EVOLUTION & DETECTION OF NON-CODING RNA, AND TRANSCRIPTOME ANALYSES OF TWO NON-MODEL SYSTEMS

by

ANUJ SRIVASTAVA

(Under the Direction of Dr. Russell L. Malmberg)

ABSTRACT

Over the last decade, with the advent of high-throughput technologies, a massive flood of biological data has occurred and is continuing to occur. These technologies generate diverse biological datasets including whole-genome sequences, transcriptome sequencing (RNA-Seq, EST sequence), epigenetics (ChIP-chip, ChIP-Seq), and other -omics. These datasets offer unprecedented opportunities to increase our understanding of the functions and dynamics of the genome and the cell. This dissertation entitled **"Evolution & Detection of ncRNA and Transcriptome Analyses of Two Non-Model Systems"** combines an evolutionary approach to study non-coding RNAs (ncRNA), and their identification in genomic data using patterns of chromatin modifications, and the analysis of transcriptomes of non-model species chosen for their evolutionary and ecological interest.

The evolutionary study of non-coding RNA involves analyzing the patterns of mutations which causes the variability's in the secondary structure of RNA. From the analysis, I found that secondary structures evolve both by whole stem insertion/deletion, and by mutations that create or disrupt stem base pairing. I analyzed the evolution of stem lengths and constructed substitution matrices describing the changes responsible for the variation in the RNA stem length. I believe that data

generated from the study will provide new insights into the evolution of RNA secondary structures and will facilitate design of improved mutational models for RNA structure evolution. I also developed a novel machine learning based approach, based upon using patterns of chromatin-modification to discriminate/detect different genomic features such as protein coding gene, RNA gene, pseudogene and transposon element gene. I implemented this approach on the model plant species *Arabidopsis* and detected 33 novel genes. I believe this approach will help in improving the annotation of newly sequenced species.

From the transcriptome analysis of two non-model systems (Pitcher plants and Songbird), I was able to identify the polymorphic loci which are fixed and shared between sub-species. I also performed functional annotation of all the genes and identified the fast evolving genes by substitution rate determination. I believe that genomic resources developed during these studies will contribute greatly to future research on these genera and their distinctive ecological adaptations.

INDEX WORDS: tmRNA, RNAseP, telomerase RNA, Evolution, histone-modifications, SVM, *Sarracenia*, Duplication, Song bird, Polymorphic loci

EVOLUTION & DETECTION OF NON-CODING RNA, AND TRANSCRIPTOME ANALYSES OF TWO NON-MODEL SYSTEMS

by

ANUJ SRIVASTAVA

B. Pharma, BBDNITM, India, 2004

M. Tech, SASTRA (Deemed University), India, 2006

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2011

© 2011

ANUJ SRIVASTAVA

All Rights Reserved

EVOLUTION & DETECTION OF NON-CODING RNA, AND TRANSCRIPTOME ANALYSES OF TWO NON-MODEL SYSTEMS

by

ANUJ SRIVASTAVA

Major Professor:	Russell L. Malmberg
Committee:	Liming Cai
	Jan Mrázek
	Ying Xu
	C. J. Tsai

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2011

DEDICATION

I dedicate this dissertation to my elder brothers, Anurag and Abhinav for their love, encouragement and teaching me to think big.

ACKNOWLEDGEMENTS

I think this is going to be the hardest part for me to write in this thesis, as there are so many people I want to thank and I have a limitation of few pages to do that. I start with thanking Dr. Liming Cai who was the graduate coordinator of Institute of Bioinformatics when I admitted to this program and later became part of my committee. I am very grateful to him and IOB committee for admitting me in the program and awarding me the graduate school assistantship, which changed my life forever.

When I entered this program, I had a doubt about myself whether I was going to be succeed in this or not and was skeptical about myself when I started my first lab rotation. I was very fortunate to start my Ph.D. work with Dr. Jan Mrázek who also became part of my committee. While working with him, I developed a lot of self-confidence and was able to publish a co-author paper in the very first semester, which set up the tone for the rest of my Ph.D. I want to thank him for his guidance and support in last four years. I also like to thank other members of committee Dr. Ying Xu and Dr. C.J. Tsai for their support. The other advisor who is not in my committee but has a significant impact on my work is Dr. Travis Glenn. I believe his 1 credit course on genome technologies, became the most valuable course for me in my Ph.D. We authored and will be authoring a number of genomics papers, and I must say that I learned a lot from him and always be grateful to him for helping me in building my career in genomics.

If one is wandering, why I did not talk about my major advisor yet, because I am out of words to describe the level of guidance and support he provided in all these years. The 25 minutes meeting with Dr. Malmberg on every Thursday is the most exciting time for me in the week. We discussed and implemented the number of different research ideas, some of them work and some did not but I never felt any pressure while working with him. Whenever we face some tough research problems, he always says 'We' will solve it or work on it together, and that 'We' makes a great difference as I always believe he is always there to help me. He provided best bioinformatics resources, books, and encouraged me to explore the other research areas away from his expertise. I sincerely thank you 'Sir' for all your support and I will miss working in this lab/group environment that you and Dr. Cai created.

I also want to use this opportunity to thank my family for all their love and encouragement. For my parents, Umesh and Kiran who raised me, continuously teaching me the importance of education and supporting me in all my pursuits. I am extremely fortunate to have two elder brothers who by themselves are remarkably successful researchers; they helped in every step of my life and always believed that I am competent to accomplishing bigger things. My sister-in-laws Shivani and Neha for their love and affection and my niece Aarushi for bringing the joy in my life.

Lastly, my time at UGA was made enjoyable in large part due to the many friends and groups that became a part of my life. I want to use this opportunity to thank Sameer, Umakanta, Wen-Chi, Tim, Will and other members of RNA-Informatics group, with whom I interacted at both fronts, personal/professional. We discussed a number of research ideas, participated in research

contest, authored papers together and played a lot of tennis and did other fun stuff; So-thank you guys and I will miss all of you.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

CHAPTER

1 INTRODUCTION AND LITERATURE REVIEW 1

1.1 SECTION I: NON-CODING RNA 1

1.2 SECTION II: TRANSCRIPTOME ANALYSIS 18

2 MUTATIONAL PATTERNS IN RNA SECONDARY STRUCTURE EVOLUTION EXAMINED IN THREE RNA FAMILIES 29

2.1 ABSTRACT 30

2.2 INTRODUCTION 31

2.3 MATERIAL AND METHODS 33

2.4 RESULTS 37

2.5 DISCUSSION 42

2.6 ACKNOWLEDGEMENTS 45

2.7 REFERENCES 45

2.8 SUPPLEMENTARY MATERIAL 60

3 PATTERNS OF CHROMATIN MODIFICATIONS DISCRIMINAT DIFFERENT GENOMIC FEATURES IN *ARABIDOPSIS* 61

3.1 ABSTRACT 62

3.2 INTRODUCTION 63

3.3 RESULTS 65
3.4 DISCUSSION 68
3.5 MATERIAL AND METHODS 72
3.6 ACKNOWLEDGEMENTS 77
3.7 REFERENCES 78
4 TRANSCRIPTOME ANALYSIS OF *SARRACENIA*, AN INSECTIVOROUS PLANT 86
4.1 ABSTRACT 87
4.2 INTRODUCTION 88
4.3 MATERIAL AND METHODS 89
4.4 RESULTS 93
4.5 DISCUSSION 97
4.6 AVAILAIBILITY 101
4.7 ACKNOWLEDGEMENTS 101
4.8 REFERENCES 101
4.9 SUPPLEMENTARY MATERIAL 116
5 PROBING THE GENOMICS OF ADAPTIVE DIVERGENCE USING COMPARATIVE TRANSCRIPTOMICS BETWEEN TWO SUB-SPECIES OF SONG BIRD 117
5.1 ABSTRACT 118
5.2 INTRODUCTION 119
5.3 MATERIAL AND METHODS 121
5.4 RESULTS 124

5.5 DISCUSSION 129

5.6 ACKNOWLEDGEMENTS 133

5.7 REFERENCES 134

5.8 SUPPLEMENTARY MATERIAL 150

6 CONCLUSION AND FUTURE WORK 154

6.1 NON-CODING RNA 154

6.2 *SARRACENIA* AND SONGBIRD 156

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1. 1 SECTION I: NON-CODING RNA

1.1.1 Brief history and importance of ncRNA:

The importance and centrality of RNA in molecular biology has become established in a series of paradigm-shifting discoveries over the last 60-70 years. Prior to the 1940s, both the genetic information and the enzymatic-catalytic functions of cells were often thought to be functions of proteins (Judson 1979). In the 1950s, it became clear that proteins were synthesized in the cytoplasm in the presence of abundant RNA (Brachet and Chantrenne 1956), and RNA served as an intermediary molecule during the transfer of genetic information to proteins (Central Dogma of molecular biology) (Crick 1958).

After the establishment of the Central Dogma, it was believed for several decades that RNA molecules come in 3 flavors (rRNA, tRNA and mRNA) and all were involved in synthesizing proteins. Later, several other types of RNAs including uridine (U)-rich U RNAs were discovered and then, it was realized that RNA comes in more than 3 flavors (Zieve 1981; Busch, Reddy et al. 1982). In the same time frame when other families of RNAs were discovered, Cech and Altman from their study of RNA splicing in the ciliated protozoan *Tetrahymena thermophila* and the bacterial RNase P complex, respectively concluded that RNA had catalytic functions as

ribozymes (Kruger, Grabowski et al. 1982; Guerrier-Takada, Gardiner et al. 1983). In the late 1990s, several families of ncRNA were discovered which were involved in regulatory functions (Voinnet and Baulcombe 1997; Fire, Xu et al. 1998; Ratcliff, MacFarlane et al. 1999) and thus the three main roles that RNA molecule can take include i) being the genetic material ii) catalyzes reaction as ribozymes iii) regulates other macromolecular processes. The biomolecules that function as RNAs, not through coding for proteins, are termed non-coding RNAs (ncRNA) and the DNA sequence from which a non-coding RNA is transcribed as the end product is often called an RNA gene or non-coding RNA gene.

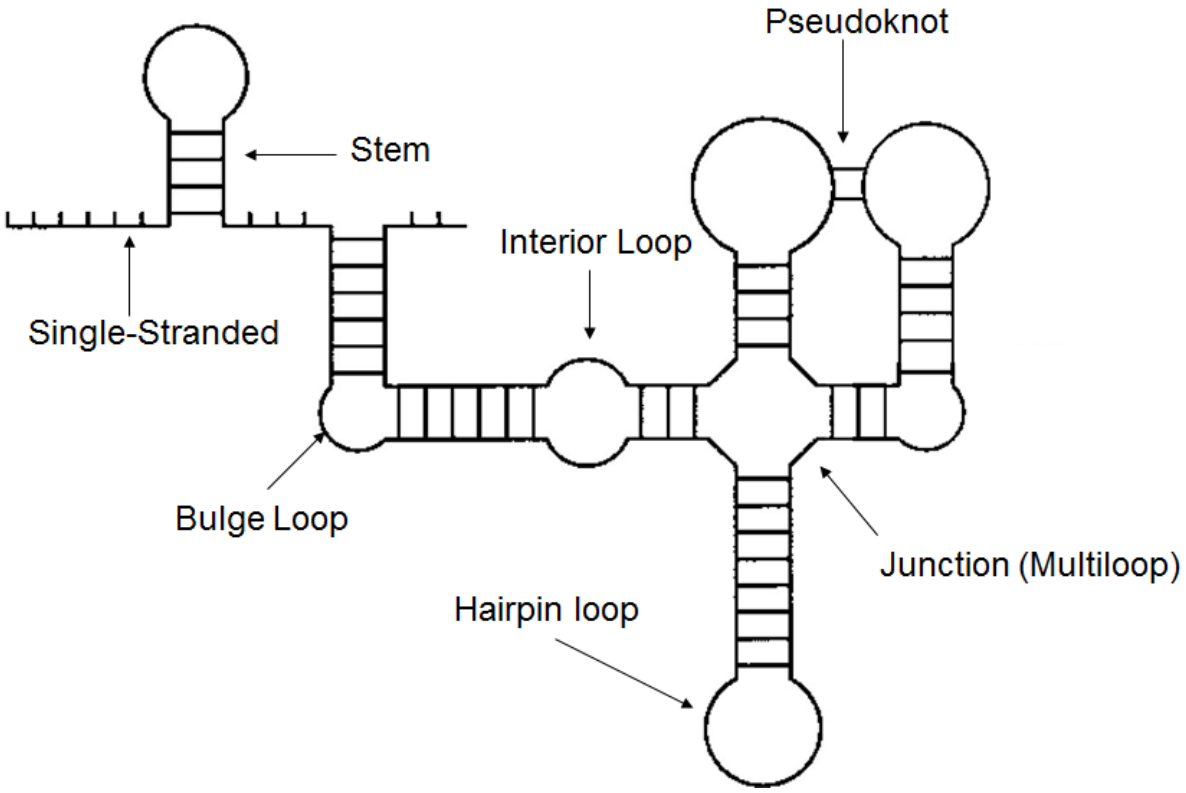
New RNA families continue to emerge and have been involved in diverse functions including having a role in chromatin structure (Kelley and Kuroda 2000). An example is the human XIST (X (inactive)-specific transcript) RNA (17-kb ncRNA) with a key role in dosage compensation and X-chromosome inactivation (Avner and Heard 2001). Due to the continuous discovery of new RNA families, the estimated number of ncRNA genes in eukaryotic genomes has fluctuated over the last few years. A few years back, several studies based primarily on microarray experiments (Birney, Stamatoyannopoulos et al. 2007; Kapranov, Willingham et al. 2007; Petherick 2008) suggested a phenomenon of "pervasive transcription", according to which 70-90 percent of eukaryotic genomes may be transcribed but only 2% is translated. However, Bakel et al. (van Bakel, Nislow et al. 2010) argued against pervasive transcription and indicated that transcripts detected in previous studies might arise due to technical artifacts and/or background biological noise. These contrasting results have attracted considerable attention (Phillips 2010; Robertson 2010; Robinson 2010) and have become a matter of debate.

One recently published study compared the result of Van Bakel et al. (Van Bakel, Nislow et al. 2010) to the others by obtaining the data from Van Bakel group and concluded that their study suffers from methodological flaw and there is strong support for pervasive transcription in mammalian genomes (Clark, Amaral et al. 2011). Van Bakel et al. responded to the criticism in the recently published article (Van Bakel, Nislow et al. 2011) and disagreed with the interpretation of their work by Clark et al. (2011) The argument about the number of different transcripts in the genome is thus still a matter of debate and it can range anywhere between 10,000 to million in number but there is no argument about the importance of ncRNA and the need for greater exploration of these molecules.

1.1.2 RNA Secondary Structure:

RNA structure is crucial for the function of the RNA. The primary structure of RNA is determined by the sequence of A, C, G, and U bases in a strand. An RNA molecule folds back to itself to form a secondary structure by forming hydrogen-bonds between complementary bases (G:C, A:U known as canonical base pairs, and G:U non-canonical base pairs). Paired and unpaired elements of RNA secondary structure are known as stem and loop, respectively. Another important element of RNA secondary (sometimes included in tertiary) structure known as pseudoknot is formed by the base pairing between the bases in the loop and a single stranded region outside the loop. The different elements of RNA secondary structure are shown in figure1.1 (Wuchty, Fontana et al. 1999).

Figure 1.1. RNA secondary structure elements



1.1.3 Bioinformatics challenges with ncRNA:

The main bioinformatics challenges associated with ncRNA are the following:

- Predict the structure of an RNA from sequence
- Align RNAs on the basis of their structure
- Find new instances of known ncRNAs or find previously unknown ncRNAs

1.1.3.1 Structure prediction:

Given an RNA sequence, it can fold into several possible secondary structures; the number of RNA structures increases exponentially with an increase in sequence length (Meyer 2007). Therefore, the challenge is to find the correct RNA structure which represents the functional RNA molecule. Popular structure prediction programs such Mfold (Zuker 2003) and RNAfold (Hofacker and Stadler 2006) are based upon the minimal free energy (MFE) method which predicts the structure with minimum free energy assuming this represents the correct structure (Kavanaugh and Dietrich 2009). These programs (Mfold and RNAfold) (Zuker 2003; Hofacker and Stadler 2006) incorporate stacking energy from neighboring base-pairs into the prediction and estimate the total free energy for each possible structure, then choosing the one with the lowest free energy. Unfortunately, the real structure is not necessarily the one with lowest free energy and structures with suboptimal free energies may be the correct structure (Higgs 2000); there may be many structures with calculated free energies which are quite close to each other.

1.1.3.2 RNA structural alignment:

The second important challenge involves finding matching RNA structures from a set of related sequences and aligning the structures to each other. Some of the important programs in this category are QRNA (Rivas and Eddy 2001), EVOFOLD (Pedersen, Bejerano et al. 2006) and RNAZ (Washietl, Hofacker et al. 2005). QRNA (Rivas and Eddy 2001) and EVOFOLD (Pedersen, Bejerano et al. 2006) are based on the stochastic grammar approach (SCFG- a probabilistic model, which can capture the long-range constraints imposed by the base-pairing sequence positions of the RNA structure) and can handle two and multiple sequences as an input, respectively. RNAz (Washietl, Hofacker et al. 2005) is based on minimal free energy (MFE)

method as well as structure conservation and can handle multiple aligned (Meyer 2007) sequences. All of these programs have a limitation that they can align only pseudoknot free structures. The recently developed tool RNASampler (Stormo, Xu et al. 2007) can predict the consensus structure with pseudoknot from multiple unaligned sequences; it is based on an algorithm which finds the common structure between two sequences by probabilistically sampling of aligned stems (Stormo, Xu et al. 2007). There are many other tools available (Bernhart, Hofacker et al. 2008; Sato, Hamada et al. 2009) for the consensus structure prediction but one of their major limitations is that they can handle only several hundred sequences of few thousand base pairs only and are not suitable for the structure prediction of very long ncRNA (greater than 10 Kb).

1.1.3.3 ncRNA gene finding:

One important challenge in bioinformatics is to develop computational methods, which can locate ncRNA within genomes. Computational ncRNA gene finding is a relatively new and more challenging field than protein coding gene finding.

The information used in the gene prediction can be divided into three major classes (Schattner 2002):

- Signals
- Content statistics
- Similarity to known genes

Sequence signals for protein coding genes include promotors, terminators, transcription factor binding sites, poly-A-addition, start and stop codon, splice sites, CpG islands (Schattner 2002) , but when it comes to ncRNA, signals are limited to some weakly conserved promoters and

terminators in some ncRNA gene or existence of splice sites in some multi-exon ncRNA (Schattner 2002).

Non-random variation in base sequences, i.e. content statistics, also provides useful signal for protein coding genes. In prokaryotes, the length of the open reading frame (ORF) alone serves as a statistical significant marker (Schattner 2002). In addition, codon based statistics (Species-specific codon usage, adjacency of amino acids) provide various signals for the detection of protein coding genes which are not available for ncRNA (Schattner 2002). Finally, due to the large number of available protein sequences in databases, it becomes increasingly likely that a new protein–coding gene will have at least some homology with an already known protein. There are far fewer ncRNA sequences available in public databases and their sequences can only be compared at the nucleotide level (not as translated amino acids) (Schattner 2002). Moreover, the sequence conservation in RNA molecule is often at the structural level i.e. instead of individual base sequence; base-pairings are conserved (Schattner 2002) (with the exception of some conserved RNA families such as rRNA), due to which conventional sequence similarity search methods (such as BLAST) are not generally useful for ncRNA gene finding (Schattner 2002).

ncRNA gene finding algorithms:

Due to the differences, between protein and ncRNA gene finding, specialized ncRNA gene finders have been developed. There are two main kinds of computational approaches to detect RNA genes:

- comparative analysis
- *ab initio*

Comparative analysis gene finding:

RNA structures are usually more conserved than their sequences. The comparative analysis approach relies on multiple alignments of ncRNA sequences with annotated structures. These methods build up a structural profile based on the alignment of annotated sequences. Afterwards, a sequence-structure alignment is used to search the genomes against the profile. Sequence segments with high alignment score indicate a structural homology. One such a model to profile RNA secondary structure is known as covariance model (CM) introduced by Eddy and Durbin (Eddy and Durbin 1994). CMs can be used for multiple structural alignments, such as those in Rfam (Griffiths-Jones, Bateman et al. 2003). A limitation of CM was that it cannot model RNA structures with pseudoknots (formed by pairing between bases in a loop region and complementary bases outside the loop). The pseudoknot structure prediction problem is NP-hard (Cai, Malmberg et al. 2003; Matsui, Sato et al. 2005) and the new pseudoknot profiling models (Holmes 2004; Menzel, Gorodkin et al. 2009; Mosig, Zhu et al. 2009), which are primarily the extension of CM, have high time and space complexities. The current version of Infernal 1.0 (Eddy, Nawrocki et al. 2009) (pseudoknot free search) and recently developed RNATOPS (Cai, Huang et al. 2008) (search with pseudoknot) address the issue of speed and are found suitable for genome wide scan of ncRNA, given a suitable structural profile.

Another problem with the profile based method is that it can find ncRNA with conserved structures only. However, RNA secondary and tertiary structures are both constant and variable

across evolution (Holmes 2004; Menzel, Gorodkin et al. 2009; Mosig, Zhu et al. 2009) (see evolution discussion below). In a given RNA gene family, some members will have stem-loops and pseudoknots that are not present in other members of the family. Additionally, some stems and loops will have substantial sequence length variation. These structural variations cause a problem in sequence structure alignment of distant homologues (Menzel, Gorodkin et al. 2009).

To address these issues, models of RNA structural evolution having the ability to deal with certain degrees of structural rearrangement between homologs have been proposed (Holmes 2004; Bradley and Holmes 2009). One method which addresses, the issue of variability is based upon modeling secondary structure conformations as graphs with small tree width and formulating sequence-structure alignment for homolog detection as graph homomorphism. It incorporates the technique of NULL stem to address the issue of optional stems that may be deleted from the structure profile or may be a misalignment. This method was incorporated into the program RNAv (Huang, Malmberg et al. 2010) and searches the genomes for RNA structural variations. In spite of these models and algorithms, structural variation is a challenging problem for RNA search programs due to the lack of methods which can incorporate these variations into the search for homologous ncRNA genes (Gruber, Findeiss et al. 2010).

***Ab initio* gene finding:**

Ab-initio gene prediction constitutes the most challenging case of RNA gene prediction. In the most difficult case, we are given a single genome sequence without any annotation and have to find the encoded RNA genes. The main method in this category is based on composition based statistics; this method works only for some specialized genomes and particular types of ncRNA.

It is based on the observation of existence of a strong correlation between the GC contents of rRNA and tRNA genes and the optimal growth temperature (Galtier and Lobry 1997). Identification of the GC rich regions in organisms (*M. jannschii and P. furiosus*) led to the detection of new and known ncRNA genes (Klein, Misulovin et al. 2002).

1.1.4 Evolution of RNA Structure:

The folds of structural RNAs are highly conserved among all kingdoms of life and have been widely used to determine phylogenic relationships between different species (Kumar and Rzhetsky 1996). The structural folds of RNA are maintained over evolution via compensatory mutations which preserve the base-pairing, in the helical region of related species. Phylogenetic comparative analysis uses the patterns of compensatory mutations to predict the structure of non-coding RNAs. Although, the double helical regions of RNA are conserved within given RNA families, there is also variation of structure elements such as stem-loops, pseudoknots across evolution. The variation includes both presence/absence of entire structural elements and change in length of the helical regions due to mutations. Holmes (Holmes 2004) proposed a model of RNA structure evolution based upon an extension of the TKF91 sequence evolution model (Thorne, Kishino et al. 1991). The TKF91 **(**time reversible model**)** describes the evolution of single sequence under the action of two kinds of mutation i) Point substitution events (act on a single residue only) and (ii) single-residue indel events, (insert or delete a single residue). The rates of both types of the event are independent of the neighboring sequence in this model. The Holmes (Holmes 2004) extension incorporates insertions and deletions of bases, base pairs, and whole stems. Another recently published evolutionary model has the important application of the inference of ancestral sequences for a set of diverged RNAs. This model is known as

evolutionary triplet model and it is based on transducer composition algorithm (Bradley and Holmes 2009). This model showed that a probabilistic grammar, which was used to model the pair of homologous RNA, could be extended to an entire phylogeny by treating the pairwise grammar as a machine (a "transducer"), which models a single ancestor-descendant relationship in the tree, transforming one RNA structure into another. In addition, there are some other studies which analyzed the patterns of compensatory mutations in RNA evolution (Dixon and Hillis 1993) and estimated rate of evolution of different RNA structural elements (Knight, Smit et al. 2007). However, there have been a lack of studies providing estimation of relative frequencies of various events which cause variability in secondary structure. The widely use substitution matrix RIBOSUM (Klein and Sr 2003) which describes the pattern of mutations, is based on rRNA family which are known to have very well conserved sequence and structure, and therefore might not be suitable finding the diverged ncRNA.

1.1.5 My objectives:

This dissertation involves two approaches to deal with ncRNA detection problem in the genomes:

The first approach (chapter 2) analyzes the patterns of RNA secondary structure evolution in the different RNA families and determines the associated mutational patterns and frequencies of major types of variation, which can lead to changes in the RNA secondary structures in closely related species. The essence of this approach is based upon treating elements of RNA secondary structures such as stem-loops and pseudoknots, as evolutionary characters and mapping of these characters in a standard way on to a phylogenetic tree, constructed primarily from rRNA sequences. From the analysis, I concluded that RNA secondary structure evolves by both whole

stem insertion/deletion and by mutations that create or disrupt base pairing. I determined the relative frequencies of these two events and recorded the associated mutational patterns in log-odds matrices. I believe that, this work will facilitate the design of improved mutational models for RNA structure evolution and will provide more sensitivity to ncRNA search programs. Moreover, the study of RNA structure evolution is interesting biology in its own right, separate from its utility in helping design better ncRNA search engines.

The second approach (chapter 3) is a novel method for the discrimination/identification of ncRNA and other genomic features inside the genomes based upon the patterns of chromatin modification and DNA methylation obtained through chromatin immunoprecipation. This is a machine learning based method and has the potential to distinguish/detect different genomic features, thus improving genome annotation. I showed the implementation of this approach on the model plant species *Arabidopsis thaliana*.

1.1.6 References:

Avner, P. and E. Heard (2001). "X-chromosome inactivation: counting, choice and initiation." Nat Rev Genet **2**(1): 59-67.

Bernhart, S. H., I. L. Hofacker, et al. (2008). "RNAalifold: improved consensus structure prediction for RNA alignments." BMC Bioinformatics **9**: 474.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Brachet, J. and H. Chantrenne (1956). "The function of the nucleus in the synthesis of cytoplasmic proteins." Cold Spring Harb Symp Quant Biol **21**: 329-337.

Bradley, R. K. and I. Holmes (2009). "Evolutionary triplet models of structured RNA." PLoS Comput Biol **5**(8): e1000483.

Bradley, R. K. and I. Holmes (2009). "Evolutionary Triplet Models of Structured RNA." Plos Computational Biology **5**(8).

Busch, H., R. Reddy, et al. (1982). "SnRNAs, SnRNPs, and RNA processing." Annu Rev Biochem **51**: 617-654.

Cai, L., R. L. Malmberg, et al. (2003). "Stochastic modeling of RNA pseudoknotted structures: a grammatical approach." Bioinformatics **19 Suppl 1**: i66-73.

Cai, L. M., Z. B. Huang, et al. (2008). "Fast and accurate search for non-coding RNA pseudoknot structures in genomes." Bioinformatics **24**(20): 2281-2287.

Clark, M. B., P. P. Amaral, et al. (2011). "The reality of pervasive transcription." Plos Biology **9**(7): e1000625.

Crick, F. H. (1958). "On protein synthesis." Symp Soc Exp Biol **12**: 138-163.

Dixon, M. T. and D. M. Hillis (1993). "Ribosomal-Rna Secondary Structure - Compensatory Mutations and Implications for Phylogenetic Analysis." Molecular Biology and Evolution **10**(1): 256-267.

Eddy, S. R. and R. Durbin (1994). "RNA sequence analysis using covariance models." Nucleic Acids Res **22**(11): 2079-2088.

Eddy, S. R., E. P. Nawrocki, et al. (2009). "Infernal 1.0: inference of RNA alignments." Bioinformatics **25**(10): 1335-1337.

Fire, A., S. Xu, et al. (1998). "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans." Nature **391**(6669): 806-811.

Galtier, N. and J. R. Lobry (1997). "Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes." J Mol Evol **44**(6): 632-636.

Griffiths-Jones, S., A. Bateman, et al. (2003). "Rfam: an RNA family database." Nucleic Acids Res **31**(1): 439-441.

Gruber, A. R., S. Findeiss, et al. (2010). "Rnaz 2.0: Improved Noncoding Rna Detection." Pac Symp Biocomput **15**: 69-79.

Guerrier-Takada, C., K. Gardiner, et al. (1983). "The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme." Cell **35**(3 Pt 2): 849-857.

Higgs, P. G. (2000). "RNA secondary structure: physical and computational aspects." Q Rev Biophys **33**(3): 199-253.

Hofacker, I. L. and P. F. Stadler (2006). "Memory efficient folding algorithms for circular RNA secondary structures." Bioinformatics **22**(10): 1172-1176.

Holmes, I. (2004). "A probabilistic model for the evolution of RNA structure." BMC Bioinformatics **5**.

Holmes, I. (2004). "A probabilistic model for the evolution of RNA structure." BMC Bioinformatics **5**: 166.

Huang, Z., R. Malmberg, et al. (2010). "RNAv: Non-coding RNA Secondary Structure Variation Search via Graph Homomorphism." Proceedings of Computational Systems Bioinformatics Conference **9**: 56-69.

Judson, H. F. (1979). "The Eighth Day of Creation: Makers of the Revolution in Biology. ." Simon & Schuster.

Kapranov, P., A. T. Willingham, et al. (2007). "Genome-wide transcription and the implications for genomic organization." Nat Rev Genet **8**(6): 413-423.

Kavanaugh, L. A. and F. S. Dietrich (2009). "Non-coding RNA prediction and verification in Saccharomyces cerevisiae." PLoS Genet **5**(1): e1000321.

Kelley, R. L. and M. I. Kuroda (2000). "Noncoding RNA genes in dosage compensation and imprinting." Cell **103**(1): 9-12.

Klein, R. J., Z. Misulovin, et al. (2002). "Noncoding RNA genes identified in AT-rich hyperthermophiles." Proc Natl Acad Sci U S A **99**(11): 7542-7547.

Klein, R. J. and E. Sr (2003). "RSEARCH: Finding homologs of single structured RNA sequences." BMC Bioinformatics **4**.

Knight, R., S. Smit, et al. (2007). "Evolutionary rates vary among rRNA structural elements." Nucleic Acids Research **35**(10): 3339-3354.

Kruger, K., P. J. Grabowski, et al. (1982). "Self-Splicing Rna - Auto-Excision and Auto-Cyclization of the Ribosomal-Rna Intervening Sequence of Tetrahymena." Cell **31**(1): 147-157.

Kumar, S. and A. Rzhetsky (1996). "Evolutionary relationships of eukaryotic kingdoms " J. Mol. Evol. **42**: 183-193.

Matsui, H., K. Sato, et al. (2005). "Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures." Bioinformatics **21**(11): 2611-2617.

Menzel, P., J. Gorodkin, et al. (2009). "The tedious task of finding homologous noncoding RNA genes." RNA **15**(12): 2075-2082.

Meyer, I. M. (2007). "A practical guide to the art of RNA gene prediction." Brief Bioinform **8**(6): 396-414.

Mosig, A., L. Zhu, et al. (2009). "Customized strategies for discovering distant ncRNA homologs." Brief Funct Genomic Proteomic **8**(6): 451-460.

Pedersen, J. S., G. Bejerano, et al. (2006). "Identification and classification of conserved RNA secondary structures in the human genome." PLoS Comput Biol **2**(4): e33.

Petherick, A. (2008). "Genetics: The production line." Nature **454**(7208): 1042-1045.

Phillips, M. L. (2010). "Existence of RNA 'dark matter' in doubt. The abundance of transcripts from the genome may have been overestimated." Nature: 10.1038/news.2010.1248.

Ratcliff, F. G., S. A. MacFarlane, et al. (1999). "Gene silencing without DNA: RNA-mediated cross-protection between viruses." Plant Cell **11**(7): 1207-1215.

Rivas, E. and S. R. Eddy (2001). "Noncoding RNA gene detection using comparative sequence analysis." BMC Bioinformatics **2**: 8.

Robertson, M. (2010). "The evolution of gene regulation, the RNA universe, and the vexed questions of artefact and noise." BMC Biol **8**: 97.

Robinson, R. (2010). "Dark matter transcripts: sound and fury, signifying nothing?" Plos Biology **8**(5): e1000370.

Sato, K., M. Hamada, et al. (2009). "CENTROIDFOLD: a web server for RNA secondary structure prediction." Nucleic Acids Res **37**(Web Server issue): W277-280.

Schattner, P. (2002). "Computational Gene-finding for Non-coding RNAs." Non-coding RNAs, J. Barciszewski and V. Erdmann (eds.).

Stormo, G. D., X. Xu, et al. (2007). "RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment." Bioinformatics **23**(15): 1883-1891.

Thorne, J. L., H. Kishino, et al. (1991). "An Evolutionary Model for Maximum-Likelihood Alignment of DNA-Sequences." Journal of Molecular Evolution **33**(2): 114-124.

van Bakel, H., C. Nislow, et al. (2010). "Most "Dark Matter" Transcripts Are Associated With Known Genes." Plos Biology **8**(5): -.

van Bakel, H., C. Nislow, et al. (2011). "Response to "the reality of pervasive transcription"." Plos Biology **9**(7): e1001102.

Voinnet, O. and D. C. Baulcombe (1997). "Systemic signalling in gene silencing." Nature **389**(6651): 553.

Washietl, S., I. L. Hofacker, et al. (2005). "Fast and reliable prediction of noncoding RNAs." Proc Natl Acad Sci U S A **102**(7): 2454-2459.

Wuchty, S., W. Fontana, et al. (1999). "Complete suboptimal folding of RNA and the stability of secondary structures." Biopolymers **49**(2): 145-165.

Zieve, G. W. (1981). "Two groups of small stable RNAs." Cell **25**(2): 296-297.

Zuker, M. (2003). "Mfold web server for nucleic acid folding and hybridization prediction." Nucleic Acids Res **31**(13): 3406-3415.

1.2 SECTION II: TRANSCRIPTOME ANALYSIS

Organisms respond to changing environmental conditions by adjusting their programs of growth and development, or by moving. Many genes that respond to changing conditions are controlled at the transcriptional level, and therefore an analysis of transcriptomes will provide a snapshot of the genes which are expressed under specific conditions. Analysis of transcriptomes has been a key area of biological investigation for decades. Research in the field has progressed from northern blotting (candidate gene-based detection of RNA) to early EST sequencing to the sequencing of entire transcriptome (RNA-Seq) (Morozova, Hirst et al. 2009). I am going to give a brief history of methods used in genomic/transcriptome analysis and then provide the objectives of my transcriptomic studies on two non-model species.

1.2.1 Candidate gene approaches:

These are the some of the earliest methods for the study of cellular transcriptomes. Northern bloat analysis is a low-throughput technique (Alwine, Kemp et al. 1977) which requires the use of radioactivity, or a chemical marker, and a large amount of input RNA. The experimental throughput has been increased, and the requirement for the quantity of RNA reduced, with the development of reverse transcription quantitative PCR (RT-qPCR) method (Becker-Andre and Hahlbrock 1989; Noonan, Beck et al. 1990). However, the throughput of this method does not approach the transcriptome-wide scale but is limited to the order of hundreds of known transcripts at a time (VanGuilder, Vrana et al. 2008). RT-qPCR is quantitative and is sometimes used as a second-level analysis after a higher throughput qualitative survey method.

1.2.2 Microarray technology:

Single gene based approaches for transcriptome analyses were, at least partially, replaced by development of microarray technology. This allowed the characterization of the expression levels of thousands of known or putative transcripts (Schena, Shalon et al. 1995) and was the first high-throughput based approach for the characterization of transcriptome. Numerous initiatives (Pollack, Perou et al. 1999; Tsukamoto, Uchida et al. 2008; Etemadmoghadam, deFazio et al. 2009; Mantripragada, Diaz de Stahl et al. 2009; Hu, Clifford et al. 2010) to characterize the expression signatures of cell types and disease states started after the development of microarray. The spotting microarray technology offered several advantages over its predecessors, but it still had the key limitation of detecting the known transcripts only (Pozhitkov, Tautz et al. 2007), which lead to development of tiling microarray technology (Mockler, Chan et al. 2005; Liu 2007). Tiling array functions on a similar principle to spotting microarrays but differ in the nature of the probe design. In tiling microarrays, short fragments are designed to cover the entire genome or contiguous segments of genome, so previously unknown transcripts can be discovered. However, tiling microarrays also do not provide base specific resolution, suffer from high background noise and still needs the reference genome for probe design (Snyder, Wang et al. 2009).

1.2.3 Sequencing-based approaches for transcriptome study:

Transcriptome analysis by DNA sequencing provides an alternative to microarray based methods. A key advantage of this method over microarrays is the ability to directly identify transcripts by sequence instead of measuring them by hybridization intensities. From sequencing of the individual cDNA clones (Stone, Rothblum et al. 1985) to constructing the cDNA libraries, representing portions of species transcriptome (Seki, Narusaka et al. 2002), sequencing based

studies have evolved considerably. Though Sanger sequencing method offers advantages over microarray, it was still not suitable for routine full-length cDNA (FLcDNA) sequencing, primarily due to the sequence cost and complexity involve with cloning. An alternative to FLcDNA sequencing is the EST (a short sub sequence of cDNA sequence, usually from the 3' end to take advantage of the polyA tail) sequencing (Boguski 1995), which provides the snapshot of expressed genes or transcripts under defined conditions in particular tissues in a cost effective manner. EST sequencing with Sanger method is cheaper than FLcDNA but is still too expensive to carry out routinely at transcriptome level. The other problem with the EST sequencing is to resolve the gene representation within the cDNA library. The presence of EST is a reliable signature that gene was present and expressed in the genome, but this outcome could not be interpreted other way round (Rudd 2003). The problem of gene representation usually occurs due to the limited depth of the sequencing, even after the normalization cDNA of libraries, the EST collection often represents highly expressed genes only (Rudd 2003). Moreover, due to the low redundancy of sequencing reads, EST's are usually not suitable for estimating transcript abundance. Recent estimates of available number of in the NCBI dbEST (Boguski, Lowe et al. 1993) database can be found at http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html. For some species, Human, Mouse, Rat, Arabidopsis, Soybean, Rice, Wheat, greater than one million ESTs are present in the dbEST and constitute a valuable resource for investigation of these species at the molecular level, and these numbers of sequences may be sufficient to make quantitative inferences.

1.2.4 Next-Generation transcriptome sequencing:

Considering the Sanger sequencing as the first generation sequencing technology; further advancement in the sequencing area referred as the next generation sequencing (NGS). These NGS technologies can be further divided to second, third and fourth (Oxford nanopore, still in the development phase) generation sequencing (Barron, Niedringhaus et al. 2011). The second generation sequencing technologies usually refers to four commercially available technologies: Roche/454, Illumina, Applied Biosystems-SOLiD, and Helicos, from HeliScope (Barron, Niedringhaus et al. 2011). In general, these sequencing technologies produce a large number of reads in high-throughput fashion and provide a cost-effective way for genome and transcriptome studies (Morozova, Hirst et al. 2009). Some of them are better than others in different aspects, for example, 454 produces a longer read compared to other second generation sequencing, but the throughput is not as high and suffers from sequencing error in homopolymer regions (Barron, Niedringhaus et al. 2011) while Illumina and SOLiD have very high throughput but the cost of the instrument is very high, the reads are shorter, and they suffer from substitution error in sequencing (Barron, Niedringhaus et al. 2011). Due to the massive amount of data generated, these methods require high-powered computers for data assembly and analysis. Helicos single molecule sequencing by synthesis technology is unique (no PCR amplification, allowing direct sequencing of targeted DNA or RNA molecules) among other second generation sequencing platforms but has very high instrument cost (Barron, Niedringhaus et al. 2011).

Third generation sequencing involves the technologies from three companies, Pacific Bioscience, Complete genomics and Ion Torrent/Life technologies. These sequencing technologies are still in an early development phase, and they will give a lot of advantages over their predecessors such

as longer read length (up to 3000 bp) from Pacific Bioscience, as well as the lowest reagent cost for re-assembling a genome among all sequencing technologies from Complete genomics and direct measurement of nucleobase incorporation events (no need for modified bases) from Ion torrent (Barron, Niedringhaus et al. 2011). Some of the key limitations include high sequencing error rates (Pacific Biosciences), sequencing in repetitive regions or homopolymeric regions (Complete genomics, Ion torrent) (Barron, Niedringhaus et al. 2011).

454 technology has been one of more widely used transcriptome sequencing methods for the last 2-3 years, but is now being gradually replaced by the Illumina RNA-Seq method. RNA-Seq is also known as Whole Transcriptome Shotgun Sequencing (WTSS) and offers up to single base resolution and provides a large number of short reads which are suitable for measuring transcript abundance (Snyder, Wang et al. 2009). One of the limitations of RNA-Seq is in the study of small RNA (sRNAs [<200 bp]). It has been found that number of reads obtained for sRNA does not necessarily correlate with their abundances and this problem arises mainly during the sample preparation and sequencing steps (Ozsolak and Milos 2011).

1.2.5 Applications of transcriptome analysis:

1.2.5.1 Protein-coding gene annotation:

Even ten years after the first human genome sequence draft, the genomes in humans or in other organisms have not yet been fully understood (Brent 2008). Knowledge of transcriptional start, polyadenylation sites, exon-intron structures, splice variants and regulatory sequences are required for a complete genome annotation. EST based approaches provided an effective means

of annotating many abundant protein coding genes. EST sequences are very useful in improvement of the genome annotation as the alignment of the EST against the reference genome can identify the exons, introns, exon-intron junctions, alternatively spliced genes and transcription boundaries of captured genes. In the non-model species, lacking genomic resources, EST sequencing has been used for annotation by comparison with genomes of closely related species (Ondov, Varadarajan et al. 2008). However, it has been noted that Sanger based EST sequencing only finds 60% of transcripts in the cell and, therefore, does not provide a complete representation of transcriptome (Brent 2008). This gap now can be filled by the emergence of second generation sequencing technologies. Moreover, 454 sequencing technology, with its increased sequencing depth and reasonably long reads can be readily assembled into the of transcriptome (Vera, Wheat et al. 2008); the newest version of 454 technology can generate 600 to 700 bp reads.

1.2.5.2 Differential expression analysis and ncRNA detection:

The RNA-Seq data obtained by from different samples (controlled/treated, Normal/Diseased) or different tissue types can be used to measures the abundance of transcripts which are over/under expressed under particular conditions/tissues (Morozova, Hirst et al. 2009). In addition, de novo assembly of transcriptomes could lead to the discovery of novel ncRNA genes. 454 sequencing technology has been used to characterize the small non-coding RNA genes in several species (Aravin, Gaidatzis et al. 2006; Axtell, Jan et al. 2006; Yao, Guo et al. 2007; Zhao, Li et al. 2007).

1.2.5.3 Transcriptome rearrangement and single nucleotide variation profiling:

Aberrant transcriptional events resulting from genome rearrangement are a common feature of human cancers. These rearrangements may include inversions, insertion/deletions and copy number variations etc. (Hanahan and Weinberg 2000). Next generation technologies involving paired end sequencing in which reads (mates in paired end) are sequenced from both ends of the fragment and orientation/distance of mates are known, provide an effective means of detecting the genomic aberrations by mapping the mates against the reference genome (Morozova, Hirst et al. 2009).

Transcriptome sequencing is also widely used for single nucleotide variation profiling. Components of genetic variation, which falls within the coding regions, and might contribute to alteration of function, could be detected by the whole genome re-sequencing but it will be too expensive to carry out routinely. In such cases, transcriptome sequencing would be the method of choice as it is focused only on the coding part of the genes (Morozova, Hirst et al. 2009). One recent advancement in single nucleotide variation profiling is the development of restriction site associated DNA sequencing (RADSeq) (Davey and Blaxter 2010), which can identify, and score thousands of genetic markers randomly distributed across the target genome, from a group of individuals using Illumina sequencing**.**

1.2.6 My objectives:

I analyzed the transcriptome of non-model species chosen primarily for their intrinsic evolutionary and ecological properties using comparative genomic approaches. The two species of interest include an insectivorous plant, pitcher plant (*Sarracenia*) (chapter 4) and a bird song

sparrow (*Melospiza melodia*) (chapter 5). The prime objective of sequencing and analyzing the transcriptome of these species was to determine the main categories of genes, identify fast evolving genes (from substitution rate estimation), find closely related species (based on homology particular for *Sarracenia*), find polymorphic loci (with-in and in-between) and to study transcriptome evolution. In order to perform this analysis, my colleagues constructed sequence libraries (cDNA) of these species and sequenced it via 454 sequencing technology. From my analysis, I gained new insights into these species, which will substantially assist the research communities working on them and their ecology.

1.2.7 References:

Alwine, J. C., D. J. Kemp, et al. (1977). "Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes." Proc Natl Acad Sci U S A **74**(12): 5350-5354.

Aravin, A., D. Gaidatzis, et al. (2006). "A novel class of small RNAs bind to MILI protein in mouse testes." Nature **442**(7099): 203-207.

Axtell, M. J., C. Jan, et al. (2006). "A two-hit trigger for siRNA biogenesis in plants." Cell **127**(3): 565-577.

Barron, A. E., T. P. Niedringhaus, et al. (2011). "Landscape of Next-Generation Sequencing Technologies." Analytical Chemistry **83**(12): 4327-4341.

Becker-Andre, M. and K. Hahlbrock (1989). "Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY)." Nucleic Acids Res **17**(22): 9437-9446.

Boguski, M. S. (1995). "The Turning-Point in Genome Research." Trends in Biochemical Sciences **20**(8): 295-296.

Boguski, M. S., T. M. Lowe, et al. (1993). "dbEST--database for "expressed sequence tags"." Nat Genet **4**(4): 332-333.

Brent, M. R. (2008). "Steady progress and recent breakthroughs in the accuracy of automated genome annotation." Nature Reviews Genetics **9**(1): 62-73.

Davey, J. W. and M. L. Blaxter (2010). "RADSeq: next-generation population genetics." Brief Funct Genomics **9**(5-6): 416-423.

Etemadmoghadam, D., A. deFazio, et al. (2009). "Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas." Clin Cancer Res **15**(4): 1417-1427.

Hanahan, D. and R. A. Weinberg (2000). "The hallmarks of cancer." Cell **100**(1): 57-70.

Hu, N., R. J. Clifford, et al. (2010). "Genome wide analysis of DNA copy number neutral loss of heterozygosity (CNNLOH) and its relation to gene expression in esophageal squamous cell carcinoma." BMC Genomics **11**: 576.

Liu, X. S. (2007). "Getting started in tiling microarray analysis." PLoS Comput Biol **3**(10): 1842-1844.

Mantripragada, K. K., T. Diaz de Stahl, et al. (2009). "Genome-wide high-resolution analysis of DNA copy number alterations in NF1-associated malignant peripheral nerve sheath tumors using 32K BAC array." Genes Chromosomes Cancer **48**(10): 897-907.

Mockler, T. C., S. Chan, et al. (2005). "Applications of DNA tiling arrays for whole-genome analysis (vol 85, pg 1, 2005)." Genomics **85**(5): 655-655.

Morozova, O., M. Hirst, et al. (2009). "Applications of New Sequencing Technologies for Transcriptome Analysis." Annual Review of Genomics and Human Genetics **10**: 135-151.

Noonan, K. E., C. Beck, et al. (1990). "Quantitative analysis of MDR1 (multidrug resistance) gene expression in human tumors by polymerase chain reaction." Proc Natl Acad Sci U S A **87**(18): 7160-7164.

Ondov, B. D., A. Varadarajan, et al. (2008). "Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications." Bioinformatics **24**(23): 2776-2777.

Ozsolak, F. and P. M. Milos (2011). "RNA sequencing: advances, challenges and opportunities." Nature Reviews Genetics **12**(2): 87-98.

Pollack, J. R., C. M. Perou, et al. (1999). "Genome-wide analysis of DNA copy-number changes using cDNA microarrays." Nat Genet **23**(1): 41-46.

Pozhitkov, A. E., D. Tautz, et al. (2007). "Oligonucleotide microarrays: widely applied--poorly understood." Brief Funct Genomic Proteomic **6**(2): 141-148.

Rudd, S. (2003). "Expressed sequence tags: alternative or complement to whole genome sequences?" Trends in Plant Science **8**(7): 321-329.

Schena, M., D. Shalon, et al. (1995). "Quantitative Monitoring of Gene-Expression Patterns with a Complementary-DNA Microarray." Science **270**(5235): 467-470.

Seki, M., M. Narusaka, et al. (2002). "Functional annotation of a full-length Arabidopsis cDNA collection." Science **296**(5565): 141-145.

Snyder, M., Z. Wang, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature Reviews Genetics **10**(1): 57-63.

Stone, E. M., K. N. Rothblum, et al. (1985). "Complete Sequence of the Chicken Glyceraldehyde-3-Phosphate Dehydrogenase Gene." Proceedings of the National Academy of Sciences of the United States of America **82**(6): 1628-1632.

Tsukamoto, Y., T. Uchida, et al. (2008). "Genome-wide analysis of DNA copy number alterations and gene expression in gastric cancer." J Pathol **216**(4): 471-482.

VanGuilder, H. D., K. E. Vrana, et al. (2008). "Twenty-five years of quantitative PCR for gene expression analysis." Biotechniques **44**(5): 619-626.

Vera, J. C., C. W. Wheat, et al. (2008). "Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing." Molecular Ecology **17**(7): 1636-1647.

Yao, Y. Y., G. G. Guo, et al. (2007). "Cloning and characterization of microRNAs from wheat (Triticum aestivum L.)." Genome Biology **8**(6): -.

Zhao, T., G. L. Li, et al. (2007). "A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii." Genes & Development **21**(10): 1190-1203.

CHAPTER 2

MUTATIONAL PATTERNS IN RNA SECONDARY STRUCTURE EVOLUTION EXAMINED IN THREE RNA FAMILIES[1]

[1] Anuj Srivastava, Liming Cai, Jan Mrázek, and Russell L. Malmberg, 2011, PLoS ONE, 6(6): e20484, doi:10.1371/journal.pone.0020484, Reprinted here with permission of publisher.

2.1 ABSTRACT

The goal of this work was to study mutational patterns in the evolution of RNA secondary structure. We analyzed bacterial tmRNA, RNaseP and eukaryotic telomerase RNA secondary structures, mapping structural variability onto phylogenetic trees constructed primarily from rRNA sequences. We found that secondary structures evolve both by whole stem insertion/deletion, and by mutations that create or disrupt stem base pairing. We analyzed the evolution of stem lengths and constructed substitution matrices describing the changes responsible for the variation in the RNA stem length. In addition, we used principal component analysis of the stem length data to determine the most variable stems in different families of RNA. This data provides new insights into the evolution of RNA secondary structures and patterns of variation in the lengths of double helical regions of RNA molecules. Our findings will facilitate design of improved mutational models for RNA structure evolution.

2.2 INTRODUCTION

Molecules of RNA perform biological functions which require that they fold into specific secondary and tertiary structures. Conservation of these structures may be as important as, or more important than, sequence conservation during the course of RNA evolution (Fox and Woese 1975; Gutell, Larsen et al. 1994). The associated base pairing in the double helical region of the RNA molecules is retained via patterns of compensatory mutations across sequences (covariation). Comparative methods for the determination of RNA secondary structures rely on detecting these compensatory mutations (Woese 1993; Gutell 1996).

Although many structural elements (stem-loops, pseudoknots) are conserved within a given RNA family, there is also variation in the presence or absence of certain stem-loops and pseudoknots across evolution, and there is variation in the length of corresponding double-helical regions (Mccormickgraham and Romero 1995; Haas, Banta et al. 1996; Williams and Bartel 1996; Chen, Blasco et al. 2000). The types of variation that might be observable when comparing RNAs thus include single base substitutions, insertions and deletions, base-pair substitutions and insertions and deletions within a conserved stem, and insertion and deletion of entire secondary structure elements.

The patterns of RNA base and base pair changes have been both studied and modeled. One of the earliest models was developed by Knudsen et al. (Knudsen and Hein 1999); it incorporates the information of evolutionary history during RNA secondary structure prediction. Other studies analyzed patterns of compensatory mutations in RNA evolution (Dixon and Hillis 1993) and showed the existence of variable rates of evolution across different rRNA structural elements

(Smit, Widmann et al. 2007). A comparison of various mutational models describing the evolution of RNA secondary structure is presented by Savill et al. (Savill, Hoyle et al. 2001)**.** The patterns of compensatory mutations in RNA structures have been summarized in a matrix called RIBOSUM by analogy with the BLOSUM series of protein matrices; this matrix was developed and used in the RNA search program *RSEARCH* (Klein and Eddy 2003).

Recently, evolutionary models that address structural variation have been proposed. Holmes (Holmes 2004) developed a model of RNA structure evolution, which incorporates insertions and deletions of bases, base pairs, and whole stems. This model was based on the TKF91 model of sequence evolution (Thorne, Kishino et al. 1991; Thorne, Kishino et al. 1992). Other recent models of RNA evolution include the non-reversible generative (birth-death) evolutionary model for insertions and deletions (Rivas and Eddy 2008), and the evolutionary triplet model based on a transducer composition algorithm (Bradley and Holmes 2009). One important potential application of the evolutionary triplet model is the inference of ancestral sequences for a set of diverged RNAs.

Our primary goal in this study was to determine the evolutionary and mutational patterns in double helical regions of RNA secondary structures that are responsible for variability in stem length, focusing on those that lead to stem-insertion and deletion. We chose to work with tmRNA (found in bacteria and organelles), RNaseP A (bacterial), RNaseP B (bacterial) and eukaryotic telomerase RNA sequences. This selection was motivated by the availability of large, well annotated databases for these RNA sequences and structures (Brown 1999; Zwieb, Gorodkin et al. 2003; Griffiths-Jones, Moxon et al. 2005; Podlevsky, Bley et al. 2008). We

mapped structural changes onto phylogenetic trees which were constructed from data independent of the tmRNA, RNaseP and telomerase RNA sequences. Mutational patterns, obtained from correlated evolution of paired bases within the same stem among the related species, were documented by creating single and double nucleotide substitution matrices. In addition to determining the mutational patterns that lead to variability within individual stems, we also examined variability attributed to each stem by principal component analyses (PCA) of the stem length data. Our results build-on and extend early analyses of RNA secondary structure for tmRNA (Williams and Bartel 1996; Zwieb, Wower et al. 1999), RNaseP (Brown 1990; Haas, Banta et al. 1996; Haas and Brown 1998) and telomerase RNA (Lingner, Hendrick et al. 1994; Mccormickgraham and Romero 1995; Chen, Blasco et al. 2000; Dandjinou, Levesque et al. 2004)**.**

2.3 MATERIAL AND METHODS

2.3.1 Alignment analysis

We obtained structural alignments for tmRNAs and RNasePs from the tmRNA database (Zwieb, Gorodkin et al. 2003) and the Ribonuclease P database (Brown 1999), respectively. Vertebrate, Ciliate and *Saccharomyces*, *Kluyveromyces* telomerase RNA structural alignments were obtained from Rfam (Griffiths-Jones, Moxon et al. 2005) and the telomerase database (Podlevsky, Bley et al. 2008), respectively. We preferred these databases over Rfam, as we believed that these databases are specialized for particular molecules and therefore contain better quality structural alignment; they provided expert annotation of the various structures (stem-loops, pseudoknots) across the sequence alignments. The alignments consisted of 268, 126, 25,

35, 22, 7 and 6 sequences for the tmRNA, RNaseP A, RNaseP B, and the Vertebrate, Ciliate, *Saccharomyces* and the *Kluyveromyces* telomerase RNAs, respectively. We chose *K. lactis* structure as a consensus for all 6 species of *Kluyveromyces*, as the telomerase database contains the annotation for the conserved segments only and Rfam has the alignment only for *Saccharomyces* species. Therefore, we used the *K .lactis* structure as a consensus and predicted additional helices in the segments which are unique to other *Kluyveromyces* species using *RNAfold* at default parameters (ViennaRNA-1.8.4) (Zuker and Stiegler 1981; McCaskill 1990; Hofacker, Fontana et al. 1994).

Except for RNaseP A and RNaseP B, the numbers of sequences used in our study are greater than or equal to the number of sequences present in the seed alignment of the Rfam database. We excluded RNaseP A and RNaseP B sequences that did not have corresponding rRNA sequences in the ribosomal database project (Cole, Wang et al. 2009). *RNApasta* (Malmberg, Shaw et al. 2010) was used to determine the length of each stem and loop and the stems involved in the RNA pseudoknot formation. This program takes predetermined RNA structural alignment as input and outputs the length of each stem and loop and information about the stems involved in the pseudoknot formation for each RNA molecule. All the alignments used in study along with their secondary structure model were included in the supplementary material (Text S1, Text S2, Text S3, Text S4, Text S5, Text S6, and Text S7).

2.3.2 Phylogenetic analysis

We obtained rRNA sequences for the same species whose sequences were in the tmRNA and RNaseP datasets from the Ribosomal Database Project (Cole, Wang et al. 2009). For the Ciliate and *Kluyveromyces* telomerase RNAs, corresponding rRNA sequences were obtained from the comparative rRNA website (Cannone, Subramanian et al. 2002). These rRNA sequences were used to create a reference phylogenic tree on which structural characters for each family of RNA were mapped. The vertebrate reference tree was obtained from the tree of life project (Maddison, Schulz et al. 2007) and final branches were adjusted manually from tree created by using the cytochrome B protein sequences. The accession number of cytochrome B sequences obtained from Swiss-Prot is given in supplementary Table S1. For *Saccharomyces*, the reference tree was obtained from the *Saccharomyces* phylogeny website (http://www.genetics.wustl.edu/saccharomycesgenomes/yeast_phylogeny.html).

The reference phylogenetic trees were built by *MrBayes3.1.2* program (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). The details of all the *MrBayes* parameters is given in supplementary Table S2 and the reference tree for each family of RNA under study is shown in the supplementary Figure S1.

We used the *Mesquite (version 2.74 (build 550))* program (Maddison and Maddison 2009) to map the tmRNA, RNaseP and telomerase RNA stem lengths onto the reference phylogenetic tree. The history of each character (the stems) was traced onto the tree using the "reconstruct ancestral state" module of *Mesquite* with maximum parsimony. Given the tree and observed character distribution, this method finds the ancestral states that minimize the number of

steps of character change. The cost of change for the continuous data from state x to state y is (x-y) which can be linear or squared; we used the default squared method as it can handle the trees with polytomies. *Dnapars(version 3.5c)* (Felsenstein 1989), a DNA parsimony program in the *Phylip* suite, was used to construct the hypothetical ancestral sequence at each node of the tree. This program counts the number of changes of bases needed on a given tree. We generated the hypothetical ancestral sequences by turning on the user tree and printing the sequences at the node of the tree options.

2.3.3 Structure evolution analysis

We manually determined which stems were variable across the phylogenetic tree; if one of the branches at the nodes containing closely related species were variable with respect to stem-loops then all the RNA sequences belonging to that particular node were used in the further analysis. We collected the underlying sequences for those stems from our alignment file. Afterwards, we created two types of base pair substitution matrices for each type of RNA under study. The base pair substitution matrices summarize information about the mutations that affect the pairing ability of the RNA molecules. The first matrix was created by counting the base changes that occur in the stem regions of extant sequences (those at the leaves of the trees). The second matrix was created by comparing the changes that occurred with respect to reconstructed ancestral sequences present at the corresponding node in the tree. Similarly, we also created two single nucleotide substitution matrices.

We transformed the counts in each cell of the matrices into observed/expected values using the formula**:** $A_{ijkl} = \log_2 ((f_{ijkl}) / (f_{ij}f_{kl}))$ where A_{ijkl} is the value in any cell of the matrix, f_{ijkl} is

the frequency of base pair change for that cell, f_{ij} and f_{kl} are frequency of individual base pair involved in that change**.** Similarly, in the single nucleotide substitution matrix, observed/ expected values were calculated by the formula**:** $A_{ij} = \log_2 ((f_{ij}) / (f_i f_j))$ where A_{ij} is the value in any cell of the matrix, f_{ij} is the frequency of single nucleotide change for that cell, f_i and f_j are frequency of single nucleotide involved in that change. The expected values were calculated by obtaining the frequencies of nucleotides/base pairs within the matrices.

We also performed principal components analysis (PCA) on the stem length data obtained from the *RNApasta* program. Prior to performing PCA, we clustered data by k-means clustering (Hartigan and Wong 1979) and then used the PCA to display the clusters. K-means clustering assigns each object (RNA molecule) into a predefined number (*k*) of clusters; we grouped the RNA molecules from different species based on similarity in their stem lengths. Both of the above analyses were performed using the R (R 2.9.1) statistical programming language.

2.4 RESULTS

2.4.1 Variable and conserved regions

We used arc diagrams (Figure S2) generated by *RNApasta* (Malmberg, Shaw et al. 2010) to display the length variability shown by each stem for all lineages in three families of RNAs. In these figures stems are divided into three categories based on their variability and colored differently. In addition, based upon the results obtained from the "reconstruct ancestral state"

module of *mesquite*, we showed the ancestral state of each stem in terms of the presence or absence of it at the root node using these arc diagrams (Figure 2.1 and Figure S3).

2.4.2 Types of changes in helical regions

We found that there are two kinds of changes which lead to variability in the presence or absence of specific stems. They are whole stem insertion/deletion and stem gain/loss due to base substitution/indels which create or disrupt secondary structure base pairs. A summary for the two types of changes for every stem in each family of RNAs is shown in Table S3. Among the more than 100 examples of stem-loop evolution listed, we selected several examples of two kinds of changes to discuss in detail.

2.4.2.1 Whole stem insertion/deletion

The first example is stem W1 of tmRNA, which is typically six base pairs long; it is involved in formation of an RNA pseudoknot (PK4) in cyanobacteria and chloroplasts' tmRNA. In cyanobacteria, this pseudoknot divides into two small pseudoknots PK4A and PK4B (Zwieb, Wower et al. 1999). When we mapped stem W1 onto the tree (Figure 2.2 A), we found that out of 14 related species, six species have this stem and out of seven cyanobacterial species, stem W1 is present in five of them. The presence/absence of structure is not certain for *Prochlorococcus marinus* and *Synechococcus* sp. *WH8102*, as this particular region is not sequenced. Interestingly, *Mesostigma viride* (fresh water algae) chloroplasts have this stem. *M. viride* represents the earliest diverging green plant lineage (Lemieux, Otis et al. 2000) and its chloroplast retains this stem which was lost in the other species' chloroplast tmRNAs. In order to determine whether this is an example of a stem insertion or deletion, we examined the

reconstructed ancestral sequence at the common node (ignoring the *Prochlorococcus* and *Synechococcus* sequence during structure reconstruction) of RNA molecules of all these species. The alignment (Figure 2.2 B) clearly suggests that this is an event of whole stem insertion as there is no sequence present at the ancestral node.

The second example is stem R of RNaseP A which is typically 10-12 base pairs long including the bulges. From the mapping of this stem onto the tree (Figure 2.3 A), we found that this stem is present in full length in *B. thetaiotaomicron, P. gingivalis, F .yabuuchiae* and completely absent in *C. limicola* and *C. tepidum*. These species belong to Bacteroidetes/Chlorobi group. A reconstruction of the ancestral sequence (Figure 2.3 B) suggests that this is an event of stem deletion in several derived sequences as there is sequence present at the ancestral node.

2.4.2.2 Stem gain/loss due to base substitutions/indels

The variability in RNA secondary structure length also occurs due to mutations that create or eliminate base pairs in a stem region. These kind of mutations involves indels and substitutions. Two examples of stem gain/loss due to changes in base pairing potential are described below:

Stem D1 of tmRNA is up to 5 base pairs long. When we mapped the variation in this stem onto the tree (Figure 2.4 A), we found that the size of the stem varies among members of the genus *Mycoplasma*. We then analyzed the underlying sequences (Figure 2.4 B) and found that the nucleotides are present for all these species in the double-helical regions but they are mutating in certain positions in such a way that they are no longer able to pair, leading to a variable length for this stem in some tmRNA molecules.

Stem G of Vertebrate telomerase RNA is typically 8 base pair long. Mapping of stem length on a tree (Figure 2.5 A) shows that this stem is variable among the species of order *Rodentia*. This stem is present in full length in *C.porcellus* and partially lost in other species. From the analysis of underlying sequences (Figure 2.5 B), we found that this is an event of stem loss primarily due to base indels.

2.4.3 Substitutions associated with structural variation

We created base pair substitution matrices (Table S4 and Table S5) and single base substitution matrices (Table 2.1 and Table 2.2) combining the mutations from all three RNA families. These matrices were created by observing the variability in the size of each stem among RNA molecules of closely related species (Table 2.1 and Table S4) and variability with respect to hypothetical ancestral sequences (Table 2.2 and Table S5). The counts in each cell of the base pair matrix were transformed into observed/expected values. The total number of events scored in the base pair matrices constructed from extant/extant and ancestral/extant sequence comparisons are 53956 and 16903, respectively.

2.4.4 Principal component analysis on stem length data

We further analyzed the variation in stem lengths by k-means clustering (Hartigan and Wong 1979) followed by principal component analysis (PCA)**.** By comparing clustering results for different values of *k*, we determined that 5, 4, 3, 3 ,3 were natural numbers of clusters for the sequences of tmRNA, RNaseP A, RNaseP B, Ciliate and Vertebrate telomerase RNA, respectively. The clustering followed the taxonomical classification of the species.

We displayed the clusters on a PCA biplot to investigate further variance in stem lengths. The first 2 principal components explain 45% of the overall variance in stem lengths for tmRNA. The biplot of the first 2 principal components for tmRNA (Figure 2.6 A) shows that stems U1 and G1 contribute most to the first and second principal components, respectively. For RNaseP A and RNaseP B, the first two components cover 78% and 80% of the variance, respectively. The biplot of the first 2 principal components for RNaseP A (Figure S4A) shows that the stems L and S contribute most to the first and second principal components, respectively. In fact, the vast majority of the stem length variance in the RNaseP A family can be attributed to these two stems. For the RNaseP B, the major contributors to the first and second principal components (Figure S4B) are stem C, K and Q, respectively. In the eukaryotic ciliate and vertebrate telomerase RNA, the first two components cover 95% and 80% of the variance, respectively. In the Vertebrates, stem F, D (Figure S4C) and in Ciliates stem E, B (Figure 2.6 B) contribute most to the first and second principal components, respectively. We were not able to perform the PCA on *Saccharomyces* and *Kluyveromyces* stem length data as the number of sequences was fewer than number of dimensions (stems). For prokaryotic tmRNA and RNaseP, we investigated possible relationships of the first two principal components with biological properties of the organisms, including oxygen requirements, temperature, energy source and motility. However, we did not find any significant relationship between the biological properties and principal components. Detailed results of the clustering and symbols representing the species are presented in the Supplementary Table S6.

2.5 DISCUSSION

Our analysis of RNA secondary structures centers on documenting the mutational patterns responsible for the variation in the double helical regions, including insertion and deletions of whole stems as well as changes in the stem lengths. Our approach differs from previous studies of tmRNA (Williams and Bartel 1996; Zwieb, Wower et al. 1999) and RNaseP ((Brown 1990; Haas, Banta et al. 1996; Haas and Brown 1998; Collins, Moulton et al. 2000; Ellis and Brown 2009; Sun and Caetano-Anolles 2010) in using a reference phylogenetic tree on to which the stem characteristics of the respective RNAs are mapped (Figure 2.2- 2.5), as well as the other methods of data analysis, and in the number of sequences used. For telomerase RNA, comparative methods were previously used to help predict the consensus structures (Lingner, Hendrick et al. 1994; Mccormickgraham and Romero 1995; Chen, Blasco et al. 2000; Dandjinou, Levesque et al. 2004), but there were no analyses of stem-loop evolution and the base pair changes that accompany it.

Based upon the variability obtained by mapping the structure characters onto the tree, we were able to determine the level of variability shown by every stem of each RNA family under study (Figure S1). We determined the relative frequency of the two categories of events responsible for the variation in the RNA secondary structure (Table 2.3). Our data suggests that models to describe RNA structure evolution have to consider both modes of stem appearance/disappearance; while stem insertion/deletion is the less common mode, the rates differs significantly among three RNA families ($\chi^2 = 16.8019$, df = 2, p-value = 0.0002247).

We constructed matrices to summarize the changes in bases and base pairs that occurred in stems that were variable across the phylogenetic tree. Since we also reconstructed the ancestral sequences, we were able to compare ancestral sequences with extant sequences as well as extant sequences with each other. All the methods available for ancestral sequence reconstruction have their limitations (Krishnan, Seligmann et al. 2004; Williams, Pollock et al. 2006); in particular parsimony and maximum likelihood may lead to sequences which contain fewer of the less common residues than they should (Krishnan, Seligmann et al. 2004; Williams, Pollock et al. 2006). We chose parsimony for the sequence reconstruction since the ancestral RNA structure reconstruction was performed by parsimony, although both are based upon an underlying tree generated by Bayesian methods. The primary effect of the parsimony bias in ancestral sequence reconstruction on our results would be that the matrices comparing ancestral and current sequences would be conservative, slightly underestimating some of the rarer changes. The previously constructed RIBOSUM matrices (Klein and Eddy 2003) are based upon rRNA structure alignments, which are highly conserved molecules and therefore might not be suitable for the analyses, where the structure of the RNA is variable among the related species. In contrast, our matrices should be well-suited for such an analysis as they were derived from alignments showing structural variability in phylogenetically related species. Thus, we also have a gap column '-' in the matrices showing the relative frequencies of indel events.

From the reconstructed ancestral state of each stem, obtained using mesquite, we found that in vertebrate telomerase RNA, stem D (Figure 2.1 A) is absent from the root node. This stem is specific to mammals and is considered to be possibly involved in binding to the TERT protein (Ly, Blackburn et al. 2003). The absence of this stem at the root node suggests that it has been

acquired in the course of evolution and the lack of this stem in species other than mammals might indicate there is an alternative way to interact with TERT protein in these species**.**

We used principal components analysis to identify co-variable stems among the RNA molecules under study. The observation that stem U1 (involved in the formation of RNA pseudoknot PK4) is variable among the tmRNAs (Figure 2.6 A) is consistent with our other observation that the PK4 pseudoknot is absent from chloroplasts and from some endosymbiont tmRNAs. Endosymbionts may be under relaxed selective pressure in order to maintain fast growth and therefore they may tolerate a less efficient stalled translation associated with a suboptimal tmRNA (Gueneau de Novoa and Williams 2004).

In the Ciliate PCA biplot (Figure 2.6 B), we found that *Tetrahymena paravorax* separates from all other species. A comparison among the ciliate telomerase RNA sequences indicates that this is due to the absence of stem B in the *T. paravorax* telomerase RNA. In the previous studies of ciliate telomerase RNA, this helix has been suggested to be a primitive telomerase RNA structural feature and deletion of this stem in *T. paravorax* and in other hypotrich telomerase RNAs is considered to be example of convergent evolution (Mccormickgraham and Romero 1995). Our ancestral arc diagram (Figure 2.1 B) also showed the presence of this stem at the root node.

In summary, we implemented a new approach to analyzing RNA structure from an evolutionary prospective. From this analysis, we conclude that different types of mutations are responsible for the variation in the lengths of double helical regions of RNA. We documented the associated

substitution patterns in log-odds matrices. We also demonstrate the usefulness of PCA in the analysis of the RNA structure alignment. PCA in combination with clustering can easily determine the outliers from the large structure alignment of RNA which can then be subjected to further analysis. Further studies like these of the evolutionary variability of RNA structure and the associated mutational patterns will be essential for improving computational programs that model RNA structures.

2.6 ACKNOWLEDGEMENTS

The authors are gratefully acknowledging the National Science Foundation (IIS 0916250), National Institutes of Health (R01GM072080-01A1) and University of Georgia Franklin College of Arts and Science's research fund, for the support of this study.

2.7 REFERENCES

Bradley, R. K. and I. Holmes (2009). "Evolutionary triplet models of structured RNA." PLoS Comput Biol **5**(8): e1000483.

Brown, J. W. (1999). "The Ribonuclease P Database." Nucleic Acids Res **27**(1): 314.

Brown, J. W., Haas, E. S., Hunt, D. A, and Pace, N. R. (1990). "*Structure, function and evolution of ribonuclease P RNA.*" Second Int. Meetings on Structure, Mechanism and Function of Ribonucleases, Sant Feliu de Guxols, Spain.

Cannone, J. J., S. Subramanian, 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.

Chen, J. L., M. A. Blasco, et al. (2000). "Secondary structure of vertebrate telomerase RNA." Cell **100**(5): 503-514.

Cole, J. R., Q. Wang, et al. (2009). "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis." Nucleic Acids Res **37**(Database issue): D141-145.

Collins, L. J., V. Moulton, et al. (2000). "Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP." J Mol Evol **51**(3): 194-204.

Dandjinou, A. T., N. Levesque, et al. (2004). "A phylogenetically based secondary structure for the yeast telomerase RNA." Curr Biol **14**(13): 1148-1158.

Dixon, M. T. and D. M. Hillis (1993). "Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis." Mol Biol Evol **10**(1): 256-267.

Ellis, J. C. and J. W. Brown (2009). "The RNase P family." Rna Biology **6**(4): 362-369.

Felsenstein, J. (1989). "PHYLIP - Phylogeny Inference Package. *Version 3.2*." Cladistics **5**: 164-166.

Fox, G. E. and C. R. Woese (1975). "The architecture of 5S rRNA and its relation to function." J Mol Evol **6**(1): 61-76.

Griffiths-Jones, S., S. Moxon, et al. (2005). "Rfam: annotating non-coding RNAs in complete genomes." Nucleic Acids Res **33**(Database issue): D121-124.

Gueneau de Novoa, P. and K. P. Williams (2004). "The tmRNA website: reductive evolution of tmRNA in plastids and other endosymbionts." Nucleic Acids Res **32**(Database issue): D104-108.

Gutell, R. R. (1996). "Comparative sequence analysis and the structure of 16S and 23S RNA." Ribosomal RNA: Structure, Evolution, Processing, and Function in Protein Biosynthesis, edited by R. A. Zimmermann and A. E. Dahlberg. CRC Press, Boca Raton: 15-27

Gutell, R. R., N. Larsen, et al. (1994). "Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective." Microbiol Rev **58**(1): 10-26.

Haas, E. S., A. B. Banta, et al. (1996). "Structure and evolution of ribonuclease P RNA in Gram-positive bacteria." Nucleic Acids Res **24**(23): 4775-4782.

Haas, E. S. and J. W. Brown (1998). "Evolutionary variation in bacterial RNase P RNAs." Nucleic Acids Res **26**(18): 4093-4099.

Hartigan, J. A. and M. A. Wong (1979). "A K-means clustering algorithm." Applied Statistics **28**: 100–108.

Hofacker, I. L., W. Fontana, et al. (1994). "Fast Folding and Comparison of RNA Secondary Structures." Monatshefte f. Chemie **125**: 167-188.

Holmes, I. (2004). "A probabilistic model for the evolution of RNA structure." Bmc Bioinformatics **5**: 166.

Huelsenbeck, J. P. and F. Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees." Bioinformatics **17**(8): 754-755.

Klein, R. J. and S. R. Eddy (2003). "RSEARCH: finding homologs of single structured RNA sequences." Bmc Bioinformatics **4**: 44.

Knudsen, B. and J. Hein (1999). "RNA secondary structure prediction using stochastic context-free grammars and evolutionary history." Bioinformatics **15**(6): 446-454.

Krishnan, N. M., H. Seligmann, et al. (2004). "Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference." Mol Biol Evol **21**(10): 1871-1883.

Lemieux, C., C. Otis, et al. (2000). "Ancestral chloroplast genome in Mesostigma viride reveals an early branch of green plant evolution." Nature **403**(6770): 649-652.

Lingner, J., L. L. Hendrick, et al. (1994). "Telomerase Rnas of Different Ciliates Have a Common Secondary Structure and a Permuted Template." Genes & Development **8**(16): 1984-1998.

Ly, H., E. H. Blackburn, et al. (2003). "Comprehensive structure-function analysis of the core domain of human telomerase RNA." Mol Cell Biol **23**(19): 6849-6856.

Maddison, D. R., K. S. Schulz, et al. (2007). "The Tree of Life Web Project." Zootaxa(1668): 19-40.

Maddison, W. P. and D. R. Maddison (2009). "Mesquite: A modular system for evolutionary analysis. *Version 2.6*."

Malmberg, R., T. Shaw, et al. (2010). "RNApasta: a tool for analysis of RNA structural alignments." Journal of Bioinformatics Research and Applications: *In Press*.

McCaskill, J. S. (1990). "The equilibrium partition function and base pair binding probabilities for RNA secondary structure." Biopolymers **29**(6-7): 1105-1119.

Mccormickgraham, M. and D. P. Romero (1995). "Ciliate Telomerase Rna Structural Features." Nucleic Acids Research **23**(7): 1091-1097.

Podlevsky, J. D., C. J. Bley, et al. (2008). "The telomerase database." Nucleic Acids Res **36**(Database issue): D339-343.

Rivas, E. and S. R. Eddy (2008). "Probabilistic phylogenetic inference with insertions and deletions." PLoS Comput Biol **4**(9): e1000172.

Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics **19**(12): 1572-1574.

Savill, N. J., D. C. Hoyle, et al. (2001). "RNA sequence evolution with secondary structure constraints: Comparison of substitution rate models using maximum-likelihood methods." Genetics **157**(1): 399-411.

Smit, S., J. Widmann, et al. (2007). "Evolutionary rates vary among rRNA structural elements." Nucleic Acids Research **35**(10): 3339-3354.

Sun, F. J. and G. Caetano-Anolles (2010). "The ancient history of the structure of ribonuclease P and the early origins of Archaea." Bmc Bioinformatics **11**: -.

Thorne, J. L., H. Kishino, et al. (1991). "An evolutionary model for maximum likelihood alignment of DNA sequences." J Mol Evol **33**(2): 114-124.

Thorne, J. L., H. Kishino, et al. (1992). "Inching toward reality: an improved likelihood model of sequence evolution." J Mol Evol **34**(1): 3-16.

Williams, K. P. and D. P. Bartel (1996). "Phylogenetic analysis of tmRNA secondary structure." RNA **2**(12): 1306-1310.

Williams, P. D., D. D. Pollock, et al. (2006). "Assessing the accuracy of ancestral protein reconstruction methods." PLoS Comput Biol **2**(6): e69.

Woese, C. R., and N. C. Pace (1993). "Probing RNA structure, function and history by comparative analysis." *The RNA World*, edited by R. F. Gesteland and J. F. Atkins. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY: 91-117

Zuker, M. and P. Stiegler (1981). "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information." Nucleic Acids Res **9**(1): 133-148.

Zwieb, C., J. Gorodkin, et al. (2003). "tmRDB (tmRNA database)." Nucleic Acids Res **31**(1): 446-447.

Zwieb, C., I. Wower, et al. (1999). "Comparative sequence analysis of tmRNA." Nucleic Acids Res **27**(10): 2063-2071.

Tables:

Table 2.1: Observed/expected value matrix combining the single nucleotide mutations from extant/extant sequences

	A	**C**	**G**	**U**	**-**
A	4.17	-	-	-	-
C	2.41	3.36	-	-	-
G	2.44	1.25	2.05	-	-
U	2.07	1.79	0.63	1.43	-
-	0.29	-0.30	-0.98	-1.01	0.30

Table 2.2: Observed/expected value matrix combining the single nucleotide mutations from ancestral/extant sequences

	A	**C**	**G**	**U**	**-**
A	2.98	-0.69	-0.16	-0.22	-3.61
C	-0.69	3.01	-1.14	0.20	-3.38
G	0.28	-0.47	2.44	-0.44	-3.23
U	-0.06	0.17	-1.07	2.76	-3.43
-	-1.21	-1.38	-1.61	-1.27	0.51

Table 2.3: Frequency of events in percentage responsible for variation in stem length in RNA secondary structure

RNA family	Whole stem insertion/deletion (%)	Base substitution/indels (%)
tmRNA	43.6	56.4
RNaseP A	27.5	72.5
RNaseP B	15	85
Vertebrate telomerase RNA	9.4	90.6
Ciliate telomerase RNA	7.6	92.3
Saccharomyces telomerase RNA	11.2	88.8
Kluyveromyces telomerase RNA	43.4	56.6

Figures:

Figure 2.1 - RNApasta arc diagram showing the ancestral state of each stem
RNA secondary structure diagram labeled with *RNApasta* annotation showing the ancestral state of each stem in terms of presence/absence of it, for A) Vertebrate telomerase RNA B) Ciliate telomerase RNA; the black and red of the each stem indicates the presence and absence, respectively. A crossing pattern of arcs indicates a pseudoknot. Each alphabet in the figure represents an RNA stem (*RNApasta* notation).

[also see supplement figure S3 for other RNA families].

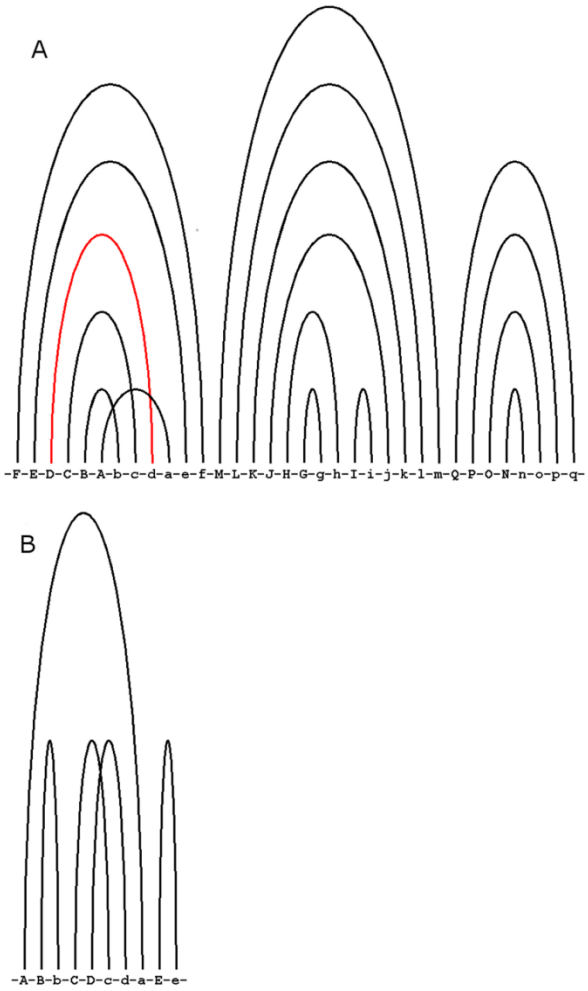
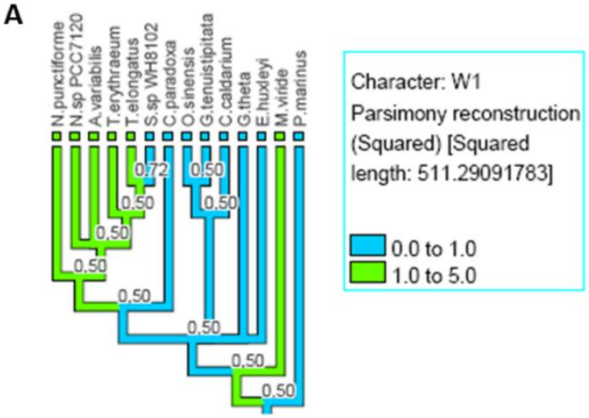


Figure 2.2 - Cyanobacteria and chloroplasts' tmRNA stem W1 length mapped on rRNA phylogenetic tree

A) rRNA phylogenetic tree for cyanobacteria and chloroplasts' for the sequences of tmRNA under study with tmRNA stem W1 length values mapped on the rRNA tree; *MrBayes* calculated posterior probabilities of partition shown on each node of the tree and every branch is colored according to its stem length. The side bar shows the color legend for stem length values mapped onto the tree by *mesquite* using the parsimony ancestral reconstruction method. B) The tmRNA sequences including the reconstructed ancestral sequence (at the top generated by *Dnapars*) for the species present on the rRNA tree in the figure 2.2 A are shown here. The '-' and '~' indicate sequence absence and non-sequenced regions, respectively.

A

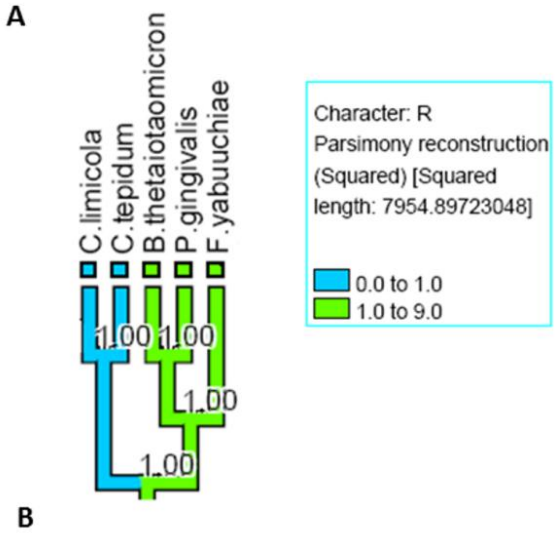


B

```
>Ancestral sequence                  ------  ------
>Nostoc punctiforme PCC73102         AUCCUA  -GGGAU
>Nostoc sp.PCC7120                   AUCCUA  -GGGAU
>Anabaena variabilis ATCC29413       AUCCUA  -GGGAU
>Trichodesmium erythraeum IMS101     AUCCCA  -GGGAU
>Thermosynechococcus elongatus BP-1  AUCCUA  -GGGAU
>Synechococcus sp. WH8102            ~~~~~~  ~~~~~~
>Cyanophora paradoxa                 ------  ------
>Odontella sinensis                  ------  ------
>Gracilaria tenuistipitata           ------  ------
>Cyanidium caldarium                 ------  ------
>Guillardia theta                    ------  ------
>Emiliania huxleyi                   ------  ------
>Mesostigma viride                   UAGAAA  -UUUUA
>Prochlorococcus marinus             ~~~~~~  ~~~~~~
```

Figure 2.3 - Bacteroidetes and Chlorobi RNaseP A stem R length mapped on rRNA phylogenetic tree

A) rRNA phylogenetic tree for Bacteroidetes and Chlorobi for the sequences of RNaseP A under study; RNaseP A stem R values mapped onto the rRNA tree; other legend are similar as figure 2.2 A. B) The RNaseP A sequences including the hypothetical ancestral sequence (at the top generated by *Dnapars*) for the species present on the rRNA tree in the figure 2.3 A are shown here. The '?' indicates that the ancestral base is not certain at that position; other alphabet notation follows the standard IUPAC nucleotide code.

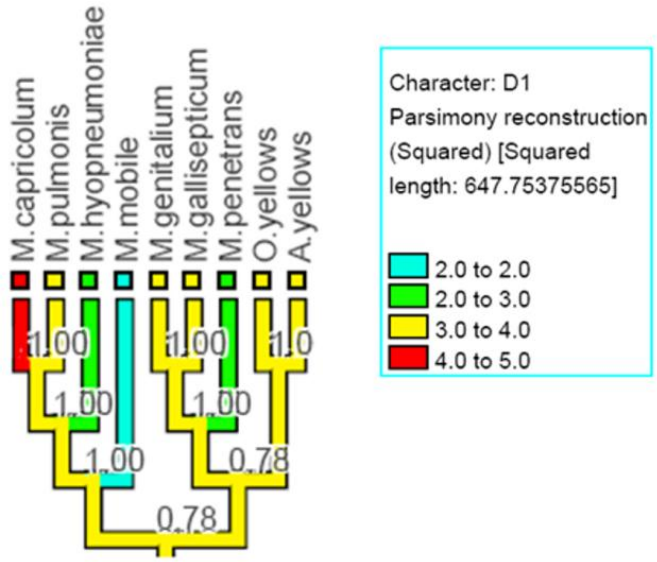


B

```
>Ancestral sequence            G?C-UK-U?G CGG-CMG--?-C
>Chlorobium limicola           ---------- ------------
>Chlorobaculum tepidum         ---------- ------------
>Bacteroides thetaiotaomicron  GUG-uC-GUG CAC-GcC--A-C
>Porphyromonas gingivalis      GCC-aU-GCG CGC-AcG--G-C
>Flavobacterium yabuuchiae     GUC-UG-UCA UGG-CAG--A-U
```

Figure 2.4 - Mycoplasma tmRNA stem D1 length mapped on rRNA phylogenetic tree
A) rRNA phylogenetic tree for *Mycoplasma* for the sequences of tmRNA under study; tmRNA stem D1 values mapped on the rRNA tree; other legend symbols are similar to figure 2.2 A**.** B) Underlying sequences of the species present in the tree shown in figure 2.4 A; the small letter in the sequences indicate those bases which are mutated in such a way that they are not able to pair any more. The '-' indicates the absence of base.

A

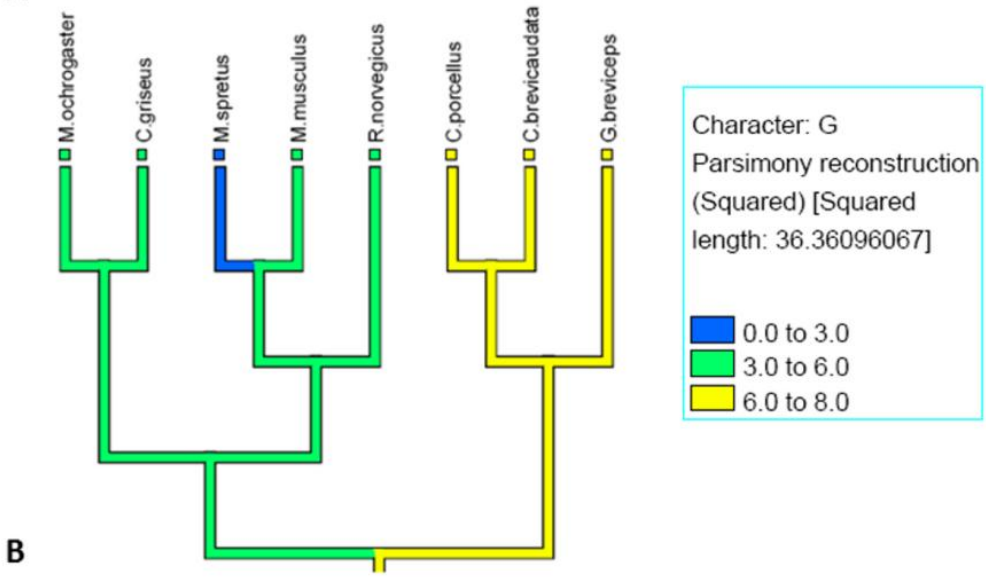


B

>*Ancestral sequence*	AGUGU	ABACK
>*Mycoplasma capricolum*	AGUGU	AUACU
>*Mycoplasma pulmonis*	AGUuU	AuGCU
>*Mycoplasma hyopneumoniae*	gGuUU	GAcCa
>*Mycoplasma mobile*	acUaU	AgAcc
>*Mycoplasma genitalium*	GGGGu	cCCUU
>*Mycoplasma gallisepticum R*	AGGG-	cCCUU
>*Mycoplasma penetrans HF-2*	aGGgU	AgCUg
>*Onion Yellows (Phytoplasma OY-M)*	CUUGg	-CAAG
>*Aster Yellows (Phytoplasma AYWB)*	CUUGg	-CAAG

Figure 2.5 - Vertebrates telomerase RNA stem G length mapped on reference tree
A) Reference phylogenetic tree for Vertebrates for the sequences of telomerase RNA under study; telomerase stem G values mapped on the reference tree; other legend symbols are similar as figure 2.2 A**.** B) Underlying sequences of the species present in the tree shown in figure 2.5A; The '-' indicates the absence of base.

A

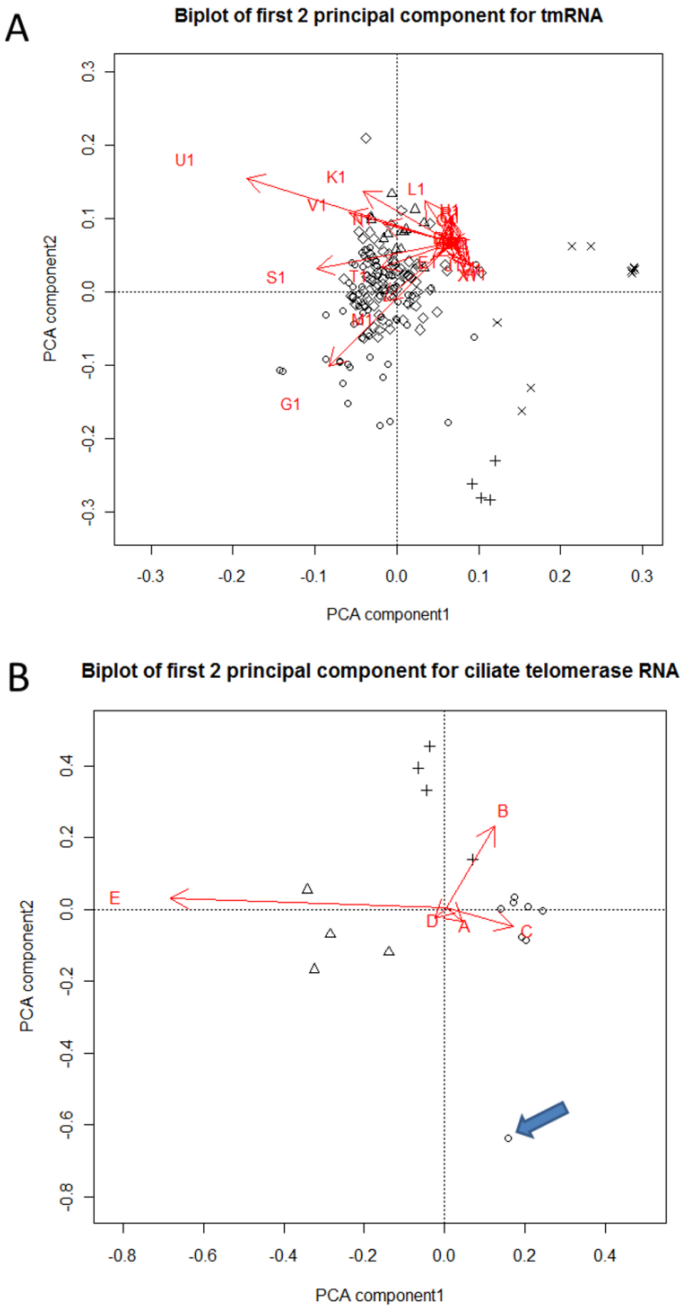


B

```
>Ancestral sequence          GGGGCUUC GAGGYGCCC
>Microtus ochrogaster        GGAG-UUC GA---CUCC
>Cricetulus griseus          GGAG--UC GG---CUCC
>Mus spretus                 GGGG--UC ------UCC
>Mus musculus                GGGG--UC GA---CUCC
>Rattus norvegicus           GGGG--Uc aA---CUCC
>Cavia porcellus             GGGGCUUC GAGGCGCCC
>Chinchilla brevicaudata     GGGGCUUC GGAGCGCC-
>Geomys breviceps            GGGGCCUC GAGGUGCCC
```

Figure 2.6 - PCA Biplot for tmRNA and Ciliate telomerase RNA

Biplot of principal components for A) tmRNA B) Ciliate telomerase RNA; points in different shape represents clusters of species; partial tmRNA sequences were excluded from the analysis. Isolated species mentioned in the discussion are indicated by arrows on biplot.



2.8 SUPPLEMENTARY MATERIAL

Supplementary Data are available at

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0020484

CHAPTER 3

PATTERNS OF CHROMATIN-MODIFICATIONS DISCRIMINATE DIFFERENT GENOMIC FEATURES IN *ARABIDOPSIS*[1]

[1]Anuj Srivastava , Xiaoyu Zhang, Sal LaMarca, Liming Cai and Russell L. Malmberg, To be submitted to Proceedings of the National Academy of Sciences.

3.1 ABSTRACT

Dynamic regulation and packaging of genetic information is achieved by the organization of DNA into chromatin. Nucleosomal core histones, which form the basic repeating unit of chromatin, are subject to various post-translational modifications such as acetylation, methylation, phosphorylation and ubiquitinylation. These modifications have effects on chromatin structure and, along with DNA methylation, regulate gene transcription. In this study, we used publically available data (ChIP-chip) for different types of histone modifications (methylation and acetylation) and for DNA methylation for *Arabidopsis thaliana* and then applied a machine learning based approach (a support vector machine) to demonstrate that patterns of these modifications are very different among different kinds of genomic features (protein, RNA, pseudogene and transposon elements). These patterns can be used to distinguish the types of genomic features. DNA methylation and H3K4me3 methylation emerged as features with most discriminative power**.** From our analysis on *Arabidopsis*, we were able to predict 33 novel genomic features, whose existence was also supported by analysis of RNA-Seq experiments. In summary, we present a novel approach which can be used to discriminate/detect different categories of genomic features based upon their patterns of chromatin modification and DNA methylation.

3.2 INTRODUCTION

In eukaryotic nuclei, DNA associates with proteins to form chromatin. The structure of chromatin plays an essential role in organization of genome, transcriptional activity and developmental state memory (Bernstein, Humphrey et al. 2002). The basic unit is the nucleosome in which 146 base pairs of DNA are wrapped around an octamer of four core histone proteins (H2A, H2B, H3 and H4) (Luger, Mader et al. 1997). Histones play a role in the building of a closed or an open chromatin state which in turn regulates the expression of genes by controlling their accessibility to transcription machinery. The structures of core histone protein are predominantly globular with the exception an unstructured amino-terminal 'tail' of 25-40 residues. A variety of post-translational modifications (acetylation, phosphorylation and methylation) occur on these unstructured tails (Zhang and Reinberg 2001) and have effects on gene expression. These changes are referred to as epigenetic modifications as changes in gene expression are caused by mechanisms other than changes in the underlying DNA sequence.

A second type of epigenetic modification is the addition of methyl groups to the DNA (DNA methylation), primarily at CpG sites, to convert cytosine to 5-methylcytosine. Cytosine DNA methylation is a conserved epigenetic silencing mechanism involved in many important biological processes including defense against transposon proliferation, heterochromatin formation, control of genome imprinting, regulation of endogenous gene expression, and silencing of transgenes (Paszkowski and Whitham 2001; Bender 2004; Zhang, Yazaki et al. 2006). Another type of epigenetic data (but not a modification) is the enrichment of RNA PolII in different genic regions (Chodavarapu, Feng et al. 2010). Relating the multitude of epigenetic modifications to their regulatory effects poses a complex and fascinating challenge.

In recent years, the use of modification-specific antibodies in chromatin immune-precipitations (ChIP) coupled to gene array technology (ChIP on CHIP) has become an important experimental tool to determine these modifications (Kouzarides 2007). An advance on ChIP-chip technology is ChIP-Seq which involves chromatin immune-precipitations followed by sequencing. ChIP-Seq offers greater coverage, less noise and higher resolution than its predecessor ChIP-chip, owing largely to advances in next generation sequencing technology (Park 2009). These technologies can generate chromatin modification maps for particular organisms which can further be used to study epigenetic regulation and other processes which involve of dynamic modification of DNA and histone proteins.

In this research, we used chromatin modification data to test their ability to be markers to discriminate or detect different classes of genomic features (Protein coding, RNA, Pseudogene and Transposable elements). The two main questions that we asked here are: 1) Are there differences between the epigenetic modification patterns of different genomic feature types? 2) Can these patterns be used to find the new instances of these features from the un-annotated regions of genome? To perform this analysis, we gathered data for different kinds of epigenetic modifications (DNA methylation, H3 methylation and H4 acetylation at different lysine residue) and also RNA PolII occupancy of *Arabidopsis thaliana* and then used a machine learning based approach (support vector machine) to distinguish/detect different genomic features.

Support vector machines (SVMs) (Vapnik 1995) are machine learning techniques widely used to solve classification problems (Bhardwaj, Langlois et al. 2005; Barutcuoglu, Schapire et al. 2006; Hoglund, Donnes et al. 2006) . An SVM generates classifications for objects (here, genomic

features) based on a set of input features (here, epigenetic data) associated with the object. In the field of bioinformatics, the SVM is a widely used classification method and has been used in studies such as prediction of DNA-binding proteins (Bhardwaj, Langlois et al. 2005), gene function (Barutcuoglu, Schapire et al. 2006) and protein subcellular localization (Hoglund, Donnes et al. 2006). We implemented an SVM which found there are substantial differences between modification patterns of genomic feature types which can be readily used to distinguish them. We also showed that these patterns can be used to identify novel genomic features from the genomic background whose existence was also confirmed by RNA-Seq experiment.

3.3 RESULTS

We gathered datasets on different types of epigenetic modifications (*DNA methylation*, H3 methylation and H4 acetylation at different lysine residue and RNA PolII occupancy) for *Arabidopsis thaliana*; these were in the form of probabilities for each type of modification for regions of the *Arabidopsis* chromosomes. The coordinates for the genomic features protein coding genes, non-coding RNA genes, pseudogenes, and transposable elements were obtained from *Arabidopsis* GFF (General File Format) file. Scripts were developed which convert the experimentally determined feature probabilities for regions of the genome to basepair coordinates to match the coordinate system of the features in the GFF files. We performed two analyses: first determining the level of bias in chromatin modification pattern which exists between different feature classes using an SVM, and then second predicting novel genomic features by applying the SVM classification method to regions currently labeled intergenic.

3.3.1 Binary and Multiple-Way Classifiers:

Binary classifiers (Two-class SVMs) were used to determine discrimination in epigenetic modification patterns of different genomic features and multiple way classifiers (four-class SVMs) were used to assign the intergenic regions to different feature classes.

We performed 6 different comparisons among 4 different feature classes (protein, RNA, pseudogenes and transposon element genes) using two-class SVMs. We obtained modification probabilities for 24872, 806, 803 and 3006 protein coding, RNA, pseudogenes and transposable element genes, respectively. In each classification, SVM based classifiers were used to separate two feature classes and an F-score was calculated to determine the discrimination power of each epigenetic feature in every comparison (Table 3.1). In general, the larger an F-score, the more discriminative the corresponding feature is. Based upon F-scores, DNA methylation and H3K4me3 emerged as the features with most discriminative power (Figure 3.1). The average value of each type of epigenetic modifications in every feature class determined using their genomic coordinates is shown in Table 3.2 and the results of the two-class SVMs, showing the discrimination between different feature classes, are given in Table 3.3. The Table 1 and Table 3.2 values are also shown in the form of bar plots in figure 3.1.

Since the numbers of each type of genomic feature were different from each other, we used an over-sampling strategy to equalize and balance the numbers. This randomly resampled the data with the smallest numbers to obtain a dataset of the same size as the largest number. Without oversampling, an SVM would classify most instances as protein, since the number of proteins in the non-oversampled dataset is much higher than ncRNAs, pseudogenes, and transposons. Our testing strategy used a 5-fold cross-validation scheme in which 80% of the data was used as a

training set, to be evaluated on the remaining 20%, repeated 5 times so that each 20% fraction was evaluated. For the over-sampling strategy with the smaller data numbers, the 80%/20% division of the data was performed prior to the resampling/amplification so that the final set was derived from the desired 80% of the sequences.

The "validation" row in table3 shows the results from 5 independent validation datasets used in 5-fold cross-validation on the SVMs trained on their corresponding 5 training datasets; the rows containing "training" show the results on the training datasets used in 5-fold cross-validation. The rows labeled as "all" are the results from the SVM trained on all of the oversampled data. The two-class SVM has the highest accuracy (0.90) and MCC (0.81) on its validation sets for transposon/ncRNA classification. The protein/transposon and protein/ncRNA classification has accuracy of 0.90, 0.74 and MCC of 0.80, 0.49, respectively and finally, the ncRNA/pseudogenes classification has the lowest accuracy (0.71) and MCC (0.43) on its validation sets.
We developed a multi-label classifier by combining the binary SVMs after exploring the optimum parameters for doing this. We first tested the ability of the SVMs to discriminate between genomic features using a cross-validation approach similar to that used previously. The results from the 4-class SVM, which is built from 6, 2–class SVMs, are shown in Table 3.4. In a four-class SVM, for each data point, each of the 6, 2-class SVMs makes a prediction and the final prediction from the 4-class is the majority vote from the 6, 2-class SVMs. In the case of a tie, the 4-class SVM will predict the class with the highest probability (Wu, Lin et al. 2004).

Table 3.4 contains three results of the 4-class SVMs in confusion matrix form i.e. validation sets, training sets results (in 5-fold cross-validation) and the 4-class SVM results (trained on all

oversampled data). As shown in Table 3.4, the 4-class SVM (trained on 5 training sets) on the validation datasets has an accuracy of 0.5870, 0.7097, 0.2192 and 0.8323 for protein, RNA, pseudogenes and transposon, respectively. The overall accuracy on the 4-class SVM's on validation set was 0.5871, which is 0.3371 above randomly classifying one class out of the four classes (i.e. 0.25). The 4-class SVMs had the highest and lowest accuracies/reliabilities for predicting transposons and pseudogenes, respectively

3.3.2 Novel feature prediction:

We then used the multi-label classifier to detect novel genomic features by using the chromatin modification probabilities of *Arabidopsis* genomic regions currently annotated as intergenic in the data set. Based upon the number of RNA-Seq reads covering the region, the intergenic data for *Arabidopsis* was divided into two parts i.e. intergenic expressed (617 sequences) and unexpressed (25331sequences). Afterwards, these two datasets were used as a testing set in multi-label classification. The predicted features from the multi-label classification were further filtered by checking their potential for coding by a coding potential calculator (CPC) (Kong, Zhang et al. 2007). The consensus results of SVM prediction and CPC were included in final prediction. In all, we were able to identify 4 protein, 21 ncRNA, 1 pseudogene, and 7 transposons, respectively in intergenic expressed category and 15 protein, 479 ncRNA, 8 pseudogenes and 734 transposons, respectively in the intergenic non-expressed category.

3.4 DISCUSSION

We used data for different types of epigenetic modification from *Arabidopsis* and then used binary SVM classifiers to discriminate the patterns of epigenetic modification among different

genomic features. The F-statistic allowed us to identify the modifications with the greatest impact on a particular classification, however the actual discriminations relied upon combinations of all 7 features considered, using the SVM kernel.

For protein/RNA binary classification, the feature with the most discriminative power is H3K4me1. From the determination of probabilities within the feature region (Table 3.2), we found that average modification probabilities for H3K4me1 are higher for protein coding genes compared to ncRNA genes. Previously, it has been found that H3K4me1 modification occurs predominantly in the transcribed region of genes and has positive correlation with length of the genes (Zhang, Bernatavichute et al. 2009). The low value of H3K4me1 for RNA genes could possibly be due to their length as in our dataset 65% of the RNAs are less than 200bp in length, compared to 2% of protein coding genes.

In the protein/transposon comparison, DNA methylation emerged as the feature with the most discriminative power (Table 3.1) and was associated with transposons (Table 3.2). A strong pattern of methylation is known to be associated with transposable element gene (Paszkowski and Whitham 2001; Bender 2004; Zhang, Yazaki et al. 2006) and it serves as a defense mechanism against proliferation of transposons in the genome. In the protein/pseudogene classification, DNA methylation was the top feature with high F-score (Table 3.1). Pseudogenes also have overall high DNA methylation value comparing to protein coding genes (Table 3.2). Similar to transposable element genes, their high value of DNA methylation is related to transcriptional silencing of pseudogenes (Zhang, Yazaki et al. 2006). However, unlike transposable element genes which are methylated to prevent their deleterious effects,

pseudogenes might be methylated to prevent the cost of transcription for the non-functional unit of genome.

A strong DNA methylation pattern associated with transposons has also the most discriminative power in ncRNA/transposon classification (Table 3.1). The second best feature H3K4me3 has a high F-score and was associated with ncRNA (Table 3.2). Several categories of ncRNA genes (tRNA, miRNA, snoRNA) were previously shown to have higher H3K4me3 methylation compared to DNA methylation and the H3K4me2 type of modifications in rice (Li, Wang et al. 2008); sixty-nine percent of the ncRNA in our dataset were comprised of these 3 types of RNA which explains the high value of H3K4me3. The H3K4me3 type of modification is found in genes known to be highly expressed (Zhang, Bernatavichute et al. 2009) and these ncRNA genes are likely to be highly expressed. In the ncRNA/pseudogene comparison, the two features DNA methylation and H3K4me3 have the most discriminative power and based on average modification probabilities, they are found to be associated with pseudogenes and RNA, respectively (Table 3.2).

For pseudogenes/ transposons DNA methylation also emerged as a feature with most power in separating two classes, similarly to the other comparisons involving transposons. The second best feature in this classification is H3K27me3, which is associated with gene silencing in *Arabidopsis* (Kong, Zhang et al. 2007), and has high average modification probabilities (along with DNA methylation) for pseudogenes. To determine, whether DNA methylation and H3K27me3 occur in tandem or mutually exclusively in pseudogenes, we calculated the correlation coefficient (Spearman) value for the pseudogenes from data extracted from the SVM

feature file and found that these two modifications are inversely related (r = -0.20, p-value <0.01). The existence of two alternate mechanisms of gene silencing which occur largely exclusively suggests the importance to the genome of silencing pseudogenes.

In the normalized F-score plot in figure1, H4K5ac has the lowest value. This epigenetic feature also has the lowest average value compared to other epigenetic modifications in all four feature classes (Table 3.2). Acetylation patterns are positively correlated with gene expression and in particular H4K5ac modification are elevated in transcribed regions of active genes in human (Wang, Zang et al. 2008); there is also enrichment of this modification at origin of replications (Costas, Sanchez et al. 2011). The lower average values indicate that this modification is not frequent as compared to others and particularly is very rare in transposons which makes sense as genomes in general try to silence transposon not activate them.

We predicted novel genomic features from epigenetic modification patterns of intergenic using the multi-label SVM. Data from an RNA-Seq experiment and CPC was used to further verify the predicted features. The higher number of ncRNA genes in the intergenic expressed dataset (RNA-Seq reads present) makes biological sense as protein coding genes are already well annotated in *Arabidopsis* and therefore expressed reads has more likelihood to be associated with ncRNA. In the non-expressed class (RNA-Seq reads absent), the number of ncRNA genes is second to transposon element genes. This is reasonable due to the abundance of transposons in genomes and the lack of transcription evidence in the RNA-Seq data.

In conclusion, we provided support for distinctive patterns of chromatin modifications being associated with different kinds of genomic features, and we demonstrated a novel approach for discriminating/detecting different genomic features based upon these modifications. We did not predict many new features in *Arabidopsis* as it has already being extensively studied. However, with the continuous progress in the field of high-throughput sequencing generating this kind of data is become simpler and cheaper and this approach might be used to discriminate/detect novel features in many newly sequenced plant species such as *Populus* and *Vitis.*

3.5 MATERIAL AND METHODS

3.5.1 Datasets:

We obtained data for 7 different types of chromatin modifications for *Arabidopsis thaliana*. These datasets were generated using biochemical methods in combination with whole-genome tiling microarrays at 35bp resolution (Zhang, Yazaki et al. 2006; Kong, Zhang et al. 2007; Zhang, Bernatavichute et al. 2009; Chodavarapu, Feng et al. 2010; Costas, Sanchez et al. 2011). The datasets came in the form of probabilities of modification of particular genomic region. These probabilities were obtained by tilemap, using a two-state hidden Markov model (HMM) based on probe-level t statistics (Ji and Wong 2005). The region probabilities in these datasets were converted to basepair coordinates using the feature coordinates based on TAIR5. The detailed methodology used in obtaining datasets can be found in these articles (Zhang, Yazaki et al. 2006; Kong, Zhang et al. 2007; Zhang, Bernatavichute et al. 2009; Chodavarapu, Feng et al. 2010; Costas, Sanchez et al. 2011). In addition, we also had the expression information obtained by an RNA-Seq experiment for *Arabidopsis thaliana* (. The RNA-Seq dataset was obtained by

using similar protocol and methods as described in Lister et al. (Lister, O'Malley et al. 2008). We converted the 35 bp region probability values in the dataset, into base specific probabilities. The region ± 20bp were assigned the same modification probabilities and values in adjacent overlapping regions were averaged together.

3.5.2 Obtaining the Genomic Features:

Prior to obtaining the genomic features, we adjusted the coordinates of the epigenetic modification probabilities using the assembly update information file obtained from TAIR database (ftp://ftp.arabidopsis.org/home/tair/Software/UpdateCoord/). This enabled us to use the feature coordinates based upon the latest TAIR release i.e. TAIR10. We obtained GFF file containing coordinates of genomic features (protein, RNA, pseudogene and transposable element gene) of *Arabidopsis* from the TAIR database (TAIR10) and assigned epigenetic modification probabilities to genomic features. Overlapping genomic features and features with less than 30% of the regions covered with modification probabilities were ignored. This cut-off was decided by plotting the number of features against different spanning thresholds.

3.5.3 Feature selection:

We used the feature selection tool provided in LIBSVM (Chang and Lin 2001) to determine which of the initially considered epigenetic features are actually useful in discriminating different genomic features. An F-score (Chang and Lin 2001) was used to measure the discriminating power of each feature value to our classification problem in different categories. The code used and detailed information regarding the F-score can be found at the LIBSVM

website (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#feature_selection_tool). However, in all categories, no single feature stands out and therefore all 7 features were used in classifications.

3.5.4 Creating datasets for SVM classification:

The 24,872 proteins, 806 RNAs, 803 pseudogenes, and 3006 transposons with their corresponding 7 epigenetic feature modification probabilities were used to create datasets for 5-fold cross-validation experiments, for 2-class and 4-class SVMs. The datasets for 5-fold cross-validation for the 4-class SVMs were created by randomly shuffling the data for proteins, RNAs, pseudogenes, and transposons. After the random shuffling, the first 20% of proteins, RNAs, pseudogenes, and transposons were extracted and a union of these 4 extractions was used to create the first validation set; the union of the remaining 80% from each class was used to create the first training set. Similarly, the second 20% of each class was extracted and combined for the second validation dataset, and the remaining 80% were used as the second training dataset, and this was continued until there were 5 independent validation sets and 5 training sets for the 5-fold cross-validation experiments for the 4-class SVM.

To equalize the number of features of each type, oversampling was applied by copying all RNAs (30 times), pseudogenes (30 times) and transposons (8 times) in each dataset, to roughly equalize the number of proteins, RNAs, pseudogenes, and transposons in each validation and training dataset for the 4-class SVMs. Thus, 5 validation and 5 training oversampled datasets were created that had the properties that the validation sets were independent of their corresponding training sets, and each set had roughly an equal number of data points consisting of each class. The “all” dataset for the 4-class SVM was created by taking the union of the 5 validation sets.

The training sets were used for training 4-class SVMs to predict which of the 4 classes a set of 7 epigenetic feature modifications belong to for 5-fold cross-validation experiments, the validation sets were used as an unbiased testing set for the trained 4-class SVMs for 5-fold cross-validation experiments, and the "all" dataset was used to train the 4-class SVM on 100% of the oversampled data.

For the 2-class SVMs, the 4-class SVM training, validation, and all oversampled datasets were split into 6 subsets consisting of 6 binary combinations of the 4 different classes: {protein, RNA}, {protein, pseudogene}, {protein, transposon}, {RNA, pseudogene}, {pseudogene, transposon}, and {RNA, transposon}. For example, the {protein, RNA} validation datasets consisted of subsets of the 4-class SVM validation datasets that contained all of the protein and RNA data points, but no data points from the other classes. This was also used to create the training and "all" datasets for the 2-class SVMs that was used to predict if a sequence is protein or RNA, and this process was repeated to create validation, training, and "all" datasets for the other binary combinations. The validation, training, and "all" datasets for the 2-class SVMs also have the properties that the number of instances of each class is roughly equal, and the validation datasets are independent of their corresponding training datasets.

3.5.5 SVM classification experiments:

Six, 2-class SVM classifiers were created using the LIBSVM package (Chang and Lin 2001) that was trained on the 2-class training datasets for each of the 6 binary combinations for 5-fold cross-validation experiments. The radial basis kernel function and SVM probability estimates were used in the LIBSVM package. The 4-class SVM was built in a similar fashion using

LIBSVM, but it used the majority vote multi-label class SVM that splits the problem of 4-classification into a 6 part, 2-classification problem.

The training of the 4-class SVM-based classifier was performed using a standard procedure provided in LIBSVM (Chang and Lin 2001) to find values of two parameters *C* and γ, where *C* controls the trade-off between training errors and classification margins, and γ determines the width of the radial basis kernel (Chang and Lin 2001). A grid search using 5-fold cross-validation was used on the training and validation sets for the 4-class SVM for values of lg *C* ranging from -14 to 10 (as shown in figure 1) and values of γ of 0.125, 2, 8, and 16. The optimal set of parameters found was $C = 1$ and $\gamma = 2$, and these parameters yielded the lowest average error (1 – average accuracy) on their independent validation sets for the 5-fold cross-validation experiments without showing signs of over-fitting. Figure 3.1 shows the results of the grid search for $\gamma = 2$, and it shows that there is little evidence of over-fitting at $C = 1$ and $\gamma = 2$ (the minimum average error) since the average error on the validation sets for 5-fold cross-validation on the 4-class SVM increased for $C > 1$ ($\lg C > 0$) while the average error for the training sets decreased for $C > 1$. These values of *C* and γ were used for training the 2-class and 4-class SVMs with LIBSVM using the radial basis kernel and probability estimates on their respective training and "all" datasets.

3.5.6 Prediction of novel genomic features:

The coordinates and sequences of intergenic regions were obtained from the TAIR database (reference). Intergenic sequences were divided into two parts i.e. intergenic expressed/unexpressed, based on RNA-Seq expression data. The intergenic regions with at least

2 RNA-Seq reads covering at least 20% of the total intergenic region length were considered to be expressed. This threshold was decided after plotting different combinations i.e. number of reads/spanning length against the number of sequences. We further ignored the intergenic regions which are less than 200bp and also ignored the sequences ±50bp from both ends of the intergenic region .

We then extracted the chromatin modification probabilities of the intergenic regions and used them to identify novel features. Potentially novel features were identified using genomic feature probabilities and intergenic region probabilities as training and testing datasets during classification, respectively. The multi-label classifier provides the probability estimates for a test data instance of its belonging to each of four feature class. We chose a probability threshold of 0.70 (determined after plotting the distribution of values of each feature classes) to assign the data instance to particular feature type. To make our prediction more reliable, we took the sequences of intergenic regions and checked the coding potential of predicted features by coding potential calculator (CPC) (Kong, Zhang et al. 2007). The regions predicted as protein coding genes and also predicted as coding by CPC were considered as protein coding and vice-versa. A similar analysis was performed for both intergenic expressed/unexpressed categories.

3.6 ACKNOWLEDGEMENTS

We gratefully acknowledge support from NSF grant IIS 0916250 to Liming Cai, and from the University of Georgia Franklin College of Arts & Science's research fund.

3.7 REFERENCES

Barutcuoglu, Z., R. E. Schapire, et al. (2006). "Hierarchical multi-label prediction of gene function." Bioinformatics **22**(7): 830-836.

Bender, J. (2004). "DNA methylation and epigenetics." Annu Rev Plant Biol **55**: 41-68.

Bernstein, B. E., E. L. Humphrey, et al. (2002). "Methylation of histone H3 Lys 4 in coding regions of active genes." Proc Natl Acad Sci U S A **99**(13): 8695-8700.

Bhardwaj, N., R. E. Langlois, et al. (2005). "Kernel-based machine learning protocol for predicting DNA-binding proteins." Nucleic Acids Research **33**(20): 6486-6493.

Chang, C.-C. and C.-J. Lin (2001). "LIBSVM: a library for support vector machines."

Chodavarapu, R. K., S. H. Feng, et al. (2010). "Relationship between nucleosome positioning and DNA methylation." Nature **466**(7304): 388-392.

Costas, C., M. D. Sanchez, et al. (2011). "Genome-wide mapping of Arabidopsis thaliana origins of DNA replication and their associated epigenetic marks." Nature Structural & Molecular Biology **18**(3): 395-U190.

Hoglund, A., P. Donnes, et al. (2006). "MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition." Bioinformatics **22**(10): 1158-1165.

Ji, H. K. and W. H. Wong (2005). "TileMap: create chromosomal map of tiling array hybridizations." Bioinformatics **21**(18): 3629-3636.

Kong, L., Y. Zhang, et al. (2007). "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine." Nucleic Acids Research **35**: W345-W349.

Kouzarides, T. (2007). "Chromatin modifications and their function." Cell **128**(4): 693-705.

Li, X. Y., X. F. Wang, et al. (2008). "High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression." Plant Cell **20**(2): 259-276.

Lister, R., R. C. O'Malley, et al. (2008). "Highly integrated single-base resolution maps of the epigenome in Arabidopsis." Cell **133**(3): 523-536.

Luger, K., A. W. Mader, et al. (1997). "Crystal structure of the nucleosome core particle at 2.8 A resolution." Nature **389**(6648): 251-260.

Park, P. J. (2009). "ChIP-seq: advantages and challenges of a maturing technology." Nature Reviews Genetics **10**(10): 669-680.

Paszkowski, J. and S. A. Whitham (2001). "Gene silencing and DNA methylation processes." Curr Opin Plant Biol **4**(2): 123-129.

Vapnik, N. V. (1995). "The Nature of Statistical Learning Theory." Springer.

Wang, Z., C. Zang, et al. (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome." Nat Genet **40**(7): 897-903.

Wu, T. F., C. J. Lin, et al. (2004). "Probability estimates for multi-class classification by pairwise coupling." Journal of Machine Learning Research **5**: 975-1005.

Zhang, X., J. Yazaki, et al. (2006). "Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis." Cell **126**(6): 1189-1201.

Zhang, X. Y., Y. V. Bernatavichute, et al. (2009). "Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana." Genome Biology **10**(6): -.

Zhang, Y. and D. Reinberg (2001). "Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails." Genes Dev **15**(18): 2343-2360.

Tables:

Table 3.1: F-score result for each binary classification category

Epigenetic Feature	Protein/RNA	Protein/Pseudo	Protein/Trans	RNA/Pseudo	RNA/Trans	Pseudogene/Trans
H3K27me3	0.000945	0.000305	0.011745	0.021394	0.01527	0.075851
H4K5ac	0.000098	0.000042	0.000689	0.000639	0.000642	0.0015
DNA methylation	0.00392	0.015529	0.632277	0.142858	0.601012	0.213082
RNA PolII	0.001991	0.001688	0.014568	0.031097	0.071851	0.006889
H3K4me1	0.008662	0.004281	0.033849	0.013567	0.001405	0.025655
H3K4me2	0.00233	0.005418	0.03183	0.007523	0.032187	0.010474
H3K4me3	0.000388	0.013048	0.062806	0.135458	0.285946	0.019144

Table 3.2: Average value of each feature obtained using their genomic coordinates

Epigenetic feature	Protein (mean ± SEM)	RNA (mean ± SEM)	Pseudogene (mean ± SEM)	Transposon (mean ± SEM)
H3K27me3	0.0728 ± 0.00003	0.0748 ± 0.0003	0.1206 ± 0.0003	0.0161 ± 0.00004
H4K5ac	0.0018 ± 0.000003	0.0013 ± 0.00002	0.0014 ± 0.00002	0.0006 ± 0.000006
DNA methylation	0.1231 ± 0.00004	0.0723 ± 0.0003	0.2467 ± 0.0004	0.6569 ± 0.0001
RNA PolII	0.0778 ± 0.00003	0.1083 ± 0.0004	0.0460 ± 0.0002	0.0175 ± 0.00004
H3K4me1	0.1414 ± 0.00004	0.0364 ± 0.0002	0.0384 ± 0.0001	0.0116 ± 0.00003
H3K4me2	0.1164 ± 0.00003	0.1000 ± 0.0004	0.0546 ± 0.0002	0.0196 ± 0.00005
H3K4me3	0.2051 ± 0.00005	0.2532 ± 0.0006	0.0473 ± 0.0002	0.0094 ± 0.00003

Table 3.3: Two-class SVM results

Results from 2-class SVMs C = 1 and γ = 2											
+	**-**	**Dataset**	**TP**	**FP**	**FN**	**TN**	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**	**MCC**
Protein	**RNA**	Validation	18352	6030	6520	18150	0.7441	0.7527	0.7379	0.7506	0.4884
		Training	73603	20670	25885	76050	0.7627	0.7807	0.7398	0.7863	0.5265
		All	18723	5580	6149	18600	0.7609	0.7704	0.7528	0.7692	0.5220
Protein	**Pseudogene**	Validation	16873	5160	7999	18930	0.7312	0.7658	0.6784	0.7858	0.4665
		Training	67654	17160	31834	79200	0.7498	0.7977	0.6800	0.8219	0.5064
		All	17007	4260	7865	19830	0.7524	0.7997	0.6838	0.8232	0.5113
Protein	**Transposon**	Validation	23199	3312	1673	20736	0.8981	0.8751	0.9327	0.8623	0.7977
		Training	92853	12880	6635	83312	0.9003	0.8782	0.9333	0.8661	0.8019
		All	23217	3232	1655	20816	0.9001	0.8778	0.9335	0.8656	0.8016
Pseudogene	**RNA**	Validation	15270	5130	8820	19050	0.7110	0.7485	0.6339	0.7878	0.4269
		Training	64140	18210	32220	78510	0.7388	0.7789	0.6656	0.8117	0.4826
		All	16050	4620	8040	19560	0.7377	0.7765	0.6663	0.8089	0.4802
Pseudogene	**Transposon**	Validation	16440	4144	7650	19904	0.7550	0.7987	0.6824	0.8277	0.5156
		Training	67740	16152	28620	80040	0.7675	0.8075	0.7030	0.8321	0.5396
		All	16860	4032	7230	20016	0.7660	0.8070	0.6999	0.8323	0.5369
Transposon	**RNA**	Validation	21184	1650	2864	22530	0.9064	0.9277	0.8809	0.9318	0.8138
		Training	85144	5460	11048	91260	0.9144	0.9397	0.8851	0.9435	0.8302
		All	21304	1380	2744	22800	0.9145	0.9392	0.8859	0.9429	0.8303

Table 3.4: Four-class SVM results

Confusion matrix for 4-class SVMs (C = 1 and γ = 2) Validation Sets in 5-fold cross-validation						
	Predicted by SVM					
		Protein	**RNA**	**Pseudogene**	**Transposon**	**Accuracy**
Observed in oversampled dataset	**Protein**	14600	5409	3539	1324	0.5870
	RNA	3720	17160	2280	1020	0.7097
	Pseudogene	4020	7440	5280	7350	0.2192
	Transposon	1024	1776	1232	20016	0.8323
	Reliability	0.6249	0.5399	0.4282	0.6737	
Average Class Accuracy	0.5053					
Average Class Reliability	0.5667					
Overall Accuracy	0.5871					
Confusion matrix for 4-class SVMs (C = 1 and γ = 2) Training Sets in 5-fold cross-validation						
	Predicted by SVM					
		Protein	**RNA**	**Pseudogene**	**Transposon**	**Accuracy**
Observed in oversampled dataset	**Protein**	58591	21508	14118	5271	0.5889
	RNA	13260	70620	8670	4170	0.7301
	Pseudogene	13560	28500	25290	29010	0.2625
	Transposon	3976	6960	5056	80200	0.8337
	Reliability	0.6555	0.5535	0.4760	0.6759	
Average Class Accuracy	0.5272					
Average Class Reliability	0.5902					
Overall Accuracy	0.6037					
Confusion matrix for 4-class SVMs (C = 1 and γ = 2) All data						
	Predicted by SVM					
		Protein	**RNA**	**Pseudogene**	**Transposon**	**Accuracy**
Observed in oversampled dataset	**Protein**	14715	5380	3464	1313	0.5916
	RNA	3360	17640	2130	1050	0.7295
	Pseudogene	3420	7140	6240	7290	0.2590
	Transposon	1008	1736	1232	20072	0.8347
	Reliability	0.6539	0.5530	0.4776	0.6753	
Average Class Accuracy	0.5267					
Average Class Reliability	0.5900					
Overall Accuracy	0.6036					

Figures:

Figure 3.1: The power of each feature in discrimination (normalized F-score) and the average value of each feature in protein, RNA, pseudogene and transposon element regions. In the normalized F-score plot, DNA methylation has the maximum F-score so it is defined to be 1.0 and other feature F-scores were divided by the DNA methylation values. Other plots have the mean and SEM (standard error of mean values) obtained using the co-ordinate of genomic features. The standard error is indicated by a fuzzy area at the top of the bar.

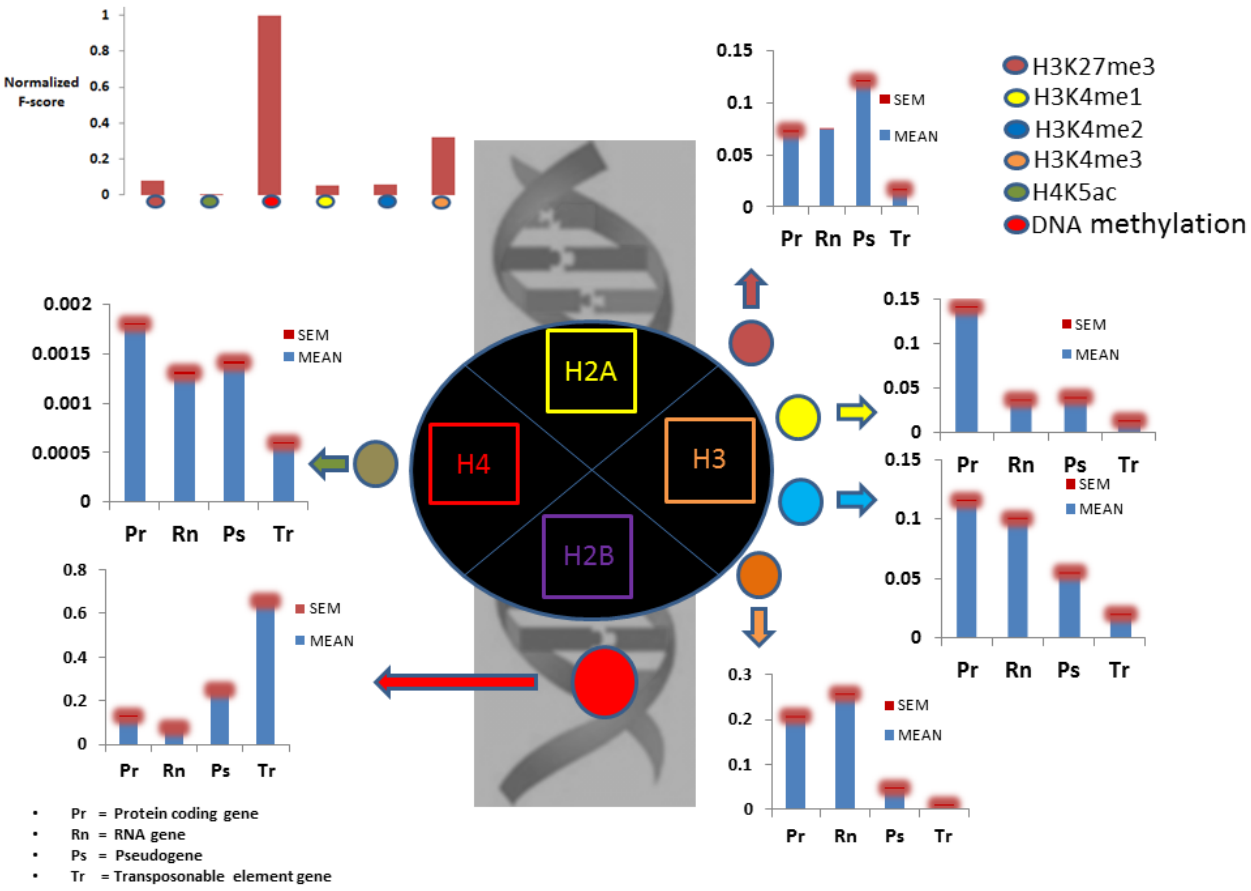
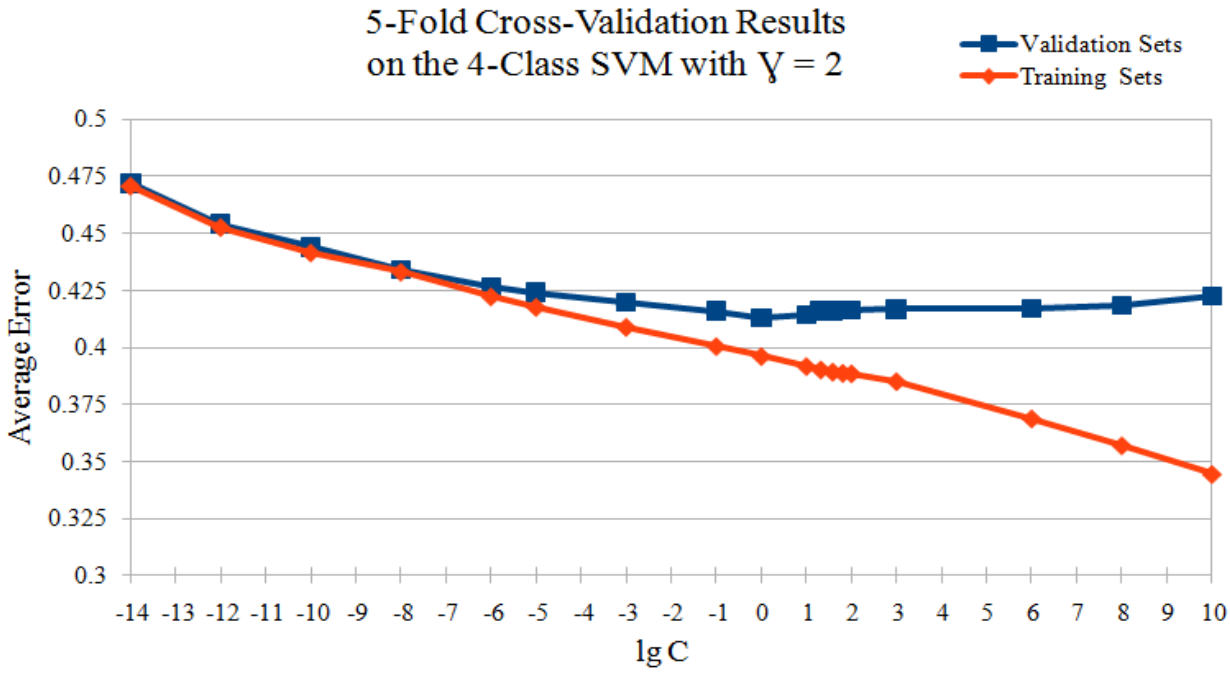


Figure 3.2: Fivefold cross-validation results indicate the lack of over fitting which might occur due to over-sampling of the data. At the optimal parameters, the average error (1- average accuracy) is same for the training and validation sets. These parameters were used during feature discrimination and detection.



CHAPTER 4

TRANSCRIPTOME ANALYSIS OF *SARRACENIA*, AN INSECTIVOROUS PLANT[1]

[1]Anuj Srivastava, Willie L. Rogers, Catherine M. Breton, Liming Cai and Russell L. Malmberg, 2011, DNA Research, doi: 10.1093/dnares/dsr014, Reprinted here with permission of publisher.

4.1 ABSTRACT

Sarracenia species (pitcher plants) are carnivorous plants which obtain a portion of their nutrients from insects captured in the pitchers. To investigate these plants, we sequenced the transcriptome of two species, *S. psittacina* and *S. purpurea*, using Roche 454 pyrosequencing technology. We obtained 46,275 and 36,681 contigs by de novo assembly methods for *S. psittacina* and *S. purpurea*, respectively, and further identified 16,163 orthologous contigs between them. Estimation of synonymous substitution rates between orthologous and paralogous contigs indicates the events of genome duplication and speciation within the *Sarracenia* genus, occurred approximately 2 million years ago. The ratios of synonymous and non-synonymous substitution rates indicated 491 contigs have been under positive selection (K_a/K_s >1). Significant proportions of these contigs were involved in functions related to binding activity. We also found that the greatest sequence similarity for both of these species was to *Vitis vinifera*, which is most consistent with a non-current classification of the order Ericales as an asterid. This study has provided new insights into pitcher plants and will contribute greatly to future research on this genus and its distinctive ecological adaptations.

4.2 INTRODUCTION

Carnivorous plants fascinate both scientists (Darwin 1875) and the general public (Audrey Jr. in: Corman, 1960). One carnivorous plant genus is *Sarracenia* (pitcher plants) which typically grow in highly acidic, nutrient poor soils that are water saturated for at least part of the year, such as bogs, grassy savannas, fens, and similar wetlands. They obtain a portion of their nutrients from prey captured in their pitchers - highly modified tubular leaves. *Sarracenia* species may digest their prey directly with secreted proteases, phosphatases, nucleases (Hepburn, John et al. 1920; Gallie and Chang 1997). However, one of our focal species, *S. purpurea* hosts a complex food web of bacteria, protozoa, and arthropods that mineralize the prey and release nutrients that are taken up by the plant (Gotelli and Ellison 2006). The other focal species, *S. psittacina* does not host such a food web. Despite the interspecific variability in pitcher form, the flowers of the different species are morphologically quite similar, and they are pollinated by a range of generalist bees and sarcophagid flies (Schnell 1983; Ne'eman, Ne'eman et al. 2006); interspecific hybrids with intermediate morphologies are common in nature and are fertile. A number of the taxa within the genus are considered endangered. As a genus, *Sarracenia* provides a wealth of opportunities for ecological and evolutionary studies.

Here, we present a comprehensive analysis of the transcriptome (cDNA) of two *Sarracenia* species (*S. psittacina* and *S. purpurea*) obtained by 454 sequencing technology (454 GS FLX). Transcriptome sequencing represents the subset of genes from the genome that are functionally active in a selected tissue and species of interest. In nonmodel organisms, lacking genomic resources such as sequenced genome, transcriptome study is an effective way to study the gene expression and address comparative genomic-level questions (Bouck and Vision 2007; Hudson 2008). Moreover, massive parallelized sequencing technologies, have made transcriptome

studies, one of the most cost-effective methods for gene discovery (Bouck and Vision 2007), even more robust and efficient. 454 sequencing was selected as its sufficiently long sequence reads can help compensate for the lack of a reference genome during de novo sequence assembly (Rokas and Abbot 2009).

As this is the first set of sequence data for any pitcher plants, we addressed a number of questions in this study ranging from: identifying the events of genome duplication, determining the level of orthology between two species, estimating the substitution rates between the orthologous contigs and investigating the contigs which show signatures of diversifying natural selection (a non-neutral rate of synonymous and non-synonymous substitutions between sequence pairs). These comparative genomics methods along with genetic mapping, expression profiling and candidate gene approaches are part of investigating the genetic basis of phenotypic variation (Ellegren and Sheldon 2008). We also performed a functional annotation, through gene ontology analysis, of all contigs in order to detect any pattern (biological process, molecular function and cellular component) which may be unique or predominant to pitcher plants.

4.3 MATERIAL AND METHODS

4.3.1 Plant material

Two species of *Sarracenia* were chosen based on morphology and differences in insect trapping/digestion. *S. purpurea* has the widest natural range among all species within the genus. It is found from Mississippi eastward up the entire U.S. east coast. It is also found in all the Great Lake states and throughout the majority of Canada. Natural populations of *S. psittacina* can be

found in all Gulf coast states except Texas and it also has populations in Georgia. Fresh juvenile leaf samples were taken from greenhouse maintained plant stocks. Samples came from identical or very similar cultivars generated from rhizome propagation.

4.3.2 RNA Extraction

After multiple extraction attempts using various kits and other wet lab techniques a successful protocol was found using the Spectrum Plant Total RNA Kit (Sigma #STRN50-1KT). Only the youngest unopened leaves were used as older tissues yield negligible RNA amounts. Certain steps call for the use of RNase-free water; this was created by adding 0.1ml of diethylpyrocarbonate to 100ml of water, incubating at 37°C for 12 hours and then autoclaving for 15 minutes to remove trace amounts of the chemical. All instruments and surfaces were cleaned thoroughly, treated with an RNase deactivating solution (100mM sodium hydroxide plus .1% sodium lauryl sulfate) and wiped dry. The mortar and pestles were frozen with liquid nitrogen for about 20 seconds prior to the start and maintained at a subzero temperature throughout tissue grinding to prevent RNA degradation. A single total RNA prep of 100 mg of young leaf tissue yielded on average 4-8 ug of high quality RNA. The extraction was repeated 40-50 times. We obtained approximately .25 mg of total RNA for each species.

4.3.3 mRNA Isolation and cDNA generation

The Oligotec mRNA Kit (Qiagen # 70022) was used to purify mRNA from 0.25 mg extracted total RNA, according to the manufacturer's recommendation. During the elution steps, 25ul of OEB buffer was used in the initial step and 25ul in the follow up step for maximum mRNA concentration in the smallest volume possible. mRNA quality was checked with a Bioanalyzer

(Agilent, Inc) and a NanoDrop 2000 (Thermo Scientific). For cDNA generation, the Evrogen cDNA synthesis kit (SK001) was used with a modification to the kit's 3' primer. We used approximately 26 ng of mRNA for each species.

Normalization of the cDNA was performed by using Evrogen Normalization Trimmer Kit (NK001) in order to minimize the repetition of transcripts. This normalization protocol is based on denaturing-reassociation of cDNA, followed by digestion with a duplex-specific nuclease (DSN). The single-stranded cDNA fraction was amplified twice by sequential PCR reactions according to the manufacturer's protocol. A Qiagen MiniElut kit (#28006) was used to purify the normalized cDNA and sterile water was used in the final elution in order to prevent the likely interference of TE with 454 processing. Normalized cDNA in 100 ul sterile water was submitted to the Georgia Genomic Facility (www.dna.uga.edu) for 454 sequencing in a Genome Sequencer TM (GS) Titanium FLX instrument (Roche Diagnostics) employing a standard protocol. *S. psittacina* and *S. purpurea* cDNAs were submitted at a concentration of 49 ng/ul and 45 ng/ul respectively.

4.3.4 Sequence assembly, contig annotation and ortholog/paralog identification:

We called the bases from the sequence data using ***pyrobayes*** **(Quinlan, Stewart et al. 2008)** from the 454 sequencer generated sff files. Vector and other contaminants were removed using ***seqclean*** (http://compbio.dfci.harvard.edu); assembly of the cleaned reads was performed by ***MIRA*** (Chevreux, Pfisterer et al. 2004). ***Blast2Go*** (B2G) (Conesa, Götz et al. 2005) was used at the default criterion to functionally annotated the contigs. In order to localize the contigs, we obtained *Vitis* genomic sequences (since this genome had the top hits in blastx searches) and gff

files from the phytozome project (http://www.phytozome.net/) and then mapped both the assemblies against the *Vitis* genome using ***blat*** (step size =11, minScore =40) (Kent 2002). Orthologous sequences between the two *Sarracenia* species were predicted using the reciprocal blast (blastn) (Altschul, Madden et al. 1997) hit method at e-value 1e-20. This stringent e-value cut-off leads to a higher identification of orthologous as opposed to paralogous sequences. Paralogous sequences were identified by doing all vs all blast (blastn) within the species. Sequences producing a significant alignment over 300 BP and 40% identity were defined as paralogs (Blanc and Wolfe 2004).

4.3.5 Detection of genome duplication and speciation events:

The approach used in the detection of genome duplication and speciation event was adapted from Blanc & Wolfe (Blanc and Wolfe 2004). Identified paralogous pairs were organized into gene families using single linkage clustering. Afterwards, synonymous substitution rates (K_s) were estimated for all possible pairs within a gene family. One potential drawback of measuring the substitution rates from transcriptome data is that multiple entries of the same gene can be present which leads to redundant K_s measures. To minimize this false peak in the K_s distribution plot, we discarded one of the sequence from paralogous pairs with $K_s = 0$ (assuming that no synonymous substitutions between sequences means they belong to same gene) and all other K_s values involving that particular sequence. We corrected for multiple K_s comparisons from gene families which contain non-overlapping incomplete sequences by a simple clustering method, as described in Blanc & Wolfe (Blanc and Wolfe 2004).

4.3.6 Estimation of substitution rates:

To estimate the synonymous (K_s) and non-synonymous substitutions (K_a) rates between paralogous and orthologous sequences pairs, we first aligned the sequence pairs using *tblastx* (Altschul, Madden et al. 1997); sequences producing significant alignments were extracted using their aligned coordinates and further analyzed by translation and then amino acid sequence alignment was performed by ***Clustalw*** *(Larkin, Blackshields et al. 2007)*. Only the longest uninterrupted reading frame was used for the analysis. Corresponding codon alignments were produced using ***PAL2NAL*** *(Suyama, Torrents et al. 2006)* and finally rates were estimated using a maximum likelihood method implemented in the ***CODEML*** program of the *PAML* package Version 4.1(Yang 2007). Pair-wise maximum likelihood analyses were performed in runmode-2. In order to minimize the statistical artifacts which could arise due to short alignments and a saturation of K_s, we further discarded those alignments which are less than 30aa in length and which had K_s greater than 2. The K_s frequency in each interval size of 0.01 within the range [0, 2.0] was plotted.

4.4 RESULTS

4.4.1 Sequencing and assembly:

The amount of raw sequences obtained was 123 Mb and 88 Mb with a mean raw read length was 249 and 258 bp for *S. psittacina* and *S. purpurea*, respectively. Raw sequences were cleaned and assembled into contigs. Since the *Sarracenia* genome sequence was not available, a de novo assembly of the cDNA sequences was performed. We obtained 46,275 and 36,681 contigs with an N50 value of 479 and 485 for *S. psittacina and S. purpurea*, respectively. The number of

assembled contigs and the mean average coverage per contig were found to be correlated with the number of cleaned reads. The complete assembly statistics are shown in Table 4.1 and assemblies are available under supplementary data file1.

4.4.2 Functional annotation of contigs:

We used B2G to functionally annotate the contigs. The B2G annotation has 3 steps which involve using blast against the public or private databases, mapping against GO resources and annotating to generate reliable functional assignments. From our data, 20,920 (45.2 %) of the *S. psittacina* and 17,821 (48.6%) *S. purpurea* cDNA sequences were shown to have significant matches to currently known proteins in the NCBI nonredundant protein database. The B2G blast hit bar plot (Fig. 4.1) shows *Vitis*, *Ricinus* and *Populus* as the top 3 species with greatest number of hits for both species. Contigs with the significant blast matches were functionally annotated. We found GO resource assignments for 19.92% and 21.19% of contigs for *S. psittacina* and *S. purpurea*, respectively. A summary of B2G mapping is given in the Table 4.2.

The first major GO division, 'biological process', associates contigs to the biological objective to which it contributes (The Gene Ontology Consortium (2000)). Within it, 11 major categories were identified and found to be similarly distributed in both species. The two most abundant categories were: (i) 'cellular and metabolic processes', to which 75% of both species' contigs were associated (*S. psittacina:* 7543 sequences and *S. purpurea*: 6285 sequences) and (ii) 'biological regulation', to which 8% of contigs were dedicated (*S. psittacina*: 802 sequences and *S. purpurea*: 684 sequences) (Fig. 4.2 A).

The second major GO division, 'molecular function', links genes to their biochemical activity (The Gene Ontology Consortium (2000)). The contig coverage was again found to be similar for both species. Most of the contigs in the molecular function division were dedicated to binding functions and catalytic activity (82% of both species; *S. psittacina*: 7367 sequences and *S. purpurea*: 6170 sequences) (Fig. 4.2 B).

The last GO division is 'cellular component', which refers to sub-cellular location where gene product is active (The Gene Ontology Consortium (2000)). In this 8 major categories were identified and again a similar type of coverage was found for both species. Gene products were mainly expressed intracellularly (52% for both species; *S. psittacina*: 3033 sequences; *S. purpurea*: 2562 sequences) or in the membrane bound/non-membrane bound organelle (29% for both species; *S. psittacina*: 1730 sequences; *S. purpurea*: 1386) (Fig. 4.2 C). The complete B2G results are shown in the supplementary excel sheet S1& S2.

4.4.3 Localization of the contigs with respect to genic features:

We used the *Vitis* genome as a reference in order to locate the contigs with respect to genomic sequence. *Vitis* was the species whose sequences were found as the top hit for sequences from both *Sarracenia* species in the blast (blastx) search. Blat was used to map the assemblies against the *Vitis* genome. Sequences were uniquely mapped to a particular feature (5'UTR, 3'UTR, CDS, 1Kb up, 1Kb down and intergenic) plotted as a bar plot (Fig. 4.3). 8501 and 7224 contigs were uniquely mapped to features under consideration for *S. psittacina* and *S. purpurea*, respectively. Based on the *Vitis* genome annotation, nearly 60 % of the mapped contigs were found to be in CDS regions and 33% of the contigs were associated with putative intergenic

region (shown as NA in figure). Only a very tiny fraction of the contigs were mapped to 5'UTR/3'UTR and regions 1Kb upstream/1Kb downstream of them.

4.4.4 Orthologous and paralogous contigs:

We identified 16,163 pairs of orthologous contigs by the reciprocal best blast hit approach; this approach has been found superior to more sophisticated orthology detection algorithms (Altenhoff and Dessimoz 2009). We used a stringent e-value cutoff (1e-20) in order to separate the paralogous sequences from the orthologous sequences. A total of 8,706 orthologous contigs matched to ORFs of known or putative proteins. The Venn diagram indicating the orthologous and unique contigs is shown in figure 4.4. We also identified 18,296 and 9,565 paralogs pairs by all vs all blast (blastn); these were organized into 2,555 and 1,708 gene families by a single linkage clustering method for *S. psittacina* and *S. purpurea*, respectively.

4.4.5 Estimation of K_a/K_s:

We calculated the K_a and K_s values for 10,715 orthologous contigs and their ratio for 4,810 contigs (a $Ks = 0$ made the ratio incalculable for some of the contigs). Similarly, we estimated the *Ks* value for each of the paralogous pairs. We further studied substitution rates and GO categorization within the orthologous contigs only. Out of these 4,810 contigs, we were able to find the functional annotation for 3,444 contigs. We identified 491 contigs which are under diversifying selection $K_a/K_s \geq 1$ (Fig. 4.5). Most of these contigs were found to be involved in the molecular function related to nucleotide binding. The complete K_a/K_s result for 3,444 contigs is shown in the supplementary excel sheet S3.

4.4.6 Genome duplication and speciation:

From the estimation of synonymous substitution rates, we were able to find signatures of genome duplication within both *S. psittacina* and *S. purpurea*. The K_s value distribution for both species is shown in the figure 4.6 A. The secondary peak in the paralogous K_s distribution plot indicates a genome duplication event (Blanc and Wolfe 2004). Estimated K_s value between orthologous pairs were also plotted along with paralogous (Fig. 4.6 B). The secondary peak in the orthologous K_s value distribution indicates the speciation events. The number of sequences involved in the genome duplication events and gene family statistics are shown in the Table 3. Considering a clock-like rates of synonymous substitution of 1.5×10^{-8} substitutions/synonymous site/year for dicots (Koch, Haubold et al. 2000), we estimated the age of these events and found that they are fairly recent (approx 2 million years (myr) for duplication and speciation). The purpose of these estimates is just to give an idea of the time scale involved because they are certainly highly approximate.

4.5 DISCUSSION

We generated the 46275 and 36681 contigs by pyrosequencing for *S. psittacina* and *S. purpurea*, respectively using young pitchers as starting material. We normalized our cDNA library in order to maximize coverage of transcripts and prevent biases due to highly expressed transcripts. Based upon the sequence similarity searches, we found *Vitis* to have the greatest number of hits to contigs from both species. This is an interesting result since *Sarracenia* is currently considered to belong to the order Ericales which is a part of clade asterids, whereas *Vitis* belongs to clade rosids (http://www.mobot.org/mobot/research/apweb/). The expectation might have been that an extensively sequenced asterid such as *Lycopersicon* would have had the greatest number of blast

hits, rather than a rosid. To test the statistical significance of this result, we divided the species with the top 20 number of blast (blastx) hits into 2 parts based upon their phylogenetic position in asterids and rosids and performed a t-test of significance (two-sample t-test for unequal variances, one-tail t-statistics $t(6) = 1.943180274$, $p \sim 0.05$ for the *S. psittacina* sequences, with a similar result for *S. purpurea* sequences). The t-test shows that observed difference is significant. Relationships among familial clades in the order Ericales have been considered as problemeatic (Judd and Olmstead 2004) and Ericales was placed in the clade dilleniidae in the previous classifications (Soltis and Soltis 2004). Our results are thus consistent with the original placement of the Ericales, as being closer to rosids than asterids.

Since a genomic sequence is not available for any pitcher plants, we used the *Vitis* genome as a reference for contig localization. About 60% of contigs aligned uniquely to the protein coding regions of *Vitis* genome in both species. The nearly equal coverage in 5' UTR and 3'UTR regions showed the success of the protocol in full length cDNA construction. We also found a number of contigs belong to intergenic regions based upon the *Vitis* genome annotation. As the *Vitis* genome annotation is still not finished and many more genes have yet to be identified; it is possible that contigs currently localized in intergenic regions might be as yet unidentified protein coding genes, or they might be non-coding RNA genes or alternatively spliced exons. To test this, we took the contigs belonging to intergenic regions and checked whether they had similarity to any known or hypothetical protein. Out of 1773 putative intergenic sequences of *S. psittacina* , 1280 sequences had shown similarity ($\geq$ 1e-5) to known/hypothetical proteins. We also obtained ncRNA sequences for *Arabidopsis* and *Oryza* from ncRNA database (Mituyama, Yamada et al. 2009) and mapped our intergenic sequences against them. Only few sequences had showed

similarity to known ncRNA. Similar patterns were obtained from *S. purpurea*. The detailed mapping results are shown in the supplementary excel sheet S4.

The paired pattern in functional annotation for all three GO divisions reflects that our library and 454 sequencing covered the transcriptome of the two species equally well. We can speculate about the significance of the GO annotations relative to the pitcher plant insectivorous adaptations. In the biological process and molecular function division, an abundance of genes were found to be related with metabolic process and catalytic activity and inside the metabolic process and catalytic activity a number of genes were found related to macromolecule metabolic process and hydrolase activity (supplementary figure 1A, B and 2A, B). The hydrolytic enzymes (protease, RNase, nuclease, phosphatase) may be required for the digestion of prey(Gallie and Chang 1997). The pitchers of pitcher plants may contain water with microorganisms; a high level of hydrolytic enzymes in the pitcher plant transcriptome may favor the hypothesis that pitcher plants do not rely solely on microorganisms to digest their insect prey(Rosa, Malek et al. 2009). More detailed information about *Sarracenia*, its prey-digestion system, and its microbial food web, can be found in a review by Ellison et al. (Ellison, Gotelli et al. 2003)

Previously we estimated the genome sizes (25 % more nuclear DNA than maize) of these two species (Rogers, Cruse-Sanders et al. 2010) and based on that we expected that the species might be polyploid. To test our hypothesis, we estimated of substitution rate for the all the paralogous contigs showing significant sequence similarity at protein level and plotted the K_s value distribution histogram. We were able to detect a moderate signature of a duplication event within this dataset. Time estimates suggest that these events were fairly recent. The lack of clear

secondary peaks can be attributed to the fact that signal of this event, provided it is relatively recent, is not dissociable from the initial peak. The orthologous contig comparison secondary peak is in the same time frame as the ortholog/paralog duplication peak. Possibly speciation within the genus *Sarracenia* might have occurred after the duplication event, which is a well-recognized pattern of plant evolution(Wood, Takebayashi et al. 2009).

A substitution rate within the orthologous contigs (identified by reciprocal blast) found 491 contigs have K_a/K_s ratio ≥ 1. This ratio is considered to be a good measure of selective pressure acting at the sequence level (Yang and Bielawski 2000; Bustamante, Fledel-Alon et al. 2005) and has been used in different studies to identify the contigs under positive/adaptive evolution (K_a/K_s >1) or under negative/purifying selection (K_a/K_s <1) (Hurst 2009). The majority of contigs under diversifying selection were found to be related to binding activity in the molecular function category of GO assignment (supplementary excel sheet S3).

In summary, our analysis showed a high degree of similarity in sequence existed between the pitcher plants, with *Vitis* as the model species outside the genus with the greatest sequence similarity. Functional annotation of all the contigs identified the major categories of genes, and substitution rate estimation identified the signatures of genome duplication and rapidly evolving genes in the pitcher plants. We believe that these sequences and analysis will greatly aid the research community working on insectivorous plants and their ecology.

4.6 AVAILAIBILITY

The data from the experiments described in this work are available from the NCBI Sequence Read Archive at http://www.ncbi.nlm.nih.gov/sra under the accession SRP006675 and SRP006677.

4.7 ACKNOWLEDGEMENTS

We gratefully acknowledge support from NSF grant IIS 0916250 to Liming Cai, and from the University of Georgia Franklin College of Arts & Science's research fund.

4.8 REFERENCES

(The Gene Ontology Consortium (2000)). "Gene Ontology: tool for the unification of biology." Nature Genetics **25**(1): 25-29.

Altenhoff, A. M. and C. Dessimoz (2009). "Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods." Plos Computational Biology **5**(1): -.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Research **25**(17): 3389-3402.

Blanc, G. and K. H. Wolfe (2004). "Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes." Plant Cell **16**(7): 1667-1678.

Bouck, A. and T. Vision (2007). "The molecular ecologist's guide to expressed sequence tags." Molecular Ecology **16**(5): 907-924.

Bustamante, C. D., A. Fledel-Alon, et al. (2005). "Natural selection on protein-coding genes in the human genome." Nature **437**(7062): 1153-1157.

Chevreux, B., T. Pfisterer, et al. (2004). "Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs." Genome Research **14**(6): 1147-1159.

Conesa, A., S. Götz, et al. (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." Bioinformatics **21**(18): 3674-3676.

Darwin, C. (1875). " Insectivorous Plants. Murray, London.".

Ellegren, H. and B. C. Sheldon (2008). "Genetic basis of fitness differences in natural populations." Nature **452**(7184): 169-175.

Ellison, A. M., N. J. Gotelli, et al. (2003). "The evolutionary ecology of carnivorous plants." Advances in Ecological Research, Vol 33 **33**: 1-74.

Gallie, D. R. and S. C. Chang (1997). "Signal transduction in the carnivorous plant Sarracenia purpurea - Regulation of secretory hydrolase expression during development and in response to resources." Plant Physiology **115**: 1461-1471.

Gallie, D. R. and S. C. Chang (1997). "Signal transduction in the carnivorous plant Sarracenia purpurea. Regulation of secretory hydrolase expression during development and in response to resources." Plant Physiol **115**(4): 1461-1471.

Gotelli, N. J. and A. M. Ellison (2006). "Food-web models predict species abundances in response to habitat change." Plos Biology **4**(10): 1869-1873.

Hepburn, J. S., E. Q. S. John, et al. (1920). "The absorption of nutrients and allied phenomena in the pitchers of the Sarraceniaceae,." J.Franklin Inst **189**: 147-184.

Hudson, M. E. (2008). "Sequencing breakthroughs for genomic ecology and evolutionary biology." Molecular Ecology Resources **8**(1): 3-17.

Hurst, L. D. (2009). "Evolutionary genomics and the reach of selection." J Biol **8**(2): 12.

Judd, S. W. and G. R. Olmstead (2004). "A SURVEY OF TRICOLPATE (EUDICOT) PHYLOGENETIC RELATIONSHIPS." American Journal of Botany **91(10)**: 1627–1644.

Kent, W. J. (2002). "BLAT - The BLAST-like alignment tool." Genome Research **12**(4): 656-664.

Koch, M. A., B. Haubold, et al. (2000). "Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae)." Molecular Biology and Evolution **17**(10): 1483-1498.

Larkin, M. A., G. Blackshields, et al. (2007). "Clustal W and clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

Mituyama, T., K. Yamada, et al. (2009). "The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs." Nucleic Acids Res **37**(Database issue): D89-92.

Ne'eman, G., R. Ne'eman, et al. (2006). "Limits to reproductive success of Sarracenia purpurea (Sarraceniaceae)." American Journal of Botany **93**(11): 1660-1666.

Quinlan, A. R., D. A. Stewart, et al. (2008). "Pyrobayes: an improved base caller for SNP discovery in pyrosequences." Nature Methods **5**(2): 179-181.

Rogers, L. W., M. J. Cruse-Sanders, et al. (2010). "Development and characterization of microsatellite markers in Sarracenia L. (pitcher plant) species." Conservation Genet Resour(75-79).

Rokas, A. and P. Abbot (2009). "Harnessing genomics for evolutionary insights." Trends in Ecology & Evolution **24**(4): 192-200.

Rosa, A. B., L. Malek, et al. (2009). "The development of the pitcher plant Sarracenia purpurea into a potentially valuable recombinant protein production system." Biotechnology and Molecular Biology Reviews **3**(5): 105-110.

Schnell, D. E. (1983). "Notes on the Pollination of Sarracenia-Flava L (Sarraceniaceae) in the Piedmont Province of North-Carolina." Rhodora **85**(844): 405-420.

Soltis, S. P. and E. D. Soltis (2004). "THE ORIGIN AND DIVERSIFICATION OF ANGIOSPERMS." American Journal of Botany **91(10)**: 1614–1626.

Suyama, M., D. Torrents, et al. (2006). "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments." Nucleic Acids Research **34**: W609-W612.

Wood, T. E., N. Takebayashi, et al. (2009). "The frequency of polyploid speciation in vascular plants." Proceedings of the National Academy of Sciences of the United States of America **106**(33): 13875-13879.

Yang, Z. (2007). "PAML 4: Phylogenetic analysis by maximum likelihood." Molecular Biology and Evolution **24**(8): 1586-1591.

Yang, Z. and J. P. Bielawski (2000). "Statistical methods for detecting molecular adaptation." Trends in Ecology & Evolution **15**(12): 496-503.

Tables:

Table 4.1: Assembly statistics for two *Sarracenia* species:

Species	Ind *	Plate†	Cleaned reads	Cleaned Bases (MB)	Contigs	Singletons	Mean avg coverage per contig	Mean GC per contig (%)
S. psittacina	8	1/2	392346	102	46275	587	3.40	40.84
S. purpurea	4	1/2	282150	75	36681	234	3.05	41.14

*Number of individual pooled prior to sequencing

†One plate represents a full Roche 454 run.

Table 4.2: Summary statistics for two *Sarracenia* species of Blast2GO assignment:

Species	Number of contigs	Number of blast hits	Number of GO mapped	Number of GO terms*	Number of functionally annotated
S. psittacina	46275	20920	9222	39500	7208
S. purpurea	36681	17821	7773	33093	6060

*Can be multiple per contigs

Table 4.3: Number of sequences and paralogs found for *S. psittacina* and *S. purpurea*:

Species	Number of contigs	Number of paralogs pairs	Percentage of paralogs	Gene families	Gene family size	Duplication event with mean *Ks*
S. psittacina	46275	18296	39.53	2555	3.53	2635
S. purpurea	36681	9565	26.07	1708	3.46	1417

Figures:

Figure 4.1: A bar plot showing the hits (blastx top hit) to previously sequenced species (displaying only top five species) for *S. psittacina and S. purpurea* contigs.

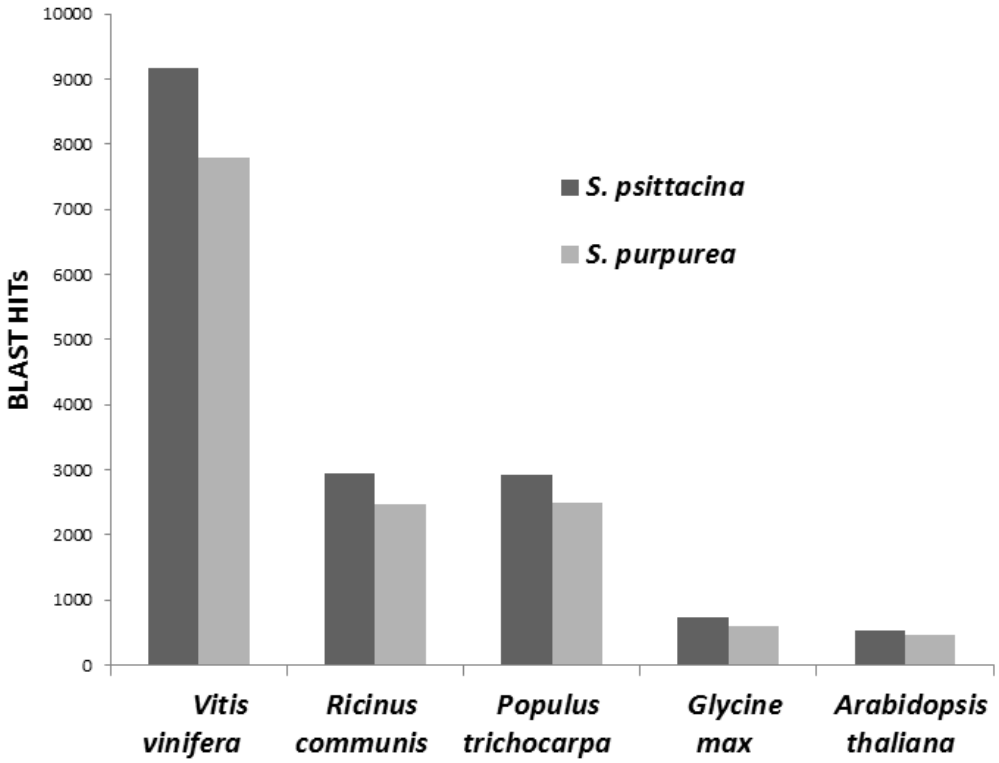
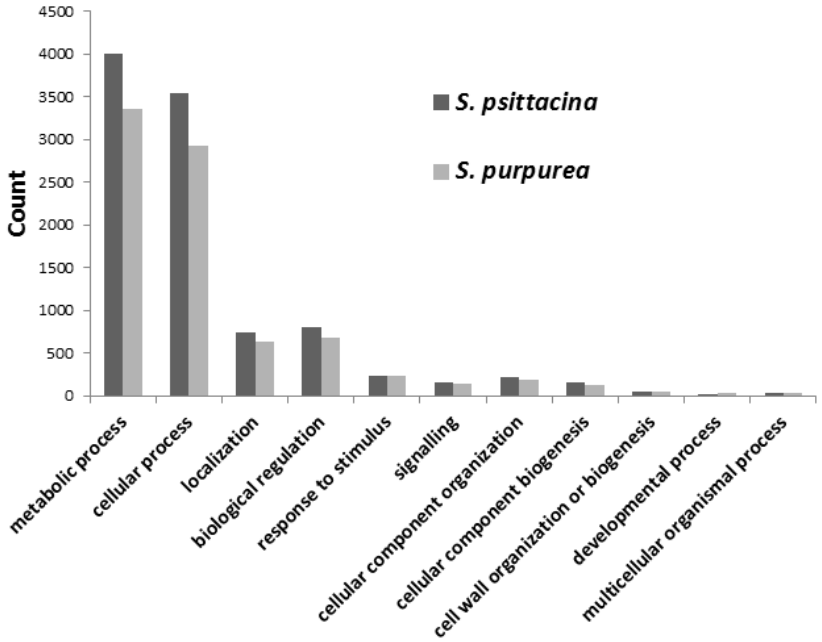
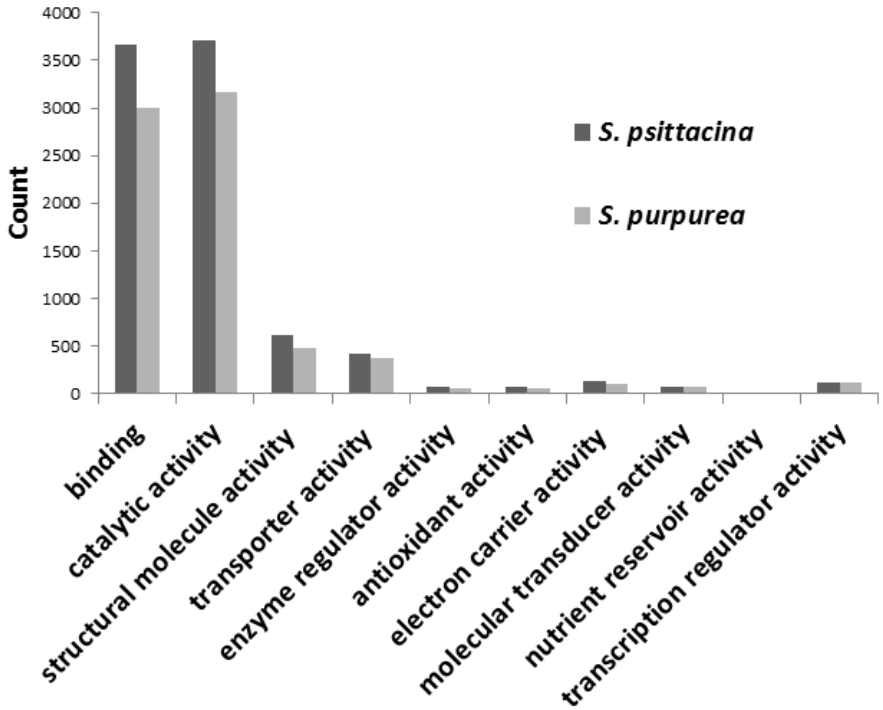


Figure 4.2: A bar plot showing the Blast2GO functional assignments in three GO categories A) Biological process B) Molecular function C) Cellular component for *S. psittacina* and *S. purpurea* contigs.




Count
4000
3500
3000
2500
2000
1500
1000
500
0
S. psittacina
S. purpurea
binding
catalytic activity
structural molecule activity
transporter activity
enzyme regulator activity
antioxidant activity
electron carrier activity
molecular transducer activity
nutrient reservoir activity
transcription regulator activity

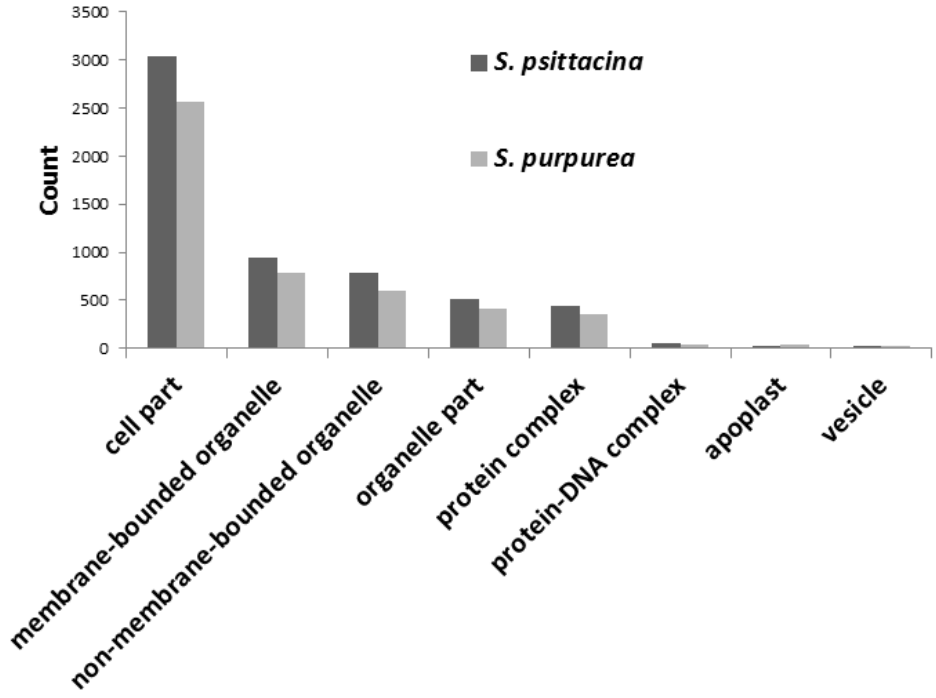

Count
3500
3000
2500
2000
1500
1000
500
0
S. psittacina
S. purpurea
cell part
membrane-bounded organelle
non-membrane-bounded organelle
organelle part
protein complex
protein-DNA complex
apoplast
vesicle

Figure 4.3: A bar plot displaying the proportion of contigs mapped to a particular region of *Vitis* genome in two species

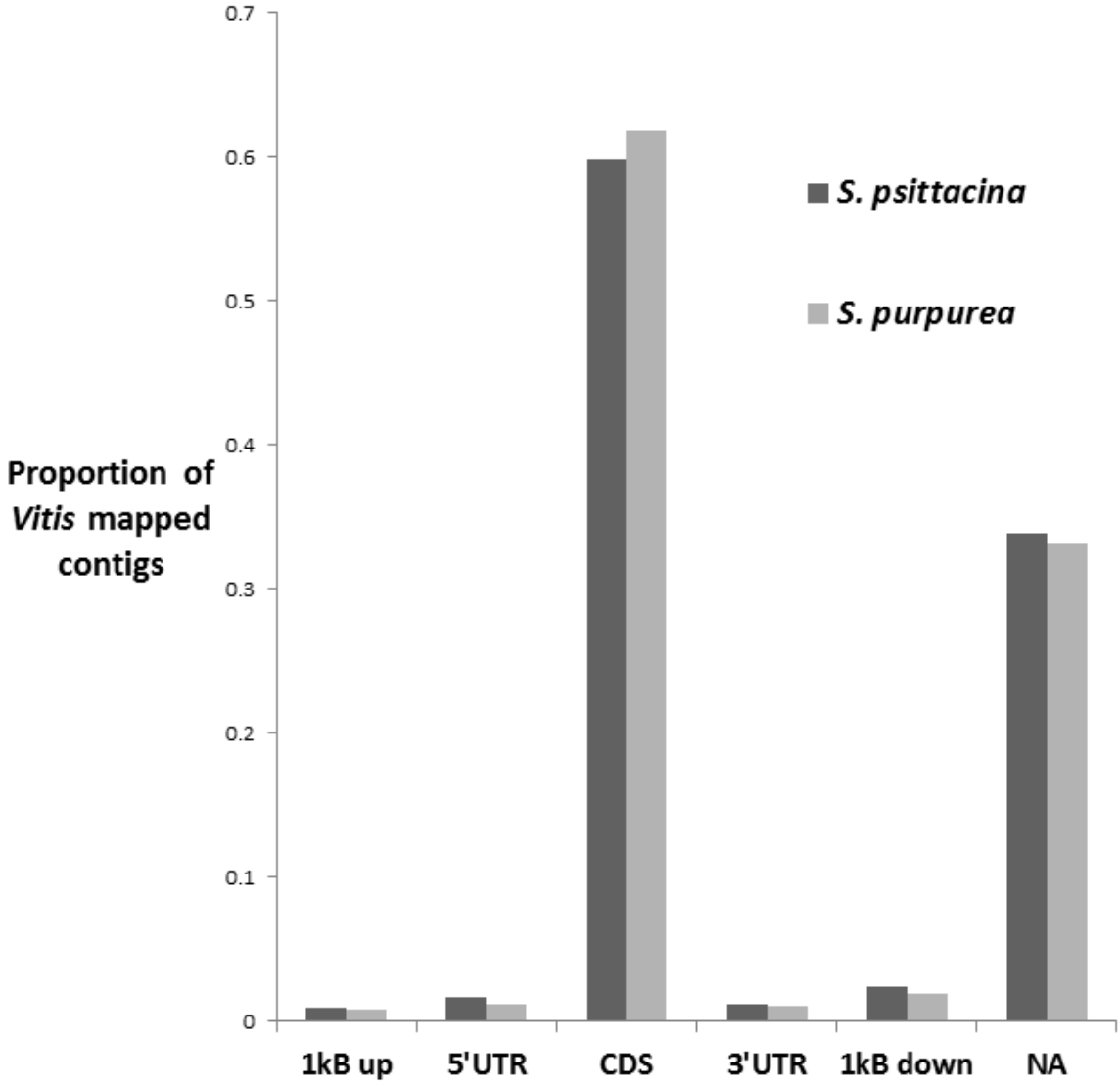


Figure 4.4: A Venn diagram showing the count of orthologous and unique contigs between two species.

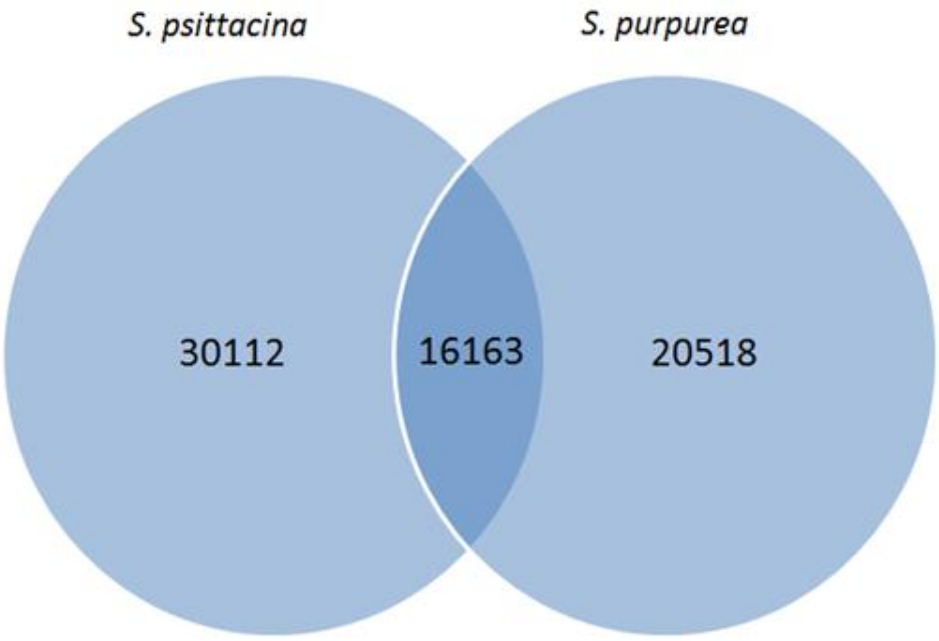


Figure 4.5: A scatter plot of K_a/K_s ratio for 491 orthologous contigs under diversifying selection. Contigs with $K_a/K_s > 1$ fall above the light grey line and greater than 2 values fall above the black line.

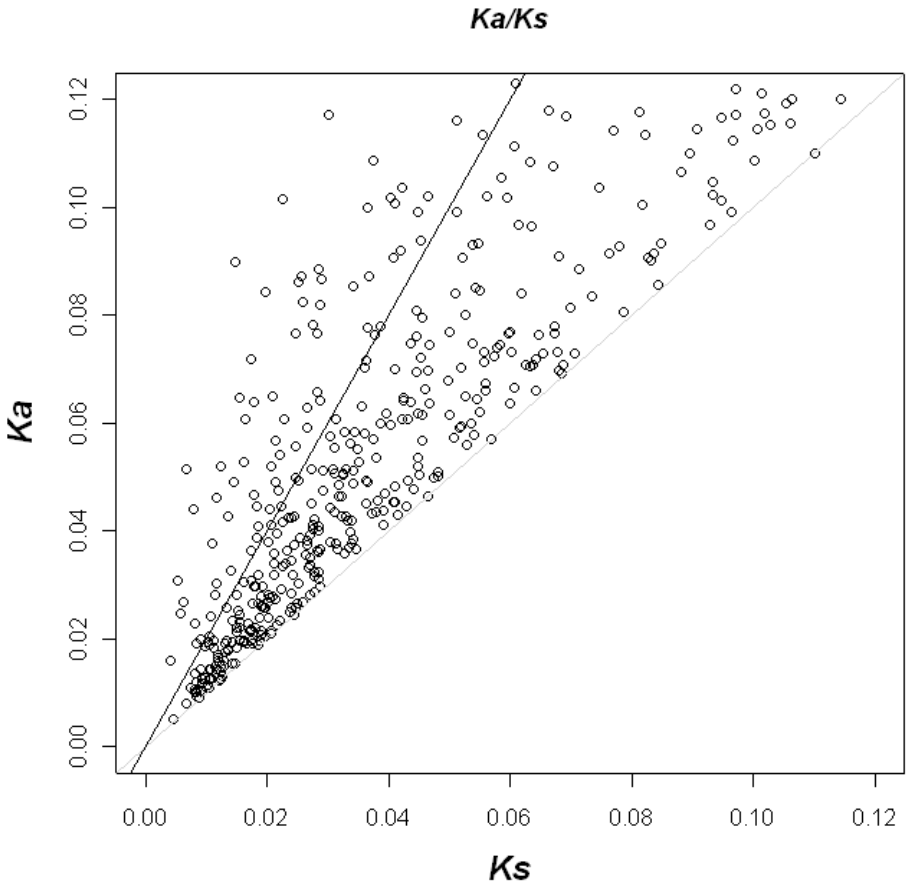
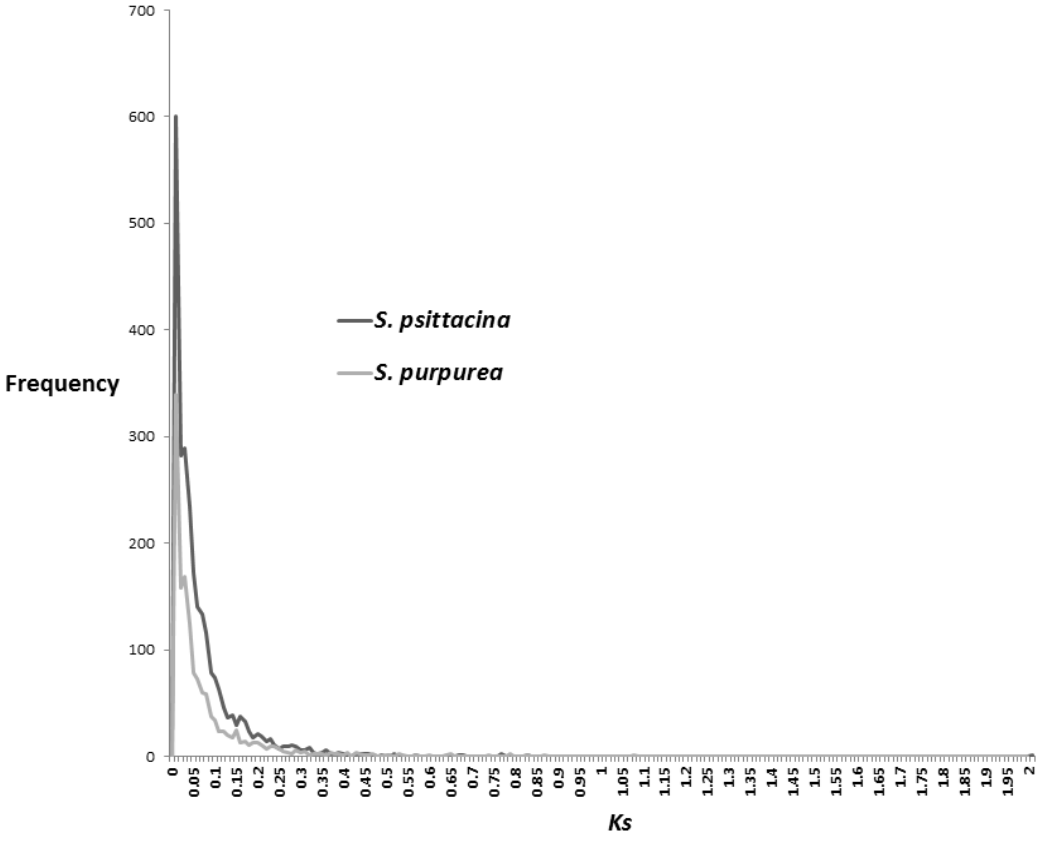
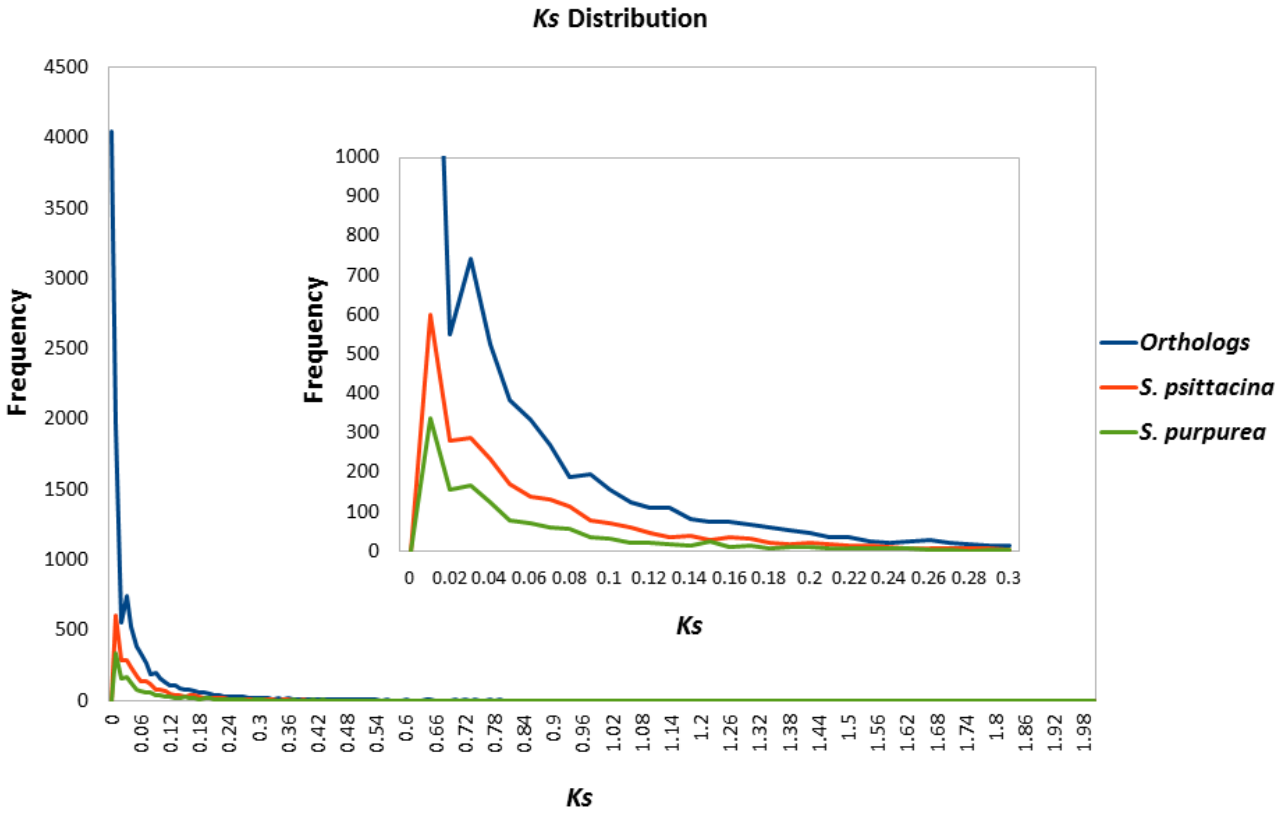


Figure 4.6: K_s value distribution between the two *Sarracenia* species for the identification of the genome duplication event and the speciation event.

A) The secondary peak (*Ks* = 0.03) in the paralogous *Ks* distribution indicates the genome duplication. B) The secondary peak (*Ks* = 0.03) in the orthologous *Ks* distribution give the indication of speciation event.




Ks Distribution
Frequency
Ks
Orthologs
S. psittacina
S. purpurea

4.9 SUPPLEMENTARY MATERIAL

Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

CHAPTER 5

PROBING THE GENOMICS OF ADAPTIVE DIVERGENCE USING COMPARATIVE TRANSCRIPTOMICS BETWEEN TWO SUBSPECIES OF SONGBIRD[1]

[1]Kevin Winker*, Anuj Srivastava*, Timothy I. Shaw, Kenneth L. Jones, and Travis C. Glenn, To be submitted to DNA Research.

5.1 ABSTRACT

Plant and animal subspecies are commonly described based on phenotypic differences between populations—differences widely presumed to have arisen through local adaptation. Island populations are especially prone to having distinctive phenotypes. We examined two subspecies of songbird (*Melospiza melodia*) that differ phenotypically in multiple characteristics that are commonly different between diverging bird populations: body size, plumage coloration, and seasonal migratory behavior. We sampled two growing individuals from each of three populations, two island populations of *M. m. maxima* (from Attu and Adak islands, Alaska) and one mainland population of *M. m. caurina* (from Cordova, Alaska). Whole-body mRNA was normalized and sequenced via 454 technology, yielding more than 100,000 reads per population (110 Mb total). Cleaned sequences were assembled into 38,539 contigs (with N50 = 482 bp), 4,574 of which were orthologous to both the zebra finch (*Taeniopygia guttata*) and chicken (*Gallus gallus*) genomes and 3,680 of which are functionally annotated. We detected 29,982 SNPs/indels, 1,402 of which were fixed between populations and subspecies; of which 392 and 410 SNP/indel were present between and within subspecies, respectively. Additional work is needed to identify the specific SNPs/indels associated with phenotypic effects, but our approach is widely applicable for efficient discovery of candidate loci, and those we identify form the basis for further hypothesis testing of differentiation among songbirds.

5.2 INTRODUCTION

The early stages of speciation occur between populations within a species. Phenotypic differences develop between diverging populations, and taxonomists have commonly described plant and animal subspecies based on such differences. Given substantial evidence, the majority of these differences are widely presumed to arise through local adaptation (Mayr 1982, Winker 2010). Determining the genetic underpinnings for such population divergences will provide insight into the speciation process, and next-generation sequencing (NGS) is propelling such studies in non-model organisms (Nadeau & Jiggins 2010). In this respect, the relative frequencies of protein-coding versus noncoding regulatory DNA polymorphisms during adaptive divergence become relevant. Although both types of polymorphisms are important in different contexts, there is growing evidence that *cis*-regulatory genetic changes dominate interspecific comparisons, whereas protein-coding changes dominate intraspecific comparisons (Hoekstra & Coyne 2007, Carroll 2008, Haygood *et al.* 2010). This situation is at odds with traditional evolutionary theory, which suggests that species-level divergences should reflect what occurs at initial stages of differentiation (Stern & Orgogozo 2008). Nevertheless, the importance of protein-coding changes suggests that focusing on transcriptome variation within and between subspecies is likely to yield novel candidate loci for studies of intraspecific adaptive divergence.

Here we examine two subspecies of songbird that differ phenotypically in multiple ways in characteristics that are commonly different between diverging bird populations: body size, plumage coloration, and seasonal migratory behavior. Island populations are especially prone to having distinctive phenotypes; one of these is island gigantism. Enlarged body size in island populations is found among many small vertebrate species (Grant 1968, 1998; Lomolino 1985,

2005). For example, in southwestern Alaska, Aleutian Island populations of song sparrow (*Melospiza melodia*) show a pronounced increase in size over mainland populations. Based on body mass (a commonly used measure of size), individuals from the Aleutian Islands are much larger than those from mainland Alaska (Pruett & Winker 2005, 2010; Patten & Pruett 2009). Most relevant here, the Aleutian subspecies *M. m. maxima* ($\bar{x}$ = 45.7 g, SE = 0.38) is about 1.6 times larger than the mainland subspecies *M. m. caurina* ($\bar{x}$ = 28.4 g, SE = 0.45; $P < 0.001$). In this study we used specimens of *M. m. maxima* from Attu and Adak islands and of *M. m. caurina* from the mainland at Cordova (Fig. 5.1).

Phenotypic attributes such as body size can be influenced by developmental (i.e., non-genetic) processes, but body size in vertebrates with determinate growth is generally considered to be both heritable and a polygenic trait; for example, there are at least 44 loci known to affect variation in human height (West-Eberhard 2003, Weedon and Frayling 2008). Body size is a dramatic phenotypic difference between these two song sparrow subspecies, but several others exist, including plumage coloration and a loss of seasonal migratory behavior in *maxima*. The differences between *caurina* and *maxima* in migration are also likely to be influenced by many loci, because migratory traits are both multiple (e.g., fattening, timing of onset, and direction and distance travelled) and polygenic (Pulido 2007). Thus, there are many potential candidate loci and signals of adaptive divergence between the genomes of these two song sparrow subspecies. The suite of divergent characteristics exhibited between these two subspecies likely represents adaptive divergence that has accrued from standing genetic variation since the last glacial maximum, or about 10-20 Kyr (Pruett and Winker 2005, 2010). However, the genetic underpinnings of these traits will be among genomic characteristics that have been influenced by

genetic drift, especially through the founder events and bottlenecks that have affected the subspecies *maxima* (Pruett and Winker 2005).

Our goals in this study were to 1) obtain a functional annotation of a substantial portion of the song sparrow transcriptome; 2) compare transcript divergence between the song sparrow and the two bird genomes sequenced and assembled to the highest quality thus far, zebra finch (*Taeniopygia guttata*) and chicken (*Gallus gallus*); and 3) obtain a set of SNP/indel markers to identify candidate genes associated with phenotypic divergence in this species. Achieving these goals will establish important baseline data for a non-model organism in a speciose group (passerines, or songbirds) that is frequently studied by molecular ecologists. Moreover, it should also lead to the identification of some of the genes and genetic variation correlated with phenotypic traits such as body size, plumage coloration, and seasonal migration that are commonly different between subspecies and species of songbirds.

5.3 MATERIAL AND METHODS

5.3.1 Samples, cDNA library, and sequencing

Two song sparrows still undergoing growth (from egg to just-fledged) were sampled from each of three Alaska populations: two island populations of *M. m. maxima* (from Attu and Adak islands; an egg and a very young nestling from Attu Island, unvouchered, and vouchers UAM 27831& 27832 from Adak Island; Attu [MID 18] samples were unvouchered, and Adak [MID 19], and Cordova [MID 13] had two pooled vouchers each) and one mainland population of *M. m. caurina* (from Cordova, vouchers UAM 27829 & 27830). The egg was homogenized,

whereas from the others six tissues (brain, liver, heart, muscle, bone, and pancreas) were taken, minced, and placed in RNAlater (Qiagen, Valencia, CA) within minutes of death and then frozen. In the lab, tissues were homogenized and total RNA was isolated using Trizol (Invitrogen, Carlsbad, CA) and subsequently cleaned using a Qiagen RNeasy column. Equal amounts of RNA from individuals of each population were pooled and a MINT universal cDNA kit (Evrogen, Moscow, Russia) with primers modified specifically for 454 (Beldade *et al.* 2006) was used to create cDNA libraries enriched for full-length transcripts. We then normalized the cDNA libraries using the TRIMMER cDNA normalization kit (Evrogen) to substantially decrease the relative abundance of common transcripts. The normalized cDNA was fragmented and prepared for sequencing using standard 454 procedures, including independent molecular identifiers (MID tags) for each of the three populations (individuals from the same population received the same tag). As each library contained a unique MID tag, libraries were pooled and sequenced as a single sample. Sequencing was performed at the University of Georgia's Georgia Genomics Facility on a Roche 454 FLX using Titanium chemistry.

5.3.2 Assembly; polymorphism and ortholog identification

Bases were called from the 454-generated sff file using Pyrobayes (Quinlan *et al.* 2008), which provides improved accuracy in the estimation of base qualities for pyrosequences. We removed MINT primer sequences and other contaminants using SeqClean (http://compbio.dfci.harvard.edu), and reads from all three populations were combined. We performed a combined assembly of reads using MIRA (Chevreux *et al.* 2004), and then used GigaBayes (Marth *et al.* 1999), a short-read SNP and short indel discovery program, to detect polymorphisms. To make the SNP/indel predictions more reliable, we used the more stringent

criteria that the minor allele must occur at least 3 times and be present at ≥10% relative to the major allele frequency. We identified orthologous contigs (against the zebra finch and chicken genomes) using the reciprocal blast approach, because it has been found to be superior to sophisticated orthology detection algorithms (Altenhoff *et al.* 2009). A stringent cutoff of 1e-20 was used to separate paralogs from orthologs. The cDNA sequences from zebra finch (taeGut3.2.4.60.cdna.all.fa) and chicken (WASHUC2.60.cdna.all.fa) were obtained from the Biomart database (www.biomart.org). Although the zebra finch is a passerine and thus more closely related to the song sparrow, the chicken database contains sequences from whole growing chicks, whereas that of the zebra finch emphasizes neural transcripts.

To identify likely genomic positions of the song sparrow contigs, we mapped them against genomic sequences of zebra finch (taeGut3.2.4.60.dna_rm.toplevel.fa) and chicken (WASHUC2.60.dna_rm.toplevel.fa) using BLAT (Kent 2002) with default criteria. We obtained feature information for protein-coding genes and ncRNA using the Ensemble (http://uswest.ensembl.org/index.html/) Xenoref and gtf files, respectively.

5.3.3 Functional annotation of contigs

We used Blast2GO (B2G; Conesa *et al.* 2005) to functionally annotate the contigs. A combined graph was generated for each GO category. For the molecular function division, a graph was obtained using default criteria, and for the other two divisions (cellular component and biological process), seq/node filter values were changed to 4/10, to prevent overloading the graphs.

5.3.4 Estimation of substitution rates

Substitution rates were estimated for contigs that were orthologous to both zebra finch and chicken. Reading frames for these contigs were identified using BLASTX (Alschul *et al.* 1997) against protein sequences of zebra finch (taeGut3.2.4.60.pep.all.fa) and chicken (WASHUC2.60.pep.all.fa) obtained from Biomart (www.biomart.org). Sequences that produced significant alignments were extracted, translated, and aligned using CLUSTALW (Larkin *et al.* 2007). Sequences that contained frame shifts were excluded from analysis. Corresponding codon alignments were produced using PAL2NAL (Suyama *et al.* 2006), and, finally, rates were estimated using a maximum likelihood method implemented in the CODEML program of the PAML package Version 4.1 (Yang 2007). Pairwise maximum likelihood analyses were performed in runmode-2. The estimated rates of non-synonymous to synonymous substitutions (K_a/K_s values) were plotted as a scatter plot in the range of 0-2.0.

5.4 RESULTS

5.4.1 Sequence assembly and SNP detection

The pooled reads from all three populations yielded 131 Mb (458,808 sequences) of raw data, which was reduced to 110 Mb (381,474 sequences) through cleaning (Table 5.1). The mean raw and cleaned read lengths were 286 and 290 bp, respectively. Poor-quality reads were often very short and were purged entirely prior to assembly. Without a reference genome for the song sparrow, *de novo* assembly was required. Cleaned sequences were assembled into 38,539 contigs with N50 and N90 values of 482 and 317bp, respectively. There were 1,417 singletons. Mean coverage per contig was 3.93 X, and mean GC content per contig was 43.6%.

We detected a total of 29,982 SNPs/indels that were spread relatively evenly within, between, and among all three populations (Fig. 5.2, Supplementary Information). A total of 1,402 SNPs/indels were fixed between populations and subspecies (Fig. 5.3; the sum of all pairwise comparisons is 1,635 because some pairwise SNPs are found in more than one pair). This provides many SNPs/indels for further study (Supplementary Information), although given our limited sampling of individuals within populations ($N = 2$) many will not be true fixed differences (i.e., they are false positives).

5.4.2 Orthology with zebra finch and chicken

The reciprocal blast approach identified 4,574 contigs as orthologous to both zebra finch and chicken. As expected because of phylogenetic relationships, more contigs were identified as orthologous to the zebra finch than the chicken: the set (unique song sparrow [orthologs] unique zebra finch) was (32,435 [6,104] 12,493), whereas the set (unique song sparrow [orthologs] unique chicken) was (32,767 [5,772] 16,518). A substantial number of orthologous contigs (3,894) were found to have the same chromosome location in the zebra finch and chicken (Supplementary Information).

5.4.3 Functional annotation

B2G, which we used to functionally annotate the contigs, has three annotation steps involving 1) a blast against databases, 2) mapping against GO resources, and 3) annotation to generate reliable functional assignments. In our data, 12,880 of the contigs (33.46 % overall, of which 8,540 were unique hits) had significant matches to currently known proteins in the NCBI nonredundant protein database. Zebra finch and chicken were identified as the top two species

with the best blast hits for our song sparrow contigs (Fig. S1). Contigs with significant blast matches were functionally annotated. GO resource assignment was found for 3,949 (10.2 %) of the total contigs (with 24,363 GO terms; there can be multiple terms per contig), of which 3,367 (8.7% of all contigs) were functionally annotated (Supplementary Information).

In the first GO division, 'biological process', 22 categories were identified. Most contigs (3,578 = 53.1 %) were involved in 'cellular and metabolic processes'. The second most abundant category was 'biological regulation and localization' (1,253 = 18.6%; Fig. S1). Within the second division, 'molecular function', 9 major categories were identified. Most of the contigs were functionally related to 'nucleotide binding' (1,966 = 43.9%) and 'catalytic activity' (1,266 28.2%; Fig. S1). Finally, the last division, 'cellular component', also had 9 categories. Gene products were primarily expressed intracellularly (2,322 = 41.9%) or in the membrane bound/non-membrane bound organelle (1,787 = 32.3%; Fig. S1).

5.4.4 Localization of contigs

The zebra finch and chicken genomes were used as references to locate the contigs. BLAT mapping of our assemblies against these genomes showed sequences that uniquely mapped to particular features (5'UTR [untranslated region], 3'UTR, CDS [coding sequence], 1 Kb upstream, 1 Kb downstream, and RNA genes; Fig. 4). Based on the zebra finch genome annotation, nearly 34% of mapped contigs (2,890 of 8,561) were found to be in CDS regions. There was a greater propensity for reads to have come from 3'UTR and 1 Kb downstream than from 5'UTR or 1 Kb upstream (Fig. 4). Similar patterns, although with slightly fewer hits, were obtained from mapping to the chicken genome. Localization of contigs containing SNPs/indels

mapped against the zebra finch and chicken genomes showed that a major proportion of polymorphisms belong to coding sequences (Fig. S2). Contigs with SNPs/indels had more blast hits to the zebra finch than to the chicken, reflecting the overall pattern of all contigs (Table 2). Few RNA genes were found by BLAT mapping (Fig. 5).

5.4.5 Estimation of K_a/K_s

Substitution rates were estimated for the 4,574 contigs orthologous to both zebra finch and chicken. After filtering (based on length of alignment and removing frame shifts), the number of contigs was reduced to 3,821. We then excluded contigs that were either identical or which had $K_s = 0$ (which made K_a/K_s incalculable). Thus, K_a/K_s was estimated for 3,252 (zebra finch) and 3,127 (chicken) contigs. Rate estimation with zebra finch identified 43 contigs with $K_a/K_s \geq 1$ and 283 with values of 0.5-1.0 (Fig. 6A). Rate estimations with chicken yielded 5 and 58 contigs with $K_a/K_s \geq 1$ and between 0.5-1.0, respectively (Fig. 6B). Afterwards, assuming the song sparrow contigs have the same chromosome organization as zebra finch and chicken, calculated ratios were organized into chromosomes (Table 3); this is not an unrealistic assumption considering the high degree of chromosomal conservation among avian genomes (Griffin *et al.* 2007, Ellegren 2010) and the fact that such a high proportion (85.1%) of our orthologous contigs were found to have shared chromosomal locations with zebra finch and chicken.

Data organized into chromosomes suggest that contigs may have undergone more selection with respect to the zebra finch than the chicken (as high K_a/K_s values are typically interpreted, though see Hughes 2007). However, we discuss below why this is likely a methodological artifact of different branch lengths; similar patterns were identified for each chromosome (Table 3).

Chromosomes 22 and 26 showed the greatest differences between zebra finch and chicken in the percentage of song sparrow contigs mapped (relative to the number of genes available in the Biomart database for zebra finch and chicken). Both of these chromosomes had significantly different frequencies of mapped-song-sparrow versus Biomart data-available genes between zebra finch and chicken ($G_{adj} = 4.4$, $P < 0.05$, and $G_{adj} = 6.9$, $P < 0.01$, respectively at 1 d.f., *G*-test with Williams' correction; Table 3). In both cases proportionally more contigs were mapped to zebra finch than to chicken given the sizes of the respective databases (Table 3).

5.4.6 Chromosomal distributions of between-subspecies candidate loci

Two findings emerged in comparing the among-chromosome locations (mapped against zebra finch) of the between-subspecies candidate loci that were mapped to chromosomes (218 SNP/indel-bearing, between-subspecies song sparrow contigs; Supplementary Information) versus all orthologous song sparrow contigs (Table 3). First, the chromosomal distribution of the candidate loci was significantly different from the distribution of all orthologous contigs ($G_{adj} = 51.5$, 27 d.f., $P < 0.005$), indicative of a nonrandom process (e.g., selection). Importantly, the chromosomal distribution of the 199 unique, mappable SNP/indel-bearing contigs between Attu and Adak islands (within the subspecies *maxima*), where we expected drift rather than selection to be more pronounced, was not significantly different from the chromosomal distribution of all orthologous contigs ($G_{adj} = 35.1$, 27 d.f., $P > 0.1$). Secondly, the greatest differences in the distribution of between-subspecies candidate loci from the distribution of all contigs occurred among chromosomes 2, 5, and Z (where proportionally fewer SNP/indel-bearing contigs occurred than expected) and chromosomes 3 and 11 (where relatively more SNP/indel-bearing contigs occurred than expected).

Finally, in contrasting our between-subspecies results with those of our between-species comparisons above, we found that seven of the SNP/indel-bearing contigs that are candidate loci between-subspecies were also contigs that exhibited evidence suggestive of selection (high K_a/K_s values) when compared with zebra finch and chicken. Each contig has one between-subspecies SNP (Supplementary Information). Three of these seven occurred on chromosome 3 and one on chromosome 11, where the between-subspecies contrasts suggested elevated levels of SNPs/indels. These contigs and their chromosomal locations may thus be important in divergence at both intra- and interspecific levels.

5.5 DISCUSSION

Among non-model organisms, determining the molecular genetic underpinnings of adaptation and divergence is a challenge. Genomic scans such as those performed here can help to find loci linked to or responsible for adaptive evolution. We chose to study two subspecies of songbird that exhibit phenotypic differences in multiple characteristics that commonly differ between diverging bird populations. Because the song sparrow is not a model species, we had no *a priori* expectations of specific genes likely to exhibit evidence either of selection between song sparrow and zebra finch/chicken or of SNP/indel differences between song sparrow subspecies. However, using our approach we did expect to discover candidate genes, or loci associated with such (i.e., through linkage), at both levels of comparison. Through comparative transcriptomics we have identified two groups of candidate loci, at both intraspecific and interspecific levels, and these sets of loci overlapped a little, with seven loci (and 7 SNPs) in common.

The dataset assembled here is a balance between efficient discovery of SNPs/indels of interest, conservative criteria for calling variants, and economics. We acknowledge that the amount of sequencing presented is insufficient to allow a high-quality assembly of the extremely diverse transcriptome that we have sampled. A large number of tissues were sampled, and these clearly contain a large and diverse set of transcripts (see *Functional annotation* above). Large transcripts were often split among multiple contigs in our assembly, which is evident from the large number of contigs (about twice as many as there are in the chicken and zebra finch Biomart files) and from the fact that 12,880 contigs mapped to only 8,540 unique proteins. Simulations indicate that transcriptomes sequenced with 454 Titanium chemistry will quickly lead to about twice as many contigs as transcripts, and additional sequences only gradually cause the number of contigs to reach the number of transcripts (i.e., the point when contigs = transcripts; data not shown). Thus, quite large numbers of sequences will be necessary to fully assemble the transcripts contained in these cDNA libraries.

We also note that although the MINT cDNA construction kit is meant to only allow amplification of full-length transcripts, we still observe a substantial bias toward contigs mapping to 3'UTR and 1 kb downstream relative to 5' UTR and 1 kb upstream. The normalized distributions clearly indicate that our libraries contain relatively few transcripts that are full-length. We also note that we have used quite stringent criteria for SNP/indel assignment. By requiring at least three reads for the minor allele, a minimum of 6x coverage is required to call a SNP. Because our average assembly depth is only about 4x, most polymorphic nucleotides in our contigs will not pass our criteria for SNP discovery. Because of this, we have biased the SNPs to be from the relatively highly expressed transcripts. None of these issues limit our ability to

achieve our stated goals, but we note them so that it is understood that we have made appropriately cautious interpretations of our results.

Although K_a/K_s (sometimes calculated as d_N/d_S or ω) is commonly misinterpreted (Hughes 2007), this ratio of rates of non-synonymous to synonymous substitutions can give some context to candidate genes and allows for subsequent hypothesis testing (e.g. Elmer *et al.* 2010, Barreto *et al.* 2011). The fact that K_a/K_s values were higher on average for the zebra finch than the chicken (Table 3) is likely a methodological artifact. Zebra finch is in the same taxonomic order as the song sparrow (Passeriformes), whereas the chicken is taxonomically distant (Galliformes). Estimates of ω necessarily classify sites with differences as nonsynonymous or synonymous, and errors in the estimation of either can profoundly affect the outcome of these analyses (Yang and Bielawski 2000). Taxonomic or lineage distance (longer branches) will affect reconstruction of synonymous substitution rates especially (through an expected increase in repeated mutations, or multiple hits), and we consider this to be a likely source of the consistent differences in apparent molecular selection between our song-sparrow-to-zebra-finch versus song-sparrow-to-chicken contrasts (Table 3; see also Schneider *et al.* 2009). Nevertheless, these contrasts are valuable in highlighting the chromosomal distributions (assuming chromosomal stability; Ellegren 2010) and relative values of ω between closer and more distant relatives of the song sparrow, providing insights into attributes of selection in the coding genome across these scales. Unfortunately, this approach is not valid within species (e.g. between subspecies; Rocha *et al.* 2006, Kryazhimskiy & Plotkin 2008).

The strong likelihood of selection operating on some of the multiple characteristics that differ between the song sparrow subspecies *caurina* and *maxima* provided an *a priori* expectation that contigs with between-subspecies SNPs/indels would represent a non-random subset of all contigs. Our comparison of these distributions as mapped against zebra finch chromosomes showed that such a difference occurred. This allows us to infer that among the 392 between-subspecies SNPs/indels some are likely to be linked to traits under selection or may themselves be under selection. Some (probably most), however, will be false positives. In the nearly linear distribution of the song sparrow in this northwestern portion of its range (Fig. 1), the Cordova population of *caurina* is the first (as one goes west) to exhibit significant bottlenecking (Pruett & Winker 2005). At least two more bottlenecks occurred that affect our data, one on the Alaska Peninsula before the species' range reaches Adak Island, and the other on Attu Island (Pruett & Winker 2005). We thus expected some SNPs/indels to appear due to the neutral loss of genetic variation, but we chose the bottlenecked Cordova population as the earliest in this historical sequential bottlenecking process to minimize the effects of drift. The significant difference between the distributions among chromosomes of all orthologous contigs versus those with between-subspecies SNPs/indels suggests that this approach was successful.

In addition to possible population bottleneck effects, our small within-subspecies sample sizes (N = 2 *caurina* and 4 *maxima*) also suggest that many of our between-subspecies SNPs/indels will likely not represent fixed differences. We do not consider this problematic, however, for several reasons. First, the most dramatic phenotypic differences, body size and seasonal migration, almost certainly involve many genes each, and fixation among even a substantial fraction of them seems unlikely (Pulido 2007, Weedon & Frayling 2008). Second, especially when dealing

with polygenic traits, we might predict that few genes of large effect (if they are present between subspecies) would exhibit fixation and be uncovered in our study of just part of the transcriptome. And, finally, increasing evidence suggests that polygenic adaptation, in which allele frequencies change modestly at many loci during adaptive change (especially when changes originate from standing genetic variation, as is likely in our case), may be a common phenomenon (Hancock *et al.* 2010, Pritchard *et al.* 2010, Pritchard & Di Rienzo 2010).

Additional sequencing of these libraries and further characterization of the candidate SNPs/indels on large numbers of individuals from multiple song sparrow populations is clearly warranted. However, the loci identified in our dataset may have immediate value to the broader molecular ecology community. Pritchard *et al.* (2010:R213) stated that "…as we learn more about gene functions, we will surely find that there are some – perhaps many – more great biological candidate loci lurking near the top of the selection scan lists." Insofar as the list of phenotypic differences between these song sparrow subspecies (i.e., body size, seasonal migration, and plumage coloration) include differences commonly found between divergent bird populations and species, we consider that the candidate loci described here will include some of broad utility for studying the genomics of avian divergence and speciation.

5.6 ACKNOWLEDGEMENTS

We thank NSF Alaska EPSCoR (EPS-0701898) and the University of Alaska Museum for supporting this research. Jack Withrow assisted in fieldwork, Roger Nilsen made the normalized cDNA libraries, Jeff Wagner sequenced the libraries, and Christin Pruett and Erik Postma provided helpful comments.

5.7 REFERENCES

Altenhoff AM, Dessimoz C (2009) Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods. *PLoS Computational Biology*, **5,** e1000262.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.

Barreto FS, Moy GW, Burton RS (2011) Interpopulation patterns of divergence and selection across the transcriptome of the copepod *Tigriopus californicus. Molecular Ecology*, **20**, 560-572.

Beldade P, Rudd S, Gruber J, Long A (2006) A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. *BMC Genomics*, **7**, 130.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, **14**, 1147-1159.

Conesa A, Götz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674-3676.

Ellegren H (2010) Evolutionary stasis: the stable chromosomes of birds. *TREE*, **25**, 283-291.

Elmer KR, Fan S, Gunter HM, Jones, JC, Boekhoff S, Kuraku S, Meyer A (2010) Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. *Molecular Ecology*, **19** (Suppl. 1), 197-211.

Götz S, García-Gómez JM, Terol J, *et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, **36**, 3420-3435.

Griffin DK, Robertson LBW, Tempest HG, Skinner BM (2007) The evolution of the avian genome as revealed by comparative molecular cytogenetics. *Cytogenetic Genome Research*, **117**, 64-77.

Hale MC, McCormick CR, Jackson JR, DeWoody JA (2009) Next-generation pyrosequencing of gonadal transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics*, **10**, 203.

Hancock M, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, et al. (2010) Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Science USA*, **107** (suppl.2), 8924-8930.

Haygood R, Babbitt CC, Fedrigo O, Wray GA (2010) Contrasts between adaptive coding and noncoding changes during human evolution. *Proceedings of the National Academy of Science, USA,* **107**, 7853-7857.

Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity*, **99**, 364-373.

Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome Research*, **12**, 656-664.

Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. *PLoS Genetics*, **4**, e1000304.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al.* (2007) Clustal W and clustal X version 2.0. *Bioinformatics*, **23,** 2947-2948.

Marth GT, Korf I, Yandell MD, Yeh RT, Gu ZJ, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR (1999) A general approach to single-nucleotide polymorphism discovery. *Nature Genetics*, **23,** 452-456.

Mayr E (1982) *The Growth of Biological Thought: Diversity, Evolution, and Inheritance.* Belknap Press, Cambridge, Massachusetts.

Nadeau NJ, Jiggins CD (2010) A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. *Trends in Genetics*, **26**, 484-492.

Pritchard JK, Di Rienzo A (2010) Adaptation – not by sweeps alone. *Nature Reviews Genetics*, **11**, 665-667.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: Hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, R208-R215.

Pruett CL, Winker K (2005) Northwestern song sparrow populations show genetic effects of sequential colonization. *Molecular Ecology*, **14**, 1421-1434.

Pruett CL, Winker K (2010) Alaska Song Sparrows (*Melospiza melodia*) demonstrate that genetic marker and method of analysis matter in subspecies assessments. *Ornithological Monographs*, **67**, 162-171.

Pulido F (2007) The genetics and evolution of avian migration. *BioScience*, **57**, 165-174.

Quinlan AR, Stewart DA, Stromberg MP, Marth GT (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods*, **5,** 179-181.

Rocha EPC, Maynard Smith J, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ (2006) Comparisons of d*N*/d*S* are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*, **239**, 226-235.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D (2009) Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biology and Evolution*, **2009**, 114-118.

Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J (2010) Adaptation genomics: the next generation. *Trends in Ecology and Evolution*, 25, 705-712.

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, **34**, W609-W612.

Weedon MN, Frayling TM (2008) Reaching new heights: insights into the genetics of human stature. *Trends in Genetics*, **24**, 595-603.

West-Eberhard MJ (2003) *Developmental Plasticity and Evolution*. Oxford University Press, Oxford.

Winker K (2010) Subspecies represent geographically partitioned variation, a goldmine of evolutionary biology, and a challenge for conservation. *Ornithological Monographs*, 67, 6-23.

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *TREE* **15**, 496-503.

Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586-1591.

Tables:

Table 5.1 Number of reads and assembly statistics for three song sparrow populations.

Subspecies	Locality	N[1]	MID	Raw reads	Cleaned reads	Cleaned Bases (MB)	Accession
M. m. caurina	Cordova	2	13	138,439	114,098	32.5	XXX
M. m. maxima	Adak	2	19	135,588	117,166	34.7	XXX
M. m. maxima	Attu	2	18	184,781	150,210	42.8	XXX
Combined	-	6	-	458,808	381,474	110	XXX

Table 5.2 Species with ≥100 Top Hits from Blast2GO.

Species	**Hits**
Taeniopygia guttata	7820
Gallus gallus	2222
Homo sapiens	235
Monodelphis domestica	193
Mus musculus	187
Ailuropoda melanoleuca	177
Ornithorhynchus anatinus	149
Canis familiaris	119
Melospiza melodia	113
Rattus norvegicus	100

Table 5.3 Number of contgis orthologous to particular zebra finch and chicken chromosomes, and mean K_a/K_s ratio for each chromosome, assuming the orthologous contigs have the same chromosomal location as zebra finch and chicken.

Chr	Contigs orthologous to particular Zebra finch chromosome	Total Number of transcripts from particular Zebra finch chromosome in Biomart file	K_a/K_s Mean ± SD	Contigs orthologous to particular Chicken chromosome	Total Number of transcripts from particular Chicken chromosome in Biomart file	K_a/K_s Mean ± SD
1	261	1124	0.2552 ± 0.2733	492	2994	0.1528 ± 0.1694
2	338	1345	0.2434 ± 0.2465	339	1995	0.1457 ± 0.1326
3	309	1169	0.2434 ± 0.2807	314	1672	0.1565 ± 0.1497
4	188	741	0.2258 ± 0.3347	252	1516	0.1374 ± 0.1274
5	229	936	0.2103 ± 0.2184	234	1299	0.1280 ± 0.1219
6	107	562	0.2447 ± 0.2112	106	781	0.1486 ± 0.1187
7	124	521	0.2220 ± 0.2103	120	767	0.1361 ± 0.1235
8	111	416	0.2581 ± 0.2196	127	723	0.1436 ± 0.1251
9	90	458	0.2286 ± 0.3839	86	598	0.1045 ± 0.1087
10	86	394	0.1784 ± 0.1738	90	599	0.1220 ± 0.1890
11	68	371	0.2330 ± 0.2978	61	499	0.1429 ± 0.1439
12	73	349	0.1799 ± 0.2206	68	427	0.1076 ± 0.1122
13	77	321	0.1845 ± 0.2319	83	499	0.0994 ± 0.1225
14	80	390	0.2541 ± 0.3448	79	578	0.1333 ± 0.1288
15	76	350	0.1817 ± 0.2299	73	531	0.0925 ± 0.1207
17	49	300	0.1705 ± 0.1597	46	432	0.0967 ± 0.0861
18	54	309	0.2230 ± 0.1950	55	428	0.1085 ± 0.0907
19	68	313	0.2004 ± 0.2982	66	443	0.0858 ± 0.0952
20	50	329	0.2419 ± 0.2444	51	476	0.1336 ± 0.1277
21	34	192	0.1470 ± 0.1569	44	346	0.0847 ± 0.1058
22	16	98	0.1000 ± 0.0976	11	160	0.0441 ± 0.0593
23	34	205	0.1783 ± 0.1828	33	288	0.0782 ± 0.0920
24	27	181	0.1961 ± 0.1906	24	270	0.1000 ± 0.0982
25	7	92	0.1161 ± 0.1069	6	169	0.0711 ± 0.1017
26	31	176	0.1148 ± 0.1081	29	341	0.0824 ± 0.0927
27	31	252	0.1471 ± 0.1438	28	345	0.0698 ± 0.0727
28	27	227	0.1102 ± 0.1256	23	284	0.0476 ± 0.0414
Z	149	745	0.2321 ± 0.2293	146	990	0.1381 ± 0.1174

Figures:

Fig. 5.1 Samples in this study came from Cordova (*Melospiza melodia caurina*, right in inset) and Adak and Attu islands (*M. m. maxima*, left in inset); gray shading indicates the species range.

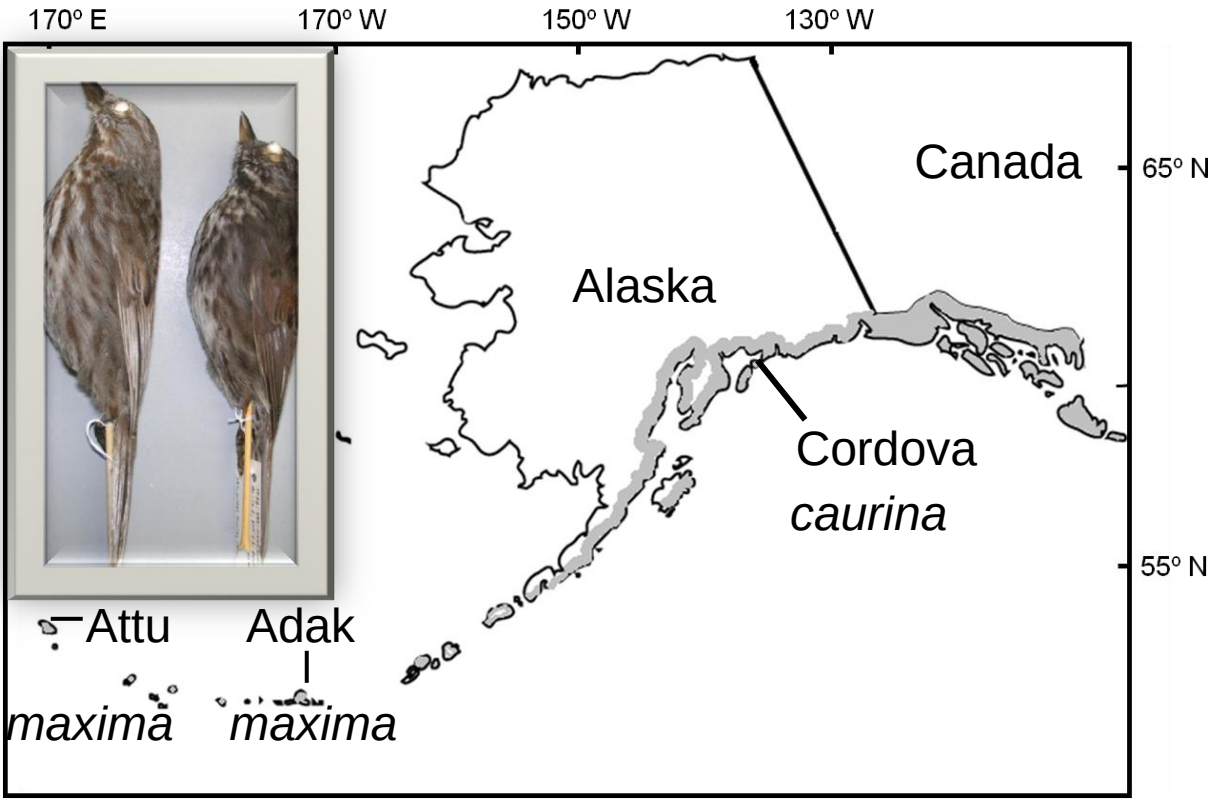


Fig. 5.2 Numbers of SNPs and indels that are within and shared between and among three populations of song sparrows.

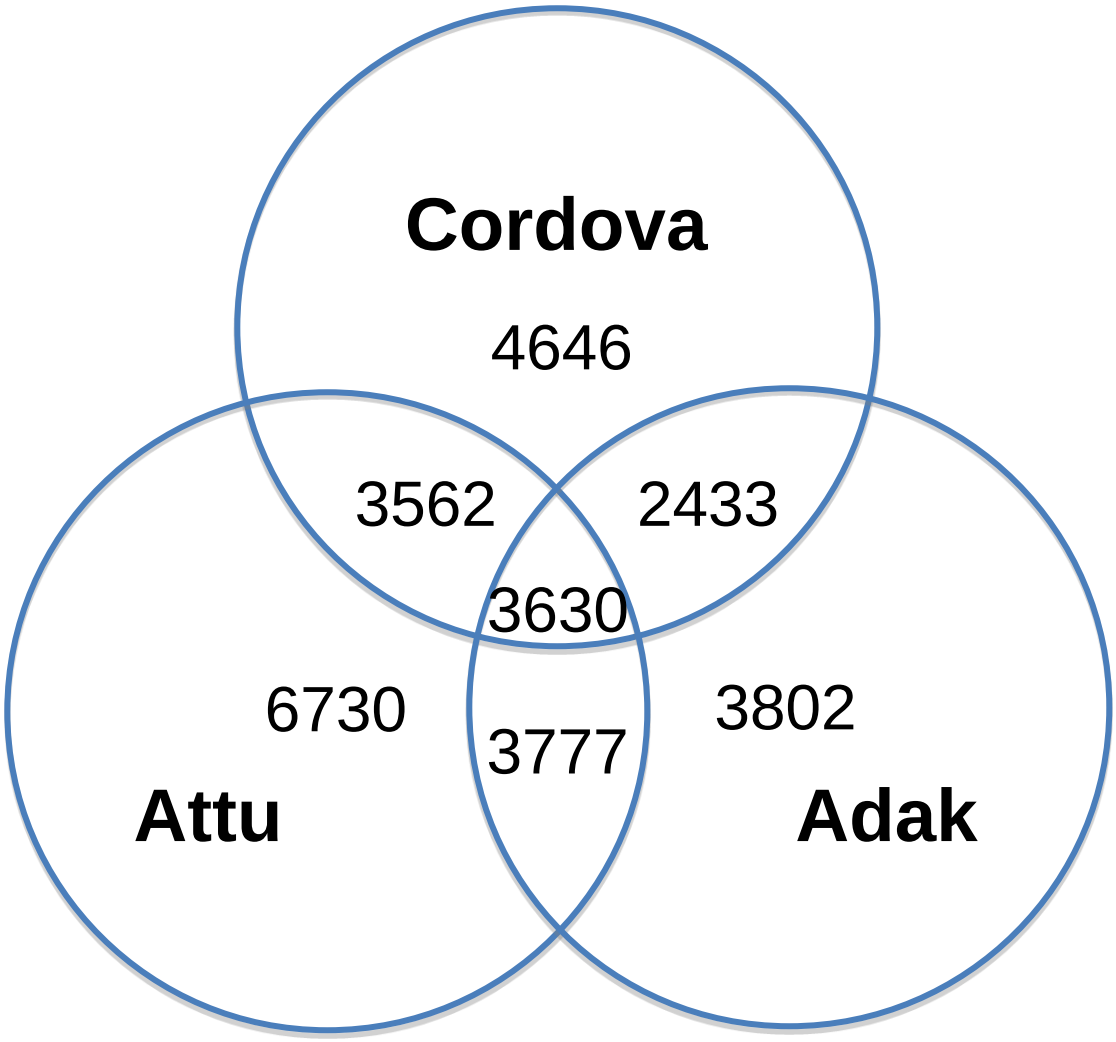


Fig. 5.3 SNPs and indels that are fixed between and among three populations of song sparrows. There are 392 SNPs/indels that are identical in Attu and Adak, but different from Cordova. Because sample sizes are small, these figures include false positives.

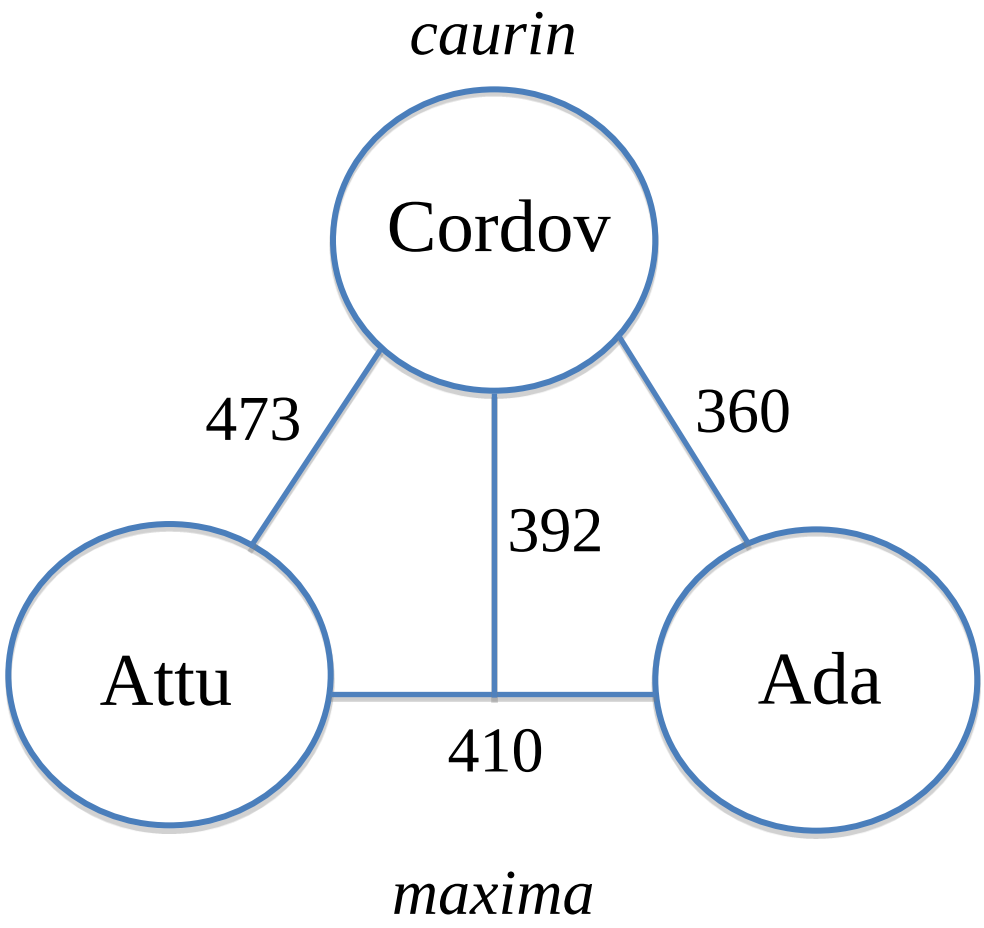
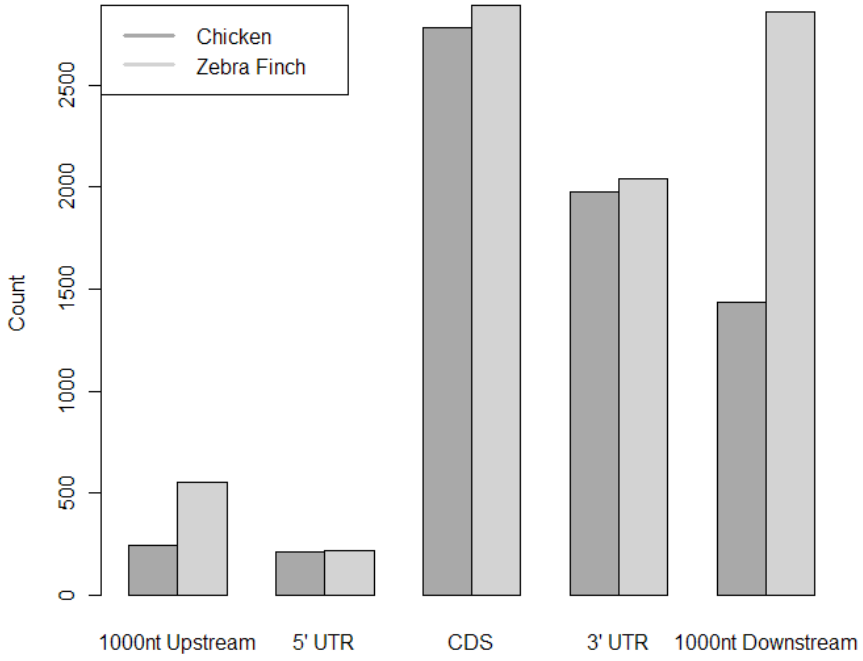


Fig. 5.4 Histogram displaying the proportion of contigs mapped to particular features of protein coding genes of zebra finch and chicken (UTR is untranslated region, and CDS is coding sequence). The upper panel displays the raw count, and the lower panel displays normalized values (the proportion discovered relative to how many could be discovered within each category).



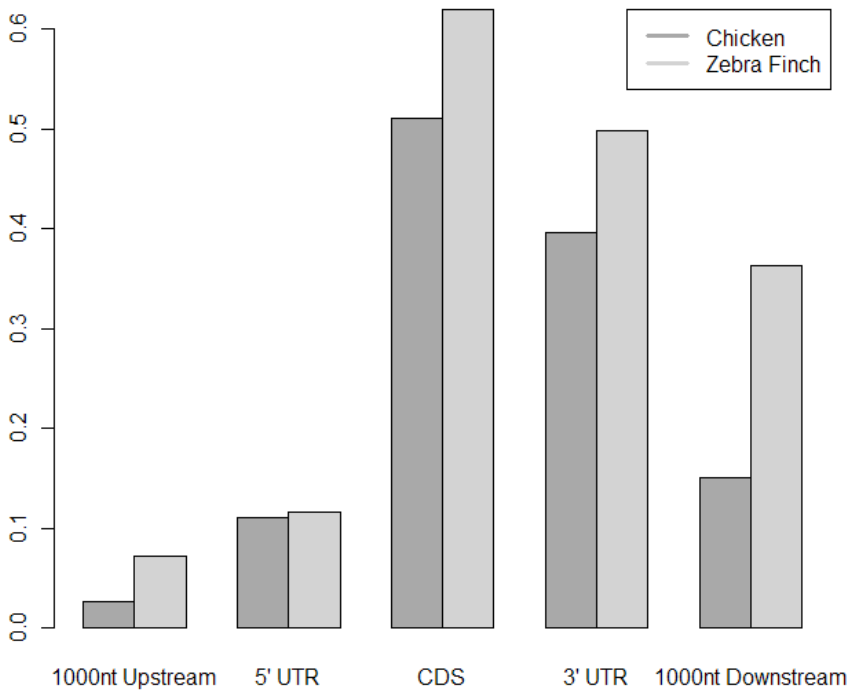
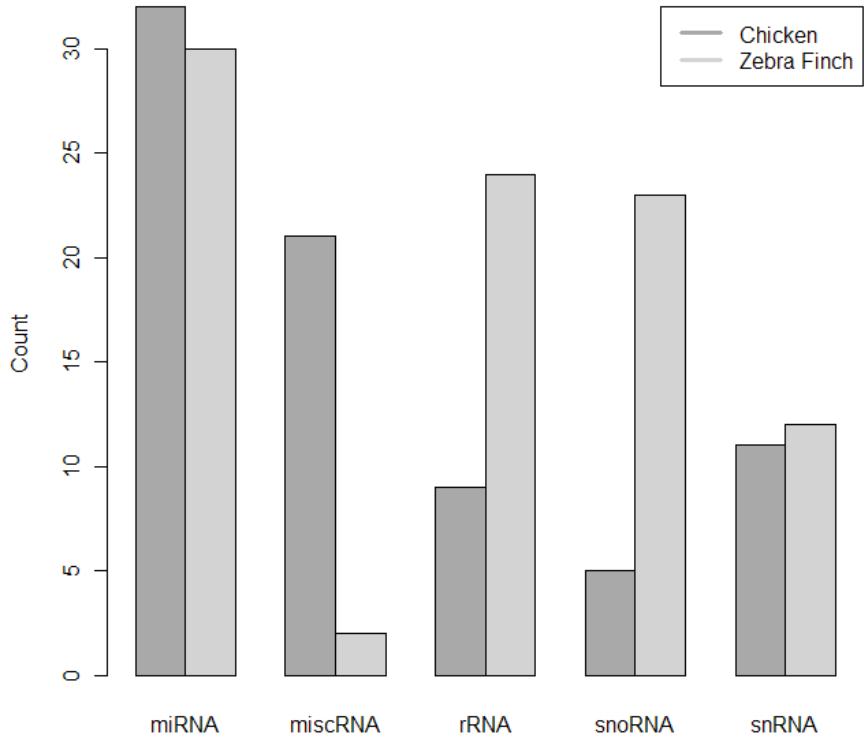

Chicken
Zebra Finch
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1000nt Upstream
5' UTR
CDS
3' UTR
1000nt Downstream

Fig. 5.5 Histogram displaying the proportion of contigs mapped to particular RNA genes of the zebra finch and chicken. The upper panel displays the raw count, and the lower panel displays normalized values (the proportion discovered relative to how many could be discovered within each category).



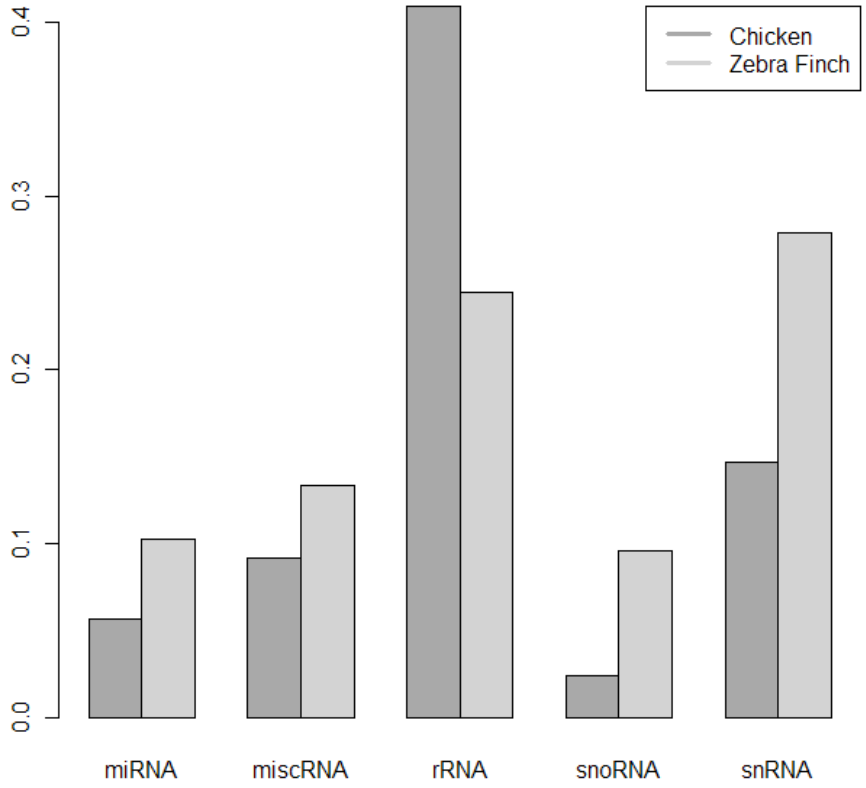
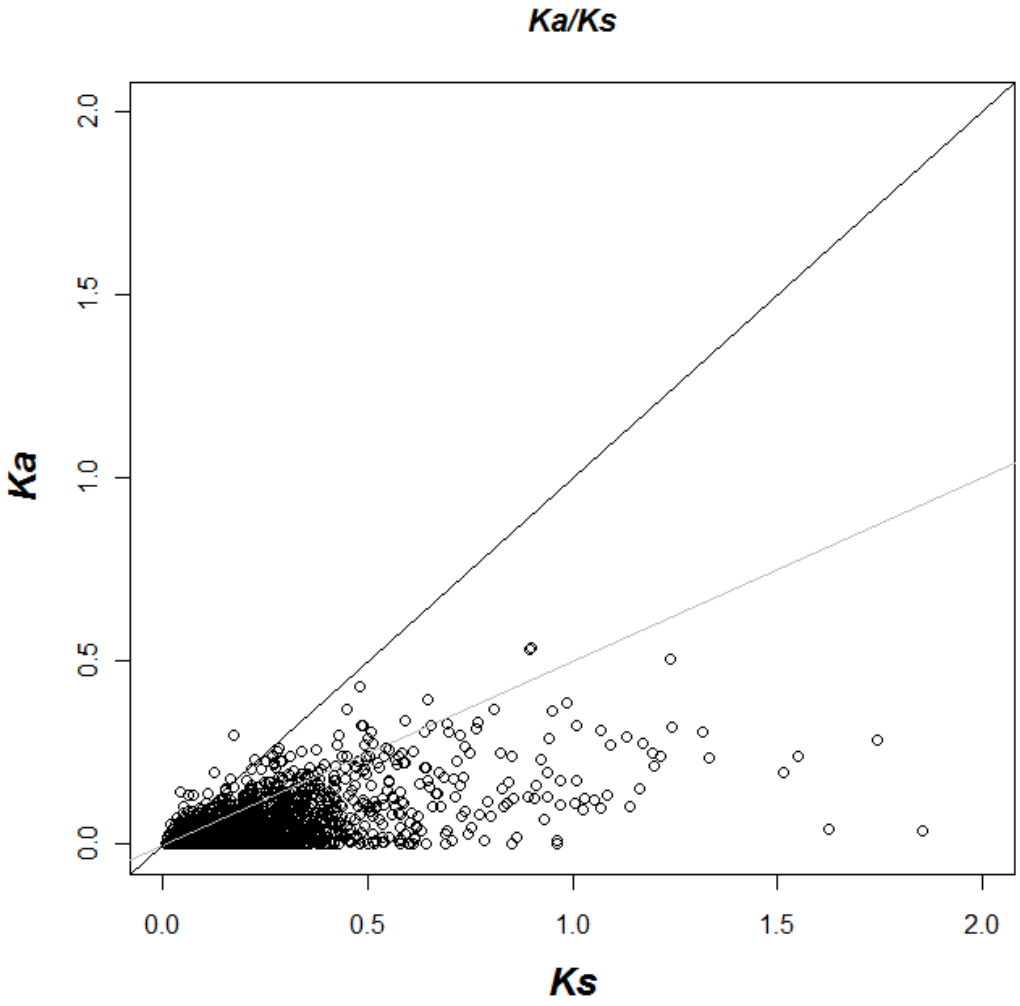
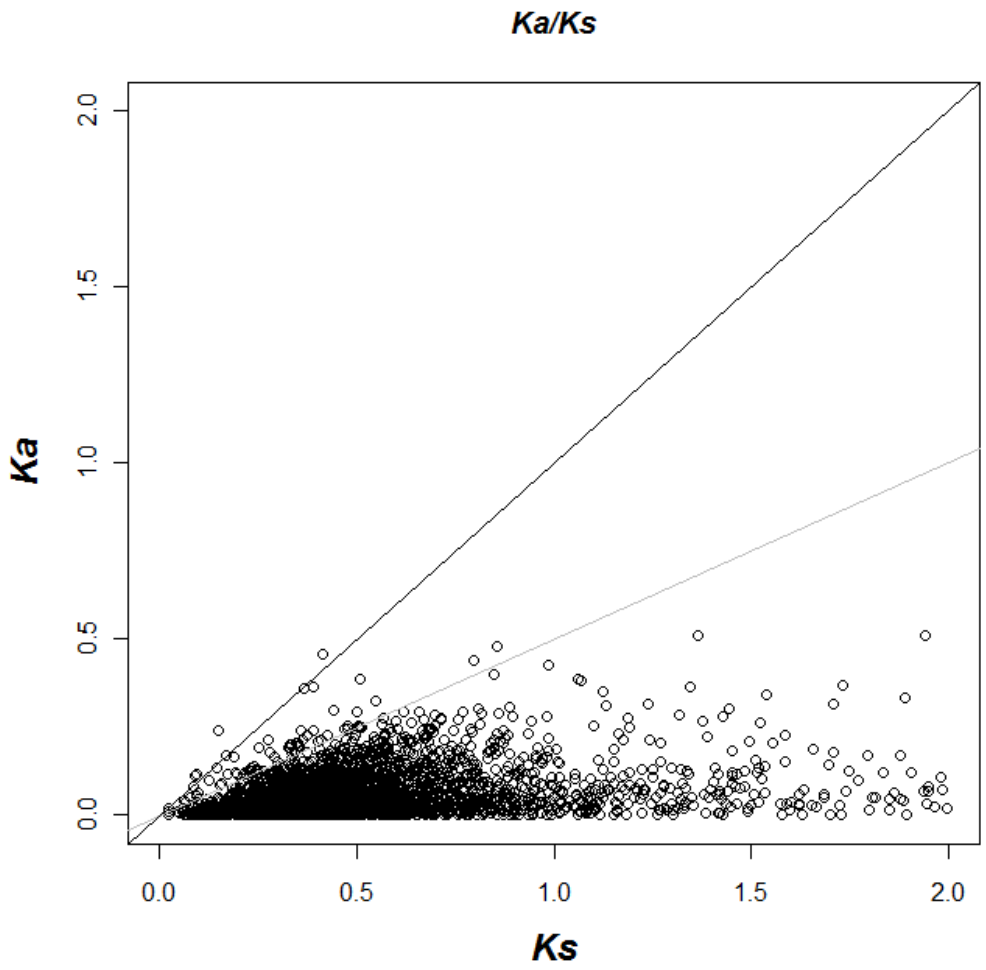

Chicken
Zebra Finch
0.0
0.1
0.2
0.3
0.4
miRNA
miscRNA
rRNA
snoRNA
snRNA

Fig. 5.6 Distribution of K_a/K_s ratio for the contigs orthologous to both zebra finch (A) and chicken (B). Contigs with K_a/K_s values of 0.5-1.0 fall above the grey line, and values > 1.0 fall above the black line.




Ka/Ks
Ka
Ks
0.0
0.5
1.0
1.5
2.0

5.8 SUPPLEMENTARY MATERIAL

Fig. S 5.1A Song sparrow GO biological process distribution:

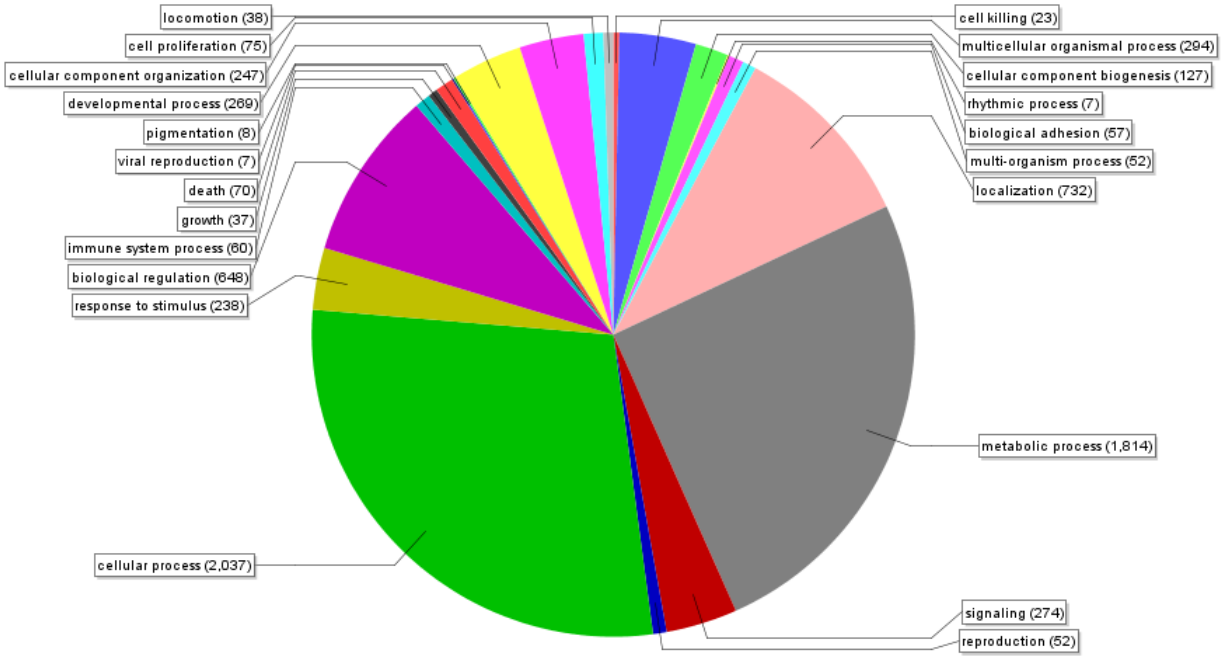


Fig. S 5.1B Song sparrow GO molecular function distribution.

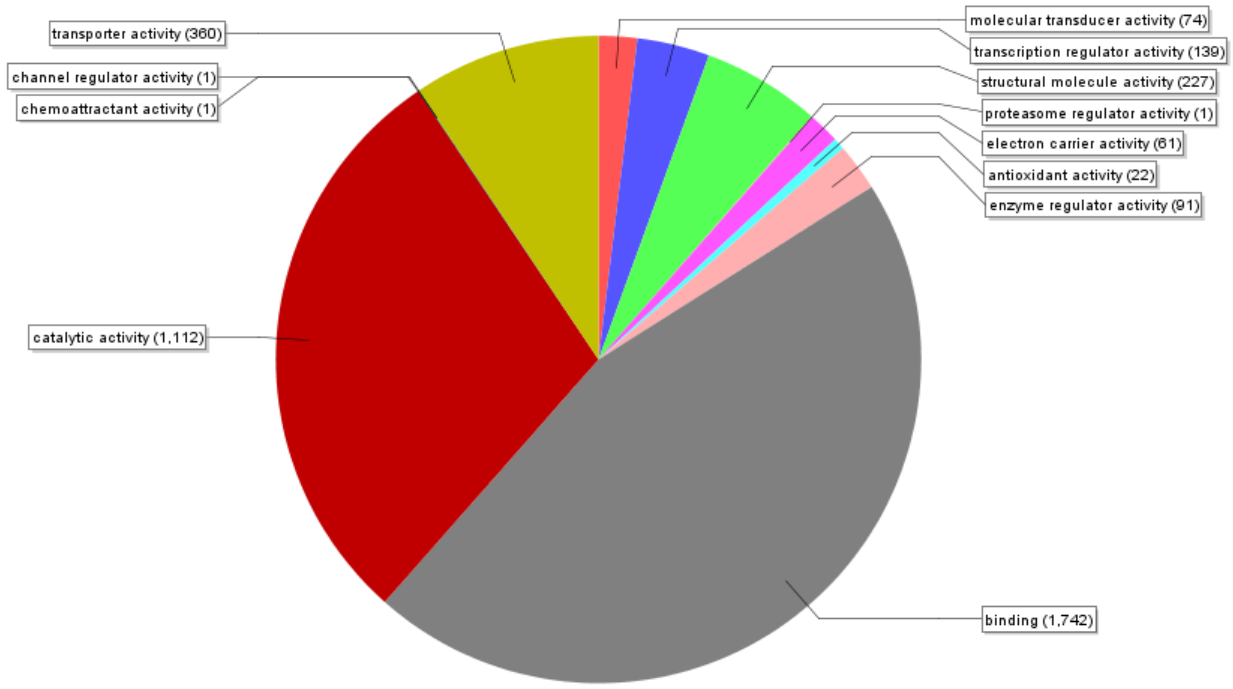


Fig. S 5.1C Song sparrow Cellular component distribution.

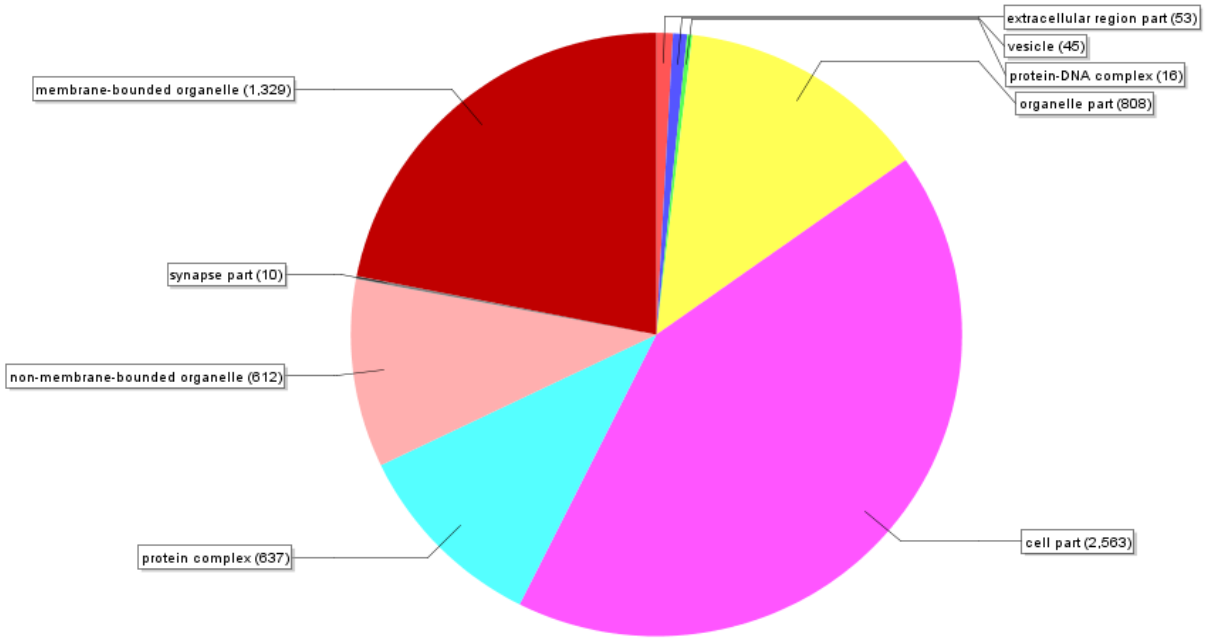
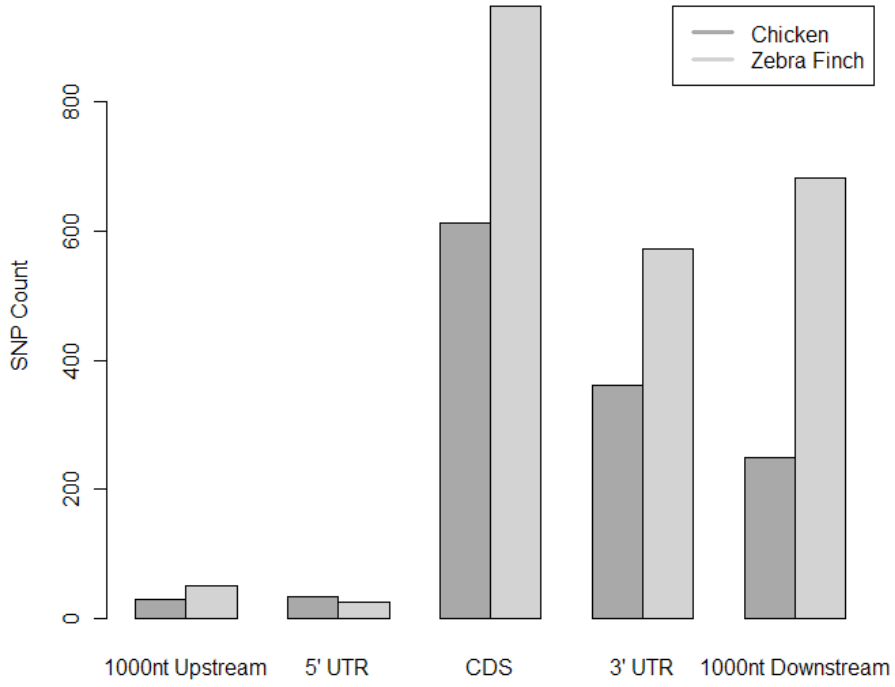
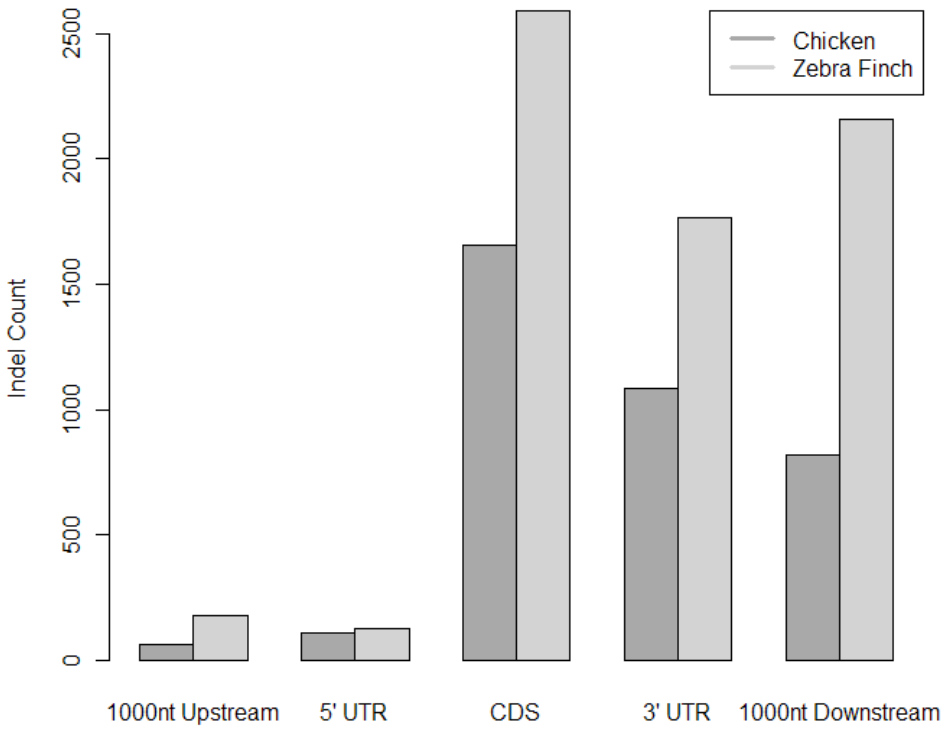


Fig. S 5.2 Histograms showing the localization of predicted SNPs and indels inside the transcripts of zebra finch and chicken (UTR is untranslated region and CDS is coding sequence). Panel A displays the raw count, and the lower panel (B) displays normalized values (the proportion discovered relative to how many could be discovered within each category).




Chicken
Zebra Finch
Indel Count
0
500
1000
1500
2000
2500
1000nt Upstream
5' UTR
CDS
3' UTR
1000nt Downstream

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 NON-CODING RNA

In section I of this dissertation, we worked on the problem of finding new instances of known non-coding RNAs in the genome. We approached the problem in two different ways; the first approach focused on improving the sensitivity of current ncRNA gene finders by developing mutational matrices which, when incorporated into ncRNA search programs, could lead to the detection of distant homologs. Additionally, the study of ncRNA evolution provided insight into the changes which cause variability in ncRNA secondary structure in related species. An extension of this project would be to apply this analysis to other families of RNA with both conserved and variable secondary structures to find the universal patterns of RNA secondary structure evolution. Some of the other families on which this analysis could be extended may involve group I Introns, group II introns, and RNase MRP, as the length of the sequences of these families are sufficiently long and their secondary structures contain both conserved and variable structural elements (Griffiths-Jones, Bateman et al. 2003), a key requirement for this type of analysis.

The second approach was based on using the patterns of chromatin-modification to discriminate and detect different genomic features, particularly ncRNAs. This is a novel approach for genomic features prediction. From our analysis, we found that there is a strong bias among the

patterns of chromatin-modification in different genomic features which can be used for the prediction of novel features from modification patterns of un-annotated regions of genomes. Although, this is an expensive and laborious approach (due to work involved in ChIP-Chip and ChIP-Seq data generation), it can still be used to improve the annotation for the species for which this dataset is already available (as shown for *Arabidopsis* in chapter 3). Future extensions of this project involves the application of this approach to other plant species such as *Oryza* for which a similar type of dataset is already obtained (Li, Wang et al. 2008), as it will give us insight into whether different patterns of chromatin modifications that we obtained for *Arabidopsis* are universal or unique to it. We also propose the extension of this approach to Human and Yeast by obtaining data from the Human Histone modification database (Zhang, Lv et al. 2010), the Histone database at National Human Genome Research Institute (NHGRI) and the chromatin DB (O'Connor and Wyrick 2007), respectively.

6.1.1 References:

Griffiths-Jones, S., A. Bateman, et al. (2003). "Rfam: an RNA family database." Nucleic Acids Research **31**(1): 439-441.

Li, X., X. Wang, et al. (2008). "High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression." Plant Cell **20**(2): 259-276.

O'Connor, T. R. and J. J. Wyrick (2007). "ChromatinDB: a database of genome-wide histone modification patterns for Saccharomyces cerevisiae." Bioinformatics **23**(14): 1828-1830.

Zhang, Y., J. Lv, et al. (2010). "HHMD: the human histone modification database." Nucleic Acids Res **38**(Database issue): D149-154.

6.2 *SARRACENIA* AND SONGBIRD:

The main objective of our transcriptomic project involves, obtaining the functional annotation of a significant portion of transcriptome, identification of polymorphic loci and estimation of the substitution rate to identify rapidly evolving genes. By using our transcriptome analysis pipeline, we were able to achieve all the aforementioned goals. We believe that these studies have provided new insights into pitcher plant and Songbird transcriptomes and will contribute significantly to future research on these genus and their distinctive ecological adaptations. Some of the possible future extensions of these projects include:

6.2.1 *Sarracenia* sequence based phylogeny:

The first extension of this project would be to construct asequence based phylogeny for species in the *Sarracenia* genus. Previously, Neyland and Merchant (Ray and Merchant 2006) constructed a phylogeny based upon internally transcribed spacer 2 and 26S large ribosomal subunit rRNA genes DNA sequences, for the family Sarraceniaceae. In their study, they found that, with-in the genus *Sarracenia*, *S. purpurea* is a sister taxa to all remaining species. However, the study was not able resolve the evolutionary relatedness among the majority of *Sarracenia* species. We are planning to construct comprehensive sequence based phylogeny for the genus *Sarracenia* by taking advantage of shared polymorphic loci detected in our transcriptome assembly. Previously, we detected 3865 SNPs/Indels which are shared between *S. psittacina* and *S. purpurea* and we purpose to use the contigs with these SNP/Indels as a starting point for the construction of phylogeny. To do this, we will construct a normalized list of contigs with the most number of SNP/Indel and then regions surrounding the loci will be sequenced in a high-throughput manner by a sequence capture technology (Roche NimbleGen). NimbleGen sequence

capture technology is a revolutionary process for the enrichment of selected genomic regions. We are planning to sequence 200-300 sequences across 10 different species and 4 different genotype of each species (10x4 = 40). Afterwards, sequences would be aligned using the alignment program Muscle (Edgar 2004) and then alignment either be manually curated in Bioedit (Hall 1999) or by the automatic alignment editor RASCAL (Thompson, Thierry et al. 2003). Mr. Bayes (Huelsenbeck and Ronquist 2001) either will be used to create a single tree (from merging separate alignments) or separate phylogenetic trees [which then would be used to construct the one super tree (http://genome.cs.iastate.edu/supertree/introduction/intro_content.html)]. We believe this project will lead us to construct the comprehensive sequence based phylogeny for *Sarracenia* and will give us insight into the evolutionary relationships of the different *Sarracenia* species.

6.2.2 Linking biodiversity to speciation genomics:

This project will be focused on improving our knowledge of the patterns of genomic change that accompany speciation. In this work, we will examine the patterns of genomic change that accompany divergence and speciation in 36 vertebrate lineages that are split at various levels (populations-to-species) by a common biogeographic barrier, the Bering and Chukchi seas, which currently meet in the Bering Strait. This barrier has appeared and disappeared several times (Hopkins DM 1967; Ruddiman, M. Raymo et al. 1986) and caused the divergence to accrue in *Beringia* (Ruddiman, M. Raymo et al. 1986), which is evident from full sister species down to different continental populations. This represents in a way a temporal transect through the speciation process, replicated in this study at several taxonomically key levels (Del Hoyo, A. Elliott et al. 1992-2001). The 36 taxon pairs represent seven avian orders with 11 population-

level, 12 subspecies-level, and 13 species-level comparisons (these categorical levels will be set aside by using two different measures of divergence using phenomics and genomics). We chose to work with birds because of following regions:

- Birds are well understood taxonomically
- The sxistence of sufficient sample to make these contrasts
- Sufficient genomic data to study large numbers of homologous loci

Our fundamental molecular approach to achieve these goals is to sequence ~1,500 homologous loci for all individuals using next-generation sequencing (NGS, on an Illumina HiSeq). For our questions this approach is superior to broader genomic approaches such as using RAD (restriction-site associated DNA) (Davey and Blaxter 2010) markers to obtain a sequence from broader, non-targeted regions of the genome. Targeting homologous subsets of the genome for all specimens in this study gives several advantages:

- It provides a rigorously equivalent comparative basis for examining the genomic effects of divergence and speciation.
- With this number of loci it provides a fairly cost-effective snapshot of the genome ($0.00000195/base for all anticipated data).
- It largely sets aside issues of ascertainment bias (bias in a dataset caused by sampling error, such as selecting only the most variable loci to study.

The other important challenge of this work is to design the probes for the enrichment of the targeted regions. The probes will be designed by the following two sources:

- A panel of ultra-conserved elements (UCEs) conserved among all amniotes;

- From the ten avian (Kunstner, Wolf et al. 2010) and two crocodilian transcriptomes currently available, as well as additional transcriptomes that will become available in 2012.

After obtainin the sequence data, data analysis will be performed using the following steps:

- Demultiplexing of reads obtained by Illumina Hi-Seq
- Removal of adaptor sequence by SCYTHE (https://github.com/vsbuffalo/scythe), and reads will be trimmed using SICKLE (https://github.com/vsbuffalo/sickle)
- Afterwards reads will be split into species specific files and a combined assembly will be performed to generate a consensus string by Velvet (Zerbino and Birney 2008).
- To identify the polymorphism reads from individual species, reads will be mapped against the consensus sequence by the BWA/bowtie (Langmead, Trapnell et al. 2009; Li and Durbin 2009) aligner
- The SAM/BAM files will be generated from the alignment, which then will be used as an input to Samtools pileup (Li, Handsaker et al. 2009), a program to call the variation from the alignment and to determine the most probable genotype at each reference position

6.2.3 References:

Davey, J. W. and M. L. Blaxter (2010). "RADSeq: next-generation population genetics." Brief Funct Genomics **9**(5-6): 416-423.

Del Hoyo, J., A. Elliott, et al. (1992-2001). "Handbook of the Birds of the World, Vols." Lynx Edicions, Barcelona.: 1-16.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Hall, T. A. (1999). " BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT." Nucleic Acids Symp **Ser. 41**: 95–98.

Hopkins DM, e. (1967). "The Bering Land Bridge." Stanford University Press, Stanford, California.

Huelsenbeck, J. P. and F. Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees." Bioinformatics **17**(8): 754-755.

Kunstner, A., J. B. Wolf, et al. (2010). "Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species." Mol Ecol **19 Suppl 1**: 266-276.

Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol **10**(3): R25.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics **25**(14): 1754-1760.

Li, H., B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Ray, N. and M. Merchant (2006). "Systematic relationships of Sarraceniaceae inferred from nuclear ribosomal DNA sequences." Madrono **53(3)**: 223-232.

Ruddiman, W. F., M. Raymo, et al. (1986). "Matuyama 41,000-year cycles: North Atlantic Ocean and northern hemispheric ice sheets. Earth and Planetary Science Letters." **80**: 117-129.

Thompson, J. D., J. C. Thierry, et al. (2003). "RASCAL: rapid scanning and correction of multiple sequence alignments." Bioinformatics **19**(9): 1155-1161.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Res **18**(5): 821-829.