

# **THE DISTRIBUTION AND EVOLUTION OF SMALL NON-CODING RNAS IN ARCHAEA IN LIGHT OF NEW ARCHAEAL PHYLA**

A thesis submitted in partial fulfilment of the  
requirements for the  
Degree of  
Master of Science  
in Biological Sciences  
at the  
University of Canterbury

by Laura A.A. Grundy  
University of Canterbury  
2016

# Table of Contents

|                                                                                                                      |           |
|----------------------------------------------------------------------------------------------------------------------|-----------|
| Table of Contents.....                                                                                               | 2         |
| Acknowledgements .....                                                                                               | 4         |
| Abstract.....                                                                                                        | 5         |
| <b>Chapter One - Introduction.....</b>                                                                               | <b>6</b>  |
| Overview .....                                                                                                       | 6         |
| The Archaeal Domain of Life .....                                                                                    | 6         |
| Metagenomics and the Availability of Genomic Data .....                                                              | 6         |
| The History of Archaeal Taxa .....                                                                                   | 7         |
| Archaeal Similarities to the Bacteria and Eukaryotes.....                                                            | 9         |
| Small Non-Coding RNA.....                                                                                            | 10        |
| RNA.....                                                                                                             | 10        |
| Riboswitches .....                                                                                                   | 11        |
| snoRNAs .....                                                                                                        | 16        |
| Searching Genomes.....                                                                                               | 18        |
| Tools for the Comparison of Genomic Data .....                                                                       | 19        |
| Databases of Sequence Features .....                                                                                 | 20        |
| Objectives of this Thesis.....                                                                                       | 21        |
| References.....                                                                                                      | 21        |
| <b>Chapter Two - The Function and Distribution of Riboswitches in Archaea.....</b>                                   | <b>30</b> |
| Introduction .....                                                                                                   | 30        |
| Methods .....                                                                                                        | 30        |
| Investigating the Distribution of Riboswitches in the Archaea .....                                                  | 30        |
| Examining Protein-Coding Genes Found Downstream of Riboswitches in the Archaea.....                                  | 32        |
| Identifying the likelihood of Horizontal Gene Transfer of Riboswitch-Gene pairs from Bacteria into the Archaea ..... | 33        |
| Results .....                                                                                                        | 35        |
| The TPP Riboswitch in Archaea .....                                                                                  | 36        |
| Genes found downstream of the TPP riboswitch in Archaea .....                                                        | 37        |
| The FMN Riboswitch in Archaea .....                                                                                  | 41        |
| Genes downstream of the FMN riboswitch in Archaea .....                                                              | 42        |
| The Fluoride Riboswitch in Archaea .....                                                                             | 44        |
| Genes downstream of the Fluoride riboswitch in Archaea .....                                                         | 45        |
| The Moco_RNA_motif in Archaea .....                                                                                  | 48        |
| Discussion.....                                                                                                      | 50        |
| The Distribution of Riboswitches in the Archaea.....                                                                 | 50        |
| Protein-Coding Genes Found Downstream of Riboswitches in the Archaea....                                             | 53        |
| Horizontal Gene Transfer of Riboswitch-Gene pairs from Bacteria into the Archaea.....                                | 54        |
| Horizontal Gene Transfer: The Bigger Picture .....                                                                   | 56        |
| Concluding Remarks .....                                                                                             | 56        |
| References.....                                                                                                      | 57        |
| <b>Chapter Three - The Distribution of Known snoRNA families in the Archaea ..</b>                                   | <b>61</b> |
| Introduction .....                                                                                                   | 61        |
| Methods .....                                                                                                        | 62        |
| Investigating the Distribution of snoRNAs and snoRNPs in the Archaea .....                                           | 62        |
| Results .....                                                                                                        | 63        |
| Distribution of C/D Box snoRNAs in the Archaea.....                                                                  | 64        |
| Distribution of H/ACA Box snoRNAs in the Archaea .....                                                               | 67        |
| Distribution of snoRNPs in the Archaea.....                                                                          | 69        |
| Discussion.....                                                                                                      | 71        |

|                                                                                 |           |
|---------------------------------------------------------------------------------|-----------|
| C/D box snoRNAs and their associated snoRNPs in the Archaea .....               | 71        |
| H/ACA box snoRNAs and their associated snoRNPs in the Archaea.....              | 72        |
| The presence of eukaryotic snoRNAs within archaeal genomes.....                 | 73        |
| Concluding Remarks .....                                                        | 74        |
| References.....                                                                 | 74        |
| <b>Chapter Four - Summary and Concluding Remarks .....</b>                      | <b>77</b> |
| Our results in the context of the currently known archaeal phyla .....          | 77        |
| New knowledge of snoRNAs and riboswitches from this study .....                 | 78        |
| Future Directions.....                                                          | 79        |
| Conclusion .....                                                                | 80        |
| References.....                                                                 | 81        |
| Appendix One - Species Comprising the Archaeal Taxa Examined in this Thesis ... | 83        |

## Acknowledgements

This thesis was made possible with the help and support of a great many people.

Firstly, I would like to thank my supervisors, Ant Poole and Paul Gardner, without the support of which I would have been lost from the beginning. Thank you for the time and effort you have put into helping me succeed in completing this thesis. I would also like to acknowledge the patience you have shown in the guidance you have given me throughout the time I have been working with you.

I would next like to acknowledge the support of those in both Ant and Paul's lab groups for the input and encouragement you have given. In particular to Sinan for your technical support and to Alannah, Alicia, Nicole, Nellie, and Fatemeh for your input and friendship during the times we have spent together.

To other members of the School of Biological Sciences and University community, the support and encouragement I have received has definitely been appreciated. Every one of you has played an important role in helping me succeed and while there are too many of you to list here your individual support of me has been of great help.

To my fellow members of the UC Gaming Guild exec over the past two years, thank you for putting up with my juggling of exec responsibilities while completing this thesis. Your willingness to step up to support and guide the club allowing me to step back and focus on my studies was a great relief.

Finally I must thank my friends, flatmates and family for always being encouraging of me throughout my journey to complete this thesis. To Charles, Charlotte, Hannes and especially to my Kip, thank you for being amazing flatmates. And special thanks must be also given to my Mum and Dad along with my brother Braden and again to Peter, you have been invaluable in your love and support of me during this time.

## Abstract

The archaea are organisms which are currently thought to constitute the third domain of life. Sharing many similarities with the other two domains, bacteria and eukaryotes, the knowledge of the diversity of the archaea has been rapidly expanding as metagenomic technologies continue to be developed. In this study we attempt to expand knowledge about two types of small non-coding RNAs (riboswitches and snoRNAs) within the archaea. Both of these types of RNA are also found in other domains of life making them an interesting point of comparison between the archaea and the bacteria and eukaryotes. We studied the distribution of both riboswitches and snoRNAs across 26 archaeal classes representing 13 archaeal phyla using homology searches based on known models from the Rfam and Pfam databases.

Our study finds that riboswitches are distributed throughout relatively few archaeal taxa. Despite this, we identified many new occurrences of the known riboswitch families and identified a potential case of the presence of a riboswitch family not previously known to be found in archaea. We examined the genes found downstream of the riboswitch occurrences within the archaea. We found that while many riboswitches are associated with genes expected from previous studies of these riboswitch families in bacteria, other genes with both known and unknown functions also appear to be associated with these riboswitches in the archaea. The association of the Fluoride riboswitch with ion transporters that are not currently known to transport fluoride ions is one example of this. We also identified at least one case of a likely horizontal gene transfer (HGT) event of a riboswitch-gene pair from the bacteria into the archaeal genus *Methanocorpusculum*. This result lend weight to the suggestion of a series of HGT thought the evolution of the archaea as an explanation for the current distribution of riboswitches in the archaeal taxa studied.

In investigating the distribution of snoRNAs within the archaea, we determined that while most archaeal taxa show evidence of known snoRNA families, these snoRNAs are not necessarily distributed through all species in each taxa. This broad distribution is consistent with a potentially evolutionary ancient origin for the snoRNAs. We identified the need for future work to determine whether the wide distribution of known snoRNPs associated with C/D box snoRNAs suggests as yet unknown RNA families within the archaea and whether a lack of an essential snoRNP for pseudouridylation in the DPANN superphylum indicates the absence of H/ACA box snoRNAs in those phyla.

# Chapter One - Introduction

## **Overview**

The archaea are a diverse group of organisms that form the third domain of life alongside the bacteria and the eukaryotes. Archaea share many traits in common with both groups. In this study we investigate two types of small non-coding RNAs (riboswitches and snoRNAs) in the archaea and draw comparisons to their presence in the other two domains of life. In doing so we seek to provide new information about the evolutionary history of these two types of RNA within the archaea and to determine how the distribution of these RNAs within the archaea relates to their properties known from the bacteria and the eukaryotes. We carry out this study in light of an increasing amount of genomic data becoming available from both established and emerging archaeal phyla. We sought to determine whether many established patterns related to these types of small RNAs from all three domains of life were reflected across all archaeal phyla. This study therefore also provides novel information about the overall biology of these emerging taxa.

## ***The Archaeal Domain of Life***

### **Metagenomics and the Availability of Genomic Data**

Advances in microbiological and genomic techniques have meant that information about the diversity of life has rapidly expanded. New organisms are regularly being described based on environmental sampling and genetic and genomic techniques (Sharon & Banfield 2013). Traditional sequencing studies have focused on sequencing of genetic material from clonal colonies cultivated in a laboratory setting. Metagenomic techniques allow scientists to study evolutionary and genetic relationships of organisms directly from environmental samples (Amann *et al.* 1995; Handelsman 2004).

Targeted amplification of genetic material to examine the 16S rRNA gene or other specific indicator genes is a useful way to examine the microbial diversity of an environmental sample (Hugenholtz *et al.* 1998). The 16S rRNA gene is a good candidate for this amplification as it contains both highly conserved regions sufficient for targeting by universal primers, as well as regions that allow distinctions to be

made between different phylogenetic groups (Weisburg *et al.* 1991; Coenye & Vandamme 2003). However, more recent metagenomic techniques focus on whole genomes of microbial life in environmental samples. This is preferable to sampling of just select genes as it can give a more complete picture of the biological functions of a sample within the context of an ecological community (Eisen 2007; Sharon & Banfield 2013). While metagenomics is a powerful tool for generating novel genomic data, there are still issues when it comes to constructing genomes of individual species from the data collected. Contamination of samples with material from other unknown organisms present may confound whether novel genetic material belongs to a particular species or not (Schmieder & Edwards 2011; Salter *et al.* 2014).

The Genomic Encyclopedia of Bacteria and Archaea (GEBA project) is an attempt to catalogue and properly classify this new diversity as it is uncovered (Wu *et al.* 2009; Kyrpides *et al.* 2014). The project and techniques and methodologies pioneered by it help to overcome limitations of previous genetic studies of microbial life. Namely, sampling based on organisms that were easily cultivated in laboratories, sampling of organisms with close associations to humans, and selective sampling of organisms based on interesting traits (Hugenholtz 2002; Wu *et al.* 2009). As a result of projects such as GEBA, there is increasingly more data being generated than can be easily analysed completely in a timely manner. Thus the use of bioinformatic techniques, such as homology searches for similarities between new genomes and the genomes of more well studied organisms, are valuable to generating a more complete picture of the biology of these new organisms. The archaea are one such group in which these new sequencing and searching technologies have progressed our understanding of their biology.

## **The History of Archaeal Taxa**

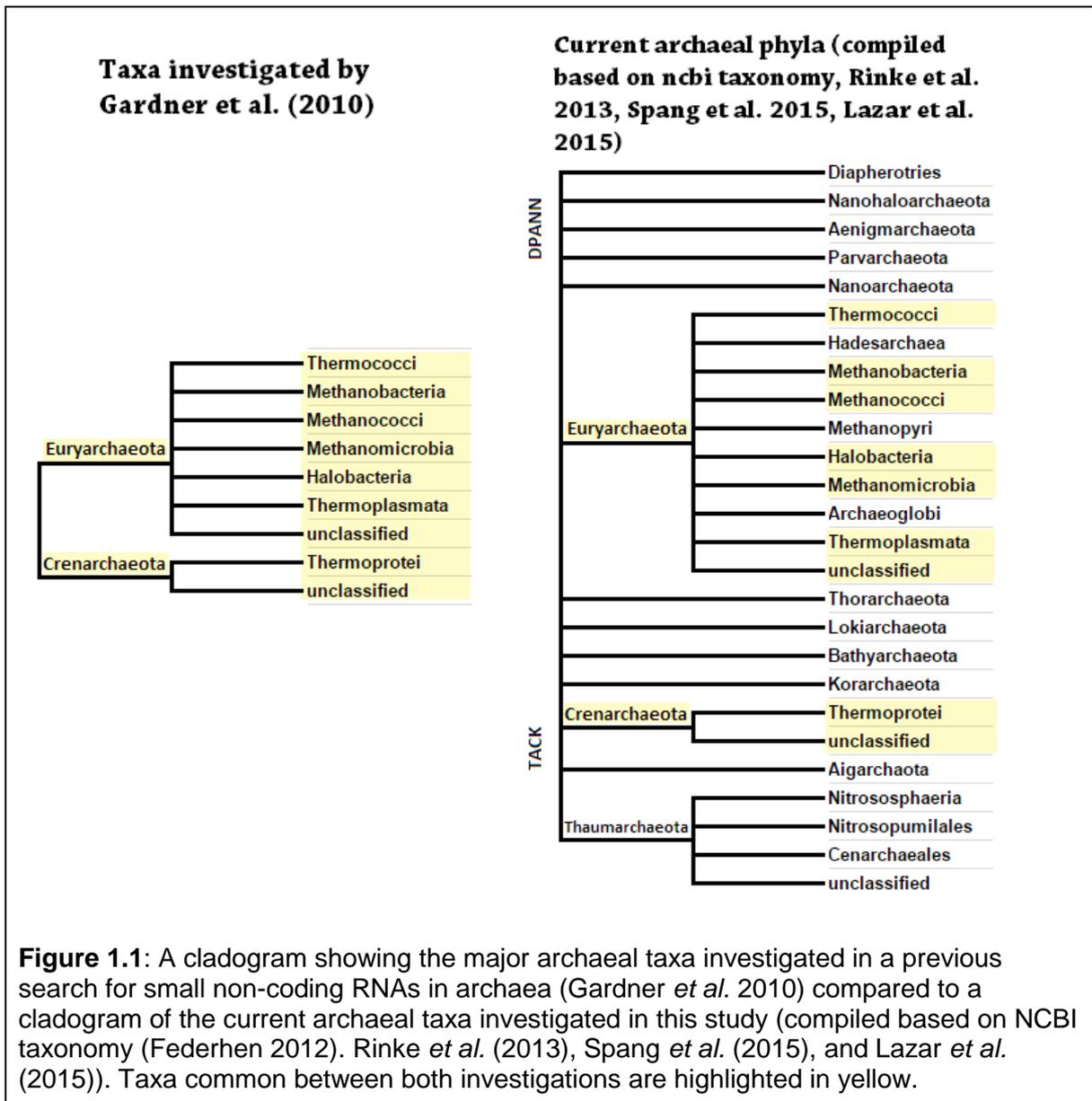
The archaea are organisms considered to constitute the third domain of life. They were previously considered part of the “prokaryotes”, a domain consisting of single celled organisms lacking a nucleus and other membrane-bound organelles (Sapp 2005). Ongoing developments in microbiological and genomic techniques have seen the evolutionary history of the archaea recognised as a distinct group from the similarly prokaryotic bacteria (Woese *et al.* 1990). On a molecular level, the archaea more closely resemble the eukaryotes than their fellow “prokaryotic” life, the bacteria (Pace 2006). In fact, many studies in phylogenetics of early life now posit

that the eukaryotes arose from within the archaeal lineage (Guy & Ettema 2011; Williams & Embley 2014; Spang *et al.* 2015; Raymann *et al.* 2015).

The archaea have seen many additions and changes in recent times. In fact, ongoing developments in sequencing and genomic data collection mean that changes and additions to the archaeal tree are likely to continue relatively rapidly in the near future. At the time at which the archaea were first defined as a separate group to the bacteria, the archaea were divided into only two main phyla, the Euryarchaeota and the thermophilic Crenarchaeota (Woese *et al.* 1990). Since then further divisions and revisions have been made to these groupings (Figure 1.1). The Korarchaeota were identified early on as being deeply divergent from both the Crenarchaeota and the Euryarchaeota (Barns *et al.* 1994; Burggraf *et al.* 1997). Together with the Thaumarchaeota, one of the next major divisions to be identified (Brochier-Armanet *et al.* 2008), and the more recently proposed Aigarchaeota (Nunoura *et al.* 2005; Nunoura *et al.* 2010), the Korarchaeota and the Crenarchaeota form the TACK superphylum (Guy & Ettema 2011). This superphylum recognises similarities of all four groups that set them apart from both the Euryarchaeota and the DPANN superphylum (Guy & Ettema 2011; Rinke *et al.* 2013).

The DPANN superphylum, originally consisting of the Diapherotrites, the Parvarchaeota, Aenigmarchaeota, the Nanoarchaeota and the Nanohaloarchaeota was first proposed by Rinke *et al.* (2013). Since then, new phyla (such as Pacearchaeota and Woesearchaeota) have also been identified as likely being part of this grouping (Castelle *et al.* 2015).

Other candidate phyla of the archaea have also been recently proposed. The Lokiarchaeota (Spang *et al.* 2015), Bathyarchaeota (Meng *et al.* 2014; Attar 2015), and Thorarchaeota (Seitz *et al.* 2016), all represent novel archaeal lineages which are currently thought to show similarities to the TACK superphylum (Spang *et al.* 2015; Lazar *et al.* 2016). As these new phyla are still coming to light, information about their evolutionary history, metabolism, and other biological features is not yet well understood. This makes these new phyla prime targets for investigations which may shed more light on their overall biology, along with their similarities to, and differences from, other more well known archaeal phyla.



## Archaeal Similarities to the Bacteria and Eukaryotes

The archaea show many similarities to both eukaryotes and the bacteria. Like the bacteria they are single-celled and lack a nucleus. However, the methods of transcription and translation in archaea also share many similarities with the eukaryotes (Bell & Jackson 1998; Kyripides & Ouzounis 1999).

Eukaryotes have been identified as a closer relative to the archaea than to the bacteria and more recent work point to evidence suggesting that the eukaryotes have in fact arisen from within the archaea rather than having diverged from them closer to the point of the evolutionary split between archaea and the bacteria as initially hypothesised (Archibald 2008; Cox *et al.* 2008; Guy & Ettema 2011; Kelly *et*

*al.* 2011; Raymann *et al.* 2015). This discovery of the relationship between the archaea and the eukaryotes means that continued investigation into the overall biology of the archaea is not only important to gaining a full understanding of the archaea themselves but also to better understanding their relationships to and the evolutionary history they share with both the bacteria and the eukaryotes.

Repertoires of RNA families are another part of archaeal biology where similarities to both the bacteria and the eukaryotes can be seen. In an attempt to characterise RNA families traceable to the last universal common ancestor, Hoepfner *et al.* (2012) found that while individual families of RNA shared little overlap between domains, there is evidence of a handful of RNA families being common to all domains of life. The majority of universally conserved RNA families were noted to be those involved in translation and protein export. However, while individual RNA families that are universally conserved were found to be scarce, many larger classes of RNA are common between domains. Two examples of this are the snoRNAs, common between archaea and eukaryotes (Omer *et al.* 2000; Gaspin *et al.* 2000; Bachellerie *et al.* 2002), and riboswitches which are found in all three domains of life (Sudarsan *et al.* 2003; Vitreschak *et al.* 2004).

## ***Small Non-Coding RNA***

### **RNA**

RNA is a diverse molecule. Aside from the encoding of genetic information like DNA, RNA also has structural properties that allow it to function in a number of different biological processes such as RNA splicing, editing, other regulation of gene expression, and modification guidance (Eddy 2001; Mattick & Makunin 2006). Investigating the diversity and distributions of these RNAs can lead to a greater understanding of their functional role in the organisms they are found in and shed light on the evolutionary history of these RNAs and the organisms in which they operate.

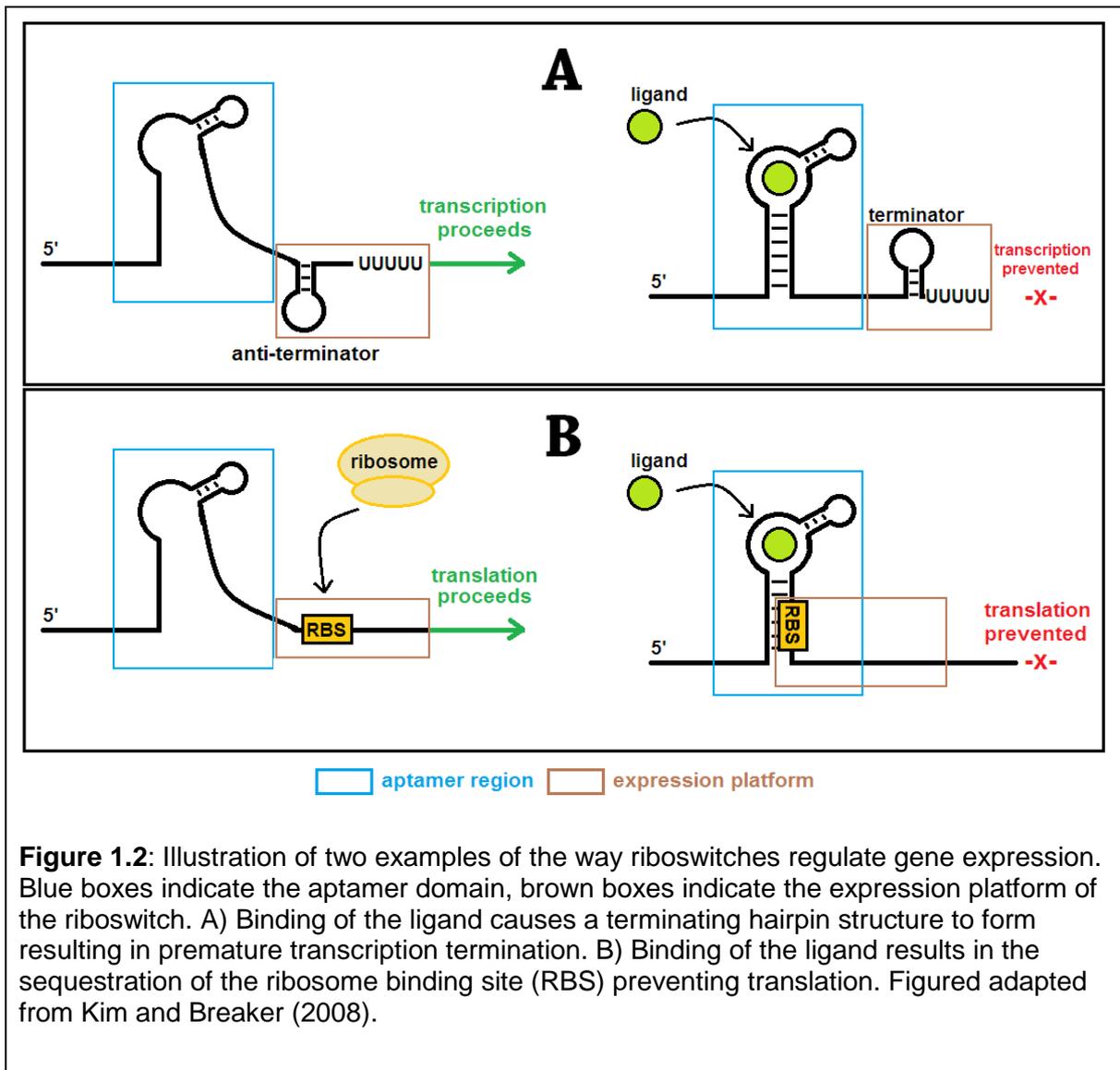
Despite RNA's importance to the biological world, many studies examining and comparing diversity at a genetic or genomic level often overlook non-coding RNAs. Instead, focus is placed on the more easily to examine protein sequences, and on other RNAs, like the 16S rRNA, proven to be useful for constructing higher level phylogenies (Weisburg *et al.* 1991; Hofstetter *et al.* 2007). While these features are useful for examining the broader picture, considering non-coding RNAs in

addition to this other information has potential for providing new and more interesting results on a finer scale (Eddy 2001).

Riboswitches and snoRNAs are two such non-coding RNAs that are interesting to look at when studying the archaea. Riboswitches are primarily known from the bacteria (Winkler *et al.* 2002), while snoRNAs are common between both archaea and eukaryotes (Bachellerie *et al.* 2002). Together, discovering more about the distributions and diversity of these two RNAs within the archaea can help us to build a more complete picture of the links between the archaeal domain and the other two domains within the tree of life.

## **Riboswitches**

Riboswitches are short RNA sequences that control their host gene expression by altering the RNA structural conformation in response to binding metabolites (Breaker 2012). This change in structural conformation can result in changes in gene expression in multiple ways. Namely, through disruption of transcription, RNA processing, or translation processes of the gene the riboswitch regulates. The way in which a riboswitch alters one of these processes is specific to the riboswitch, its sequence and the metabolite it binds (Breaker 2011). Riboswitches consist of two main functional parts. The aptamer domain is responsible for the binding of the metabolite. The aptamer domain of a riboswitch is highly specific to the ligand molecule which the riboswitch senses (Ellington & Szostak 1990). In response to changes in the aptamer, the expression platform part of the riboswitch undergoes structural changes resulting in gene expression being regulated (Winkler *et al.* 2002). The expression platform can either turn off or turn on gene expression in response to the binding of a ligand at the aptamer region. Some ways in which this is achieved include: the ribosome-binding site being sequestered inhibiting translation (Winkler *et al.* 2002; Vitreschak *et al.* 2002), formation of hairpin RNA structures causing premature transcription termination (Vitreschak *et al.* 2002; Sudarsan *et al.* 2003), and indirectly such as a riboswitch on a neighbouring mRNA interfering with transcription of a ribozyme (Andre *et al.* 2008). Figure 1.2 illustrates two of these methods.



**Figure 1.2:** Illustration of two examples of the way riboswitches regulate gene expression. Blue boxes indicate the aptamer domain, brown boxes indicate the expression platform of the riboswitch. A) Binding of the ligand causes a terminating hairpin structure to form resulting in premature transcription termination. B) Binding of the ligand results in the sequestration of the ribosome binding site (RBS) preventing translation. Figure adapted from Kim and Breaker (2008).

While riboswitches have been identified in all domains of life, the majority of knowledge has primarily come from studies on bacteria (Tucker & Breaker 2005; Serganov and Nudler 2013). The role of riboswitches in archaeal gene expression is poorly understood with few known riboswitches having been identified in archaeal genomes (Figure 1.3). There are currently two riboswitches which have been described in archaea, the TPP riboswitch (Rodionov *et al.* 2002; Barrick & Breaker 2007) and the *crcB* RNA motif which is now also known as the Fluoride riboswitch (Weinberg *et al.* 2010; Baker *et al.* 2012). The presence of a third riboswitch, the FMN riboswitch, has been detected bioinformatically (Nawrocki *et al.* 2014). Figure 1.4 gives an overview of the known riboswitch families currently known from each domain of life.

## **The TPP Riboswitch**

The TPP riboswitch has been described in all three domains of life. However experimental work on its structure and function has come predominantly from bacterial samples (Winkler *et al.* 2002; Serganov *et al.* 2006; Lang *et al.* 2007). The TPP riboswitch is known to bind the compound thiamine pyrophosphate (TPP). As such this riboswitch plays a role in the thiamine biosynthesis pathways of the organisms in which it is found (Winkler *et al.* 2002; Rodionov *et al.* 2002). In the archaea, the TPP riboswitch is known to be associated with genes responsible for both biosynthesis and transport of thiamine (Rodionov *et al.* 2002). In the bacteria *E. coli* regulation of gene expression by the TPP riboswitch has been shown to occur by both sequestration of the ribosome binding site and by premature transcription termination (Sudarsan *et al.* 2003).

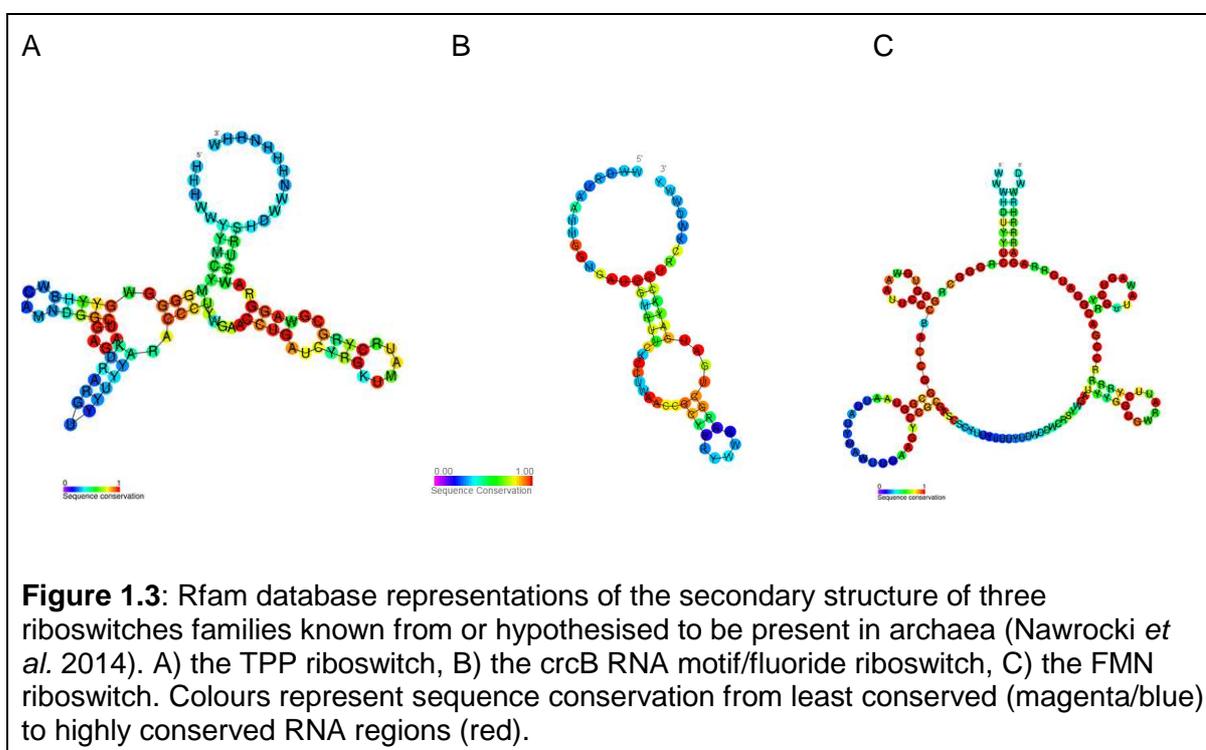
## **The Fluoride Riboswitch**

The *crcB* RNA motif, now also known as the fluoride riboswitch is not currently marked in Rfam 12.0 as a riboswitch entry. However, this RNA is now widely recognised as functioning as a riboswitch in response to fluoride molecules and upregulating gene expression of genes that can mitigate the toxicity of high levels of environmental fluoride (Baker *et al.* 2012). The mechanism by which the change in structural conformation increases expression of these genes has been described by Baker *et al.* (2012) where binding of the fluoride molecules was shown to change sites of spontaneous cleavage such that translation and expression of downstream genes increased. Genes known to be associated with the fluoride riboswitch include those involved with DNA repair, ion transport, and gene known as CRCB. The CRCB protein is hypothesised to function as a fluoride exporter and as such increased expression of this gene in response to higher concentrations of fluoride may be beneficial to an organism in avoiding fluoride toxicity (Baker *et al.* 2012; Stockbridge *et al.* 2012).

The occurrence of the fluoride riboswitch has been previously documented in the archaea, where it is known to be associated with genes encoding proteins mostly associated with ion transport (Weinberg *et al.* 2010; Baker *et al.* 2012).

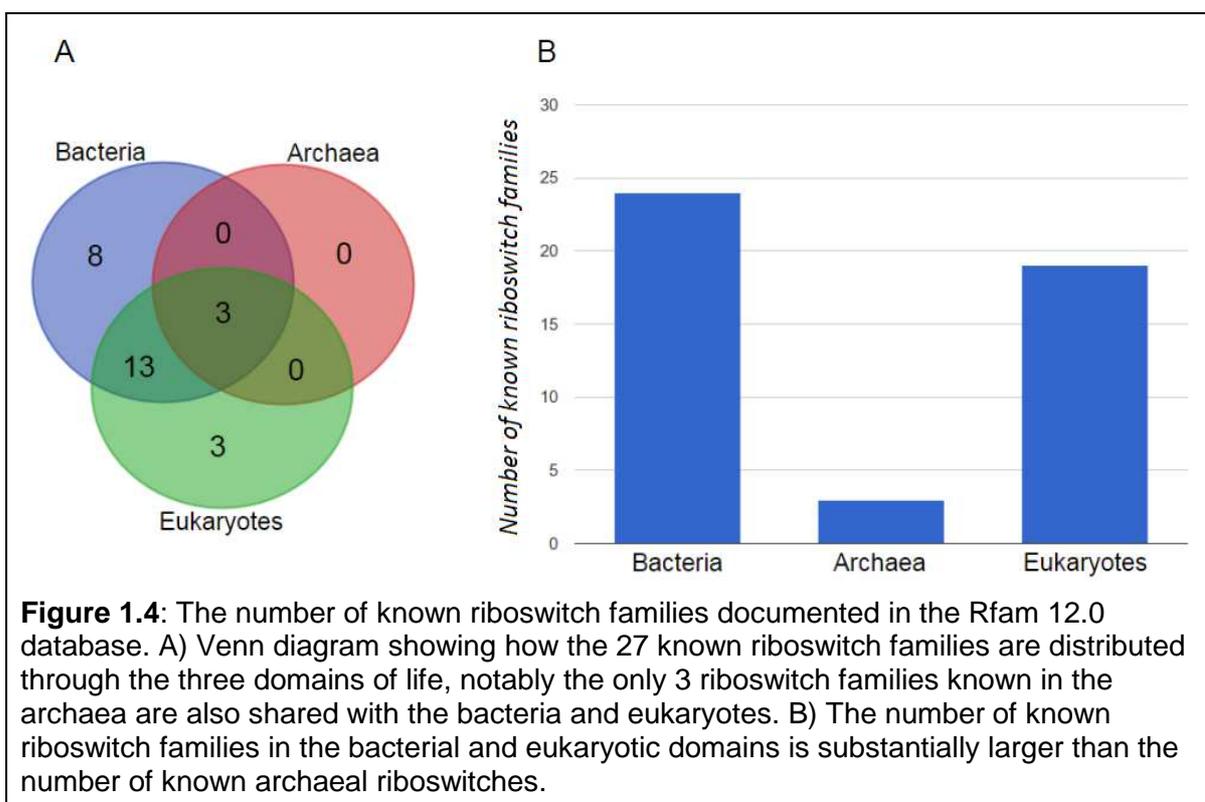
## The FMN Riboswitch

The Flavin mononucleotide (FMN) riboswitch is responsible for the regulation of the riboflavin pathway. It is known to regulate gene expression in *Bacillus subtilis* by a combination of causing premature transcription termination and precluding access to the ribosome-binding site (Pedrolli *et al.* 2015; Vitreschak *et al.* 2002). The riboswitch is known to be associated with a number of proteins involved in riboflavin biosynthesis and transport (Gutiérrez-Preciado *et al.* 2015). The Rfam database entry for this riboswitch suggests it is found in both some eukaryotes and in many strains of a single species of archaea, *Methanobrevibacter smithii* which inhabits the human gut (Samuel *et al.* 2007). However, bacterial species form the predominant knowledge of this riboswitch (Vitreschak *et al.* 2002; Gutiérrez-Preciado *et al.* 2015). Despite Rfam listing the occurrence of this riboswitch in archaea, no studies have as yet explicitly described the presence of the FMN riboswitch in archaea. As such, further examination of archaeal genomes for the presence of this riboswitch is advisable.



From the knowledge we do have of archaeal riboswitches, horizontal gene transfer (HGT) is thought to have had involvement in the distribution of riboswitches

throughout archaeal lineages (Hoepfner *et al.* 2012). HGT is a process by which genetic material is laterally transmitted between organisms rather than being passed down through vertical inheritance (Olendzenski & Gogarten 2009). While HGT can be implicated in the occurrence of riboswitches within the archaea, the pattern of inheritance of riboswitches and the genes they regulate is less well understood. Determining whether riboswitches and genes are inherited together or whether riboswitches can be inherited and function independently of the genes they are known to regulate is therefore of interest to increasing our overall understanding of how riboswitches behave on an evolutionary scale. A comparison of riboswitches and associated genes found in archaea to those found in other domains may also highlight important differences of archaeal riboswitches that may help in identifying new riboswitches or confirming candidate riboswitches. As new genetic data from the archaea becomes increasingly available, opportunities to search for new riboswitches and improve our understanding of the role of known riboswitches within the archaea also become more pronounced.



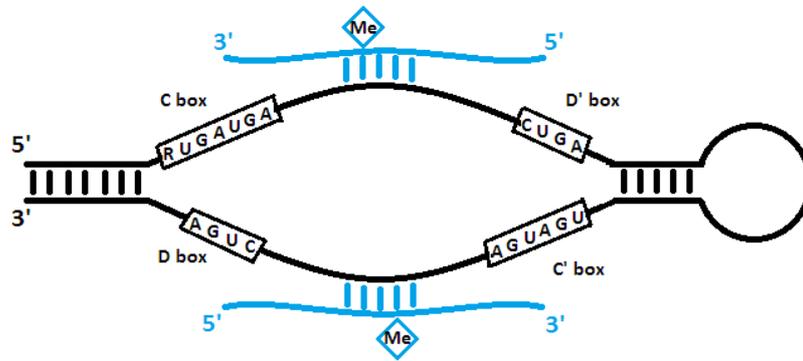
## snoRNAs

Small nucleolar RNAs or snoRNAs are another type of small non-coding RNA found in archaea. These RNAs are responsible for guiding the processes of site-specific methylation and pseudouridylation involved in pre-RNA processing for ribosomal RNA (Kiss-László *et al.* 1998). It has also been noted that this role extends to guiding the same processes for other small RNAs such as tRNAs and snRNAs (Kiss 2001).

There are two known types of snoRNA, classified by conserved features in both their sequence and secondary structure. C/D box snoRNAs have conserved sequences that include the “C” motif RUGAUGA located near the 5' end and the “D” motif CUGA, located near the 3' end of the snoRNA. The nucleotides adjoining these motifs usually form a stem-box structure and help position the target RNA. Another section of conserved nucleotides is complementary to the target RNA and in the case of the C/D box snoRNA forms an RNA duplex with the target RNA enabling and guiding the methylation of the target RNA (Samarsky *et al.* 1998). Proteins known to be associated with the C/D box snoRNA in archaea are fibrillar (Nop1p), NOP56, NOP58, and L7Ae (Yip *et al.* 2013).

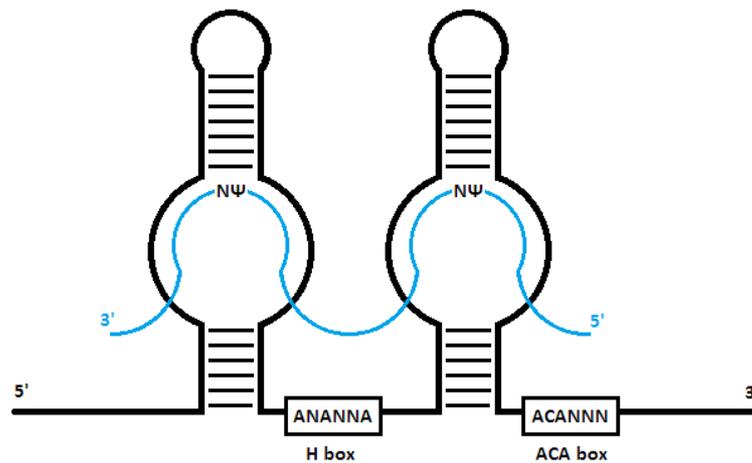
H/ACA snoRNAs enable and guide pseudouridylation of target RNAs. Similar to C/D box snoRNAs, H/ACA snoRNAs include conserved sequences known as the H motif (sequence ANANNA) and the ACA motif (sequence ACA). H/ACA box snoRNAs generally have a more complex secondary structure than C/D box snoRNAs with a secondary structure consisting of a two hairpins and two single-stranded regions (Ganot *et al.* 1997). H/ACA box snoRNAs are known to be associated with the proteins L7Ae, Cbf5p, GAR1, NHP2, and NOP10 (Rozhdestvensky *et al.* 2003; Yip *et al.* 2013). A diagram showing the structure and function of both C/D and H/ACA box snoRNAs can be seen in figure 1.5 below.

## C/D Box snoRNA



Associated snoRNPs: Fibrillarin, Nop5, L7Ae

## H/ACA Box snoRNA



Associated snoRNPs: Cbf5 (TruB\_N), Gar1, Nop10p, L7Ae

**Figure 1.5:** Structure and function of C/D box and H/ACA box snoRNAs. Target RNAs on which modifications are performed are shown in blue. C/D box snoRNAs have conserved C and D box motifs and guide methylation. H/ACA box snoRNAs guide pseudouridylation and have conserved H box and ACA box motifs. Figure adapted from Gardner *et al.* (2010).

Knowledge of snoRNAs is primarily based on studies of eukaryotes (Samarsky *et al.* 1998; Bachellerie *et al.* 2002; Dupuis - Sandoval *et al.* 2015) with archaeal examples of snoRNAs only being identified later (Gaspin *et al.* 2000; Omer *et al.* 2000; Randau 2015). It was first thought that these RNAs only operated within the eukaryotic nucleolus, a cell structure which archaea lack. Despite this, archaea do also carry out analogous processes for preparation of ribosomal RNA and as

such, these small RNAs are still referred to as “snoRNAs” in archaea (Bertrand & Fournier 2004).

In archaea, known examples of snoRNAs have been detected in both the Euryarchaeota and the Crenarchaeota. A bioinformatic approach to snoRNA identification in archaea was undertaken with known or predicted snoRNAs being detected in 33% of crenarchaeal groups and 60% of euryarchaeal groups (Gardner *et al.* 2010). However, since this work was done, the amount of genomic data available for archaea has increased significantly. As discussed above, many new archaeal phyla have been either identified or hypothesised and many more archaeal genomes from within pre-existing archaeal groups published. While there are now greater amounts of data available, there have been no recent studies that have made use of this data to look at patterns of snoRNA distribution. It is therefore prudent that such investigations be made; potentially helping to make sense of classification of this new data and overall evolutionary patterns in all domains. Since snoRNAs appear to be found broadly across eukaryotic diversity (Gardner *et al.* 2010), if snoRNAs are also found to be distributed throughout all archaeal diversity this may lend weight to the idea that snoRNAs evolved early in the evolutionary history of these groups. A distribution that includes some archaeal lineages but not others may suggest that archaeal groups lacking the snoRNP machinery have either undergone lineage-specific losses or represent life that is evolutionarily older than snoRNAs.

## ***Searching Genomes***

The vast amounts of genetic data generated by new techniques such as metagenomics have traditionally been stored in databases such as NCBI's GenBank after they are published (Benson *et al.* 2013). In addition to whole genomes, NCBI also stores short nucleotide sequence and protein sequence data and curates a database of non-redundant sequence data, RefSeq (Pruitt *et al.* 2007; Benson *et al.* 2013). Storing genetic data in databases such as these makes the data readily accessible to those interested in analysis and provides a point of comparison for researchers' own samples. Data in these databases is valuable for the functioning of tools like BLAST, a tool traditionally used to compare sequence features between organisms (Altschul *et al.* 1990; Johnson *et al.* 2008). Comparisons of sequence features, whether they be nucleotide or protein sequences, are valuable as they provide clues as to the function of sequence features of interest, help with

identification of unknown sequence data, and can provide clues about the evolutionary relationships between sets of similar sequences.

## **Tools for the Comparison of Genomic Data**

BLAST is a sequence comparison tool that uses heuristic techniques to find sequences similar to the input/target sequence without requiring full alignments of the sequence data involved. Consequently, BLAST is faster and less computationally expensive than prior techniques such as the Smith–Waterman algorithm and FASTA (Altschul *et al.* 1990; Gish & States 1993). BLAST is useful for finding similar sequences not only of the same sequence type but also of other sequence types through the different BLAST programs such as Blastn, Blastp, Blastx and tBlastn (Altschul *et al.* 1990; Camacho *et al.* 2009). BLAST however is just one technique used to make these kinds of comparisons. Homology-based searches, such as those using covariance models like the INFERNAL package and Hidden-Markov-Models like HMMER, are increasingly being used as powerful tools to analyse sequence features within genetic data.

Hidden Markov Models, or HMMs are a method for describing distributions of probabilities (Eddy 2004). Profile HMMs can be used to model probability distributions within sequence data such as in DNA, RNA and proteins using a position-specific scoring system (Eddy 1998). Profile HMMs represent a technological improvement over the BLAST algorithm, being again less computationally intensive while also having and improved accuracy (Johnson *et al.* 2010).

Profile HMMs use consensus information from multiple sequence alignments to assess sequence similarities. As such, better models are able to be built where more examples of homologous sequences are available (Eddy 1998). This tool can also be used in an iterative fashion, building up a model from data retrieved during each subsequent pass of the HMM search. This allows for similar sequences that may not have been detected during an initial search with a less optimal model to be detected in subsequent searches as the model becomes refined (Eddy 2011).

Online servers for these programs can be used to make searching online genetic databases including UniProt, SwissProt and Ensembl Genomes more accessible. With web interfaces to the tools in question, results can be presented in a

more interactive format; helping to link various data together (Finn *et al.* 2011; Finn *et al.* 2015a).

Covariance models or CMs are similar to HMMs in that they are used to describe probability distributions within sequence data (Eddy and Durbin 1994). CMs differ from HMMs in their ability to also describe probability of secondary structure based on the sequence data. This has particular use in the modeling and comparison of RNA molecules (Nawrocki 2009, Nawrocki & Eddy 2013a). Both riboswitches and snoRNAs have important secondary structures that play a role in their biological function (Breaker 2011; Ganot *et al.* 1997; Samarsky *et al.* 1998). Being able to detect these structures based on sequence characteristics is therefore important and useful within the scope of this project. The INFERNAL package provides a useful toolset for working with covariance models to search sequence databases for candidate RNAs (Nawrocki & Eddy 2013b).

## **Databases of Sequence Features**

Databases that store information about known sequences and structures and the organisms they are found in are invaluable to the world of bioinformatics. Such databases are in fact crucial to the effective functioning of tools such as those outlined above. Pfam and its sister database Rfam are two such databases that store information about protein sequences and RNA sequences and structure respectively.

Pfam, the protein families database stores sequence and similarity information about an abundance of known proteins. Specifically, Pfam focuses on grouping proteins by “family” or rather by regions that share significant sequence similarity (Finn *et al.* 2013). Pfam itself uses the above outlined HMMER program to detect these sequence similarities (Finn *et al.* 2015b). For this study, the use of the Pfam HMM models of similarity for different protein families, is useful for both searching target archaeal genomes for evidence of proteins known to be associated with non-coding RNAs of interest, and also for identifying proteins that are found in nearby genomic regions where evidence for a target non-coding RNA has been found.

Rfam is a database of similar premise to Pfam but relating to RNA families rather than protein families. Data stored in Rfam about known RNA “families” (again groupings of sequences with significant sequence similarity) not only includes sequence information but also information about secondary structure of the RNA (Nawrocki *et al.* 2014). Covariance models for each Rfam family provide a useful starting place for searches for families of interest within novel genomes. The data

Rfam stores on the species that have contributed to the sequence data and alignments for the covariance models is also useful for providing clues about whether or not a given Rfam family would be expected to be found in the results of a search of a given genome. Rfam families that list a particular species as contributing to an alignment used for the families covariance model should be expected to be found in genome sequences from that species.

## ***Objectives of this Thesis***

Using the tools discussed above, this project gives new insights into previously understudied aspects of the archaea. We seek to show an updated distribution of both the riboswitches and the snoRNAs within the currently described archaeal taxa. We look to provide new information about the function of riboswitches within the archaea by determining the genes commonly associated with each riboswitch family occurring within archaeal species. We also investigate the potential for horizontal transfer of riboswitch-gene pairs to have shaped the distribution of riboswitch families within the archaea. Finally, we examine the distribution and prevalence of known snoRNAs families and their associated protein families within the archaeal taxa. We explore limitations of using only current information of known archaeal snoRNA families to predict their presence in novel archaeal taxa.

## ***References***

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403-410.

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59(1), 143-169.

Andre, G., Even, S., Putzer, H., Burguiere, P., Croux, C., Danchin, A., Martin-Verstraete, I, & Soutourina, O. (2008). S-box and T-box riboswitches and antisense RNA control a sulfur metabolic operon of *Clostridium acetobutylicum*. *Nucleic acids research*, 36(18), 5955-5969.

Archibald, J. M. (2008). The eocyte hypothesis and the origin of eukaryotic cells. *Proceedings of the National Academy of Sciences*, 105(51), 20049-20050.

- Attar, N. (2015). Archaeal genomics: A new phylum for methanogens. *Nature Reviews Microbiology*, 13(12), 739-739.
- Bachellerie, J. P., Cavaillé, J., & Hüttenhofer, A. (2002). The expanding snoRNA world. *Biochimie*, 84(8), 775-790.
- Baker, J. L., Sudarsan, N., Weinberg, Z., Roth, A., Stockbridge, R. B., & Breaker, R. R. (2012). Widespread genetic switches and toxicity resistance proteins for fluoride. *Science*, 335(6065), 233-235.
- Barns, S. M., Fundyga, R. E., Jeffries, M. W., & Pace, N. R. (1994). Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proceedings of the National Academy of sciences*, 91(5), 1609-1613.
- Barrick, J. E., & Breaker, R. R. (2007). The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome biology*, 8(11), 1.
- Bell, S. D., & Jackson, S. P. (1998). Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. *Trends in microbiology*, 6(6), 222-228.
- Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic acids research*, 41(D 1), D36-D42.
- Bertrand, E., & Fournier, M. J. (2004). The snoRNPs and Related Machines: Ancient Devices That Mediate Maturation of rRNA. *The nucleolus*, 223.
- Breaker, R. R. (2011). Prospects for riboswitch discovery and analysis. *Molecular cell*, 43(6), 867-879.
- Breaker, R. R. (2012). Riboswitches and the RNA world. *Cold Spring Harbor perspectives in biology*, 4(2), a003566.
- Brochier-Armanet, C., Boussau, B., Gribaldo, S., & Forterre, P. (2008). Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology*, 6(3), 245-252.
- Burggraf, S., Heyder, P., & Eis, N. (1997). A pivotal Archaea group. *Nature*, 385(6619), 780.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, 10(1), 1.

- Castelle, C. J., Wrighton, K. C., Thomas, B. C., Hug, L. A., Brown, C. T., Wilkins, M. J., Frischkorn, K.R., Tringe, S.G., Singh, A., Markillie, L.M., & Taylor, R. C. (2015). Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Current Biology*, 25(6), 690-701.
- Coenye, T., & Vandamme, P. (2003). Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiology Letters*, 228(1), 45-49.
- Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R., & Embley, T. M. (2008). The archaeobacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences*, 105(51), 20356-20361.
- Dupuis - Sandoval, F., Poirier, M., & Scott, M. S. (2015). The emerging landscape of small nucleolar RNAs in cell biology. *Wiley Interdisciplinary Reviews: RNA*, 6(4), 381-397.
- Eddy, S. R., & Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic acids research*, 22(11), 2079-2088.
- Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9), 755-763.
- Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12), 919-929.
- Eddy, S. R. (2004). What is a hidden Markov model?. *Nature biotechnology*, 22(10), 1315-1316.
- Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput Biol*, 7(10), e1002195.
- Eisen, J. A. (2007). Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol*, 5(3), e82.
- Ellington, A. D., & Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *nature*, 346(6287), 818-822.
- Federhen, S. (2012). The NCBI taxonomy database. *Nucleic acids research*, 40(D1), D136-D143.
- Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, gkr367.
- Finn, R. D., Bateman, A., Clements, J., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2013). Pfam: the protein families database. *Nucleic acids research*, gkt1223.

Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F., Bateman, A., & Eddy, S. R. (2015a). HMMER web server: 2015 update. *Nucleic acids research*, gkv397.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., & Salazar, G. A. (2015b). The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*, gkv1344.

Ganot, P., Caizergues-Ferrer, M., & Kiss, T. (1997). The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes & Development*, 11(7), 941-956.

Gardner, P. P., Bateman, A., & Poole, A. M. (2010). SnoPatrol: how many snoRNA genes are there?. *Journal of biology*, 9(1), 1.

Gaspin, C., Cavallé, J., Erauso, G., & Bachellerie, J. P. (2000). Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the *Pyrococcus* genomes. *Journal of molecular biology*, 297(4), 895-906.

Gish, W., & States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nature genetics*, 3(3), 266-272

Gutiérrez-Preciado, A., Torres, A. G., Merino, E., Bonomi, H. R., Goldbaum, F. A., & García-Angulo, V. A. (2015). Extensive identification of bacterial riboflavin transporters and their distribution across bacterial species. *PLoS one*, 10(5), e0126124.

Guy, L., & Ettema, T. J. (2011). The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends in microbiology*, 19(12), 580-587.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews*, 68(4), 669-685.

Hoepfner, M. P., Gardner, P. P., & Poole, A. M. (2012). Comparative analysis of RNA families reveals distinct repertoires for each domain of life. *PLoS Comput Biol*, 8(11), e1002752.

Hoepfner, M. P., & Poole, A. M. (2012). Comparative genomics of eukaryotic small nucleolar RNAs reveals deep evolutionary ancestry amidst ongoing intragenomic mobility. *BMC evolutionary biology*, 12(1), 1.

Hofstetter, V., Miadlikowska, J., Kauff, F., & Lutzoni, F. (2007). Phylogenetic comparison of protein-coding versus ribosomal RNA-coding sequence data: a case study of the Lecanoromycetes (Ascomycota). *Molecular phylogenetics and evolution*, 44(1), 412-426.

Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology*, 180(18), 4765-4774.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome biology*, 3(2), 1.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic acids research*, 36(suppl 2), W5-W9.

Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics*, 11(1), 1.

Kelly, S., Wickstead, B., & Gull, K. (2011). Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1708), 1009-1018.

Kim, J. N., & Breaker, R. R. (2008). Purine sensing by riboswitches. *Biology of the Cell*, 100(1), 1-11.

Kiss, T. (2001). Small nucleolar RNA - guided post - transcriptional modification of cellular RNAs. *The EMBO journal*, 20(14), 3617-3622.

Kiss - László, Z., Henry, Y., & Kiss, T. (1998). Sequence and structural elements of methylation guide snoRNAs essential for site - specific ribose methylation of pre - rRNA. *The EMBO Journal*, 17(3), 797-807.

Kyrpides, N. C., & Ouzounis, C. A. (1999). Transcription in archaea. *Proceedings of the National Academy of Sciences*, 96(15), 8545-8550.

Kyrpides, N. C., Hugenholtz, P., Eisen, J. A., Woyke, T., Göker, M., Parker, C. T., Amann, R., Beck, B. J., Chain, P. S., Chun, J., & Colwell, R. R. (2014). Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biol*, 12(8), e1001920.

Lang, K., Rieder, R., & Micura, R. (2007). Ligand-induced folding of the thiM TPP riboswitch investigated by a structure-based fluorescence spectroscopic approach. *Nucleic acids research*, 35(16), 5370-5378.

Lazar, C. S., Baker, B. J., Seitz, K., Hyde, A. S., Dick, G. J., Hinrichs, K. U., & Teske, A. P. (2016). Genomic evidence for distinct carbon substrate preferences and ecological niches of Bathyarchaeota in estuarine sediments. *Environmental microbiology*.

Mattick, J. S., & Makunin, I. V. (2006). Non-coding RNA. *Human molecular genetics*, 15(suppl 1), R17-R29.

Meng, J., Xu, J., Qin, D., He, Y., Xiao, X., & Wang, F. (2014). Genetic and functional properties of uncultivated MCG archaea assessed by metagenome and gene expression analyses. *ISME J* 8: 650-659. *Isme Journal*, 8, 650-659.

Nawrocki, E. (2009). Structural RNA homology search and alignment using covariance models.

Nawrocki, E. P., & Eddy, S. R. (2013a). Computational identification of functional RNA homologs in metagenomic data. *RNA biology*, 10(7), 1170-1179.

Nawrocki, E. P., & Eddy, S. R. (2013b). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22), 2933-2935.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., & Finn, R. D. (2014). Rfam 12.0: updates to the RNA families database. *Nucleic acids research*, gku1063.

Nunoura, T., Hirayama, H., Takami, H., Oida, H., Nishi, S., Shimamura, S., Suzuki, Y., Inagaki, F., Takai, K., Nealon, K.H., & Horikoshi, K. (2005). Genetic and functional properties of uncultivated thermophilic crenarchaeotes from a subsurface gold mine as revealed by analysis of genome fragments. *Environmental microbiology*, 7(12), 1967-1984.

Nunoura, T., Takaki, Y., Kakuta, J., Nishi, S., Sugahara, J., Kazama, H., Chee, G.J., Hattori, M., Kanai, A., Atomi, H., & Takai, K. (2010). Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic acids research*, gkq1228.

Olendzenski, L., & Gogarten, J. P. (2009). Evolution of Genes and Organisms. The tree/web of life in light of horizontal gene transfer. *Annals of the New York Academy of Sciences*, 1178(1), 137-145.

Omer, A. D., Lowe, T. M., Russell, A. G., Eberhardt, H., Eddy, S. R., & Dennis, P. P. (2000). Homologs of small nucleolar RNAs in Archaea. *Science*, 288(5465), 517-522.

Pace, N. R. (2006). Time for a change. *Nature*, 441(7091), 289-289.

Pedrolli, D., Langer, S., Hobl, B., Schwarz, J., Hashimoto, M., & Mack, M. (2015). The ribB FMN riboswitch from *Escherichia coli* operates at the transcriptional and translational level and regulates riboflavin biosynthesis. *FEBS journal*, 282(16), 3230-3242.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(suppl 1), D61-D65.

Randau, L. (2015). Evolution of small guide RNA genes in hyperthermophilic archaea. *Annals of the New York Academy of Sciences*, 1341(1), 188-193.

Raymann, K., Brochier-Armanet, C., & Gribaldo, S. (2015). The two-domain tree of life is linked to a new root for the Archaea. *Proceedings of the National Academy of Sciences*, 112(21), 6670-6675.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., & Dodsworth, J. A. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459), 431-437.

Rodionov, D. A., Vitreschak, A. G., Mironov, A. A., & Gelfand, M. S. (2002). Comparative genomics of thiamin biosynthesis in procaryotes New genes and regulatory mechanisms. *Journal of Biological chemistry*, 277(50), 48949-48959.

Rozhdestvensky, T. S., Tang, T. H., Tchirkova, I. V., Brosius, J., Bachellerie, J. P., & Hüttenhofer, A. (2003). Binding of L7Ae protein to the K - turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic acids research*, 31(3), 869-877.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N.J., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology*, 12(1), 1.

Samarsky, D. A., Fournier, M. J., Singer, R. H., & Bertrand, E. (1998). The snoRNA box C/D motif directs nucleolar targeting and also couples snoRNA synthesis and localization. *The EMBO Journal*, 17(13), 3747-3757.

Samuel, B. S., Hansen, E. E., Manchester, J. K., Coutinho, P. M., Henrissat, B., Fulton, R., Latreille, P., Kim, K., Wilson, R.K., & Gordon, J. I. (2007). Genomic and metabolic adaptations of *Methanobrevibacter smithii* to the human gut. *Proceedings of the National Academy of Sciences*, 104(25), 10643-10648.

Sapp, J. (2005). The prokaryote-eukaryote dichotomy: meanings and mythology. *Microbiology and molecular biology reviews*, 69(2), 292-305.

Schmieder, R., & Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one*, 6(3), e17288.

Seitz, K. W., Lazar, C. S., Hinrichs, K. U., Teske, A. P., & Baker, B. J. (2016). Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *The ISME journal*.

Serganov, A., Polonskaia, A., Phan, A. T., Breaker, R. R., & Patel, D. J. (2006). Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature*, 441(7097), 1167-1171.

- Serganov, A., & Nudler, E. (2013). A decade of riboswitches. *Cell*, 152(1), 17-24.
- Sharon, I., & Banfield, J. F. (2013). Genomes from metagenomics. *Science*, 342(6162), 1057-1058.
- Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521(7551), 173-179.
- Stockbridge, R. B., Lim, H. H., Otten, R., Williams, C., Shane, T., Weinberg, Z., & Miller, C. (2012). Fluoride resistance and transport by riboswitch-controlled CLC antiporters. *Proceedings of the National Academy of Sciences*, 109(38), 15289-15294.
- Sudarsan, N., Barrick, J. E., & Breaker, R. R. (2003). Metabolite-binding RNA domains are present in the genes of eukaryotes. *Rna*, 9(6), 644-647.
- Tucker, B. J., & Breaker, R. R. (2005). Riboswitches as versatile gene control elements. *Current opinion in structural biology*, 15(3), 342-348.
- Vitreschak, A. G., Rodionov, D. A., Mironov, A. A., & Gelfand, M. S. (2002). Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic acids research*, 30(14), 3141-3151.
- Vitreschak, A. G., Rodionov, D. A., Mironov, A. A., & Gelfand, M. S. (2004). Riboswitches: the oldest mechanism for the regulation of gene expression?. *TRENDS in Genetics*, 20(1), 44-50.
- Weinberg, Z., Wang, J. X., Bogue, J., Yang, J., Corbino, K., Moy, R. H., & Breaker, R. R. (2010). Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome biology*, 11(3), 1.
- Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of bacteriology*, 173(2), 697-703.
- Williams, T. A., & Embley, T. M. (2014). Archaeal "dark matter" and the origin of eukaryotes. *Genome biology and evolution*, 6(3), 474-481.
- Winkler, W., Nahvi, A., & Breaker, R. R. (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, 419(6910), 952-956.
- Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12), 4576-4579.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., Kunin, V., Goodwin, L., Wu, M., Tindall, B.J., & Hooper, S. D. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, 462(7276), 1056-1060.

Yip, W. S. V., Vincent, N. G., Baserga, S. J. (2013). Ribonucleoproteins in archaeal pre-rRNA processing and modification. *Archaea*, 2013.

# Chapter Two - The Function and Distribution of Riboswitches in Archaea

## *Introduction*

As the number of archaeal genomes available to study has expanded (Rinke *et al.* 2013; Spang *et al.* 2015; Lazar *et al.* 2015), investigations into the prevalence and functions of many RNAs within these genomes has not kept pace. In this study we sought to determine the distribution of all known riboswitch families currently documented in the Rfam 12.0 database (Nawrocki *et al.* 2014) across the currently known archaeal genomes available from Genbank (Benson *et al.* 2013). This investigation provides new information about how commonplace riboswitches are in archaea compared to their prevalence bacterial and eukaryotic genomes.

We also sought to determine the nature of genes found downstream of each riboswitch identified in the archaea. We show cases of genes found downstream of a given riboswitch differing from both those previously documented as being associated with a particular riboswitch. This finding may be used to enhance knowledge about the both the evolutionary and functional properties of riboswitches in general.

Finally, in cases where the identification of downstream genes suggested high similarity of the archaeal riboswitch-gene pairing to bacterial species, further analysis was carried out to examine whether horizontal gene transfer (HGT) from bacterial species into archaea could be a possible explanation for the occurrence of the riboswitch-gene pairing in archaea. Our findings suggest that there is some evidence that the presence of some gene-riboswitch pairings within archaea may be explained by a horizontal transfer event from bacteria.

## *Methods*

### **Investigating the Distribution of Riboswitches in the Archaea**

#### **Dataset**

To examine the distribution of riboswitches across a representative range of archaeal diversity, genomic data for all archaeal species listed in NCBI's Taxonomy

browser (Federhen 2012), that had whole genome data available and were not classed as “environmental samples”, were collected from Genbank (Benson *et al.* 2013) to form a dataset of archaeal genomes. This dataset included nucleotide sequences from 463 archaeal species representing 13 archaeal phyla or candidate phyla. For a list of all species represented in the dataset see Appendix 1.

A dataset of covariance models for 27 known riboswitch families were downloaded from the RNA families database Rfam 12.0 (Nawrocki *et al.* 2014). Rfam identifiers for included riboswitch families are listed in Table 2.1 below.

**Table 2.1:** Riboswitch families for which cm models were retrieved from Rfam

| Rfam Accession | Description                                     |
|----------------|-------------------------------------------------|
| RF00050        | FMN riboswitch (RFN element)                    |
| RF00059        | TPP riboswitch (THI element)                    |
| RF00162        | SAM riboswitch (S box leader)                   |
| RF00167        | Purine riboswitch                               |
| RF00168        | Lysine riboswitch                               |
| RF00174        | Cobalamin riboswitch                            |
| RF00234        | glmS glucosamine-6-phosphate activated ribozyme |
| RF00504        | Glycine riboswitch                              |
| RF00521        | SAM riboswitch (alpha-proteobacteria)           |
| RF00522        | PreQ1 riboswitch                                |
| RF00634        | S-adenosyl methionine (SAM) riboswitch          |
| RF01054        | preQ1-II (pre queuosine) riboswitch             |
| RF01055        | Moco (molybdenum cofactor) riboswitch           |
| RF01056        | Magnesium Sensor                                |
| RF01057        | S-adenosyl-L-homocysteine riboswitch            |
| RF01482        | AdoCbl riboswitch                               |
| RF01510        | M. florum riboswitch                            |
| RF01689        | AdoCbl variant RNA                              |
| RF01725        | SAM-I/IV variant riboswitch                     |
| RF01727        | SAM/SAH riboswitch                              |
| RF01734        | Fluoride riboswitch                             |
| RF01767        | SMK box translational riboswitch                |
| RF01786        | Cyclic di-GMP-II riboswitch                     |
| RF01787        | drz-agam-1 riboswitch                           |
| RF01788        | drz-agam-2-2 riboswitch                         |
| RF01826        | SAM-V riboswitch                                |
| RF01831        | THF riboswitch                                  |

## **Analysis Methods**

Homology searches are powerful tools for determining the presence of known sequence features within sequence data (Nawrocki & Eddy 2013a). In this study we use covariance models to determine the presence of riboswitches across the archaeal phyla in our dataset and analyse the genes found adjacent to identified riboswitches.

To determine presence of riboswitches across archaeal phyla, the INFERNAL package was used to perform a cmsearch of all genomes in the dataset against all riboswitch families (Nawrocki & Eddy 2013b). Default settings for bit-score and e-value were used for reporting significance in the cmsearch (Nawrocki & Eddy 2014). Positions of each significant hit within the genome were recorded.

## **Examining Protein-Coding Genes Found Downstream of Riboswitches in the Archaea**

### **Analysis Methods**

To analyse the genes which each detected riboswitch is likely to influence expression of, Prodigal 2.6.2 with default settings was first used to create protein translations from the nucleotide sequences of all genomes in the dataset (Hyatt *et al.* 2010). The translated protein sequences for the first three open reading frames (ORFs) downstream of the location of each previously located riboswitch occurrence were then analysed using the online phmmer server (<https://www.ebi.ac.uk/Tools/hmmer/search/phmmer>) to search the UniProt database of proteins (Finn *et al.* 2015a; UniProt Consortium 2015). For each translated protein searched in this way, the Pfam domain(s) detected of the closest gene match were recorded (Finn *et al.* 2015b; Finn *et al.* 2013). Also recorded were the most closely related protein matches from UniProt and which domain of life these related proteins were found in. A combination of the names of the closest gene matches and the closest Pfam domains identified were then researched to identify the likely metabolic pathway the associated riboswitch was involved with along with likely evolutionary relationships of the protein-riboswitch systems.

## Identifying the likelihood of Horizontal Gene Transfer of Riboswitch-Gene pairs from Bacteria into the Archaea

### Dataset

Three types of riboswitch-gene pairings were identified through the phmmer search of the UniProtKB database as having high similarity to bacterial protein sequences rather than to other archaeal protein sequences. For each of these three riboswitch-gene pairings, a dataset comprised of protein sequences with high similarity to the downstream gene were compiled.

For each dataset, protein sequences for all genes of the same type found downstream of the same riboswitch in archaeal species were added to an alignment which also contained approximately 75 additional highly similar protein sequences from both other archaeal species and bacterial species retrieved from UniProtKB. The datasets for each riboswitch-gene pairing are described in detail below.

### *Dataset for the TPP riboswitch-Thi4 gene pairing*

Four sequences from downstream of the TPP riboswitch in archaea were retrieved by translating ORFs from downstream of the riboswitch location using Prodigal (Hyatt *et al.* 2010). Translated ORFs for other archaeal phyla were then searched for similar protein sequences using an hmm built from the initial four sequences (Eddy 2013). 18 similar protein sequences (the best match to the hmm from each phyla with hmmsearch e-value  $< 1.0 \times 10^{-10}$ ) were subsequently added to the dataset. A further 87 sequences similar to the Thi4 gene, representing both bacterial and archaeal species, were also added to the initial dataset. These were selected using an hmmsearch (<https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch>) of the UniProtKB database for sequences similar to an alignment of the initial Thi4 ORFs downstream of the TPP riboswitch in the archaeal species it was detected in with a bit score significance cutoff of 300.

### *Dataset for the FMN riboswitch-DHBP<sub>synthase</sub> gene pairing*

Nine sequences from downstream of the FMN riboswitch in archaea were retrieved by translating ORFs from downstream of the riboswitch location using Prodigal (Hyatt *et al.* 2010). Translated ORFs for other archaeal phyla were then searched for similar protein sequences using an hmm built from the initial nine sequences (Eddy 2013). 16 similar protein sequences (the best match to the hmm

from each phyla with hmmsearch e-value  $< 1.0 \times 10^{-10}$ ) were subsequently added to the dataset. A further 75 sequences similar to the DHBP\_synthase gene, representing both bacterial and archaeal species, were also added to the initial dataset. These were selected using an hmmsearch (<https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch>) of the UniProtKB database for sequences similar to an alignment of the initial DHBP\_synthase ORFs downstream of the FMN riboswitch in the archaeal species it was detected in with a bit score significance cutoff of 333.

#### ***Fluoride riboswitch-Na\_H\_exchanger gene pairing***

17 sequences from downstream of the Fluoride riboswitch in archaea were retrieved by translating ORFs from downstream of the riboswitch location using Prodigal (Hyatt *et al.* 2010). Translated ORFs for other archaeal phyla were then searched for similar protein sequences using an hmm built from the initial 17 sequences (Eddy 2013). 19 similar protein sequences (the best match to the hmm from each phyla with hmmsearch e-value  $< 1.0 \times 10^{-10}$ ) were subsequently added to the dataset. Again, a further 56 sequences similar to the Na\_H\_exchanger gene, representing both bacterial and archaeal species, were also added to the initial dataset. These were selected using a hmmsearch (<https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch>) of the UniProtKB database for sequences similar to an alignment of the initial Na\_H\_exchanger ORFs downstream of the Fluoride riboswitch in the archaeal species it was detected in with a bit score significance cutoff of 300.

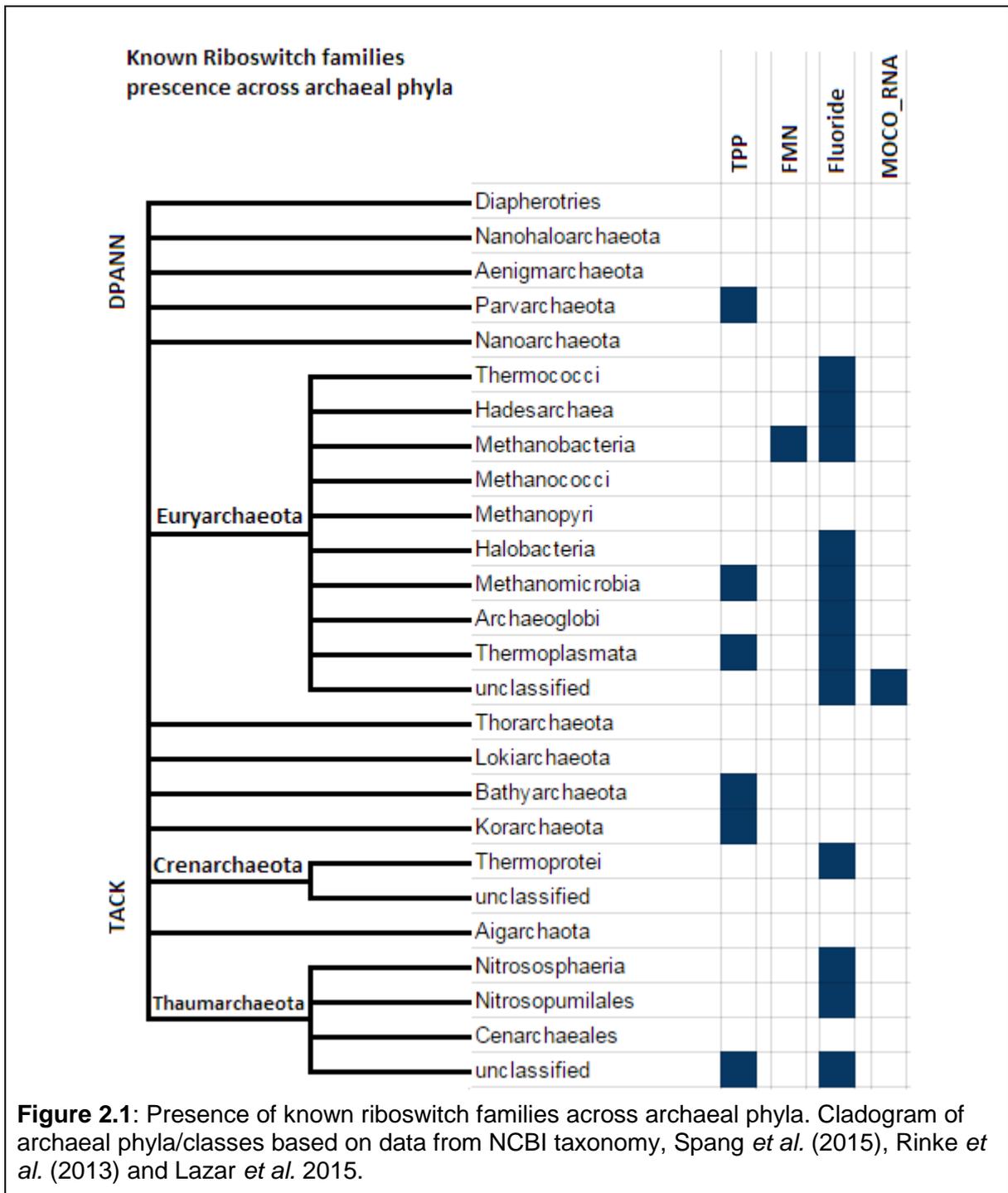
#### **Analysis Methods**

Preliminary phylogenetic trees for each of the three datasets were created using the <http://www.phylogeny.fr> online one-click method (Dereeper *et al.* 2008). Results of each tree analysis were used to eliminate sequences from the UniProt set of bacterial/archaeal samples with high redundancy from each dataset before a more thorough phylogenetic analysis was carried out. After this step, the remaining datasets consisted of 37 Thi4 sequences, 40 DHBP\_synthase sequences and 51 Na\_H\_exchanger sequences. Genome data from the remaining bacterial species from each dataset were examined for presence of the riboswitch paired with that dataset's gene using the methods described in the section "*Investigating the Distribution of Riboswitches in the Archaea*" above.

For each resulting dataset the protein sequences were aligned using MUSCLE (Edgar 2004), with conserved regions of the resulting alignment selected for use in phylogenetic analysis using the G-blocks program with the settings “Allow smaller final blocks” and “Allow gap positions within the final blocks” selected (Castresana 2000). PhyML was then used to perform phylogenetic analysis of the protein alignments with branch support tested using the Approximate Likelihood-Ratio Test (Guindon & Gascuel 2003; Anisimova & Gascuel 2006). TreeDyn was then used to provide a visual representation of the trees generated with PhyML (Chevenet *et al.* 2006). The relationships between species found in the trees generated were then visually compared to species relationships in reference trees for archaea (Rinke *et al.* 2013; Spang *et al.* 2015; Lazar *et al.* 2015) and to NCBI taxonomy (Federhen 2012) to determine the likelihood of gene transfer event compared to the likelihood of a vertical inheritance pattern for the protein involved.

## **Results**

To determine the overall distribution of known riboswitches throughout the archaea, covariance models of each known riboswitch family were used to search genomes representing 26 archaeal classes within 13 archaeal phyla. Of the 27 known riboswitch families represented, only four families were found to be present in archaeal genomes. The overall distribution of these families throughout the archaea was limited, with only 15 of the archaeal classes within 6 of the archaeal phyla studied showing evidence of riboswitch presence. The presence of each riboswitch family in each taxa of archaea studied is summarised in Figure 2.1 below.



## The TPP Riboswitch in Archaea

The TPP riboswitch is one of two known riboswitch families previously documented in archaea (Rodionov *et al.* 2002; Barrick & Breaker 2007). However, its distribution in more recently described archaeal phyla is not yet documented. Our investigation using covariance models to search both well known and newer archaeal

phyla revealed that while this riboswitch is found in many archaeal species, it is not widely distributed through all archaeal phyla.

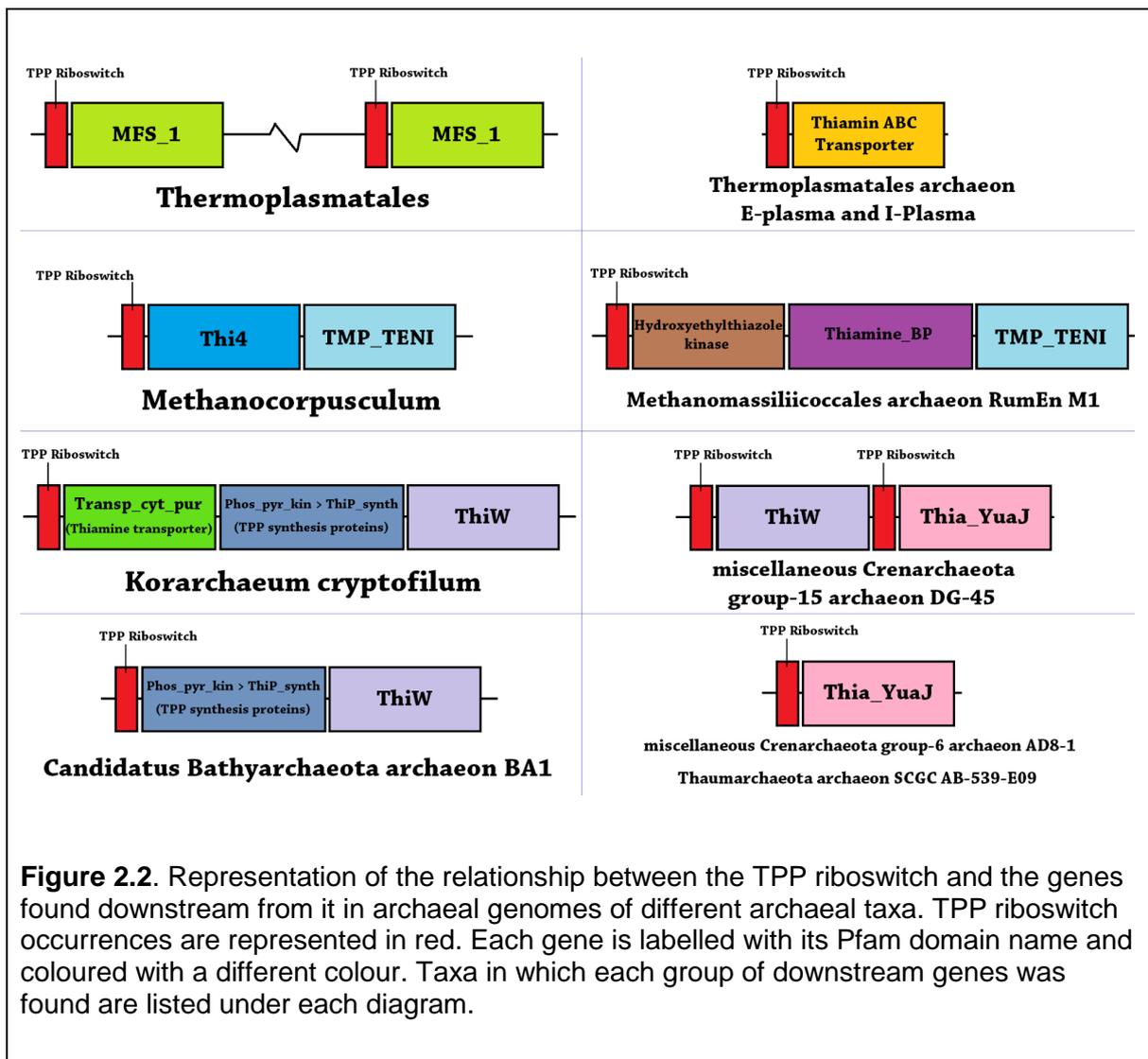
The TPP Riboswitch family was found in five of the archaeal phyla studied. These phyla include the Euryarchaeota, Parvarchaeota, Thaumarchaeota, Korarchaeota and Bathyarchaeota. Newly identified occurrences of this riboswitch include its presence in three Bathyarchaeota species, one Thaumarchaeota species, and seven additional species from the Euryarchaeota classes Thermoplasmata and Methanomicrobia from which the TPP riboswitch was previously known.

Coverage of each archaeal taxa the TPP riboswitch was identified in was incomplete except in the case of the Korarchaeota in which only one species represents the entire phyla. The TPP riboswitch was identified in only three of the nine species of Bathyarchaeota, one of the 23 species of unclassified Thaumarchaeota, one of the three species of Parvarchaeota, 13 of the 23 species of Thermoplasmata and two of the 61 species of Methanomicrobia studied. However, in the case of the Methanomicrobia, the riboswitch was located in both species that currently represent the genus *Methanocorpusculm* and no other methanomicrobial species. This suggests the riboswitch is restricted to this one genus among the Methanomicrobia.

## **Genes found downstream of the TPP riboswitch in Archaea**

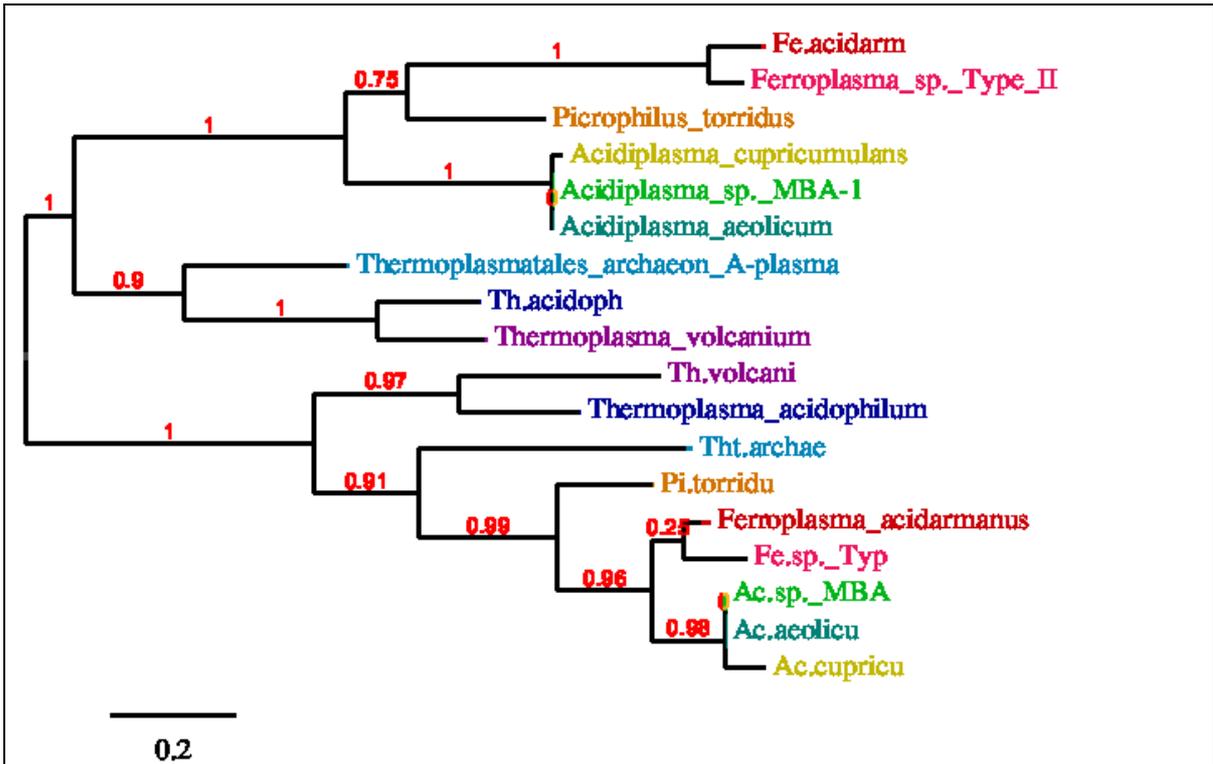
The genes associated with the TPP riboswitch in archaea were examined by comparing translated nucleotide sequences of open reading frames (ORFs) detected downstream of each riboswitch occurrence to the UniProtKB proteins database (Finn *et al.* 2015a; UniProt Consortium 2015).

Genes found downstream of the TPP riboswitch include the transporter proteins MFS\_1/Sugar\_tr along with proteins involved in thiamine biosynthesis and transport.



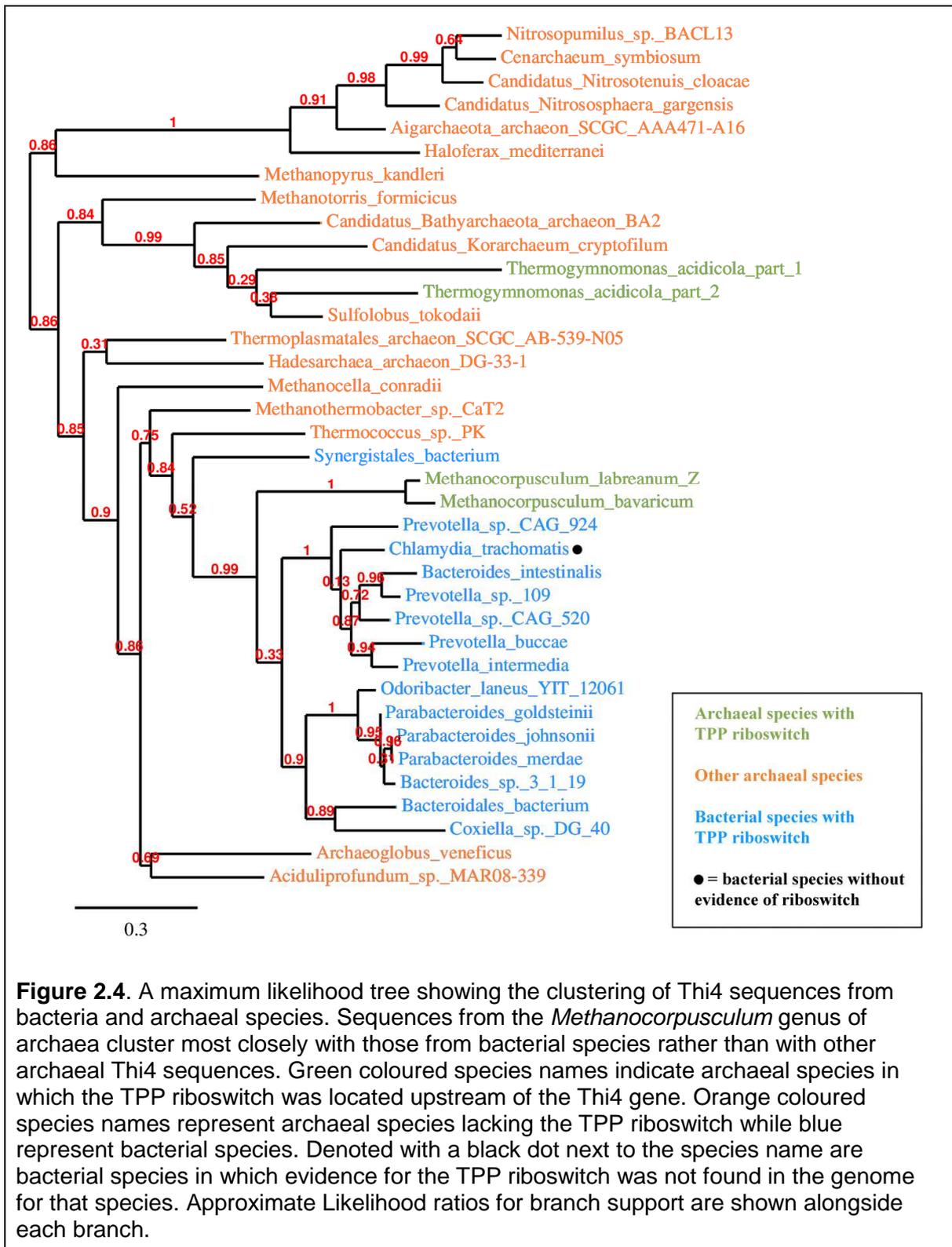
The protein sequences translated from ORFs located downstream of the TPP riboswitches identified can be broadly classed into two categories: transporter proteins, and thiamine biosynthesis proteins. The thiamine biosynthesis proteins detected include Thi4, TMP-TENI, Phos\_pyr\_kin and ThiP\_synth. The transporter proteins include MFS\_1/Sugar\_tr, Transp\_cyt\_pur, Thiamin ABC Transporter, ThiW, Hydroxyethylthiazole kinase, and Thia\_YuaJ. For more detail on each of these protein families see Table 2.2

Many species of archaea in the class Thermoplasmata were found to contain two copies of the TPP riboswitch. In the majority of these cases the protein found immediately downstream of this riboswitch was a transporter protein either MFS\_1 or Sugar\_tr. We compared the sequences thought to code for the MFS\_1/Sugar\_tr protein in these species with a preliminary phylogenetic analysis and found a distinct split in similarity between the two copies from each genome (Figure 2.3).



**Figure 2.3.** Maximum likelihood tree generated using approximate likelihood ratio tests and the WAG evolutionary model of protein substitution. Each species is coded with a different colour to show that the two copies of the MFS\_1/Sugar\_tr protein in each genome have distinct differences from one another (note that second occurrences of each species name have been truncated).

The translated ORFs downstream of the TPP riboswitch in the genus *Methanococcus* showed high similarity to protein sequences of bacterial Thi4 proteins during the process of identifying candidate genes from downstream ORFs. We therefore investigated the relationship of the ORFs corresponding to Thi4 in *Methanococcus* to protein sequences coding for Thi4 in both bacterial and archaeal species.



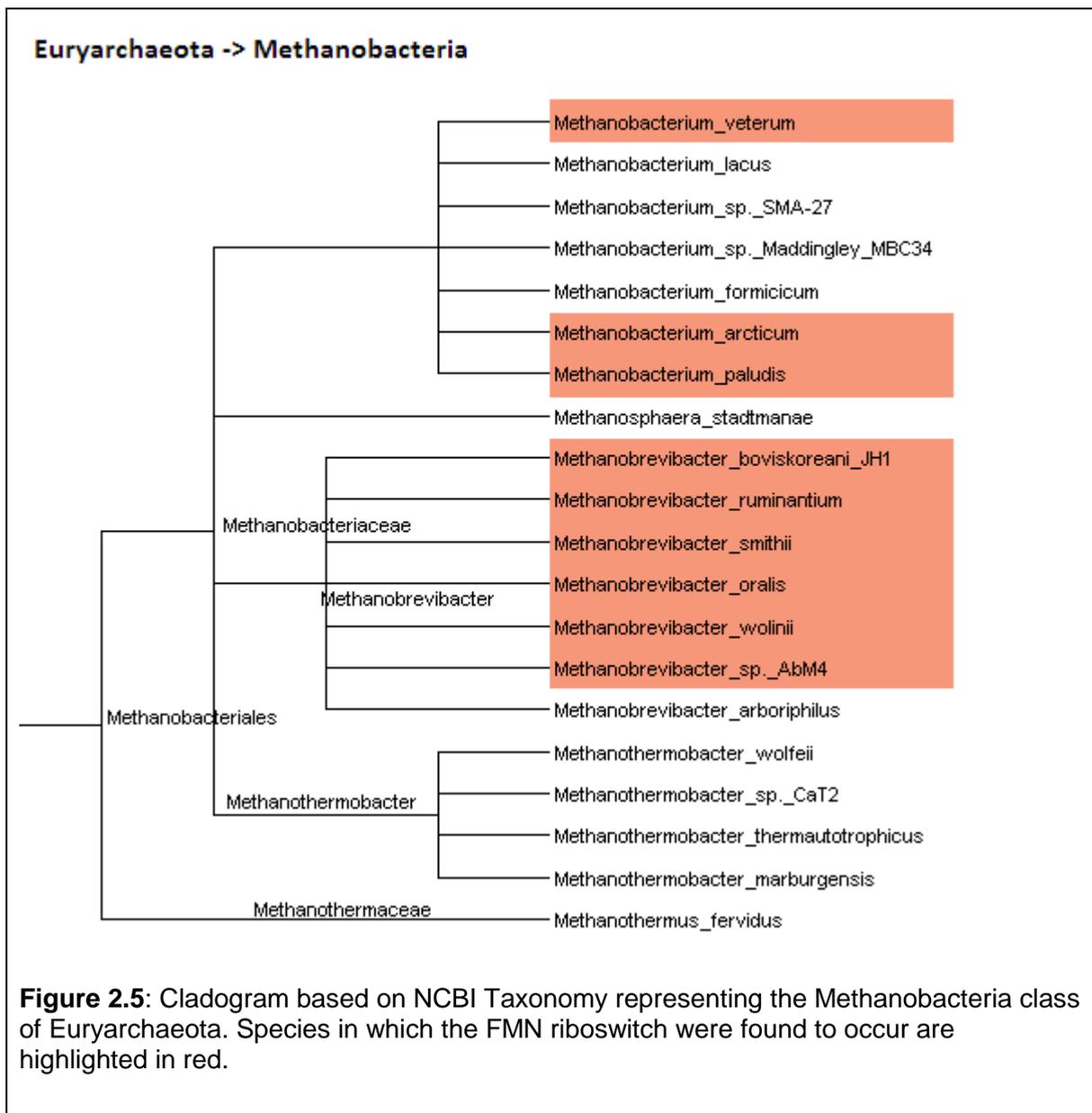
It was found that the Thi4-corresponding ORFs from *Methanocorpusculum* clustered most closely with Thi4 sequences from bacterial species rather than with Thi4 sequences from other archaeal species (Figure 2.4). It was also noted that while other archaeal Thi4 protein sequences did not show evidence of the TPP riboswitch

being located upstream, the most similar bacterial species showed evidence of the riboswitch located upstream. This suggests the possibility that there has been a horizontal transfer event of the TPP riboswitch-Thi4 gene system from bacterial species into the *Methanocorpusculum* genus.

## The FMN Riboswitch in Archaea

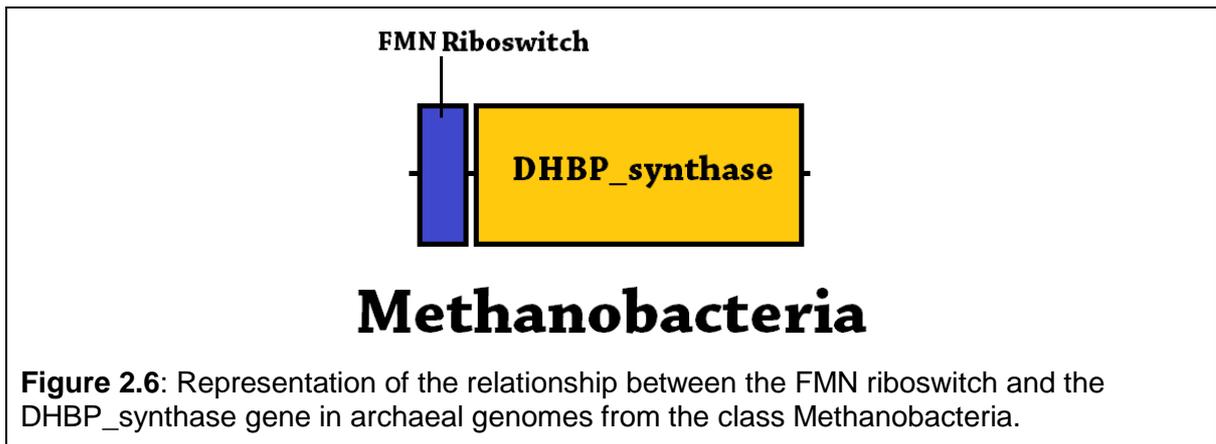
While the FMN or riboflavin mononucleotide riboswitch is documented in the Rfam database as being present in strains of the methanobacterial species *Methanobrevibacter smithii* (Nawrocki *et al.* 2014), its presence in other archaeal species is not well understood. We investigated the distribution of this riboswitch across 13 archaeal phyla using covariance models to search both well known and newer archaeal genomes. This analysis revealed that the distribution of the FMN riboswitch appears to be restricted to the archaeal class Methanobacteria.

Previously only known from *Methanobrevibacter smithii*, the FMN riboswitch was identified in nine of 20 species of Methanobacteria. This included six of seven species in the genus *Methanobrevibacter* and two of six species in the genus *Methanobacterium*. No examples of the FMN riboswitch were conclusively identified in the genera *Methanothermobacter*, *Methanosphaera*, or *Methanothermus* (Figure 2.5).

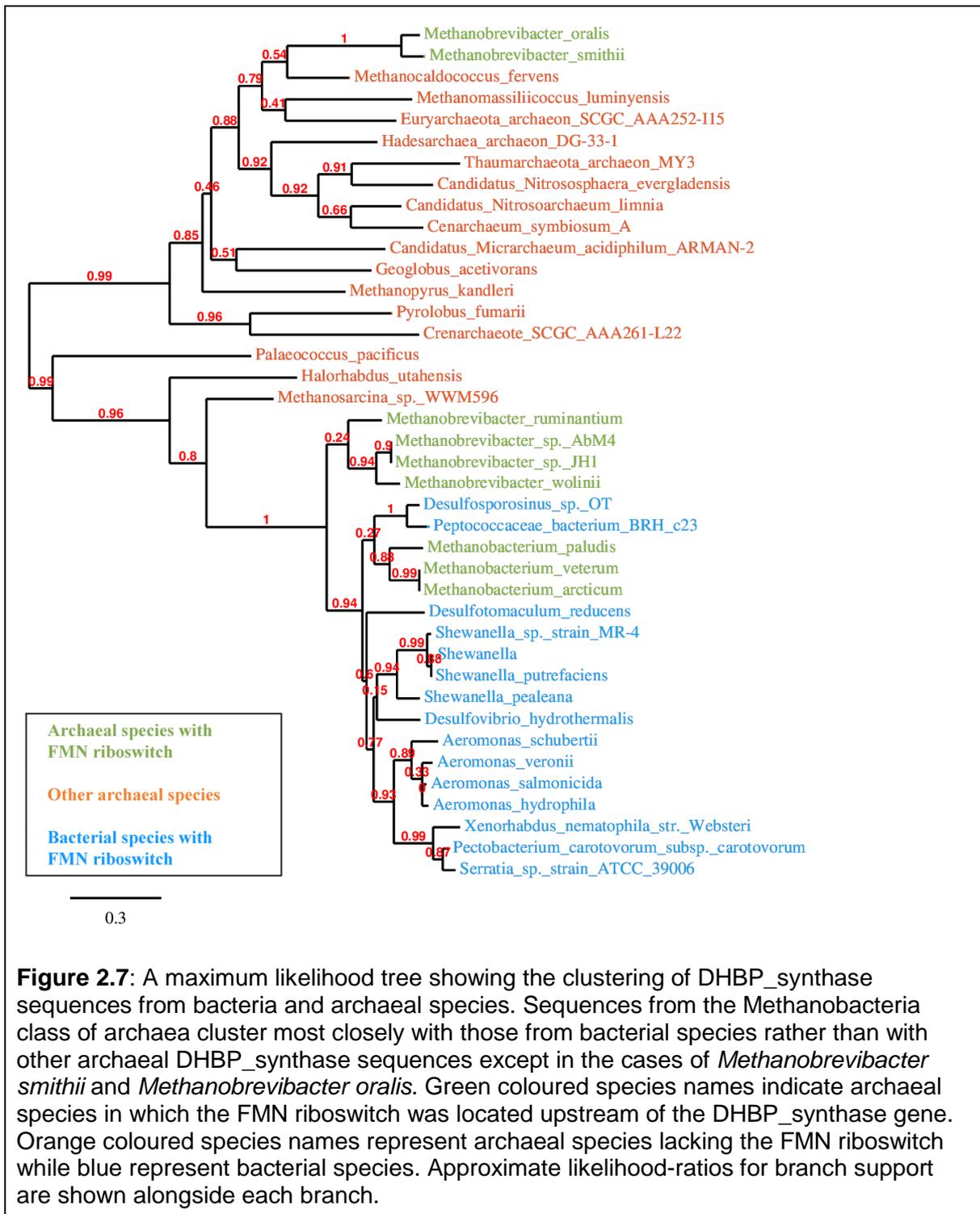


## Genes downstream of the FMN riboswitch in Archaea

ORFs detected downstream of the FMN riboswitch in the Methanobacteria species the riboswitch was found to occur in were translated to protein sequences and compared using profile hidden markov models to the UniProtKB database of proteins (Finn *et al.* 2015a; UniProt Consortium 2015). It was found that in all cases where the riboswitch was located, downstream ORFs corresponded to sequences for the riboflavin precursor synthesis protein 3,4-dihydroxy-2-butanone-4-phosphate synthase (DHBP\_synthase). This protein is associated with the biosynthesis pathway of flavin mononucleotide, the metabolite the FMN riboswitch is known to sense.



The majority of the translated downstream ORFs sequences were also noted to be most similar to DHBP\_synthase protein sequences from bacterial species when searching the UniProt database. A subsequent phylogenetic analysis of these downstream ORFs compared to both bacterial DHBP\_synthase protein sequences and protein sequences from archaeal species matching the DHBP\_synthase protein model revealed that the DHBP\_synthase protein sequences translated from ORFs downstream of the FMN riboswitch in the Methanobacteria clustered into two groups. One group, comprised of protein sequences translated from *Methanobrevibacter smithii* and *Methanobrevibacter oralis* cluster most closely with other archaeal protein sequences. The other group clustered more closely to the bacterial protein sequences included in the analysis. However, the clustering of the second group shows some variance in which bacterial sequences the methanobacterial ORFs are most similar to (Figure 2.7). Further analysis of the bacterial species the DHBP\_synthase protein sequences originated from suggests that the FMN riboswitch is commonly found upstream of this protein sequence in the bacterial species it occurs in. Archaeal species outside of the Methanobacteria do not show evidence of the FMN riboswitch upstream of this protein sequence. A horizontal gene transfer event of the FMN riboswitch-DHBP\_synthase gene pair from bacteria into the Methanobacteria is one possible explanation for the pattern seen in this case.



## The Fluoride Riboswitch in Archaea

The Fluoride riboswitch is the second of two known riboswitch families previously recorded in archaea (Baker *et al.* 2012). However, the extent of its distribution in light of recently described archaeal phyla has not been well studied. We investigated the distribution of this riboswitch across the currently known

archaeal phyla. We found evidence that this riboswitch is the most commonly occurring riboswitch in the archaea. However, there are still many archaeal phyla which lack any evidence of the occurrence known riboswitch families.

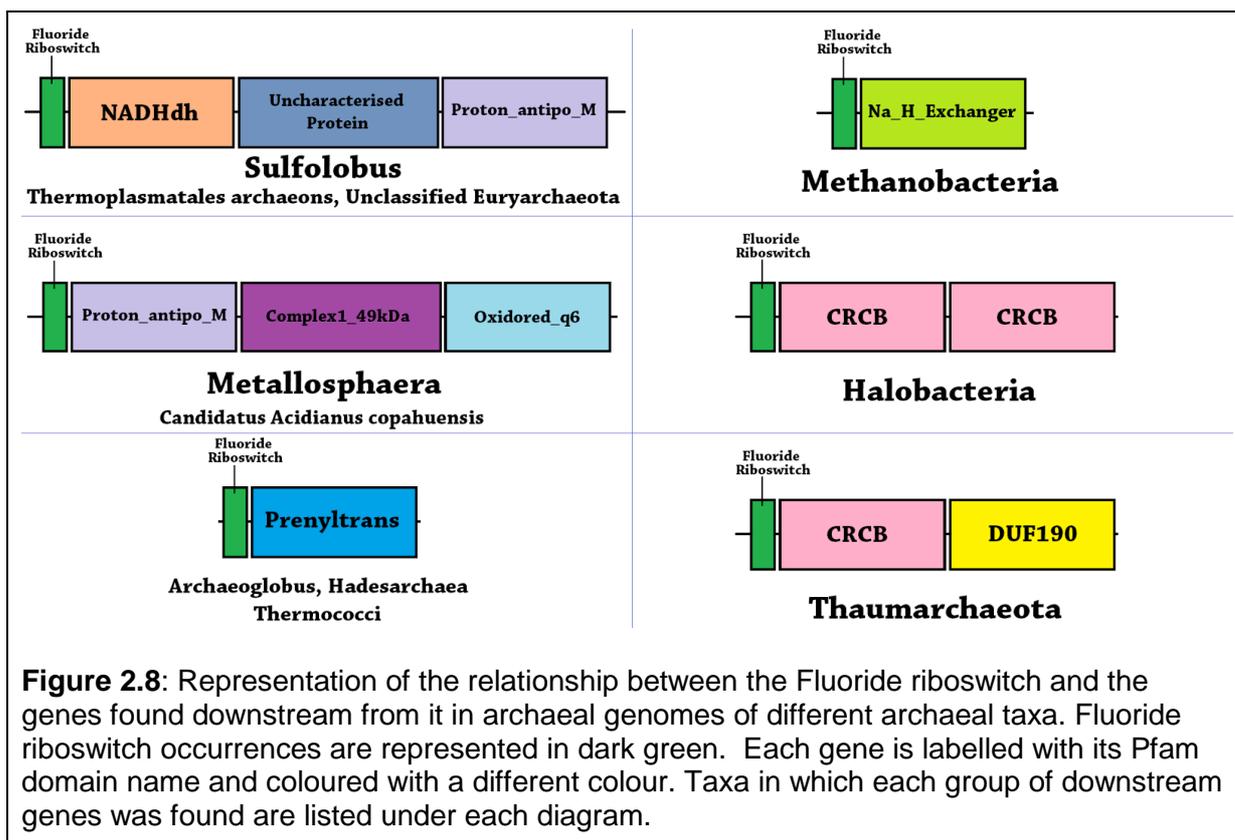
The Fluoride riboswitch family was found in three of the archaeal phyla studied. These phyla include the Euryarchaeota, Thaumarchaeota, and Crenarchaeota. While the presence of this riboswitch was previously documented in both the Euryarchaeota and the Crenarchaeota as listed in Rfam 12.0 (Nawrocki *et al.* 2014), its presence in additional species in each of the taxa was discovered. Other new discoveries include its presence within the Thaumarchaeota and in one species from the class Hadesarchaea within the Euryarchaeota.

With the exception of the Thaumarchaeota class Nitrososphaera where the Fluoride riboswitch was found in all three species representing the class, coverage of each archaeal taxa the Fluoride riboswitch was identified in was incomplete. The Fluoride riboswitch was identified in only 3 of the 36 other species of phylum Thaumarchaeota and only in 10 of the 117 Thermoprotei species within the Crenarchaeota phylum. In the Euryarchaeota, while the riboswitch was found in 26 of the 28 species of the Thermococci, its presence was only detected in 50 of the remaining 266 species making up the Euryarchaeota.

## **Genes downstream of the Fluoride riboswitch in Archaea**

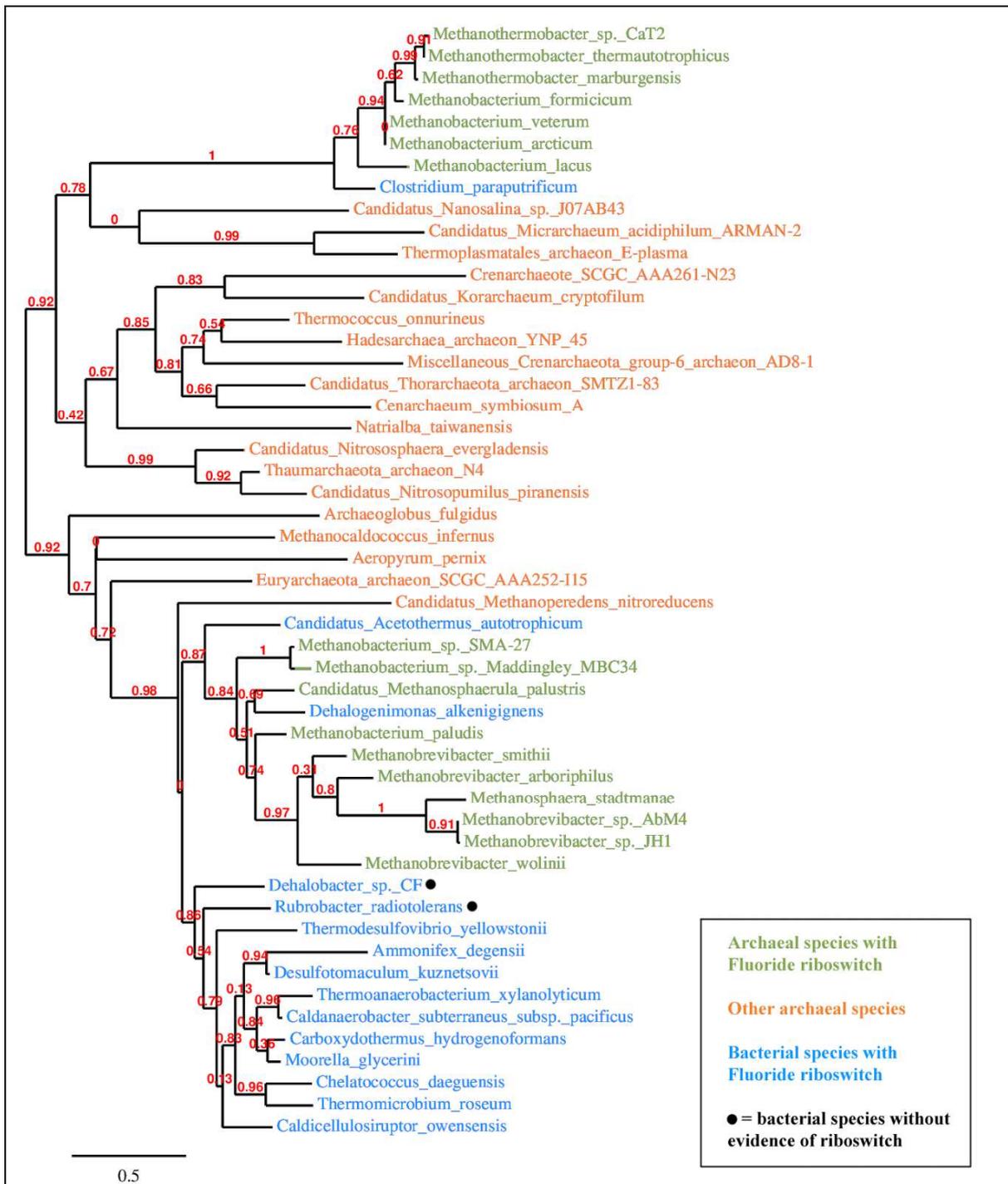
As in the case of the other riboswitch families detected in the archaea, ORFs occurring downstream of the Fluoride riboswitch in the species the riboswitch was identified in were translated to protein sequences and compared using profile hidden markov models to the UniProtKB database of proteins (Finn *et al.* 2015a; UniProt Consortium 2015). The genes found downstream of the riboswitch differed among the different archaeal taxa the riboswitch was found in although in the majority of cases the downstream gene was consistent within each taxa. A summary of the downstream genes identified and the taxa each riboswitch-gene(s) pairing was associated with is presented in Figure 2.8.

The downstream genes detected were mostly associated with ion transport. The presence of CRCB, a fluoride ion transporter, is a positive indication that the riboswitch located in these genomes is indeed regulating gene expression in response to binding of the fluoride ion. However, transporters of other ions such as Na<sup>+</sup> and H<sup>+</sup> were also found downstream of the riboswitch in many of the taxa the riboswitch was detected in.



ORFs downstream of the riboswitch in the Archaeoglobi, Hadesarchaea and Thermococci were a partial match in many cases to the Pfam domain Prenyltrans. However, the sequence coverage of this match was poor and other results from these classes most closely matched uncharacterised archaeal proteins. It is therefore possible that these ORFs represent a protein of as yet unknown function rather than coding for the Prenyltransferase protein.

Translated ORFs downstream of the Fluoride riboswitch in the Methanobacteria most closely matched both bacterial and archaeal examples of the Na<sub>2</sub>H<sub>2</sub> exchanger protein when searching the UniProtKB database based on sequence similarity. For this reason, further investigation was carried out to attempt to determine whether a horizontal gene transfer event may be a possible explanation for the similarity to bacterial sequences. It was found that while some of the Na<sub>2</sub>H<sub>2</sub> exchanger corresponding ORFs from the Methanobacteria did cluster closely with examples of the Na<sub>2</sub>H<sub>2</sub> exchanger protein from bacteria, many of the sequences were a closer match to other sequences from archaeal species (Figure 2.9).



**Figure 2.9:** A maximum likelihood tree showing the clustering of Na<sub>H</sub> exchanger sequences from bacteria and archaeal species. Green coloured species names indicate archaeal species in which the Fluoride riboswitch was located upstream of the Na<sub>H</sub> exchanger gene. Orange coloured species names represent archaeal species lacking the Fluoride riboswitch while blue represent bacterial species. A black dot next to the species name denotes bacterial species in which evidence for the Fluoride riboswitch was not found in the genome for that species. Approximate likelihood-ratios for branch support are shown alongside each branch. While many of archaeal species with evidence of the fluoride riboswitch cluster most closely with other archaeal species, others cluster closer to bacterial examples of the gene.

## The Moco\_RNA\_motif in Archaea

Moco\_RNA\_motif is a presumed riboswitch that is thought to bind molybdenum cofactor (Regulski *et al.* 2008). Rfam 12.0 currently lists this riboswitch as occurring in both Bacteria and Eukaryotes (Nawrocki *et al.* 2014). Our investigation detected a single probable occurrence of this riboswitch in an Euryarchaeota genome, *Euryarchaeote* SCGC AAA261-G15. The single example detected had a relatively high bit score in terms of other riboswitch matches found (77.5, e-value  $2.3 \times 10^{-16}$ ) but was found in a genome where genomic context for confirming whether the riboswitch occurrence was likely to be genuine was limited.

The riboswitch was reported as being on the sense strand while the closest ORF reported by prodigal was located on the antisense strand. There was also a 55 base-pair distance between the end of the riboswitch found by cmsearch and the start of the ORF found by Prodigal. No genes corresponding to ORFs detected by Prodigal close to the riboswitch location were able to be identified using phmmer. Nearby ORFs were most similar to uncharacterised protein sequences from the bacterial species *Paenibacillus polymyxa*. A subsequent search of this species's genome using cmsearch suggests that the Moco\_RNA\_motif is not present in this species and is therefore unlikely to be a source of the riboswitch detected in *Euryarchaeote* SCGC AAA261-G15.

An alternative approach was next used to attempt to identify the genomic context of the Moco\_RNA\_motif in this case. Blastn with default settings was used to match the raw nucleotide sequence the riboswitch hit was found in, to other organisms in the NCBI database (Altschul *et al.* 1990; Johnson *et al.* 2008). This search yielded a match to three bacterial species in the *Desulfotomaculum* genus. The matches found had e-values of between  $5 \times 10^{-12}$  and  $4 \times 10^{-13}$ . The sequence similarities found covered only 4% of the inputted sequence data and matches were found only around 200 base-pairs downstream of the site the riboswitch occurrence was located.

Blastx (again using default settings) was also used to again try to identify any protein matches from the original sequence data the riboswitch occurrence was found in (Altschul *et al.* 1990; Johnson *et al.* 2008). Three significant protein hits were discovered in the sequence using this method, having e-values of  $6 \times 10^{-11}$ ,  $6 \times 10^{-11}$  and  $1 \times 10^{-08}$  respectively. These hits matched hypothetical proteins from the bacterial family *Peptococcaceae*. This is the same bacterial family from which the earlier discovered *Desulfotomaculum* genus is from. However, none of the protein hits found

were from this same genus. The protein hits found also only covered 12% of the inputted sequence and were more than 600 base-pairs downstream of the location of the riboswitch as reported by cmsearch. Several other non-significant protein matches were also found to hypothetical bacterial proteins.

**Table 2.2:** Genes associated with riboswitches detected in the archaea. A longer description of each of the genes corresponding to downstream ORFs mentioned in Figures 2.2, 2.6 and 2.8 above is detailed in column 3. Column 2 lists the Pfam domain associated with the downstream ORF and column 1 lists the riboswitch family associated with the gene.

| Associated Riboswitch | Pfam Domain             | Protein description                                                  |
|-----------------------|-------------------------|----------------------------------------------------------------------|
| TPP                   | MFS_1                   | Major Facilitator Superfamily                                        |
|                       | Sugar_tr                | Sugar (and other) transporter                                        |
|                       | Thi4                    | Thiamine biosynthesis protein                                        |
|                       | TMP-TENI                | Thiamine monophosphate synthase/Thiazole tautomerase                 |
|                       | Transp_cyt_pur          | Permease for cytosine/purines, uracil, thiamine, allantoin           |
|                       | ThiW                    | Thiamine-precursor transporter protein                               |
|                       | Thiamin ABC transporter | Thiamine transporter                                                 |
|                       | HK                      | Hydroxyethylthiazole kinase (thiamine metabolism-associated protein) |
|                       | Thiamine_BP             | Thiamine-binding protein                                             |
|                       | Thia_YuaJ               | Thiamine transporter protein                                         |
| Fluoride              | NADHdh                  | NADH dehydrogenase                                                   |
|                       | Proton_antipo_M         | Proton-conducting membrane transporter                               |
|                       | Complex1_49kDa          | Respiratory-chain NADH dehydrogenase, 49 Kd subunit                  |
|                       | Oxidored_q6             | NADH ubiquinone oxidoreductase, 20 Kd subunit                        |
|                       | Prenyltrans             | Prenyltransferase and squalene oxidase repeat                        |
|                       | Na_H_exchanger          | Sodium/hydrogen exchanger family                                     |
|                       | CRCB                    | Camphor Resistance protein/putative fluoride ion transporter         |
|                       | DUF190                  | Domain of unknown function                                           |
| FMN                   | DHBP_synthase           | Catalytic enzyme for the biosynthetic precursor to riboflavin        |

## ***Discussion***

### **The Distribution of Riboswitches in the Archaea**

In this study we found that the distribution of riboswitches across archaeal phyla is limited. Occurrences of only four of the 27 known riboswitch families were found. The TPP riboswitch was found in members of the Parvarchaeota, Methanomicrobia, Thermoplasmata, Bathyarchaeota, Korarchaeota, and Thaumarchaeota. This extends previous knowledge of its distribution into two additional archaeal phyla. We noted the presence of the Fluoride riboswitch in several classes of Euryarchaeota along with examples from the Crenarchaeota and Thaumarchaeota. Presence of the FMN riboswitch was also found in eight further species in addition to its occurrence in *Methanobrevibacter smithii* as listed in Rfam 12.0 (Nawrocki *et al.* 2014).

Our finding of an occurrence of the Moco\_RNA\_motif, a riboswitch not previously known from the archaea, was restricted to a single archaeal genome and convincing genomic context to support the validity of the result was not able to be obtained. Partial matches from Blast searches of the archaeal genome involved in both the protein and nucleotide spaces suggest the genetic data this match for the Moco\_RNA\_motif riboswitch was found in could be linked to members of the bacterial family *Peptococcaceae*. The species of *Peptococcaceae* that provide the closest Blast matches to the sequence data from the archaeal genome are all listed in Rfam as having at least one occurrence of the Moco\_RNA\_motif riboswitch. This suggests that these bacteria may be a potential origin of the riboswitch found in our archeal sequence. However, it is still unclear whether this could be because of a gene transfer event or contaminated sequence data that was incorrectly assembled from an environmental sample. This could be tested further by examining sequence similarity and gene ordering in both the archaeal genome and in the candidate source bacteria. While the presence of the Moco\_RNA\_motif riboswitch makes sense in the context of the Blast matches in this case, the presence of both this riboswitch and the sequence data matching closely to bacterial species do not necessarily make sense in the context of the archaeal genome. It is important to note that one limitation of Blast searches is similarity can only be compared to known sequences in the database (Koski & Golding 2001). As the archaeal genome the Moco\_RNA\_motif was found to occur in has few close neighbours to compare to, drawing comparisons using Blast should be regarded with caution.

The three riboswitch families for which good support of their occurrence was available, also showed limited distributions throughout the archaeal phyla. We posit a possible explanation for this distribution: Riboswitches that are found in archaea may be exceptions to the domain otherwise lacking these RNAs with those that are found representing acquisitions of riboswitch-gene pairs from the bacteria into specific archaeal taxa; both recently and much earlier in the evolution of the archaea. Below we consider this scenario in light of the evidence collected during this investigation.

### **The Distribution of the TPP Riboswitch in Archaea**

The TPP Riboswitch family was found in five of the archaeal phyla studied. These phyla include the Euryarchaeota, Parvarchaeota, Thaumarchaeota, Korarchaeota and Bathyarchaeota. This result is partially consistent with the known occurrences of this riboswitch family in the archaeal phyla Korarchaeota and Parvarchaeota along with the Euryarchaeota orders Thermoplasmatales and Methanomicrobiales as listed in Rfam 12.0 (Nawrocki *et al.* 2014). Within each of these phyla, occurrences of the TPP riboswitch were restricted to only a few genomes. The seemingly discrete groupings of this riboswitch within these taxa lend weight to the theory of acquisition by HGT. In the case of the Parvarchaeota and Thermoplasmata, we see the riboswitch being associated with the same gene. Preliminary phylogenetic analysis (Figure 2.3 above) suggests that this riboswitch-gene pair may have undergone a duplication event in the ancestor of the Thermoplasma clade. However, duplication was not detected within the Parvarchaeota. This, combined with only one species of Parvarchaeota showing evidence of the MFS\_1-TPP riboswitch pair may suggest that the pairing was acquired separately from or prior to the possible genome duplication event in the Thermoplasmata.

### **The Distribution of the Fluoride Riboswitch in Archaea**

Like the TPP riboswitch, the Fluoride riboswitch was also found to have a limited distribution throughout the archaeal phyla studied. Again, we suggest that this limited distribution may be explained by occurrences of the Fluoride riboswitch in archaea representing cases of separate acquisitions from other domains rather than widespread losses of the riboswitch from the archaeal taxa. Genomic context of the

Fluoride riboswitches found to occur in archaea lends further weight to this theory as discussed below.

### **The role of iterative searching in discovering riboswitches within the archaea**

One possible explanation for the limited distribution of riboswitches found in the archaea in this study is that riboswitches may be more well distributed throughout the archaea than we are currently able to detect with the covariance models used in this study.

Homology search programs such as INFERNAL and HMMER are notable for their ability to search iteratively. New sequences found in an initial search can be aligned back to the search model, improving the accuracy of subsequent passes over the data (Nawrocki & Eddy 2013b; Eddy 2011; Nawrocki 2009). We performed a trial of an iterative search for the TPP riboswitch in archaea, including sequences with a significant match to the initial model in the model used for a second search of the archaeal genomes. The second search failed to identify any significant subsequent occurrences of the TPP riboswitch in any of the genomes searched. However, as the distribution of all the riboswitches identified in the archaea was found to be limited, using iteration should still be considered in future work of this type.

Additionally, taxa in which known riboswitch families were not located may contain riboswitch families which are yet to be described, or variants of currently known riboswitch families which the covariance models cannot account for. In the former case, further experimental investigation of archaeal genomes both in laboratory settings and bioinformatically with programs such as RNAseq (Wang *et al.* 2009) may be required in order to identify new candidate riboswitches which may or may not be present in archaeal genomes. However, care must be taken in sampling to separate genuine results from transcriptional noise (Lindgreen *et al.* 2014). Use of iteration, as discussed above, has potential to improve results from any new investigation carried out.

## **Protein-Coding Genes Found Downstream of Riboswitches in the Archaea**

We demonstrated in this study that the genes found downstream of a particular riboswitch are not always consistent between all taxa that the riboswitch is found in.

### **Genes found downstream of the TPP riboswitch in Archaea**

A number of genes were found to be associated with the TPP riboswitch in archaea. Mostly these genes were identified as thiamine biosynthesis and transport proteins which we would expect to see associated with the riboswitch given that it is known to bind the molecule thiamine pyrophosphate (Serganov *et al.* 2006). However, one surprising result is of note in this case. In both the Thermoplasmata and Parvarchaeota, the TPP riboswitch was found to be associated with the transporter protein MFS\_1 or Sugar\_tr. This result has been previously observed when studying the thiamine biosynthesis pathways in archaea (Rodionov *et al.* 2002). Literature on the function of the MFS\_1/Sugar\_tr proteins suggest that this class of transporter protein may be somewhat generalized (Pao *et al.* 1998; Saier *et al.* 1999). We therefore suggest (as also suggested by Rodionov *et al.* (2002)) the possibility that this protein may act as a transporter for thiamine pyrophosphate or its components in the cases where it is found downstream of the TPP riboswitch in these archaeal genomes. As discussed above, two copies of the TPP riboswitch-MFS\_1 gene pairing were found to occur in each genome of many of the Thermoplasmata. Differences found between the two copies of the gene in each genome support the theory of a duplication event having occurred; suggesting the possibility of slight differences in function between the two copies. However, more extensive investigation should be carried out in this case before any further conclusions about the relationship or function of these genes can be drawn. Experimental verification of the function of these genes as thiamine transporters in the Thermoplasmata would strengthen the conclusions drawn here.

### **Genes found downstream of the Fluoride riboswitch in Archaea**

Like in the case of the TPP riboswitch, genes found downstream of the Fluoride riboswitch varied with the taxa the riboswitch was found in. The riboswitch was found to be mostly associated with sequences coding for proteins involved in ion

transport. CRCB, a transporter of the fluoride ion has been previously described to be associated with the Fluoride riboswitch in archaea (Baker *et al.* 2012). Baker *et al.* (2012) also noted the presence of other ion transporter proteins downstream of archaeal examples of the Fluoride riboswitch, lending weight to the integrity of the findings in our study. Many of the downstream ion transported identified are not known to transport fluoride ions. Fluoride is a small molecule and the fluoride riboswitch is smaller with regions that are not as well conserved as the sequence features noted in other more classic examples of riboswitches. These facts raise the possibility that this riboswitch may be more generalist in its function than previously thought; perhaps responding not only to fluoride ions but to other negatively charged ions participating in ion exchange processes within the cell. Flexibility in the ability of a riboswitch to sense metabolites has been noted in some other cases (Li *et al.* 2016).

We also identified a case where the function of the gene found downstream of the Fluoride riboswitch was unknown. In the Thermococci class of the Euryarchaeota, the majority of ORFs located directly downstream of the Fluoride riboswitch were close matches to uncharacterised archaeal proteins. While these proteins show similarity to the Pfam domain Prenyltrans, their function is as yet unverified. We hypothesise that these proteins may be involved in ion transport based on their proximity to the Fluoride riboswitch and the prevalence of ion transport proteins located downstream of this riboswitch in other archaeal species we examined.

Based on the findings of this study we recommend further experimental investigation into the function of this riboswitch and the genes it is found to be associated with in archaea be carried out to test this hypothesis and potentially discover new information about the functional capacity of this riboswitch family in general.

## **Horizontal Gene Transfer of Riboswitch-Gene pairs from Bacteria into the Archaea**

Three cases of potential HGT events were identified in this study. Subsequent analysis of these cases revealed strong evidence for at least one evolutionarily recent acquisition of a riboswitch-gene pairing by an archaeal genus from a bacterial

source. Two other potential HGT events were less well supported by the subsequent evidence collected. Evidence for HGT events occurring between the bacteria and the archaea have been previously noted (Nelson *et al.* 1999; Garcia-Vallvé *et al.* 2000). In particular, Hoepfner *et al.* (2012) showed evidence that the distribution of the TPP riboswitch throughout all three domains of life may be at least in part explained by HGT events between the three domains.

### **Integrity of phylogenetic trees constructed**

Three phylogenetic trees were constructed with varying success in attempts to determine the likelihood of horizontal gene transfer of riboswitch-gene pairs from bacteria into the archaea. While good support for our hypothesised transfer of the TPP riboswitch-Thi4 pair from bacteria into the archaea was found, support for transfer events of the FMN riboswitch-DHBP\_synthase and Fluoride riboswitch-Na\_H\_exchanger pairings from bacteria into the archaea was less convincing. Branch support values for trees constructed in the latter cases (see Figures 2.7 and 2.9 above) were poorer than those in the tree of Thi4 sequences (Figure 2.4 above). Integrity of the trees generated in this case may be verified by additional investigation of the relationships generated using both Bayesian methods and Maximum likelihood methods (Smith & Naylor 1987). As both methods have potential to generate different trees from the same data (Douady *et al.* 2003), care must be taken to verify if relationships found are realistic and that the sequence data being analysed is .

While good branch support values were obtained in most cases for both the Thi4 tree and the DHBP\_synthase tree, branch support values in the Na\_H\_exchanger tree were poor in many cases. This poor support is also reflected by the percentage of conserved sites after G-blocks curation of the Na\_H\_exchanger alignment being just 17% conserved positions rather than the higher values of 59% and 31% conservation for the Thi4 and DHBP\_synthase trees respectively. As the support for this phylogeny was limited, we are unable to make a conclusive determination as to the likelihood that the presence of the Fluoride riboswitch-Na\_H\_exchanger pairing represents a transfer of genetic material from the bacteria into the Methanobacteria. Further analysis, including with other tree construction methods (Douady *et al.* 2003) and the use of tests to verify the best-fit model of protein evolution (Abascal *et al.* 2006), is recommended for improving future attempts at constructing trees from this data to demonstrate HGT.

## Horizontal Gene Transfer: The Bigger Picture

Horizontal gene transfer events appear to be a possible explanation for the distribution of riboswitches throughout the archaea in light of the evidence we have collected. We have demonstrated not only that riboswitches found within the archaea are restricted to few taxa, but also that the genes associated with each riboswitch are also dependent upon the taxa in which that riboswitch was found. We have shown evidence of potential support for HGT events of bacterial riboswitch-gene pairs into archaea. Namely, our phylogenetic analysis of the Thi4-TPP riboswitch gene pair in the genus *Methanocorpusculum* suggests a bacterial origin of this riboswitch-gene pair. We therefore suggest that the current distribution of riboswitches in the archaea may be explained by a series of separate HGT events over the course of the evolution of the archaea, although we recommend future investigations into the role of HGT in explaining the distribution of other archaeal riboswitch-gene pair to further refine this conclusion.

## Concluding Remarks

This study has examined the distribution and function of the riboswitches and explored possibilities of the evolutionary history of this class of RNA within the archaeal domain. This work has provided an important update to knowledge previously obtained about riboswitches in the archaea in light of genomic data from many new archaeal taxa becoming available (Barrick & Breaker 2007; Baker *et al.* 2012; Rinke *et al.* 2013, Spang *et al.* 2015, Lazar *et al.* 2015). We find that while some of the more recently identified archaeal taxa show evidence of known riboswitch families, overall the presences of riboswitch elements in the archaea appears to be limited.

We reveal evidence that while some of the riboswitches that are found in archaea are associated with genes that are expected to be regulated by the riboswitch in question, there are other cases in which genes associated with a given riboswitch suggest that the function of that riboswitch may be more general than previously recognised.

We identify the opportunity for future work to be carried out investigating the function and gene associations of the fluoride riboswitch in the archaea. We also

suggest that more extensive study into the role of HGT in the distribution of the riboswitches in archaea would be beneficial to our understanding of the evolutionary history of small non-coding RNAs in the archaea.

## References

- Abascal, F., Zardoya, R., & Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, 21(9), 2104-2105.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403-410.
- Anisimova, M., & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic biology*, 55(4), 539-552.
- Barrick, J. E., & Breaker, R. R. (2007). The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome biology*, 8(11), 1.
- Baker, J. L., Sudarsan, N., Weinberg, Z., Roth, A., Stockbridge, R. B., & Breaker, R. R. (2012). Widespread genetic switches and toxicity resistance proteins for fluoride. *Science*, 335(6065), 233-235.
- Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic acids research*, 41(D 1), D36-D42.
- Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution*, 17(4), 540-552.
- Chevenet, F., Brun, C., Bañuls, A. L., Jacq, B., & Christen, R. (2006). TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC bioinformatics*, 7(1), 439.
- Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., & Claverie, J. M. (2008). Phylogeny. fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids research*, 36(suppl 2), W465-W469.
- Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F., & Douzery, E. J. (2003). Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution*, 20(2), 248-254.
- Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput Biol*, 7(10), e1002195.

- Eddy, S. R., & Wheeler, T. J. (2013). HMMER User's Guide.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792-1797.
- Federhen, S. (2012). The NCBI taxonomy database. *Nucleic acids research*, *40*(D1), D136-D143.
- Finn, R. D., Bateman, A., Clements, J., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2013). Pfam: the protein families database. *Nucleic acids research*, gkt1223.
- Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F., Bateman, A., & Eddy, S. R. (2015a). HMMER web server: 2015 update. *Nucleic acids research*, gkv397.
- Finn, R. D., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., & Salazar, G. A. (2015b). The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*, gkv1344.
- Garcia-Vallvé, S., Romeu, A., & Palau, J. (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Research*, *10*(11), 1719-1725.
- Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, *52*(5), 696-704.
- Hoepfner, M. P., Gardner, P. P., & Poole, A. M. (2012). Comparative analysis of RNA families reveals distinct repertoires for each domain of life. *PLoS Comput Biol*, *8*(11), e1002752.
- Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, *11*(1), 1.
- Johnson, M., Zaretskaya, I., Raytselis, Y., Merezuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic acids research*, *36*(suppl 2), W5-W9.
- Koski, L. B., & Golding, G. B. (2001). The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution*, *52*(6), 540-542.
- Lazar, C. S., Baker, B. J., Seitz, K., Hyde, A. S., Dick, G. J., Hinrichs, K. U., & Teske, A. P. (2016). Genomic evidence for distinct carbon substrate preferences and ecological niches of Bathyarchaeota in estuarine sediments. *Environmental microbiology*.
- Li, S., Hwang, X. Y., Stav, S., & Breaker, R. R. (2016). The yjdB riboswitch candidate regulates gene expression by binding diverse azaaromatic compounds. *RNA*, *22*(4), 530-541.

Lindgreen, S., Umu, S. U., Lai, A. S. W., Eldai, H., Liu, W., McGimpsey, S., McGimpsey, S., Wheeler, N.E., Biggs, P.J., Thomson, N.R., Barquist, L., Poole, A. M., & Gardner P. P. (2014). Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Comput Biol*, 10(10), e1003907.

Pao, S. S., Paulsen, I. T., & Saier, M. H. (1998). Major facilitator superfamily. *Microbiology and molecular biology reviews*, 62(1), 1-34.

Nawrocki, E. (2009). Structural RNA homology search and alignment using covariance models.

Nawrocki, E. P., & Eddy, S. R. (2013a). Computational identification of functional RNA homologs in metagenomic data. *RNA biology*, 10(7), 1170-1179.

Nawrocki, E. P., & Eddy, S. R. (2013b). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22), 2933-2935.

Nawrocki, E., & Eddy, S. (2014). INFERNAL User's Guide.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., & Finn, R. D. (2014). Rfam 12.0: updates to the RNA families database. *Nucleic acids research*, gku1063.

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A. & McDonald, L. (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, 399(6734), 323-329.

Regulski, E. E., Moy, R. H., Weinberg, Z., Barrick, J. E., Yao, Z., Ruzzo, W. L., & Breaker, R. R. (2008). A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism. *Molecular microbiology*, 68(4), 918-932.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., & Dodsworth, J. A. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459), 431-437.

Rodionov, D. A., Vitreschak, A. G., Mironov, A. A., & Gelfand, M. S. (2002). Comparative genomics of thiamin biosynthesis in procaryotes New genes and regulatory mechanisms. *Journal of Biological chemistry*, 277(50), 48949-48959.

Saier, M. H., Beatty, J. T., Goffeau, A., Harley, K. T., Heijne, W. H., Huang, S. C., Jack, D.L., Jahn, P.S., Lew, K., Liu, J., & Pao, S. S. (1999). The major facilitator superfamily. *J Mol Microbiol Biotechnol*, 1(2), 257-279.

Serganov, A., Polonskaia, A., Phan, A. T., Breaker, R. R., & Patel, D. J. (2006). Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature*, 441(7097), 1167-1171.

Smith, R. L., & Naylor, J. C. (1987). A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Applied Statistics*, 358-369.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521(7551), 173-179.

UniProt Consortium. (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1), D204-D212.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1), 57-63.

## Chapter Three - The Distribution of Known snoRNA families in the Archaea

### *Introduction*

The relationship between the archaeal and eukaryotic domains of life has long been an intriguing point of study for those interested in the evolutionary history of life. A common feature between both domains are the snoRNAs. Establishing the evolutionary history of the snoRNAs in the archaea, especially in light of more recently available data from many emerging archaeal phyla (Rinke *et al.* 2013; Spang *et al.* 2015; Lazar *et al.* 2015), is important to a more complete understanding of the relationship between the archaea and the eukaryotes.

In this study we sought to determine the distribution of all known snoRNA families currently documented in the Rfam 12.0 database (Nawrocki *et al.* 2014) across the currently known archaeal genomes available from Genbank (Benson *et al.* 2013). This investigation provides new information about how commonplace the known snoRNAs are in archaea and how their distribution may relate to their evolutionary history within both the archaea and the eukaryotes.

To provide additional information in cases where the short sequence length of the snoRNAs themselves may have impeded efforts to detect them, we also investigated the distribution of known snoRNA associated proteins throughout the same archaeal phyla. We find that the distribution snoRNPs within the archaea may be a better reflection of the true distribution of snoRNAs within the archaeal domain. We note that while snoRNPs were present in all archaeal phyla, cases where particular snoRNPs were not found to occur in particular phyla were apparent

## **Methods**

### **Investigating the Distribution of snoRNAs and snoRNPs in the Archaea**

#### **Dataset**

We sought to determine the distribution of snoRNAs throughout the archaeal domain. To represent the currently diversity of the archaeal phyla, nucleotide sequences from 463 species across 13 archaeal phyla or candidate phyla were downloaded from Genbank (Benson *et al.* 2013), for a list of all species represented see Appendix 1. Genomic data was chosen from all archaeal species listed in NCBI's Taxonomy browser (Federhen 2012) that had whole genome data available and were not classed as "environmental samples".

A second dataset that was comprised of covariance models for 729 known snoRNA families was created by downloading covariance models from the RNA families database Rfam 12.0 (Nawrocki *et al.* 2014). All Rfam families that were marked with the entry type "CD-box" or as "HACA-box" snoRNAs were included in this dataset.

To examine the presence of snoRNA-associated proteins across the archaeal phyla, a third dataset comprised of profile hidden markov models (hmms) was downloaded from the protein families database Pfam (Finn *et al.* 2013). The following known archaeal snoRNPs were included in this dataset: Fibrillarin, Gar1, Nop, Nop10p, L7Ae, and TruB\_N (Gardner 2010).

#### **Analysis Methods**

The use of homology searches, which are powerful tools for determining the presence of known sequence features within sequence data (Nawrocki and Eddy 2013a), was employed in order to determine the presence of both snoRNA families and sno-associated proteins across the archaeal phyla in our dataset. In this study we used a combination of profile hidden markov models and covariance models for this purpose.

To determine presence of snoRNAs across archaeal phyla, the INFERNAL package was used to perform a cmsearch of all genomes in the dataset against all snoRNA families (Nawrocki and Eddy 2013b). Default settings for bit-score and e-

value were used for reporting significance in the cmsearch (Nawrocki and Eddy 2013b). For all significant hits found, sequence data corresponding to the hit was extracted from the genome using INFERNAL's cmfetch function (Nawrocki and Eddy 2013b). These sequences were then analysed for evidence of C and D box or H and ACA box features based on the type of snoRNA family the hit was for. Any sequences that did not show evidence of either at least one C-box and at least one D-box or at least one H-box and at least one ACA-box were discarded from the results as a false positive. The presence of each snoRNA family in each species was then reported on based on the remaining sequences.

To analyse the presence of corresponding archaeal sno-associated proteins, Prodigal 2.6.2 with default settings was first used to create protein translations from the nucleotide sequences of all genomes in the dataset (Hyatt *et al.* 2010). The HMMER package was then used to search the translated genome sequences for all genomes against all hmms for sno-associated proteins in the dataset using hmmsearch (Eddy 2011). Default values for bit-score and e-value were used as the threshold for significant hits (Eddy & Wheeler 2013). Presence or absence of each sno-associated protein in each phyla was then reported based on positive hits found in each species.

## **Results**

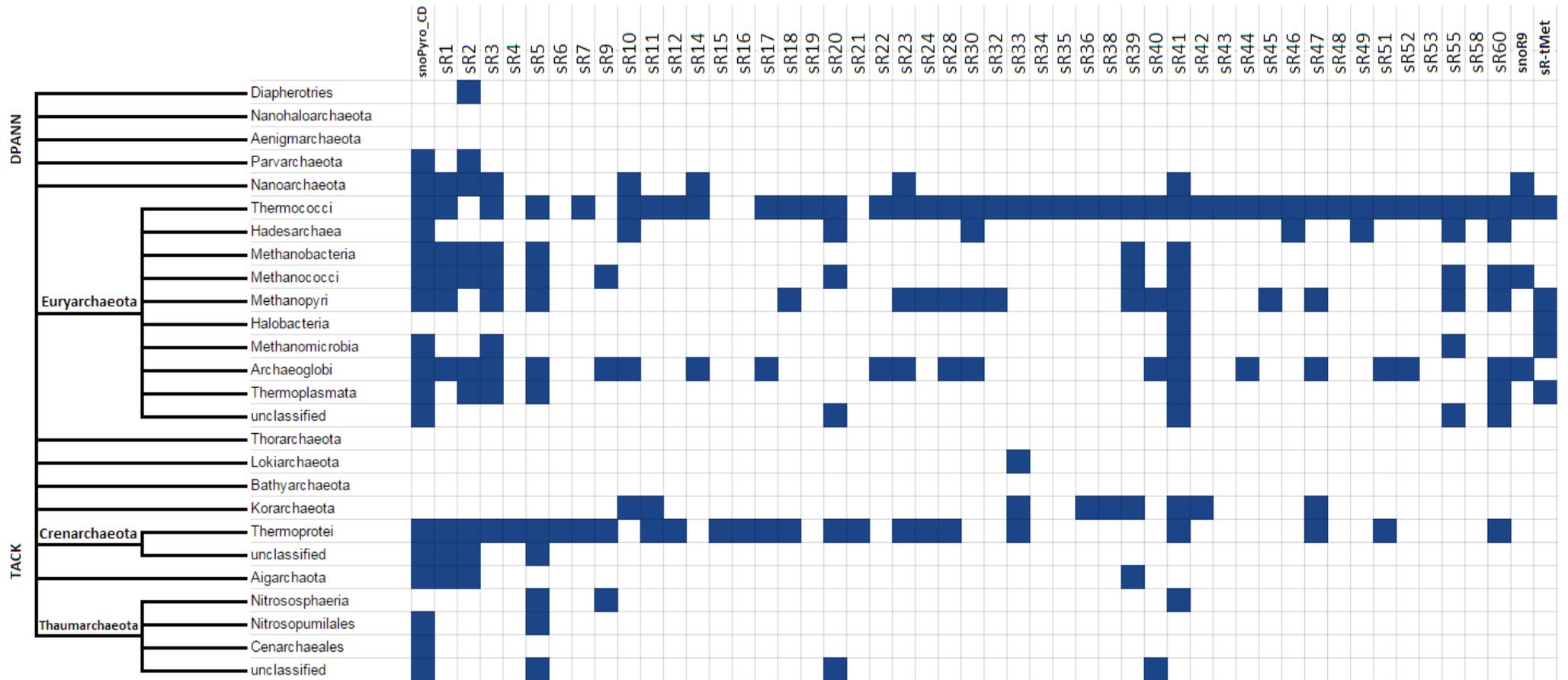
To determine the overall distribution of known snoRNAs throughout the archaea, covariance models of each known snoRNA family were used to search genomes representing 26 archaeal classes within 13 archaeal phyla. The overall distribution of these families throughout the archaea was limited. No known archaeal C/D box snoRNA families were found in the Nanohaloarchaeota phylum or the Aenigmarchaeota phylum. No known archaeal snoRNA families were found in the more recently described phyla (Bathyarchaeota, Lokiarchaeota, Thorarchaeota). However, some evidence of snoRNA families that are known from eukaryotes was found in species from these phyla. Evidence for snoRNA families known from eukaryotes was also found in the Thaumarchaeota. The presence of each snoRNA family in each taxa of archaea studied is summarised in Figure 3.1 below.

## Distribution of C/D Box snoRNAs in the Archaea

We used covariance models of 463 C/D box snoRNA families to search across genomes representing 13 archaeal phyla for evidence of these families. Fifty families of snoRNA that were previously known from archaeal genomes were located within the archaeal genomes studied. The most commonly occurring of these families include snoPyro\_CD, sR1, sR2, sR3, sR5, and sR41. Archaeal taxa with the most C/D-box snoRNA families located include the Thermococci class from the Euryarchaeota (43 families), the Thermoprotei class from the Crenarchaeota (25 families), and the Archaeoglobi class of the phylum Euryarchaeota (21 families). The single genome representing the Methanopyri class of the phylum Euryarchaeota was found to show evidence of 13 C/D box snoRNA families previously known from archaeal species.

In addition, potential matches to C/D box snoRNA families known previously only from eukaryotes were identified in archaeal species. These included an occurrence of SNORD78 in the Lokiarchaeota, snosnR61 and cen40 in the Thorarchaeota, SNORD15 in the Bathyarchaeota, and snoR113, SNORD59, and snoMBII-202 in Thaumarchaeota genomes (Figure 3.2).

In many taxonomic groups considered in this study we found that the snoRNAs detected only occurred in a minority of species representing that taxa. Proportions of species for each taxa showing evidence of any C/D box snoRNA are summarised in Table 3.1 and in Figure 3.2 below. Notably, all species of Archaeoglobi showed evidence for the C/D box snoRNAs snoPyro\_CD and sR3 and all but one species from the Nitrososporales class of Thaumarchaeota showed evidence of the snoRNA family snoPyro\_CD. All members of the following Euryarchaeota classes were found to show evidence of C/D box snoRNAs although not all snoRNA families detected were found in every species: Thermococci, Methanobacteria, Methanococci, Methanopyri, Methanomicrobia, and Archaeoglobi (Table 3.2).



**Figure 3.1:** Presence of C/D snoRNAs families previously known from archaeal species across archaeal phyla and major classes. Many known snoRNAs families were only detected in the class Thermococci of the Euryarchaeota or the Thermoprotei in the Crenarchaeota. Of the 50 known archaeal snoRNA families from Rfam, only snoPyro\_CD, sR1, sR2, sR5 and sR41 appear to be somewhat widely distributed across many archaeal phyla. Cladogram of archaeal phyla/classes based on data from NCBI taxonomy, Spang *et al.* (2015), Lazar *et al.* (2015), and Rinke *et al.* (2013)

**Table 3.1.** Proportion of species in each archaeal taxa studied that showed evidence of known snoRNA families from Rfam. Counts for snoRNA families detected that were previously known from archaea are displayed in column 3, while those families known from eukaryotic species are shown in column 4. The number of species that represent each taxa studied are shown in column 2. A percentage of these species which show evidence of any snoRNA family searched for is calculated in column 5 from the data in columns 2-4.

| Archaeal Taxon             | Number of species | Number of species with archaeal C/D box snoRNAs detected | Number of species with eukaryotic C/D box snoRNAs detected | Percentage of species with evidence of any C/D box snoRNA detected |
|----------------------------|-------------------|----------------------------------------------------------|------------------------------------------------------------|--------------------------------------------------------------------|
| Diapherotries              | 3                 | 1                                                        | 0                                                          | 33%                                                                |
| Nanohaloarchaeota          | 4                 | 0                                                        | 0                                                          | 0%                                                                 |
| Aenigmarchaeota            | 3                 | 0                                                        | 0                                                          | 0%                                                                 |
| Parvarchaeota              | 3                 | 1                                                        | 0                                                          | 33%                                                                |
| Nanoarchaeota              | 10                | 2                                                        | 1                                                          | 20%                                                                |
| Thermococci                | 27                | 27                                                       | 0                                                          | 100%                                                               |
| Hadesarchaea               | 4                 | 2                                                        | 1                                                          | 50%                                                                |
| Methanobacteria            | 20                | 20                                                       | 0                                                          | 100%                                                               |
| Methanococci               | 15                | 15                                                       | 2                                                          | 100%                                                               |
| Methanopyri                | 1                 | 1                                                        | 0                                                          | 100%                                                               |
| Halobacteria               | 124               | 7                                                        | 0                                                          | 6%                                                                 |
| Methanomicrobia            | 61                | 61                                                       | 0                                                          | 100%                                                               |
| Archaeoglobi               | 7                 | 7                                                        | 0                                                          | 100%                                                               |
| Thermoplasmata             | 23                | 8                                                        | 0                                                          | 35%                                                                |
| Unclassified Euryarchaeota | 10                | 5                                                        | 0                                                          | 57%                                                                |
| Thorarchaeota              | 3                 | 0                                                        | 2                                                          | 67%                                                                |
| Lokiarchaeota              | 1                 | 0                                                        | 1                                                          | 100%                                                               |
| Bathyarchaeota             | 9                 | 0                                                        | 1                                                          | 11%                                                                |
| Korarchaeota               | 1                 | 1                                                        | 0                                                          | 100%                                                               |
| Thermoprotei               | 62                | 46                                                       | 0                                                          | 74%                                                                |
| Unclassified Crenarchaeota | 14                | 11                                                       | 0                                                          | 79%                                                                |
| Aigarchaeota               | 19                | 10                                                       | 0                                                          | 53%                                                                |
| Nitrososphaeria            | 3                 | 1                                                        | 1                                                          | 67%                                                                |
| Nitrosopumilales           | 12                | 11                                                       | 5                                                          | 92%                                                                |

|                             |    |    |   |      |
|-----------------------------|----|----|---|------|
| Cenarchaeales               | 1  | 1  | 1 | 100% |
| Unclassified Thaumarchaeota | 23 | 11 | 3 | 52%  |

### Distribution of H/ACA Box snoRNAs in the Archaea

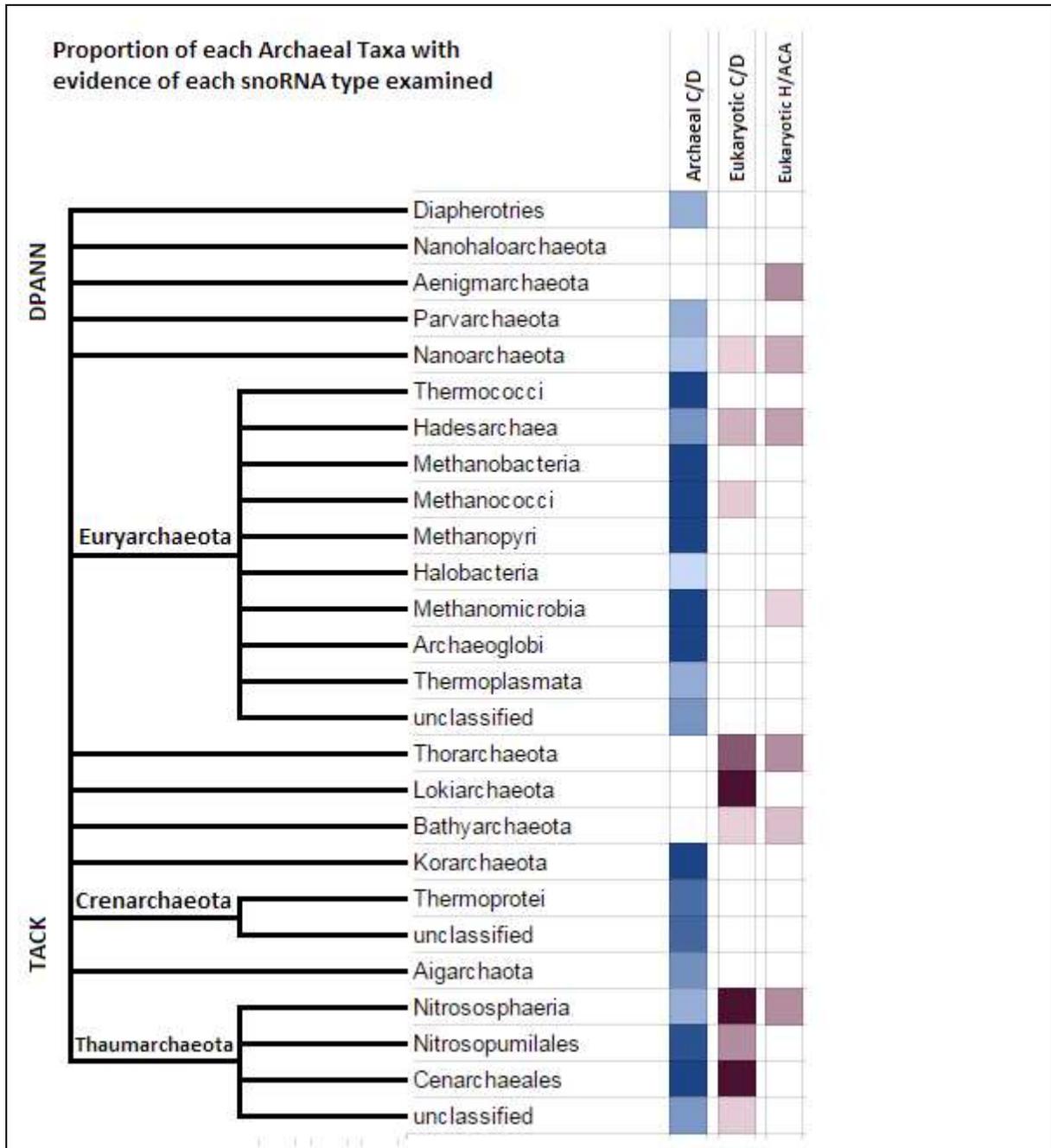
Rfam 12.0 does not currently list any known families of H/ACA box snoRNAs as being found in archaeal species. In our investigation we searched the genomes of archaeal species for evidence of the 266 H/ACA box snoRNAs known from Rfam with limited success. We detected evidence for eight families of H/ACA box snoRNAs within the archaeal species studied. These included two families (snR83 and TB11Cs5H1) in the Nitrososphaeria, and one family in each of the following archaeal taxa: Aenigmarchaeota, Nanoarchaeota, Hadesarchaea, Methanomicrobia, Thorarchaeota, and Bathyarchaeota. In each of these cases only a single match was found in a single genome with the exception of the Nanoarchaeota where one H/ACA box snoRNA family was found in two of the genomes studied.

**Table 3.2.** Summary of H/ACA box snoRNA occurrences detected in the archaea. Only taxa where possible occurrences of H/ACA box snoRNAs were found are shown. The snoRNA family of each detected occurrence is noted along with the primary eukaryotic taxon that snoRNA is associated with.

| Archaeal Taxon  | Number of species in which snoRNA detected | H/ACA snoRNA family detected | Primary eukaryotic taxon RNA family known from |
|-----------------|--------------------------------------------|------------------------------|------------------------------------------------|
| Aenigmarchaeota | 1                                          | SNORA32                      | Mammalia                                       |
| Nanoarchaeota   | 2                                          | snoR138                      | Streptophyta                                   |
| Hadesarchaea    | 1                                          | SNORA14                      | Mammalia                                       |
| Methanomicrobia | 1                                          | snR42                        | Saccharomycetaceae                             |
| Thorarchaeota   | 1                                          | SNORA61                      | Mammalia                                       |
| Bathyarchaeota  | 1                                          | snoR104                      | Streptophyta                                   |
| Nitrososphaeria | 1                                          | snR83<br>TB11Cs5H1           | Saccharomycetaceae<br>Trypanosoma              |

After examining the combined distributions of both C/D box and H/ACA box snoRNAs screened for in this study, only the Nanohaloarchaeota were not found to show evidence of any known snoRNA family, be they snoRNAs known from archaea or those known from eukaryotes. However, many other taxa were found to have only

very limited evidence for known snoRNA families. All members of the DPANN superphylum along with the Halobacteria, and Bathyarchaeota show poor distributions of known snoRNA families in terms of proportion of species with evidence of known snoRNA families.

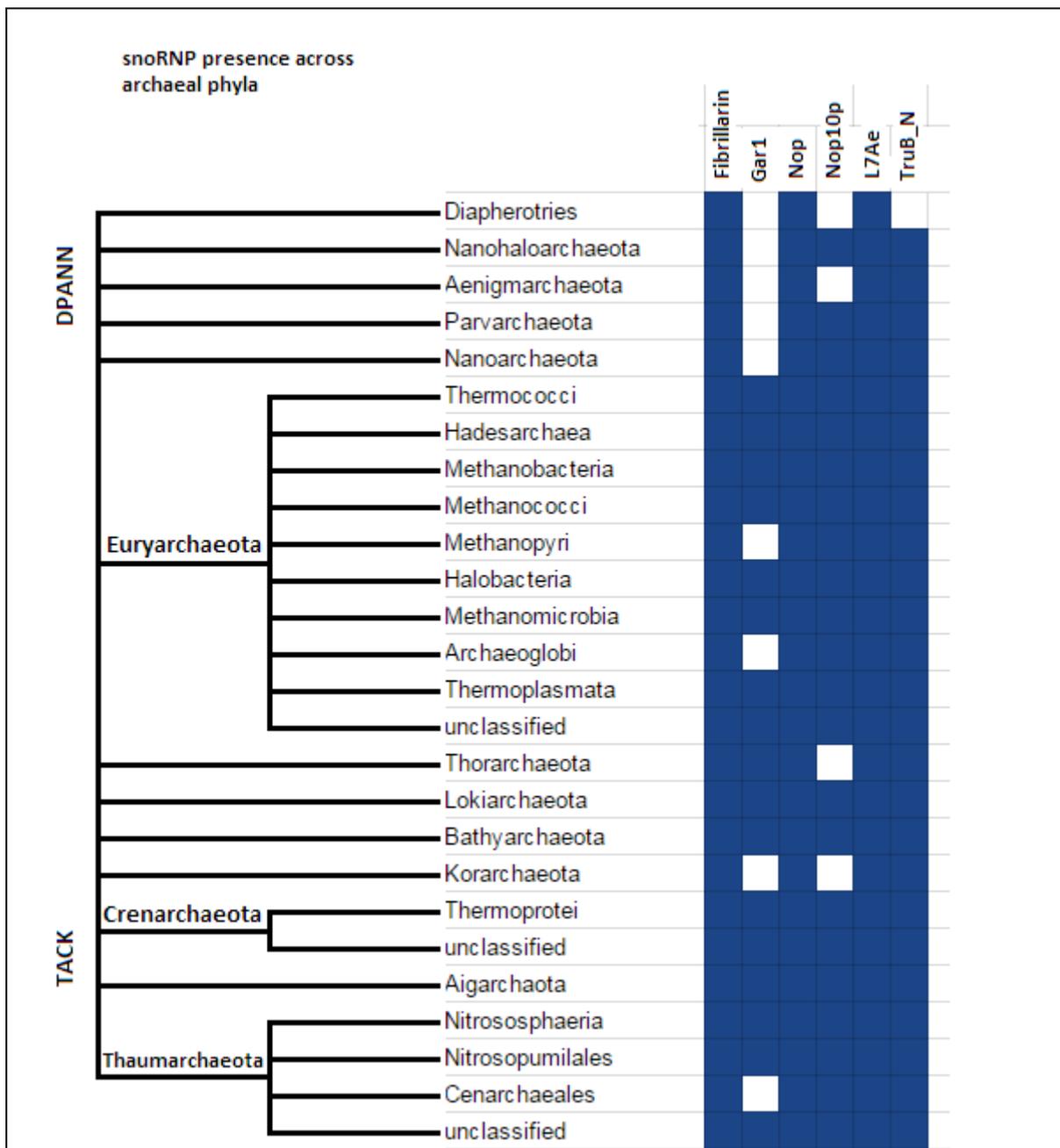


**Figure 3.2.** Proportion of each archaeal taxa studied showing evidence of each broad category of snoRNAs examined. Colour scale represents fewest representatives of a taxon showing evidence of that snoRNA type (lightest), to all species in a taxon showing evidence of that snoRNA type (darkest). Blue represents snoRNA families previously known from archaeal species, red represents snoRNA families known from eukaryotes. Cladogram of archaeal phyla/classes based on data from NCBI taxonomy, Spang *et al.* (2015), Rinke *et al.* (2013), and Lazar *et al.* (2015).

## ***Distribution of snoRNPs in the Archaea***

To investigate whether the patterns found in the survey of snoRNAs in the archaea were robust, we also examined the archaeal taxa in question for presence of the snoRNA-associated proteins known to interact with snoRNAs within the archaea.

Our investigation revealed that while most snoRNPs are found in the majority of archaeal taxa, coverage of both the proteins Gar1 and Nop10p was incomplete. Of the snoRNPs searched for, Fibrillarin, Nop and L7Ae were detected in all archaeal groups surveyed and TruB\_N was only not detected in the Diapherotrites. Gar1 was not detected in any genome from the DPANN superphylum nor the groups Methanopyri, Korarchaeota, or Cenarchaeales (where each of the latter three groups are each only represented by one species). Gar1 was also not detected in the Archaeoglobi (represented by genomes of seven species). The snoRNP Nop10p was also not detected in the Diapherotrites, Aenigmarchaeota, Thorarchaeota or the Korarchaeota. For a summary of the distributions of each of the snoRNPs examined, see Figure 3.3.



**Figure 3.3:** Presence of known snoRNPs across archaeal phyla and major classes. The Gar1 protein was not detected in the DPANN superphylum or the single known genome from the Korarchaeal phylum. While the proteins, Fibrillarlin, Nop and L7Ae we found in all groups surveyed, TruB\_N and Nop10p were also not detected in at least one group surveyed. Cladogram of archaeal phyla/classes based on data from NCBI taxonomy, Spang *et al.* (2015), Lazar *et al.* (2015), and Rinke *et al.* (2013).

## ***Discussion***

### **C/D box snoRNAs and their associated snoRNPs in the Archaea**

We found that, overall, C/D box snoRNAs were distributed through almost all of the archaeal taxa examined. However, in many cases very few of the known archaeal snoRNA families were detected. The greatest representation of archaeal C/D box snoRNA families occurred in the Thermoprotei class of the Crenarchaeota and the Thermococci class of the Euryarchaeota. This result reflects the fact that the majority of experimental investigations into archaeal snoRNA families has been carried out on members of these taxa (Gaspin *et al.* 2000, Omer *et al.* 2000, Dennis & Omer 2005). In fact, the covariance models currently stored by Rfam for all archaeal snoRNA species are built from snoRNA sequences taken only from these two taxa.

In line with a previous survey of snoRNA presence across the archaea (Gardner *et al.* 2010), we found that there was a limited presence of snoRNA families within each of the taxa examined. Only nine of the 26 taxa examined showed evidence of snoRNA presence in each species which made up that group. We hypothesize that since C/D box snoRNAs are of short sequence length and in some cases do not have strong sequence conservation between species outside of the C and D box motifs (Omer *et al.* 2000), our ability to detect known snoRNAs with the current models available may be limited. This may be more likely in species that are of greater evolutionary distance from the species the models were built from. We suggest that iterative techniques applied to the homology searches described here may mean future searches will be more successful. Using an iterative approach to extend the quality of the initial model by aligning detected matching sequences back to the model after each search gives a greater capacity for the model to more accurately detect sequences that are similar to newly found matches from the initial search (Nawrocki and Eddy 2013a).

We also suggest that the limited number of species in which snoRNAs were detected may be explained by the presence of as-yet unknown snoRNA families in these species. An extensive search targeting the identification of new snoRNA families across the archaea has not been undertaken since Gaspin *et al.* (2000) and Omer *et al.* (2000) although smaller scale searches have occurred since then (Tang *et al.* 2002; Klein *et al.* 2002; Dennis & Omer 2005). We therefore identify a need for

future studies of this type to be carried out in order to address the lack of current knowledge in this area. Particular focus should be placed on identification of potential snoRNA families within the taxa (those outside of the Euryarchaeota and Crenarchaeota) which have not yet been well examined. Combined with better quality models of existing snoRNA families, models that can accurately identify new or proposed snoRNA families may be used in future bioinformatic-based investigations of snoRNA distributions in the archaea to help fill gaps that this study identified.

When examining the distribution of snoRNPs, we discovered good evidence for widespread distribution of the known archaeal C/D box associated snoRNP families Fibrillarin, Nop, and L7Ae (Rashid *et al.* 2003 , Rozhdestvensky *et al.* 2003). This reinforces our finding that C/D box snoRNAs are present in almost all archaeal taxa and suggests that other, as yet unknown, C/D box snoRNAs families may still be present in cases where we did not detect them using covariance models.

## **H/ACA box snoRNAs and their associated snoRNPs in the Archaea**

During our investigation we discovered limited evidence for H/ACA box snoRNAs within the archaea. This is most likely due to models of H/ACA box snoRNAs known from archaea not currently being present in or classified as snoRNAs in Rfam 12.0. The H/ACA box snoRNAs that were detected were therefore those known from eukaryotic species. The detection of presence of these eukaryotic snoRNAs is discussed in more detail in the section below.

Support for archaeal H/ACA box snoRNAs being present is therefore represented in this study by the presence of H/ACA-associated snoRNPs throughout all archaeal taxa. The snoRNP Pfam domains TruB\_N (Cbf5), Gar1, Nop10p and L7Ae are all known to be associated with H/ACA box snoRNAs in the archaea (Henras *et al.* 2004; Watanabe & Gray 2000; Rozhdestvensky *et al.* 2003). Of these snoRNPs, we found that only L7Ae was distributed throughout all major archaeal taxa studied. However, this protein is also known to be associated with C/D box snoRNAs (Rozhdestvensky *et al.* 2003), a property that means the likelihood that this protein is not associating with H/ACA box snoRNAs in all cases where it occurs cannot be ruled out. Of the other three proteins, both TruB\_N and Nop10p were detected in almost all taxa studied. Cases where they were not detected include taxa where fewer than four genomes represent each each phylum. This suggests the

possibility that detection of the snoRNPs in these phyla may be hampered by limited genomic data available. However, two taxa where one or both of these proteins were not detected fall within the DPANN superphylum in which no evidence of the last H/ACA box snoRNP studied (Gar1) was detected. The Gar1 protein is known to be essential to the process of pseudouridylation guided by H/ACA box snoRNAs (Girard *et al.* 1992, Rashid *et al.* 2006). The absence of Gar1 detected throughout all genomes of this superphylum raises questions over how prevalent H/ACA box snoRNAs may be in these species. We identify an opportunity for further work, including experimentation considering the pseudouridylation process in these species, to be undertaken to address the unknowns regarding this result.

## **The presence of eukaryotic snoRNAs within archaeal genomes**

Our examination of the distribution of known snoRNAs families within archaeal genomes revealed evidence to suggest the presence of several snoRNAs families that are only known from the eukaryotes. As snoRNA families appear to be domain specific (Hoeppner *et al.* 2012, Hoeppner & Poole 2012), it seems unlikely that these occurrences are genuine examples of these eukaryotic snoRNAs. However, as evidence of both C and D box motifs or H and ACA box motifs were identified in all of these occurrences, further investigation should be carried out to confirm whether these examples are a result of contamination of the genomes in which they occur, false positives, or cases of genuine matches to snoRNAs. A comparison of the snoRNA sequences found to complementing known target sites on the RNAs these snoRNAs are known to interact with would be a good step towards determining whether or not these occurrences are false positives. Subsequently, analysis of the genomic context of regions of the genome where the match was discovered could be used to rule out genomic contamination. Genuine matches to snoRNAs in these archaea may indicate either sequence convergence of previously unknown archaeal C/D box or H/ACA snoRNA families with analogous eukaryotic C/D or H/ACA snoRNA families, or snoRNAs that trace back to an ancestor common to both archaeal and eukaryotic species (Hoeppner & Poole, 2012). In either of these cases, a confirmed archaeal occurrence of either a C/D box snoRNA or H/ACA box snoRNA family based on a model created from a eukaryotic snoRNA example would be a clear indication that further experimental investigation is needed into the functions of these snoRNAs within the archaea.

## Concluding Remarks

The snoRNAs are found in the majority of archaeal taxa examined in this study. However, distribution of the snoRNAs at the species level was found to be less complete. This finding is similar to previous work looking at the distribution of snoRNAs across both the archaea and eukaryotes, where limited distribution of snoRNAs in the Euryarchaeota and Crenarchaeota was found (Gardner *et al.* 2010). We hypothesise that the limited distribution of snoRNAs found in our study may be due to two related factors. Firstly, snoRNAs in archaea may be highly specific to the taxa they are found in. Secondly, as the covariance models of archaeal snoRNAs used in this study were built largely from examples from only a limited number of archaeal species, our power to detect snoRNAs with these models may be inhibited. We recommend the use of iterative methods, building on the initial models with data collected in each pass of the search, to address this in any future studies.

In our examination of the distribution of snoRNPs throughout the archaea we discovered that while the distribution of snoRNAs themselves was incomplete, this was not necessarily reflected in the snoRNPs. All major taxa showed evidence of snoRNP families being present. However, the H/ACA-associated snoRNP Gar1 was notably absent from the DPANN superphylum. We suggest that the absence of this protein in this case, and the absence of other H/ACA-associated snoRNPs in other taxa, should be investigated more thoroughly with aim to determine whether this finding does in fact indicate a lack of H/ACA snoRNAs in these taxa.

## References

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic acids research*, 41(D 1), D36-D42.

Dennis, P. P., & Omer, A. (2005). Small non-coding RNAs in Archaea. *Current opinion in microbiology*, 8(6), 685-694.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput Biol*, 7(10), e1002195.

Eddy, S. R., & Wheeler, T. J. (2013). HMMER User's Guide.

Federhen, S. (2012). The NCBI taxonomy database. *Nucleic acids research*, 40(D1), D136-D143.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2013). Pfam: the protein families database. *Nucleic acids research*, gkt1223.

Gardner, P. P., Bateman, A., & Poole, A. M. (2010). SnoPatrol: how many snoRNA genes are there?. *Journal of biology*, 9(1), 1.

Gaspin, C., Cavaillé, J., Erauso, G., & Bachellerie, J. P. (2000). Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the *Pyrococcus* genomes. *Journal of molecular biology*, 297(4), 895-906.

Girard, J. P., Lehtonen, H., Caizergues-Ferrer, M., Amalric, F., Tollervey, D., & Lapeyre, B. (1992). GAR1 is an essential small nucleolar RNP protein required for pre-rRNA processing in yeast. *The EMBO Journal*, 11(2), 673.

Henras, A. K., Capeyrou, R., Henry, Y., & Caizergues-Ferrer, M. (2004). Cbf5p, the putative pseudouridine synthase of H/ACA-type snoRNPs, can form a complex with Gar1p and Nop10p in absence of Nhp2p and box H/ACA snoRNAs. *RNA*, 10(11), 1704-1712.

Hoepfner, M. P., Gardner, P. P., & Poole, A. M. (2012). Comparative analysis of RNA families reveals distinct repertoires for each domain of life. *PLoS Comput Biol*, 8(11), e1002752.

Hoepfner, M. P., & Poole, A. M. (2012). Comparative genomics of eukaryotic small nucleolar RNAs reveals deep evolutionary ancestry amidst ongoing intragenomic mobility. *BMC evolutionary biology*, 12(1), 1.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11(1), 1.

Klein, R. J., Misulovin, Z., & Eddy, S. R. (2002). Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proceedings of the National Academy of Sciences*, 99(11), 7542-7547.

Lazar, C. S., Baker, B. J., Seitz, K., Hyde, A. S., Dick, G. J., Hinrichs, K. U., & Teske, A. P. (2016). Genomic evidence for distinct carbon substrate preferences and ecological niches of Bathyarchaeota in estuarine sediments. *Environmental microbiology*.

Nawrocki, E. P., & Eddy, S. R. (2013a). Computational identification of functional RNA homologs in metagenomic data. *RNA biology*, 10(7), 1170-1179.

Nawrocki, E. P., & Eddy, S. R. (2013b). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22), 2933-2935.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., & Finn, R. D. (2014). Rfam 12.0: updates to the RNA families database. *Nucleic acids research*, gku1063.

Omer, A. D., Lowe, T. M., Russell, A. G., Eberhardt, H., Eddy, S. R., & Dennis, P. P. (2000). Homologs of small nucleolar RNAs in Archaea. *Science*, 288(5465), 517-522.

Rashid, R., Aittaleb, M., Chen, Q., Spiegel, K., Demeler, B., & Li, H. (2003). Functional requirement for symmetric assembly of archaeal box C/D small ribonucleoprotein particles. *Journal of molecular biology*, 333(2), 295-306.

Rashid, R., Liang, B., Baker, D. L., Youssef, O. A., He, Y., Phipps, K., Terns, R.M., Terns, M.P., & Li, H. (2006). Crystal structure of a Cbf5-Nop10-Gar1 complex and implications in RNA-guided pseudouridylation and dyskeratosis congenita. *Molecular cell*, 21(2), 249-260.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., & Dodsworth, J. A. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459), 431-437.

Rozhdestvensky, T. S., Tang, T. H., Tchirkova, I. V., Brosius, J., Bachellerie, J. P., & Hüttenhofer, A. (2003). Binding of L7Ae protein to the K - turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic acids research*, 31(3), 869-877.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521(7551), 173-179.

Tang, T. H., Bachellerie, J. P., Rozhdestvensky, T., Bortolin, M. L., Huber, H., Drungowski, M., M., Elge, T., Brosius, J., & Hüttenhofer, A. (2002). Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proceedings of the National Academy of Sciences*, 99(11), 7536-7541.

Watanabe, Y. I., & Gray, M. W. (2000). Evolutionary appearance of genes encoding proteins associated with box H/ACA snoRNAs: cbf5p in *Euglena gracilis*, an early diverging eukaryote, and candidate Gar1p and Nop10p homologs in archaeobacteria. *Nucleic Acids Research*, 28(12), 2342-2352.

## Chapter Four - Summary and Concluding Remarks

In this study we sought to update existing knowledge about the distribution of both snoRNAs and riboswitches within archaeal species in light of new genomic data for emerging archaeal phyla becoming available. We also examined the relationship of riboswitches to genes found downstream of them within archaeal genomes. We found that while riboswitches not well distributed throughout the currently known archaeal phyla, snoRNAs families are much more widespread. We identified novel riboswitch-gene pairings within the archaea, and found evidence of a horizontal gene transfer of a riboswitch-gene pair from the bacteria into the archaea. We identified the need for further investigation into HGT of riboswitch-gene pairs from bacteria into the archaea and for the implications of a possible lack of the H/ACA box snoRNA-associated protein Gar1 to be further examined.

### ***Our results in the context of the currently known archaeal phyla***

Our study revealed several new findings about recently described archaeal phyla and the archaeal domain as a whole. We found that riboswitches are not well distributed throughout the archaeal domain. As evidence of horizontal gene transfer (HGT) of riboswitch-gene pairs from bacteria into the archaea was detected, we suggest that the distribution of known riboswitch families we have identified may be explained by a series of such independent HGT events. This finding is supported by a previous examination of the role of HGT in the distribution of non-coding RNA across all three domains (Hoepfner *et al.* 2012)

Of the known riboswitch families studied, our examination has revealed their presence in several archaeal taxa where they were not previously known. In particular, the more recently described Bathyarchaeota (Meng *et al.* 2014; Attar 2015), showed evidence of the TPP riboswitch, and the Fluoride riboswitch was detected in many of the Thaumarchaeota (Brochier-Armanet *et al.* 2008) genomes studied.

In our investigation of the distribution of snoRNAs, we discovered that while almost all archaeal taxa studied show evidence of known snoRNA families, these snoRNAs were not often found across every species within each taxa. This result is similar to the findings of Gardner *et al.* (2010) in which only 33% of Crenarchaeota

groups and 60% of Euryarchaeota groups showed evidence of known snoRNA families. In particular, known snoRNA families were underrepresented in the more recently described of the archaeal phyla (Rinke *et al.* 2013; Spang *et al.* 2015; Lazar *et al.* 2016) studied. No evidence of known archaeal snoRNA families was found in the Nanoarchaeota, the Aenigmarchaeota, the Lokiarchaeota, the Thorarchaeota or the Bathyarchaeota. We conclude that this finding may suggest a limited power to detect known snoRNA families based on the currently available models of these families combined with the possibility that snoRNAs that do occur in these phyla are of yet to be described RNA families.

We also investigated the distribution of snoRNA-associated protein families within the archaeal domain. We found that while C/D box associated snoRNPs are found in every archaeal phyla, the distribution of H/ACA box associated snoRNPs was less complete. We identify a noticeable absence of the Gar1 snoRNP, essential to the pseudouridylation process H/ACA box snoRNAs carry out, within the DPANN superphylum. We identify a need for further investigation of the prevalence of pseudouridylation modifications in species from these phyla.

### ***New knowledge of snoRNAs and riboswitches from this study***

Our study has uncovered new information about both the riboswitches and the snoRNAs in the archaea. In our examination of the genes found downstream of known riboswitch families, we found that there were some cases where the Fluoride riboswitch was not necessarily associated with the genes it would otherwise be expected to regulate (Weinburg *et al.* 2010; Baker *et al.* 2012). This finding raises questions about both the function of the riboswitch and the function of the genes it appears to regulate. While we did identify unusual cases of riboswitch-gene associations with the Fluoride riboswitch, other riboswitch families our study detected in the archaea were found to be associated with genes that those riboswitches are known to regulate (Rodionov *et al.* 2002; Pedrolli *et al.* 2015). Our investigation also revealed potential evidence for a riboswitch family not previously known from archaeal species. However, as genomic data for the species this potential occurrence of the Moco\_RNA\_motif riboswitch was found in is dissimilar to other material in the databases we can use to give context to the match, caution must be taken in determining that this is a genuine occurrence of this riboswitch within the archaea.

During this investigation, we have also discovered new information about the presences of snoRNAs within the archaea. While the prevalence of snoRNA and snoRNPs across the archaea detected in our study is consistent with an evolutionary ancient origin for these RNAs (Gardner *et al.* 2010; Hoepfner & Poole 2012), we also identified several cases where the knowledge of these RNAs within the archaea is incomplete. In particular, we found that coverage of known snoRNA families in newly identified archaeal phyla is poor. We suggest that, due to the presence of snoRNP detected in these taxa, as yet unknown snoRNA families may be found in these species. We also identified what appears to be a striking lack of the Gar1 snoRNP within the DPANN superphylum. Gar1 is an essential H/ACA box snoRNA-associated protein that guides pseudouridylation modifications (Girard *et al.* 1992; Rashid *et al.* 2006). Its absence in this superphylum may suggest a lack of H/ACA snoRNAs within these taxa and we suggest that investigating this finding further, both experimentally and bioinformatically would be of interest to a greater understanding of archaeal snoRNA families.

## ***Future Directions***

Through our findings, we have identified several opportunities for future work to build upon, and enhance knowledge, surrounding our conclusions. As the prevalence and function of classes of non-coding RNAs that are shared between multiple domains of life can tell us more about the evolutionary histories of the domains involved, it is important that investigations of these RNAs continue to be carried out. Many of the unknowns identified in this study present opportunities for targets of such future work. We suggest that both experimental and bioinformatic research be undertaken to increase our understanding of the function and gene-association of the fluoride riboswitch. Particularly, laboratory investigations may be used to help identify the function of the ORFs thought to be regulated by this riboswitch. Further and more precise bioinformatic study, including more robust phylogenetic investigation, may be used to confirm whether horizontal transfer has played a role in the riboswitch-gene distributions detected in this thesis. This also extends to other riboswitch families our examination has detected and identified as likely candidates for the occurrence of horizontal gene transfer events.

We also identify opportunities for further investigation of the snoRNA families found in the archaea. The limited prevalence of these RNAs in many archaeal taxa despite a broad distribution suggests the possibility that new families of snoRNA may

be able to be identified within taxa where fewer known snoRNAs have been located. Both bioinformatic probing and laboratory experimentation could be used to accomplish this. A more robust bioinformatic investigation of known families using models refined through iterative techniques could also help to generate a more precise picture of the distribution and prevalence of snoRNAs within the archaea. The distribution and prevalence of snoRNA-associated proteins should not be overlooked during such investigations. Our study demonstrates that while our power to detect the snoRNAs themselves may have been somewhat limited, the inclusion of consideration of the snoRNPs revealed specific targets for future study that may further our knowledge of archaeal snoRNAs. We identify a need for further investigation into pseudouridylation modifications within the DPANN superphylum to confirm our hypothesis of a lack of H/ACA snoRNA families in these species based on a lack of the Gar1 snoRNP throughout this superphylum.

## ***Conclusion***

In this study we have shown that while snoRNAs are widely distributed across the currently known archaeal phyla, the distribution of riboswitches across many archaeal taxa is limited. Our study suggests that the riboswitches may exist in the archaea through horizontal transfer events from bacteria rather than being an intrinsic part of the archaeal domain. We find riboswitches associated with genes mostly of previously known and expected function. However, we also identified opportunities for future work to confirm cases where the function of a known riboswitch within the archaea seems unclear based on the identity of ORFs located downstream of the riboswitch sequence.

We have found evidence that while particular families of snoRNA may not be distributed throughout all archaeal taxa, the snoRNAs themselves do appear to be an evolutionary ancient RNA class, common to both archaea and eukaryotes. We suggest further investigation be carried out to examine the lack of a H/ACA box-associated snoRNP, essential to pseudouridylation modifications, in the DPANN superphylum.

In light of recent expansions to our knowledge of the archaeal diversity, our investigations have provided a well needed update to both our knowledge of the distributions of, and knowledge about the evolutionary background involved in, the presence of both riboswitches and snoRNAs across the archaeal domain of life.

## References

- Attar, N. (2015). Archaeal genomics: A new phylum for methanogens. *Nature Reviews Microbiology*, 13(12), 739-739.
- Baker, J. L., Sudarsan, N., Weinberg, Z., Roth, A., Stockbridge, R. B., & Breaker, R. R. (2012). Widespread genetic switches and toxicity resistance proteins for fluoride. *Science*, 335(6065), 233-235.
- Brochier-Armanet, C., Boussau, B., Gribaldo, S., & Forterre, P. (2008). Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology*, 6(3), 245-252.
- Gardner, P. P., Bateman, A., & Poole, A. M. (2010). SnoPatrol: how many snoRNA genes are there?. *Journal of biology*, 9(1), 1.
- Girard, J. P., Lehtonen, H., Caizergues-Ferrer, M., Amalric, F., Tollervey, D., & Lapeyre, B. (1992). GAR1 is an essential small nucleolar RNP protein required for pre-rRNA processing in yeast. *The EMBO Journal*, 11(2), 673.
- Hoepfner, M. P., Gardner, P. P., & Poole, A. M. (2012). Comparative analysis of RNA families reveals distinct repertoires for each domain of life. *PLoS Comput Biol*, 8(11), e1002752.
- Hoepfner, M. P., & Poole, A. M. (2012). Comparative genomics of eukaryotic small nucleolar RNAs reveals deep evolutionary ancestry amidst ongoing intragenomic mobility. *BMC evolutionary biology*, 12(1), 1.
- Lazar, C. S., Baker, B. J., Seitz, K., Hyde, A. S., Dick, G. J., Hinrichs, K. U., & Teske, A. P. (2016). Genomic evidence for distinct carbon substrate preferences and ecological niches of Bathyarchaeota in estuarine sediments. *Environmental microbiology*.
- Meng, J., Xu, J., Qin, D., He, Y., Xiao, X., & Wang, F. (2014). Genetic and functional properties of uncultivated MCG archaea assessed by metagenome and gene expression analyses. *ISME J* 8: 650-659. *Isme Journal*, 8, 650-659.
- Pedrolli, D., Langer, S., Hobl, B., Schwarz, J., Hashimoto, M., & Mack, M. (2015). The ribB FMN riboswitch from *Escherichia coli* operates at the transcriptional and translational level and regulates riboflavin biosynthesis. *FEBS journal*, 282(16), 3230-3242.

Rashid, R., Aittaleb, M., Chen, Q., Spiegel, K., Demeler, B., & Li, H. (2003). Functional requirement for symmetric assembly of archaeal box C/D small ribonucleoprotein particles. *Journal of molecular biology*, 333(2), 295-306.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., & Dodsworth, J. A. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459), 431-437.

Rodionov, D. A., Vitreschak, A. G., Mironov, A. A., & Gelfand, M. S. (2002). Comparative genomics of thiamin biosynthesis in procaryotes New genes and regulatory mechanisms. *Journal of Biological chemistry*, 277(50), 48949-48959.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521(7551), 173-179.

Weinberg, Z., Wang, J. X., Bogue, J., Yang, J., Corbino, K., Moy, R. H., & Breaker, R. R. (2010). Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome biology*, 11(3), 1.

# Appendix One - Species Comprising the Archaeal Taxa Examined in this Thesis

## EURYARCHAEOTA

### EURYARCHAEOTA - HALOBACTERIA

*Candidatus Halobonum tyrrellensis*  
*Haladaptatus cibarius*  
*Haladaptatus paucihalophilus*  
*Halalkalicoccus jeotgali*  
*Halanaeroarchaeum sulfurreducens*  
*Halapricum salinum*  
*Halarchaeum acidiphilum*  
*haloarchaeon 3A1\_DGR*  
*Haloarcula amylolytica*  
*Haloarcula argentinensis*  
*Haloarcula californiae*  
*Haloarcula japonica*  
*Haloarcula salaria*  
*Haloarcula sinaiensis*  
*Haloarcula sp. CBA1115*  
*Haloarcula sp. SL3*  
*Haloarcula vallismortis*  
*Halobacterium salinarum*  
*Halobacterium sp. DL1*  
*Halobellus rufus*  
*Halobiforma lacisalsi*  
*Halobiforma nitratireducens*  
*Halococcus hamelinensis*  
*Halococcus morrhuae*  
*Halococcus saccharolyticus*  
*Halococcus salifodinae*  
*Halococcus sediminicola*  
*Halococcus agarilyticus*  
*Halococcus thailandensis*  
*Haloferax alexandrinus*  
*Haloferax denitrificans*  
*Haloferax elongans*  
*Haloferax gibbonsii*  
*Haloferax larsenii*  
*Haloferax lucentense*  
*Haloferax mediterranei*  
*Haloferax mucosum*  
*Haloferax prahovense*  
*Haloferax sp. ATB1*  
*Haloferax sp. ATCC BAA-644*  
*Haloferax sp. ATCC BAA-645*  
*Haloferax sp. ATCC BAA-646*  
*Haloferax sp. Arc-Hr*  
*Haloferax sp. BAB2207*  
*Haloferax sulfurifontis*  
*Haloferax volcanii*  
*Halogeometricum borinquense*  
*Halogeometricum pallidum*  
*Halogranum salarium*  
*Halolamina pelagica*  
*Halolamina rubra*  
*Halolamina sediminis*  
*Halomicrobium katesii*  
*Halomicrobium mukohataei*  
*Halonotius sp. J07HN4*  
*Halonotius sp. J07HN6*  
*halophilic archaeon J07HB67*  
*halophilic archaeon J07HX5*  
*halophilic archaeon J07HX64*  
*Halopiger salifodinae*  
*Halopiger djelfamassiliensis*  
*Halopiger goleamassiliensis*  
*Halopiger xanaduensis*  
*Haloplanus natans*  
*Haloquadratum sp. J07HQX50*  
*Haloquadratum walsbyi*  
*Halorhabdus tiamatea*  
*Halorhabdus utahensis*  
*Halorubrum aidingense*  
*Halorubrum arcis*  
*Halorubrum californiense*  
*Halorubrum coriense*  
*Halorubrum distributum*  
*Halorubrum ezzemoulense*  
*Halorubrum halophilum*  
*Halorubrum hochstenium*  
*Halorubrum kocurii*  
*Halorubrum lipolyticum*  
*Halorubrum litoreum*  
*Halorubrum saccharovororum*  
*Halorubrum sp. 5*  
*Halorubrum sp. AJ67*  
*Halorubrum sp. BV1*  
*Halorubrum sp. J07HR59*  
*Halorubrum sp. SD626R*  
*Halorubrum sp. T3*  
*Halorubrum tebenquichense*  
*Halorubrum terrestre*  
*Halosimplex carlsbadense*  
*Halostagnicola larsenii*  
*Halostagnicola sp. A56*  
*Haloterrigena jeotgali*  
*Haloterrigena limicola*  
*Haloterrigena salina*  
*Haloterrigena sp. H13*  
*Haloterrigena thermotolerans*  
*Haloterrigena turkmenica*  
*Halovivax asiaticus*  
*Halovivax ruber*  
*Natrialba aegyptia*  
*Natrialba asiatica*  
*Natrialba chahannaoensis*  
*Natrialba hulunbeirensis*  
*Natrialba magadii*  
*Natrialba taiwanensis*  
*Natrinema altunense*  
*Natrinema gari*  
*Natrinema pallidum*  
*Natrinema pellirubrum*  
*Natrinema sp. J7-1*  
*Natrinema sp. J7-2*  
*Natrinema versiforme*  
*Natronobacterium gregoryi*  
*Natronococcus amylolyticus*  
*Natronococcus jeotgali*

*Natronococcus occultus*  
*Natronolimnobius baerhuensis*  
*Natronolimnobius innermongolicus*  
*Natronomonas moolapensis*  
*Natronomonas pharaonis*  
*Natronorubrum bangense*  
*Natronorubrum sulfidifaciens*  
*Natronorubrum tibetense*  
*Salinarchaeum sp. Harcht-Bsk1*

#### EURYARCHAEOTA - ARCHAEOGLOBI

*Archaeoglobus fulgidus*  
*Archaeoglobus profundus*  
*Archaeoglobus sulfaticallidus*  
*Archaeoglobus veneficus*  
*Ferroglobus placidus*  
*Geoglobus acetivorans*  
*Geoglobus ahangari*

#### EURYARCHAEOTA - HADESARCHAEA

*Hadesarchaea archaeon DG-33*  
*Hadesarchaea archaeon DG-33-1*  
*Hadesarchaea archaeon YNP\_45*  
*Hadesarchaea archaeon YNP\_N21*

#### EURYARCHAEOTA - METHANOBACTERIA

*Methanobacterium arcticum*  
*Methanobacterium formicicum*  
*Methanobacterium lacus*  
*Methanobacterium paludis*  
*Methanobacterium sp. Maddingley MBC34*  
*Methanobacterium sp. SMA-27*  
*Methanobacterium veterum*  
*Methanobrevibacter arboriphilus*  
*Methanobrevibacter oralis*  
*Methanobrevibacter ruminantium*  
*Methanobrevibacter smithii*  
*Methanobrevibacter sp. AbM4*  
*Methanobrevibacter boviskoreani*  
*Methanobrevibacter wolinii*  
*Methanosphaera stadmanae*  
*Methanothermobacter marburgensis*  
*Methanothermobacter sp. CaT2*  
*Methanothermobacter thermotrophicus*  
*Methanothermobacter wolfeii*  
*Methanothermus fervidus*

#### EURYARCHAEOTA - METHANOCOCCI

*Methanocaldococcus fervens*  
*Methanocaldococcus infernus*  
*Methanocaldococcus jannaschii*  
*Methanocaldococcus sp. FS406-22*  
*Methanocaldococcus bathoardescens*  
*Methanocaldococcus villosus*  
*Methanocaldococcus vulcanius*  
*Methanococcus aeolicus*  
*Methanococcus maripaludis*  
*Methanococcus vannielii*  
*Methanococcus voltae*  
*Methanothermococcus okinawensis*  
*Methanothermococcus thermolithotrophicus*  
*Methanotorris formicicus*  
*Methanotorris igneus*

#### EURYARCHAEOTA - METHANOMICROBIA

*Methanosphaerula palustris*  
*Methanocella arvoryzae*

*Methanocella conradii*  
*Methanocella paludicola*  
*Methanococcoides burtonii*  
*Methanococcoides methylutens*  
*Methanocorpusculum labreanum*  
*Methanocorpusculum bavaricum*  
*Methanoculleus bourgensis*  
*Methanoculleus chikugoensis*  
*Methanoculleus marisnigri*  
*Methanoculleus sp. CAG:1088*  
*Methanoculleus sp. MH98A*  
*Methanoculleus sediminis*  
*Methanoculleus sp. SDB*  
*Methanofollis liminatans*  
*Methanogenium cariaci*  
*Methanohalobium evestigatum*  
*Methanohalophilus mahii*  
*Methanolacinia paynteri*  
*Methanolinea sp. SDB*  
*Methanolinea tarda*  
*Methanolobus psychrophilus*  
*Methanolobus tindarius*  
*Methanomethylovorans hollandica*  
*Methanomicrobium mobile*  
*Candidatus Methanoperedens nitroreducens*  
*Candidatus Methanoperedens sp. BLZ1*  
*Methanoplanus limicola*  
*Methanolacinia petrolearia*  
*Methanoregula formicica*  
*Methanoregula boonei*  
*Methanosaeta concilii*  
*Methanosaeta harundinacea*  
*Methanosaeta sp. SDB*  
*Methanosaeta thermophila*  
*Methanosalsum zhilinae*  
*Methanosarcina acetivorans*  
*Methanosarcina barkeri*  
*Methanosarcina horonobensis*  
*Methanosarcina lacustris*  
*Methanosarcina mazei*  
*Methanosarcina siciliae*  
*Methanosarcina soligelidi*  
*Methanosarcina sp. 1.H.A.2.2*  
*Methanosarcina sp. 1.H.T.1A.1*  
*Methanosarcina sp. 2.H.A.1B.4*  
*Methanosarcina sp. 2.H.T.1A.15*  
*Methanosarcina sp. 2.H.T.1A.3*  
*Methanosarcina sp. 2.H.T.1A.6*  
*Methanosarcina sp. 2.H.T.1A.8*  
*Methanosarcina sp. 795*  
*Methanosarcina flavescens*  
*Methanosarcina sp. Kolksee*  
*Methanosarcina sp. MTP4*  
*Methanosarcina sp. WH1*  
*Methanosarcina sp. WWM596*  
*Methanosarcina thermophila*  
*Methanosarcina vacuolata*  
*Methanospirillum hungatei*  
*Methermicoccus shengliensis*

#### EURYARCHAEOTA - METHANOPYRI

*Methanopyrus kandleri*

#### EURYARCHAEOTA - THERMOCOCCI

*Palaeococcus ferrophilus*  
*Palaeococcus pacificus*  
*Pyrococcus abyssi*

*Pyrococcus furiosus*  
*Pyrococcus horikoshii*  
*Pyrococcus* sp. NA2  
*Pyrococcus* sp. ST04  
*Pyrococcus yayanosii*  
*Thermococcus barophilus*  
*Thermococcus celer*  
*Thermococcus cleftensis*  
*Thermococcus eurythermalis*  
*Thermococcus gammatolerans*  
*Thermococcus kodakarensis*  
*Thermococcus litoralis*  
*Thermococcus nautilii*  
*Thermococcus onnurineus*  
*Thermococcus peptonophilus*  
*Thermococcus sibiricus*  
*Thermococcus* sp. 4557  
*Thermococcus* sp. AM4  
*Thermococcus* sp. EP1  
*Thermococcus paralvinellae*  
*Thermococcus* sp. JCM 11816  
*Thermococcus* sp. PK  
*Thermococcus thio reducens*  
*Thermococcus zilligii*

#### EURYARCHAEOTA - THERMOPLASMATA

*Acidiplasma aeolicum*  
*Acidiplasma cupricumulans*  
*Acidiplasma* sp. MBA-1  
*Ferroplasma acidarmanus*  
*Ferroplasma* sp. Type II  
*Candidatus Methanomassiliicoccus intestinalis*  
*Methanomassiliicoccales* archaeon RumEn M1  
*Methanomassiliicoccales* archaeon RumEn M2  
*Methanomassiliicoccus luminyensis*  
*Candidatus Methanomethylophilus alvus*  
*Candidatus Methanoplasma termitum*  
*Picrophilus torridus*  
*Thermogymnomonas acidicola*  
*Thermoplasma acidophilum*  
*Thermoplasma volcanium*  
*Thermoplasmatales* archaeon A-plasma  
*Thermoplasmatales* archaeon BRNA1  
*Thermoplasmatales* archaeon E-plasma  
*Thermoplasmatales* archaeon Gpl  
*Thermoplasmatales* archaeon I-plasma  
*Thermoplasmatales* archaeon SCGC AB-539-C06  
*Thermoplasmatales* archaeon SCGC AB-539-N05  
*Thermoplasmatales* archaeon SCGC AB-540-F20

#### EURYARCHAEOTA - UNCLASSIFIED

*Aciduliprofundum boonei*  
*Aciduliprofundum* sp. MAR08-339  
*Euryarchaeota* archaeon SCGC AAA252-I15  
*Euryarchaeota* archaeon SCGC AAA286-E23  
*euryarchaeote* SCGC AAA261-E04  
*euryarchaeote* SCGC AAA261-G15  
*Marine Group II euryarchaeote* SCGC AB-629-J06  
*Marine Group III euryarchaeote* SCGC AAA007-O11  
*Marine Group III euryarchaeote* SCGC AAA288-E19

*Marine group II euryarchaeote* SCGC AAA288-C18

#### DPANN GROUP

AENIGMARCHAEOTA  
*Candidatus Aenigmarchaeota* archaeon JGI 0000106-F11  
*Candidatus Aenigmarchaeota* archaeon SCGC AAA011-F07  
*Candidatus Aenigmarchaeum subterraneum* SCGC AAA011-O16

#### DIAPHEROTRITES

*Candidatus Iainarchaeum andersonii* SCGC AAA011-E11  
*Diapherotrites* archaeon SCGC AAA011-K09  
*Diapherotrites* archaeon SCGC AAA011-N19

#### NANOARCHAEOTA

*Nanoarchaeota* archaeon JGI OTU-1  
*Nanoarchaeota* archaeon JGI OTU-2  
*Nanoarchaeota* archaeon SCGC AAA011-D5  
*Nanoarchaeota* archaeon SCGC AAA011-G17  
*Nanoarchaeota* archaeon SCGC AAA011-J2  
*Nanoarchaeota* archaeon SCGC AAA011-K22  
*Nanoarchaeota* archaeon SCGC AAA011-L15  
*Nanoarchaeota* archaeon SCGC AAA011-L22  
*nanoarchaeote* Nst1  
*Nanoarchaeum equitans* Kin4-M  
NANOHALOARCHAEOTA  
*Candidatus Haloredivivus* sp. G17  
*Candidatus Nanosalina* sp. J07AB43  
*Candidatus Nanosalinarum* sp. J07AB56  
*Nanohaloarchaea* archaeon AB578-D14

#### PARVARCHAEOTA

*Candidatus Micrarchaeum acidiphilum* ARMAN-2  
*Candidatus Parvarchaeum acidiphilum* ARMAN-4  
*Candidatus Parvarchaeum acidophilum* ARMAN-5

#### TACK GROUP

##### THAUMARCHAEOTA - NITROSOSPHERIA

*Candidatus Nitrososphaera evergladensis*  
*Candidatus Nitrososphaera gargensis*  
*Nitrososphaera viennensis*

##### THAUMARCHAEOTA - NITROSOPUMILALES

*Candidatus Nitrosoarchaeum koreensis*  
*Candidatus Nitrosoarchaeum limnia*  
*Candidatus Nitrosopumilus salaria*  
*Candidatus Nitrosopumilus* sp. AR2  
*Candidatus Nitrosopumilus piranensis*  
*Candidatus Nitrosopumilus adriaticus*  
*Nitrosopumilus maritimus* SCM1  
*Nitrosopumilus* sp. AR  
*Nitrosopumilus* sp. BA CL13 MAG-120910-bin56  
*Nitrosopumilus* sp. BA CL13 MAG-121220-bin23  
*Nitrosopumilus* sp. PRT-SC01  
*Nitrosopumilus* sp. SJ

THAUMARCHAEOTA - CENARCHAEALES

*Cenarchaeum symbiosum*

THAUMARCHAEOTA - UNCLASSIFIED

*Candidatus Nitrosopelagicus brevis*  
*Marine Group I thaumarchaeote* SCGC AAA799-B03  
*Marine Group I thaumarchaeote* SCGC AAA799-D07  
*Marine Group I thaumarchaeote* SCGC AAA799-D11  
*Marine Group I thaumarchaeote* SCGC AAA799-E16  
*Marine Group I thaumarchaeote* SCGC AAA799-N04  
*Marine Group I thaumarchaeote* SCGC AAA799-O18  
*Marine Group I thaumarchaeote* SCGC AAA799-P11  
*Marine Group I thaumarchaeote* SCGC AB-629-A13  
*Marine Group I thaumarchaeote* SCGC AB-629-I23  
*Marine Group I thaumarchaeote* SCGC RSA3  
*Thaumarchaeota archaeon* CSP1-1  
*Thaumarchaeota archaeon* MY2  
*Thaumarchaeota archaeon* MY3  
*Thaumarchaeota archaeon* N4  
*Candidatus Nitrosotenuis cloacae*  
*Thaumarchaeota archaeon* SCGC AAA007-O23  
*Thaumarchaeota archaeon* SCGC AAA282-K18  
*Thaumarchaeota archaeon* SCGC AAA287-E17  
*Thaumarchaeota archaeon* SCGC AAA287-I03  
*Thaumarchaeota archaeon* SCGC AAA287-N16  
*Thaumarchaeota archaeon* SCGC AB-179-E04  
*Thaumarchaeota archaeon* SCGC AB-539-E09

AIGARCHAEOTA

*Aigarchaeota archaeon* JGI 0000001-A7  
*Aigarchaeota archaeon* JGI 0000001-B8  
*Aigarchaeota archaeon* JGI 0000001-H6  
*Aigarchaeota archaeon* JGI 0000106-J15  
*Aigarchaeota archaeon* SCGC AAA471-A16  
*Aigarchaeota archaeon* SCGC AAA471-B22  
*Aigarchaeota archaeon* SCGC AAA471-D15  
*Aigarchaeota archaeon* SCGC AAA471-E14  
*Aigarchaeota archaeon* SCGC AAA471-E16  
*Aigarchaeota archaeon* SCGC AAA471-F17  
*Aigarchaeota archaeon* SCGC AAA471-G05  
*Aigarchaeota archaeon* SCGC AAA471-I13  
*Aigarchaeota archaeon* SCGC AAA471-J07  
*Aigarchaeota archaeon* SCGC AAA471-J08  
*Candidatus Caldiarchaeum subterraneum*  
*Thaumarchaeota archaeon* JGI OTU-1  
*Thaumarchaeota archaeon* JGI OTU-2  
*Thaumarchaeota archaeon* JGI OTU-3  
*Thaumarchaeota archaeon* JGI OTU-4

CRENARCHAEOTA - THERMOPROTEI

*Candidatus Acidianus copahuensis*  
*Acidianus hospitalis*  
*Acidilobus saccharovorans*  
*uncultured Acidilobus* sp. CIS

*uncultured Acidilobus* sp. JCHS  
*uncultured Acidilobus* sp. MG  
*uncultured Acidilobus* sp. OSP8  
*Aeropyrum camini*  
*Aeropyrum pernix*  
*Caldisphaera lagunensis*  
*Caldivirga maquilingensis*  
*Desulfurococcus amylolyticus*  
*Desulfurococcus fermentans*  
*Desulfurococcus kamchatkensis*  
*Desulfurococcus mobilis*  
*Desulfurococcus mucosus*  
*Fervidicoccus fontis*  
*Hyperthermus butylicus*  
*Ignicoccus hospitalis*  
*Ignisphaera aggregans*  
*Metallosphaera cuprina*  
*Metallosphaera hakonensis*  
*Metallosphaera sedula*  
*Metallosphaera yellowstonensis*  
*Pyrobaculum aerophilum* str. IM2  
*Pyrobaculum arsenaticum*  
*Pyrobaculum calidifontis*  
*Pyrobaculum islandicum*  
*Pyrobaculum oguniense*  
*Pyrobaculum ferrireducens*  
*Pyrobaculum* sp. WP30  
*Pyrodictium delaneyi*  
*Pyrolobus fumarii*  
*Staphylothermus hellenicus*  
*Staphylothermus marinus*  
*Sulfolobales archaeon* AZ1  
*Sulfolobales archaeon* Acd1  
*Sulfolobus acidocaldarius*  
*Sulfolobus islandicus*  
*Sulfolobus metallicus*  
*Sulfolobus solfataricus*  
*Sulfolobus shibatae*  
*Sulfolobus* sp. JCM 16833  
*Sulfolobus tokodaii*  
*Thermocladium modestius*  
*Thermofilum carboxyditrophus*  
*Thermofilum pendens*  
*Thermofilum* sp. 1807-2  
*Thermofilum* sp. 1910b  
*Thermogladius cellulolyticus*  
*Pyrobaculum neutrophilum*  
*Thermoproteus* sp. AZ2  
*Thermoproteus tenax*  
*Thermoproteus uzoniensis*  
*Thermosphaera aggregans*  
*Vulcanisaeta distributa*  
*Vulcanisaeta moutnovskia*  
*Vulcanisaeta souniana*  
*Vulcanisaeta* sp. AZ3  
*Vulcanisaeta* sp. JCM 14467  
*Vulcanisaeta* sp. JCM 16159  
*Vulcanisaeta* sp. JCM 16161

CRENARCHAEOTA - UNCLASSIFIED

*Crenarchaeota archaeon* SCGC AAA471-B05  
*Crenarchaeota archaeon* SCGC AAA471-B23  
*Crenarchaeota archaeon* SCGC AAA471-C03  
*Crenarchaeota archaeon* SCGC AAA471-L13  
*Crenarchaeota archaeon* SCGC AAA471-L14  
*Crenarchaeota archaeon* SCGC AAA471-O08  
*crenarchaeote* JGI-OTU-1

*crenarchaeote* SCGC AAA261-C22  
*crenarchaeote* SCGC AAA261-F05  
*crenarchaeote* SCGC AAA261-G18  
*crenarchaeote* SCGC AAA261-L14  
*crenarchaeote* SCGC AAA261-L22  
*crenarchaeote* SCGC AAA261-N13  
*crenarchaeote* SCGC AAA261-N23

#### KORARCHAEOTA

*Candidatus Korarchaeum cryptofilum*

#### **CANDIDATE PHyla**

##### LOKIARCHAEOTA

*Lokiarchaeum* sp. GC14\_75

##### THORARCHAEOTA

*Candidatus Thorarchaeota* archaeon SMTZ-45

*Candidatus Thorarchaeota* archaeon SMTZ1-45

*Candidatus Thorarchaeota* archaeon SMTZ1-83

##### BATHYARCHAEOTA

*Candidatus Bathyarchaeota* archaeon BA1

*Candidatus Bathyarchaeota* archaeon BA2

*miscellaneous Crenarchaeota* group archaeon SMTZ-80

*miscellaneous Crenarchaeota* group archaeon SMTZ1-55

*miscellaneous Crenarchaeota* group-1

archaeon SG8-32-1

*miscellaneous Crenarchaeota* group-1

archaeon SG8-32-3

*miscellaneous Crenarchaeota* group-15

archaeon DG-45

*miscellaneous Crenarchaeota* group-6

archaeon AD8-1

*uncultured miscellaneous Crenarchaeota* group