University of Texas at Austin. The trimmed reads were mapped to the *E. coli* genome (NC_000913.2) using CLC Genomics Workbench to identify genes that were differentially expressed. The RNA-seq reads are available on NCBI SRA (accession: SRP044074).

Crosslink-seq was adapted from Lustig et al. 2010 and Liu et al. 2015. EcsR2-deletion strain containing either empty pBAD (control) or pBAD with cloned EcsR2 (test) were grown in LB aerobically to OD<sub>600</sub> of ~0.5 and cultures were supplemented with arabinose (0.2%) for 10 min to induce the expression of EcsR2. Cells were washed twice with PBS, resuspended in 8 mL of fresh PBS, and 0.2 mg/mL 4′-Aminomethyltrioxsalen (Cayman Chemicals) was added. The cells were incubated on ice for 10 min, and were irradiated with UV light at 365 nm for 1 h on ice. The cells were washed once with PBS and total RNA was isolated using TRI reagent. RNA treated with DNase was mixed in hybridization buffer (20 nM HEPES pH8, 5 mM MgCl<sub>2</sub>, 300 mM KCl, 0.01% NP-40, 1 mM DTT) and heated at 80 °C for 2 min followed by immediate cooling on ice. 10 nmol biotinylated oligonucleotides that were antisense to a portion of EcsR2 were added to the samples and incubated at room temperature overnight. 150 µL of NeutrAvidin agarose resin (Thermo Fisher) was washed twice in WB100 buffer (20 mM HEPES pH 8, 10 mM MgCl<sub>2</sub>, 100 mM KCl, 0.01% NP-40, 1 mM DTT) followed by blocking the beads for 2 h (blocking buffer: WB100, 50 µL BSA (10 mg/mL), 40 µL tRNA (10 mg/mL), 10 µL glycogen (20 mg/mL)). The blocked beads were once again washed with blocking buffer and added to the hybridized RNAs bound to the biotinylated oligos. Samples were incubated for 4 h at 4 °C and then washed five times with WB400
buffer (20 mM HEPES pH 8, 10 mM MgCl2, 400 mM KCl, 0.01% NP40, 1 mM DTT). The hybridized RNAs bound to the beads were isolated using TRI reagent. The affinity-selected, crosslinked mRNAs were released from EcsR2 using UV light at 254 nm on ice for 15 min. The RNA samples were deep-sequenced at Oregon Health and Science University Massively Parallel Sequencing Shared Resource (Illumina HiSeq 2500, 100 cycles, single-end), and the trimmed reads were mapped to *E. coli* genome (NC_000913.2) using CLC Genomics Workbench to determine the genes that were enriched in test samples (expressing EcsR2) in comparison to controls (no EcsR2). Gene expression was calculated from two independent experiments, and the RNA-seq reads are available on NCBI SRA (accession: SRP074317).

For qRT-PCR confirmation, EcsR2-deletion strain containing empty pBAD, or pBAD with cloned full-length EcsR2, or pBAD with EcsR2 in which the +51 to +80 region was deleted using inverse PCR were used. Bacteria were grown in LB aerobically to OD$_{600}$ of ~0.5. Cultures were supplemented with arabinose (0.2%) for 10 min to induce the expression of EcsR2, and 0.2 volume stop solution was immediately added, and the cells were harvested by centrifugation. Total RNA was extracted using TRI reagent, treated with DNase, and qRT-PCR was performed as previously described (Raghavan et al. 2011).

**Expression and Evolution of sRNAs**

We used previously published RNA-seq data to determine the expression of sRNAs (Raghavan et al. 2011; Kröger et al. 2013). To identify the homologs of 92 sRNAs
described in *E. coli* K-12 MG1655 (NC_000913.2) (Raghavan et al. 2011, 2015), we used BLASTn (*E* value < $10^{-5}$ and target length ≥60% of query length) to search 146 fully sequenced *E. coli* genomes available on NCBI. We ultimately chose 81 sRNAs that were conserved in 85 *E. coli* genomes in order to maximize the number of genomes and sRNAs (Supplemental dataset 1). Similarly, we searched 151 *S. enterica* genomes to identify the homologs of 170 sRNAs described in *S. enterica* Typhimurium SL1344 (NC_016810.1) (Kröger et al. 2013), and chose 127 sRNAs that are conserved in 112 *S. enterica* genomes for further analyses (Supplemental dataset 1). Sequences were aligned using Clustal Omega (Sievers et al. 2011), and nucleotide differences were quantified using nucleotide diversity index $\pi$ (Nei 1987; Jovelin and Cutter 2014) with DnaSP 5.10 (Librado and Rozas 2009). Briefly, $\pi$ was calculated by summing, over all distinct pairs of sequences in the sample, the proportion of different nucleotides between a pair of sequences multiplied by the respective frequencies of those sequences. To calculate the average nucleotide differences throughout the *yagU–ykjG* IGR, we used a sliding window of 35 bp and a step size of 15 bp using DnaSP. RNA secondary structure and minimum free energy were predicted using Vienna RNA webserver (Gruber et al. 2008) and Mfold webserver (Zuker 2003), and EcsR2–AnsB interaction was modeled using IntaRNA (Wright et al. 2014).

To determine whether the sRNAs in *E. coli* and *S. enterica* were present in other enteric bacteria, sRNA gene sequences were searched (BLASTn, *E* < $10^{-5}$ and target length ≥60% of query length) against the following representative genomes (as denoted by NCBI Genome database): *Citrobacter freundii* (NZ_CP007557.1), *Klebsiella*
pneumoniae (NC_016845.1), Serratia marcescens (NZ_HG326223.1), and Yersinia enterocolitica (NC_008800.1). PMCMR R package was used to perform both Kruskal–Wallis test (non-parametric 1-way ANOVA) to assess differences in expression and nucleotide diversity between sRNA age classes, and the post hoc pairwise comparison Dunn’s test. For analyzing S. enterica expression data from 22 growth conditions (Kröger et al. 2013), Permutational ANOVA (non-parametric 2-way ANOVA) was conducted using perm.anova, and post hoc pairwise comparisons were conducted using pairwise.perm.t.test with FDR correction in the RVAideMemoire R package. To analyze sRNA evolution in more detail, 38 sRNAs that were conserved in at least 50% of currently available complete genomes of E. coli (74 strains) and S. enterica (102 strains) were chosen (Table C.1). To detect purifying selection, within-species polymorphism and between-species divergence were calculated using DnaSP; four-fold degenerate sites (in hundred randomly selected genes; Table C.2) were used as control because they are considered to evolve neutrally (Ochman and Wilson 1987).
Table C.1 – Genomes of *E. coli* (85 strains) and *S. enterica* (102 strains) used in analysis

<table>
<thead>
<tr>
<th>Sample</th>
<th>Genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> genomes</td>
<td>NC_000913.2, NC_004431.1, NC_007779.1, NC_007946.1, NC_008563.1, NC_008901.1, NC_010468.1, NC_010498.1, NC_011415.1, NC_011741.1, NC_011748.1, NC_011751.1, NC_011993.1, NC_013353.1, NC_013941.1, NC_016902.1, NC_017625.1, NC_017626.1, NC_017628.1, NC_017631.1, NC_017632.1, NC_017633.1, NC_017634.1, NC_017638.1, NC_018500.1, NC_018658.1, NC_018661.1, NC_020163.1, NC_022364.1, NC_022370.1, NC_CP005903.1, NC_CP006027.1, NC_CP006262.1, NC_CP006632.1, NC_CP007133.1, NC_CP007136.1, NC_CP007149.1, NC_CP007265.1, NC_CP007390.1, NC_CP007391.1, NC_CP007392.1, NC_CP007393.1, NC_CP007442.1, NC_CP007491.1, NC_CP007594.1, NC_CP008801.1, NC_CP009072.1, NC_CP009104.1, NC_CP009106.2, NC_CP009166.1, NC_CP009273.1, NC_CP009685.1, NC_CP010344.1, NC_CP011320.1, NC_CP011321.1, NC_CP011331.1, NC_CP011342.2, NC_CP011343.2, NC_CP011461.1, NC_CP011495.1, NC_CP012125.1, NC_CP012126.1, NC_CP012127.1, NC_CP012625.1, NC_CP012631.1, NC_CP012635.1, NC_CP012868.1, NC_CP012869.1, NC_CP012870.1, NC_CP013025.1, NC_CP013112.1, NC_CP014251.1, NC_CP014597.1, NC_CP008790.1, NC_CP009081.1, NC_CP009082.1, NC_CP009083.1, NC_CP009084.2, NC_CP010279.1, NC_CP010280.1, NC_CP010281.1, NC_CP010282.1, NC_CP010283.1, NC_CP010284.1, NC_CP011394.1, NC_CP011396.1, NC_CP011428.1, NC_CP011790.1, NC_CP011791.1, NC_CP012144.1, NC_CP012347.1, NC_CP012513.1, NC_CP012514.1, NC_CP012681.1, NC_CP012921.1, NC_CP012924.1, NC_CP012930.1, NC_CP013097.1, NC_CP014051.1, NC_CP014356.1, NC_CP014358.1, NC_CP014536.1, NC_CP014538.1, NC_CP014539.1</td>
</tr>
<tr>
<td><em>S. enterica</em> genomes</td>
<td>NC_003197.1, NC_006905.1, NC_010102.1, NC_01080.1, NC_011083.1, NC_011205.1, NC_011247.1, NC_011294.1, NC_012125.1, NC_016810.1, NC_016831.1, NC_016854.1, NC_016856.1, NC_016857.1, NC_016863.1, NC_017046.1, NC_017623.1, NC_021151.1, NC_021810.1, NC_021812.2, NC_021844.1, NC_021902.1, NC_022221.1, NC_022525.1, NC_022544.1, NC_022569.1, NC_AP014565.1, NC_CP005995.1, NC_CP007175.1, NC_CP007216.1, NC_CP007235.1, NC_CP007245.1, NC_CP007246.1, NC_CP007247.1, NC_CP007248.1, NC_CP007249.2, NC_CP007250.1, NC_CP007251.1, NC_CP007252.1, NC_CP007253.1, NC_CP007254.1, NC_CP007255.1, NC_CP007256.1, NC_CP007258.1, NC_CP007259.1, NC_CP007260.1, NC_CP007261.1, NC_CP007262.1, NC_CP007263.1, NC_CP007267.2, NC_CP007269.1, NC_CP007271.1, NC_CP007280.2, NC_CP007282.2, NC_CP007283.2, NC_CP007285.2, NC_CP007286.2, NC_CP007288.2, NC_CP007292.1, NC_CP007294.2, NC_CP007312.2, NC_CP007319.1, NC_CP007333.1, NC_CP007425.1, NC_CP007426.1, NC_CP007434.2, NC_CP007507.1, NC_CP007523.1, NC_CP007528.1, NC_CP007534.1, NC_CP007559.1, NC_CP007581.1, NC_CP007598.1, NC_CP008928.1, NC_CP009083.1, NC_CP009084.2, NC_CP009085.2, NC_CP009086.1, NC_CP009087.1, NC_CP009088.1, NC_CP009089.1, NC_CP009090.1, NC_CP009091.1, NC_CP009092.1, NC_CP009093.1, NC_CP009102.1, NC_CP009561.1, NC_CP009768.1, NC_CP010279.1, NC_CP010280.1, NC_CP010281.1, NC_CP010282.1, NC_CP010283.1, NC_CP010284.1, NC_CP011394.1, NC_CP011396.1, NC_CP011428.1, NC_CP011790.1, NC_CP011791.1, NC_CP012144.1, NC_CP012347.1, NC_CP012513.1, NC_CP012514.1, NC_CP012681.1, NC_CP012921.1, NC_CP012924.1, NC_CP012930.1, NC_CP013097.1, NC_CP014051.1, NC_CP014356.1, NC_CP014358.1, NC_CP014536.1</td>
</tr>
</tbody>
</table>
Table C.2 – 100 genes selected in *E. coli* (NC_000913) and *S. enterica* (NC_016810) selected for four-fold degenerate site analysis

<table>
<thead>
<tr>
<th>Sample</th>
<th>Genes</th>
</tr>
</thead>
</table>
C.4 RESULTS

Newly Evolved sRNAs have Low Expression and Rapid Rate of Evolution

We identified the homologs of 81 *E. coli* sRNAs (Raghavan et al. 2011, 2015) and 127 *S. enterica* Typhimurium sRNAs (Kröger et al. 2013) in *Citrobacter freundii, Klebsiella pneumoniae, Serratia marcescens,* and *Yersinia enterocolitica* using BLASTn. This approach has been shown to be effective at identifying sRNA homologs within Enterobacteriaceae (Skippington and Ragan 2012; Peer and Margalit 2014). We utilized maximum parsimony to estimate the age of each sRNA along a 16S rDNA phylogenetic tree that encompasses the six enteric bacteria (Figure C.1), as done previously to study miRNA evolution (Lyu et al. 2014). We assigned the sRNAs into three age groups: old (those that originated in the common ancestor of all six bacteria), middle-aged (those that originated in the common ancestor of *E. coli, S. enterica, C. freundii,* and *K. pneumoniae*), and young (those that originated in the *E. coli, S. enterica* branch) (Figure C.1). Based on this classification, *E. coli* contained 21 young, 27 middle-aged, and 33 old sRNAs, whereas *S. enterica* contained 53 young, 48 middle-aged, and 26 old sRNAs (Additional_file_A10.csv, Additional_file_A11.csv).

We analyzed the expression of sRNAs using RNA-seq data for exponential phase growth of *E. coli* in Lysogeny Broth (Raghavan et al. 2011), and *S. enterica* Typhimurium in Lennox Broth (Kröger et al. 2013), and discovered that sRNA expression correlated with sRNA age: younger sRNAs have significantly lower expression than older sRNAs (Figure C.2, Additional_file_A10.csv, Additional_file_A11.csv).
Figure C.1 – sRNA age groups. sRNAs were binned based on their presence in six enteric bacteria using maximum parsimony.
**Figure C.2 – Younger sRNAs have lower expression.** (A) Mean expression (±SE) of 81 sRNAs in *E. coli* grown in Lysogeny Broth, and (B) 127 sRNAs in *S. enterica* grown in Lennox Broth are shown. sRNA expression was measured using RNA-seq (Raghavan et al. 2011; Kröger et al. 2013). sRNAs were binned as described in Figure C.1 and *p*-values were calculated using Kruskal–Wallis test and Dunn’s test for multiple comparisons.
Additional_file_A11.csv). In order to rule out the possibility that the observed relationship between sRNA age and expression is an artifact of the growth conditions, we expanded our analysis of *S. enterica* sRNAs to all 22 “infection-relevant” growth conditions described by Kroger et al. (2013). Significantly reduced expression of young sRNAs in comparison to older sRNAs was observed under all growth conditions (Table S1, found in Additional_file_A9.pdf), showing that the relationship between sRNA age and expression is not dependent on the growth condition (*p* = 0.818, Permutational ANOVA). Furthermore, to confirm that our conclusions are independent of the BLAST algorithm’s ability to locate sRNA homologs in other enteric species, we performed a similar analysis using 49 *E. coli* sRNAs described in the Rfam database, which uses a different approach (covariance model) to identify homologous sRNAs (Nawrocki et al. 2015). As shown previously (Peer and Margalit 2014), we got comparable results using either BLASTn or Rfam (Table C.3, Additional_file_A12.csv).

We also measured the rate of evolution of each sRNA by calculating the nucleotide diversity index π, the average number of nt differences per site, using homologs in 85 *E. coli* and 112 *S. enterica* strains (Supplemental dataset 1) (Nei 1987; Jovelin and Cutter 2014). We discovered that the rate of sRNA evolution inversely correlated with age i.e., younger sRNAs evolved at significantly higher rates than older sRNAs (Figure C.3; Additional_file_A10.csv, Additional_file_A11.csv). Although we examined sRNAs only in enteric bacteria, the observed relationship among sRNA age, expression, and rate of evolution is likely to be a widespread phenomenon because previous studies in humans, nematodes, and *Drosophila* have shown that younger
Table C.3 – Comparison of mean expression and nucleotide polymorphism of 49 rfam sRNAs in *E. coli* grown in Lysogeny Broth. sRNA expression was measured using RNA-seq (Raghavan et al. 2011; Kröger et al. 2013). sRNAs were binned as described in Figure C.1 and *p*-values were calculated using Kruskal–Wallis test and Dunn’s test for multiple comparisons.

<table>
<thead>
<tr>
<th>Gene</th>
<th>sRNA Expression</th>
<th></th>
<th>Nucleotide Polymorphism</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chi-Square</td>
<td><em>p</em>-value</td>
<td>Chi-Square</td>
<td><em>p</em>-value</td>
</tr>
<tr>
<td>Overall</td>
<td>13.13</td>
<td>0.001</td>
<td>10.69</td>
<td>0.005</td>
</tr>
<tr>
<td>young vs. middle</td>
<td>0.74</td>
<td>0.389</td>
<td>1.86</td>
<td>0.172</td>
</tr>
<tr>
<td>middle vs. old</td>
<td>7.84</td>
<td>0.005</td>
<td>4.36</td>
<td>0.037</td>
</tr>
<tr>
<td>young vs. old</td>
<td>9.48</td>
<td>0.002</td>
<td>9.33</td>
<td>0.002</td>
</tr>
</tbody>
</table>
Figure C.3 – Younger sRNAs have higher rates of evolution. (A) Average (±SE) π values for 81 sRNAs in *E. coli*, and (B) 127 sRNAs in *S. enterica* are shown. sRNAs were binned as described in Figure C.1 and *p*-values were calculated using Kruskal–Wallis test and Dunn’s test for multiple comparisons.
miRNAs have lower expression and faster rate of evolution than older miRNAs (Chen and Rajewsky 2007; Jovelin and Cutter 2014; Lyu et al. 2014).

**A Young sRNA in *E. coli* Originated from a Degraded Bacteriophage Gene**

One of the young sRNAs that had low expression and rapid rate of evolution was EcsR2 (Figure C.4), an sRNA found exclusively in *E. coli* (Raghavan et al. 2015). In order to understand the origin of EcsR2, we traced the evolutionary history of the *yagU–ykgJ* intergenic region (IGR) that contains this sRNA. The arrangement in which *yagU* neighbors *ykgJ* is found only in *E. coli*, whereas an alternate, and likely ancestral, gene order (*ycIC–ykgJ–ompW*) is present in most other enteric bacteria (Figure C.5). The ancestral gene arrangement is also present in *E. albertii*, which is one of *E. coli*’s closest relatives, indicating that *ykgJ* moved to its current location in *E. coli* after the two bacteria split from a common ancestor. Additionally, the *ykgJ* ORF (open reading frame) is smaller in *E. coli* than in *E. albertii*, and ~ 90 bp remnant of the gene’s 3’ end is still recognizable in the *ycIC-ompW* IGR in *E. coli*, confirming that *ykgJ* was translocated recently to its current location in *E. coli* to create the unique *yagU–ykgJ* IGR (Figure C.5).

Due to their rapid rate of evolution, it is usually difficult to trace the ancestry of sRNAs (Gottesman and Storz 2011). However, because EcsR2 emerged in an IGR that was formed recently, we were able to identify through sequence alignment that the sRNA evolved from a vestigial bacteriophage gene (Figure C.6). To identify genes that are potentially regulated by EcsR2, we used RNA-seq to detect changes in mRNA levels in
Figure C.4 – EcsR2 has low expression and rapid evolution. (A) Expression levels, and (B) nucleotide divergence index ($\pi$) of EcsR2 compared to two old sRNAs (RyhB and SsrS).
Figure C.5 – EcsR2 evolved in a novel IGR in *E. coli*. (A) The *yagU–ykgJ* IGR is present only in *E. coli*. (B) Relocation of *ykgJ* next to *yagU* in *E. coli* resulted in the formation of a novel IGR that contains EcsR2. Gene locations based on *E. albertii* (NZ_CP007025.1) and *E. coli* (NC_000913.2).
Figure C.6 – Origination of EcsR2 from a degraded prophage gene. (A) A pseudogenized version of a prophage gene (*tfaR*) evolved into EcsR2 by gaining a promoter-like sequence and an intrinsic terminator. Putative −10 and −35 sequences are in *green*, and the transcription start site of EcsR2 is in bold. (B) Alignment of EcsR2 gene and its homologous region in *tfaR* gene. The putative mRNA-binding region and the intrinsic terminator of EcsR2 are shown in *red* and *purple*, respectively. The *tfaR* stop codon is in bold. The 5′ end of EcsR2 (nucleotides 1–107) has ~75% identity, whereas the 3′ end (nucleotides 108–166) has ~50% identity to *tfaR*.
cells transiently expressing EcsR2 (cells with empty vector was used as control). This approach has been used previously to identify sRNA targets because pulse expression of sRNA limits indirect regulatory effects (e.g., Zhang et al. 1998; Wang et al. 2015). The RNA-seq analysis identified 26 genes that were significantly downregulated in the EcsR2-expressing strain (Table S2, found in Additional_file_A9.pdf). Further, we combined in vivo RNA crosslinking (Lustig et al. 2010; Liu et al. 2015) with RNA-seq to identify mRNAs that could directly interact with EcsR2 (Fig. S1, found in Additional_file_A9.pdf). This approach (Crosslink-seq) identified nine mRNAs that were potentially bound to EcsR2 (Table S3, found in Additional_file_A9.pdf).

One gene that was identified through both RNA-seq and Crosslink-seq as a potential direct target of EcsR2 was ansB (downregulated ~16 fold in RNA-seq, and enriched >3 fold in Crosslink-seq). In silico modeling predicted that EcsR2 could bind to AnsB mRNA via nucleotides within an unstructured region: positions +52 to +83 (Fig. S2, found in Additional_file_A9.pdf); coincidentally, using a sliding-window analysis that mapped the rate of polymorphism across EcsR2, we identified the same region to be evolving at a much lower rate than the rest of the sRNA (Figure C.7). Previous studies have shown that mRNA-binding sites are the most conserved regions within sRNAs (Peer and Margalit 2011; Richter and Backofen 2012), indicating that the +50 to +80 region is the potential AnsB-binding site. Additionally, although this region is highly conserved among E. coli strains (Figure C.7), it seems to have evolved considerably from its progenitor tfaR gene (Figure C.6, Fig. S3, found in Additional_file_A9.pdf), potentially due to its functional importance. To verify EcsR2’s ability to regulate ansB expression,
Figure C.7 – EcsR2 sequence conservation and predicted structure. (A) Nucleotide differences calculated for the \textit{yagU–ykgJ} IGR using a sliding-window analysis is represented by the \textit{black line}, and the flanking \textit{green lines} indicate the 95\% confidence interval. The locations of the 3’ ends of \textit{yagU} and \textit{ykgJ} genes are shown in \textit{yellow}. The most conserved region (\textit{blue}) within the IGR correlates with \textit{yagU}'s intrinsic terminator. The location of EcsR2 within the IGR is highlighted in \textit{purple}. The \textit{blue} and red \textit{arrows} on \textit{y}-axis indicate the average nucleotide polymorphism for EcsR2 and IGR, respectively. (B) The predicted secondary structure of EcsR2 was generated using mfold web server.
we pulse induced the expression of full-length EcsR2 and a version of EcsR2 in which
the putative binding site was deleted. We quantified ansB expression using qRT-PCR and
observed significant reduction in ansB expression only with the full-length version of
EcsR2 (Figure C.8), suggesting that the putative mRNA-binding region is required for
gene regulation.

**Conserved sRNAs are Under Purifying Selection**

To assess the impact of natural selection on sRNAs, we analyzed a subset of sRNAs
(n = 38) that are conserved in *E. coli* and *S. enterica* (74 and 102 strains, respectively)
(Supplemental dataset 1, found in Additional_file_A13.csv). As expected, younger
sRNAs have higher rates of polymorphism and divergence than older sRNAs (Figure
C.9). Interestingly, both young and old sRNAs are evolving at significantly lower rates
than genome-wide four-fold degenerate sites (proxy for neutral evolution), showing that
purifying selection is acting to preserve sRNAs in both bacteria, probably due to their
contribution to bacterial fitness, as shown previously for miRNAs in humans,
*Drosophila*, and *Caenorhabditis* (Quach et al. 2009; Jovelin and Cutter 2014; Lyu et
al. 2014).
Figure C.8 – EcsR2 downregulates *ansB* expression. Fold difference in *ansB* expression in the presence of full-length EcsR2 or EcsR2 without nucleotides +51 to +80, in comparison to a strain lacking EcsR2 (normalized to 1, *dashed line*). Data represent the means of three experiments ± standard deviations. *Asterisk* indicates $p < 0.01$ (unpaired *t* test).
Figure C.9 – sRNAs are under purifying selection. Selective constrains are stronger on sRNAs ($n = 38$) than on neutral sites (4-fold degenerate sites; $n = 100$). sRNAs present in the common ancestor of all six species were considered old and the rest as young. $p$-values calculated using Kruskal–Wallis test and Dunn’s test for multiple comparisons.
C.5 DISCUSSION

Although sRNAs are critical to gene regulation, we lack a clear understanding of how they originate and evolve in bacteria. In this study, we show that young sRNAs are expressed at low levels and evolve at faster rates than older sRNAs, thereby uncovering a novel process that potentially facilitates the establishment of new sRNAs in bacterial genomes. We also discovered that an sRNA (EcsR2) emerged from a degraded bacteriophage protein-coding gene, thus revealing the first known source for sRNA origination in bacteria. Similar to the origination of EcsR2, new ncRNA genes in eukaryotes have arisen from the remnants of protein-coding genes by gaining regulatory motifs (Kaessmann 2010; Ruiz-Orera et al. 2015). Additionally, in eukaryotes, the evolution of a spurious transcript into a functional ncRNA is associated with changes in the RNA’s secondary structure (Heinen et al. 2009). In concordance with this observation, in EcsR2, the putative mRNA-binding region appears to have become more unstructured, whereas the intrinsic terminator likely became more structured (Figure C.6, Fig. S3, found in Additional_file_A9.pdf). Analogous to EcsR2, another E. coli-specific sRNA IsrA (McaS) (Jørgensen et al. 2013) is also evolving at a rapid rate ($\pi = 0.050$), validating the observation that young sRNAs evolve swiftly in bacteria. Most of the other sRNAs with similarly high rates of evolution are antisense RNAs that are part of toxin–antitoxin systems (Fozo et al. 2008). Interestingly, SgrS, an sRNA present in several Gammaproteobacteria (Horler and Vanderpool 2009), displayed an elevated rate of evolution ($\pi = 0.060$). However, a closer examination revealed that the sRNA has
diverged considerably in 30 out of the 85 *E. coli* strains used in our analysis. If we consider only the other 55 genomes, the nucleotide diversity value for SgrS falls within the expected range for older sRNAs ($\pi = 0.0029$). The reasons for the disparity in SgrS evolutionary rates in the two *E. coli* cohorts are unknown.

Similar to protein-coding genes, most sRNA genes are evolving under purifying selection, suggestive of their importance to bacterial fitness; however, young sRNAs are evolving much more rapidly than evolutionarily older sRNAs, and young sRNAs are expressed at significantly lower levels than established sRNAs. One of the probable causes for the low expression could be that their promoters are not yet fully functional. In bacteria, promoter-like sequences arise spontaneously through point mutations, especially in IGRs (Mendoza-Vargas et al. 2009), and inefficient transcription from these promoters is the main source of pervasive transcripts i.e., RNAs originating from all across the genome (Dornenburg et al. 2010; Raghavan et al. 2012; Thomason et al. 2015). The functions, if any, of these genome-wide transcripts are not yet understood, but they could serve as the raw material for the emergence of new functional RNAs (Gottesman and Storz 2011; Wade and Grainger 2014; Lybecker et al. 2014). Pervasive transcription has been observed in all domains of life, and recently it was shown that new functional RNAs could evolve from such transcripts in humans (Ruiz-Orera et al. 2015). Our data also point towards such a scenario where the emergence of a promoter-like sequence resulted in the production of a transcript that evolved into EcsR2 by gaining regulatory motifs.

To be functional, an sRNA only requires a small seed sequence with partial complementarity to an mRNA; therefore, several such target mRNAs should occur in a
bacterial genome just through chance. Although a few nascent sRNA–mRNA interactions might have positive outcomes, most are likely deleterious, which could be mitigated by the low expression of incipient sRNAs, while new beneficial interactions could arise through rapid sRNA evolution. Similar to what we show in young enterobacterial sRNAs, low expression and rapid evolution have been observed in young miRNAs (Chen and Rajewsky 2007; Jovelin and Cutter 2014; Lyu et al. 2014), suggesting that this a universal process that facilitates the emergence of new non-coding regulatory RNAs in all domains of life.
C.6 REFERENCES


Horler RSP, Vanderpool CK. Homologs of the small RNA SGRS are broadly distributed in enteric bacteria but have diverged in size and sequence. Nucleic Acids Res. 2009;37:5465–5476.


APPENDIX D

Accumulation and expression of multiple antibiotic resistance genes in *Arcobacter cryaerophilus* that thrives in sewage

D.1 ABSTRACT

We explored the bacterial diversity of untreated sewage influent samples of a wastewater treatment plant in Tucson, AZ and discovered that *Arcobacter cryaerophilus*, an emerging human pathogen of animal origin, was the most dominant bacterium. The other highly prevalent bacteria were members of the phyla Bacteroidetes and Firmicutes, which are major constituents of human gut microbiome, indicating that bacteria of human and animal origin intermingle in sewage. By assembling a near-complete genome of *A. cryaerophilus*, we show that the bacterium has accumulated a large number of antibiotic resistance genes (ARGs) probably enabling it to thrive in the wastewater. We also determined that a majority of ARGs was being expressed in sewage, suggestive of trace levels of antibiotics or other stresses that could act as a selective force that amplifies multidrug resistant bacteria in municipal sewage. Because all bacteria are not eliminated even after several rounds of wastewater treatment, ARGs in sewage could affect public health due to their potential to contaminate environmental water.
D.2 INTRODUCTION

Non-coding Over the past few decades, based on numerous studies that examined the bacterial composition of wastewater during varying stages of treatment, there is growing evidence that sewage is an important hub for horizontal gene transfer (HGT) of antibiotic resistance genes (e.g., Baquero, Martínez & Cantón, 2008; Zhang, Shao & Ye, 2012; Rizzo et al., 2013; Pehrsson et al., 2016). Additionally, studies have shown that discharge of treated sewage allows these concentrated communities to spread into environmental water (Okoh et al., 2007; Varela & Manaia, 2013). The Arcobacter genus is commonly detected in sewage treatment plants around the world (Collado et al., 2008; Zhang, Shao & Ye, 2012; Varela & Manaia, 2013). This sparsely studied Epsilonproteobacteria is frequently associated with veterinary diseases, and is closely related to Campylobacter, and is considered an emerging human pathogen that causes enteritis and bacteremia (Kabeya et al., 2004; Morita et al., 2004; Collado et al., 2008). In addition, Arcobacter is known to be resistant to a wide array of commonly used antibiotics, with varying resistance profiles observed in different species (Houf et al., 2004; Abay et al., 2012; Rahimi, 2014), but the genes that enable antibiotic resistance are mostly unknown (Abdelbaqi et al., 2007; Miller et al., 2007).

In this study, we examined the bacterial diversity and the presence and expression of antibiotic resistance genes (ARGs) and virulence factors in untreated sewage. Our analyses revealed that an A. cryaerophilus strain that contained multiple ARGs was a major constituent of the sewage microbiome. In addition, we detected a large number of
expressed ARGs and virulence factors in *A. cryaerophilus* and in the rest of the sewage microbiome, which portends potential public health risk if bacteria carrying these genes contaminate public water resources.
D.3 MATERIALS AND METHODS

**Sewage sample collection, DNA and RNA extraction, and deep-sequencing**

Three untreated sewage influent samples (50 ml) were collected at the Roger Road Wastewater Reclamation Facility, Tucson, Arizona (March 2012) and immediately transferred on ice to the laboratory and stored at −80 °C until further use. The sewage samples were spun down (12,000 × G, 15 min, 4 °C) and the pellets were suspended in 1 ml of TRI reagent (Life Technologies). Total RNA from each sample was extracted from the aqueous phase and corresponding DNA was isolated from the interphase using a protocol provided by the manufacturer. RNA samples were treated with TurboDNAse (Life Technologies) to remove contaminating DNA, and PCR reactions using 16S rDNA primers were performed to confirm complete DNA removal. Furthermore, RNA samples were depleted of ribosomal RNA using RiboZero Bacteria and RiboZero Human kits (Illumina). Around 100 ng of RNA from each sample was used to prepare directional mRNA-seq libraries using the Illumina Small RNA Sample Preparation Kit and Directional mRNA-seq Sample Preparation protocol provided by Illumina Inc. DNA samples were further purified using DNeasy kit (Qiagen) and around 5 μg of DNA from each sample was used to prepare paired-end DNA-seq libraries using the Paired-End sample Preparation Kit (Illumina). All RNA libraries were pooled into a single lane, and all DNA samples were pooled into another lane of Illumina HiSeq 2000 and were sequenced at the Yale Center for Genome Analysis (RNA-seq: single-end, 75 cycles; DNA-seq: paired end, 2 × 75 cycles) using standard adapters. All DNA and RNA reads
have been deposited at NCBI (BioProject PRJNA354077).

**Taxonomic classification**

DNA reads were cleaned by removing adapters and were filtered by quality (≥Q20) and length (≥50 bp) using Trimmomatic v0.32 (Bolger, Lohse & Usadel, 2014). Each library was assembled into contigs using 17,000,000 reads and IDBA-UD (Peng et al., 2012), and normalized to the smallest library size (6,000 contigs) using Seqtk (https://github.com/lh3/seqtk) (Perner et al., 2014) (Table S1, found in Additional_file_A14.pdf). All contigs were searched against the NCBI nt database using BLAST and analyzed in MEGAN (Huson et al., 2007), requiring at least 70% of the query sequence to align with the subject sequence with ≥70% identity to be assigned to a given phylum. Contigs classified at the phylum level (48%, 57%, and 54% of contigs from the three samples, respectively) were used to determine their detailed taxonomic positions. Remaining contigs either did not have significant BLAST hits or mapped to unidentified environmental samples; however, all contigs were used in downstream analyses (detection of antibiotic resistance genes, virulence factors etc.).

**Identification of antibiotic resistance genes, virulence factors, transposases, and bacteriophage genes**

Contigs were run through MetaProdigal (Hyatt et al., 2012) to identify encoded ORFs, which were annotated by mapping to antibiotic resistance genes, virulence factors, bacterial transposases, and prophages obtained respectively from CARD, PATRIC,
InterPro, and PHAST databases using PHMMER, with an $E$-value of at least $1 \times 10^{-10}$ as the cutoff (Finn, Clements & Eddy, 2011; Zhou et al., 2011; McArthur et al., 2013; Wattam et al., 2014; Mitchell et al., 2015). RNA reads from each sample were filtered by quality ($\geq Q20$) and length ($\geq 50$ bp) using Trimmomatic v0.32 and normalized using Seqtk to 50,000,000 reads. They were mapped to annotated ORFs using CLC Genomic Workbench v6.5. A strict mapping criterion (at least 95% of each read should map with at least 95% identity to the mapped region) was used in order to minimize non-specific mapping. Genes were filtered and considered expressed based on at least 10 reads mapping to each ORF. Statistical analysis was conducted using SAS Studio v3.4 (SAS Institute, Cary NC).

**Arcobacter genome assembly**

DNA reads from all three samples were pooled to gain enough coverage depth, and were assembled into contigs using IDBA-UD (Peng et al., 2012). All contigs were searched against the NCBI nt database using BLASTN and analyzed in MEGAN (Huson et al., 2007). All *Arcobacter* gene sequences were downloaded from NCBI, and using PHMMER, *Arcobacter* contigs present in our data were identified with at least $1 \times 10^{-10}$ $E$-value as the cutoff. These contigs were extracted and run through the differential coverage binning procedure for metagenomic data, as described previously (Albertsen et al., 2013). In brief, contigs were binned based on coverage, tetranucleotide frequency, GC%, and length, then examined for presence of essential single copy genes. Phylogenetic analyses were conducted on nucleotide sequences using several
housekeeping genes to identify the bins containing *A. cryaerophilus* (marked in blue in Fig. S1, found in Additional_file_A14.pdf). One genome bin with $\sim 200 \times$ coverage that contained all *A. cryaerophilus* housekeeping genes was selected for secondary refinement and finishing (top right cluster in Fig. S2, found in Additional_file_A14.pdf). This cluster of contigs was isolated and all original trimmed DNA reads were mapped against them using Bowtie2 v2.1.0 (Langmead & Salzberg, 2012). All mapped reads were reassembled into contigs using IDBA-UD. These contigs were combined with all original trimmed DNA reads for scaffold extension using SSPACE (Boetzer et al., 2011) into a final scaffold of $\sim 1.8$ Mb over 456 contigs. To check for completeness of the assembled *A. cryaerophilus* genome, we used a single-copy gene database (Albertsen et al., 2013), and as a control we performed the same analysis with the *A. butzleri* (CP000361.1) genome. Visual representation of draft genome was created using Circos (Krzywinski et al., 2009). The draft genome has been deposited at NCBI under the accession LNTC0000000.

**Detection of HGT**

Horizontally acquired genes were identified using HGTector (Zhu, Kosoy & Dittmar, 2014). *Arcobacter* was set as self-group, and Campylobacterales was set as exclusion group. This method captured HGT events where only *Arcobacter* has acquired a particular gene from outside of Campylobacterales and ignored any events where the genes could also have been transferred elsewhere within the order. This conservative approach was used due to the dearth of annotated genomes within Campylobacterales.
BLASTN parameter thresholds were set at 70% identity and an $E$-value of at least $1e^{-5}$. Several putative HGT genes were examined using phylogenetic analysis to validate the HGTecor data (Fig. S3, found in Additional_file_A14.pdf).

**Phylogenetic analysis**

Nucleotide sequence alignment for all trees was performed using Clustal Omega (Sievers et al., 2011), and ambiguously aligned regions were removed using Gblocks (Talavera & Castresana, 2007). The evolution model GTR+I+G (General Time Reversible plus Invariant sites plus Gamma distribution) used for all trees was selected using jModelTest2 (Darriba et al., 2012). Bayesian trees were constructed using MrBayes as implemented in Geneious (Huelsenbeck & Ronquist, 2001; Kearse et al., 2012). A chain length of 1,000,000 was used with a burn-in fraction of 25% and sampling every 100 trees. Maximum Likelihood trees were constructed using RAxML (Stamatakis, Hoover & Rougemont, 2008) as implemented in Geneious with 1,000 bootstrap replicates to confirm Bayesian topologies. *Helicobacter pylori* (AJ558222.1) was used to root all phylogenetic trees.
D.4 RESULTS AND DISCUSSION

*A. cryaerophilus* thrives in sewage

For the three sewage samples, taxonomic labels were assigned to at least the phylum level for all contigs with significant BLAST hits. There was no significant difference in bacterial distribution between the three samples (Figure D.1); hence, average values are presented hereafter. Members of the phylum Proteobacteria (67% of total hits) was the most prevalent bacteria, followed by Bacteroidetes (23%) and Firmicutes (9%). A more comprehensive study that examined several sewage samples from across the USA observed a similar pattern of bacterial phyla abundance (Shanks et al., 2013). However, at the genus level, *Arcobacter* (an Epsilonproteobacteria) was the most dominant bacterium in our study, making up 39% of all annotated contigs, unlike members of Gammaproteobacteria (38% of all pyrotags) in the previous study (Shanks et al., 2013).

*Arcobacter* is commonly associated with both humans and farm animals (Collado et al., 2008), the latter perhaps more relevant to this specific wastewater treatment plant because agriculture accounts for the largest use of water at around 70% of all water demand within the state of Arizona (ADWR, 2009). In addition, in Tucson, the wastewater treatment plant served both agricultural and municipal areas (PAG, 2006). Members of the phyla Bacteroidetes and Firmicutes, two of the most abundant bacteria in human gut, were also abundant in the sewage samples (Cho & Blaser, 2012; Jandhyala et al., 2015). Taken together, these data highlight the important role that sewage systems play as an arena where bacteria of human and animal origin interact, which could
Figure D.1 – Bacterial composition of sewage samples. Percentage of sewage contigs that were assigned to each bacterial phylum is shown. Data represent means of three samples ± standard deviations.
promote the exchange of genes between the two groups (Baquero, Martínez & Cantón, 2008; Gaze et al., 2013; Rizzo et al., 2013). After conducting genomic binning of the *Arcobacter* contigs utilizing various factors such as coverage depth, GC%, and tetranucleotide frequency (Albertsen et al., 2013), we were able to identify ~80% of these contigs as belonging to *A. cryaerophilus*, an emerging human pathogen that is commonly associated with diseases such as bovine reproductive disorders, diarrhea and hemorrhagic colitis in cattle and sheep (Schroeder-Tucker et al., 1996; Ho, Lipman & Gaastra, 2006).

**Presence and expression of multiple ARGs in *A. cryaerophilus***

To better characterize *A. cryaerophilus*, we assembled a near-complete genome from the DNA-seq reads (Figure D.2). Based on the presence of 100 out of 106 single copy genes (Albertsen et al., 2013) with zero redundant copies, we estimate that the *A. cryaerophilus* genome is ~95% complete and contains 2,419 ORFs (including partial genes at the ends of contigs) (Table S2, found in Additional_file_A14.pdf). Among these ORFs, 115 (5% of ORFs) encode antibiotic resistance genes (ARGs) belonging to 25 categories as defined by the CARD database (Table S3, found in Additional_file_A14.pdf) (McArthur et al., 2013). Macrolide resistance made up the majority of annotated ARGs (26, 23%), (Table S3, found in Additional_file_A14.pdf), with fluoroquinolones (18, 16%), aminocoumarin (17, 15%) and vancomycin (13, 11%) resistance genes being the next largest groups. Because gene expression is a good representation for functional gene activity, we analyzed the expression of *A. cryaerophilus* genes using RNA-seq and discovered that all 115 putative ARGs genes were being expressed (Figure D.3A;
Figure D.2 – Draft genome of *A. cryaerophilus*. Two outer rings show ORFs (purple) on forward and reverse strands, respectively. Black blocks represent horizontally acquired genes. Each tick mark represents 10,000 bp. Middle two rings show positions of features annotated in the center. Inner blue and grey rings show DNA-seq coverage (mean of three samples) and RNA-seq transcription levels (mean of three samples), respectively. Note that 456 original contigs were randomly assigned to 14 equal fragments for easy visualization.
Figure D.3 – Abundance of antibiotic resistance genes. Number of annotated (black) and expressed (white) antibiotic resistance categories in (A) *A. cryaerophilus* and (B) total sewage sample is shown. Data represent means of three samples ± standard deviations.
Additional_file_A15.csv). In comparison, *Helicobacter pylori*, a closely related Epsilonproteobacteria contain 59 ARGs (4% of genes); however, in both bacteria around 50% of ARGs consisted of efflux pumps (Paulsen, Sliwinski & Saier, 1998). It should be noted that although all ARGs were found to be expressed in *A. cryaerophilus*, the median level of expression of single copy genes (3,887 reads mapped) was found to be 10× higher than the median level of expression of ARGs (363 reads mapped), probably because higher expression of many ARGs requires strong induction.

To determine the prevalence of ARGs in the total sewage, we extended our analysis to all contigs assembled in our study. Out of the 60,723 ORFs encoded in the sewage contigs, 2,606 ORFs matched 42 antibiotic resistance categories (Table S4, found in Additional_file_A14.pdf). Using RNA-seq, we determined that 2,106 (81%) of these ORFs were expressed (Figure D.3B). Of the 2,106 putative antibiotic resistance genes expressed in the sewage samples, macrolide resistance genes made up the largest portion (538, 26%) (Table S4, found in Additional_file_A14.pdf). The next two largest groups were fluoroquinolone resistance (378, 18%) and tetracycline resistance (339, 16%) genes. The expression of antibiotic resistance genes could be due to the presence of numerous antibiotics in urban wastewater (Heberer, 2002; Rizzo et al., 2013), which could select for multidrug resistant bacteria, thereby aggravating an already dire situation (Baquero, Martínez & Cantón, 2008; Rizzo et al., 2013; Wellington et al., 2013; Amos et al., 2014)). It is also possible that ARGs were being expressed constitutively or in response to stress (Poole, 2012). Additionally, previous studies have shown that ARGs are expressed in a wide variety of environments even in the absence of known anthropogenic
antibiotic pressure (Udikovic-Kolic et al., 2014; Versluis et al., 2015; Noyes et al., 2016); hence, further study is required to determine the stimuli for the observed ARG expression. Interestingly, although the total sewage contigs contained 23× more ARGs than in *A. cryaerophilus* contigs (2,606 vs. 115), 44% of DNA reads mapped to *A. cryaerophilus* ARGs, indicating that while the sewage contained high diversity of ARGs, most non-*A. cryaerophilus* ARGs were of low abundance.

**Signatures of HGT in *A. cryaerophilus* genome**

We compared our draft genome of *A. cryaerophilus* to the published genome of *A. butzleri* (CP000361.1), a closely related human and animal pathogen that has been studied much more extensively than *A. cryaerophilus* (Vandenberg et al., 2004; Miller et al., 2007; Collado et al., 2008) (Fig. S4, found in Additional_file_A14.pdf). As observed previously in other members of this genus (Karadas et al., 2013; Merga et al., 2013), the two *Arcobacter* species only shared 1,337 genes (∼50%) (Table D.1, Additional_file_A16.csv). A comparison of the two genomes was also conducted using RAST (Overbeek et al., 2014), which showed that merely 846 genes with known functions were shared between *A. butzleri* and *A. cryaerophilus* (Additional_file_A17.csv). This sizable variation in gene content between the two species indicates that HGT could have played a prominent role in shaping the genomes of *Arcobacter* species. Concomitantly, even after using a very conservative threshold, we detected 209 (9%) and 228 (10%) horizontally acquired genes in *A. cryaerophilus* and *A. butzleri*, respectively (Table D.1). While similar in scale, only 73 HGT-origin genes were
Table D.1 - Comparison of *Arcobacter* genomes.

<table>
<thead>
<tr>
<th>Features</th>
<th><em>A. cryaerophilus</em></th>
<th><em>A. butzleri</em></th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total ORFs(^a)</td>
<td>2,419</td>
<td>2,259</td>
<td>1,337</td>
</tr>
<tr>
<td>Horizontally Acquired ORFs</td>
<td>209</td>
<td>228</td>
<td>73</td>
</tr>
<tr>
<td>Antibiotic Resistance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Categories</td>
<td>25</td>
<td>29</td>
<td>23</td>
</tr>
<tr>
<td>Genes</td>
<td>115</td>
<td>140</td>
<td>54</td>
</tr>
<tr>
<td>Virulence Factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Categories</td>
<td>24</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>Genes</td>
<td>232</td>
<td>185</td>
<td>92</td>
</tr>
<tr>
<td>Transposases</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Categories</td>
<td>7</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Genes</td>
<td>61</td>
<td>57</td>
<td>15</td>
</tr>
<tr>
<td>Prophages</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genes</td>
<td>290</td>
<td>320</td>
<td>173</td>
</tr>
<tr>
<td>GenBank Accession</td>
<td>LNTC000000000</td>
<td>CP000361.1</td>
<td></td>
</tr>
</tbody>
</table>
shared between the two genomes, indicating that parallel HGT events have molded the
genomes of the two *Arcobacter* species.

HGT is known to promote ARG dissemination between bacteria (Hawkey &
Jones, 2009; Gaze et al., 2013; Pehrsson et al., 2016); hence, we compared ARGs present
in *A. cryaerophilus* to those present in *A. butzleri* in order to identify those that are of
possible HGT origin. We identified 140 putative genes belonging to 29 antibiotic
resistance categories in *A. butzleri*, and out of the 25 antibiotic resistance categories
present in *A. cryaerophilus*, 23 were present in *A. butzleri*, with two categories
(Glycylecycline and Roxithromycin resistance) found only in *A. cryaerophilus*, and six
categories (Bicyclomycin, Elfamycin, Isoniazid, Kanamycin, Streptomycin, and
Teicoplanin resistance) exclusive to *A. butzleri* (*Table D.1*). However, within each
category large differences in gene content was observed between the two bacteria, with
only 54 genes shared between *A. cryaerophilus* and *A. butzleri*. These data show that
even though the antibiotic resistance capabilities of both bacteria overlap, their respective
gene repertoires were largely assembled through independent HGT events. Transposons
and bacteriophages are important agents of HGT in bacteria, and we found several
transposases and bacteriophage ORFs in *A. cryaerophilus* (61 transposase ORFs, 290
phage ORFs) and *A. butzleri* (57, 320) (*Figure D.2* and *Table D.1*, Table S2, found in
*Additional_file_A14.pdf*). Additionally, we discovered that several of the ARGs in *A.
cryaerophilus* were located in close proximity to transposases or bacteriophage genes
(*Figure D.4*), suggestive of a role for these mobile genetic elements in the accumulation
of ARGs in this pathogen.
Figure D.4 – Location of ARGs indicates horizontal acquisition. Several antibiotic resistance genes (white) in *A. cryaerophilus* that are flanked by prophage genes (black) and transposases (grey) are shown. Nucleotide positions within each contig are also provided.
Presence and expression of virulence factors in sewage

In addition to ARGs, another class of genes in *A. cryaerophilus* that could potentially impact human health is virulence factors. Most of the previous work at the molecular level has focused on nine putative virulence genes first described in *Arcobacter butzleri* strain RM4018. The presence of these nine virulence genes in *Arcobacter* genomes is highly variable and are all rarely found together in the same genome (Miller et al., 2007; Douidah et al., 2012). In general, the ability to adhere to and invade cells varies widely between *Arcobacter* species, with some of the most invasive strains isolated from feces or sewage samples (Ho et al., 2007; Karadas et al., 2013; Levican et al., 2013). Using the PATRIC database we identified 232 putative virulence genes (24 virulence categories) in *A. cryaerophilus* (Table S5, found in Additional_file_A14.pdf), out of which 231 were expressed (Additional_file_A18.csv). In PATRIC, virulence factors are assigned the category “virulence” if their mode of action is not specified in an associated study. Among the expressed virulence factor genes, 101 were annotated with a category other than “virulence.” Of these, “intracellular survival and replication” was the largest group (30, 30%) (Table S5, found in Additional_file_A14.pdf). The next largest groups present were “cellular metabolism” (22, 22%), “adhesion” (18, 18%), and “invasion” (11, 11%) (Figure D.5A). In the total sewage contigs, we identified 4,440 putative virulence factor genes (38 virulence categories (Table S6, found in Additional_file_A14.pdf), out of which, 3,776 were expressed (Additional_file_A18.csv)). Excluding the “virulence” category, 1,812 genes belonging to 37 other virulence categories were identified in the sewage microbiome. Of these, 1,589 genes from 35 categories were expressed, with
Figure D.5 – Abundance of virulence factors. Number of annotated (black) and expressed (white) virulence factor categories in (A) *A. cryaerophilus* and (B) total sewage sample is shown. Data represent means of three samples ± standard deviations.
“intracellular survival and replication” (548, 35%), “invasion” (318, 20%) and “adhesion” (229, 14%) being the top three categories (Figure D.5B; Table S6, found in Additional_file_A14.pdf).

Our data suggest that untreated sewage contains several genes that potentially promote bacterial antibiotic resistance and virulence, and *A. cryaerophilus*, a potential human pathogen that contains multiple drug resistance and virulence factors, is a major component of this sewage system. Because we analyzed only a limited number of samples, further study is required to determine whether the dominance of *A. cryaerophilus* was a short-term phenomenon or whether this bacterium is a long-term resident of this sewage system (McLellan et al., 2010; Shanks et al., 2013). Although its cause is not understood, as observed in our study, *Arcobacter* has been shown to be highly prevalent in other sewage systems (Fisher et al., 2014). A possible explanation is the formation of biofilm on pipe surfaces and in deposited sediments along the sewer system (Chen, Leung & Hung, 2003), another possibility is that the presence of multiple antibiotics, heavy metals or xenobiotics in wastewater, even at very low concentrations is selecting for *A. cryaerophilus*, which contains multiple ARGs (Heberer, 2002; Hawkey & Jones, 2009; Gullberg et al., 2014; Jutkina et al., 2016). Similar to our observation, selection for antibiotic resistant bacteria has been described from other wastewater treatment plants (Goñi Urriza et al., 2000; Czekalski et al., 2012; Mao et al., 2015); consequently, constant monitoring of both pre- and post-treatment sewage is warranted because of the risk of reintroducing bacteria replete with ARGs and virulence factors into
natural environments (Fahrenfeld et al., 2013; Czekalski, Gasco & Burgmann, 2014; Mao et al., 2015; Pehrsson et al., 2016).
D.5 REFERENCES


Pima Association of Governments (PAG) Existing wastewater treatment facilities and other point source NPDES discharges. 2006. https://www.pagnet.org/documents/Water/PC208/Ch5_Apr06.pdf


APPENDIX E

Whole-genome sequence of *Coxiella burnetii* Nine Mile RSA 439 (phase II, clone 4), a laboratory workhorse strain

E.1 ABSTRACT

Here, we report the whole-genome sequence of *Coxiella burnetii* Nine Mile RSA439 (phase II, clone 4), a laboratory strain used extensively to investigate the biology of this intracellular bacterial pathogen. The genome consists of a 1.97-Mb chromosome and a 37.32-kb plasmid.
E.2 GENOME ANNOUNCEMENT

Coxiella burnetii is a Gram-negative intracellular bacterium that causes an influenza-like illness in humans called Q fever (Eldin et al. 2017). Most infections occur through inhalation of aerosols originating from domestic livestock operations. Within the host cell, the pathogen becomes metabolically activated upon delivery into an acidic lysosome-like vacuole (Voth and Heizen 2007). The only C. burnetii virulence factor established in an immunocompetent animal model of infection is full-length lipopolysaccharide (LPS), which is synthesized by virulent phase I bacteria (Moos and Hackstadt 1987). Upon serial in vitro passage, phase I bacteria convert to avirulent phase II bacteria, which produce truncated LPS lacking O antigen and several core sugars (Amano and Williams 1984; Hackstadt 1986; Fiset 1957; Hackstadt 1990; Kersh et al. 2011).

The Nine Mile RSA439 (phase II, clone 4) strain (NMII) was derived from the Nine Mile strain, which was originally isolated in 1935 from the tick Dermacentor andersoni in Montana (Davis and Cox 1938). The Nine Mile strain was passaged 94 times in embryonated hen’s eggs and then plaque purified to generate NMII (Wike et al. 1972; Williams et al. 1981). NMII has an ~26-kb chromosomal deletion that eliminates several LPS biosynthetic genes and is associated with the production of a severely truncated LPS (Hackstadt et al. 1985; Vodkin and Williams 1986; Hoover et al. 2002; Beare et al. 2006; Denison et al. 2007). Because of clonality, avirulence in a guinea pig model of infection, and lack of phase reversion, NMII is considered a biosafety level 2
(BSL-2) bacterium (Moos and Hackstadt 1987; Hackstadt 1996; Samuel and Hendrix 2009). Other *C. burnetii* strains are considered BSL-3 bacteria and are regulated as select agents by the U.S. Centers for Disease Control and Prevention (Samuel and Hendrix 2009).

The NMII genome has not yet been sequenced; consequently, most researchers use the published genome of the Nine Mile RSA493 phase I strain (NMI) for reference (Seshadri et al. 2003). This occasionally leads to inconclusive results; for instance, the gene *caeA* is not annotated as a functional protein-coding gene in the NMI genome, but transcriptome analysis of NMII indicates its presence (Bisle et al. 2016; Raghavan 2016). Thus, a fully annotated genome of the widely used NMII laboratory strain is needed to better understand the unique biology of this intracellular pathogen.

NMII was grown in ACCM-2 at 37°C in a 2.5% O₂/5% CO₂ environment in a tri-gas incubator (New Brunswick Scientific, NJ), as described previously (Omsland et al. 2011). DNA was isolated from a 500-ml 7-day culture using phenol-chloroform with gentle cell disruption using a vortex adapter (Qiagen, CA) in order to minimize DNA fragmentation. DNA was sequenced using the PacBio RS II platform (Pacific Biosciences, USA), which generated a library containing 86,731 reads with an average length of 7,565 bp. Reads were assembled using HGAP 2.3.0 (Chin et al. 2013), which returned three contigs. The two chromosome contigs and the plasmid contig were closed in SSPACE 2.0 (Boetzer et al. 2011) using trimmed Illumina MiSeq 75-bp paired-end reads (5.50 million). Finally, CLC Genomics Workbench 6.5 (Qiagen) was used to map all Illumina and PacBio reads to the NMII chromosome (410× and 290× coverage,
respectively) and plasmid (240× and 30× coverage, respectively) scaffolds to generate the consensus genome sequence. As expected, relative to the genome of NMI, homologs of genes CBU_0679 to CBU_0697 were completely deleted from the NMII genome (Moos and Hackstadt 1987; Vodkin and Williams 1986; Hoover et al. 2002; Beare et al. 2006; Denison et al. 2007). In addition, partial deletions of homologs of CBU_0678, CBU_0698, and CBU_0918 and several single nucleotide polymorphisms were observed.

**Accession numbers**

The complete genome sequence of *C. burnetii* Nine Mile RSA439 (phase II, clone 4) has been deposited in GenBank under the accession numbers CP020616 (for chromosome) and CP020617 (for plasmid).
E.3 REFERENCES


