Abstract

Computational Discovery of Structured Non-coding RNA Motifs in Bacteria

Kenneth Ivan Brewer

2021

This dissertation describes a range of computational efforts to discover novel structured non-coding RNA (ncRNA) motifs in bacteria and generate hypotheses regarding their potential functions. This includes an introductory description of key advances in comparative genomics and RNA structure prediction as well as some of the most commonly found ncRNA candidates. Beyond that, I describe efforts for the comprehensive discovery of ncRNA candidates in 25 bacterial genomes and a catalog of the various functions hypothesized for these new motifs. Finally, I describe the Discovery of Intergenic Motifs PipeLine (DIMPL) which is a new computational toolset that harnesses the power of support vector machine (SVM) classifiers to identify bacterial intergenic regions most likely to contain novel structured ncRNA and automates the bulk of the subsequent analysis steps required to predict function. In totality, the body of work will enable the large scale discovery of novel structured ncRNA motifs at a far greater pace than possible before.

Computational Discovery of Structured Non-coding RNA Motifs in Bacteria

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by

Kenneth Ivan Brewer

Dissertation Director: Ronald R. Breaker

June 2021

Table of Contents

Abstract ..... **i**

Table of Contents ..... **v**

List of Figures ..... **viii**

List of Tables ..... **xi**

Acknowledgements ..... **xii**

Chapter One **1**

Introduction 2

Background 2

Central Dogma of Molecular Biology 2

RNA World 4

Riboswitches 5

Approaches to Riboswitch Discovery 6

Searching for unexplained gene regulatory phenomena 6

Computational search by sequence clustering 8

Computational search by GC/length filtering of IGRs 10

Tables and Figures 13

Figures 13

Chapter Two **18**

Summary 19

Introduction 20

Results and Discussion 22

Genomes and IGRs selected for comparative sequence analysis 22

Classification of newly discovered nucleic acid motifs 24
Strong riboswitch candidates 25
Weak riboswitch candidates 31
An unusual gene control element incorporating the *odc1* motif. 33
Additional notable motifs identified with the GC-IGR pipeline 36
Comparison with transcriptomics-based approaches 41
Summary of the motifs discovered and implications for future searches 43
Concluding Remarks 44
Materials and Methods 50
Databases and bioinformatics. 50
Chemical and oligonucleotides. 50
Bacterial strains and growth conditions. 50
Reporter Gene Construct Design. 51
Liquid-Based β-galactosidase Assays. 51
Tables and Figures 53
Tables 53
Figures 56
Chapter Three 80
Summary 81
Introduction 81
Results and Discussion 82
Pipeline Overview 82
Details on SVM Enrichment 84

Usage Guide 84
Workstation Setup and Installation 85
Compute Station Setup 86
Notebook 1: Genome IGR Selection 88
Notebook 2: BLAST data processing 90
Notebook 3: IGR Report 92
Conclusion 94
Tables and Figures 95
Figures 95
References 100

List of Figures

Figure 1-1: Riboswitch Ligands and Abundances 13

Figure 1-2: Riboswitch class abundances 14

Figure 1-3: Plot of IGRs from *Pelagibacter ubique* 15

Figure 1-4: Overview of GC-IGR Computational Pipeline 16

Figure 2-1: Consensus sequence and structural models for strong riboswitch candidates identified in this study. 56

Figure 2-2: Consensus sequence and structural models for the weak riboswitch candidates identified in this study. 57

Figure 2-3: The *odc1* motif consensus model and gene associations. 58

Figure 2-4: Expression of proteins from the two *odc1* translation start codons. 59

Figure 2-5: Consensus sequence and secondary structure models for additional structured nucleic acid motifs that are representative of those identified in this study.. 61

Figure 2-6: Comprehensive summary of classification of selected unknown IGRs from the analysis of 26 bacterial genomes chosen for this study. 62

Figure 2-7. Motifs identified from the IGRs containing the “riboregulators” identified via term-Seq. 63

Figure 2-8: Additional motifs identified from the IGRs containing the “riboregulators” identified via term-Seq. 65

Figure 2-9. Plot of the IGRs from the L. monocytogenes genome. 67

Figure 2-10: Plot of the IGRs from the *F. nucleatum* genome. 67

Figure 2-11: Plot of the IGRs from the *Ruegeria* genome 68

Figure 2-12: Plot of the IGRs from the *Clostridium perfringens* genome. 68

Figure 2-13: Plot of the IGRs from the *P. pentosaceus* genome. 69

Figure 2-14: Plot of the IGRs from the *C. fetus* genome. 69

Figure 2-15: Plot of the IGRs from the *P. necessarius* genome. 70

Figure 2-16: Plot of the IGRs from the *A. laidlawii* genome..... 70

Figure 2-17: Plot of the IGRs from the *L. biflexa* genome. 71

Figure 2-18: Plot of the IGRs from the *T. africanus* genome. 71

Figure 2-19: Plot of the IGRs from the *Exiguobacterium* genome. 72

Figure 2-20: Plot of the IGRs from the *M. mobilis* genome. 72

Figure 2-21: Plot of the IGRs from the *V. parvula* genome. 73

Figure 2-22: Plot of the IGRs from the *C.R. pediculicola* genome. 73

Figure 2-23: Plot of the IGRs from the *A. nitrofigilis* genome. 74

Figure 2-24: Plot of the IGRs from the *M. fermentans* genome. 74

Figure 2-25: Plot of the IGRs from the *H. maritima* genome. 75

Figure 2-26: Plot of the IGRs from the *T. geofontis* genome. 75

Figure 2-27: Plot of the IGRs from the *T. asinigenitalis* genome. 76

Figure 2-28: Plot of the IGRs from the *Z. mobilis* genome. 76

Figure 2-29: Plot of the IGRs from the *C. K. oncopelti* genome. 77

Figure 2-30: Plot of the IGRs from the *Hydrogenobaculum* genome. 77

Figure 2-31: Plot of the IGRs from the *B. proteobacterium* genome. 78

Figure 2-32: Plot of the IGRs from the *S. salinus* genome. 78

Figure 2-33: Plot of the IGRs from the *C. pecorum* genome..... 79

Figure 2-34: Plot of the IGRs from the *C. B. massiliensis* genome. 79

Figure 3-1: Overview of DIMPL Process 95

Figure 3-2: Plot of IGRs from the genome *Campylobacter jejni* generated by DIMPL. . 96

Figure 3-3: Plot of IGR selection from *Campylobacter jejuni* enriched by DIMPL 97

Figure 3-4: Graphical depiction of blast hit processing with default parameters 98

Figure 3-5: Sample hit from genetic context report generated by DIMPL 99

List of Tables

Table 2-1: List of all 26 genomes included in analysis. 53

Table 2-2: Sequences of synthetic DNAs used in this study. 54

Acknowledgements

I would like to begin by thanking the many individuals that have contributed to my scientific training through the years before I reached Yale. That begins with Mr. Dave Shelton and Mr. Alan Myrup who taught AP Chemistry and AP Biology at Timpview High School, and continues with Dr Merrit Andrus at Brigham Young University who allowed me to participate in the organic chemistry research in his lab while I took concurrent enrollment courses and Dr. Brian Woodfield whose Honors chemistry course was an excellent tutorial in the complexities of scientific thinking. I'd also like to thank Dr. Greg Tucci at Harvard for guiding me through the Chemistry program at Harvard and Dr. Alan Saghatelian who gave me my first exposure to biologically focused chemistry research in his lab.

At Yale, I would like to thank Dr. Mark Solomon for the excellent training in scientific thinking I gained in the Methods and Logics in Molecular Biology course. I would like to thank Dr. Breaker for taking me into his lab and encouraging me through the ups and downs of my thesis research. I'd also like to thank the many members of the Breaker Lab whose conversations and advice have stimulated my own thinking on many occasions. I would also be remiss not to thank my thesis committee members Dr. Scott Strobel and Dr. Karla Neugebauer whose advice has helped me focus my research efforts in the most impactful directions on many occasions.

Finally, I'd like to thank my parents for helping me develop a questioning mind that takes nothing for granted and persistent desire to strive for excellence. And my spouse, Kelly, whose love and support has been invaluable as I've completed this journey.

Chapter One

Introduction

Introduction

At the core of molecular biology are three primary biopolymers: deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins. This dissertation is focused on the second of those three biomolecules, RNA. This first chapter will seek to establish the value of finding new structured RNA motifs and provide the background to explain how this search can be accomplished. An example of such discovery efforts will be provided in the second chapter which includes results from an in-depth search of 26 bacterial genomes for new structured RNA motifs. Finally, this dissertation's third chapter will present a new computational toolset, called Discovery of Intergenic Motifs PipeLine (DIMPL), which will make the future discovery of structured RNA motifs far easier than ever before. Taken together, these chapters will seek to make the argument that finding new examples of structured ncRNA motifs is of great scientific value and that the combination of methods and tools presented here can be used to uncover many more functions for this versatile biomolecule.

Background

Central Dogma of Molecular Biology

Nearly 65 years ago, Francis Crick gave a lecture[1] that has been called "one of the most significant lectures in the history of biology"[2]. In this lecture, Crick discussed the flow of genetic information from DNA to RNA to proteins in a set of ideas that was referred to as the "Sequence Hypothesis" or "The Central Dogma"[3]. According to this dogma, DNA plays a role primarily as an information storehouse keeping genetic information safe and secure. RNA's primary role is that of an intermediary, or messenger, storing temporary

copies of the genetic information before the information is used to assemble amino acids in the proper order required for a specific protein.

Although this foundational lecture was largely correct about most biological genetic information flow, the complexity of biology has required this "dogma" to be reassessed and revised numerous times[3–5]. The biological macromolecule from the central dogma that has likely required the most substantial reassessments and revisions is RNA. The original monolithic role of RNA solely as the information-carrying intermediate in the central dogma was replaced with the idea that RNA comes in three "flavors"[6]: 1) messenger RNA (mRNA) the temporary copy of genetic information, 2) transfer RNA (tRNA) which act as adaptors that mediate the triplet genetic code and the encoded amino acid and 3) ribosomal RNA (rRNA) which was largely assumed to have a purely structural role in protein synthesis.

The intervening decades have shown that even this tripartite classification of RNA's role in biology is far too simplistic. Largely responsible for this latest round of reassessments is the growing appreciation for the diverse range of non-coding RNA (ncRNA) genes that are active at the level of RNA itself[6]. Some of the diverse ncRNA classes that are active in broad domains of life include (i) RNase P[7,8], an RNA enzyme (or ribozyme) which performs site-specific phosphoester hydrolysis to process many precursor RNA transcripts, (ii) the signal recognition particle RNA[9,10], which helps route certain proteins to the cell membrane and (iii) 6S RNA[11], which serves as a regulator of RNA polymerase activity in bacteria. There are also a number of biochemical and biological functions that ncRNA classes fulfill that are rarer, but no less remarkable. These rarer classes include ribozymes, which have RNA processing activity[12,13], and riboswitches,

which can control gene expression through direct metabolite binding[14–16]. Riboswitches and ribozymes in particular required substantial reassessment of RNA's role in biology by fulfilling functions (catalysis and gene control) that had been previously viewed as the exclusive domain of proteins.

RNA World

Taken together, the diverse roles that structured ncRNA elements fill in biology also hint at a tantalizing possibility. With RNA capable of storing of genetic information (as shown by mRNA), controlling gene expression in response to changes in the environment (as shown by riboswitches) and chemical catalysis (as shown by ribozymes), there is a strong case that RNA is the parent biomolecule found at the dawn of evolution and molecular life. This idea, often called the "RNA World" theory, was seen as quite controversial when it was first proposed[17]. However, support for this theory has only grown as researchers have made a number of key insights and discoveries.

One such insight came from the study of protein chemistry, where the observation that proteins rely so much on ribonucleotide-based enzymatic cofactors led to these cofactors to be considered "fossils of an earlier metabolic state"[18]. Another key piece of evidence was the discovery that ribonucleotides can indeed arise spontaneously in prebiotically plausible conditions[19]. However, no single experiment makes an argument in favor of the RNA World as powerful as the great diversity of roles that ncRNA has been found to fulfill. Finding even more examples of structured ncRNA taking on these unusual roles will further strengthen this body of evidence in support of the RNA World theory and may provide a window into the earliest form of molecular life.

Riboswitches

One class of ncRNA motifs that likely descended from the RNA World but are still extensively used by bacteria today are riboswitches. These remarkable ncRNAs are capable of controlling gene expression through direct binding to a metabolite or ion. Riboswitches accomplish this task by forming an aptamer domain responsible for ligand binding and an effector domain which contains a mechanism for gene control[20]. Over 50 classes of riboswitches have been discovered[21–33] to date binding a wide range of ligands (**Figure 1-1**). The bulk of the ligands discovered to date are a variety of RNA derivatives although a many ions, amino acids and other metabolites have also been identified as riboswitch ligands.

Besides just varying in their ligand specificity, riboswitch classes also exhibit great variety in their abundances (**Figure 1-2A**). The riboswitch class with the most unique representatives is the TPP riboswitch class with 16,701 representatives. The rarest riboswitch class is an FMN-variant riboswitch class with only two known representatives. Plotting the abundances of riboswitch classes in rank order on a log-log plot (**Figure 1-2B**) allows predictions based on the power law[21,34,35] about the number of total riboswitch classes one might expect to find in our current sequence databases. The predicted total number of classes is on the order of thousands, far exceeding the roughly 50 classes currently known today. The same equations that predict thousands of riboswitch classes yet to be discovered, also predict that the remaining classes will be increasingly rare. Developing the methods and tools to discover these increasingly rare undiscovered riboswitch classes is one of the central objectives of the work described in this dissertation.

Approaches to Riboswitch Discovery

Having established that a great number of riboswitch classes remain to be found and finding them will provide valuable insights into the dawn of molecular life, it is worth carefully considering the various approaches that have been used to discover these motifs in the past and whether or not these approaches are likely to be useful in discovering new motifs today.

Searching for unexplained gene regulatory phenomena

Many of the earliest riboswitch candidates were discovered by searching existing scientific literature for motifs of unknown function that had been identified in the 5' untranslated region (5'-UTR) of metabolic genes. This was the case for the B_{12} box[36], *thi* box[37] and the *RFN* element[38] which were previously known motifs but were later proven to be portions of the first three validated riboswitch classes, namely the AdoCbl[39], TPP[40,41], and FMN[41,42] riboswitch classes respectively. The most promising candidates for riboswitch validation identified via this approach would often include several pieces of evidence already present in the scientific literature hinting at possible riboswitch function. For example, the fifth riboswitch class to be discovered, the lysine riboswitch[43,44], was preceded by (i) evidence showing transcription repression of the lysine metabolism gene *LysC* by high concentrations of lysine itself[45], (ii) evidence showing that this repression was not caused by protein partners[46], (iii) examples of bacterial mutants with a constitutively active *LysC* gene where the mutation was mapped to the 5'-UTR[47] and (iv) evidence of sequence homology between the *B. subtilis* and *E. coli lysC* 5'-UTRs[48]. With perfect hindsight and a conceptual understanding of riboswitches, this body of preliminary evidence readily

available in the scientific literature almost makes the eventual validation of the lysine riboswitch seem like a forgone conclusion. This example also perfectly illustrates the exceptional utility that searching scientific literature for unexplained gene regulatory phenomena provided researchers seeking to discover the first several riboswitch classes.

Unfortunately, the prospects for discovering new riboswitch classes using this same approach today seem quite poor. Beyond the obvious challenges of conducting a literature-search approach to riboswitch candidate discovery in a systematic fashion, this approach is also unlikely to be productive simply because riboswitches are now widely recognized as relatively common bacterial regulatory mechanisms[21]. Modern researchers working to understand the regulation of a particular bacterial gene in a collaborative environment are unlikely to assemble a body of evidence similar to that which preceded the validation of the lysine riboswitch, without independently realizing that a riboswitch may be involved. This means that any attempts to find new riboswitch candidates via unexplained gene regulatory phenomena in the scientific literature may, by necessity, limit themselves to literature published before a certain date. These would-be riboswitch discoverers may eventually find themselves studying increasingly obscure references of little practical utility and nothing to show for their efforts.

Another, perhaps more critical, reason that a literature search for unexplained gene control phenomena is not a sensible approach for riboswitch candidate discovery today is because only a few model bacterial organisms, such as *B. subtilis* and *E. coli,* attract the intense study that would motivate researchers to publish extensively on the particulars of gene control mechanisms for specific genes. For those model organisms, the extensive study that has already been focused on their gene regulatory mechanisms make it less likely

that undiscovered riboswitch classes remain to be found within their genomes. The "low-hanging fruit" has likely already been plucked from these highly studied genomes. However, even if the scientific community were to begin studying gene control in new bacterial organisms with the same intensity historically devoted to *E. coli* and *B. subtilis*, the odds are not good that those studies would find new riboswitches candidates. This can be deduced from a notable paradox of the power-law predictions for riboswitch class distribution. While the number of undiscovered riboswitch classes present in current sequence databases likely numbers in the low thousands[21,34,35], these same sequence databases contain several thousand more bacterial genomes. On average, one might expect any given genome to have only one (or zero) new classes hidden somewhere in the 5'-UTR of one of its genes. The odds of that particular gene becoming the focus of study and publication that can later be found by would-be riboswitch candidate discoverers are vanishingly small. This is especially true for the increasingly rare riboswitch candidates that the power law predicts remain to be found.

Computational search by sequence clustering

As discussed in the previous section, the prior identification of sequence homology across species was one of the most powerful pieces of evidence used to find early riboswitch candidates in the scientific literature. It is therefore intuitive that the second wave of riboswitch candidate discovery efforts was focused on harnessing computers to find similar sets of homologous sequences in a systematic rather than serendipitous manner.

The first major effort[49] to discover new riboswitch candidates using sequence homology relied on clustered sequences from the 5'-UTRs extracted from the *B. subtilis* genome as the search seeds and the full set of intergenic regions (IGRs) extracted from 91

bacterial genomes as the search database. This discovery effort was extremely successful. Six motifs were identified as riboswitch candidates, all of which were eventually validated[50–55], although in some cases[54,55], it was not until several years later.

This first computational discovery effort to find riboswitch candidates has been followed by a wave of successors[56–61] that have continually expanded their search databases and have had various approaches to generating the clusters of similar sequences used as the search seeds. The first set of these computational searches used search seeds generated from the 5'-UTRs of an individual bacterial organism[49,56]. The next iteration generated their search seeds by clustering the 5'-UTRs found in front of the all representatives of the same protein family[57,58]. This was followed by a succession of search efforts[59–61] that relied on BLAST to generate cluster-based search seeds without relying on protein domain annotations. The more recent of these computational searches[57–61] have all used CMfinder[62] to generate possible motif covariation models from the search seeds and used Infernal[63] to leverage those covariation models in structure-aware search queries. Collectively, these computational search approaches have been enormously successful and have been the primary source for the discovery of dozens of validated riboswitch classes[21,64,65]. Cognate computational search approaches focused on very large ncRNAs[61,66] and ribozyme[67,68] discovery have also been successful when clusters of extremely long IGRs were used as search seeds for the former and clusters of IGRs near known ribozymes were used for the latter.

Given the success of this sequence clustering approach to ncRNA discovery for the past several years, one might expect the approach to continue as a productive source of new riboswitch candidates for years to come as more bacterial genomes are sequenced.

Unfortunately, there appear to be some unexpected limitations to generating search seeds from sequence clusters when it comes to finding the increasingly rare riboswitch classes that power-law predictions[21] of riboswitch abundances suggest remain to be found. As the size of search databases increased, clustering-based approaches have required higher and higher thresholds for the number of initial sequences required to form a search seed. Without raising those thresholds, the sheer number of possible sequence clusters that can be formed from these larger databases makes the computationally expensive homology searches impossible. In short, the "signal-to-noise" ratio when using these clustering-based approaches is too poor to find rare riboswitch classes and an additional noise "filter" is required.

Computational search by GC/length filtering of IGRs

The noise "filter" this dissertation advances as a method to identify rare riboswitch classes relies on a few distinctive attributes of bacterial genomes. One way that bacterial genomes differ from other kingdoms of life is the high percentage of their genomes dedicated to protein-coding regions. On average, bacterial genomes are 88% protein-coding[69], and all the necessary ncRNAs required for a bacteria must be crammed into the remaining 12% of their genomes which consist of their IGRs. However, not all these IGRs have an equal probability of containing ncRNA motifs. Specifically, most bacterial IGRs are relatively short and do not contain enough space to contain a functional ncRNA motifs. Additionally, bacterial organisms whose genomes are enriched for A and T nucleotides will typically have their structured ncRNAs in IGRs with a higher than average GC content[70,71] due to the stronger base-pairing interactions formed between these two nucleotides. These two

characteristics, IGR length and GC content, form the basis of a powerful search strategy for identifying structured ncRNA motifs.

This GC-IGR search strategy was first piloted[72] in 2009 by analyzing the long, GC-rich IGRs for the organism *Pelagibacter ubique*. The plot of the IGRs from this organism (**Figure 1-3**) is a clear visual example of how this approach works. All of the IGRs containing known structured ncRNAs are clustered in the region of plot corresponding to long, GC-rich IGRs. Performing homology searches on the IGRs with no known ncRNAs revealed the presence of the SAM-V riboswitch class[73] in this organism. This same approach was later revived and converted into a systematic search strategy by analyzing the genomes of five additional bacterial genomes[74]. This revived approach sought not only to identify new riboswitch candidates, but to systematically categorize (**Figure 1-4**) every selected IGR with a hypothetical function whenever possible. This second trial of five genomes also proved successfully in identifying new riboswitches as demonstrated by the HMP-PP riboswitch class[30] that was validated from the list of candidates.

The work described in this dissertation represents both a continuation and an expansion of this GC-IGR search approach. The continuation will be described in Chapter 2 which describes the analysis of an additional 26 bacterial genomes that aided in the identification of over 150 new ncRNA motif candidates including 10 riboswitch candidates. One of these riboswitch candidates, the *pnuC* motif, has since been validated as the NAD-II riboswitch[25]. The expansion of the GC-IGR search approach is discussed in Chapter 3, which describes a new computational toolset called the Discovery of Intergenic Motifs PipeLine (DIMPL) that uses algorithms from machine learning to help identify genomic regions enriched for structured ncRNA and automates many of the subsequent

computational tasks required to predict a motif's function. Taken together, the work described in this dissertation demonstrates the strength of the GC-IGR search strategy for ncRNA motif discovery and provides a powerful toolset to make future efforts using this discovery approach faster and easier than ever before.

Tables and Figures

Figures

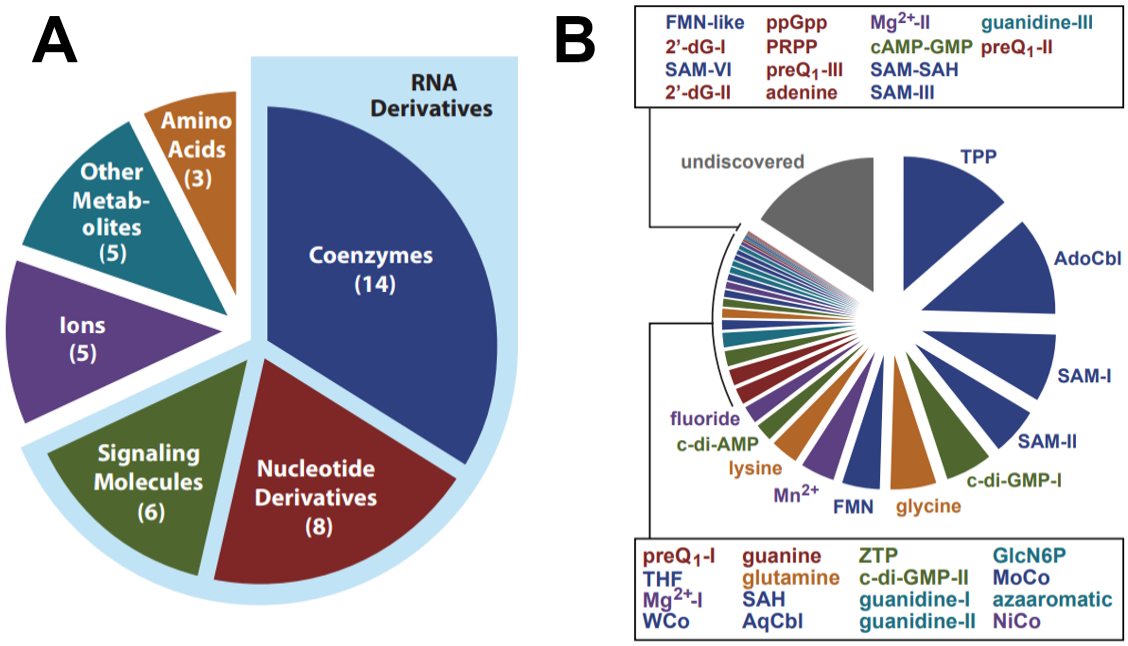

Figure 1-1: Riboswitch Ligands and Abundances

(A) Categorization of the different ligands with known riboswitch classes. (B) Abundances of riboswitch classes starting the most abundant class TPP (16,701 unique representatives). Number of representatives for each class was determined by searching NCBI's collection of reference genomes (RefSeq76)[75]. Number of undiscovered riboswitch class representatives is estimated using the trendline in **Figure 1-2B**. Figure adapted with updated data from McCown *et al*[21].

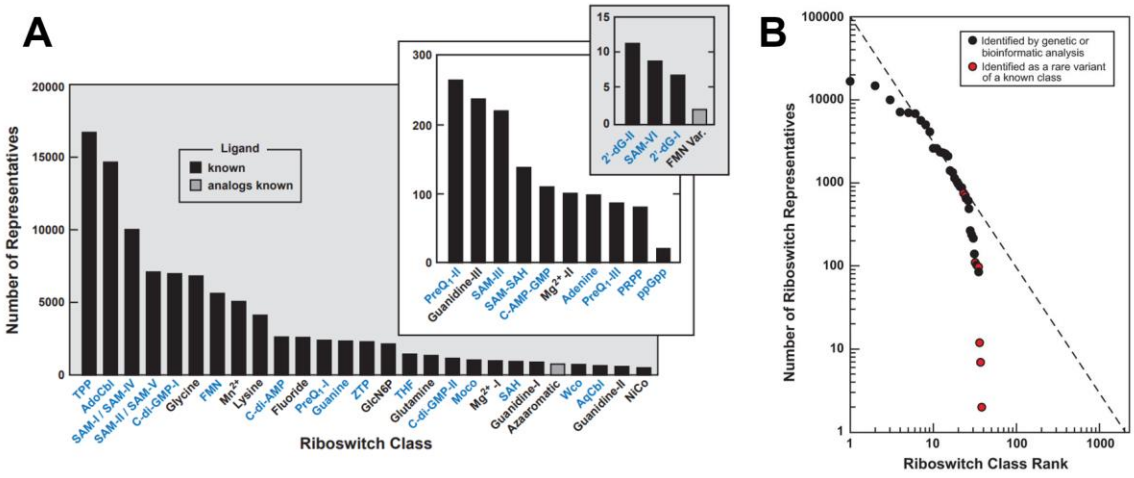

Figure 1-2: Riboswitch class abundances

(A) Number of representatives for each riboswitch class listed presented in rank order from the most abundant (TPP) to the least abundant (FMN Variant). (B) The same data presented in a log-log scale with a trendline predicting the total number of riboswitch classes with at least 1 representative (x -intercept) and the total number of undiscovered riboswitch representatives remaining (area under trendline). Figure adapted from McCown *et al*.[21]

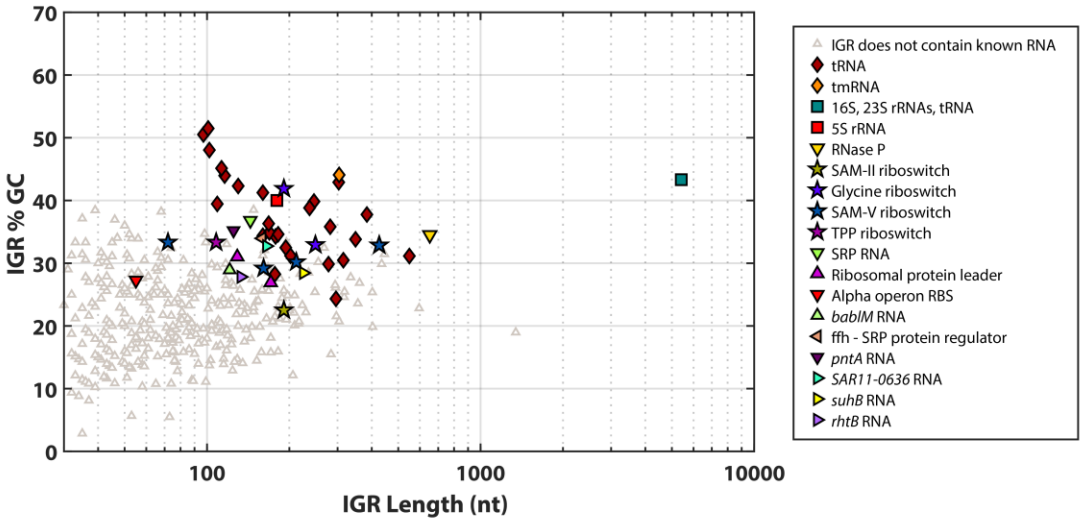

Figure 1-3: Plot of IGRs from *Pelagibacter ubique*

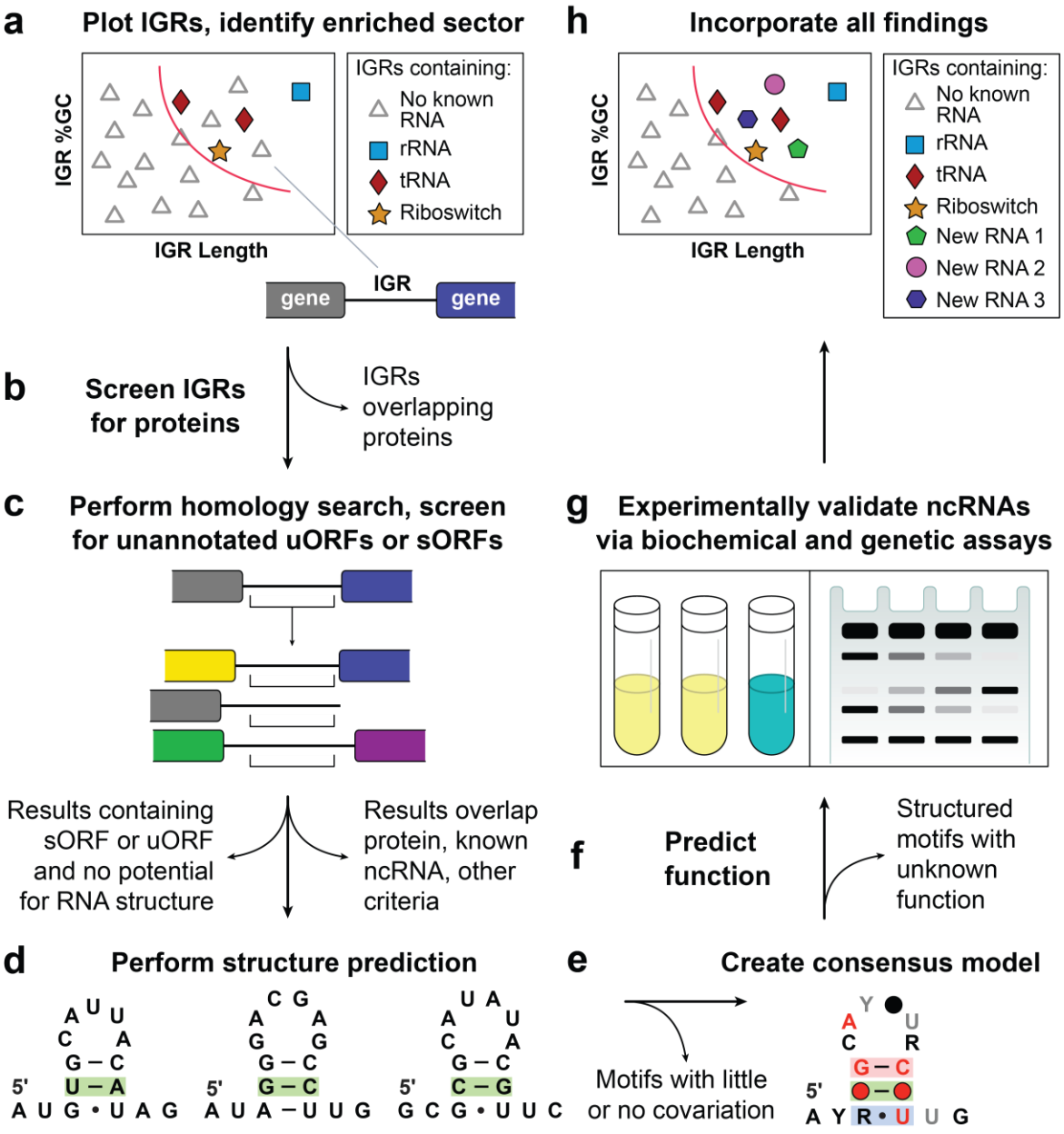

Figure 1-4: Overview of GC-IGR Computational Pipeline

(A) The IGRs from a bacterial genome are plotted by %GC content and length and IGRs containing known structured ncRNA motifs are labeled. IGRs with no known ncRNA motifs but similar length and GC content are selected for analysis. (B) Selected IGRs are first screened using BLAST[76] to ensure they don't actually contain unannotated open reading frames (ORFs). (C) A computational search using Infernal[63] is performed on each

IGR to search for homologous regions from other bacterial genomes. (D) The collection of homologous sequences are then used for sequence prediction. Covariation, as shown by the green shaded nucleotide pair, is particularly important in establishing structure (E) Example consensus model that identifies highly conserved nucleotides (in red) that may be of great importance for motif function. (F) Motifs are assigned a predicted function based on the structural information contained in the consensus model and the information from the prevailing gene associations. (G) Predicted motif functions can be validated experimentally. (H) Validated and predicted motif functions can be incorporated into annotation databases. Figure adapted from Stav *et al.*[74]

Chapter Two

Discovery of structured noncoding RNAs in 25 bacterial genomes

Largely adapted from the following manuscript:

Brewer, K. I., Greenlee, E. B., Higgs, G., Yu, D., Mirihana Arachchilage, G., Chen, D., King, N., White, N. & Breaker, R. R. Comprehensive discovery of novel structured noncoding RNAs in 26 bacterial genomes. *Manuscript In Preparation* (2021).

Summary

Comparative sequence analysis methods are highly effective for uncovering novel classes of structured noncoding RNAs (ncRNAs) from bacterial genomic DNA sequence datasets. Previously, we developed a computational pipeline to more comprehensively identify structured ncRNA representatives from individual bacterial genomes. This search process exploits the fact that genomic regions serving as templates for the transcription of structured RNAs tend to be present in longer than average noncoding 'intergenic regions' (IGRs) that are enriched in G and C nucleotides compared to the remainder of the genome.

In the present study, we apply this computational pipeline to identify structured ncRNA candidates from 26 diverse bacterial species. Numerous novel structured ncRNA motifs were discovered, including several riboswitch candidates, one whose ligand has been identified and others that have yet to be experimentally validated. Our findings support recent predictions that hundreds of novel riboswitch classes and other ncRNAs remain undiscovered among the limited number of bacterial species whose genomes have been completely sequenced.

Introduction

Each bacterial species carries genes expressing structured noncoding RNAs (ncRNAs) whose nucleotide sequences and folded shapes are critical for their biological and biochemical functions. These characteristics are most prominently manifested by molecules such as tRNAs and rRNAs that collaborate with mRNAs to catalyze the synthesis of all genetically encoded polypeptides[77]. However, other classes of structured ncRNAs are likewise essential for the survival of nearly all bacterial species. These include ncRNA classes such as (i) the RNase P ribozyme[7] whose site-specific phosphoester hydrolysis activity processes many precursor RNA transcripts, (ii) the signal recognition particle RNA[9], which is instrumental in routing certain proteins to cell membranes, and (iii) 6S RNA[11], which serves as a regulator of RNA polymerase activity. Members of another class of highly folded RNA molecules, called tmRNA[78], exploit both a complex tertiary structure and a protein coding region to simultaneously induce the release of ribosomes from broken mRNAs while tagging truncated proteins with a polypeptide sequence designating them for disposal. This list of fundamental biochemical and biological functions of bacterial ncRNA classes is made larger when including less widespread classes that also exhibit remarkable functions, including various RNA processing (ribozymes)[12,13] and gene control (riboswitches)[14–16] activities.

A focus of our research team in recent years has been the discovery of riboswitches that sense specific metabolites or inorganic ions to regulate the expression of genes whose protein products synthesize, transport, utilize, or respond to the ligands being monitored. Over 50 distinct classes of riboswitches have been experimentally validated to date[21–33], and many of these sense nucleotide-derived compounds that have been proposed to have

been present during an era of biology before the evolutionary emergence of proteins[18,79–82]. Thus, uncovering additional classes of riboswitches promises to reveal more about how modern cells regulate critical biochemical processes, and perhaps also to provide insight into the functions of ancient RNA-based regulatory systems[35,83].

It has been proposed[21,34,35] that many bacterial riboswitch classes remain hidden in the existing bacterial genomic datasets, and that these undiscovered representatives are likely to be rare compared to those from the most abundant classes that have already been experimentally validated. If true, we currently are in an era wherein many hundreds of sparsely represented riboswitch classes remain undiscovered in existing genomic DNA datasets. Unfortunately, this rarity creates a considerable barrier for those who seek to uncover novel riboswitch classes. Past computational search campaigns that led to the discovery of multiple novel riboswitch classes[49,56,57,60,61] typically have exploited comparative sequence analysis algorithms. These algorithms seek similarity in sequence and possible secondary structure features of one intergenic region (IGR) with that of all or a subset of other IGRs in the database under examination. These approaches have uncovered riboswitch classes that are abundant and phylogenetically diverse. However, they typically fail to identify rare riboswitch candidates, particularly when only few highly similar representatives are present in the database.

To overcome this problem, we developed a computational pipeline[72,74] that substantially improves the chances that rare riboswitch candidates and other structured ncRNAs will be uncovered. This “GC-IGR” pipeline takes advantage of the fact that structured ncRNAs are encoded in IGRs that have unique characteristics compared to most other IGRs in certain bacterial genomes. Specifically, organisms whose genomes are rich

in A and T nucleotides usually carry structured ncRNAs in IGRs that have higher than average GC content[70,71]. Also, most IGRs are relatively short unless they serve as templates for the production of ncRNA motifs. By sorting 'unknown' IGRs (those that are not known to carry a genetic element) based on both length and GC content, IGRs that carry previously undiscovered ncRNA motifs are likely to cluster with IGRs that are already known to serve as templates for structured ncRNAs. By analyzing the IGRs with these unique characteristics, we have been able to discover various novel classes of ncRNAs, such as riboswitches and other structured RNA and DNA motifs[72,74]. The detailed methodology employed in the GC-IGR computational pipeline has been demonstrated previously by examining the IGRs of five bacterial genomes[74]. This same methodology has now been employed herein to search for structured ncRNA motifs from an additional 26 bacterial genomes. We report the discovery of a variety of novel ncRNA motifs, which is consistent with earlier predictions[21,34,35] that numerous functional RNAs, including distinct candidate riboswitch classes, remain to be discovered in the genomes of many bacterial species.

Results and Discussion

Genomes and IGRs selected for comparative sequence analysis

In continuation of our campaign to discover novel ncRNA motifs by using the GC-IGR bioinformatics pipeline[72,74], we analyzed an additional 26 bacterial genomes. The genomes chosen for analysis were from the following phyla: Aquificae, Chlamydiae, Firmicutes, Fusobacteria, Proteobacteria, Spirochaetes, Tenericutes, Thermodesulfobacteria, and Thermotogae (**Table 2-1**). This collection of genomes expands the analyzed phyla from three in the initial work (Firmicutes, Proteobacteria, and Synergistetes) to nine.

Additionally, the classes of Proteobacteria were expanded from three to five. Within this collection of 26 bacterial genomes, there is substantial variability in the total number of IGRs found within a genome, in the average IGR length, and in the average IGR GC content (**Table 2-1** and **Figures 2-9 to 2-34**). This constitutes a broader distribution of IGR characteristics compared to those examined in the previous studies, which provides an opportunity to further evaluate the robustness of the GC-IGR pipeline.

Certain IGRs from these 26 were chosen for in-depth analysis in the same manner as previously described[74]. Briefly, the IGRs of each genomes genome were plotted using two parameters: the length of the IGR and its GC content. Genomes that exhibit substantial clustering of IGRs carrying known structured ncRNAs were favored for further analysis. IGRs lacking a known function, called "unknown IGRs", were selected for analysis if they cluster near to IGRs carrying known structured ncRNAs. Potential structural elements residing in unknown IGRs of interest were then sought. For example, the presence of conserved nucleotide sequences between two or more IGRs is suggestive of a conserved RNA architecture. Likewise, the presence of nucleotide covariation, such that a mutation at one position always occurs in conjunction with a mutation at a second site that retains base-pairing potential, is strongly indicative of a conserved secondary structure feature.

By identifying conserved sequences and secondary structures, we can focus additional attention on the unknown IGRs that carry novel candidate ncRNA motifs. However, some of the motifs identified in this manner represent previously discovered, but misannotated or unannotated ncRNAs, such as tRNAs, riboswitches, or other structured RNAs. Similarly, some conserved features reflect the presence of conserved open reading frames (ORFs), which are identified and classified based on bioinformatic evidence of their ability

to code for protein. IGRs that carry truly distinct motifs are tentatively named based on their apparent potential to function as a noncoding RNA, including riboswitch candidates whose functions might be predicted based on their gene associations and genomic orientations. Specifically, strong riboswitch candidates typically have robust sequence and structural conservation, a genomic orientation that consistently positions the motif upstream of its associated protein-coding genes, and gene associations that are indicative of regulation by a small molecule or inorganic ion ligand[21]. Importantly, each proposed function for a novel motif should be considered highly speculative, and meant only to create a preliminary organization of the findings. In each instance, biochemical or genetic validation experiments are subsequently needed to confirm the proposed functions.

Classification of newly discovered nucleic acid motifs

In a previous study[74], we defined five general categories into which we classified all IGRs examined by implementing the GC-IGR pipeline. This exercise is not intended to serve as an error-free evaluation of each motif, but rather to provide a preliminary assessment that might help researchers make decisions regarding future experimental analyses. For the current study we follow these same guidelines to make predictions regarding the possible functions of novel motifs, which are repeated below.

(i) Unnamed: Insufficient evidence to classify.

(ii) Low-ranking candidate (LRC): Typically fewer than 5 unique representatives and a poor consensus model.

(iii) Medium-ranking candidate (MRC): Typically fewer than 20 unique representatives and/or a poor consensus model.

(iv) High-ranking candidate (HRC): Many representatives and a good consensus model, but insufficient information regarding possible function.

(v) Named candidate: Could be rare, but usually has many representatives with a good consensus model and some evidence supporting a hypothesis for function.

Some sequences annotated as IGRs in the various DNA sequence databases actually function as ORFs for ordinary-sized proteins, or serve as templates for the translation of known ncRNAs. These misannotated IGRs are removed from our candidate list before sorting the IGRs by the criteria just described. Note that category (v) includes any motifs where we judged that sufficient information was evident to make a prediction of the motif's function, including riboswitches.

Strong riboswitch candidates

Riboswitches are one of many types of ncRNAs that are expected to be uncovered by the GC-IGR pipeline, and these will contribute to the "named candidate" category as defined above. They are also one of the more straightforward candidate ncRNA types to generate hypotheses regarding their functions. Many of the known riboswitch classes exhibit characteristics that are conspicuous due to the necessities of forming metabolite-sensing (aptamer) and gene-control (expression platform) domains. In the current study, we have arranged the riboswitch candidates into strong and weak groups to give readers a general indication of the merits of each motif. Strong riboswitches candidates typically exhibit extensive sequence and structural conservation, associate with genes for related biochemical or biological processes, and are oriented in a manner that is consistent with a gene control function, among other attributes.

However, some newly found ncRNA motifs imperfectly reflect these characteristics, and therefore we consider them weak riboswitch candidates. Regardless, we recommend that researchers who wish to pursue the experimental validation of even the strong riboswitch candidates presented herein should proceed with appropriate skepticism regarding our preliminary classifications. With these caveats in mind, below we describe the characteristics of the six motifs we judged to be strong riboswitch candidates (SRCs) on the basis of their structural characteristics and gene associations.

The *pnuC* (SRC-10-1) motif

Approximately 140 distinct representatives of a conserved motif were identified located almost exclusively upstream of and in the same orientation as *pnuC* genes in species within the genus *Streptococcus*. Published evidence indicates that some PnuC proteins transport nicotinamide riboside (NR)[84–86], which is a component of the ubiquitous coenzyme nicotinamide adenine dinucleotide (NAD^+). Given that numerous riboswitch classes respond to nucleotide-based coenzymes[21], and that a recently reported riboswitch class was found to function with NAD^+[32,87], the *pnuC* motif RNA was immediately considered a promising riboswitch candidate.

Additional bioinformatic and biochemical analyses[25] identified a central core of the motif that is required for its function. These efforts trimmed the collection of non-redundant representatives to 43, which were used to create a consensus sequence and structural model (**Fig. 2-1A**). Each *pnuC* motif RNA appears to carry highly conserved nucleotides within an internal bulge formed by two stems. The RNA also has the potential to form a pseudoknot with the purine-rich ribosome binding site (RBS)[88] of the adjacent coding region, suggesting that binding of a ligand related to NAD^+ metabolism might block

ribosome binding to inhibit translation initiation[89]. This function would be analogous to that of a previously discovered riboswitch class for NAD^+[32]. These characteristics solidified our preliminary classification of the *pnuC* motif as a strong riboswitch candidate. Indeed, binding assays confirmed that *pnuC* motif RNAs selectively bind various nicotinamide derivatives, including NR and NAD^+[25]. Thus, each *pnuC* motif RNA serves as the aptamer for a second riboswitch class for this coenzyme, which we have named the NAD^+-II riboswitch as published elsewhere[25].

The *icd*-II (SRC-12-2) motif

The *icd*-II motif (**Fig. 2-1B**) is represented by only 25 unique examples that are found mostly in the *Polynucleobacter* genus and metagenomic environmental sequences. Although originally found in *Polynucleobacter necessarius*, the same motif was also found through a search starting from one of the IGRs (SRC-28-2) of *Beta proteobacterium CB*, which has some degree of genetic similarity to members of the *Polynucleobacter* genus. This motif is located upstream of *icd* gene, which codes for an $NADP^+$-dependent isocitrate dehydrogenase (IDH) enzyme[90]. IDH is an important enzyme of the citric acid cycle, and it participates in managing the carbon flux through this energy metabolism pathway by supplying the cell with 2-oxoglutarate and NADPH[91]. The proposed secondary structure model consists of a three-stem junction[92], wherein the RBS for the adjacent ORF is predicted to participate in forming the first base-paired stem (pairing element 1, or "P1"). Thus, the *icd*-II motif appears to form an architecture that can regulate translation initiation of its associated ORF. However, the left shou1der of P1 is similar to a -10 region of an RNA polymerase promoter sequence[93], and therefore we currently cannot be certain of the nucleotides that form the complete motif.

The *icd*-II motif is the second riboswitch candidate that has been discovered to associate with *icd* genes. A previously reported motif, called *icd*[60,65], is also found in species of Proteobacteria, although the representatives of these two motifs are present in different classes of this phylum. Both motifs are predicted to form a three-stem junction, but they differ both in their conserved sequences and the number of base-paired stems. Potential ligands for the *icd* motif are also potential ligands for the *icd*-II motif, such as citrate, oxaloacetate, glyoxylate, 2-oxoglutarate, or related compounds.

The *carA* (SRC-12-1) motif

The *carA* motif (**Fig. 2-1C**) has 30 unique representatives, each located immediately upstream of a *carA* gene. Again, although originally found in *Polynucleobacter necessarius*, it was also found in a search starting from one of the IGRs (SRC-28-1) of *Beta proteobacterium CB*. The gene *carA* codes for the small subunit of carbamoyl phosphate synthase, which is an enzyme that catalyzes the first committed step in pyrimidine and arginine biosynthesis[94]. The *carA* motif is found in β-proteobacteria, mostly in the *polynucleobacter* genus, along with metagenomic environmental datasets. The proposed secondary structure consists of an extended imperfect hairpin, although the sparse representation provides little covariation support for this structure.

In *Escherichia coli*, the *carAB* operon is regulated by transcription factors binding to two promoters directly upstream of *carA*[95]. However, in genomes that contain the *carA* motif, there appears to be a promoter located upstream of this structure, but there does not appear to be a second promoter. Thus, it seems possible that this motif functions as a riboswitch, and that it substitutes for the function of a transcription factor system in organisms like *E. coli*. Notably, the conserved sequence and structural features are located

immediately upstream of the predicted RBS and start codon for the adjoining ORF. This architecture is consistent with a possible *cis*-regulatory function where ligand binding regulates translation initiation. Carbamoyl phosphate or endpoints of biosynthetic pathways using this compound are potential ligands for this riboswitch candidate.

The *gltD* (SRC-18-1) motif

There are 95 distinct representatives of the *gltD* motif (**Fig. 2-1D**), although the majority of these are from bacterial DNA sequence information derived from metagenomic environmental samples that are not assigned to specific species. Representatives in sequenced organisms are exclusively derived from species of the Selenomonadales and the Veillonellales orders. This motif is frequently preceded by a predicted RNA polymerase promoter sequence, is always directly upstream of a predicted *gltD* gene, and is followed by a possible RNA hairpin structure that appears to occlude the RBS of the associated gene. All of these features are consistent with an RNA motif that has a *cis*-regulatory function. The *gltD* gene encodes glutamate synthase which catalyzes the conversion of L-glutamine and 2-oxoglutarate into two molecules of L-glutamate[96]. A possible ligand for this riboswitch candidate might be related to either of the amino acids in this pathway. Intriguingly, a known riboswitch class for glutamine[97–99] is occasionally associated with glutamate synthase genes, but representatives of that previously discovered riboswitch class are not present in the same species that carry the *gltD* motif. This lack of phylogenetic overlap strengthens the possibility that glutamine may be the ligand for the *gltD* riboswitch candidate.

The *ldh2* (SRC-18-2) motif

The *ldh2* motif (**Fig. 2-1E**) has 34 unique representatives found in the *Negativicutes* class and in various environmental DNA samples. It is almost exclusively found in front of the genes associated with the *Ldh_2* conserved protein domain family whose representatives encode malate/L-lactate dehydrogenases[100]. The structure of this motif includes a three-stem junction containing highly conserved nucleotides and a long P3 stem with a central bulge that may cause it to kink back to form a binding pocket. A possible expression platform is present in the form of an RBS that may be hidden by formation of an additional P4 stem. It is difficult to define a short list of candidate ligands because of the uncertainty of the substrate and product of the associated gene product.

The *proX* (SRC-29-1) motif

The *proX* motif (**Fig. 2-1F**) has 17 unique representatives found primarily in the *Spiribacter* genus and in various environmental DNA samples from saltwater origin. It is found almost exclusively in front of the gene *proX*, which encodes a predicted glycine betaine/proline betaine transporter[101]. A highly conserved RBS that can be occluded by a corresponding sequence on the P2 stem provides a mechanism for plausible translational control. Given the importance of glycine betaine and proline betaine for osmoregulation of saltwater organisms, we speculate that this riboswitch candidate might respond to changes in the concentration shifts of one of these two molecules, or other osmoprotectants. Alternatively, the signaling molecule c-di-AMP is frequently used to signal osmotic distress[102–104], and a common riboswitch class for this compound has been reported previously[52].

Weak riboswitch candidates

The following four riboswitch candidates described below are generally rare, and may be associated with genes whose functions are not readily apparent. In some cases we cannot provide a strong hypothesis for the possible metabolite or inorganic ion ligand that might be sensed by the candidate. Nevertheless, we judged them to be supported by sufficient structural and gene context evidence to classify them as weak riboswitch candidates (WRCs) and worthy of bringing to the attention of the research community. We advise those interested in pursuing the experimental validation of the motifs below to first pursue additional bioinformatics studies to strengthen their hypotheses.

The *ilvD* (WRC-18-2) motif

There are 73 distinct representatives of the *ilvD* motif (**Fig. 2-2A**) found exclusively in the *Veillonella* genus of the Firmicutes phylum. The associated gene for all representatives is annotated as *ilvD*, which is predicted to encode a dihydroxyacid dehydratase enzyme that is a key contributor to the branched-chain amino acid biosynthesis pathway[105]. Some representatives have a potential intrinsic transcription terminator stem[106–108] (not depicted), which is positioned between the conserved, putative aptamer region and its associated ORF located downstream. This is a common characteristic of riboswitches that operate using an expression platform that permits ligand-mediated transcription termination[109,110].

Unfortunately, we cannot be certain that the *ilvD* motif functions as a riboswitch for several reasons. For example, the motif includes base-paired stems that are uncharacteristically long compared to most validated riboswitch classes. Furthermore, there is little evidence for covariation, which reduces our confidence in the proposed secondary structure. However, this latter weakness is not unexpected for a rare motif that

lacks phylogenetic diversity among its representatives. In total, the characteristics of the *ilvD* motif suggest it likely acts as a cis-regulatory element, although it could possibly serve as a small RNA (sRNA)[111,112] that regulates genes related to the biochemistry promoted by the *ilvD* gene product. We also cannot yet rule out the possibility that the *ilvD* motif functions as a protein-binding regulatory motif. However, we do not observe evidence typical of certain protein-binding nucleic acids, such as palindromic or short repetitive sequences.

Given the various points above, we have categorized the *ilvD* motif as a weak riboswitch candidate. Ligand candidates could be drawn from the list of compounds related to branched-chain amino acid biosynthesis pathways. This list also should include the bacterial signaling molecule ppGpp[113], which is sensed by a previously discovered riboswitch class that broadly controls branched-chain amino acid biosynthesis genes[27].

The *ilvB* (WRC-14-1) motif

There are only 7 unique representatives of the *ilvB* motif (**Fig. 2-2B**) and they are found exclusively in various species of the *Leptospira* genus. Due to the very low number of representatives, the consensus structure for this motif consists mostly of highly conserved nucleotides. This limites the number of nucleotides that may be expected to show covariation in stems. This motif is nevertheless notable because it is always found in front of the gene *ilvB*, which is another member of the branched chain amino acid biosynthesis pathway discussed above for the *ilvD* motif. Similarly, the list of ligand candidates can be drawn from compounds related to branch-chain amino acid biosynthesis or the signaling molecule ppGpp.

The *sucC-II* (WRC-14-2) motif

There are only six representatives of the *sucC* motif (**Fig. 2-2C**) and they are also found exclusively in various species from the *Leptospira* genus. Despite the small number of hits there is some evidence of covariation supporting the formation of the P1 and P3 stems. These motifs are exclusively found in front of the *sucC* gene which encodes the succinyl-CoA synthetase beta subunit[114], which suggests a wide range of possible ligand candidates such as members of the citric acid cycle as well as a variety of secondary signaling molecules involved in maintaining cellular energy homeostasis.

The *potE* (WRC-18-1) motif

There are 144 representatives of the *potE* motif (**Fig. 2-2D**) and they are found exclusively in members of the *Veillonella* genus and environmental sequence samples. They are almost always found in front of the gene *potE* which encodes a putrescine-ornithine antiporter[115]. Unfortunately, the motif's secondary structure model is not well supported by covariation and it lacks an obvious expression platform. In addition, the motif is generally located some 200 nt away from the downstream *potE* gene. However, the compelling gene association, possibly related to polyamine production[116], and the fact that a riboswitch for adenosylcobalamin (AdoCbl)[39] is almost always located in the IGR immediately downstream of *potE* causes us to consider this motif a weak riboswitch candidate.

An unusual gene control element incorporating the *odc1* motif

A particularly unusual gene control candidate called the *odc1* (uORF-25-1) motif (**Fig. 2-3A**) was uncovered using the GC-IGR pipeline. This motif, which is well represented with 272 examples from species of the α-proteobacteria class, is predicted to form a hairpin with

an extended stem structure (P1) closed by a relatively large loop. A pseudoknot appears to form between nucleotides near the 5′ terminus of the motif and nucleotides in the loop. The RNA domain is named for the most commonly annotated gene associated with the motif (**Fig. 2-3B**), which presumably codes for the enzyme ornithine decarboxylase. This enzyme converts ornithine into putrescine, which is the committed step to produce polyamines[117].

A predicted RBS forms part of the right shoulder of P1, and therefore is sequestered within the predominant secondary structure element depicted in the consensus model (**Fig. 2-3A**). Also included among the many highly conserved nucleotides that are characteristic of this motif is a predicted noncanonical UUG start codon for the main ORF. Intriguingly, another possible UUG start codon is always located immediately upstream of the main start codon, wherein only two nucleotides separate these two codons. Thus, the architecture of this motif indicates that the *odc1* motif likely has *cis*-regulatory function that controls the site of translation initiation.

Notably, the start codon for the upstream open reading frame (uORF) is predicted to initiate synthesis of an 11 amino acid peptide wherein most amino acid positions are highly conserved (**Fig. 2-3C**). There is no evidence for variation of peptide length with any of the *odc1* motif representatives. These observations suggest that there is a functional role for this short peptide product. However, the well-conserved sequence and structural features of the complete *odc1* motif, largely encompassing nucleotides located upstream of the conserved uORF, suggests that the RNA might serve a larger role in choosing between the use of the two start codons during translation initiation. One possibility is that the motif acts as a metabolite-binding riboswitch. Perhaps, in the presence of ligand, the RBS is

sequestered by formation of the P1 stem to suppress expression from both start codons. In the absence of ligand, an alternative structure might favor one or the other possible start codons, which would be utilized to produce either the short peptide from the uORF or the larger protein from the main ORF.

To investigate the general mechanistic features of the *odc1* motif, a reporter construct was created wherein the representative from the gram-positive bacterium *Sphingomonas echinoides* (**Fig. 2-4A**) was fused upstream of a β-galactosidase reporter gene. Two specific reporter fusion arrangements were prepared (**Fig. 2-4B**) to position the reporter gene either in the same reading frame as the main ORF by using the second start codon at positions 80 to 82, or in the same reading frame as the uORF by using the first start codon at nucleotide positions 75 to 77. In addition, certain mutant constructs (called M1 through M6, **Fig. 2-4A**) were made to examine the importance of highly conserved sequences and structural features of the *odc1* motif.

E. coli cells were used as a surrogate host organism to assess the function of the reporter constructs in vivo, which revealed several notable findings. For example, fusion of the reporter gene to either the main ORF or the uORF start codon of the wild-type (WT) *S. echinoides odc1* motif yields measurable but modest levels of expression (**Fig. 2-4C**). Expression from the uORF start codon is approximately 3-fold greater than that observed for the main ORF start codon, regardless of whether rich or minimal medium is used. Notably, both ORFs exhibit higher gene expression levels in minimal media (M9) than in rich media (LB). The modest amount of gene expression from either reporter construct is largely eliminated when a disruptive mutation is made to its corresponding start codon (M1 through M3, **Fig. 2-4C**), which confirms that protein synthesis is being driven by the

predicted sites for translation initiation. Importantly, these results reveal that both start codons can be utilized by the mRNA, indicating that the proposed 11 amino acid peptide (**Fig. 2-4C**) is produced from the start codon associated with the uORF.

Next, we examined whether the motif serves as a genetic control element and whether highly conserved nucleotides near the pseudoknot are important for this function. Mutant constructs M4 through M6 were prepared that carry changes to strictly conserved positions immediately preceding the nucleotides predicted to form part of the pseudoknot (**Fig. 2-4A**). All three mutations caused a loss of expression of the reporter gene when fused in-frame with the main ORF start codon. In contrast, mutations present in M4 and M5 had little effect on reporter gene expression when initiated by the start codon of the uORF. Again, although the level of gene expression is modest, these results suggest that certain conserved nucleotides in the *odc1* motif are important for regulating translation, particularly for the main ORF. However, additional experiments will be needed to determine how these nucleotides participate in affecting the level of gene expression from the adjacent start codons.

Additional notable motifs identified with the GC-IGR pipeline

The GC-IGR bioinformatics pipeline yielded numerous additional motifs that we have not included on our list of candidate riboswitch classes. Some motifs are rare, and thus lack sufficient clues for us to responsibly speculate on their possible functions. However, some candidates have properties that are suggestive of functions that are different than gene control via riboswitch action. Several of the most intriguing candidates are briefly described below.

The PBC-28-3 motif

A total of 21 unique examples of the PBC-28-3 motif (**Fig. 2-5A**) are found upstream of genes that encode proteins closely related to AhpC, a peroxiredoxin, which forms an important component of the bacterial defense system against toxic peroxides[118]. The motif appears to be cis-regulatory because it is almost always oriented in the same direction as the downstream gene. Also, the start codon for the adjacent ORF is located immediately next to the motif, suggesting that regulation of translation might be its function. The PBC-28-3 motif RNA consists of a long covarying stem with a highly conserved loop. This loop includes regions of pyrimidine nucleotides, which suggests a possible role associated with protein factors that prefer polypyrimidine binding sites. This characteristic, in addition to the long base-paired substructures, suggests this RNA might function as a protein binding motif rather than as a riboswitch that binds a small molecule ligand.

The HRC-8-4 motif

The consensus sequence and secondary structure model of the HRC-8-4 motif (**Fig. 2-5B**) was determined from 160 unique representatives exclusively uncovered from species of the Rhodobacterales family of the α-proteobacteria class. The motif is primarily located in between the *ccoNOQP* and *ccoGHIS* operons that both code for genes necessary for the maturation of cbb3-type cytochrome c oxidase complex[119]. Usually the *ccoP* gene is found immediately upstream and the *ccoG* gene immediately downstream, but in some instances there are other genes adjacent to the motif between the two cytochrome c oxidase operons. The most commonly inserted gene is *hvrB* that encodes a LysR-family transcriptional regulator[120]. When the *hvrB* gene is present, it is always located immediately downstream

of the *ccoG* motif. There are also some arrangements where the motif is upstream of the *hvrB* gene when it is not adjacent to cytochrome c operons.

Structurally, the *ccoG* motif vaguely resembles the recently validated HMP-PP[30] and guanidine-IV[23,24] riboswitch classes. Like these experimentally validated riboswitch classes, the *ccoG* motif generally conforms to the classic architecture of an intrinsic terminator stem[106–108], which includes a strong stem followed by a run of U nucleotides. However, they also carry well conserved nucleotides at the tip of the stem-loop structure, which is unusual for terminator stems. HMP-PP and guanidine-IV riboswitches appear to exploit these conserved nucleotides to form the aptamers for their target ligands. This precludes the formation of the terminator stem when ligands bind and thus they function as a genetic 'ON' switches[21,23,24]. We speculate that the *ccoG* might perform a similar function, although we again cannot be certain that the ligand is a small molecule. The frequent association of the *hvrB* gene might indicate that this nucleic acid binding protein might bind to this terminator stem to regulate its own expression.

The HRC-28-1 motif

The HRC-28-1 motif (**Fig. 2-5C**) is represented by 33 unique examples mostly found in the species of *polynucleobacter*. The most commonly associated upstream gene encodes a provisionally annotated oxidative damage protection protein that is a member of the CDD family PRK05408[121]. The most commonly associated downstream gene is the ribose-5-phosphate isomerase *RpiA*[122], which is always oriented in the opposite direction as the HRC-28-1 motif. The proposed secondary structure model for the motif includes six stems, where the final base-paired substructure appears to be an intrinsic terminator. The motif is unlikely to be a riboswitch due to its orientation relative to adjacent genes, but rather may

be a structured RNA element located in the 3′-UTR of the mRNA encoded by the upstream gene.

The uORFC-8-2 (*cysS)* motif

There are 145 unique examples of the uORFC-8-2 motif (**Fig. 2-5D**) that are found exclusively upstream of the *cysS* gene of the Rhodobacterales family in the α-proteobacteria class. The *cysS* gene encodes a cysteinyl-tRNA synthetase, and thus it is not surprising that this motif has characteristics expected for a ribosome-mediated attenuation sequence[123]. Specifically, we note the presence of a conserved AUG start codon for a uORF, wherein two highly conserved cysteine codons appear in tandem usually at positions 14 and 15 of the resulting short polypeptide (**Fig. 2-5D**, orange shading). The UGC codon for cysteine dominates, but there are examples of the UGU codon as well.

Structurally, the motif forms two prominent stems with the first containing the cysteine codons. In some cases there is the possibility that a third base-paired region can form to create a three stem junction. However, these cases are rare and lack evidence for covariation. In all cases, the consensus model includes a run of U nucleotides at the 3′ terminus, which is consistent with the formation of an intrinsic terminator stem. Presumably, ribosome stalling at the cysteine codons due to inadequate levels of aminoacylated $tRNA^{Cys}$ permits transcriptional read-through and production of the mRNA encoding the CysS protein, which generates more cysteinyl-$tRNA^{Cys}$.

The uORFC-6-1 (*hflX*) motif.

The uORFC-6-1 motif (**Fig. 2-5E**) is a uORF candidate with 51 unique representatives predominantly found in the Bacillales order of Firmicutes. It is primarily found upstream of the gene *hflX*, which encodes an RNA helicase involved in rescuing stalled and heat-

damaged ribosomes[124]. The RNA secondary structure of this motif is entirely located downstream of the predicted start codon for the uORF and is supported by strong evidence of covariation. In addition, the peptide sequence has a conserved Arg-Leu-Arg motif (**Fig. 2-5E**, orange shading), which has been found previously in antibiotic resistance leader peptides that rely on antibiotic-induced ribosome stalling to turn on expression of downstream genes[125]. The presence of this ORF within the predicted RNA structure suggests a similar mechanism whereby the presence of stalled ribosomes on the transcript at the uORFC-6-1 motif might influence transcription of the downstream main ORF coding for the HflX protein.

The uORFC-13-1 (*aroF*) motif

The uORFC-13-1 (*aroF*) motif is found only in species from the genus *Acholeplasma* and from environmental sequences where the organisms are unknown. There are 24 unique sequence representatives, which are located exclusively upstream of the *aroF* gene encoding the enzyme 3-deoxy-7-phosphoheptulonate synthase. This enzyme produces the first compound in the shikimate pathway that leads into the biosynthesis of the amino acids phenylalanine, tyrosine, and tryptophan. The motif is located immediately upstream of an apparent intrinsic transcription terminator stem, suggesting this motif regulates transcription termination. Features of the consensus model (**Fig. 2-5C**) reflect its likely function as a uORF, including possible -10 and -35 promoter regions, an RBS and start codon, and a run of U nucleotides that code for at least one phenylalanine residue. These features suggest that gene regulation might involve the speed of phenylalanine incorporation into a peptide encode by the uORF, in a process that might control the formation of the adjacent terminator stem. Comparative sequence analysis also supports

the formation of another base-paired substructure, and therefore the motif might have additional functional features. A similar attenuation based mechanism for the control of aroF in *E.coli* has been previously hypothesized[126], although the uORFC-13-1 motif differs from the 5' UTR found in that organism.

The OTH-22-1 motif

The OTH-22-1 motif (**Fig. 2-5F**) has 1887 unique representatives found in several bacterial phyla as well as in Archaea and environmental sequences. The RYYYAAC consensus sequence, found at the 5′ terminus of the motif is also a distinguishing characteristic of *attC* structured DNA elements[127]. Because this motif contains the same consensus sequence with comparable secondary structure, this motif is likely another variant of this type of structured DNA element. However, none of the representatives of this motif overlap with annotations of known classes of *attC* sites, indicating that this is likely a novel form of these single-stranded DNA motifs.

Comparison with transcriptomics-based approaches

The inclusion of *L. monocytogenes* among the collection of 26 genomes analyzed in the current study provided the opportunity to contrast the results from our computational approach (GC-IGR search)[74] with the transcriptomics-based experimental approach (term-seq)[128]. The term-seq approach generates sequencing reads on a genome-wide scale that establish the natural 3′ termini of RNA transcripts. This data can then be examined to detect signatures of regulatory RNAs such as riboswitches. For example, a riboswitch that terminates transcription immediately upstream of its associated mRNA ORF region will exhibit a higher abundance of sequencing reads spanning its aptamer region compared to

those for the downstream ORF. Indeed, term-seq data yielded such signatures for 28 previously known riboregulators in *L. monocytogenes*, including 13 riboswitches, 12 T-box leader sequences, and three protein-binding leader sequences. Also, the authors concluded that 12 additional novel regulatory RNA regions are detected by the term-seq approach, although 10 of these 12 novel regulatory regions had been previously identified as expressed sRNAs[129–131].

Interestingly, none of the 12 proposed novel regulatory RNAs from term-seq passed through our initial IGR filtering process. Specifically, the IGRs corresponding to eight of the riboregulator candidates were excluded from analysis because they were already annotated in Rfam[132], and thus are not considered unknown IGRs. The remaining four, including two previously identified sRNAs whose annotations were not present in Rfam, were excluded from the latter stages of our computational pipeline because of insufficient IGR length or GC content. In other words, they do not exhibit the characteristics typical of most other structured bacterial ncRNAs such as riboswitches.

To further assess these 12 term-seq candidates, we subjected them to comparative sequence analysis in search of conserved nucleotide sequences or secondary structure features. A more detailed description of each of these motifs is included in the supplementary information (**Fig. 2-7 and 2-8**). None of the 12 IGRs exhibited the combination of conserved nucleotides, structural complexity, and a genomic context theme that are typical of strong riboswitch candidates or other RNAs that rely on sophisticated secondary or tertiary structures for their function. Rather, the IGRs mostly appear to carry an intrinsic terminator stem[106,107] without evidence for complex structure formation beyond this genetic element. These RNAs might carry the regulatory equivalent of a riboswitch

expression platform, but they appear to lack the accompanying aptamer domain that is needed to selectively bind the target ligand to trigger changes in gene expression. These findings demonstrate that the GC-IGR search strategy can be used to uncover ncRNAs in a genome-wide manner, wherein the resulting candidate list is enriched for RNAs that exhibit complex conserved sequences and secondary structures.

Summary of the motifs discovered and implications for future searches

The GC-IGR analysis of 26 bacterial genomes has led to the discovery of a diverse collection of novel structured ncRNA motifs and other genetic elements (**Fig. 2-6)**. Approximately half of all IGRs analyzed in detail were found to be previously misannotated, and actually carry coding regions for known proteins or serve as transcription templates for known classes of ncRNAs. This extent of misannotation is similar to that observed with our previous effort to examine five bacterial genomes[74]. Perhaps as automated annotation algorithms improve, these inappropriately classified IGRs will be reduced in number, which would help accelerate genome analysis using the GC-IGR pipeline.

Another finding that is consistent with our previous study is that a large portion of IGRs that cluster with those carrying known ncRNAs do not exhibit sequence or structural homology with other IGRs, sometimes even from closely related species. Perhaps additional representatives for these rare IGR types will be found as the bacterial genomic DNA sequence databases expand. However, some of these IGRs could represent exceedingly rare ncRNA classes, which are predicted to exist based on the distribution of abundances for known riboswitch classes[21]. Alternatively, some of these IGRs might have

no function derived from their sequences and structures, and thus truly represent junk DNA sequences that have no value to the cell.

Despite the large number of misannotated and unassigned IGR functions, a variety of novel regulatory RNA candidates were uncovered by the GC-IGR pipeline. Approximately 13% of all analyzed IGRs are predicted to carry regulatory RNAs. Of particular interest to us are riboswitches, and this study has revealed the existence of at least 10 reasonable candidates (**Fig. 2-1** and **Fig. 2-2**). Indeed, one of these, initially called the *pnuC* motif (**Fig. 1A**), has proven to function as a riboswitch for NAD^{+}[25]. Many additional candidates also have characteristics suggestive of gene control functions, but additional experiments will be necessary to establish the true functions of these conserved ncRNA structures. However, if this collection of riboswitch candidates indeed yields several validated riboswitch classes, these findings would be consistent with our prediction that many more regulatory RNAs such as riboswitches remain hidden among the sequenced bacterial genomes[21,34,35]. Given that many thousands of bacterial genomes have the properties needed for successful application of the GC-IGR pipeline, we believe that many hundreds of classes of structured ncRNA candidates could be uncovered just among the bacterial genomes that have been sequenced to date.

Concluding remarks

In the current study, we have expanded the number of phyla represented by the species subjected to the GC-IGR search pipeline to nine, from the three phyla sampled in our previous study[74]. In addition to the new RNA discoveries, our findings again demonstrate the utility of the GC-IGR pipeline. This computational system offers a means to efficiently

identify nearly all ncRNA motifs in a given genome by examining unknown IGRs that cluster near known ncRNA representatives based on length and GC content characteristics. The GC-IGR analysis can be applied to multiple different genomes without the need for experimental manipulation, which permits the large-scale analysis of many different genomes. Indeed, the GC-IGR pipeline has advantages over transcriptomics-based methods (e.g. term-seq[128]) in several ways. For example, a bioinformatics approach only requires a sequenced genome, and does not require the ability to culture the species of interest to isolate and prepare RNA transcripts for subsequent analysis. In addition, the GC-IGR method can identify structured ncRNA motifs regardless of whether they are transcribed only under certain growth conditions.

The findings from the 26 genomes reported in this study, along with the results from the analysis of five genomes previously[74], demonstrate that a great diversity of novel ncRNA motifs, and even some structured single-stranded DNA elements, can be discovered by applying a more exhaustive bioinformatic analysis of sequenced bacterial genomes. Through the current study, along with our pilot applications of the GC-IGR pipeline[72,74], we conclude that it is possible to scale-up the application of the search method. However, there are certain barriers that must be overcome to discover candidate regulatory RNAs and other ncRNA motifs that are of most interest to RNA or microbiology researchers. Over half of the IGRs that were analyzed in the current project eventually could be ruled out as novel structured ncRNA candidates (**Fig. 2-6**). Substantial time is required to manually improve the genome sequence annotations to remove IGRs that code for known proteins or that function as templates for the transcription of known ncRNAs. Improvements to the pipeline that reduce the required manual contributions to the search

process will speed the analytical process. However, it is already practical to apply the GC-IGR pipeline to hundreds or even thousands of bacterial genomes. If implemented, such searches would undoubtedly yield many novel ncRNA classes.

Although we cannot precisely define the size of the undiscovered pool of ncRNAs in the bacterial domain of life, this collection must be vast given the great number and diversity of bacterial species on the planet, along with their propensity to use structured ncRNAs to achieve various tasks. Hundreds of additional riboswitch classes are predicted[21,34,35] to await discovery and validation among the bacterial genomes whose sequences are already available, although the vast majority of these classes are expected to be rare and therefore only narrowly distributed. Rarer riboswitch classes will pose challenges for discovery by search strategies that rely on comparative sequence analysis algorithms. However, the abundance of sequenced bacterial genomes coupled with the application of search strategies such as the GC-IGR pipeline should continue to yield numerous candidate riboswitch classes and many other structured ncRNA motifs. Thus, despite the conclusion that numerous structured ncRNA classes remain to be discovered, the rarity of each individual motif means that the genome of each bacterial species is likely to carry very few if any novel ncRNA classes.

The genomes that produced the most riboswitch candidates when subjected to the GC-IGR analysis pipeline belong to the two α-proteobacterial species from among the 26 species that were analyzed in the current study. Notably, Proteobacteria constitute the phylum with the second most abundant representation of riboswitches[21]. Thus, our GC-IGR pipeline is uncovering rare and obscure riboswitch candidates primarily in organisms from bacterial domains that previously yielded the most candidates. This suggests that

future riboswitch discoveries will likely be concentrated in species from these riboswitch-favoring lineages.

As noted above, our findings also highlight the growing difficulty for those who seek to discover novel riboswitch classes. Distinct riboswitch candidates account for only about 1% of all analyzed IGRs, even after the IGRs are sorted based on IGR length and GC content to favor the analysis of those with characteristics most similar to known riboswitch classes. At the current pace, we are encountering an average of one new riboswitch candidate for every two to three bacterial genomes analyzed based on our current findings, on previous publications[72,74], and on unpublished observations. This ratio likely could be improved by biasing our searches towards genomes from bacterial lineages that are known to be enriched for riboswitches. However, as novel riboswitch candidates are uncovered, it is expected that this ratio will continue to decline as the more common and noticeable classes are identified. The remaining hidden classes should trend towards the rare and the structurally obscure. Even with this substantial and growing technical burden, a large number of novel regulatory RNAs are accessible through the use of the GC-IGR pipeline. Although we are working to implement new computational tools to accelerate the GC-IGR pipeline, it is already possible using the current pipeline to systematically analyze all the suitable bacterial genomes available to uncover hundreds of reasonable riboswitch candidates.

Given the capability of this search strategy to yield many additional riboswitch candidates, and the abundance of existing orphan riboswitch candidates[65], it seems appropriate to question the academic value and practical utility of discovering and validating more classes. However, we anticipate that numerous surprising findings await

the discovery of additional riboswitch classes. Some of these novel motifs are rare and/or exceedingly simple in architecture, but might perform sophisticated ligand sensing and gene control functions. For example, a notable riboswitch candidate uncovered by our recent implementation of the GC-IGR pipeline is the "*thiS*" motif[74]. This simple hairpin structure has proven to function as a riboswitch for the thiamin pyrophosphate precursor called HMP-PP[30]. Experimental validation of this unusual ncRNA candidate has demonstrated a unique riboswitch regulatory mechanism as well as revealed another form of biochemical control of the thiamin pyrophosphate biosynthesis pathway. Given the architectural similarity between HMP-PP riboswitches and the HRC-8-4 motif uncovered in the current study (**Fig. 2-5B**), we are optimistic that additional riboswitches exist that make use of very simple terminator-embedded aptamers to control gene expression. If true, the GC-IGR pipeline is particularly well suited to uncover additional versions that exploit this riboswitch mechanism.

Furthermore, numerous other types of structured ncRNA and ncDNA elements are uncovered by implementing the GC-IGR pipeline (**Fig. 2-5**). Many of these will be readily assigned to a few well-understood classes of RNA or DNA genetic elements. However, others will have unusual characteristics, or be modestly represented, such that predicting their functions without additional information will be problematic. Currently, we group these motifs into high-, medium- and low-ranking candidates, although future experiments or additional bioinformatics information might allow them to be grouped into a known class. However, it is very likely that some of these undefined motifs will represent entirely new classes of nucleic acids that carry out novel functions.

Despite our arguments supporting the projection that thousands of riboswitch classes remain to be discovered[21,34,35], it is unrealistic to expect that any single bacterial species will have a dozen or more novel classes of metabolite-responsive riboswitches. The huge number of riboswitch classes proposed to remain undiscovered can be reached even if each bacterial species has an average of less than one novel riboswitch class. Therefore, transcriptomics methods such as term-seq[128] are unlikely to provide a practical means to uncover these hidden classes more broadly among bacterial species. Specifically, it is not possible to culture all bacterial species in a laboratory setting, and even cultured species might need to be grown under many different conditions to generate a transcriptomics pattern that reveals the presence of a rare riboswitch class. In contrast, DNA sequences can be obtained for all bacterial species whose genome can be sampled, and bioinformatics methods such as the GC-IGR search pipeline can reveal relatively rare riboswitch classes.

It is important to note that the term-seq approach can uncover metabolite-binding riboswitch representatives, as was demonstrated by the use of this approach to rediscover 13 previously known examples[128]. However, given the relative rarity of each undiscovered riboswitch class, and the low likelihood of finding a novel riboswitch class in any specific bacterial genome, experimental methods such as term-seq will produce a far greater collection of short RNA transcripts that have functions distinct from metabolite-binding riboswitches. As we have noted previously[74], the GC-IGR search strategy is not well suited to uncover simpler RNA motifs such as are characteristic of sRNAs[112] or regions that code for short peptides. This, coupled with the enrichment of candidate IGRs by length and GC content provides a more effective means to uncover novel riboswitch classes from a large number of bacterial species.

Materials and Methods

Databases and bioinformatics

Bacterial genome sequences from Reference Sequence (RefSeq)[75] database (release 76) and metagenomic datasets previously described[133] were used for the initial searches. In some instances, additional representatives were obtained by conducting homology searches using the Infernal 1.1[63] [92] software package and RefSeq version 80, taking into consideration the sequences flanking the original detected element in both 5' and 3' directions. CMfinder[62] [93] was then used to generate RNA structural alignments to determine if the motif contained additional or alternative structures not previously identified. The first gene downstream of the motif was used to generate genetic context graphics, as this is the gene most likely controlled by a *cis*-regulatory element. If no genetic information is available for a particular sequence, then that representative was excluded from the genetic context data. Consensus sequence and structural models were generated with the program R2R[134].

Chemical and oligonucleotides

All chemicals and chemically synthesized oligonucleotides were purchased from Sigma-Aldrich. Enzymes were purchased from New England BioLabs unless otherwise noted.

Bacterial strains and growth conditions

E. coli BW25113 was obtained from the Coli Genetic Stock Center (Yale University). Reporter vector pRS414 was a gift from W. W. Simons (UCLA). Cells were grown in

Lysogeny Broth (LB) or in M9 broth (1X M9 salts [42 mM Na_2HPO_4, 24 mM KH_2PO_4, 9 mM NaCl, 19 mM NH_4Cl], 1 mM $MgSO_4$, 0.1 mM $CaCl_2$, 2% glucose, 0.5 µg/mL thiamin) purchased from Teknova. When required, growth medium was supplemented with carbenicillin (100 µg/mL).

Reporter gene construct design

Sequences of DNA primers used for cloning are included in **Supplemental Table 2-2**. In-frame plasmid reporter fusion constructs were created via PCR of a DNA fragment that contains the *B. subtilis lysC* promoter and the region encompassing the *odc1* riboswitch candidate from *S. echinoides* ATCC 14820, extending through the first 8 codons of the main ORF. This DNA segment was cloned into the translational reporter vector pRS414, where the 8th codon of the main ORF associated with *odc1* was fused with the 7th codon of the *lacZ* reporter gene. The analogous uORF plasmid reporter fusion construct was created by including an additional G nucleotide between nucleotides 103 and 104 of the natural sequence (**Fig. 2-3B**). The resulting plasmid was transformed into *E. coli* BW25113. Mutant reporter strains were prepared from these two parent constructs using synthetic primers containing the relevant mutations.

Liquid-based β-galactosidase assays

Reporter gene assays were conducted as previously described[55]. For liquid-based β-galactosidase assays, a single colony of the relevant *E. coli* reporter strain was picked and grown overnight in LB medium supplemented with carbenicillin. Cells were then washed twice with phosphate buffered saline (PBS), then diluted 1:200 in either LB or M9 minimal medium and incubated for 6 h at 37^{o}C in various growth conditions. Visual detection of

reporter gene expression is achieved by supplementing liquid media with X-gal (50 μg mL^{-1}). To measure reporter gene expression, 80 μL of each resulting culture was added to a black Costar 96-well clear-bottom assay plate and the absorbance at 595 nm was measured using a Tecan Synergy 2 plate reader. Cells in each well were then mixed with 80 μL of Z buffer (60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 10 mM KCl, 1 mM $MgSO_4$), after which 40 μL of 4-methylumbelliferyl-β-D-galactopyranoside (40 μl of a 1 mg mL^{-1} solution) was added and mixed thoroughly. Plates were incubated at room temperature for 15 min, and the reaction was quenched by the addition of 40 μL of 1 M Na_2CO_3 solution. Excitation and emission were measured at 360/460 nm using a Tecan Synergy 2 plate reader, and fluorescence units were calculated as previously described[29,55].

Tables and Figures

Tables

Table 2-1: List of all 26 genomes included in analysis.

The genome numbering for this series begins at 6 due to the five genome analyses already published. The plot of the IGRs in the first genome, *Listeria monocytogenes* corresponds with **Figure 2-9**, the plot of the last genome, *Candidatus babela massiliensis*, corresponds with **Figure 2-34**, respectively.

Series #	Name	Accession #	Taxonomy	# IGRs analyzed	# IGRs	Avg IGR length	Avg IGR %GC
6	*Listeria monocytogenes*	NC_003210.1	Firmicutes	94	2439	131	34
7	*Fusobacterium nucleatum*	NC_003454.1	Fusobacteria	98	1679	155	23
8	*Ruegeria sp. TM1040*	NC_008044.1	Alphaproteobacteria	74	2450	152	59
9	*Clostridium perfringens*	NC_008261.1	Firmicutes	60	2142	241	21
10	*Pediococcus pentosaceus*	NC_008525.1	Firmicutes	29	1072	189	30
11	*Campylobacter fetus*	NC_008599.1	Epsilonproteobacteria	44	1090	138	24
12	*Polynucleobacter necessarius*	NC_009379.1	Betaproteobacteria	64	1769	84	38
13	*Acholeplasma laidlawii*	NC_010163.1	Tenericutes	61	1039	125	27
14	*Leptospira biflexa*	NC_010602.1	Spirochaetes	142	2501	100	36
15	*Thermosipho africanus*	NC_011653.1	Thermotogae	36	1338	129	31
16	*Exiguobacterium sp. AT1b*	NC_012673.1	Firmicutes	44	1830	157	41
17	*Methylotenera mobilis*	NC_012968.1	Betaproteobacteria	16	1994	119	37
18	*Veillonella parvula*	NC_013520.1	Negativicutes	74	1576	176	32
19	*Candidatus Riesia pediculicola*	NC_014109.1	Gammaproteobacteria	25	407	228	21
20	*Arcobacter nitrofigilis DSM 7299*	NC_014166.1	Epsilonproteobacteria	51	2275	103	22
21	*Mycoplasma fermentans*	NC_014921.1	Tenericutes	25	681	155	21
22	*Hippea maritima*	NC_015318.1	Deltaproteobacteria	32	968	106	37
23	*Thermodesulfobacterium geofontis*	NC_015682.1	Thermodesulfobacteria	28	1147	90	24
24	*Taylorella asinigenitalis MCE3*	NC_016043.1	Betaproteobacteria	50	1124	107	29
25	*Zymomonas mobilis* sp. mobilis	NC_017262.1	Alphaproteobacteria	82	1497	182	38
26	*Candidatus kinetoplastibacterium oncopeltii*	NC_020299.1	Betaproteobacteria	8	625	162	20
27	*Hydrogenobaculum*	NC_020411.1	Aquificae	34	954	93	27
28	*Beta proteobacterium*	NC_020417.1	Betaproteobacteria	56	1697	90	39
29	*Spiribacter salinus M19-40*	NC_021291.1	Gammaproteobacteria	53	1151	80	63
30	*Chlamydia pecorum*	NC_022440.1	Chlamydiae	22	745	118	35
31	*Candidatus babela massiliensis*	NC_023003.1	Deltaproteobacteria	33	888	159	20

Table 2-2: Sequences of synthetic DNAs used in this study.

Name	Sequence (5′ to 3′)	Annotation
WT	TACGAC**GAATTC**CAAAAATA ATG**TTGATCCTTTTAAATAA GTCTGATAAAATGTGAACTA** AAUCUCGCCCAUCCGCUGCC CCAUGGGACUUCCAACCGCA AGGCAGCUUUUUUCACCGUA GGCGCAAGCCACUUGGAGGU CCCUUGAGUUGCACAAGCAU CAUCGCGCGCU**GGATCC**AAA GGA	The *odc1* reporter with *lysC* promoter for making the wild-type reporter construct plasmid, from EcoRI to BamHI restriction sites. Restriction sites are bolded, lysC promoter is in blue, and *odc1* temple template from *S. echinoides* is highlighted in gray. The two alternative start codons are highlighted in red. Nucleotides before and after the restriction enzyme sites are fillers for PCR.
M1	TACGAC**GAATTC**CAAAAATA ATG**TTGATCCTTTTAAATAA GTCTGATAAAATGTGAACTA** AAUCUCGCCCAUCCGCUGCC CCAUGGGACUUCCAACCGCA AGGCAGCUUUUUUCACCGUA GGCGCAAGCCACUUGGAGGU CCCUUGAGUUCCACAAGCAU CAUCGCGCGCU**GGATCC**AAA GGA	The template for making the M1 mutant *odc1* reporter. Annotations are described above.
M2	TACGAC**GAATTC**CAAAAATA ATG**TTGATCCTTTTAAATAA GTCTGATAAAATGTGAACTA** AAUCUCGCCCAUCCGCUGCC CCAUGGGACUUCCAACCGCA AGGCAGCUUUUUUCACCGUA GGCGCAAGCCACUUGGAGGU CCCUUCAGUUGCACAAGCAU CAUCGCGCGCU**GGATCC**AAA GGA	The template for making the M2 mutant *odc1* reporter. Annotations are described above.
M3	TACGAC**GAATTC**CAAAAATA ATG**TTGATCCTTTTAAATAA GTCTGATAAAATGTGAACTA** AAUCUCGCCCAUCCGCUGCC CCAUGGGACUUCCAACCGCA AGGCAGCUUUUUUCACCGUA GGCGCAAGCCACUUGGAGGU CCCUUCAGUUCCACAAGCAU CAUCGCGCGCU**GGATCC**AAA GGA	The template for making the M3 mutant *odc1* reporter. Annotations are described above.

M4	TACGAC**GAATTC**CAAAAATA ATG**TTGATCCTTTTAAATAA GTCTGATAAAATGTGAACTA** AAUCUCGCCCAUCCGCUGCC CCAUGGGACUUCCAACAGCA AGGCAGCUUUUUUCACCGUA GGCGCAAGCCACUUGGAGGU CCCUUGAGUUGCACAAGCAU CAUCGCGCGCU**GGATCC**AAA GGA	The template for making the M4 mutant *odc1* reporter. Annotations are described above.
M5	TACGAC**GAATTC**CAAAAATA ATG**TTGATCCTTTTAAATAA GTCTGATAAAATGTGAACTA** AAUCUCGCCCAUCCGCUGCC CCAUGGGACUUCCAACCGAA AGGCAGCUUUUUUCACCGUA GGCGCAAGCCACUUGGAGGU CCCUUCAGUUGCACAAGCAU CAUCGCGCGCU**GGATCC**AAA GGA	The template for making the M5 mutant *odc1* reporter. Annotations are described above.
M6	TACGAC**GAATTC**CAAAAATA ATG**TTGATCCTTTTAAATAA GTCTGATAAAATGTGAACTA** AAUCUCGCCCAUCCGCUGCC CCAUGGGACUUCCAACCGCC AGGCAGCUUUUUUCACCGUA GGCGCAAGCCACUUGGAGGU CCCUUCAGUUGCACAAGCAU CAUCGCGCGCU**GGATCC**AAA GGA	The template for making the M6 mutant *odc1* reporter. Annotations are described above.

PCR Forward Primer

TAC GAC GAA TTC CAA AAA TAA TGT TGA TCC (56.1)

PCR Reverse Primer

TCC TTT GGA TCC AGC GC (55.3)

PCR Reverse Primer (For making +1 frame reporter constructs)

TCC TTT GGA TCC CAG CGC

Figures

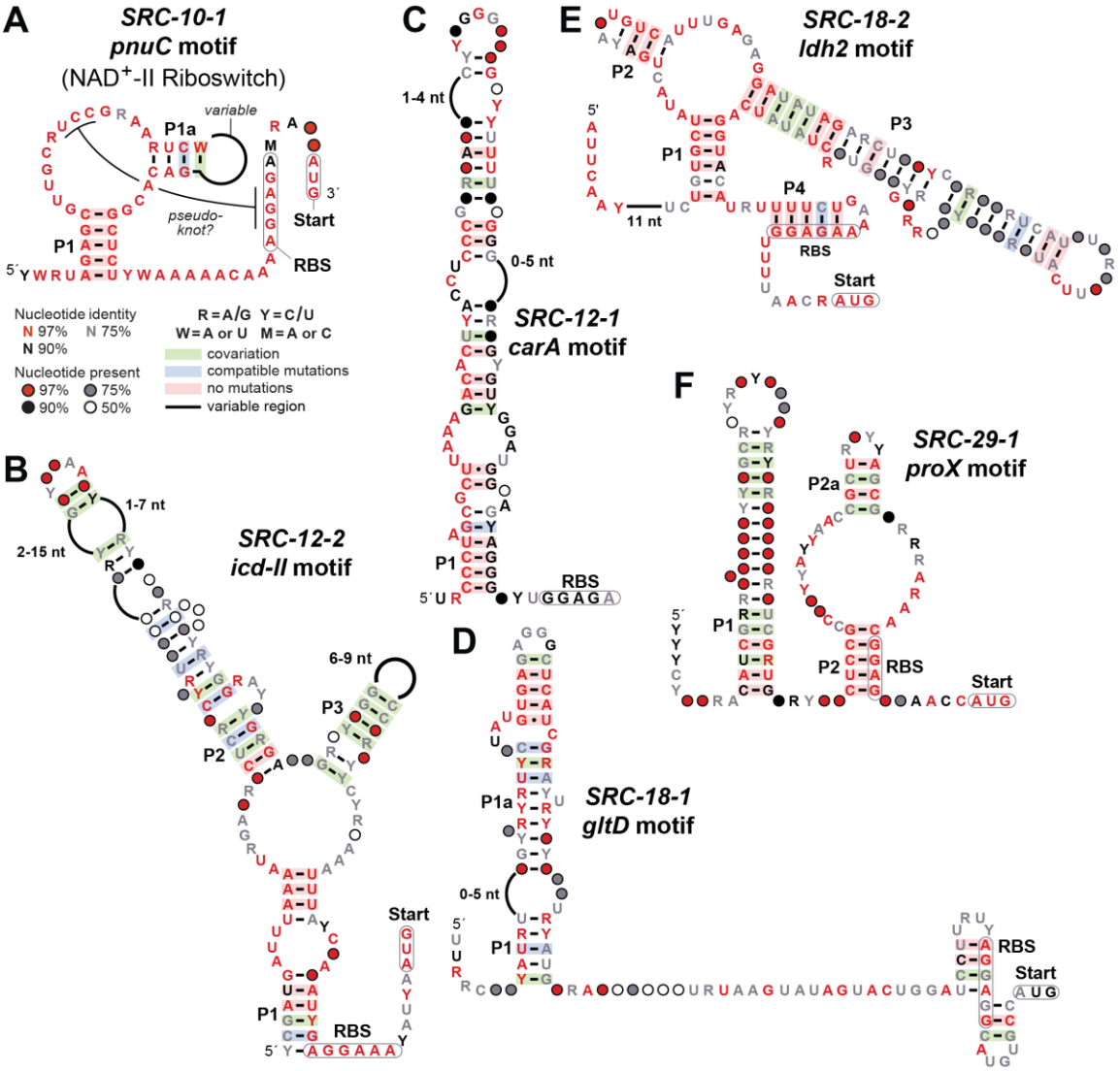

Figure 2-1: Consensus sequence and structural models for strong riboswitch candidates identified in this study.

(A) *pnuC* (NAD+-II riboswitch)[25], (B) *icd*-II, (C) *carA*, (D) *gltD*, (E) *ldh2* and (F) *proX* motif RNAs. Annotations are as defined in the key depicted in A.

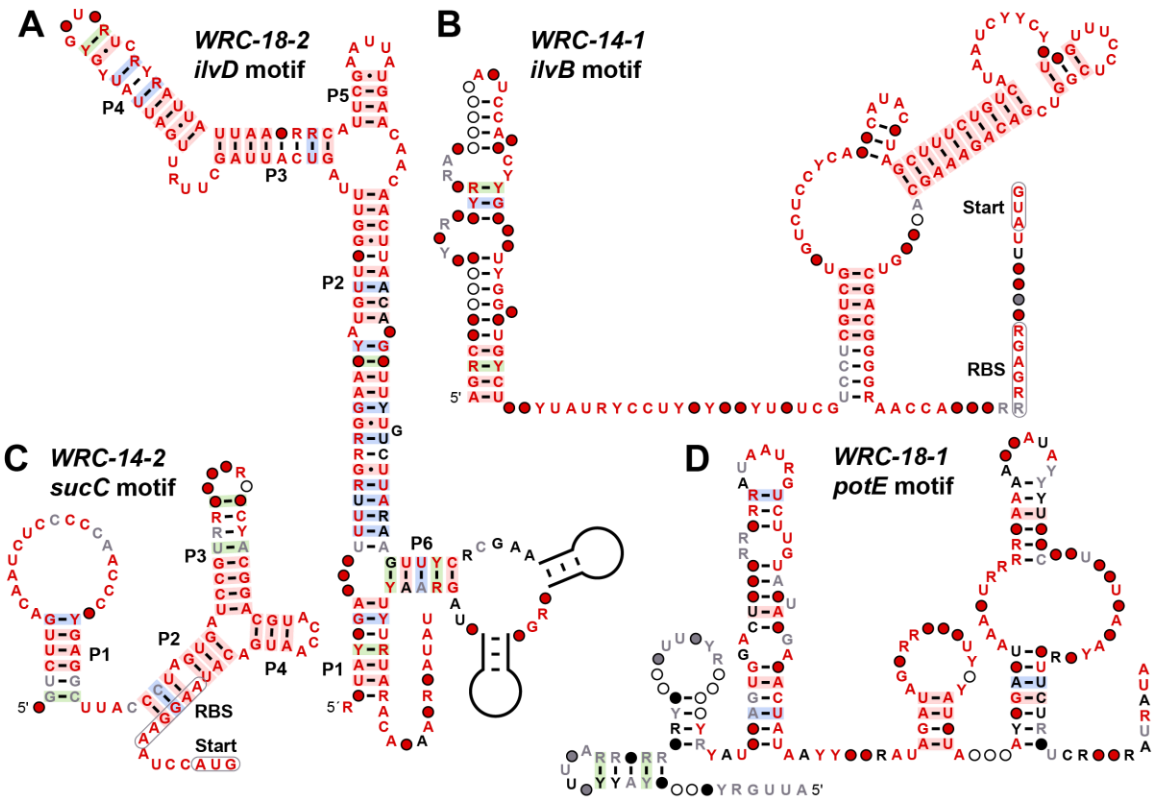

Figure 2-2: Consensus sequence and structural models for the weak riboswitch candidates identified in this study.

(A) *ilvD*, (B) *ilvB*, (C) *sucC*, and (D) *potE* motif RNAs. Annotations are as defined in the key depicted in **Figure 2-1A**.

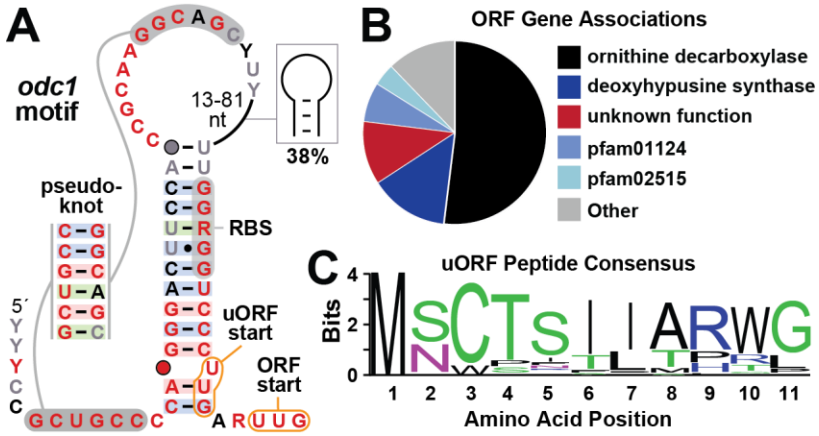

Figure 2-3: The *odc1* motif consensus model and gene associations.

(A) Conserved nucleotide sequences and secondary structure model based on 272 representatives of *odc1* motif RNAs. Annotations are as defined in the key depicted in **Figure 2-1A**. (B) Distribution of gene associations for the *odc1* motif. Included are the first three genes downstream of the motif, which might constitute an operon. Genes annotated pfam01124 code for proteins of unknown function, and genes for pfam02515 code for putative CoA transferase enzymes also of uncertain function. (C) Consensus sequence (logo plot) for the peptide produced from the uORF start codon in all representatives of *odc1* motif RNAs.

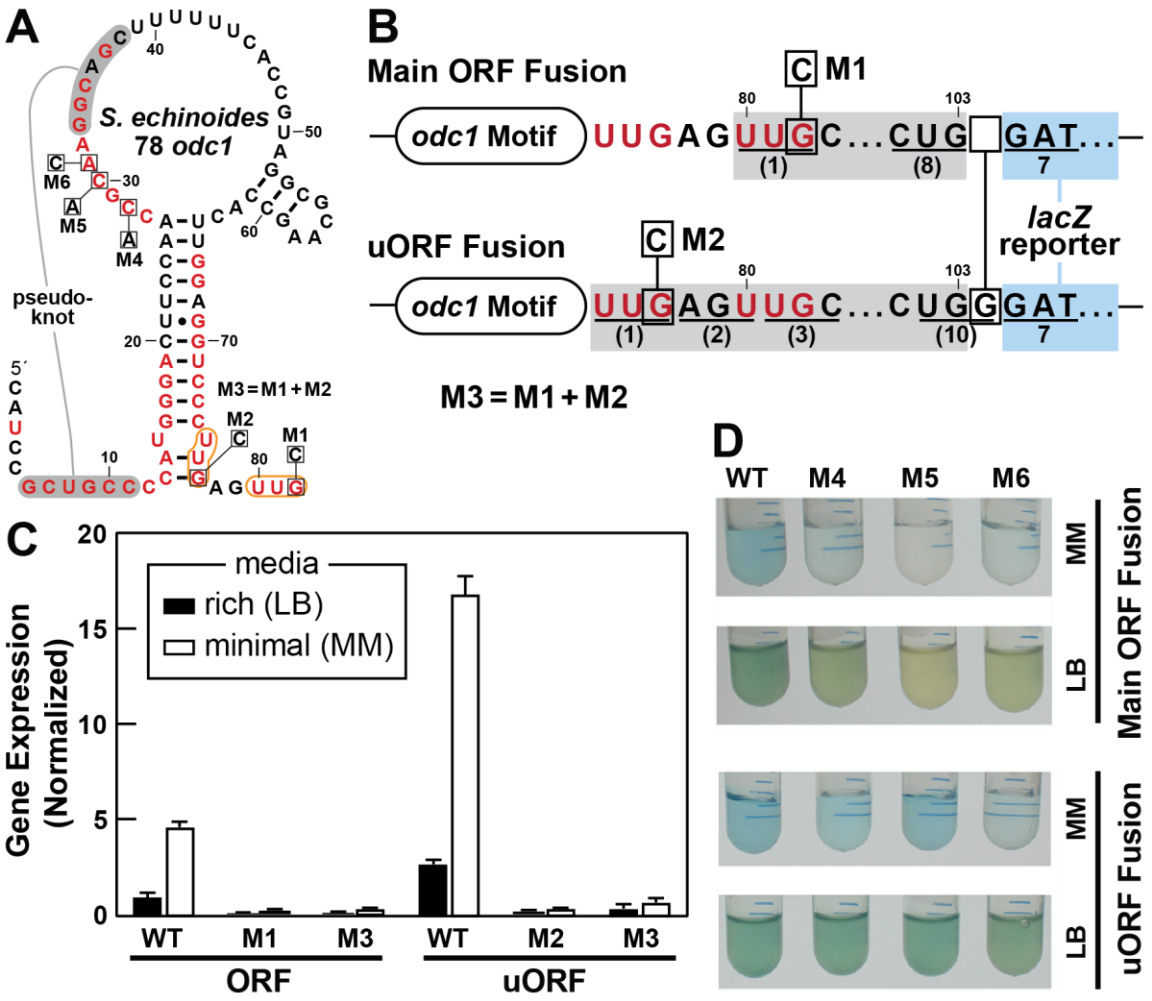

Figure 2-4: Expression of proteins from the two *odc1* translation start codons.

(A) Sequence and predicted secondary structure of the wild-type *odc1* motif RNA from the *S. echinoides* representative. Red nucleotides identify conserved nucleotides of the *odc1* motif consensus as depicted in **Fig. 2-3A**. Nucleotide changes to create mutant constructs used in reporter gene assays are identified by boxed letters. Mutant construct M3 is the combination of mutations M1 and M2. (B) Reporter constructs were created by joining the *S. echinoides odc1* motif representative to an *E. coli lacZ* reporter gene. Top: Sequence of the construct fusing first 8 codons of the *odc1* gene (gray; codons underlined and numbered in parentheses) with the 7th codon of the *lacZ* ORF (blue). Bottom: Sequence of the

construct fusing the uORF start codon in frame with the *lacZ* ORF. Note that an insertion of a single G nucleotide after nucleotide 103 creates an in-frame fusion between the 10 codons of the *odc1* motif uORF and the 7th codon of *lacZ*. Additional annotations are as described for A. (C) Plot of gene expression values for *E. coli* cells carrying various reporter fusion constructs as indicated and grown in media characterized as rich (LB) or minimal (MM). Values were normalized to that measured for cells carrying the WT main ORF construct and grown in rich medium. Bars represent the average of three replicates at the error bars indicate standard deviation. (D) Images of LB or MM liquid cultures supplemented with x-gal that were inoculated with *E. coli* cells carrying various reporter constructs as indicated.

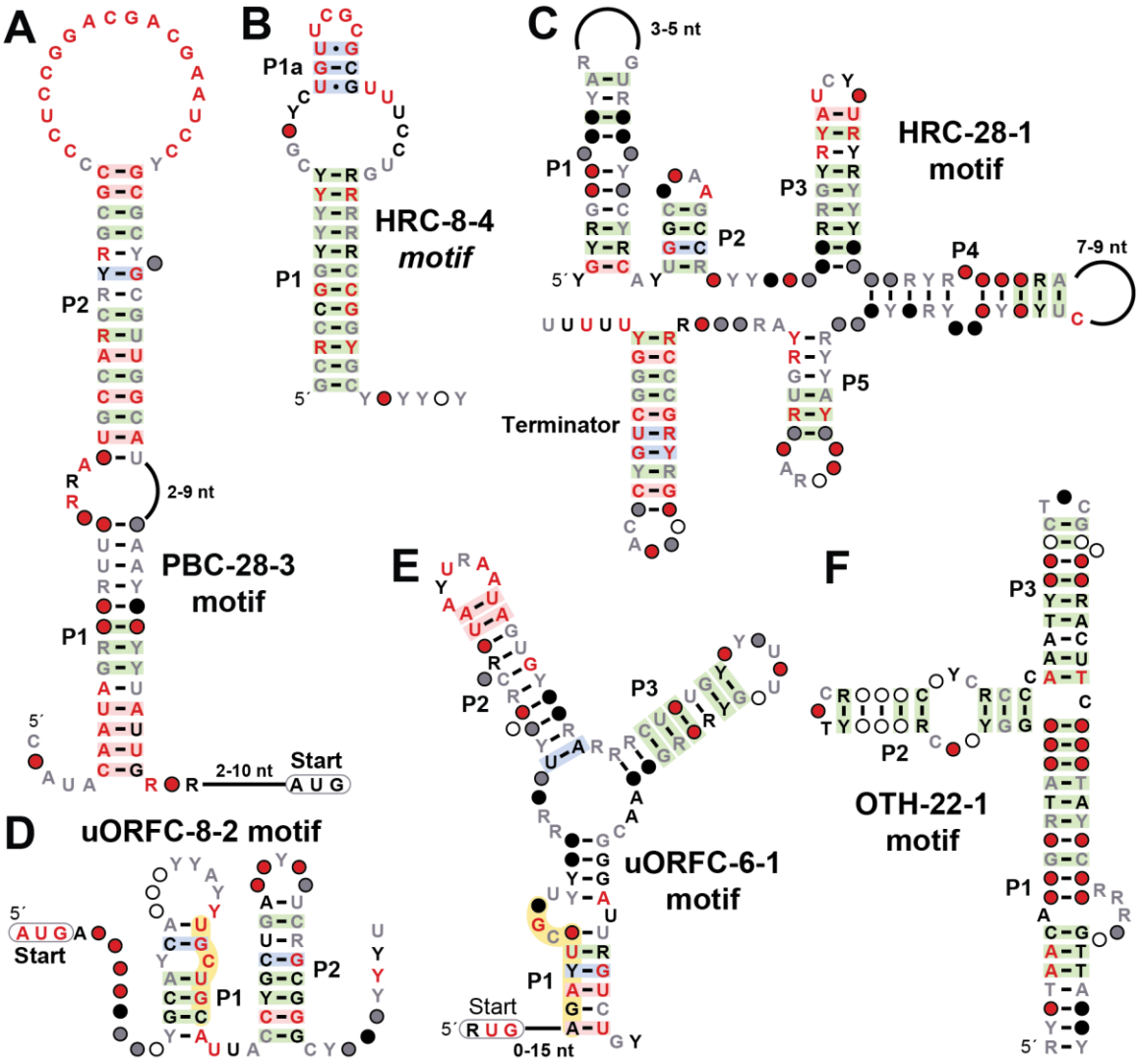

Figure 2-5: Consensus sequence and secondary structure models for additional structured nucleic acid motifs that are representative of those identified in this study. (A) a putative protein binding candidate PBC-28-3, two high-ranking candidates (B) HRC-8-4 and (C) HRC-28-1, two uORF candidates (D) uORFC-8-2, and (E) uORFC-6-1, and (F) a predicted structured DNA motif OTH-22-1. Annotations are as defined in the key depicted in **Figure 2-1A**.

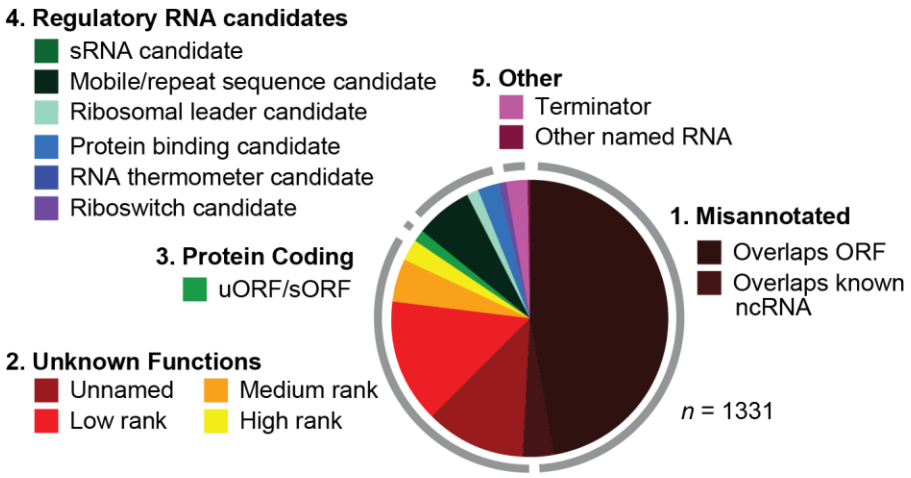

Figure 2-6: Comprehensive summary of classification of selected unknown IGRs from the analysis of 26 bacterial genomes chosen for this study.

IGR classifications are placed into five groups. (1) Originally annotated IGRs that either code for ordinary-length proteins or serve as templates for known types of ncRNAs. (2) IGRs that are judged to have varying degrees of promise as structured ncRNAs but lack sufficient evidence to assign a possible function. (3) Originally annotated as IGRs but are now predicted to code for short peptides. (4) IGRs that likely serve as templates for the transcription of ncRNAs, including regulatory RNA candidates and selfish (mobile/repeat) sequences. (5) Transcription terminator structures or other sequences.

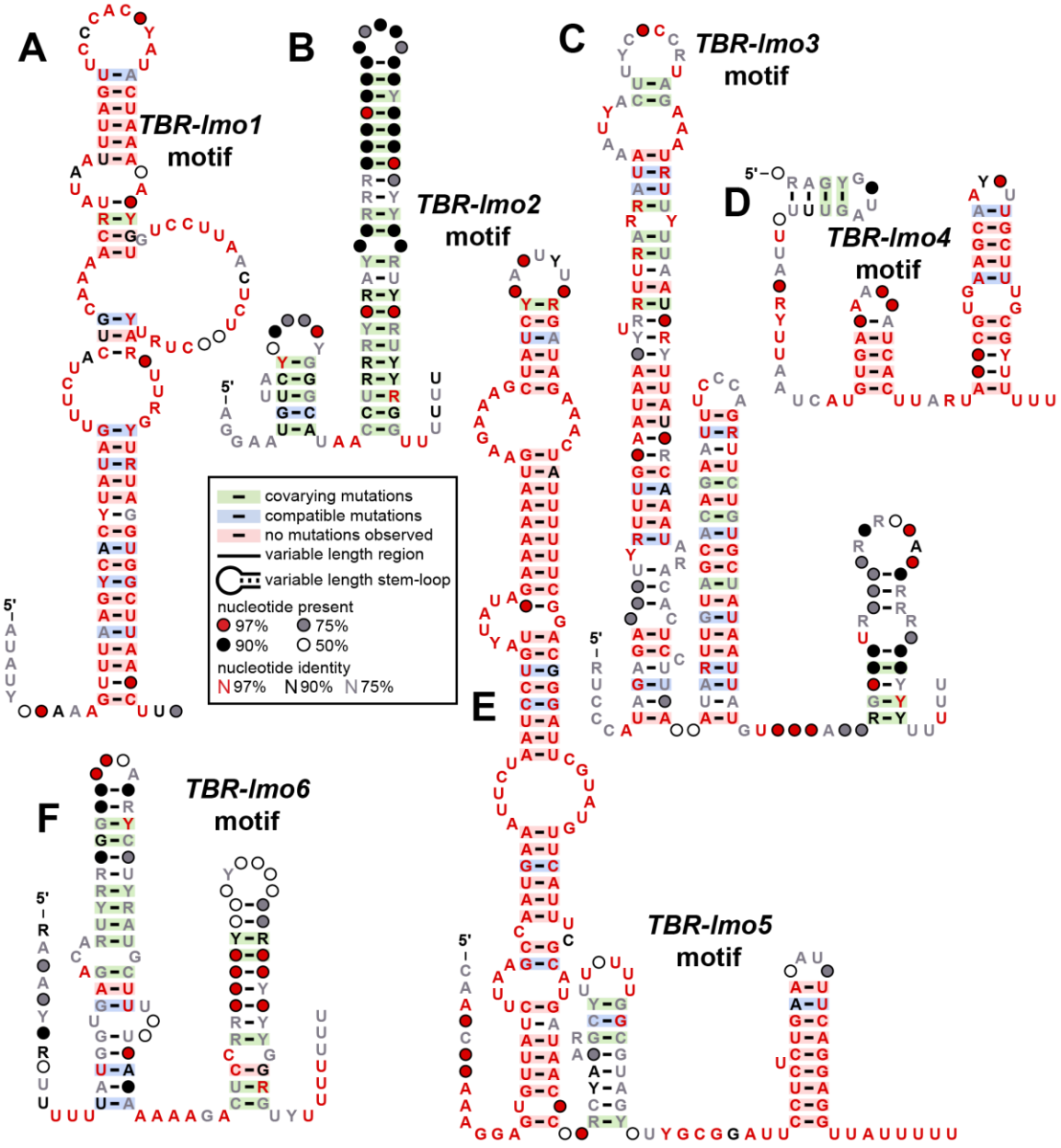

Figure 2-7. Motifs identified from the IGRs containing the "riboregulators" identified via term-Seq.

The IGR for each riboregulator was added to the analysis pipeline despite most having been previously filtered out on the basis of IGR GC-content, length, or the presence of known structured RNAs. The most promising structured ncRNA motif resulting from the pipeline's iterative cycle of Infernal search and CMFinder structure prediction is shown.

(**A**) TBR-lmo1. This motif consists of 27 unique sequences found exclusively in various species of the *Listeria* genus. The structure appears to be an unremarkable terminator stem, although the predicted base-pairing is not supported by strong co-variation. The motif is found exclusively in front of genes coding for predicted Zn-dependent metalloproteases. (**B**) TBR-lmo2. This motif consists of 73 unique sequences found in several families of Bacillales. The structure is dominated by an intrinsic terminator stem that is strongly supported by co-variation. The motif can be found in front of collection of disaccharide-specific transporter and phosphatase encoding genes. (**C**) TBR-lmo3. This motif consists of 41 unique sequences found only in the *Listeria* genus. The structure consists of two adjacent stems moderately supported by covariation followed by a short terminator stem. The motif is with few exceptions found in front of genes encoding predicted divalent metal ion transporters. (**D**) TBR-lmo4. This motif consists of 31 unique sequences found only in the *Listeria* genus. The structure consists of three-particularly short stems poorly supported by covariation. Only the final stem appears to have length and stretch of U's to indicate activity as a possible terminator. The motif is always found in front of predicted sulfate permease encoding genes. (**E**) TBR-lmo5. This motif consists of 25 unique sequences found only in the *Listeria* genus. The predicted structure consists of one long stem poorly supported by covariation, one short stem strongly supported by covariation and a short terminator stem. The motif is usually found near genes generically annotated as ABC transporter ATPase. (**F**) TBR-lmo6. This motif consists of 38 unique sequences found mostly in the *Listeria* genus. The predicted structure consists of two covariation supported stems, one of which is a clear terminator. The motif is always found upstream of genes

encoding the uncharacterized membrane-anchored protein YitT. Annotations are as defined in the key depicted in **Figure 2-1A**.

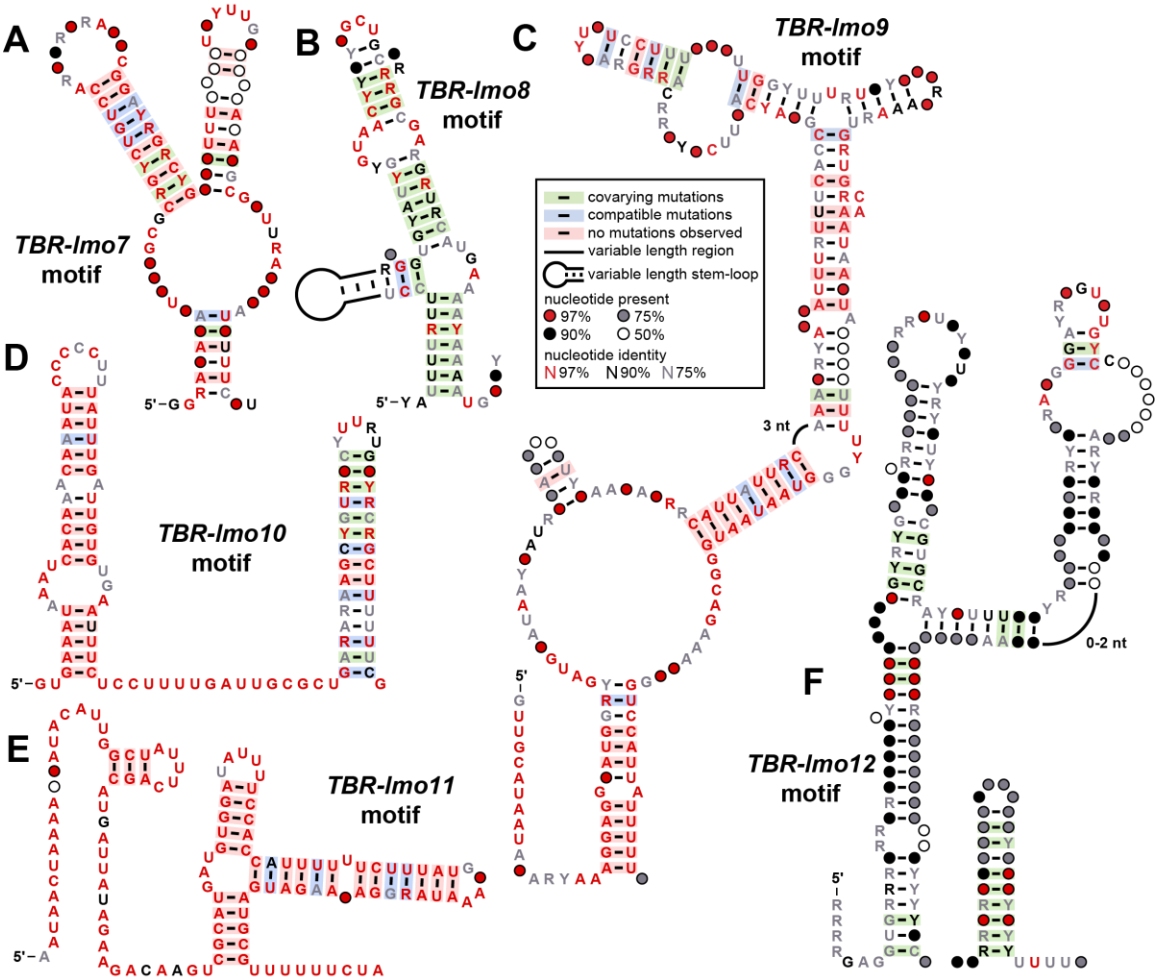

Figure 2-8: Additional motifs identified from the IGRs containing the "riboregulators" identified via term-Seq.

Analysis performed as described in **Figure 2-7**. (**A**) TBR-lmo7. This motif consists of 20 unique sequences found almost exclusively in the *Listeria* genus. The predicted structure for this motif is one of the few from this collection that contain a multi-stem junction. However, this structure has only some support from covariation. The motif is always found upstream of genes encoding acetyl-CoA carboxylase. (**B**) TBR-lmo8. This motif consists

of 466 unique sequences found in a wide variety of Frimicutes. The genetic context upstream 30S of ribosomal protein S4 is indicative of a ribosomal leader candidate. (**C**) TBR-lmo9. This motif consists of 36 unique sequences found only in the *Listeria* genus. The structure of this motif is unusual with large loops in between co-variation supported stems and no clear terminator. This motif is typically found upstream of genes encoding putative membrane-bound multidrug transporters. (**D**) TBR-lmo10. This motif consists of 62 unique sequences found in multiple families of Bacillales. The structure of this motif consists of two simple stems one of which may be a terminator. This motif is nearly always found upstream of DUF3116, a domain of unknown function that appears to be restricted to Bacillales. (**E**) TBR-lmo11. This motif consists of 18 unique sequences found only in the genus *Listeria*. The narrow distribution of nearly identical sequences in this motif leave the predicted structure completely unsupported by covariation. There does, however, appear to be a plausible termination stem. This motif is always associated with inosine 5'-monophosphate dehydrogenase. (**F**) TBR-lmo12. This motif consists of 359 unique sequences found in multiple families of Bacilli and Clostridia. Covariation strongly supports the presence of a terminator stem at the end of the motif, and there is some support from covariation for the three-stem multistem junction. This motif is always found upstream of genes encoding putative membrane-bound multidrug transporters. Annotations are as defined in the key depicted in **Figure 2-1A**.

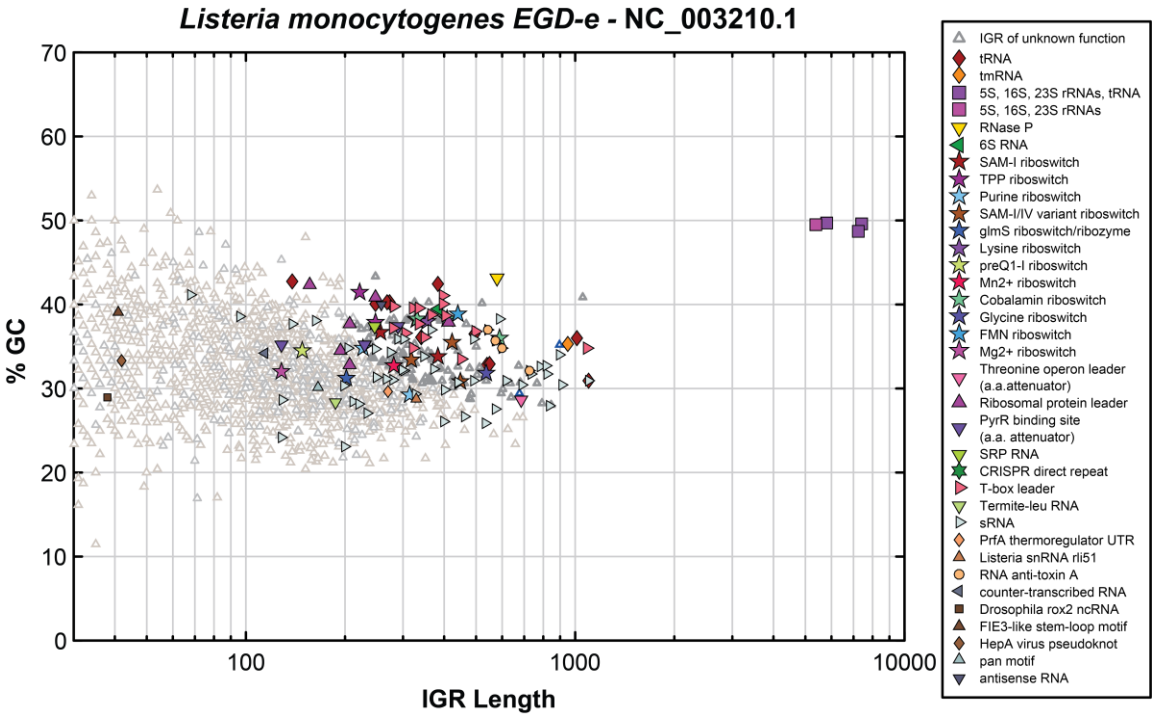

Figure 2-9. Plot of the IGRs from the L. monocytogenes genome.

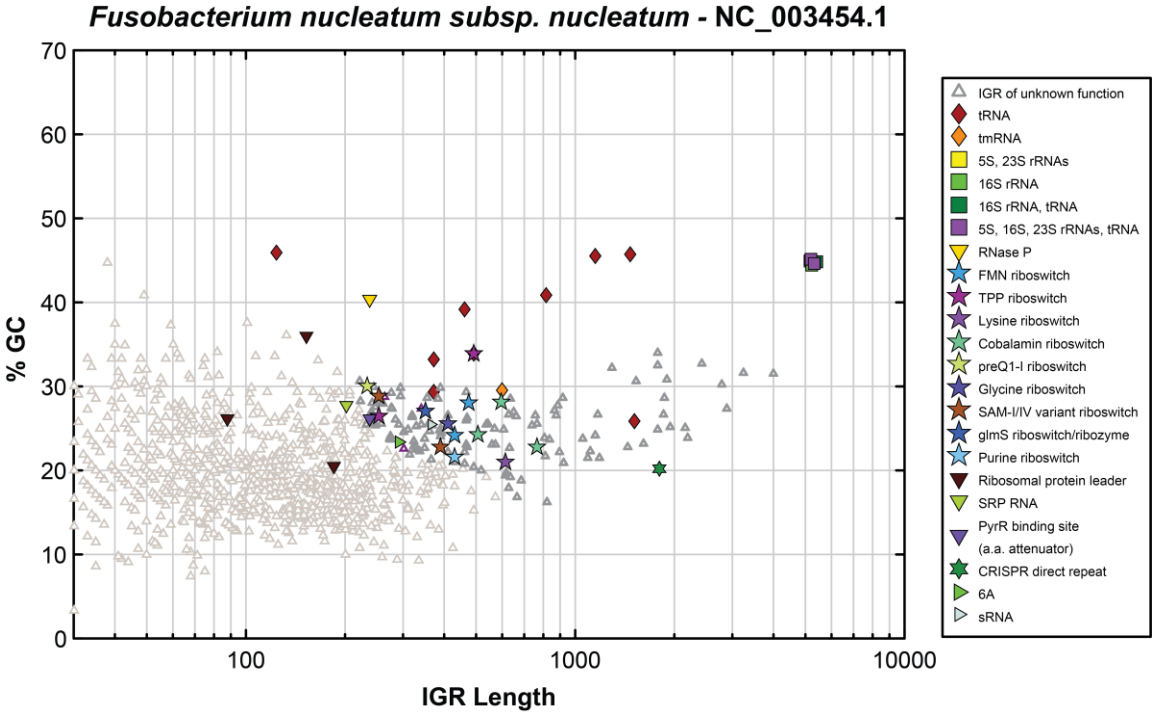

Figure 2-10: Plot of the IGRs from the *F. nucleatum* genome.

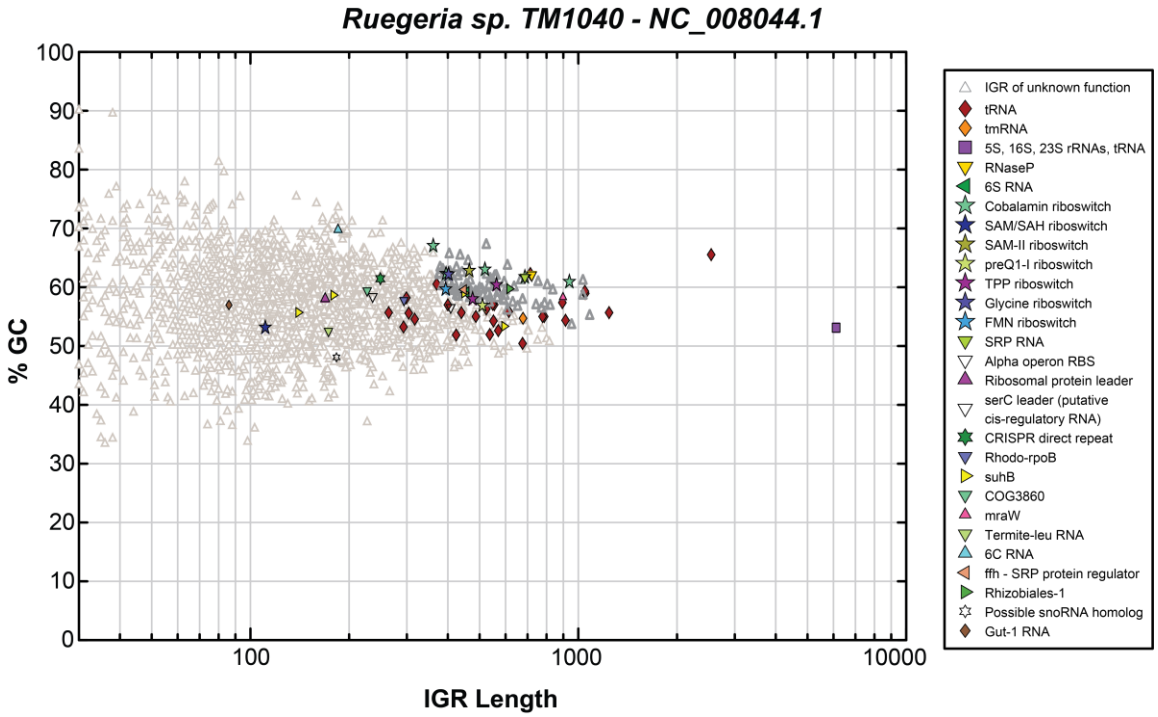

Figure 2-11: Plot of the IGRs from the *Ruegeria* genome.

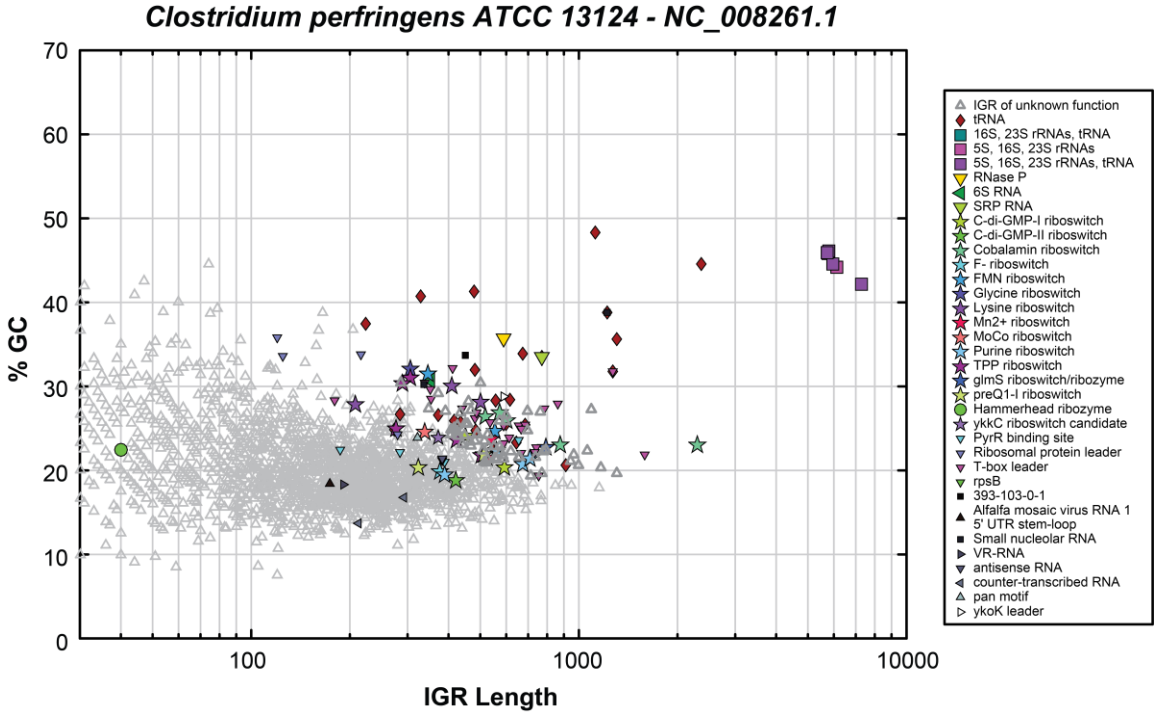

Figure 2-12: Plot of the IGRs from the *Clostridium perfringens* genome.

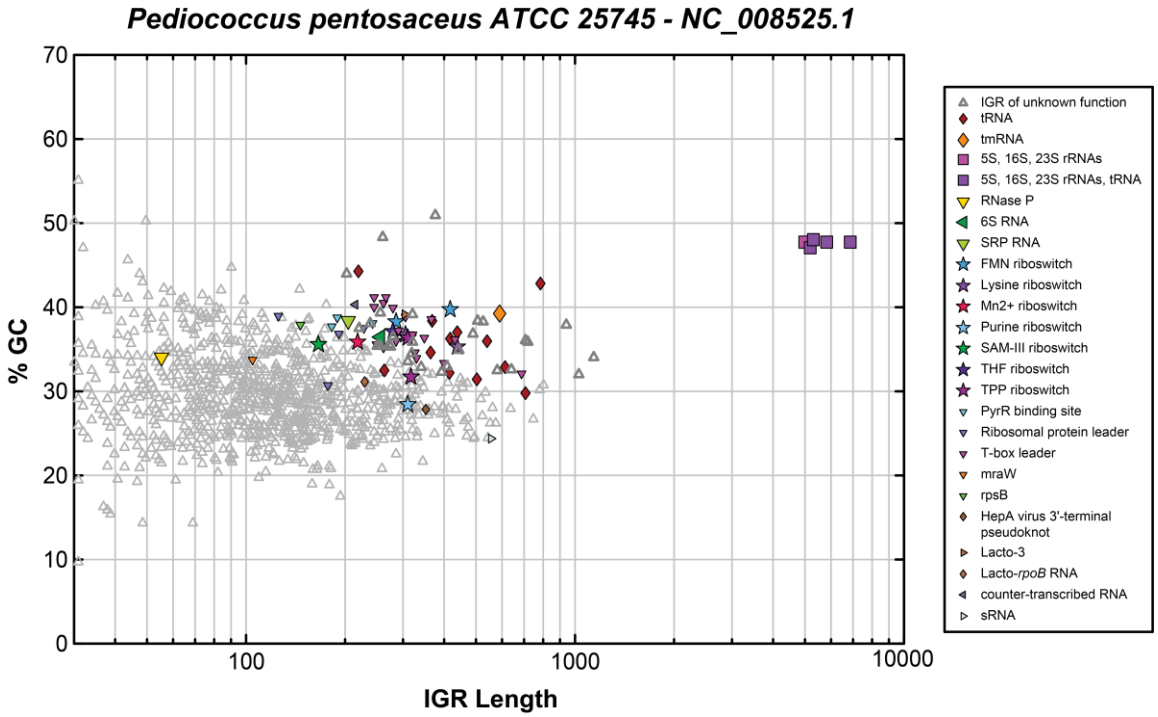

Figure 2-13: Plot of the IGRs from the *P. pentosaceus* genome.

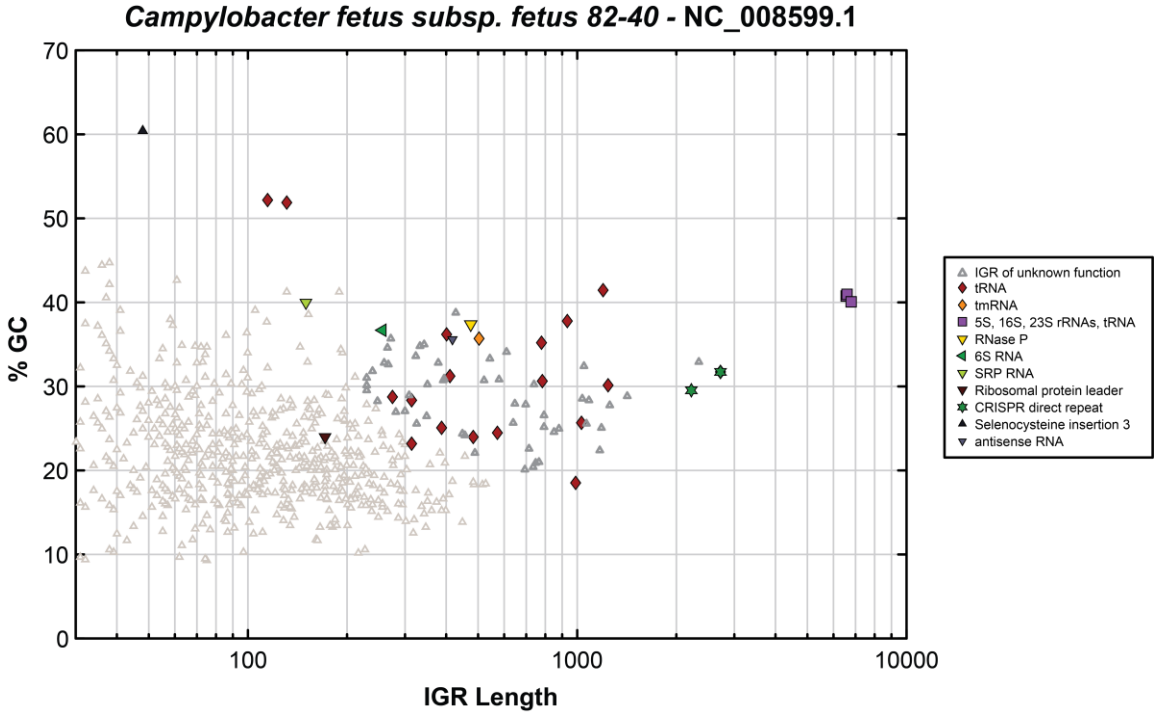

Figure 2-14: Plot of the IGRs from the *C. fetus* genome.

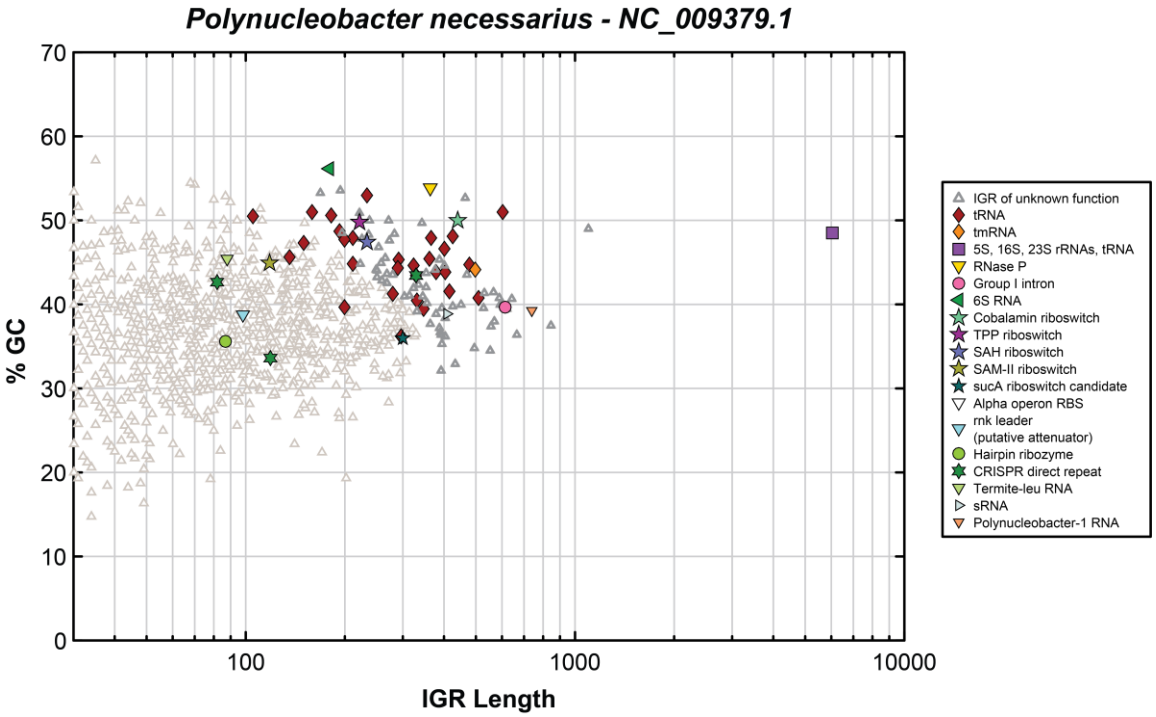

Figure 2-15: Plot of the IGRs from the *P. necessarius* genome.

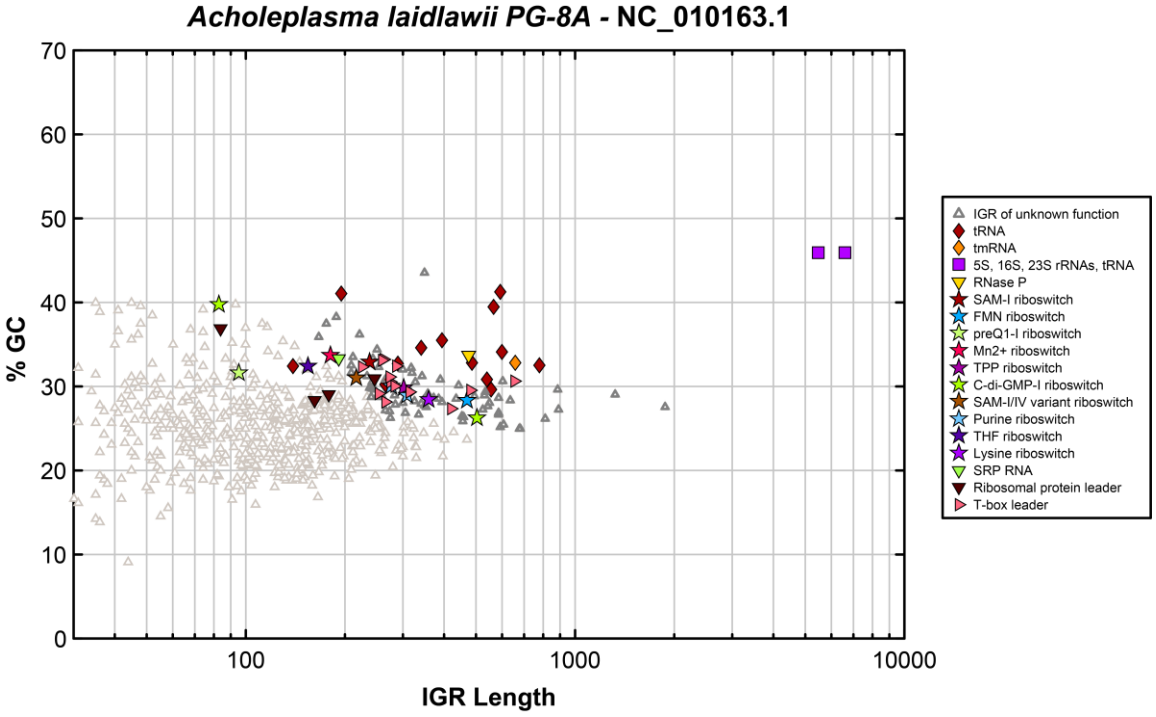

Figure 2-16: Plot of the IGRs from the *A. laidlawii* genome.

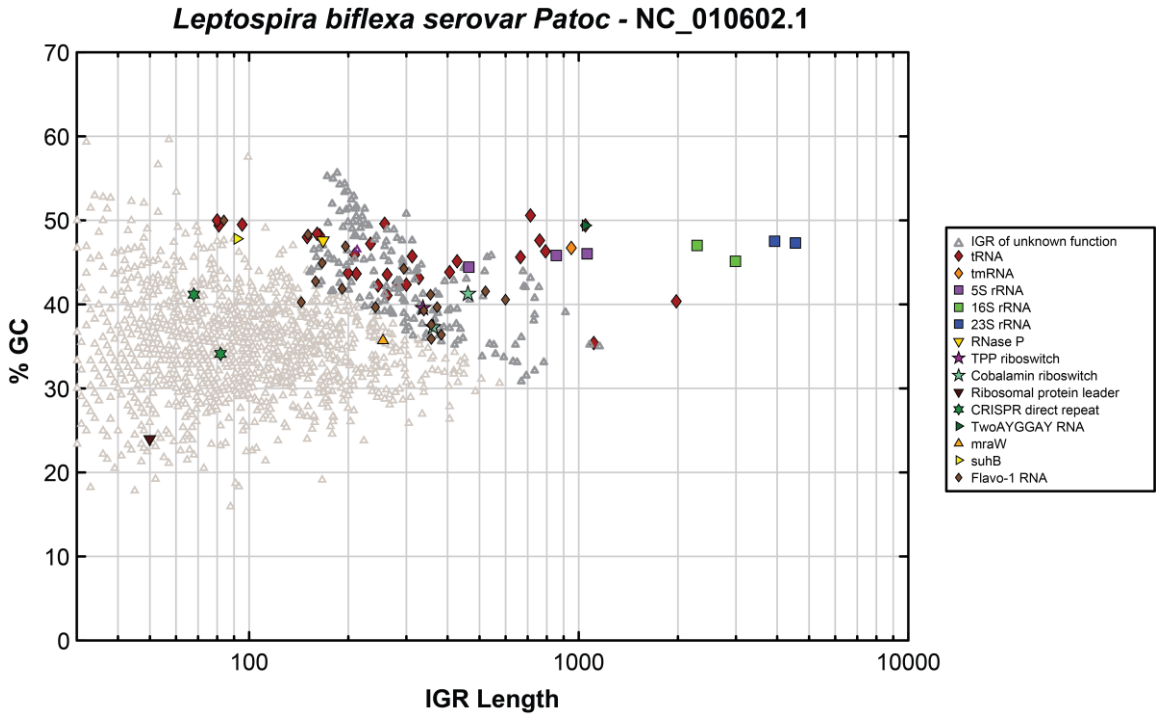

Figure 2-17: Plot of the IGRs from the *L. biflexa* genome.

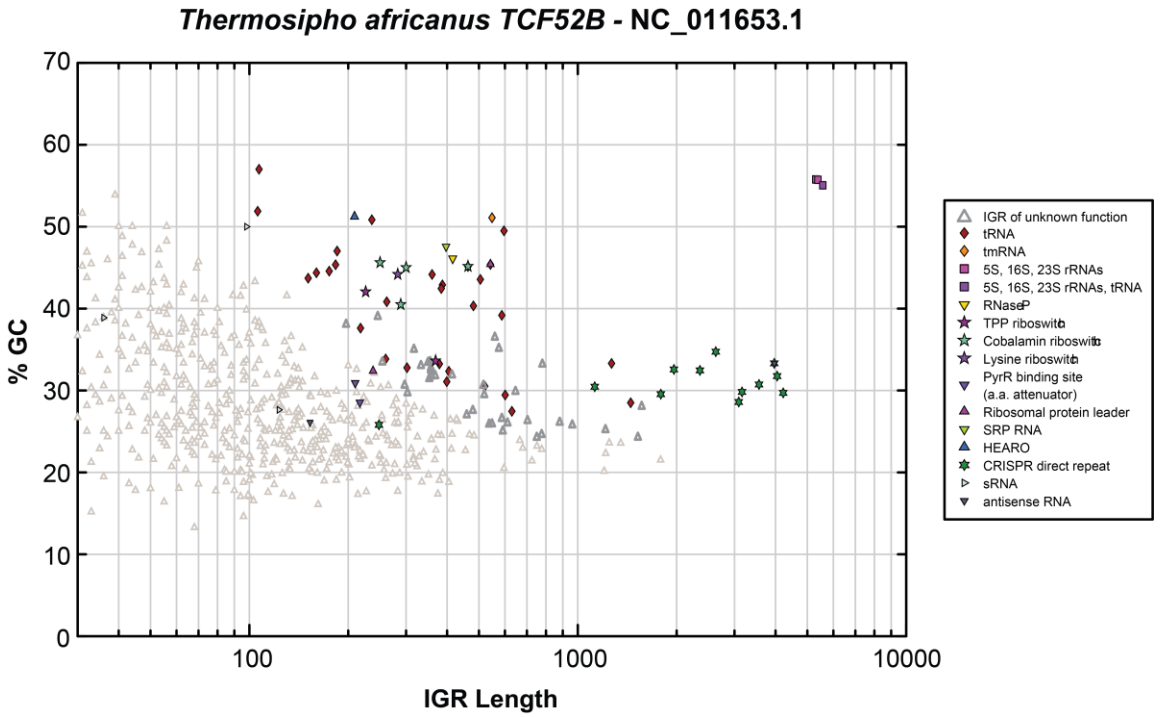

Figure 2-18: Plot of the IGRs from the *T. africanus* genome.

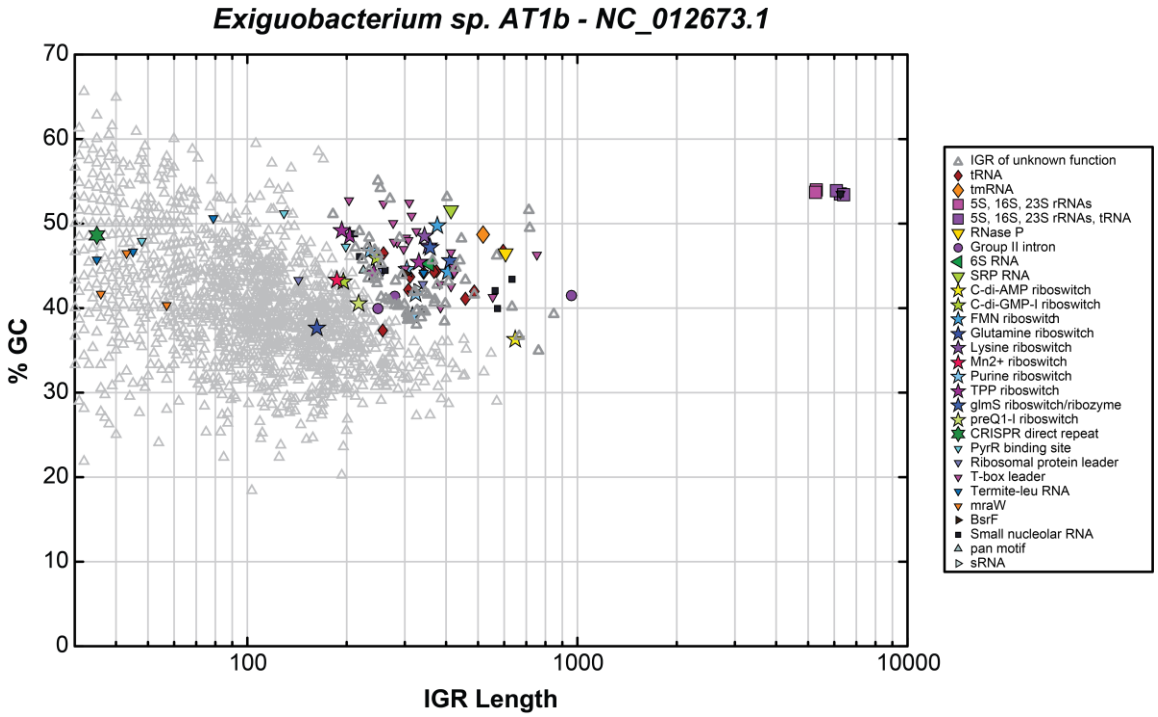

Figure 2-19: Plot of the IGRs from the *Exiguobacterium* genome.

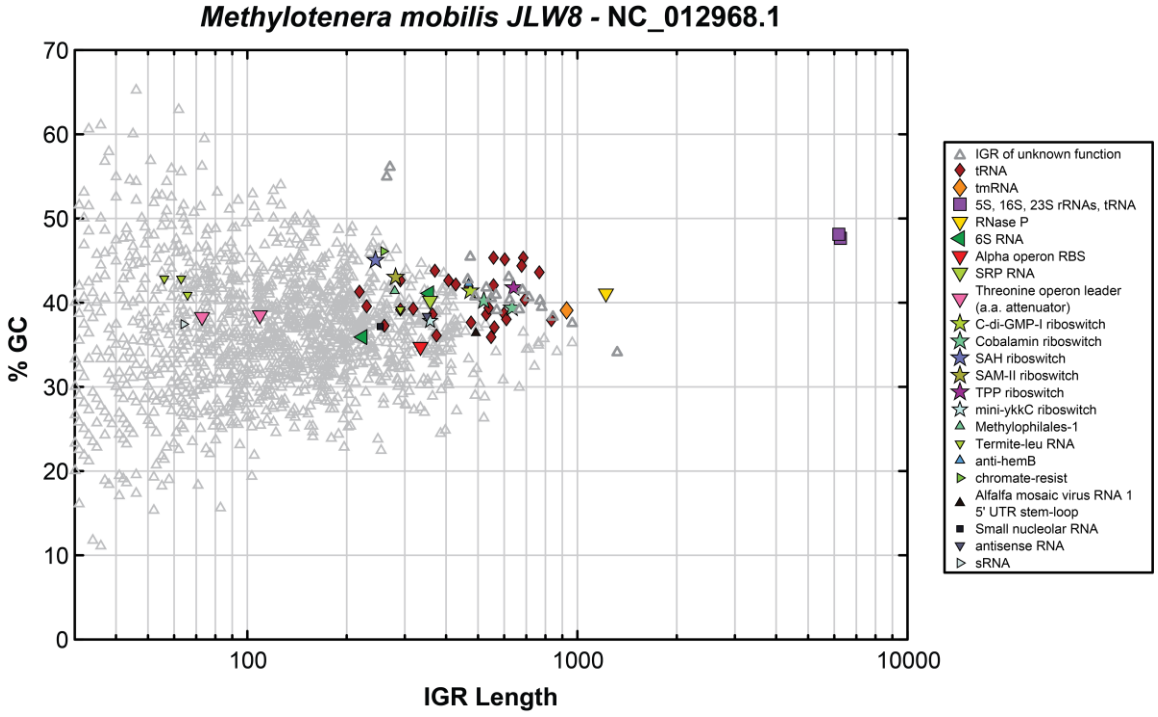

Figure 2-20: Plot of the IGRs from the *M. mobilis* genome.

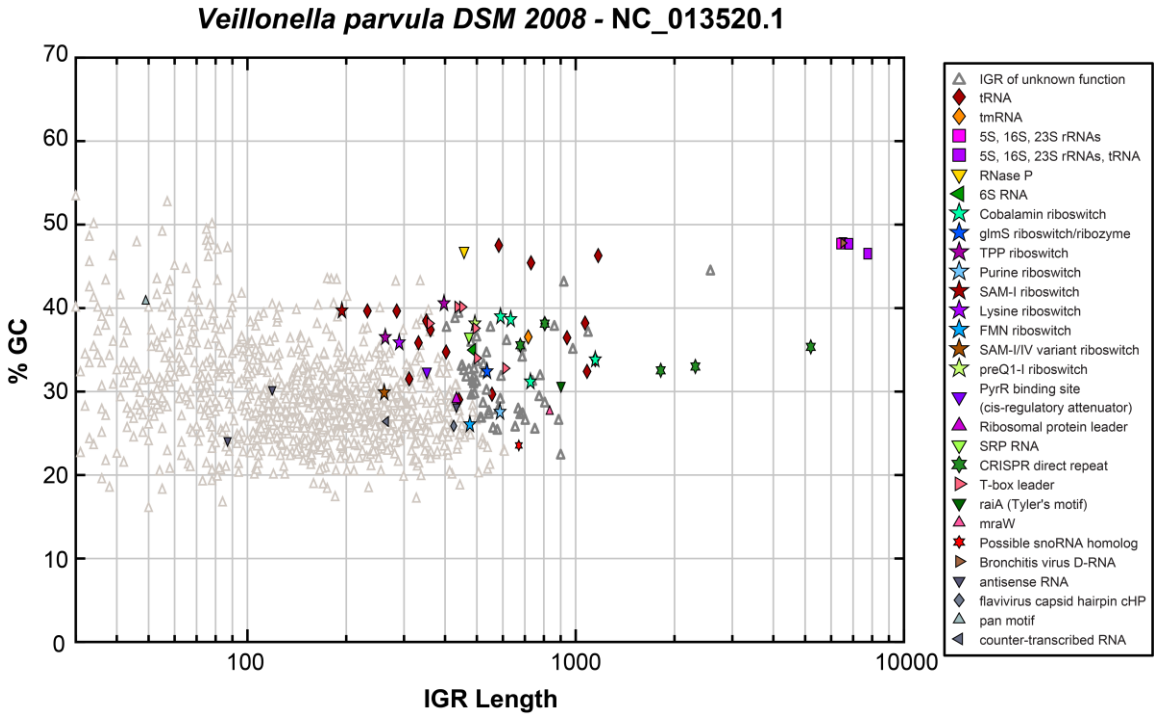

Figure 2-21: Plot of the IGRs from the *V. parvula* genome.

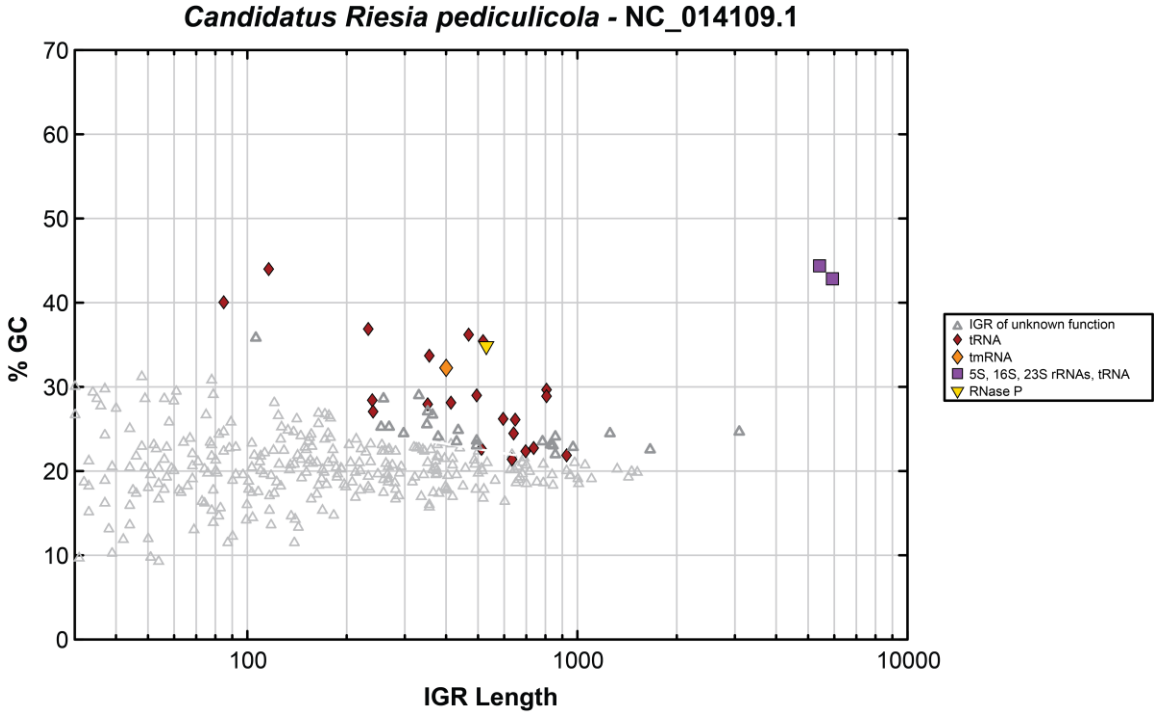

Figure 2-22: Plot of the IGRs from the *C.R. pediculicola* genome.

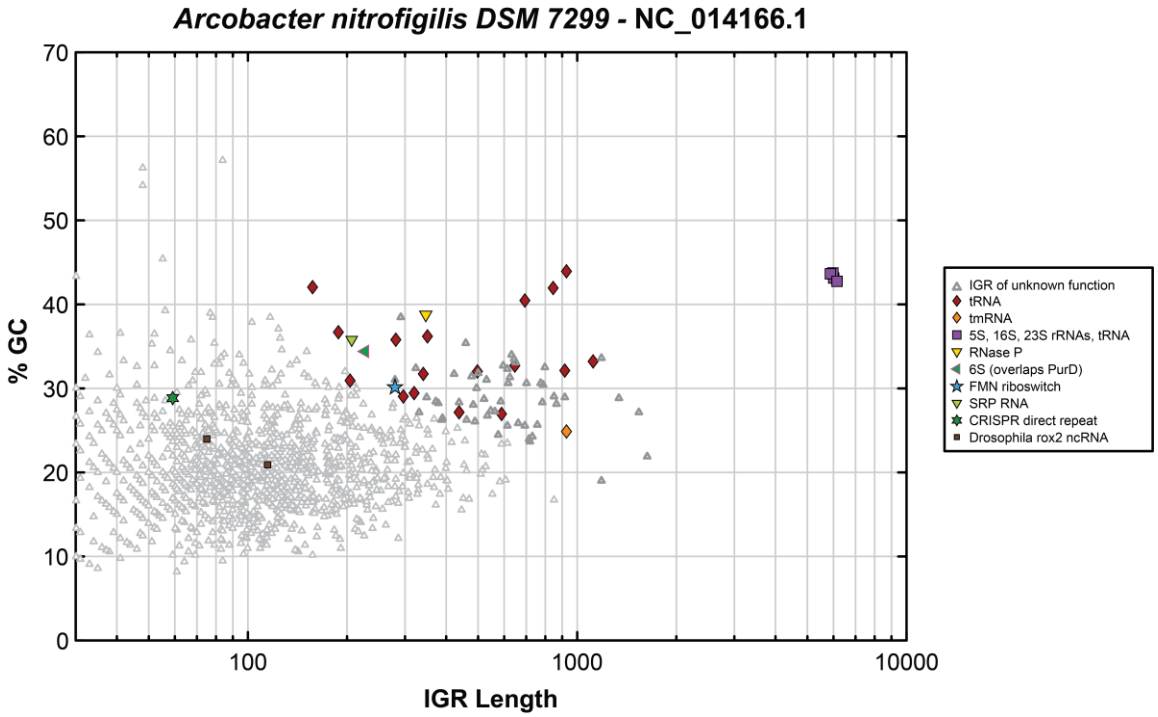

Figure 2-23: Plot of the IGRs from the *A. nitrofigilis* genome.

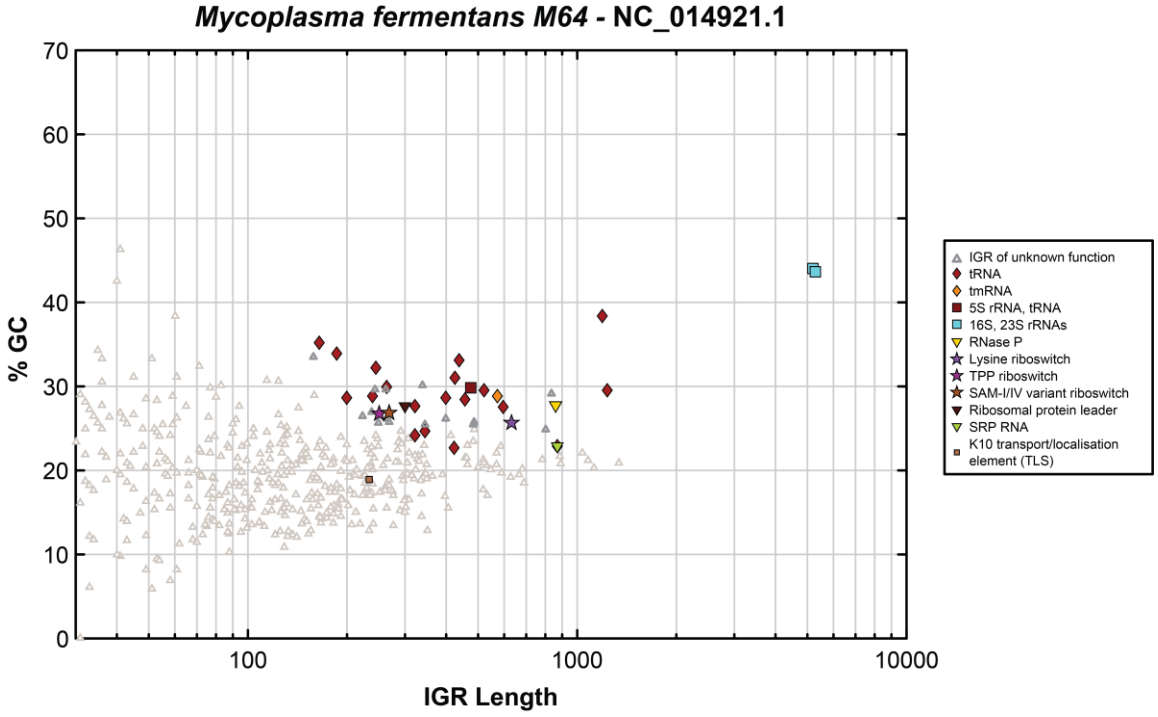

Figure 2-24: Plot of the IGRs from the *M. fermentans* genome.

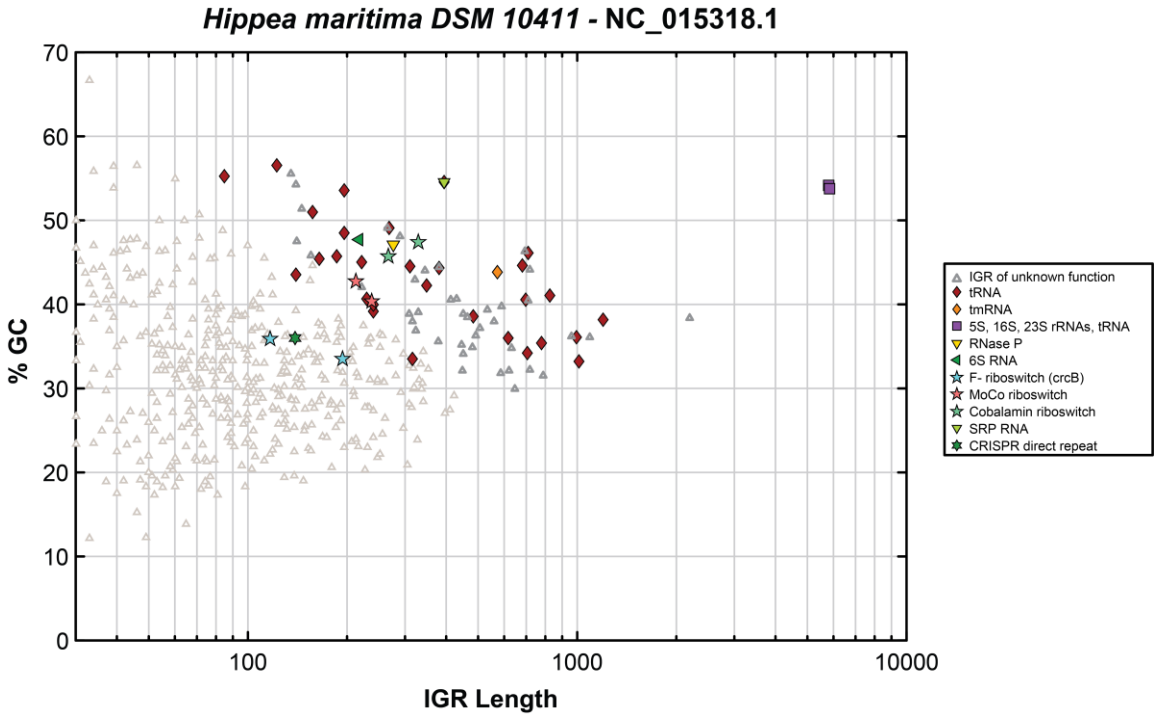

Figure 2-25: Plot of the IGRs from the *H. maritima* genome.

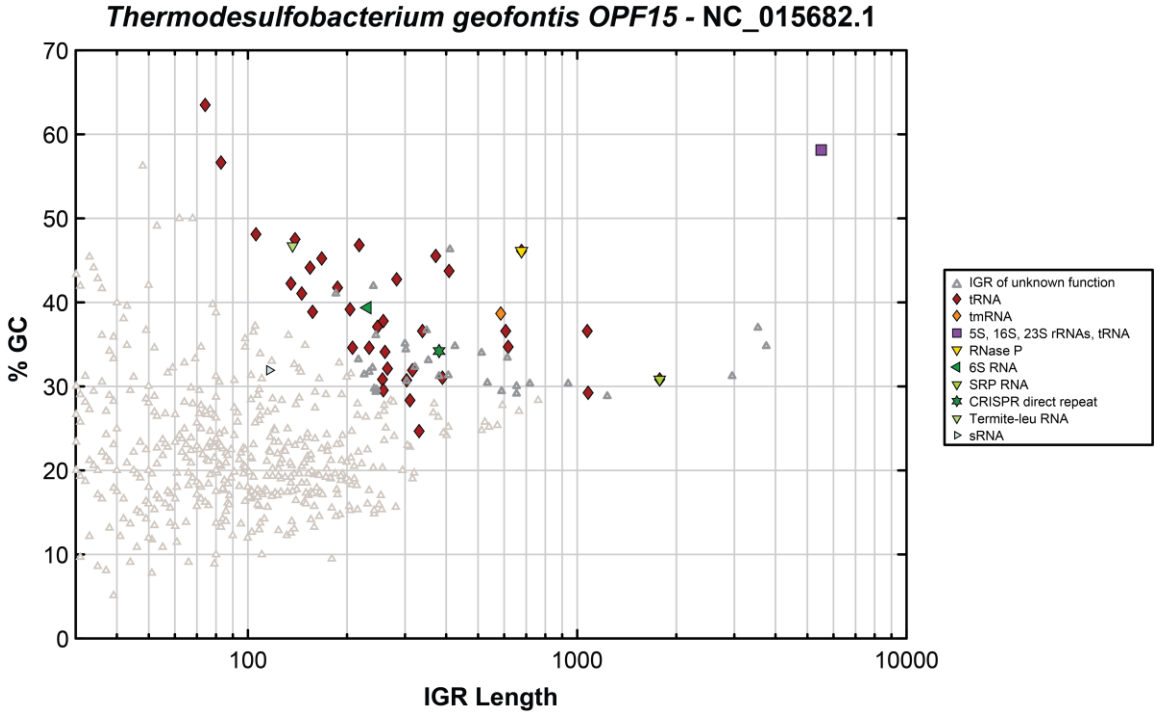

Figure 2-26: Plot of the IGRs from the *T. geofontis* genome.

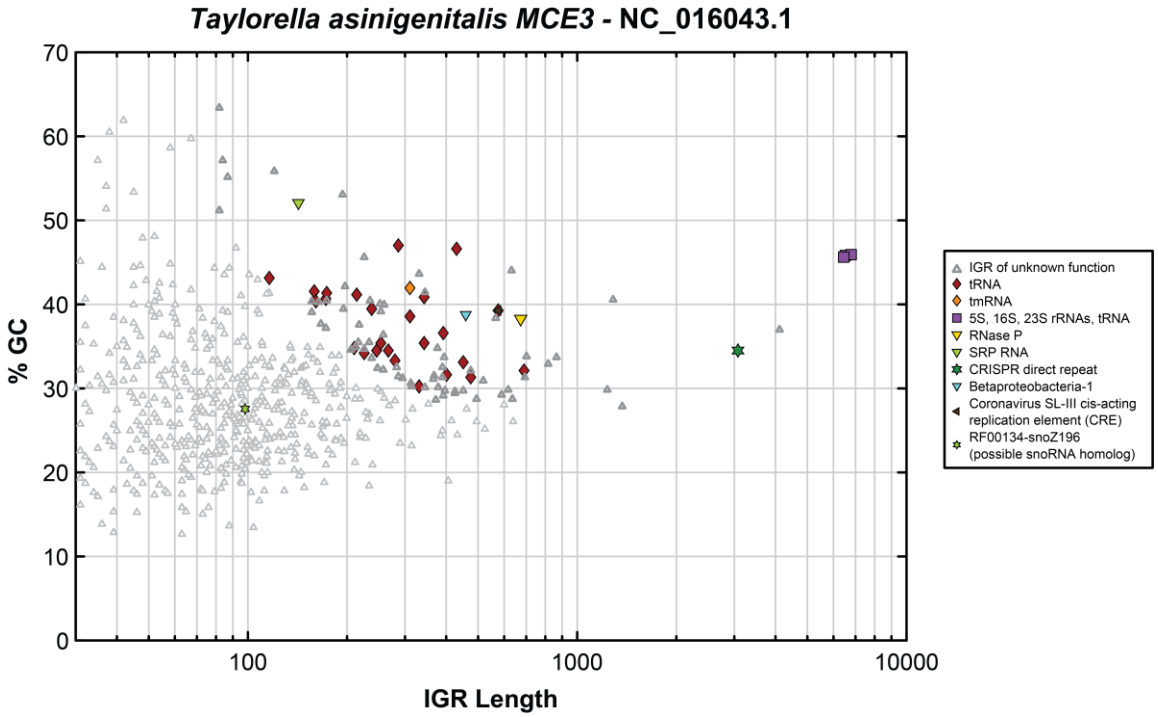

Figure 2-27: Plot of the IGRs from the *T. asinigenitalis* genome.

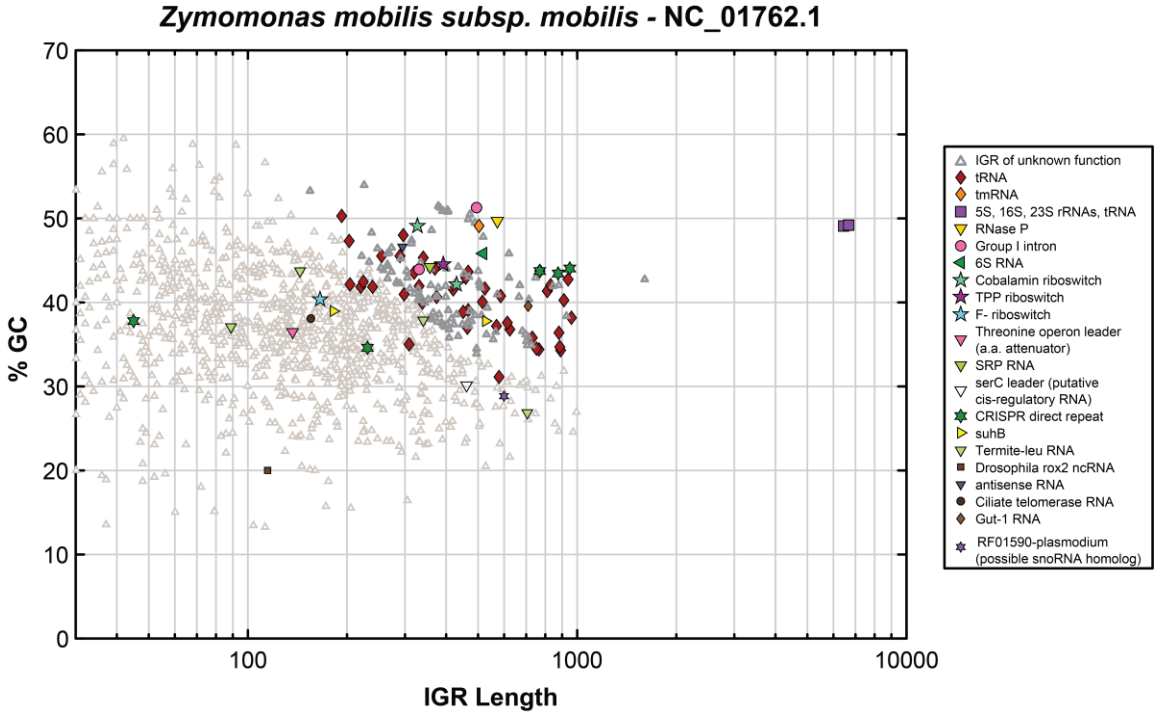

Figure 2-28: Plot of the IGRs from the *Z. mobilis* genome.

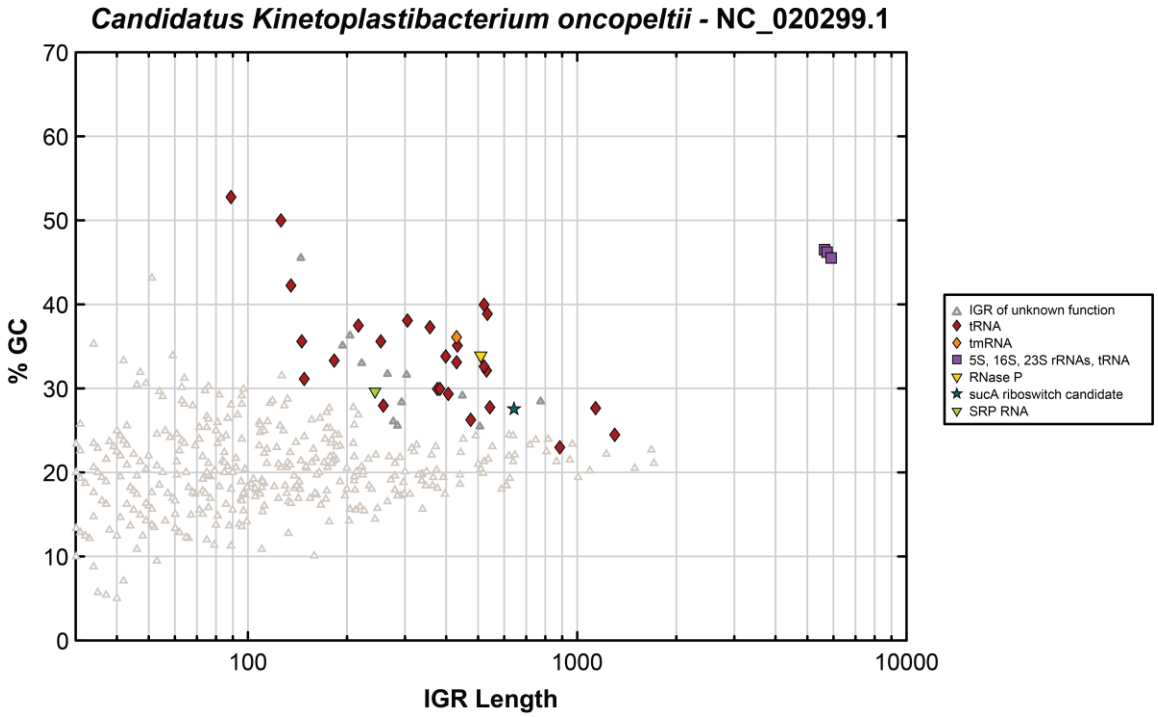

Figure 2-29: Plot of the IGRs from the *C. K. oncopelti* genome.

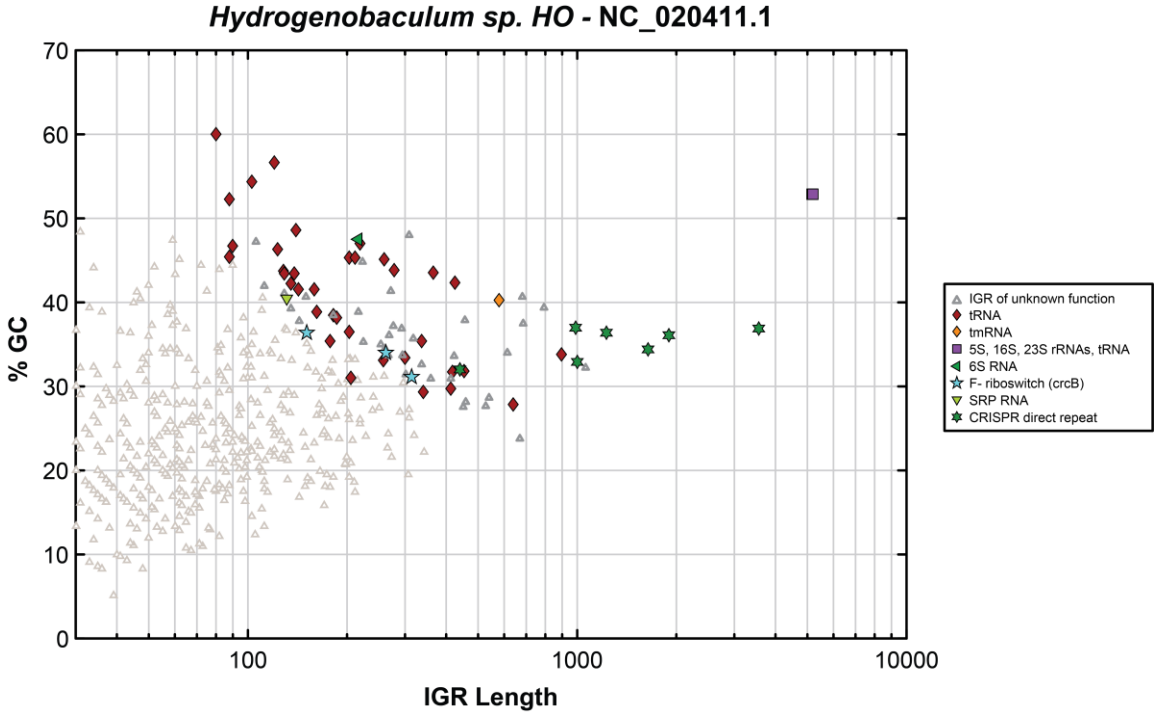

Figure 2-30: Plot of the IGRs from the *Hydrogenobaculum* genome.

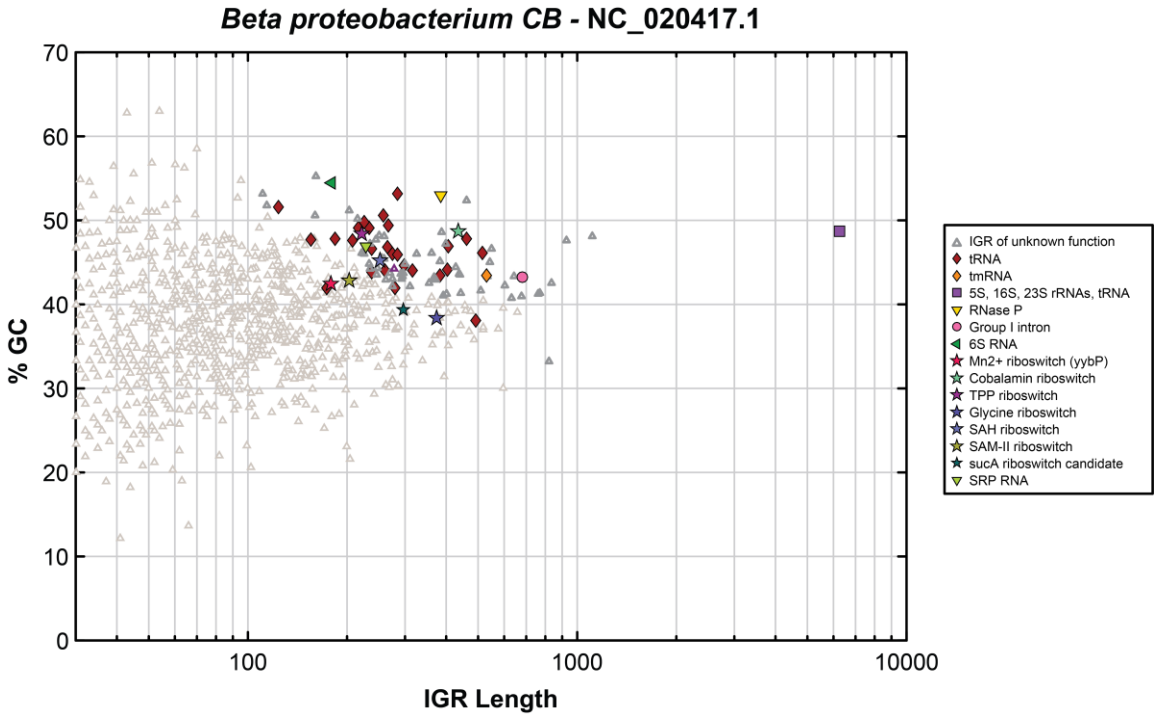

Figure 2-31: Plot of the IGRs from the *B. proteobacterium* genome.

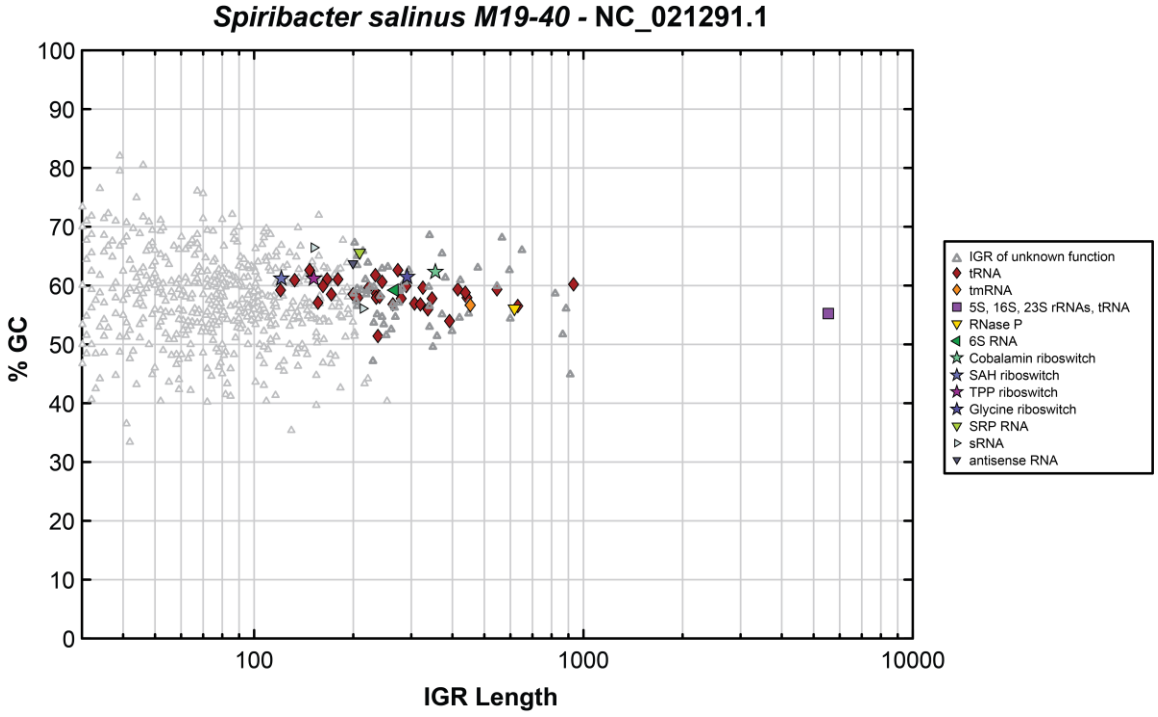

Figure 2-32: Plot of the IGRs from the *S. salinus* genome.

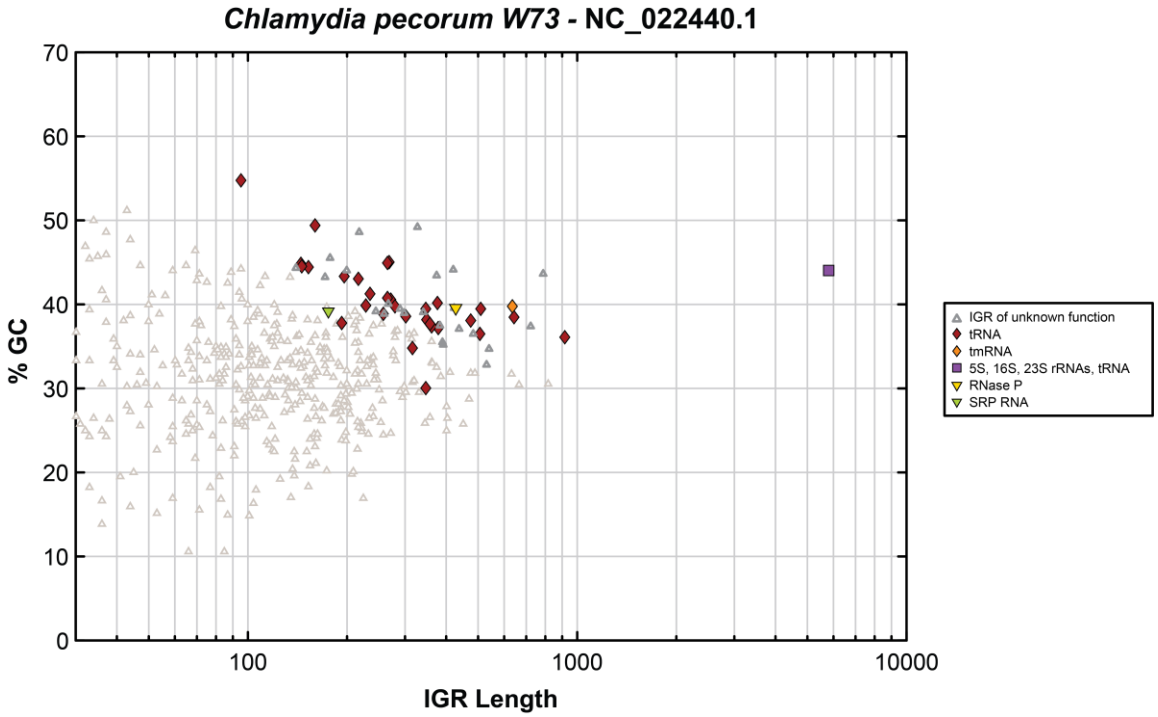

Figure 2-33: Plot of the IGRs from the *C. pecorum* genome.

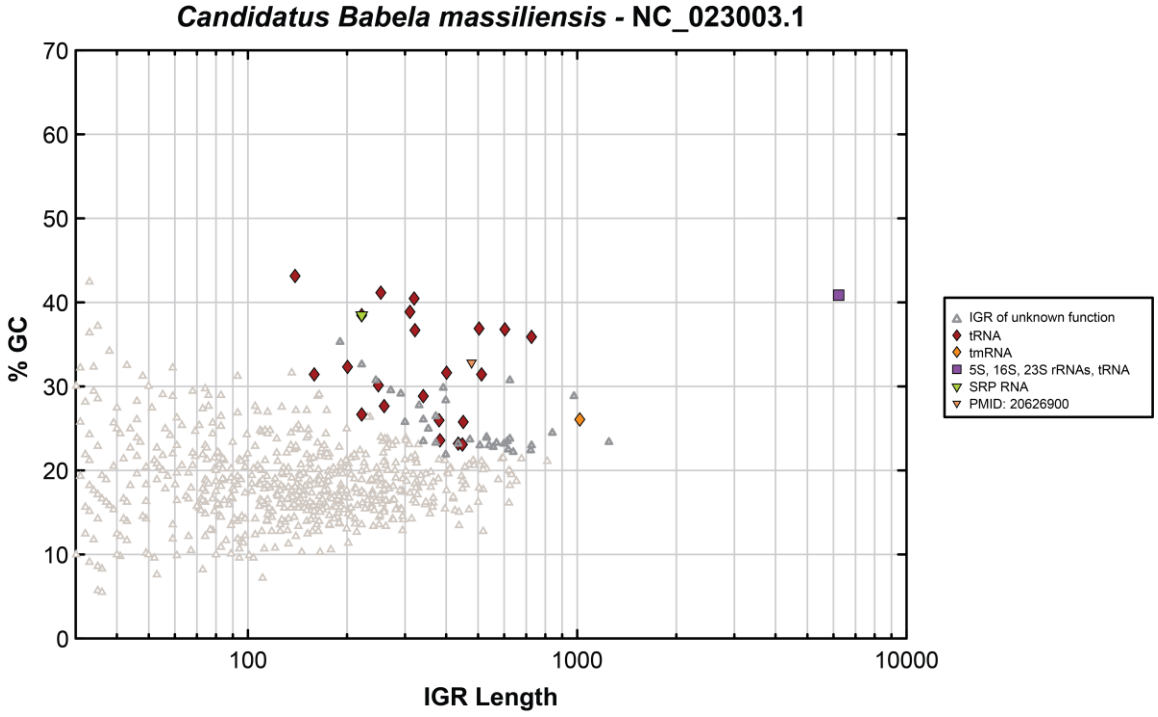

Figure 2-34: Plot of the IGRs from the *C. B. massiliensis* genome.

Chapter Three

DIMPL: A bioinformatics pipeline for the discovery of structured noncoding RNA motifs in bacteria

Largely adapted from the following manuscript:

Brewer, K. I., Gaffield, G. J., Puri, M. & Breaker, R. R. DIMPL: A Bioinformatics pipeline for the discovery of structured noncoding RNA motifs in bacteria. *Manuscript in Revisions* (2021).

Summary

Recent efforts to identify novel bacterial structured noncoding RNA (ncRNA) motifs through searching long, GC-rich intergenic regions (IGRs) have revealed several new classes, including the recently validated HMP-PP riboswitch. The DIMPL discovery pipeline described herein enables rapid extraction and selection of bacterial IGRs that are enriched for structured ncRNAs. Moreover, DIMPL automates the subsequent computational steps necessary for their functional identification.

Introduction

Discovery and validation of the over 45 known classes of metabolite- or elemental ion-binding riboswitches[21] have relied extensively on large-scale computational approaches based on comparative sequence analysis[57,60,61]. However, these large-scale approaches may struggle to identify new classes of riboswitches, which are predicted to exist by the thousands but are likely much rarer than known classes[14,21]. Genome-level filtering of bacterial intergenic regions by nucleic acid composition and length[72,74] was developed to address the challenges of discovering these rarer riboswitch classes. This approach has already enabled the discovery and validation of the SAM-V[72,73], HMP-PP[30], and NAD-II[25] riboswitch classes and the discovery of dozens of new intergenic motif candidates in the first genomes analyzed. However, until now this approach has required time-consuming manual analysis using several bioinformatic tools and lacked well-defined techniques to define genomic regions for further analysis that are enriched for noncoding RNAs.

In this chapter, I introduce DIMPL (Discovery of Intergenic Motifs PipeLine), a bioinformatics pipeline which automates the process of total genome analysis by extracting

intergenic regions, filtering them by length and nucleic acid composition, and collecting the data necessary to identify candidate motifs and assign their possible functions. DIMPL also provides reproducible techniques for identifying genomic regions enriched for ncRNA through support vector machine (SVM) classifiers. Although my primary objective in creating DIMPL was to accelerate the discovery of novel riboswitch classes, it can also be used to identify a wide-range of other intergenic nucleic acid and protein motifs such as upstream open reading frames (uORFs), short open reading frames (sORFs), ribosomal protein leader sequences, selfish genetic elements and other structured RNA motifs of unknown function.

Results and Discussion

Pipeline Overview

The DIMPL computational pipeline consists of two primary stages: 1) genome analysis, and 2) draft motif analysis. The genome analysis stage of DIMPL starts by integrating the genomic sequence and protein annotations (**Figure 3-1A**) accessible via NCBI Entrez[135] with corresponding RNA family annotations provided by the Rfam MySQL Database[132]. All intergenic regions (IGRs) located between protein-coding open reading frames (ORFs) are then extracted and labeled (**Figure 3-1B-C**) with their percentage of G and C nucleotides relative to the total nucleotides in the IGR (%GC content), length, and the presence of any known ncRNA motifs. DIMPL then generates an interactive graph (**Figure 1D**) showing the IGRs plotted by their %GC content and length with labels for IGRs with known RNA families. In the next step, the pipeline uses a support vector machine (SVM)

classifier (**Figure 3-1E**) to identify IGRs with no annotated ncRNAs that have similar features to other IGRs with known structured ncRNAs. DIMPL then performs a BLASTX search[76] on the selected IGRs to ensure they do not contain unannotated protein coding regions. Any unannotated protein coding regions discovered in the search are removed from the selected IGRs, which are discarded in their entirety if the remaining IGR no longer meets the length and %GC content requirements for the selection.

The draft motif analysis portion of DIMPL is performed in parallel on all IGRs that have met the selection criteria. The process begins by using Infernal 1.1.3[63] to search each selected IGR's sequence (**Figure 3-1F**) against a database of all microbial intergenic regions derived from NCBI's RefSeq[75]. The collection of homologous sequences from a single IGR search forms the 'draft motif' that is further analyzed in several steps. First, representatives with identical nucleotide sequences are removed. Next, the draft motifs are analyzed via CMfinder 0.4.18[62] to look for possible RNA secondary structure features (**Figure 3-1G**). All realigned motifs generated by CMfinder are evaluated for evidence of statistical significance for predicted nucleotide covariations. Subsequently, the consensus sequence and structural model for each motif is generated (**Figure 3-1H**) using R-scape 1.4.0[136], which integrates the RNA drawing algorithm R2R[134]. Draft motifs are also checked for the presence of coding regions using RNAcode[137]. Finally, for each draft motif, DIMPL uses GenomeView[138] to visualize the genetic contexts (**Figure 3-1I**) of the motif's representatives to aid in determining a possible function for the candidate RNA motif. A draft motif's most strongly supported alignment can then be analyzed by one or more additional cycles of Infernal homology searches, which take advantage of the proposed secondary structure to expand the number of representatives found.

Details on SVM Enrichment

The SVM enrichment of IGRs in DIMPL uses a radial basis-function (RBF) kernel and is implemented with scikit-learn[139]. The SVM classifier is trained *de novo* for each genome analyzed using the IGR %GC content and nucleotide length as the features, the presence/absence of a structured RNA as the class labels and a set of hyperparameters that have been weighted to select a contiguous region of a genome's %GC versus length plot. The primary purpose of the SVM classifier is to perform an enrichment of IGRs that reduces the number subjected to the more computationally intensive steps in the pipeline. Applying the SVM-RBF algorithm allows DIMPL to accomplish this goal in a systematic and reproducible manner.

Usage Guide

The DIMPL pipeline is built primarily in Python and is distributed as a Docker image[140] with all the necessary tools already installed. Along with the Docker image, DIMPL includes a set of detailed Jupyter notebooks that walk users through the steps of the pipeline, display interactive graphs and assemble results from analysis tools. For computationally intensive steps such as BLAST, Infernal and CMfinder that are typically performed on a high-performance computing cluster, DIMPL exports compressed tar files containing the necessary bash scripts and data files that can be configured for a custom compute environment. Detailed instructions are included below.

Workstation Setup and Installation

Prerequisites:

- Install the Docker program (https://www.docker.com/products/docker-desktop)
- Ensure Python version 3 is installed on your machine.
- Ensure you have an NCBI account and API Key.

Setup

- Download the 5.7 GB Docker image for dimpl using the command:

```
docker pull breakerlab/dimpl
```

- Configure the settings of docker to allow the virtual machine to access local files:
    - Docker settings → Resources → File sharing → Select the local drives you want to be available to the docker image.
- Download the archive of DIMPL source from github.com/breakerlab/dimpl and extract to folder accessible to docker.

```
wget https://github.com/BreakerLab/dimpl/archive/v1.0.0.tar.gz
tar xzvf v1.0.0.tar.gz
```

- There will now be a dimpl folder in this repository.

```
cd dimpl-1.0.0; ls
```

Starting DIMPL

- To start dimpl, run the start.sh script:

```
./start.sh
```

    - For first time DIMPL users, a prompt will appear asking for NCBI account information (email address and API key). This information is found in your NCBI account, in account settings. If you do not already have an API key created, you can create one on this page.
- A URL will be output to access the jupyter notebook: (e.g. localhost:8888/lab… etc.). Copy and paste this url into your web browser.

Using Jupyter Notebooks

By default, DIMPL launches a JupyterLab interface that includes an integrated file browser, terminal, and access to the three computational notebooks that walk users through the DIMPL analysis pipeline. The "work" directory contains all the folders accessible to DIMPL from your local machine. The three main pipeline notebooks can be found in the "notebooks" directory. Each notebook is intended to be run one after the other, running each python code block in order and following the built-in instructions about which variables should be modified. An overview of each of the three computational notebooks is included here with more detailed instructions built into each notebook itself.

Compute Station Setup

Compute Station Prerequisites

The following programs need to be installed and available on each node of the cluster that you plan to use:

- Slurm[141] Job scheduler
- Dead Simple Queue[142] Job array tool
- Infernal[63] Covariance model search tool
- BLAST+[76] Sequence similarity search tool
- CMfinder[62] RNA motif predictor

Data to be processed (tar files) needs be placed in a location that is shared by all compute nodes.

Memory required per node:

- Blast step 70GB
- Infernal step 32GB

Downloading Necessary Files

- Download the IGR search database (filename: s50.igr.fasta) from the link on the DIMPL Github to your cluster using Globus FTP[143]. This file is 91GB.
- Ensure the availability of the BLAST "nr" database on your computational cluster. This database is currently 547GB.

Modifying Configuration File

The cluster configuration file found at dimpl/src/shell/cluster.conf is used to modify your PATH environmental variables on your cluster such that all of the utilities mentioned under "Prerequisites" can be run. It also defines a few variables that define your cluster partition(queue) and database locations. Although this file is used on the compute cluster, it should be modified within the workstation dimpl installation as the config file is packaged with each tarfile of computational instructions assembled by DIMPL.

Notebook 1: Genome IGR Selection

Introduction

The purpose of DIMPL Notebook 1 is to walk the user through the steps of selecting a bacterial genome, extracting and graphing the intergenic regions, and refining the Support Vector Machine (SVM) selection parameters used to enrich the intergenic regions for structured non-coding RNAs.

Step 1: Review bacterial/archaeal genomes annotated by Rfam

Run the first code block to make all the necessary python imports, then run the second code block to create a connection to the Rfam mySQL database[132] and request a list of annotated genomes from the Bacteria and Archaea kingdoms. This code block demonstrates some of the various filtering capability for genome selection that can be accomplished using the built-in functionality of the SQLAlchemy classes used to query Rfam's database. By default, this code block will output a table with the information for the first 10 genomes, but this can be modified by changing the indices in the code block. Take note of the Uniprot ID (UPID) of the genome you wish to analyze and enter it in the code block below.

Step 2: Extract and graph IGRs

For this example, we will select the genome of *Campylobacter jejuni* (UPID: UP000000799). Replace the `upid` variable in code block 3 with the value for this genome and then run the code block. This code block will automatically identify the correct genome from NCBI, download the necessary files, extract the genome's IGRs, annotate the IGRs using data available from Rfam and then generate the plot of the IGRs as shown in **Figure 3-2.** This particular genome shows some of the ideal characteristics for genomes that are

relatively easy to analyze using the GC-IGR analysis approach. There is a relatively low overall GC content, which results in the IGRs containing known structured ncRNAs having good separation on the plot from other IGRs.

Step 3: SVM selection of IGRs enriched for ncRNAs.

Running the next code block will create a similar IGR plot for the selected genome, but will use a radial basis function (RBF) kernel with a support vector machine classifier to create a selection of intergenic regions that are enriched for structured noncoding RNAs (**Figure 3-3**). Above the plot are sliders for adjusting three hyperparameters that go into defining the IGR selection: 1) *class_weight,* 2) *gamma_exp* and 3) *c_exp*. Researchers are seeking a greater understanding of the role of these hyperparameters in SVM classifiers are advised to consult the documentation of Scikit-learn[139]. In short, a proper selection of hyperparameters is necessary to prevent overfitting while training an SVM classifier. DIMPL performs the SVM classifier training on a genome-by-genome basis, however, we have found that the default hyperparameters shown in this code block generate reasonable enrichment for a wide-range of genomes. Generally, we advise users to rely on these default parameters unless analyzing a genome with truly unusual IGR characteristics.

Step 4: Finalize IGR Selection and export blast tarfile

Running the final two code blocks in Notebook 1 will re-create the selection and write the data into a comma separated value file called 'annotated_igrs.csv'. DIMPL will then generate appropriate fasta files and scripts necessary to perform a BLAST search on the selected IGRs to look for unannotated ORFs. These fasta files and scripts are packaged into a tarfile that is placed in DIMPL's data/export folder. Transfer this tarfile to your computational cluster, unzip the tarfile and execute the script "`./blast_run.sh`" to start

the series of blast searches on the selected IGRs. Progress of all the different tasks can be monitored in slurm using the "`squeue -u username`" command. Once all computational tasks are completed, run the "`./make_tar.sh`" command to package the results folder for the next processing step.

Notebook 2: BLAST data processing

Introduction

The purpose of DIMPL's notebook 2 is to process the BLAST results generated using Notebook 1 to identify IGRs that contain unannotated known ORFs and should either be discarded or trimmed. After importing the Blast results, DIMPL determines how IGRs need to be trimmed by calculating a customizable "`orf_score`" at each nucleotide position in the IGR. This `orf_score` is a sum of quantity and quality of BLAST hits that overlap that position. After removing all portions of an IGR that meet the `orf_score` threshold, the notebook checks if any of the remaining IGR or IGR fragment still falls within the original IGR selection boundary. Finally, the notebook builds a new set of scripts required to start the Infernal searches on each remaining IGR and assembles those into a tarfile.

Step 1: Import BLAST search files

Transfer the Blast results tarfile generated at the end of Notebook 1's instructions and place it in the `data/import` folder. After running code block 1, change the variables in code block 2. The `import_tar_name` should be changed to the name of the tarfile placed in `data/import`. For the *Campylobacter jejuni* example described above the `assembly_acc` variable should be set to "GCA_000009085.1" and `selection_name` may be left

unchanged. Code blocks 2, 3, and 4 should then be run to extract the BLAST results from the tarfile, and rebuild the original selection used for a genome.

Step 2: Process BLAST results

Run code block 5 to analyze the blast results for each IGR and calculate an `orf_score` at each nucleotide position as depicted in **Figure 3-4**. The parameters used for trimming IGR portions can be customized by changing the arguments of the `process_blast()` function in this code block. For poorly annotated genomes where the vast majority of IGRs contain unannotated ORFs, there may be very few of the resulting trimmed IGRs that meet the original GC/length selection criteria as defined by the SVM decision surface. In these cases, selection criteria may be relaxed by decreasing the `svm_decision_cutoff` from 0 to a slight negative value such as -0.3. DIMPL will visualize the impact of the slightly relaxed selection criteria by replotting the selected genome with the revised criteria. Any processed IGRs that fall in the highlighted area according to their recalculated length and GC content will be included in the subsequent steps.

Step 3: Build data and script tarfiles

Run code block 6 to assemble all of the processed IGRs into a new tarfile along with the necessary scripts to perform an Infernal[63] search on each IGR. You may customize the `step_dir` variable if you anticipate performing multiple rounds of infernal searches on this genome each of which can be placed into an appropriately named subfolder. Note that this codeblock calls the `build_infernal_commands()` function with the `no_secondary_structure` parameter set to "True". This parameter instructs Infernal to use its nhmmer fallback option to handle these alignment files which contain only one sequence each and no predicted secondary structure. After running code block 6, find the

resulting tarfile in the `data/export` folder, transfer it to your high-performance computing cluster, unpack the tar archive, and then run the command `./infernal_step1_run.sh`. After all the computational steps have completed, assemble the data using the `./make_tar.sh` command and transfer it back to DIMPL for analysis in Notebook 3.

Notebook 3: IGR Report

Introduction

The purpose of Notebook 3 is to display and assemble the results of the Infernal search for each IGR from your selected genome. In addition to showing a results table with the number of hits and their locations, this notebook will display a variety of possible structures for each motif candidate predicted using CMfinder[62] and R-scape[136]. It will also run an analysis using RNACode[137] to check if the collection of hits is predictive of a novel protein-coding region. Finally, DIMPL generates genome context images that show the predominant gene associations upstream and downstream of each representative. The combination of structural information and genetic context information aids in the categorization of each motif candidate.

Step 1: Import infernal search results

Place the tarfile generated at the end of the instructions for Notebook 2 in the `data/import` folder. Run code block 1 and then change the variables in code block 2 to reflect the information for the genome being imported just as previously described for Notebook 2. Running code blocks 2 and 3 will generate a drop-down menu with an option for each IGR's search results. Select the search results you are interested in viewing from this menu.

Step 2: View results table

Run code block 4 to generate an abbreviated results table that shows the number of unique representatives found for this motif along with the e-values and locations of these hits.

Step 3: View possible structures predicted by Rscape and CMFinder

Running code block 5 will generate a structure prediction using R-scape's CaCoFold (Cascade variation/covariation Constrained Folding) algorithm[144]. The next code block will display the collection of possible submotifs predicted by CMFinder. The most well-supported structure may be one of these submotifs that might only be present in a subset of the homologous sequences that Infernal initially found. After viewing all of the possible structures, select the most promising structure and find the corresponding Stockholm alignment file in DIMPL's folder architecture for further analysis.

Step 4: Perform RNAcode analysis to search for possible protein-coding regions

Run code block 7 to start an RNAcode analysis on this motif candidate to check if there are indications of novel protein-coding regions in this motif.

Step 5: View motif's genome context

Running code block 7 will generate genome context images (**Figure 3-5**) for the motif. This allows a user to view the genes present upstream and downstream of each representative. A link to the NCBI genome browser of the genomic location is also provided which allows easy access to more detailed information about neighboring genes.

Notebook 4: Motif Refinement

Introduction

The purpose of DIMPL's notebook 4 is to perform follow-up searches on the initial motifs generated.

Conclusion

DIMPL provides an integrated collection of tools to streamline the process of identifying novel structured ncRNA motifs, including new riboswitch candidates, on a genome-wide scale. It relies on established methods of enriching bacterial IGRs for ncRNA motif discovery and quickly assembles the combination of structural and genetic context information that are key to identifying the function of the newly discovered motifs. This pipeline should permit the rapid analysis of bacterial genomes for novel and rare ncRNA classes and can help accelerate the pace of riboswitch and ribozyme discovery.

Tables and Figures

Figures

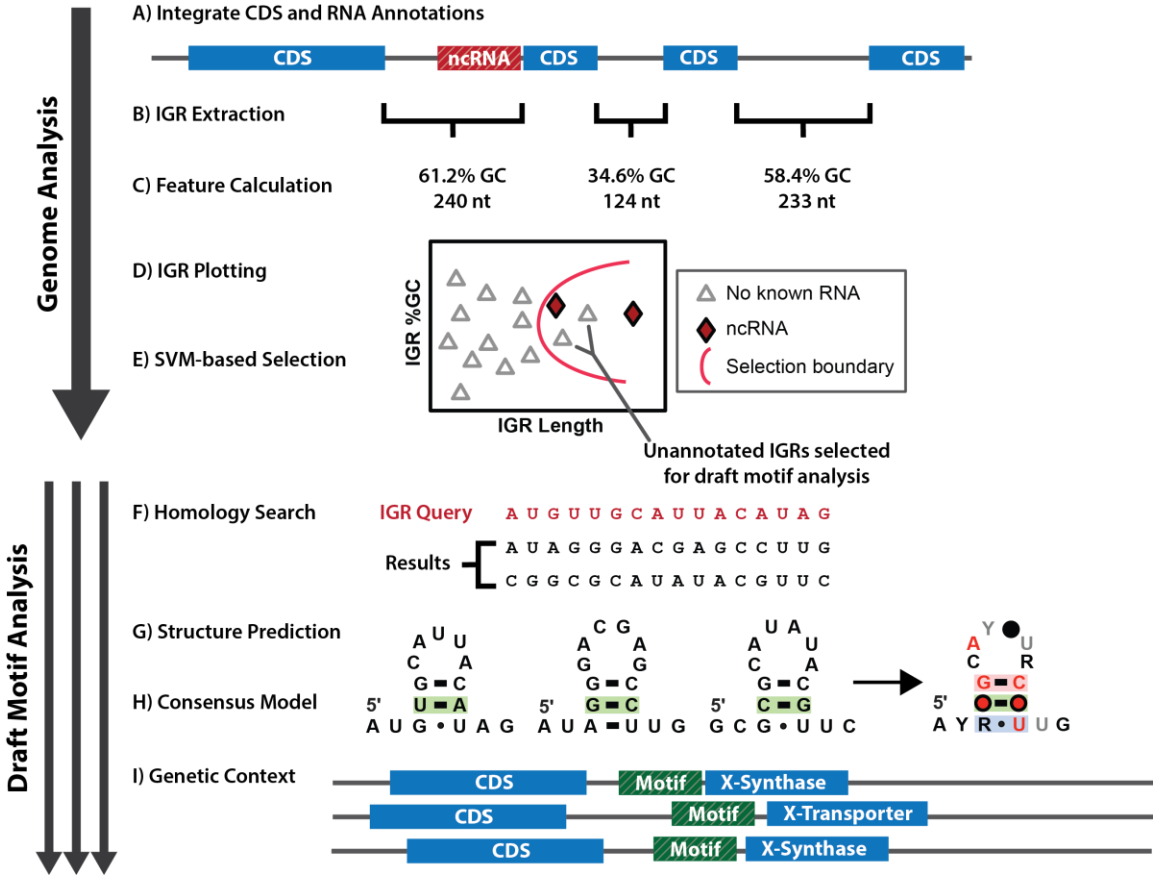

Figure 3-1: Overview of DIMPL Process

Overview of DIMPL process divided into two stages: genome analysis (**A-E**) and draft motif analysis (**F-I**).

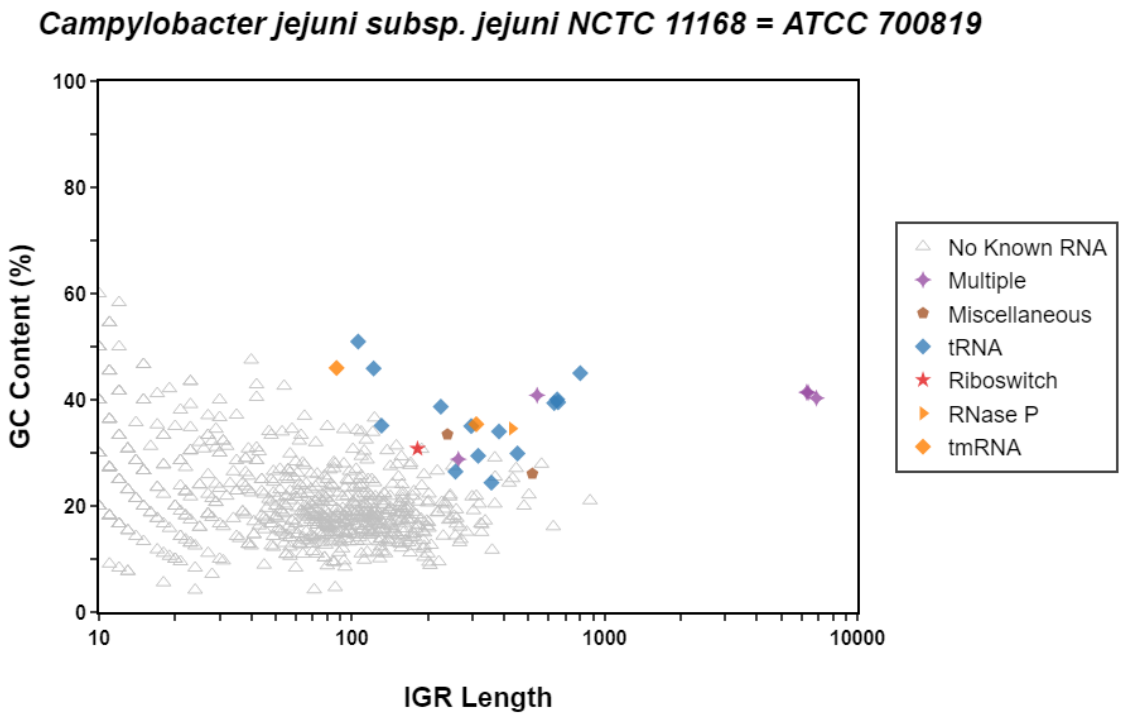

Figure 3-2: Plot of IGRs from the genome *Campylobacter jejni* generated by DIMPL. Example of the genome plot generated by DIMPL When viewed in the DIMPL's ingegrated Jupyter notebook this plot is interactive and specific IGRs annotations can be viewed by hovering over relevant points.

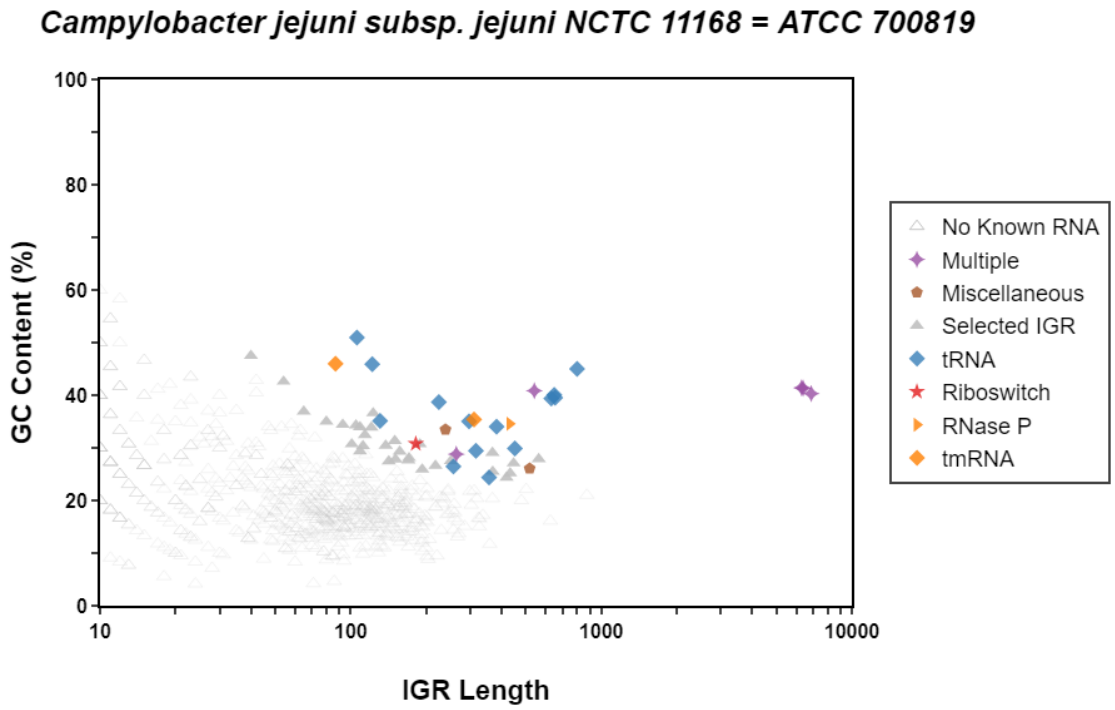

Figure 3-3: Plot of IGR selection from *Campylobacter jejuni* enriched by DIMPL

In contrast to **Figure 3-2**, this figure has IGRs with no annotated ncRNA highlighted if they are identified by DIMPL as having similar GC and length characteristics as those IGRs that do contain structured ncRNAs. This selection was generated using the default DIMPL hyperparameters of 0.50 for class_weight. 2.00 for c_exp and -2.00 for gamma_exp. In the Jupyter notebook that generates these selections, DIMPL also displays statistics about the selection including the number of IGRs with known ncRNAs included in the selection area (25 known ncRNAs out of 25 total ncRNAs or 100%), the number of unknown IGRs included in the selection regions (30 selected unknown IGRs out of 1008 total unknowns or 3.0%) and the fold enrichment for the selection (17.98 greater likelihood of finding ncRNA in selected region vs a random genomic IGR).

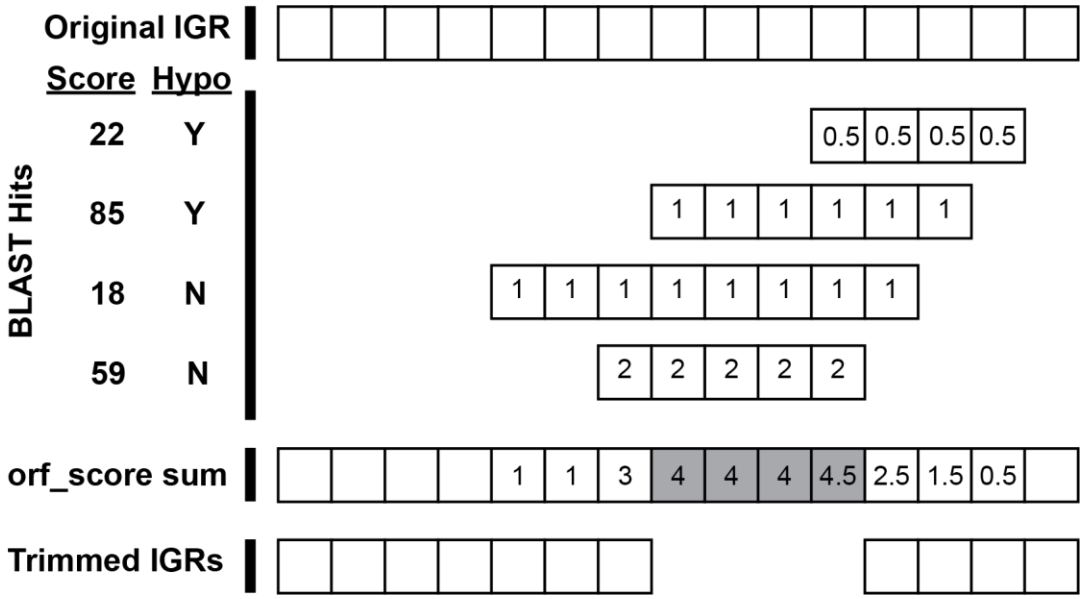

Figure 3-4: Graphical depiction of blast hit processing with default parameters

In this toy example, the BLAST search results for a 15 nt long IGR include four hits, each of which has data about the score, annotation, and region of overlap. Each hit is initially worth `score_increment` points (2) but that is multiplied by the `poor_score_weight` (0.5) if the quality of the hit is below the `poor_blast_score` (40) threshold. This is further multiplied by `hypothetical_weight` (0.5) if the hit includes variations of "hypothetical" in the annotation text. The sum of the resulting points at each nucleotide position is calculated and the central portion of the IGR which has an `orf_score` greater than `orf_score_cutoff` (4) is removed giving two new trimmed IGRs whose %GC content and length can be compared to the IGR selection parameters for possible inclusion.

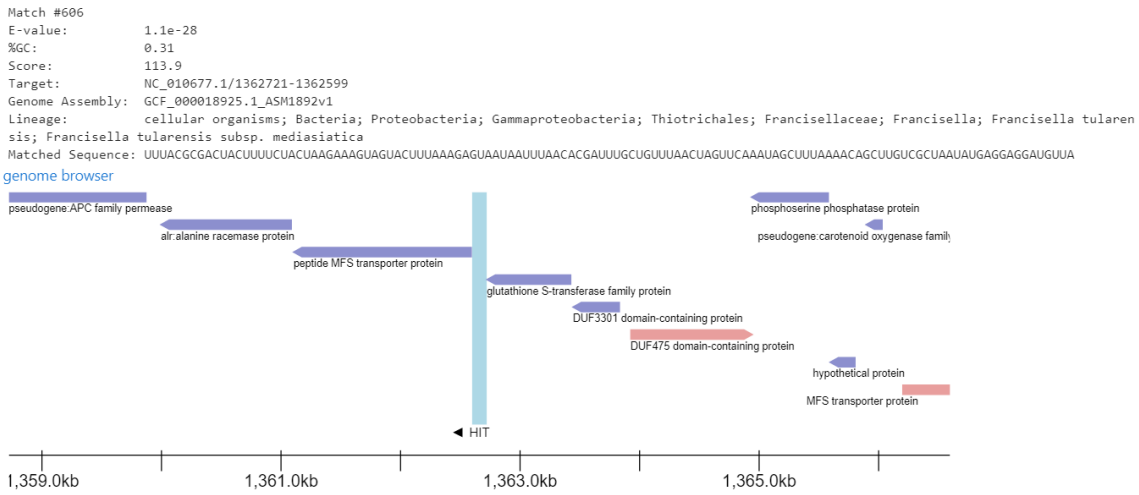

Figure 3-5: Sample hit from genetic context report generated by DIMPL

For each unique representative identified via an Infernal search DIMPL generates a genetic context report listing the e-value, genomic location, organism name and the sequence of the hit at that location. The motif is located blue-shaded region and is oriented with the 3'end towards the right. If the annotated gene names are uninformative additional detail is available by clicking the automatically generated link that will display the appropriate location in NCBI's online genome browser[135].

References

1. Crick, F. H. C. On protein synthesis. in *Symp Soc Exp Biol* vol. 12 8 (1958).
2. Cobb, M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol.* **15**, e2003243 (2017).
3. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).
4. Shapiro, J. A. Revisiting the central dogma in the 21st century. *Ann. N. Y. Acad. Sci.* **1178**, 6–28 (2009).
5. Thieffry, D. & Sarkar, S. Forty years under the central dogma. *Trends Biochem. Sci.* **23**, 312–316 (1998).
6. Eddy, S. R. Non–coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2**, 919–929 (2001).
7. Kazantsev, A. V & Pace, N. R. Bacterial RNase P: a new view of an ancient enzyme. *Nat. Rev. Microbiol.* **4**, 729–740 (2006).
8. Stark, B. C., Kole, R., Bowman, E. J. & Altman, S. Ribonuclease P: an enzyme with an essential RNA component. *Proc. Natl. Acad. Sci.* **75**, 3717–3721 (1978).
9. Pool, M. R. Signal recognition particles in chloroplasts, bacteria, yeast and mammals. *Mol. Membr. Biol.* **22**, 3–15 (2005).
10. Walter, P. & Blobel, G. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature* **299**, 691–698 (1982).
11. Cavanagh, A. T. & Wassarman, K. M. 6S RNA, a global regulator of transcription in Escherichia coli, Bacillus subtilis, and beyond. *Annu. Rev. Microbiol.* **68**, 45–60 (2014).

12. Doherty, E. A. & Doudna, J. A. Ribozyme structures and mechanisms. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 457–475 (2001).

13. Jimenez, R. M., Polanco, J. A. & Lupták, A. Chemistry and biology of self-cleaving ribozymes. *Trends Biochem. Sci.* **40**, 648–661 (2015).

14. Breaker, R. R. Prospects for Riboswitch Discovery and Analysis. *Molecular Cell* vol. 43 867–879 (2011).

15. Serganov, A. & Nudler, E. A decade of riboswitches. *Cell* **152**, 17–24 (2013).

16. Sherwood, A. V & Henkin, T. M. Riboswitch-mediated gene regulation: novel RNA architectures dictate gene expression responses. *Annu. Rev. Microbiol.* **70**, 361–374 (2016).

17. Rich, A. On the problems of evolution and biochemical information transfer. *Horizons Biochem.* 103–126 (1962).

18. White, H. B. Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.* **7**, 101–104 (1976).

19. Powner, M. W., Gerland, B. & Sutherland, J. D. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **459**, 239–242 (2009).

20. Baker, J. L. *et al.* Widespread genetic switches and toxicity resistance proteins for fluoride. *Science (80-.).* **335**, 233–235 (2012).

21. McCown, P. J., Corbino, K. A., Stav, S., Sherlock, M. E. & Breaker, R. R. Riboswitch diversity and distribution. *RNA* **23**, 995–1011 (2017).

22. Mirihana Arachchilage, G., Sherlock, M. E., Weinberg, Z. & Breaker, R. R. SAM-VI RNAs selectively bind S-adenosylmethionine and exhibit similarities to SAM-III riboswitches. *RNA Biol.* **15**, 371–378 (2018).

23. Lenkeit, F., Eckert, I., Hartig, J. S. & Weinberg, Z. Discovery and characterization of a fourth class of guanidine riboswitches. *Nucleic Acids Res.* **48**, 12889–12899 (2020).

24. Salvail, H., Balaji, A., Yu, D., Roth, A. & Breaker, R. R. Biochemical Validation of a Fourth Guanidine Riboswitch Class in Bacteria. *Biochemistry* (2020).

25. Panchapakesan, S. S. S., Corey, L., Malkowski, S. N., Higgs, G. & Breaker, R. R. A second riboswitch class for the enzyme cofactor NAD+. *RNA* **27**, 99–105 (2020).

26. Sherlock, M. E., Sudarsan, N., Stav, S. & Breaker, R. R. Tandem riboswitches form a natural Boolean logic gate to control purine metabolism in bacteria. *Elife* **7**, e33908 (2018).

27. Sherlock, M. E., Sudarsan, N. & Breaker, R. R. Riboswitches for the alarmone ppGpp expand the collection of RNA-based signaling systems. *Proc. Natl. Acad. Sci.* **115**, 6052–6057 (2018).

28. Sherlock, M. E., Sadeeshkumar, H. & Breaker, R. R. Variant bacterial riboswitches associated with nucleotide hydrolase genes sense nucleoside diphosphates. *Biochemistry* **58**, 401–410 (2018).

29. Atilho, R. M., Perkins, K. R. & Breaker, R. R. Rare variants of the FMN riboswitch class in Clostridium difficile and other bacteria exhibit altered ligand specificity. *RNA* **25**, 23–34 (2019).

30. Atilho, R. M., Arachchilage, G. M., Greenlee, E. B., Knecht, K. M. & Breaker, R. R. A bacterial riboswitch class for the thiamin precursor HMP-PP employs a terminator-embedded aptamer. *Elife* **8**, e45210 (2019).

31. Chen, X., Arachchilage, G. M. & Breaker, R. R. Biochemical validation of a second

class of tetrahydrofolate riboswitches in bacteria. *RNA* **25**, 1091–1097 (2019).

32. Malkowski, S. N., Spencer, T. C. J. & Breaker, R. R. Evidence that the nadA motif is a bacterial riboswitch for the ubiquitous enzyme cofactor NAD+. *RNA* **25**, 1616–1627 (2019).

33. Yu, D. & Breaker, R. R. A bacterial riboswitch class senses xanthine and uric acid to regulate genes associated with purine oxidation. *RNA* rna-075218 (2020).

34. Ames, T. D. & Breaker, R. R. Bacterial Riboswitch Discovery and Analysis. *The Chemical Biology of Nucleic Acids* 433–454 (2010) doi:https://doi.org/10.1002/9780470664001.ch20.

35. Breaker, R. R. Riboswitches and the RNA world. *Cold Spring Harb. Perspect. Biol.* **4**, a003566 (2012).

36. Lundrigan, M. D., Köster, W. & Kadner, R. J. Transcribed sequences of the Escherichia coli btuB gene control its expression and regulation by vitamin B12. *Proc. Natl. Acad. Sci.* **88**, 1479–1483 (1991).

37. Miranda-Ríos, J., Navarro, M. & Soberón, M. A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl. Acad. Sci.* **98**, 9736–9741 (2001).

38. Gelfand, M. S., Mironov, A. A., Jomantas, J., Kozlov, Y. I. & Perumov, D. A. A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. *Trends Genet.* **15**, 439–442 (1999).

39. Nahvi, A. *et al.* Genetic control by a metabolite binding mRNA. *Chem. Biol.* **9**, 1043–1049 (2002).

40. Winkler, W., Nahvi, A. & Breaker, R. R. Thiamine derivatives bind messenger

RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952–956 (2002).

41. Mironov, A. S. *et al.* Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111**, 747–756 (2002).

42. Winkler, W. C., Cohen-Chalamish, S. & Breaker, R. R. An mRNA structure that controls gene expression by binding FMN. *Proc. Natl. Acad. Sci.* **99**, 15908–15913 (2002).

43. Sudarsan, N., Wickiser, J. K., Nakamura, S., Ebert, M. S. & Breaker, R. R. An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes Dev.* **17**, 2688–2697 (2003).

44. Grundy, F. J., Lehman, S. C. & Henkin, T. M. The L box regulon: lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. *Proc. Natl. Acad. Sci.* **100**, 12057–12062 (2003).

45. Kochhar, S. & Paulus, H. Lysine-induced premature transcription termination in the lysC operon of Bacillus subtilis. *Microbiology* **142**, 1635–1639 (1996).

46. Liao, H.-H. & Hseu, T.-H. Analysis of the regulatory region of the lysC gene of Escherichia coli. *FEMS Microbiol. Lett.* **168**, 31–36 (1998).

47. Boy, E., Borne, F. & Patte, J.-C. Isolation and identification of mutants constitutive for aspartokinase III synthesis in Escherichia coli K 12. *Biochimie* **61**, 1151–1160 (1980).

48. Patte, J.-C., Akrim, M. & Méjean, V. The leader sequence of the Escherichia coli lysC gene is involved in the regulation of LysC synthesis. *FEMS Microbiol. Lett.* **169**, 165–170 (1998).

49. Barrick, J. E. *et al.* New RNA motifs suggest an expanded scope for riboswitches in

bacterial genetic control. *Proc. Natl. Acad. Sci.* **101**, 6421–6426 (2004).

50. Winkler, W. C., Nahvi, A., Roth, A., Collins, J. A. & Breaker, R. R. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428**, 281–286 (2004).

51. Mandal, M. *et al.* A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science (80-.).* **306**, 275–279 (2004).

52. Nelson, J. W. *et al.* Riboswitches in eubacteria sense the second messenger c-di-AMP. *Nat. Chem. Biol.* **9**, 834 (2013).

53. Dann III, C. E. *et al.* Structure and mechanism of a metal-sensing regulatory RNA. *Cell* **130**, 878–892 (2007).

54. Dambach, M. *et al.* The ubiquitous yybP-ykoY riboswitch is a manganese-responsive regulatory element. *Mol. Cell* **57**, 1099–1109 (2015).

55. Nelson, J. W., Atilho, R. M., Sherlock, M. E., Stockbridge, R. B. & Breaker, R. R. Metabolism of Free Guanidine in Bacteria Is Regulated by a Widespread Riboswitch Class. *Mol. Cell* **65**, 220–230 (2017).

56. Corbino, K. A. *et al.* Evidence for a second class of S-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. *Genome Biol.* **6**, R70 (2005).

57. Weinberg, Z. *et al.* Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res.* **35**, 4809–4819 (2007).

58. Yao, Z. *et al.* A computational pipeline for high-throughput discovery of cis-regulatory noncoding RNA in prokaryotes. *PLoS Comput Biol* **3**, e126 (2007).

59. Tseng, H.-H., Weinberg, Z., Gore, J., Breaker, R. R. & Ruzzo, W. L. Finding non-coding RNAs through genome-scale clustering. *J. Bioinform. Comput. Biol.* **7**, 373–388 (2009).

60. Weinberg, Z. *et al.* Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.* **11**, R31 (2010).

61. Weinberg, Z. *et al.* Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic regions. *Nucleic Acids Res.* **45**, 10811–10823 (2017).

62. Yao, Z., Weinberg, Z. & Ruzzo, W. L. CMfinder - A covariance model based RNA motif finding algorithm. *Bioinformatics* **22**, 445–452 (2006).

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

64. Meyer, M. M. *et al.* Challenges of ligand identification for riboswitch candidates. *RNA Biol.* **8**, 5–10 (2011).

65. Greenlee, E. B. *et al.* Challenges of ligand identification for the second wave of orphan riboswitch candidates. *RNA Biol.* **15**, 377–390 (2018).

66. Weinberg, Z., Perreault, J., Meyer, M. M. & Breaker, R. R. Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. *Nature* **462**, 656–9 (2009).

67. Roth, A. *et al.* A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nat. Chem. Biol.* **10**, 56–60 (2014).

68. Weinberg, Z. *et al.* New classes of self-cleaving ribozymes revealed by comparative genomics analysis. *Nat. Chem. Biol.* **11**, 606–610 (2015).

69. Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. *Functional and Integrative Genomics* vol. 15 141–161 (2015).

70. Klein, R. J., Misulovin, Z. & Eddy, S. R. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Acad. Sci.* **99**, 7542–7547 (2002).

71. Schattner, P. Searching for RNA genes using base-composition statistics. *Nucleic Acids Res.* **30**, 2076–2082 (2002).

72. Meyer, M. M. *et al.* Identification of candidate structured RNAs in the marine organism 'Candidatus Pelagibacter ubique'. *BMC Genomics* **10**, 1–16 (2009).

73. Poiata, E., Meyer, M. M., Ames, T. D. & Breaker, R. R. A variant riboswitch aptamer class for S-adenosylmethionine common in marine bacteria. *RNA* **15**, 2046–2056 (2009).

74. Stav, S. *et al.* Genome-wide discovery of structured noncoding RNAs in bacteria. *BMC Microbiol.* **19**, 66 (2019).

75. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

76. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

77. Steitz, T. A. & Moore, P. B. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem. Sci.* **28**, 411–418 (2003).

78. Janssen, B. D. & Hayes, C. S. The tmRNA ribosome-rescue system. in *Advances in protein chemistry and structural biology* vol. 86 151–191 (Elsevier, 2012).

79. Benner, S. A., Ellington, A. D. & Tauer, A. Modern metabolism as a palimpsest of

the RNA world. *Proc. Natl. Acad. Sci.* **86**, 7054–7058 (1989).

80. Nelson, J. W. & Breaker, R. R. The lost language of the RNA World. *Sci. Signal.* **10**, (2017).

81. Hernández-Morales, R., Becerra, A. & Lazcano, A. Alarmones as vestiges of a bygone RNA world. *J. Mol. Evol.* **87**, 37–51 (2019).

82. Vitreschak, A. G., Rodionov, D. A., Mironov, A. A. & Gelfand, M. S. Riboswitches: the oldest mechanism for the regulation of gene expression? *TRENDS Genet.* **20**, 44–50 (2004).

83. Breaker, R. R. Riboswitches: from ancient gene-control systems to modern drug targets. *Future Microbiol.* **4**, 771–773 (2009).

84. Zhu, N., Olivera, B. M. & Roth, J. R. Genetic characterization of the pnuC gene, which encodes a component of the nicotinamide mononucleotide transport system in Salmonella typhimurium. *J. Bacteriol.* **171**, 4402–4409 (1989).

85. Kemmer, G. *et al.* NadN and e (P4) are essential for utilization of NAD and nicotinamide mononucleotide but not nicotinamide riboside in Haemophilus influenzae. *J. Bacteriol.* **183**, 3974–3981 (2001).

86. Sauer, E., Merdanovic, M., Mortimer, A. P., Bringmann, G. & Reidl, J. PnuC and the utilization of the nicotinamide riboside analog 3-aminopyridine in Haemophilus influenzae. *Antimicrob. Agents Chemother.* **48**, 4532–4541 (2004).

87. Huang, L., Wang, J. & Lilley, D. Structure and ligand binding of the ADP-binding domain of the NAD+ riboswitch. *RNA* rna-074898 (2020).

88. Laursen, B. S., Sørensen, H. P., Mortensen, K. K. & Sperling-Petersen, H. U. Initiation of protein synthesis in bacteria. *Microbiol. Mol. Biol. Rev.* **69**, 101–123

(2005).

89. Breaker, R. R. Riboswitches and translation control. *Cold Spring Harb. Perspect. Biol.* **10**, a032797 (2018).

90. Ishii, A. *et al.* Genes encoding two isocitrate dehydrogenase isozymes of a psychrophilic bacterium, Vibrio sp. strain ABE-1. *J. Bacteriol.* **175**, 6873–6880 (1993).

91. Thauer, R. K. Citric-acid cycle, 50 years on: Modifications and an alternative pathway in anaerobic bacteria. *Eur. J. Biochem.* **176**, 497–508 (1988).

92. Lescoute, A. & Westhof, E. Topology of three-way junctions in folded RNAs. *Rna* **12**, 83–93 (2006).

93. Haugen, S. P., Ross, W. & Gourse, R. L. Advances in bacterial promoter recognition and its control by factors that do not bind DNA. *Nat. Rev. Microbiol.* **6**, 507–519 (2008).

94. Cunin, R., Glansdorff, N., Pierard, A. & Stalon, V. Biosynthesis and metabolism of arginine in bacteria. *Microbiol. Rev.* **50**, 314 (1986).

95. Charlier, D. *et al.* Molecular interactions in the control region of the carAB operon encoding Escherichia coli carbamoylphosphate synthetase. *J. Mol. Biol.* **204**, 867–877 (1988).

96. Miller, R. E. & Stadtman, E. R. Glutamate synthase from Escherichia coli an iron-sulfide flavoprotein. *J. Biol. Chem.* **247**, 7407–7419 (1972).

97. Ames, T. D. & Breaker, R. R. Bacterial aptamers that selectively bind glutamine. *RNA Biol.* **8**, 82–89 (2011).

98. Ren, A. *et al.* Structural and dynamic basis for low-affinity, high-selectivity binding

of L-glutamine by the glutamine riboswitch. *Cell Rep.* **13**, 1800–1813 (2015).

99. Klähn, S. *et al.* A glutamine riboswitch is a key element for the regulation of glutamine synthetase in cyanobacteria. *Nucleic Acids Res.* **46**, 10082–10094 (2018).

100. Jendrossek, D., Kratzin, H. D. & Steinbüchel, A. The Alcaligenes eutrophus ldh structural gene encodes a novel type of lactate dehydrogenase. *FEMS Microbiol. Lett.* **112**, 229–235 (1993).

101. Haardt, M., Kempf, B., Faatz, E. & Bremer, E. The osmoprotectant proline betaine is a major substrate for the binding-protein-dependent transport system ProU of Escherichia coli K-12. *Mol. Gen. Genet. MGG* **246**, 783–796 (1995).

102. Zeden, M. S. *et al.* Cyclic di-adenosine monophosphate (c-di-AMP) is required for osmotic regulation in Staphylococcus aureus but dispensable for viability in anaerobic conditions. *J. Biol. Chem.* **293**, 3180–3200 (2018).

103. Stülke, J. & Krüger, L. Cyclic di-AMP signaling in bacteria. *Annu. Rev. Microbiol.* **74**, 159–179 (2020).

104. Yin, W. *et al.* A decade of research on the second messenger c-di-AMP. *FEMS Microbiol. Rev.* (2020).

105. Amorim Franco, T. M. & Blanchard, J. S. Bacterial branched-chain amino acid biosynthesis: structures, mechanisms, and drugability. *Biochemistry* **56**, 5849–5865 (2017).

106. Wilson, K. S. & von Hippel, P. H. Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proc. Natl. Acad. Sci.* **92**, 8793–8797 (1995).

107. Yarnell, W. S. & Roberts, J. W. Mechanism of intrinsic transcription termination and antitermination. *Science (80-.).* **284**, 611–615 (1999).

108. Merino, E. & Yanofsky, C. Transcription attenuation: a highly conserved regulatory strategy used by bacteria. *Trends Genet.* **21**, 260–264 (2005).

109. Barrick, J. E. & Breaker, R. R. The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biol.* **8**, 1–19 (2007).

110. Garst, A. D., Edwards, A. L. & Batey, R. T. Riboswitches: structures and mechanisms. *Cold Spring Harb. Perspect. Biol.* **3**, a003533 (2011).

111. Oliva, G., Sahr, T. & Buchrieser, C. Small RNAs, 5′ UTR elements and RNA-binding proteins in intracellular bacteria: impact on metabolism and virulence. *FEMS Microbiol. Rev.* **39**, 331–349 (2015).

112. Waters, L. S. & Storz, G. Regulatory RNAs in bacteria. *Cell* **136**, 615–628 (2009).

113. Hauryliuk, V., Atkinson, G. C., Murakami, K. S., Tenson, T. & Gerdes, K. Recent functional insights into the role of (p) ppGpp in bacterial physiology. *Nat. Rev. Microbiol.* **13**, 298–309 (2015).

114. Bridger, W. A. *et al.* The subunits of succinyl-coenzyme A synthetase--function and assembly. in *Biochemical Society Symposium* vol. 54 103–111 (1987).

115. Kashiwagi, K., Miyamoto, S., Suzuki, F., Kobayashi, H. & Igarashi, K. Excretion of putrescine by the putrescine-ornithine antiporter encoded by the potE gene of Escherichia coli. *Proc. Natl. Acad. Sci.* **89**, 4529–4533 (1992).

116. Miller-Fleming, L., Olin-Sandoval, V., Campbell, K. & Ralser, M. Remaining mysteries of molecular biology: the role of polyamines in the cell. *J. Mol. Biol.* **427**, 3389–3406 (2015).

117. Tabor, C. W. & Tabor, H. Polyamines. *Annu. Rev. Biochem.* **53**, 749–790 (1984).

118. Seaver, L. C. & Imlay, J. A. Alkyl hydroperoxide reductase is the primary scavenger

of endogenous hydrogen peroxide in Escherichia coli. *J. Bacteriol.* **183**, 7173–7181 (2001).

119. Kulajta, C., Thumfart, J. O., Haid, S., Daldal, F. & Koch, H.-G. Multi-step assembly pathway of the cbb3-type cytochrome c oxidase complex. *J. Mol. Biol.* **355**, 989–1004 (2006).

120. Buggy, J. J., Sganga, M. W. & Bauer, C. E. Nucleotide sequence and characterization of the Rhodobacter capsulatus hvrB gene: HvrB is an activator of S-adenosyl-L-homocysteine hydrolase expression and is a member of the LysR family. *J. Bacteriol.* **176**, 61–69 (1994).

121. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203 (2017).

122. Hove-Jensen, B. & Maigaard, M. Escherichia coli rpiA gene encoding ribose phosphate isomerase A. *J. Bacteriol.* **175**, 5628–5635 (1993).

123. Yanofsky, C. Attenuation in the control of expression of bacterial operons. *Nature* **289**, 751–758 (1981).

124. Dey, S., Biswas, C. & Sengupta, J. The universally conserved GTPase HflX is an RNA helicase that restores heat-damaged Escherichia coli ribosomes. *J. Cell Biol.* **217**, 2519–2529 (2018).

125. Sothiselvam, S. *et al.* Macrolide antibiotics allosterically predispose the ribosome for translation arrest. *Proc. Natl. Acad. Sci.* **111**, 9804–9809 (2014).

126. Camakaris, H., Camakaris, J. & Pittard, J. Regulation of aromatic amino acid biosynthesis in Escherichia coli K-12: control of the aroF-tyrA operon in the absence of repression control. *J. Bacteriol.* **143**, 613–620 (1980).

127. Stokes, H. W., O'gorman, D. B., Recchia, G. D., Parsekhian, M. & Hall, R. M. Structure and function of 59-base element recombination sites associated with mobile gene cassettes. *Mol. Microbiol.* **26**, 731–745 (1997).

128. Dar, D. *et al.* Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science (80-.).* **352**, (2016).

129. Mraheil, M. A. *et al.* The intracellular sRNA transcriptome of Listeria monocytogenes during growth in macrophages. *Nucleic Acids Res.* **39**, 4235–4248 (2011).

130. Toledo-Arana, A. *et al.* The Listeria transcriptional landscape from saprophytism to virulence. *Nature* **459**, 950–956 (2009).

131. Wurtzel, O. *et al.* Comparative transcriptomics of pathogenic and non-pathogenic Listeria species. *Mol. Syst. Biol.* **8**, 583 (2012).

132. Kalvari, I. *et al.* Non-Coding RNA Analysis Using the Rfam Database. *Curr. Protoc. Bioinforma.* **62**, 51 (2018).

133. Weinberg, Z., Nelson, J. W., Lünse, C. E., Sherlock, M. E. & Breaker, R. R. Bioinformatic analysis of riboswitch structures uncovers variant classes with altered ligand specificity. *Proc. Natl. Acad. Sci.* **114**, E2077–E2085 (2017).

134. Weinberg, Z. & Breaker, R. R. R2R-software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics* **12**, 3 (2011).

135. Agarwala, R. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **44**, D7–D19 (2016).

136. Rivas, E., Clements, J. & Eddy, S. R. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods* **14**, 45–48 (2016).

137. Washietl, S. *et al.* RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* **17**, 578–594 (2011).

138. Spies, N., Zook, J. M., Sidow, A. & Salit, M. GenomeView - An extensible python-based genomics visualization engine. *BioRxiv* 355636 (2018).

139. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

140. Merkel, D. Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux J.* 2 (2014).

141. Yoo, A. B., Jette, M. A. & Grondona, M. Slurm: Simple linux utility for resource management. in *Workshop on Job Scheduling Strategies for Parallel Processing* 44–60 (Springer, 2003).

142. Evans, B. & Bjornson, R. DeadSimpleQueue. https://github.com/ycrc/dsq (2021).

143. Allcock, W., Bresnahan, J., Kettimuthu, R. & Link, M. The Globus striped GridFTP framework and server. in *SC'05: Proceedings of the 2005 ACM/IEEE conference on Supercomputing* 54 (IEEE, 2005).

144. Rivas, E. RNA structure prediction using positive and negative evolutionary information. *PLOS Comput. Biol.* **16**, e1008387 (2020).