

Dissecting the Genetic Basis of Convergent Complex Traits Based on Molecular

Homoplasy

by

Rui Wang

Graduate Program in Bioinformatics and Genome Technology  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Erich D. Jarvis, Supervisor

\_\_\_\_\_  
Gregory A. Wray

\_\_\_\_\_  
Alexander Hartemink

\_\_\_\_\_  
Paul Magwene

\_\_\_\_\_  
Sridhar Raghavachari

Dissertation submitted in partial fulfillment of  
the requirements for the degree of Doctor of Philosophy in the  
Graduate Program in Bioinformatics and Genome Technology in the Graduate School  
of Duke University

2011

ABSTRACT

Dissecting the Genetic Basis of Convergent Complex Traits Based on Molecular

Homoplasy

by

Rui Wang

Graduate Program in Bioinformatics and Genome Technology  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Erich D. Jarvis, Supervisor

\_\_\_\_\_  
Gregory A. Wray

\_\_\_\_\_  
Alexander Hartemink

\_\_\_\_\_  
Paul Magwene

\_\_\_\_\_  
Sridhar Raghavachari

An abstract of a dissertation submitted in partial fulfillment of  
the requirements for the degree of Doctor of Philosophy in the  
Graduate Program in Bioinformatics and Genome Technology in the Graduate School  
of Duke University

2011

Copyright by  
Rui Wang  
2011

## Abstract

The goal of my thesis is to understand the genetics of a complex behavioral trait, vocal learning, which serves as a critical substrate for human spoken language. With the available genomes of 23 mammals, I developed a novel approach based on molecular homoplasy to reveal Single Non-random Amino Acids Patterns (SNAAPs) that are associated with convergent traits, a task that proved intractable for standard approaches, e.g. dN/dS analyses. Of 73 genes I identified in mammalian vocal learners, ~25% function in neural connectivity, auditory or speech processing. Remarkably, these include a group of 6 genes from the ROBO1 axon guidance pathway. In birds, I found ROBO1 and its ligand SLIT1 show convergent differential expression in the motor output song nucleus of the three independent lineages of vocal learners but not in analogous brain areas of vocal non-learners, and ROBO1 is developmentally regulated during song learning critical periods in songbirds. In a different set of genes, I came across an unexpected discovery of the excess sharing of homoplastic substitutions in humans and domesticated species. I revealed biased nucleotide transitions (mostly favoring A/G mutation) for above amino acid substitutions and found that this rule was significantly relaxed during domestication for artificial selection. Overall, my thesis has resulted in a novel approach for studying convergent complex traits and provided

critical insights into the evolution of vocal learning specifically, and complex traits generally.

## **Dedication**

This thesis is dedicated to my parents and my wife for their endless love and support. They have been a great source of motivation and inspiration.

# Contents

Abstract .....	iv
List of Tables .....	xi
List of Figures .....	xiii
Acknowledgments .....	xvii
Chapter 1. Introduction.....	1
1.1 Vocal learning, a complex convergent behavioral trait .....	2
1.2 Limited knowledge on the genetics for vocal learning .....	8
1.3 Genetic parallelism underlying convergent complex traits .....	10
1.4 Hypothesis: genetic parallelisms among vocal learners .....	11
1.5 Current approaches for studying genetic parallelism.....	12
Chapter 2. dN/dS analysis for dissecting the genetics of vocal learning .....	18
2.1 Introduction.....	18
2.2 Materials and methods .....	19
2.2.1 Selection of species and sequences .....	19
2.2.2 Pipeline for identifying orthologous transcripts.....	20
2.2.3 Search for genes with convergent elevated dN/dS ratios .....	23
2.2.4 Bayesian phylogenetic reconstruction of gene trees.....	25
2.3 Results .....	25
2.3.1 Quality control results of orthologous transcripts.....	25
2.3.2 Genes with dS trees that violate the species tree .....	26

2.3.3 Analyses of dN/dS ratios in vocal learners.....	29
2.4 Brief discussion.....	33
Chapter 3. A novel method for detecting trait associations with molecular homoplasy .	34
3.1 Introduction.....	34
3.2 Materials and methods .....	35
3.2.1 Overview of the SNAAP approach.....	35
3.2.2 The evolutionary model and calculation of tree likelihoods.....	39
3.2.3 Hypothesis testing for the phylogenetic screening step .....	41
3.3 Results .....	42
3.3.1 Evaluation of the phylogenetic screening (steps 2 and 3).....	45
3.3.2 Low false discovery rates of phylogenetic screening .....	47
3.3.3 Little influence from sequencing error or SNPs.....	48
3.4 Brief discussion.....	51
Chapter 4. SNAAPs associated with natural vs. artificial selection.....	53
4.1 Introduction.....	53
4.2 Materials and methods .....	53
4.2.1 Computation of Shannon's entropy surrounding SNAAPs.....	53
4.2.2 Gene ontology enrichment analysis.....	54
4.2.3 Enrichment analysis of genes with high dN/dS ratios in primates.....	54
4.2.4 Ancestral sequence reconstruction .....	55
4.2.5 Enrichment analysis of amino acid substitutions .....	55
4.2.6 Computing mutation biases by codon usage and similar amino acid types ....	56

4.3 Results .....	57
4.3.1 H+VL type 1 SNAAPs occur in more conserved regions .....	61
4.3.2 Excess SNAAPs for human and two domesticated species .....	66
4.3.3 Many type 1 SNAAP genes are functionally relevant to the trait of interest....	70
4.3.4 Enriched GO terms for genes with H+VL but not H+DOM type 1 SNAAPs ...	74
4.3.5 Eriched AE/RSC in genes with H+VL but not H+DOM type 1 SNAAPs .....	78
4.3.6 Gains and Losses in shaping H+VL type 1 SNAAPs in mammals.....	81
4.3.7 Biased transitional nucleotide changes shape type 1 SNAAPs .....	86
4.3.8 SNAAP substitutions present in other vocal learners.....	91
4.3.9 Sharing of H+VL type 1 SNAAP substitutions in mammals.....	94
4.3.10 The H+VL and H+DOM type 1 SNAAPs in Neanderthals.....	95
4.4 Brief discussion.....	95
Chapter 5. Expression studies on the ROBO1 axon guidance pathway in birds.....	98
5.1 Introduction.....	98
5.2 Materials and methods .....	101
5.2.1 Animals.....	101
5.2.2 Cloning ROBO1/2 and SLIT1/2/3 .....	102
5.2.3 Radioactive <i>in situ</i> hybridizations.....	103
5.2.4 Gene expression quantifications. ....	104
5.3 Results .....	105
5.3.1 ROBO/SLITs show differential expression in song nuclei of songbirds.....	105
5.3.2 Convergent expression patterns of ROBO1/SLIT1 in avian song nuclei.....	109

5.3.3 ROBO1/SLIT1 expression in juvenile zebra finch brain.....	112
5.4 Brief discussion.....	116
Chapter 6. Discussions and future directions .....	120
6.1 Comparison of dN/dS and SNAAP analyses .....	120
6.2 New insights into vocal learning evolution in primates.....	122
6.3 Co-option of the ROBO1 axon guidance pathway in vocal learning.....	126
6.4 Non-exclusive SNAAP substitutions in vocal learners.....	129
6.5 The legacy of domestication and early human evolution.....	132
6.6 RNA editing as a potential mechanism for creating SNAAPs.....	135
6.7 Future directions.....	136
6.7.1 Extensions of the computational approach .....	136
6.7.2 Exploration on more domesticated species .....	138
6.7.3 Experimental validation of genes with H+VL type 1 SNAAPs .....	139
References .....	142
Biography .....	163

## List of Tables

Table 1: The species used in this study. All sequences were obtained from Ensembl v49, except that the zebra finch genome was from Ensembl v53, the Neanderthal genome was from the UCSC browser (alignment sequences with human sequences from NCBI build 36/hg18), and the parrot sequence was from the ongoing genome sequencing efforts in our lab. * Genomic sequences made available by the Mammalian Genome Project.....	17
Table 2: The 24 genes with dS-based trees violating the known species tree.....	28
Table 3: The genes under positive selection or accelerated evolution shared among the three vocal learners, relative to cow, dog and (A) chimpanzee or (B) monkey. Underline: significant sharing of AE genes among the three vocal learners. P values: (A) 6.7E-8 or (B) 1.5E-8. Asterisk: overlapping AE gene in both primate tests. ....	32
Table 4: The numbers of type 1-4 putative SNAAP sites and SNAAP sites identified for different species trios in Hom-Mac and Hom-Pan tests, respectively. ....	44
Table 5: The percentages of sites removed by step 2, step 3 or step 4, respectively, in all sites removed. The results for three example species trios for the Hom-Mac test are shown.....	44
Table 6: The average percentages of all species that have the same amino acid types as the species of interest at the putative SNAAP sites removed and not removed by step 3, respectively. Results for three example species trios for the Hom-Mac test are shown...	46
Table 7: Identified genes with H+VL type 1 SNAAPs in Hom-Mac and Hom-Pan tests.	73
Table 8: Groups of genes with type 1 SNAAPs for each trio of species that are found with significantly enriched gene ontology terms. ....	77
Table 9: Genes under PS/AE/RSC with H+VL type 1 SNAAPs. NHP, the non-human primate used in inferring PS/AE/RSC events. Outgroups, the outgroup species used in inferring PS/AE/RSC events. Events, the PS/AE/RSC in specified lineages. # Events, the number of PS/AE/RSC events inferred in specified lineages. # Overlaps, the number of genes with both PS/AE/RSC events and type-1 H+VL SNAAPs. P-values, assessment of the significance of overlap using Fisher Exact Tests. P-value in the parenthesis, assessment after removing the PLEKHH1 whose SNAAP is a human SNP variant. ....	80

Table 10: Ancestral reconstruction results of H+VL type 1 SNAAPs. Position, the SNAAP site at the corresponding translated protein sequences. Ancestor, the amino acid type in the common ancestor of placental mammals. VL, the amino acid type in the three vocal learners. # total species, the number of species used in ancestral reconstruction. Category, the "Gain" or "Loss" category type..... 83

Table 11: Significantly enriched amino acid substitutions in each species trio..... 88

## List of Figures

Figure 1: Phylogenetic trees (A) of mammals and (B) of birds respectively (Murphy et al. 2001; Murphy et al. 2007; Hackett et al. 2008). Red: vocal learning species. Dashed lines: branches of outgroup species drawn not to scale for display purposes. Trees are drawn in unrooted form to highlight the early branching that may indicate independent gains or losses of traits. .... 4

Figure 2: Summary diagrams of (A) vocal learning systems in songbirds and (B) the proposed pathway in humans. Both possess forebrain pre-motor circuits, including cortico-striatal-thalamic loops (black and white lines) and a direct forebrain projection to phonatory motor neurons in the brainstem (red arrows). The direct cortico-bulbar projection is absent in vocal non-learners such as (C) chickens and (D) monkeys. All diagrams show the sagittal view. Brain regions and connections in the same color indicate homology of general brain subdivisions or vocal areas. Dashed lines indicate proposed connections. For reviews see (Jarvis 2004; Jürgens 2009). H: hindbrain; M: midbrain; T: thalamus; FMC: face motor cortex; PFC: prefrontal cortex; ProM: promoter area; XII: hypoglossal motor nucleus; Amb: nucleus ambiguus; RA: robust nucleus of arcopallium; HVC: high vocal center; LMAN: lateral magnocellular nucleus of the anterior neostriatum; AT: anterior thalamic nucleus; Hp: hippocampus; PAG: periaqueductal gray; ASt: anterior striatum; RF: reticular formation; V: ventricle; VL: ventral lateral thalamus; DM: dorsomedial nucleus of ICo; DLM: medial nucleus of the dorsolateral thalamus. .... 7

Figure 3: (A) The orthologous transcript identification pipeline. The labels: START, END or (a) - (d) correspond to the numbering of steps described in section 2.2.2. (B) The distribution of protein sizes of the human proteome and the proteins encoded by the human transcripts after my orthology identification. .... 21

Figure 4: Overview of the SNAAP approach, using the trio of vocal learners as an example. In step 1, different symbols indicate different amino acid types at the site, filled or open symbols indicate the same or different amino acid types in the trio of species, respectively. In step 2, each  $X_i$  denotes an amino acid type from a species in the alignment, and those in red are labeled as being from vocal learners. .... 38

Figure 5: The z-scores for randomly generated amino acid alignments in step 2. The x-axis denotes the number of species whose amino acid type information was available for an alignment. Red dashed line: the z-score cutoff (-1.96), corresponding to  $p=0.025$ . Values below this line are considered significant. .... 46

Figure 6: Density distributions and plots of the Phred quality scores of type 1 SNAAPs for vocal learners trio. Dots on x-axis are plotted phred scores of each SNAAP in each species and the curves are approximated density plots of these scores. Dots or curves in different colors indicate the phred scores of the three nucleotide at the SNAAP site from different species: human, monkey or chimpanzee, elephant, microbat, cow and dog..... 50

Figure 7: The numbers of type 1-4 SNAAPs identified in each trio of species using either monkeys (Mac; Hom-Mac test) or chimpanzees (Pan; Hom-Pan test) in the pre-selection of species. \* = significantly more SNAAPs of the H+DOM trio than those of other trios (Grubbs outlier test,  $p = 0.0023$  and  $0.0018$ )..... 58

Figure 8: Alignments for example type 1 SNAAPs and surrounding amino acids identified in the screen with the three vocal learners. (A) ROBO1; (B) E2F3; (C) PARP1; (D) CASP8AP2. Dots in alignments indicate the same whereas letters indicate different amino acid residues relative to human. Listed are the average entropy scores for the 20 amino acids surrounding the SNAAPs and sequence identities for displayed alignments except SNAAP sites. Arrows, the SNAAP site. Residues in yellow, the human substitution. Species in red, known vocal learners. Diagrams of protein structure with predicted domains by PROSITE ([expasy.org/prosite](http://expasy.org/prosite)) are displayed for ROBO1, PARP1 and CASP8AP2. The dyslexia and/or speech sound disorder susceptible regions, DYX2 and D6S109-D6S506 (locations indicated by vertical lines), that contain E2F3 and two known dyslexia susceptibility genes (DCDC2 and KIAA0319, colored in red) are also shown..... 60

Figure 9: The mean Shannon entropy scores of the 20 amino acids (aa) surrounding type 1-4 SNAAPs for each species trio (symbols/colors), when using either monkey (Hom-Mac test; closed symbols) or chimpanzee (Hom-Pan test; open symbols) as the NHP control. Arrows, the H+VL trios in the Hom-Pan and Hom-Mac tests..... 62

Figure 10: Mean Shannon entropy scores of the 20 amino acids surrounding type 1 SNAAPs in each trio that contains human (Hom.A.B) compared to their swap control trios ((Mac.A.B) or (Pan.A.B)), where A and B = species listed on the x-axis. \* = significantly lower scores for the H+VL trio (Hom.Ele.Mic) ( $p = 0.046$  for Mac.A.B comparison, blue bars;  $p=4.8E-4$  for Pan.A.B comparison, brown bars; Mann-Whitney U test). Error bars: standard error..... 63

Figure 11: Diagonal plots of the percentages of type 1 SNAAPs whose surrounding 20 amino acid entropy scores are lower than specified cutoffs (0.2, 0.4, 0.6, 0.8; indicated in upper left graphs), in both (A) the Hom-Pan test and (B) the Hom-Mac (bottom panel)

tests. Shaded regions, much higher proportions of SNAAPs in the (Human.A.B) trio than in their swap control trios, i.e. far off the diagonal. .... 65

Figure 12: The Mac:Pan ratios of the number of type 1-4 SNAAPs for each species trio. \* = significantly lower Mac:Pan ratio (Grubbs outlier test,  $p = 0.038$ ). Dashed line, average Mac:Pan ratios across all SNAAP types, without outlier. Shaded area, the ratio range (3.0-8.5) of divergence time estimates of human-monkey versus human-chimpanzee from their common ancestors from multiple studies (Steiper and Young 2006). Such a large range reflects uncertainties across studies..... 68

Figure 13: The Hom:NHP swap ratios of the number of type-1 SNAAPs for the human-monkey and human-chimpanzee test. .... 69

Figure 14: Analysis of nucleotide substitutions underlying type 1 SNAAPs. (A) Frequencies of all six single nucleotide changes (bidirectional) for type 1 SNAAPs from each species trio. Human-chimpanzee (Hom-Pan) test is above the x-axis and human-monkey (Hom-Mac) is below. (B) Transversion:Transition (Tv:Ts) ratios of nucleotide substitutions for SNAAPs in each species trio that contains humans and two non-primates. \* Grubbs outlier test,  $p = 0.029$  (Hom-Mac test) and  $1.1E-4$  (Hom-Pan test). (C) The Tv:Ts ratios in the control analyses that put dog and two other species on one side, and the rest and cow on the other side. .... 89

Figure 15: Frequencies of all six single nucleotide changes for non-synonymous mutations in the 1st or 2nd codon positions influenced by (A) codon usage bias in humans, mice and pigs or (B) the putative preference of similar amino acid properties: polar, H-bond forming or active site involvement..... 90

Figure 16: Ratios of the percentage of the Gain or Loss category type 1 SNAAPs with the same human substitutions in zebra finch versus in chicken. A ratio  $\sim 1$  indicates no difference (dashed line),  $> 1$  indicates enrichment of such substitutions in zebra finch, and  $< 1$  would be enrichment in chicken. \* Grubbs outlier test,  $p = 0.05$ . .... 93

Figure 17: The expression patterns of ROBO1/2 and SLIT1/2/3/ in the RA nucleus and surrounding areas of adult male zebra finches. (A) The anatomic location of RA nucleus; (B) ROBO1; (C) ROBO2; (D) SLIT1; (E) SLIT2; (F) SLIT3. A: arcopallium. White arrows: RA nucleus. .... 107

Figure 18: The expression patterns of ROBO1/2 and SLIT1/2/3/ in the XII nucleus and its surrounding areas of adult male zebra finches. (a) The anatomic location of XII nucleus;

(b) ROBO1; (c) ROBO2; (d) SLIT1; (e) SLIT2; (f) SLIT3. White arrows: XII nucleus. Yellow arrows: SSP nucleus..... 108

Figure 19: (A) The expression pattern of ROBO1 and SLIT1 in the frontal sections of the arcopallium (yellow dashed lines) of avian vocal learners and vocal non-learners. (B) Quantification of ROBO1 and SLIT1 expression in the song nuclei vs. surrounding arcopallium in songbirds, parrots, hummingbirds, ring doves and quails. \*  $p < 0.05$ ; \*\*  $p < 0.01$ ; \*\*\*  $p < 0.001$  (one tailed t tests). error bar: standard deviation..... 110

Figure 20: (A) The high-resolution pictures of nXII and SSP with ROBO1 or ROBO2 expression in zebra finches. Enriched expression of ROBO2 is only seen in nXII. White arrows: nXII. (B) Quantification of ROBO2 expression level in nXII vs. SSP across vocal learning and non-learning birds. \*  $p < 0.05$  (paired t-test). Error bar: standard error.... 111

Figure 21: (A) The expression of ROBO1 and SLIT1 in RA of male and female zebra finches during developmental stages: Day 20, 35 and 65. (B) High resolution pictures of RA nucleus. (C) Quantification of ROBO1 and SLIT1 expression in RA vs. RA surrounding arcopallium in male (filled bars) and female (open bars) zebra finches. \*  $p < 0.05$ ; \*\*  $p < 0.01$ ; \*\*\*  $p < 0.001$  (one tailed t tests). Error bar: standard deviation..... 115

## Acknowledgments

It would not have been possible to finish this thesis without the help from many people. It is my great pleasure here to have the opportunity to thank them all.

First and foremost, I wish to thank my Ph.D. advisor, Dr. Erich D. Jarvis. It has been a challenging but pleasant experience and a cherished memory to be Erich's graduate student. I appreciate all his contributions of time, critiques, suggestions, resources, and funding. I also appreciate his supportive and respectful discussions, flexibility, higher level of thinking and open-mindedness to support the completion of my thesis work. I am grateful to be trained in both areas of computational biology and experimental neurobiology.

I thank the members of the Jarvis lab who have contributed immensely to my personal and professional development at Duke. Particularly, I am grateful to Jason Howard who helped with the molecular cloning, Dr. Miriam Rivas and Dr. Osceola Whitney who helped with some initial experiments, and Andreas Pfenning and Dr. Ganeshkumar Ganapathy for useful discussions. I thank Erina Hara for her help with the cross-species in-situ experiments and Dr. Chun-chun Chen for her help with the developmental experiments on juvenile zebra finches. I am indebted to Dr. Ricardo Rossello for critical reading of a manuscript based on my thesis, discussions, and suggestions for presenting and interpreting the findings. Lastly, I thank Tony

Zimmermann and Theresa Renuart who kept us organized and were always ready to help.

I thank many people outside of the lab, including my committee members for their time and useful comments, Dr. Olivier Fedrigo in the Wray lab for generating some preliminary intriguing results from his promoter region analyses (not included in this thesis work), Dr. Fan Wang in the Department of Cell Biology of Duke for her suggestions on RNA editing and the ROBO/SLIT family, and Dr. Simon Fisher in the Wellcome Trust Center, University of Oxford, UK, and Max Planck Institute for Psycholinguistics, The Netherlands for his critical feedback.

I am also grateful for time spent with many friends who became a part of my life and have made my time at Duke even more enjoyable.

Lastly, I thank my parents Chunmin Wang and Zeng'e Shang who raised me and supported me in all my pursuits, and my lovely wife Huimeng Lei for her love and encouragement. To them, I dedicate this thesis.

## Chapter 1. Introduction

A central question in biology is to understand how novel traits can arise through the accumulation of small heritable mutations during evolution (Griffiths et al. 2007). Past studies have achieved reasonable success for traits whose phenotypic variations are closely correlated with mutations in single genes (Glazier et al. 2002). But most other traits are complex, including many important human diseases, and dissecting their genetic basis has been regarded as one of the grand challenges to modern biology (Glazier et al. 2002; Carlborg and Haley 2004; Valdar et al. 2006; Mackay and Anholt 2007).

Complex traits have no simple genotype-phenotype correspondence. Instead, a single complex trait may be influenced by many genes, known as "genetic heterogeneity". For example, nearly 100 genes affecting pigmentation in mice have been cloned, where changes in different genes, e.g. *Mc1r*, *Asip* and *Oca2*, can cause a similar light-color phenotype or albinism (Hoekstra 2006). Another characteristic of complex traits is the "pleiotropic" effect, where one gene may control many traits. For example, the scribble locus of fruit flies controls the establishment of polarity in epithelial cells during embryonic development (Bilder and Perrimon 2000), bristle number (Lyman et al. 1996; Norga et al. 2003) and olfactory behavior (Anholt et al. 1996; Ganguly et al. 2003). Lastly, non-genetic or environmental factors may interfere with the expression of the phenotype, too (Glazier et al. 2002; Carlborg and Haley 2004; Valdar et al. 2006; Mackay

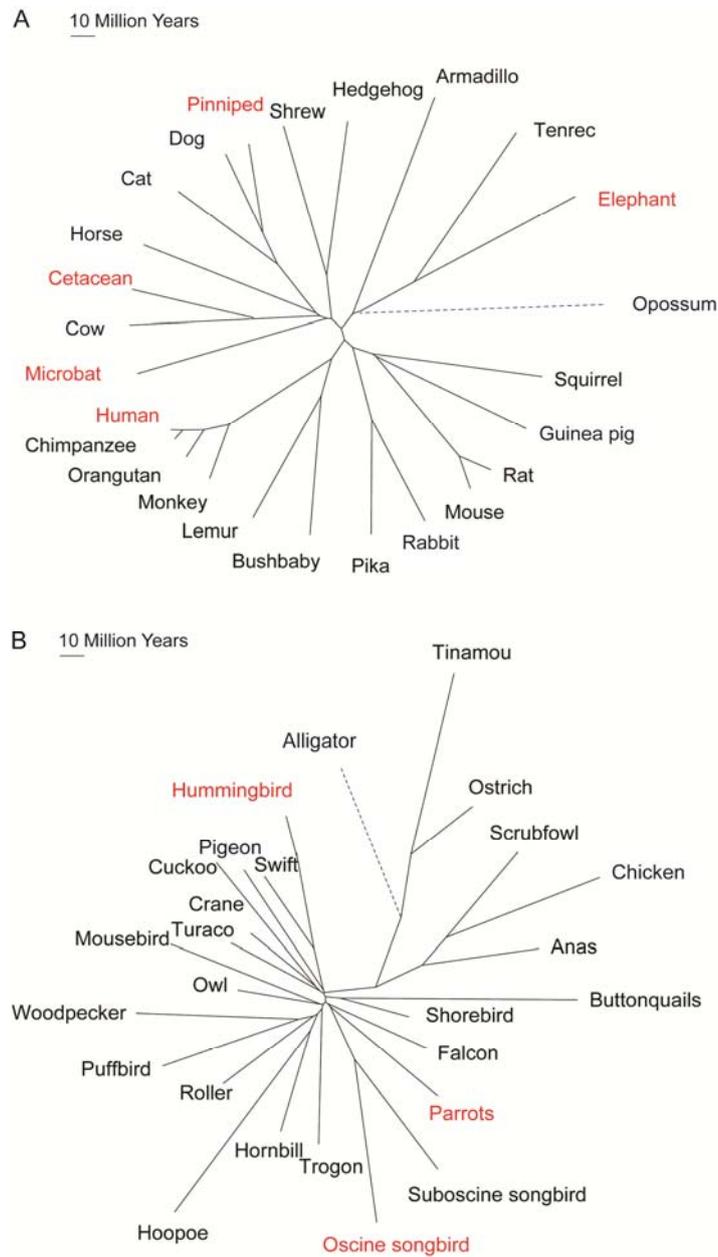
and Anholt 2007). For example, the risk of breast cancer in women with BRCA1 mutations increases from 37% by age 40 to 85% by age 80 (Easton et al. 1993), implying a strong influence from age. Consequently, for most complex traits, there are no general approaches or protocols to study. The challenge is to combine the existing knowledge from all aspects to design toolkits best suited for studying the trait of interest.

### ***1.1 Vocal learning, a complex convergent behavioral trait***

For my thesis work, I focused on a complex behavioral trait, vocal learning. Vocal learning is the ability to imitate sounds and is a critical substrate for human spoken language (Wilbrecht and Nottebohm 2003). In humans, the learned vocalizations include speech or spoken-language. But distinguishing the mechanisms for speech, i.e. production of learned vocalizations, from those for spoken-language, i.e. the ability of vocally combining individual communicative elements and ordering them with meaning, has often been difficult, particularly at the neural level (Fitch 2000; Liberman and Whalen 2000; Jarvis 2004). Therefore, most research on the biology of vocal learning in humans used a combined term, speech-language (Lewis et al. 2006). Besides human, vocal learning so far is only found in four distantly related mammalian groups: cetaceans (Lilly 1965; Foote et al. 2006), pinnipeds (Ralls et al. 1985; Sanvito et al. 2007), elephants (Poole et al. 2005) and microbats (Boughman 1998), and three distantly related avian groups: oscine songbirds (Nottebohm 1972), parrots (Gramza 1970; Pepperberg 1981; Farabaugh et al. 1994) and hummingbirds (Baptista and Schuchmann 1990) (Figure

1). Some non-human vocal learners, such as corvid songbirds and parrots, can imitate human speech and even understand the meaning of the imitated words (Pepperberg 2006). The majority of other vertebrate groups, including those closely related to vocal learners, such as non-human primates (Egnor and Hauser 2004) and sub-oscine songbirds (Kroodsma and Konishi 1991), are considered vocal non-learners, even though some of them, e.g. chimpanzee, can comprehend the meaning of speech sounds with parallels to humans (Heimbauer et al. 2011).

While all vertebrates have the ability of auditory learning to form auditory memories and make sound associations, the vocal learners can further distinguish self-generated vocalizations from others sounds and compare the developing vocal imitations with a memorized template (Brainard and Doupe 2000). This is achieved by auditory feedback, a unique feature that distinguishes vocal learners from vocal non-learners (Konishi 1965). Based on this knowledge, the standard experimental paradigm for testing the presence of vocal learning includes: (1) depriving juveniles of auditory experience from tutor vocalizations by either social isolation or deafening to determine if they develop abnormal vocalizations in adulthood; (2) deafening adults to determine if their vocalizations deteriorate with time; and (3) examining vocal imitation of unrelated individuals or of novel sounds (Janik and Slater 1997).



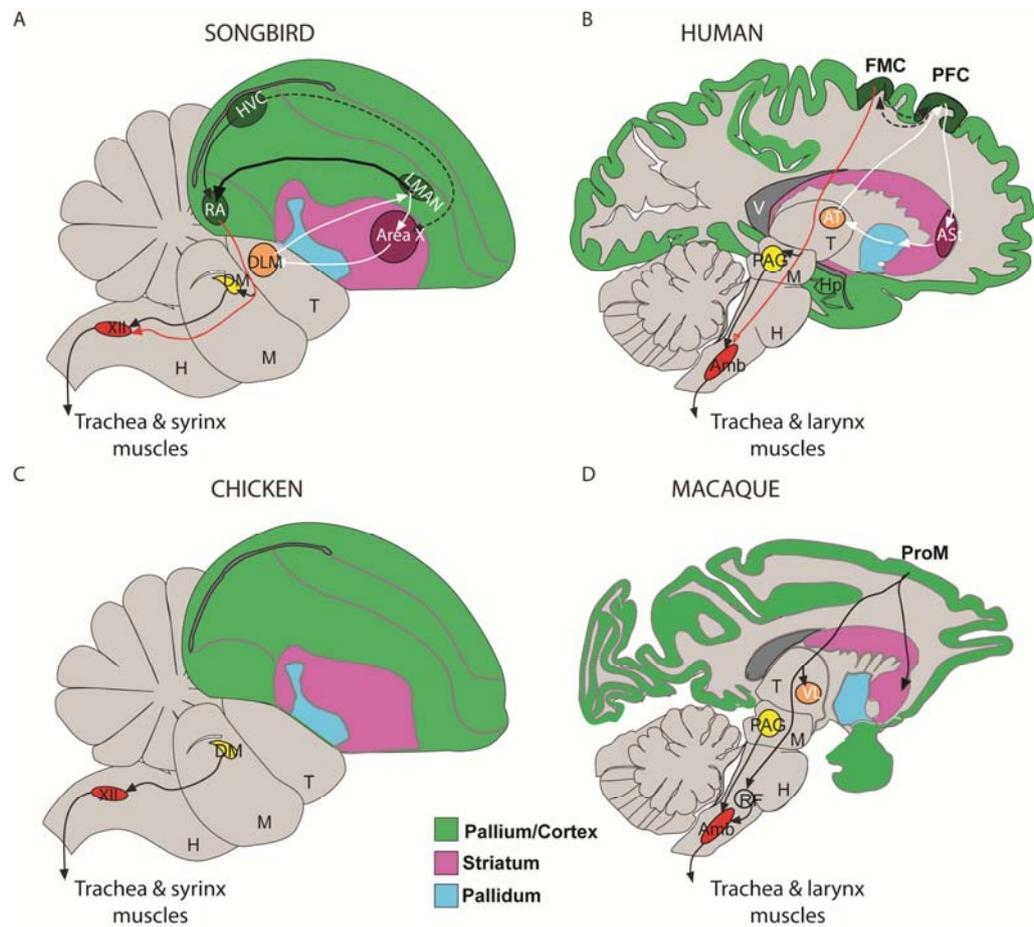
**Figure 1: Phylogenetic trees (A) of mammals and (B) of birds respectively (Murphy et al. 2001; Murphy et al. 2007; Hackett et al. 2008). Red: vocal learning species. Dashed lines: branches of outgroup species drawn not to scale for display purposes. Trees are drawn in unrooted form to highlight the early branching that may indicate independent gains or losses of traits.**

Vocal learning behavior is controlled by specialized forebrain pathways not present in vocal non-learners (Nottebohm et al. 1976; Gahr 2000; Jarvis 2004) (Figure 2). In vocal learning birds, this pathway, called the song system, consists of seven comparable nuclei that make up: 1) an anterior vocal pathway necessary for song learning, which resembles mammalian cortical-basal ganglia-thalamo-cortical loops used for motor learning; and 2) a posterior vocal pathway necessary for song production, which resembles mammalian descending motor pathways used for voluntary movement control. The posterior vocal pathway sends out a unique projection to the brainstem motor neurons, nXIIts, which innervate the avian vocal organ, or syrinx. In humans, there is a proposed analogous anterior cortical-basal ganglia-thalamo-cortical pathway, including Broca's area, the anterior striatum (ASt) and anterior thalamus (AT), and an analogous posterior vocal pathway, including the facial motor cortex (FMC) that makes a direct projection to the vocal motor neurons of nucleus ambiguus (Amb), in the brainstem (Kuypers 1958; Iwatsubo et al. 1990; Jarvis 2004). These forebrain song/speech regions and the direct projection from the cortex to the brainstem vocal motor neurons are not present in vocal non-learning species, including non-human primates and sub-oscine songbirds, chickens and pigeons (Kuypers 1958; Wild 1994; Jürgens 2002; Simonyan and Jürgens 2003; Jarvis 2004).

These forebrain song and speech vocal learning systems receive inputs from a conserved homologous auditory pathway that controls auditory learning in both vocal

learners and non-learners (Jarvis 2004). Similarly, the vocal organ (larynx in mammals and syrinx in birds) as well as its brainstem input (vocal motor neurons and several midbrain regions) are also found in all mammals and birds, respectively, which are involved in producing innate vocalization. No specializations in these brainstem pathway and vocal organs are present in vocal learners (Wild 1994; Jarvis 2004; Jürgens 2009). Therefore, in parallel to the behaviors, the auditory pathway and brainstem vocal pathway in vocal learning birds and mammals are considered to have been inherited from a common ancestor with vocal non-learners. In contrast, the remarkably unique vocal learning forebrain pathways in the three song learning birds, humans and possibly also other mammalian vocal learners are considered to have resulted from convergent evolution (Jarvis et al. 2000; Jarvis 2004).

The convergence is remarkable given that vocal learning mammals diverged at least 80 million years ago (MYA) from a common ancestor and vocal learning birds did so at least 65 MYA (Figure 1). Each vocal learning group is well separated by many closer vocal non-learning relatives, e.g. chimpanzee for humans and sub-oscine songbirds for songbirds. Songbirds and parrots were recently proposed to be close relatives (Hackett et al. 2008), but this is still in debate (Nabholz et al. 2011). Thus, one may ask, if the behavior and associated neural anatomical circuits are highly convergent, what about the genetics?



**Figure 2: Summary diagrams of (A) vocal learning systems in songbirds and (B) the proposed pathway in humans. Both possess forebrain pre-motor circuits, including cortico-striatal-thalamic loops (black and white lines) and a direct forebrain projection to phonatory motor neurons in the brainstem (red arrows). The direct cortico-bulbar projection is absent in vocal non-learners such as (C) chickens and (D) monkeys. All diagrams show the sagittal view. Brain regions and connections in the same color indicate homology of general brain subdivisions or vocal areas. Dashed lines indicate proposed connections. For reviews see (Jarvis 2004; Jürgens 2009). H: hindbrain; M: midbrain; T: thalamus; FMC: face motor cortex; PFC: prefrontal cortex; ProM: promoter area; XII: hypoglossal motor nucleus; Amb: nucleus ambiguus; RA: robust nucleus of arcopallium; HVC: high vocal center; LMAN: lateral magnocellular nucleus of the anterior neostriatum; AT: anterior thalamic nucleus; Hp: hippocampus; PAG: periaqueductal gray; AS: anterior striatum; RF: reticular formation; V: ventricle; VL: ventral lateral thalamus; DM: dorsomedial nucleus of IC; DLM: medial nucleus of the dorsolateral thalamus.**

## **1.2 Limited knowledge on the genetics for vocal learning**

It is difficult to answer the above question, given the limited knowledge of the genetics of vocal learning currently, including for birdsong and human spoken language (Scharff and White 2004). In humans, studies on persons with speech-language deficits, e.g. stuttering, speech-sound disorders, specific language impairment, etc., led to the identification of several susceptible genes, e.g. FOXP2, CNTNAP2, CMIP and ATP2C2 (Newbury et al. 2010). Of them, the FOXP2 gene has been the most extensively studied. It encodes a highly conserved transcription factor that contains a forkhead-box DNA-binding domain, a polyglutamine tract, a zinc finger and a leucine zipper. In humans, mutations of this gene cause a severe speech-language disorder (Lai et al. 2001). In songbirds, its expression in the striatal song nucleus Area X is correlated with song learning critical periods (Haesler et al. 2004; Teramitsu and White 2006; Teramitsu et al. 2010) and knockdown of this gene in Area X leads to an incomplete and inaccurate vocal imitation (Haesler et al. 2007).

More importantly, FOXP2 had a recent accelerated evolution (Arbiza et al. 2006; Nickel et al. 2008) or positive selection (Enard et al. 2002; Zhang et al. 2002; Clark et al. 2003) in the human lineage, resulting in two unique substitutions not found in non-human primates. Another vocal learner, the microbat, exhibited accelerated evolution on this gene, too (Webb and Zhang 2005). Combined with its known functional association with vocal learning, it was tempting to imagine that FOXP2 might have recruited genetic

changes important for vocal learning evolution. But the same human substitutions were not found in other mammalian or avian vocal learners (Webb and Zhang 2005). The current consensus is that FOXP2 plays a general role for neural circuit formation by regulating genes involved in axon guidance (Spiteri et al. 2007; Vernes et al. 2007), but its role is not specialized to circuits for vocal learning (Fisher and Scharff 2009).

In addition, several studies in birds discovered genes that exhibit convergent differential gene expression in song nuclei in at least two vocal learning species. The genes include cell adhesion molecules, e.g. cadherins, neuropilin 1 and plexin A4, expressed in several song nuclei in both songbirds and parrots, with differential patterns not present in quails or ring doves (Matsunaga and Okanoya 2008; Matsunaga and Okanoya 2009). Although these genes have expression changes well correlated with the presence of vocal learning brain nuclei, it is not known yet if such specialized expression patterns are associated with their own genetic changes or those of upstream genes in the same molecular pathway. Therefore, their roles in vocal learning evolution remain obscure.

Nonetheless, these findings do suggest that the evolution of vocal learning abilities may involve the same genes in different species. If so, a systematic comparative study of known vocal learners would help identify the genes. No one has attempted such an analysis before. Therefore, before doing so, I asked if there are other existing

studies that have revealed convergent molecular changes associated with convergent complex traits, and if so, what approaches they used.

### ***1.3 Genetic parallelism underlying convergent complex traits***

Given its genetic heterogeneity, the convergent evolution of a complex trait may arise via different combinations of small polygenic changes in different species. But in many cases, repeated changes on the same genes, i.e. genetic parallelisms, have been reported. For example, despite the fact that hundreds of genes are involved in pigmentation production, most of the convergent color coat/skin variations found in vertebrates are strongly associated with changes in the coding sequence of one gene, MC1R, the melanocortin 1 receptor (Hoekstra 2006). Similarly, convergent pelvic reduction in three-spine stickleback fish and loss of hindlimbs in the manatee, a marine mammal, are suggested to be associated with mutations in the same Pitx1 transcription factor gene (Shapiro et al. 2006; Shapiro et al. 2009). Convergent evolution of the electric organ in two distantly related groups of teleost fishes, the mormyriforms of Africa and the gymnotiforms of South America, involved mutations in the same functional domains of the same Na<sup>+</sup> ion channel protein expressed in electric organ cells (Zakon et al. 2006).

Whether convergent evolution involves genetic parallelism does not seem to depend on the closeness of the different populations or species bearing the trait. For example, for independent electric organ evolution, similar changes are observed within the same protein domains in two distantly related electric fishes that diverged at least

140 MYA from a common ancestor (Alves-Gomes 1999; Zakon et al. 2006). On the other hand, for the more closely related species, threespine and ninespine sticklebacks that diverged at 10-16 MYA, their shared skeletal traits were found due to changes in different genes (Shapiro et al. 2009). One explanation could be that these novel complex adaptations emerged by re-using or extending old homologous anatomical and genetic structures, so-called “deep homology”, which were deeply shared at different degrees (Wray and Abouheif 1998; Shubin et al. 2009).

#### ***1.4 Hypothesis: genetic parallelisms among vocal learners***

The song systems of vocal learning bird lineages were shown to be embedded within a parallel brain pathway that controls movement of body and limbs present in both vocal learning and non-learning birds, leading to a motor theory of vocal learning origin whereby vocal learning circuits are proposed to have arisen independently by duplication out of an old homologous motor learning brain pathway, followed by divergence for the novel vocal learning function (Feenders et al. 2008). Therefore, it is possible that the vocal learning circuits are deeply homologous, being co-opted from the same ancestral substrate in the common ancestor of all vocal learners.

In light of these findings, I hypothesize that the genetic changes for vocal learning should coincide in at least some of the same genes in all vocal learners. Further, some of these genes should be involved in forming the forebrain neural connectivity differences between vocal learners and vocal non-learners.

## **1.5 Current approaches for studying genetic parallelism**

Currently, there are a few methods to dissect genetic parallelisms for convergent traits. One approach is to analyze the genotype-phenotype variations in one species with the trait to identify a list of candidate genes, and then check if there are also genetic changes on these genes in other species with the trait. If the answer is yes, one may analyze the expression patterns or conduct functional studies of the discovered genes across species to see if they share similar patterns or roles. This approach is particularly useful when the trait exhibits wide and quantitative variations in natural populations or can readily undergo genetic crosses in the lab within a species. One can genotype the genetic markers and study their co-segregation patterns with the trait through pedigrees to infer the linkage between the marker(s) and the trait locus. The strength of linkage can be measured by the maximal likelihood estimate of the recombination fraction  $\theta$  when the inheritance model of the trait is known, or by model free (non-parametric) methods, e.g. excess of shared identical-by-descent (IBD) alleles in sibling pairs (Lander and Schork 2006). Successful examples of this approach include the identification of the master genes that control color patterns in butterflies (Joron et al. 2006) and the *Pitx1* transcription factor for independent evolution of pelvic reduction in sticklebacks, manatee, and experimental loss in mice (Shapiro et al. 2004; Shapiro et al. 2006; Shapiro et al. 2009).

In the case of vocal learning, however, there is no known complete loss of the circuits in the natural populations of vocal learning species. In many songbird species, females do not have a vocal learning ability. These females hatch with a song learning circuit that atrophies before song learning starts. High doses of estrogen can prevent the atrophy of the song learning circuit, but cannot induce formation of a vocal learning circuit in males or females of vocal non-learners (Gurney and Konishi 1980). Even if it is possible to perform genetic crosses to create the desired loss of the vocal learning phenotype in the lab, it is much more difficult to measure such a complex behavioral trait compared to others like skin color, limb loss, or eyes, as the phenotypic variations are not obvious and the trait is controlled by structures internal to the brain. Therefore, the implementation of genetic mapping in a vocal learning species may not be effective.

An alternative way to obtain a candidate list of genes is by educated guesses. For example, in the electric organ evolution studies, the authors focused on six Na<sup>+</sup> ion channel genes that had been previously shown relevant for electric organ discharge patterns (Zakon et al. 2006). Expression analyses of these genes in electric and non-electric fishes led to the discovery of the Nav1.4a gene that exhibits electric organ specific expression in all electric fishes. Coincidentally, repeated changes within the same functional domains of this gene were found in both electric fish groups, which then established the correlation between genetic changes and the expression patterns of this gene in distantly related electric fishes. However, this strategy might also have a

low success rate in studying vocal learning. Compared to the relatively dedicated electric organ whose primary structural unit is the myogenic cell, the specialized forebrain circuits for learned vocalization are more complex, being embedded within the brain and involving many neural and non-neural cell types, all organized by long-distance neural connectivity.

Another approach is large scale gene expression profiling. Our and other labs have conducted large-scale profiling in birds and primates to reveal either singing regulated genes in song learning circuits, or specialized expression in song learning circuits and speech brain areas (Cáceres et al. 2003; Oldham et al. 2006; Wada et al. 2006; Li et al. 2007; Benjamin et al. 2008; Lovell et al. 2008; Oldham et al. 2008; Pinaud et al. 2008; Naurin et al. 2011). Although a few candidate genes are starting to emerge, this approach requires knowing the brain regions in advance and may need material from multiple developmental stages, which is costly in general. Both are problematic for most vocal learning mammals.

A variant of the educated guess approach is to focus on genes with significant non-neutral genetic changes between species with the trait and species without the trait. With the increasing capability and better quality of genome sequencing technology, it is now possible to conduct a genome-wide comparison to get a complete list of such genes, which may underlie many species-specific adaptations, hopefully including the trait of interest. Current analysis methods in this approach are heavily based on computing the

relative rates of nucleotide mutations accumulated at synonymous (resulting in no amino acid change) and non-synonymous (resulting in amino acid change) sites (Nielsen 2001). As synonymous mutations are neutral or at least have a much smaller effect on fitness than non-synonymous mutations, a gene with an excess of non-synonymous mutations, i.e. with significantly increased dN/dS ratio (dN, non-synonymous mutation rate; dS, synonymous mutation rate) relative to other lineages, is considered to be under accelerated evolution (abbreviated as AE). Moreover, if the dN/dS ratio is greater than 1, the gene is further inferred to be under positive selection (abbreviated as PS), i.e. selection for functional diversification which increases the fitness. Multiple cases of genes under AE or PS are found relevant to the species-specific trait of interest, including the Nav1.4a gene for convergent electric organ discharge patterns (Zakon et al. 2006), the FOXP2 gene for speech-language deficits and bat echolocation (Zhang et al. 2002), and the MCPH1 gene that regulates brain size during development (Evans et al. 2005).

While dN/dS analysis focuses on coding region changes, there are also some recent attempts to detect non-neutral changes in non-coding regions (Bush and Lahn 2005; Haygood et al. 2007; Bush and Lahn 2008). However, non-coding regions are far less conserved across species. It is challenging to obtain their alignment and requires sequenced genomes of both high quality and a close evolutionary relation (Haygood et

al. 2007). Therefore, these constraints restrict the usage of non-coding region analysis for convergent complex traits across distantly related species.

During my dissertation research tenure, intermediate coverage for genomes of many mammalian and non-mammalian species became available, including three mammalian vocal learners (human, elephant, microbat) and several close relatives (e.g. non-human primates). The complete list of species used in this study is in Table 1. With this sequence availability, in my thesis work, I started by conducting a genome-wide dN/dS analysis in vocal learners to detect non-neutrally selected genes and assessed their relevance to vocal learning; I present the results in Chapter 2. Then I developed a novel approach (the SNAAP approach) that aimed to detect genes with more subtle changes associated with convergent evolution in Chapter 3, explored their associations with vocal learning and a surprising association with domestication, and analyzed their underlying evolutionary scenarios in Chapter 4. In Chapter 5, I present experimental findings on one of the discovered genes, *ROBO1*, and its associated molecular partners from the same axon guidance pathway. Lastly, in Chapter 6, I discuss the overall findings, highlight their implications in understanding vocal learning evolution, as well as natural vs. artificial selection, and list a few proposed ideas for future directions.

**Table 1: The species used in this study. All sequences were obtained from Ensembl v49, except that the zebra finch genome was from Ensembl v53, the Neanderthal genome was from the UCSC browser (alignment sequences with human sequences from NCBI build 36/hg18), and the parrot sequence was from the ongoing genome sequencing efforts in our lab.**

**\* Genomic sequences made available by the Mammalian Genome Project.**

Scientific name	Common name	Domestication
Species used in the SNAAP analysis		
<i>Bos taurus</i>	Domesticated cattle	10000 yrs
<i>Canis familiaris</i>	Domestic dog	17000 yrs
<i>Cavia porcellus</i> *	Guinea Pig	7000 yrs
<i>Dasypus novemcinctus</i> *	Nine-banded armadillo	
<i>Echinops telfairi</i> *	The lesser hedgehog tenrec	
<i>Equus caballus</i>	Domestic horse	6000 yrs
<i>Erinaceus europaeus</i> *	Western European hedgehog	
<i>Felis catus</i> *	Domestic cat	9500 yrs
<i>Homo sapiens</i>	Human	
<i>Loxodonta africana</i>	African elephant	
<i>Macaca mulatta</i>	Rhesus monkey	
<i>Microcebus murinus</i> *	Mouse lemur	
<i>Monodelphis domestica</i>	Grey short-tailed opossum	
<i>Mus musculus</i>	Laboratory mouse	recent
<i>Myotis lucifugus</i> *	Microbat; Little brown bat	
<i>Ochotona princeps</i> *	American pika	
<i>Oryctolagus cuniculus</i> *	European rabbit	1400 yrs
<i>Otolemur garnettii</i> *	Bushbaby	
<i>Pan troglodytes</i>	Chimpanzee	
<i>Pongo pygmaeus</i>	Sumatran orangutan	
<i>Rattus norvegicus</i>	Brown rat	recent
<i>Sorex araneus</i> *	European shrew	
<i>Spermophilus tridecemlineatus</i> *	Thirteen-lined ground squirrel	
Species used in other analysis of this study		
<i>Anolis carolinensis</i>	Anole lizard	
<i>Gallus gallus</i>	Red jungle fowl	8000 yrs
<i>Homo neanderthalensis</i>	Neanderthal	
<i>Melopsittacus undulatus</i>	Common pet parakeet	recent
<i>Taeniopygia guttata</i>	Zebra finch	recent
<i>Tursiops truncatus</i> *	Dolphin	

## **Chapter 2. dN/dS analysis for dissecting the genetics of vocal learning**

### ***2.1 Introduction***

The available genomes of mammalian vocal learners and vocal non-learners (Table 1) allow a genome-wide comparison to reveal genes with elevated dN/dS ratios in vocal learners. However, since in many cases the evolutionary distance between a vocal learner and a vocal non-learner is large, the resultant gene sets are usually too large to test experimentally. In one study, the analysis of 13,198 genes revealed 108 genes in humans and 577 genes in chimpanzee under positive selection (PS) since their divergence (Arbiza et al. 2006). In other words, this means that over 5% of the genes shared by these two species have been positively selected since ~5 MYA. Considering that the vocal learners diverged from each other at least ~65 MYA, the number of such genes would be even higher, not to mention that PS genes are only a subset of genes with elevated dN/dS ratios. It is infeasible to experimentally test all these genes to see if they are part of the genetic parallelism or relevant to the vocal learning trait.

To tackle this issue, my strategy is to first identify genes with elevated dN/dS ratios in each vocal learner, i.e. the AE genes (including those PS genes), then to test if different vocal learners share a statistically significant amount of such genes, and if true, then to determine if any of the shared genes have neural functions relevant to vocal learning behavior.

## **2.2 Materials and methods**

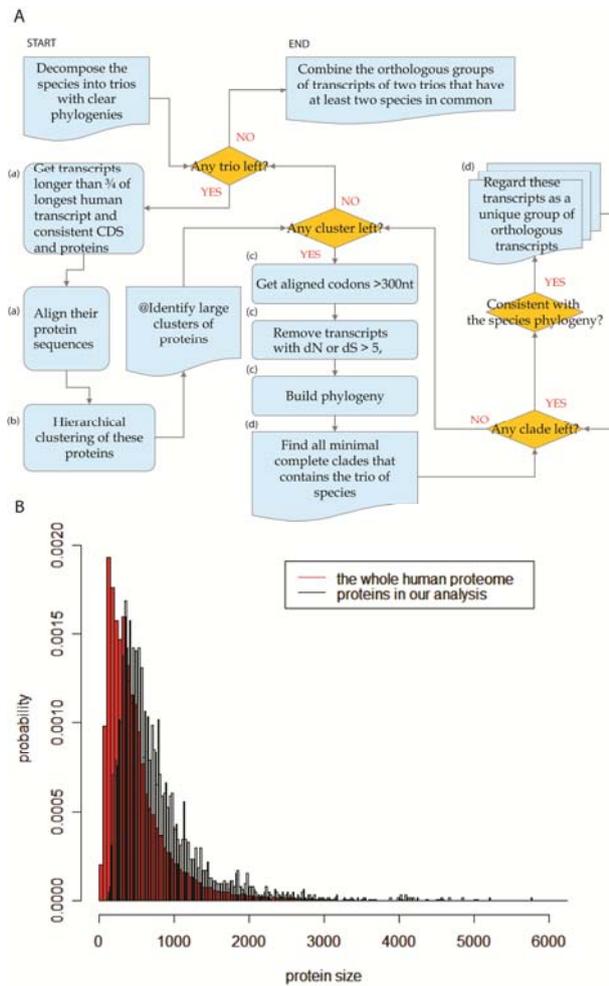
### **2.2.1 Selection of species and sequences**

I chose three vocal learners: *Homo sapiens* (humans), *Myotis lucifugus* (microbats) and *Loxodonta africana* (elephants); and four vocal non-learners: *Pan troglodytes* (chimpanzees), *Macaque mulatta* (monkeys), *Canis familiaris* (dogs), and *Bos taurus* (cows) for this dN/dS analysis. All seven species are terrestrial mammals. I did not include dolphin, another mammalian vocal learner, since it lives in a very different environment, marine, and can emit learned sounds without using the larynx (Mackay and Liaw 1981). Also, its genome at the time was not released to the public. The other four species were chosen because they produce much simpler vocalizations, their genomes have been sequenced at relatively higher coverage (>5x coverage) than most other mammals, and they are widely recognized as vocal non-learners (Egnor and Huaser 1991). The two different non-human primate species (chimpanzee and monkey) were used to perform two parallel tests with each primate to cross-validate the results. I did not include mouse or rat because they can produce complex ultrasonic vocalizations that are suspected to be learned (Holy and Guo 2005), which is still in debate in light of the recent findings from studies in our (Arriaga et al submitted) and other labs (Grimsley et al. 2011; Kikusui et al. 2011). I did not include more mammalian species, not only because they have genomes with lower coverage (2x), but also because given their incomplete genomes, the more species included, the fewer orthologous genes would be

available for systematic analyses in the study. From the seven species, I identified 6678 genes with one-to-one orthology relationship among the chosen species as predicted by ENSEMBL v49 (<http://mar2008.archive.ensembl.org>) and downloaded all transcripts.

### **2.2.2 Pipeline for identifying orthologous transcripts**

Using the above orthologous genes and their alignments of protein coding sequences (CDS), I developed a stringent pipeline to obtain orthologous transcripts (Figure 3A). The pipeline involves two key steps from the PhyOP's algorithm that proved effective in sorting orthologous and paralogous transcripts from the same gene in mammals (Goodstadt and Ponting 2006): the hierarchical clustering of transcripts based on sequence identity and the use of dS (synonymous mutation rate) based trees. The original PhyOP algorithm was invented to only handle 3-4 species at a time. To adapt it to my study for more species, I decomposed all the seven species into trios with clear phylogenies like ((Hom, A), B), where A is chimpanzee if B is monkey, or A is monkey if B is any of the remaining species. I then processed each trio through a series of steps (step a-d) described below, and combined the orthologous transcripts identified from any two trios, e.g. ((Hom, A), B) and ((Hom, A), B'), if they contained the same transcripts in both Hom and the other species. The steps of processing of each species trio consist of the following:



**Figure 3: (A) The orthologous transcript identification pipeline. The labels: START, END or (a) - (d) correspond to the numbering of steps described in section 2.2.2. (B) The distribution of protein sizes of the human proteome and the proteins encoded by the human transcripts after my orthology identification.**

(a) Get alignments: Coding sequences (CDS) and translated protein sequences of each group of orthologous genes in the trio of species from Ensembl. Those that have inconsistent lengths between CDS and protein sequences or whose transcripts were shorter than 3/4th of the longest human transcript were excluded. The remaining transcripts of each gene were aligned using Dialign2 (Subramanian et al. 2008).

(b) Hierarchical clustering: To group transcripts with similar sequence identities into large clusters, I defined a similarity score  $S(i, j)$  between two individual transcripts  $i$  and  $j$  by the percentage of their aligned part in the longer transcript. I then defined a single linkage similarity score between two groups of transcripts  $[A]$  and  $[B]$  by the smallest  $S(A_i, B_j)$ , where  $A_i$  and  $B_j$  are any transcripts from  $[A]$  and  $[B]$  respectively. All transcripts are then clustered hierarchically by repeatedly combining the closest two transcripts or groups, until there is no group with a single linkage distance over 60%. Singletons were assigned to their closest group with at least a 50% similarity score. These two percentage cutoffs were previously used in the original PhyOP algorithm and proved effective in analyzing mammalian sequences.

(c) Build dS-based phylogenies: For each cluster of transcripts identified above, I obtained their alignment of codon sequences. Any columns of the alignment with gaps were removed. After the removal of gaps, if the remaining aligned codon sequences are shorter than 300 nucleotides, i.e. 100 codons, they were excluded from subsequent analyses in order to avoid the potentially biased dN/dS estimates by short and partial

sequences. The remaining sequences have at least 100 codons aligned in all three species. I then calculated the pairwise synonymous (dS) and non-synonymous rates (dN) of each transcript against the longest human transcript within the same cluster using the codeml program of the PAML package (Yang 2007). Those with either  $dS > 5$  or  $dN > 5$  were removed for putative sequencing or genome assembly errors. Lastly, I calculated all pairwise dS values to build distance-based trees of the remaining transcripts by the Kitsch program of the Phylip package (<http://evolution.genetics.washington.edu/phylip.html>).

(d) Identify complete clades: The dS-based trees of orthologous transcripts should agree with the phylogeny of ((Hom, A), B). I defined a clade as complete if it contained leaves from all three species following the phylogeny topology of ((Hom, A), B) and is not the subset of any other complete clade. I obtained all complete clades on the tree built in step (c). Each clade of sequences is considered an orthologous set of transcripts.

### **2.2.3 Search for genes with convergent elevated dN/dS ratios**

To search for genes with convergent elevated dN/dS ratios among the vocal learning species, I performed a two-step analysis:

First, of the final collection of orthologous groups of transcripts, I first identified genes whose transcript coding sequences have elevated dN/dS ratios (i.e. AE genes, including PS genes) in each vocal learner, e.g. human, relative to three vocal non-

learners (chimpanzee or monkey, cow and dog). I used the improved branch-site model tests implemented in the codeml program of the PAML package to analyze the aligned transcripts (Zhang et al. 2005). If dN/dS in human lineage is significantly greater than 1 and also greater than all other lineages (i.e. the background lineages), I consider this gene under PS in human, which is also a form of AE. If dN/dS is significantly greater than the background lineages but not greater than 1, I consider this gene under AE only in human. The significance of dN/dS relative to background lineages or 1 was assessed by the two branch-site tests implemented in codeml: Test I and Test II, respectively. The cutoffs for the log-likelihood ratio in Test I is 3.37 and in Test II is 2.95, both corresponding to a false positive rate of 5% (Zhang et al. 2005). Similarly, I conducted the above analysis on the other two vocal learners, elephant and microbat to identify their PS and AE genes, respectively.

Second, I assessed if the three vocal learners had a significant sharing of such genes by the log-linear model test ( $p < 0.05$ ), an extension of the Fisher's exact test for more than 2 factors. By using either monkey or chimpanzee in the above tests, I constructed two parallel tests to see (1) if the PS or AE genes are significantly shared among all three vocal learners in both tests, and (2) if the two tests lead to the same shared PS or AE genes in vocal learners.

## **2.2.4 Bayesian phylogenetic reconstruction of gene trees**

To compare results of genes with dS based trees that violated known phylogeny (step c in section 2.2.2), I used PhyloBayes program (Lartillot et al. 2009) to build gene trees based on the coding sequence alignments. This program uses an explicit probabilistic model to account for site specific evolutionary rates. It can alleviate the artifacts due to long branch attraction where rapidly evolving lineages are inferred as closely related. The uni-rate model was used. I ran three independent runs to diagnose the convergence ( $\text{maxdiff} < 0.1$ ) and took the majority rule consensus trees.

## **2.3 Results**

### **2.3.1 Quality control results of orthologous transcripts**

Step (a) of the orthology identification pipeline removed 1247 genes (including the FOXP2 gene) whose transcript sequences were less than 3/4th the size of the longest human transcript in at least one species, mostly due to the low coverage of microbat or elephant genomes, and removed 66 genes that had inconsistent CDS and protein length sequences. Steps (b) and (c) resulted in the removal of 2102 genes for which there is no sequences in at least one species in any clusters of species, possibly due to not having sufficient coverage in several species or incomplete prediction of transcripts by Ensembl. Step (c) also removed 24 genes whose dS-trees violated known phylogeny, i.e. bringing a

non-primate as a closer relative to human than non-human primates, but none of them brought only vocal learners together.

The final collection consisted of 3218 genes, of which 3197 genes contained one group of orthologous transcripts, and 21 genes contained two very different groups of orthologous transcript variants, across all seven species: 3239 orthologous transcripts total. On average, each orthologous group of transcripts contains 1.62 transcripts from humans, 1.35 from chimpanzee, 1.46 from macaque, 1 from elephant, 1 from microbat, 1.27 from dog, and 1.26 from cow. The sizes of their encoded human proteins range from 106 aa to 8995 aa with a median of 597 aa, which is comparable to the entire human proteome, though it only includes about 1/5-1/6 of all possible human genes (Figure 3B).

As an independent validation of my orthology pipeline, I searched for the mouse orthologs of the above identified transcript groups using the same pipeline. If these transcripts are sufficiently conserved, I should find the mouse orthologs for all of them. The reason of choosing mouse is because no rodent species was included in above seven species and mouse has one of the best annotated genomes. As expected, I found almost all transcripts but two (0.06%) had orthologous transcripts in mouse, suggesting my orthology identification pipeline was sufficiently stringent.

### **2.3.2 Genes with dS trees that violate the species tree**

Of the 24 genes excluded for having inconsistent dS-based phylogenies with the species tree (Table 2), four (RTN4IP1, RHBDL2, RAPSN and KCNIP4) brought at least

one non-human vocal learner (elephant or microbat ) closer to human than monkey, but not chimpanzee. All four genes have been implicated for a role in central nervous system function. Particularly, RAPSN regulates nicotinic acetylcholine receptor (AChR) clustering at the motor endplate and deficits in it are causal for congenital myasthenic syndrome type 1d (CMS1D) that is often associated with a fatigable speech (Ohno et al. 2002).

The smaller dS distance between human and the non-human vocal learner than that between human and monkey may be due to a faster evolution throughout the whole sequence of the monkey gene or the two vocal learners' genes, or alternatively, rapid changes on a few sites of the monkey gene. To explore on this matter, I constructed the gene trees for these 4 genes based on their whole nucleotide sequence identity using the PhyloBayes program (Lartillot et al. 2009). I found the resultant gene trees for three of the genes (RAPSN, RTN4IP1, and RHBDL2) were consistent with the species tree but not the dS based tree. This result is consistent with the latter scenario where a few sites in the monkey gene may be under accelerated evolution. Further consistent with this idea, I found all three genes exhibit significant accelerated evolution ( $dN/dS > 1$ ) in monkey relative to human with elephant and microbat as outgroups. In contrast, there is no accelerated evolution in human relative to monkey with cow and dog as outgroups.

**Table 2: The 24 genes with dS-based trees violating the known species tree.**

Gene Ids	Gene Name	Annotation	Memo
ENSG00000169181	GSG1L	Germ cell-specific gene 1-like protein	
ENSG00000197353	LYPD2	Ly6/PLAUR domain-containing protein 2	
ENSG00000118292	C1orf54	Uncharacterized protein C1orf54	
ENSG00000049883	PTCD2	Pentatricopeptide repeat-containing protein 2	
ENSG00000130347	RTN4IP1	Reticulon-4-interacting protein 1	Expressed in brain; Appears to be a potent inhibitor of regeneration following spinal cord injury
ENSG00000153347	FAM81B	Protein FAM81B	
ENSG00000166225	FRS2	Fibroblast growth factor receptor substrate 2	Highly expressed in brain
ENSG00000158315	RHBDL2	Rhomboid-related protein 2	Involved in regulated intra-membrane proteolysis and the subsequent release of functional polypeptides from their membrane anchors. Known substrate: EFNB3
ENSG00000185774	KCNIP4	Kv channel-interacting protein 4	Predominantly expressed in brain, may regulate neuronal excitability in response to intracellular Ca <sup>2+</sup> ; Top GWAS candidates in schizophrenia and bipolar disorder
ENSG00000088766	CRLS1	Cardiolipin synthetase	Patients with Leigh Syndrome were homozygous for P193L in CRLS1; Leigh Syndrome, a early onset developmental and motor recession
ENSG00000166359	WDR88	WD repeat-containing protein 88	
ENSG00000135697	BCMO1	Beta,beta-carotene 15,15'-monooxygenase	
ENSG00000165917	RAPSN	43 kDa receptor-associated protein of the synapse	Thought to anchor or stabilize the nicotinic acetylcholine receptor at synaptic sites; CMS1D syndrome (abnormal speech)
ENSG00000020633	RUNX3	Runt-related transcription factor 3	Up-regulated in autism
ENSG00000101079	NDRG3	Protein NDRG3	Highly expressed in brain
ENSG00000139793	MBNL2	Muscleblind-like protein 2	Expressed in brain; May play a role in myotonic dystrophy pathophysiology (DM)
ENSG00000155367	PPM1J	Protein phosphatase 1J	
ENSG00000160087	UBE2J2	Ubiquitin-conjugating enzyme E2 J2	
ENSG00000176715	ACSF3	Acyl-CoA synthetase family member 3	
ENSG00000138463	DIRC2	Disrupted in renal carcinoma protein 2	
ENSG00000112053	SLC26A8	Testis anion transporter 1	
ENSG00000132406	TMEM128	Transmembrane protein 128	
ENSG00000169599	NFU1	NFU1 iron-sulfur cluster scaffold homolog	
ENSG00000136933	RABEPK	Rab9 effector protein with Kelch motifs	

For the fourth gene, KCNIP4, however, neither the dS-tree nor the gene tree was consistent with the ((Hom, Mac), NPM) topology. I found that this is due to a unique KCNIP4-Ia variant in human, elephant and microbat, and the first exon (out of 8 exons) of this variant is skipped in monkey, cow and dog (ENSEMBL v49). If the Ensembl predictions for alternative splicing forms in this gene are accurate and complete, it represents a difference in alternative splicing between these two trios of species (vocal learners versus vocal non-learners). However, I also found this KCNIP4-Ia variant in other primates (chimpanzee, orangutan, mouse lemur and bushbaby) and placental mammals (cat, kangaroo rat, sloth and armadillo) and even marsupial mammals (wallaby). So this variant form is not selective in vocal learners. Considering poor reliability of current methods in predicting splicing variants, it is difficult to test if there are distinctive splicing forms between vocal learners and non-learners in general.

### **2.3.3 Analyses of dN/dS ratios in vocal learners**

On the final collection of 3239 orthologous groups of transcripts from 3218 genes, I identified 3-4 times more genes with AE than PS in each vocal learner (Table 3). Microbat had the most PS/AE genes, followed by elephant, and then human. Relative to the three vocal non-learners, microbat is considered to have diverged ~80 MYA and elephant ~100 MYA, while human diverged no earlier than 40 MYA (Figure 1). This may explain why human has the least PS/AE genes identified. However, it can not explain why microbat has more than elephant. One possibility is the more PS/AE genes in

microbats than in elephant could be part of the changes adapted to the flight, which is unique in microbats relative to all the species examined. This hypothesis needs further testing. The number of identified PS/AE genes in microbat or elephant remain the exactly the same, regardless of using monkey or chimpanzee as the vocal non-learner, suggesting none of these PS/AE changes in these two vocal learners are shared with one non-human primate but not another. In contrast, the number of PS or AE genes identified in humans is ~50% more when using monkey instead of chimpanzee, suggesting there are more shared PS/AE genes between human and chimpanzee than those between human and monkey. This is consistent with the more recent ancestry of human and chimpanzee than that of human and monkey.

I did not find shared PS genes among vocal learners when using chimpanzee as the non-human primate vocal non-learner, and found a non-significant sharing of two genes (CR1 serine protease [C1R] and neuroguidin [NGDN]) when using monkey (Table 3). In contrast, I found a significant sharing of AE genes among vocal learners due to 5 and 8 AE genes using chimpanzee and monkey, respectively. However, only one of these shared AE genes (C4orf21, encoding a membrane protein of unknown function) was found in both parallel tests (Table 3). Only two were neural related (NGDN and AP3B2). NGDN encodes an EIF4E-binding protein that functions as a translational regulatory protein during nervous system development in vertebrates (Jung et al. 2006). AP3B2 is exclusively expressed in neurons and thought to serve neuron-specific

functions such as neurotransmitter release (Grabner et al. 2006). This weak association of AE and neural genes in vocal learners relative to only monkey but not chimpanzee does not make them a convincing case for further studies.

**Table 3: The genes under positive selection or accelerated evolution shared among the three vocal learners, relative to cow, dog and (A) chimpanzee or (B) monkey. Underline: significant sharing of AE genes among the three vocal learners. P values: (A) 6.7E-8 or (B) 1.5E-8. Asterisk: overlapping AE gene in both primate tests.**

(A)

Relative to Chimpanzee, Cow, Dog					
positive selection (PS)			accelerated evolution (AE)		
human	elephant	microbat	human	elephant	microbat
23	99	223	75	393	628
-			<u>APC, CASCL, C4orf21*, MRC2, SUPT6H</u>		

(B)

Relative to Monkey, Cow, Dog					
positive selection (PS)			accelerated evolution (AE)		
human	elephant	microbat	human	Elephant	microbat
35	99	223	109	393	628
C1R,NGDN			<u>C1R, NGDN, AP3B2, C4orf21*, DTX3L, MYO5A, PLIN, TRIM47</u>		

## **2.4 Brief discussion**

The findings of genes with no or weakly convergent lineage-specific PS and AE in the three vocal learners could be explained by a number of reasons: (1) convergent selection may not act on the coding regions to evolve the vocal learning trait; (2) such coding region changes are not present in the collection of 3239 orthologous transcripts studied here; or (3) they exist but cannot be identified by the dN/dS analysis.

The third reason may be associated with three different scenarios: (3.1) the branch-site model test focuses on the whole length of the protein coding region and may be too stringent to reveal the changes for complex traits that are supposedly modest or even weak in each gene; (3.2) some potential sites of genes are under PS/AE in vocal learners while some other sites of the same genes are also under PS/AE in vocal non-learners for other traits, canceling out their detection, as the standard method focuses on lineage specific events. (3.3) there are over 10 times more PS and AE events inferred in elephant and microbat than in human, owing to the greater evolutionary distance of these two species from the remaining species (~80-100 MYA for elephant/microbat versus ~5-25 MYA for human), most of which are supposedly not relevant to the trait of vocal learning and thus could have masked the signals from those relevant events.

Given the above scenarios, the possible role of PS or AE events in vocal learning should not be ruled out, but there are many caveats as to why the dN/dS method for detecting such events may fail.

## Chapter 3. A novel method for detecting trait associations with molecular homoplasy

### 3.1 Introduction

Significant fractions of species-specific amino acid substitutions have been fixed by natural selection. For example, 10-20% of the amino acid differences since the divergence of human and chimpanzee are thought to be fixed by positive selection (Boyko et al. 2008). Moreover, identical genetic substitutions can occur for the phenotypic convergence in different lineages, many of which are found in coding regions. Examples include the various combinations of the same five amino acid substitutions in opsins for similar shifts in the light absorption maximum in vertebrates (Yokoyama and Radlwimmer 2001), identical amino acid replacements at three sites within digestive RNases for lowering its optimal pH in Asian and African leaf monkeys (Zhang 2006; Yu et al. 2010), convergent substitutions in the prestin protein for echolocation in bats, dolphins and whales (Liu et al. 2010) and in the ribulose-bisphosphate carboxylase encoded by the *rbcl* gene for photosynthesis in C4 plants (Christin et al. 2008).

However, most amino acid substitutions may be subject to a small selection coefficient. For example, studies in two *Drosophila* species revealed that about 95% of their amino acid differences were fixed by positive selection, but only with  $N_e * s = 2.5$  on average (Sawyer et al. 2007), where  $N_e$  is the effective population size and  $s$  is the selection coefficient. According to population genetics, mutations fixed with so small a

selective advantage will still be determined largely by random genetic drift (Nei 2005). Therefore, such substitutions would perhaps be missed by the dN/dS analysis, even if they are involved in the evolution of the complex trait.

In light of these findings and even before some of the above studies were published, I developed a dN/dS-free approach that aims to capture subtle changes in amino acid substitutions during complex trait evolution. The approach identifies sites with Single Non-random Amino Acid Patterns (SNAAPs), most of which appear convergent, and evaluates their significance of being non-randomly associated with a given group of species that share convergent complex traits (i.e. the species of interest).

## **3.2 *Materials and methods***

### **3.2.1 Overview of the SNAAP approach**

The SNAAP approach consists of four sequential steps. Sites that passed all these steps are considered SNAAP sites for the species of interest (Figure 4). In the following, I used the vocal learners trio as an example to describe each step in more detail.

- Step 1: systematic screening. I take sequences from a pre-selected group of species with both good genome coverage and clear knowledge of the presence or absence of the convergent trait of interest. I then align the protein sequences translated from their orthologous transcripts and scan each column of the alignments to identify putative SNAAP sites with one of the following four specific substitution patterns (Figure 4):

- Type 1) An identical amino acid (AA) for all vocal learners and a different identical AA for all non-learners;
- Type 2) An identical AA for all vocal learners but more than one AA for all non-learners;
- Type 3) More than one AA for all vocal learners and an identical AA for all non-learners;
- Type 4) More than one AA for vocal learners and non-learners, respectively.

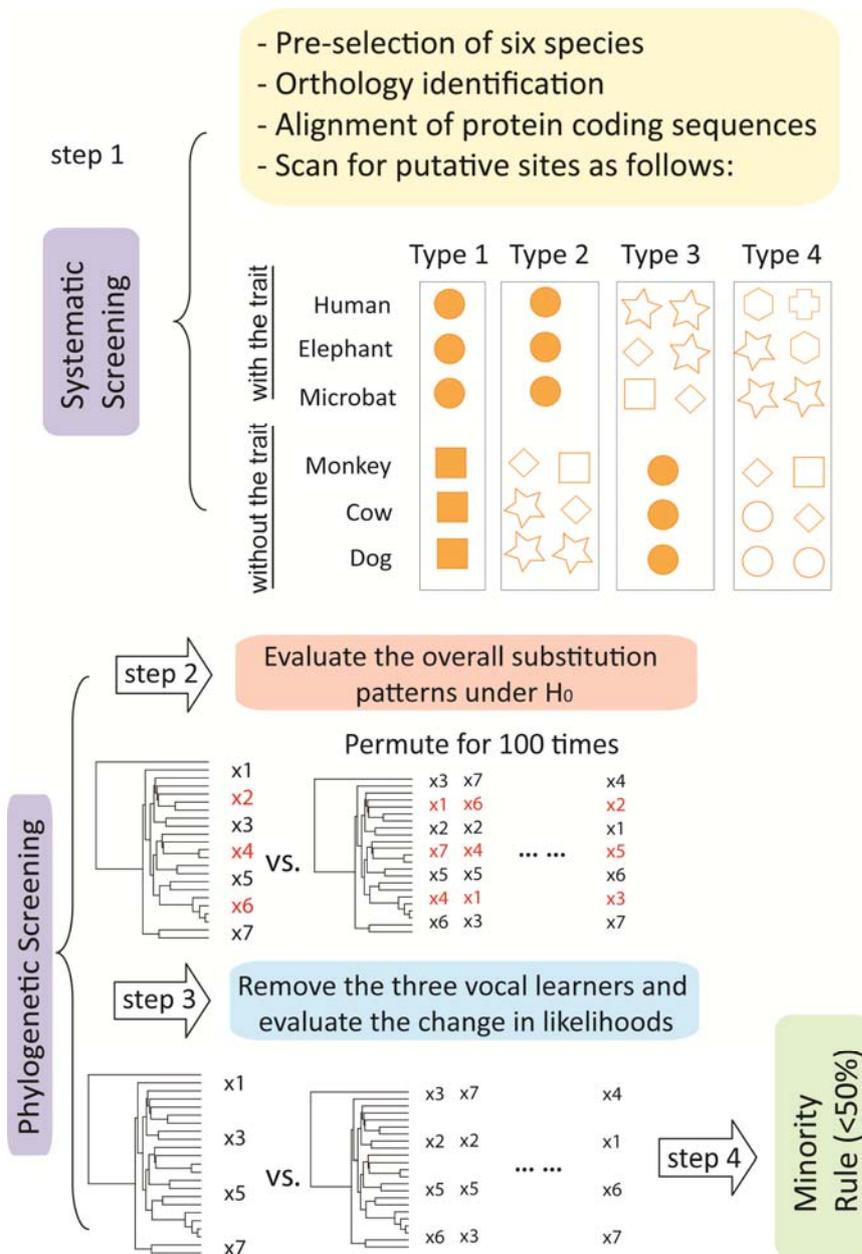
Note that for all four patterns, the vocal learners have distinct amino acid types from the vocal non-learners. Types 1-3 in vocal learners and/or vocal non-learners contain putatively homoplastic substitutions, i.e. the same amino acid types that may not be inherited from their most recent common ancestor. Such homoplastic substitutions have been suggested as a good signature for natural selection (Zhang and Kumar 1997; Rokas and Carroll 2008).

- Step 2: phylogenetic screening for non-random evolution. I determined if the substitutions in all species at each of the putative SNAAP sites identified above were evolved from a random substitution process described by a simple Poisson model. For this purpose, orthologous sequences from more species were added. To do so, for each of the putative SNAAP sites, I obtained the orthologous amino acid types in up to an additional 16 mammalian species (Table 1; Figure 1), which are

supposedly vocal non-learners, using the same orthology identification pipeline described in Section 2.1.2. Depending on sequence availability, each site may have aligned amino acids from a different number of species.

- Step 3: phylogenetic screening for convergent evolution. I assessed whether the substitutions in the species of interest evolved significantly differently from other species.
- Step 4: the minority rule. Considering the convergent traits are most likely rare, I required the overall frequency of the amino acid type(s) in species with the trait to appear in less than 50% of all species.

The resultant sites that satisfy all of the above criteria were identified as type 1-4 SNAAPs, whose evolution can't be explained by the random substitution process and were significantly relevant to the species of interest. Because of the stochastic aspect of the permutation tests, different runs of the above analyses can result in slightly different numbers of SNAAPs (and thus the genes carrying them). I reported those SNAAPs that showed up in at least two of three independent runs.



**Figure 4: Overview of the SNAAP approach, using the trio of vocal learners as an example. In step 1, different symbols indicate different amino acid types at the site, filled or open symbols indicate the same or different amino acid types in the trio of species, respectively. In step 2, each  $X_i$  denotes an amino acid type from a species in the alignment, and those in red are labeled as being from vocal learners.**

### 3.2.2 The evolutionary model and calculation of tree likelihoods

The random substitution process of step 2 and step 3 was described by a simple Poisson model (the null hypothesis  $H_0$ ) with the following assumptions: (a) each amino acid site evolved independently at the same rate of  $10^{-9}$  substitutions per year, which is the average rate in vertebrates (Kumar and Subramanian 2002); (b) the substitution rate between any two amino acid types is set to be the same, which implies time reversibility of this model and thus allows us to approximate the stationary probability of each amino acid type by its observed frequency in extant species. Therefore, the probability at any site for one amino acid mutating into another during an infinitesimal period  $dt$  is:

$\Pr(t + dt) = \Pr(t) (I + Q dt)$ , where  $Q$ , the substitution rate matrix, is the identity matrix. Solving it yields  $\Pr(t) = \exp(Qt) + \text{constant}$ . Here we chose to use the identity matrix as  $Q$  as it is “non-informative” and the simplest choice available, while there are a few other choices, e.g. PAM/BLOSUM matrices (Henikoff and Henikoff 2000), to incorporate certain knowledge.

With above information, the likelihood for the observed substitutions at any given site under the null hypothesis  $H_0$  can be computed. To do this, for each putative SNAAP site, I generated a new tree with the  $m$  available species at this site from the following master species tree summarized from prior studies (Murphy et al. 2001; Murphy et al. 2007) (the numbers below denote branch lengths in million years of divergence from a common ancestor):

(((((Microbat:85,(Cow:82,(Horse:80,(Cat:53,Dog:53):27):2):3):2,(Shrew:68,Hedgehog:68):19):6,((Squirrel:74,(GuineaPig:72,(Rat:15,Mouse:15):57):2):9,(Rabbit:50,Pika:50):33):5,(Bushbaby:61,Lemur:61):17,((Orangutan:14,(Human:5,Chimpanzee:5):9):11,Monkey:25):53):10):5):8,Armadillo:101):4,(Tenrec:79,Elephant:79):26):75,Opossum:180)

The branch lengths of the new tree are generated by mapping the two nodes back to the master tree to identify their divergence time. Each leaf or internal node of the new tree represents an extant species ( $X_i$ ) or ancestral species ( $T_i$ ), which has a 20-dimension vector whose value is the probability of seeing the 20 AA type in this species. For example, for the leaves, i.e. in extant species, this vector has the value of 1 for the observed AA type but otherwise 0.

The overall tree's log-likelihood  $L$  for each putative SNAAP site under  $H_0$  is given by recursively computing the likelihoods of the two direct descendant trees ( $T_1$  and  $T_2$ ) of the root ( $T$ ) using Felsenstein's pruning algorithm (Felsenstein 1981) according to the following equations:

$$L = \log \sum_T \pi(T) \Pr(X_i \dots X_m | T)$$

$$= \log \sum_T \pi(T) (\sum_{T_1} \Pr(T_1 | T) \Pr(X_1 \dots X_m | T_1) + \sum_{T_2} \Pr(T_2 | T) \Pr(X_1 \dots X_m | T_2))$$

where  $\sum_T$ ,  $\sum_{T_1}$  and  $\sum_{T_2}$  denote the marginalization of probabilities over all 20 possible amino acids at the root  $T$  and its two ancestral descendants  $T_1$  and  $T_2$ , respectively,  $\pi(T)$  is the stationary probability of having a specific amino acid type at the root, and  $\Pr(T_1 | T)$

T) is the probability of deriving the amino acid type at T1 from the amino acid type at T over the divergence time between T and T1.

### 3.2.3 Hypothesis testing for the phylogenetic screening step

The phylogenetic screening (step 2 and step 3) consists of the following two hypothesis tests to evaluate the statistical significance of the substitution pattern at each putative SNAAP site using its likelihood under  $H_0$ .

In step 2, the statistical significance of the observed substitutions for rejecting the null hypothesis is evaluated by permuting the species labels 100 times, which is equivalent to permuting their amino acid types, and computing the log-likelihood value after each permutation  $(L_p)_{1...100}$ , the variability of which represents the level of uncertainty associated with the random substitution process. If the observed substitutions were well modeled by the null hypothesis  $H_0$ , no significant difference between  $L$  and  $(L_p)_{1...100}$  should be expected; otherwise,  $L$  should be significantly smaller. As the distribution of  $(L_p)_{1...100}$  can be approximated by a normal distribution for most sites, I retained the sites with z-scores  $(L - \langle L_p \rangle) / \sigma(\langle L_p \rangle) < -1.96$ , where  $\langle L_p \rangle$  is the average of  $(L_p)_{1...100}$  and  $\sigma(L_p)$  is the standard deviation. Roughly, these sites reject the null hypothesis at a significance level of 0.025 (uncorrected for multiple hypothesis testing).

In step 3, to determine if the  $L$  for the SNAAP sites of each tree was weakened by convergence in vocal learners, the three vocal learners were taken out to calculate the new log-likelihoods for substitutions in the remaining species. The resultant difference

$D$  in log-likelihood before and after removing the vocal learners is computed. The hypothesis is that if the SNAAP substitutions in vocal learners are convergent, then the removal of the vocal learners should help fit the null hypothesis  $H_0$  better. However, the removal of species would always lead to a greater log-likelihood value under  $H_0$  due to a reduction in the model parameters. To normalize against this affect, I computed the permuted log-likelihood difference  $(D_p)_{1...100}$  for  $(L_p)_{1...100}$ , where  $(D_p)_{1...100}$  represents the null distribution of changes by randomly removing three species at a time. I retained sites with z-scores  $(D - \langle D_p \rangle) / \sigma(D_p) < -1.96$ , where  $\langle D_p \rangle$  is the average of  $(D_p)_{1...100}$  and  $\sigma(D_p)$  is the standard deviation, which roughly corresponds to a significance level of 0.025 (uncorrected for multiple hypothesis testing).

### **3.3 Results**

I applied the SNAAP approach in two parallel tests, which differ in their pre-selection of nonhuman primate species (using either monkey or chimpanzee) in step 1 (the systematic screening). These are referred to as Hom-Mac test (using monkey) and Hom-Pan test (using chimpanzee), respectively. Each test was performed on six species trios: (human, elephant, microbat), (human, elephant, cow), (human, elephant, dog), (human, microbat, cow), (human, microbat, dog) and (human, cow, dog). In each test, all six trios used the same pre-selection of species: human, elephant, microbat, monkey or chimpanzee, cow and dog, whose orthologous transcripts were identified in the

previous dN/dS analyses, in order to identify the putative SNAAP sites. The above algorithms were coded in Java and run in parallel fashion on the Duke Shared Cluster Resource (DSCR) to speed up the analysis of hundreds of genes or sites simultaneously.

I discovered the presence of all four types of SNAAPs for vocal learners and the other five trios of species in both Hom-Mac and Hom-Pan tests (Table 4). In getting these SNAAPs, the systematic screening (step 1) narrowed the ~1.5 million aligned sites to tens or hundreds of putative sites for each type (Table 4). Interestingly, the phylogenetic screening step 2 hardly removed any sites, while the phylogenetic screening step 3 removed many sites for type 1 and 2 but not type 3 and 4 (Table 5). Moreover, sites removed by step 2 were a subset of those removed by step 3, suggesting step 2 may be superfluous. Lastly, step 4 (the minority rule) removed a considerable number of sites for all types. The sites removed by step 3 and step 4 typically overlapped by 70-80% (of the smaller set). These findings of uneven removal of sites in phylogenetic screening (step 2 and step 3) led us to wonder if the results were due to technical or biological reasons.

**Table 4: The numbers of type 1-4 putative SNAAP sites and SNAAP sites identified for different species trios in Hom-Mac and Hom-Pan tests, respectively.**

Number of putative SNAAP sites (after step 1)								
	Type 1		Type 2		Type 3		Type 4	
	Hom-Mac	Hom-Pan	Hom-Mac	Hom-Pan	Hom-Mac	Hom-Pan	Hom-Mac	Hom-Pan
human,cow,dog	153	79	412	101	272	63	701	126
human,elephant,cow	126	25	283	53	309	60	682	123
human,elephant,dog	137	21	387	76	275	56	676	134
human,elephant,microbat	139	31	329	53	291	73	736	148
human,microbat,cow	149	33	282	46	314	45	692	120
human,microbat,dog	140	29	347	72	262	62	750	133

Number of SNAAP sites (after step 4)								
	Type 1		Type 2		Type 3		Type 4	
	Hom-Mac	Hom-Pan	Hom-Mac	Hom-Pan	Hom-Mac	Hom-Pan	Hom-Mac	Hom-Pan
human,cow,dog	84	59	111	43	209	52	406	79
human,elephant,cow	69	16	91	19	224	50	370	70
human,elephant,dog	59	11	95	18	180	42	361	61
human,elephant,microbat	70	14	92	13	218	57	418	86
human,microbat,cow	80	22	81	18	231	39	418	62
human,microbat,dog	58	12	72	18	190	47	423	72

**Table 5: The percentages of sites removed by step 2, step 3 or step 4, respectively, in all sites removed. The results for three example species trios for the Hom-Mac test are shown.**

	Step 2	Step 3	Step 4	Step 2	Step 3	Step 4	Step 2	Step 3	Step 4
	human, cow, dog			human, elephant, microbat			human, elephant, dog		
Type 1	3%	38%	85%	0%	52%	87%	1%	43%	90%
Type 2	0%	74%	80%	1%	80%	85%	0%	74%	86%
Type 3	2%	2%	98%	0%	2%	100%	0%	7%	98%
Type 4	0%	6%	99%	0%	7%	100%	0%	7%	100%

### 3.3.1 Evaluation of the phylogenetic screening (steps 2 and 3)

To evaluate step 2, I performed a simulation study. I generated a random set of  $N$  species and a random set of stationary frequencies for 20 AA types to randomly assign an AA type to these species. The resulting random alignments of substitutions are a special case of the random substitution process, where the substitutions are independent of the ancestral states and only determined by the stationary probabilities. More importantly, they were not generated from the Poisson model used in the permutation tests which helped avoid circular reasoning. I performed this simulation 100 times. As can be seen in Figure 5, each black dot denotes a z-score calculated from step 2 for a random alignment. These z-scores spread widely across the y-axis but with no obvious correlation with the number of species used on the x-axis. Interestingly, none of them were lower than -1.96, the cutoff I used for the step 2 hypothesis testing, i.e.  $p < 0.025$ . This result shows that step 2 can screen out random substitutions and highly suggests that step 2 of my SNAAP approach is a very stringent statistical test. The fact that almost all putative SNAAP sites passed step 2 in real cases indicate that their evolution has significantly deviated from the random process implied by my Poisson model.

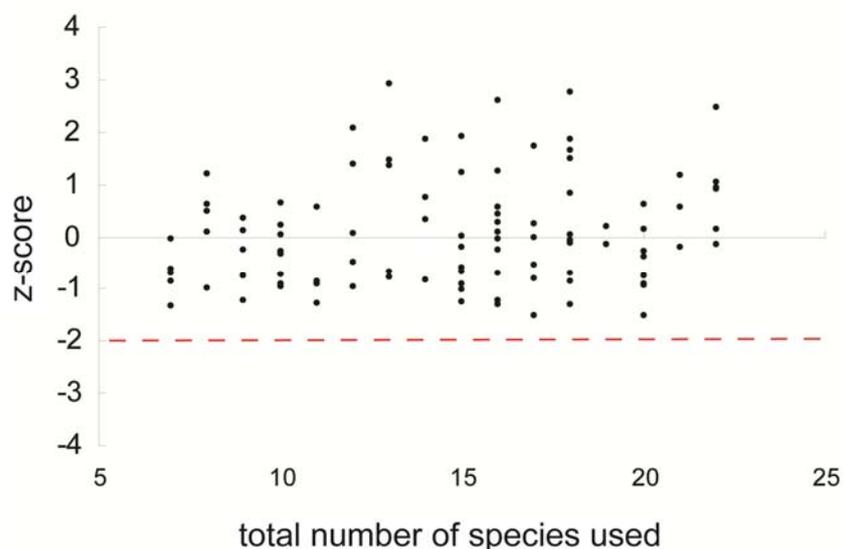


Figure 5: The z-scores for randomly generated amino acid alignments in step 2. The x-axis denotes the number of species whose amino acid type information was available for an alignment. Red dashed line: the z-score cutoff (-1.96), corresponding to  $p=0.025$ . Values below this line are considered significant.

Table 6: The average percentages of all species that have the same amino acid types as the species of interest at the putative SNAAP sites removed and not removed by step 3, respectively. Results for three example species trios for the Hom-Mac test are shown.

	human, elephant, microbat		
Sites	type 1	type 2	type 4
removed	59.50%	60.40%	61.70%
not removed	39.40%	40.20%	44.00%
	human, cow, dog		
Sites	type 1	type 2	type 4
removed	65.5%	60.7%	63.0%
not removed	40.3%	42.4%	44.2%
	human, elephant, dog		
Sites	type 1	type 2	type 4
removed	67.2%	60.3%	63.3%
not removed	44.6%	42.7%	45.5%

To evaluate step 3, I wondered if there was any systematic difference between sites removed and sites not removed. Specifically, if the step 3 was effective in identifying sites whose substitutions are associated with a convergent trait, which is mostly likely rare, the corresponding amino acid types in these species of the identified sites should be rarely seen in all species as well, compared to those sites that were removed by step 3. I tested this idea for the vocal learning trio. Consistent with my expectation, I found the percentages of species having the same amino acid types as vocal learners are significantly lower for sites identified by step 3 than sites removed by step 3 (Table 6; Wilcoxon test; Hom-Mac test p-values:  $1E-11$  for type 1,  $2E-16$  for type 2,  $6E-7$  for type 4; Hom-Pan test p-values: 0.02 for type 1,  $2E-5$  for type 2, 0.02 for type 4). Note that few type 3 sites were removed by step 3, indicating a technical limitation, so that it is hard to make a similar statistical comparison. The similarly significant differences were also seen for other five trios. Table 6 shows the average percentages for three example trios from the Hom-Mac test. Consistent with these results, as mentioned previously, I found the sites removed by step 3 and step 4 typically overlapped by 70-80% (of the smaller set). This finding suggested that the 50% rule in step 4, which was arbitrarily chosen, may be considered an optional step in the future.

### **3.3.2 Low false discovery rates of phylogenetic screening**

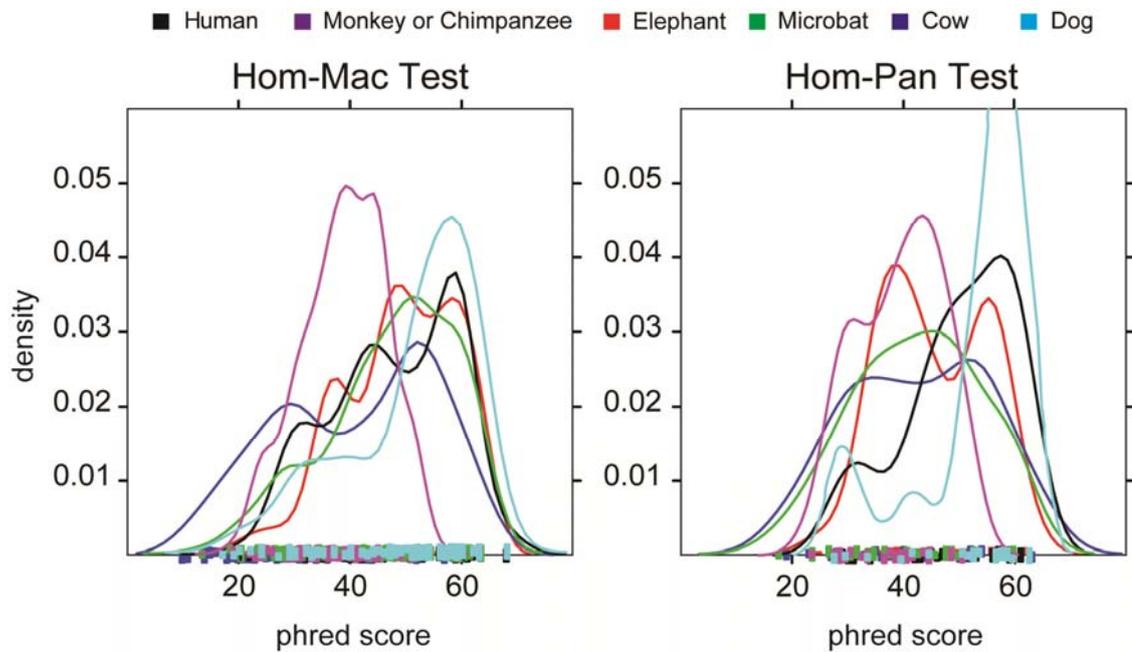
Because of the multiple tests for many sites, the phylogenetic screening step 2 and step 3 may be associated with a certain false discovery rate (FDR). To explore this

matter, I obtained the p-values from step 2 and step 3, assuming that the z-scores, i.e.  $(L - \langle L_p \rangle) / \sigma(L_p)$  and  $(D - \langle D_p \rangle) / \sigma(D_p)$ , follow a unit normal distribution. The larger value of these two p-values was used to compute the false discovery rate following the Simes procedure, which is a popular modification of the Bonferroni correction (Simes 1986; Newson and The 2003). For the vocal learners' trio, the estimated FDR of the phylogenetic screening is 2.8% for type 1 SNAAPs and 4.6% for type 2 SNAAPs in the Hom-Mac test and 2.7% for type 1 SNAAPs and 1.8% for type 2 SNAAPs in the Hom-Pan test. These results suggested the phylogenetic screening was able to control the false discovery rates under the widely used 5% cutoff of FDR for type 1 and type 2 sites. In contrast, the much fewer type 3 and type 4 sites removed by phylogenetic screening made it difficult to reliably estimate the corresponding FDRs.

### **3.3.3 Little influence from sequencing error or SNPs**

I wondered if the identified SNAAPs might result from sequencing errors or naturally occurring single nucleotide polymorphisms (SNPs). I examined the raw read sequences bearing type 1 SNAAPs for the vocal learning trio at <http://trace.ensembl.org/> and found that the vast majority (>98%) of them in all seven species (human, chimpanzee, monkey, cow, dog, microbat and elephant) had phred quality scores greater than 20, i.e. with the base call accuracy > 99% (Figure 6). Only one SNAAP (1.3%) was due to known human SNPs in the PLEKHH1 gene (NCBI SNPdb accession # rs11158685) with two alleles (arginine 'R' and histidine 'H') about equally distributed in

humans, but yet with no reported individuals with homozygous "RR". At this SNAAP site, monkey has an arginine while chimpanzee has a histidine. These findings indicated a minimal influence from SNPs or sequencing errors on the SNAAPs identified.



**Figure 6: Density distributions and plots of the Phred quality scores of type 1 SNAAPs for vocal learners trio. Dots on x-axis are plotted phred scores of each SNAAP in each species and the curves are approximated density plots of these scores. Dots or curves in different colors indicate the phred scores of the three nucleotide at the SNAAP site from different species: human, monkey or chimpanzee, elephant, microbat, cow and dog.**

### **3.4 Brief discussion**

Moving from a dN/dS approach that analyzes whole protein coding sequences to an analysis of individual amino acid sites, the amount of data increased dramatically by several orders of magnitude, from thousands of genes to millions of aligned amino acid sites. Therefore, to detect the shared genetics for convergent evolution, an efficient bioinformatics tool with a clear biological basis was needed. Here my SNAAP approach met this need by using the current knowledge of homoplasy on convergent evolution as a heuristic to only focus on sites with distinct amino acid types in the species of interest. It evaluated each site quantitatively, making use of the well established models and algorithms for amino acid evolution. To this end, it successfully narrowed down to a still sizable list of genes enriched with sites whose substitutions might have been shaped by convergent phenotypic evolution. I showed that this method is stringent and achieves low false discovery rates.

The different performance of step 3 on type 1/2 and type 3/4 sites was likely due to technical limitations. The major difference is that the removed species of interest have only 1 AA type for type 1/2 sites but at least 2 AA types for type 3/4 sites (Figure 4, step 1). Therefore, in the latter case, the removal of more amino acid types from the tree will also automatically improve the new log-likelihood  $L$  so that the resultant log-likelihood difference  $D$  becomes more negative, i.e. the site being less likely to be removed. I

therefore suggest that type 1 and type 2 SNAAPs are more reliably filtered than types 3 and 4 in the phylogenetic screening.

Now it is crucial to conduct a follow-up analysis to further test the functional association of the revealed SNAAPs with the convergent trait of interest. I note that genes that do not bear such SNAAP sites may still play a functional role and may be identified by analyzing the interaction partners of those genes that do bear SNAAPs.

## **Chapter 4. SNAAPs associated with natural vs. artificial selection**

### ***4.1 Introduction***

In this chapter, I evaluated the significance of the SNAAPs associated with vocal learning, explored the functional relevance of genes bearing such SNAAPs to this trait, and compared the results with those for other species trios. These analyses led to the discovery of a list of genes that function in development of neural connectivity or auditory or speech processing, including six genes from the ROBO1 axon guidance pathway. I also came across an unexpected but significant association of SNAAPs with the species trio that contains human and two domesticated species (cow and dog). I suggested two possible types of evolutionary forces in shaping the SNAAP substitutions and tested the idea also in avian species. I also discovered a biased pattern in nucleotide changes for SNAAP substitutions, which was relaxed for the trio of human, cow and dog.

### ***4.2 Materials and methods***

#### **4.2.1 Computation of Shannon's entropy surrounding SNAAPs**

Shannon's entropy is a simple metric for measuring diversity and has been widely used in measuring the conservation of nucleotide sites in genes (Schneider et al. 1986) or amino acid diversity in proteins (Zou and Saven 2000). Here given a multiple alignment of the protein sequences, I calculated the surrounding conservation value (C)

of a SNAAP site  $k$  by  $C = \sum H_j / (2n)$  where  $n$  = number of residues examined on either side;  $j = k-n, k-n+1, \dots, k-1, k+1, \dots, k+n$ , and the Shannon entropy  $H_j = - \sum_i P_{ij} \log P_{ij}$ , where  $P_{ij}$  is the observed frequency of amino acid  $i$  at site  $j$ . In this study, I calculated  $2n= 10$  or  $20$  amino acids surrounding the SNAAP sites.

#### **4.2.2 Gene ontology enrichment analysis**

I performed a systematic gene ontology (GO) analysis on genes with type 1 SNAAPs using a bioinformatics tool, GeneCodis 2.0 (Nogales-Cadenas et al. 2009). The goal is to see if any GO terms are enriched and functionally relevant to vocal learning or other shared traits. The enrichment analysis was normalized to all 3218 genes by a hyper-geometric test at a false discovery rate of 0.05. I used four types of annotations: biological process, molecular function, and cellular component at the lowest GO levels as well as the transcription factor binding motifs.

#### **4.2.3 Enrichment analysis of genes with high dN/dS ratios in primates**

Using the branch-site models in PAML4.2, I inferred the following evolutionary events in primates on genes with type 1 SNAAPs: (1) positive selection (PS) and accelerated evolution (AE) in humans, or (2) slightly relaxed selection (RSC) in non-human primates. AE and PS events were inferred in humans in a similar way as described in Chapter 2, where a tree of four species was used: human, a nonhuman primate (chimpanzee or monkey, depending on whether the SNAAPs were from the Hom-Mac or Hom-Pan test), and two non-primate species without the trait of interest

(e.g. cow and dog for testing the vocal learning trio). To infer RSC events in nonhuman primates, I used a different tree which consists of human, a nonhuman primate and two non-primate species that have the trait of interest. The transcripts with  $\omega$  ratios in the non-human primate that was higher than the rest of the lineages but not 1 were considered RSC events in nonhuman primates. The log likelihood ratio test cutoffs were initially set to be 3.37 and 2.95 for Test I and Test II, and were relaxed to 1.5 and 1.5 in a subsequent test. Lastly, I used the Fisher Exact Test to determine if there is any significant association between genes with such events and genes with type 1 SNAAPs for specific species trios.

#### **4.2.4 Ancestral sequence reconstruction**

Ancestral reconstruction is widely used to infer characteristics or the genetic form associated with the ancestral nodes in a phylogenetic tree. Here I used the maximum likelihood method implemented in the codeml program of the PAML package (Yang 2007) (with default parameters) to conduct a marginal reconstruction of ancestral protein sequences in the common ancestor of placental mammals. Ancestral reconstructions using a parsimonious method were also conducted with the GASP program (Edwards and Shields 2004) and yielded the same results.

#### **4.2.5 Enrichment analysis of amino acid substitutions**

I performed an enrichment analysis of bi-directional amino acid substitutions for type 1 SNAAPs identified from each species trio as follows. For a given species trio, I

counted the frequencies ( $P_F$ ) of each type of bi-directional substitution. For the control, I swapped human and a nonhuman primate to get the corresponding background frequencies ( $P_B$ ). Then for each type of substitution, I calculated the frequency difference  $\Delta P = P_F - P_B$ , and transformed this value into  $\delta = \log(1 + \Delta P) - \log(1 - \Delta P)$ . Since  $\delta$  is in the  $(-\infty, +\infty)$  range, I calculated the average  $\langle \delta \rangle$  and the standard deviation  $\sigma(\delta)$ . A substitution type would be considered significantly enriched if and only if its  $\delta$  is greater than  $\langle \delta \rangle + 2\sigma(\delta)$ .

#### **4.2.6 Computing mutation biases by codon usage and similar amino acid types**

For codon usage bias analyses, I collected the frequencies of each codon type in humans, mice and pigs (Jorgensen et al. 2005) and calculated the joint probability of obtaining a pair of codons that differ in the 1st or 2nd codon positions for encoding different amino acids. I then classified these pairs of codons into the six substitution categories: A/G, A/T, A/C, G/T, G/C, T/C. Likewise, I examined the amino acids with similar properties in the polar group (N Q S T K R H D E), the H-bond forming group (C W N Q S T Y K R H D E) and the active site group (H C S K T N R Q E D A M) (Barnes 2007). Taking the polar group of amino acids for example, I enumerated the pairs of codons that differ in the 1st or 2nd codon positions for encoding different amino acids within this polar group, and counted the frequencies for each category of nucleotide substitution.

### **4.3 Results**

We found that for all species combinations, there were progressively more SNAAPs (also in more genes) from types 1 to 4 (Table 3, Figure 7). The type 1 SNAAPs showed limited choices of changes among  $3.28 \pm 1.25$  alternating amino acids across  $21.5 \pm 5.0$  species on average (see examples in Figure 8), suggesting a strong negative selection over most substitutions at these sites.

I found that very few type 1 SNAAP substitutions of the vocal learning trio (i.e. human + the two vocal learners, abbr. H+VL) were exclusive to the three mammalian vocal learners. Further, no such substitutions are exclusively shared in the six known vocal learners whose sequences are available (human, elephant, microbat, dolphin, zebra finch, parrots). One SNAAP close to this criterion is in the CASP8AP2 gene (Figure 8D). There is a glutamine (Q) at this SNAAP site in all known vocal learning mammals (humans, microbats, elephants and dolphins), an arginine (R) in other placental mammals including the non-human primates and megabats that do not rely on echolocation but vision, and a histidine (K) in marsupials (opossum and wallaby). In birds, there is an arginine (R) in chickens and two vocal learners, zebra finches and parrots.

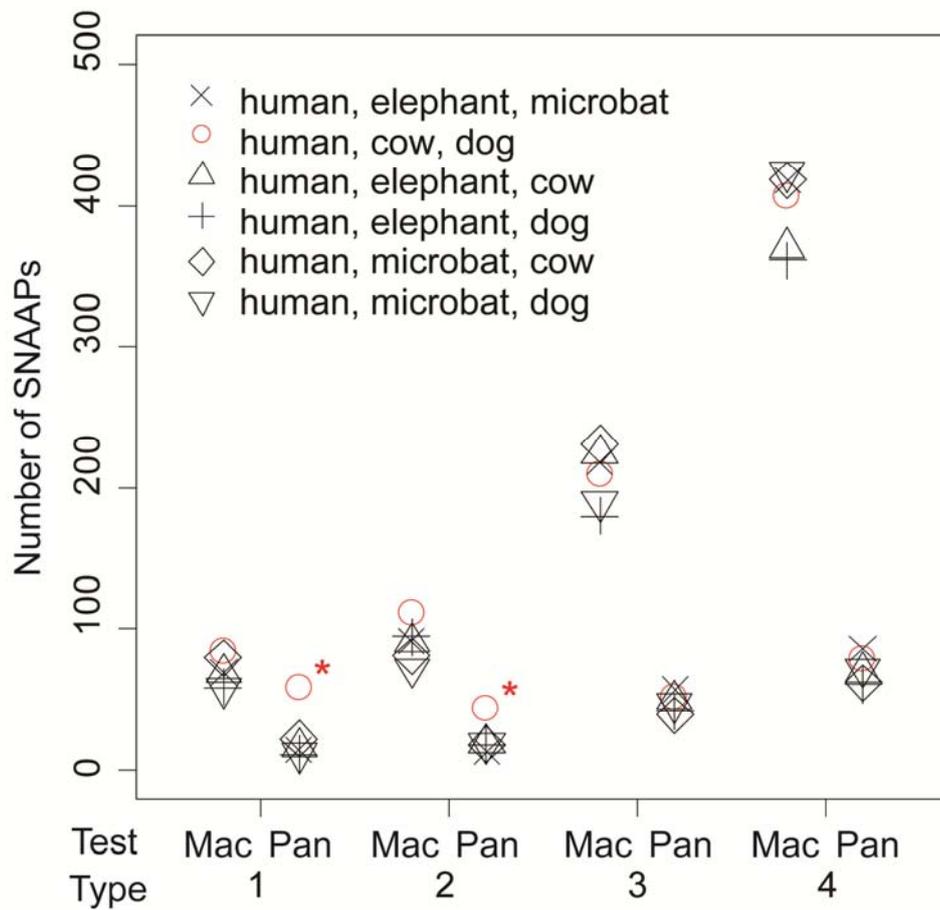
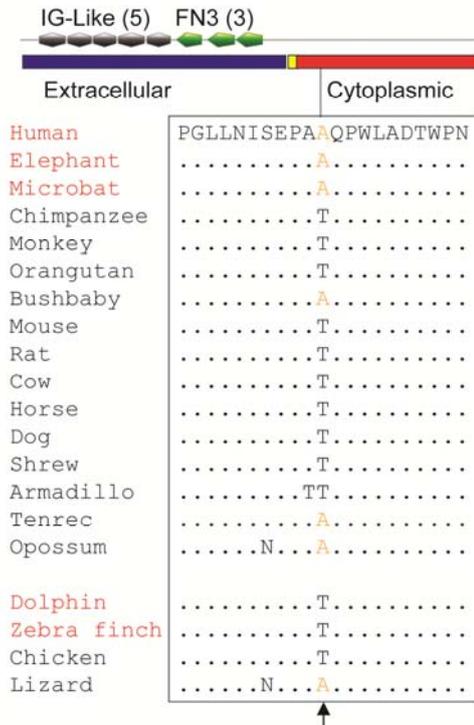
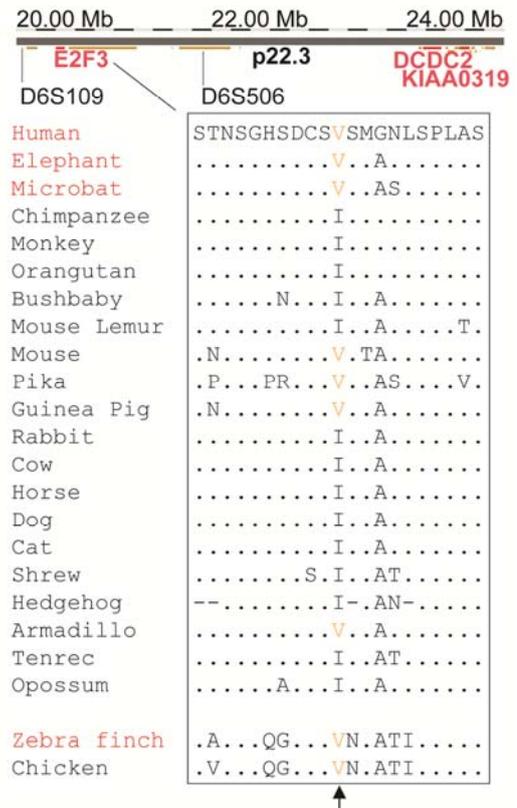


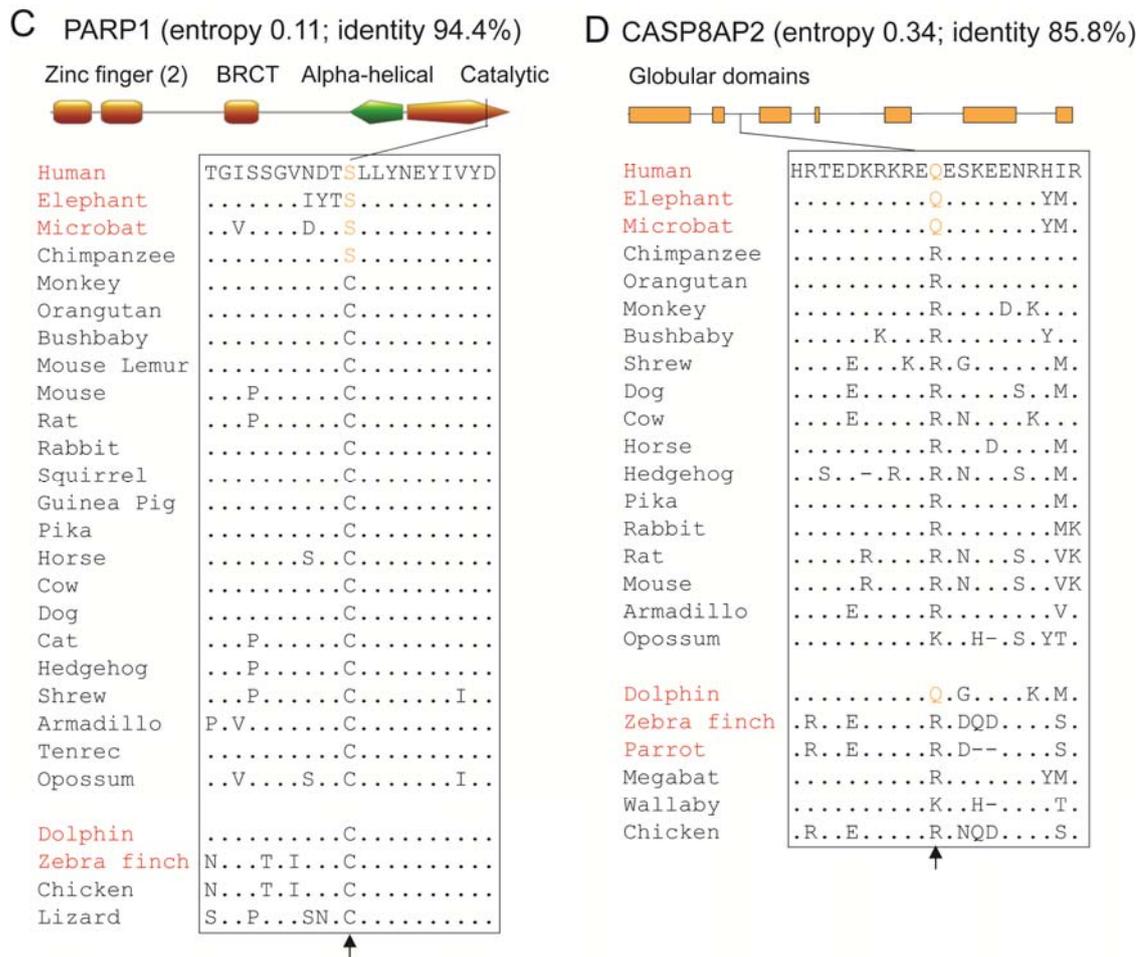
Figure 7: The numbers of type 1-4 SNAAPs identified in each trio of species using either monkeys (Mac; Hom-Mac test) or chimpanzees (Pan; Hom-Pan test) in the pre-selection of species. \* = significantly more SNAAPs of the H+DOM trio than those of other trios (Grubbs outlier test,  $p = 0.0023$  and  $0.0018$ ).

**A** ROBO1 (entropy 0.05; identity 99.4%)



**B** E2F3 (entropy 0.19; identity 93.3%)

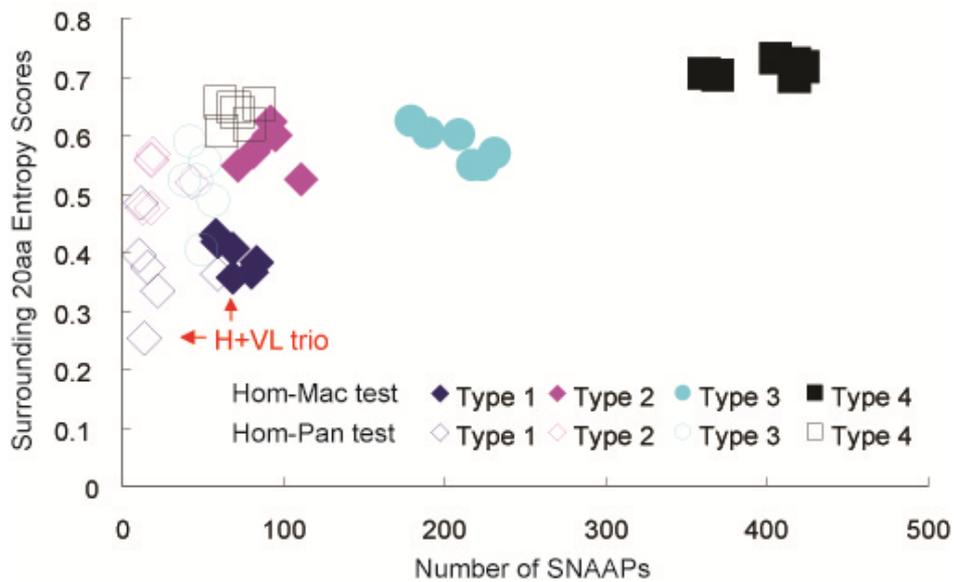




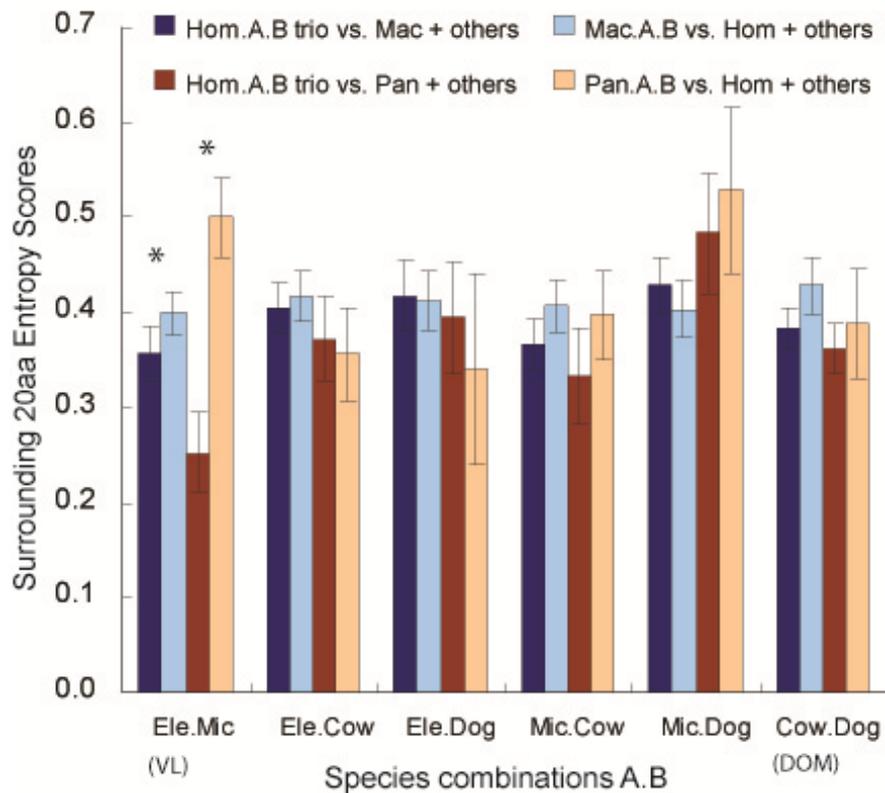
**Figure 8: Alignments for example type 1 SNAAPs and surrounding amino acids identified in the screen with the three vocal learners. (A) ROBO1; (B) E2F3; (C) PARP1; (D) CASP8AP2. Dots in alignments indicate the same whereas letters indicate different amino acid residues relative to human. Listed are the average entropy scores for the 20 amino acids surrounding the SNAAPs and sequence identities for displayed alignments except SNAAP sites. Arrows, the SNAAP site. Residues in yellow, the human substitution. Species in red, known vocal learners. Diagrams of protein structure with predicted domains by PROSITE ([expasy.org/prosite](http://expasy.org/prosite)) are displayed for ROBO1, PARP1 and CASP8AP2. The dyslexia and/or speech sound disorder susceptible regions, DYX2 and D6S109-D6S506 (locations indicated by vertical lines), that contain E2F3 and two known dyslexia susceptibility genes (DCDC2 and KIAA0319, colored in red) are also shown.**

### **4.3.1 H+VL type 1 SNAAPs occur in more conserved regions**

Intuitively, mutations in more conserved regions may have a higher chance to be functionally relevant. I calculated Shannon entropy scores to measure the 20 amino acid sequence conservation surrounding each SNAAP site. Lower surrounding entropy scores indicate more conserved surrounding regions. To put these scores in a perspective, entropy scores from 0.05-0.20 may correspond to 99%-93% sequence identity across species, but it can further measure the amino acid heterogeneity at a site. I found that the H+VL trio type 1 SNAAPs were in the most conserved regions compared to SNAAPs identified from the other five species combinations, especially in Hom-Pan test (Figure 9, arrows). To test whether this relationship was specific to human but not nonhuman primates, for each of the six trios, I swapped humans with a non-human primate (either chimpanzee or monkey) to conduct a similar SNAAP identification analysis (type 1 - 4), and then compared the surrounding entropy scores of the resultant SNAAPs with those for the original trios. I found that the type 1 SNAAPs identified from the H+VL trio were the only case whose surrounding amino acids were significantly more conserved than those identified from their human to non-human primate swap controls in both Hom-Mac and Hom-Pan tests (Figure 10).

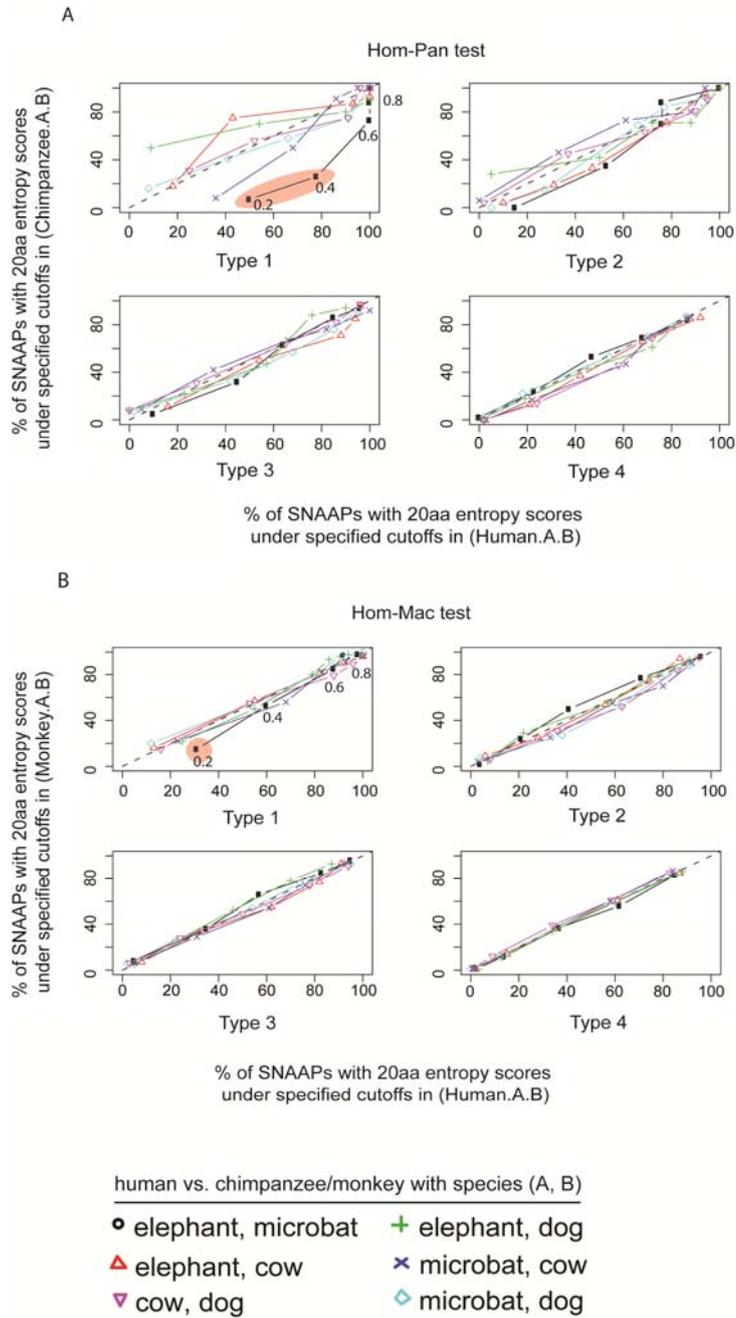


**Figure 9: The mean Shannon entropy scores of the 20 amino acids (aa) surrounding type 1-4 SNAAPs for each species trio (symbols/colors), when using either monkey (Hom-Mac test; closed symbols) or chimpanzee (Hom-Pan test; open symbols) as the NHP control. Arrows, the H+VL trios in the Hom-Pan and Hom-Mac tests.**



**Figure 10: Mean Shannon entropy scores of the 20 amino acids surrounding type 1 SNAAPs in each trio that contains human (Hom.A.B) compared to their swap control trios ((Mac.A.B) or (Pan.A.B)), where A and B = species listed on the x-axis. \* = significantly lower scores for the H+VL trio (Hom.Ele.Mic) ( $p = 0.046$  for Mac.A.B comparison, blue bars;  $p=4.8E-4$  for Pan.A.B comparison, brown bars; Mann-Whitney U test). Error bars: standard error.**

To examine the level of conservation in further detail, I compared the percentages of SNAAPs below different entropy cutoffs (1.0, 0.8, 0.6, 0.4, or 0.2) for different combinations of species before and after the swap. The results are plotted in Figure 11. Values off the diagonal line to the bottom indicated a bias at the corresponding conservation level for a species trio (with human) vs. the control trio (with a non-human primate) after the human-nonhuman primate swap. The further off the diagonal line, the stronger the bias is. I found that the only species trio with consistent bias in both Hom-Mac and Hom-Pan tests is the H+VL trio for type 1 SNAAPs. There was a strong bias of a higher proportion of type 1 H+VL SNAAPs in highly conserved regions (entropy  $<0.2$  and  $<0.4$ ) relative to the SNAAPs of the non-human primate swap controls (Figure 11; shaded values). This bias was stronger in the SNAAP regions identified using chimpanzee as the non-human primate control than when using monkey. In real numbers, for the Hom-Mac test, 22 of the 70 H+VL type 1 SNAAPs had entropy scores lower than 0.2 vs. 12 of 76 for Monkey+VL SNAAPs, while in Hom-Pan test it was 6 of the 14 for H+VL vs. 2 of 26 for Chimpanzee+VL. Very little bias was seen in type 2 SNAAPs, and nothing out of the ordinary was seen for the other types. Taken together, the above findings suggested that a unique and prominent influence on type 1 SNAAP substitutions for vocal learners in more conserved regions.



**Figure 11: Diagonal plots of the percentages of type 1 SNAAPs whose surrounding 20 amino acid entropy scores are lower than specified cutoffs (0.2, 0.4, 0.6, 0.8; indicated in upper left graphs), in both (A) the Hom-Pan test and (B) the Hom-Mac (bottom panel) tests. Shaded regions, much higher proportions of SNAAPs in the (Human.A.B) trio than in their swap control trios, i.e. far off the diagonal.**

### **4.3.2 Excess SNAAPs for human and two domesticated species**

I found there were fewer SNAAPs identified from the Hom-Pan test than those from the Hom-Mac test (Table 3). Most ratios of the numbers of SNAAPs from the Hom-Mac test versus Hom-Pan test (i.e. Mac:Pan ratios) ranged from 2.6 to 7.1 (Figure 12), which are in the range of ratios of divergence time estimates (3.0-8.5) of human and monkey (20-34 MYA) versus human and chimpanzee (4-8 MYA) from their common ancestor (Steiper and Young 2006) . This finding may suggest a continuous accumulation of human SNAAP substitutions from monkey to chimpanzee with time. However, this suggestion has to be taken cautiously, as the range of divergence time estimates is large and thus not yet highly reliable.

Interestingly, a significant outlier from the ratio range of divergence times was found for the type 1 SNAAPs of the trio of species that included humans, cows, and dogs (i.e. Human + two domesticated species, abbr. H+DOM trio; Figure 12). Their Mac:Pan ratio of SNAAP numbers was much lower (1.4 for type 1 and 2.6 for type 2) than those for any other species combination. In addition, there were significantly more type 1 and type 2 SNAAPs in more genes for this trio than for any other trio in both Hom-Mac vs Hom-Pan tests (Figure 7). In contrast to the continuous accumulation of SNAAPs during primate evolution for other species trios or types, this result highly suggested a recent accumulation of these SNAAP substitutions in humans, but not in either monkey or chimpanzee.

To test this hypothesis, I replaced human with either chimpanzee or monkey, to form a control trio (i.e. chimpanzee/monkey.elephant.microbat) to identify their SNAAPs and calculated the ratio of the numbers of SNAAPs before and after such human-nonhuman primate swap (i.e. Hom:NHP swap ratios). The null hypothesis is that this human-nonhuman primate swap should not result in a significant change in the number of SNAAPs (i.e. swap ratio  $\sim 1$ ), unless there is a higher representation of such changes in humans but not nonhuman primates shared with cows and dogs.

I found that the Hom:NHP swap ratios for types 1-2 SNAAPs of the H+DOM trio were all greater than 1 (range 1.2-3.7), and the highest among all trios for type 1 SNAAPs in both Hom-Mac and Hom-Pan tests (Figure 13). Further, the swap ratios of the H+DOM trio increased significantly more than any other trio (three times for type 1) from the Hom-Mac test to the Hom-Pan test. In contrast, the swap ratios for type 4 SNAAPs were nearly all  $\sim 1$ , suggesting that they are more influenced by a continuous substitution process. These findings provided further support for an excess of substitutions in humans but not the two nonhuman primates shared with the two domesticated species.

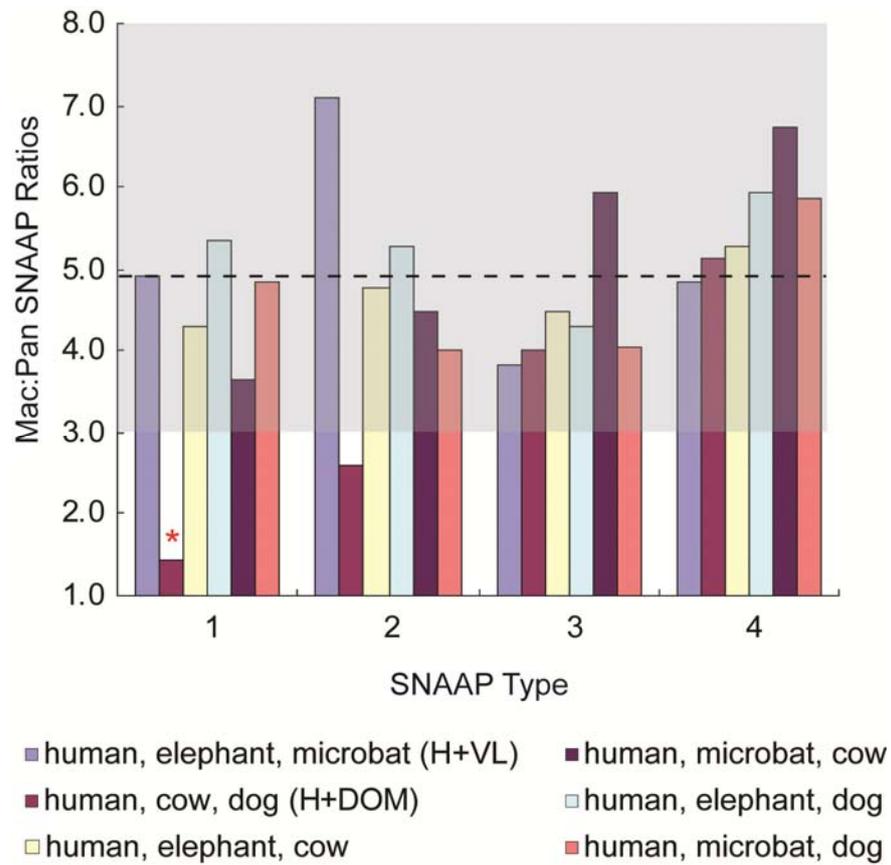
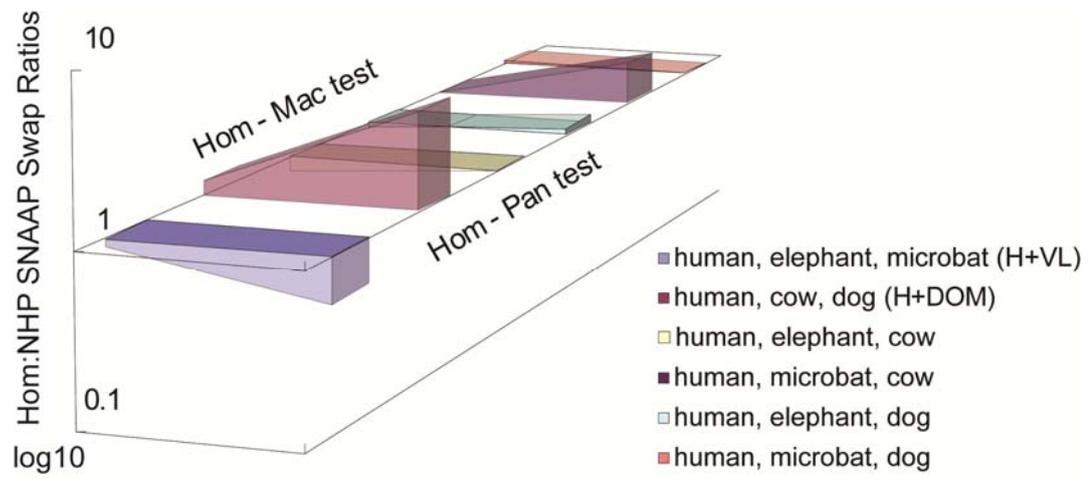


Figure 12: The Mac:Pan ratios of the number of type 1-4 SNAAPs for each species trio. \* = significantly lower Mac:Pan ratio (Grubbs outlier test,  $p = 0.038$ ). Dashed line, average Mac:Pan ratios across all SNAAP types, without outlier. Shaded area, the ratio range (3.0-8.5) of divergence time estimates of human-monkey versus human-chimpanzee from their common ancestors from multiple studies (Steiper and Young 2006). Such a large range reflects uncertainties across studies.



**Figure 13: The Hom:NHP swap ratios of the number of type-1 SNAAPs for the human-monkey and human-chimpanzee test.**

### **4.3.3 Many type 1 SNAAP genes are functionally relevant to the trait of interest**

Humans, elephants and microbats have large trait differences, including body size, metabolism, and mode of locomotion. Therefore, with the one rare convergent trait they do share, vocal learning, one might expect to find at least a subset of genes that contain such H+VL type 1 SNAAPs to be functionally relevant to this trait. Moreover, since human is the only known vocal learner in primates, the Hom-Mac and Hom-Pan tests are expected to overlap substantially in their identified SNAAPs. Consistent with this idea, I identified 70 H+VL type 1 SNAAPs in 68 genes from Hom-Mac test and 14 H+VL type 1 SNAAPs in 14 genes from Hom-Pan test (Table 7), and unlike the AE and PS analyses in Chapter 2, the two lists of SNAAPs overlapped significantly (Fisher exact test,  $P = 8E-13$ ) by 9 genes, resulting in 73 unique genes in total.

I found that at least 4 (44%) of the 9 overlapped genes have specific neural functions with close functional ties to speech or language abilities. These include: (1) *ROBO1*, which encodes an axon guidance receptor critical for forming longitudinal axonal projections (Devine and Key 2008) and when mutated leads to susceptibility for reading dyslexia (RD) and speech sound disorder (SSD) (Hannula-Jouppi et al. 2005); (2) *E2F3*, which lies in a small RD susceptible region spanning 7 protein-coding genes and adjacent to the *DYX2* locus for both RD and SSD (Turic et al. 2003) and is part of the pRB/E2F pathway for maintaining the inner ear hair cells; (3) *CASP8AP2*, which is necessary for activating Casp8 and in turn Casp3 that when knocked out causes neural

degeneration of inner ear hair cells and thus hearing loss (Morishita et al. 2001); and (4) The Usher syndrome 1C binding protein 1 (USHBP1), which binds to harmonin encoded by Usher syndrome 1C gene in inner ear hair cells, and when harmonin is mutated it causes Usher's syndrome, a progressive hearing (and vision) loss (Ishikawa et al. 2001).

Of the H+VL type 1 SNAAPs found only in Hom-Mac test (i.e. chimpanzees have the same SNAAP substitutions as vocal learners), I also found 9 in genes related to speech/language abnormalities or ROBO1 mediated axon guidance. These include: FRAS1 (Mcgregor et al. 2003) in which mutations cause abnormalities in the larynx (Fraser syndrome); GDAP1 (Sevilla et al. 2003) in which mutations cause neuropathy with severe vocal cord paresis and diaphragmatic dysfunction; PARP1 (Tapia-Paez et al. 2008) that regulates the dyslexia susceptible gene DYX1C1 for neuronal migration in the developing cortex; HAL (Hyanek and Raisova 1985), in which mutations can cause severe histidinaemia, a rare metabolic disorder with symptoms of delayed speech and language development; and most interestingly, five genes in the ROBO1 axon guidance pathway – PTPRB (Sun et al. 2000), PTPNA (Xie et al. 2005), CKAP5, PCM1, and CEP192. The last three genes are part of the centrosome assembly that interacts with SLIT1, the ligand of ROBO1, necessary for proper neuron migration and neural process extension (Higginbotham et al. 2006).

I also searched for possible evidence of genes with H+DOM type 1 SNAAPs for domestication. I found that one SNAAP from the Hom-Mac test is in an ATP-binding

cassette gene, ABCA12, which is known to be influenced by domestication in cows (Charlier et al. 2008). In both cows and humans, individuals with homozygous mutations suffer from similar skin disorders (Charlier et al. 2008). Two other genes from the same gene family, ABCA3 and ABCA4, also contain type 1 H+DOM SNAAPs.

These results suggest that my SNAAP approach may have identified a subset of genes relevant to the convergent traits of interest. They further suggest that relative to the three vocal learners, both chimpanzee and monkey appear to lack convergent SNAAPs in genes that function in neural connectivity, vocal production and audition, and monkey further lacks more of them than chimpanzees. The findings, however, do not indicate whether or not the individual SNAAP site is responsible for the functional differences between species.

**Table 7: Identified genes with H+VL type 1 SNAAPs in Hom-Mac and Hom-Pan tests.**

Hom-Mac Test				Hom-Pan Test	
Ensembl ID	Gene name	Ensembl ID	Gene name	Ensembl ID	Gene name
157426	AASDH	91128	LAMB4	152611	CAPSL
159461	AMFR	115850	LCT	118412	CASP8AP2
53900	ANAPC4	140400	MAN2C1	112242	E2F3
103569	AQP9	140398	NEIL1	110756	HPS5
167080	B4GALNT2	67141	NEO1	114648	KLHL18
152785	BMP3	95319	NUP188	155087	ODFP1
127720	C12orf26	162600	OMA1	54690	PLEKHH1
182795	C1orf116	145623	OSMR	118898	PPL
162598	C1orf87	143799	PARP1	101040	PRKCBP1
118412	CASP8AP2	78674	PCM1	108344	PSMD3
101639	CEP192	107815	PEO1	169855	ROBO1
175216	CKAP5	204138	PHACTR4	174175	SELP
187955	COL14A1	174238	PITPNA	136731	UGCGL1
174080	CTSF	101040	PRKCBP1	130307	USHBP1
114395	CYB561D2	7062	PROM1		
167969	DCI	108344	PSMD3		
104808	DHDH	127329	PTPRB		
112242	E2F3	172053	QARS		
203965	EFCAB7	108961	RANGRF		
90776	EFNB1	23287	RB1CC1		
121053	EPX	165476	REEP3		
180210	F2	140519	RHCG		
187790	FANCM	159753	RLTPR		
64763	FAR2	169855	ROBO1		
138759	FRAS1	74660	SCARF1		
104381	GDAP1	103168	TAF1C		
143167	GPA33	184786	TCTE3		
84110	HAL	169903	TM4SF4		
116882	HAO2	153214	TMEM87B		
171495	HEATR7B2	166479	TXNDC10		
110756	HPS5	136731	UGCGL1		
168418	KCNG4	142207	URB1		
114648	KLHL18	163625	WDFY3		
158552	ZFAND2B	130307	USHBP1		

#### **4.3.4 Enriched GO terms for genes with H+VL but not H+DOM type 1 SNAAPs**

The above analyses revealed two species trios with type 1 SNAAPs that stood out as outliers from the rest species trios: human and vocal learning species (H+VL trio) for amino acid changes in highly conserved protein coding regions; and unexpectedly humans and domesticated animals (H+DOM trio) with non-randomly associated changes in many more genes. Despite with quite a few genes of known functional relevance to the corresponding traits, the significance of these discoveries still needs systematical evaluation to confirm a non-random functional association with convergent traits. Ideally, this can be assessed by testing if the list of vocal learning or domestication related genes are over-represented. This is not feasible given the current limited knowledge in these aspects. Therefore, to address this question, I turned to the Gene Ontology (GO) analysis on these type 1 SNAAP genes.

I found 7 genes in 3 GO enrichment groups from Hom-Mac test and 3 genes in 1 GO enrichment group from Hom-Pan test, which overlapped by 2 genes (E2F3 and CASP8AP2; Table 8). Remarkably, these 8 genes included the aforementioned ROBO1, E2F3 and CASP8AP2 that were identified to have SNAAPs in both Hom-Mac and Hom-Pan tests. Besides ROBO1, two more axon guidance related genes (NEO1 and EFNB1) were also on the list. NEO1 encodes a membrane receptor for netrins to mediate axon guidance, whereas EFNB1 is a cell adhesion ligand protein expressed on post-synaptic dendritic spine nerve fibers to provide guidance signals, cell adhesion, and synaptic

plasticity (Bianchi and Gray 2002; Migani et al. 2009). Another gene on the enrichment list is RB1CC1, which together with E2F3, is part of the pRB pathway that is critical for maintaining inner ear auditory hair cells (Mantela et al. 2005).

Taken together, the enrichment of genes that function in axon guidance (ROBO1, EFNB1 and NEO1) is consistent with the known neural connectivity differences between vocal learners and non-learners (Figure 2). In contrast, the three genes necessary for maintaining inner ear hair cells (RB1CC1, E2F3, and CASP8AP2) may imply another critical difference.

To test how likely the above GO results are due to randomness, I conducted the same analysis on 50 random sets of genes chosen from the 3218 genes. Each of these random sets contained the same number of genes as those with H+VL type 1 SNAAPs. In the Hom-Mac test, I found the random gene sets had an average of 1.37 enriched groups. 45 of the total 50 random sets (i.e. 90%) had fewer than 3 enriched groups as in the case for genes with H+VL type 1 SNAAPs (32 random sets with 0 enrichment group; 11 with 1 group; 3 with 2 groups). In the Hom-Pan test, the random gene sets had an average of 0.41 enriched groups. 40 of the total 50 random sets (i.e. 80%) had fewer than 3 such groups as in the case for genes with H+VL type 1 SNAAPs. These results suggested the observed GO enrichment results for H+VL type 1 SNAAP genes may occur only in 10-20% of random cases (p-value: 0.1-0.2).

Further, the GO terms “integral to plasma” (GO:0005887) and “cell adhesion” (GO:0007155) were found to appear in genes with H+VL SNAAPs but not in any of the random gene sets. Particularly, cell adhesion molecules are considered necessary for developing neural connectivity and synapse formation (Dalva et al. 2007).

I conducted the same analysis for genes bearing type 1 SNAAPs for H+DOM trio and the other control species trios. Two of the four control species trios (the trio of human, elephant and cow; the trio of human, microbat and cow) also have significant GO enrichment, but only in Hom-Mac test. Surprisingly, I did not observe any enriched groups in H+DOM trio in either test. The full results were listed in Table 8.

**Table 8: Groups of genes with type 1 SNAAPs for each trio of species that are found with significantly enriched gene ontology terms.**

Trio	Gene Ontology	Genes
Hom-Mac Test		
Hom.Ele.Mic	protein binding nucleus cell cycle	CASP8AP2, E2F3, RB1CC1
	protein binding integral to plasma membrane cell adhesion	NEO1, EFNB1, ROBO1
	integral to plasma membrane	NEO1, EFNB1, AQP9
	Hom.Cow.Dog	-
Hom.Ele.Cow	receptor activity plasma membrane protein binding	IL7R, PDGFRB, SLC1A5
	Hom.Ele.Dog	-
Hom.Mic.Cow	Nucleus	CHD6, HIVEP3, NEDD9, SERTAD4
Hom.Mic.Dog	-	-
Hom-Pan Test		
Hom.Ele.Mic	protein binding	CASP8AP2, E2F3, UGCGL1
Hom.Cow.Dog	-	-
Hom.Ele.Cow	-	-
Hom.Ele.Dog	-	-
Hom.Mic.Cow	-	-
Hom.Mic.Dog	-	-

#### **4.3.5 Enriched AE/RSC in genes with H+VL but not H+DOM type 1 SNAAPs**

Although the dN/dS analysis in Chapter 2 did not reveal significantly shared AE or PS genes across the three vocal learners, I wondered if it can help understand the evolution of the type 1 SNAAPs in the primates only. I considered two possible selective forces that could have shaped these SNAAPs: (1) positive selection (PS) or accelerated evolution (AE) in human and (2) relaxed selective constraints (RSC) in non-human primates. These alternatives correspond to the two long-standing hypotheses on the origin of vocal learning: independent gains of vocal learning and the less commonly believed many losses of vocal learning from a common ancestor that had the trait (Jarvis 2004). Here I examined if genes with these SNAAPs were enriched with PS events (gains) in humans or RSC events (losses) in non-human primates using the improved branch-site model as described in the Methods. I did not check for PS or RSC events in the other two vocal learners (elephant and microbat) or non-learners (cow and dog), since there are no high coverage sequenced genomes of their close relatives as for humans and NHPs. In this case, it is hard to tell if the resulted PS/AE/RSC genes are lineage specific to the vocal learner or non-learners.

Among genes with the H+VL type 1 SNAAPs, none had PS in human but a few had AE in human and RSC in chimpanzee when using chimpanzees as the nonhuman primate (Table 9). Genes with all three types of events (human PS/AE, monkey RSC) were found when using monkey as the nonhuman primate (Table 9). Moreover, such AE

and RSC events were significantly enriched when using chimpanzee but not monkey as the nonhuman primate with the standard stringent cutoffs (Table 9). When relaxing cutoffs to 1.5 to increase the sensitivity, I found significant enrichment of AE and RSC events using either chimpanzee or monkey as the non-human primate, but still no enriched PS events (Table 9). These results indicate that certain AE events in humans and RSC events in nonhuman primates are associated in genes with the H+VL type 1 SNAAPs, but not PS events in humans. However, only several aforementioned genes were under such events: USHBP1 and RB1CC1 (AE), ROBO1 (RSC in chimpanzee, but AE in human relative to monkey), EFNB1 (RSC) and UGCGL1 (RSC).

I also did the same analysis for genes with H+DOM type 1 SNAAPs but did not find any significant association with PS or AE events in humans or RSC events in NHPs using either stringent or relaxed cutoffs.

**Table 9: Genes under PS/AE/RSC with H+VL type 1 SNAAPs. NHP, the non-human primate used in inferring PS/AE/RSC events. Outgroups, the outgroup species used in inferring PS/AE/RSC events. Events, the PS/AE/RSC in specified lineages. # Events, the number of PS/AE/RSC events inferred in specified lineages. # Overlaps, the number of genes with both PS/AE/RSC events and type-1 H+VL SNAAPs. P-values, assessment of the significance of overlap using Fisher Exact Tests. P-value in the parenthesis, assessment after removing the PLEKHH1 whose SNAAP is a human SNP variant.**

NHP	Chimpanzee			Monkey		
Outgroups	Cow, Dog		Ele, Mic	Cow, Dog		Ele, Mic
Events	Human PS	Human AE	NHP RSC	Human PS	Human AE	NHP RSC
using stringent cutoffs: 3.37 (Test I), 2.95 (Test II)						
# Events	23	75	63	35	109	32
# Overlaps	0	3	3	1	3	2
Overlapping Genes		PLEKHH1, USHBP1, PSMD3	ROBO1, PPL, UGCGL1	PHACTR4	PHACTR4, CKA5, HAL	EFNB1, DH
p-values	1	0.00391	0.00237	0.54146	0.72773	0.15446
Using relaxed cutoffs: 1.5 (Test I), 1.5 (Test II)						
# Events	35	352	263	85	366	167
# Overlaps	0	7	5	3	17	11
p-values	1	0.00035 (0.0016)	0.0042 (0.018)	0.42853	0.00199	0.00086
Overlapping genes		CAPSL, E2F3, KLHL18, PLEKHH1, PSMD3, USHBP1, ZMYND8	PLEKHH1, PPL, PSMD3, ROBO1, UGCGL1	DHDH, PHACTR4, RHCG	C12orf26, C1orf116, CKAP5, CYB561D2, DHDH, FANCM, HAL, MAN2C1, NUP188, PCM1, PHACTR4, PSMD3, QARS, RB1CC1, REEP3, RHCG, ROBO1	C1orf87, CEP192, CYB561D2, DHDH, E2F3, EFNB1, FRAS1, PHACTR4, QARS, TCTE3, UGCGL1

#### 4.3.6 Gains and Losses in shaping H+VL type 1 SNAAPs in mammals

One caveat of above PS/AE/RSC analyses is that it searches for non-neutral selection over the entire protein sequence, which does not necessarily act upon the specific SNAAP sites. In addition, I found that certain genes (ROBO1, PSMD3, PLEKHH1, CYB561D2, DHDH, PHACTR4 and QARS) showed both AE in humans and RSC in nonhuman primates, making it less obvious as to which events directly shaped the SNAAPs in these genes. To investigate individual SNAAP sites, I reconstructed the ancestral amino acid types in the most recent common ancestor of the three vocal learners (see Methods) and asked if the vocal learners evolved a novel amino acid type (Gain category) or preserved the ancestral amino acid type (Loss category) (Table 10). To avoid potential biases by small sample sizes, 7 SNAAPs sequenced in <12 mammals were excluded.

Remarkably, I found that the SNAAPs in almost all genes related with the ROBO1 axon guidance pathway (ROBO1, NEO1, PITPNA, CKAP5, PCM1, and CEP192) were in the Loss category. The only exception was PTPRB, which interacts with the axon guidance receptor EFNB1 (Holland et al. 1998), and both of them have SNAAPs in the Gain category. SNAAPs in genes related with hearing (E2F3, CASP8AP2, RB1CC1) or when mutated lead to susceptibility to speech-language disorders (FRAS1 and HAL) were also in the Gain category (Table 10). One gene, WDFY3, had SNAAPs in both the Gain and Loss categories, is critical for motor neuron survival in flies (Lim and Kraut

2009) and has a 4-fold higher expression in cerebral cortex in humans relative to chimpanzees (Kehrer-Sawatzki et al. 2005). The above findings indicated that SNAAPs in different but functionally related genes, particularly those of the ROBO1 axon guidance pathway, may have been shaped through the same evolutionary trajectories.

One Gain SNAAP in the PARP1 gene resides in the catalytic domain responsible for Poly (ADP-ribose) polymerization. Interestingly, the codons at this SNAAP site in vocal learners are of three distinctive forms: TCT (Hom), AGT (Ele) and TCC (Mic), all of which encode a serine, as opposed to a cysteine in most mammalian species (Figure 8C). Moreover, this substitution occurs in a highly conserved region where the surrounding 20 amino acids are of ~95% identity in all mammalian species. These findings demonstrate a convergent, and possibly positive, selection for a rare amino acid type (serine) in these vocal learners.

**Table 10: Ancestral reconstruction results of H+VL type 1 SNAAPs. Position, the SNAAP site at the corresponding translated protein sequences. Ancestor, the amino acid type in the common ancestor of placental mammals. VL, the amino acid type in the three vocal learners. # total species, the number of species used in ancestral reconstruction. Category, the "Gain" or "Loss" category type.**

Hom.Ele.Mic, i.e. H+VL	Ensembl		Position	Ancestor	VL	# total species	Category
	gene id	protein id					
Hom-Mac test							
WDFY3	163625	295888	382	I	V	18	Gain
PARP1	143799	355759	999	C	S	23	Gain
KLHL18	114648	232766	568	R	C	22	Gain
CTSF	174080	310832	466	I	V	14	Gain
HAL	84110	261208	135	H	R	20	Gain
PSMD3	108344	264639	173	M	V	13	Gain
RANGRF	108961	226105	58	Q	R	18	Gain
HEATR7B2	171495	296803	1084	V	I	21	Gain
RANGRF	108961	313307	58	Q	R	18	Gain
LAMB4	91128	205386	958	Q	H	15	Gain
E2F3	112242	262904	433	I	V	21	Gain
C1orf116	182795	352447	391	R	K	21	Gain
DHDH	104808	221403	182	L	I	21	Gain
EFNB1	90776	204961	172	V	I	19	Gain
COL14A1	187955	247781	1225	A	T	17	Gain
URB1	142207	270201	525	I	V	19	Gain
NEIL1	140398	347170	203	R	H	20	Gain
QARS	172053	307567	108	Q	R	17	Gain
FRAS1	138759	264895	1507	S	K	19	Gain
C10orf2	107815	309595	50	E	D	21	Gain
CASP8AP2	118412	237177	508	K	Q	19	Gain
EPX	121053	225371	170	S	D	17	Gain
CA14	118298	358107	344	R	Q	18	Gain
RB1CC1	23287	25008	1044	V	I	19	Gain
ZMYND8	101040	262975	652	R	Q	18	Gain
PTPRB	127329	261266	905	Q	K	19	Gain
HPS5	110756	265967	746	Q	R	20	Gain
REEP3	165476	298249	187	G	S	19	Gain
ANAPC4	53900	318775	775	T	S	21	Gain
TM4SF4	169903	305852	155	H	N	19	Gain
PHACTR4	204138	362941	525	S	P	14	Gain

Table 10 (continued)

AASDH	157426	205214	572	L	P	19	Gain
GPA33	143167	356842	183	D	N	17	Gain
TMEM87B	153214	283206	171	K	M	17	Gain
F2	180210	308541	226	R	Q	17	Gain
OSMR	145623	274276	771	I	V	19	Gain
C1orf87	162598	360244	61	L	M	17	Gain
ROBO1	169855	305626	991	A	A	16	Loss
PCM1	78674	315008	1158	S	S	20	Loss
WDFY3	163625	295888	1585	S	S	18	Loss
NUP188	95319	349125	684	I	I	19	Loss
NEO1	67141	261908	526	H	H	19	Loss
KCNG4	168418	312129	281	I	I	19	Loss
ZFAND2B	158552	289528	251	R	R	19	Loss
GDAP1	104381	220822	113	E	E	17	Loss
TMX3	166479	299608	162	V	V	16	Loss
OMA1	162600	360270	254	I	I	19	Loss
RLTPR	159753	334958	814	R	R	17	Loss
PITPNA	174238	316809	152	V	V	18	Loss
UGCGL1	136731	259253	513	M	M	15	Loss
EFCAB7	203965	360129	677	I	I	16	Loss
FAR2	64763	182377	275	V	V	18	Loss
AQP9	103569	219919	68	C	C	18	Loss
CKAP5	175216	310227	566	A	A	20	Loss
SCARF1	74660	263071	162	H	H	15	Loss
MAN2C1	140400	267978	327	Y	Y	18	Loss
AMFR	159461	290649	536	S	S	16	Loss
LCT	115850	264162	788	N	N	19	Loss
B4GALNT2	167080	300404	124	R	R	15	Loss
RHCG	140519	268122	46	E	E	16	Loss
FANCM	187790	267430	1422	K	K	20	Loss
PROM1	7062	265014	111	V	V	13	Loss
BMP3	152785	282701	94	R	R	18	Loss
CEP192	101639	317156	312	I	I	14	Loss
C12orf26	127720	248306	305	N	N	19	Loss
Hom-Pan test							
ROBO1	169855	305626	991	A	A	16	Loss
CAPSL	152611	282501	199	I	I	20	Loss
UGCGL1	136731	259253	513	M	M	15	Loss
KLHL18	114648	232766	568	R	C	22	Gain

Table 10 (continued)

PSMD3	108344	264639	173	M	V	13	Gain
PLEKHH1	54690	330278	964	R	H	19	Gain
E2F3	112242	262904	433	I	V	21	Gain
CASP8AP2	118412	237177	508	K	Q	19	Gain
ZMYND8	101040	262975	652	R	Q	18	Gain
HPS5	110756	265967	746	Q	R	20	Gain
SELP	174175	356764	541	H	R	15	Gain

#### **4.3.7 Biased transitional nucleotide changes shape type 1 SNAAPs**

I noted from visual qualitative scanning of the SNAAP substitutions, there appeared to be specific amino acid substitutions that were more common than others. Using the quantitative enrichment analysis as described in the Methods, I found the H+VL type 1 SNAAPs from both the Hom-Mac and Hom-Pan tests were significantly enriched with Q/R (bidirectional) substitutions (Table 11). Most of the remaining trios exhibited enriched amino acid substitutions only in the Hom-Mac test. An exception was the H+DOM trio, which showed no enriched substitutions in either test (Table 11). These results are consistent with the GO enrichment analysis.

I wondered if the underlying nucleotide changes for these SNAAPs were also enriched. For all species combinations, the vast majority of SNAAP substitutions (87% in the Hom-Mac test and 93% in the Hom-Pan test) were attributed to a single nucleotide change rather than 2 or 3 nucleotide changes. And these single nucleotide changes occur at either 1<sup>st</sup> or 2<sup>nd</sup> codon position, which changed the identity of the amino acid type. Moreover, I found transitional (Ts) nucleotide changes (A/G or C/T) were 10-20 times more frequent than transversional (Tv) changes (A/T, A/C, C/G or G/T), i.e. Tv:Ts ratios of 0.05-0.1 (Figure 14A-B).

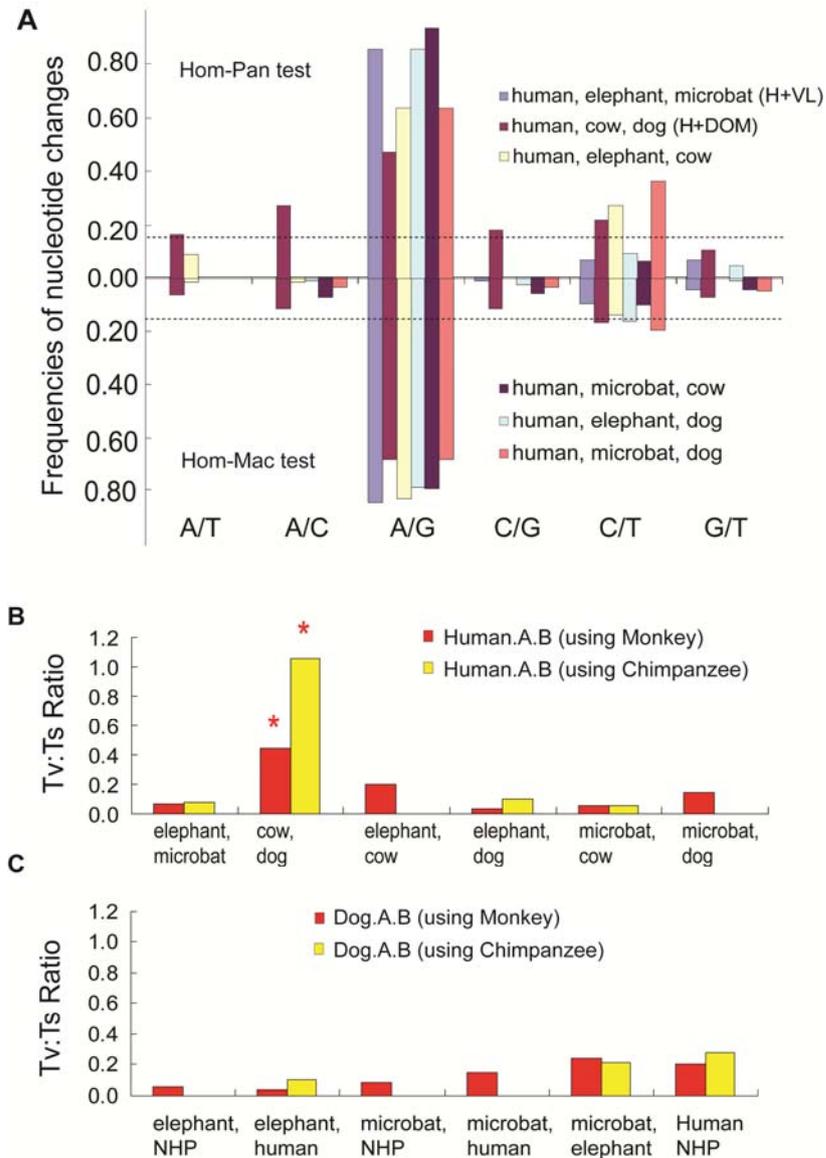
Although transitions are known to be more common than transversions, they have not been found to occur at this high rate. These Tv:Ts ratios were 20-40 times smaller than the random expectation (~2) for non-synonymous changes resulting from

1st or 2nd codon position changes and 5-10 times smaller than the average ratios (~0.5) identified for mammalian nuclear DNA (Belle et al. 2005). Moreover, I found A/G transitions were ~3 times more frequent than C/T transitions (Figure 14A), which is in sharp contrast, for example, with the mitochondria cytochrome b gene in mammals (commonly considered free of selection pressure) where A/G transitions are only ~0.67 times as frequent as C/T transitions (Belle et al. 2005). Lastly, the Tv:Ts ratio distributions (Figure 14A) were also very different from those that result from codon usage bias in mammals (Figure 15A) or preferred substitution biases for amino acids with similar properties (Figure 15B).

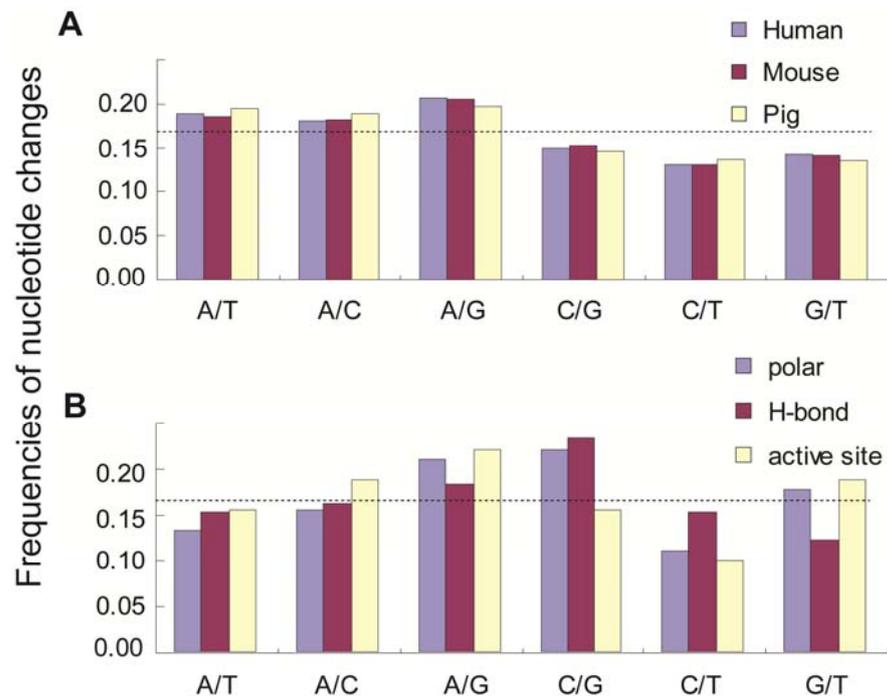
A relaxation to this biased transition rule was seen for type 1 SNAAPs of the H+DOM trio (Figure 14A-B). These SNAAPs had significantly higher Tv:Ts ratios of 0.44 (~1:1.3) and 1.05 (~1:1) in Hom-Mac and Hom-Pan tests, respectively, than for all other species combinations (Figure 14B). To try to falsify the relative uniqueness of this result for the H+DOM trio, I performed a control analysis that identified SNAAPs in species trios where the two domesticated species (dog and cow) were never grouped together. The Tv:Ts ratios of all such trios were low and the highly biased A/G transitions were seen again, with no outliers found (Figure 14C). This result supports my conclusion that the biased transitional changes are significantly reduced specifically for the H+DOM trio.

**Table 11: Significantly enriched amino acid substitutions in each species trio.**

Trios	Enriched amino acid substitutions	
	Hom-Mac test	Hom-Pan test
Hom.Ele.Mic	I/V, R/H, Q/R	Q/R
Hom.Cow.Dog	-	-
Hom.Ele.Cow	I/V, R/H	-
Hom.Ele.Dog	T/A	-
Hom.Mic.Cow	Q/R	-
Hom.Mic.Dog	S/N	-
Hom.Ele.Mic		-



**Figure 14: Analysis of nucleotide substitutions underlying type 1 SNAAPs. (A)** Frequencies of all six single nucleotide changes (bidirectional) for type 1 SNAAPs from each species trio. Human-chimpanzee (Hom-Pan) test is above the x-axis and human-monkey (Hom-Mac) is below. **(B)** Transversion:Transition (Tv:Ts) ratios of nucleotide substitutions for SNAAPs in each species trio that contains humans and two non-primates. \* Grubbs outlier test,  $p = 0.029$  (Hom-Mac test) and  $1.1E-4$  (Hom-Pan test). **(C)** The Tv:Ts ratios in the control analyses that put dog and two other species on one side, and the rest and cow on the other side.



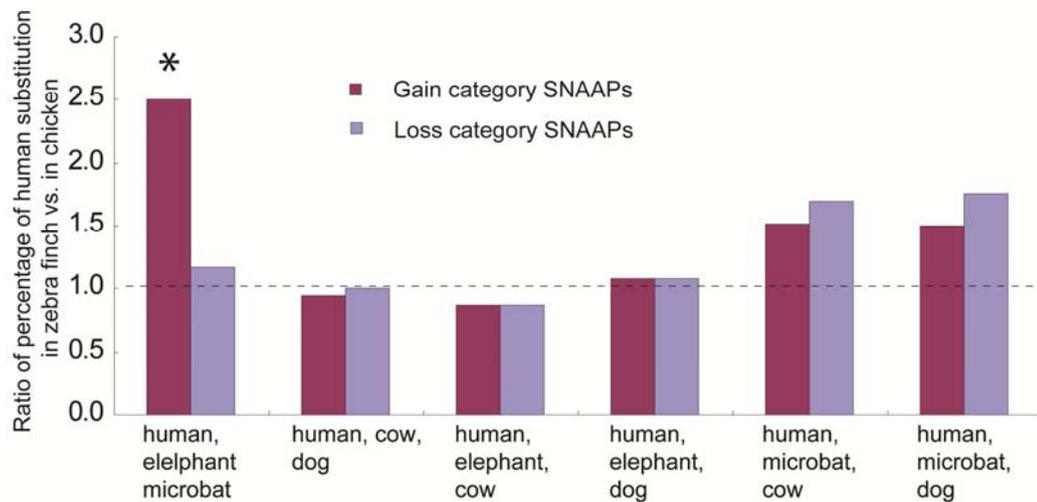
**Figure 15: Frequencies of all six single nucleotide changes for non-synonymous mutations in the 1st or 2nd codon positions influenced by (A) codon usage bias in humans, mice and pigs or (B) the putative preference of similar amino acid properties: polar, H-bond forming or active site involvement.**

#### 4.3.8 SNAAP substitutions present in other vocal learners

With such a high prevalence of convergent SNAAPs in trios of species that share a trait, we expected to find some of them in other species with that trait that were not included in our computational screening due to the stringent selection criterion of genome coverage. This would be even more remarkable if they were present in another vertebrate order. To test this hypothesis, I analyzed the recently available sequences (Ensembl v57) of another mammalian vocal learner, the dolphin, one avian vocal learner, the zebra finch (a songbird) (Warren et al. 2010), and one avian vocal non-learner, the chicken (Hillier et al. 2004). I would expect to see an over-representation of human SNAAP substitutions in dolphin and zebra finch for Gain category than for the Loss category SNAAPs, but not in chicken.

Significant evidence met my expectation with SNAAPs from the Hom-Mac test. In dolphin, I found a higher proportion of type 1 H+VL SNAAPs (20%, 7 of 35) in the Gain category shared with humans, which was twice that of the Loss category (11%, 3 of 28). In zebra finch, I also found higher proportions (33%, 8 of 24) of Gain category SNAAP substitutions shared with humans relative to chicken (13%, 4 of 30), but comparable proportions of the Loss category SNAAPs in zebra finch (44%, 11 of 25) and chicken (38%, 9 of 24). Further, this zebra finch (33%) versus chicken (13%) ratio for H+VL type 1 SNAAPs in the Gain category was also over 1.5 times higher than that for any of the five species trio combinations, and an outlier (Figure 16).

Interestingly, I did not find higher representations of H+DOM SNAAPs in chicken relative to zebra finch (Figure 16), despite long-term domestication of chickens. Yet it should be noted that the reference chicken genome is from the red-jungle fowl that is considered the wild type of the species (Hillier et al. 2004).



**Figure 16: Ratios of the percentage of the Gain or Loss category type 1 SNAAPs with the same human substitutions in zebra finch versus in chicken. A ratio ~1 indicates no difference (dashed line), > 1 indicates enrichment of such substitutions in zebra finch, and <1 would be enrichment in chicken. \* Grubbs outlier test,  $p = 0.05$ .**

#### 4.3.9 Sharing of H+VL type 1 SNAAP substitutions in mammals

I also examined the shared human H+VL type 1 SNAAP substitutions in the vocal non-learning mammals in my analyses except the three species: monkey, cow and dog. I found most mammals have fewer sharing for both categories than those of zebra finch (33% for Gain category; 44% for Loss category), even though they are phylogenetically closer to human, elephant and microbat. The higher sharings (relative to zebra finch) were seen only in the apes (chimpanzee: 85%; orangutan: 60%) for the Gain category, and in some more species (chimpanzee 90%, tenrec 60%, opossum 57%, squirrel 50%, orangutan 50%, guinea pig 50%, armadillo 44%) for the Loss category, most of which (except the two apes) diverged in the early stage of mammalian evolution. Interestingly, I found the opossum, a marsupial, only shares 3% of Gain SNAAP substitutions with the three mammalian vocal learners, which is much lower than that for zebra finch (33%), contradicting with the closer relationship of marsupials to mammalian vocal learners. On the other hand, it has more Loss category SNAAPs (57%) than zebra finch (44%) shared with the three mammalian vocal learners. Taken together, these findings imply the possibility that the Gains were rapidly and independently accumulated in apes, elephant and microbat, as well as zebra finch, while the Losses occurred earlier and gradually with time of divergence in most mammals.

#### **4.3.10 The H+VL and H+DOM type 1 SNAAPs in Neanderthals**

Recently, ~60% of the Neanderthal genome was sequenced (Green et al. 2010), which allowed us to ask if one of our closest and extinct relatives shared SNAAPs overrepresented in vocal learners and domestics. Of the 14 H+VL type 1 SNAAPs from the Hom-Pan test, 11 were sequenced in Neanderthals and all had the same human substitutions. Likewise, of the 44 H+DOM type 1 SNAAPs, 30 were sequenced in Neanderthals and all had the human substitutions.

#### **4.4 Brief discussion**

Our SNAAP approach revealed certain nonrandom functional association with genes carrying the H+VL SNAAPs. Importantly, they occur at genes from both Hom-Mac and Hom-Pan tests for neural connectivity and audition, which are relevant to the vocal learning ability. Relative to the three vocal learners, both chimpanzee and monkey appear to lack SNAAPs in such genes, and that monkey further lacks more of them than chimpanzees as well as in genes involved in vocal production. Of these genes, the *ROBO1* and its partner genes in axon guidance pathway are particularly interesting, as their SNAAP patterns seem to have developed through the same evolutionary scenario, possibly via losses in early mammalian evolution. This finding, if true, may suggest the existence of ancestral substrate that was re-used by vocal learners. And in turn, it would support a deep homology for the convergent evolution of vocal learning.

The SNAAP nucleotide transition bias I found (heavily favoring A/G) is not due to neutral mutational mechanisms or a general natural selection over protein sequences, but more likely due to a special selective force. Further, the relaxation of this rule for type 1 SNAAPs shared among human and two domesticated species highlighted a difference between natural selection and artificial selection. The unusually high rate of transversional changes, which are considered to induce more dramatic genetic alternations than transitions (Zhang 2000), for type 1 H+DOM SNAAPs indicate that most of them could be deleterious. If true, this is consistent with the findings that these species have accumulated more deleterious mutations than their wild close relatives (Eyre-Walker and Keightley 1999; Wang et al. 2004; Björnerfeldt et al. 2006; Cruz et al. 2008). I suggest one likely cause is the reduced effective population size observed during both human evolution and the domestication of species.

Lastly, it should be noted that for swap analysis, one may replace any vocal learner with a vocal non-learner, not limited to the human-nonhuman primate swap. However, caution needs to be taken for choosing the pair of swapped species, as confounding systematic factors may be introduced. I required the choice of swap: (1) not to bring in unbalanced phylogenies in either trio, i.e. the phylogenetic relationship of species in both trios should not be very different, and (2) to make sure the pair of swapped species are not too far away in their relationship, since neutral mutations accumulate more rapidly than non-neutral ones and thus may make it hard to detect the

non-neutral changes. Overall, swapping human & NHP turns out to be the best choice, as the two species form a monophyly with a much more recent common ancestor than any other choice.

## **Chapter 5. Expression studies on the ROBO1 axon guidance pathway in birds**

### ***5.1 Introduction***

The specialized forebrain neural circuits for learned vocalizations and their direct projections onto the brainstem vocal nuclei are thought to be a definitive feature of vocal learning. Particularly, the forebrain direct projections have been shown experimentally in both humans and three vocal learning birds, and were not found in even very closely related vocal non-learners, e.g. chimpanzee and suboscine songbirds (Jarvis 2004). Genes involved in evolving such novel projections are largely unknown. Interestingly, in my SNAAP analysis, many genes with H+VL type 1 SNAAPs fall in the functional category of neural connectivity. It is possible that some of them could be responsible in forming the forebrain direct connections in mammals, and maybe in birds, too. Since mammalian vocal learners are generally difficult to study in the lab, I tested this idea in avian vocal learners.

The most promising candidate from my list is the ROBO1 gene that encodes an axon guidance receptor to navigate the growth of longitudinal axons. The members of the ROBO gene family include ROBO1 and ROBO2, play an important role to control the development of ascending or descending major axon tracts to or from the forebrain, and interneuron migration in the forebrain, through interactions with the ligands, SLIT1, SLIT2 and SLIT3 (Andrews et al. 2006; López-Bendito et al. 2007; Farmer et al. 2008; Dugan et al. 2011). Mutations on the ROBO1 gene are associated with human dyslexia

and speech sound disorder (Hannula-Jouppi et al. 2005). Its ligand SLIT1 serves a direct downstream target for the speech-language related gene FOXP2 (Vernes et al. 2007; Konopka et al. 2009), while the other ligand, SLIT2, has been suggested to have had significant positive selection in its promotor region in humans relative to other primates (Haygood et al. 2007). The ROBO1 SNAAP is among the 9 H+VL type 1 SNAAPs where neither chimpanzee nor monkey has the vocal learners' substitution. In addition, six more genes in the same axon guidance pathway have H+VL type 1 SNAAPs identified from Hom-Mac test. As discussed in Chapter 4, the implicated extensive loss of ancestral amino acid type at these SNAAP sites suggests that the ROBO1 axon guidance pathway might have served an ancestral substrate for vocal learning evolution, which already existed in the common ancestor of placental vocal learning mammals. Therefore, it is tempting to imagine that the ROBO1 axon guidance pathway may have played a similar role in vocal learning birds.

The product of ROBO1 gene was first discovered in *Drosophila melanogaster* to interact with Slit proteins to control midline crossing of central nervous system axons (Seeger et al. 1993; Kidd et al. 1998). This gene is highly conserved in zebrafish, mouse and human (Kidd et al. 1998). The human ROBO1 has two main isoforms ROBO1a and ROBO1b (DUTT) that differ in the starting exons and introns, with ROBO1a being longer (Clark et al. 2002). The two isoforms have distinct brain expression patterns in mouse (Clark et al. 2002; Nural et al. 2007) and humans, including the up-regulation of specific

splicing variants in fetal auditory cortex and Broca's area (Johnson et al. 2009). The developmental dyslexia and speech sound disorder is associated with a translocation breakpoint of the ROBO1 affecting only a ROBO1a splice variant but not ROBO1b variant (Hannula-Jouppi et al., 2005).

In chickens, both ROBOs (ROBO1/2) and their ligands SLITs (SLIT1/2/3) have been identified and their expression patterns were studied in developing limbs (Vargesson et al. 2001) and inner ears (Battisti and Fekete 2008). Like in mammals, ROBO1/2 also regulate commissural axon guidance in the avian spinal cord (Reeber et al. 2008).

ROBO1 has an extracellular domain consisting of five immunoglobulin-like (Ig) and three fibronectin type III repeats (Fn3), where the Ig domains are sufficient for binding to SLIT1 (Ba-Charvet et al. 2001; Battye et al. 2001; Chen et al. 2001). The cytoplasmic part of ROBO1 is highly unstructured and determines the repulsive response to SLIT1 (Bashaw and Goodman 1999). The SNAAP site of human ROBO1 is in the cytoplasmic part, ~ 54 aa downstream to the trans-membrane part.

As I showed in Chapter 3, the H+VL type 1 SNAAP substitution of ROBO1 in zebra finch is the same as that in chicken and different from those of the mammalian vocal learners. However, this does not refute the possible role for vocal learning of this gene in birds. To understand this better, it is useful to refer to the lessons learned from the FOXP2 studies. The two unique and functional substitutions in human FOXP2 not

found in other primates are not present in songbirds, either. Nonetheless, FOXP2 exhibits similar brain expression patterns, particularly in the striatum, and is necessary for learning complex vocalizations in both humans and songbirds (Lai et al. 2003; Haesler et al. 2004; Haesler et al. 2007). These prior findings suggested that the amino acid substitutions in some vocal learners could be useful indicators on genes co-opted for vocal learning in general. Likewise, here I suggest that the discovery of the ROBO1 SNAAP for vocal learners in mammals implies the possible co-option of this gene in vocal learning in general, considering the highly conserved nature of this gene. One possible scenario is that like FOXP2, ROBO1 in vocal learning birds may exhibit convergent expression different with unique spatial-temporal patterns. This may extend to other members of the ROBO1 axon guidance pathway, too. In this Chapter, I tested this possibility in vocal learning birds.

## ***5.2 Materials and methods***

### **5.2.1 Animals**

I collected the brains of three vocal learners and two vocal non-learners. The vocal learners were 3 male adult zebra finches (a songbird), 18 juvenile zebra finches (9 males and 9 females), 3 male budgerigars (a parrot), and 3 male Anna's hummingbirds. The vocal non-learners were 3 male ring doves, and 3 male quails. Except for hummingbirds, all animals were obtained from our breeding colonies at The Duke University Medical Center (USA). Anna's hummingbirds were obtained with the help of

Dr. Douglas Altshuler at the University of California, Riverside (USA) in a previous study (Feenders et al. 2008). For the developmental study, since it is not always possible to determine sex of the animals at young ages, as the full sex-specific plumage develops later in adulthood (>90 days), I collected zebra finch brains and toes at post hatch day 20, 35 and 65. From the toes, I extracted DNA and used PCR reactions on the CHD gene to identify the gender of those juvenile zebra finches following a described protocol (Wada et al. 2006). I then processed three male and female zebra finch brains from each developmental stage. All animal procedures were approved by the Duke University Institutional Animal Care and Use Committee.

### **5.2.2 Cloning ROBO1/2 and SLIT1/2/3**

I cloned cDNA fragments of Robo1/2 and Slit1/2/3 using specific primers on zebra finch brain cDNA. The primer sequences were designed from the zebra finch transcriptome (Wada et al. 2006) and genome (Warren et al. 2010) in our databases (songbirdtranscriptome.org and aviangenomes.org). For ROBO1, the primers 5' primer (5'- AGTCCCGTCTTTTACCTTCAC-3') and 3' (5'- CCCAGCCATTGATCATGGA -3') primers, which generated a 1.3kb PCR product. For ROBO2, the primers were 5' (5'- CAGGAGTACAAGATCTGGTGC -3') and 3' (5'- ATGCTGCTGTAAATGGCTCC -3'), which generated a ~600bp PCR product. For Slit1, the primers were 5' (5'- TCTCTCACTGCTCTCACTCT -3') and 3' (5'- GGTGGCGAGCAGGTTTCAGT -3'), which generated a 777bp PCR product. For Slit2, the primers were 5' (5'-

CTTGAATCTTCTTTCTTTGT -3') and 3' (5'- GATTGGCCAAGAGGTTTAGC -3'), which generated a 783bp PCR product. And for Slit3, the primers were 5' (5'- TCTCAAACCTGTTATCTCTTT -3') and 3' (5'- AGTTAGCCAGCAAATTAATT -3'), which generated a 700bp PCR product. PCR was performed with a 68–60°C touchdown protocol followed by: 3-min denaturation at 95°C, 16 cycles of 30-sec denaturation at 95°C, and 30-sec annealing (68 – 60°C touchdown), and 15-min extension at 72°C. These reactions yielded single products, as revealed by gel electrophoresis. PCR products were examined on 1% agarose gels, extracted from the gels, ligated into the pGEM-T Easy plasmid (Promega, Madison, WI), and transformed into XL-1 blue E. coli cells. Plasmid DNA was isolated, and the inserted cDNA was sequenced from the 5' and 3' ends, by using plasmid sequencing primers. The cloned sequences had 100% identity to the predicted ROBO and SLIT sequences from the draft zebra finch genome.

### **5.2.3 Radioactive *in situ* hybridizations**

Radioactive *in situ* hybridization was performed on brain sections as previously described (Wada et al. 2006). Dissected brains were immediately frozen in tissue Tek OCT (Sakura, USA) and stored at -80°C. The brains were then frontally sliced at 12µm thickness. To generate the riboprobes, the ROBO1 and SLIT1 inserts in the pGEM-T Easy vector were PCR amplified and purified. With the amplified DNA, either T7 or SP6 RNA polymerase was used to synthesize the antisense 35S-riboprobes, and the other was used to synthesize the sense 35S-riboprobes. Then 10<sup>6</sup> cpm of the 35S-probe was added to the

hybridization solution. Sections were fixed in 4% paraformaldehyde in PBS (pH 7.4) and hybridized at 65°C with the antisense 35S-UTP labeled riboprobes. The hybridized sections were first exposed to X-ray film (Biomax MR; Kodak) for 2 days, then dipped into autoradiographic emulsion (NTB2; Kodak), incubated for 10 to 14 days at 4°C, processed with Kodak developer (D-19) and fixer, and Nissl stained with cresyl violet acetate solution (Sigma, St. Louis, MO). All sections were mounted on Permount (Sigma-Aldrich). All dark field or bright field pictures were taken with an Olympus dissecting microscope (MVX10) and images were processed in Adobe Photoshop 7. The brain nomenclature used in the present study follows the new definitions (Feenders et al. 2008).

#### **5.2.4 Gene expression quantifications.**

Quantifications were conducted as previously described (Wada et al. 2006). Brain images were digitally scanned from a dissecting microscope connected to the Olympus MVX10 camera using the associated imaging software (Olympus Instruments, Inc.). Care was taken to use the same light settings across all images for quantifications. I used Adobe Photoshop 7.0 to measure the mean pixel intensities in the brain areas of interest from at least two adjacent sections on a 256 grey scale. Ratios of pixel density, representing differential expression, were then calculated from within a song nucleus and the region lateral to it in the arcopallium. The statistical significance of the mean differences was assessed by unpaired Student's t-test.

## **5.3 Results**

### **5.3.1 ROBO/SLITs show differential expression in song nuclei of songbirds**

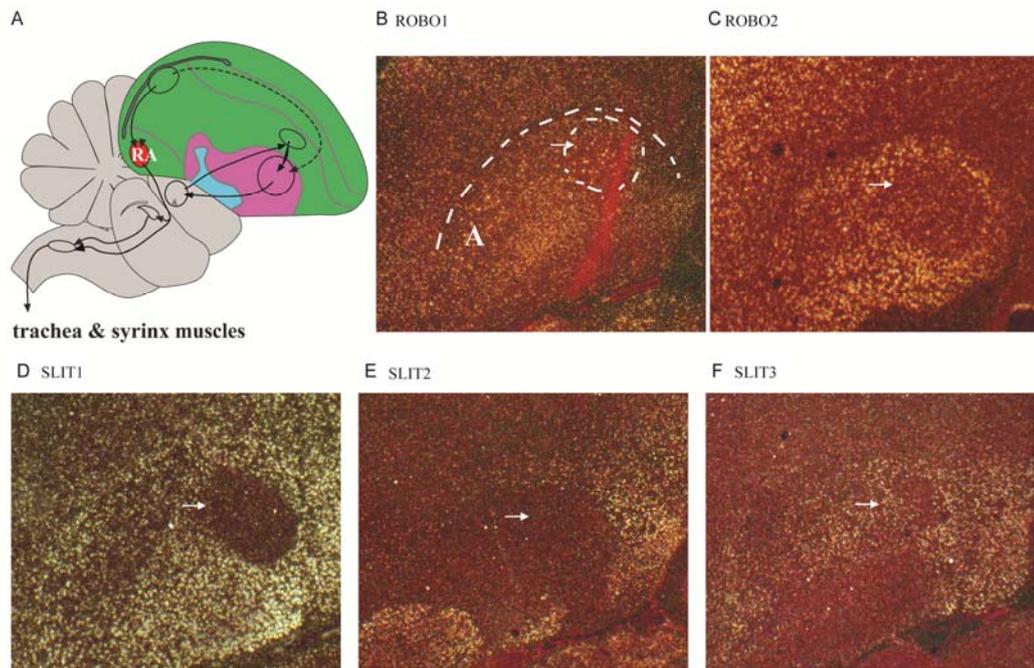
I hybridized the radioactive riboprobes of the ROBO and SLIT cDNAs to brain sections that contain either the arcopallium song nucleus RA or the 12th motor nucleus (nXII) in male adult zebra finches. The arcopallium song nucleus RA makes a direct projection to the brainstem (Figure 2A). This projection is believed to be unique in the vocal learners, as direct projections to other motor neurons are not found from the surrounding arcopallium and vocal non-learners do not have any forebrain direct projections to vocal motor neurons (Figure 2 and Chapter 1). The 12th motor nucleus contains neuronal somata that innervate the syrinx and the tongue of birds. As a control, I analyzed expression in nucleus supraspinalis (SSP) motor neurons that control neck muscles and do not receive a direct projection from the arcopallium (Devoogd et al. 1991).

I found high ROBO1 mRNA expression in the isolated neurons of RA, relative to the surrounding arcopallium, which had more uniform ROBO1 expression as in the other brain regions (Figure 17B). In contrast, ROBO2 was expressed at much lower levels in RA relative to the high expression in the surrounding arcopallium (Figure 17C). For the ligands, SLIT1 mRNA expression was nearly absent in RA relative to the surrounding arcopallium (Figure 17D). SLIT2 and SLIT3 had little difference of

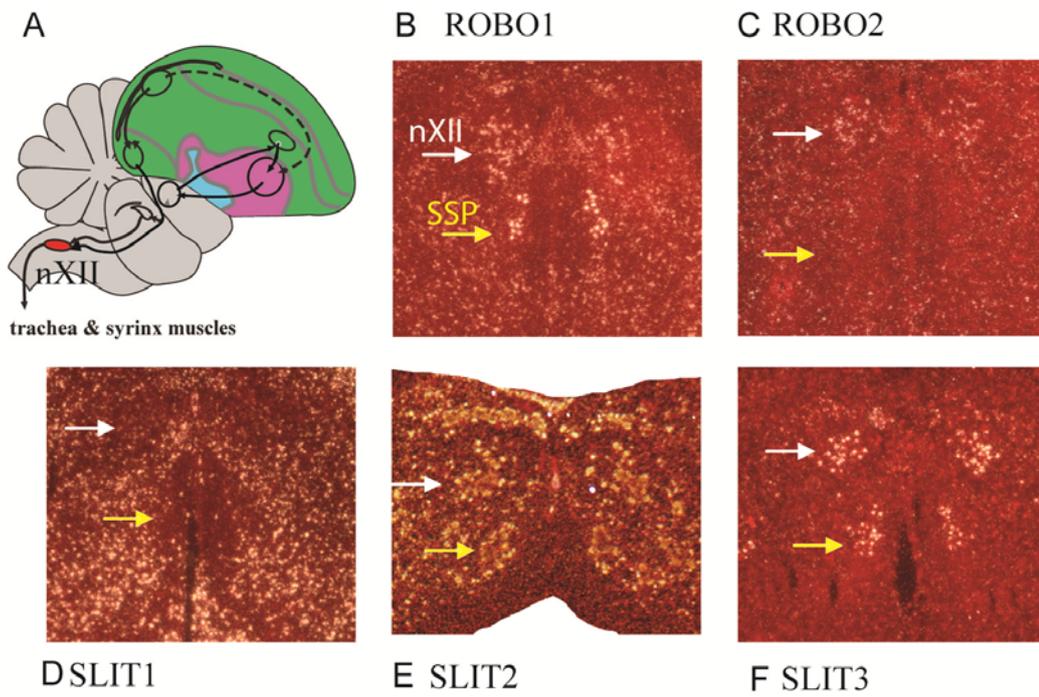
expression in RA relative to the surrounding arcopallium (Figure 17E, 17F), indicating that the differences found for ROBO1, ROBO2 and SLIT1 are specific to these genes. The differential expression of these genes is remarkable, since we know that from many other studies, most genes expressed in RA are similar to the surrounding levels when the birds are silent (Wada et al. 2006).

In the brainstem, I found that ROBO1 is highly expressed in both nXII and SSP motor nuclei (Figure 18B), while ROBO2 is expressed in nXII but nearly absent in SSP (Figure 18C). There is little SLIT1 mRNA expression (Figure 18D) but high expression of both SLIT2 (Figure 18E) and SLIT3 (Figure 18F) in the nXII and SSP. Further, SLIT3 appears to be relatively specific to motor neurons.

These findings demonstrate that at least in a songbird, several genes of the ROBO/SLIT gene family have evolved specialized regulation in the vocal pre-motor song nucleus and brainstem vocal motor neurons that have specialized neural connectivity. This made us wonder if such specializations occurred independently in other avian vocal learners, and whether I would see absence of such specializations in vocal non-learning species. I performed such experiments, as described in the next section.



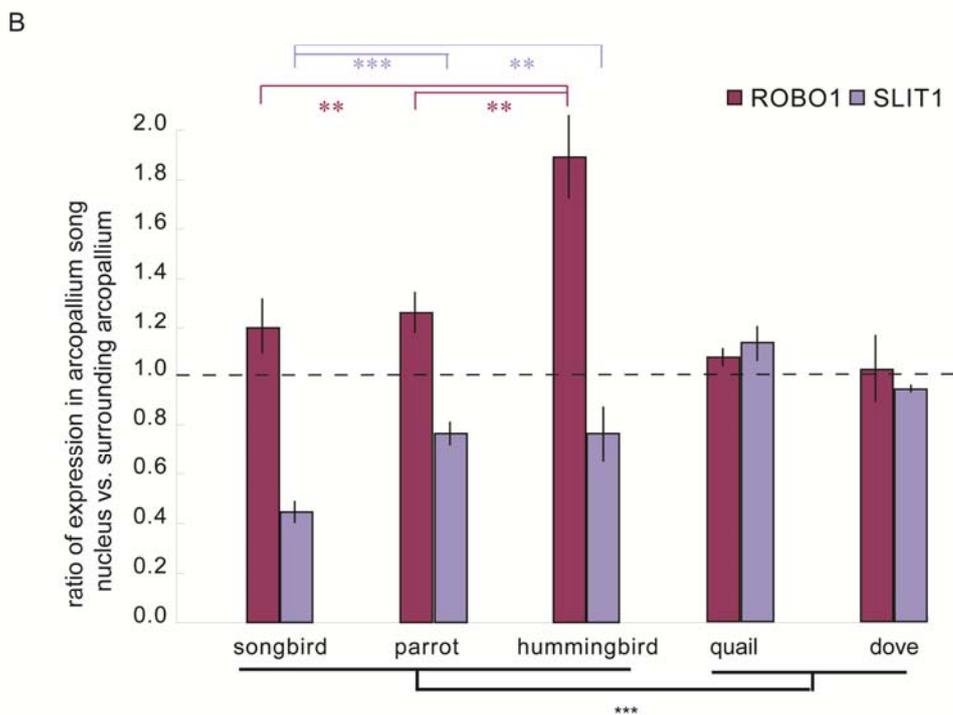
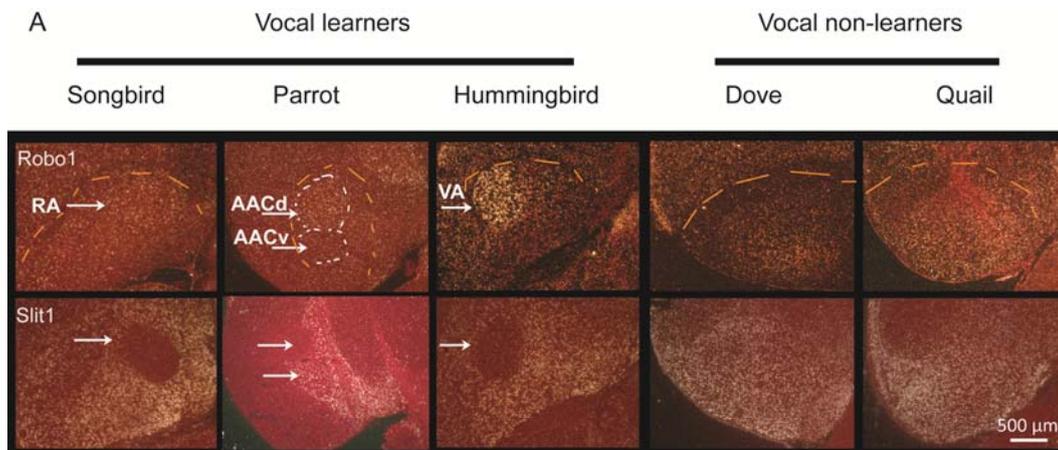
**Figure 17: The expression patterns of ROBO1/2 and SLIT1/2/3/ in the RA nucleus and surrounding areas of adult male zebra finches. (A) The anatomic location of RA nucleus; (B) ROBO1; (C) ROBO2; (D) SLIT1; (E) SLIT2; (F) SLIT3. A: arcopallium. White arrows: RA nucleus.**



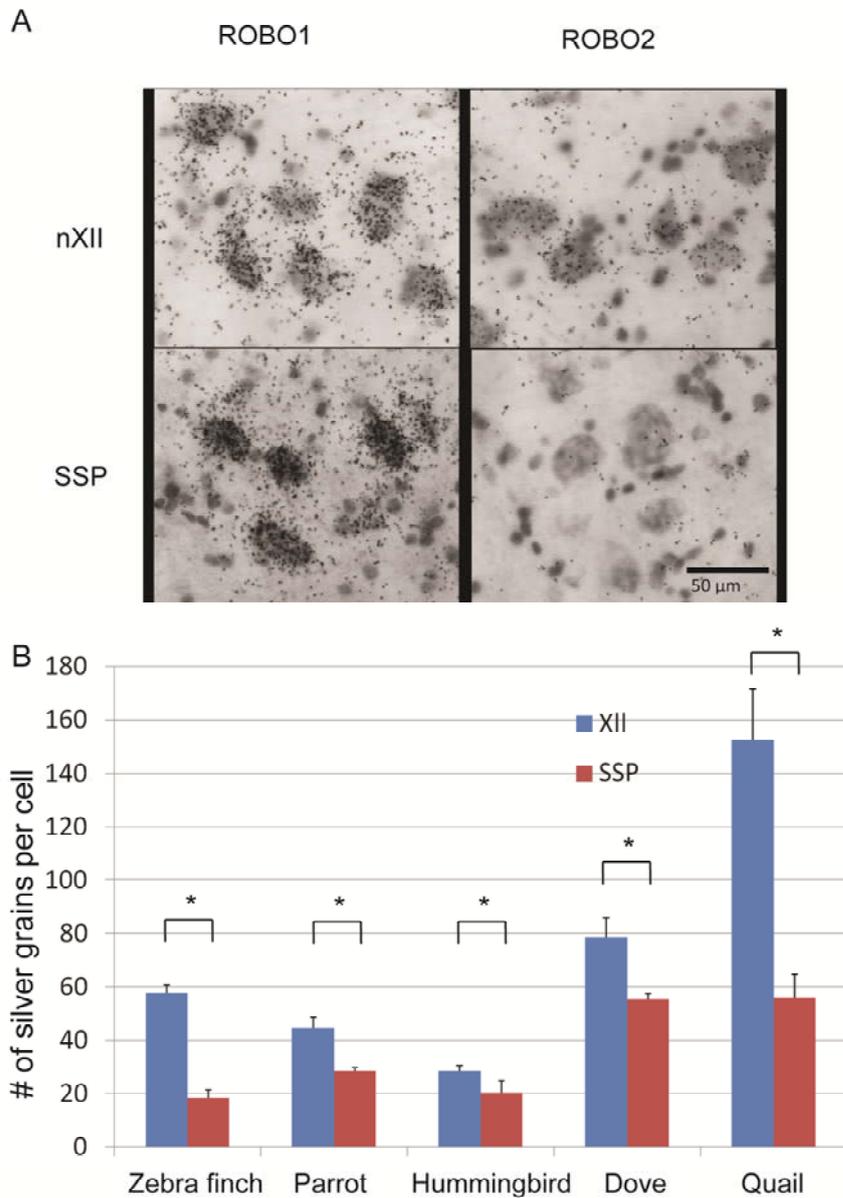
**Figure 18: The expression patterns of ROBO1/2 and SLIT1/2/3 in the XII nucleus and its surrounding areas of adult male zebra finches. (a) The anatomic location of XII nucleus; (b) ROBO1; (c) ROBO2; (d) SLIT1; (e) SLIT2; (f) SLIT3. White arrows: XII nucleus. Yellow arrows: SSP nucleus.**

### **5.3.2 Convergent expression patterns of ROBO1/SLIT1 in avian song nuclei**

Here I focused on the genes that showed the most robust differential expression in either the RA nucleus or the nXII motor nucleus: ROBO1, ROBO2, and SLIT1. Similar to the patterns in songbirds, I found higher ROBO1 mRNA expression and a lower expression of its repulsive ligand SLIT1 in the arcopallium song nucleus (AAC in parrots, and VA in hummingbirds) than the surrounding arcopallium (Figure 19A). More strikingly, in parrots, AAC has two subdivisions, a ventral subdivision that projects to the other forebrain song nuclei (AACv) and a dorsal subdivision (AACd) that makes the specialized direct projection to the vocal motor neurons as does songbird RA and hummingbird VA (Durand et al. 1997; Gahr 2000; Jarvis 2004). Only dorsal AAC showed the differential expression of ROBO1 and SLIT1. Quantifications revealed that Anna's hummingbirds has the most differential up-regulation of ROBO1 expression in the arcopallium song nucleus, while zebra finches have the most down-regulation of SLIT1 expression in its arcopallium song nucleus (Figure 19B). In contrast, no large differential ROBO1 and SLIT1 expressions differences were seen in the arcopallium in ring doves and quails (Figure 19). On the other hand, I found the exclusive expression of ROBO2 in nXII nucleus but not adjacent non-vocal motor neurons in both vocal learning and non-learning all bird species (Figure 20). These findings suggest a convergent differential expression of ROBO1 and SLIT1 in the arcopallium song nuclei of the three vocal learners, and a conserved high expression of ROBO2 in brainstem vocal motor neurons.



**Figure 19: (A)** The expression pattern of ROBO1 and SLIT1 in the frontal sections of the arcopallium (yellow dashed lines) of avian vocal learners and vocal non-learners. **(B)** Quantification of ROBO1 and SLIT1 expression in the song nuclei vs. surrounding arcopallium in songbirds, parrots, hummingbirds, ring doves and quails. \*  $p < 0.05$ ; \*\*  $p < 0.01$ ; \*\*\*  $p < 0.001$  (one tailed t tests). error bar: standard deviation.



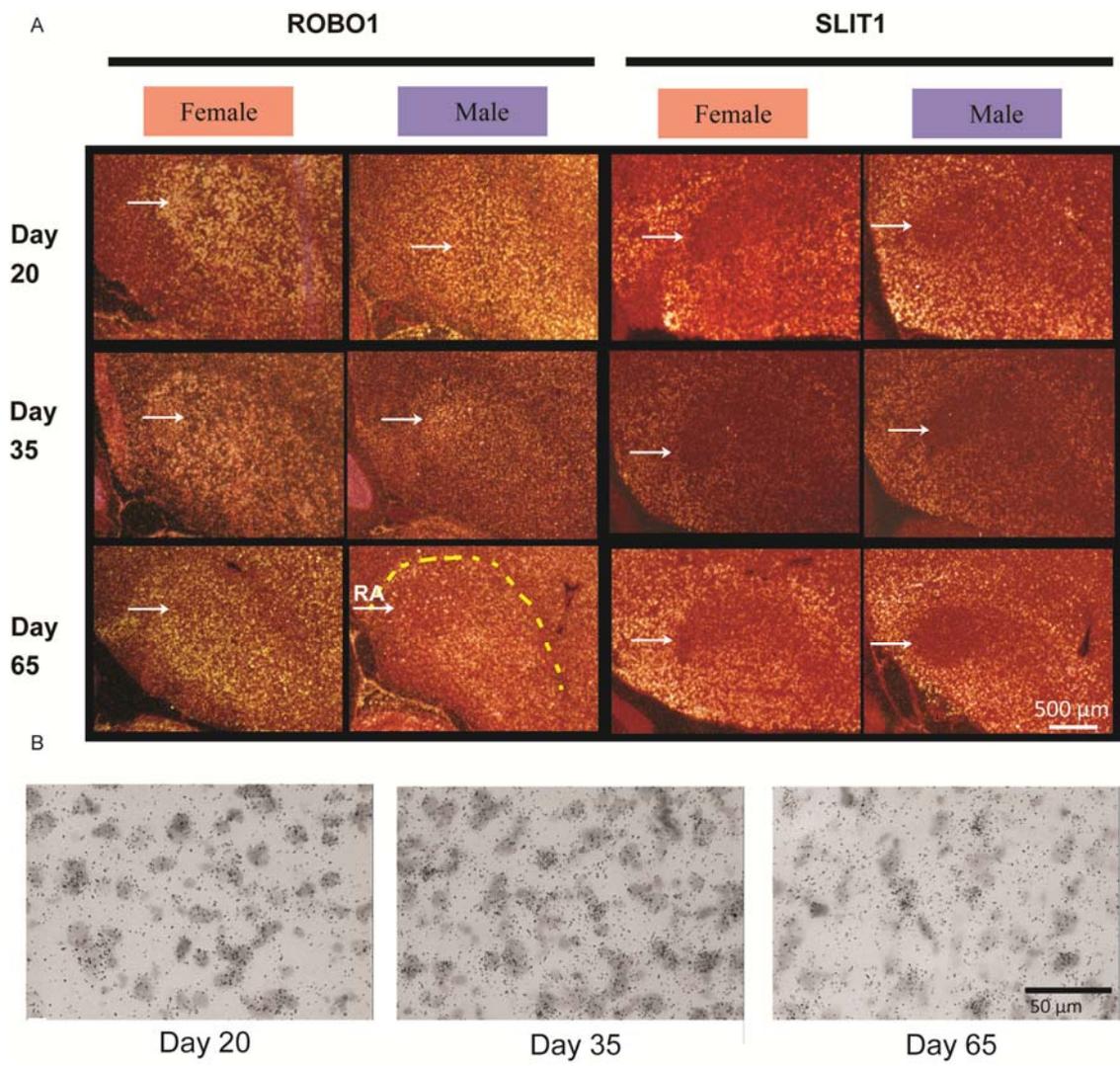
**Figure 20: (A)** The high-resolution pictures of nXII and SSP with ROBO1 or ROBO2 expression in zebra finches. Enriched expression of ROBO2 is only seen in nXII. White arrows: nXII. **(B)** Quantification of ROBO2 expression level in nXII vs. SSP across vocal learning and non-learning birds. \*  $p < 0.05$  (paired t-test). Error bar: standard error.

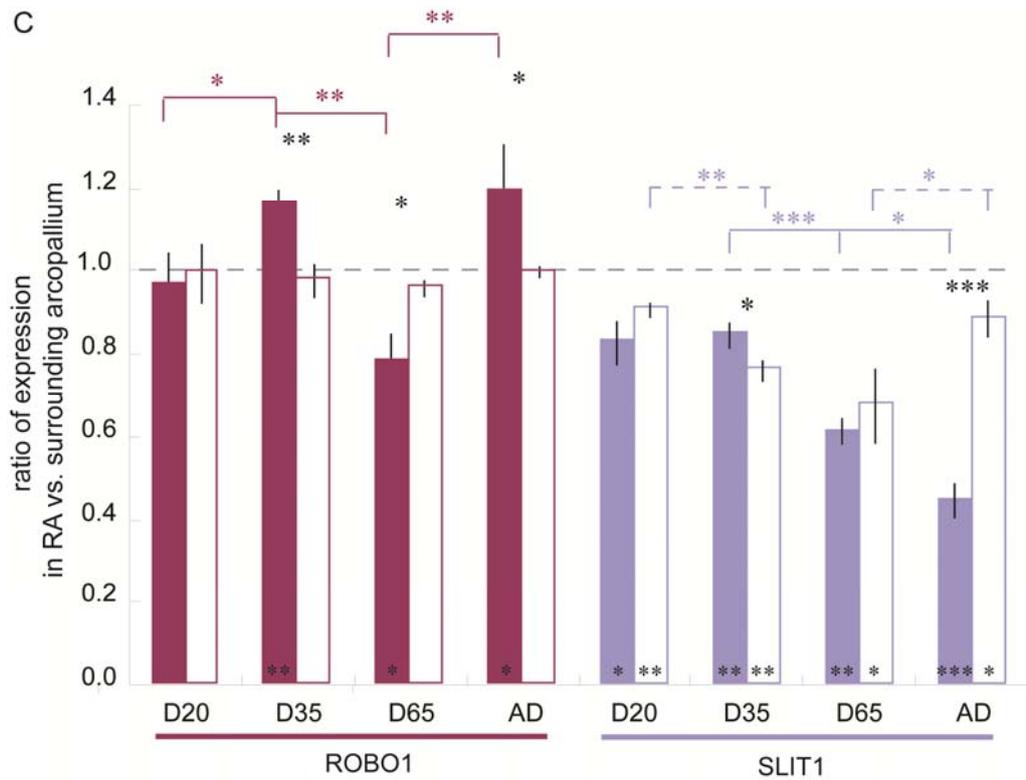
### **5.3.3 ROBO1/SLIT1 expression in juvenile zebra finch brain**

In zebra finches, only the males have the vocal learning ability and associated functioning brain pathways. Most sex differences develop at post-hatch stages when the critical period of vocal learning starts (around post hatch day 30; D30). The RA to nXII projection is already formed in both sexes by D18 post hatch, and as early as at D30, they continue to thicken in males but shrink significantly with a developmental loss of RA neurons in females (Johnson and Sellix 2000). This provide a means to determine if there is a correlation of ROBO1 and SLIT1 expression in the developing RA-nXII projections by testing if their expression are correlated with sex dimorphic, developmental changes. For this purpose, I collected the forebrain sections of juvenile male and female zebra finches at D20, D35 and D65. The projection at D20 is similar to D18. While the male birds start sensorimotor learning, i.e. hearing and trying to imitate the tutor songs, the sex differences in the RA-nXII projection emerge after D30 (Johnson and Sellix 2000).

I found both males and females exhibited similarly strong ROBO1 expression in the arcopallium by D20, and there is no differential expression between RA and the surrounding arcopallium (Figure 21). Significant differential expression patterns of ROBO1 between RA and the surrounding arcopallium were present in later stages (D35, D65 and adulthood) and only in males (Figure 21). In contrast, both males and females showed lower SLIT1 expression in RA relative to the surrounding arcopallium throughout all measured later stages of juvenile development into adulthood (Figure 21).

In addition, there was a general reduction in ROBO1 expression in the arcopallium region excluding the RA nucleus in adulthood relative to the developmental stages (D20-65), whereas the SLIT1 expression in this area remained similar (Figure 21).





**Figure 21: (A) The expression of ROBO1 and SLIT1 in RA of male and female zebra finches during developmental stages: Day 20, 35 and 65. (B) High resolution pictures of RA nucleus. (C) Quantification of ROBO1 and SLIT1 expression in RA vs. RA surrounding arcopallium in male (filled bars) and female (open bars) zebra finches. \*  $p < 0.05$ ; \*\*  $p < 0.01$ ; \*\*\*  $p < 0.001$  (one tailed t tests). Error bar: standard deviation.**

## **5.4 Brief discussion**

The avian arcopallium contains large regions with parallels to the mammalian cortex. Particularly, RA projection neurons are proposed to be analogous to the layer V neurons of the facial motor cortex of mammals that sends direct projections to the brainstem vocal nucleus in humans (Jarvis 2004), while the arcopallium surrounding RA may be analogous to the layer V neurons of mammalian motor cortex (Feenders et al. 2008). Here I detected the expression of all ROBOs and SLITs in RA and the surrounding arcopallium in the adult male zebra finch brain. As a comparison, three regions (layer V, entorhina cortex and piriform cortex) of the adult rat cortex are found to express all five ROBOs/SLITs genes (Marillat et al. 2002). Of them, only layer V contains the descending motor pathway neurons. Therefore, these findings supported the analogy between the RA surrounding arcopallium to the motor cortex (layer V) of mammals. It remains to examine the detailed expression patterns of these molecules in layer V.

I found specialized expression patterns of ROBO1, SLIT1 and ROBO2, but not SLIT2 and SLIT3, in the RA nucleus of adult male zebra finch, relative to the surrounding arcopallium. Only ROBO1 is up-regulated in RA, while ROBO2 and SLIT1 are down-regulated in RA. This difference between ROBO1 and ROBO2 in RA is remarkable, since the two ROBOs typically act together for axon guidance in the forebrain (López-Bendito et al. 2007) and in rats, they co-express in most telencephalon regions (Marillat et al. 2002). Recent studies suggested that in mice, ROBO1, but not

ROBO2, directs the migration of interneurons through the striatum and into the cortex by a ROBO1-NRP1 interaction that modulates semaphoring signaling (Hernández-Miranda et al. 2011). It is possible that a similarly ROBO2-independent ROBO1 axon guidance mechanism may be adapted in the RA nucleus. The down regulation of ROBO1's ligand, SLIT1 in this song nucleus could be due to the same adaptive changes, as the complementary expression pattern of ROBO1 and SLIT1 in RA is consistent with the fact that they interact to mediate a repulsive effect in forebrain development (Bagri et al. 2002). Finally, as illustrated in my cross-species studies of multiple avian vocal learners and vocal non-learners, the expression of ROBO1 and SLIT1 patterns are tightly correlated with the presence of this arcopallium song nucleus in general.

I found that only the dorsal division (AACd) of this arcopallium song nucleus (AAC) in parrots has the specialized pattern for ROBO1 and SLIT1, but not the ventral division (AACv). Both AACd and AACv receive the same input from ventral part of the oval song nucleus of the anterior nidopallium (NAO), the magnocellular song nucleus of the dorsomedial thalamus (DMM) and the central song nucleus of the lateral neostriatum (NLC) (Durand et al. 1997; Jarvis 2004). The two divisions differ in their output projections: AACd sends out descending projections to brainstem respiratory and syringeal motoneurons including those that project directly to nXII, while AACv gives rise to ascending projections back to AACd and two other telencephalon nuclei

(Durand et al. 1997; Jarvis 2004). Therefore, the specialized expression patterns of ROBO1/SLIT1 may be relevant to this anatomical difference.

I found SLIT1, but not SLIT2/3, was low in nXII that is the target of RA direct projections. Abundant in vivo evidence showed that both SLIT1 and SLIT2 regulate the corticofugal, callosal, thalamocortical, serotonergic, and dopaminergic projections that presumably express ROBO1 or/and ROBO2 in the embryonic forebrain by preventing axonal entry into certain regions (Bagri et al. 2002). The roles of the two SLITs may overlap only partially. In some cases (e.g. corticofugal or thalamocortical projections), the loss of SLIT2 activity may uncover the function of the SLIT1 (Bagri et al. 2002). It is possible that the down regulation of SLIT1 provides the permissive environment for RA-projections that presumably express ROBO1 to enter nXII. However, since the RA-XII projection forms as early as in D18, to explore this idea, it is necessary to look at the earlier stages that I did not include in my study.

Nonetheless, I found that ROBO1 is developmentally regulated and this is only seen in male but not female zebra finches. Such differential expression between RA and the surrounding arcopallium did not occur at D20, i.e. before the critical period of vocal learning. Such differential expressions seen in males could be due to different reasons: the D20-35 and D35-65 differential expression patterns in RA and surrounding arcopallium were largely attributed to the changes in expression of ROBO1 that occurred in RA, whereas the D65 differential expression pattern was most likely due to

the decrease of ROBO1 in the surrounding arcopallium. In contrast, the down-regulation of SLIT1 in RA appeared as early as D20 in both sexes and persists through all stages. The developmental results suggest that the expression of ROBO1 but not SLIT1 could be associated with pruning the RA-nXII projections during developmental critical periods.

## **Chapter 6. Discussions and future directions**

In this final chapter of my thesis, I discuss the broader implications of my findings and SNAAP approach. I highlight the strength and weaknesses of my findings, and what type of studies can be conducted in the future to address them.

### ***6.1 Comparison of dN/dS and SNAAP analyses***

Our dN/dS analysis and SNAAP analysis were both motivated by the same assumption that the convergent evolution of vocal learning involved coding region changes in the same genes and both used explicit evolutionary models to assess the significance of such change. While the dN/dS analysis failed to reveal any significant results, this notion was supported by the SNAAP analysis results which revealed significant changes in coding regions associated with species that share convergent complex trait evolution. I suggest two major reasons may underlie this success of SNAAP approach over dN/dS approach.

First, a single protein may be subject to more than one type of adaptive changes in species with and without the trait, respectively. In this case, even if selection exists for convergent evolution on a gene, it might not be identified by a dN/dS approach that takes the changes over the whole protein sequences into account. For example, up to 90% of the proteins in eukaryotes consists more than one domain (Apic et al. 2001). Each domain is a distinct, compact and stable protein structural unit, can fold independently and acts as an independent evolutionary unit during protein evolution (Koonin et al.

2002). For the transmembrane receptor-like kinases (RLKs) which have important roles in growth, development and defense responses, their extracellular domains that perceive signals evolve much faster than the intracellular domains which propagate the signals, and different extracellular domains can harbor changes in different member genes of this family (Shiu et al. 2004). In this case, the dN/dS approach cannot distinguish the selection from convergent evolution over one domain from the irrelevant selection at other domains within the same protein. This may be improved by focusing on the dN/dS ratios for each domain, yet the identification of protein domains poses another challenge. In contrast, this issue was circumvented in the SNAAP approach by focusing on individual sites.

Secondly, it is possible that the dN/dS approach was too stringent to detect the subtle changes for complex traits. Now it is recognized that the adaptation might not occur as once thought via the sequential fixation of individual beneficial mutations, as many mutations exhibit signs of epistasis in which case the beneficial effect of a mutation are conditional on certain genetic backgrounds, i.e. other genes (Weinreich et al. 2005). Modeling such dependent sites among different genes are not made possible in present evolutionary models. These weak signals of selection are very likely to be missed by the contemporary computational approaches for detecting positive selection in protein-coding genes, including the branch-site model and others, which are heavily biased to sites with high ratios of non-synonymous vs. synonymous mutation rates, as

were shown both experimentally and computationally (Yokoyama et al. 2008; Nozawa et al. 2009). In support of this idea, as I revealed in this study, the genes with a systematic accumulation of mutations in humans but not in non-human primates, possibly from a relaxation of selective constraints, were not well captured by the dN/dS approach.

I caution, however, our analysis does not determine if the revealed SNAAP substitutions are directly causal for that trait. Quite a few SNAAPs showed up in functionally important domains or regions. For example, the E2F3 SNAAP is in a potential dimerization region that determines interaction with the differential regulation of transcription factor proteins (Giangrande et al. 2003). Another SNAAP in PARP1 lies in the catalytic domain responsible for Poly (ADP-ribose) polymerization (Simonin et al. 1990). The 10aa region upstream of the EFNB1 SNAAP includes 4 sites with SNPs unique to patients with craniofrontonasal syndrome, a disease caused by defects in EFNB1 (Twigg et al. 2006). However, the roles of individual SNAAPs like these would have to be further explored with experimental manipulations.

## ***6.2 New insights into vocal learning evolution in primates***

Using the SNAAP approach, I identified significant association with vocal learners for type 1 SNAAPs. Both monkey and chimpanzee lack the SNAAP substitutions in human in genes that participate in neural connectivity development and in genes that are critical in processing auditory information. Our GO analysis further showed that some of these genes belong to GO groups that were non-randomly enriched,

including the cell adhesion category that is directly relevant with development of neural connectivity. Interestingly, monkey lacks the human SNAAP substitutions in more of these genes, as well as those in genes that when mutated can cause disabilities in vocal tracts, e.g. FRAS1 and GDAP1. These results may suggest (1) a possibly continuous evolution of advanced vocal development from monkey, to chimpanzee and human, or (2) a pre-adaptation of substrates that were later used in humans for vocal learning.

All primates are social animals and communicate frequently with their conspecifics. Primate vocal development has been studied with some of the standard experimental paradigms used in songbird research, e.g. deafening, social isolation or cross-fostering. Changes in the vocalizations during development are reported in early studies on both Old World (Seyfarth and Cheney 1986; Gouzoules and Gouzoules 1990) and New World monkeys (Lieblich et al. 1980; Elowson et al. 1992). But they are most likely due to maturational factors rather than acquired modifications (Hammerschmidt et al. 2001). Deafening either the infant or adults did not affect normal calls in squirrel monkey (Talmage-Riggs et al. 1972; Winter et al. 1973). Both short and long-term vocal convergence were observed in adult monkeys but are largely limited to subtle changes within innate specific call types (Sugiura 1998; Snowden and Elowson 1999). These findings suggest little evidence for vocal learning in monkeys. On the other hand, apes may have more vocal plasticity, and they are more quiet and do not vocalize as much. They can imitate gestures, e.g. orofacial movements, body postures, and locomotion

patterns, and this ability used to be thought absent outside of humans and apes but was lately revealed in monkey infants (Ferrari et al. 2006). Analyses of the acoustic variations in primates suggest that apes have "dialects", i.e. geographic variations that are both phonologically and lexically different, which are not found in monkeys. Abnormal vocalizations were observed in isolate-reared chimpanzees (reviewed in (Egnor and Hauser 2004)) but not in monkeys (Hammerschmidt et al. 2001), yet the influence from changes in social behavior cannot be excluded. Convergence in acoustic structures was also observed in paired chimpanzees (Mitani and Gros-Louis 1998), and remarkably, it is reported that a novel distinct syllable (the Bronx cheer variant) can be introduced to a captive colony of chimpanzees (Marshall et al. 1999). However, this sound is generated by the lips and not the vocal organ. Lastly, at the anatomic level, putative homologues in terms of sound processing (not production) of the anterior speech-language zone (Broca's region) in the human ventrolateral frontal lobe have been proposed in both apes and macaque monkeys (Petrides et al. 2005). Moreover, it was suggested that left-right hemispheric asymmetries in Broca's area might be present in the corresponding regions of other great apes besides humans (Cantalupo and Hopkins 2001).

Interestingly, both monkey and chimpanzee lack the vocal learners' SNAAP changes in genes expressed in the inner ear hair cells (USHBP1, E2F3 and CASP8AP2). The inner ear hair cells sense sounds and deflect to trigger the release of transmitters to the terminal of the afferent fibers, converting the sound to electric signals. In both

songbirds and humans, loss of inner ear hair cells leads to song/speech deterioration similar to those after surgical deafening (Woolley and Rubel 1999). It remains to be seen if there is an actual difference in the inner hair cells between vocal learners and non-learners. Alternatively, these genes expressed in the inner ear cells may also play a role in the central auditory pathway. For example, one potassium channel gene *KCNQ4*, mutated in a form of dominant deafness, is expressed in both the inner ear cells and the central auditory cells but not in most other parts of the brain (Kubisch et al. 1999). In fact, it was recently shown that, unlike other sensory stimuli, monkeys have a weak long-term central representation of the auditory stimuli relative to human (Fritz et al. 2005). As a third possibility, there may be a direct link between inner ear hair cells and auditory memory and thus vocal learning. The study on patients with hearing loss and spontaneous musical auditory perceptions suggested that the dysfunction of their inner ear might remove the inhibition on the neuronal groups storing auditory memory (Goycoolea et al. 2006).

It remains possible that these SNAAP substitutions of vocal learners are for some other traits shared in them. For example, I searched the literature for known hearing differences between human and non-human primates and found that humans maintain a uniquely high sensitivity from 2-4 kHz, a region that contains relevant acoustic information in spoken language (Martínez et al. 2004). The specific tuning in perceiving auditory information is also seen in some bats whose ears are tuned to a second

sensitivity peak at low frequencies (8–20 kHz) besides ultrasonic sounds, and this secondary sensitivity peak is presumably an adaptation to wide-range social communication calls (Gridi-Papp and Narins 2009). Also, elephants are highly sensitive to low-frequency sound and can both produce and detect seismic signals that transmit from the ground to the feet and up to the ears through bone conduction (Gridi-Papp and Narins 2009). These specializations of the three vocal learners might underlie their shared SNAAP changes, and there may be other possibilities, too. It is possible that they have been linked to, or even used as critical substrate for the further evolution of vocal learning ability. An alternative interpretation is that they are associated with other shared traits. Resolving these interpretations again requires experimental studies.

### ***6.3 Co-option of the *ROBO1* axon guidance pathway in vocal learning***

Our results revealed convergent amino acid substitutions of multiple genes in the *ROBO1* axon guidance pathway under non-neutral selection in the terrestrial vocal learning mammals. More interestingly, almost all these SNAAPs were in the Loss category, meaning that vocal learners maintained the ancestral amino acid types at these sites while most vocal non-learning mammals did not.

I envision two possible scenarios about when the relaxed constraints on ancestral amino acid types could have occurred in mammals to shape the Loss category SNAAPs. In one scenario, the ancestral amino acid types were lost multiple times, while three vocal learners and their ancestors always maintained the ancestral form. In the other

scenario, these ancestral types were lost once at a much earlier time point and were recapitalized independently in the vocal learners through reversal mutations. Reverse mutations refers to the situation where natural selection first favors a mutation, then favors its removal, and later still favors its ultimate restoration. This type of mutations can contribute substantially to the efficient evolution of novel complex functions of a protein (Crill et al. 2000; Depristo et al. 2007; Rokas and Carroll 2008).

Here I argue that the second scenario could be more likely, because almost all vocal non-learning placental mammals in my analyses (except the chimpanzee) has less shared Loss category SNAAP substitutions with the three vocal learners than opossum, a marsupial mammal. Without a single loss event occurring at an earlier time point, one would expect to see higher percentages of shared substitutions than opossum, especially for species that are very close to the vocal learners, e.g. orangutan. But in fact, orangutan only has a shared percentage of 50%, still smaller than the 57% for opossum.

Interestingly, most vocal non-learning species with higher shared percentages diverged early in mammalian evolution. They included two rodents (squirrel and guinea pig) have shared percentages (50%) that are at least 2.5 times of that of the other 4 rodents (rat, mouse, rabbit, pika; 18%-20%). This may indicate a secondary loss in other rodents. But it is also possible that these two rodents have independent reversal mutations, not to the level of that in humans.

If the above postulated loss of ancestral amino acid types in early mammalian evolution is true, then it is intriguing about how they got fixed by occurring in highly conserved regions which may result in deleterious effects. In addition, the Loss category SNAAPs showed no difference of sharing with zebra finch vs. chicken, supporting that these changes are less likely to be positively selected. In the course of early mammalian evolution, one key factor for tolerating such deleterious mutations could be the decline in population size. In mammals, there is a general evolutionary trend toward larger body size (Alroy 1998; Van Valkenburgh et al. 2004), while body size is often inversely related to the population size (Damuth 1981). For example, mitochondrial protein-coding genes of large mammals have a higher rate of accumulation of the more harmful nonsynonymous substitutions (as opposed to synonymous substitutions) and radical substitutions (as opposed to conservative substitutions) than small mammals (Popadin et al. 2007).

For birds, our expression studies in birds suggest ROBO1 is co-opted in the specialized vocal motor output nucleus of vocal learning birds. Whether it participates in forming the unique direct projections to the brainstem vocal nucleus, needs more investigation. It would be interesting to explore the ROBO1 expression patterns in the analogous face motor cortex area in human and other mammalian vocal learner or non-learners.

In this regard, recent studies showed that a splice variant of ROBO1, called ROBO1a, was highly enriched in the temporal auditory neocortex and/or temporal association neocortex, while ROBO1b was enriched in the prefrontal neocortex where face motor cortex and Broca's area develops (Johnson et al. 2009). Since the Broca's area is proposed to be similar to songbird IMAN song nucleus, not the RA (Figure 2; Jarvis 2004), it remains to be seen if ROBO1 is also differentially expressed in the songbird IMAN nucleus, and this appears true in my preliminary study (data not shown). Also, it is possible that ROBO1 is co-opted in the vocal communication areas in both songbirds and humans, but in different parts of the system. Another scenario is that ROBO1 may be differentially expressed in the facial motor cortex in humans at some other time points to set up the critical circuitry for vocal learning in humans. As an example, it was found that FOXP2 was differentially regulated only during fetal development in the human frontal neocortex (Johnson et al. 2009).

#### **6.4 Non-exclusive SNAAP substitutions in vocal learners**

I found few H+VL type 1 SNAAP substitutions are exclusive to the three mammalian vocal learners, and none are exclusively shared in all six known vocal learners whose sequences are available (human, elephant, microbat, dolphin, zebra finch, parrots). These findings may indicate that one or more other untested species could have the trait, or that many vocal non-learning species have developed different degrees of the genetics necessary but not sufficient for the trait, or that different vocal learners have

developed their own adaptations through various combinations of substitutions, some manifested as convergent SNAAP substitutions.

Here I argue that the impact from first and second scenarios, if exist, may be only minimal. In the phylogenetic screening step, I performed the permutations of species labels for 100 times, which is only a tiny fraction of all possible permutations ( $23! \approx 1022$ ). Therefore, if the assumed vocal non-learners have different degrees of vocal learning ability or the required genetics, it will introduce a substantial bias to the resultant SNAAPs. In this case, different runs of SNAAP analyses would lead to large differences in the resultant sets of SNAAPs from each run. However, I only observed <1% differences for type 1, 3-4 and <10% differences for type 2 across multiple runs for H+VL SNAAPs.

The third scenario may be very likely the answer, i.e. due to differences in the strategies for evolving voluntary control of the specialized vocal apparatus in different vocal learners. Specifically, in terrestrial mammals, sounds are generated as the lungs expel air through the larynx to cause the vocal folds to vibrate, which are subsequently filtered by the shape and movements of the upper respiratory tract, either from nose or mouth (Jürgens and Ploog 1981). Compared to the non-human primates, humans lost the air sacs in their evolution but have a much more flexible supralaryngeal vocal tract to manipulate and a better control of lungs that can produced controlled air bursts within an extended exhalation (Maclarnon and Hewitt 2004; Ghazanfar and Rendall

2008). As marine mammals, cetaceans also use the vocal folds of the larynx for generating most or all of the initial sound vibrations. But their use of air is in such a highly restricted way that the airflow from the lungs at vocalization is captured into internal reservoirs and then back to the lungs, so as to make multiple or long vocalizations efficient (Reidenberg and Laitman 2010). Moreover, the air sacs in cetaceans may participate in generating underwater vocalizations (Reidenberg and Laitman 2008), and certain sounds may be emitted without use of the larynx at all in cetaceans (Mackay and Liaw 1981). Unlike mammals, birds generate sounds with a unique vocal organ, the syrinx that bears functional similarities and dissimilarities with the mammalian larynx. They also have a more rigid lung, whose ventilation comes from the bellow-like activities of air sacs (Riede and Goller 2010). As one example, in songbirds, the vocal apparatus has been adapted for high speeds, for which temporal patterns and fast modulation of sound features become more important (Riede and Goller 2010). It is reasonable that the diving adaptations in these species might have resulted in a different mechanism of voluntary control of vocalizations, e.g. involving a deliberate coordination of those tracheal, pharyngeal, laryngeal and nasal air sacs. Therefore, the differences in respiratory mechanisms and vocal organs may underlie a sufficient difference in fine motor control of the vocalizations among different vocal learners, and thus differences in their genetic mechanisms. What may be important for

the evolution of vocal learning is not a single or a few commonly shared SNAAPs, but specific combinations of multiple SNAAPs from an ancestral pool of substrate.

Lastly, coding region changes are not the only way to evolve a novel complex trait. The regulatory non-coding changes of a gene may play a role, too. In fact, the non-coding regulatory elements are considered better targets for changes, as they are organized in a modular fashion and can tolerate more changes (Wray 2007). Also, the effect of these changes are most likely co-dominant, i.e. quickly becoming the target of natural selection, but appear limited in scope so that no greater tradeoff of the pre-existing function will be made (Carroll 2005; Wray 2007). In our case, for example, the zebra finch ROBO1 does not have the SNAAP substitution shared by human, elephant and microbat, but exhibited the differential expression between the arcopallium song nuclei and the surrounding arcopallium, which is correlated with the presence of the vocal learning trait in birds. I suggest such temporal and spatial changes in song nucleus expression in vocal learning birds may probably reflect a change in regulatory regions.

### ***6.5 The legacy of domestication and early human evolution***

The excess of homoplastic substitutions shared by humans, cow and dog, and particularly, and their excess of underlying transversional nucleotide changes, the absence of enriched GO ontology or amino acid substitutions, suggest to us a sign for a widespread relaxation of negative selection on these SNAAPs. These findings are consistent with the elevated deleterious mutation rates found in these species, relative to

their wild close relatives (Eyre-Walker and Keightley 1999; Wang et al. 2004; Björnerfeldt et al. 2006; Cruz et al. 2008).

In domesticated animals like dogs and cows, two major factors may account for this increased rate: (1) the relaxation of selective constraints due to demographic factors or selective breeding; and (2) a reduction of effective population size at loci linked to those under artificial positive selection. Analyses in the Bovini tribe (including domestic cattle, buffalo, yak and bison) revealed higher dN/dS ratios in the lineages of domestic cows, which were comparable to those dN/dS ratios due to within species polymorphisms, thus excluding the role of positive selection (Maceachern et al. 2009). Further, the bison lineage, though with a recent population bottleneck, only exhibits a low dN/dS rate, which indicates a reduction in effective population size alone is not sufficient. In light of these findings, I argue that the selective breeding among small populations is likely an indispensable component in elevating the mutation rates.

On the other hand, the size of ancestral human populations is known to shrink and expand often. It is estimated that the effective population size ( $N_e$ ) of human ancestors before 1.2 million years ago was 18,500, which is a strikingly small population for a species spread across multiple continents, compared to the effective population sizes of chimpanzees (21,000) and gorillas (25,000) that inhabit only part of a single continent (Huff et al. 2010). There is also evidence for population bottlenecks following the out-of-Africa expansion (Reich et al. 2001; Zhang et al. 2004; Tenesa et al. 2007). Most

interestingly, the recent analyses of the Neanderthals genome indicated possible interbreedings between Neanderthals and humans (Green et al. 2010). The ancestor of Neanderthals started migrating out of Africa as early as ~600,000 yr ago and later arrived in Europe, while modern humans migrated out of Africa at least ~200,000 yr ago and reached Europe about ~40,000 yr ago to eventually replace the Neanderthal population (Finlayson 2005).

These findings could have created similar conditions for ancestral humans and domesticated animals, in that selective inter-breedings among small populations presumably occurred in both cases. Our finding that all 30 H+DOM type 1 SNAAPs shared by both modern humans and Neanderthals may suggest a hypothesis that the common ancestor of them experienced such interbreedings, which would date at least over 600,000 yr ago, i.e. before the moving out of Neanderthal from Africa. Such interbreeding among ancestral human populations, if exist, would indicate an evolutionary process similar to the selective breeding of small populations of domesticated animals, which may lend support to the once popular idea that human ancestors had "self-domesticated" themselves (Brune 2007).

Our findings of the H+VL type 1 SNAAPs in the Neanderthal genome also support their shared traits with humans for vocal learning. The human specific FoxP2 mutation thought to be associated with the evolution of speech is also present in Neanderthals (Krause et al. 2007). Combined with my results on H+VL and H+DOM

type 1 SNAAPs, this suggest the possibility that the common ancestor of modern human, Neanderthals, and presumably also other hominid species, had evolved the genetics for complex vocal communication abilities. I postulate that this could have facilitated or/and been a result of inter-breeding.

### ***6.6 RNA editing as a potential mechanism for creating SNAAPs***

Most type 1 SNAAPs are shaped by transitional nucleotide changes at the 1st or 2nd codons, with A/G changes being the most frequent for all six trios. Even though transitions can change amino acid properties less radically than tranversions (Zhang 2000), there is a still a great possibility for them to be removed quickly, particularly in the highly conserved regions as in the cases of H+VL type 1 SNAAPs, which may require a special mechanism to be fixed more efficiently. Consistent with this idea, my analysis further demonstrated that such bias cannot be explained by the mutational mechanism or the general natural selection on nuclear genes and thus implied a special mechanism. Here I suggest the RNA editing could be an important mechanism.

RNA editing refers to the naturally occurring process of modifying the RNA molecules to alter their information content (Bass 2002). It involves two common forms: C-U editing and A-I editing. Particularly, the A-I editing can be site specific and the enzymes responsible for A-to-I editing, the adenosine deaminases acting on RNA (ADARs), are ubiquitously expressed in mammals (Bass 2002). Since inosine (I) is read as guanosine (G) by the translation machinery, A-to-I editing often leads to single amino

acid changes equivalent to an A/G nucleotide mutation, as observed in several neurotransmitter receptors(Seeburg and Hartner 2003; Valente and Nishikura 2005). The deficiency or misregulation of A-to-I RNA editing has been implicated in the etiology of neurological diseases, such as epilepsy, amyotrophic lateral sclerosis (ALS), and depression in mammals (Mehler and Mattick 2007). The loss of A-to-I editing upon the genetic inactivation of ADARs in mammals can result in neurological dysfunctions, too (Higuchi et al. 2000; Wang et al. 2000). They also tend to occur at the 1st and 2nd codons in highly conserved coding regions, introduce certain substitutions more frequently like Q/R, and has been proposed to sometimes become fixed in a genome (Bass 2002). I suggest that some extra diversity at transcriptomic level created by RNA editing might have occurred before the SNAAP substitutions, which may bring certain advantages, thus relaxed the negative selection at the site, and eventually makes it easier for the subsequent fixation of a novel genetic mutation.

## ***6.7 Future directions***

### **6.7.1 Extensions of the computational approach**

Our approach consists of the following four key steps: (1) selection of species and identification of orthologous transcripts; (2) selection of sites with homoplasy; (3) phylogenetic screening to remove substitutions due to random changes; and (4) reconstruction of the ancestral amino acid type. For each of them, there is room to generalize for identifying sites associated with convergent evolution among any

distantly related species even with unknown shared traits. Further, such a hypothesis-free approach would not be limited by the dichotomic classification between vocal learners and non-learners, and may even provide a more quantitative classification.

In future directions, for step 1, more species with higher coverage genome sequences, and more description of vocal learning phenotypes among vertebrates will be needed. This work is currently being done with the advent of next-generation sequencing.

For step 2, as one of the goals in this study is to compare with the positive selection methods, it restricted us to the analyses of predicted coding sequences (CDS). One caveat is that the CDS predictions for most species' genomes are not as complete or accurate as those well annotated ones, e.g. human and mouse genomes, resulting in missing exons or too short sequences in certain species which may prevent the whole gene from the subsequent analyses. In the future, I suggest to use exon or the whole genome alignments to maximize the use of available data. The identification of amino acid translations from these alignments can be guided by using the corresponding protein sequences from one or a few of the best annotated genomes. Also for step 2, my current specifications call for initial selection of homoplastic sites using a small initial selection of six species to define the four types of patterns. In the future, for a hypothesis-free method, I suggest to define the patterns by sampling many more species, particularly with high coverage genomes. In other words, one can pre-specify N

selections of species, and for each of these N selection of species, define M types of patterns that requires homoplasy among trios, quads or more of species combinations. The relationship among the N selections of species can be modeled and help interpret or even merge the resultant SNAAPs.

For step 3, the current phylogenetic filtering step relied on pre-specified assumptions or parameters on the evolution of substitutions. In the future, more complex evolutionary models can be introduced and the parameters can be estimated from the data to obtain a maximum likelihood estimate or a Bayesian estimate to evaluate the changes against random chance. This could improve the sensitivity and specificity of the method.

For step 4, currently the ancestral reconstruction step is independent from step 3. In the future, I suggest combining the two steps together, i.e. estimating the tree likelihood and the ancestral state of amino acid types simultaneously, where the same evolutionary model will be used.

### **6.7.2 Exploration on more domesticated species**

Our results suggest an excess of shared homoplastic amino acid replacements in human, cow and dog. However, for making a general conclusion for domestication, more domesticated species need to be analyzed and should not be restricted to using only trios of species. Moreover, species with longer domestication history, e.g. >1000 years, should be distinguished from those with a more recent history, and livestock

should be distinguished from companion animals. I hypothesize that we will see significantly more sharing of homoplastic amino acids among the domesticated species with humans, the longer they have been domesticated by humans. I also expect a relaxed rule of the transitional bias in nucleotide changes for their homoplasy. It will be of interest to explore whether or not there is substantially more sharing among domesticated species for similar selected traits, independent of the length of their domestication history. The results will provide critical insights into distinguishing the two common consequences from domestication: the deleterious mutations due to relaxed selective constraints and the positively selected sites for the desirable trait for humans.

### **6.7.3 Experimental validation of genes with H+VL type 1 SNAAPs**

I hypothesize that many of the genes with H+VL type 1 SNAAPs will exhibit different brain expression patterns between humans and non-human primates, e.g. those that may play a role in the central auditory pathway or regulate neural development. Our hypothesis is based on the findings thus far with FOXP2 and ROBO1, as well as known correlation of coding mutations and differential gene regulation (Zakon et al. 2006). Further, their expression patterns are worth exploring during developmental regulation in multiple songbird species whose sensitive periods are of different durations and timing, to establish a precise temporal and spatial correlation with vocal plasticity.

I further hypothesize that that ROBO1 and SLIT1 may be critical for formation of direct forebrain projections to vocal motor neurons, and that this projection can evolved either through the SNAAP change or by expression level changes of ROBO1 and its ligand SLIT1. To test this hypothesis, future experiments would need to manipulate the SNAAP site and manipulate expression in vocal learning and vocal nonlearning species, and assess neural connectivity of vocal brain regions. I expect abrupt changes in expression levels of the ROBO1 gene or its critical interacting partners (e.g. SLIT1) will cause abnormal connections between RA and XIIIts in vocal learning birds. To do so, one can either directly over-express the gene(s) in the appropriate vocal nucleus or surrounding areas, or knock-down expression by RNAi. Alternatively, one can isolate and culture the RA projecting neurons in vitro at certain stages (e.g. D20, D35 or D65), and see if the outgrowth of axons will be influenced by changes of its ligand(s) in the target.

Lastly, I hypothesize that the genetic variations on these genes or in their adjacent loci could be associated with speech deficits in humans. This idea can be tested by generating a list of the known common SNPs within these regions and sequencing them in the patients with speech deficits to see if any significant non-random patterns can be found, including of the SNAAPs I found. In summary, my computational analyses and experiments provided some reasonable candidate genes to test for their

role in evolving and controlling unique features of song/spoken-language and the underlying neural pathways.

## References

- Alroy, J. (1998). "Cope's Rule and the Dynamics of Body Mass Evolution in North American Fossil Mammals." Science **280**(5364): 731-734.
- Alves-Gomes, J. A. (1999). "Systematic biology of gymnotiform and mormyriiform electric fishes: Phylogenetic relationships, molecular clocks and rates of evolution in the mitochondrial rRNA genes." Journal of Experimental Biology **202**(10): 1167-1183.
- Andrews, W., A. Liapi, et al. (2006). "Robo1 regulates the development of major axon tracts and interneuron migration in the forebrain." Development **133**(11): 2243-2252.
- Anholt, R. R. H., R. F. Lyman, et al. (1996). "Effects of single P-element insertions on olfactory behavior in *Drosophila melanogaster*." Genetics **143**(1): 293-301.
- Apic, G., J. Gough, et al. (2001). "Domain combinations in archaeal, eubacterial and eukaryotic proteomes." Journal of Molecular Biology **310**(2): 311-325.
- Arbiza, L., J. Dopazo, et al. (2006). "Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome." Plos Computational Biology **2**(4): 288-300.
- Ba-Charvet, K. T. N., K. Brose, et al. (2001). "Diversity and Specificity of Actions of Slit2 Proteolytic Fragments in Axon Guidance." J. Neurosci. **21**(12): 4281-4289.
- Bagri, A., O. MarIn, et al. (2002). "Slit Proteins Prevent Midline Crossing and Determine the Dorsovenral Position of Major Axonal Pathways in the Mammalian Forebrain." Neuron **33**(2): 233-248.
- Baptista, L. F. and K. L. Schuchmann (1990). "SONG LEARNING IN THE ANNA HUMMINGBIRD (CALYPTE-ANNA)." Ethology **84**(1): 15-26.
- Barnes, M. R. (2007). Bioinformatics for geneticists : a bioinformatics primer for the analysis of genetic data. Chichester [u.a.], Wiley.

- Bashaw, G. J. and C. S. Goodman (1999). "Chimeric Axon Guidance Receptors: The Cytoplasmic Domains of Slit and Netrin Receptors Specify Attraction versus Repulsion." Cell **97**(7): 917-926.
- Bass, B. L. (2002). "RNA EDITING BY ADENOSINE DEAMINASES THAT ACT ON RNA." Annual Review of Biochemistry **71**(1): 817-846.
- Battisti, A. C. and D. M. Fekete (2008). "Slits and robos in the developing chicken inner ear." Developmental Dynamics **237**(2): 476-484.
- Battye, R., A. Stevens, et al. (2001). "Repellent Signaling by Slit Requires the Leucine-Rich Repeats." J. Neurosci. **21**(12): 4290-4298.
- Belle, E. M. S., G. Piganeau, et al. (2005). "An investigation of the variation in the transition bias among various animal mitochondrial DNA." Gene **355**: 58-66.
- Benjamin, A., M. Kashem, et al. (2008). "Proteomics of the Nucleus Ovoidalis and Field L Brain Regions of Zebra Finch." Journal of Proteome Research **7**(5): 2121-2132.
- Bianchi, L. M. and N. A. Gray (2002). "EphB receptors influence growth of ephrin-B1-positive statoacoustic nerve fibers." European Journal of Neuroscience **16**(8): 1499-1506.
- Bilder, D. and N. Perrimon (2000). "Localization of apical epithelial determinants by the basolateral PDZ protein Scribble." Nature **403**(6770): 676-680.
- Björnerfeldt, S., M. T. Webster, et al. (2006). "Relaxation of selective constraint on dog mitochondrial DNA following domestication." Genome Research **16**(8): 990-994.
- Boughman, J. W. (1998). "Vocal learning by greater spear-nosed bats." Proceedings of the Royal Society of London. Series B: Biological Sciences **265**(1392): 227-233.
- Boyko, A. R., S. H. Williamson, et al. (2008). "Assessing the evolutionary impact of amino acid mutations in the human genome." Plos Genetics **4**(5).
- Brainard, M. S. and A. J. Doupe (2000). "Auditory feedback in learning and maintenance of vocal behaviour." Nat Rev Neurosci **1**(1): 31-40.
- Brune, M. (2007). "On human self-domestication, psychiatry, and eugenics." Philosophy, Ethics, and Humanities in Medicine **2**(1): 21.

- Bush, E. and B. Lahn (2005). "Selective Constraint on Noncoding Regions of Hominid Genomes." PLoS Computational Biology **1**(7): e73.
- Bush, E. and B. Lahn (2008). "A genome-wide screen for noncoding elements important in primate evolution." BMC Evolutionary Biology **8**(1): 17.
- Cáceres, M., J. Lachuer, et al. (2003). "Elevated gene expression levels distinguish human from non-human primate brains." Proceedings of the National Academy of Sciences **100**(22): 13030-13035.
- Cantalupo, C. and W. D. Hopkins (2001). "Asymmetric Broca's area in great apes." Nature **414**(6863): 505-505.
- Carlborg, O. and C. S. Haley (2004). "Epistasis: too often neglected in complex trait studies?" Nature Reviews Genetics **5**(8): 618-U614.
- Carroll, S. B. (2005). "Evolution at Two Levels: On Genes and Form." PLoS Biol **3**(7): e245.
- Charlier, C., W. Coppieters, et al. (2008). "Highly effective SNP-based association mapping and management of recessive defects in livestock." Nature Genetics **40**(4): 449-454.
- Chen, J.-h., L. Wen, et al. (2001). "The N-terminal Leucine-Rich Regions in Slit Are Sufficient To Repel Olfactory Bulb Axons and Subventricular Zone Neurons." J. Neurosci. **21**(5): 1548-1556.
- Christin, P.-A., N. Salamin, et al. (2008). "Evolutionary Switch and Genetic Convergence on rbcL following the Evolution of C4 Photosynthesis." Molecular Biology and Evolution **25**(11): 2361-2368.
- Clark, A. G., S. Glanowski, et al. (2003). "Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios." Science **302**(5652): 1960-1963.
- Clark, K., E. Hammond, et al. (2002). "Temporal and spatial expression of two isoforms of the Dutt1/Robo1 gene in mouse development." FEBS Letters **523**(1-3): 12-16.
- Crill, W. D., H. A. Wichman, et al. (2000). "Evolutionary Reversals During Viral Adaptation to Alternating Hosts." Genetics **154**(1): 27-37.

- Cruz, F., C. Vila, et al. (2008). "The Legacy of Domestication: Accumulation of Deleterious Mutations in the Dog Genome." Molecular Biology and Evolution **25**(11): 2331-2336.
- Dalva, M. B., A. C. McClelland, et al. (2007). "Cell adhesion molecules: signalling functions at the synapse." Nat Rev Neurosci **8**(3): 206-220.
- Damuth, J. (1981). "Population density and body size in mammals." Nature **290**(5808): 699-700.
- DePristo, M. A., D. L. Hartl, et al. (2007). "Mutational Reversions During Adaptive Protein Evolution." Molecular Biology and Evolution **24**(8): 1608-1610.
- Devine, C. A. and B. Key (2008). "Robo-Slit interactions regulate longitudinal axon pathfinding in the embryonic vertebrate brain." Developmental Biology **313**(1): 371-383.
- DeVoogd, T. J., D. J. Pyskaty, et al. (1991). "Lateral asymmetries and testosterone-induced changes in the gross morphology of the hypoglossal nucleus in adult canaries." The Journal of Comparative Neurology **307**(1): 65-76.
- Dugan, J. P., A. Stratton, et al. (2011). "Midbrain dopaminergic axons are guided longitudinally through the diencephalon by Slit/Robo signals." Molecular and Cellular Neuroscience **46**(1): 347-356.
- Durand, S. E., J. T. Heaton, et al. (1997). "Vocal control pathways through the anterior forebrain of a parrot (*Melopsittacus undulatus*)." The Journal of Comparative Neurology **377**(2): 179-206.
- Easton, D., D. Ford, et al. (1993). "INHERITED SUSCEPTIBILITY TO BREAST-CANCER." Cancer Surveys **18**: 95-113.
- Edwards, R. and D. Shields (2004). "GASP: Gapped Ancestral Sequence Prediction for proteins." BMC Bioinformatics **5**(1): 123.
- Egnor, S. E. R. and M. D. Hauser (2004). "A paradox in the evolution of primate vocal learning." Trends in Neurosciences **27**(11): 649-654.
- Elowson, A. M., C. T. Snowdon, et al. (1992). "Ontogeny of trill and J-call vocalizations in the pygmy marmoset, *Cebuella pygmaea*." Animal Behaviour **43**(5): 703-715.

- Enard, W., M. Przeworski, et al. (2002). "Molecular evolution of FOXP2, a gene involved in speech and language." Nature **418**(6900): 869-872.
- Evans, P. D., S. L. Gilbert, et al. (2005). "Microcephalin, a Gene Regulating Brain Size, Continues to Evolve Adaptively in Humans." Science **309**(5741): 1717-1720.
- Eyre-Walker, A. and P. D. Keightley (1999). "High genomic deleterious mutation rates in hominids." Nature **397**(6717): 344-347.
- Farabaugh, S. M., A. Linzenbold, et al. (1994). "VOCAL PLASTICITY IN BUDGERIGARS (MELOPSITTACUS-UNDULATUS) - EVIDENCE FOR SOCIAL-FACTORS IN THE LEARNING OF CONTACT CALLS." Journal of Comparative Psychology **108**(1): 81-92.
- Farmer, W. T., A. L. Altick, et al. (2008). "Pioneer longitudinal axons navigate using floor plate and Slit/Robo signals." Development **135**(22): 3643-3653.
- Feenders, G., M. Liedvogel, et al. (2008). "Molecular Mapping of Movement-Associated Areas in the Avian Brain: A Motor Theory for Vocal Learning Origin." Plos One **3**(3).
- Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." Journal of Molecular Evolution **17**(6): 9.
- Ferrari, P. F., E. Visalberghi, et al. (2006). "Neonatal Imitation in Rhesus Macaques." PLoS Biol **4**(9): e302.
- Finlayson, C. (2005). "Biogeography and evolution of the genus Homo." Trends in Ecology & Evolution **20**(8): 457-463.
- Fisher, S. E. and C. Scharff (2009). "FOXP2 as a molecular window into speech and language." Trends in genetics : TIG **25**(4): 166-177.
- Fitch, W. T. (2000). "The evolution of speech: a comparative review." Trends in Cognitive Sciences **4**(7): 258-267.
- Foote, A. D., R. M. Griffin, et al. (2006). "Killer whales are capable of vocal learning." Biology Letters **2**(4): 509-512.

- Fritz, J., M. Mishkin, et al. (2005). "In search of an auditory engram." Proceedings of the National Academy of Sciences of the United States of America **102**(26): 9359-9364.
- Gahr, M. (2000). "Neural song control system of hummingbirds: Comparison to swifts, vocal learning (Songbirds) and nonlearning (Suboscines) passerines, and vocal learning (Budgerigars) and nonlearning (Dove, owl, gull, quail, chicken) nonpasserines." The Journal of Comparative Neurology **426**(2): 182-196.
- Ganguly, I., T. F. C. Mackay, et al. (2003). "Scribble is essential for olfactory Behavior in *Drosophila melanogaster*." Genetics **164**(4): 1447-1457.
- Ghazanfar, A. A. and D. Rendall (2008). "Evolution of human vocal production." Current Biology **18**(11): R457-R460.
- Giangrande, P. H., T. C. Hallstrom, et al. (2003). "Identification of E-Box Factor TFE3 as a Functional Partner for the E2F3 Transcription Factor." Mol. Cell. Biol. **23**(11): 3707-3720.
- Glazier, A. M., J. H. Nadeau, et al. (2002). "Finding genes that underlie complex traits." Science **298**(5602): 2345-2349.
- Goodstadt, L. and C. P. Ponting (2006). "Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human." Plos Computational Biology **2**(9): 1134-1150.
- Gouzoules, H. and S. Gouzoules (1990). "Matrilineal Signatures in the Recruitment Screams of Pigtail Macaques, *Macaca Nemestrina*." Behaviour **115**: 327-347.
- Goycoolea, M., I. Mena, et al. (2006). "Spontaneous musical auditory perceptions in patients who develop abrupt bilateral sensorineural hearing loss. An uninhibition syndrome?" Acta Oto-laryngologica **126**(4): 368-374.
- Grabner, C. P., S. D. Price, et al. (2006). "Regulation of large dense-core vesicle volume and neurotransmitter content mediated by adaptor protein 3." Proceedings of the National Academy of Sciences **103**(26): 10035-10040.
- Gramza, A. F. (1970). "VOCAL MIMICRY IN CAPTIVE BUDGERIGARS MELOPSITTACUS-UNDULATUS." Zeitschrift fuer Tierpsychologie **27**(8): 971-983.

- Green, R. E., J. Krause, et al. (2010). A Draft Sequence of the Neandertal Genome. **328**: 710-722.
- Gridi-Papp, M. and P. M. Narins (2009). "Environmental influences in the evolution of tetrapod hearing sensitivity and middle ear tuning." Integrative and Comparative Biology **49**(6): 702-716.
- Griffiths, A., S. Wessler, et al. (2007). Introduction to Genetic Analysis (Introduction to Genetic Analysis (Griffiths)), W. H. Freeman.
- Grimsley, J. M. S., J. J. M. Monaghan, et al. (2011). "Development of Social Vocalizations in Mice." PLoS ONE **6**(3): e17460.
- Gurney, M. E. and M. Konishi (1980). "Hormone-Induced Sexual Differentiation of Brain and Behavior in Zebra Finches." Science **208**(4450): 1380-1383.
- Hackett, S. J., R. T. Kimball, et al. (2008). "A Phylogenomic Study of Birds Reveals Their Evolutionary History." Science **320**(5884): 1763-1768.
- Hackett, S. J., R. T. Kimball, et al. (2008). A Phylogenomic Study of Birds Reveals Their Evolutionary History. **320**: 1763-1768.
- Haesler, S., C. Rochefort, et al. (2007). "Incomplete and Inaccurate Vocal Imitation after Knockdown of *FoxP2* in Songbird Basal Ganglia Nucleus Area X." PLoS Biol **5**(12): e321.
- Haesler, S., K. Wada, et al. (2004). "FoxP2 expression in avian vocal learners and non-learners." Journal of Neuroscience **24**(13): 3164-3175.
- Hammerschmidt, K., T. Freudenstein, et al. (2001). "VOCAL DEVELOPMENT IN SQUIRREL MONKEYS." Behaviour **138**: 1179-1204.
- Hannula-Jouppi, K., N. Kaminen-Ahola, et al. (2005). "The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia." Plos Genetics **1**(4): 467-474.
- Haygood, R., O. Fedrigo, et al. (2007). "Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution." Nat Genet **39**(9): 1140-1144.

- Heimbauer, Lisa A., Michael J. Beran, et al. (2011). "A Chimpanzee Recognizes Synthetic Speech with Significantly Reduced Acoustic Cues to Phonetic Content." Current Biology In Press, Corrected Proof.
- Henikoff, S. and J. G. Henikoff (2000). Amino acid substitution matrices. Advances in Protein Chemistry. B. Peer, Academic Press. **Volume 54**: 73-97.
- Hernández-Miranda, L. R., A. Cariboni, et al. (2011). "Robo1 Regulates Semaphorin Signaling to Guide the Migration of Cortical Interneurons through the Ventral Forebrain." The Journal of Neuroscience **31**(16): 6174-6187.
- Higginbotham, H., T. Tanaka, et al. (2006). "GSK3 beta and PKC zeta function in centrosome localization and process stabilization during Slit-mediated neuronal repolarization." Molecular and Cellular Neuroscience **32**(1-2): 118-132.
- Higuchi, M., S. Maas, et al. (2000). "Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2." Nature **406**(6791): 78-81.
- Hillier, L. W., W. Miller, et al. (2004). "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution." Nature **432**(7018): 695-716.
- Hoekstra, H. E. (2006). "Genetics, development and evolution of adaptive pigmentation in vertebrates." Heredity **97**(3): 222-234.
- Holland, S. J., E. Peles, et al. (1998). "Cell-contact-dependent signalling in axon growth and guidance: Eph receptor tyrosine kinases and receptor protein tyrosine phosphatase [beta]." Current Opinion in Neurobiology **8**(1): 117-127.
- Holy, T. E. and Z. Guo (2005). "Ultrasonic Songs of Male Mice." PLoS Biol **3**(12): e386.
- Huff, C. D., J. Xing, et al. (2010). "Mobile elements reveal small population size in the ancient ancestors of Homo sapiens." Proceedings of the National Academy of Sciences **107**(5): 2147-2152.
- Hyaneek, J. and V. Raisova (1985). "Speech and Language Disorders in Histidinemia and Other Amino-Acid Disturbances." Journal of Inherited Metabolic Disease **8**: 130-130.

- Ishikawa, S., I. Kobayashi, et al. (2001). "Interaction of MCC2, a novel homologue of MCC tumor suppressor, with PDZ-domain Protein AIE-75." Gene **267**(1): 101-110.
- Iwatsubo, T., S. Kuzuhara, et al. (1990). "CORTICOFUGAL PROJECTIONS TO THE MOTOR NUCLEI OF THE BRAIN-STEM AND SPINAL-CORD IN HUMANS." Neurology **40**(2): 309-312.
- Jürgens, U. (2002). "Neural pathways underlying vocal control." Neuroscience and Biobehavioral Reviews **26**(2): 235-258.
- Jürgens, U. (2009). "The Neural Control of Vocalization in Mammals: A Review." Journal of Voice **23**(1): 1-10.
- Jürgens, U. and D. Ploog (1981). "On the neural control of mammalian vocalization." Trends in Neurosciences **4**: 135-137.
- Janik, V. M. and P. J. B. Slater (1997). "Vocal learning in mammals." Advances in the Study of Behavior, Vol 26 **26**: 59-99.
- Jarvis, E. D. (2004). Learned birdsong and the neurobiology of human language. Behavioral Neurobiology of Birdsong. H. P. Zeigler and P. Marler. **1016**: 749-777.
- Jarvis, E. D., S. Ribeiro, et al. (2000). "Behaviourally driven gene expression reveals song nuclei in hummingbird brain." Nature **406**(6796): 628-632.
- Johnson, F. and M. Sellix (2000). "Reorganization of a telencephalic motor region during sexual differentiation and vocal learning in zebra finches." Developmental Brain Research **121**(2): 253-263.
- Johnson, M. B., Y. I. Kawasawa, et al. (2009). "Functional and Evolutionary Insights into Human Brain Development through Global Transcriptome Analysis." Neuron **62**(4): 494-509.
- Jorgensen, F., A. Hobolth, et al. (2005). "Comparative analysis of protein coding sequences from human, mouse and the domesticated pig." BMC Biology **3**(1): 2.
- Joron, M., R. Papa, et al. (2006). "A conserved supergene locus controls colour pattern diversity in Heliconius butterflies." Plos Biology **4**(10): 1831-1840.

- Jung, M.-Y., L. Lorenz, et al. (2006). "Translational Control by Neuroguidin, a Eukaryotic Initiation Factor 4E and CPEB Binding Protein." Mol. Cell. Biol. **26**(11): 4277-4287.
- Kehrer-Sawatzki, H., C. Sandig, et al. (2005). "Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (*Pan troglodytes*)." Human Mutation **25**(1): 45-55.
- Kidd, T., K. Brose, et al. (1998). "Roundabout Controls Axon Crossing of the CNS Midline and Defines a Novel Subfamily of Evolutionarily Conserved Guidance Receptors." Cell **92**(2): 205-215.
- Kikusui, T., K. Nakanishi, et al. (2011). "Cross Fostering Experiments Suggest That Mice Songs Are Innate." PLoS ONE **6**(3): e17721.
- Konishi, M. (1965). "The role of auditory feedback in the control of vocalization in the white-crowned sparrow." Z. Tierpsychol. **22**: 770-783.
- Konopka, G., J. M. Bomar, et al. (2009). "Human-specific transcriptional regulation of CNS development genes by FOXP2." Nature **462**(7270): 213-217.
- Koonin, E. V., Y. I. Wolf, et al. (2002). "The structure of the protein universe and genome evolution." Nature **420**(6912): 218-223.
- Krause, J., C. Lalueza-Fox, et al. (2007). "The derived FOXP2 variant of modern humans was shared with neandertals." Current Biology **17**(21): 1908-1912.
- Kroodsma, D. E. and M. Konishi (1991). "A SUBOSCINE BIRD (EASTERN PHOEBE, SAYORNIS-PHOEBE) DEVELOPS NORMAL SONG WITHOUT AUDITORY-FEEDBACK." Animal Behaviour **42**: 477-487.
- Kubisch, C., B. C. Schroeder, et al. (1999). "KCNQ4, a Novel Potassium Channel Expressed in Sensory Outer Hair Cells, Is Mutated in Dominant Deafness." Cell **96**(3): 437-446.
- Kumar, S. and S. Subramanian (2002). "Mutation rates in mammalian genomes." Proceedings of the National Academy of Sciences of the United States of America **99**(2): 803-808.

- Kuypers, H. G. J. (1958). "CORTICOBULBAR CONNEXIONS TO THE PONS AND LOWER BRAIN-STEM IN MAN - AN ANATOMICAL STUDY." Brain **81**(3): 364- &.
- Kuypers, H. G. J. M. (1958). "Some projections from the peri-central cortex to the pons and lower brain stem in monkey and chimpanzee." The Journal of Comparative Neurology **110**(2): 221-255.
- López-Bendito, G., N. Flames, et al. (2007). "Robo1 and Robo2 Cooperate to Control the Guidance of Major Axonal Tracts in the Mammalian Forebrain." The Journal of Neuroscience **27**(13): 3395-3407.
- Lai, C. S. L., S. E. Fisher, et al. (2001). "A forkhead-domain gene is mutated in a severe speech and language disorder." Nature **413**(6855): 519-523.
- Lai, C. S. L., D. Gerrelli, et al. (2003). "FOXP2 expression during brain development coincides with adult sites of pathology in a severe speech and language disorder." Brain **126**(11): 2455-2462.
- Lander, E. S. and N. J. Schork (2006). "Genetic Dissection of Complex Traits." Focus **4**(3): 442-458.
- Lartillot, N., T. Lepage, et al. (2009). "PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating." Bioinformatics **25**(17): 2286-2288.
- Lewis, B. A., L. D. Shriberg, et al. (2006). "The Genetic Bases of Speech Sound Disorders: Evidence From Spoken and Written Language." J Speech Lang Hear Res **49**(6): 1294-1312.
- Li, X., X.-J. Wang, et al. (2007). "Genomic resources for songbird research and their use in characterizing gene expression during brain development." Proceedings of the National Academy of Sciences **104**(16): 6834-6839.
- Liberman, A. M. and D. H. Whalen (2000). "On the relation of speech to language." Trends in Cognitive Sciences **4**(5): 187-196.
- Lieblich, A. K., D. Symmes, et al. (1980). "Development of the Isolation Peep in laboratory-bred squirrel monkeys." Animal Behaviour **28**(1): 1-9.

- Lilly, J. C. (1965). "Vocal Mimicry in Tursiops: Ability to Match Numbers and Durations of Human Vocal Bursts." Science **147**(3655): 300-301.
- Lim, A. and R. Kraut (2009). "The Drosophila BEACH Family Protein, Blue Cheese, Links Lysosomal Axon Transport with Motor Neuron Degeneration." Journal of Neuroscience **29**(4): 951-963.
- Liu, Y., J. A. Cotton, et al. (2010). "Convergent sequence evolution between echolocating bats and dolphins." Current Biology **20**(2): R53-R54.
- Lovell, P. V., D. F. Clayton, et al. (2008). "Birdsong "Transcriptomics": Neurochemical Specializations of the Oscine Song System." PLoS ONE **3**(10): e3440.
- Lyman, R. F., F. Lawrence, et al. (1996). "Effects of single P-element insertions on bristle number and viability in *Drosophila melanogaster*." Genetics **143**(1): 277-292.
- MacEachern, S., J. McEwan, et al. (2009). "Molecular evolution of the Bovini tribe (Bovidae, Bovinae): Is there evidence of rapid evolution or reduced selective constraint in Domestic cattle?" BMC Genomics **10**(1): 179.
- Mackay, R. S. and H. M. Liaw (1981). "Dolphin Vocalization Mechanisms." Science **212**(4495): 676-678.
- Mackay, T. F. C. and R. R. H. Anholt (2007). "Ain't misbehavin? Genotype-environment interactions and the genetics of behavior." Trends in Genetics **23**(7): 311-314.
- Maclarnon, A. and G. Hewitt (2004). "Increased breathing control: Another factor in the evolution of human language." Evolutionary Anthropology: Issues, News, and Reviews **13**(5): 181-197.
- Mantela, J., Z. Jiang, et al. (2005). "The retinoblastoma gene pathway regulates the postmitotic state of hair cells of the mouse inner ear." Development **132**(10): 2377-2388.
- Marillat, V., O. Cases, et al. (2002). "Spatiotemporal expression patterns of slit and robo genes in the rat brain." The Journal of Comparative Neurology **442**(2): 130-155.
- Marshall, A. J., R. W. Wrangham, et al. (1999). "Does learning affect the structure of vocalizations in chimpanzees?" Animal Behaviour **58**: 825-830.

- Martínez, I., M. Rosa, et al. (2004). "Auditory capacities in Middle Pleistocene humans from the Sierra de Atapuerca in Spain." Proceedings of the National Academy of Sciences of the United States of America **101**(27): 9976-9981.
- Matsunaga, E. and K. Okanoya (2008). "Expression analysis of cadherins in the songbird brain: Relationship to vocal system development." The Journal of Comparative Neurology **508**(2): 329-342.
- Matsunaga, E. and K. Okanoya (2009). "Vocal control area-related expression of neuropilin-1, plexin-A4, and the ligand semaphorin-3A has implications for the evolution of the avian vocal system." Development, Growth & Differentiation **51**(1): 45-54.
- McGregor, L., V. Makela, et al. (2003). "Fraser syndrome and mouse blebbed phenotype caused by mutations in FRAS1/Fras1 encoding a putative extracellular matrix protein." Nature Genetics **34**(2): 203-208.
- Mehler, M. F. and J. S. Mattick (2007). "Noncoding RNAs and RNA Editing in Brain Development, Functional Diversification, and Neurological Disease." Physiological Reviews **87**(3): 799-823.
- Migani, P., C. Bartlett, et al. (2009). "Regional and cellular distribution of ephrin-B1 in adult mouse brain." Brain Research **1247**: 50-61.
- Mitani, J. C. and J. Gros-Louis (1998). "Chorusing and call convergence in chimpanzees: Tests of three hypotheses." Behaviour **135**: 1041-1064.
- Morishita, H., T. Makishima, et al. (2001). "Deafness Due to Degeneration of Cochlear Neurons in Caspase-3-Deficient Mice." Biochemical and Biophysical Research Communications **284**(1): 142-149.
- Murphy, W. J., E. Eizirik, et al. (2001). "Molecular phylogenetics and the origins of placental mammals." Nature **409**(6820): 614-618.
- Murphy, W. J., E. Eizirik, et al. (2001). Molecular phylogenetics and the origins of placental mammals. **409**: 614 - 618.
- Murphy, W. J., T. H. Pringle, et al. (2007). Using genomic data to unravel the root of the placental mammal phylogeny. **17**: 413-421.

- Nabholz, B., A. Künstner, et al. (2011). "Dynamic evolution of base composition: causes and consequences in avian phylogenomics." Molecular Biology and Evolution.
- Naurin, S., B. Hansson, et al. (2011). "The sex-biased brain: sexual dimorphism in gene expression in two species of songbirds." BMC Genomics **12**(1): 37.
- Nei, M. (2005). "Selectionism and Neutralism in Molecular Evolution." Molecular Biology and Evolution **22**(12): 2318-2342.
- Newbury, D., S. Fisher, et al. (2010). "Recent advances in the genetics of language impairment." Genome Medicine **2**(1): 6.
- Newson, R. and A. S. T. The (2003). "Multiple-test procedures and smile plots." Stata Journal **3**(2): 109-132.
- Nickel, G. C., D. L. Tefft, et al. (2008). "An Empirical Test for Branch-Specific Positive Selection." Genetics **179**(4): 2183-2193.
- Nielsen, R. (2001). "Statistical tests of selective neutrality in the age of genomics." Heredity **86**(6): 641-647.
- Nogales-Cadenas, R., P. Carmona-Saez, et al. (2009). "GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information." Nucleic Acids Research **37**(suppl 2): W317-W322.
- Norga, K. K., M. C. Gurganus, et al. (2003). "Quantitative analysis of bristle number in *Drosophila* mutants identifies genes involved in neural development." Current Biology **13**(16): 1388-1397.
- Nottebohm, F. (1972). "ORIGINS OF VOCAL LEARNING." American Naturalist **106**(947): 116-&.
- Nottebohm, F., T. M. Stokes, et al. (1976). "CENTRAL CONTROL OF SONG IN CANARY, *SERINUS-CANARIUS*." Journal of Comparative Neurology **165**(4): 457-486.
- Nozawa, M., Y. Suzuki, et al. (2009). "Reliabilities of identifying positive selection by the branch-site and the site-prediction methods." Proceedings of the National Academy of Sciences.

- Nural, H. F., W. Todd Farmer, et al. (2007). "The Slit receptor Robo1 is predominantly expressed via the Dutt1 alternative promoter in pioneer neurons in the embryonic mouse brain and spinal cord." Gene Expression Patterns 7(8): 837-845.
- Ohno, K., A. G. Engel, et al. (2002). "Rapsyn Mutations in Humans Cause Endplate Acetylcholine-Receptor Deficiency and Myasthenic Syndrome." The American Journal of Human Genetics 70(4): 875-885.
- Oldham, M. C., S. Horvath, et al. (2006). "Conservation and evolution of gene coexpression networks in human and chimpanzee brains." Proceedings of the National Academy of Sciences 103(47): 17973-17978.
- Oldham, M. C., G. Konopka, et al. (2008). "Functional organization of the transcriptome in human brain." Nat Neurosci 11(11): 1271-1282.
- Pepperberg, I. M. (1981). "FUNCTIONAL VOCALIZATIONS BY AN AFRICAN GREY PARROT (PSITTACUS-ERITHACUS)." Zeitschrift Fur Tierpsychologie-Journal of Comparative Ethology 55(2): 139-160.
- Pepperberg, I. M. (2006). "Cognitive and communicative abilities of Grey parrots." Applied Animal Behaviour Science 100(1-2): 77-86.
- Petrides, M., G. Cadoret, et al. (2005). "Orofacial somatomotor responses in the macaque monkey homologue of Broca's area." Nature 435(7046): 1235-1238.
- Pinaud, R., C. Osorio, et al. (2008). "Profiling of experience-regulated proteins in the songbird auditory forebrain using quantitative proteomics." European Journal of Neuroscience 27(6): 1409-1422.
- Poole, J. H., P. L. Tyack, et al. (2005). "Elephants are capable of vocal learning." Nature 434(7032): 455-456.
- Popadin, K., L. V. Polishchuk, et al. (2007). "Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals." Proceedings of the National Academy of Sciences 104(33): 13390-13395.
- Ralls, K., P. Fiorelli, et al. (1985). "VOCALIZATIONS AND VOCAL MIMICRY IN CAPTIVE HARBOR SEALS, PHOCA-VITULINA." Canadian Journal of Zoology-Revue Canadienne De Zoologie 63(5): 1050-1056.

- Reeber, S. L., N. Sakai, et al. (2008). "Manipulating Robo Expression In Vivo Perturbs Commissural Axon Pathfinding in the Chick Spinal Cord." The Journal of Neuroscience **28**(35): 8698-8708.
- Reich, D. E., M. Cargill, et al. (2001). "Linkage disequilibrium in the human genome." Nature **411**(6834): 199-204.
- Reidenberg, J. S. and J. T. Laitman (2008). "Sisters of the Sinuses: Cetacean Air Sacs." The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology **291**(11): 1389-1396.
- Reidenberg, J. S. and J. T. Laitman (2010). Generation of sound in marine mammals. Handbook of Behavioral Neuroscience. M. B. Stefan, Elsevier. **Volume 19**: 451-465.
- Riede, T. and F. Goller (2010). "Peripheral mechanisms for vocal production in birds - differences and similarities to human speech and singing." Brain and Language **115**(1): 69-80.
- Rokas, A. and S. B. Carroll (2008). "Frequent and widespread parallel evolution of protein sequences." Molecular Biology and Evolution **25**(9): 1943-1953.
- Sanvito, S., F. Galimberti, et al. (2007). "Observational evidences of vocal learning in southern elephant seals: A longitudinal study." Ethology **113**(2): 137-146.
- Sawyer, S. A., J. Parsch, et al. (2007). "Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*." Proceedings of the National Academy of Sciences **104**(16): 6504-6510.
- Scharff, C. and S. A. White (2004). "Genetic Components of Vocal Learning." Annals of the New York Academy of Sciences **1016**(1): 325-347.
- Schneider, T. D., G. D. Stormo, et al. (1986). "Information-Content of Binding-Sites on Nucleotide-Sequences." Journal of Molecular Biology **188**(3): 415-431.
- Seeburg, P. H. and J. Hartner (2003). "Regulation of ion channel/neurotransmitter receptor function by RNA editing." Current Opinion in Neurobiology **13**(3): 279-283.

- Seeger, M., G. Tear, et al. (1993). "Mutations affecting growth cone guidance in drosophila: Genes necessary for guidance toward or away from the midline." Neuron **10**(3): 409-426.
- Sevilla, T., A. Cuesta, et al. (2003). "Clinical, electrophysiological and morphological findings of Charcot-Marie-Tooth neuropathy with vocal cord palsy and mutations in the GDAP1 gene." Brain **126**: 2023-2033.
- Seyfarth, R. M. and D. L. Cheney (1986). "Vocal development in vervet monkeys." Animal Behaviour **34**(6): 1640-1658.
- Shapiro, M. D., M. A. Bell, et al. (2006). "Parallel genetic origins of pelvic reduction in vertebrates." Proceedings of the National Academy of Sciences **103**(37): 13753-13758.
- Shapiro, M. D., M. E. Marks, et al. (2004). "Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks." Nature **428**(6984): 717-723.
- Shapiro, M. D., B. R. Summers, et al. (2009). "The Genetic Architecture of Skeletal Convergence and Sex Determination in Ninespine Sticklebacks." Current Biology **19**(13): 1140-1145.
- Shiu, S.-H., W. M. Karlowski, et al. (2004). "Comparative Analysis of the Receptor-Like Kinase Family in Arabidopsis and Rice." The Plant Cell Online **16**(5): 1220-1234.
- Shubin, N., C. Tabin, et al. (2009). "Deep homology and the origins of evolutionary novelty." Nature **457**(7231): 818-823.
- Simes, R. J. (1986). "An improved Bonferroni procedure for multiple tests of significance." Biometrika **73**(3): 751-754.
- Simonin, F., J. Ménessier-de Murcia, et al. (1990). "Expression and site-directed mutagenesis of the catalytic domain of human poly(ADP-ribose)polymerase in Escherichia coli. Lysine 893 is critical for activity." Journal of Biological Chemistry **265**(31): 19249-19256.
- Simonyan, K. and U. Jürgens (2003). "Efferent subcortical projections of the laryngeal motorcortex in the rhesus monkey." Brain Research **974**(1-2): 43-59.

- Snowdon, C. T. and A. M. Elowson (1999). "Pygmy Marmosets Modify Call Structure When Paired." Ethology **105**(10): 893-908.
- Spiteri, E., G. Konopka, et al. (2007). "Identification of the Transcriptional Targets of FOXP2, a Gene Linked to Speech and Language, in Developing Human Brain." The American Journal of Human Genetics **81**(6): 1144-1157.
- Steiper, M. E. and N. M. Young (2006). "Primate molecular divergence dates." Molecular Phylogenetics and Evolution **41**(2): 384-394.
- Subramanian, A. R., M. Kaufmann, et al. (2008). "DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment." Algorithms for Molecular Biology **3**: 11.
- Sugiura, H. (1998). "Matching of acoustic features during the vocal exchange of coo calls by Japanese macaques." Animal Behaviour **55**: 673-687.
- Sun, Q., S. Bahri, et al. (2000). "Receptor tyrosine phosphatases regulate axon guidance across the midline of the Drosophila embryo." Development **127**(4): 801-812.
- Talmage-Riggs, G., P. Winter, et al. (1972). "Effect of Deafening on the Vocal Behavior of the Squirrel Monkey *(Saimiri sciureus)*." Folia Primatologica **17**(5-6): 404-420.
- Tapia-Paez, I., K. Tammimies, et al. (2008). "The complex of TFII-I, PARP1, and SFPQ proteins regulates the DYX1C1 gene implicated in neuronal migration and dyslexia." Faseb Journal **22**(8): 3001-3009.
- Tenesa, A., P. Navarro, et al. (2007). "Recent human effective population size estimated from linkage disequilibrium." Genome Research **17**(4): 520-526.
- Teramitsu, I., A. Poopatanapong, et al. (2010). "Striatal *FoxP2* Is Actively Regulated during Songbird Sensorimotor Learning." PLoS ONE **5**(1): e8548.
- Teramitsu, I. and S. A. White (2006). "FoxP2 regulation during undirected singing in adult songbirds." Journal of Neuroscience **26**(28): 7390-7394.
- Turic, D., L. Robinson, et al. (2003). "Linkage disequilibrium mapping provides further evidence of a gene for reading disability on chromosome 6p21.3-22." Molecular Psychiatry **8**(2): 176-185.

- Twigg, S. R. F., K. Matsumoto, et al. (2006). "The Origin of EFNB1 Mutations in Craniofrontonasal Syndrome: Frequent Somatic Mosaicism and Explanation of the Paucity of Carrier Males." The American Journal of Human Genetics **78**(6): 999-1010.
- Valdar, W., L. C. Solberg, et al. (2006). "Genetic and environmental effects on complex traits in mice." Genetics **174**: 959-984.
- Valente, L. and K. Nishikura (2005). ADAR Gene Family and A-to-I RNA Editing: Diverse Roles in Posttranscriptional Gene Regulation. Progress in Nucleic Acid Research and Molecular Biology. M. Kivie, Academic Press. **Volume 79**: 299-338.
- Van Valkenburgh, B., X. Wang, et al. (2004). "Cope's Rule, Hypercarnivory, and Extinction in North American Canids." Science **306**(5693): 101-104.
- Vargesson, N., V. Luria, et al. (2001). "Expression patterns of Slit and Robo family members during vertebrate limb development." Mechanisms of Development **106**(1-2): 175-180.
- Vernes, S. C., E. Spiteri, et al. (2007). "High-Throughput Analysis of Promoter Occupancy Reveals Direct Neural Targets of FOXP2, a Gene Mutated in Speech and Language Disorders." The American Journal of Human Genetics **81**(6): 1232-1250.
- Wada, K., J. T. Howard, et al. (2006). "A molecular neuroethological approach for identifying and characterizing a cascade of behaviorally regulated genes." Proceedings of the National Academy of Sciences **103**(41): 15212-15217.
- Wang, Q., J. Khillan, et al. (2000). "Requirement of the RNA Editing Deaminase ADAR1 Gene for Embryonic Erythropoiesis." Science **290**(5497): 1765-1768.
- Wang, X., S. D. Thomas, et al. (2004). "Relaxation of selective constraint and loss of function in the evolution of human bitter taste receptor genes." Human Molecular Genetics **13**(21): 2671-2678.
- Warren, W. C., D. F. Clayton, et al. (2010). "The genome of a songbird." Nature **464**(7289): 757-762.

- Webb, D. M. and J. Zhang (2005). "FoxP2 in Song-Learning Birds and Vocal-Learning Mammals." Journal of Heredity **96**(3): 212-216.
- Weinreich, D. M., R. A. Watson, et al. (2005). "PERSPECTIVE: SIGN EPISTASIS AND GENETIC CONSTRAINT ON EVOLUTIONARY TRAJECTORIES." Evolution **59**(6): 1165-1174.
- Wilbrecht, L. and F. Nottebohm (2003). "Vocal learning in birds and humans." Mental Retardation & Developmental Disabilities Research Reviews **9**(3): 135-148.
- Wild, J. M. (1994). "The Auditory-Vocal-Respiratory Axis in Birds." Brain, Behavior and Evolution **44**(4-5): 192-209.
- Winter, P., P. Handley, et al. (1973). "Ontogeny of Squirrel Monkey Calls Under Normal Conditions and Under Acoustic Isolation." Behaviour **47**: 230-239.
- Woolley, S. M. N. and E. W. Rubel (1999). "High-Frequency Auditory Feedback Is Not Required for Adult Song Maintenance in Bengalese Finches." The Journal of Neuroscience **19**(1): 358-371.
- Wray, G. A. (2007). "The evolutionary significance of cis-regulatory mutations." Nat Rev Genet **8**(3): 206-216.
- Wray, G. A. and E. Abouheif (1998). "When is homology not homology?" Current Opinion in Genetics & Development **8**(6): 675-680.
- Xie, Y., Y. Q. Ding, et al. (2005). "Phosphatidylinositol transfer protein-alpha in netrin-1-induced PLC signalling and neurite outgrowth." Nature Cell Biology **7**(11): 1124-1132.
- Yang, Z. (2007). "PAML 4: Phylogenetic Analysis by Maximum Likelihood." Molecular Biology and Evolution **24**(8): 1586-1591.
- Yokoyama, S. and F. B. Radlwimmer (2001). "The Molecular Genetics and Evolution of Red and Green Color Vision in Vertebrates." Genetics **158**(4): 1697-1710.
- Yokoyama, S., T. Tada, et al. (2008). "Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates." Proceedings of the National Academy of Sciences **105**(36): 13480-13485.

- Yu, L., X.-y. Wang, et al. (2010). "Adaptive Evolution of Digestive RNASE1 Genes in Leaf-Eating Monkeys Revisited: New Insights from Ten Additional Colobines." Molecular Biology and Evolution **27**(1): 121-131.
- Zakon, H. H., Y. Lu, et al. (2006). "Sodium channel genes and the evolution of diversity in communication signals of electric fishes: Convergent molecular evolution." Proceedings of the National Academy of Sciences of the United States of America **103**(10): 3675-3680.
- Zhang, J. (2000). "Rates of Conservative and Radical Nonsynonymous Nucleotide Substitutions in Mammalian Nuclear Genes." Journal of Molecular Evolution **50**(1): 56-68.
- Zhang, J. (2006). "Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys." Nat Genet **38**(7): 819-823.
- Zhang, J., D. M. Webb, et al. (2002). "Accelerated Protein Evolution and Origins of Human-Specific Features: FOXP2 as an Example." Genetics **162**(4): 1825-1835.
- Zhang, J. Z. and S. Kumar (1997). "Detection of convergent and parallel evolution at the amino acid sequence level." Molecular Biology and Evolution **14**(5): 527-536.
- Zhang, J. Z., R. Nielsen, et al. (2005). "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level." Molecular Biology and Evolution **22**(12): 2472-2479.
- Zhang, W., A. Collins, et al. (2004). "Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps." Proceedings of the National Academy of Sciences of the United States of America **101**(52): 18075-18080.
- Zou, J. M. and J. G. Saven (2000). "Statistical theory of combinatorial libraries of folding proteins: Energetic discrimination of a target structure." Journal of Molecular Biology **296**(1): 281-294.

## **Biography**

Rui Wang was born on September 26, 1982, in Kenli county, Shandong province, China. He obtained his Bachelor degree of Science at Tsinghua University, Beijing, China in 2003. He won the first place in the National College Entrance Examination of Science in Shandong province in 1999, the first-class excellence in academia fellowship from Tsinghua University in 2000 and a silver medal for Outward Bound program of Tsinghua University as one of the principal organizers in 2001, and was the Phase I prize winner at Duke startup challenge as the team leader in 2006. He married Huimeng Lei in 2009, who is also an alumna of Tsinghua University and Duke University.