UNIVERSITY OF CALIFORNIA
SANTA CRUZ

EVOLUTION OF STRUCTURE AND FUNCTION OF KINK-TURN CONTAINING RNAs IN THE DOMAIN ARCHAEA

A dissertation submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Lauren Michelle Lui

September 2015

The Dissertation of Lauren Michelle Lui is approved:

__________________________________________________________________________________
Professor Todd Lowe, Chair

__________________________________________________________________________________
Professor Manuel Ares, Jr

__________________________________________________________________________________
Professor Jeremy Sanford

__________________________________________________________________________________
Professor Kevin Karplus

__________________________________________________________________________________
Tyrus Miller

Vice Provost and Dean of Graduate Studies
# Table of Contents

List of Figures vii  
List of Tables x  
Abstract xi  
Dedication xiii  
Acknowledgments xiv  
1 Introduction 1  
2 Background and Significance 5  
  2.1 The Archaea 5  
  2.1.1 Archaea: the third domain of life 5  
  2.1.2 Phylogeny and traits of the Archaea 8  
  2.2 The RNA Kink-Turn Motif 12  
  2.2.1 Structure and Characteristics of the RNA Kink-Turn Motif 12  
  2.2.2 Role of the kink-turn in human diseases 17  
  2.3 Computational prediction of RNA structure 19  
  2.3.1 Tools that Take User-Defined Motifs and Patterns 20  
  2.3.2 Tools That Use 3D Structural Data 20  
  2.3.3 Stochastic-Context Free Grammars (SCFGs) and Covariance Models 21  
3 Structure, function, and evolution of archaeal C/D box sRNAs as revealed by a comprehensive set from the *Pyrobaculum genus* 23  
  3.1 Abstract 23  
  3.2 Collaboration in this study 24  
  3.3 Introduction and Background 25  
  3.4 Materials and Methods 30  
  3.4.1 Computational prediction of C/D box sRNAs 30
## 5 Non-canonical H/ACA box sRNAs are Widespread Across the Archaea

### 5.1 Abstract

### 5.2 Introduction and Background

#### 5.2.1 Features of Pseudouridine and H/ACA box sRNAs

#### 5.2.2 Current state of H/ACA RNA annotations in the Domain Archaea

#### 5.2.3 Other computational prediction methods for archaeal H/ACA box sRNAs

#### 5.2.4 Collaboration in this study

### 5.3 Development of covariance models

#### 5.3.1 Four core models of H/ACA box RNAs in the Archaea

#### 5.3.2 Models and parameters for Infernal search program

#### 5.3.3 Iterative development of covariance models

### 5.4 Results

#### 5.4.1 Non-canonical forms are widespread across the archaeal domain

#### 5.4.2 Conservation of H/ACA box sRNAs can extend through an entire order

#### 5.4.3 Model idiosyncrasies, versions and forms

#### 5.4.4 Guide-extraction

### 5.5 Discussion

### 5.6 Materials and Methods

#### 5.6.1 Small RNA sequencing data

#### 5.6.2 Covariance models

#### 5.6.3 Homolog detection

#### 5.6.4 Guide Extraction

#### 5.6.5 Target Prediction

## 6 Survey of L7Ae-associated RNAs in *Thermococcus kodakaraensis*

### 6.1 Abstract

### 6.2 Introduction and Background

#### 6.2.1 RNA kink-turn structure and protein partners

#### 6.2.2 Evolutionary perspective of kink-turn-containing RNAs and L7Ae

#### 6.2.3 Example of a highly conserved kink-turn-protein interaction: L30e binds to its mRNA and the ribosome

#### 6.2.4 Synthetic biology applications of kink-turns

#### 6.2.5 Previous study of L7Ae-Associated sRNAs in archaea

#### 6.2.6 *Pyrobaculum* Ribosomal Protein L14e

#### 6.2.7 Summary of this chapter

### 6.3 Methods and Materials

#### 6.3.1 *T. kodakaraensis* strains

#### 6.3.2 Media and growth conditions
6.3.3 Co-immunoprecipitation of kink-turn-containing RNAs using HA antibodies .......................................................... 162
6.3.4 RNA preparation and fragmentation .......................................................... 163
6.3.5 Preparation for sequencing .......................................................... 163
6.3.6 Processing and analysis of sequencing data .......................................................... 164
6.4 Results ........................................................................................................ 164
6.4.1 Description of samples and analysis of enrichment .......................................................... 164
6.4.2 Enrichment of known non-coding RNAs .......................................................... 166
6.4.3 L7Ae mRNA (TK1311) ........................................................................... 168
6.4.4 Enriched mRNA fragments ........................................................................... 168
6.4.5 Enriched antisense small RNAs ........................................................................... 172
6.5 Discussion ....................................................................................................... 173

7 Conclusion ......................................................................................................... 174

A Description of supplementary files ........................................................................ 178
A.1 Pyrobaculum supplementary files ........................................................................ 178
  A.1.1 Structural alignment for Pyrobaculum covariance model ......................................... 178
  A.1.2 Target finding program .................................................................................. 179
  A.1.3 C/D box sRNA annotations of Pyrobaculum and Pyrococcus ........................................ 179
  A.1.4 Predicted methylation targets ........................................................................ 180
  A.1.5 Pyrobaculum Ribosomal RNA alignments .......................................................... 180
  A.1.6 Overlap of sRNAs with protein-coding genes ...................................................... 181
A.2 Supplementary material for Panarchaeal C/D box sRNA study .................................. 181
  A.2.1 C/D box sRNA Annotations ......................................................................... 181
  A.2.2 K-turn Analysis .......................................................................................... 182
  A.2.3 Ribosomal RNA ........................................................................................ 182
  A.2.4 Alignment of L7Ae protein sequence across diverse archaea ................................... 182
A.3 Supplementary material for H/ACA box sRNA study .................................................. 183

Bibliography ........................................................................................................ 185
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Proposed phylogenetic tree of the three domains using universally conserved</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>genes</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>Unrooted Bayesian tree of the Archaea based on 57 ribosomal proteins.</td>
<td>10</td>
</tr>
<tr>
<td>2.3</td>
<td>Two-dimensional Representations of the Kink Turn</td>
<td>12</td>
</tr>
<tr>
<td>2.4</td>
<td>Three-dimensional Structure of the Kink Turn with L7Ae</td>
<td>14</td>
</tr>
<tr>
<td>2.5</td>
<td>Illustration of the stabilizing trans Hoogsteen/sugar edge interaction</td>
<td>15</td>
</tr>
<tr>
<td>2.6</td>
<td>Variants of the kink-turn</td>
<td>16</td>
</tr>
<tr>
<td>2.7</td>
<td>Example of an unusual kink-turn, KT-11</td>
<td>16</td>
</tr>
<tr>
<td>2.8</td>
<td>The RNase MRP kink-turn</td>
<td>18</td>
</tr>
<tr>
<td>3.1</td>
<td>Conserved features of archaeal C/D box sRNAs</td>
<td>27</td>
</tr>
<tr>
<td>3.2</td>
<td>Tree of <em>Pyrobaculum</em> species</td>
<td>29</td>
</tr>
<tr>
<td>3.3</td>
<td>Most <em>Pyrobaculum</em> C/D box RNAs are highly conserved</td>
<td>35</td>
</tr>
<tr>
<td>3.4</td>
<td>C/D box RNA secondary structural alignment used to train the covariance</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>model</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>Sensitivity and Specificity of <em>Pyrobaculum</em> C/D box sRNA covariance</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>model</td>
<td></td>
</tr>
<tr>
<td>3.6</td>
<td>Comparison of <em>Pyrobaculum</em> and <em>Pyrococcus</em> C/D box sRNA K-turns and</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>K-loops</td>
<td></td>
</tr>
<tr>
<td>3.7</td>
<td>Nucleotide frequencies of the position 5' of the D and D' boxes (position</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>5n)</td>
<td></td>
</tr>
<tr>
<td>3.8</td>
<td>Features of <em>Pyrobaculum</em> C/D box sRNAs guides</td>
<td>48</td>
</tr>
<tr>
<td>3.9</td>
<td>Example of a <em>Pyrobaculum</em> sRNA family where one member has a mismatch</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>base pair at the “N+five” position</td>
<td></td>
</tr>
<tr>
<td>3.10</td>
<td><em>Pyrobaculum</em> sRNA family that has a mismatch base pair at the “N+five”</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>position</td>
<td></td>
</tr>
<tr>
<td>3.11</td>
<td>Mapping of the predicted sites of sRNA directed methyl modification</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>onto the conserved secondary structure of archaeal 16S rRNA</td>
<td></td>
</tr>
<tr>
<td>3.12</td>
<td>Mapping of the predicted sites of sRNA directed methyl modification</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>onto the conserved secondary structure of archaeal 23S rRNA</td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>3.13</td>
<td><em>Pyrobaculum</em> sRNAs that exhibit alternative long range guide-target interactions with rRNA</td>
<td>58</td>
</tr>
<tr>
<td>3.14</td>
<td>Examples of C/D box sRNA duplication within a species</td>
<td>63</td>
</tr>
<tr>
<td>3.15</td>
<td>Proliferation and evolutionary interconnections between sRNA genes</td>
<td>65</td>
</tr>
<tr>
<td>3.16</td>
<td>MITE-like element in the <em>Pca</em> genome</td>
<td>67</td>
</tr>
<tr>
<td>3.17</td>
<td>Polycistronic <em>Pyrobaculum</em> C/D box sRNAs</td>
<td>70</td>
</tr>
<tr>
<td>3.18</td>
<td>Evidence of sRNAs co-transcribed with tRNAs</td>
<td>71</td>
</tr>
<tr>
<td>3.19</td>
<td>The overlap between sRNA genes and protein-coding genes</td>
<td>73</td>
</tr>
<tr>
<td>3.20</td>
<td>The distance between tRNAs and sRNAs with protein-coding genes</td>
<td>76</td>
</tr>
<tr>
<td>3.21</td>
<td>Boxplots of distances for C/D box sRNAs, tRNAs, and mRNAs to the closest or overlapping mRNA</td>
<td>77</td>
</tr>
<tr>
<td>3.22</td>
<td>Boxplots of distances for C/D box sRNAs, tRNAs, and mRNAs to the closest or overlapping mRNA divided by operon membership</td>
<td>80</td>
</tr>
<tr>
<td>4.1</td>
<td>Published annotations of archaeal C/D box sRNAs mapped onto a phylogenetic tree</td>
<td>88</td>
</tr>
<tr>
<td>4.2</td>
<td>Summary of C/D box sRNA Annotations</td>
<td>94</td>
</tr>
<tr>
<td>4.3</td>
<td>Conserved C/D box sRNA in the Thaumarchaeota</td>
<td>95</td>
</tr>
<tr>
<td>4.4</td>
<td>Mapping of predicted 2′-O-methylation for seven diverse archaeal species onto 23S rRNA</td>
<td>96</td>
</tr>
<tr>
<td>4.5</td>
<td>C/D box sRNA candidates in the Lokiarchaeota</td>
<td>97</td>
</tr>
<tr>
<td>4.6</td>
<td>Conserved C/D box sRNAs in the methanogens</td>
<td>100</td>
</tr>
<tr>
<td>4.7</td>
<td>Halophile C/D box sRNAs</td>
<td>104</td>
</tr>
<tr>
<td>4.8</td>
<td>Variability of K-motifs of C/D box sRNAs between the <em>Pyrobaculum</em> and those found in the rest of the Archaea</td>
<td>107</td>
</tr>
<tr>
<td>4.9</td>
<td>Portion of alignment of L7Ae from diverse archaea</td>
<td>109</td>
</tr>
<tr>
<td>4.10</td>
<td>Comparison of <em>Pyrobaculum</em> K-motifs of C/D box sRNAs with those found in rest of the Archaea split by number of C/D box sRNAs</td>
<td>110</td>
</tr>
<tr>
<td>4.11</td>
<td>C/D box sRNAs with K-turns instead of K-loops</td>
<td>113</td>
</tr>
<tr>
<td>4.12</td>
<td>Models of C/D box sRNP formation</td>
<td>115</td>
</tr>
<tr>
<td>5.1</td>
<td>Diagram of uridine and pseudouridine</td>
<td>122</td>
</tr>
<tr>
<td>5.2</td>
<td>Diagrams of canonical and non-canonical forms of archaeal H/ACA box sRNAs</td>
<td>123</td>
</tr>
<tr>
<td>5.3</td>
<td>Previously annotated archaeal H/ACA box sRNAs mapped onto a phylogenetic tree</td>
<td>126</td>
</tr>
<tr>
<td>5.4</td>
<td>Core models of canonical and non-canonical forms of archaeal H/ACA box sRNAs</td>
<td>131</td>
</tr>
<tr>
<td>5.5</td>
<td>Summary of archaeal H/ACA box sRNA annotations</td>
<td>135</td>
</tr>
<tr>
<td>5.6</td>
<td>Lokiarchaeota H/ACA box sRNA</td>
<td>138</td>
</tr>
<tr>
<td>5.7</td>
<td>Example of conserved non-canonical H/ACA box sRNA</td>
<td>140</td>
</tr>
<tr>
<td>5.8</td>
<td>K-loop H/ACA box sRNAs</td>
<td>143</td>
</tr>
<tr>
<td>5.9</td>
<td>Multiple targets of <em>Pyrobaculum</em> H/ACA box sRNAs</td>
<td>145</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>6.1 Examples of synthetic biology applications of kink-turns</td>
<td>155</td>
<td></td>
</tr>
<tr>
<td>6.2 Small RNAs of <em>S. sofataricus</em></td>
<td>157</td>
<td></td>
</tr>
<tr>
<td>6.3 Predicted Cis-acting regulatory element of <em>Pyrobaculum</em> L14e predicted structure</td>
<td>159</td>
<td></td>
</tr>
<tr>
<td>6.4 Evidence that the predicted Cis-acting regulatory element is in the 5′-UTR of <em>Pyrobaculum</em> L14e</td>
<td>160</td>
<td></td>
</tr>
<tr>
<td>6.5 Enriched HACA box sRNA</td>
<td>166</td>
<td></td>
</tr>
<tr>
<td>6.6 SRP RNA is slightly enriched</td>
<td>167</td>
<td></td>
</tr>
<tr>
<td>6.7 Gel-shift assay of L7Ae mRNA in <em>Pyrococcus furiosus</em></td>
<td>169</td>
<td></td>
</tr>
<tr>
<td>6.8 Kink-turn of L7Ae mRNA is enriched</td>
<td>170</td>
<td></td>
</tr>
<tr>
<td>6.9 Enriched mRNA fragments</td>
<td>171</td>
<td></td>
</tr>
<tr>
<td>6.10 Enriched antisense small RNAs</td>
<td>172</td>
<td></td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Examples of C/D box sRNAs not predicted by the covariance model and Infernal v1.1</td>
<td>43</td>
</tr>
<tr>
<td>3.2</td>
<td>Composite and transposed C/D box sRNAs</td>
<td>62</td>
</tr>
<tr>
<td>4.1</td>
<td>Summary of species for panarchaeal C/D box RNA analysis</td>
<td>91</td>
</tr>
<tr>
<td>4.2</td>
<td>Summary of small RNA seq samples</td>
<td>92</td>
</tr>
<tr>
<td>4.3</td>
<td>Number of C/D box sRNAs with K-turns instead of K-loops</td>
<td>114</td>
</tr>
<tr>
<td>5.1</td>
<td>Previously annotated archaeal H/ACA box sRNAs</td>
<td>125</td>
</tr>
<tr>
<td>5.2</td>
<td>H/ACA box sRNA homologs of the Desulfuroccales order</td>
<td>139</td>
</tr>
<tr>
<td>6.1</td>
<td>RNPs that contain kink-turns</td>
<td>152</td>
</tr>
<tr>
<td>6.2</td>
<td>Conserved Kink-Turns in Ribosomal RNA</td>
<td>153</td>
</tr>
</tbody>
</table>
Abstract

Evolution of structure and function of Kink-turn containing RNAs in the Domain Archaea

by

Lauren Michelle Lui

The Domain Archaea is comprised of prokaryotic organisms known for often living in extreme environments. Despite potential biotechnological uses and their importance to the carbon and nitrogen cycles, archaea are understudied compared to eukaryotes and bacteria. Considering that archaea share features of bacterial RNA regulation (e.g., CRISPRs) and have transcription and translation machinery similar to eukaryotes, elucidating mechanisms of archaeal RNA biology may help us understand the evolution of RNA in the tree of life. For my dissertation I studied archaeal RNAs that contain the kink-turn, an RNA structural motif, with an emphasis on C/D box sRNAs and H/ACA box sRNAs.

I used high-throughput RNA sequencing data and RNA bioinformatics to (1) improve detection and explore biological roles of archaeal C/D box sRNAs, (2) improve detection and analysis of H/ACA box sRNAs across the archaea, and (3) survey kink-turn-containing RNAs in archaea. C/D box and H/ACA box sRNAs guide the 2′-O-methylation and pseudouridylation, respectively, of transfer and ribosomal RNAs in archaea and are likely under-annotated. To improve computational detection of C/D box and H/ACA box sRNAs, I incorporated the RNA kink-turn motif into covariance
models. My results are supported by comparative genomics, promoter predictions, and small RNA sequencing from 20 diverse archaeal species. I increased the annotations across the archaeal domain and, in particular, increased annotations by 20-30% in the hyperthermophilic species. I also demonstrated that non-canonical forms of H/ACA box sRNAs are not unique to the *Pyrobaculum* genus, but are widespread across the archaeal domain and compose approximately half of all H/ACA box sRNAs.

During my work with C/D box sRNAs, I found that the sequences comprising archaeal kink-turns are likely more diverse than previously thought. The kink-turn motif is a three-dimensional RNA motif ubiquitous across the three domains of life and is a key component of important ribonucleoprotein (RNP) complexes, such as the ribosome, riboswitches, and small nucleolar RNPs. Recent discoveries of kink-turns in archaeal RNase P and the unusual form of *Pyrobaculum* H/ACA box sRNAs indicate that there may be kink-turns in RNAs that we have not yet discovered. To improve understanding of this motif, provide data for computational models, and find novel kink-turn containing RNAs, I sequenced RNAs co-immunoprecipitated with the archaeal kink-turn binding protein L7Ae, which has a broad-binding specificity. I used a strain of *Thermococcus kodakarensis* with a HA-tagged L7Ae for these experiments. I found evidence that kink-turns occur in mRNAs and some annotated non-coding RNAs.
This work is dedicated to my mother and my sister.

For their love, support, guidance, and most of all, putting up with my shenanigans.
Acknowledgments

There are many people for me to thank on my road to getting a PhD. This end product, my dissertation, is not just the result of my efforts but also reflects the mentoring and support I have received. First and foremost, I want to thank my adviser Todd Lowe, and my committee members Manny Ares, Kevin Karplus, and Jeremy Sanford for their generous support and guidance over the years.

Many of the Biomolecular Engineering faculty have also helped me including Mark Akeson who gave me freedom to design and teach the BME150 lab with David Bernick, which gave me valuable teaching experience; Josh Stuart for mentoring me while I was an undergraduate in the UC LEADS program and introduced me to Bioinformatics; Ed Green for advice on writing; and Rebecca Dubois for advice that led to my postdoc. I also thank Tracie Tucker and Carol Mullane for their administrative support.

I thank Lowe Lab members, past and present for helpful discussions and advice: David Bernick, Andrew Uzilov, Andrea Corredor, Andrew Smith, Andrew Holmes, Brian Lin, Patricia Chan, Eva Robinson, Julie Murphy, Aaron Cozen, Janani Ravikrishnan, Allysia Mak, Titash Chatterjee. I especially want to acknowledge David and Eva for their mentorship and assistance in the wetlab and Andrew Uzilov for introducing me to covariance models. David and Andrew both provided helpful discussions on computational methods and worked on the initial Pyrobaculum C/D box sRNA annotation and analysis with me.
I had the wonderful opportunity to mentor students in both molecular biology and bioinformatics. These students helped me with my doctoral work: Rahul Parulkar, Annie Apffel, Alex Ayon, Matt Speir, and Ashmitha Rajendran. Matt Speir helped me scan and analyze the H/ACA box sRNA predictions in Chapter 5. Matt and Ashmitha also assisted in the development of the target prediction program.

Patrick Dennis has been an amazing collaborator and mentor. He provided extensive analytic and editorial support for the study of Pyrobaculum C/D box sRNAs in Chapter 3. He has also provided general advice for my work on H/ACA box sRNAs and kink-turn RNAs.

Other labs have been generous and hosted me so that I could learn new techniques. Tom Santangelo hosted me for two weeks at Ohio State to learn Thermococcus kodakaraensis genetics, and Jeremy Sanford in his lab to learn CLIP-seq with Jon Howard. As a student from a computational background, these experiences gave me valuable knowledge and training in molecular biology methods.

I would like to give a special acknowledgement to Zia Isola and the officers of UCSC Women in Science and Engineering group. These smart, innovative women joined me in my pursuit to create dialog about diversity and support other women in science and engineering, but they also became my friends and provided invaluable support as I worked toward my degree. I thank the Graduate Division, Research Mentoring Institute, and Women’s Center for giving us funds for our programming—in particular, Dean Tyrus Miller, who has been very supportive and allowed us to try out new programming.

Over the course of my graduate career I have been supported by different
fellowships and I thank the organizations for their support: ARCS fellowship, NASA-ARP grant, Google Anita Borg Scholarship, NIH training grant, and ACM-W travel grant.

Finally, I would like to thank my family and friends for their support. Without them this would not have been possible.
Chapter 1

Introduction

For my dissertation, I developed new prediction methods and greatly improved annotation of archaeal C/D box and H/ACA box sRNAs, which guide the 2′-O-methylation and pseudouridylation of target ribonucleotides in rRNA and tRNAs, respectively. By incorporating the RNA kink-turn into my computational models, I not only improved the computational detection of archaeal C/D and H/ACA box sRNAs, but I also revealed patterns of the kink-turn variation in archaea. With a comprehensive set of sRNAs, I studied domain-wide and clade-specific patterns of C/D box and H/ACA box sRNAs in regards to number, structure, and function.

At the core of my work is the relationship between RNA sequence and structure, and how these two elements relate to non-coding RNA function. I chose to use covariance models for RNA gene prediction, because they incorporate structure and sequence information and can easily tolerate insertions and deletions compared to the consensus model. Covariance models are described in more depth in Section 2.3.3.
Development of these models included iterative development and using C/D box and H/ACA box sRNAs discovered from small RNA sequencing (RNAseq) data across 20 diverse archaeal species. This data provided true positives for my models and indicated sRNAs that were not detected by previous models. These computational methods will provide a valuable resource for scientists to annotate newly sequenced archaeal genomes, especially if they do not have RNAseq data.

My work focuses exclusively on archaea, microorganisms primarily known for living at the extremes of life. Archaea have likely existed for more than three billion of years, adapting to a variety of Earth’s diverse environments from deep ocean vents [86] to the inside of the human gut [55]. Archaea are estimated to be 20% of marine biomass, participating in the nitrogen and carbon cycles [13, 16, 65]. Despite their long history, archaea were not categorized into their own distinct slice of life until the late 1970s [202], partially due to how difficult they are to culture and the sequencing technology available at the time. The physiology of archaea make them attractive for biotechnology and astrobiology studies, but our understanding of their inner workings remains poor, reflected by the fact that nearly half of archaeal genes do not have an annotated function [2]. This gap in our knowledge indicates that we are missing information valuable to understanding the evolutionary history of the three domains of life.

My first two chapters (Chapters 2 and 3) describe my studies of archaeal C/D box sRNAs. To study archaeal C/D box sRNA evolution and function, I compiled a comprehensive set of C/D box sRNAs within the hyperthermophilic genus *Pyrobaculum*. With improvements to computational prediction methods and the use of small RNAseq
data and comparative genomics, my set is likely a complete or near complete set of C/D box sRNAs within an archaeal genus. With this set I studied the structure, targets, and genomic context of archaeal C/D box sRNAs. I demonstrated that the covariance model could predict C/D box sRNAs in other archaeal species efficiently. Next, I extended my computational prediction methods and analysis to the rest of the archaea. This archaea-wide search for C/D box sRNAs confirmed and revealed clade-specific patterns. In addition, my findings suggest that not all thermophiles have a large number of C/D box sRNAs (greater than 50). Highly conserved guides, those conserved across entire orders or multiple phyla, have targets in functionally important sites of the ribosome.

Chapter 4 discusses the discovery of non-canonical H/ACA box sRNAs in all major archaeal clades. Previously, our lab discovered non-canonical forms of H/ACA box sRNAs in the *Pyrobaculum* genus. I created covariance models of these forms and scanned genomes across the Archaea. Now close to half of all annotations (49.5%) are of non-canonical forms. The lack of awareness and detection methods for non-canonical forms is a major reason for the under-annotation of this RNA class in archaea.

My final chapter asks, does the kink-turn hold any other secrets in the archaea? To answer this question I sequenced RNAs co-immunoprecipitated (co-IP) with L7Ae. The RNA kink-turn is a ubiquitous RNA motif found in all three domains of life and is bound by the L7Ae/L30e family of proteins. In archaea, the kink-turn is a characteristic of multiple RNA types, including the ribosome, C/D box sRNAs, H/ACA box sRNAs, and SRP RNA. I found enriched fragments of mRNAs and small RNAs that had the potential to form kink-turns.
Much of this work has been in collaboration with Matt Speir, with Dr. Andrew Uzilov, formerly of the Lowe lab, and with Dr. Patrick Dennis at Janelia Farm Research Campus. Target prediction and C/D box sRNA annotation in the archaea was originally a joint effort with Uzilov. Dennis provided extensive editorial and analytical support for my first chapter and we will be co-first author when it is published. Matt Speir is a major collaborator to the H/ACA box RNA work. I guided him to run predictions of the H/ACA box RNAs and he helped write some of the target analysis pipeline.

As the “third domain,” Archaea is often compared to Bacteria and Eukarya, the other domains of life where we know more. Archaea give us another facet of life to consider and remind us the diversity life that can emerge. What have we learned since the discovery of archaea 45 years ago? What have we learned about archaeal C/D and H/ACA box sRNAs since their discovery 15 years ago? This dissertation provides perspective on the Domain Archaea in the context of small non-coding RNAs containing the kink-turn.
Chapter 2

Background and Significance

2.1 The Archaea

2.1.1 Archaea: the third domain of life

The Domain Archaea was born in the late 1970s when Carl Woese and colleagues proposed that archaebacteria should be in their own distinct group of life separate from bacteria and eukaryotes [202]. Previously scientists classified archaea with bacteria since both are prokaryotes. However, after analyzing small subunit (SSU) rRNA from a methanogenic microbe, Woese and colleagues found that the sequence looked neither bacterial nor eukaryotic (modern reflection on the discovery discussed in [149]). In addition, Woese and George Fox also pointed out that although methanogens resembled bacteria, they do not have peptidoglycan cell walls and have different patterns of modification in rRNA and tRNA. With this evidence, Woese and colleagues convinced the scientific community (although not without resistance) that there should
be a three domain system. The creation of the archaeal domain represents a triumph of using sequence analysis, rather than only physical characteristics, to reveal evolution.

In recent years scientists have debated whether the term prokaryote should be retired since it implies a particular phylogeny that divides organisms into eukaryotes and not eukaryotes [147, 148, 199, 198].

In addition to their unique characteristics, archaea have similarities to eu-
karyotes and bacteria. Similarities between archaea and bacteria are mostly structural (circular chromosomes, no nucleus, operons), but both also share CRISPR-mediated immunity and horizontal gene transfer. Archaea do not use sigma factors to initiate transcription like bacteria, but archaeal mRNAs do have Shine-Dalgarno sequences for
ribosomal binding [166]. Like eukaryotes, archaeal mRNA promoters also have TATA boxes and may have B recognition elements [159]. Archaea have homologous genes with eukaryotic informational processing (transcription, translation, and replication), as well as other eukaryotic-like proteins scattered throughout the Archaea [29]. The post-transcriptional modification system using C/D and H/ACA box sRNAs to precisely target ribonucleotides for 2′-O-methylation and pseudouridylation is also conserved between Archaea and Eukarya [144, 74, 179, 116]. Some eukaryotic features are restricted to particular phyla, such as topoisomerase IB (Thaumarchaeota), the subunit of Rpb8 of RNA polymerase (Crenarchaeota and Korarchaeota [106]), and the ubiquitin modification system (Aigarchaeota) [29]. These features make the phylogenetic relationship between Archaea and Eukarya more uncertain in regards to whether eukaryotes arose from an archael lineage or Eukarya and Archaea are sister lineages.

Scientists have generated multiple hypotheses on how archaea fit into the origin of eukaryotes. The ‘eocyte hypothesis’ proposes that eukaryotes arose from a member of the Crenarchaeota phylum [46], while the hydrogen [122] and syntrophic [128] hypotheses suggest that eukaryotes originated from the fusion of a euryarchaeal species and a bacterium. More evidence supporting the hypothesis that eukaryotes arose from an archael ancestor comes from analysis of a new archael phylum, Lokiarchaeota. Phylogenetic analysis indicates that Lokiarchaeota is monophyletic with Eukarya and eukaryotic-like membrane deformation proteins are present in its members [172]. As the number and diversity of archael genomes sequenced increases, the relationship between the three domains of life will become more clear.
2.1.2 Phylogeny and traits of the Archaea

Here I describe the major clades of the archaeal domain for reader’s reference. I discuss clade-specific patterns of C/D box and H/ACA box sRNAs in Chapters 4 and 5.

2.1.2.1 Major phyla of the Archaea

The three-domain system was formally proposed in 1990 and also introduced the two major phyla of the archaea, Crenarchaeota and Euryarchaeota (Figure 2.1). Archaeal phylogeny has been proposed and analyzed based on comparing SSU rRNA (such as in the 1977 study by Woese and colleagues [202]), but also by analysis of ribosomal proteins and other core proteins [29]. Currently there are eight proposed phyla in the Domain Archaea, but nearly all archaeal species can be classified into Euryarchaeota or Crenarchaeota. Thaumarchaeota is also a widely accepted phylum, whose members are known for being ammonia-oxidizers and playing a role in the nitrogen cycle [192, 28]. Three other phyla each contain a single deeply-branching member: Aigarchaeota, Korarchaeota, and Nanoarchaeota. The sole member of Nanoarchaeota, Nanoarchaeum equitans, is a symbiont and will only grow in the presence of crenarchaeote Ignicoccus hospitalis. The last phylum, Lokiarchaeota, was proposed this year by Spang et al. based on a metagenomic sample gathered at a hydrothermal vent [172]. Phylogenetic analysis of Lokiarchaeota indicates that it is a deeply branching member of the ‘TACK’ superphylum and supports the original intent for creating the ‘TACK’ superphylum. The Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota
phyla were grouped into the superphylum to reconcile phylogenetic analysis of various archaeal phyla and eukaryotes, but also in anticipation that eukaryotes originated from the ancestor of multiple archaeal phyla [78].

2.1.2.2 Major growth and phenotypic characteristics of archaeal phyla and orders

Members of the Domain Archaea are found in diverse environments, including those at extremes in salinity, temperature, acidity, or pH. Often species are grouped together based on what characterizes their growth conditions or phenotypes. Some of these groups, such as thermophiles (optimal growth at >60° C) and hyperthermophiles (optimal growth at >80° C) are found across the domain, but some are only found in certain orders.

The Euryarchaeota phylum contains methanogens, halophiles, thermophiles, and the ARMAN (Archaeal Richmond Mine Acidophilic Nanoorganisms) species, which have microgenomes [10]. The Halobacterales order contains all of the halophiles (‘salt loving’), and many of these species have more than one chromosome or have plasmids [146]. Methanogens reduce carbon dioxide into methane and are typically split into two classes, although some studies have argued for three [5]. Class I is composed of the Methanobacteriales and Methanococcales orders. The Methanosarcinales, Methanocellales, and Methanomicrobiales make up Class II. Recently Anderson et al. proposed that the Methanosarcinales be their own class [5]. Thermophiles and hyperthermophiles occur in the Thermococcales and Archaeoglobales orders.
Figure 2.2: Unrooted Bayesian tree of the Archaea based on 57 ribosomal proteins. Reproduced from [29]. Major phyla and orders are shown.
The divisions of Crenarchaeota are less obvious than in Euryarchaeota, since all crenarchaeal species are thermophiles or hyperthermophiles. However, these species live in a wide variety of environments including freshwater, marine, and environments with acidic or neutral pH. Crenarchaeal species may be aerobic or anaerobic and reduce variety of compounds including sulfur, nitrate, and iron.

Although archaea are usually characterized for living at extremes, most of the Thaumarchea are considered mesophiles and are found in marine, freshwater, and soil environments. However, \textit{Cenarchaeum symbiosum} is considered a psychrophile (‘cold-loving’) species and grows at approximately 10° C. Few other psychrophiles are known and occur in the Euryarchaeota phylum (\textit{Methanococcoides burtonii} and \textit{Methanogenium frigidum}) [2].

As I will discuss in Chapters 4 and 5, there are clade-specific patterns of abundance and form in archaeal C/D box and H/ACA box sRNAs. Very few sno-like RNAs are found in the halophiles and the Class II methanogens [20 51], and my findings agree with the hypothesis that these archaea do not have many C/D or H/ACA box sRNAs. In addition, although hyperthermophiles in both the Crenarchaea and Euryarchaea usually have greater than 50 C/D box sRNAs [51], from my work it appears that this may not be true for all hyperthermophilic methanogens and other species closely related to the methanogens such as those in the order Archaeoglobales. In regards to form, I found that the recently discovered non-canonical H/ACA box sRNA forms [17] occur in different proportions among the various archaeal clades.
Figure 2.3: Two-dimensional Representations of the Kink Turn. (A) Figure from [102] of the consensus kink turn. (B) 2D representation of the kink-turn motif from [170]. ‘C’ indicates the ‘canonical stem,’ which is also known as stem I and is shown in yellow. ‘NC’ indicates the ‘non-canonical stem,’ which is also known as stem II and is shown in grey. The stacked, sheared G:A base pairs are shown in green. The Ln positions represent the bulge positions.

2.2 The RNA Kink-Turn Motif

In this section I discuss the aspects of the kink-turn motif that were incorporated into my computational models, the role of the kink-turn in RNA function, and the biology of kink-turn binding proteins.

2.2.1 Structure and Characteristics of the RNA Kink-Turn Motif

The archetypal kink-turn motif consists of approximately 15 nucleotides that form an asymmetrical bulge closed by two helical regions known as stem I and stem II, or the ‘canonical stem’ and ‘non-canonical’ stem, respectively (Figure 2.3) [102, 195]. Hydrogen bonding of stacked, sheared G:A base pairs adjacent to the asymmetrical bulge result in a sharp ‘kink’ in the RNA backbone that is roughly 60 degrees (Figure 2.4). In kink-turn studies, the base pairs in the extended stem II are numbered starting with the G:A base pairs and the first letter in the pair is from the strand with the bulge [170, 113].
Nucleotides in this strand have a \( b \) after their position number. The non-bulged strand is referred to as \( n \). Thus the first two positions in stem II are typically position 1b1n being G:A and position 2b2n being A:G (Figure 2.3).

In addition to the cross-strand purine stack, the \textit{trans} Hoogsteen/sugar edge interaction between the adenines and guanines of positions 1b1n and 2b2n and an A-minor interaction help stabilize the kink-turn structure [102, 45, 109, 136] (Figure 2.5). Experiments with the human L7Ae homolog, 15.5 kD protein, indicate the importance of the stacked purines in protein recognition. Mutations or deletions of any of these nucleotides results in the loss of protein binding [139]. The characteristic sharp axial bend of this motif is stabilized by the presence of monovalent or divalent metal ions [76, 124]. Turner \textit{et al.} demonstrated that the binding of the L7Ae protein also helps stabilize the motif [186], so likely binding of other L30e/L7Ae family proteins can help stabilize the kink-turn structure as well. Molecular dynamics simulations demonstrate the potential for the kink-turn to adopt a more extended geometry and to act as a flexible hinge in the absence of these stabilizing factors [44, 158]. The stabilization of kink-turn structure by binding partners hints at the importance of proteins that bind to this motif.

Variants of the kink-turn exist, even those that defy the seemingly necessary rule to have tandem sheared G:A base pairs. The K-loop variant, which commonly occurs between the C' and D' boxes of C/D box RNAs in archaeal species, has stem I replaced by a terminal loop (Figure 2.6). Hamma and Ferre-D’Amare characterized a minimal kink-turn that does not require the formation of stem I to be recognized.
Figure 2.4: **Three-dimensional Structure of the Kink Turn with L7Ae.** The crystal structure of a L7Ae-C/D box RNA complex. L7Ae is labeled with secondary structure elements and the nucleotides for a 25-mer RNA. G6:A20 and G19:A7 are the stacked, sheared G:A base pairs. G17 is also participating in the stacking. It is common for a purine in this position to stack and for a uridine (in this example U18) to protrude due to this stacking. U18 and G17 correspond to the L1 and L2 positions from figure 1A, respectively. G1 through U5 and U21 through C25 participate in the non-canonical stem. G15, G14, C8, and C9 participate in the canonical stem. Adapted from [127].
Figure 2.5: Illustration of the stabilizing trans Hoogsteen/sugar edge interaction of adenine and guanine. These interactions that occur between the GA base pairs of the kink turn motif help stabilize the kink turn structure. Modified from [109].

by L7Ae [79]. This minimal kink-turn occurs in H/ACA box RNAs and in C/D box RNAs [79, 137]. Some variants replace one or both of the G:A base pairs. A kink-turn in the *Thermus thermophilus* 30S ribosomal subunit, Kt-23, has an A:U pair replacing the typical A:G pair distal to the bulge of the kink-turn. This interaction has been confirmed by single-molecule FRET experiments [169]. A very unusual kink-turn that has U:G/G:C replacing the canonical G:A base pairs has been found in the 5e domain of the human signal recognition particle (SRP) RNA and confirmed by crystral structure analysis [87]. The unusual features of this kink-turn are also conserved across multiple archaeal clades [210]. More complex interactions that stray from the typical features that stabilize kink-turn structure exist, such as in Kt-11 and Kt-15 which are both found in crystal structures of the ribosome [170, 197, 21] (Figure 2.7). These examples that stretch our understanding of kink-turn structure indicate that there may be additional diversity in kink-turn structure and sequence yet to be discovered.
Figure 2.6: **Variants of the kink-turn** (A) The consensus secondary structure of the kink-turn. Stem I (the ‘canonical stem’) (green) and stem II (the ‘noncanonical stem’) (blue) flank the tandem-sheared G-A base pairs (red) and the nucleotide that commonly protrudes once the structure is formed (yellow). (B,C) Examples of the minimal kink-turn variant that has a loop instead of stem I. (C) A kink-turn variant from ribosomal RNA that has the typical G-A base pairs replaced with a more complicated molecular interaction. Modified from [183].

Figure 2.7: **Example of an unusual kink-turn, KT-11.** The typical A:G pair distal to the bulge of the kink-turn is replaced with a more complicated molecular interaction. Not only is the 2b2n position an A:A pair instead of the typical A:G, the 2n nucleotide is not immediately adjacent to the 1n nucleotide. From kink-turn database ([170], http://www.dundee.ac.uk/biocentre/nasg/kturn).
2.2.2 Role of the kink-turn in human diseases

Loss or mutation of specific kink-turn-containing RNA types plays a role in human disease. The loss of particular C/D box RNAs, a type of kink-turn-containing RNA typically associated with RNA modification, is implicated in the etiology of the neurogenetic disorder Prader-Willi Syndrome (PWS). This congenital disease affects approximately 1 in 8,000-20,000 people and is characterized by mental retardation, hyperphagia, and short adult height [98, 132]. Dysregulation of two C/D box snoRNA clusters, SNORD116 and SNORD115, has been implicated in the etiology of PWS. Multiple studies provide evidence that a microdeletion resulting in the loss of the SNORD116 cluster (HBII-85) causes the PWS [53, 164, 50, 57]. Another C/D box snoRNA cluster implicated in PWS is SNORD115 (HBII-52) and studies indicate that SNORD115 is missing in Prader-Willi patients [98]. Evidence suggests that snoRNAs from these two clusters regulate the alternative splicing of a serotonin receptor mRNA and other mRNAs [99, 98]. PWS provides an example of how kink-turn-containing RNAs are involved in human disease and that C/D box sRNAs may have other unknown roles beyond 2'-O-methylation.

Human RNase mitochondrial RNA processing (MRP) RNA is another RNA predicted to contain a proposed kink-turn with a possible link to a variety of human diseases. Mutations of this RNA have been implicated in various inherited immunodeficiency and developmental disorders: cartilage-hair hypoplasia, metaphyseal dysplasia without hypotrichosis, anauxetic dysplasia, kyphomelic dysplasia, and Omenn syndrome.
Figure 2.8: The RNase MRP kink-turn. Mutations known to lead to cartilage-hair hypoplasia are in red. The nucleotides participating in the putative kink-turn are contained within the green box. The purines that make up the stacked, sheared G-A base pairs are connected with black lines.

(reviewed in [121]). There is evidence that the RNase MRP complex has a role in ribosome biogenesis [117] and CLB2 mRNA degradation in the regulation of mitosis [75] in addition to its mitochondrial function. The proposed kink-turn [102, 170] is the binding site of human protein Rpp38, which is known to contain an L7Ae/L30e domain [91]. Mutations leading to cartilage-hair hypoplasia occur in the predicted kink-turn, which involves positions 146–154 and 164–169 of RNase MRP RNA [121].

The negative consequences of modification or loss of kink-turn-containing RNAs emphasizes their importance to cellular machinery.
2.3 Computational prediction of RNA structure

Many RNA folding algorithms view the base-pairing of an RNA structure as a type of structural palindrome. Instead of two corresponding positions in the string having the same letter, these two positions can form complementary base pairs. The whole RNA structure can be viewed as a set of nested, pairwise interactions. With this representation, dynamic programming algorithms can be used to calculate the structure(s) that maximize the number of base-pairs. However, it is more realistic to assume that an RNA will adopt a minimum free energy structure than simply one that maximizes base-pairing. To improve the accuracy of dynamic program algorithms, current methods incorporate thermodynamics [85, 209] and comparative sequence information [104, 58]. In this section I focus on methods that use user-defined motifs and probabilistic models, since these methods are more easily extended to include statistical sequence information than ones that use RNA thermodynamics.

I incorporated the kink-turn motif in the computational models I created in my work. To model the kink-turn I chose to use covariance models because they can (1) tolerate insertions and deletions in the consensus, (2) model the G:A base pairs of the kink-turn, and (3) incorporate sequence and structural information. In addition to describing covariance models, I briefly describe other methods I considered to model the kink turn.
2.3.1 **Tools that Take User-Defined Motifs and Patterns**

Current RNA secondary structure prediction tools that can detect kink-turns use varying amounts of structural and sequence information and are diverse in the amount of motif variation (e.g., length of loops) that can be modeled in the algorithm. RNAmot [73] and PatScan [56] are easy to use and simply take an input motif, but cannot take position-specific base probabilities into consideration and cannot easily handle non-canonical base pairs. RNAmotif [118] is more flexible although it also does not use base probabilities either. It performs well on large RNA elements (more than 100 nt), but has difficulty finding motifs that are less than 20 nt long and would not be feasible for detecting the K-loop variant of the kink-turn [47]. RNAcontext [94] and ERPIN [107] take into account both user-designed structural motifs and base position probabilities, but they cannot handle insertions of arbitrary length at specific positions.

2.3.2 **Tools That Use 3D Structural Data**

RMDetect is a promising tool that can detect 3D structural motifs from sequence [47], but it is limited by the type of training data that it can use. This algorithm uses Bayesian network models and base-pair probability detection. To build a new model, it is necessary to have PDB files of known crystal structures. The MCfold program [150] is also able to use RNA sequence as an input to predict a 3D structure, but the computation time and complexity for sequences longer than approximately 30 nt would make it infeasible for a whole transcriptome or genome search.
2.3.3 Stochastic-Context Free Grammars (SCFGs) and Covariance Models

SCFGs are algorithms that can model secondary structures in a probabilistic manner, capable of capturing long-range interactions and predicting RNA structures from sequence (reviewed in [54, 58]). In general, grammars are finite formal machines that can generate sequences and also recognize and score strings based on the model. Context-free grammars generate a string of symbols based on user-defined production rules (an in-depth description of how the production rules of context-free grammars differ from other grammars can be found in [58]). To create an SCFG, a probability is associated with each production rule in a CFG. SCFGs are easier to implement and can be nearly as accurate as energy-minimization algorithms in predicting RNA secondary structure [54].

Typically, SCFGs are parameterized using known sequences and structures (training). After it is parameterized, the SCFG can be used to calculate the probability of a given sequence given the model (scoring predictions). To model RNA families based on a consensus secondary structure, a special case of SCFGs, the covariance model, can be used [59]. More formally, covariance models are profile-SCFGs and describe secondary structure and the sequence consensus of an RNA structure family. With a multiple sequence alignment annotated with secondary structure, you can build a covariance model that can be used for database searching. A collection of covariance models for RNA families has been compiled in the Rfam database [70].
The Infernal software package \cite{134,133} contains programs that build covariance models from sequence and secondary structure alignments and use covariance models to search sequence databases. Infernal v1.1 was released in late 2013 and introduced truncated hit detection and decreased search time using profile HMM methods and constrained covariance model alignment algorithms. The approximately 100-fold acceleration in search time comes at a slight cost to sensitivity \cite{133}. In the work that I describe in Chapters 3, 4, and 5 \cite{3,4,5}, I noted this loss of sensitivity in my predictions of C/D box and H/ACA box sRNAs. In general, only a few sRNAs were missed by Infernal v1.1. Final annotation sets I describe include predictions from both Infernal v1.0 and v1.1 scans. For general scans of archaeal sRNAs, I found that Infernal v1.1 is better, because it reduces the number of false positives and has a faster run time, usually minutes compared to hours. The reduction in false positives is useful especially if no small RNA sequencing data is available to verify predictions. However for comprehensive scans, Infernal v1.0 is better; I found some predictions made by Infernal v1.0 cannot be picked up by Infernal v1.1 with the same model, even if the bitscore threshold is set to -50.
Chapter 3

Structure, function, and evolution of archaeal C/D box sRNAs as revealed by a comprehensive set from the *Pyrobaculum genus*

3.1 Abstract

Archaea use C/D box sno-like RNAs (sRNAs) to guide precise 2’-O-methyl modification in ribosomal and transfer RNAs. Although C/D box sRNAs are the most numerous archaeal small RNA class, many are not annotated. To study archaeal C/D box sRNA function and evolution, I compiled, to my knowledge, the most comprehensive set of C/D box sRNAs of an archaeal genus. I used high-throughput small RNA sequencing data, computational methods and comparative genomics to catalogue 526 C/D
box sRNAs from six related species within the hyperthermophilic genus *Pyrobaculum* (phylum Crenarchaeota) into 110 homologous families based on sequence conservation of their guide regions and their guide-target complementarity to ribosomal and transfer RNAs. My computational model is the first to incorporate the RNA kink-turn, and testing suggests that it can predict >90% of C/D box sRNAs in an archaeal genome. My comprehensive analysis adds more evidence for the importance of guide-target interactions in ribosomal assembly (I find that approximately 45% of *Pyrobaculum* C/D box sRNAs are double guides with targets within 100 nucleotides of each other). I also find that the dynamic nature of C/D box sRNA sequence and C/D box sRNA proximity to other types of genes influences C/D box sRNA function and may impact the processing of archaeal protein-coding genes.

### 3.2 Collaboration in this study

I did much of this work in collaboration with Dr. Patrick Dennis, who will be co-first author when this chapter is published. Dennis and I analyzed the computational predictions of methylation targets that I generated. Dennis mapped target predictions onto rRNA secondary structure and conducted analysis of the long-range interactions of sRNA guides. Dennis, Dr. David Bernick and Dr. Andrew Uzilov helped me annotate the C/D box sRNAs and hand-curate box features. Bernick ran the Northern blot of the polycistronic C/D box sRNA in Pae. I developed the target prediction program with Uzilov and Andrea Corredor. Corredor analyzed potential mRNA, C/D box sRNA,
H/ACA box sRNA, and CRISPR targets for orphan guides under my guidance.

3.3 Introduction and Background

In eukaryotic cells, ribosome assembly occurs in the nucleolus, a specialized structure located within the nucleus. This is the site where ribosomal RNA (rRNA) is transcribed, modified, processed, folded, and assembled along with ribosomal proteins into the large and small ribosomal subunits. The nucleolus also contains a large number of small (sno)RNAs that have been implicated in the modification, folding, and maturation of rRNA (reviewed in [116]). These snoRNAs are incorporated into dynamic ribonucleoprotein (RNP) complexes that act as molecular machines along the ribosomal assembly line. The snoRNAs in these complexes use guide sequences to base pair with rRNA and target modifications to specific nucleotide locations within the region of complementarity. The snoRNAs divide into two classes: C/D box snoRNAs which guide 2′-O-methylation of ribose residues and H/ACA box snoRNAs which guide the conversion of uridine to pseudouridine [101, 68]. Although archaeal cells do not contain an organized nucleolar structure, they possess and utilize both C/D box and H/ACA box sno-like RNAs (sRNAs) in the modification of rRNA and assembly of ribosomal subunits (reviewed in [116, 193]).

Archaeal C/D box sRNAs are generally about 50 nucleotides (nts) in length and contain highly conserved C (RUGAUGA consensus) and D (CUGA consensus) box sequences at the 5′ and 3′ end of the molecule and less conserved versions (designated
C′ and D′) near the center of the molecule [12]. These RNAs fold into a balloon-like hairpin as a result of the formation of a kink-turn (K-turn) structural motif through the interaction of the C and D box sequences and a K-loop motif through the interaction of the D′ and C′ box sequences (Figure 3.1). The K-turn and the K-loop are each recognized by a copy of the protein L7Ae [137]. The binding of L7Ae stabilizes the RNA structure and allows two copies each of Nop56/58 (also called Nop5) and fibrillarin to bind, completing the active ribonucleoprotein (RNP) complex [143]. The fibrillarin protein is a S-adenosyl methionine dependent RNA methylase and is responsible for the catalytic activity of the RNP complex. Because of the strong protein-protein interactions, a single K-motif within the sRNA may be sufficient to assemble an active complex. These interactions explain how sRNA box sequences can be mutated without completely losing in vitro methylation activity [145, 67].

The two guide regions between the C and D′ boxes and between the C′ and D boxes are unstructured and each is available to base pair with an approximately 8–12nt target sequence (Figure 3.1). In addition to rRNA targets, a significant proportion of archaeal sRNAs have guide regions that are complementary to transfer RNA (tRNA) [143, 43]. Methyl modification in the target RNA occurs at the nucleotide position that base pairs with the guide five nucleotides upstream from the start of the D′ or D box sequence. This is known as the “N+five” rule and methylation targets are referred to as the D and D′ targets [101]. Many C/D box sRNAs with strong box features have guides that lack complementarity to rRNA and tRNA sequences; these guides may target other RNAs, but to date no conserved targets to alternative RNAs have been identified within
Figure 3.1: **Conserved features of archaeal C/D box sRNAs** The typical structure of an archaeal C/D box sRNA contains two K-motifs: the K-turn formed by the interaction between the C and D box sequences and the K-loop formed by the interaction between the D’ and C’ motifs (black dashed lines). Positions of the K-motifs are referenced by the positions paired, e.g., 1b1n is typically a GA pair. The two guide regions, located respectively between the D’ and C boxes and between the D and C’ boxes, base pair with the target RNA, and methylation (yellow hexagon) occurs in the target nucleotide that base pairs with the guide five nts upstream from the start of the D’ or D box sequence. This is the “N+five” rule.

Although C/D box sRNAs are the most prevalent small RNA class in Archaea, they remain under-annotated and many computational prediction methods do not model all archaeal C/D box sRNA features, namely the K-turn and K-loop [69]. Originally, archaean C/D box RNAs were annotated with computational search methods and verified by Northern blots [71, 144, 64] or identified by Sanger sequencing of cDNAs [177, 206].
With the availability of high-throughput small RNA sequencing (RNAseq), many C/D box sRNAs can be identified from transcripts [17, 157, 174]. This data has also revealed that current computational models may miss a significant portion of archaeal C/D box sRNAs. In collaboration with Dr. David Bernick and Dr. Andrew Uzilov, I found that at least 23 sRNAs were missed from the original annotation of 65 in *P. aerophilum* [18]. Thus, for a comprehensive annotation of C/D box sRNAs in an archaeal genome, small RNAseq is necessary. Comparative genomics may also be required to detect sRNAs that are not expressed in the sample [18]. The under-annotation of this RNA class may prevent us from fully understanding the roles and evolution of archaeal C/D box sRNAs.

To conduct a thorough analysis of archaeal C/D box sRNA evolution and function, I compiled the first complete or nearly complete set of C/D box sRNAs present in an archaeal genus. I built upon previous studies from the Lowe lab that used small RNAseq data, computational methods, and comparative genomics to identify C/D box sRNAs in six species of the hyperthermophilic genus *Pyrobaculum* [18, 19]. A few studies have used comparative genomics to find and analyze sRNAs in other archaeal genera (*Sulfolobus* and *Pyrococcus*) [71, 144, 206], but these sets are incomplete due to lack of RNAseq data and insensitivity of computational methods available at the time [144, 51]. Other studies with small RNAseq data did not conduct comparative genomics in closely related species [157, 174, 8]. In this study, I used small RNAseq from *P. aerophilum* (Pae), *P. arsentaticum* (Par), *P. calidifontis* (Pca), *P. islandicum* (Pis), and new small RNAseq data for *P. oguniense* (Pog). The species *P. neutrophilus* comb. nov.
Figure 3.2: **Tree of Pyrobaculum species** Reproduced from [19]. Numbers are bootstrap percentages and nucleotide mutation rate is indicated by branch length (scale bar upper right corner). Species studied in this chapter are highlighted with blue boxes. *Thermoproteus neutrophilus* has been reclassified into the *Pyrobaculum* genus and named *P. neutrophilus* comb. nov. [34].

(Pne; formerly *Thermoproteus neutrophilum* [34]) was used to supplement comparative genomics analyses (Figure 3.2). With small RNA sequencing data and comparative genomics, I was able to find C/D box sRNAs that were not detected computationally and used the data to improve my computational method.

After compiling the *Pyrobaculum* C/D box sRNA set, I cataloged the sRNAs carefully into homologous groups based on sequence similarity and methylation target prediction. In addition to this comprehensive annotation, Dennis and I hand-curated computationally predicted methylation targets to rRNAs and tRNAs to aid our analysis and study orphan guides (guides with no targets to tRNA or rRNA). I used this extensive data set to examine (i) the sRNA sequence and guide variation, (ii) the function of these sRNAs in modification of rRNA and tRNA, and the assembly of ribosomes, (iii) the genomic context of sRNA genes, and (iv) the proliferation, mobility, plastic-
ity and evolutionary divergence of sRNA genes as viewed within and between the six
Pyrobaculum genomes.

3.4 Materials and Methods

3.4.1 Computational prediction of C/D box sRNAs

To generate computational predictions, I used covariance models to scan genomes
and small RNA sequencing data of Pae, Par, Pca, Pis, and Pog (data reported in [18]
except for Pog). The covariance model was created by using a hand-curated multiple
structural alignment of Pyrobaculum aerophilum C/D box RNAs (Appendix A.1.1) as
input to cmbuild from the Infernal v1.1 software package [133] with the hand-curated
option --hand specified. The covariance model was calibrated with cmcalibrate. A
final covariance model was built from a complete set of Pae sRNAs found from ex-
amining sequencing data and using comparative genomics with the other Pyrobaculum
species (see Appendix A.1). RNA structure formed by the box features (Figure 3.1),
including the two G:A pairs of the K-turn and K-loop, was annotated in the alignment.
The variable regions of archaeal C/D box sRNAs, the guides and variable loop, were
annotated to be any nucleotide. I used cmsearch to scan the genomes and small RNA
sequencing data using the glocal (-g) and no HMM filter (--nohmm) options. Candidates
from genome scans that overlapped by more than 80% with Genbank RefSeq genes or
with other annotated non-coding RNAs were discarded. With a low bitscore cutoff of 1
to detect as many sRNAs as possible, filtering by overlap with Genbank RefSeq genes
and other non-coding RNAs removed between 75–93% of false positives. On average, filtering removed 88% of false positives (criteria for classifying true and false positives are defined in Section 3.5.2).

To find additional orthologs of sRNAs in the Pyrobaculum genomes not found by the covariance model, sRNAs were used as queries to BLASTN [30]. Top hits were manually curated, based on predicted promoters, conservation, and sequencing evidence.

3.4.2 Organization of C/D box sRNAs into homolog families

C/D box sRNA homolog families were created and named based on sequence similarity of guide sequences and predicted targets. C/D box sRNA numbers indicate to which family each belongs. Thus, Pae sR01, Par sR01, etc. belong to the sR01 family. C/D box sRNAs were first grouped into families using the original annotation numbering in Pae (1-65, [64]). All other C/D box sRNAs were grouped into families starting at number 100.

3.4.3 Prediction of methylation targets

To identify the putative sites of 2′-O-methylation guided by Pyrobaculum C/D box sRNAs, I scanned mature rRNA and tRNA sequences for regions of complementarity to the D and D′ guides of the sRNAs with a Python program (findAntisense.py, Appendix A.1) and evaluated top hits. The first nucleotide of the D and D′ boxes were allowed to participate in the guide-target interaction.

To obtain a uniform numbering system for sites of methylation, Dennis and I
first made an alignment of the 16S and 23S rRNAs from the six species (See Appendix [A.1]). Dennis then mapped the predicted positions of modification on the alignment and assigned a position for each prediction based on the Pae numbering.

We considered a prediction to be considered credible, if it had at least nine continuous Watson-Crick base pairs of complementarity centering at or near the “N+five” position. These criteria were relaxed in two specific instances. First, if the majority of members in an sRNA group met the prediction criteria, the prediction was extended to minority members that nearly met the criteria (for example, matches containing a mismatch or G:U base pair). Second, it has been noted that many sRNAs use their two guide regions to direct methylations to closely spaced nucleotides within the target RNA. Presumably this enhances target identification and creates greater stabilization of the guide target interaction within the RNP complex. Consequently, when one guide exhibits strong complementarity to the target, the criterion for the second guide match is relaxed if (i) it is 100 nts of the first complementarity (ii) the weaker complementarity contains no more than one mismatch (iii) and the combined bit score for the two complementarities was 32 or higher (where a Watson-Crick base pair is 2, a G:U base pair is 1 and a mismatch is -2).

### 3.4.4 Northern Blot of polycistronic sRNAs

Bernick prepared northern blots as described in [18]. The following DNA oligomers (Integrated DNA technologies) were used as probes:

Pae sR21 sense (GCCAGTGTCGGAAAATTGACGAGCTCACCCCTTTGC)
Pae sR21 antisense (GCAAAGGGTGAGCTCGTCAATTTTCGGACACTGGC). We used these blots to confirm the expression of the polycistronic sRNA described in Section 3.5.6.6. These blots also allowed Bernick and me to analyze the processing of the polycistron into individual C/D box sRNAs.

3.4.5 Small RNA sequencing and read processing of *Pyrobaculum* species

Bernick prepared Pae, Par, Pca, and Pis small RNA libraries with an in-house protocol. The libraries were sequenced on the 454 platform. Description of sample preparation for small RNA sequencing and read mapping are described in [18].

Sample preparation of small RNA libraries for Pog are described in [18, 188] and carried out by Bernick. Briefly, the small RNA size fraction was isolated by running total RNA in denaturing gel electrophoresis and extracting the region below tRNAs. Libraries were sequenced by the UC Davis Sequencing Facility on Illumina HiSeq 2000 to produce 2x75 nt paired-end sequencing reads. Barcodes and linkers were removed by a program written by Bernick (pairParser.pl). Reads were mapped using BLAT [95] and the resulting PSL file was processed with pairPicker.pl, also written by Bernick, to determine paired reads and generate a BED file.
3.5 Results and Discussion

3.5.1 Most Pyrobaculum C/D box sRNAs homologous families have members in all six species

We identified 526 C/D box sRNAs from six species of Pyrobaculum using evidence from (i) small RNA sequencing data from Pae, Par, Pca, Pis and Pog, (ii) a computational covariance prediction model, and (iii) comparative genomics. Nearly all of the sRNAs are represented in the RNAseq libraries and have strong phylogenetic conservation within the Pyrobaculum. The new covariance model incorporates the K-turn and K-loop structural information and allowed me to detect sRNAs where both guides lack complementarity to rRNA or tRNA sequences [116][115]. The model predicts the sRNA box features; I manually checked and adjusted the box features when required. Dennis, Bernick, and Uzilov also corrected some of the box features.

I organized the 526 sRNAs into 110 different homologous families based on sequence conservation of their guide regions and predicted targets of methylation in tRNA and rRNA. The sequences, family organization, and genomic location of these sRNAs can be found at the Lowe Lab Archaeal snoRNA-like C/D box RNA Database [114]. I found 26 additional sRNAs in this study (two in Pae, three in Par, four in Pca, three in Pis, four in Pog, and ten in Pne) and the number of C/D box sRNAs in individual species now ranges between 84 and 92 (Figure 3.3A). Grouping the sRNAs into families allowed me to study more easily the evolutionary origins and relationships of C/D box sRNAs genes within the genus and for Dennis and me to more accurately
Figure 3.3: **Most Pyrobaculum C/D box RNAs are highly conserved** (A) The number of identified sRNAs in each of the six species of the *Pyrobaculum* genus. (B) Most *Pyrobaculum* C/D box sRNAs have homologs in other *Pyrobaculum* species. The majority of sRNAs fall into families with representatives on all six species. C/D box sRNAs were organized into 110 homologous families based on sequence similarity of the guides and predicted targets in rRNA and tRNAs, and 70 of these families had members in all six *Pyrobaculum* species. A notable number of families (18) have only one member. Most of these families occur in *P. calidifontis*, which has a transposon-like sRNA (see Section 3.5.6.4).

Most of the homolog families are conserved, with 70 of the 110 families having representative sRNAs encoded in each of the six *Pyrobaculum* genomes. The remaining families (40) have representatives missing from one or more of the six genomes (Figure 3.3B). Eighteen of the families are unique with the representative present in only a single species. Each of these 18 sRNAs have small RNA sequencing reads and 15 have at least one target to rRNA or tRNA. Within a family it is common for both guide regions to exhibit a high degree of sequence similarity indicative of a common ancestry. For example, of the 70 families that have a representative sRNA form each of the six species, 62 exhibit a recognizable degree of sequence similarity (>70%) in both the D and the D′
guide regions among all members whereas the remaining eight families have a conserved
sequence across all species in only one of the two guide regions (see URL list [114] for
numerous examples). Even when a particular guide region is conserved, it is frequently
punctuated by nucleotide insertions or deletions or by nucleotide substitutions primarily
at the 5′ or 3′ end of the guide that are less likely to impact the guide-target interaction.
Even with the high degree of guide sequence similarity, not a single guide is perfectly
conserved in any of the 70 families with representatives in all six species.

The insertions, deletions, or mutations between guide sequences within a ho-
mologous C/D box sRNA family may not co-vary with the predicted target. As previ-
ously mentioned, in some cases the differences are at the termini of the guide and do not
participate in the guide-target interaction. In other cases, the changes will change the
predicted target and usually shift the target a few nucleotides. These types of changes
likely lead to the creation of new sRNA families (see Section 3.5.6.3 on C/D box sRNA
super families).

3.5.2 New covariance model can predict archaeal C/D box sRNAs
with high sensitivity and specificity

Archaeal C/D box sRNAs are known to be under-annotated [69], indicating
that better computational models are needed. Some issues with existing computational
prediction methods are that they are not specific to archaea [15, 116] or they require a
predicted target to rRNA or tRNA [115]. Filtering out sRNAs that do not have plausible
targets reduces the number of false positives, but sRNAs that do not have targets
(orphan guides) will not be reported. Current C/D box sRNA prediction methods also do not model the kink-turn and only model box-feature sequence variation, guide length, loop length, and the stem that sometimes forms at the 3’ and 5’ termini of the C/D box sRNA \cite{115}. Both snoScan \cite{115} and snoSeeker \cite{205} use probabilistic models and hidden Markov models to predict the box features. SnoReport \cite{83} uses support vector machines and RNA secondary structure prediction, but it requires that the region from the start of the C box to the end of the D box be unpaired. Even if the secondary structure prediction method used could predict kink-turns, SnoReport loses the information carried in the kink-turn between the C and D boxes with the required unpaired region. Finally, Rfam, the database of RNA families represented by covariance models, can be used to search for C/D box RNAs. However, none of the C/D box models incorporate the kink-turn, and many are specific to eukaryotes \cite{116}. A search using these models did not find many more C/D box RNA candidates in Archaea \cite{69}.

In the previous study where Bernick generated small RNAseq data for four of the *Pyrobaculum*, Bernick and I reviewed unannotated transcripts and discovered that several of them were plausible conserved C/D box sRNAs whose box features (specifically the K-turn motif formed by box base pairing) were divergent from the canonical C/D box model \cite{18}. These divergent features likely resulted in these sRNAs being missed in prior computational annotations.

To create a general method that can predict orphan guides and model C/D box sRNA structure, Uzilov and I developed a covariance model that models box sequences, K-turn and K-loop structure, and length of spacers (guides and variable loop) (Figure
3.4). Uzilov created the initial model and I refined the model by iterative development. I converted the model from Infernal v1.0 format to Infernal v1.1 and tested differences in the output between the different versions of Infernal. The length of the spacers in the covariance model is based on the longest observed length in the training set. I trained the final model on a structural alignment of all known Pae sRNAs. Since all of the C/D box RNAs used in the model are conserved or have small RNA sequencing data confirming their transcription, I know that this is a high-quality model that most likely does not have spurious members.

I classified predictions as true positives based on a combination of small RNA sequencing data, bitscores, comparative genomics, and promoter prediction. Predictions were automatically classified as true positives if they were spanned by small RNA sequencing reads and had a bitscore ≥5. Otherwise, predictions were still classified as positives if they had box features with strong kink-turns (purine cap above a uridine and G:A or G:G base pairs in the 1b1n and 2b2n positions), conservation in at least one other species and a good predicted promoter. If a prediction was conserved with an sRNA with small RNA sequencing data, the requirement on box features was relaxed (no purine cap allowed or unusual base pairs in the 1b1n and 2b2n position of the kink-turn).

The predictions generated from the final covariance model indicate that (1) removing predictions that have >80% overlap with mRNAs and overlap with known non-coding RNAs reduces the number of false positives, on average, by 88% and (2) thresholds can be used with the model to evaluate de novo predictions across different species.
Figure 3.4: **C/D box RNA secondary structural alignment used to train the covariance model.** (A) Part of the secondary structural alignment used as input to **cmBuild** to create a covariance model. This alignment is input in the form of a Stockholm file. The **SS_cons** line annotates base pairs, where open and close parentheses indicate a base pair and periods indicate unpaired regions. The **RF** line indicates insertion positions with periods; these positions are not required to be in outputted predictions. Positions marked with ‘x’ are required to be in predictions, and include the box features and minimum guide length in this model. Note that the nucleotide one position 5’ of the D and D’ box is included in the annotated structure. C and C’ boxes are highlighted in yellow and D and D’ boxes are highlighted in blue. (B) Graphical representation of the covariance model.
The average reduction of false positives is not only true of the Pyrobaculum species, but also for species from the Pyrococcus genus, a member of the Euryarchaeaa phylum. Thus, the additional filter of overlap with other annotated genes will also likely reduce the number of false positives in other archaeal species.

Different bitscore thresholds can be used depending on the sensitivity and specificity desired and if predictions are removed based on overlap with mRNAs and other non-coding RNAs. A bitscore threshold of 17 results in a sensitivity of 68–80% and a specificity of 98–100% for the Pyrobaculum species (Figure 3.5A). In comparison, the original C/D box sRNA annotation of Pae had a sensitivity of 72% [64]. Decreasing the threshold to 8 increases the sensitivity to 86–99%, but the specificity drops to approximately 55%. To reduce the number of false positives (and improve specificity) for lower bitscore thresholds, predictions that overlap annotated non-coding RNAs and overlap annotated mRNAs by more than 80% can be removed (Figure 3.5B). With filtering, and at a bitscore threshold of 13, sensitivities >80% are achieved with specificities between 72-92% for the Pyrobaculum species. Lower thresholds can be used to increase sensitivity, but manual curation may be required to evaluate lower-scoring predictions.

To test how well this model would work on less closely related species to Pae, I used it to search three species in the euryarchaeal genus Pyroccoccus: P. abyssi (Pab), P. furiosus (Pfu), and P. horikoshii (Pho). The Pyroccoccus genus has been a model for C/D box sRNA study and has many C/D box sRNA predictions [71] [144]. I was surprised to find nine new C/D box sRNA among these species (two in Pab, six in Pfu, and one in Pho). The model has better specificity in the Pyroccoccus than in
Figure 3.5: **Sensitivity and Specificity of *Pyrobaculum* C/D box sRNA covariance model.** True positives were determined based on combination of small RNA sequencing data, bitscores, comparative genomics, and promoter prediction. These criteria are described in more detail in Section 3.5.2. (A, C) Histograms of true positives and false positives predicted in the *Pyrobaculum* species and *Pyrococcus* species by the covariance model. (B, D) Histograms of true positives and false positives after filtering predictions that overlap with non-coding RNA annotations and Genbank RefSeq genes by more than 80%. (E) ROC curves from scanning six *Pyrobaculum* and three *Pyrococcus* species, with and without filtering results that overlap with non-coding RNAs and Genbank RefSeq genes by ≥80%. Plots made with the pROC package [161].
the *Pyrobaculum* (Figure 3.5C-E). Without filtering for overlap with mRNAs and non-coding RNAs, the AUC for *Pyrococcus* species is 0.936 and for *Pyrobaculum* species is 0.897 (Figure 3.5E). Upon closer inspection, I found that the *Pyrococcus* C/D box sRNAs have much less variation in their box sequences and more canonical K-turns compared to the *Pyrobaculum* (see Section 3.5.3.1 for more discussion). The larger variation in the *Pyrobaculum* may make this model ideal for scanning other Archaea.

Approximately 3–11% of known C/D box sRNAs in the *Pyrobaculum* and *Pyrococcus* are not predicted with this covariance model. Even with a bitscore threshold as low as -50, these sRNAs are not predicted. We examined the false negatives and found that in most cases one or more box features were unusual and often did not have canonical K-turns (see Table 3.1 for examples). For example, the sR42 family has members in all six *Pyrobaculum* species, but Pae sR42 has an unusual D’ box of GCAA and C’ box of AUGGCGU. Not only do these box sequences diverge from the consensus (CUGA for D box and RUGAUGA for C box), they also create kink-turns that do not have the canonical GA/AG base pairs. To capture these unusual C/D box sRNAs, another model may be needed. Adding some of these missed sRNAs to the model reduces the sensitivity of the model for sRNAs with more canonical features.

These sRNAs were originally picked up using Infernal v1.0, which is known to be more sensitive than Infernal v1.1 [133]. I found that Infernal v1.1 tends to miss one to three candidates per species that were predicted by Infernal v1.0. However, using Infernal v1.1 reduces the number of false positives by a hundred-fold and reduces search time. Thus for comprehensive scans, Infernal v1.0 is better, but for casual scans, Infernal v1.1
sRNA C box D′ guide D′ box Loop C′ box D guide D box
Par sR15 GUGAUGA GCCUCACGGGUAG GCGA GAG GUCUGCA CAACGCCCUUCA CCGA
Pae sR42 AUGAGGA GCCGAUGGCUGGCC GCA CUC AUUGCGU UAAAGCUUGGCGGGUU CUGA
Pca sR116 GUGG CUG AAAGGGUUUGGCGG CAGA GCC GUGAAGA AAAGCCCUACACG AGGA

Table 3.1: **Examples of C/D box sRNAs not predicted by the covariance model and Infernal v1.1.** All of these examples are conserved in five or more *Pyrobaculum* species. These C/D box sRNAs likely were not picked up by the model because they have unusual box sequences that do not form canonical kink-turns. Nucleotides described are underlined. Typically the 1b1n and 2b2n positions of the k-motifs formed by the C and D and C′ and D′ boxes are G:A and A:G, respectively. The D′ and C′ boxes of Par sR15 have C:A and U:G base pair for the 1b1n and 2b2n positions, respectively. The D′ and C′ boxes of Pae sR42 has A:G base pair for the 1b1n position of the K-loop, which is usually G:A. Pca sR116 has a G:G base pair for the 2b2n position of the kink-turn formed between the C and D boxes. Par sR15 and Pca sR116 were predicted by Infernal v1.0 but not v1.1, so for comprehensive scans Infernal v1.0 should be used. Although it was not picked up by the covariance models, Pae sR42 was discovered using comparative genomics.

is better. These differences appear to be the result of an algorithmic change and are discussed in Section 2.3.3.

**3.5.3 Structural features of *Pyrobaculum* C/D box sRNAs**

**3.5.3.1 Large sequence variation in *Pyrobaculum* box features compared to those in *Pyrococcus***

The sequence conservation of the C, D, C′, and D′ boxes are due to the formation of the K-turn and K-loop motifs [137] [145]. Past studies show that the formation of intact K-turn and K-loop structural motifs by the respective C/D and C′/D′ box sequences are necessary for efficient methylation of a target ribonucleotide [185] [145]; methylation activity of the D and D′ guides is reduced *in vitro* when either the K-turn or K-loop is destroyed by mutating box features [185], although this decrease may be modest [145]. The K-turn motif appears to be more important for the formation of the
RNP and efficient methylation than the K-loop. If at least one K-motif is intact, the RNP complex will still form because of strong protein-protein interactions, but methylation activity is generally reduced [185]. In addition, normal methylation levels can still occur even if L7Ae protein is mutated so it can no longer bind to the K-loop [67].

The formation of the K-turn and K-loop is also crucial for specificity of the modification site [80].

In the study of C/D box sRNAs in three *Pyrococcus* species, Gaspin et al. noted that the C, D, C', and D' boxes were highly homogenous in sequence [71]. In the *Pyrobaculum*, the opposite appears to be true, with the box sequences being more variable, resulting a larger number of base pair types that are formed in the K-turn and K-loop (Figure 3.6B and C). At all positions of the K-turn and K-loop, the *Pyrobaculum* C/D box sRNAs have more base pair types represented (out of a possible 16). The 1b1n and 2b2n positions that form the critical G:A base pairs of the kink-turn have less variation than the other positions, but still have more variation in *Pyrobaculum* species than in *Pyrococcus* species. The most dramatic difference between the two genera is in the 3b3n-5b5n positions of the K-loop (Figure 3.6B and C).

Suspecting that there may be unannotated sRNAs in the *Pyrococcus* species that could account for the differences in observed sequence frequencies, I used comparative genomics to look for additional sRNAs, but only found one new sRNA in each species that was not predicted by the covariance model. These findings did not have a large effect on the sequence analysis of the box features. Based on conservation of the sRNAs in both genera, I believe that the box features are correct, although in the
Figure 3.6: **Comparison of Pyrobaculum and Pyrococcus C/D box sRNA K-turns and K-loops.** (A) Diagram of K-turn and K-loop positions. In K-turn studies, the base pairs are numbered starting with the G:A base pairs and the first letter in the pair is from the C box \(^{170, 113}\). The strand with the bulge and the C box is referred to as ‘b,’ and the non-bulged strand with the D box is referred to as ‘n.’ Thus the first two positions are typically position 1b1n being G:A and position 2b2n being A:G. (B) Sequence logos of the box features of the Pyrobaculum and Pyrococcus genera. Created with RILogo \(^{125}\). (C) Representation of the variability of base pair type found at each position of the K-turn or K-loop. Darker boxes indicate that position is more variable.
Figure 3.7: Nucleotide frequencies of the position 5′ of the D and D′ boxes (position 5n).

*Pyrobaculum*, a few of the sRNAs have ambiguous D boxes. Both genera have a large number of sRNAs compared to other archaeal species, so these differences aren’t due to the large proliferation of C/D box sRNAs in hyperthermophiles exploring sequence space, but likely due to differences in L7Ae binding specificity between species.

### 3.5.3.2 Consensus of D and D′ boxes as GCUGA

When Uzilov and I created the covariance model, we hypothesized that the nucleotide 5′ of the D and D′ box also participated in the structure of the C/D box sRNA (position 5n in K-turn nomenclature), so we included it as part of the sequence and structure in the model (Figure 3.4). To see if the 5n nucleotide provided extra information in the model, I analyzed nucleotide frequencies of the annotated C/D box sRNAs at that position. I found that guanine is preferred at this position 44–62% of the time, in relation to the D and D′ box and in both the *Pyrobaculum* and *Pyrococcus* species (Figure 3.7). Adenine occurs approximately 17% of the time, cytosine 4–24%, and uridine 15–21% of the time. This preference suggests that the consensus sequence for the D and D′ box should be extended.
The 5n position is considered to be part of the sRNA guide, so it is possible that the preference is a result of its target RNA composition, rather than a result of its role in K-motifs of the C/D box sRNAs. However, the nucleotide frequencies in ribosomal and transfer RNA, the main targets of C/D box sRNAs, do not support this hypothesis. In \textit{Pyrobaculum} species in this study, the approximate frequencies of nucleotides in rRNA are 20% A, 30% C, 37.5% G, and 12.5% U and in tRNA the approximate frequencies are 14.5% A, 34% C, 36.5% G, and 15% U. If the 5n position was influenced solely by targeting ribosomal and transfer RNA, we would expect the nucleotide frequencies to show a preference for C and then a slightly lower preference for G. Thus, it is more likely that the 5n position plays a role in the structure of archaeal C/D box sRNAs and should be included in computational models.

3.5.3.3 Guide length and symmetry

The features of \textit{Pyrobaculum} C/D box sRNA guides are generally similar to the ones found in C/D box sRNAs of other archaeal species. The sRNAs are approximately 50 nts in length and two guides designated the D guide and D’ guide are located between the C’ and D, and between the C and D’ boxes, respectively (Figure 3.1). The guides are on average about 12 ± 1.5 nucleotides (nts) long (Figure 3.8).

Of the \textit{Pyrobaculum} sRNAs, 90.5% of D and D’ guide lengths are within 2nt of each other. However, some sRNAs with D and D’ length differences greater than 2nt exist, such as the sR26 family (Figure 3.8B). The five core members of the sR26 family have fairly canonical box features and conserved guides with targets in tRNA^{Trp}.

47
Figure 3.8: **Features of *Pyrobaculum* C/D box sRNAs guides** (A) Distribution of guide lengths. The length is defined as the number of nucleotides separating the end of the C (or C') box from the beginning of the D (or D') box. The length distributions of sRNA guides for each of the six of *Pyrobaculum* species are depicted. (B) Distribution of symmetry of guide lengths. The difference in length between the D and D' guides within each sRNA is depicted.

but the D' guide is 16–17nt and the D guide is 12nt (see Appendix A.1.3 for sR26 member annotations and Appendix A.1.4 for predicted targets). The short D guides are predicted to target position C34 or C35 in the anticodon loop, whereas the D' guides are less conserved and have a less certain target at position C70 in the acceptor stem. Some guides that are longer or shorter than 12nt have decreased methylation activity *in vitro* [184], so it is possible that the length of the sR26 family D' guide may interfere with methylation function. Since at least some members of the sR26 family appear to be double-guide sRNAs, it is possible that the D' guide functions not in methylation, but
rather to assist in the folding the tRNA, or to provide additional target RNA binding stability in order to increase methylation efficiency of the D guide target [189]. These sRNAs with extremes in guide length variation that we observe in *Pyrobaculum* may have escaped detection in other archaeal species, where sRNA annotation has been less comprehensive.

3.5.4 Prediction of methylation targets in ribosomal RNA and transfer RNAs

To predict methylation targets in rRNA and tRNA, Dennis and I used the “N+five” rule [101] (Figure 3.1) and ranked hits based on extended complementarity between guide sequences and target RNAs (Appendix Appendix A.1.4). Using the criteria described in Section 3.4.3, we were able to predict targets for nearly 75% (767/1052) of the sRNA guides, and we found that 89% (468/526) of the sRNAs had predicted targets for one or both of the guides.

The rRNA predictions were mapped onto 16S and 23S rRNA sequence alignments (Appendix A.1.5). In 16S rRNA there are 60 different positions of modification/interaction of which 40 occur in all 6 species, 7 occur in 2–5 of the species, and 9 are unique and occur in only a single species. In 23S rRNA there are 114 different positions of modification of which 54 are modified in all 6 species, 34 are modified in 2–5 species, and 22 are unique and modified in only a single species. The absence of modification at a particular position that is modified in another species is either the consequence of the absence of an sRNA homolog or sequence divergence of the guide within
the sRNA family. Finally, 235 of the 526 sRNAs use their D and D’ guides to target sites that are within 100 nts of each other in the primary 16S or 23S rRNA sequences (Appendix A.1). Other studies have suggested that this dual interaction at two closely positioned sites plays an important role in mediating the folding and stabilization of the nascent rRNAs and their assembly onto ribosomal subunits \[208\].

### 3.5.4.1 Instances of mismatched base pairs at the “N+five” position

In several instances (two positions in 16S and five positions in 23S) Dennis and I found a mismatch at the “N+five” position in the region of guide-target complementarity (Appendix A.1.4). *In vivo* and *in vitro* studies have demonstrated that a Watson-Crick base pair at this position is essential for methylation of the target RNA \[143, 33, 6\]. Nonetheless, a mismatch at the site of methylation within a conserved region of guide-target complementarity implies that the interaction may be beneficial but that the modification is either not needed or harmful to the function of the target RNA. We have found cases where there is a mismatch in only one member of the family and other cases where there is a conserved mismatch among all of the members of the family.

Mismatches in one member of a family occur in the sR09, sR44, and sR106 families. These families have at least five members, and the D and D’ guides are highly conserved. The Par sR09 D guide has a mismatch at the “N+5” position with its target (Figure 3.9A), and the Pae sR44 and the Pis sR106 D’ guides also have mismatches at the critical position. For the rest of discussion I focus on the sR09 family as an example. Both Par sR09 guides have complementarity to 23S rRNA and the D guide
and D′ guides to target positions U912 in helix 32 and U879 in helix 35, respectively (Figure 3.9B). I suspect that these guide-target interactions play an important role in the localized folding of the 23S rRNA in this region. The Par sR09 D guide contains an A-to-U nucleotide substitution at the critical “N+five” position, changing the guide-target interaction to U:U at this position. Based on in vivo and in vitro studies demonstrating a requirement for a canonical base pair at “N+five” position [143, 33, 6], I predict that Par sR09 base pairs with both 23S rRNA targets, but that the D guide interaction does not result in modification of U912.

Four sRNA families have at least four members with mismatches at the site of methylation: sR25, sR33, sR56, and sR116. All of these families are highly conserved. As an example, I discuss the sR33 family. The sR33 family is conserved in all six Pyrobaculum species, and the D and D′ guides exhibit complementarity to two closely spaced 23S rRNA sequences (Figure 3.10A,B). The D guide is predicted to direct methylation to position G2021 in helix 68 whereas the D′ guide is predicted to be incapable of methylation at C2045 located in helix 69 because of a C:U mismatch at the “N+five” position. I consider the D′ guide-target interaction credible because of its strong conservation and its close proximity to the D guide interaction. The fact that the D′ guide interaction is conserved without the potential for methyl modification highlights the potential role of sRNAs as chaperones for productive and efficient folding of the rRNA during the ribosome assembly process [116, 193]. The proper folding of helix 69 is critical because of its interaction with the anticodon stems of A site and P site tRNAs during protein synthesis [173].
Figure 3.9: Example of a *Pyrobaculum* sRNA family where one member has a mismatch base pair at the “N+five” position. Sequence alignment of sRNAs in the sR09 family (A) is presented and the guide complementarities to 23S rRNA are indicated below. The critical “N+five” nucleotide in the guide regions is highlighted in blue when there is a Watson-Crick base pair between the guide and target and in rose at a mismatch base pair in Par. The yellow highlight represents the “N+five” position in the rRNA target. (B) The secondary structure of helices 32–35 in 23S rRNA is depicted showing the complementarity of the sR09 D guide near position U912 and the D’ guide near position U879. Par sR09 has a mismatch at position U912 at the “N+five position” which may result in that nucleotide remaining unmethylated. Secondary structure generated by SSU-ALIGN package [134].
Figure 3.10: *Pyrobaculum* sRNA family that has a mismatch base pair at the “N+five” position. Sequence alignment of sRNAs in the sR33 family (A) is presented and the guide complementarities to 23S rRNA are indicated below. The critical “N+five” nucleotide in the guide regions is highlighted in blue when there is a Watson-Crick base pair between the guide and target and in rose when there is a mismatch base pair. The yellow highlight represents the “N+five” position in the rRNA target. (B) The secondary structure of helices 68 and 69 in 23S rRNA is depicted showing the complementarity of the sR09 D guide near position G2021 and the D’ guide near position C2045. All six sR33 members have a mismatch at position 2045 at the “N+five” position in the guide-target interaction. Secondary structure generated by SSU-ALIGN package [134].
3.5.5 Methylation sites cluster within functionally important regions of rRNA.

Dennis and I mapped the positions of predicted methyl modification on the 16S and 23S rRNA secondary structures in order to visualize their clustering within functionally important regions. (Figures 3.11 and 3.12). In general, functionally important regions contain a high density of modifications whereas less important regions contain a lower density of modifications. Comparisons with positions of predicted modification in species outside of Pyrobaculum indicate that the precise sites of modification are, with a few notable exceptions, generally not conserved although the clustering pattern is conserved (Dennis et al, in preparation).

3.5.5.1 Long range interactions between sRNAs and rRNAs

Several sRNAs have D and D′ guides that have complementarities and predicted methylation targets that are more than 100 nts apart in the primary rRNA sequence but are close in the secondary structure. I suspect that these long-range interactions play an important role in the folding of rRNA during the assembly process.

In 16S rRNA, the D guide of sR53 is complementary to the loop region of helix 18 and is predicted to methylate A509 (Figure 3.13A). This entire stem loop 18 has been implicated in translational fidelity. The G507 (G530 in E coli 16S rRNA) is intimately associated with the interaction of the A site tRNA anticodon with the mRNA codon; site directed mutations at this position are lethal [153]. Other mutations in this region affect translational fidelity and resistance to the antibiotic streptomycin. The D′
Figure 3.11: **Mapping of the predicted sites of sRNA directed methyl modification onto the conserved secondary structure of archaeal 16S rRNA.** The predicted sites of methyl modification by each sRNA family (circled in red) were mapped onto the secondary of Pca 16S rRNA. The family number followed by the number of species that modify at that site are indicated. For example 19-6 at position U15 is predicted to be modified by all six members of the sR019 family. An * indicates that there is a mismatch between the guide and target at the predicted site of methylation. Secondary structure created by the SSU-ALIGN package [134] and derived from CRW database (http://www.rna.ccbb.utexas.edu).
Figure 3.12: Mapping of the predicted sites of sRNA directed methyl modification onto the conserved secondary structure of archaeal 23S rRNA. The predicted sites of methyl modification by each sRNA family (circled in red) were mapped onto the secondary of Pca 23S rRNA. Labeling of predicted sites of modification is as described in the legend to Figure 3.11. An * indicates that there is a mismatch between the guide and target at the predicted site of methylation. Secondary structure created by the SSU-ALIGN package [13] and derived from CRW database (http://www.rna.ccb.utexas.edu).
guide of sR53 has two separate complementarities to 16S sequences and is predicted to methylate at position C514 in the 3′ strand of helix 3 as well as at position C34 in the 5′ strand of helix 4. There are eight additional predicted sites of methyl modification in this region that are mediated by other *Pyrobaculum* sRNA families. It is unclear how these multiple sR53 guide interactions might occur within the nascent rRNA transcript and how they impact the folding and structural stability of the translational fidelity stem-loop.

Two other three-way interactions have been identified around the helix 26-27 junction in 23S rRNA (Figure 3.13B) and in the core region around helix 28 that serves as the connection point for the four domains in 16S rRNA (Figure 3.13C). In the first instance sR2 is predicted to use its D guide to modify position G655 and its D′ guide to modify both positions C667 and C781 in 23S rRNA. The second instance involving the core region in 16S includes helix 2, a complex pseudoknot that forms between the loop of helix 1 and the connector region between domain 2, and the core helix 28. The D guide of sR56 is predicted position U877 immediately 5′ to the helix 2 pseudoknot structure. The D′ guide is predicted to modify both positions G908 and G1337 in 16S rRNA. These interactions likely facilitate the complex folding events that arrange the four 16S domains around the central core helix 28.

### 3.5.5.2 One-third of *Pyrobaculum* sRNAs target tRNAs

It has been shown previously that archaeal C/D box RNA can target modification to tRNAs as well as rRNAs [51]. The tRNA methylation targets are at structurally
Figure 3.13: *Pyrobaculum* sRNAs that exhibit alternative long range guide-target interactions with rRNA. (A) Secondary structure of the translational fidelity region of Pca 16S rRNA. The D guide of sR53 exhibits complementarity near position A493. The D' guide exhibits complementarity to two different regions, one centered on C514 and the second on C34. Asterisks indicate the eight other predicted sites of sRNA-guided methyl modification in this region. (B) Secondary structure near helicies 26-27 in Pca 23S rRNA. The D guide of sR2 exhibits complementarity near position G655. The D' guide exhibits complementarity to two different regions, one centered on C677 in the 5' strand of helix 26 and the second on C781 in the 3' strand of helix 67. The asterisk indicates the other predicted site of methyl modification in this region. (C) Secondary structure of the connector helix 28 that serves as the hub for the four domains of Pca 16S rRNA. The D guide of sR56 exhibits complementarity near position U877 in helix 27. The D' guide exhibits complementarity to two different regions, one centered on G908 in helix 29 and the second on 1337 in helix 43. Asterisks indicate the eight other predicted sites of methyl modification in this region.
conserved positions that are modified by tRNA methylases in other organisms. In the collection of 110 *Pyrobaculum* sRNA families that I annotated, 32 are predicted to target modification to 23 different positions in tRNAs (Appendix A.1.4).

The number of different tRNAs that can be targeted by a particular sRNA guide varies over a wide range and reflects the fact that some sequences in tRNAs are unique whereas others are shared among many different tRNA isoacceptors. The region surrounding position C34, the wobble base in the anticodon, is an example of a variable sequence. Guides from four different sRNAs target position C34 or U34, and each has only a single tRNA target (sR26:C34Trp, sR27:U34Gln, sR45:C34Val, sR46:U34Thr and sR51:C24Glu). Other guides have multiple tRNA targets; for example, the D guide of Pae sR64 exhibits complementarity to a conserved sequence in sixteen different tRNA families and directs modification to position G51 in the TΨC stem.

### 3.5.5.3 Guides with no predicted targets in tRNA or rRNA (orphan guides)

In the 526 different *Pyrobaculum* sRNAs (representing 1052 guides) that I have identified, there are 285 guides (27%) that show no significant complementarity to either rRNA or tRNA sequences (see Appendix A.1). Andrea Corredor and I searched for mRNA targets for these orphan guides, but no significant and conserved complementarities were observed. There are a number of instances within an sRNA family where one guide target-interaction is conserved in all members and where the second guide-target interaction in one or more family members is disrupted by nucleotide substitutions in the guide sequence. For example, in the sR30 family, the D guide of all six members
is predicted to target C2724 in 23S rRNA, whereas the D′ guide is predicted to target C2708 in only four of the members. The two sRNAs containing the disrupted D′ guides occur in Pog and Par, a sub-lineage within the *Pyrobaculum* genus. In addition to mutational drift, genomic arrangements that result in an overlap between an sRNA gene and a protein-coding gene could potentially disrupt guide function. Of the sRNAs that overlap the 5′ or 3′ end of a protein-coding gene in the sense orientation, 88% and 61% of the respective overlapping guides do not have predicted targets in rRNA or tRNA (see Section 3.5.7 for more discussion).

There are only three families with six members (sR43, sR50 and sR108) where neither the D nor the D′ guides have significant complementarity to rRNA or tRNA. In the sR43 family both guides are highly conserved among members and would be expected to recognize the same target sequence whereas in the other two families, the guide sequences are only moderately conserved and would likely not all recognize identical the same target sequence. The lack of conservation for the sR50 family may be partially explained by overlap with the promoter of a nearby protein-coding gene.

3.5.6 Proliferation, mobility, plasticity and evolutionary divergence of C/D box sRNA genes within the *Pyrobaculum* genus

Grouping the *Pyrobaculum* C/D box sRNAs into homologous families greatly facilitated my analysis of target conservation and the origins and evolution of these sRNAs. Here I describe sequence similarity of guide sequences between different homologous sRNA families.
3.5.6.1 Composite and transposed sRNAs

Of the 18 sRNA single-member families, five have a guide that shares some resemblance to a guide in a different sRNA family. These are designated as either transposed or composite sRNAs (Table 3.2). Transposed sRNAs share one guide with another defining family, but typically the guide has been transposed from D to D′ or visa versa, from D′ to D position, compared to the defining family. Composite sRNAs have D and D′ guides that each match one of the guides in two different families. For example, Pca sR12/45 has a D guide that is similar to the D guide of the sR12 family and a D′ guide that is similar to the D′ guide of the sR45 family (Figure 3.14). I suggest that the genes encoding these composite and transposed sRNAs are generated by genetic rearrangements between different sRNAs or sRNA genes.

3.5.6.2 Duplication of sRNAs

Duplication of a full-length sRNA gene can also occur as evidenced by the highly similar Pae sR113a and 113b (Figure 3.14A). No other species have members in this sRNA family. The 5′ flanking regions in front of the two genes are unrelated. In contrast, the 3′ flanking regions are identical for 14 bp with sequence similarity extending a further 30 bp. The sequence 3′ to the sR113a gene encodes sR08 on the opposite strand, while the sequence 3′ to the sR113b gene contains what appears to be the remnant of the sRNA gene that has been obliterated by the presence of ORF3005. The sR113a and sR08 genes are convergently transcribed and separated by a 1 bp intergenic space.
<table>
<thead>
<tr>
<th>Composite C/D box sRNA</th>
<th>D' guide</th>
<th>D guide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pca sR12/45</td>
<td>Shared with D guide of sR45 family</td>
<td>Shared with D guide of sR12 family</td>
</tr>
<tr>
<td>Pca sR103/109</td>
<td>Shared with D' guide of sR103 family</td>
<td>Shared with D guide of sR109 family</td>
</tr>
<tr>
<td>Transposed C/D box sRNA</td>
<td>D' guide</td>
<td>D guide</td>
</tr>
<tr>
<td>Pca sR26a</td>
<td>Shared with D guide of sR26 family</td>
<td>Not shared</td>
</tr>
<tr>
<td>Par and Pog sR13a</td>
<td>Shared with D guide of sR13 family</td>
<td>Not shared</td>
</tr>
<tr>
<td>Pca sR57a</td>
<td>Shared with D guide of sR57 family</td>
<td>Not shared</td>
</tr>
<tr>
<td>Pae sR57b</td>
<td>Not shared</td>
<td>Shared with D guide of sR57 family</td>
</tr>
</tbody>
</table>

Table 3.2: **Composite and transposed C/D box sRNAs.** Two unusual types of sRNAs were identified. Composite sRNAs have D guide that shows sequence similarity to a guide in one sRNA family and D' guide that show sequence similarity to a guide in another sRNA family. These are given both family numbers, with the smallest family number first, separated by a forward slash (/). Transposed sRNAs have either a D guide that is shared with the D' guide of the defining family or visa versa, a D' guide that is shared with the D guide of the defining family. Transposed sRNAs are identified with the number of the defining family followed by a lower case a or b. The Pae sR57b is considered as a transposed sRNA since the D' guide normally associated with the sR57 is not present.

Other examples of duplication are of Pog sR46 and Pae sR62. The sR46 family has members in all six species. Pog sR46a is a duplication of Pog sR46 and has a nearly identical D' guide and a D guide with 3 substitutions (Figure 3.14B). Another apparent duplication involves Pae sR62 (Figure 3.14C). This gene is located at position 2104084-4133 on the chromosome. A highly similar sequence presumably representing an sR62 pseudogene remnant occurs at position 2101402-1469.
Figure 3.14: Examples of C/D box sRNA duplication within a species. (A) Duplication of the sR113 gene in Pae. The sR113 family has two members and only occurs in Pae. The sequence similarity between the two nearly identical Pae sR113a (red) and sR113b (blue) genes and their 3′ flanking regions is illustrated (grey highlight). The 3′ flanking regions of the sR113a gene contains the sR08 gene (green) that is transcribed on the opposite strand and is separated from sR113a by a single nucleotide. The 3′ flanking region of the sR113b gene contains a remnant of an sR08-like gene that is partially buried in the PAE 3005 protein ORF that is separated from the sR113b gene by 14 nucleotides. There is no sequence similarity between the 5′ flanking regions of the sR113a and 113B genes. (B) Duplication of sR46 in Pog. Two nearly identical sRNAs that belong to the sR46 family occur in Pog. Grey highlight indicates the conserved guide regions of the sR46 family. (C) The Pae chromosome contains imperfect duplicate sequences that are separated by 2Mbps. Both copies retain a promoter-like sequence (blue) that is likely used to drive expression of the Pae sR62 gene (red). The first sequence contains an apparent remnant of the sR62 gene as suggested by the regions of sequence identity (grey highlight).
3.5.6.3 Super families of sRNAs

The tracking of sequence similarity between guides from different families can reveal ancient origins and evolutionary relationships between the homologous sRNA gene families. We have uncovered evidence suggesting that the sR45, 12, 56 and 57 families share a complex evolutionary history (Figure 3.15). The sR56 and sR57 families appear to represent an ancient duplication appearing early within the *Pyrobaculum* lineage. Only a homolog of sR56 appears in the closely related species, *Thermoproteus tenax* (Tte). Each family has representatives in all six *Pyrobaculum* species and one of the families (sR57) is a circular permutation of the other (sR56). The D and D’ guides of sR56 target modifications to 16S U877 and G908 respectively and the D and D’ guides of sR57 target modifications to 16S G906 and A879 respectively. The shared core sequence between the D guide of sR57 and the D’ guide of sR56 is UUCACC and the shared core sequence between the D guide of sR56 and the D’ guide of sR57 is AUUUCCU. These cores sequences are offset by two nucleotides due to indels within the respective guides and this accounts for the two nt shift in target specificities. The two aberrant (transposed) members of the sR57 family (Pae sR57a and Pca sR57b) are circular permutations of each other and share only the single guide UC-CC-CUU with the D guide of the core sR57 family.

The sR12 family is also implicated in this complex interconnection of families. It has a D’ guide that exhibits sequence similarity to the D’ guide of the sR56 family (CU-UC-CCUC). Indels in the sR12 D’ guide changes the target specificity to position

64
Figure 3.15: Proliferation and evolutionary interconnections between sRNA genes. Interconnected guide sequence similarity between different families of sRNAs. The colored sequences (green, magenta, blue and orange) indicate different sequence similarities in the guide regions of sRNAs of the interconnected sRNA families - sR45, sR12/45, sR56, and sR57. The sRNA in the outgroup species Thermoproteus tenax is related to the sR57 family. The relationships between these sRNA families illustrate how insertions, deletions, and rearrangements can contribute to the evolution of archaeal C/D box sRNAs.
As mentioned above, the D guide in the sR12 family is shared with the D guide of the composite sR12/45. The second D′ guide of sR12/45 is derived from the sR45 family; this guide is predicted to target methylation to position C34 in the anticodon loop of tRNA<sub>Val</sub>.

The relationships between these four related families illustrate several important aspects of sRNA gene evolution including: (i) gene duplication; (ii) altered or abolished guide-target interactions by guide migration (resulting from insertion/deletion) or divergence (resulting from nucleotide substitution); (iii) rearrangements, including guide replacement and/or circular permutation.

### 3.5.6.4 MITE-like elements resembling sRNAs

Many of the families with only one sRNA member occur in Pca (Figure 3.3B). This species exhibits modular duplications and rearrangements between and within sRNA families as evidenced by the transposed and composite sRNAs described above (see Table 3.2). A careful analysis has also revealed the presence of a MITE-like element present in at least 15 copies within the Pca genome (Figure 3.16). MITEs are miniature inverted-repeat transposon elements that are characterized by a combination of terminal inverted repeats and internal sequences too short to encode proteins. These elements are Class II transposons that occur in plants and other archaea [62].

These elements in Pca have characteristics of both sRNAs and MITEs. Each copy contains 5′ C box and 3′ D box sequences and highly degenerate internal D′ and C′ sequences. The guides located between the C and D′ and C′ and D boxes exhibit
<table>
<thead>
<tr>
<th>sR141</th>
<th>C box</th>
<th>D box</th>
<th>C' box</th>
<th>D box</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
</tr>
<tr>
<td>sR131</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
<tr>
<td>sR133</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
<tr>
<td>sR137</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
<tr>
<td>copy1</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
<tr>
<td>copy2</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
<tr>
<td>copy3</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
<tr>
<td>copy4</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
<tr>
<td>copy5</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
<tr>
<td>copy6</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
<tr>
<td>copy7</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
<tr>
<td>copy8</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
<tr>
<td>copy9</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
<tr>
<td>copy10</td>
<td>GUA</td>
<td>GGG GAGG</td>
<td>UACGCUCCG</td>
<td>GGG GAGG</td>
</tr>
</tbody>
</table>

Consensus: GUA | GGG GAGG     | UACGCUCCG     | GGG GAGG     | UACGCUCCG     |

Figure 3.16: **MITE-like element in the Pca genome.** The chromosome of Pca contains fifteen copies of a MITE-like sRNA element. The sequences are aligned to illustrate the high degree of conserved sequence similarity in the 5′ and 3′ flanking inverted repeat sequences (blue highlight). The sRNA-like sequences (yellow highlight) contain canonical C and D boxes but generally degenerate D′ and C′ boxes (boxed in red). The conservation between the D and D′ guide sequences in the 15 elements is moderate with a consensus sequence at the bottom. Five of these elements were cataloged as authentic sRNAs.

modest sequence similarity across the 15 copies. Highly conserved imperfect inverted repeat sequences flank the C and D boxes (Figure 3.16). The elements are located in insertion regions of the Pca genome and have a large average distance (322bp) from the nearest protein-coding gene compared to other sRNAs (22bp). Word count analysis of the regions flanking these MITE-like elements reveal an enrichment for the inverted repeat sequences compared to the rest of the genome.

Five of the element copies were classified as C/D box sRNAs (sR131, sR133, sR137, sR139, sR141) and contain moderately degenerate internal box sequences. One of the sRNAs, sR141, is very highly expressed—13.6% of the uniquely mapped RNASeq reads in Pca are generated from this locus. The other copies have expression levels similar to other sRNAs. The MITE-like sRNAs also usually have a higher percentage...
of antisense reads compared to other Pca sRNAs. On average, 39% of reads from a MITE-like sRNA locus are antisense, while on average 9.3% of reads from other Pca sRNAs are antisense.

These MITE-like sRNAs and their copies do not have homologs in the other Pyrobaculum species, so their proliferation may be due to a transposon-like mechanism since they have MITE characteristics. I suggest that similar MITE elements may play a role in the generation, mobilization and proliferation of archaeal C/D box sRNAs or their modular components since MITEs are known to occur in other archaea, although the MITE-like element in Pca might be a unique occurrence. I did not observed these MITE-like elements in the other five species, but they may exist in lower copy numbers and with less sequence conservation.

3.5.6.5 Genomic association of C/D box sRNA and tRNA genes

Most archaeal C/D box sRNAs are independently transcribed, but in a few cases C/D box sRNA genes are known to be polycistronic [116]. Transcription of archaeal C/D box sRNAs genes with protein-coding genes has been reported in Sulfolobus solfataricus and the Pyrococcus genus [71, 51]; in Nanoarchaeum equitans, a few instances of di-cistronic C/D box sRNA-tRNA transcripts have also been reported [157]. In Pyrobaculum, I find conserved instances of C/D box RNAs co-transcribed with other sRNAs, a tRNA, and protein-coding genes (Figure 3.17 and 3.18).
3.5.6.6 Operons containing sRNA and tRNA genes

My analysis indicates that the transcriptional relationships between sRNA and tRNA genes within the *Pyrobaculum* genus are extremely fluid. Bernick and I identified a novel archaeal transcript in Pae, Pis, and Pca that contains three C/D box sRNAs (sR101, sR21, and sR100). These three genes are polycistronic based on genomic proximity, northern hybridization, and overlapping RNA-Seq reads (Figure 3.17A and B). In Pne, Par, and Pog, there is no homolog of sR100, but sR21 and sR101 are still syntenic. In Par and Pog the two genes are approximately 180 nts from each other and appear to be expressed from separate promoters. In Pis and Pne the sR34 and sR40 genes are also co-transcribed (based on RNA-seq reads) and separated by 10, and -4 nts respectively. In Pne the D box of sR34 is located within the C box of sR40 (four nt overlap); it is unclear how this overlap affects the maturation of the two sRNAs. In Par, Pog, and Pca the genes are separated by 16, 16, and 78 nts respectively and are convergently transcribed whereas in Pae the two genes are separated by more than 2000 nts.

Plant species and the archaeon *Nanoarchaeum equitans* have C/D box sRNA genes that are reported to be co-transcribed with tRNAs (15). In the *Pyrobaculum* genus we find one case of a C/D box sRNA that is likely co-transcribed with elongator tRNA^{Met}. In Pae, Pis, and Pne, the sR44 gene is positioned 8 bp or less from the 3′ end of tRNA^{Met} gene (Figure 3.18). In Par and Pog, the sR44 and tRNA^{Met} genes share the same synteny, but the genes are separated by about 100 nts and their expression
Figure 3.17: **Polycistronic Pyrobaculum C/D box sRNAs.** (A) Genomic organization of the sR101 (blue), sR21 (red) and sR100 (yellow) genes in the six species of *Pyrobaculum*. The 16S rRNA phylogenetic tree with *Thermoproteus tenax* (Tte) as the outgroup, is illustrated on the left; the sRNA gene locations above a bp distance scale is illustrated on the left for the *Pyrobaculum* species. There is no representative of the sR100 gene in Pne, Pog, Par. (B) Northern hybridization by Bernick using RNA extracted from Pae cells with probes to Pae sR21. The position of molecular size markers is indicated in nts on the left and the identity of the four detectable transcripts is indicated on the right. (I) full length polycistron (sR101+sR21+sR100); (II) sR101+sR21; (III) sR21+sR100; (IV) sR21.

appears to be driven from separate promoters. In Pca, the sR44 gene is approximately 13,000 nt downstream of the tRNA\textsuperscript{Met} gene. In Tte, there is no homolog of sR44 and none of its orthologs of tRNA\textsuperscript{Met} are linked to C/D box sRNA genes.

These two examples demonstrate fluidity of C/D box sRNAs genes within the *Pyrobaculum* genus. None of the four sRNAs discussed (sR21, sR100, sR101, and sR44) have homologs in Tte (Figure 3.17B). Within the polycistronic example, the sR100 was lost from the transcription unit in the Par/Pog/Pne lineage and in Pog and Par the
Figure 3.18: **Evidence of sRNAs co-transcribed with tRNAs.** Linkage of tRNA and sRNA genes. In Pae, Pis, and Pne the sR44 genes (green arrows) is located eight nts or less from the 3′ end of a tRNA\textsuperscript{Met} gene (black arrows). In Pog and Par the distance between the tRNA\textsuperscript{Met} and sR44 gene is increased to about 100 nts. In Pca, sR44 is located approximately 13,000 nt downstream of the tRNA\textsuperscript{Met} gene. There is no representative of the sR44 family in Tte.

remaining sR100 and sR101 genes developed individual promoters. Similarly, the sR44 gene appears to have become linked to the tRNA\textsuperscript{Met} gene in the ancestor of Pae, Pis, Pne, Pog, and Par lineage; Pca is an outgroup to these species and does not have the same sRNA-tRNA linkage (Figure 3.18). Separate promoters for the two genes first appear in the Pog/Par sub-lineage.

### 3.5.7 Overlap of C/D box sRNA genes and protein-coding genes

In a previous study from the Lowe lab [18], Bernick and I noted that *Pyrobaculum* C/D box sRNAs, compared to tRNAs, are over 40-fold more likely to have conserved overlap with orthologous protein-coding genes. Other studies have also noted the 3′ antisense overlap of C/D box sRNAs with protein-coding genes [51]. I looked closer at
the relationship between C/D box sRNAs with protein-coding genes since overlap could impact the function of both gene types. In addition, antisense interactions suggest the possibility of C/D box sRNA guided modification of mRNAs.

Of the C/D box sRNAs in this study, I found that 97 exhibit either partial or complete overlap with protein-coding genes (Figure 3.19 and Appendix A.1.6). For this analysis, I considered only overlaps that extend either into the $D'$ guide region (eight nts or more beyond the 5' end of the sRNA gene) or into the D guide region (five nts or more beyond the 3' end of the sRNA gene) since shorter overlaps ending in the D box or C box were not expected to impact target specificity. I classified these overlaps into five categories (Figure 3.19A-E; Appendix A.1.6). I manually checked instances of overlap with the 5' end of an mRNA to confirm that the start codon of the mRNA was called correctly; I adjusted start codons based on conservation if they were annotated incorrectly.

The first and largest category involves sRNA genes that overlap a protein-coding gene and was divided into three subcategories: (i) overlap at the 5' end in the sense orientation; (ii) overlap at the 3' end in the sense orientation and (iii) overlap at the 3' end in the antisense orientation. There were no sRNAs genes that overlapped the 5' end of a protein-coding gene in the antisense orientation. In the first subcategory only two of the 17 overlapping guides (12%) were predicted to have methylation targets in rRNA or tRNA. In contrast, the overlapping guides in the sRNAs at 3' end of the protein-coding gene had numerous predicted targets (40 of 47 for antisense sRNAs guides and 7 of 18 for sense sRNA guides). This disparity suggests that the N-terminal end of
Figure 3.19: **The overlap between sRNA genes and protein-coding genes.** The overlap between sRNA genes and protein-coding genes is divided into five categories (A–E). One category, where an sRNA overlaps the 5' ends of two mRNAs that are antisense to each other, is not included because no sRNAs were found in this category. The protein genes are shown as black arrows with the 5' and 3' polarity indicated. Overlapping sRNA genes are shown in red with polarity indicated by the internal arrows; the C, D', C' and D box sequences are indicated as shown in the top left sRNA. The number of sRNA genes, the number of families that they represent and the number that have predicted targets is indicated for each type of overlap. Details on the amount of overlap are given in Appendix A.1.6.
a protein-coding gene cannot generally be usurped for sRNA guide function. In contrast the 3′ end of protein-coding genes appears more flexible and accommodates in many instances both C-terminal amino acid sequence coding and sRNA guide function.

The high proportion of sRNA located near or overlapping the 3′ end of protein-coding genes may suggest that they play a role in gene regulation and possibly mRNA stability. Sense-strand sRNAs that are co-transcribed with mRNA need to be excised and rescued from decaying mRNA transcripts; sRNAs that are antisense will base pair with the mRNA to form an RNA/RNA duplex that may either stabilize or destabilize the mRNA. We also note in our RNAseq reads that many sRNA genes generate both sense-strand and antisense-strand transcripts. In other archaea, small antisense RNAs have been shown to regulate gene expression by binding to 3′ UTRs (reviewed in [8]). A role for these antisense sRNA transcripts has not been defined.

The second category represents sRNAs that are contained completely within protein-coding genes (Figure 3.19B). Nine of the ten of these are in the antisense category and all have at least one guide that has a target in rRNA or tRNA. These internal sRNAs are located near the 3′ end of the protein-coding gene, again suggesting that this region is flexible and can accommodate both amino acid coding and guide function without detriment.

The last three categories represent sRNA genes that span between two protein-coding genes (Figure 3.19C-E). The Pne sR42 gene is on the sense strand and spans the 4 nt intergenic space between two co-transcribed protein-coding genes. In other species there is a longer intergenic space and the sR42 members overlap only the 3′ end of the
upstream NAD dependent deacetylase gene. In the final two categories, four members of the sR127 family are located at the convergence of two protein-coding genes. In Par and Pog the Uridine phosphorylase gene contains a 3’ extension not found in Pne and Pis that extends through the entire sR127 gene.

### 3.5.8 *Pyrobaculum* C/D box sRNAs overlap potential 3’ UTRs

To explore the relationship between C/D box sRNAs and protein-coding genes further, for each C/D box sRNA I extended the analysis to look at potential overlap with UTRs. For each sRNA, I found the overlapping or closest protein-coding gene and classified the relationship based on the orientation of the C/D box sRNA to the protein-coding gene, the same way I classified strictly overlapping sRNAs (3’ sense, 3’ antisense, 5’ sense, and 5’ antisense). Next, I determined the amount of overlap or the distance to the nearest protein-coding gene. C/D box sRNAs whose entire length overlapped an mRNA were excluded from this analysis. I conducted the same analysis with tRNAs instead of sRNAs for comparison with another non-coding RNA with structure (Figure 3.20). In general, I find that in comparison to tRNAs, C/D box sRNAs tend to not only overlap the 3’ ends of protein-coding genes, but also their potential 3’ UTRs.

In general, *Pyrobaculum* C/D box sRNAs tend to overlap or be closer to the 3’ end of mRNAs than the 5’ end. Less than 25% of the distances in either the 5’ sense or 5’antisense categories are within 30nt. At this distance, the C/D box sRNA will not overlap with the promoter of the mRNA. In contrast, more than 50% of the distances in the 3’ distance categories are within 30nt. There is a large peak at -4 nt (29
Figure 3.20: **The distance between tRNAs and sRNAs with protein-coding genes.** Histograms of the distances between sRNAs or tRNAs to the nearest or overlapping mRNAs. Distances are categorized by orientation to the mRNA. See Figure 3.19 for diagram of orientations. Grey vertical line indicates a distance of zero. In general, compared to tRNAs, C/D box sRNAs tend to overlap the 3′ end of mRNAs (two top panels). C/D box sRNAs and tRNAs rarely overlap the 5′ end of mRNAs, but C/D box sRNAs that are 5′ antisense to mRNAs tend to be closer than tRNAs in this orientation (lower left panel).

sRNAs) for the 3′ sense orientation category and 0 nt (43 sRNAs) for the 3′ antisense orientation category. The peak at -4nt is likely the result of C boxes starting with either the sequence GUGA or AUGA, both containing UGA that serves as the stop codon for the mRNA. Forty-three of 48 guides 0 nt away from an mRNA are 3′ antisense, creating a potential antisense interactions with 3′ UTRs but not influencing the coding region of the mRNA.
Figure 3.21: **Boxplots of distances for C/D box sRNAs, tRNAs, and mRNAs to the closest or overlapping mRNA.** Boxplots divide the distribution into four quartiles (Q1-Q4). Within the box, a horizontal line is drawn at Q2, the median of the data set. Two vertical lines, called whiskers, extend from the front and back of the box. The front whisker goes from Q1 to the smallest non-outlier, and the back whisker goes from Q3 to the largest non-outlier. Note that distances greater than 1000nt are not shown to provide a clearer view of the distributions near 0. Outliers were computed using the statistical software R [156] when the plots were created using the `geom_boxplot` command of the `ggplot2` library [201]. The smallest non-outlier is calculated by multiplying the interquartile range (Q3-Q1) by 1.5 and subtracting the product from Q1. The largest non-outlier is calculated by multiplying the interquartile range by 1.5 and adding the product to Q3. C/D box sRNAs tend to be closer to mRNAs in all orientation categories (described below the plot) than tRNAs. However, the overlap distribution of C/D box sRNAs with mRNAs in the 3' orientations look similar to the overlap distribution of mRNAs with each other in these categories. This suggests that the overlap of C/D box sRNAs with the 3' end of mRNAs may simply be a result of the ability of the 3' end of *Pyrobaculum* mRNAs to tolerate overlap.
In all categories, C/D box sRNAs tend to cluster nearer to protein-coding genes than tRNAs do (Figures 3.20 and 3.21). For C/D box sRNAs, 3’ antisense group had on average the smallest distance to an mRNA, but 3’ sense C/D box sRNAs were nearly as close, especially compared to tRNAs (40% of sRNAs compared to 14.7% of tRNAs are within 20nt of 3’ end of a protein-coding gene). Again, I suggest that these potential interactions between C/D box sRNAs and UTRs may play a role in gene regulation and possibly mRNA stability, especially the antisense interactions. Both C/D box sRNAs and tRNAs tend to be further from the 5’ end of mRNAs, likely to avoid disrupting the promoter of the mRNA. However, C/D box sRNAs are still much closer on average to the 5’ end of mRNAs (14.6% of sRNAs compared to 1.4% of tRNAs are within 30nt of 5’ end of a protein-coding gene). It is unclear whether the pattern of C/D box sRNAs in the 5’ antisense category indicates 5’ UTR interactions, divergent promoters, or another factor. A 5’ UTR interaction is less likely as many archaea have been shown to have leaderless transcripts, including Pae \[26, 66\].

3.5.8.1 Why do C/D box sRNAs cluster near mRNAs?

I explored reasons why C/D box sRNAs could cluster near protein-coding genes. I suspected that the distance distributions of C/D box sRNAs to the nearest mRNA might simply be the result of the tolerance of protein-coding genes for 3’ overlap within the *Pyrobaculum*. First, I looked at how much *Pyrobaculum* mRNAs typically overlap with each other (Figure 3.21). The distance distributions for mRNAs are similar to the C/D box sRNA distributions in the 3’ antisense and sense categories. However,
the distance distributions for the C/D box sRNAs and mRNAs were slightly different for the 5' antisense and sense categories. In the 5' antisense category, C/D box sRNAs are slightly closer to the nearest mRNA than mRNAs were to each other. The opposite is true for the 5' sense category, where mRNAs are slightly closer to the nearest mRNA than C/D box sRNAs are. The pattern for the 5' sense category is likely the result of how mRNAs are located in relation to each other within operons. These results suggest that the tolerance of mRNAs for 3' overlap plays a large role in how C/D box sRNAs are situated in the genome.

Next I tried to test whether the C/D box sRNA distance distributions are a result of random insertions rather than a selection mechanism. I randomly shuffled C/D box sRNA positions around mRNAs and calculated the distance distributions. I found that the results were inconclusive, because the distributions change based on the amount of allowed overlap with mRNAs.

Genes are more likely to be closer together within operons, so I further split the distributions based on membership within an operon (Figure 3.22). I used Arkin operon predictions, because the efficacy of the method in archaea was verified with microarray evidence [154]. I found that outside of operons C/D box sRNAs (i) still show the same patterns of being closer to mRNAs than tRNAs are and (ii) appear to be closer than mRNAs in the 3' sense and 5' antisense orientations than mRNAs to each other. The difference in the C/D box sRNA 3' antisense and 3' sense orientation distributions indicate that there is a more complex relationship between C/D box sRNAs and mRNAs than toleration of overlap.
Figure 3.22: **Boxplots of distances for C/D box sRNAs, tRNAs, and mRNAs to the closest or overlapping mRNA divided by operon membership.** See description of boxplots in caption of Figure 3.21. Note that distances greater than 500nt are not shown. I used the Arkin operon predictions for the operon set [154]. I re-plotted the data in Figure 3.21 based on whether the genes belonged to an operon to see whether most overlap between mRNAs and other genes occurred in operons. A large number of C/D box sRNAs and mRNAs still appear to overlap the 3' end of mRNAs outside of operons.
As expected, distributions in the operon category indicated either overlapping or small distances. There are some mRNAs in the antisense categories, because in some instances the mRNAs are closer to mRNAs outside of the operon rather than within the operon. No C/D box sRNAs in an operon were 5′ antisense to mRNAs. In general, C/D box sRNAs in operons showed similar distance distributions to mRNAs in the 3′ sense and antisense orientations. These results indicate that operon membership plays a role in where C/D box sRNAs are located in relation to protein-coding genes, but did not clarify factors that influence C/D box sRNA location outside of operons.

A variety of factors influence the location of C/D box sRNAs within the genome. I find that tRNAs tend to not overlap other genes, likely because structure and sequence are so integral to their function. In contrast, mRNAs and C/D box sRNAs appear to be more flexible and have a complex relationship.

3.6 Discussion

I used RNAseq data, computational methods and comparative genomics to identify a likely comprehensive set of 526 C/D box sRNAs from six species within the genus *Pyrobaculum*. I organized these sRNAs into 110 homologous families based on sequence similarity of their D and D′ guides and used extended guide-target complementarity and the “N+five” rule to predict different sites methylation in 16S and 23S rRNA, respectively. With this set of families and predicted targets, I was able to improve computational predictions, explore known and hypothesized functions of C/D box
sRNAs, study their impact on the genome, and the evolution of C/D box sRNAs.

I created a covariance model to incorporate K-turn structure into a computational prediction method for C/D box sRNAs. I showed that different bitscore thresholds used for evaluating *Pyrobaculum* predictions are can also be used for evaluating predictions within *Pyrococcus*, a genus from a different archaeal phylum—this means that the covariance model can be used to predict C/D box sRNAs in other species without having to re-evaluate thresholds. In addition, I found that filtering by overlap with mRNAs greatly reduces the number of false positives. This filter is especially useful if the genome does not have any small RNA sequencing data to help validate predictions.

Patterns of the predicted methylation sites suggest that the modifications may be important for the stability and function of the rRNA and that sRNAs may help chaperone rRNA folding. The sites of modification cluster primarily in regions of rRNA that are of functional importance and are largely devoid of ribosomal protein contacts (e.g., the decoding center in the small subunit and the peptidyl transferase center in the large subunit). Many sRNAs (235 of 526) use their D and D′ guides to base pair with rRNA sequences that are within 100 nts; I suggest that these dual interactions play an important role in the localized folding of the rRNAs and their assembly into ribosomal subunits. Dennis and I also identified instances where D and D′ guides pair with sites that are distant in the primary rRNA sequence but close in the secondary structure; these long-range interactions could help arrange localized regions into the larger global 16S or 23S rRNA structures. A recent computational study simulating C/D box sRNA chaperone function to rRNA folding suggests that double guide sRNAs
may be especially important for proper long-range interactions in rRNAs [168].

Predicting targets also allowed me to identify orphan guides. The function of these guides continues to be a mystery, especially for a conserved double orphan guide such as sR43. Although some of the orphan guides may be the result of overlap or proximity to protein-coding genes, not all can be accounted for by this explanation.

I used the grouping of the *Pyrobaculum* sRNAs into homologous families as a resource to address questions relating to the evolutionary origins, propagation, and diversification of sRNA genes. I identified instances of gene duplication, gene rearrangement, guide translocations, and guide divergence caused by deletion/insertion or nucleotide substitution events. By tracking low-level sequence similarity between guides from different homologous families, I was able to identify a superfamily consisting of five different families that exhibit deep evolutionary connections within the *Pyrobaculum* lineage. Finally, I carried out a detailed examined the genomic context of sRNA genes. I identified several instances where sRNAs are co transcribed with each other or with a tRNA gene. These arrangements were generally unstable and not conserved between all six of the *Pyrobaculum* species. I also noted that sRNA genes are frequently located near to or overlapping protein-coding genes. The sRNA guides that overlap the 5′ end of a protein-coding gene generally lack complementarity to rRNA or tRNA whereas sRNAs guides that overlap the 3′ end of protein genes have guide function. I suggest that the close association of sRNAs with the 3′ ends of protein genes may play a role in gene regulation.

In summary, this comprehensive effort of identify the complete set of C/D
box sRNA genes from six species within the hyperthermophilic genus *Pyrobaculum* has provided unique and valuable insights into (i) their structure and function, (ii) their role in ribosome subunit biogenesis, (iii) their evolutionary origin, propagation and divergence and (iv) their role in shaping overall genome architecture.
Chapter 4

Computational and Experimental
Detection of C/D box RNAs Across the Archaea

4.1 Introduction and Background

In the previous chapter, I compiled a comprehensive set of C/D box sRNAs in the *Pyrobaculum* genus to conduct a comparative genomics study of the structure and evolution of C/D box sRNAs. A brief comparison of C/D box sRNAs between the *Pyrobaculum* and *Pyrococcus* genera led to questions that are explored in this chapter, including (i) will the new model predict more C/D box sRNAs in other archaeal species? (ii) what is the abundance of C/D box sRNAs across the archaeal domain? (iii) what is the kink-turn sequence diversity across the Archaea? (do the *Pyrobaculum* sRNAs have more flexibility than those found in the rest of the Archaea?) (iv) how conserved are
sRNAs among different clades? In this chapter I extend my work in the *Pyrobaculum* genus to create a comprehensive view of C/D box sRNA characteristics across the Archaea. These results indicate future directions for computational models and study of C/D box sRNA evolution and function.

4.1.1 State of C/D box sRNA annotations in the Archaea

Although there are now over a hundred sequenced archaeal genomes [29], only 24 species have published C/D box sRNA annotations. Of these species, there is a heavy concentration in particular genera; six of these species are in the *Pyrobaculum* genus and three are in the *Pyrococcus* genus, as discussed in Chapter 3 [18, 71] (Figure 4.1). Five of the species are halophiles with only 1–2 predicted candidates based on comparative genomics [77, 171]. One C/D box sRNA was found in *Halobacterium salinarum* NRC-1 by sequencing RNAs bound to proteins associated with C/D box sRNAs [196]. Sequencing of cDNA clones has yielded annotations in *Sulfolobus acidocaldarius* [144, 51], *Sulfolobus solfataricus* [178, 206], and *Archaeoglobus fulgidus* [177]. Computational efforts have also produced annotations in *S. solfataricus*, *A. fulgidus*, *Aeropyrum pernix*, and *Methanocaldococcus jannaschii* [144, 51]. More recently, high-throughput RNA sequencing has identified C/D box sRNAs in *Methanopyrus kandleri* [174], *Nanoarchaeum equitans* [157], *S. solfataricus* [204] and *Thermococcus kodakaraensis* [89]. Small RNA sequencing has been one of the most effective ways to annotate archaeal C/D box sRNAs and validate computational predictions [18].

Based on the presence of proteins associated with C/D box sRNAs (L7Ae,
fibrillarin, Nop5), we would expect to find C/D box sRNAs in every archaeal species \[63, 69\]. As discussed in Section 3.5.2 and Appendix A.1, limitations of current computational models partially account for the underannotation, but annotation of C/D box sRNAs in archaeal genomes has also been sporadic. Many studies have focused exclusively on well-annotated species, such as those in the *Pyrococcus* genus, or did a cursory look for C/D box sRNAs if studying the transcriptome of a species. Fewer than 10 studies of have been completed with archaeal cDNA clone, transcriptome, or small RNA sequencing data, so most studies have relied on computational methods for annotations \[8\]. Many C/D box sRNAs have been found in thermophilic species, but very few or no C/D box sRNAs have been identified in the Class II methanogens and halophiles (Figure 4.1), which suggests that there might be other undiscovered forms of C/D box sRNAs that are used to guide 2′-O-methylation \[77\]. Sequence-homology-based searches have had limited success because archaeal small RNAs are rarely conserved outside of a genus \[8\]. Two exceptions are the highly conserved C/D box sRNA in the tRNA\(^{Trp}\) intron found in many species of Euryarchaeota \[171\] and the highly conserved sRNA found in the halophiles and other euryarchaea \[77\].

The approximately 30% increase in C/D box sRNAs in the *Pyrobaculum* species from small RNAseq data \[18\] and my computational work indicates that there are likely more C/D box sRNAs in the archaea, especially for crenarchaeal species. The upper range of C/D box sRNAs is represented by *I. hospitalis* (129 sRNAs) and *M. kandleri* (127 sRNAs), a crenarchaeal and an euryarchaeal species, respectively. The lower range (for species with annotations) is one to two C/D box sRNAs. We
Figure 4.1: Published annotations of archaeal C/D box sRNAs mapped onto a phylogenetic tree. The tree is a simplified version of the trees found in [29] and Figure 2.2. Bolded clades are major phyla of the Domain Archaea. Species with C/D box sRNA annotations are listed. If two or more species from a genus are listed, the genus and number of sRNAs typically found in each species is listed. Most C/D box sRNA annotations are in hyperthermophilic species, which usually have greater than 50 C/D box sRNAs. However, the hyperthermophiles *M. jannaschii* and *A. fulgidus* have fewer than 10 C/D box sRNAs. No C/D box sRNA annotations have been published from thaumarchaeal or Class II methanogen species and very few have been found in halophilic species. The best studied genera are *Pyrococcus* and *Pyrobaculum* [18] [71].
expect to have a large range for the number of C/D box sRNAs in various archaeal species, especially considering the ranges in eukaryotes [52], but missing annotations hinders studies into the impact of C/D box RNAs on the genome and evolution of their function. Functions for C/D box sRNAs other than 2′-O-methylation have been found in eukaryotes, and it is possible that these types of functions have evolved in archaea as well. Eukaryotic C/D box RNAs or fragments of C/D box RNAs have been found to behave like microRNAs and cause gene silencing in human [61, 25], the protozoan parasite Giardia lamblia [167], mouse, chicken, fly, Arabidopsis, and fission yeast [175]. Evidence indicates that a few eukaryotic C/D box sRNAs may even regulate alternative splicing of mRNAs [98]. Better annotations will assist in exploration of archaeal C/D box sRNA functions beyond post-transcriptional modification.

4.1.2 Summary of this chapter

To improve annotations of C/D box sRNAs in the Archaea, I applied the co-variance model I developed in Chapter 3 to 21 diverse archaeal species (Table 4.1) and combined these with hand-curated snoscan predictions [115, 144], a C/D box sRNA prediction method previously used by the Lowe lab. To augment the search and validate predictions, I used small RNA sequencing data in these species. In some cases I used comparative genomics to validate predictions where there was no small RNA sequencing evidence. With this set of C/D box sRNAs, I looked at clade-specific patterns of abundance and structure of C/D box sRNAs. These annotation efforts revealed that (i) not all thermophiles have a large number of C/D box sRNAs (greater than 50), (ii) at least
seven C/D box sRNAs exist in the thaumarchaeal species *Nitrosopumilus maritimus*, (iii) there are likely no other C/D box sRNAs in the halophiles or methanogens or there are C/D box sRNAs with forms not detectable by current models in these species, and (iv) there are a few highly conserved guides with targets in functionally important sites of the ribosome.

Next I analyzed the structural diversity by analyzing (i) the kink-turn motifs in these sRNAs and (ii) the prevalence of a kink-turn instead of a K-loop forming between the C′ and D′ boxes. These results reveal the diversity of archaeal C/D box sRNA evolution in terms of structure and function.

### 4.2 Small RNA sequencing data and annotations

Small RNA sequencing used in this study is from 44 samples across 19 archaeal species that the Lowe lab prepared (Table 4.2). Samples were collected in various growth stages and conditions by either the Lowe lab or collaborators. Size-selected libraries were prepared by the Lowe lab using an in-house protocol as described in the Methods section.

Small RNA sequencing data from *S. acidocaldarius* is from a collaboration with Lennart Randau’s lab (Max Planck Institute for Terrestrial Microbiology). Annotations in *Methanococcus maripaludis* C5, *Ignicoccus hospitalis*, *S. acidocaldarius*, and *Thermoproteus tenax* are a result of efforts by me, Randau, Vanessa Tripp, and Patrick Dennis and will be published in BMC Genomics (Dennis *et al.*, in review). Andrew Uzilov also
<table>
<thead>
<tr>
<th>Species and Strain</th>
<th>Abbr.</th>
<th>Growth Temp (C)</th>
<th>Phylum</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Aeropyrum pernix</em> K1</td>
<td>Ape</td>
<td>90-95</td>
<td>Crenarchaeota</td>
<td>Desulfurcoccales</td>
</tr>
<tr>
<td><em>Archeaoglobus fulgidus</em> DSM4304</td>
<td>Afu</td>
<td>60-95</td>
<td>Euryarchaeota</td>
<td>Archaeoglobales</td>
</tr>
<tr>
<td><em>Halobacterium sp</em> NRC-1</td>
<td>Hsp</td>
<td>42</td>
<td>Euryarchaeota</td>
<td>Halobacteriales</td>
</tr>
<tr>
<td><em>Haloferax volcanii</em> DS2</td>
<td>Hvo</td>
<td>37-42</td>
<td>Euryarchaeota</td>
<td>Halobacteriales</td>
</tr>
<tr>
<td><em>Ignicoccus hospitalis</em> KIN4-1</td>
<td>Iho</td>
<td>90</td>
<td>Crenarchaeota</td>
<td>Desulfurcoccales</td>
</tr>
<tr>
<td><em>Lokiarchaeum</em></td>
<td>Lok</td>
<td>Thermophile</td>
<td>Lokiarchaeota</td>
<td>N/A</td>
</tr>
<tr>
<td><em>Methanocaldococcus jannaschii</em> DSM 2661</td>
<td>Mja</td>
<td>85</td>
<td>Euryarchaeota</td>
<td>Methanococcales</td>
</tr>
<tr>
<td><em>Methanococcus maripaludis</em> S2</td>
<td>Mma</td>
<td>18-47</td>
<td>Euryarchaeota</td>
<td>Methanococcales</td>
</tr>
<tr>
<td><em>Methanococcus maripaludis</em> C5</td>
<td>Mma C5</td>
<td>18-47</td>
<td>Euryarchaeota</td>
<td>Methanococcales</td>
</tr>
<tr>
<td><em>Methanopyrus kandleri</em></td>
<td>Mka</td>
<td>84-110</td>
<td>Euryarchaeota</td>
<td>Methanopyrales</td>
</tr>
<tr>
<td><em>Methanosarcina acetivorans</em> C2A</td>
<td>Mac</td>
<td>35-40</td>
<td>Euryarchaeota</td>
<td>Methanosarcinales</td>
</tr>
<tr>
<td><em>Methanosarcina barkeri str. Fusaro</em></td>
<td>Mba</td>
<td>35-40</td>
<td>Euryarchaeota</td>
<td>Methanosarcinales</td>
</tr>
<tr>
<td><em>Methanosarcina mazei</em> Go1</td>
<td>Mmz</td>
<td>30-37</td>
<td>Euryarchaeota</td>
<td>Methanosarcinales</td>
</tr>
<tr>
<td><em>Methanospirillum hungatei</em> JF-1</td>
<td>Mhu</td>
<td>37</td>
<td>Euryarchaeota</td>
<td>Methanomicrobiales</td>
</tr>
<tr>
<td><em>Nanoarchaeum equitans</em></td>
<td>Neq</td>
<td>80-100</td>
<td>Nanoarchaeota</td>
<td>Nanoarchaeales</td>
</tr>
<tr>
<td><em>Nitrosopumilus maritimus</em> SCM1</td>
<td>Nma</td>
<td>21-23</td>
<td>Thaumarchaeae</td>
<td>Nitrosopumilales</td>
</tr>
<tr>
<td><em>Staphylothermus marinus</em> F1</td>
<td>Sma</td>
<td>85-98</td>
<td>Crenarchaeota</td>
<td>Desulfurcoccales</td>
</tr>
<tr>
<td><em>Sulfolobus acidocaldarius</em> DSM 3638</td>
<td>Sac</td>
<td>70-80</td>
<td>Crenarchaeota</td>
<td>Sulfolobales</td>
</tr>
<tr>
<td><em>Sulfolobus solfataricus</em> P2</td>
<td>Sso</td>
<td>80</td>
<td>Crenarchaeota</td>
<td>Sulfolobales</td>
</tr>
<tr>
<td><em>Thermococcus kodakaraensis</em> KOD1</td>
<td>Tko</td>
<td>60-100</td>
<td>Euryarchaeota</td>
<td>Thermococcales</td>
</tr>
<tr>
<td><em>Thermoproteus tenax</em></td>
<td>Tte</td>
<td>70-97</td>
<td>Crenarchaeota</td>
<td>Thermoproteales</td>
</tr>
</tbody>
</table>

Table 4.1: Summary of species for panarchaeal C/D box RNA analysis. Growth temperatures for each species and three-letter name abbreviations used in this study are listed. Data sources are the UCSC Archaeal Browser and the DSMZ culture collection. In general, species are considered thermophilic if they have an optimal growth temperature $\geq 60$°C and hyperthermophilic if they have an optimal growth temperature of $\geq 80$°C. *Lokiarchaeum* is a composite genome and optimal growth temperatures for *Lokiarchaeal* species is unknown. The metagenomic sample was collected near a hydrothermal vent, so *Lokiarchaeal* members are likely thermophiles or hyperthermophiles.
<table>
<thead>
<tr>
<th>Species</th>
<th>Size Selection</th>
<th>Growth Condition</th>
<th>Sequencing date</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afu</td>
<td>small RNA</td>
<td>exp, stat</td>
<td>2-Nov-2010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tRNA-5S</td>
<td></td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td>Ape</td>
<td>small RNA</td>
<td>exp, stat</td>
<td>15-Jul-2010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tRNA-5S</td>
<td></td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td>Hsp</td>
<td>small RNA</td>
<td>exp, stat</td>
<td>2-Nov-2010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tRNA-5S</td>
<td></td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td>Hvo</td>
<td>small RNA</td>
<td>exp, stat</td>
<td>2-Nov-2010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tRNA-5S, 5S-500</td>
<td></td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td>Ihö</td>
<td>small RNA</td>
<td>Independent co-culture w/ Neq</td>
<td>25-Jul-2011</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>25-Jul-2011</td>
<td></td>
</tr>
<tr>
<td>Mac</td>
<td>small RNA</td>
<td>stat</td>
<td>2-Nov-2010</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>25-Jul-2011</td>
<td></td>
</tr>
<tr>
<td>Mba</td>
<td>small RNA</td>
<td>exp, stat</td>
<td>2-Nov-2010</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>25-Jul-2011</td>
<td></td>
</tr>
<tr>
<td>Mhu</td>
<td>small RNA</td>
<td>exp, stat</td>
<td>2-Nov-2010</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>25-Jul-2011</td>
<td></td>
</tr>
<tr>
<td>Mja</td>
<td>small RNA, 5S-500</td>
<td>stat</td>
<td>25-Jul-2011</td>
<td></td>
</tr>
<tr>
<td></td>
<td>small RNA</td>
<td>exp, stat</td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td>Mka</td>
<td>small RNA</td>
<td>N/A</td>
<td>Published in [174]</td>
<td></td>
</tr>
<tr>
<td>Mma</td>
<td>small RNA</td>
<td>NH4 limited</td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>H2 limited</td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PO4 limited</td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>exp, stat</td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td>Mma C5</td>
<td>small RNA</td>
<td>N/A</td>
<td>Published in [160]</td>
<td></td>
</tr>
<tr>
<td>Mmz</td>
<td>small RNA</td>
<td></td>
<td>25-Jul-2011</td>
<td></td>
</tr>
<tr>
<td>Neq</td>
<td>small RNA</td>
<td>co-culture w/ Ihö</td>
<td>25-Jul-2011</td>
<td></td>
</tr>
<tr>
<td>Nma</td>
<td>small RNA</td>
<td>exp, stat, keto</td>
<td>2-Nov-2010</td>
<td></td>
</tr>
<tr>
<td>Pfu</td>
<td>tRNA-5S</td>
<td></td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5S-500</td>
<td></td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td></td>
<td>small RNA</td>
<td>exp, stat</td>
<td>25-Jul-2011</td>
<td></td>
</tr>
<tr>
<td>Sac</td>
<td>small RNA</td>
<td>N/A</td>
<td>Collab. w/ Randau lab (MPI)</td>
<td></td>
</tr>
<tr>
<td>Sma</td>
<td>tRNA-5S</td>
<td></td>
<td>26-Mar-2011</td>
<td></td>
</tr>
<tr>
<td>Sso</td>
<td>small RNA</td>
<td>STIV(+)</td>
<td>25-Jul-2011</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>STIV(-)</td>
<td>25-Jul-2011</td>
<td></td>
</tr>
<tr>
<td>Tko</td>
<td>small RNA</td>
<td>exp, stat</td>
<td>15-Jul-2010</td>
<td></td>
</tr>
<tr>
<td>Tte</td>
<td>small RNA</td>
<td>N/A</td>
<td>Published in [152]</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: **Summary of small RNA seq samples.** Samples used to generate the small RNA sequencing data used in this study are described here. Species abbreviations are in Table 4.1. Growth phase at time of harvest or other special growth conditions (e.g. limited resource such as H2) are listed under “Growth Condition.” *Exp* and *stat* represent exponential and stationary growth stages. Collaborators that provided material for small RNA sequencing or small RNA sequencing data are listed.
provided new annotations for *I. hospitalis* (53 sRNAs), *S. solfataricus* (7), *S. acidocaldarius* (2), and *A. pernix* (2) from his circular RNA work \[188\]. Additional annotations in *T. kodakaraensis* and *S. solfataricus* are from transcriptome and small RNA studies \[89, 204, 178\].

Predictions were generated by scans with the covariance model developed in Chapter 3. Initial scans used Infernal v1.0 software and species were later re-scanned with Infernal v1.1 software when it was released. In some cases, our lab had already generated predictions using snoscan software \[115\], and these predictions are available via the “Genbank ncRNA” track in the UCSC Archaeal Genome Browser \[35\]. I used naming from snoscan predictions for consistency. The snoscan predictions were generated with low thresholds and have not been verified with RNAseq data or manual curation. Some manual curation of snoscan predictions was done by Andrew Uziel. Novel predictions from the covariance model were named starting at sR101.

### 4.3 Results

#### 4.3.1 Summary of new C/D box sRNA predictions across the Domain Archaea

There was a general increase of annotations across species, mostly in the crenarchaeal thermophiles. (See Section 2.1.2 for a detailed description of archaeal phylogeny and traits.) The new covariance model predicted approximately 30% more C/D box sRNAs in these species than snoscan predictions, echoing the increase seen the *Pyrobac-
Figure 4.2: Summary of C/D box sRNA Annotations. New annotations are a combination of my predictions, hand-curated snoscan predictions, and other unpublished sRNAs found by the Lowe lab. Exceptions are those found in *S. acidocaldarius*, *T. tenax*, *I. hospitalis*, and *M. maripaludis C5*, which were hand-curated predictions from small RNAseq data in collaboration with the Randau lab.

ulum (examples are *A. pernix* and *S. marinus*). Six new sRNAs were found in *T. kodakarensis*, a euryarchaeal thermophile where 54 C/D box sRNAs were confirmed by transcriptome sequencing [89]. Most of these sRNAs have small RNA sequencing data from our samples, so small RNA sequencing data, rather than transcriptome sequencing may be better for small RNA annotations. I found that very few C/D box sRNAs (1–2) in the halophiles and I discuss these findings in Section 4.3.3.

4.3.1.1 C/D box sRNAs in previously unannotated phyla

No C/D box sRNAs were previously annotated in the Thaumarchaeota phylum, and I found seven sRNAs with small RNA sequencing evidence in *Nitrososphaeraceae marinus SCM1*. Five of these are conserved with two other thaumarchaeal species, *Nitrososphaeraceae yellowstonii HL72* and *Cenarchaeum symbiosum A*. Notably, the D guide
Figure 4.3: **Conserved C/D box sRNA in the Thaumarchaea.** Conservation of Nma sR16 with other thaumarchaeal species, with species from the Desulfurococcales order, and Pca sR40 is shown. Homologous guide sequence are highlighted in blue and red. Nucleotides modified in rRNA are in bold and base pairing between the guide and rRNA is shown by ‘|’. Species abbreviations: Nma - *Nitrosopumilus maritmus*, Csy - *Cenarchaeum symbiosum A*, Nye - *Nitrosocaldus yellowstonii HL72*, Sma - *S. marinus*, Dmu - *Desulforcoccus mucosus*, Tag - *Thermosphaera aggregans*, She - *Staphylothermus hellenicus DSM 12710*, Dka - *Desulfurococcus kamchatkensis*, Hbu - *Hyperthermus butylicus*.


of Nma sR16 is conserved in not only other Thaumarchaea, but also in species of the Desulfurococcales order of the Crenarchaea phylum (Figure 4.3). These guides are homologous to the D’ guide of Pca sR40 and target nucleotides in helix 70 of 23S rRNA, which is part of the interface with the small ribosomal subunit.[32] In my work with the Randau lab, we found that the region containing helices 68-71 is a methylation hotspot after mapping methylation targets of *S. acidocaldarius*, *T. tenax*, *M. kandleri*, *P. calidifontis*, *M. maripaludis C5*, *I. hospitalis* and *N. equitans* (Figure 4.4).

Candidates were also found in the C/D box sRNAs found in the Lokiarchaeal composite genome (504 contigs, Genbank Accession JYIM00000000, [172]). Thirty-two predictions had a bitscore of 15 or higher. This cutoff was chosen based on the results from Chapter 3. Approximately 70% of the predictions are 60nt or longer. This is larger than typical archaeal C/D box sRNAs, which usually range from 47 to 55nt. The greater
Figure 4.4: **Mapping of predicted 2'-O-methylation for seven diverse archaeal species onto 23S rRNA.** Predicted sites of sRNA directed 2'-O-methylation from *T. tenax*, *S. acidocaldarius*, *M. kandleri*, *I. hospitalis*, *M. maripaludis C5*, and *N. equitans* were mapped onto the consensus secondary structure of the archaeal 23S rRNA [31]. The multiple occurrences of methylation at a given nucleotide position is indicated by increasing dot size (*i.e.*, 1 methylation targets found in one to seven organisms). The color of the dots (from blue to red) indicates increasing methylation frequency within a nine-nucleotide window. Note that there are hotspots at functionally important regions of 23S rRNA, including helices 68-71 of domain IV and helices 90-93 of domain V, which belong to the interface between the large and small ribosomal subunits and the peptidyl transferase center, respectively. This figure was created by Vanessa Tripp from the Randau lab. Data used to create this figure was generated by Tripp, Patrick Dennis, and me and is in a paper that is under review.
Figure 4.5: **C/D box sRNA candidates in the Lokiarchaea.** Top scoring C/D box sRNA candidates found in the Lokiarchaeum composite genome. Some of the C/D box sRNA candidates in this genome have longer than normal guides, such as the D guide of cand1 (33nt) and the D' guide of cand3 (35nt). Archaeal C/D box sRNA guides are typically 10–14nt [184], while eukaryotic ones are longer.

Length of lokiarchaeal C/D box sRNAs may indicate that the sRNAs in these species are more eukaryotic-like compared to C/D box sRNAs found in other archaea. Eukaryotic C/D box sRNAs typically have longer guides and loop regions than those found in archaea. Analysis by Spang et al. indicates that Lokiarchaea form a monophyletic group with eukaryotes and that Lokiarchaea have homologs to eukaryotic signature proteins [172], so the eukaryotic-like features may extend to C/D box sRNAs.

### 4.3.1.2 C/D box sRNAs in the methanogens

Little has been described of C/D box sRNAs in the methanogens. Published description and study of methanogen C/D box sRNAs have been of Class I methanogen species *M. jannaschii* and *M. kandleri*. Eight sRNAs were predicted in *M. jannaschii* in the first computational analysis of archaeal sRNAs [144], and a record number of 127 sRNAs were discovered in *M. kandleri* [174]. The disparity between the number of C/D box sRNAs between these two species might be simply a result of phylogeny. *M. kandleri* is an outgroup to the rest of the methanogens and its phylogenetic placement is uncertain [27], so in regards to other methanogens it is unusual. The large number of *M. kandleri* C/D box sRNAs supports the hypothesis that the number of C/D box
sRNAs is positively correlated with optimal growth temperature [51, 138]. *M. kandleri* has an optimal growth temperature of 98°C and can survive at 122°C [176], which are extreme temperatures even compared to other archaeal hyperthermophiles. However, the hypothesis correlating growth temperature and number of C/D box sRNAs does not hold up in regards to *M. jannaschii*, which only has 10 C/D box sRNAs but is also a hyperthermophile (optimal growth at 85°C). This discrepancy is discussed further in Section 4.3.2.

A few new sRNAs were found in the methanogens, with species totals ranging from 5 to 15 after new annotations were added. Few novel high-confidence sRNAs were found by the covariance model compared to what was predicted by snoscan, but one new sRNA was found in *M. maripaludis S2*, one in *M. acetivorans*, and three in *M. barkerii* with the new covariance model. No predictions in *M. hungatei* had small RNA sequencing, and some of the snoscan predictions are low confidence based on overlap with predicted mRNAs or poor box features. The predictions in *M. hungatei* also did not have any conservation with the other methanogens in this study nor the most closely related species with an available genome, *Methanoregula boonei*. The apparent lack of conservation of *M. hungatei* C/D box sRNAs may be due to the dearth of closely related species with sequenced genomes.

In general there is very little conservation of C/D box sRNA guides across the methanogens. However, there is one sRNA family conserved in all Class I methanogens in this study and two in the Methanosarcinales order of the Class II methanogens. For the Class I methanogens, Mka sR47 has homologs with Mja sR07, Mma sR03,
and Mma_C5 sR104 and is homologous to Pca sR33 (Figure 4.6A). These sRNAs have targets to helix 68 of 23S rRNA which is, as described in Section 4.3.1.1, a hotspot for methylation and is at the interface between the large and small ribosomal subunits. Despite the large number of sRNAs in *M. kandleri*, the only other sRNA in this species that has a homolog in other methanogens is Mka sR48, and its homolog is Mja sR13. Since *M. kandleri* is an outgroup to the rest of the methanogens, the low number of C/D box sRNAs that are conserved with other known methanogen C/D box sRNAs is not unexpected. The *M. maripaludis* strains and *M. jannaschii* have a higher proportion of homologs: one family is comprised of Mma sR05, Mma_C5 sR105, and Mja sR09 and another family is comprised of Mma sR08, Mma_C5 sR107, Mja sR01 (Figure 4.6B). This means that approximately one-third of Mja, Mma, and Mma_C5 C/D box sRNAs belong to homologous families with C/D box sRNAs found in other Class I methanogens.

The C/D box sRNAs sR02 and sR16 in Class II methanogen *M. acetivorans* has homologs in *M. barkeri*, *M. mazei*, and *M. burtonii* (Figure 4.6B). The D’ guide of Mac sR02 also targets helix 68, similar to the conserved sRNA in the Class I methanogens. The D guide of Mac sR02 targets helix 89 of 23S rRNA which is in the peptidyl transferase center and was also found to be a hotspot of methylation in my work with the Randau lab (Figure 4.4). The comparison of C/D box sRNA conservation and targets across and within major phyla indicate that the general region of methylation of ribosomal RNA, rather than a specific nucleotide, is important.
Figure 4.6: **Conserved C/D box sRNAs in the methanogens.** Only a few C/D box sRNAs families are conserved across multiple methanogen species. Portions of guide sequences identical across multiple species are highlighted. (A) Conserved sRNA family in all the Class I methanogens in this study. This family is homologous to *Pyrobaculum* C/D box sRNA family sR33 and thus is a rare case of guides conserved between species of different phyla. Guide-target interactions with 23S rRNA are shown and nucleotides predicted to be methylated are in bold. *Pyrobaculum calidifontis* (Pca) sR33 is shown as the representative from the *Pyrobaculum* sR33 family. Mma sR03 and Mma_C5 sR104 do not have a predicted D guide target. (B) The two conserved sRNA families in the Class II methanogens. Homologs from *Methanococcoides burtonii* (Mbu) are shown. The D’ guide of Mac sR02 targets 23S rRNA G1934 and the D guide targets 23S U2566. The D’ guide of Mac sR16 has no predicted target and the D guide targets 16S A929.
4.3.2 Not all hyperthermophiles have a large number of C/D box sRNAs

Past studies of archaeal C/D box sRNAs have noted a positive correlation between optimal growth temperature and the number of C/D box sRNAs in a species, which suggests that increased RNA methylation is advantageous for life at high temperature [138, 51]. Ribose methylation appears to be beneficial for bacterial thermophiles, as *Bacillus stearothermophilus* demonstrates increased methylation of tRNAs as growth temperature increases [1]. In archaea, most hyperthermophiles have more than 50 C/D box sRNAs. For example, each of the *Pyrobaculum* species have approximately 90 C/D box sRNAs and *Pyrococcus* species have approximately 60 (Chapter 3 [18]). Based on our analysis of small RNAseq data, *Ignicoccus hospitalis* has 132 (optimal growth at 90°C), and *Methanopyrus kandleri* has 127 (optimal growth at 98°C) [174]. Archaeal species that grow at moderate temperatures (25–40°C), such as halophiles, methanogens (mostly Class II), and Thaumarchaea members have either very few or no predicted C/D box sRNAs [51].

There are three species that do not support the hypothesis that higher optimal growth temperatures correlate with a large number of C/D box sRNAs. These archaeal hyperthermophiles are *Nanoarchaeum equitans* (26 C/D box sRNAs [157], optimal growth at 90°C), *Archaeoglobus fulgidus* (5 [144, 177], optimal growth at 83°C), and *Methanocaldococcus jannashchii* (8 [144], optimal growth at 85°C). The sRNAs in *N. equitans* have been confirmed by examining small RNAseq data, and I did not detect
any plausible candidates with the covariance model, so likely this species does not have many more sRNAs, if any, than what has been reported. The low number of sRNA in this species compared to other hyperthermophiles can be explained by its reduced genome (approximately 0.49 Mbp) and its reliance on *Ignicoccus hospitalis* to survive. Since no high-throughput small RNAseq data had been examined for C/D box sRNAs in *A. fulgidus* and *M. jannaschii*, I scanned these species to see if I could find more C/D box sRNAs.

I found only two new C/D box sRNAs were found in *M. jannaschii* and four in *A. fulgidus*, for a total of ten and eight sRNAs in each species, respectively. To find new sRNAs, I gathered snoScan predictions from the UCSC Archaeal Genome Browser (Genbank ncRNAs track) and predictions using the covariance model. I removed predictions that had greater than 80% overlap with mRNAs. If a prediction did not have small RNA sequencing reads, it was removed unless it had homologs in closely related species.

Although *M. jannaschii* and *A. fulgidus* do not have many C/D box sRNAs, these species may have other mechanisms for ribose methylation. Site-specific methyltransferases that function independently of C/D box sRNA exist. These methyltransferases are used in bacteria, which do not have sno-like RNAs. Thus *M. jannaschii* and *A. fulgidus* could use site-specific methyltransferases to introduce ribose methylation beyond the targets of their C/D box sRNAs. To definitively determine if these two species support the hypothesis that a higher rate of ribose methylation occurs in hyperthermophiles compared to mesophilic archaea, biochemical assays that detect 2′-
Methylation are needed [119].

*M. jannaschii* belongs to the Methanococcales order and *A. fulgidus* is closely related to Class II methanogens, so phylogenetically they group near other species that are not known to have many C/D box sRNAs. Thus, it is also possible that the correlation between C/D box sRNAs and growth temperature depends on where species fall in the phylogenetic tree, rather than being a universal relationship.

### 4.3.3 Very few C/D box sRNA candidates predicted in the halophiles

Very few sRNAs have been detected in the halophiles, and I did not have very much success either. Before this study, only three C/D box sRNAs were predicted in the Halobacteriales order by comparative genomics [77, 92, 171]. One C/D box sRNA is located in the intron of tRNA<sup>Trp</sup>, and homologs are present throughout the Euryarchaea phylum (homologous to sR40 in *P. abyssi*) [43, 171]. Another sRNA, sR-t<sup>Met</sup>, modifies elongator tRNA<sup>Met</sup>. Northern blots have confirmed the presence of sR-t<sup>Met</sup> in *H. volcanii* [92]. The third C/D box sRNA was found in *H. volcanii* and predicted to target 23S rRNA G1950 (in helix 69). This sRNA is conserved across the halophiles, Thermococcales species (*Pyrococcus* and *Thermococcus* genera), and in Class I and Class II methanogens [77]. The homolog of this sRNA in *H. salinarum* NRC-1 was also detected by sequencing cDNA clones from co-immunoprecipitated RNA with the methyltransferase fibrillarin [196].

To look for C/D box sRNAs in *H. volcanii* and *Halobacterium* sp. (the two halophilic species with sequencing data), I scanned the genomes with the covariance
Figure 4.7: Halophile C/D box sRNAs. Tracks showing small RNA reads coverage from stationary and exponential phase are shown below the C/D box sRNA. Numbers to the left of the coverage tracks indicate maximum coverage in the current region. Strand is indicated by (+) or (-). (A) New C/D box sRNA in *H. volcanii*. This C/D box sRNA is located at chr:1299145 – 1299199 on the plus strand. (B) Confirmed expression of *H. volcanii* C/D box sRNA predicted to target 23S rRNA G1950 [77]. This C/D box sRNA is located at chr:1781284 – 1781334 on the minus strand. (C) Questionable C/D box sRNA prediction in *H. volcanii*. This questionable C/D box sRNA is located at chr:2125967 – 2126012 on the plus strand. (D) Questionable C/D box sRNA prediction in *Halobacterium* sp. This questionable C/D box sRNA is located at chr:743812 – 743862 on the minus strand and does not have any small RNA reads. The C/D box sRNAs are represented by the black bar with white arrow heads indicating the direction of the sRNA. The box features are outlined with black rectangles. In (B) the stem that forms from the flanking regions is underlined.

model created in Chapter 3. This yielded two plausible candidates in *H. volcanii* and one plausible and one questionable candidate in *Halobacterium* sp. I created a halophile-specific covariance model with sRNAs from four different halophiles, but it did not yield any new results.
The plausible C/D box sRNA candidate in *H. volcanii* has a few small RNAseq reads (Figure 4.7A), but it is not found in other halophiles. The plausible prediction in *Halobacterium* sp. (chr:1937216 – 1937267) has small RNA sequencing reads and is the homolog to the highly conserved C/D box sRNA in *H. volcanii* that targets 23S rRNA G1950. The questionable candidate in *H. volcanii* has only 3 partial reads and is also not found in other halophiles 4.7C). The one questionable candidate in *Halobacterium* sp. has good box features, but it overlaps with a chemotaxis protein (VNG0976G) on the 5′ end and does not have any small RNA reads (Figure 4.7C).

Although only a few new C/D box sRNAs were predicted, the sequencing data did confirm the expression of a *H. volcanii* C/D box sRNA computationally predicted by Grosjean et al.. They used comparative genomics to find this C/D box sRNA and predicted its target to be 23S rRNA G1950 (Figure 4.7B). It is possible that there are other C/D box sRNAs of different forms that are not detected by the current models or there are simply very few C/D box sRNAs in the halophilic branch of the Archaea. To rule out this possibility, we would need to categorize all of the sRNAs in the small RNA sequencing data for each species, and then try to characterize the remaining, unannotated sRNAs.

### 4.3.4 Patterns of K-turn variability in archaeal C/D box sRNAs

In the last chapter, I compared the base-pair variability of the K-turn (formed by the C and D boxes) and K-loop (formed by the C′ and D′ boxes) between *Pyrobaculum* and *Pyrococcus* species and found that *Pyrobaculum* C/D box sRNAs appeared to have
more variability (Section 3.5.3.1). To compare the variability, I looked at each position of the K-turn or K-loop (1b1n–5b5n) between the *Pyrobaculum* and *Pyrococcus* species. Using the data collected in this study, I analyzed the variability of *Pyrobaculum* C/D box sRNA K-motifs compared to those found in other archaea. At first glance, some positions of the K-motifs appear to be more variable in the *Pyrobaculum*. However, after taking into account the number of C/D box sRNAs in each species, *Pyrobaculum* C/D box sRNA K-motifs to not appear to be significantly different from most archaea.

### 4.3.4.1 K-turn and K-loop variability

To compare the *Pyrobaculum* species to the rest of the Archaea, I created a comparison group of archaeal species contains representatives from across the archaeal phyla studied in this chapter. In cases where the box features could not be unambiguously assigned, the sRNA was not included in the analysis. Positions 1b1n, 2b2n, and 5b5n of the K-turn and K-loop were found to be significantly different in the *Pyrobaculum* compared to the rest of the Archaea using the Wilcoxon rank-sum test and a p-value cutoff of 0.05 with Bonferroni correction. The p-values for the K-turn 1b1n, 2b2n, and 5b5n positions are 1.48e-05, 3.06e-03, and 3.36e-03, respectively. These values are still significant with Bonferroni correction to account for multiple hypothesis testing (p-value cutoff with correction is 5e-03), although the p-values for the 2b2n and 5b5n are close to the cutoff so the significance is questionable. The K-loop p-values for the 1b1n, 2b2n, and 5b5n positions are 5.99e-04, 7.35e-04, and 2.61e-03, respectively.

The 1b1n and 2b2n positions are typically G:A and A:G pairs that stabilize the
Figure 4.8: Variability of K-motifs of C/D box sRNAs between the *Pyrobaculum* and those found in the rest of the Archaea. *Pyrobaculum* species have a greater number of possible base pair types at nearly all K-turn and K-loop positions compared to other species across the Domain Archaea. Using the Wilcoxon rank-sum test and Bonferroni correction, the *Pyrobaculum* K-turns and K-loops are significantly different from the rest of the Archaea in positions 1b1n, 2b2n, and 5b5n (p-values are 1.48e-05, 3.06e-03, and 3.36e-03, respectively). (A) Diagram of C/D box sRNA with K-turn notation marked. See Figure 3.6 caption for detailed description of notation. (B) Plot of *Pyrobaculum* K-turn and K-loop variability compared to other archaea (panarchaeal group). The diagram shows number of different base-pair types at each position of the K-turn or K-loop for each species. The *Pyrobaculum* group is composed of Pae, Par, Pca, Pis, Pne, and Pog. The panarchaeal group is composed of Ape, Afu, Hsp, Hvo, Iho, Mac, Mba, Mja, Mka, Mma_C5, Mmz, Neq, Nma, Pfu, Sma, and Tko. See Table 4.1 for species abbreviations.
K-turn and K-loop, so the difference in variability is somewhat surprising. K-turns that do not have G:A or A:G pairs do exist, so it is possible that the unusual *Pyrobaculum* C/D box sRNAs still form K-turns (see Section 2.2.1 for discussion of unusual K-turns). Another explanation is that protein-protein interactions may assist the formation of the RNP despite a poor K-loop structure [67]. In the *Pyrobaculum* the kink-turn binding protein L7Ae has conserved N-terminal and C-terminal flanking sequences that are not found outside of the family Thermoproteaceae and we hypothesize these sequences help bind L7Ae to the kink-turn and help it tolerate more variation in kink-turn sequence and structure (Figure 4.3.4.1).

I suspected that the small number of C/D box sRNAs in some species, such as the halophiles, could be skewing the results. In species where the number of C/D box sRNAs is small, the 16 possible base pair types cannot be represented. In the same token, species with larger numbers of C/D box sRNAs have a higher probability of exploring the sequence space. I split the species into two groups based on whether they had 25 C/D box sRNAs or greater. Comparing the 25 or greater number of sRNA group to the *Pyrobaculum* (Figure 4.10), only the 1b1n position of the K-turn showed statistical significance with Bonferroni correction. However, the p-value for the 1b1n position was 1.22e-03, so its significance is questionable. Thus, the *Pyrobaculum* do not appear to have more variable K-motifs in C/D box sRNAs compared to diverse archaeal species, at least for those with more than 25 C/D box sRNAs.

Although the K-motif variation of *Pyrobaculum* C/D box sRNAs turned out not to be significantly different from those found in other archaea, this analysis revealed
Figure 4.9: **Portion of alignment of L7Ae from diverse archaea.** The L7Ae protein from *Pyrobaculum* species and other closely related species have conserved N-terminal and C-terminal flanking sequences that are not found outside of the family Thermoproteaceae. *Pyrobaculum* species have a conserved 10-15 amino acid N-terminus extension and a 6 amino acid C-terminus extension. These extensions are also present in some of the other members of *Thermoproteaceae*, such as *Vulcanisaeta moutnovskia* and *Thermoproteus uzoniensis*. These regions are outlined by red boxes. Species from the family Thermoproteaceae are highlighted in light blue. Other colors indicate different clades, such as green for the species of Halobacteriales. The full alignment in FASTA format can be found in Appendix A.2.4.
Figure 4.10: Comparison of *Pyrobaculum* K-motifs of C/D box sRNAs with those found in rest of the Archaea split by number of C/D box sRNAs. Species are the same as those in Figure 4.8. Plot of *Pyrobaculum* K-turn and K-loop variability compared to other archaea (panarchaeal group). Blue group - *Pyrobaculum* species, pink group - species with less than 25 C/D box sRNAs, green group - species with greater than 25 C/D box sRNAs excluding the *Pyrobaculum*. There does not appear to be a significant difference in base pair variability in the *Pyrobaculum* species compared to species with more than 25 C/D box sRNAs.
that there is a distribution of K-motif variability across archaeal C/D box sRNAs.

4.3.4.2 Preferences in K-motif base pair types

Past studies have observed base pair preferences at each position that contribute to varying degrees to the structural features that define the K-turn structural motif [102, 169, 145]. Substitutions compared to the consensus are tolerated [102], but likely some base pairs may be deleterious to kink-turn formation.

In general, the base-pair types of the K-motifs matched the consensus, but there was notable deviation in the 3b3n position. If the box features match the consensus, the 3b3n position should be a U:U mismatch. However A:U, C:C, and G:C are also common (occurring more than 10% of the time). The G:C pair is consistent with previous findings; it can be inserted in the 3b3n position with little change in methylation activity [145]. It occurs as the prevalent 3b3n base pair of the K-loop in many species, including *P. islandicum* (16%), *A. pernix* (30%), and *I. hospitalis* (48%). Deviations from the U:U and G:C pairs as the prevailing type include the A:U pair as the prevalent type in the 3b3n base pair type in K-loops of *P. arsenaticum, P. islandicum, and P. oguniense*, and C:C is the prevalent type in the K-turn and K-loops of *P. aerophilum* and halophilic species.

The prevalent base pair at the 5b5n position is almost always A:G, typically at a rate of 25–50%. This is consistent with the consensus of the C or C’ box ending in an A and the finding in Chapter 3 where the 5n position, the nucleotide 5’ of the D or D’ box, is G a majority of the time. G:G, A:A, and A:U are also common base-pair types.
occurring more than 10% of the time. Since certain base pair types are favored and others are rare or absent at the 5b5n position, there possible there is some structural selection at this position. One other explanation is that there is protein recognition for an A at the end of the C or C′ box and a G 5′ of the D box or D′.

Variability in the formation of the kink-turn base pairs may reflect specificity of the L7Ae proteins in the different species, but also brings into question whether L7Ae can bind to these deviant K-motifs. Studies have found that protein-protein interactions between the core C/D box sRNA proteins (L7Ae, Nop5, and Fibrillarin) can compensate for poor binding to the K-loop and allow the formation of a functional sRNP [67, 145]. There are also several sRNAs in *I. hospitalis* identified from small RNA sequencing data that do not have recognizable C′ or D′ boxes (e.g. Iho sR02 and sR43). In some cases, the C′ or D′ boxes may not be needed. In the future, computational methods with only the C and D boxes modeled could detect more C/D box sRNAs.

### 4.3.5 C/D box sRNAs with K-turns instead of K-loops

In *H. volcanii*, an unusual C/D box RNA with a K-turn instead of a K-loop was found that targets elongator pre-tRNA^{Met} [92] (Figure 4.11A). Typically the spacer between the C′ and D′ is about 3–6nt long in archaeal C/D box sRNAs (Figure 4.11B). With the increase in C/D box sRNA annotations, I noticed that there is a larger number of sRNAs with spacers much longer than 6nt that could form K-turns with the C′ and D′ boxes. I investigated whether this form occurred in other archaeal species.

To look for a K-turn instead of a K-loop, I used criteria that a stem I of at
Figure 4.11: **C/D box sRNAs with K-turns instead of K-loops.** C and C' boxes highlighted in yellow, D and D' boxes highlighted in blue. (A) First C/D box sRNA with a K-turn in place of the K-loop found by Joardar et al. (B) Example of an archaeal C/D box sRNA with a K-loop for comparison. This C/D box sRNA is from *Pyrobaculum aerophilum*. (C,D) Examples of the C/D box sRNAs I found that had K-turns instead of K-loops. These examples represent the C/D box sRNAs with the longest stem I’s that I found. (C) Examples from Class I methanogens. Stem I and stem II of the K-turn is labeled. Mma - *Methanococcus maripaludis*, Mka - *Methanopyrus kandleri* (D) Examples from the Desulfurococcales order. Sma - *Staphylothermus marinus*, Ape - *Aeropyrum pernix*, Iho - *Ignicoccus hospitalis*. Structures created with VARNA. 

113
Table 4.3: **Number of C/D box sRNAs with K-turns instead of K-loops.** Some archaeal species have C/D box sRNAs with K-turns forming between the C′ and D′ boxes instead of a K-loop. K-turns have an additional stem, called stem I, that K-loops are missing (see Figure 4.11C). In this table, species are ordered from largest to smallest based on the number of C/D box sRNAs with a stem I length ≥3bp. The number of C/D box sRNAs with a stem I of 2bp is also shown, although stems of this length are less likely to form. *M. kandleri*, *S. marinus*, and *A. pernix* are the species with the most C/D box sRNAs with a stem I of length ≥3bp or 2bp. Of the C/D box sRNAs of this form, 65/127 (51%) occur in the Desulfurococcales order.

<table>
<thead>
<tr>
<th>Species</th>
<th>Order</th>
<th>sRNAs with ≥3bp stem I</th>
<th>sRNAs with 2bp stem I</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M. kandleri</em></td>
<td>Methanopyrales</td>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td><em>S. marinus</em></td>
<td>Desulfurococcales</td>
<td>21</td>
<td>8</td>
</tr>
<tr>
<td><em>A. pernix</em></td>
<td>Desulfurococcales</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td><em>M. maripaludis S2</em></td>
<td>Methanococcales</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td><em>M. maripaludis C5</em></td>
<td>Methanococcales</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td><em>I. hospitalis</em></td>
<td>Desulfurococcales</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td><em>M. jannaschii</em></td>
<td>Methanococcales</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><em>T. kodakaraensis</em></td>
<td>Thermococcales</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td><em>M. hungatei</em></td>
<td>Methanosarcinales</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td><em>A. fulgidus</em></td>
<td>Archaeoglobales</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>N. maritimus</em></td>
<td>Nitrosopumilales</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>S. acidocaldarius</em></td>
<td>Thermoproteales</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><em>M. acetivorans</em></td>
<td>Methanosarcinales</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>M. barkeri</em></td>
<td>Methanosarcinales</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

least 3 bp could form with a loop of at least 3 nt (see Figure 4.11C for placement of stem I). I found that a K-turn forming between the C′ and D′ appears mostly in the Class I methanogens and the Desulfurococcales order (Figure 4.11C,D). There is also one sRNA in *T. kodakaraensis* and one in *M. hungatei* that fit this criteria. The highest prevalence of this form occurs in *S. marinus* and *A. pernix* at approximately one-third of all sRNAs (Table 4.3). The longest stem occurs in Mma sR103 at 11bp. In other species, the spacers are either too small or do not have possible stems longer than 2 bp.

Relaxing the requirement of the stem to be at least 2bp instead of 3bp, more
Figure 4.12: **Models of C/D box sRNP formation.** Different proposed models of C/D box sRNP formation. Reproduced from [22]. Nop5 - blue, fibrillarin - orange, L7Ae - yellow. (A) Docking of *M. jannaschii* L7Ae and *P. furiosus* Nop5-Fibrillarin crystal structures onto the isodensity map created from single-particle electron microscopy map of a reconstituted *M. jannaschii* C/D RNP (B) Proposed C/D box di-sRNP model based on (A) in comparison to the conventional model of C/D box sRNP formation in (C). In the conventional model, one C/D box sRNA is bound by two L7Ae proteins, two Nop5 proteins, and two Fibrillarin proteins. In the di-sRNP model, there are two C/D box sRNAs and double the number of each protein that compose the sRNP.

C/D box sRNAs of this form were found in *M. kandleri, S. marinus, A. pernix, I. hospitalis, T. kodakarensis,* and *M. hungatei* (Table 4.3). One sRNA of this form was found in each of the following species: *N. maritimus, A. fulgidus, M. acetivorans,* and *M. barkeri.* Since this form of the C/D box sRNA appears to be more prevalent in certain orders, it might simply be more tolerated in some species.

If this extra stem does has a function, it could also act as a binding site for other proteins or provide added stability to the sRNA. Depending on the model of C/D box sRNP, the longer stem could influence the assembly of the sRNP. (Figure 4.12). With the conventional sRNP model (Figure 4.12C), L7Ae nucleates the assembly of the RNP by binding to K-turn and K-loop. This complex is then bound by Nop5 [116, 123]. In the di-sRNP model (Figure 4.12B), L7Ae still binds to the K-motifs, but Nop5 may play a more active role in positioning the sRNAs [22]. The di-sRNP model is based
on electron microscopy maps where the location of the sRNA is ambiguous, so it is unclear which part of the sRNA interacts with Nop5. Whether or not these extra stems have influence on the function or assembly of the sRNP, the possible length of the stem should be taken into account in future computational models.

### 4.4 Discussion

In this chapter I improved annotations for species across the Archaea, including those with no published C/D box sRNA annotations, using a new model and small RNA sequencing data. I present a comprehensive view of C/D box sRNAs across the Domain Archaea in terms of abundance, structure and box feature variation from this improved annotation set.

In general, I increased the number of annotated C/D box sRNAs in crenarchaeal thermophiles by 30%. The final annotation set, which includes curated noscan predictions, indicates that most thermophiles and hyperthermophiles have more than 50 C/D box sRNAs. In mesophilic species, such as the halophiles and Class II methanogens, I did not find very many C/D box sRNAs. These findings agree with the hypothesis that an abundance of 2′-O-methyls assists growth at higher growth temperatures. Halophilic rRNA is not known to have many modifications [77], so these species may not need many C/D box sRNAs.

Two alternative hypotheses for the lack of C/D box sRNAs in mesophilic species are (1) C/D box sRNAs exist in these species that are not detected by current
models and (2) ribose methylation in these species is primarily catalyzed by site-specific methyltransferases. The discovery of multiple C/D box sRNAs with K-turns instead of K-loops indicates that there may be C/D box sRNAs that deviate from known variation of archaeal C/D box sRNA features, so there may be C/D box sRNAs that I did not detect. In regards to the second hypothesis, homologs of methyltransferases that can catalyze the formation of 2′-O-methyls independent of C/D box sRNAs have been found in \textit{H. volcanii} \[77\], so these methyltransferases may also occur in methanogens and other mesophilic species.

In conflict with the hypothesis that the number of 2′-O-methyls is positively correlated with growth temperature are my results in \textit{A. fulgidus} and \textit{M. jannaschii}. These species have ten or fewer C/D box sRNAs, suggesting that they do not have many ribose methylations despite the fact that they are hyperthermophiles. However, it is still possible that these two species have site-specific methyltransferases that carry out additional ribose methylation.

I rarely found guides conserved outside of phylogenetic orders. Typically one or zero guide families in an order had homologs outside of the order. If a guide was conserved beyond its order, it targeted a functionally important region of ribosomal RNA. These result confirm conclusions from past studies that functionally important regions of the ribosome, rather than particular sites, are targeted for methylation \[51\]. Despite the rapid evolution of C/D box sRNA guides, regions of modification are consistent.

I also investigated the structure of C/D box sRNAs and found that the kink-turn is more variable in some species and, as previously mentioned, that some C/D box
sRNAs have two K-turns instead of a K-turn and K-loop. These structural variations should be accounted for in future computational models.

4.5 Methods and Materials

4.5.1 Small RNA sequencing libraries

Library preparation from purified RNA was carried out by David Bernick, Julie Murphy, and Andrew Smith as described in Section 3.4.5. Total RNA was either purified from cells cultured in-house at the indicated growth stage, or we used total RNA sent by collaborators (Table 4.2). The small RNA, tRNA-5S, and 5S-500nt fractions were obtained by running total RNA on denaturing PAGE and excising the proper region. Libraries were sequenced on Illumina HiSeq 2000 at UC Davis to produce 2 × 75 nt paired-end sequencing reads.

4.5.2 Computational prediction of C/D box sRNAs

The final covariance model based on Pae C/D box sRNAs from Chapter 3 was used to scan whole genomes with the cmsearch program from the Infernal v1.1 package. Previous annotations were culled if they had greater than 80% overlap with predicted mRNAs (Genbank RefSeq). If any of these culled annotations were overlapped by small RNA sequencing data, they were manually re-added to the set.
4.5.3 Detection of homologs

Homologous C/D box sRNAs were found using BLAST [30]. Sequences of sRNAs were used as input to blastn to search whole genomes. The following parameters were used: -word_size 7, -evalue 0.001, -penalty -1, -reward 1, -gapopen 0, -gapextend 2. Hits were checked for box features and guide similarity to the query.

4.5.4 Calculation of Wilcoxon rank sum values

The wilcox.text function in the R software package [156] was used to calculate p-values for the Wilcoxon rank sum test. Default parameters were used (continuity correction was automatically applied).

4.5.5 Prediction and Analysis of Methylation Targets

The antisense prediction program (findAntisense.py) developed in Chapter 3 was used to predict methylation targets (as described in Section 3.4.3). The same parameters were used (--minLen=10 --maxMis=2 --maxGU=2 --open=0 --extend=0). Ribosomal RNA used for predictions can be found in Appendix A.2.
Chapter 5

Non-canonical H/ACA box sRNAs are Widespread Across the Archaea

5.1 Abstract

Pseudouridine is a ubiquitous post-transcriptional modification that adds stability to RNA and is found in functionally important regions of non-coding RNAs. In archaean, H/ACA box sRNAs guide the pseudouridylation of target nucleotides in ribosomal and transfer RNAs. Despite the fact that proteins associated with H/ACA box RNAs is found in all archaean species, very few H/ACA box sRNAs have been confirmed in archaean, and in particular crenarchaeal species. Non-canonical forms of archaean H/ACA box sRNA structures have been reported that are missing either the lower stem, the ACA box, or both, suggesting that the missing annotations are due to the lack of proper models for these forms.
I present a comprehensive look at H/ACA box sRNAs in nearly 50 archaeal species, with 87 and 144 new annotations of the canonical and non-canonical forms, respectively. These new annotations were predicted using covariance models and are supported by conservation and small RNA sequencing data. These models also found new canonical H/ACA box sRNAs in species that have been well studied, such as *Pyrococcus furiosus*. In general, crenarchaeal species have a larger proportion of non-canonical forms than euryarchaeal species.

I also created new guide extraction software that can process the non-canonical forms to facilitate prediction of pseudouridylation targets.

### 5.2 Introduction and Background

#### 5.2.1 Features of Pseudouridine and H/ACA box sRNAs

Pseudouridine (Ψ, 5-ribosyluracil) is one of the most common post-transcriptional modifications in non-coding RNA and clusters in functionally important regions of ribosomal RNA (rRNA), small nuclear RNA and transfer RNAs (tRNA) [162, 36]. Isomerization of uridine to pseudouridine results in an increase of hydrogen bonding capacity of the nucleoside (Figure 5.1), allowing pseudouridines to stabilize RNA structure and enhance base-stacking of RNA duplexes [36, 97]. This modification is likely necessary for proper ribosome function as removal of pseudouridylation capability via mutations in the pseudouridine synthase Cbf5 reduces growth rate in yeast [207].

Eukaryotes and archaea use ribonucleoprotein (RNP) complexes composed of
Figure 5.1: **Diagram of uridine and pseudouridine.** Reproduced from [36]. Differences in hydrogen bonding capacity and linking of nitrogenous base to ribose between uridine (U) and pseudouridine (Ψ, 5-β-D-ribofuranosyluracil). Uracil (left) is isomerized to pseudouridine (right) when the base is rotated along the N₃-C₆ axis (circular arrow), resulting in the linking position to ribose to change from N₁ to C₅ (arrowhead). Isomerization increases hydrogen capacity of the nucleoside, as uridine has one hydrogen bond acceptor (a) and one donor (d), while pseudouridine has one acceptor and two donors.

four core proteins and H/ACA box small RNAs (sRNAs) to precisely target ribonucleotides for pseudouridylation [116]. The structure of H/ACA box sRNAs and associated proteins (Cbf5, Gar1, L7Ae, and Nop10) are homologous between Eukarya and Archaea. Canonical H/ACA box sRNAs are hairpin RNAs with two bulges and an ACA element at the 3′ end (Figure 5.2A). In some species the consensus sequence of the ACA element is ANA [14, 17]. A crystal structure of an archaeal H/ACA box sRNA indicates that the ACA box and the lower stem are bound by the PUA domain of Cbf5 [112]. The upper bulge near the apical loop is a result of a kink-turn that forms in the upper stem. The lower bulge is the pseudouridylation pocket, which contains the guide sequence complementary to the target. The guide sequence is bipartite, with the
Figure 5.2: **Diagrams of Canonical and non-canonical forms of archaeal H/ACA box sRNAs.** The kink-turn motif is shown in yellow. The nucleotide that is pseudouridylated is represented by Ψ in a green hexagon. If the RNA has an ACA sequence, a feature typically found in H/ACA box RNAs, it is indicated in light blue. Dashed lines indicate if feature is missing in some H/ACA box sRNAs. (A) Canonical archaeal H/ACA box sRNA. (B) Non-canonical H/ACA box sRNAs. Three forms are represented: missing ANA box, missing lower stem, or both.

5′ and 3′ regions separated by the upper stem and apical loop. When the target RNA binds to the H/ACA box sRNA, the first ribonucleotide at the base of the upper stem is pseudouridylated (Figure 5.2). Often the target uridine does not pair with the guide. Eukaryotic H/ACA box sRNAs can exist as double guides with two hairpins separated by an H box (hinge sequence, consensus ANANNA). Most archaeal H/ACA box sRNAs occur as single hairpins, but two and three linked hairpins have been found in archaea.\[177\] [130].
5.2.2 Current state of H/ACA RNA annotations in the Domain Archaea

Archaeal H/ACA box sRNA annotations remain poor. Half of the species with annotations have only one or two H/ACA box sRNAs noted (Table 5.1) and most orders have annotations in only one or two species. Notably, no H/ACA box sRNA annotations exist in Class II methanogens or the Thaumarchaea (Figure 5.3). Focused computational studies of archaeal H/ACA box sRNAs have expanded annotations mostly in euryarchaeal species, particularly for Pyrococcus furiosus, Thermococcus kodakaraensis, and Methanocaldococcus jannashchii [129, 181, 77].

Small RNA studies have had mixed results in annotating archaeal H/ACA box sRNAs. The first four archaeal H/ACA box sRNAs were identified in a small RNA study of Archaeoglobus fulgidus by Sanger sequencing of cDNA [177] and one H/ACA box sRNA was identified in Sulfolobus solfataricus using similar methods [206]. More recently, high-throughput sequencing of small RNAs has produced annotations in Methanopyrus kandleri [174], Nanoarchaeum equitans [157], Pyrococcus abyssi [182], and from our lab, species of the Pyrobaculum genus [17]. However, not all small RNA studies have reported new H/ACA box sRNAs [90, 204]. The presence of the four proteins of archaeal H/ACA box sRNPs across the archaeal domain [63, 180, 69] and the number of mapped pseudouridines in archaeal rRNA [130] indicate that we should find more H/ACA box sRNAs in more archaeal species so the lack of archaeal H/ACA box sRNAs is puzzling.
<table>
<thead>
<tr>
<th>Species</th>
<th>Phylum</th>
<th>Order</th>
<th>Canon, Non-canon Guides</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Aeropyrum pernix K1</em></td>
<td>Crenarchaea</td>
<td>Desulfurococales</td>
<td>1, -</td>
<td>131</td>
</tr>
<tr>
<td><em>Archeaoglobus fulgidus DSM 4304</em></td>
<td>Euryarchaea</td>
<td>Archeoglobales</td>
<td>6, -</td>
<td>177,163</td>
</tr>
<tr>
<td><em>Haloarcula marismortui ATCC 43049</em></td>
<td>Euryarchaea</td>
<td>Halobacteriales</td>
<td>2, -</td>
<td>77</td>
</tr>
<tr>
<td><em>Halobacterium salinarum R1</em></td>
<td>Euryarchaea</td>
<td>Halobacteriales</td>
<td>2, -</td>
<td>77</td>
</tr>
<tr>
<td><em>Halofexx volcanii DS2</em></td>
<td>Euryarchaea</td>
<td>Halobacteriales</td>
<td>2, -</td>
<td>77, 20</td>
</tr>
<tr>
<td><em>Halterubrum lucusprofundi ATCC 49229</em></td>
<td>Euryarchaea</td>
<td>Halobacteriales</td>
<td>2, -</td>
<td>77</td>
</tr>
<tr>
<td><em>Ignicoccus hospitalis KIN4-I</em></td>
<td>Crenarchaea</td>
<td>Desulfurococales</td>
<td>2, -</td>
<td>77</td>
</tr>
<tr>
<td><em>Metallosphaera sedula DSM 5348</em></td>
<td>Crenarchaea</td>
<td>Sulfolobales</td>
<td>1, -</td>
<td>131</td>
</tr>
<tr>
<td><em>Methanocaldoccoccus jannaschii DSM 2661</em></td>
<td>Euryarchaea</td>
<td>Methanococcales</td>
<td>8, -</td>
<td>131,129</td>
</tr>
<tr>
<td><em>Methanopira kandleri AV19</em></td>
<td>Euryarchaea</td>
<td>Methanopryales</td>
<td>11, -</td>
<td>174</td>
</tr>
<tr>
<td><em>Nanooarchaeum equitans Kin4-M</em></td>
<td>Nanoarchaeota</td>
<td>Nanoarchaeales</td>
<td>2, -</td>
<td>131</td>
</tr>
<tr>
<td><em>Natronomonas pharaonis DSM 2160</em></td>
<td>Euryarchaea</td>
<td>Halobacteriales</td>
<td>2, -</td>
<td>77</td>
</tr>
<tr>
<td><em>Pyrobaculum aerophilum str. IM2</em></td>
<td>Crenarchaea</td>
<td>Thermoproteales</td>
<td>5, 5</td>
<td>17</td>
</tr>
<tr>
<td><em>Pyrobaculum arsenaticum DSM 13514</em></td>
<td>Crenarchaea</td>
<td>Thermoproteales</td>
<td>5, 5</td>
<td>17</td>
</tr>
<tr>
<td><em>Pyrobaculum calidofontis JCM 11548</em></td>
<td>Crenarchaea</td>
<td>Thermoproteales</td>
<td>5, 5</td>
<td>17</td>
</tr>
<tr>
<td><em>Pyrobaculum islandicum DSM 4184</em></td>
<td>Crenarchaea</td>
<td>Thermoproteales</td>
<td>5, 5</td>
<td>17</td>
</tr>
<tr>
<td><em>Pyrobaculum neutrophilum comb. nov</em></td>
<td>Crenarchaea</td>
<td>Thermoproteales</td>
<td>5, 5</td>
<td>17</td>
</tr>
<tr>
<td><em>Pyrobaculum aquimense TE7</em></td>
<td>Crenarchaea</td>
<td>Thermoproteales</td>
<td>5, 4</td>
<td>17</td>
</tr>
<tr>
<td><em>Pyrococcus abyssi GE5</em></td>
<td>Euryarchaea</td>
<td>Thermococcales</td>
<td>11, 4</td>
<td>131,37,111,103,139,77</td>
</tr>
<tr>
<td><em>Pyrococcus horokoshii OT3</em></td>
<td>Euryarchaea</td>
<td>Thermococcales</td>
<td>11, -</td>
<td>131,37,111,103,139,77</td>
</tr>
<tr>
<td><em>Pyrococcus furiosus DSM 3638</em></td>
<td>Euryarchaea</td>
<td>Thermococcales</td>
<td>11, -</td>
<td>131,37,111,103,139,77</td>
</tr>
<tr>
<td><em>Sulfolobus acidocaldarii DSM 639</em></td>
<td>Crenarchaeae</td>
<td>Sulfolobales</td>
<td>1, -</td>
<td>131</td>
</tr>
<tr>
<td><em>Sulfolobus solfataricus P2</em></td>
<td>Crenarchaeae</td>
<td>Sulfolobales</td>
<td>2, -</td>
<td>131,206</td>
</tr>
<tr>
<td><em>Sulfolobus tokodaii str. 7</em></td>
<td>Crenarchaeae</td>
<td>Sulfolobales</td>
<td>1, -</td>
<td>131</td>
</tr>
<tr>
<td><em>Thermococcus kodakaranensis KOD1</em></td>
<td>Euryarchaea</td>
<td>Thermococcales</td>
<td>17, -</td>
<td>139,77</td>
</tr>
</tbody>
</table>

Table 5.1: **Previously annotated archaeal H/ACA box sRNAs.** Number of canonical and non-canonical sRNAs are indicated in the “Guides” column (canonical, non-canonical). If there are no annotations in a category, it is indicated by “-”. Note that most of the annotations are in *Thermococcus kodakaranensis* and the *Pyrococcus* and *Pyrobaculum* genera. The number of guides is listed, so if an annotation is a double hairpin and thus has two guides, it counts as two.
Figure 5.3: **Previously annotated archaeal H/ACA box sRNAs mapped onto a phylogenetic tree.** Tree is a simplified version from [29]. Species with H/ACA box sRNA annotations are listed. If three or more species from a genus are listed, the genus and number of sRNAs typically found in each species is listed. Bolded clades are major phyla of the Domain Archaea. Most archaeal H/ACA box sRNA annotations are in the Thermococcales order and the *Pyrobaculum* genus. In contrast, most other archaeal species have very few H/ACA box sRNA annotations. No H/ACA box sRNAs have been reported in the Thaumarchaeota or the Class II methanogens. Very few H/ACA box sRNAs have been reported in the Halobacteriales and Archaeoglobales orders. Other than the *Pyrobaculum* species, very few H/ACA box sRNAs annotations exist in crenarchaeal species. Before the study by Bernick *et al.* [17] and this study, the dearth of H/ACA box sRNA annotations across most of the Archaea was perplexing.
Work by Bernick et al. on the *Pyrobaculum* genus suggests that the lack of H/ACA box sRNA annotations in archaea may be the result of non-canonical forms [17]. Bernick et al. identified 10 H/ACA box sRNA families conserved within the genus and most have transcription data supporting their expression in the four species with small RNA sequencing. Some members of this RNA family lack some of the typical conserved sequence features of H/ACA box sRNAs (the 3′ terminal ACA sequence, lower stem, and 5′ guide region), which may account for the low number of H/ACA box sRNA annotations in archaea (Table 5.1). These small RNAs have guide complementarity to rRNA and tRNA targets at conserved pseudouridine positions and the rRNA modifications were confirmed by CMCT assays. A later study found these non-canonical forms in *Pyrococcus abyssi* [182], which provided evidence that these forms are not unique to *Pyrobaculum* species.

Hypothesizing that non-canonical forms are found in other species, I created computational models of the non-canonical forms (Figure 5.2B). I present a comprehensive look at H/ACA box sRNAs across 48 archaeal species, with 87 and 144 new annotations of the canonical and non-canonical forms, respectively. Assisted by Matthew Speir, I created these new annotations from covariance model predictions supported by conservation and small RNA sequencing data. Surprisingly, I found new H/ACA box sRNAs in species that have been well studied, including *Pyrococcus furiosus*. After adding these new annotations to the known H/ACA box sRNAs, I found that nearly half of all archaeal H/ACA box sRNAs are of the non-canonical form.
5.2.3 Other computational prediction methods for archaeal H/ACA box sRNAs

Few methods exist to predict archaeal H/ACA box sRNAs. One method is to create models as input to MilPat software, which models RNA motifs and structure using constraint networks [181]. Thébault et al. used this method to predict known sRNAs in Pyrococcus species and five new candidates in M. jannaschii. Some of the candidates in M. jannaschii had already been verified as small RNAs in other studies, but not classified as H/ACA box sRNAs. With this method, predictions can be filtered based on if potential RNA interactions (in this case, pseudouridylation targets) can be found. However, even with this filter, a large number of false positives are still reported. Of the 118 candidates predicted for M. jannaschii, Thébault et al. only considered five high confidence.

Muller et al. created a pipeline using the profile-based ERPIN software [129, 72]. Their method was successful in predicting known H/ACA box sRNAs in Pyrococcus species (7 sRNA families) and new candidates in Thermococcus kodakarensis (16 sRNA families). A rotation student in our lab, Wendy Lee, used the method designed by Muller et al. to predict canonical and non-canonical H/ACA box sRNAs in seven crenarchaeal species and one euryarchaeal species. Lee used the Pyrobaculum H/ACA box sRNAs to create ERPIN profiles. The resulting models yielded only five plausible candidates in four of the species. Lee’s results highlighted a limitations of ERPIN; often models for subfamilies must be made for more sensitive prediction [107]. Another limitation of
ERPIN is that it cannot tolerate insertions of arbitrary length at some positions of the model[107]. For example, the variation in length of the stem I and II of the kink-turn in H/ACA box sRNAs cannot be modeled by ERPIN. Despite the limitations of Muller’s method, prior to this study it was the most successful computational search method for archael H/ACA box sRNAs.

5.2.4 Collaboration in this study

Two of my mentees, Matt Speir and Ashmitha Rajendran, assisted me in this work. Matt Speir was a major contributor. He made some of the covariance models under my direction and he helped me evaluate our computational predictions. Speir and Rajendran both wrote parts of the target prediction program that iterates through possible targets and finds the best scoring targets.

Small RNA sequencing data generated by the Lowe lab and collaborators is the same as what was used in Chapter 4 and described in Table 4.2.

5.3 Development of covariance models

In this section I describe the iterative development of four core models (and clade specific versions of each model) that were used to represent the different forms of archael H/ACA box sRNAs. First, I describe the models and their features for reference. Next, I describe their development to explain why each model is necessary and why one covariance model is not sufficient to efficiently detect all archael H/ACA box sRNAs.
5.3.1 Four core models of H/ACA box RNAs in the Archaea

Matthew Speir created some of the covariance models described in this section under my guidance and helped scan genomes with the covariance models. Speir and I created four general models: (1) canonical, (2) non-canonical without ANA box, (3) non-canonical without the lower stem, and (4) non-canonical without the lower stem or ANA box (Figure 5.4A–D). In all models the kink-turn structure is included and the critical G:A base pairs are explicitly modeled (kink-turn structure described in Section 2.2.1). The canonical model is composed of the apical loop, upper stem that has the kink-turn, pseudouridylation pocket, lower stem, and ANA box. The non-canonical models are missing the lower stem, the ANA box, or both. In all models, guides are modeled by insertion states, so the sequence and length can vary. This is also true of the other single-stranded regions of the H/ACA box sRNA, such as the apical loop and nucleotides in the bulge of the kink-turn below the upper stem.

For each model, we created versions specific to the two main archaeal phyla: Crenarchaeota and Euryarchaeota. Species outside of these phyla (Nitrosopumilus maritimus and Nanoarchaeum equitans) were scanned with both models. Using the Infernal software package [133], covariance models were built from structural alignments of known archaeal H/ACA box sRNAs and calibrated (see Method section 5.6.2 for more detail). Genomes were then scanned with these models.
Figure 5.4: Core models of canonical and non-canonical forms of archaeal H/ACA box sRNAs. The portion of the kink-turn (K-turn) motif modeled is shown in yellow. Modeled base-pairs are indicated by black lines. ‘R’ in the kink-turn represents the purine ‘cap.’ Guide region is shown in green. Sometimes the 5′ portion of the guide is missing (indicated by dashed lines). (A) Canonical archaeal H/ACA box sRNA. (B) Non-canonical H/ACA box sRNA without ANA box. (C) Non-canonical H/ACA box sRNA without lower stem. (D) Non-canonical H/ACA box sRNA without lower stem or ANA box. Note that this model does not model the lower stem, but will partially capture some hits that have lower stems but no ANA box. Although this model can find some of the same hits from the model in (B), the other model is still used to more accurately predict the termini of H/ACA box sRNAs and classify their form as an H/ACA with a lower stem but no ANA box. (E) Part of the alignment used to generate the canonical model for crenarchaeal species. Guide region is modeled as any base (‘n’). This alignment is input in the form of a Stockholm file. The bottom line is the SS_cons line, where open and close parentheses indicate base pairs and periods indicate unpaired regions.
5.3.2 Models and parameters for Infernal search program

I decided to use the four H/ACA box sRNA models and chose the parameters for cmsearch, the Infernal program that scans a database with a covariance model, after running tests and analyzing predictions. I evaluated predictions based on conservation, good structure, predicted promoter evidence, less than 80% overlap with protein-coding genes, and small RNA sequencing data. I decided to use the glocal (-g) option to prevent large insertions or deletions in predictions compared to the models. I also found that turning off the hidden markov model (HMM) filter (-nohmm option) yielded more hits that I believed to be real H/ACA box sRNAs. Often the HMM filter is used to reduce search time, but it also filters out more hits. This decision also lead me to use multiple non-canonical models, because Infernal automatically turns off the search for truncated hits when the HMM filter is turned off (Infernal v1.1 User’s Guide, http://selab.janelia.org/software/infernal/Userguide.pdf). Using the four general covariance models also helped me determine whether predictions had a lower stem or ANA box.

5.3.3 Iterative development of covariance models

Iterative development of the models was necessary since the only non-canonical H/ACA box sRNAs known were in the Pyrobaculum genus (the study reporting non-canonical forms in Pyrococcus abyssi [182] came out after I started). The Pyrobaculum H/ACA box sRNAs are highly similar in sequence, so the likelihood of creating a general archaeal model from these sRNAs was small. To find more H/ACA box sRNAs that
would introduce more sequence and structural diversity into the general models, Speir
and I created models for each of the 10 *Pyrobaculum* guide families and scanned closely
related species that are also in the order Thermoproteales: *Thermoproteus uzonien-
sis*, *Thermoproteus tenax*, *Vulcanisaeta moutnovskia*, *Vulcanisaeta distributa*, and Cal-
divirga maquilingensis. We hand-curated predictions based on conservation, good struc-
ture, predicted promoter evidence, less than 80% overlap with protein-coding genes, and
small RNA sequencing data (only in *C. maquilingensis* and *T. tenax*).

With a larger, more diverse set of H/ACA box sRNAs, I created general co-
variance models for the canonical form and the non-canonical forms. Speir helped me
compile sRNA sets for each form and cull highly similar sequences to prevent over-
training. I converted each set into a structural alignment, and the guide regions were
generalized to be composed of any base (‘n’ in the alignment) (Figure 5.4D). Finally, I
created the new, general covariance models from these alignments.

Speir and I used the general models to scan a wide variety of archaeal species.
The models were generally successful in crenarchaeal species, detecting approximately
5–10 H/ACA box sRNAs in each species. The models also predicted nearly all of
the previously annotated H/ACA box sRNAs. The previous annotations that were
not detected either were canonical forms with K-loops (discussed in Section 5.4.3.1) or
overlapped tRNAs or mRNAs. Unfortunately, the models had mixed results in species
of the Euryarchaeota phylum. For the two euryarchaeal species with the best H/ACA
box sRNA annotations, *Pyrococcus furiosus* and *Thermococcus kodakaraensis*, Speir and
I found approximately 5–10 new sRNAs, but did not find many previously annotated
ones. The models also performed poorly in methanogenic and halophilic species and only yielded 1–3 plausible predictions. Very few H/ACA box sRNAs have been predicted in methanogens and halophiles (typically 0–3), but I knew that the models may have been overtrained on crenarchaeal species. Speir and I made euryarchaeal-specific models for the non-canonical forms by truncating known H/ACA box sRNAs. This model was successful, and we used the newly predicted sRNAs to create models that contained only known non-canonical forms, rather than manufactured ones. We believe that these sRNAs are real, based on small RNA sequencing data, conservation with closely related species, promoter prediction, and less than 80% overlap with mRNAs. These models predicted 3–15 more sRNAs in each species.

Idiosyncrasies of model versions and other form-specific models are discussed in the Results section.

5.4 Results

5.4.1 Non-canonical forms are widespread across the archaeal domain

Speir and I scanned species from thirteen of the major archaeal orders where we had sequencing data and found non-canonical forms of H/ACA box sRNAs in nearly all species we scanned (Figure 5.5). I did not scan any species from Aigarchaeota phylum (*Caldarchaeum subterraneum* is the only species), the Korarchaeota (*Koarchaeum cryptofilum* is the only species), or the Thermoplasmales order because of the lack of sequencing data and dearth of closely related genomes for comparative genomics. With
Figure 5.5: Summary of archaeal H/ACA box sRNA annotations. Only species scanned in this study are shown. Species are grouped by order, which is listed on the left. Predictions for each species are shown as a stacked bar chart. Stars indicate species with small RNA sequencing data. There are 110 previously annotated canonical forms (dark blue, including species not shown), and 49 previously annotated non-canonical (light blue). There are 87 new canonical forms (dark green) and 144 new non-canonical forms. Approximately 87 candidates are questionable, nearly all of non-canonical forms.
the addition of the new annotations, I find that non-canonical forms now comprise half of all archaeal H/ACA box sRNAs (193/390, 49.5%). Most of the new annotations are either conserved or have small RNA sequencing evidence. We found homologs of candidates in other species by using BLAST [30] and aligned them using MAFFT [93]. Generally one or two guides are conserved across an order, but most are conserved only within their genus or with closely related genera. Annotations that have good structure, but do not have sequencing data or conservation, I have labeled as questionable.

In general, crenarchaeal species have a higher proportion of non-canonical forms and the euryarchaeal species have a higher proportion of canonical forms. This result correlates with the findings in the Pyrobaculum species [17] and the ease of finding H/ACA box sRNAs in the euryarchaea. On average, 68% of predicted H/ACA box sRNAs are of a non-canonical form in crenarchae species, and 27% in euryarchaeal species. An exception to the euryarchaeal pattern is in the Class II methanogens, where on average 64% of H/ACA box sRNA predictions are non-canonical. Up until this point, no H/ACA box sRNAs have been reported in these species.

Notably, no high confidence non-canonical H/ACA box sRNAs were found in the halophiles, although some new canonical H/ACA box sRNAs were predicted and most candidates in the “questionable” category are non-canonical. As discussed in Section 4.3.3, halophiles do not have many post-transcriptional modifications compared to other archaea [77] and may not need many sno-like RNAs. In addition, the pseudouridine synthase component of the H/ACA box RNP, Cbf5, has been shown to be non-essential [20] in H. volcanii. Instead, the independent pseudouridine synthase Pus10
is essential, suggesting that halophiles rely more on site-specific pseudouridine synthases than H/ACA box sRNA machinery.

Scans of species in the Thaumarchaea and Nanoarchaea phyla yielded few results. The sole member of Nanoarchaea, *N. equitans*, had two canonical and two non-canonical predictions. *N. equitans* has a reduced genome (0.49 Mb) and is an obligate symbiote with *I. hospitalis*, so it may not need many H/ACA box sRNAs. Two canonical predictions and one non-canonical prediction were found in *N. martimus* of the Thaumarchaeaa. One prediction from each form is conserved with *Cenarchaeum sybiosum*. *N. maritmus* also have very few C/D box sRNAs, so this species may also not use many H/ACA box sRNAs.

Initial scans of the Lokiarchaeal composite genome (504 contigs, Genbank Accession JYIM00000000) \[172\] did not yield any results. Even at a bitscore cutoff of 5, no hits were returned. As discussed in Section 2.3.3, Infernal v1.0 is slightly more sensitive than Infernal v1.1 due to algorithmic differences, so I rescanned the genomes using Infernal v1.0 software and gathered a few weak hits. Only three non-canonical H/ACA box sRNAs were found with weak structures (Figure 5.6). Since C/D box sRNAs were predicted (Chapter 4) in the genome and a homolog of Cbf5 was predicted (pseudouridine synthase, arCOG00987) we expect that there are H/ACA box sRNAs in the Lokiarchaeaa. The missing part of the genome (estimated at 8%) could contain some H/ACA box sRNAs. It is also possible that the lokiarchaeal H/ACA box sRNAs may not be detectable with our current models.
Conservation of H/ACA box sRNAs can extend through an entire order

Similar to C/D box sRNAs, many H/ACA box sRNAs demonstrate sequence conservation of guides only with species within the same genus or closely related genera. However, I found that some H/ACA box sRNAs were conserved throughout an entire order. Examples of these patterns can be seen in the Desulfurococcales order of the Crenarchaeota phylum (Table 5.2). The sR205, sR206, and sR208 families are only conserved within the *Staphylothermus* genus, but the sR201 and sR204 families are conserved in all eight species of the order in this study.
Table 5.2: H/ACA box sRNA homologs of the Desulfurococcales order. Rarely are sno-like sRNA conserved throughout an entire order. In the Desulfurococcales order, two canonical and two non-canonical H/ACA box sRNAs are conserved in at least seven of the eight species scanned in this study. “X” represents the presence of a member of the sRNA in that species. For species tree see Figure 5.7A. Only sRNAs that are conserved between at least two species are shown.

The sR204 family provides an example of a non-canonical H/ACA sRNA conserved throughout an entire order (Figure 5.7). There is some variance in the length and sequence of the stems, but the core guide sequence of approximately 15nt is identical through the entire family (Figure 5.7A). The ANA box occurs in five of the eight species. In the species where we have sequencing data, this sRNA is expressed. The target of this family is 23S U2287 (A. pernix numbering) in helix 69, a core part of the interface between the ribosomal subunits. In eukaryotes and archaea, this region is known to have many pseudourididine modifications based on biochemical assays [74, 77]. Further analysis of the conservation of H/ACA box sRNAs within clades and throughout the archaea will give us a better understanding of which pseudouridines are essential in archaea and the evolution of archaeal H/ACA box sRNAs.
Figure 5.7: Example of conserved non-canonical H/ACA box sRNA. (A) Alignment of the sR204 family of the Desulfurococcales order in the Crenarchaeota phylum. Yellow - Kink-turn, Pink - upper stem, blue - intermediate stem (lower part of upper stem), Red - guide, Green - ACA box. (B) Read coverage for A. pernix sR204. Black arrow above the read coverage track represents the H/ACA box sRNA. Colored boxes correspond to those in (A). (C) Secondary structure of A. pernix sR204. Target to 23S rRNA shown and uridine modified shown as Ψ. (D) Location of predicted modification of A. pernix sR204. Guide shown in red.
5.4.3 Model idiosyncrasies, versions and forms

Although the models helped me greatly expand archaeal H/ACA box sRNA annotations, I found the need for clade-specific forms and models (unlike for C/D box sRNAs). Often there were trade-offs for different model versions and methods, including (1) switching to Infernal v1.1 from v1.0 reduced false positives but lowered sensitivity (such as for the Lokiarchaea) and (2) different models for non-canonical forms increased sensitivity, but would sometimes detect canonical forms, and other non-canonical forms necessitating hand-curation.

With a larger set of archaeal H/ACA box sRNAs, we can see more of the structural variation that exists. Here I discuss two other models I briefly tested to improve sensitivity for more unusual forms, such as ones with K-loops instead of kink-turns.

5.4.3.1 H/ACA box sRNA K-loop model for canonical form

Although canonical H/ACA box sRNAs were not the focus of this study, I checked how well the canonical covariance model predicted previously annotated H/ACA box sRNAs. I noted that a sub-type of previously annotated H/ACA box sRNAs was rarely predicted by our canonical models. Approximately half of the annotations our models missed had a K-loop instead of a K-turn in the upper stem (Figure 5.8). Species with previously annotated H/ACA box sRNAs that were not detected by my models were A. fulgidus (1 not detected), M.kandleri (1), T. kodakaraensis (4), and the three Pyrococcus species with H/ACA box sRNA annotations (3 in P. furiousus). The com-
putational models by Muller et al. [129, 131] and Thebault et al. [181] did not explicitly model base-pairing above the kink-turn G:A base pairs (stem I, Figure 2.3), in effect modeling a K-loop for the upper stem (Figure). Since my canonical model includes stem I, the sensitivity for K-loop forms was reduced. The only K-loop sRNAs that my canonical models predicted were in *N. equitans* (Figure 5.8). To create a K-loop model, I removed stem I from the euryarchaeal canonical structural alignments and created a new covariance model.

The new model picked up about half of the missing K-loop H/ACA box sRNAs. One sRNA in *A. fulgidus*, one in *T. kodakarensis*, and all three *P. furiosus* K-loop sRNAs were picked up by this model. Upon closer inspection of the K-loop sRNAs still not predicted, I found that the other *T. kodakaraensis* K-loop sRNAs are less likely to be real because they are completely overlapped by mRNAs.

I also noted that most of the other H/ACA box sRNAs with stem I were predicted by the K-loop model. In some cases the K-loop model did not predict all of the annotations the original canonical model could, such as Mj-3 in *M. jannaschii* [181]. However, the success of this model to pick up H/ACA box sRNAs with stem I in addition to the K-loop H/ACA box sRNAs indicates that a more general model that encapsulates both forms may be possible. The presence of canonical K-loop forms also suggests that K-loops might be present in the non-canonical forms as well. I did not create any non-canonical models with K-loops, but this line of inquiry should be pursued in future studies.
Figure 5.8: **K-loop H/ACA box sRNAs.** Except for the Neq sRNAs, these H/ACA box sRNAs were not detected by the original canonical models Speir and I created. Boxes indicate location of K-loop. These structures differ from other canonical H/ACA forms because they do not have an additional stem above the G:A pairs. The Neq sR201 and sR202 are the only two canonical K-loop H/ACA box sRNAs detected by the original canonical models. It is unclear why some H/ACA box sRNAs of this form were picked up by the models and not others. Afu-46 was detected in a cDNA study of *A. fulgidus* [177], Mka- Nr-5 from a small RNA study in *M. kandleri* [174], and Tko21 (RC1) in a computational study [129]. The Neq sRNAs were found in this study.
Guide extraction and target prediction in H/ACA box sRNAs is not as simple as with the C/D box sRNAs since the H/ACA box guide is bipartite. Sometimes the guide can extend into the upper or lower stem (Figure 5.9). Opening a stem can also result in another target, as confirmed in the study of *Pyrobaculum* H/ACA box sRNAs (Figure 5.9B,C) \[17\]. I wrote a guide-extraction program that can output multiple guides based on opening the stem (Section 5.6.4). However, when testing the guide-target interaction prediction program, I did not find the same targets as were reported by Bernick *et al.* for the *Pyrobaculum* H/ACA box sRNAs \[17\]. The targets reported for the *Pyrobaculum* H/ACA box sRNAs were often found by reverse search; conserved positions of pseudouridylation were compared against the guides (personal communication, David Bernick). The targets I predicted may also be biological targets, but it is unclear without doing biochemical assays to confirm the presence of the pseudouridines at these positions. Until then, manual curation is necessary for high-confidence targets.

5.5 Discussion

With the predicted H/ACA box sRNAs identified in this study, I confirmed that the non-canonical forms are widespread and prevalent across the archaea. This annotation effort also allowed me to see clade-specific patterns of H/ACA box sRNA abundance and form across the Archaea. Among my observations I found (1) very few H/ACA box sRNAs in the halophiles, (2) that in crenarchaeal species, typically the
non-canonical forms outnumber the canonical forms found in a species, and (3) that canonical forms typically outnumber non-canonical forms in euryarchaeal species. With this comprehensive annotation set, more questions about archaeal H/ACA box sRNA biology can be answered.

In addition to providing an overall look at archaeal H/ACA box sRNAs, this study also has produced computational models that can be used to annotate newly sequenced genomes or small RNA transcriptomes. Annotating H/ACA box sRNAs will
be especially useful for studies identifying novel small RNAs with unknown function since these studies need to rule out known small RNAs. Future work includes making clade-specific models to improve sensitivity and creating non-canonical models with K-loops and studying the pseudouridylation targets of the H/ACA box sRNA predictions.

This study also raises question of whether the non-canonical forms are found in eukaryotes. Special models would need to be developed because eukaryotic H/ACA box sRNAs are not known to have kink-turns. Loss of this feature will likely result in increased false positives, so prediction may be more difficult. If non-canonical forms are not found in the eukaryotes, it suggests that this feature was lost in the Domain Eukarya since eukaryotes likely arose from an archaeal ancestor [172, 46].

5.6 Materials and Methods

5.6.1 Small RNA sequencing data

Small RNA sequencing data used to support computational predictions is described in the Methods of Chapter 4 and Section 4.2.

5.6.2 Covariance models

Covariance models were created by using a hand-curated multiple structural alignments of archaeal H/ACA box sRNAs as input to cmbuild from the Infernal v1.1 software package [133]. The covariance model was calibrated with cmcalibrate. I used cmsearch to scan the genomes and small RNA sequencing data using the glocal (-g)
and no HMM filter (--nohmm) options. I used the -global option to minimize large insertions.

Structural alignments (Stockholm format) used to create covariance models can be found in Appendix A.3. Models of each non-canonical form and clade-specific models are included.

5.6.3 Homolog detection

Homologs were detected using BLAST [30] and the parameters -task blastn, -word_size 7, -penalty -1, -reward 1, -gapopen 0, -gapextend 2, -evalue 1. Results were aligned using MAFFT [93] with default parameters and evaluated manually. The predicted homologs were compared against H/ACA box sRNA predictions to see if they had already been predicted as an H/ACA box sRNA. Otherwise I evaluated the structure of the prediction to see if it could form an H/ACA box sRNA structure.

5.6.4 Guide Extraction

The python program hacaGuideExtract.py was written to extract guides from canonical and non-canonical H/ACA box sRNA forms. Input can be Vienna or Stockholm format for multiple RNAs with the same format. The program will output multiple guides based on opening the base of the upper stem.

To extract the guide, the program first finds the kink-turn, allowing for GA or GG or AA base-pairs. This step is necessary to find the guide sequences in the non-canonical forms. Then it finds the pseudouridylation pocket based on the location of
the kink-turn. If multiple guide output is selected, the number of base-pairs the stem
should be opened must be specified by the user.

5.6.5 Target Prediction

Potential guide-target interactions were found with findAntisense.py (see Section
3.4.3 for more description). The following parameters were used: minimum guide-
target interaction length of 10bp, a maximum of 2 mismatches, a maximum of 3 G:U
base-pairs, and an open and extend gap penalty of 0. Parameters were chosen by ex-
amining the guide-target interactions of Pyrobaculum H/ACA box sRNAs.
Chapter 6

Survey of L7Ae-associated RNAs in

Thermococcus kodakaraensis

6.1 Abstract

Kink-turns are ubiquitous in all life and play roles in important, conserved biological roles, yet a comprehensive study of kink-turns in a transcriptome has yet to be completed. To find previously unknown kink-turn containing RNAs, I sequenced RNAs that co-immunoprecipitated with Thermococcus kodakaraensis L7Ae (RIP-seq). The archaeal L7Ae protein has a broad RNA-binding specificity, making it an ideal tool to survey the kink-turn motif and the types of ribonucleoprotein (RNP) complexes in which it is present. L7Ae can even bind to the K-loop variant that its eukaryotic homolog cannot [35]. In archaea, the RNA components of the ribosome, C/D box and H/ACA box RNP complexes, and RNase P contain kink-turns that are L7Ae binding
targets [127, 79, 40]. This multi-functional archaeal protein can also bind to kink-turns that are found in eukaryotes and bacteria [38, 82].

For my RIP-seq experiments, I used a strain of T. kodakaraensis that had a hemagglutinin-tag (HA-tag) on L7Ae and used HA antibodies for the immunoprecipitation. I used a strain that did not have a HA-tag on L7Ae as a control. From my sequencing results, I found various RNAs enriched compared to the control strain, including some mRNA fragments with potential kink-turns and antisense RNAs to mRNAs. Other known RNAs with kink-turns, such as C/D box sRNAs and H/ACA box sRNAs were also enriched.

6.2 Introduction and Background

6.2.1 RNA kink-turn structure and protein partners

The kink-turn is a secondary RNA structural motif that is characterized by stacked-sheared G:A base pairs and an asymmetric bulge flanked by two stems (Figures 2.3 and 2.4). Often within the bulge there is a flipped-out uridine and a purine that participates in the base-stacking of the G:A base-pairs [102]. The interactions of the G:A base-pairs help stabilize the structure and result in the approximately 60° ‘kink’ of the RNA backbone. A minimal kink-turn that has a loop instead of a stem above the asymmetric bulge is known as a K-loop [137, 79] (Figure 2.6). For a more detailed review of kink-turn structure and known variation see Section 2.2.1.

Klein et al. first defined the kink-turn in 2001 after looking at the crystal
structure of the *Haloarcula marismortui* 50S ribosomal subunit. They found seven kink-turns in 23S rRNA and nine ribosomal proteins that interact with these structures. Using the consensus sequence, they also found that the kink-turn also occurs in other RNAs, including human RNase mitochondrial RNA-processing (MRP) RNA, C/D box RNAs, and in the 5' UTR of ribosomal protein mRNAs. Prior to this study, a kink-turn had also been noted in yeast L30e mRNA based on NMR spectra, but was thought to be a unique structure \[120\].

Since their initial characterization, kink-turns have been found in a variety of RNP complexes and are often bound by members of the L7Ae/L30 protein family \[105\]. In eukaryotes, the 15.5kD protein is a core protein in C/D box RNPs and the human spliceosomal U4 small nuclear RNP \[194\] \[105\] \[190\]. In archaeal species, L7Ae is a core protein in both C/D box and H/ACA box RNPs. In addition to the regulation of L10 in *Escherichia coli* by a kink-turn in its 5'UTR \[102\] \[88\], two other ribosomal proteins, L30 and L1, autoregulate their translation by binding to kink-turns formed in their mRNA \[102\] \[120\] \[135\]. Another L7Ae/L30 member that binds to mRNA is eukaryotic SBP2, which recognizes the SECIS element and allows for selenocysteine incorporation into selenoproteins \[42\]. In eight different L7Ae/L30 protein family members, the conserved RNA-binding region is different for each family, indicating adaptation of each protein family for the particular kink-turn variant it recognizes \[67\].
Table 6.1: **RNPs that contain kink-turns.** Bold entries indicate that L7Ae binds to kink-turns in archaea. Kink-turns occur ubiquitously across the tree of life and in important RNPs such as the ribosome. In eukaryotes, almost every RNP listed has a distinct kink-turn-binding protein. In contrast, archaea use L7Ae in multiple RNPs. Abbreviations in the domain column: A - Archaea, B - Bacteria, E - Eukarya.

### 6.2.2 Evolutionary perspective of kink-turn-containing RNAs and L7Ae

The role of kink-turn-containing RNAs in multiple, critically conserved RNPs creates an interesting evolutionary perspective (Table 6.1). The kink-turn is an ancient motif, as evidenced by the conservation of multiple kink-turns in the rRNA across the three domains (Table 6.2) [110]. Different groups have suggested that C/D and H/ACA box RNPs, which guide 2’-O-methylation and pseudouridylation of RNA respectively, originated from a primitive translation apparatus. This proposition is based on the observation that in archaea, the kink-turn binding protein L7Ae occurs in both these...
Table 6.2: **Conserved Kink-Turns in Ribosomal RNA.** Multiple kink-turns have been found in rRNA from the crystal structures of the ribosome. Many of these kink-turns are conserved across the three domains of life, indicating that the kink-turn structure is an ancient motif. The presence of the kink-turns listed in this table among the different domains is indicated in the ‘Conservation’ column (A - Archaea, B - Bacteria, E - Eukarya). Data from [110].

<table>
<thead>
<tr>
<th>Kink-turn</th>
<th>Molecule</th>
<th>Nucleotides</th>
<th>PDB file</th>
<th>Conservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>116S Kt-11</td>
<td>T.th. 16S</td>
<td>242-247/277-284</td>
<td>1J5E</td>
<td>A,B,E</td>
</tr>
<tr>
<td>16S Kt-23</td>
<td>T.th. 16S</td>
<td>683-688/699-707</td>
<td>1J5E</td>
<td>A,B,E</td>
</tr>
<tr>
<td>23S Kt-7</td>
<td>H.m. 23S</td>
<td>77-81/93-100</td>
<td>1S72</td>
<td>A, B</td>
</tr>
<tr>
<td>23S Kt-42</td>
<td>H.m. 23S</td>
<td>1147-1155/1212-1216</td>
<td>1S72</td>
<td>A,B,E</td>
</tr>
<tr>
<td>23S Kt-46</td>
<td>H.m. 23S</td>
<td>1312-1319/1338-1342</td>
<td>1S72</td>
<td>A,B,E</td>
</tr>
<tr>
<td>23S Kt-4/5</td>
<td>H.m. 23S</td>
<td>111-115/147-149</td>
<td>1S72</td>
<td>A,B,E</td>
</tr>
<tr>
<td>23S Kt-77/78</td>
<td>T.th. 23S</td>
<td>2124-2127/21612163 2172-2828</td>
<td>1MZP</td>
<td>A,B,E</td>
</tr>
</tbody>
</table>

The discovery that L7Ae also participates in archael RNase P and SRP RNA adds other entries to the list of RNPs that contain kink-turns and have function related to translation-related RNPs [40, 206, 210]. Based on this observation that L7Ae is linked to translation, Cho *et al.* suggested that L7Ae could influence coordinate regulation, especially in consideration of a study that implicates chaperone Hsp90 as a master regulator of cell growth by assisting the folding of human and yeast L7Ae homologs [24]. In comparison to the multiple roles of L7Ae in archaea, eukaryotes have evolved distinct L7Ae homologs for each functionally different RNP [105, 195, 38]. This increase in diversity implies a need for fine-tuning or adaptation of RNA-binding specificity in eukaryotes [38, 67]. A comprehensive study of kink-turns and kink-turn-binding proteins may reveal more details about the evolutionary history of this ancient motif and its function.
6.2.3 Example of a highly conserved kink-turn-protein interaction: L30e binds to its mRNA and the ribosome

An interesting example of a highly conserved kink-turn-protein interaction is L30e to its own transcript and ribosomal binding site. This interaction has been preserved for more than a billion years [191], originating from an ancestor of Archaea and Eukarya. In *Saccharomyces cervisiae*, L30e binds to a kink-turn in ribosomal RNA and also a kink-turn in its own transcript to autoregulate its translation. Vilardell and colleagues demonstrated that the L30e homolog in *Sulfolobus acidocaldarius* can limit the production of the L30e *S. cervisiae* protein, illustrating that the recognition of the L30e kink-turn by the L30e protein family is highly conserved [191]. Although the ribosomal kink-turn is conserved in all three domains, there is no L30e homolog in Bacteria. The bacterial protein that binds to the conserved kink-turn is S15, which is missing in Archaea and Eukarya. Despite the extreme conservation of the ribosomal kink-turn that is bound by L30e, the protein that binds to this kink-turn has not been conserved in all three domains. Given that halophilic archaea are also missing L30e, the full evolutionary story of this binding site of ribosomal RNA remains unclear.

6.2.4 Synthetic biology applications of kink-turns

In addition to the many biological roles of kink-turns and L7Ae, synthetic biology labs have developed artificial constructions using kink-turns and L7Ae. Saito *et al.* created genetic switches that reduced or activated translation [165]. For the construction that reduced the translation of an mRNA, a kink-turn was inserted at translation
Figure 6.1: **Examples of synthetic biology applications of kink-turns.** Examples of how the kink-turn and L7Ae can be used to regulate translation of a protein. (A) In the translational “OFF” system, a kink-turn is inserted into the 5′ end of an mRNA. In this case a kink-turn from an archaeal C/D box sRNA was used (red box). When L7Ae is added to the system, it binds the kink-turn and prevents the ribosome from translating the RNA. (B) In the translational “ON” system, a sequence complementary to the regulator small RNA (blue box) is inserted into the 5′ end of the mRNA. This small RNA has a kink-turn. When L7Ae is introduced, it sequesters the regulatory RNA and allows translation of the mRNA. Modified from [165].

Introduction of L7Ae either to an *in vitro* system or by expression *in vivo* repressed translation by about tenfold. The switch to induce translation used a small RNA containing a kink-turn (Figure 6.1B). The sequence antisense to this RNA was inserted after the start codon of the target mRNA; the binding of the small RNA to the mRNA repressed translation. In the presence of L7Ae, the small RNA is sequestered and the mRNA is free to be translated [165]. In another study, Ohno *et al.* used kink-turn-containing RNAs and L7Ae proteins to create a triangular nano-scale structure. This type of molecule could be used for regulation of cellular signaling and potentially could be produced *in vivo* [141, 142].

The ability of kink-turns to regulate translation in an artificial system suggests
that this could be happening naturally in the cell. As previously mentioned, ribosomal proteins that bind to kink-turns in rRNA are often auto-regulatory and have kink-turns in the 5′ end of their mRNA. It is possible that there are other mRNAs with kink-turns and small RNAs with kink-turns that bind to mRNAs to control translation. A study in *Sulfolobus solfataricus*, described in the next section, found some experimental evidence that suggests that this type of regulation exists in natural systems [206].

### 6.2.5 Previous study of L7Ae-Associated sRNAs in archaea

To investigate the role of L7Ae in various archaeal RNPs, Zago *et al.* constructed a cDNA library containing small RNAs co-immunoprecipitated from *Sulfolobus solfataricus* cell extracts using anti-L7Ae antibodies. This study indicated that unknown kink-turns exist in the transcriptome of archaea and may be bound by L7Ae. Analysis of this library revealed unexpected RNAs with kink-turns that overlap or are encoded within open reading frames (ORFs), are contained within intergenic regions, or are antisense RNAs with partial or full complementary to ORFs (Figure 6.2A). Interestingly, many of the mRNAs detected to have kink-turns are transposases. This observation has also been made in *Pyrococcus abyssi*, a euryarchaeon [151]. Some small RNAs found in this study were also found to be antisense to transposases.

Zago *et al.* found kink-turns in variants of known types of RNAs. Atypical C/D box RNAs were discovered, as well as an H/ACA-like RNA that was missing the 5′ guide region but had a 3′ terminal ACA sequence, bound to L7Ae, and produced pseudouridine modification. The fragment of a kink-turn-containing RNA was found
Figure 6.2: Small RNAs of *S. sofataricus*. Small RNAs that associate with L7Ae from [206]. Boxes indicate the presence of a kink-turn. (A) Small RNA overlapping the 5′ end of the mRNA of a formate hydrogenlyase (SSO1026, hycD). This is sR110 from [206]. Gel-shift assays containing L7Ae (L), aNOP56 (N), and fibrillarin (F) is shown. These assays demonstrate the binding of L7Ae and other C/D box sRNA associated proteins. (B) A small RNA of unknown function (sR125). This sRNA is predicted to be a non-canonical H/ACA box sRNA from the models I created in Chapter 5.

in the spacer regions of 16S and 23S rRNA. The kink-turn in the recovered fragment is located in the 23S rRNA processing stem and binds to L7Ae. Another study also confirmed the presence of this kink-turn and its conservation in other *Sulfolobus* species [179]. Zago *et al.* also found an intergenic RNA positioned between two convergently transcribed ORFs with secondary structure and kink-turn, but did not have features that would indicate its function (Figure 6.2B). This RNA is predicted to be a non-canonical H/ACA from my work in Chapter 5. In another example of L7Ae as a component of important RNPs, band shift assays demonstrated L7Ae binding to a kink-turn in 7S RNA, a component of the signal recognition particle (SRP) that is conserved in all three domains of life. The possible role of L7Ae in rRNA processing and SRP, as
well as its multifunctionality in archaea raises the possibility that this protein has roles yet unknown in other RNPs.

6.2.6 *Pyrobaculum* Ribosomal Protein L14e

Another example supporting the hypothesis that there are unknown kink-turns that occur in RNAs other than sno-like RNAs and the ribosome is a predicted kink-turn in ribosomal protein L14e. Patrick Dennis and I found that a kink-turn may play a part in the regulation of the *Pyrobaculum* ribosomal protein L14e, reminiscent of how the translation of ribosomal proteins L30 and L1 is regulated by protein-binding to kink-turns formed by their transcripts. In six *Pyrobaculum* species, this gene appears to have a highly conserved, 5′-end cis-acting regulatory RNA containing a kink-turn (Figure 6.3). This RNA has a strong predicted promoter and small RNA reads in four of the species confirm the expression of this RNA (Figure 6.4). If the predicted kink-turn is contained in the 5′ UTR, L14e could be regulated by a kink-turn binding protein or a ligand like the SAM riboswitch in bacteria [82]. L14e is not known to bind to kink-turns, but it has a putative RNA binding sequence [60].

6.2.7 Summary of this chapter

After modeling the kink-turn in C/D and H/ACA box sRNAs, learning about the multiple archaeal RNPs with L7Ae as a component, and reading the study by Zago *et al.*, I decided to conduct a high-throughput study of L7Ae-associated RNAs in an archaeal transcriptome. The study in *S. solfataricus* only obtained 45 distinct sequences
Figure 6.3: Cis-acting regulatory element of *Pyrobaculum* L14e predicted structure. I hypothesize that this secondary structure forms in the 5' UTR of L14e and may play a role in its regulation in *Pyrobaculum* species. This secondary structure is found in all six *Pyrobaculum* from the study in Chapter 3. Yellow bases indicate the kink-turn. Pae - *P. aerophilum*, Par - *P. arsenaticum*, Pis - *P. islandicum*, Pne - *P. neutrophilums*, Pca - *P. calidifontis*, Pog - *P. oguniense*
Figure 6.4: **Evidence that the predicted Cis-acting regulatory element is in the 5'-UTR of *Pyrobaculum* L14e.**

Browser shot showing the conservation and promoter prediction of the element. The gene for L14e is in pink, PAE0862. The red rectangle is the predicted element, and the black arrow indicates the predicted promoter. *Pyrobaculum* promoters usually appear 30 nt before the start of transcription, and that is where the predicted promoter is found in this case.
from the cDNA library. Using high-throughput sequencing I could gain a broader view of L7Ae-associated RNAs in archaea. Our discovery of non-canonical H/ACA box sRNAs in archaea and a conserved kink-turn in the 5' UTR of Pyrobaculum L14e also indicated that there are other kink-turn-containing RNAs to be found. This study is also of a euryarchaeal species, T. kodakaraensis, instead of one from the Crenarchaeae like S. solfataricus, and will provide evolutionary perspective.

To obtain L7Ae-associated RNAs, I sequenced RNAs co-immunoprecipitated with L7Ae. I used a strain of Thermococcus kodakaraensis that had a hemagglutinin (HA) tag inserted into the L7Ae mRNA prior to the stop codon. HA antibodies were used to immunoprecipitate L7Ae from cell extracts. The results were similar to what was found in the S. solfataricus study. Potential kink-turns were found in mRNAs, and some small RNAs antisense to mRNAs were also enriched. RNAs with known kink-turns, such as C/D box sRNAs, were also enriched. Further study will be needed to confirm the location of the kink-turns and binding of L7Ae to these RNAs.

6.3 Methods and Materials

6.3.1 T. kodakaraensis strains

A strain of T. kodakaraensis with hemagglutinin (HA)-tagged L7Ae (TK1311), 1311C, and the parent strain, TS559, were a gifts from Dr. Thomas Santangelo at Ohio State University. Methods for the creation of this strain are described in [84]. Both strains require agamatine for growth. The parent strain, TS559, was used as a control.
6.3.2 Media and growth conditions

Cultures were grown anaerobically at 85°C in nutrient-rich artificial salt water media with elemental sulfur (ASW-YT-S°) at pH 7.0 \cite{84, 89}. The media contained per liter 5g tryptone, 5g yeast extract, DSMZ 141 trace minerals, 2g elemental sulfur, 500 µM agamatine sulfate, 1mg resazurin (oxygen indicator) and sodium sulfide to 0.625%. Cells were harvested at late exponential phase (10^6 cells/mL). Cells were spun down and washed with tris-buffered saline (TBS) twice. After washing, cells were pelleted and suspended in modified RIPA lysis buffer pH 7.2 (50 mM Tris-Cl pH 7.2, 100 mM NaCl, 1% NP-40, 0.25% sodium deoxycholate). For experiments testing whether Mg^{2+} ions would improve kink-turn detection, MgCl_2 was added to a concentration of 2.5 mM in the wash and lysis buffers.

6.3.3 Co-immunoprecipitation of kink-turn-containing RNAs using HA antibodies

For all steps low-retention tubes were used. NEB Murine RNase inhibitor and Pierce mini-tablet EDTA-free protease inhibitor were added to cell pellets before lysis. The NEB RNase inhibitor regent contains DTT so that the RNase inhibitor is active. Cells were lysed by sonication. The cell lysate was centrifuged to remove cellular debris and the soluble fraction saved. The Pierce HA-Tag IP/Co-IP kit (26180) was used for the immunoprecipitation. Briefly, soluble fraction was pre-cleared with agarose beads for an hour at 4°C with end-over-end mixing. Anti-HA agarose beads were incubated with the sample overnight at 4°C with end-over-end mixing. Beads were washed three
times with TBS-T (Tris-buffered saline with Tween 20). Sample was treated with DNase I (NEB) to remove DNA. RNA was extracted directly from the beads and separated from L7Ae by TRIzol using the Zymo Directzol kit \cite{81,206}.

### 6.3.4 RNA preparation and fragmentation

RNase III (NEB) was used to cleave the RNA into fragments less than 300bp in preparation for sequencing on the Illumina platform. I briefly considered shearing the RNA by sonication, and ran tests to find the conditions to obtain the desired size range. Shearing requires end-repair prior to library preparation, so I decide to use RNase III which leaves 5′ phosphate and 3′ OH end chemistry and thus RNA can go directly into library preparation after fragmentation \cite{7}. The sample was cleaned up with Zymo II columns.

### 6.3.5 Preparation for sequencing

Samples were analyzed using a Bioanalyzer before library construction. Libraries were prepared by Eva Robinson for sequencing using the NEBNext multiplex small RNA library prep kit for the Illumina platform. Samples were sequenced by the UC Davis Sequencing Facility on Illumina MiSeq to produce 2 × 300 nt paired-end sequencing reads.
6.3.6 Processing and analysis of sequencing data

Andrew Holmes performed the initial processing and mapping of reads to the *T. kodakaraensis* genome. Afterwards I created a computational pipeline to perform all of the steps. Briefly, Holmes’ program *demultiplex.py* was used to demultiplex the samples and output FASTQ files. Next, SeqPrep ([https://github.com/jstjohn/SeqPrep](https://github.com/jstjohn/SeqPrep)) was used to trim adapters and merge reads. Reads were mapped to the genome using Bowtie2 [108]. Using the Bedtools suite [155], samtools [111], and UCSC Genome Browser tools [96], bam files were converted to bigwig format to enable visualization of results as coverage tracks on the UCSC Archaeal Genome browser.

6.4 Results

6.4.1 Description of samples and analysis of enrichment

To find enriched RNAs compared to the control sample, a basic enrichment analysis was conducted. The control sample represents non-specific binding of RNAs during the co-immunoprecipitation. Read counts were normalized based on the total number of reads in each sample. Regions with log-fold change greater than two of read coverage were identified and analyzed.

The control sample had phenol-contamination and had to be re-extracted. This result in many fewer reads in this sample and only had approximately a fourth of the number of reads in the other samples. For this reason my examples focus on regions with greater than 50 reads in the experimental samples.
6.4.1.1 Addition of Mg\textsuperscript{2+} reduced regions of enrichment

The kink-turn is stabilized by divalent cations and protein binding—in a FRET study, the 50% transition to a kink-turn confirmation was 35–78 μM of Mg\textsuperscript{2+} ions \cite{169}. To aid the formation of the kink-turn, I added MgCl\textsubscript{2} to the lysis and wash buffers. I tested if adding Mg\textsuperscript{2+} ions would improve detection of kink-turns with two samples I compared to the control: (i) RNA co-immunoprecipitated with L7Ae in the presence of magnesium (Mg\textsuperscript{+} sample) and (ii) RNA co-immunoprecipitated with L7Ae without magnesium (Mg\textsuperscript{−} sample).

I found that adding MgCl\textsubscript{2} reduced the number of enriched regions since the Mg\textsuperscript{−} sample had 546 enriched regions compared to the 133 enriched regions in the Mg\textsuperscript{+} sample. It is possible that the addition of Mg\textsuperscript{2+} ions allowed more RNases to be active and digest the RNA. Many RNases are known to need divalent cations to be active, such as RNase P and RNase H \cite{39,9}. Since I did not have replicates, it is also possible that what I am seeing is within normal variation of samples.

After I sequenced the samples, I found that sodium ions can also induce kink-turn formation \cite{76}. A FRET study demonstrated that the 50% transition to the kink-turn occurred at 72 mM NaCl \cite{76}. Since the concentration of NaCl in the lysis buffer and wash buffer was greater than this amount, the kink-turn may have also been stabilized by the Na\textsuperscript{+} ions in addition to L7Ae binding.
Figure 6.5: **Enriched HACA box sRNA** Example of an enriched RNA known to contain a kink-turn. A non-canonical H/ACA box sRNA at chr:1982738-1982824 is enriched in the Mg+ and Mg− sample compared to the control strain. The blue arrow shows the location of the H/ACA and that it is on the minus strand. Coverage tracks for the control and Mg+ sample are shown with maximum coverage in the window labeled on the left. Reverse strand genomic sequence is shown.

### 6.4.2 Enrichment of known non-coding RNAs

Based on previous studies and data, I made hypotheses about the types of RNAs that would be collected in this study. RNAs known to have kink-turns, such as C/D box RNAs, H/ACA-like box RNAs, and ribosomal RNA, should be in the set of RNAs co-immunoprecipitated and sequenced. Other non-coding RNAs, such as tRNAs, should not be enriched. In general, I found the results were as expected for non-coding RNAs.

Various non-coding RNAs with known kink-turns were enriched, including
Figure 6.6: **SRP RNA is slightly enriched.** Enriched regions are indicated by black rectangles with white chevrons. The enriched regions are in the general region of the predicted kink-turn, which is indicated by blue rectangles. Enrichment does not line up with the coverage peak near the 3′ end of the RNA because of fluctuations in the control data in that area. The predicted kink-turn location based on the data in [206]. Maximum coverage for the region is indicated on the left for each RNAseq track. The location of SRP RNA is from the RFAM track in UCSC Archaeal Genome Browser.

H/ACA box sRNAs (Figure 6.5), C/D box sRNAs, and ribosomal RNA. No tRNAs were enriched, as expected. Although C/D box and H/ACA box sRNAs were enriched, only about half of the annotated C/D box sRNAs were enriched and three H/ACA box sRNAs were enriched out of the known 24. Some sno-like RNAs may only be expressed in certain conditions, so that may explain why not all were enriched. Regions of ribosomal RNA were significantly enriched with a greater than 32-fold change in some regions. Approximately 10% and 30% of the enriched regions were 16S or 23S rRNA in the Mg+ and Mg− samples, respectively. Since L7Ae is a ribosomal protein and 16S and 23S ribosomal RNA are some of the most abundant RNAs in the cell, the large proportion of rRNA in the sample was expected. In both samples, 5S rRNA is also enriched. 5S rRNA is not known to have kink-turns, but it might be pulled down with the rest of the ribosome.

As I previously mentioned, L7Ae has been shown to bind *S. solfataricus* SRP RNA [206]. This RNA was also appears to be slightly enriched in *T. kodakaraensis* (regions had between 4-fold to 6-fold difference from the control). Enrichment was only
found in the Mg+ sample (Figure 6.6). Regions of enrichment overlap region of the predicted SRP kink-turn.

### 6.4.3 L7Ae mRNA (TK1311)

I expected to find the 5′ end of L7Ae mRNA to be enriched based on predicted kink-turns in other species. In other species, often the predicted kink-turn of the L7Ae mRNA had small RNA reads. Our collaborators, Venkat Gopalan and Cory Schwartz, confirmed the kink-turn in *Pyrococcus furiosus* L7Ae mRNA by gel-shift assays (Figure 6.7). Here I found that the 5′ end of *T. kodakaraensis* L7Ae mRNA was enriched in the region of the predicted kink-turn (Figure 6.8). Since L7Ae is a ribosomal protein, it is not unexpected that it could be autoregulatory and contain a kink-turn in its mRNA to control translation.

### 6.4.4 Enriched mRNA fragments

Some mRNA fragments were enriched in both the Mg+ and Mg− samples, but there were many more enriched fragments in the Mg− sample. In the Mg− sample there were 206 mRNAs with enriched fragments compared to the 22 enriched in the Mg+ sample. However, both samples had 16 of the same mRNAs with enriched fragments. These mRNAs had diverse predicted functions, including transposases, seryl-tRNA synthase (TK1140), thioredoxin reductase (TK2100), prephenate dehydrogenase (helps catalyze phenylalanine to tyrosine, TK0259), ATPase (TK0915), histone A and B (TK1413, TK2289), and a cysteine desulfurase (TK1990). Two of these mRNAs are related to
Figure 6.7: Gel-shift assay of L7Ae mRNA in *Pyrococcus furiosus*. I found various instances of L7Ae mRNAs with kink-turns across diverse archaea. These kink-turns had small RNA reads. Our collaborators, Venkat Gopalan and Cory Schwartz, did a gel-shift assay with the predicted kink-turn of the *Pyrococcus furiosus* mRNA. Results suggested binding by L7Ae. (A) Small RNA reads overlapping the 5′ end of *P. furiosus* L7Ae mRNA and the location of a potential kink-turn (B) Gel shift assay by Cory Schwartz with L7Ae and the RNA sequence indicated in (C). Experiments were carried out in 10 mM Mg\(^{2+}\) (C) Predicted structure of the kink-turn in *P. furiosus* mRNA. Box indicates kink-turn.

cysteine: thioredoxin reductase helps form disulfide bonds and cysteine desulfurase removes sulfur from cysteine to form L-alanine. The enriched fragments in these mRNAs are located in the center of the respective mRNA. Since I grew *T.kodakaraensis* with elemental sulfur, there may be a connection to the growth conditions.

The two histone mRNAs that have enriched fragments appear to have enriched fragments at the 5′ end and near the 3′ end (Figure 6.9A,B). It is unclear if a kink-turn could form between these two region or if each region has a kink-turn. Since both of these mRNAs have similar function, the similar pattern of enriched fragments may not
Figure 6.8: **Kink-turn of L7Ae mRNA is enriched.** The region of enrichment for *T. kodakaraensis* L7Ae, Tk1311, is shown. Maximum coverage for the region is indicated on the left for the Mg− track. Structure of predicted kink-turn in the region is shown. Black box indicates kink-turn location in the RNA.

Two of the six transposases in the *T. kodakaraensis* genome had enriched fragments in both samples (TK0850 and TK0654). In both cases, the enriched fragments were at the 3′ end (Figure 6.9C). Fragments of the 3′ end of transposases were also found in the *S. solfataricus* study. The fragment of TK0850 overlaps HgcC, a predicted non-coding RNA of unknown function [103], and has strong secondary structure (Figure 6.9D). In relation to these observations, it is worth noting that homologs to the H/ACA box sRNA sR202 in *Haloferax volcanii* predicted in Chapter 5 in other halophiles also overlap the 3′ end of a transposase. The strong structure and kink-turn at the 3′ end of these transposases may serve a function and happen to resemble a H/ACA box sRNA.
Figure 6.9: **Enriched mRNA fragments.** Maximum coverage for the region is indicated on the left for each RNAseq track. Cheverons indicate direction of genomic elements. (A) Histone B (TK2289). (B) Histone A (TK1413). (C) Enriched 3′ end of a transposase that also overlaps a predicted non-coding RNA HgcC. The non-coding RNA does not have a predicted function. (D) Portion of HgcC that looks like a H/ACA box sRNA. Kink-turn is indicated by the black box.
A few small RNAs antisense to mRNAs were enriched, but none had obvious kink-turns (Figure 6.10). Many of the enriched antisense RNAs were C/D box sRNAs overlapping the 3' end of mRNAs. Since the antisense RNAs did not have any kink-turns that I could find manually, gel-shift assays could be used to confirm binding by L7Ae and the location of the kink-turn.
6.5 Discussion

The examples of enriched RNAs in the L7Ae pull-down data suggest that unannotated kink-turns occur in mRNAs and small RNAs. Gel-shift assays are needed to confirm binding of L7Ae and to determine the location of the kink-turns. Biochemical and genetic disruptions would be useful to help define the role of L7Ae to RNA processing, but are beyond the scope of this work. Disrupting kink-turns in the enriched RNAs and studying the resulting phenotype would provide information on how critical L7Ae is to the function of the RNAs.

Additional analysis of the data using computational methods such as DESeq [3] and Pirana [187] that model the underlying distribution of reads may provide better peak-calling. Future experiments can be modified to provide more information. Removing ribosomes prior to immunoprecipitation will increase sequencing depth for non-ribosomal RNAs and facilitate detection of L7Ae binding in low abundance RNAs. Harvesting the cells in experimental and stationary phase could provide additional data for RNAs that are only expressed or regulated in particular growth phases. Instead of RIP-seq, cross-linking and digestion of the RNA (CLIP-seq) could be used to home in on the L7Ae binding site and location of the kink-turns [18].

Overall, the results in T. kodakaraensis were similar to what was found in S. solfataricus, indicating that L7Ae may have other unknown functions in Archaea.
Chapter 7

Conclusion

With my dissertation, I have presented a comprehensive view of archaeal sno-like RNAs from the genus level to the domain level. To achieve this view, I developed covariance models to better predict archaeal C/D and H/ACA box sRNAs and used these models to scan species across the Domain Archaea. To aid my evaluation of predictions, I used small RNA sequencing data, conservation, and promoter predictions. For the C/D box sRNA model, I added the kink-turn, which is not found in other computational prediction methods for archaeal C/D box sRNAs. For H/ACA box sRNAs, I created new models for non-canonical forms that had previously only been found in the *Pyrobaculum* genus.

After compiling C/D and H/ACA box sRNA annotations predicted by these models, I found that there are clade-specific patterns of these sno-like RNAs terms of abundance, structure, and function. I found that abundance of sno-like RNAs within a species usually depends on where that species falls into archaeal phylogeny. Halophiles
and methanogens tend to have very few sno-like RNAs. An exception to this rule is *M. kandleri*, which has a large number of C/D and H/ACA box sRNAs. However, it is an outgroup to the rest of the methanogen species, so this may account for its number of sno-like RNAs in comparison to other methanogens.

Hyperthermophiles, no matter the clade, usually had an abundance of C/D box sRNAs. However, *M. janaschii* and *A. fulgidus*, two hyperthermophiles, have less than ten C/D box sRNAs. *M. jannaschii* is a methanogen and *A. fulgidus* is phylogenetically near methanogen clades, so that may account for their lack of C/D box sRNAs.

For H/ACA box sRNAs, I found that the non-canonical forms are widespread across archaeal species and now comprise half of all archaeal H/ACA box sRNA annotations. The abundance of non-canonical forms also appears to have clade-specific patterns. Non-canonical forms are more prevalent in the Crenarchaea than the Eurarchaea. This helps explain why so few H/ACA box sRNAs were found in crenarchaeal species in previous studies.

Very few C/D and H/ACA box sRNA guides were conserved beyond their genus or closely related genera. I found that the few guides that were highly conserved target functionally important regions of rRNA.

Although I annotated many C/D and H/ACA box sRNAs, the models can still be optimized, and likely more sRNAs exist that I did not find. More clade-specific models can be made for both the C/D box and H/ACA box sRNAs to improve specificity and sensitivity. For the C/D box sRNA models, new models specific to C/D box sRNAs with K-turns instead of K-loops (Section 4.3.5) may also yield more annotations. Other
structural variants of the H/ACA box sRNA models can also be made by incorporating a K-loop into the non-canonical forms. A comprehensive analysis of the methylation and pseudouridylation targets may also reveal more about the function of these sRNAs and suggest new post-transcriptional modification sites. I only analyzed highly-conserved guides; a comprehensive analysis will provide a broader view of modification patterns in ribosomal and tRNAs.

Finally, I surveyed RNAs bound to L7Ae by sequencing RNAs co-immunoprecipitated with the protein. I wanted to see if I could find other RNPs that contained L7Ae and thus also likely kink-turns. I became interested in what other RNAs contained the kink-turn after modeling the kink-turn in C/D and H/ACA box sRNAs and learning about the multiple archaeal RNPs with L7Ae as a component. I found some enriched mRNAs with potential kink-turns and some enriched RNAs antisense to mRNAs. To confirm that L7Ae bound some of these RNAs, gel-shift assays are needed. Repeating the experiment in triplicate would also provide more statistical power to analyze if L7Ae binds without doing gel-shift assays. If L7Ae does bind to the antisense RNAs, it could play a role in regulating mRNAs similar to the artificial system developed by Saito et al. [165]. Although I did not find any definitive evidence that L7Ae occurs in RNPs that what is currently known, my data hints that L7Ae has other unknown functions.

As with many studies, my findings answer some questions and raise new ones. My work provided a comprehensive view of archaeal sno-like RNA annotations and structure, but questions about function remain. Annotation and prediction provides a stepping stone to to fully understand the function and impact of these RNAs within the
cell. Undiscovered sites of modification are likely, and roles beyond post-transcriptional modification for archaeal C/D and H/ACA box sRNAs are also possible. Other questions raised are "do non-canonical H/ACA box sRNAs occur in eukaryotes?" and "could L7Ae regulate the function of mRNAs?" My work provides starting material for these new questions.
Appendix A

Description of supplementary files

These files are submitted as Supplementary Files with this dissertation.

A.1 Pyrobaculum supplementary files

A.1.1 Structural alignment for Pyrobaculum covariance model

File: pyrAer1_cdrna.sto

Description: This structural alignment is in Stockholm format and I used as input to cmbuild to create a covariance model for archaeal C/D box sRNAs. The --hand option is specified when using Infernal v1.1 and the --rf option when using Infernal v1.0. I modeled the structure of the K-turn and K-loop (SS_cons line). The SS_cons line annotates base pairs, where open and close parentheses indicate a base pair and periods indicate unpaired regions. The RF line indicates insertion positions with periods; these positions are not required to be in outputted predictions. Positions
marked with ‘x’ are required to be in predictions, and include the box features and minimum guide length in this model. Note that the nucleotide 5′ of the D and D′ boxes are included in the structure. Guide regions are annotated to be “n” to represent any base and make the model general to all archaeal C/D box sRNAs.

A.1.2 Target finding program

**Files:** findAntisense.zip, main program is findAntisense.py

**Description:** The zip folder contains findAntisense.py and supporting files. This python program was developed by Andrew Uzilov, Andrea Corredor, and me. Portions of the program are written in C to improve speed. Cython was used to interface the C code with the main program (.pyx file and .pyd header file). Description of how to use this program can be obtained by using the -h option. This algorithm cannot detect bulges.

A.1.3 C/D box sRNA annotations of *Pyrobaculum* and *Pyrococcus*

**File:** Pyrobaculum-Pyrococcus-Annotations.xlsx

**Description:** Annotations of *Pyrobaculum* and *Pyrococcus* C/D box sRNAs. Included bitscores are output from cmsearch using the covariance model built from pyrAer1.cdrna.sto. If an annotation has a bitscore of 0, it is a manual annotation found by either comparative genomics or manual curation of small RNA sequencing reads.
A.1.4 Predicted methylation targets

File: PyrobaculumTargets.xlsx

Description: Excel file with the predicted sites of 2′-O-methyl modification in 16S rRNA, 23S rRNA and tRNA for the D and D′ guides of each annotated C/D box sRNA in the six species of *Pyrobaculum*. This file was created by hand from findAntisense predictions. The top scoring findAntisense prediction was recorded if it had at least 9 consecutive Watson-Crick base pairs in the guide-target interaction. An exception to this rule is for ‘double guide’ C/D box sRNAs, whose D and D′ targets are within 100 nt of each other. In the case of double guides, the target for each guide would be recorded if it had a maximum of one mismatch, a maximum of two G:U base pairs, and the combined bit score for both guide-target complementarities was 32 or higher (where a Watson-Crick base pair is 2, a G:U base pair is 1 and a mismatch is -2).

A.1.5 *Pyrobaculum* Ribosomal RNA alignments

Files: Pyrobaculum16Salign.doc, Pyrobaculum16Salign.fasta

Description: A sequence alignment of the 16S rRNAs from the six species of *Pyrobaculum* in a Word document and FASTA file. The Word file contains annotations of the predicted sites of C/D box sRNA mediated methyl modification indicated in the alignment. This file was made in collaboration with Patrick Dennis. The FASTA file is the alignment in FASTA format.

Files: Pyrobaculum23Salign.doc, Pyrobaculum23Salign.fasta

Description: A sequence alignment of the 23S rRNAs from the six species of
*Pyrobaculum* in a Word document and FASTA file. The Word file contains annotations of the predicted sites of C/D box sRNA mediated methyl modification indicated in the alignment. The FASTA file is the alignment in FASTA format.

### A.1.6 Overlap of sRNAs with protein-coding genes

**File:** PyrobaculumCDmRNAOverlap.xlsx

**Description:** Excel file that lists the *Pyrobaculum* C/D box sRNAs that overlap protein-coding genes divided into the categories described in Section 3.5.7.

### A.2 Supplementary material for Panarchaeal C/D box sRNA study

#### A.2.1 C/D box sRNA Annotations

**File:** CD-Annots-Crenarchaea.xlsx

**Description:** Excel file of C/D box sRNA annotations for crenarchaeal species. Boxes are annotated. If there is structure in the loop region, it is annotated in dot-bracket notation below the sRNA sequence. In some cases there is no clear C′ box and thus it is not annotated. This list includes previously annotated sRNAs as indicated in Section 4.2.

**File:** CD-Annots-Euryarchaea.xlsx

**Description:** Same as above for euryarchaeal species.

**File:** CD-Annots-Other.xlsx
Description: Same as above for species in phyla other than the Crenarchaeota and Euryarchaeota. Annotations for *Nitrosopumilus maritimus* (Thaumarchaea), *Nanoarchaeum equitans* (Nanoarchaeota), and the composite genome of *Lokiarchaeum*.

### A.2.2 K-turn Analysis

File: *KturnAnalysis.xlsx*

Description: Excel file with analysis of K-turns formed between the C and D boxes and C’ and D’ boxes. Frequencies of different base pair types for each position of the K-turn or K-loop are listed. Wilcoxon Rank-Sum p-values discussed in Chapter 3 are included.

### A.2.3 Ribosomal RNA

Files: *rRNAs_aerPer1.bed*, *rRNAs_aerPer1.fasta*

Description: Ribosomal RNA files for *Aeropyrum pernix*. The BED file contains the coordinates of 16S and 23S rRNA and the FASTA file contains the sequences. Positions were determined from alignment with *Pyrobaculum calidifontis* rRNA.

Files: *rRNAs_stapMari1.bed*, *rRNAs_stapMari1.fasta*

Description: Same as above for *Staphylothermus marinus*.

### A.2.4 Alignment of L7Ae protein sequence across diverse archaea

File: *L7Ae-alignment.fasta*

Description: Full alignment of L7Ae from diverse archaea described in
Figure 4.3.4.1 in FASTA format. The *Pyrobaculum* species have a conserved 10-15 amino acid N-terminus extension and a 6 amino acid C-terminus extension. These extensions are also present in some of the other members of *Thermoproteales*, such as *Vulcanisaeta moutnovskia* and *Thermoproteus uzoniensis*.

### A.3 Supplementary material for H/ACA box sRNA study

File: *HACA-Annots-Crenarchaea.xlsx*

**Description:** Excel file of H/ACA box sRNA annotations for crenarchaeal species separated by phylogenetic order. Small RNA sequencing reads overlapping the position are included, if applicable. If the structure was manually examined by hand it is included. In some cases, the numbering of annotations within an order represents homologs between species, but not always. A ”ca” after the name indicates that the annotation is of the canonical form. An ”na” indicates that it is a non-canonical form with an ”ANA” box, while ”nn” indicates that it is a non-canonical form without an ”ANA” box. Annotations without conservation or small RNA reads are considered ”questionable.”

File: *HACA-Annots-Euryarchaea.xlsx*

**Description:** Same as above for euryarchaeal species.

File: *HACA-Annots-Other.xlsx*

**Description:** Same as above for species in phyla other than the Crenarchaeota and Euryarchaeota. Annotations for *Nitrosopumilus maritimus* (Thaumarchaeota),
Nanoarchaeum equitans (Nanoarchaeota), and the composite genome of Lokiarchaeum.

**File:** haca_models.zip

**Description:** Contains the structural alignments in Stockholm format used to create the H/ACA box sRNA covariance models.
Bibliography


[19] David L Bernick, Kevin Karplus, Lauren M Lui, Joanna K C Coker, Julie N Murphy, Patricia P Chan, Aaron E Cozen, and Todd M Lowe. Complete genome


[38] C Charron, X Manival, A Cléry, V Senty-Ségault, B Charpentier, N Marmier-Gourrier, C Branlant, and A Aubry. The archaeal sRNA binding protein L7Ae has a 3D structure very similar to that of its eukaryal counterpart while having a broader RNA-binding specificity. *Journal of Molecular Biology*, 342(3):757–73, September 2004.


[40] I-Ming Cho, Lien B Lai, Dwi Susanti, Biswarup Mukhopadhyay, and Venkat Gopalan. Ribosomal protein L7Ae is a subunit of archaeal RNase P. *Proceedings of
the National Academy of Sciences of the United States of America, 107(33):14573–8, August 2010.


[63] Robert D Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E Pollington, O Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, and Alex Bateman. The


