Studying trends of non-coding RNA function and evolution

By

Jeremy J. Widmann
B.A. MCDB, University of Colorado, Boulder, 2004

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Chemistry and Biochemistry

2012

This thesis entitled:

Studying trends of non-coding RNA function and evolution

written by Jeremy J. Widmann
has been approved for the Department of Chemistry and Biochemistry

_____
Rob Knight

_____
Robert Batey

_____
Michael Yarus

_____
Tom Cech

_____
Jim Goodrich

Date_____

Y The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Abstract

Widmann, Jeremy J (Ph. D., Biochemistry, University of Colorado, Boulder)
Studying trends of non-coding RNA function and evolution.

Thesis directed by Professor Rob Knight

RNA is a special type of molecule in the sense that it is an information carrier, and is also able to catalyze chemical reactions. It is consequently believed that RNA predated protein and DNA as a catalyst and information carrier in an "RNA World". A greater understanding of evolutionary and functional features of non-coding RNA is not only fundamental to elucidating the evolutionary mechanisms that give rise to RNA function, perhaps giving insight into the origin of life in an RNA World, but is necessary for the advancement of RNA biotechnology and RNA based therapeutics. Recent advancements in high-throughput sequencing technologies have provided the ability to study the function of non-coding RNAs at an unprecedented depth, producing millions to billions of sequences from a single experiment. This poses new challenges to researchers, as traditional biochemical and computational techniques are unable to scale to the massive amounts of data each experiment produces.

In this work, I present new computational tools, methods, and their applications in the study of non-coding RNA evolution. I have assembled a gold standard set of non-coding RNA alignments that have been manually curated and aligned to their known 3d structures. These manual alignments address the need for RNA alignments with structural annotation that current automated alignment algorithms do not provide. Next, I present an application of alignments to the study of tRNA evolution. tRNAs, an integral part of the modern translation machinery, are believed to be poor phylogenetic markers. Using UniFrac to cluster genomes based on the collection of tRNAs they contain, I show that these tRNA trees are similar to trees constructed from rRNA from the same organisms, congruent with universal phylogeny. Finally, I describe a technique developed to simultaneously measure the dissociation constant (K_D) of a pool of thousands of amino acid binding RNA aptamers obtained by *in vitro* selection, improving over the traditional laborious process of determining K_D one sequence at a time.

Acknowledgements

Rob Knight
Michael Yarus
Rob Batey
Tom Cech
Jim Goodrich
Irene Majerfeld
Jesse Stombaugh
Knight Lab
Robert, Kathleen, and Erika Widmann
Raeghan Mueller

Table of Contents

Table of Figures

Table of Tables

Chapter 1: Introduction

1.1 RNA World

1.1.1 RNA is believed to predate protein and DNA as catalyst and information carrier, respectively.

RNA is a special type of molecule in the sense that it is an information carrier, like DNA, and is also known to catalyze reactions, like protein enzymes (1). We know that many viruses use an RNA genome as their sole method of genetic information transmission. An expanding repertoire of biological RNAs has recently been found to be catalytic. For example, Cech and colleagues found that an intron in the pre-rRNA of *Tetrahymena thermophila* could be excised *in vitro* without the presence of any cell extract (2,3). Around the same time, the Altman and Pace labs discovered that the RNA subunit of RNase P, an enzyme responsible for tRNA processing, can in some species catalyze the reaction without the protein subunit (4,5). Additionally, we now know that the entirety of the catalytic center of the ribosome consists of RNA (6-11). It is consequently believed that RNA predated protein and DNA as catalyst and information carrier, respectively, in an “RNA World” (12-14).

In a prebiotic world, nucleic acids could have arisen from simple organic molecules *de novo* (15,16). More recently, through artificial selection techniques such as Systematic Evolution of Ligands by EXponential Enrichment (SELEX) (17-19), it has been shown that there are RNAs that are able to cleave, ligate and synthesize RNA, perform peptide bond formation, and create metal nanoparticles (20-22). It is thus important to study how RNAs have evolved functions such as amino acid binding, small molecule binding, reaction catalysis in order to learn more about the origins of life and

the capabilities that likely existed in the RNA World. This understanding can help provide insight into how modern RNA has evolved its multitude of functions and operates in nature.

1.1.2 Understanding non-coding RNA can give insight into how RNA functions have evolved.

Non-coding RNAs have evolved functions of binding, catalysis, and interactions with other cellular components, often regulatory interactions. Comparing differences between related RNAs gives clues to how functions can evolve and change during evolution. We are learning more and more about the role of non-coding RNA in development, gene regulation, and the immune system. Although the first non-coding RNAs to be discovered were involved in translation (rRNA, tRNA), suggesting that RNA might be restricted to information transmission roles, we now know that catalytic and regulatory RNAs play an important role in many cellular functions. This understanding will be increasingly important for understanding the role of RNA in various diseases.

1.2 Sequence Alignment and Phylogeny

1.2.1 Sequence alignments are essential for inferring structural, functional, and evolutionary relationships between sequences.

A sequence alignment can be viewed as a matrix of residues, where each row represents an individual sequence and each column represents a position in that sequence. Each column is considered to be evolutionarily or functionally related, although this picture can be complicated by insertions or deletions of entire structural elements that cannot be aligned at the single-residue level (23). Aligned sequences are necessary for comparing evolutionary or functionally related residues. We know that

the structure of non-coding RNAs is essential for their function. The primary sequence of a non-coding RNA is often not highly conserved, although the secondary and tertiary structures are highly conserved. In order to determine which residues and structural interactions are important for RNA function, these sequences must be aligned to one another, using different techniques depending on the problem to be solved. In cases where sequences are related evolutionarily, alignments can provide insight into how the sequences have evolved functions through mutations (substitutions, insertions, or deletions). For example, tmRNA has evolved dual functions as a tRNA and mRNA, where it rescues stalled ribosomes and aids in degradation of incomplete protein products (24,25). Alignments are also essential in inferring the evolutionary relationships between these sequences and functions. We also use alignments to construct phylogenies, which are trees that relate the sequences to one another according to the amount of sequence change between each sequence and its inferred common ancestor.

1.2.2 Pairwise versus multiple sequence alignment

A pairwise alignment is the simplest type of alignment, and is performed by aligning a single sequence to another single sequence. Because there is no sense of direction, from a pairwise alignment it cannot be inferred which residues in either sequence are evolutionarily more ancient, i.e. whether a change (including insertions and deletions) happened in one of the sequences or the other since their common ancestor.

Multiple sequence alignments are alignments that contain three or more sequences. A multiple sequence alignment can be performed in many ways. One method is a progressive alignment method, in which all of the sequences to be aligned are first

clustered by a hierarchical method (neighbor-joining, UPGMA) based on sequence similarity. Then the most similar sequences are aligned to each other first, and the alignment grows as increasingly dissimilar sequences are aligned, based on their position in the guide tree. This results in an alignment that can be turned into a profile, which represents a probability matrix “i” x “n” where, “i” represents the characters in the alphabet (nucleotides for DNA/RNA, amino acids for protein) and “n” is the position in the alignment. The values of the matrix represent the probability of finding each residue at each position in the alignment, and are useful for summarizing the variability in the alignment.

Multiple sequence alignments can also be performed using Hidden Markov Models (HMMs), which are essentially finite state machines, or Stochastic Context-Free Grammars (SCFGs). These methods construct a model from a known multiple sequence alignment, which is then used to align new sequences to the model, or to search for new sequences that match the model in a database of unmatched sequences (e.g. in a complete genome).

1.2.3 Global versus local alignment

Global alignment is often used to compare two sequences across their entire length, e.g. to find all changes between two sequences. A global alignment consists of the best alignment along the entire length of two sequences. A Needleman-Wunsch alignment is performed by calculating a matrix of length N by M, where N and M represent the length of two different sequences to be aligned. All possible match, mismatch, and gap insertion scores are calculated for each position in one sequence to each position in another. Based on a given scoring metric, one position is either aligned to another, or a

gap is inserted in the aligned sequence, representing an insertion or deletion in either sequence. This is a common method for aligning sequences that have changed slowly, or closely related sequences, or sequences where conservation are also useful for aligning sets of sequences that come from the same species, but where one isolate may have a mutation in the gene that yields a mutant phenotype and is responsible for a disease state.

Local alignment is useful for finding conserved regions that span only a small portion of a longer sequence. A local alignment consists of the best subsequence alignment between two sequences. A Smith-Waterman alignment is algorithmically similar to global alignment with Needleman-Wunsch, with a few minor modifications to the scoring scheme, which prevents long runs of gaps when aligning two dissimilar sequences. The result is an alignment of segments of the two sequences.

BLAST is a simplification of the Smith-Waterman local alignment algorithm. A common application is to find all matches for part of a gene sequence in all the genomes in GenBank. BLAST is also useful for finding short conserved regions within a gene. These regions are often important for gene activity, and have been maintained through natural selection in each genome during evolution. For instance, conserved regions may be active sites in an enzyme that contain the residues responsible for catalysis, or conserved structural elements, or elements required for interaction with other proteins or RNAs. When looking for these conserved regions, the surrounding sequence context is often irrelevant, justifying the use of local rather than global alignment. Local alignments can be useful for searching for genes that have evolved rapidly and therefore do not have high conservation over their entire length (26).

1.2.4 Review of current multiple sequence alignment algorithms

Two of the most commonly used multiple sequence alignment algorithms are ClustalW (27) and MUSCLE (28). These algorithms are simply based on the primary sequence, and use a pairwise progressive alignment mechanism. They consider all sequences to be evolutionarily related, arising from a common ancestor. However, this assumption is not justified for many non-coding RNAs, such as the hammerhead ribozyme (discussed later in this chapter) (29), which may arise multiple times during evolution.

Infernal (30) constructs covariance models (CM) from primary and secondary structural features, then aligns sequences to this covariance model. This method is useful for aligning RNAs that have a more highly conserved secondary structure than primary structure, which is the case with many non-coding RNAs such as tRNA, rRNA and RNAse P. Infernal relies on the prior knowledge of the secondary structure of the model, therefore performing a standard progressive alignment when structural information is not given.

LocARNA (31) is a local multiple alignment method that aligns sequence and secondary structure. It uses an improved implementation of the Sankoff algorithm (32), which is used to simultaneously align and predict the structure of two sequences. The multiple sequence alignment is performed as pairwise progressive alignment of predicted secondary structures, then alignment and scoring based on primary and secondary structural features. This algorithm does not rely on a CM, like Infernal, so it is useful for aligning RNAs where the true structure is unknown and cannot easily be inferred from a pre-existing alignment. Major drawbacks of LocARNA are the reduced

performance when aligning sequences which vary in length or have low sequence conservation (31).

1.2.5 Need to incorporate additional annotation

Since we know that RNA secondary structure is more conserved than primary sequence, we need a better way to annotate of alignments with structural information. This prompted development of BoulderALE (33), which is a manual alignment editor that scores the alignment (based on the IsoDiscrepancy Index (IDI) to a reference sequence) and annotation of secondary and tertiary structural interactions.

1.2.6 Why sequence alignments are critical for phylogeny

Phylogenetic trees are essential for defining evolutionary relationships between organisms. In sequence-based phylogenies, these relationships are determined by comparing analogous sequences shared between these organisms. All methods of inferring phylogenies rely on a multiple sequence alignment where the analogous residues (columns) of an alignment are considered to be evolutionarily related. Therefore, a good sequence alignment is fundamental for inferring these phylogenetic relationships.

1.2.7 Review of current phylogeny algorithms

Current algorithms for inferring phylogeny include: distance methods, maximum parsimony, maximum likelihood, and Bayesian methods (34). Distance methods calculate the pairwise distance between all sequences using a defined nucleotide substitution model. This model may assume equal substitution between all pairs of nucleotides, or different substitution rates based on known mutation rates (e.g. transitions occur more frequently than transversions). Algorithms such as neighbor-

joining algorithm are then used to construct a tree based on the distance matrix, progressively grouping more distantly related taxa. Distance based methods are fast and efficient, but can perform poorly when comparing highly divergent sequences.

Phylogenies can also be inferred from maximum parsimony. Maximum parsimony is a method for generating a phylogenetic tree using the minimal amount of evolutionary changes necessary to explain a given set of sequences Maximum parsimony is a non-parametric method, which does not rely on a known distribution of the relative rates of change from each type of nucleotide to each other type (although it does implicitly assume one). Of all the possible ways to construct a tree of multiple sequences, assigning each internal node of the tree with a character state representing the minimum number of changes to reach that state, the most parsimonious tree will have the least number of changes across the entire tree. In practice, because the number of possible trees is very large, heuristics are used to reduce the search space. Maximum parsimony benefits from simplicity, but suffers from the inability to account for multiple substitutions at the same site or to produce meaningful branch lengths.

Maximum likelihood is a parametric method for phylogenic tree optimization. Unlike maximum parsimony, it scores a tree based on known nucleotide substitution models. Maximum likelihood gives the probability of a character changing at any node on the tree. By calculating the probability of the sequence alignment given the tree, you can calculate the likelihood of the tree given the data. The tree with the highest probability of producing the alignment is chosen as the most probable tree. A benefit of this algorithm is that it gives meaningful branch lengths, where the branch lengths represent the average probability of characters changing along a given branch. However, it is

computationally expensive, and the results can be sensitive to the model used to score the tree.

Bayesian methods for phylogenetic inference are similar to the likelihood method in that they use known models of evolutionary change. However, it uses Bayes' theorem to calculate the probability of the tree given the alignment by examining a number of trees and constructing a probability distribution, usually by using Markov Chain Monte Carlo (MCMC) to sample from the posterior distribution, which derived from the data and prior probability distribution. Bayesian methods are faster than likelihood but can be influenced by the prior probability distribution, the choice of which is often subjective, and can provide statistical overconfidence if the Markov chain has not reached steady state.

1.3 SELEX

1.3.1 What is SELEX?

Systematic Evolution of Ligands by EXponential Enrichment (SELEX) is a technique for selecting nucleic acid sequences from a random pool of 10^{12} to 10^{15} different sequences that perform a given function (17-19). SELEX involves iterative rounds of selection and enrichment resulting in a pool of high-affinity RNA aptamers, which are nucleic acid sequences that bind to target molecules. This process can also be used to generate pools of ribozymes, RNAs capable of catalyzing reactions. Improvements have been made to the SELEX process by: reducing number of cycles needed, reducing time for analysis, developing *in vivo* techniques and aptamer delivery systems (35).

1.3.2 Repertoire of *in vitro* selected RNAs.

SELEX has been used to isolate aptamers for small molecules, amino acids, cofactors, proteins, and whole cells. Ribozymes that catalyze reactions such as self-cleavage, peptide bond formation, and carbon-carbon bond formation have also been isolated using this process (20,21,35-38).

1.3.3 Determination of functional motifs from SELEX pools.

Analysis of resulting pool from a SELEX experiment involves identification of nucleotides involved in the RNA's function (motifs). This is often performed in a number of ways. Comparing chemical probing data when RNA is incubated with and without ligand can identify regions where ligand binding occurs, as the RNA cannot be modified where the ligand is bound (39-42). Performing a multiple sequence alignment of a SELEX pool can also be used to identify conserved residues, which are often involved in RNA function.

1.3.4 SELEX and the genetic code

In an effort to test the RNA World hypothesis, much work has been done to experimentally determine the possible mechanisms of a primitive translation system solely comprised of RNA. Yarus and colleagues have performed numerous selections for amino acid binding RNA aptamers, observing an association between codon/anticodon sequences and the binding site of the cognate amino acids (arginine, glutamine, histidine, isoleucine, leucine, phenylalanine, tryptophan, and tyrosine (43)). Their research suggests that a primitive genetic code could have arisen by this stereochemical mechanism, where RNA codons/anticodons have an affinity to amino acids. The modern genetic code has been proposed to have evolved by a mechanism in which the codon/anticodon triplets "escaped" from being part of direct amino acid

binding sites and are used by the modern translation mechanism in mRNA and tRNA to encode protein (44).

1.3.5 Relating SELEX RNAs to biological RNAs

The hammerhead ribozyme is a small ribozyme that catalyzes the site-specific cleavage of a phosphodiester bond. The hammerhead ribozyme was originally discovered in viroids and plant viruses, where it processes multimeric RNA transcripts (45). Since its discovery, the hammerhead has also been found in bacteria (46), the genome of the newt (47,48), in schistosomes (49) and in cave crickets (50). The hammerhead ribozyme has also been independently derived in various labs using SELEX. The Szostak lab has used SELEX to obtain hammerhead ribozyme sequences *in vitro* (29), as has the Breaker lab (51). Figure 1.1 shows a phylogenetic tree built from an alignment of the hammerhead ribozyme sequences from Rfam (52) (light blue) and those isolated *in vitro* from the Szostak's group (dark blue). This tree shows that hammerhead motif is phylogenetically indistinguishable between artificially selected and naturally occurring ribozymes.

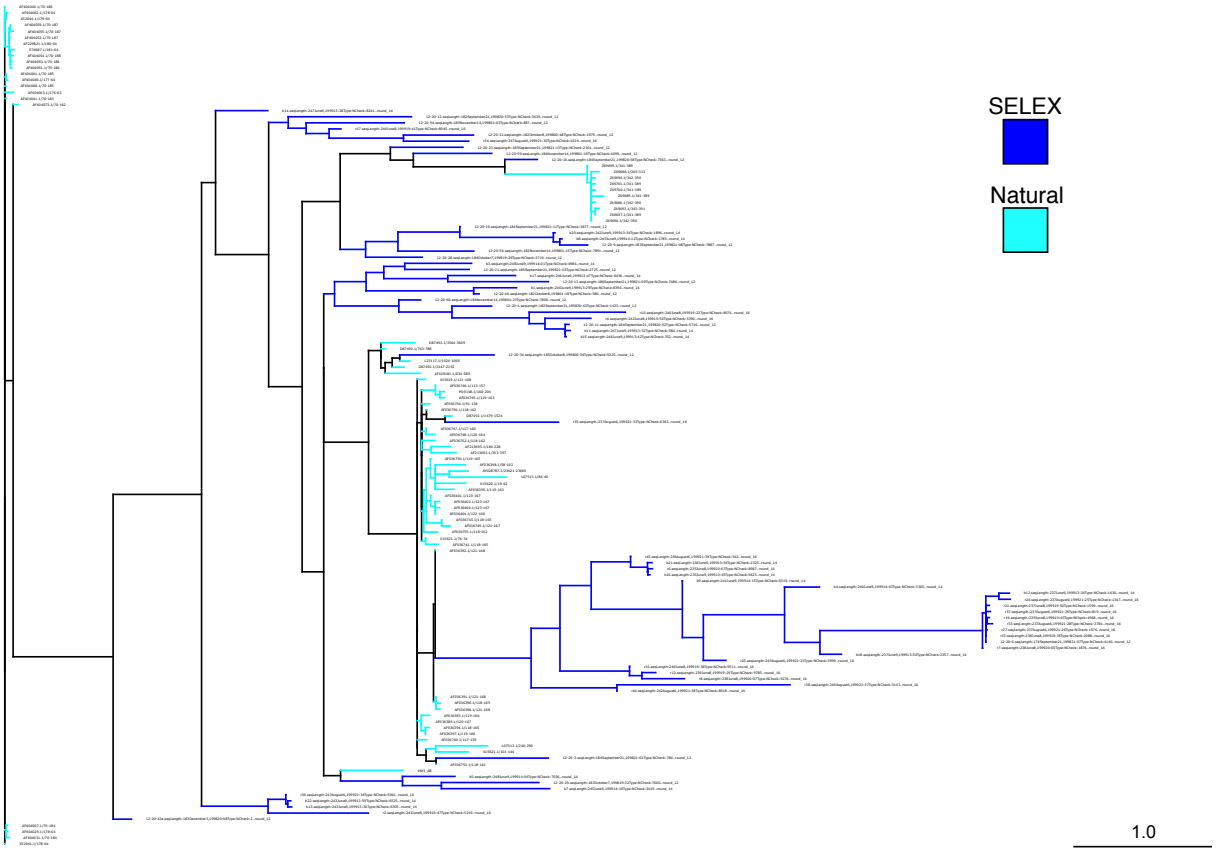


Figure 1.1: Phylogenetic tree of hammerhead ribozyme sequences. The branches colored in dark blue correspond to sequences obtained from SELEX, while the branches colored in light blue correspond to naturally occurring hammerhead sequences.

Advances in high-throughput DNA sequencing now allow us to probe the SELEX process to an unprecedented level. The ability to obtain millions of reads from a single sequencing run have given researchers the ability to monitor enrichment of pools over time (53,54), reduce the number of rounds of selection required (53,55), estimate fitness landscapes of the selection process, and examine the diversity of the resulting pool at a level never before realized (53-57). This ability to obtain such vast amounts of data necessitates the development of computational tools to aid in analysis.

Further understanding of evolutionary and functional features of non-coding RNAs is not only fundamental to elucidating the evolutionary mechanisms that give rise to RNA function, perhaps giving insight into the origin of life in an RNA World, but is necessary for the advancement of RNA biotechnology and RNA-based therapeutics. This work describes the development and application of computational tools to achieve this goal. First I describe my work assembling a gold standard set of non-coding RNA alignments that have been manually curated and aligned to their known 3d structures. These manual alignments address the need for RNA alignments with structural annotation that current automated alignment algorithms do not provide. These high-quality alignments will be useful for motif searching in increasingly expanding sequence databases, improvement in alignment and structure prediction algorithms, and applications to phylogenetic study of RNA functional evolution.

Next, I present an application of alignments to the study of tRNA evolution. tRNAs, an integral part of the modern translation machinery, are believed to be poor phylogenetic markers. Using UniFrac to cluster genomes based on the collection of tRNAs they contain, I show that these tRNA trees are similar to trees constructed from rRNA from the same organisms, congruent with universal phylogeny.

Finally, I describe a technique developed to simultaneously measure the dissociation constant of a pool of amino acid aptamers obtained from SELEX. I was able to measure the K_D of thousands of sequences in parallel, compared to the traditional laborious process of determining K_D one sequence at a time.

Chapter 2: RNASTAR: An RNA STructural Alignment Repository that provides insight into the evolution of natural and artificial RNAs

2.1 Summary

In the last chapter, I highlighted the importance and utility of sequence alignments. With the rapid improvement of high-throughput sequencing technologies, sequence databases are expanding exponentially, and the need for high quality alignments with structural annotation is becoming more apparent. With the increased number of RNA structures in the Protein Data Bank (PDB), which, despite its name, also contains RNA structures, it is clear that alignments need to exploit these rich structure resources and capture more of the structure information in defining new annotations. Additionally, current automated RNA alignment algorithms often fail to recapture the essential conserved sites that are critical for function. In this chapter I present a manually curated set of 148 alignments with a total of 9600 unique sequences, in which each alignment was backed by at least one crystal or NMR structure. These alignments included both naturally and artificially selected molecules. I use principles of isostericity to improve the alignments by maximizing isosteric base pairs.

2.2 Introduction

2.2.1 Importance of quality alignments in understanding RNA evolution and structure.

Multiple sequence alignments are critical for understanding evolutionary principles including phylogenetic relationships among sequences (23,58) and functional principles such as critical active sites, or even elements of 3D structures, revealed through patterns of conservation (59). The alignment can even have more of an influence on the

inferred phylogeny than does the phylogeny inference method (60,61). In studies of RNA, large alignments such as those in the CRW (62) and in Rfam (63) have been useful for identifying new family members and inferring the secondary structures they contain. Tools such as INFERNAL (30) have greatly assisted in this endeavor, especially as the databases continue to grow.

Improved alignments of natural and artificial RNAs will also increase our ability to test hypotheses about RNA evolution and architecture. Clear patterns of nucleotide composition have been noted in both natural and artificial RNA families (64-70), and one fascinating question is thus whether RNAs shaped by natural selection share similar features with those artificially selected in the lab. Comparing natural and artificial RNAs is important because such comparisons tell us whether we are seeing contingent features of organisms as they have evolved on Earth, or universal principles of RNA architecture (71). Artificial RNAs also provide ideal test cases for homology comparison methods because they provide a test set of sequences that are known to be non-homologous with each other, or with any natural RNA. A key question is whether tertiary motifs (72,73) reliably recur among different classes of RNAs, and can be used as universal building blocks for synthetic biology of functional RNAs (74,75).

There has been substantial progress towards automated alignment methods, although they are still relatively inaccurate, especially for distantly related RNAs (76). Most alignment programs do not incorporate features such as isostericity (77) and compositional preference (67) that are known to be important in RNA evolution. BoulderALE (78) incorporates both of these features, allowing construction of manually

curated, high quality alignments that can be used to improve algorithms for automated methods.

When evaluating an alignment of RNA molecules, nucleotides are aligned based on conservation, which can be at the level of the nucleotide or at the level of structure (secondary or 3D). The evaluation of an alignment at the level of structure consists of understanding the nucleotide interactions and the effects of nucleotide mutations on those interactions. For instance, if NT1 and NT2 form a specific base pairing, a mutation of NT1 could affect the base pairing interaction to NT2. Leontis et al. classified all RNA base pairing interactions into 12 geometric families. Then, using qualitative methods, they identified the base pairs within a family that could be easily substituted for one another without disrupting the structure, otherwise known as isostericity (79). In a more recent publication, Stombaugh et al. extended this notion by developing the IsoDiscrepancy Index (IDI) (80), a quantitative method for classifying each base pair into an isosteric group. When determining the IDI between two base pairs, the method determines three attributes: 1) if the C1'-C1' distance between the interacting nucleotides are nearly identical; 2) if the corresponding nucleotides form hydrogen bonds between equivalent atoms; and 3) if the rotational matrices between corresponding nucleotides are nearly identical (80). Using these three attributes, the IDI between two base pairs can be calculated, where lower IDIs (< 2) refer to isosteric base pairs.

Therefore we constructed a large collection of crystal and NMR structures that were related to multiple sequence alignments using the procedure shown in Figure. 2.1. This collection of manually curated alignments backed by experimentally determined atomic-

resolution structures provides us both with an ideal training set for further algorithm development, and with seeds for more sensitive database searches.

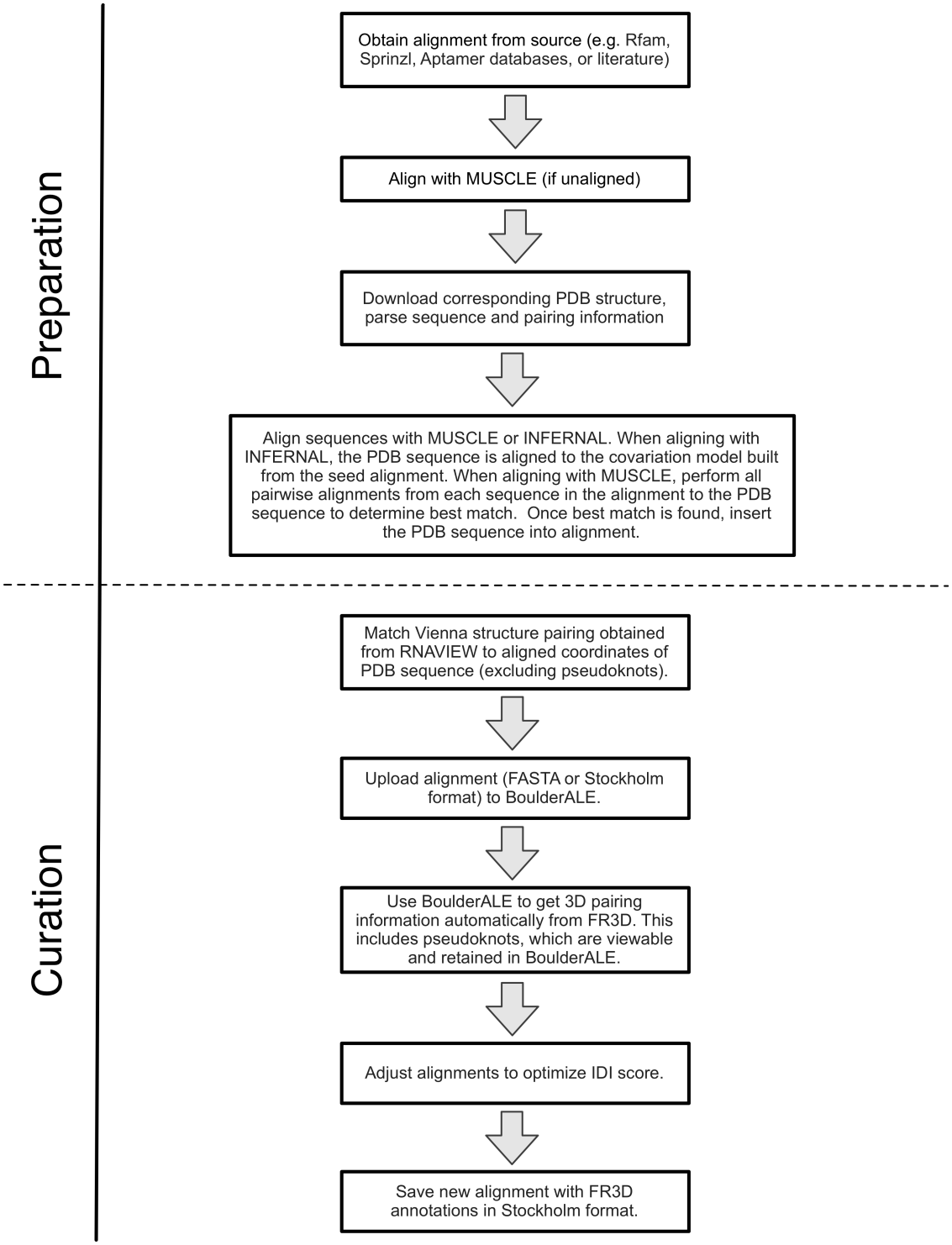


Figure 2.1: Overview of workflow for alignment

2.3 Results

2.3.1 Comparison of automated and manually curated alignments

We chose the 3D structures by manually examining all NMR and atomic-resolution crystal structures in the PDB containing RNA with a resolution < 4.1Å up to October 2011 (except for the 5S rRNA PDB 1YL3 which was 5.5 Å). Base pair information was derived from each structure using FR3D (81). Redundant sequences, defined as structures with identical base composition and base pairing, were dropped from the dataset, typically by choosing the most recent and/or highest resolution structure. Structures for which no homologous sequences could be found in Rfam (63), the tRNA database (82), the Aptamer Database (83), or as readable figures in the literature (64,84-104) were excluded from the analysis.

The manually curated alignments were substantially improved over automated alignments produced using MUSCLE (105) or INFERNAL (30), with essentially all showing an improvement in the fraction of non-isosteric base pairs (Figure 2.2). An example of the Hammerhead ribozyme MUSCLE alignment versus the manually curated alignment is shown in Figure 2.3. For these alignments, the crystal structure sequence (PDB: 379D) was aligned to a homologous set of sequences from Rfam (RF00163): note the substantially lower number of gaps and increased number of aligned positions in the manually curated alignment, which improve the IDI scores for a given alignment. On average, alignments in which the manual curation affected a greater number of positions also improved more substantially (Figure 2.4), as measured by IDI score (80)**:** the curated alignments had an average IDI score of 0.94, compared to an average score of 0.87 for INFERNAL alignments and 0.83 for the two MUSCLE

alignment methods (see **Materials and Methods**). Relative to the automatically generated alignments, the IDI scores of the curated alignment improved 32% of the sequences relative to INFERNAL,

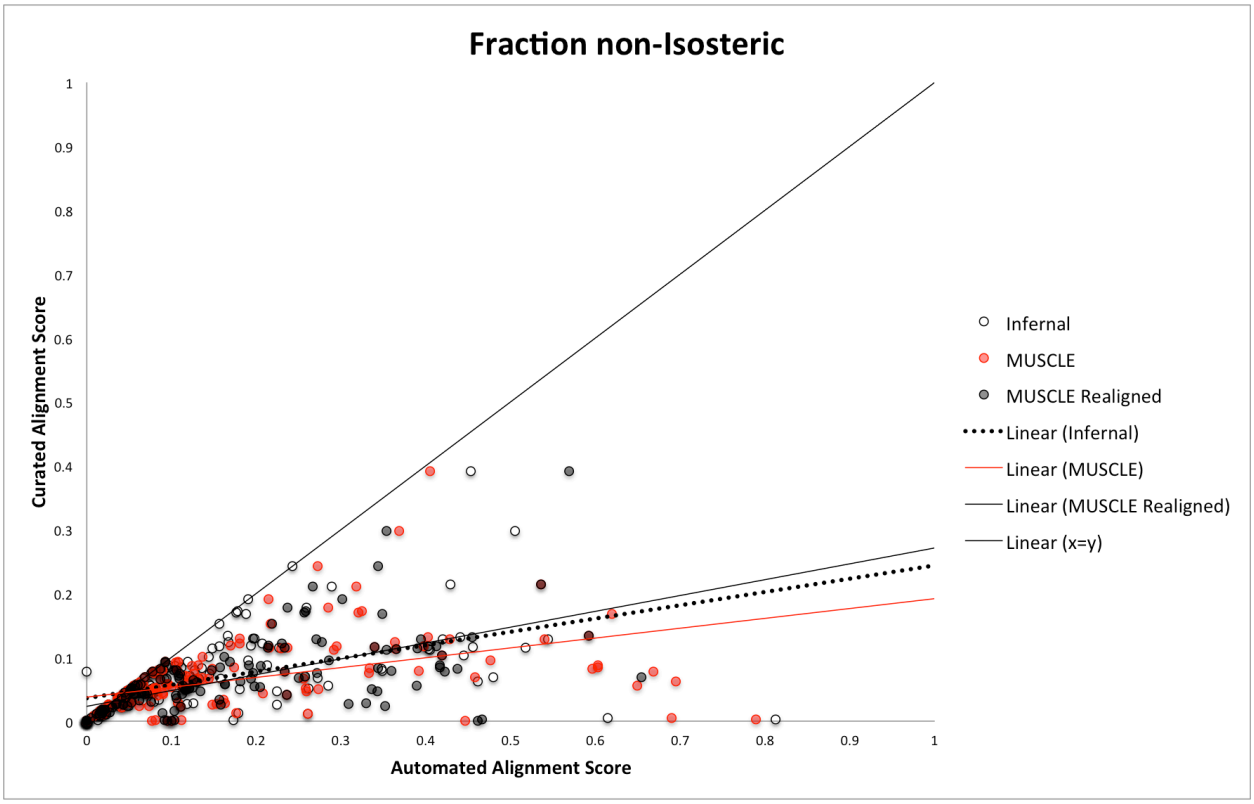


Figure 2.2: Comparison of original and improved alignments. The manually curated alignment scores (y-axis) are compared to each of three kinds of automated alignment (x-axis): inserting the PDB sequence with INFERNAL, inserting the PDB sequence with MUSCLE, and building the alignment *de novo* with MUSCLE. Scores are based on fraction of non-isosteric base pairs.

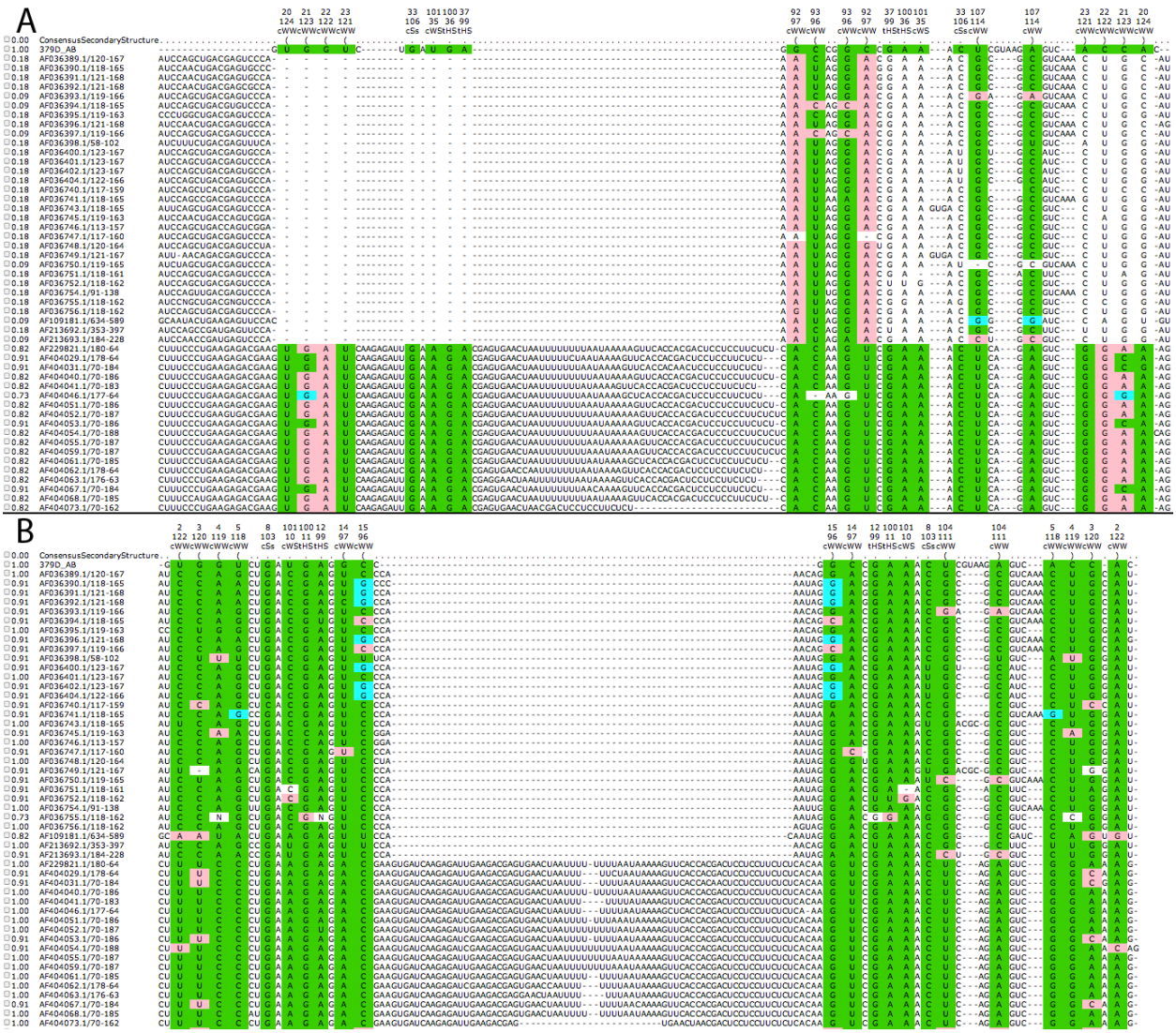


Figure 2.3: BoulderALE screenshots showing Hammerhead ribozyme alignment (Rfam: RF00163) where MUSCLE was used to align the corresponding crystal structure (PDB: 379D) sequence (A) versus the manually curated alignment (B). The colors from BoulderALE highlight isosteric (green), non-isosteric (pink) and not allowed (blue) covariations with respect to the 3D structure. For this alignment, there is an element expansion and as you can see in (A), MUSCLE aligned the x-ray crystal structure to a portion of the insertion. For the manual alignment (B), we shifted the x-ray crystal structure to align with the appropriate corresponding region.

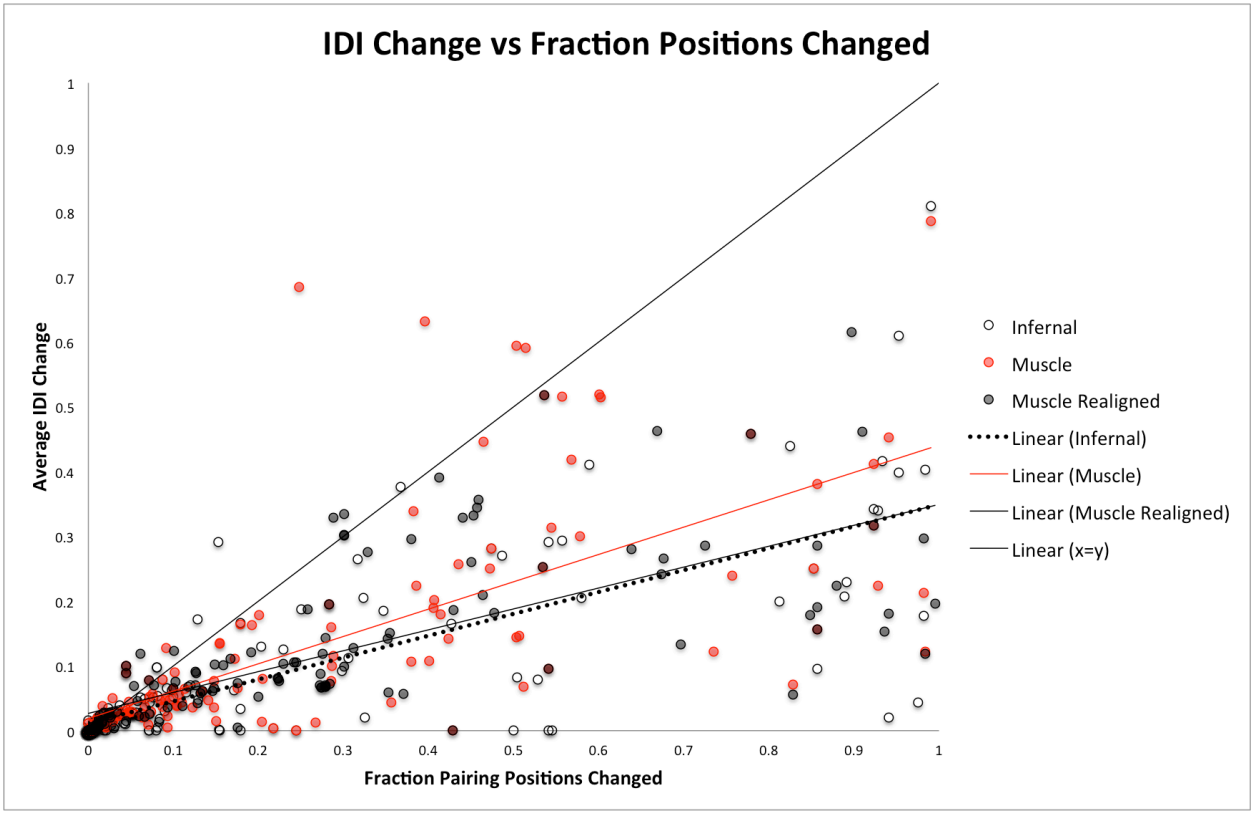


Figure 2.4: IDI change versus fraction positions changed. Alignments that underwent greater change during the manual curation process also improve more. Y-axis shows average IDI score change, x-axis shows average changed base-pairing positions within an alignment.

39% relative to MUSCLE, and 49% relative to the MUSCLE realigned method. The IDI scores of the curated alignment decreased in 1.6% of the sequences relative to INFERNAL, 1.0% relative to MUSCLE and 1.6% relative to the MUSCLE realigned method. Finally, the SPuNC scores, which calculate how well the RNA secondary structure predicted by an alignment matches the known compositional preferences for that secondary structure type (67), were substantially improved in the curated alignments over the automated alignments (Table 2.1). Consequently, manual curation substantially improved the overall alignment quality, as shown by two distinct measures.

Alignment	Unweighted_Extreme	Unweighted_Trained	Weighted_Extreme	Weighted_Trained
Curated	1.361414599	0.911091926	1.659870363	1.099245743
Infernal	1.376352659	0.958821011	1.680768259	1.141649617
Muscle	1.414364163	0.988129238	1.737054868	1.159960456
Muscle Realigned	1.597322008	1.07525067	1.962070865	1.308403591

Table 2.1: Averages from SPuNC output for manually curated alignments, INFERNAL automated alignments, MUSCLE automated alignments, and MUSCLE realigned alignments. The manually curated alignments as a whole comply far better with overall compositional preferences in different RNA structure regions as reported by Smit et al. 2008.

Overall, 146 of the 148 alignments showed equal or improved IDI scores. The two exceptions were special cases. The Valine tRNA alignment (RST00143.sto), which applies the basepairing information from PDB ID: 1J2B only aligns optimally when base pairing information from all available crystal structures is taken into consideration, possibly suggesting structural variation. For this alignment there were 3 corresponding x-ray crystal structures, so we inserted all 3 sequences from those structures and applied the FR3D base pairing information for each structure independently to determine the quality of the manually curated alignment. The VS ribozyme alignment (RST00145.sto) using the basepairing information from PDB ID: 1HWQ has a different issue: the automated alignment has the first base pair at the start of extremely long sequences, then inserts about 100 bases until the next base pair on both sides, thereby getting a perfect IDI score. In the curated alignment, the closing base pair is next to all the other base pairs, producing a non-isosteric substitution. However, this substitution is more likely as the true alignment, and is sterically acceptable at the end of the helix.

2.3.2 Comparison of structural composition of natural and artificial RNAs

As an example of the utility of a structure-backed alignment database incorporating both natural and artificial RNAs and using consistent methodology, we compared natural RNA families to artificial RNA families in terms of their rates of change of GC content across specific structural categories. On average, the total GC content did not differ substantially between natural (Figure 2.5A) and artificial (Figure 2.5B) RNAs ($t = 1.29$, $p = 0.198$). When we look at the responses to altered GC in the multiple sequence alignment within each category, we see a remarkable degree of universality in the response. Figure 2.6 shows the scatter plots of total GC content of natural sequences (Figure 2.6A) and artificial sequences (Figure 2.6B) against GC content of each structural category (stems, loops, bulges). For each structural category, the slopes of regression were determined and represented as histograms in Figure 2.7 separated by structural category (stem (Figure 2.7A), loop (Figure 2.7B), bulge (Figure 2.7C)). The t-test comparing natural versus artificial distributions of slopes shows that the difference in responses in stems is significant between natural and artificial RNA families ($t = 2.63$, $p < 0.01$), but the difference in responses in bulges and loops is not significant ($p > 0.6$ in both cases). The apparent difference in stem responses is likely driven by the greater range of mutation pressures that genomes experience relative to artificial RNA pools. A more sophisticated ANCOVA analysis, which separates out the effects of covariation in each category, suggests that interaction effects are at best weak (uncorrected interaction P-values are 0.02 for stems, 0.51 for loops, and 0.15 for bulges: none are statistically significant when corrected for multiple comparisons).

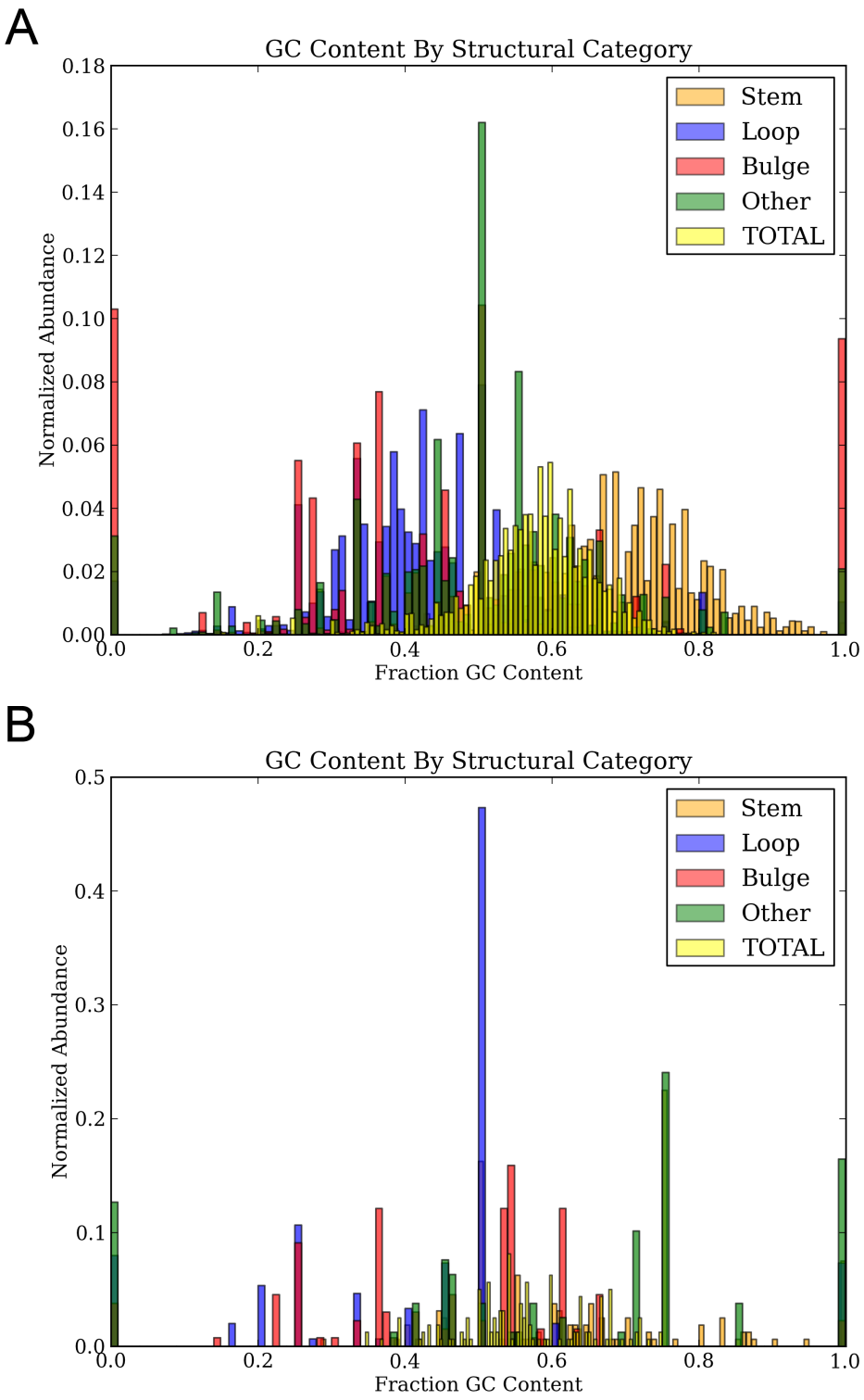


Figure 2.5: (A) Histogram of average GC content split up by structural category for naturally occurring sequences. (B) Histogram of average GC content split by structural category for artificially occurring sequences.

Consequently, the results are consistent with the idea that universal patterns of compositional change under GC content variation hold for both natural and artificial RNA families.

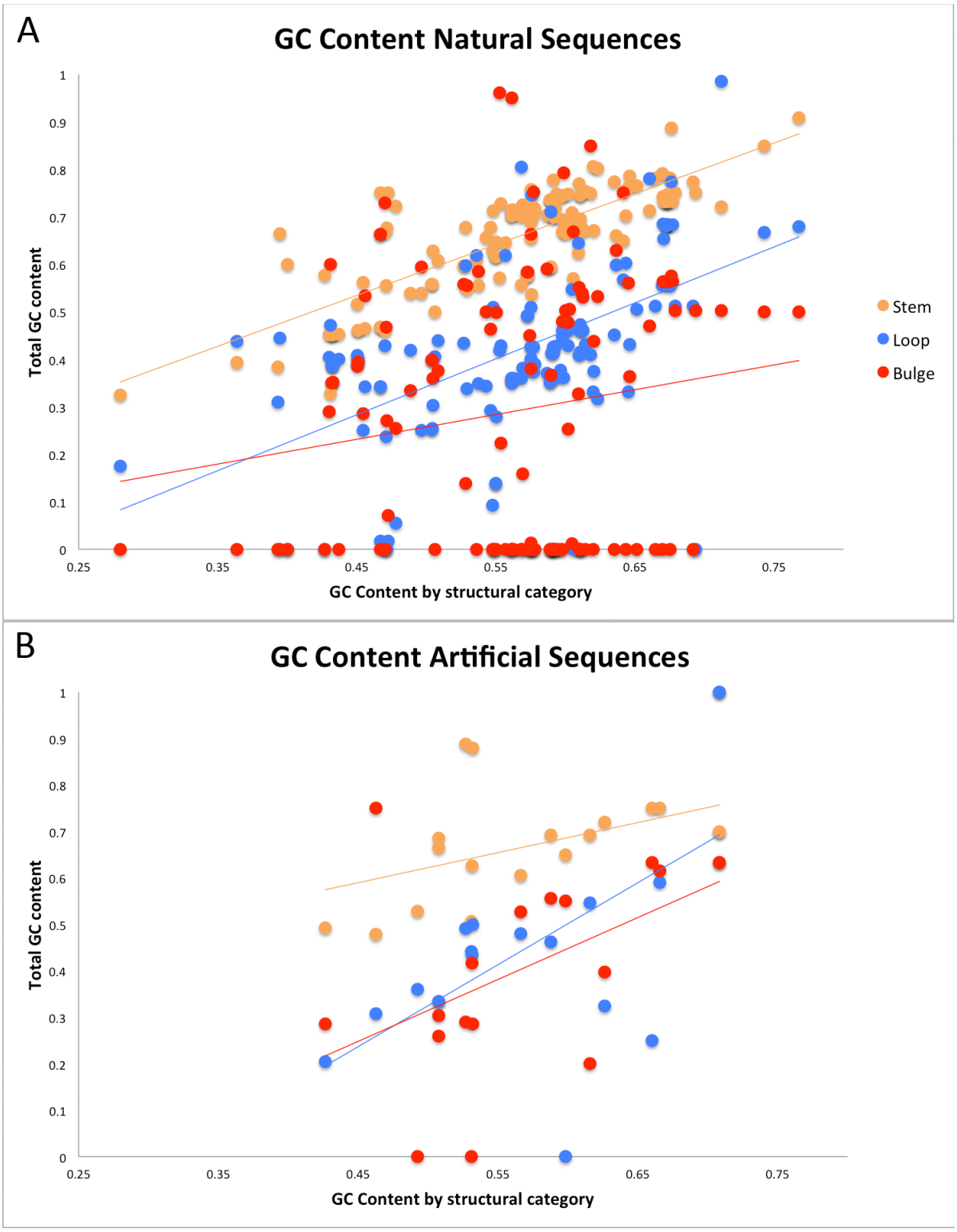


Figure 2.6: (A) Scatter plot of total GC content of natural sequences (x-axis) against GC content of each structural category (stem, loop, bulge) of the same sequences on the y-axis. (B) Scatter plot showing total GC content of artificial sequences (x-axis) against GC content of each structural category (stem, loop, bulge) on the y-axis.

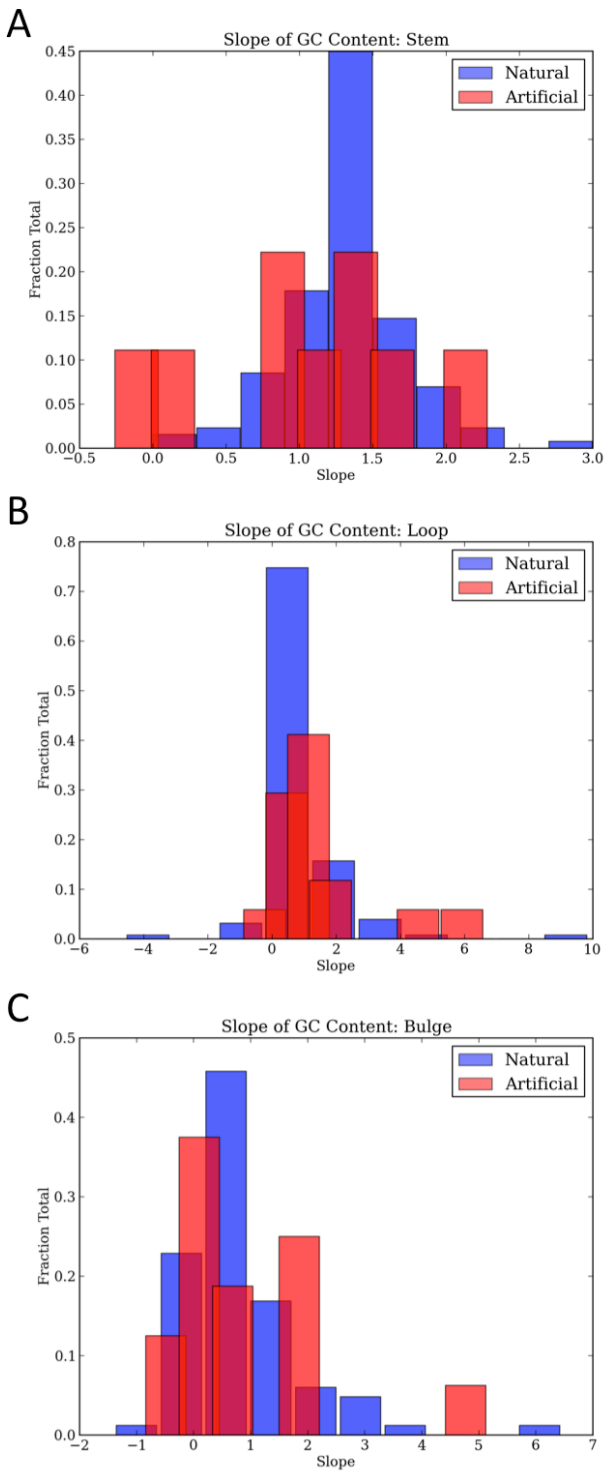


Figure 2.7: Histograms showing slopes of regression lines of GC content for each structural category (stem (**A**), loop (**B**), bulge (**C**)) versus total GC content. The responses to changes in GC content are extremely similar between natural and artificial RNA families.

2.4 Discussion

Manual alignments, especially those backed by crystal structures, still substantially outperform automated techniques by a range of metrics, suggesting that substantial improvement in algorithms is still possible. Since scoring schemes such as IDI and SPuNC can detect the improvement in manually curated alignments, incorporation of these metrics of isostericity and sequence composition into automated alignment software will likely lead to improvements in automated techniques.

IDI scores could provide an important filter for motif searching in large sequence databases, such as those now generated by sequencing SELEX pools or by metagenomics. More broadly, improved manually curated alignments will assist with benchmarking different RNA alignment and structure prediction algorithms, and provide a training set for ongoing development of these algorithms as well as providing us insight into how RNA molecules evolve.

2.5 Closing Statement

In this chapter I demonstrated the use of the principles of isostericity to align and annotate a large set of RNA alignments to their 3d structure. This is a necessary step in the advancement of tools for RNA alignment and structure prediction. Additionally, these alignments will be especially useful for the study of the function and evolution of these non-coding RNAs. In the next chapter I describe the application alignments to study the evolution of tRNAs, an essential part of the modern translation machinery.

2.6 Materials and Methods

2.6.1 Automated alignment preparation

Our choice of alignments was based on a requirement that there was a corresponding crystal structure or NMR structure in the Protein Data Bank (PDB). We did not accept poor resolution (> 4.1 Å with the exception of 5S rRNA PDB 1YL3 which was 5.5 Å) or cryo-EM structures for our reference structures. Redundant structures and those superseded by newer structures were not included in the curated alignments.

Base pair lists corresponding to each x-ray crystal structure were downloaded from the "Find RNA 3D" (FR3D) website ((81), http://rna.bgsu.edu/FR3D/AnalyzedStructures/). FR3D classifies all canonical and non-canonical base pair interactions for a given RNA 3D structure using the Leontis and Westhof (106) base pair nomenclature, which has been adopted by the RNA Ontology Consortium as the standard annotation scheme for RNA base pair interactions (107).

Structures that were identical or superseded by newer structures were eliminated from the analysis. Redundant sequences were eliminated from the analysis. Sequences introducing gaps in >95% of the positions in the alignment were also eliminated from the analysis. After these filter criteria, we ended up with 9600 non-redundant sequences corresponding to 148 unique structures.

Sequences were aligned using INFERNAL 1.0.2 and MUSCLE 3.7. For the INFERNAL alignments containing a secondary structure, we aligned the PDB sequence to the alignment with default parameters. For the cases where no secondary structure was present, we built a CM with cmbuild (using the --ignorant flag) and used an unpaired placeholder for the consensus secondary structure, then aligned the PDB sequence to this alignment with cmalign. MUSCLE alignments were produced using two methods: 1) finding the best pairwise match in an existing alignment to the PDB

sequence, then inserting the PDB sequence into the full alignment with MUSCLE and aligning it to it's best match; 2) using an existing alignment, remove all gaps in all sequences, then use MUSCLE to realign the entire alignment and insert the PDB sequence into this alignment using MUSCLE in the same way as the first method.

2.6.2 Manual curation of alignments to maximize isosteric base pairs

Curation of alignments were done using BoulderALE, where we were able to apply Watson-Crick and non-Watson-Crick base pair information onto the alignment. Using the base pairing information, we were able to manually curate the alignment to optimize isostericity. For some cases, manual inspection of the x-ray structure was necessary to determine the reliability of specific base pair interactions and for insight into the appropriate location for insertion/deletions.

2.6.3 Methods for scoring alignments.

We used several scoring schemes to assess the quality of the curated versus the automated alignments. The simplest way to score the alignments was to calculate the total entropy of the alignment. This is done by using the frequency of all nucleotides in each position (column) of the alignment to calculate the Shannon entropy for that position. The entropy values for each position can vary from 0 (absolutely conserved) to 2 (completely degenerate). These values were then summed for the entire alignment. However, we found that this simple method lacked statistical power to discriminate even among visually very good and very bad alignments (data not shown).

We also scored the alignments based on isostericity of base pairs that are known to form in the crystal/NMR structures. Using the 3D base interaction annotations from FR3D (81) we were able to assess the quality of the pairing regions of the alignments.

Using the PDB sequence as a reference, for each sequence, each base pair was assigned a value of 1 for isosteric and near-isosteric or a value of 0 for non-isosteric or not allowed. The sequence was then given a score that represented the fraction isosteric/near-isosteric base pairs. The alignment score is the average of each sequence's score, ranging from 0.0 (completely non-isosteric/not allowed) to 1.0 (perfectly isosteric/near-isosteric).

Chapter 3: Stable tRNA-based phylogenies using only 76 nucleotides.

3.1 Summary

In the previous chapter, I presented a gold-standard set of manually curated RNA alignments backed by a known 3d structure. These alignments outperformed automated alignment techniques by a number of metrics. This alignment collection will be especially useful for improvement of algorithms, motif searching, and determining evolutionary relationships between functional RNAs. In this chapter I extend the use of alignments to examine the evolution of tRNAs. tRNAs are often thought to be poor phylogenetic markers because they are short, often subject to horizontal gene transfer, and easily change specificity. Here I use an algorithm now commonly used in microbial ecology, UniFrac, to cluster 175 genomes spanning all three domains of life based on the phylogenetic relationships among their complete tRNA pools. Starting from an alignment of genomic tRNA sequences, I use UniFrac to cluster the genomes based on their entire tRNA pools to test whether there is enough phylogenetic signal present in tRNAs to recapture the universal phylogeny of the organisms that contain them.

3.2 Introduction

Transfer RNAs (tRNAs) are thought to be among the oldest biological sequences, present at the dawn of life in the Last Universal Common Ancestor (LUCA). tRNAs provide a critical step in translation, enforcing the genetic code by linking anticodon to amino acid (108), and are widely speculated to be among the most ancient RNA molecules (109-114). The availability of large tRNA databases (115-117), containing tens of thousands of tRNA sequences from hundreds of complete genomes, has

allowed the development of the new field of “tRNAomics” (117), in which the analysis of complete tRNA pools can be used to reveal selective pressures on the evolution of the translation apparatus. The overall structure of the tRNA molecules is well conserved at both the secondary and tertiary levels, with some exceptions for specific identity elements such as the variable loops (117,118).

3.2.1 tRNAs are believed to be poor phylogenetic markers.

Most bioinformatics studies of tRNA evolution to date were aimed at identifying tRNA identity elements (117,119) or sequence patterns associated with other functions of tRNA in translation (120), but not the overall pattern of tRNA evolution *per se*. Despite interest in tRNA phylogeny as a source of information about the evolution of the genetic code (110,112-114,121-125), and although tRNAs were among the first nucleic acid sequences to be used for phylogenetic reconstruction (126,127), the phylogenetic trees obtained from tRNAs are often radically different from the trees relating the species. tRNAs are now considered especially poor candidates for phylogenetic studies for several reasons. First, the sequences are short (the canonical tRNA sequence is 76 nucleotides [nt]), including invariant regions such as the terminal CCA and regions under strong selective pressure such as the anticodon loop and nucleotides involved in tertiary interactions. Additional pressures conserving tRNA structure may be imposed by the sequence requirements of other components of the translation machinery that interact with tRNAs: for example, conserved nucleotide patterns in bacterial tRNAs that correlate with the anticodon sequences was recently identified (128). Second, tRNAs are often involved in horizontal gene transfer, in part because mobile elements such as prophages carrying their own tRNAs are better able to express their genes after transfer

(129), and partly because many mobile elements preferentially integrate into or near tRNA genes (130). Indeed, these processes are so predictable that proximity to tRNAs has been exploited in computational methods for finding both prophages (131) and other genomic islands (132). Third, tRNAs can change specificity by as little as a single point mutation in an anticodon (120), suggesting that membership in a given tRNA isoacceptor family is not necessarily an evolutionary stable trait. Fourth, tRNAs have extensive paralogy through gene duplication, making the pattern of species evolution difficult to see through the tangle of duplications and losses of individual tRNA genes. Thus, it is reasonable to expect that the phylogenies of individual tRNA isoacceptor families might fail to match the organismal phylogeny. However, the question remains: do more closely related organisms tend to have more similar tRNA pools?

3.2.2 UniFrac can be used to compare organisms based on their gene content.

An algorithm that we developed that has been widely applied in microbial ecology, UniFrac, addresses this kind of question (133,134) (Figure 3.1). UniFrac works by measuring distances between groups of sequences on a phylogenetic tree in terms of the amount of evolution (measured by branch length within the tree) that is unique to each group. It then uses hierarchical clustering (135) to relate the groups based on these distances. Although it was originally developed for the analysis of microbial communities, in which the groups represent different environmental samples of 16S rRNA or other functional genes amplified from environmental samples (134,136), it can be applied to a wide range of other problems.

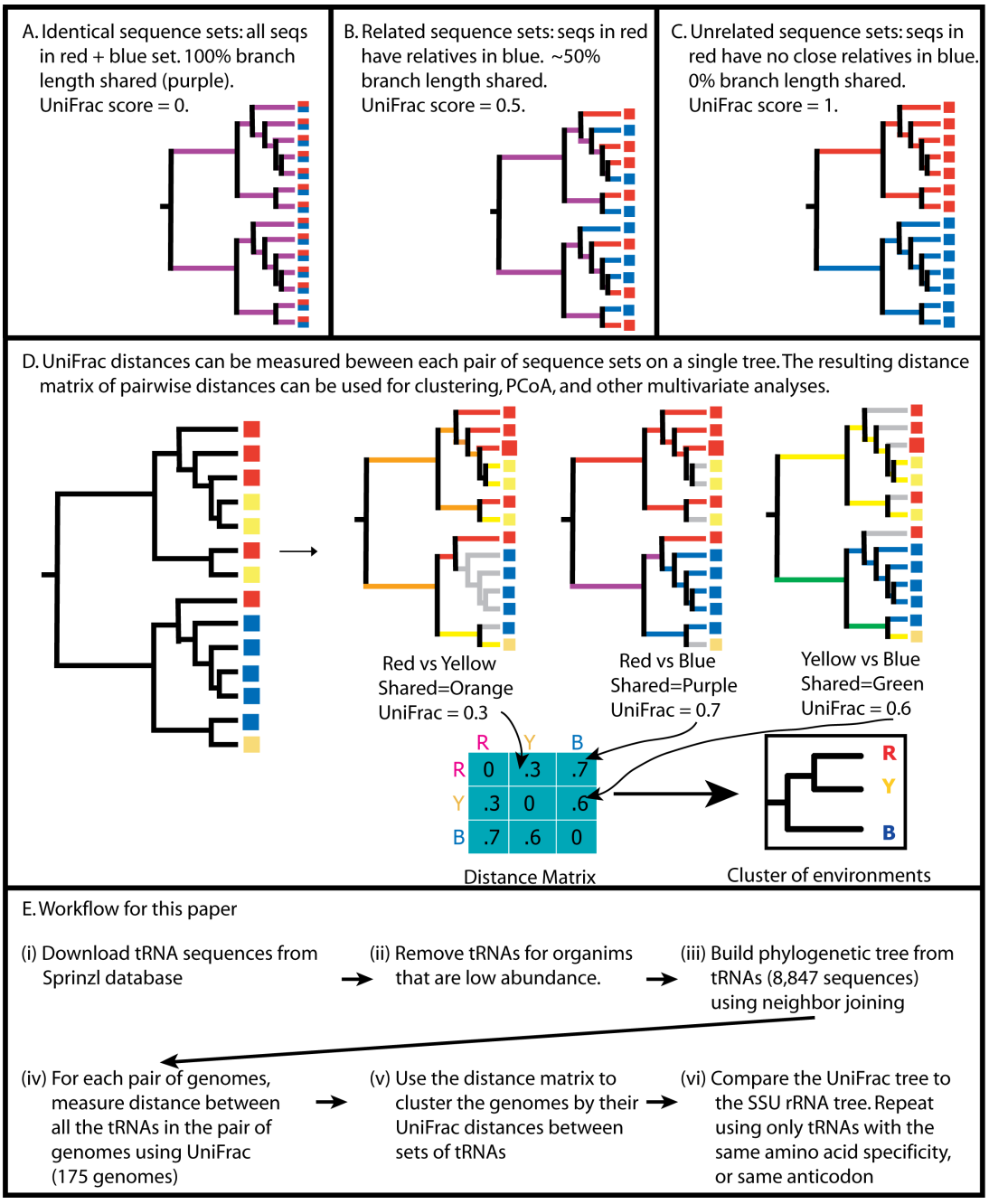


Figure 3.1: Overall tRNA tree-building procedure, including UniFrac clustering. UniFrac measures the fraction of branch length that is not shared between two groups of sequences, so that two identical groups of sequences (A) have a UniFrac score of 0, two completely dissimilar sets of sequences (C) have a UniFrac score of 1, and two related groups of sequences (B) have an intermediate UniFrac score. For a tree with many groups (here, the groups are genomes), the distance between each pair of groups can be calculated separately and summarized in a distance matrix (D). The overall workflow, including UniFrac steps, is shown in (E). These analyses were run using the weighted version of the UniFrac algorithm, which corrects for the abundance of each sequence (137).

For instance, we also recently used it to cluster genomes based on their pools of carbohydrate-active enzymes, including glycosyltransferases and glycoside hydrolases, and showed that bacteria and archaea that inhabit the human gut have converged in gene content for these groups compared to their relatives that live in other environments (138). In the present work, we again use UniFrac to cluster genomes, but this time we treat each genome as a group of tRNA sequences (its tRNA pool).

In other studies, we have found that UniFrac is able to relate complex data sets containing dozens of different microbial lineages to one another, revealing patterns in the data such as the divide between saline and non-saline aquatic communities (134) and the dominance of founder effects in establishing mouse gut microbial communities (136). Here, where the "communities" are genomes, we expect to be able to detect the total amount of tRNA evolution in each lineage, which may or may not track the organismal phylogeny depending on whether the tRNA complement is largely inherited or largely under selection. For example, we might expect unrelated lineages with similar codon usage, such as GC-rich Gram positive and Gram negative bacteria, to appear more similar to one another rather than to their relatives; similarly, we might expect archaea and bacteria that have Class I lysyl-tRNA synthetases, or that are extreme thermophiles, to cluster together. Our goal is thus to test whether the overall pattern of tRNA evolution is phylogenetically stable, or whether genomes that are similar in some other respect have convergently evolved similar tRNA pools.

3.3 Results

3.3.1 The overall pattern of tRNA evolution is phylogenetically stable.

The neighbor-joining phylogenetic tree relating all 8,847 tRNA sequences was

difficult to interpret directly. Although there were blocks of isoacceptors that appeared more or less consistent with organismal phylogeny, in general amino acid specificity, isoacceptor identity, and genome were mixed together. Figure 3.2 shows an excerpt of 35 tRNAs from the full tree of 8,847. Even in this small sample, several different amino acid specificities and a range of bacterial taxa are mixed together.

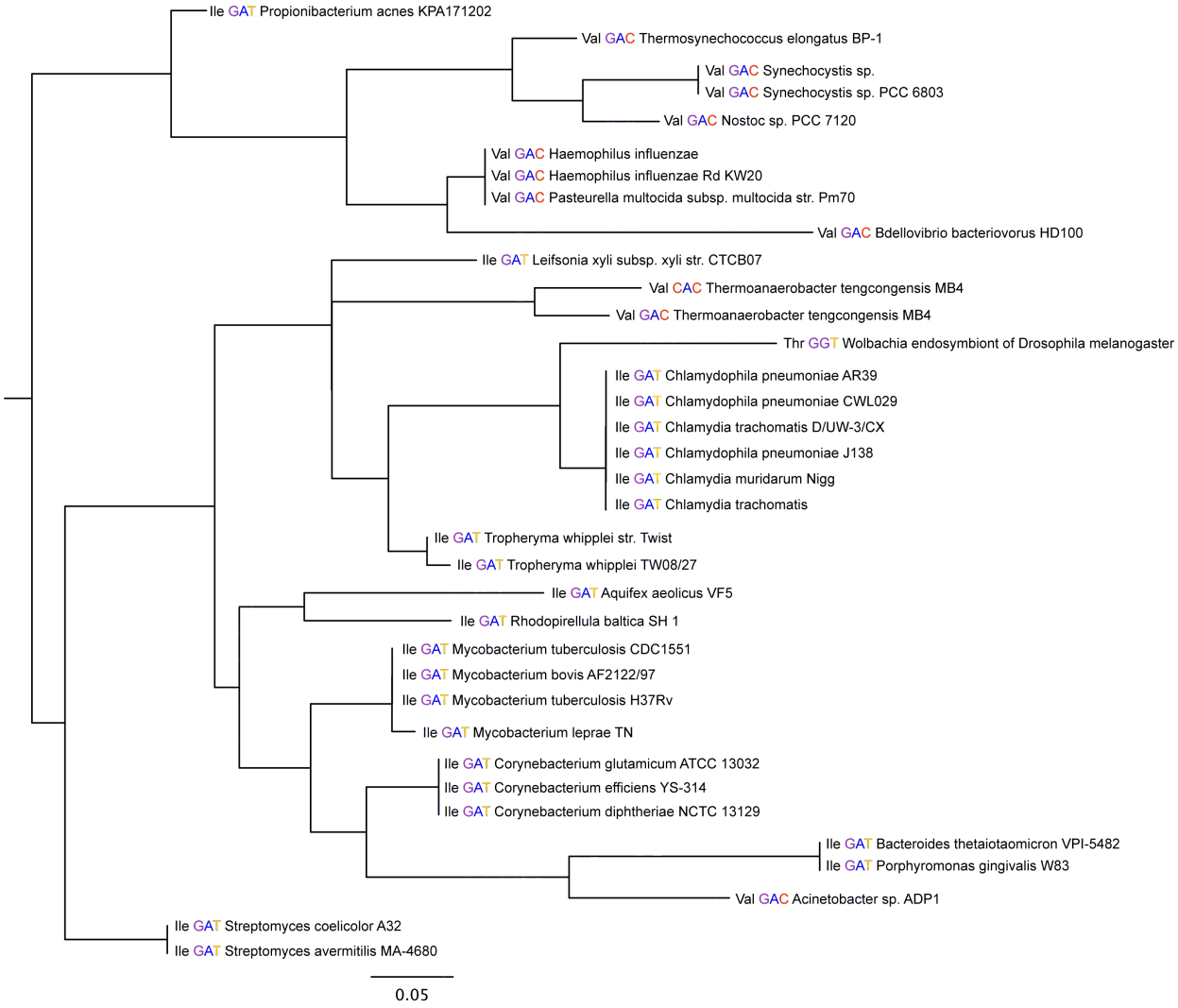


Figure 3.2: Small excerpt from the neighbor-joining phylogenetic tree containing 8,847 tRNA sequences. Each tRNA is labeled with its amino acid specificity, its anticodon, and the organism name. This tree containing only 35 tRNAs shows a mixture of several different amino acid specificities and different microbial lineages, reflecting the difficulty of using individual tRNA sequences for phylogeny. Scale bar shows 0.05 substitutions per site.

In contrast, the tree produced by applying UniFrac clustering to the tRNA pools from each genome reflected organismal phylogeny much better (Figure 3.3). The monophyly of each of the three domains of life (the eukaryotes, the archaea, and the bacteria) is recovered, and in general taxonomic groups of organisms (genera, families, etc.) cluster together. The clustering can also be represented as a scatterplot by projecting the distance matrix relating all genomes down onto the *n* dimensions that best explain the variation in the data using a multivariate technique called Principal Coordinates Analysis (PCoA) (Figure 3.4 shows the first three dimensions). These scatterplots show the same pattern: monophyly of each of the three domains of life, and eukaryotes and archaea are grouped together to the exclusion of the bacteria. Specifically, the first principal component separates the bacteria from the other two domains; the second separates groups of bacteria from one another (primarily the Gram negatives, at the top, from the Gram positives), and the third separates the archaea from the eukaryotes. The split between Gram negatives and Gram positives in the bacteria is possibly an interesting feature because these are not monophyletic groups and suggests that cell wall structure has the potential to cause a convergence in tRNA pools. Counter to our initial hypotheses, we did not find that thermophilic archaea and bacteria clustered together or that clustering was driven by GC content. Similarly, at the level of the overall tRNA pools, spirochetes with the Class I lysyl-tRNA synthetase such as *Borrelia burgdorferi* (139) clustered with the bacteria rather than with the archaea.

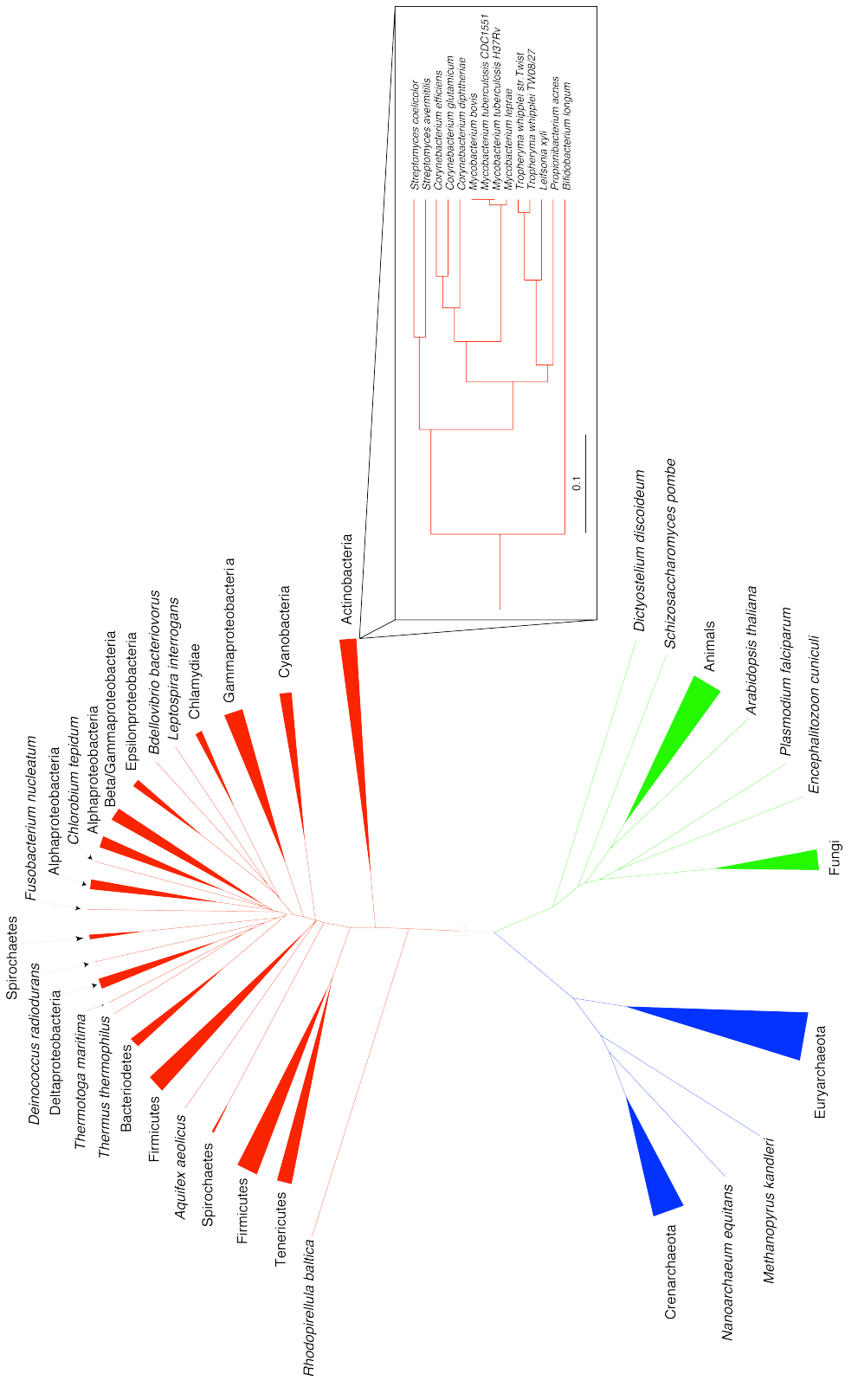


Figure 3.3: Weighted UniFrac tree of the tRNA pools in 175 genomes. The clustering recovers the monophyly of the Eukaryotes (green), the Archaea (blue), and the Bacteria (red), along with a large number of genus-level and other taxonomic groupings. Inset shows grouping at the genus level within the actinobacteria.

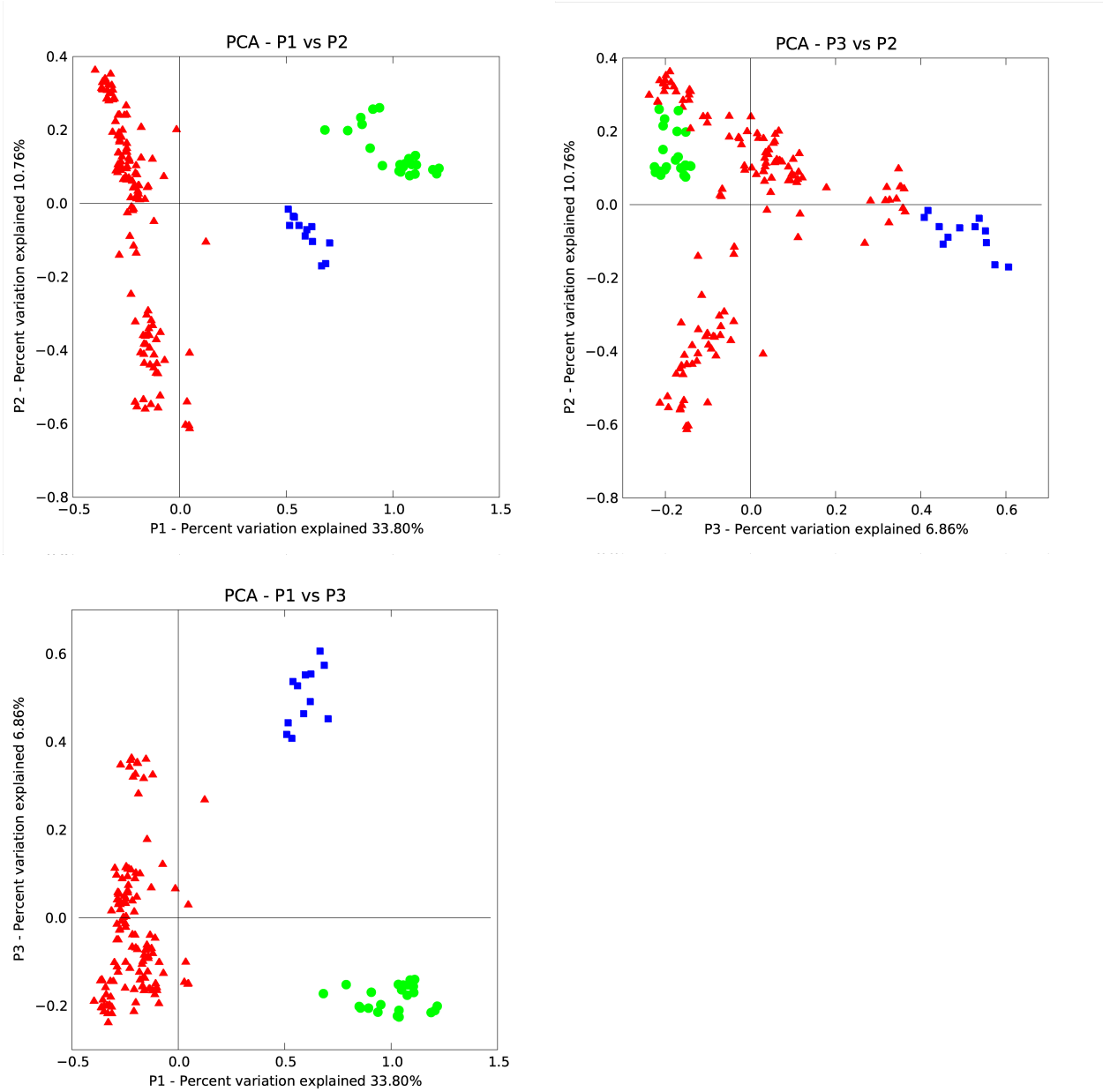


Figure 3.4: UniFrac PCoA of global tRNA pools showing clustering within the Archaea (blue squares), Eukaryotes (green circles), and Bacteria (red triangles). The scatterplots show P1 against P2 (top-left), P3 against P2 (top-right), and P1 against P3 (bottom-left); axes are aligned for direct comparison of the same components. The first principal component separates the bacteria from the other two domains; the second separates groups of bacteria from one another (primarily the Gram negatives, at the top, from the Gram positives, at the bottom); and the third separates the archaea from the eukaryotes. This clustering was performed using the weighted UniFrac algorithm as implemented on the UniFrac web site (133).

In principal coordinates analyses, the axes are chosen to maximize the variability in the data set and can thus be dominated by the most abundant categories (in this case, the bacteria). Although the separation of bacterial groups along PC axis 2 suggests that, when all species in the database are considered, the bacteria have much more variation in tRNA content than do either the eukaryotes or the archaea, there are many more bacterial genomes in this data set than archaea and eukaryotes, and, when an equal number of genomes is sampled from each domain, the effect disappears. Reinforcing this point, the total amount of sequence divergence in each of the three domains is comparable (i.e., the diversity, in terms of branch length, in Figure 3.3 does not reveal the bacteria to be far more diverse than the other domains). Thus, there is a clear split within the bacteria, but this split does not imply more variability overall in this domain than within the other two domains.

We tested the similarity of the tRNA pool cluster to a SSU rRNA tree using two approaches: the Mantel test (140), and MAST (141). The Mantel test is a permutation test that asks whether two distance matrices are correlated by permuting the row and column labels, calculating the correlation coefficient between the two matrices, and deriving an empirical distribution for the correlation expected by chance in the permuted matrices. It then tests whether the correlation coefficient for the true matrix is an outlier from the distribution of correlation coefficients from the permuted matrices. The Mantel test showed the correlation between the ARB 16S rRNA reference tree and the tree obtained from the full tRNA pool clustering to be highly significant ($P < 10^{-6}$). The correlation coefficient between the tRNA pool tree and the reference 16S rRNA tree was high ($r = 0.83$), approaching the mean value of $r=0.88$ for the correlation between the

ARB tree and the bootstrapped NJ rRNA trees (Figure 3.5). In contrast, the mean correlation coefficients from the trees based on clustering tRNAs, with UniFrac, from individual isoacceptor families, or from individual amino acid specificities, were much lower (r=0.78 and r=0.79 respectively). Interestingly, the tRNA isoacceptor clusters and amino acid clusters both outperformed on average trees build from arbitrarily sampled 76-nucleotide regions of the 16S rRNA itself (Figure 3.5).

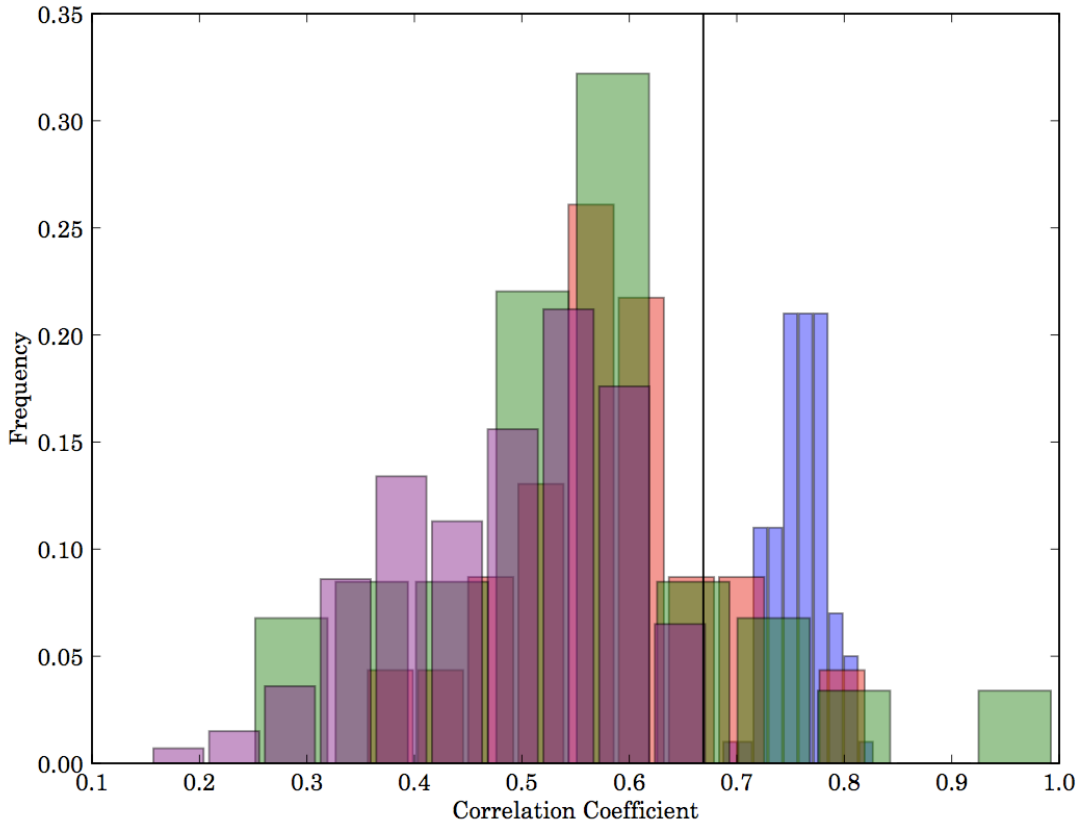


Figure 3.5: Distribution of correlation coefficients of distance matrices between the ARB tree and bootstrapped rRNA trees (blue), amino acid specificity clusters (red), isoacceptor clusters (green), and trees constructed from randomly sampled 76 nucleotide rRNA slices (purple). Each element in a matrix corresponds to the branch length traversed when moving from one genome to another genome in the corresponding tree using the shortest possible path (the tip-to-tip distance). The correlation coefficient for the full tRNA pool clustering, 0.67, is shown as a black line.

3.3.2 Individual tRNA families and isoacceptors reflect organismal phylogeny poorly.

No individual amino acid specificity tree matched the rRNA tree especially closely (the best was selenocysteine, r=0.91). The amino acid specificities ranged fairly evenly from r=0.6 to r=0.9. (Figure 3.6A) However, the isoacceptor trees were far more variable (Figure 3.6B). The Leu-IAG tree correlates almost perfectly with the rRNA tree (r=0.97, better than most bootstrapped rRNA trees). This strong correlation cannot be explained by restricted phylogenetic range (Leu-IAG tRNA is not found in archaea), because other tRNAs with similar phylogenetic distribution do not have similarly high correlations with the rRNA phylogeny. Sec-UCA, Ser-UCA, Pro-AGG, Glu-CUC, Val-UAC, and Ala-CGC all had r>0.80 (note that A at the first position of the anticodon is typically modified to I in tRNAs). In contrast, Ser-GCU, Val-AAC, Ile-AAU, Thr-UGU, and Ala-UGC all had r < 0.70. Similar variability in tRNA conservation was recently observed by Saks and Connor (128). It is unclear why the evolutionary rate of certain isoacceptors is higher then the others. It is unlikely that the observed variation in the evolutionary rate of the different tRNA sequences is correlated with the evolution of the corresponding aminoacyl-tRNA synthetases, because their recognition patterns are the same among all the isoacceptor members of tRNA family and seem to be generally well conserved, at least in bacteria (128).

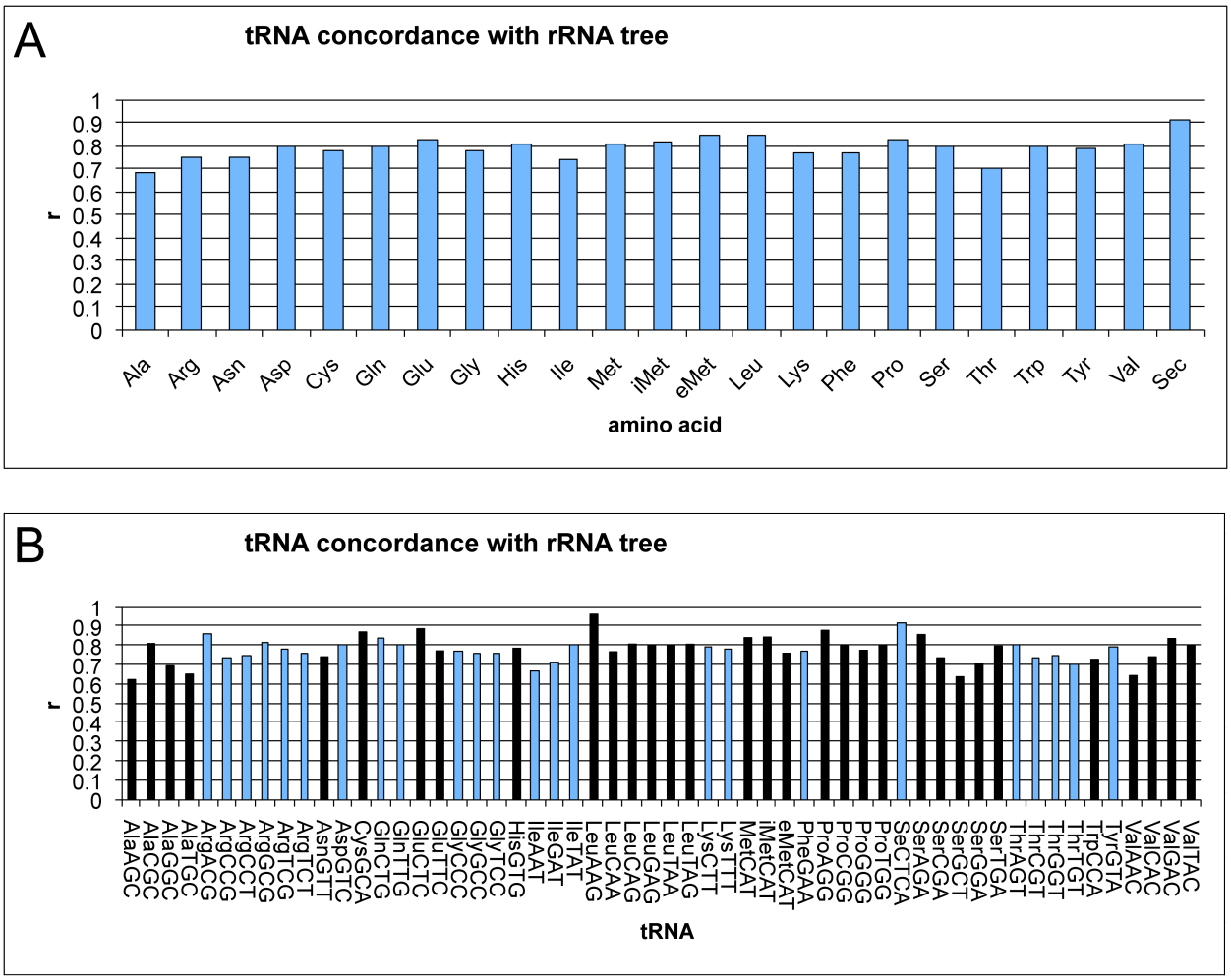


Figure 3.6: Concordance of individual tRNA trees with the rRNA tree for the full set of tRNAs for each amino acid (top), and for each isoacceptor family of tRNAs separately (bottom). Y-axis values range from 0 (no correlation with tRNA tree) to 1 (perfect correlation). iMet and eMet refer to initiator methionine and elongator methionine tRNAs separately. In the tRNA graph (bottom), the tRNAs with each amino acid specificity are colored the same way, alternating dark and light by family for clarity.

The poor correlation between the rate of tRNA and rRNA evolution might be caused either by higher or lower degrees of sequence conservation. The initiator tRNA-Met is by far more highly conserved than other tRNAs (117). The higher conservation of initiator tRNAs may be explained by the additional functional pressure applied to these tRNA by interactions with the additional components of translation initiation machinery, such as initiation factor 2 (142,143).

Interestingly, the bacterial initiator tRNA is more conserved than either the archaeal or eukaryotic initiator tRNA. This conservation may be due to the requirement for formylation of this tRNA in bacteria. Other tRNAs with low r-values are not generally highly conserved. No particular pattern seems to link these anticodons: there is a mixture of GC contents, first anticodon position base identity, etc. However, the difference in phylogenetic stability between different amino acid and anticodon identities presumably has some biochemical basis, perhaps in terms of interactions with other components of the translation apparatus. There may be not a single factor that explains all the differential rates of evolution in different tRNA isoacceptors. We note that both $tRNA^{Asn}$ and $tRNA^{Gln}$ fall in the group of tRNAs that correlate poorly with rRNA phylogeny. Both these tRNAs are considered to be later additions to the genetic code (112,113,121), and have to be adapted to the indirect transamidation pathway in most archaea and bacteria (144). Using tRNA phylogeny as a guide, we can now begin to explore the corresponding changes in translation machinery, with the hope of establishing causal relationships between changes in different lineages of interacting molecules.

3.4 Discussion

Our results demonstrate that UniFrac is able to derive biologically meaningful patterns even from trees with considerable levels of horizontal gene transfer and statistical error in their reconstruction (in this case, due to short sequences), and has considerable promise for other applications. In particular, it expands upon previous work with carbohydrate-active enzymes (138) to show that UniFrac can meaningfully cluster

genomes based on subsets of functional genes, to determine whether the content of the pools of these genes is driven by phylogenetic relationships or by the organism's lifestyles or habitats. It thus further supports the potential for the application of UniFrac to genomic and metagenomic data, in order to account to phylogenetic relationships in addition to presence/absence of genes while relating organisms or communities of organisms based on their gene content.

tRNAs were traditionally viewed as inadequate tracers of evolutionary events, primarily due to the their short length and frequent horizontal transfer between genomes. Our analyses have demonstrated that although most tRNA families and individual isoacceptors reflect the organismal phylogeny poorly, some isoacceptors, and the overall set of tRNAs in each genome, reflect the organismal phylogeny very well. Thus, the overall pattern of tRNA evolution is phylogenetically stable, and deviations from this reference pattern may reveal interesting biological features. Although the tRNA phylogenies are not quite as consistent as rRNA bootstrapped phylogenies, they may, like breakpoint phylogenies (145), provide an additional source of information to help address poorly resolved relationships throughout the tree of life.

Why is UniFrac able to recover phylogenies using the complete tRNA pools, when the trees recovered from individual isoacceptors perform so poorly? We suspect that the answer is that although individual tRNAs have idiosyncratic histories, these histories differ from one another, and thus these individual effects disappear when UniFrac effectively averages the results over all tRNAs. Because the overall pattern of similarities in tRNA pools is consistent with organismal phylogeny, it is meaningful when organisms resemble each other in specific tRNA features. In future studies, application

of the phylogenetic techniques may allow us to detect convergent evolution in response to specific factors, such as the gain or loss of a modifying enzyme that certain tRNAs must fold into a different structure to interact with (146), or gain or loss of an aminoacyl-tRNA synthetase. In particular, factors such as the use of a class I or class II lysyl-tRNA synthetase (147), or direct tRNA synthesis versus transamidation for Asn and Gln (148), may be reflected in the history and conservation of specific groups of tRNA isoacceptors. Comparative evolutionary studies of tRNA may thus provide a clue to better understanding the evolution of the rest of the translation machinery.

We expected to find the effect of major events in evolution of tRNA aminoacylation machinery, such as introduction of the Asp and Glu transamidase pathways, indirect formation of Cys-tRNA Cys (149) or the presence of class 1 lysyl-tRNA synthetase may be a significant factor in tRNA evolution. In our current analysis we did not find these events as a major factors affecting the evolution in the corresponding tRNA isoacceptor families. This finding agrees with notions derived from the study of the effect of the presence of an indirect pathway for Cys-tRNA Cys formation (150). In this paper, the authors did not find any effect of the presence of Sep-tRNA synthetase on the identity features of tRNACys, and concluded that formation of tRNACys identity preceded consequent evolution of aminoacylation machinery. The apparent lack of a visible effect of recruitment of novel tRNA recognition proteins on the phylogeny of potentially affected tRNAs implies that adaptation to the recruitment event occurs mostly on the protein side. It seems that adaptation of a newly recruited protein to pre-existing framework of tRNAs is evolutionary simpler than introducing changes into tRNA sequence, as the latter is already adapted to multiple interactions with other

components of translational and RNA processing machinery. This finding is another confirmation of the notion that tRNA system may has been established very early in evolution preceding formation of the modern aminoacylation system and divergence of aminoacyl-tRNA synthetases into the two modern classes.

3.5 Closing Statement

In this chapter I showed that overall pattern of tRNA evolution is consistent with universal phylogeny, despite the fact that they are short and subject to horizontal gene transfer. This shows UniFrac can be used to find biologically meaningful patterns where traditional methods fall short. This research can be extended to the study of evolution of other functional gene families. An important step in the study of non-coding RNA evolution is to be able to associate these phylogenetic relationships with their functional profiles. In the next chapter I describe a method, utilizing high-throughput sequencing technology, to determine the dissociation constant of thousands of amino acid binding RNAs aptamers obtained from *in vitro* selection, in parallel.

3.6 Materials and Methods

3.6.1 Constructing and comparing tRNA and rRNA trees.

We used the Sprinzl genomic tRNA compilation (82,116) as our source for tRNA sequences. We identified 175 genomes where (a) the complete genome was available in the Sprinzl database, and (b) the full-length rRNA sequence was available from the Silva Arb database (151). tRNA sequences with unknown characters were removed from the alignment. Genomes with less than 20 tRNA genes were also removed from the full alignment. This procedure resulted in a final dataset of 8,847 tRNA sequences, of which 6,047 were unique.

The reference small subunit (SSU) rRNA tree was obtained by the following procedure. First, the full SSU rRNA tree (SSU Ref 100) was obtained from the Silva Arb database. This tree consists of more than 400,000 sequences from all three domains. To construct the final tree for comparison with the tRNA tree, all sequences other than those corresponding to the 175 genomes were removed from the tree full tree.

Bootstrapped SSU rRNA alignments were built with the PyCogent (152) package, using a character matrix exported from ARB and the highly variable regions were removed using the LaneMaskPH mask available for download at the Greengenes web site (153). One thousand bootstrapped alignments were constructed and neighbor-joining trees were built using FastTree. We compared the ARB reference tree and the population of bootstrapped SSU rRNA trees to the population of tRNA trees described below.

We built two distinct types of tRNA-based trees (compare steps iii and v in Figure 3.1E). First, we performed neighbor-joining (NJ) on the full 8,847 sequence tRNA alignment. Second, we used weighted UniFrac clustering as implemented in the web interface (133) to cluster the genomes according to the tRNA pool that each genome contained. For these analyses, we excluded the variable loop and the anticodon domain of the tRNAs, and added CCA to the ends of sequences in which the CCA was not encoded in the genomic sequence. The anticodon was excluded so that similarities between tRNAs would not be influenced by similarities in amino acid identity, which was the criterion used to group the tRNAs. Similarly, CCA is an invariant sequence in the mature tRNA molecule, and whether this sequence is genomically encoded or added after transcription is likely to be a distracting factor rather than a meaningful criterion for

grouping. The variable loop was excluded to prevent artificial clustering of sequences based on differences in the length of this region.

Trees were compared using two methods: the Mantel test for distance matrix correlation, performed using the matrix of tip-to-tip distances relating each pair of taxa between a given pair of trees as implemented in the PyCogent package, and the subset distance which calculates the fraction of overlapping subsets where two trees differ (also implemented in the PyCogent package).

Chapter 4: High-throughput Kd determination through use of massively parallel sequencing

4.1 Summary

In the previous chapter, I demonstrated that universal phylogeny can be recaptured when clustering genomes based on their tRNA repertoire. Using this methodology with high-quality alignments can be extended to the evolutionary study of functional gene clusters. An important step in analysis of the evolution of RNA functions is the ability to quantify that function. The function of an RNA aptamer to a specific ligand is measured by the RNA's dissociation constant (K_D). In this chapter I present a new method for determining the K_D of thousands of sequences obtained from a SELEX experiment. I will demonstrate the use of high-throughput DNA sequencing combined with computational analysis to determine the K_D of over 4,000 sequences in parallel, with the same accuracy seen when the calculations are performed serially.

4.2 Introduction

4.2.1 SELEX can be used to obtain many high affinity aptamers from a random pool of RNAs

Aptamers are single-stranded DNA or RNA sequences that bind with high affinity and specificity to a molecular target. Aptamers can be obtained *in vitro* from a random pool of up to 10^{12}- 10^{15} nucleic acid sequences in a process called Systematic Evolution of Ligands by EXponential Enrichment (SELEX) (17-19). This process involves multiple cycles of partitioning and amplification to obtain a subset of sequences with high affinity to a target (Figure 4.1A). Aptamers have been isolated for many diverse targets, including cells (154-156), proteins (157,158), and small molecules such as amino acids

(41,159,160) and cofactors (84,161,162). SELEX is time-consuming and labor-intensive, involving 10-15 rounds of selection-amplification, which can take several days to weeks, but downstream characterization of the resulting sequences is even more time-consuming and represents the rate-limiting step in the process as it is currently practiced. Following enrichment of high-affinity aptamers, sequences from the enriched pool are isolated and screened for activity. This process typically involves cloning the resulting aptamer sequences into plasmids, and transfecting bacteria with the plasmids. The bacteria are then plated, colonies are picked, and their DNA sequenced for further analysis (39,163). These sequences are then transcribed, and binding affinities are measured individually for each clone (Figure 4.1B). This process is time consuming, inefficient, and expensive, often taking months to years. Additionally the resulting subset of cloned sequences may not be representative of the enriched pool, in part due to biases in amplification (164). Another main limitation with this process is that we are only able to sample a few sequences from the final selected pool.

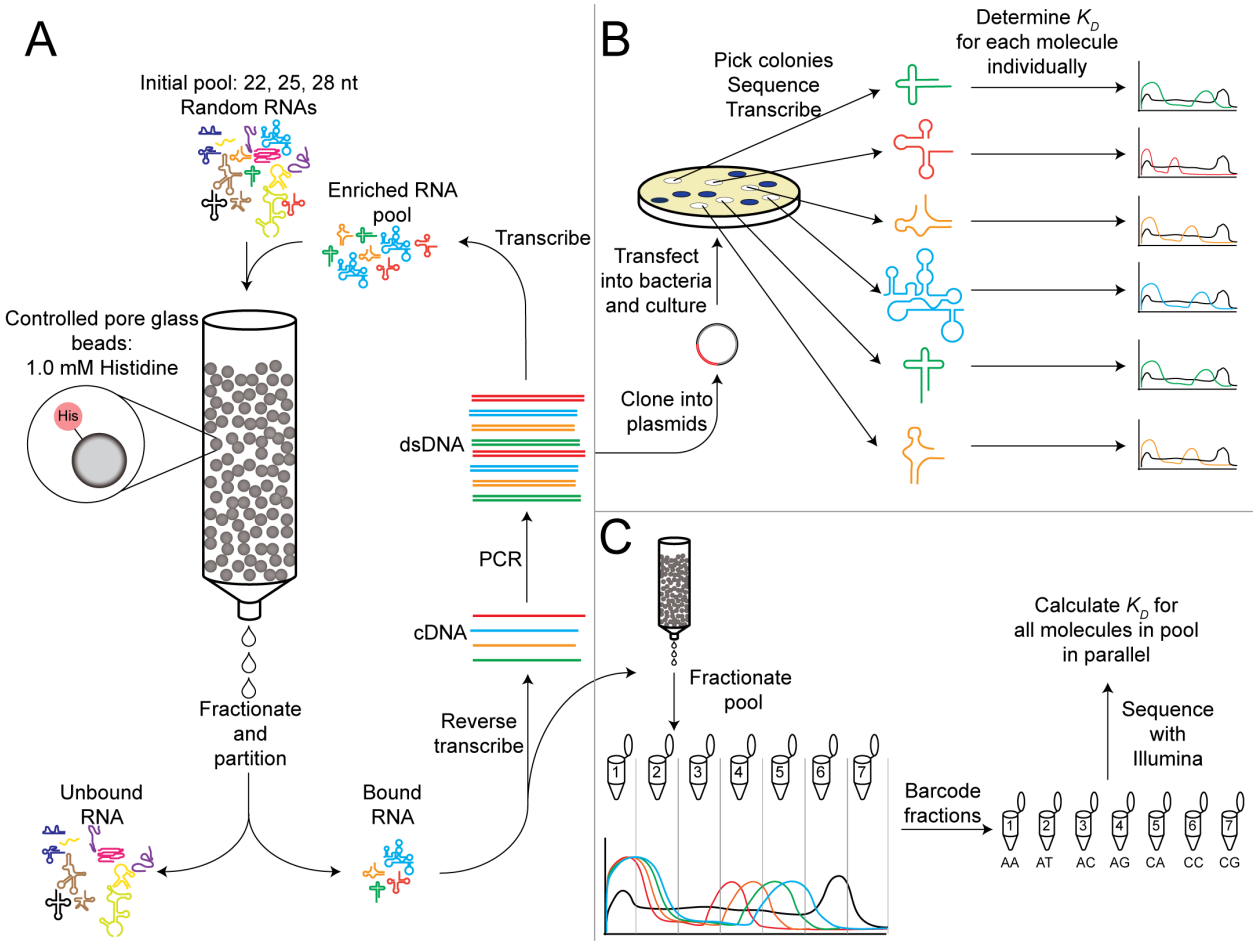


Figure 4.1. **Selection procedure and downstream analysis**. (A) Outline of SELEX procedure used to obtain histidine binding RNA aptamers. (B) Traditional low-throughput method for downstream analysis of final selected pool. This process involves cloning aptamer sequences into plasmids, transfecting into bacteria and growing bacteria in culture. Colonies are then picked, sequences transcribed, and K_Ds calculated individually. (C) High-throughput method taking final selected pool, fractionating, barcoding fractions, sequencing with Illumina, then calculating K_Ds for entire in parallel.

4.2.2 The use of high-throughput sequencing technology allows deeper understanding of SELEX pool evolution.

Traditional methods for affinity selections have been performed using nitrocellulose filter binding (17,165) and affinity chromatography (18,161,166,167). There has been much effort to make this process more efficient by reducing the selection time from

weeks to days or even hours (168,169) and number of cycles reduced from 10-15 to as few as 2 or 3 (53-55,168). Advances in capillary electrophoresis (168,169) and microfluidics (170-172) technologies have helped to make the SELEX process more efficient, reducing the time and number of required cycles, although, as noted above, analysis of the SELEX results is typically the bottleneck, and less attention has been paid to accelerating this downstream process. Recently, the development of high-throughput sequencing methods has revolutionized our ability to interrogate the SELEX process. Several groups have used high-throughput sequencing of the intermediate rounds of selection to show that the diversity decreases (54,56) and enrichment of highly active sequences increases (53,55) dramatically after only 3 or 4 rounds of selection. The major bottleneck is the lack of activity measurements of all the resulting aptamers from the selection simultaneously, preventing characterization of the overall distribution of activities in the pool. Many studies have measured activity for each round of selection, but only in bulk (53,54,169,171). Detailed examination of dissociation constants has only been performed for a handful of sequences in these studies.

4.2.3 Extension of well-established methods for affinity measures.

Previously, we used isocratic affinity chromatography to determine dissociation constants for amino acid ligands. This equilibrium method compares elution from the affinity matrix in the presence and absence of the ligand, and has the same logic as equilibrium dialysis in which there is a moving and an immobilized phase, therefore, it can be used to measure K_D (173,174). Even though isocratic affinity chromatography is not a traditional method for K_D determination, comparison of this method against well-known methods have shown to give results equivalent to ultrafiltration, equilibrium

dialysis (42), protections (40,41), and surface plasmon resonance (175). Therefore, this is considered an accurate and precise method for K_D determination.

Here we present a method for calculating the K_Ds of many RNA sequences simultaneously using the same procedure. This affinity chromatography method is crucial in the experimental design, because it allows physical separation of sequences by K_D range. Using barcoded Illumina sequencing (176-178) instead of radiolabel, we are able to calculate multiple K_Ds in parallel. The pool is fractionated through columns with and without ligand and each fraction is identified by attachment of a 12-nucleotide barcode to its sequences. . We can use the barcode of each sequence as its own label to calculate the K_D of the RNA directly from the sequence data. Figure 4.1 (panels B and C) contrasts the traditional low-throughput process with the new process based on high-throughput sequencing. This process is only possible due to the high throughput of the new sequencing platform (millions of sequences rather than dozens).

4.3 Results

4.3.1 Kds can be predicted with a high level of accuracy.

We calculated the K_D for each of the 4198 most abundant unique sequences. From those sequences, we chose 17 clones to test and compare to the K_Ds we calculated from the pooled sequence data. We chose clones that were representative of the entire sequence pool, covering the full range of abundance and K_D. These clones were therefore chosen based on range of abundances in the sequence pool (~500 to >50,000) and range of predicted K_D (~15 to ~300 μM). We measured the K_D of the clones individually using the traditional affinity chromatography method for comparison to K_Ds calculated with our new high-throughput method. The measured K_D values were

compared to the high-throughput K_Ds at the 4 degeneracy levels. We found that the 90% degeneracy level showed the least amount of error in calculated values and used this as our standard allowed degeneracy. We were able to predict the K_Ds of histidine binding RNAs by our new high-throughput method within a factor of 4 (mean=1.65, standard deviation=0.797) for all of the 17 tested RNAs (Table 4.1). We noticed one of the 17 clones (clone 118) deviates by a factor of 4 with the measured K_D, while all other clones deviate by a factor of 3 or less. This is consistent with the variation between individual affinity chromatography K_D determinations, where a three-fold difference among replicated measurements of the same RNA is not uncommon (39-42).

Sequence ID	Total Count	Predicted K_D (M)	Measured K_D (M)	Difference
6	533	8.52E-05	3.92E-05	2.17
22	10234	1.36E-05	7.00E-06	1.94
46	29943	1.87E-05	2.72E-05	1.45
47	53896	2.49E-05	2.50E-05	1.01
49	26232	1.56E-04	2.62E-04	1.68
54	2696	4.90E-05	4.67E-05	1.05
55	618	6.39E-05	1.20E-04	1.88
59	37767	1.99E-05	6.90E-06	2.89
60	8487	2.30E-04	2.16E-04	1.07
68	11209	5.12E-05	4.54E-05	1.13
71	21625	1.82E-05	1.31E-05	1.39
87	1192	1.42E-05	2.50E-05	1.76
95	3380	2.85E-04	3.60E-04	1.26
113	1100	8.42E-05	7.50E-05	1.12
118	15832	2.90E-04	7.20E-05	4.03
120	1024	6.30E-05	7.30E-05	1.16
144	12414	4.74E-05	4.85E-05	1.02

Table 4.1. Dissociation constants for 17 clones calculated from sequencing data compared to values measured individually.

4.3.2 Kds can be predicted for low abundance sequences.

To test whether the calculations depended on a minimum sequence abundance, we tested for correlations between the abundance of a given sequence and its calculated K_D. We calculated the K_D for each sequence at each rarefied level in the rarefaction analysis. The range of sequence abundances for the 17 tested clones was ~200 to >50,000 copies. There was no significant correlation ($r=-0.0136$, $p=0.51$) between abundance and K_D (Figure 4.2), demonstrating that we could confidently calculate the K_D even for low-abundance sequences. A correlation between sequence abundance and calculated K_D would suggest that K_D was a function of abundance, and there would be a threshold abundance for which we could accurately calculate K_D. This confirmed that we could not use absolute abundance as a method for filtering sequences from the analysis.

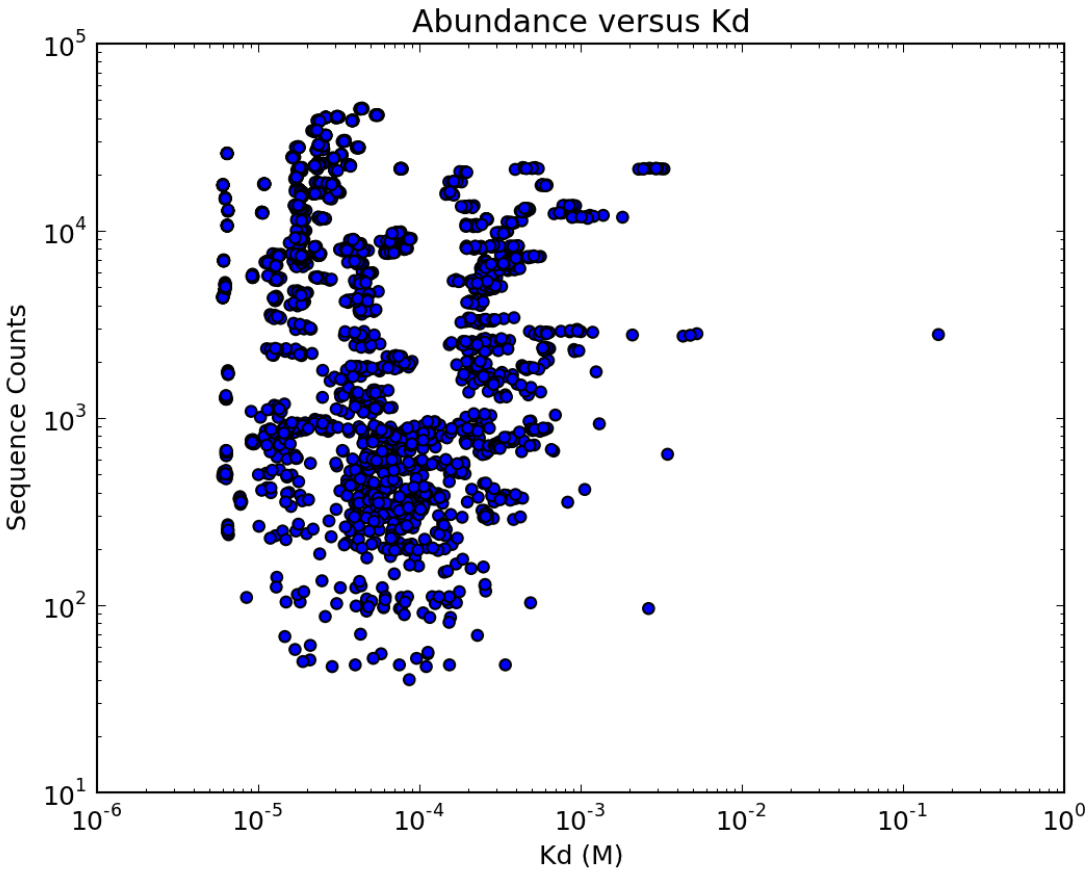


Figure 4.2. Scatterplot of sequence counts versus calculated K_D for each sequence at each rarefied level in the rarefaction analysis. Axes are on a log-log scale. We see no correlation between sequence abundance and K_D.

Since absolute abundance was not clear source of error in K_D calculation, we examined error arising from sequence abundance in each fraction collected from the affinity column. We calculated K_D for each sequence, for each of the 220 rarefaction experiments. In order to visualize the error in different rarefaction experiments, we plotted the elution profiles of the RNAs through the affinity column. An elution profile is a plot of percentage RNA eluted versus volume eluted. Figure 4.3 shows the histidine elution profile of clone 60: A) measured by radioactivity (measured K_D = 216 μM), B) the elution profile for a single rarefaction experiment where there were no sequences

missing in any fraction (shown in blue, measured K_D = 217 μM), and the elution profile for a single rarefaction experiment where there were no sequences represented in several fractions (shown in green, measured K_D = 501 μM). The arrows point to the median elution volume (Vel), which is used to calculate the K_D of the sequence. We see that the calculated K_D for the sequence where there are no counts in several fractions tends to be inaccurate compared to the sequence represented in every fraction, resulting from incorrect estimation of the median elution volume. Figure 4.4A shows the average L-His elution profile for the same clone for all 220 rarefaction experiments with error bars. When we filter out all experiments for this clone that contain fractions in which the sequence was not observed, we see a large reduction in error (Figure 4.4B), and the elution profile more closely resembles the expected elution seen in Figure 4.3A.

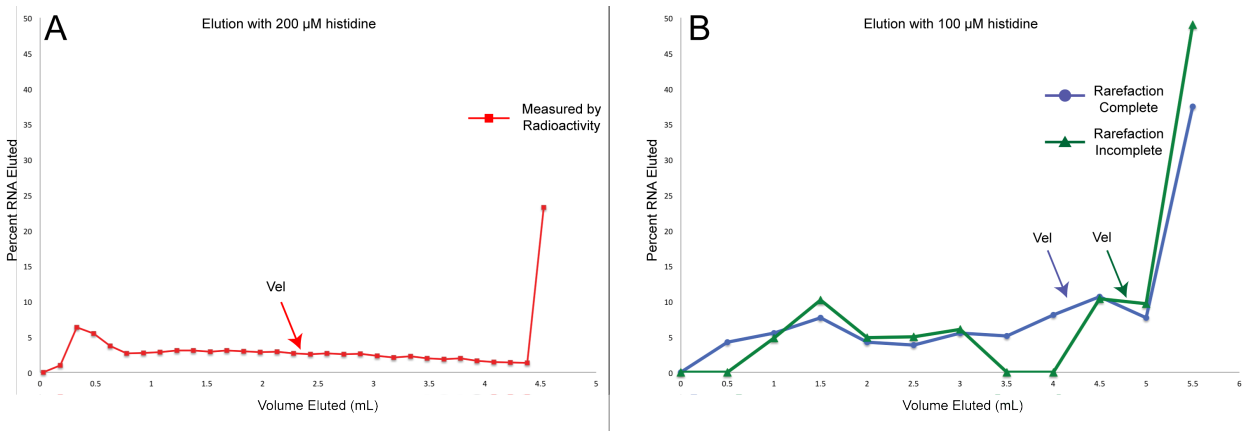


Figure 4.3. Affinity chromatography elution with free histidine for clone 60. A) Elution with 200 μM histidine measured by radioactivity. B) Elution with 100 μM histidine for a single rarefaction experiment no sequences were missing in any fraction (blue), and a single rarefaction experiment where there were no sequences represented in several fractions (green). The arrows point to the median elution volume (Vel), which is used to calculate the K_D of the sequence.

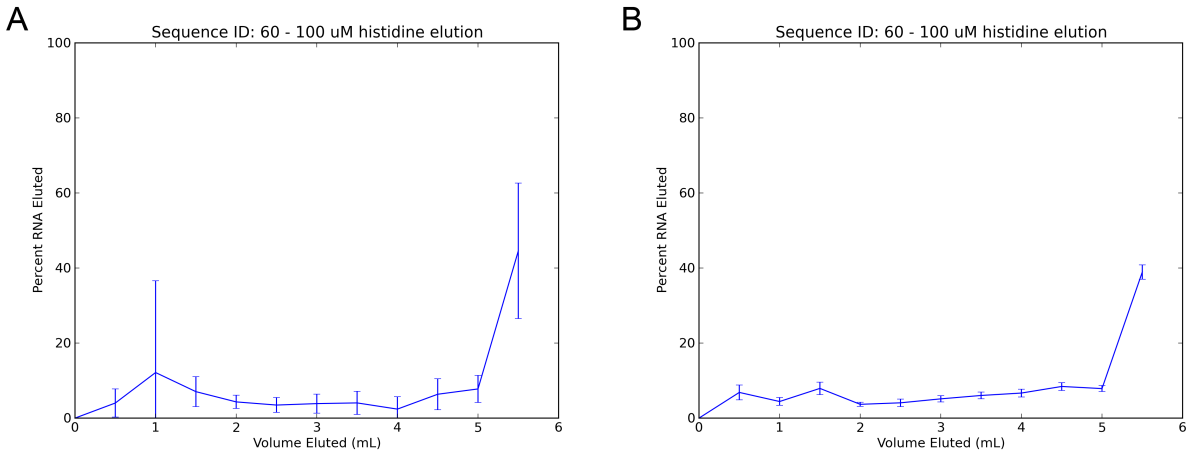


Figure 4.4. (A) Average histidine elution for the clone 60, combining all 220 rarefaction experiments with error bars. (B) Average histidine elution for clone 60, when we filter out all experiments for this clone where there are any fractions without sequence counts, showing a large reduction in error.

This trend, that fractions in which a given sequence is missing lead to high error rates for that sequence, can be demonstrated by plotting the number of fractions where no sequence counts were observed against the calculated K_D for each rarefaction experiment. Figure 4.5 shows examples of how calculated K_D is affected by missing sequences in multiple fractions. In some rare cases, we can still accurately calculate K_D even when the sequence is missing from multiple fractions, although these are the exception to the rule (Figure 4.5A). The accuracy of K_D calculation usually decreases when the sequence is missing from any fraction (Figure 4.5B). For all K_D calculations, we chose to filter out any sequences that were not represented in every fraction of the affinity elution.

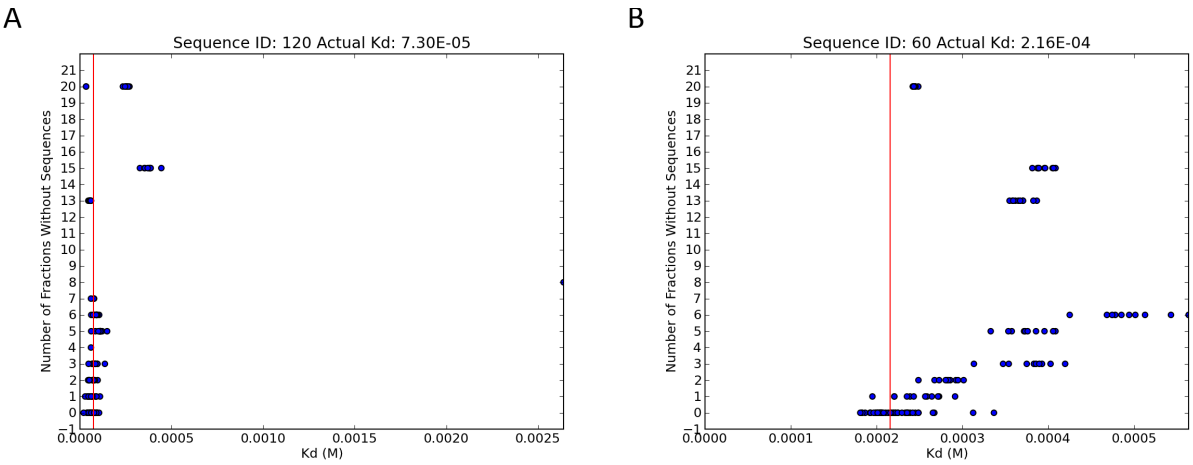


Figure 4.5. Plot of number of fractions with not sequence counts versus K_D from rarefaction experiments for two different clones. The red vertical lines show actual measured K_D. (A) A rare example where we can still accurately calculate K_D even when the sequence is missing from multiple fractions. (B) Typical example of how the accuracy of K_D calculation usually decreases when the sequence is missing from any fraction.

Therefore, our hypothesis that a minimum sequence abundance is required in order to accurately calculate K_D, was incorrect: we saw no overall correlation between K_D and

abundance. Instead, the main source of error comes from trying to calculate the K_D of sequences that are absent from several fractions from a column.

4.4 Discussion

We have shown here that we are able to accurately calculate K_Ds for thousands of sequences in parallel. For the sequences we tested, we were able to calculate K_Ds in parallel within a factor of 4 or less compared to measured values for each sequence individually (Table 4.1). This level of accuracy is notable, especially for an experiment performed without replication for a single ligand concentration, and with relatively granular fractionation of eluted RNA compared to traditional methods.

One feature of this work was that the abundance of a sequence and its K_D are uncorrelated: in other words, the sequences with the best K_Ds do not become dominant in the pool, as would be expected in a selection for high performance binding. This is a feature of the way the selection experiment was designed. The selection was performed in a way to maximize variability in the resulting aptamer pool in order to find the simplest motif that would bind histidine. We expected that there might be some minimum abundance requirements that we could use as a filtering method to remove sequences where we would be unable to calculate K_D accurately. We also anticipated an alternative possibility, in which we would see more sequences with higher K_Ds on the assumption that such sequences would be more common because the requirements are less stringent, but there proved to be no correlation between abundance and K_D across our tested sequences ($r = -0.201$, $p = 0.439$). The ability to determine activity for low-abundance sequences is especially encouraging, suggesting that the methods we introduce here will be useful for characterizing a wide range of the kinds of binding sites

present in a given SELEX pool. The main factor causing error in our K_D calculations instead comes from the number of sequences that are absent from multiple fractions of a column, and filtering by this criterion greatly reduces time spent on attempting to perform calculations that in the end will not be accurate. However, although we can accurately calculate K_Ds for low-abundance sequences, increasing sequencing depth would likely alleviate errors in calculations due to sparse representation of particular low-abundance sequences by sequencing error.

An important consideration for future experimental design is to take into account the possible systematic errors with the calculation of K_D using this method. Since the K_D calculation depends on determining the median elution volume of an RNA with and without free ligand in buffer, the total volume eluted must be sufficient to 50% of all RNAs present. RNAs with a higher affinity for the column will require larger elution volumes. RNAs with a lower affinity for the column will elute faster, potentially resulting in an error in the K_D calculation. Taking smaller fraction volumes and varying the free ligand concentration in the competitive elution should address this issue.

The results we report here were obtained via a single experiment at only one ligand concentration, which is substantially less laborious than typical K_D experiments that use multiple ligand calculations. Repeating the calculations at multiple ligand concentrations and/or performing replicate experiments at the same ligand concentration might reduce the error further. However, reductions in K_D error determination would not affect any of the major conclusions of the present work, although they might be useful for other applications. The results shown here demonstrate the feasibility of using high-throughput sequencing to rapidly characterize

the range of activities present in a pool of aptamers, and could potentially be used to trace the process of evolution of a wide range of RNA activities.

4.5 Closing Statement

In this chapter I showed that I was able to accurately calculate the K_D for thousands of sequences derived from SELEX. An exciting feature of this work was the ability to quantify the activities of these RNAs at an unprecedented depth. We were also able to determine the K_D for even low abundance sequences, which suggests this technique will be useful for analyzing selections where the motif of interest may not be the most abundant. Combining this technique to relate activity to active site motifs, elucidated by quality structural alignments, will aid in the understanding of how RNA evolves function.

4.6 Materials and Methods

4.6.1 RNA selection pools of histidine aptamers

Starting from randomized RNA pools and enriching by affinity chromatography we have previously selected for histidine aptamers (41,163). The cycle 5 pool from the selection described in (163) was used in the experiment described here. This pool contains sequences with initially randomized regions 28,25 and 22 nucleotides long. In addition, an isolate from the selection described in (41), selected from an originally 70 nucleotide long randomized region, was used in the control experiment described below.

4.6.2 Affinity chromatography support

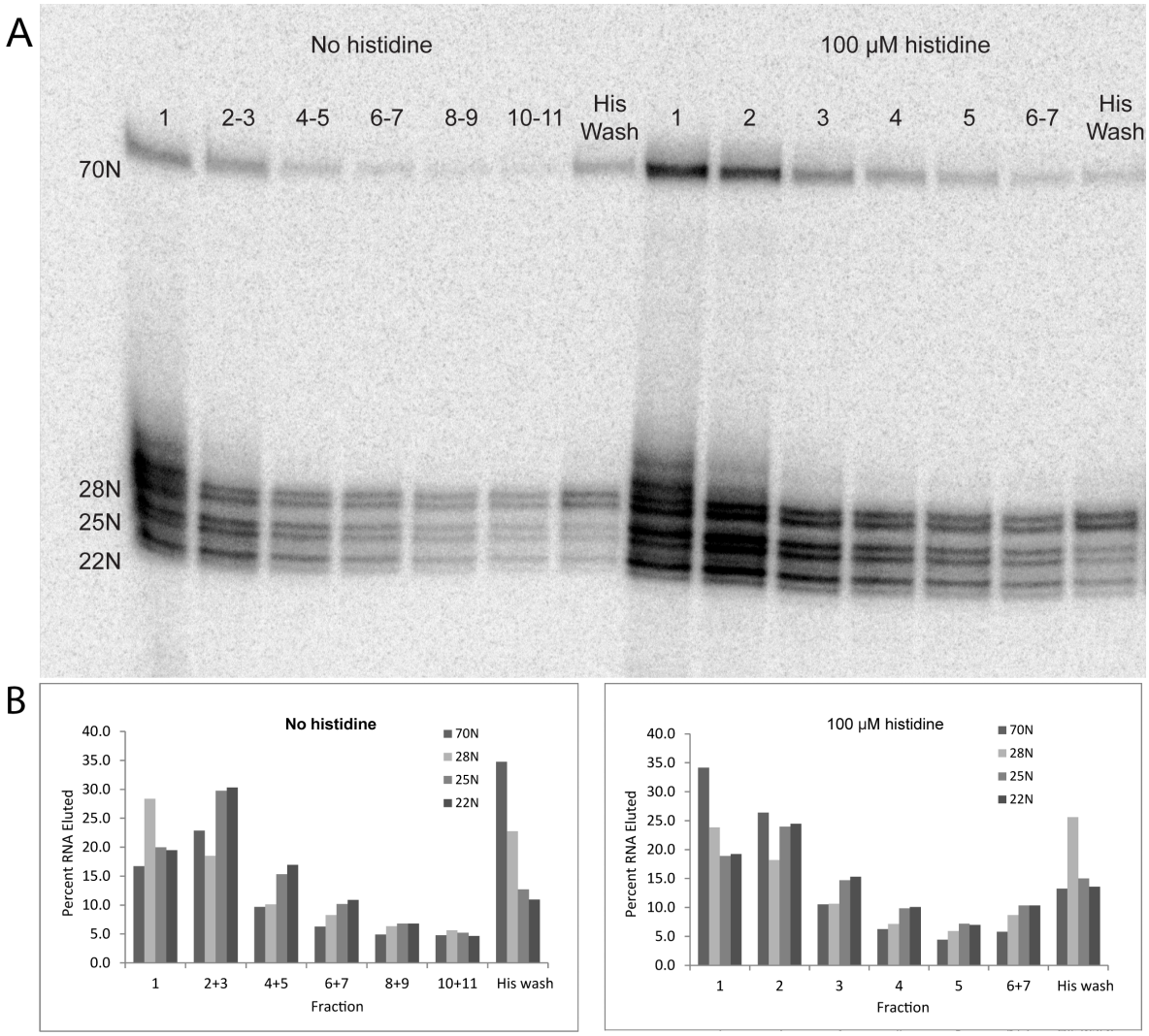
F-moc protected histidine was coupled to Controlled-Pore Glass (CPG) beads (Millipore, 125-177 μm) as previously described (40). The concentration of histidine was approximately 1 mM.

4.6.3 Biochemical K_D determination

Dissociation constants for histidine were determined by isocratic elution from the affinity matrix with and without ligand using $K_D = L (Vel – Vn)/ (Ve – Vel)$ where L is the concentration of free ligand in the buffer, Ve and Vel are the median elution volumes of the RNA in the absence and presence of the ligand in the buffer and Vn is the median elution volume in the absence of any affinity (173). Buffers used in all experiments were as previously described (163).

4.6.4 Control experiment

As a proof-of-concept for the idea that we could calculate the K_D for many sequences at once, we performed a control experiment to calculate the K_D for 4 clones in one set of reactions. In this experiment, we selected 4 histidine aptamers of different lengths (22, 25, 28, and 70 nucleotides) and eluted the RNAs through the affinity chromatography column. These beads were achiral, important for experiments in which the chirality of the substrate might potentially be of interest. The RNA was eluted through the affinity column both with and without free ligand in elution buffer. We collected 11, 0.6 mL fractions without free ligand, and four 0.6 mL fractions with 5 mM histidine to wash remaining RNA from the column. We collected 7, 0.6 mL fractions with 100μM histidine and four 0.6 mL fractions with 5 mM histidine as a final wash. To quantify RNA eluted in each fraction, we separated the RNAs by electrophoresis on a denaturing acrylamide gel, imaged using a phosphorimager (Figure 4.6). Because each RNA was a different length, we could visualize and quantify each RNA individually. The K_Ds were then calculated for each of these 4 clones, and were found to agree with previous K_D calculations (Figure 4.6C).



C

	Calculated K_D (M)	Measured K_D (M)
70N	2.7e-5	8.0e-6
28N	2.0e-4	2.16e-4
25N	3.6e-4	3.60e-4
22N	3.0e-4	not determined

Figure 4.6. Column chromatography of a mixture of four internally labelled [P^{32}]RNA samples of different lengths in the absence and presence of histidine in the elution buffer. (A) Autoradiogram of collected fractions after gel electrophoresis in denaturing acrylamide, imaged using a phosphorimager. Numbers of the left refer to the RNA initially randomized length. For the column without histidine in buffer, two fractions were generally combined and the 5 mM final histidine wash in both columns represents four fractions combined. Shorter sequences appear as doublets or triplets due to the frequent addition of non-templated nucleotides by T7 RNA polymerase. (B) Bar graph of phosphorimager quantitation of bands, corrected for different volumes. (C) The dissociation constants obtained are shown in comparison to values independently obtained for the sequences individually.

4.6.5 Fractionation and Illumina sequencing

Having demonstrated the principle that individual RNAs with known K_Ds would not interfere with one another, we then validated our high-throughput K_D determination pipeline via high-throughput sequencing using RNAs from the 5th cycle of a previously published histidine binding selection (163). Although his selection was originally performed over 6 cycles, we chose to use the 5th cycle pool in order to measure a wider range of K_Ds than would be expected in sequences surviving to the 6th cycle. This selection was started from sequences with randomized regions between 22 and 28 nucleotides long. The affinity matrix used for K_D determination was the same as used in the original selections.

Dissociation constants were determined by isocratic affinity chromatography, an equilibrium method, which has been previously shown to give results equivalent to ultrafiltration, equilibrium dialysis, and protections (41,42). RNA was eluted through the affinity column both with and without free ligand in buffer. We collected 20, 0.5 mL fractions without free ligand present in buffer and one final 2.5mL fraction with 5mM histidine to wash remaining RNA from the affinity column. We collected 10, 0.5mL fractions with free 100μM histidine in buffer and one final 2.5mL fraction with 5mM histidine to wash remaining RNA from the affinity column. The RNA was reverse transcribed and each of these fractions was prepared for sequencing on the Illumina HiSeq 2000 platform, as previously described (177). A unique 12-nucleotide error-correcting barcode (176) was attached to the sequences in each of the 32 total fractions. This allowed us to associate the sequences with the appropriate fraction they

eluted from, which is necessary for calculating K_D. We obtained 19,109,089 reads from the Illumina sequencing run before filtering. Quality-filtering was performed using the QIIME software package (179) as described previously (177). In this filtering process, we excluded truncated sequences and sequences that contained a barcode that could not be mapped to any of the 32 known barcodes, resulting in 10,544,189 reads (55% of the total; this inefficiency was likely due to the relatively low complexity of the sequence pool, which causes problems on the Illumina platform due to undersaturation of spots with some base flows).

4.6.6 Computational K_D determination

The entire collection of sequences was organized first by barcode sequence (each barcode corresponded to an individual fraction from the experiment), then by column (with or without free histidine in the elution buffer). We tabulated counts of each sequence in each fraction. All sequences that were not observed (zero counts) any of the total 32 fractions were removed from the analysis. For each sequence, the median elution volume was determined in the presence and absence of free histidine in buffer. This median elution volume is calculated by determining the total number of each individual sequence in the column. The K_D of the sequence was then calculated as described above.

4.6.7 Combining similar sequences

When calculating the abundance of each sequence in each fraction, we originally chose only count sequences that were identical. We found that many sequences were not represented in several fractions of the affinity column and therefore we were unable to calculate K_Ds for very low abundance sequences. In order to increase the number of

counts per sequence in each fraction, we chose to combine similar sequences into a single degenerate “master sequence”. Sequences were combined at 95%, 90%, 86% and 81% identity allow 1-4 mismatches respectively using the QIIME package (179). We performed K_D calculations for all sequences at these degeneracy levels.

4.6.8 Rarefaction analysis

A rarefaction analysis is typically performed in ecological studies to assess species richness in each environmental sample. In a rarefaction analysis, a pool of sequences is randomly re-sampled a number of times at decreasing intervals and the abundance of individual sequences is calculated. The results of this analysis are used to determine whether sampling is complete: a rarefaction curve that reaches an asymptote implies that additional sampling will not uncover new types of organisms (or sequences) (180,181). We performed a rarefaction analysis in order to determine the minimum number of sequences required for accurate K_D calculation, and to test how sampling error affected these calculations. The rarefaction analysis was performed using the QIIME package (179). The entire sequence pool obtained from Illumina sequencing was separated by barcode, corresponding to which fraction it eluted from in the affinity chromatography experiment. For each fraction, we randomly sampled n sequences, where n ranges from 450,000 to 18,000 in intervals of 20,000. This sampling was performed 10 times at each interval, resulting in 220 rarefaction experiments.

Chapter 5: Conclusions and future directions

5.1 Development of high-quality RNA alignments and computational tools is fundamental for understanding function and evolution of RNA.

As the work in this thesis demonstrates, multiple sequence alignments are fundamental for studying the structure, function and evolution of RNAs, and improvements in alignment techniques, especially applied to the very large numbers of sequences now available, provide insights not previously attainable. Building a collection of RNA alignments using Watson-Crick and noncanonical base pair information from known structures obtained by NMR and x-ray crystallography demonstrably improves the quality of the alignments, and these alignments will be important for use in further analyses of RNAs and for validating new structure-aware alignment tools. In Chapter 2, I showed results from manual curation of 148 non-coding RNA alignments. For these alignments, I first associated a 3d structure to its corresponding family of sequences. Then, using the base pairing information in the 3d structure, I manually aligned the sequences utilizing IDI/isostericity. This manual curation was a time-consuming, but necessary step for the end goal of providing a training set to allow development of computational tools to make better alignments and better predict RNA structures. These alignments can be used for assessing the performance of current alignment algorithms, and can also give a clear direction for further algorithm development. The utilization of IDI/isostericity for scoring RNA sequence alignments and structure predictions should also prove invaluable for future development of RNA computational tools.

5.2 Growth of GenBank and of sequencing technologies, and challenges encountered in aligning large numbers of sequences

As sequencing technology becomes cheaper and cheaper, the development of computational tools to handle these large amounts of sequences will be fundamental to the study of RNAs. New sequencing technologies can produce millions to billions of sequences from a single experiment, which now makes it impossible to perform the manual curation of alignments as we have done in the past. Additionally, the size of sequence repositories is rapidly growing due to this influx of sequence data. GenBank, which currently holds over 150 billion sequences, has consistently doubled in size every 18 months. Having a quality set of alignments with useful structural annotation can be leveraged to help alleviate this problem, because new sequences can be inserted into these existing high-quality alignments. The development of tools that can handle these more complete alignments and integrate related sequences without exacerbating the pitfalls of traditional alignment methods is becoming necessary.

5.3 Implications for evolution from understanding structure-backed sequence alignments

These high-quality, structure-backed alignments can also have implications for studying evolution and for the identification of RNA motifs associated with various functions. The ability to identify these motifs will allow us to compare different non-coding RNAs to find similarities in functions based on the motifs they share (rather than by similarities at the level of primary sequence), and should give insight into how easily RNA can evolve a function. The inference of phylogenetic relationships between sequences is fundamentally dependent upon the sequence alignment. With high-quality

alignments, for example, we will be able to test with more accuracy whether a function has evolved slowly over time from a single common ancestor, or if an RNA function has independently arisen multiple times throughout the course of evolution.

5.4 Implications for evolution from understanding tRNAs

In Chapter Three, I discussed the implications for evolution of a better understanding of relationships among tRNAs, where I used UniFrac to show phylogenetically meaningful relationships among genomes based on the relationships among the tRNAs they contain. This work provides model for multi-gene phylogenies from whole-genome sequences. It was thought that tRNAs would be poor candidates for phylogeny because they are too short, they are often duplicated (i.e. paralogy is extensive), they can change specificity by as little as a single-nucleotide change, and can be involved in horizontal gene transfer e.g. as targets for retrovirus and other mobile element insertion. Therefore, it would not be expected that a phylogenetic signal relating the organisms could be recovered through the tRNAs alone. However; using UniFrac, we were able to show that the tRNAs, when aggregated and labeled according to the genome that they came from, produce a UniFrac dendrogram that clusters like the rRNAs from the same organisms: the rRNA trees are currently the gold standard for determining universal phylogeny. This same technique can be extended to any other biological system. By using UniFrac to cluster functional genomic data, we can discover if gene families have evolved phylogenetically or through some other selective pressure, e.g. convergent evolution, despite substantial noise in the data.

5.5 Using high-throughput activity measures to group functionally related motifs can help to determine RNA function from sequence.

In Chapter Four, I discussed how using SELEX and high-throughput sequencing to measure activity. I was able to determine the K_D for histidine for over 4,000 sequences in parallel, providing a substantial improvement in speed relative to the traditional process of determining K_D one sequence at a time. These predictions were within 4-fold of the experimentally determined K_D, and were thus comparable to the variability among replicate experiments using the traditional method. With the advances in sequencing technologies, we can begin using high-throughput activity measures to pool functionally related motifs, which can help to determine RNA function from sequence. With the ability to determine the activity of thousands of sequences from a given pool, we will be able to develop a metric for comparing motifs. This metric of comparison will also allow us to find the limits of variation between RNA motifs that affect function. Having the ability to associate activity level to sequences, thereby allowing us to identify functional motif categories, is a necessary step for predicting the function of an RNA based on its sequence. Combining this high-throughput activity determination with the ability to build high-quality structure backed alignments will allow for an unprecedented ability to build motif models and search genomes for similar functions. Additionally, one could search an annotated motif alignment collection for matches to a newly discovered RNA to predict the function of that RNA.

5.6 Prospects for improving SELEX

The development of high-throughput sequencing technology has allowed us to examine a SELEX pool at a level of depth never before achieved. This allowed us to

identify rare sequences/motifs in the pool, which could not have been seen by traditional cloning/Sanger methods. We were able to obtain enough data to observe the process of evolution during a SELEX experiment: this ability to measure activity in high-throughput will allow us to determine the fundamental parts of the RNA, and the relationships among these parts, that are responsible for differences in activity.

These techniques can be extended to emerging fields utilizing SELEX to obtain aptamer biosensors as biomarkers for disease states, where rapid and deep analysis of a SELEX pool will dramatically improve the rate of development of these biomarkers. The ability to determine the range of binding activities for all aptamers to a target will also aid in aptamer-based drug development, where tuning the affinity of a drug to target is essential for efficacy. The ability to classify RNAs by activity level, combined with the advancements in sequence alignment presented here, is an important step towards understanding how RNA evolves function. A direct extension of this research would thus be to analyze each cycle of an *in vitro* selection by these techniques, to monitor which structural motifs are enriched and how variations in these motifs modulate activity.

Works Cited

1. Cech, T.R. (2002) Ribozymes, the first 20 years. *Biochemical Society transactions*, **30**, 1162-1166.

2. Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E. and Cech, T.R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, **31**, 147-157.

3. Cech, T.R., Zaug, A.J. and Grabowski, P.J. (1981) In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, **27**, 487-496.

4. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. and Altman, S. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **35**, 849-857.

5. Guerrier-Takada, C., Haydock, K., Allen, L. and Altman, S. (1986) Metal ion requirements and other aspects of the reaction catalyzed by M1 RNA, the RNA subunit of ribonuclease P from Escherichia coli. *Biochemistry*, **25**, 1509-1515.

6. Steitz, T.A. and Moore, P.B. (2003) RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends in biochemical sciences*, **28**, 411-418.

7. Ban, N., Nissen, P., Hansen, J., Moore, P.B. and Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science (New York, N.Y.)*, **289**, 905-920.

8. Nissen, P., Hansen, J., Ban, N., Moore, P.B. and Steitz, T.A. (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science (New York, N.Y.)*, **289**, 920-930.

9. Schmeing, T.M., Seila, A.C., Hansen, J.L., Freeborn, B., Soukup, J.K., Scaringe, S.A., Strobel, S.A., Moore, P.B. and Steitz, T.A. (2002) A pre-translocational intermediate in protein synthesis observed in crystals of enzymatically active 50S subunits. *Nature structural biology*, **9**, 225-230.

10. Hansen, J.L., Schmeing, T.M., Moore, P.B. and Steitz, T.A. (2002) Structural insights into peptide bond formation. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11670-11675.

11. Noller, H.F., Hoffarth, V. and Zimniak, L. (1992) Unusual resistance of peptidyl transferase to protein extraction procedures. *Science (New York, N.Y.)*, **256**, 1416-1419.

12. Crick, F.H. (1968) The origin of the genetic code. *Journal of molecular biology*, **38**, 367-379.

13. Orgel, L.E. (1968) Evolution of the genetic apparatus. *Journal of molecular biology*, **38**, 381-393.

14. Cech, T.R. (1993) The efficiency and versatility of catalytic RNA: implications for an RNA world. *Gene*, **135**, 33-36.

15. Joyce, G.F. and Orgel, L.E. (1993) Prospects for Understanding the Origin of the RNA World. *In the RNA World*, (eds. Gestland R.F, Atkins J.F), pp. 1-25 Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

16. Orgel, L.E. (2004) Prebiotic chemistry and the origin of the RNA world. *Critical reviews in biochemistry and molecular biology*, **39**, 99-123.

17. Tuerk, C. and Gold, L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science (New York, N.Y.)*, **249**, 505-510.

18. Ellington, A.D. and Szostak, J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818-822.

19. Robertson, D.L. and Joyce, G.F. (1990) Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature*, **344**, 467-468.

20. Famulok, M. and Jenne, A. (1998) Oligonucleotide libraries--variatio delectat. *Current opinion in chemical biology*, **2**, 320-327.

21. Gold, L., Janjic, N., Jarvis, T., Schneider, D., Walker, J.J., Wilcox, S.K. and Zichi, D. (2012) Aptamers and the RNA world, past and present. *Cold Spring Harbor perspectives in biology*, **4**.

22. Famulok, M. (1999) Oligonucleotide aptamers that recognize small molecules. *Current opinion in structural biology*, **9**, 324-329.

23. Brown, J.W., Birmingham, A., Griffiths, P.E., Jossinet, F., Kachouri-Lafond, R., Knight, R., Lang, B.F., Leontis, N., Steger, G., Stombaugh, J. *et al.* (2009) The RNA structure alignment ontology. *RNA (New York, N.Y.)*, **15**, 1623-1631.

24. Keiler, K.C., Waller, P.R. and Sauer, R.T. (1996) Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science (New York, N.Y.)*, **271**, 990-993.

25. Williams, K. (2004) Evolutionary resealing of a split RNA: Reversal of gene permutation. *RNA (New York, N.Y.)*, **10**, 555-557.

26. Korf, I., Yandell, M. and Bedell, J. (2003) *BLAST*. O'Reilly \& Associates, Inc.

27. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, **22**, 4673-4680.

28. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**, 1792-1797.

29. Salehi-Ashtiani, K. and Szostak, J.W. (2001) In vitro evolution suggests multiple origins for the hammerhead ribozyme. *Nature*, **414**, 82-84.

30. Nawrocki, E.P., Kolbe, D.L. and Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics (Oxford, England)*, **25**, 1335-1337.

31. Will, S., Reiche, K., Hofacker, I.L., Stadler, P.F. and Backofen, R. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS computational biology*, **3**, e65.

32. Sankoff, D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics.*, **45**, 810-825.

33. Stombaugh, J., Widmann, J., McDonald, D. and Knight, R. (2011) Boulder ALignment Editor (ALE): a web-based RNA alignment tool. *Bioinformatics (Oxford, England)*, **27**, 1706-1707.

34. Yang, Z. and Rannala, B. (2012) Molecular phylogenetics: principles and practice. *Nature reviews. Genetics*, **13**, 303-314.

35. Aquino-Jarquin, G. and Toscano-Garibay, J.D. (2011) RNA aptamer evolution: two decades of SELEction. *International journal of molecular sciences*, **12**, 9155-9171.

36. Helm, M., Petermeier, M., Ge, B., Fiammengo, R. and Jaschke, A. (2005) Allosterically activated Diels-Alder catalysis by a ribozyme. *Journal of the American Chemical Society*, **127**, 10492-10493.

37. Kang, K.N. and Lee, Y.S. (2012) RNA Aptamers: A Review of Recent Trends and Applications. *Advances in biochemical engineering/biotechnology*.

38. Famulok, M., Hartig, J.S. and Mayer, G. (2007) Functional aptamers and aptazymes in biotechnology, diagnostics, and therapy. *Chemical reviews*, **107**, 3715-3743.

39. Majerfeld, I. and Yarus, M. (1998) Isoleucine:RNA sites with associated coding sequences. *RNA (New York, N.Y.)*, **4**, 471-478.

40. Illangasekare, M. and Yarus, M. (2002) Phenylalanine-Binding RNAs and Genetic Code Evolution. *Journal of molecular evolution*, **54**, 298-311.

41. Majerfeld, I., Puthenvedu, D. and Yarus, M. (2005) RNA affinity for molecular L-histidine; genetic code origins. *Journal of molecular evolution*, **61**, 226-235.

42. Majerfeld, I. and Yarus, M. (2005) A diminutive and specific RNA binding site for L-tryptophan. *Nucleic acids research*, **33**, 5482-5493.

43. Yarus, M., Widmann, J.J. and Knight, R. (2009) RNA-amino acid binding: a stereochemical era for the genetic code. *Journal of molecular evolution*, **69**, 406-429.

44. Yarus, M., Caporaso, J.G. and Knight, R. (2005) Origins of the genetic code: the escaped triplet theory. *Annual review of biochemistry*, **74**, 179-198.

45. Symons, R.H. (1992) Small catalytic RNAs. *Annual review of biochemistry*, **61**, 641-671.

46. Perreault, J., Weinberg, Z., Roth, A., Popescu, O., Chartrand, P., Ferbeyre, G. and Breaker, R.R. (2011) Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. *PLoS computational biology*, **7**, e1002031.

47. Pabon-Pena, L.M., Zhang, Y. and Epstein, L.M. (1991) Newt satellite 2 transcripts self-cleave by using an extended hammerhead structure. *Molecular and cellular biology*, **11**, 6109-6115.

48. Zhang, Y. and Epstein, L.M. (1996) Cloning and characterization of extended hammerheads from a diverse set of caudate amphibians. *Gene*, **172**, 183-190.

49. Ferbeyre, G., Smith, J.M. and Cedergren, R. (1998) Schistosome satellite DNA encodes active hammerhead ribozymes. *Molecular and cellular biology*, **18**, 3880-3888.

50. Rojas, A.A., Vazquez-Tello, A., Ferbeyre, G., Venanzetti, F., Bachmann, L., Paquin, B., Sbordoni, V. and Cedergren, R. (2000) Hammerhead-mediated processing of satellite pDo500 family transcripts from Dolichopoda cave crickets. *Nucleic acids research*, **28**, 4037-4043.

51. Tang, J. and Breaker, R.R. (2000) Structural diversity of self-cleaving ribozymes. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 5784-5789.

52. Gardner, P.P., Daub, J., Tate, J., Moore, B.L., Osuch, I.H., Griffiths-Jones, S., Finn, R.D., Nawrocki, E.P., Kolbe, D.L., Eddy, S.R. *et al.* (2010) Rfam: Wikipedia, clans and the “decimal” release. *Nucleic acids research*.

53. Cho, M., Xiao, Y., Nie, J., Stewart, R., Csordas, A.T., Oh, S.S., Thomson, J.A. and Soh, H.T. (2010) Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 15373-15378.

54. Schutze, T., Wilhelm, B., Greiner, N., Braun, H., Peter, F., Morl, M., Erdmann, V.A., Lehrach, H., Konthur, Z., Menger, M. *et al.* (2011) Probing the SELEX process with next-generation sequencing. *PloS one*, **6**, e29604.

55. Hoon, S., Zhou, B., Janda, K.D., Brenner, S. and Scolnick, J. (2011) Aptamer selection by high-throughput sequencing and informatic analysis. *BioTechniques*, **51**, 413-416.

56. Pitt, J.N. and Ferre-D'Amare, A.R. (2010) Rapid construction of empirical RNA fitness landscapes. *Science (New York, N.Y.)*, **330**, 376-379.

57. Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpaa, M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome research*, **20**, 861-873.

58. Thompson, J.D., Holbrook, S.R., Katoh, K., Koehl, P., Moras, D., Westhof, E. and Poch, O. (2005) MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. *Nucleic Acids Res*, **33**, 4164-4171.

59. Cruz, J.A. and Westhof, E. Sequence-based identification of 3D structural modules in RNA with RMDetect. *Nat Methods*, **8**, 513-521.

60. Loytynoja, A. and Goldman, N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632-1635.

61. Wong, K.M., Suchard, M.A. and Huelsenbeck, J.P. (2008) Alignment uncertainty and genomic analysis. *Science*, **319**, 473-476.

62. Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.

63. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res*, **31**, 439-441.

64. Wang, H.C. and Hickey, D.A. (2002) Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes. *Nucleic Acids Res*, **30**, 2501-2507.

65. Smit, S., Yarus, M. and Knight, R. (2006) Natural selection is not required to explain universal compositional patterns in rRNA secondary structure categories. *RNA*, **12**, 1-14.

66. Smit, S., Widmann, J. and Knight, R. (2007) Evolutionary rates vary among rRNA structural elements. *Nucleic Acids Res*, **35**, 3339-3354.

67. Smit, S., Knight, R. and Heringa, J. (2009) RNA structure prediction from evolutionary patterns of nucleotide composition. *Nucleic Acids Res*, **37**, 1378-1386.

68. Knight, R., De Sterck, H., Markel, R., Smit, S., Oshmyansky, A. and Yarus, M. (2005) Abundance of correctly folded RNA motifs in sequence space, calculated on computational grids. *Nucleic Acids Res*, **33**, 5924-5935.

69. Gevertz, J., Gan, H.H. and Schlick, T. (2005) In vitro RNA random pools are not structurally diverse: a computational analysis. *RNA*, **11**, 853-863.

70. Gan, H.H., Pasquali, S. and Schlick, T. (2003) Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res*, **31**, 2926-2943.

71. Yarus, M. and Welch, M. (2000) Peptidyl transferase: ancient and exiguous. *Chem Biol*, **7**, R187-190.

72. Leontis, N.B. and Westhof, E. (2002) The annotation of RNA motifs. *Comp Funct Genomics*, **3**, 518-524.

73. Batey, R.T., Rambo, R.P. and Doudna, J.A. (1999) Tertiary Motifs in RNA Structure and Folding. *Angew Chem Int Ed Engl*, **38**, 2326-2343.

74. Jossinet, F., Ludwig, T.E. and Westhof, E. Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics*, **26**, 2057-2059.

75. Jaeger, L. and Chworos, A. (2006) The architectonics of programmable RNA and DNA nanostructures. *Curr Opin Struct Biol*, **16**, 531-543.

76. Gardner, P.P., Wilm, A. and Washietl, S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*, **33**, 2433-2439.

77. Leontis, N.B. and Westhof, E. (1998) Conserved geometrical base-pairing patterns in RNA. *Q Rev Biophys*, **31**, 399-455.

78. Stombaugh, J., Widmann, J., McDonald, D. and Knight, R. Boulder ALignment Editor (ALE): a web-based RNA alignment tool. *Bioinformatics*, **27**, 1706-1707.

79. Leontis, N.B., Stombaugh, J. and Westhof, E. (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res*, **30**, 3497-3531.

80. Stombaugh, J., Zirbel, C.L., Westhof, E. and Leontis, N.B. (2009) Frequency and isostericity of RNA base pairs. *Nucleic Acids Res*, **37**, 2294-2312.

81. Sarver, M., Zirbel, C.L., Stombaugh, J., Mokdad, A. and Leontis, N.B. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J Math Biol*, **56**, 215-252.

82. Juhling, F., Morl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F. and Putz, J. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic acids research*, **37**, D159-162.

83. Lee, J.F., Hesselberth, J.R., Meyers, L.A. and Ellington, A.D. (2004) Aptamer database. *Nucleic Acids Res*, **32**, D95-100.

84. Burgstaller, P. and Famulok, M. (1994) Isolation of RNA Aptamers for Biological Cofactors by In Vitro Selection. *Angewandte Chemie International Edition in English*, **33**, 1084-1087.

85. Collins, R.A. (2002) The Neurospora Varkud satellite ribozyme. *Biochem Soc Trans*, **30**, 1122-1126.

86. Famulok, M. and Huttenhofer, A. (1996) In vitro selection analysis of neomycin binding RNAs with a mutagenized pool of variants of the 16S rRNA decoding region. *Biochemistry*, **35**, 4265-4270.

87. Giedroc, D.P., Theimer, C.A. and Nixon, P.L. (2000) Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J Mol Biol*, **298**, 167-185.

88. Guo, H.C., De Abreu, D.M., Tillier, E.R., Saville, B.J., Olive, J.E. and Collins, R.A. (1993) Nucleotide sequence requirements for self-cleavage of Neurospora VS RNA. *J Mol Biol*, **232**, 351-361.

89. Jenison, R.D., Gill, S.C., Pardi, A. and Polisky, B. (1994) High-resolution molecular discrimination by RNA. *Science*, **263**, 1425-1429.

90. Jiang, F., Kumar, R.A., Jones, R.A. and Patel, D.J. (1996) Structural basis of RNA folding and recognition in an AMP-RNA aptamer complex. *Nature*, **382**, 183-186.

91. Jiang, L., Majumdar, A., Hu, W., Jaishree, T.J., Xu, W. and Patel, D.J. (1999) Saccharide-RNA recognition in a complex formed between neomycin B and an RNA aptamer. *Structure*, **7**, 817-827.

92. Jiang, L., Suri, A.K., Fiala, R. and Patel, D.J. (1997) Saccharide-RNA recognition in an aminoglycoside antibiotic-RNA aptamer complex. *Chem Biol*, **4**, 35-50.

93. Kim, Y.G., Maas, S., Wang, S.C. and Rich, A. (2000) Mutational study reveals that tertiary interactions are conserved in ribosomal frameshifting pseudoknots of two luteoviruses. *RNA*, **6**, 1157-1165.

94. Kim, Y.G., Su, L., Maas, S., O'Neill, A. and Rich, A. (1999) Specific mutations in a viral RNA pseudoknot drastically change ribosomal frameshifting efficiency. *Proc Natl Acad Sci U S A*, **96**, 14234-14239.

95. Lafontaine, D.A., Norman, D.G. and Lilley, D.M. (2002) Folding and catalysis by the VS ribozyme. *Biochimie*, **84**, 889-896.

96. Licis, N. and van Duin, J. (2006) Structural constraints and mutational bias in the evolutionary restoration of a severe deletion in RNA phage MS2. *J Mol Evol*, **63**, 314-329.

97. Pan, T., Dichtl, B. and Uhlenbeck, O.C. (1994) Properties of an in vitro selected Pb2+ cleavage motif. *Biochemistry*, **33**, 9561-9565.

98. Seelig, B. and Jaschke, A. (1999) A small catalytic RNA motif with Diels-Alderase activity. *Chem Biol*, **6**, 167-176.

99. Wallis, M.G., von Ahsen, U., Schroeder, R. and Famulok, M. (1995) A novel RNA motif for neomycin recognition. *Chem Biol*, **2**, 543-552.

100. Wang, Y., Killian, J., Hamasaki, K. and Rando, R.R. (1996) RNA molecules that specifically and stoichiometrically bind aminoglycoside antibiotics with high affinities. *Biochemistry*, **35**, 12338-12346.

101. Wang, Y. and Rando, R.R. (1995) Specific binding of aminoglycoside antibiotics to RNA. *Chem Biol*, **2**, 281-290.

102. Yang, Y., Kochoyan, M., Burgstaller, P., Westhof, E. and Famulok, M. (1996) Structural basis of ligand discrimination by two related RNA aptamers resolved by NMR spectroscopy. *Science*, **272**, 1343-1347.

103. Zimmermann, G.R., Jenison, R.D., Wick, C.L., Simorre, J.P. and Pardi, A. (1997) Interlocking structural motifs mediate molecular discrimination by a theophylline-binding RNA. *Nat Struct Biol*, **4**, 644-649.

104. Wilson, C., Nix, J. and Szostak, J. (1998) Functional requirements for specific ligand recognition by a biotin-binding RNA pseudoknot. *Biochemistry*, **37**, 14410-14419.

105. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792-1797.

106. Leontis, N.B. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499-512.

107. Hoehndorf, R., Batchelor, C., Bittner, T., Dumontier, M., Eilbeck, K., Knight, R., Mungall, C.J., Richardson, J.S., Stombaugh, J., Westhof, E. *et al.* (2011) The RNA Ontology (RNAO): An ontology for integrating RNA sequence and structure data. *Applied Ontology*, **6**, 53-89.

108. Crick, F.H.C. (1957) Discussion, in The structure of nucleic acids and their role in protein synthesis. *Biochemical Soc Symp*, **14**, 25-26.

109. Crick, F.H., Brenner, S., Klug, A. and Pieczenik, G. (1976) A speculation on the origin of protein synthesis. *Orig Life*, **7**, 389-397.

110. Eigen, M. and Winkler-Oswatitsch, R. (1981) Transfer-RNA: the early adaptor. *Naturwissenschaften*, **68**, 217-228.

111. Szathmary, E. (1993) Coding coenzyme handles: a hypothesis for the origin of the genetic code. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 9916-9920.

112. Di Giulio, M. (1994) The phylogeny of tRNA molecules and the origin of the genetic code. *Orig Life Evol Biosph*, **24**, 425-434.

113. Di Giulio, M. (2004) The origin of the tRNA molecule: implications for the origin of protein synthesis. *J Theor Biol*, **226**, 89-93.

114. Fitch, W.M. and Upper, K. (1987) The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harb Symp Quant Biol*, **52**, 759-767.

115. Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research*, **25**, 955-964.

116. Sprinzl, M. and Vassilenko, K.S. (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic acids research*, **33**, D139-140.

117. Marck, C. and Grosjean, H. (2002) tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA (New York, N.Y.)*, **8**, 1189-1232.

118. Giege, R., Sissler, M. and Florentz, C. (1998) Universal rules and idiosyncratic features in tRNA identity. *Nucleic acids research*, **26**, 5017-5035.

119. Ardell, D.H. and Andersson, S.G. (2006) TFAM detects co-evolution of tRNA identity rules with lateral transfer of histidyl-tRNA synthetase. *Nucleic acids research*, **34**, 893-904.

120. Saks, M.E., Sampson, J.R. and Abelson, J. (1998) Evolution of a transfer RNA gene through a point mutation in the anticodon. *Science (New York, N.Y.)*, **279**, 1665-1670.

121. Di Giulio, M. (1995) The phylogeny of tRNAs seems to confirm the predictions of the coevolution theory of the origin of the genetic code. *Orig Life Evol Biosph*, **25**, 549-564.

122. Di Giulio, M. (1999) The non-monophyletic origin of the tRNA molecule. *J Theor Biol*, **197**, 403-414.

123. Di Giulio, M. (2006) The non-monophyletic origin of the tRNA molecule and the origin of genes only after the evolutionary stage of the last universal common ancestor (LUCA). *J Theor Biol*, **240**, 343-352.

124. Eigen, M., Lindemann, B.F., Tietze, M., Winkler-Oswatitsch, R., Dress, A. and von Haeseler, A. (1989) How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science (New York, N.Y.)*, **244**, 673-679.

125. Eigen, M. and Winkler-Oswatitsch, R. (1981) Transfer-RNA, an early gene? *Naturwissenschaften*, **68**, 282-292.

126. Cedergren, R.J., LaRue, B., Sankoff, D., Lapalme, G. and Grosjean, H. (1980) Convergence and minimal mutation criteria for evaluating early events in tRNA evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **77**, 2791-2795.

127. Sankoff, D., Cedergren, R.J. and McKay, W. (1982) A strategy for sequence phylogeny research. *Nucleic acids research*, **10**, 421-431.

128. Saks, M.E. and Conery, J.S. (2007) Anticodon-dependent conservation of bacterial tRNA gene sequences. *RNA (New York, N.Y.)*.

129. Canchaya, C., Fournous, G. and Brussow, H. (2004) The impact of prophages on bacterial chromosomes. *Mol Microbiol*, **53**, 9-18.

130. Williams, K.P. (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic acids research*, **30**, 866-875.

131. Fouts, D.E. (2006) Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic acids research*.

132. Ou, H.Y., Chen, L.L., Lonnen, J., Chaudhuri, R.R., Thani, A.B., Smith, R., Garton, N.J., Hinton, J., Pallen, M., Barer, M.R. *et al.* (2006) A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria. *Nucleic acids research*, **34**, e3.

133. Lozupone, C., Hamady, M. and Knight, R. (2006) UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, **7**, 371.

134. Lozupone, C. and Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*, **71**, 8228-8235.

135. Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical Taxonomy*. Freeman, San Francisco.

136. Ley, R.E., Backhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D. and Gordon, J.I. (2005) Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 11070-11075.

137. Lozupone, C.A., Hamady, M., Kelley, S.T. and Knight, R. (2007) Quantitative and qualitative {beta} diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol*.

138. Lozupone, C.A., Hamady, M., Cantarel, B.L., Coutinho, P.M., Henrissat, B., Gordon, J.I. and Knight, R. (2008) The convergence of carbohydrate active gene repertoires in human gut microbes. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 15076-15081.

139. Ibba, M., Bono, J.L., Rosa, P.A. and Soll, D. (1997) Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete Borrelia burgdorferi. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 14383-14388.

140. Bonnet, E. and Van De Peer, Y. (2002) zt: a software tool for simple and partial Mantel tests. *Journal of Statistical Software*, **7**, 1-12.

141. Swofford, D. (1998). 4 ed. Sinauer Associates, Sunderland, Massachussetts.

142. Varshney, U., Lee, C.P. and RajBhandary, U.L. (1993) From elongator tRNA to initiator tRNA. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 2305-2309.

143. Kolitz, S.E. and Lorsch, J.R. Eukaryotic initiator tRNA: finely tuned and ready for action. *FEBS Lett*, **584**, 396-404.

144. Tumbula, D.L., Becker, H.D., Chang, W.Z. and Soll, D. (2000) Domain-specific recruitment of amide amino acids for protein synthesis. *Nature*, **407**, 106-110.

145. Sankoff, D. and Blanchette, M. (1998) Multiple genome rearrangement and breakpoint phylogeny. *J Comput Biol*, **5**, 555-570.

146. Ishitani, R., Nureki, O., Nameki, N., Okada, N., Nishimura, S. and Yokoyama, S. (2003) Alternative tertiary structure of tRNA for recognition by a posttranscriptional modification enzyme. *Cell*, **113**, 383-394.

147. Ibba, M., Morgan, S., Curnow, A.W., Pridmore, D.R., Vothknecht, U.C., Gardner, W., Lin, W., Woese, C.R. and Soll, D. (1997) A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. *Science (New York, N.Y.)*, **278**, 1119-1122.

148. Curnow, A.W., Hong, K.W., Yuan, R. and Soll, D. (1997) tRNA-dependent amino acid transformations. *Nucleic Acids Symp Ser*, 2-4.

149. Sauerwald, A., Zhu, W., Major, T.A., Roy, H., Palioura, S., Jahn, D., Whitman, W.B., Yates, J.R., 3rd, Ibba, M. and Soll, D. (2005) RNA-dependent cysteine biosynthesis in archaea. *Science (New York, N.Y.)*, **307**, 1969-1972.

150. Hohn, M.J., Park, H.S., O'Donoghue, P., Schnitzbauer, M. and Soll, D. (2006) Emergence of the universal genetic code imprinted in an RNA record. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 18095-18100.

151. Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J. and Glockner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research*, **35**, 7188-7196.

152. Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J.G., Easton, B.C., Eaton, M., Hamady, M., Lindsay, H., Liu, Z. *et al.* (2007) PyCogent: a toolkit for making sense from sequence. *Genome Biol*, **8**, R171.

153. DeSantis, T.Z., Jr., Hugenholtz, P., Keller, K., Brodie, E.L., Larsen, N., Piceno, Y.M., Phan, R. and Andersen, G.L. (2006) NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic acids research*, **34**, W394-399.

154. Bayrac, A.T., Sefah, K., Parekh, P., Bayrac, C., Gulbakan, B., Oktem, H.A. and Tan, W. (2011) In vitro Selection of DNA Aptamers to Glioblastoma Multiforme. *ACS chemical neuroscience*, **2**, 175-181.

155. Hamula, C.L., Zhang, H., Guan, L.L., Li, X.F. and Le, X.C. (2008) Selection of aptamers against live bacterial cells. *Analytical chemistry*, **80**, 7812-7819.

156. Hamula, C.L., Le, X.C. and Li, X.F. (2011) DNA aptamers binding to multiple prevalent M-types of Streptococcus pyogenes. *Analytical chemistry*, **83**, 3640-3647.

157. Tuerk, C., MacDougal, S. and Gold, L. (1992) RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 6988-6992.

158. Bock, L.C., Griffin, L.C., Latham, J.A., Vermaas, E.H. and Toole, J.J. (1992) Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature*, **355**, 564-566.

159. Burgstaller, P., Kochoyan, M. and Famulok, M. (1995) Structural probing and damage selection of citrulline- and arginine-specific RNA aptamers identify base positions required for binding. *Nucleic acids research*, **23**, 4769-4776.

160. Yarus, M. (1998) Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. *Journal of molecular evolution*, **47**, 109-117.

161. Sassanfar, M. and Szostak, J.W. (1993) An RNA motif that binds ATP. *Nature*, **364**, 550-553.

162. Burke, D.H. and Gold, L. (1997) RNA aptamers to the adenosine moiety of S-adenosyl methionine: structural inferences from variations on a theme and the reproducibility of SELEX. *Nucleic acids research*, **25**, 2020-2024.

163. Illangasekare, M., Turk, R., Peterson, G.C., Lladser, M. and Yarus, M. (2010) Chiral histidine selection by D-ribose RNA. *RNA (New York, N.Y.)*, **16**, 2370-2383.

164. Legiewicz, M., Lozupone, C., Knight, R. and Yarus, M. (2005) Size, constant sequences, and optimal selection. *RNA (New York, N.Y.)*, **11**, 1701-1709.

165. Jensen, K.B., Atkinson, B.L., Willis, M.C., Koch, T.H. and Gold, L. (1995) Using in vitro selection to direct the covalent attachment of human immunodeficiency virus type 1 Rev protein to high-affinity RNA ligands. *Proceedings of the National Academy of Sciences*, **92**, 12220-12224.

166. Lorsch, J.R. and Szostak, J.W. (1994) In vitro selection of RNA aptamers specific for cyanocobalamin. *Biochemistry*, **33**, 973-982.

167. Geiger, A., Burgstaller, P., von der Eltz, H., Roeder, A. and Famulok, M. (1996) RNA Aptamers That Bind l-Arginine with Sub-Micromolar Dissociation Constants and High Enantioselectivity. *Nucleic acids research*, **24**, 1029-1036.

168. Jing, M. and Bowser, M.T. (2011) Isolation of DNA aptamers using micro free flow electrophoresis. *Lab on a chip*, **11**, 3703-3709.

169. Berezovski, M., Musheev, M., Drabovich, A. and Krylov, S.N. (2006) Non-SELEX selection of aptamers. *Journal of the American Chemical Society*, **128**, 1410-1411.

170. Lou, X., Qian, J., Xiao, Y., Viel, L., Gerdon, A.E., Lagally, E.T., Atzberger, P., Tarasow, T.M., Heeger, A.J. and Soh, H.T. (2009) Micromagnetic selection of aptamers in microfluidic channels. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 2989-2994.

171. Qian, J., Lou, X., Zhang, Y., Xiao, Y. and Soh, H.T. (2009) Generation of highly specific aptamers via micromagnetic selection. *Analytical chemistry*, **81**, 5490-5495.

172. Zhang, W.Y., Zhang, W., Liu, Z., Li, C., Zhu, Z. and Yang, C.J. (2012) Highly parallel single-molecule amplification approach based on agarose droplet polymerase chain reaction for efficient and cost-effective aptamer selection. *Analytical chemistry*, **84**, 350-355.

173. Ciesiolka, J., Illangasekare, M., Majerfeld, I., Nickles, T., Welch, M., Yarus, M. and Zinnen, S. (1996) Affinity selection-amplification from randomized ribooligonucleotide pools. *Methods in enzymology*, **267**, 315-335.

174. Lozupone, C., Changayil, S., Majerfeld, I. and Yarus, M. (2003) Selection of the simplest RNA that binds isoleucine. *RNA (New York, N.Y.)*, **9**, 1315-1322.

175. Bala, J., Bhaskar, A., Varshney, A., Singh, A.K., Dey, S. and Yadava, P. (2011) In vitro selected RNA aptamer recognizing glutathione induces ROS mediated apoptosis in the human breast cancer cell line MCF 7. *RNA biology*, **8**, 101-111.

176. Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J. and Knight, R. (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature methods*, **5**, 235-237.

177. Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N. and Knight, R. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*, **108 Suppl 1**, 4516-4522.

178. Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M. *et al.* (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME journal*, **6**, 1621-1624.

179. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, **7**, 335-336.

180. Gotelli, N.J. and Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379-391.

181. Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., Arrieta, J.M. and Herndl, G.J. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 12115-12120.