EVOLUTIONARY NEUROGENOMIC APPROACHES PROVIDE INSIGHT INTO THE MOLECULAR BASIS OF VOCAL LEARNING

By

Morgan Wirthlin

A DISSERTATION

Presented to the Department of Behavioral Neuroscience
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

September 2016

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

_____

This is to certify that the PhD dissertation of
Morgan Wirthlin
has been approved

_____
Claudio Mello, Mentor/Advisor

_____
Lucia Carbone, Oral Exam Committee Chair

_____
Stephen David, Member

_____
Jacob Raber, Member

_____
John Brigande, Member

TABLE OF CONTENTS

List of Tables ii
List of Figures iii
Acknowledgments iv
Abstract v
1 Introduction 1
1.1 Models for the study of vocal learning 1
1.2 The neural basis for vocal learning 4
1.3 The molecular and genetic basis for vocal learning 8
2 The role of novel genes in the evolution of vocal learning 10
2.1 Background 10
2.2 Results & Discussion 11
2.3 Methods 41
3 'Core sets' of genes expressed in vocal learning circuits 48
3.1 Background 48
3.2 Results & Discussion 51
3.3 Methods 68
4 The role of genomic regulatory elements in the evolution of vocal learning 70
4.1 Background 70
4.2 Results & Discussion 71
2.1 Methods 96
5 New models for understanding the evolution of a complex behavior: a general discussion 101
5.1 Current models of the evolution of learned vocal behavior 101
5.2 The 'Stepwise Hypothesis,' a new model for the evolution of vocal learning 104
6 Conclusions 111
6.1 Conclusions 111
6.2 Future Directions 113
Bibliography 117

List of Tables

Table 1: Novel genes in songbirds 16
Table 2. Comparative ISH analysis of HVC-like nuclei 56
Table 3. Comparative ISH analysis of RA-like nuclei 57
Table 4. Comparative ISH of song system markers in suboscines 65
Table 5. Promoters of song system marker genes with high quality ISH data 74
Table 6. GC content and repetitive sequence in zebra finch brain-expressed gene promoters 80
Table 7. TF motifs associated with song system gene sets 82
Table 8. Brain-expression TFs associated with HVC projection neuron and RA expressed gene sets 84
Table 9. TFs enriched in promoters of song system marker gene sets in zebra finch and chicken 94

List of Figures

Figure 1. Phylogeny of major extant avian lineages 3
Figure 2. Simplified schematic of the zebra finch song system 5
Figure 3. Chromosomal rearrangement leading to songbird gene duplication 18
Figure 4. Syntenic and protein functional domain analysis of YTHDC2L1 and YTHDC2L5 22
Figure 5. Schematic model depicting the predicted structure of songbird novel gene TMRA 24
Figure 6. Expression of URB1 in the zebra finch brain 26
Figure 7. Expression of CASC1 gene duplicatios in the zebra finch brain 28
Figure 8. Expression of YTHDC2L1 gene duplications in the zebra finch brain 30
Figure 9. Expression of TMRA in the zebra finch brain 31
Figure 10. Schematic of analogous song-activated brain nuclei in avian vocal learners 49
Figure 11. Conserved gene markers of zebra finch HVC and Anna's hummingbird VLN 53
Figure 12. Comparative ISH of PVALB in HVC-like nuclei 59
Figure 13. Phylogeny of songbirds, parrots, and suboscine families 61
Figure 14. Comparative neuroanatomy of songbird RA and a suboscine arcopallial domain 66
Figure 15. Expression of song system markers in Amazonian antshrike 67
Figure 16. FISH demonstrating select expression of gene markers in HVC projection neurons 77
Figure 17. ISH of markers of HVC projection neurons 77
Figure 18. Position of TBP and EGR1 motifs in brain-expressed zebra finch promoters 80

Acknowledgments

I wish to thank the three most important mentors of my life, Manfred Ruddat, Daniel Margoliash, and Claudio Mello. It is through your inspiration, training, and belief in me that I have made it to this point.

I am deeply indebted to the members of the Mello lab who have shared their time and expertise over the years, especially Peter Lovell, Chris Olson, and Julia Carleton. Without their help, especially in contributing microarray data and assisting with many of the hybridization experiments, this work would not have been possible.

I would also like to thank my wife, Andria, for her love, encouragement, and support. This past year has been one of many milestones, and I can't imagine having anyone else by my side for the journey.

This dissertation is dedicated to the memory of the birds whose lives were ended in the course of these studies. It is my sincere hope that through passionate public outreach, we can inspire future generations to prioritize conservation, that we might continue to share the planet with these fascinating fellow lifeforms for many generations to come.

Abstract

How are complex, learned behaviors - such as the ability to speak and sing - encoded in our genes, and how did they evolve? Human speech is dependent on our ability to learn and control our vocalizations, a process that is poorly understood at the molecular level. Further, this trait is exceedingly rare in the animal kingdom, and notably absent even in our closest primate relatives, as well as in rodent laboratory models. I sought to gain insight into the neural and genetic bases for vocal learning by performing a series of comparative experiments across more than 50 avian species. This involved devising novel computational analysis pipelines, adapting laboratory practices to a field setting, and contributing extensively to our knowledge of neuroanatomical gene expression and regulation in birds. Broadly, these findings have sought to explore three possible mechanisms for the evolution of a new brain circuit to control a complex sensorimotor task. First, I performed extensive annotation and comparison of avian genomes, in order to demonstrate the existence of novel genes found uniquely in songbirds. I found that some of these were active specifically in the song system, suggesting that brain structures for vocal learning can evolve partially through larger processes of genomic evolution, such as gene duplication by chromosomal reshuffling. Second, I showed that the three lineages of birds that share vocal learning behavior—parrots, songbirds, and hummingbirds—also share expression of a 'core' set of genes in brain areas critical for the production of learned vocalizations. Intriguingly, I also discovered that a fourth group of birds related to parrots and songbirds, where there is mixed evidence for vocal learning, shares expression of some of these core sets, suggesting that vocal learning might evolve in a 'stepwise' manner. Third, I identified gene regulatory elements associated with differential expression in select vocal learning nuclei in zebra finch. By comparing regulatory elements shared with chicken, I showed that some of these gene regulatory motifs may be broadly conserved in birds, whereas others were found only in finch. This suggests that new brain structures for behavior could emerge both through the evolution of new gene regulatory associations, as well as through co-opting preexisting ones. In sum, this work provides a roadmap for integrating knowledge from genomics, gene expression analysis, neuroanatomy, physiology, and phylogenetics to provide novel insights into the evolution of behavior.

1 Introduction

1.1 Models for the study of vocal learning

Human language is, arguably, the defining trait of our species. It has been theorized to have been the evolutionary driver of some of our most essential characteristics. Our exceptionally expanded cortex, advanced theory of mind, capacity to organize into complex societies, and ability to pass on individual knowledge of essential survival skills such as tool-making, have all been suggested to be by-products of the evolution of language [1-4]. Thus, to gain insight into the biology of vocal learning is to peer through a window into the very nature of what makes us human.

To understand the bases for human language, as with any other human trait, the search must begin with our closest non-human relatives: the Great and Small Apes and other non-human primates. Attempts to identify proto-linguistic features of non-human primate communication stretch back nearly a century, ranging from long-term studies of behavior in naturalistic settings to misguided attempts to rear individual Great Apes in human families in order to determine whether any latent linguistic capability may be present [5-7]. Perhaps the most elucidating line of research is that on Great Ape gestural communication, in which the capacity for symbolic and even semantic communication has been suggested [8-11]. On the other hand, studies of Great Ape vocal behavior have failed to reveal any vocal behaviors that are dependent on learning or that change meaningfully over the course of development [5, 12]. Some recent studies have reported a limited capacity of Great Apes to use naturally occurring calls differentially depending on behavioral context and audience [13-17]. However, these represent examples not of learned vocalizations but of learned contextual usage of an unlearned call [18, 19], not unlike the manner in which domestic dogs learn to use their unlearned but naturally variable barking to coax the provision of food out of its owner.

Beyond humans, we can find only a few examples of *vocal learning*– the broader biological term for learned, imitative vocal behavior that provides the basis for human speech acquisition [18-21]. It has been perhaps most convincingly and elegantly demonstrated in cetaceans [12, 22], although strong signatures of learning have also been found in the social vocalizations of bats [23, 24]. Anecdotal evidence has been presented of vocal imitation in elephants [25] and pinnipeds [26, 27]. In all of these cases, the underlying brain circuitry and molecular mechanisms are unknown. Further, invasive mechanistic studies of vocal behavior in mammals are either not tractable experimentally (e.g., cetaceans, pinnipeds, elephants, humans) or not well-characterized in terms of learned vocal behavior (e.g., bats [23]), and traditional experimental organisms like rodents and non-human primates appear to have no capacity for vocal learning [28].

Of more than 30 extant orders of birds, vocal learning has been conclusively demonstrated in just 3 lineages: songbirds, parrots, and hummingbirds (Figure 1, [20, 29]). Previously, these three lineages of avian vocal learners were thought to be only distantly related [30]. The implication of this was that vocal learning had evolved three times in birds, independently, presenting a classic case study of the convergent evolution of a behavioral trait [20]. However, advances in molecular phylogenetics overturned this thinking, when it was conclusively demonstrated that parrots and songbirds represented near sister taxa, with only the non-learning suboscine passerines to separate them [31, 32]. Thus, it may be possible that vocal learning has evolved only twice in birds—once in hummingbirds and once in the last common ancestor of parrots and songbirds. A more detailed exploration of these possibilities is presented in Chapters 3 and 5.

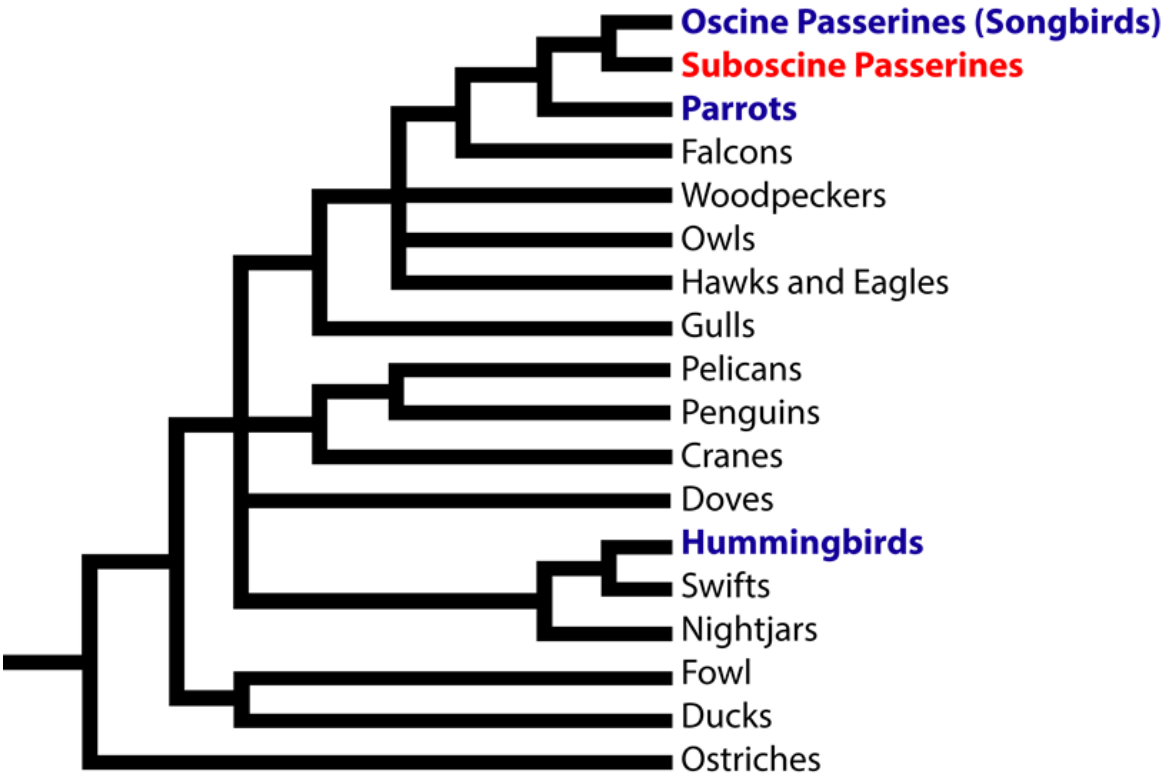

Figure 1. Phylogeny of major extant avian lineages. Lineages in blue possess the behavioral trait of vocal learning its attendant neural circuitry, known as the song system. Suboscine passerines, in red, possess mixed evidence of learned vocal behavior (see Chapter 3 for further discussion). Lineages in black are thought to be vocal non-learning. Phylogeny adapted from Jarvis et al, 2014 [33].

The study of important songbird models, including the canary (*Serinus canaria*), zebra finch (*Taeniopygia guttata*), and European starling (*Sturnus vulgaris*), has contributed tremendously to our understanding of sexual dimorphism in behavior and its correlates in the brain, the role of sleep in neuroplasticity and memory integration, the discovery and study of adult neurogenesis, and, perhaps most of all, the neurobiology of vocal learning [34, 35]. Songbirds possess complex syringeal morphology to control this behavior, not found in non-learning birds [36-38]. This mirrors the human condition, where our descended larynx shows morphological adaptations for fine control not found in non-human primates or, it has been proposed, Neanderthals [39].

As with human infants, a songbird that *must* learn its vocalizations from a tutor in much the same way a child learns to speak by copying adults [18, 40]. Without this input during a critical period, either through isolation from a tutor or deafening, the bird will not be able to learn species-typical song [18, 41]. Further, when cross-fostered by tutors from a similar species, songbirds will learn the song of their adoptive parents, just as a child born of Swedish parents would learn Finnish if adopted by a Finnish family [42]. This is a hallmark feature of vocal learning. In vocal non-learning animals, all species-typical vocalizations develop normally in deafened or isolated animals, and cross-fostered animals fail to learn the vocalizations of their adoptive tutors [43-45]. As in humans and cetaceans, these learned vocalizations are transmitted culturally, across successive generations [18, 46].

Essential to their usefulness as a behavioral model, model songbirds thrive in captivity [47] and readily learn species-typical vocalizations in a laboratory setting [48-50]. Although no animal model perfectly replicates all aspects of human language, songbirds present an excellent, well-studied starting point in the search for the neural and genetic basis of vocal learning, the substrate for human speech acquisition [34, 35, 51].

1.2 The neural basis for vocal learning

Vocal behavior requires the coordinated activity of vocal-motor and respiratory centers in the midbrain and medulla that are present in all vocalizing vertebrates—from mammals and birds to fish—whether or not they are vocal learners [36, 52-55]. In addition to these structures, all vocal learning birds (songbirds, parrots, and hummingbirds) have been demonstrated to possess a set of discrete, interconnected nuclei that appear to be devoted solely to vocal learning, termed the song system (Figure 2) [41, 56, 57]. Although less well-studied in parrots and hummingbirds, in songbirds the presence of a song system has been demonstrated to be broadly conserved across

the taxon [58-61]. All other vocal non-learning birds examined thus far (e.g., suboscine Passerines, swifts, owls, doves, and fowl) appear to lack these areas [40, 56, 57, 59, 62-69]. On the basis of extensive cytoarchitectonic similarities and extensive shared gene expression, it has been determined that these nuclei fall within avian pallial brain areas homologous to mammalian cortex, basal ganglia, and thalamus [70, 71]. This is an important distinction, as the presence of forebrain vocal areas implies that these structures represent an avian equivalent of cortical and basal ganglia circuitry that in mammals is associated with complex sensorimotor learning. Thus, vocal learning birds have cortical and basal ganglia control over their vocalizations, a feature analogous to humans that is clearly absent in vocal non-learning birds.

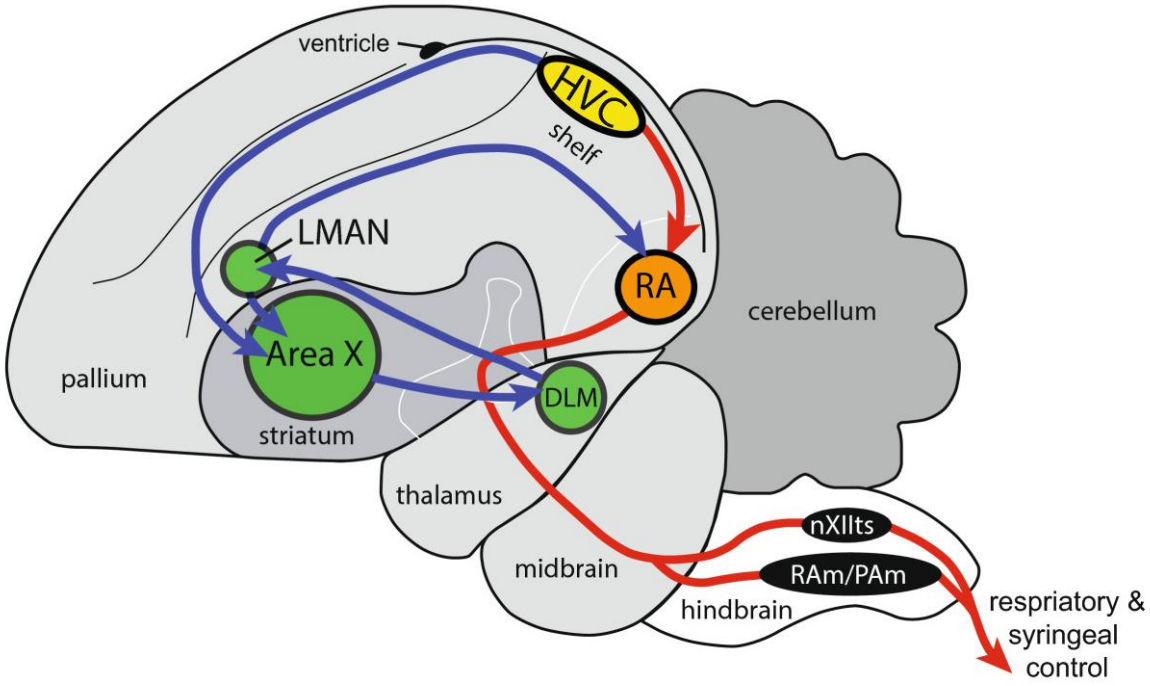

Figure 2. Simplified schematic of the Zebra finch nuclei specialized for vocal learning and their connections. Red projection: the posterior vocal motor pathway for vocal control, originating in HVC, in yellow, continuing to RA, in orange, to the hindbrain vocal motor nuclei, in black. Blue projection: the anterior forebrain pathway for vocal learning, originating in HVC, which projects to a set of interconnected nuclei (Area X, LMAN, DLM) analogous to mammalian cortical-basal ganglia-thalamo-cortical loops for sensorimotor learning. Abbreviations: DLM, medial part of the dorsal lateral nucleus of the thalamus; LMAN, lateral magnocellular nucleus of the anterior nidopallium; HVC, proper name; nXIIts, tracheosyringeal portion of the hypoglossal nucleus; PAm/RAm, nucleus para-ambiguus/retroambiguus; RA, robust nucleus of the arcopallium.

Elements of the vocal control circuit appear to be functionally analogous to human speech production areas, including Broca's Area, and decades of research on the songbird vocal control circuit have produced insights into the functional organization of pathways for the acquisition and production of learned vocalizations [20, 21]. In songbirds, the vocal control circuit is comprised of two pathways: first, a vocal-motor pathway, essential for the production of learned vocalizations [72, 73] (Figure 2, in red), and second, an anterior forebrain pathway (AFP), essential for the learning and maintenance of imitative vocal behavior[74-76] (Figure 2, in blue).

The vocal motor pathway consists of a pre-motor vocal nucleus (HVC), which projects to a vocal-motor cortical area (RA), which projects to vocal centers in the midbrain (DLM) and medulla (nXIIts, RAm/PAm) [41, 59]. Direct corticospinal projections, such as those from RA to nXIIts, have been shown in mammals to be critical for controlling movements requiring exceptional skill and flexibility, as well as enabling volitional control of movement [77, 78]. Notably, direct corticospinal vocal-motor connections do not appear to exist in non-learning birds or non-humans primates [79-82].

As a bird sings, each RA projection neurons fires in a unique, highly stereotyped bursting pattern that is tightly time-locked to specific song motifs [83]. Interestingly, it has been shown that these firing patterns are 'replayed' in RA as the bird sleeps [84], and that this replay and its effects on synaptic plasticity are necessary for song learning [85, 86], lending unique insight into the role of sleep in learning [87].

The AFP is comprised of a set of interconnected nuclei (Area X to LMAN to DLM to Area X), analogous to mammalian cortical-basal ganglia-thalamo-cortical sensorimotor loops [65, 88, 89]. This analogy is also functional, as the AFP is also essentially a circuit for the learning of a complex sensorimotor sequence [90-92]. The AFP is especially active in driving neuroplasticity during the zebra finch subsong period, a critical period for sensory template acquisition and vocal motor

practice akin to the babbling phase of human speech development [93, 94]. This knowledge motivated renewed interest in the role of the human basal ganglia in speech, yielding further insight into the neural basis for language acquisition [21].

The premotor vocal control nucleus HVC represents the main premotor nucleus of the adult song system, containing cells critical to the neural encoding of learned song, and representing the main entry point for auditory information that provides song performance feedback to the system [95, 96]. HVC is a nodal area in the songbird brain, as it is the origin for both the learning and motor pathways, two distinct neuronal populations that project to the motor output nucleus, RA (HVC-RA), or to the AFP basal ganglia nucleus, Area X (HVC-X) [97, 98]. These projection neuron types are highly distinct in terms of their electrophysiology, cellular morphology, dynamics of replacement in adulthood, and gene maker expression [99-103]. These neurons fire sparsely as the bird vocalizes, with any particular HVC projecting neuron bursting only once per song [95, 104]. Electrophysiological recordings reveal that as a population, these neurons propagate calcium-mediated excitatory bursts along a synaptically connected chain [105]. However, to what degree these bursts function as a sort of simple clock to control the timing of song [106] or provide a feedforward model of the sequence of motor gestures underlying song [107] remains to be resolved. Both types of projection neurons are required for the production of learned vocal behavior, as has been demonstrated through selective lesioning as well as electrical and optogenetic inactivation experiments [108, 109]. HVC-RA neurons undergo adult replacement, their precursors migrating into the nucleus from the subventricular zone [110-112]. This adult neurogenesis is androgen-dependent, providing a neuroanatomical and molecular mechanism for the seasonal neurogenesis observed in wild songbirds, where HVC is observed to double in size and cell density during the breeding season [72, 98].

1.3 The molecular and genetic basis for vocal learning

Speech production impairments disrupt the lives of millions of individuals. Ranging from mild stuttering to complete speech loss, the cost these disabilities incur in terms of lost productivity and therapeutic care is estimated at US$30 billion to $154 billion [113]. At present, our understanding of the genetic basis of speech acquisition, which could lead to improved treatments and therapies for communication disorders, is poor. A significant genetic component to vocal learning has been demonstrated through studies in both humans [114] and zebra finch [115, 116], which indicate that variation in the propensity to learn song and language is highly heritable. Only a handful of genes involved in vocal motor production have been identified by costly genome-wide association studies in humans with language disorders. Most famously, mutations to the transcription factor FOXP2 result in significant language impairments [117]. A handful of additional genes have been tied to the delayed language acquisition observed in association with autism [118, 119] and stuttering [120]. Although studies of these rare mutations are of value in demonstrating the link between single genes and language, they represent an inefficient strategy for determining the full molecular pathways supporting speech, which are likely to involve large networks of genes acting in concert in discrete centers of language production in the brain [121].

Although long held as the choice animal model of human speech production, it has only recently been demonstrated that specific elements of the songbird vocal control circuit are not only functionally analogous to human speech production areas, but also exhibit shared gene expression with their human counterparts [122]. This has motivated extensive efforts to characterize the molecular and genomic substrates of the song system, including the sequencing of the zebra finch genome [123]. Genomics and large-scale brain gene expression studies using microarrays and *in situ* hybridization (ISH) have identified hundreds of markers of the vocal control system of songbirds [124-127]. Some of these genes are under strong evolutionary pressure in

songbirds [128-130], with several being prominent markers of song nuclei, a subset showing selective expression in sparse cell populations [128].

Interestingly, the human language-linked gene FOXP2 is also a marker of select song system nuclei in songbirds, parrots, and hummingbirds, in analogous locations within homologous brain divisions [131-133]. Further, knockdown of FOXP2 in songbird Area X appears to detrimentally affect birds' ability to accurately learn a tutor's song [134]. A critical regulatory target of FOXP2, CNTNAP2, is also expressed differentially in the zebra finch song system [135, 136]. In human studies, it has been suggested that CNTNAP2 is the primary effector of the deficits observed in patients with disabling FOXP2 alleles – mutations in CNTNAP2 result in profound language deficits similar to those observed in FOXP2 patients [118, 137], but even CNTNAP2 variant alleles found in healthy populations have been shown to affect performance on language-related tasks [138]. Songbird researchers have noted differential expression and protein abundance of CNTNAP2 in song nuclei, including RA [135]. This suggests that transcriptional regulatory networks play a critical role in the normal function of vocal learning circuits. However, to date, only one attempt has been made to study the large-scale regulatory dynamics in the song system, which revealed transcriptional mechanisms underlying song-activated gene transcription [139].

2 The role of novel genes in the evolution of vocal learning

2.1 Background

The completion of the Zebra finch genome in 2010 [123] made it possible to search for genomic features that might be unique to songbirds. In addition to vocal learning, passerines are distinguished by a number of traits including their distinctive foot anatomy adapted for perching, an altricial pattern of offspring growth, and an exceptionally high metabolic rate [140, 141]. The emergence of novel genes has been shown to provide a genetic substrate for lineage-specific adaptations and the evolution of new functional traits [142, 143]. It is thus possible that novel genes might also be associated with the evolution of characteristic songbird traits like vocal learning. The initial comparative analysis between Zebra finch and chicken (Galliformes: *Gallus gallus*), a previously sequenced non-learning species [144], led to the identification of candidate novel gene duplications and expansions in Zebra finches [123]. However, without additional genomes that span the large phylogenetic distance between Galliformes and Passeriformes, it was impossible to conclude whether these features are specific to Zebra finch, or originated more basally in finches, oscines, Passeriformes, or elsewhere within Neoaves. Moreover, it has not been determined whether these genomic features are associated with vocal learning and related circuitry, or with other phenotypic differences between oscines and non-oscines. In the few cases where this question has been examined, there is only very limited evidence of regional or differential gene expression in the brain [123, 145].

A requirement for accurate comparative genomics is the generation of well-curated sets of 1-to-1 orthologs among the organisms being compared. Obtaining the true set of orthologs is an ongoing problem, and many genomes remain partial and/or insufficiently annotated, in part due to errors by *ab initio* gene predictive algorithms to identify all exons of a given ortholog, even when these are present in the genome assembly [146]. These issues have led to erroneous

annotations of Zebra finch models as novel, non-detection of orthologs that are in fact present in the assembly, and erroneous conclusions about gene duplications and expansions [128]. Thus, the search for genomic features unique to Zebra finches, or more broadly to songbirds, and that may relate to their distinct traits, is still very incomplete.

Here, I curated the Zebra finch Ensembl gene models annotated as novel or duplicated, as well as several previously identified novel expansions in the Zebra finch genome (Supplementary Figure 3 in [123]), and then determined their presence or absence in 45 recently sequenced, high-coverage (30 – 120X) genomes spanning the avian phylogeny (phylogeny and genomes described in [33, 147]). As a result, I have identified a set of genes that are clearly novel in songbirds, and provide evidence for chromosomal rearrangement as a potential mechanism for their origin. I also show that some of these genes are expressed in the brain, and that a small subset are differentially expressed in song nuclei relative to surrounding areas, representing molecular specializations of the songbird vocal control system; others are expressed in various non-neural tissues. In contrast, I found that several hundred putative, previously reported novel songbird gene models [145] represent artifacts, previously known genes, or genes that are present in non-songbird species. These results provide an approach for improved genome annotation, as well as identifying novel targets for investigating genes unique to a lineage or trait, including vocal learning and its associated brain circuitry in songbirds.

2.2 Results & Discussion

I implemented a comprehensive and exhaustive annotation pipeline to identify genes that evolved in songbirds subsequent to their divergence from all other birds. Specifically, I first focused on the best-annotated songbird species, the Zebra finch, retrieving all Zebra finch Ensembl gene models annotated as novel, duplicated, or expanded (>7,000 models, e59), and removing models Ensembl

considered to have orthologs in other species (n = 5,459), as well as those mapped to chromosome Unknown (n = 1,179), which were likely to represent allelic variants. The mRNA and protein sequences of the remaining models (n = 876) were aligned using the BLAST-like alignment tool (BLAT) to chicken, currently the best-annotated avian species, and Zebra finch, and synteny for all hits of sufficiently high score was verified to identify conserved orthologs. Loci present in Zebra finch but not chicken were then aligned to additional genomes—turkey (*Meleagris gallopavo*), lizard (*Anolis carolinensis*), frog (*Xenopus tropicalis*), Zebrafish (*Danio rerio*), mouse (*Mus musculus*), and human (*Homo sapiens*)—to distinguish songbird novel genes from losses in the galliform lineage (i.e., lost in chicken and turkey genomes only).

I found a distinct group of putatively novel or duplicated genes in Zebra finch that do not present an ortholog in Ensembl (e75) and NCBI databases (n = 61). This included several expanded gene families located on various Zebra finch chromosomes, suggesting that they do not represent local assembly artifacts. I added to this group a set of genes (n = 13) previously reported as expanded in Zebra finch (Supplementary Table 3 in [123]), and an additional set of duplicated loci that lack a predictive model (n = 17), detected in the course of BLAT-alignments during the candidate novel gene curation analysis.

To determine whether these genes represented true novel genomic features unique to and shared across songbirds, I conducted BLAST searches of 48 avian genomes, including 45 newly sequenced species, representing a broad sampling that covers all major extant radiations of avian diversity (complete list of species and genome assembly in [33, 147]). This provided an initial confirmation of the existence of genes unique to songbirds. However, as the BLAST output was to unannotated scaffolds, putative songbird novel gene sets were analyzed in more detail using custom optimized BLAT-alignment algorithm and syntenic analysis in the IGV browser (see Methods). This novel gene-finding strategy was applied to a subset of the new avian genomes:

two additional songbirds, the Medium ground finch (*Geospiza fortis*) and the American crow (*Corvus brachyrhynchos*); a suboscine passerine, the Golden-collared manakin (*Manacus vitellinus*); a basal passerine, the Rifleman (*Acanthisitta chloris*); and the Budgerigar (*Melopsittacus undulatus*), a parrot, the sister taxon to Passeriformes. This analysis allowed me to discard spurious hits from the BLAST search. In some cases it also allowed for the identification of the "parent" gene, shown through syntenic conservation to be orthologous between songbirds and non-songbirds, and which underwent duplication to give rise to the novel genes. Confirmed novel genes were categorized into several subsets based on their presumed phylogenetic origin (Table 1, column 2).

I discovered 10 genes that are present in the 3 songbird species studied and absent in all 45 avian non-songbird and 4 non-avian organisms examined, thus representing novel genes that are unique to songbirds. These included 7 cases of gene duplication resulting in a single novel paralog unique to songbirds and 1 case where the gene duplication resulted in two songbird-specific novel paralogs (Table 1; paralogs more generally conserved in passerines or unique to Zebra finch are also included). In some cases the parent gene is known (novel gene names have a terminal 'L'), in others the parent gene could not be established (gene names have a dash and number). I also identified a novel gene which appears to have arisen *de novo* in songbirds. Altogether, the parent genes and the expanded and *de novo* novel loci comprise a total of 38 genes (Table 1). I describe here their characterization, followed by further details on the general curation effort.

Gene name ¥	Phylogeny	Gene location	Brain-derived ESTs	Non-brain ESTs	Non-brain RNA-SEQ
A4GALT-related*					
A4GALT-1	SONGBIRDS†	chr1A:65,194,903-65,195,964	No	No	None
A4GALT-2	SONGBIRDS†	chr1A:65,202,780-65,203,835	No	No	None
A4GALT-3	SONGBIRDS†	chr1A:65,210,658-65,211,719	No	No	None
CASC1-related					
CASC1-1	SONGBIRDS†	chr1A:65,157,078-65,159,677	FE723736√, DV948439	JV165872, JV165873	Embryo, spleen, testes
CASC1-2	SONGBIRDS†	chr1A:66,373,778-66,384,478	DV948439	JV184784, JV165872, JV165873	Embryo, liver, muscle, testes
FN3KRP-related					
FN3KRP	ALL BIRDS‡	chr18:6,574624-6,580,544	DV959265√	No	Embryo
FN3KRPL1	SONGBIRDS	chrZ:24,858,422-24,862,943	No	No	Embryo, liver, skin
FN3KRPL2	PASSERINES	chrZ:69,583,022-69,590,574	DV955139√, FE727948√, DV955139√	JV168705√, JV168706√, JR864904√	Embryo, liver, spleen, testes
HYDIN-related*					
HYDIN	ALL BIRDS‡	chr11:5,451,491-5,475,997	No	No	Liver, muscle, skin, spleen
HYDINL1	SONGBIRDS	chr11:16,377,691-16,404,553	No	JV172391	Muscle, skin, spleen
HYDINL2	PASSERINES	chr11:16,633,304-16,663,944	No	No	Liver, muscle, testes
HYDINL3	ZEBRA FINCH	chr11:17,005,696-17,033,616	No	No	Muscle, testes
HYDINL4	ZEBRA FINCH	chr11:18,229,584-18,238,051	No	No	None
HYDINL5	ZEBRA FINCH	chr11:19,907,326-19,914,490	No	JV172391	Testes

Gene name ¥	Phylogeny	Gene location	Brain-derived ESTs	Non-brain ESTs	Non-brain RNA-SEQ
NOVEL (TMRA)					
TMRA	SONGBIRDS	chr1A:66,486,182-66,494,397	CK302958√	JV159445√, JV159451√	Embryo, liver, muscle, spleen, testes
RIOK2-related					
RIOK2	ALL BIRDS‡	chrZ:24,872,816-24,883,105	DV956882√	JV172474√, JR863880√	Embryo, liver, muscle, skin, spleen, testes
RIOK2L	SONGBIRDS	chrZ:69,578,247-69,578,886	No	No	Skin
RNF4-related*					
RNF4	ALL BIRDS‡	chr4:62,477,216-62,484,738	DV951366√	JV183872√, JR867734√	Liver, skin, testes
RNF4L1	SONGBIRDS	chr4:8,201,210-8,201,765	No	No	None
RNF4L2	PASSERINES	chr4:20,660,938-20,751,958	No	No	None
RNF4L3	PASSERINES	chr4:22,411,072-22,433,579	No	No	Embryo, muscle, skin, spleen
RNF4L4	ZEBRA FINCH	chr4:22,445,102-22,478,943	No	No	Testes
RNF4L5	ZEBRA FINCH	chr4:22,507,187-22,517,977	No	No	Muscle, spleen
RNF4L6	ZEBRA FINCH	chr4:22,538,729-22,547,838	No	No	None
RNF4L7	PASSERINES	chr4:41,650,214-41,650,348	No	No	None
URB1-related					
URB1	ALL BIRDS‡	chr1:97,555,543-97,592,096/ chr1_random:362,665-365,267	FE722167√	No	Embryo, liver, muscle, skin, testes
URB1L1	ZEBRA FINCH	chr1A:60,924,812-60,941,816	CK301434, CK303889, DV957700	No	Embryo, liver
URB1L2	ZEBRA FINCH	chr1A:63,520,638-63,528,208	CK301434, CK303889	No	None

Gene name ¥	Phylogeny	Gene location	Brain-derived ESTs	Non-brain ESTs	Non-brain RNA-SEQ
URB1L3	SONGBIRDS	chr5:4,764,796-4,772,823	CK301434, CK303889, DV957700	No	Skin
URB1L4	ZEBRA FINCH	chr7:1,064,071-1,065,024	No	No	None
URB1L5	ZEBRA FINCH	chr23:2,319,460-2,335,606	DV957700	No	Liver, muscle, spleen, testes
YTHDC2-related*					
YTHDC2	ALL BIRDS‡	chrZ:21,509,497-21,511,474	No	No	None
YTHDC2L1	SONGBIRDS	chr2_random:378,730-383,125	CK309358√	No	None
YTHDC2L2	ZEBRA FINCH	chr3_random:766,156-785,803	No	No	None
YTHDC2L3	PASSERINES	chrZ:10,792,293-10,809,782	No	JV174477, JV177272	Embyro, liver, muscle, spleen
YTHDC2L4	ZEBRA FINCH	chrZ:29,309,810-29,332,222	No	No	Embryo, liver, muscle, skin, spleen, testes
YTHDC2L5	ZEBRA FINCH	chrZ:55,988,800-56,020,229	DV947064√	JV174477√	Embryo, liver, muscle, skin, spleen, testes
YTHDC2L6	ZEBRA FINCH	chrZ:57,443,052-57,444,728	No	No	Embryo

Table 1: Families of genes where at least one member has been determined to be present uniquely in all songbirds. Tandem duplicates, where clear orthology cannot be determined, are distinguished with dashed numbers (e.g., CASC1-1, CASC1-2). Other duplicates are named after being 'like' their parent gene (e.g., RIOK2L). Chromosomal locations refer to the zebra finch genome (taeGut1). Expressed sequence tags (ESTs) providing evidence of gene expression in the brain [57, 81, 82] and RNA-seq data derived from other tissues [27] are from zebra finch, non-brain ESTs are derived from Dark-eyed junco [37]. ¥: Gene symbols have been corrected based on my curation of Ensembl annotation, *: Gene family previously reported as being expanded in Zebra finch, ‡: Parent gene, †: Parent gene cannot be determined, √: EST is specific to this gene locus.

Association of novel genes with genomic rearrangement sites

To determine whether the identified novel songbird genes are located close to regions of chromosomal instability in birds, I first generated a map of the sites of avian syntenic disruption (SD) likely representing chromosomal breakpoints, by comparing the syntenic order of Zebra finch vs. chicken genes using SyntenyTracker [148]. I manually verified the consistency of the synteny groups with previous reports [123, 149], using my curated Ensembl models to allow for a more precise identification of genes in the vicinity of the SDs. I also analyzed chromosomes 11-28 and Z, which were not included in previous studies (Table S1.6). Finally, I examined the syntenic arrangement of these blocks in several outgroups (lizard, mouse, human) to distinguish SDs that are specific to the chicken lineage, where the flanking regions of Zebra finch blocks are either identical or highly similar to those in non-avian species (highlighted in pink in Table S1.6), from SDs specific to the songbird lineage, where the flanking regions of Zebra finch blocks differ from that of the other species examined (highlighted in blue-green in Table S1.6). Compared to previous studies [123, 149], this analysis resulted in a refined and comprehensive list of SDs representing likely chromosomal breakpoints specific to the songbird lineage (Table S1.6).

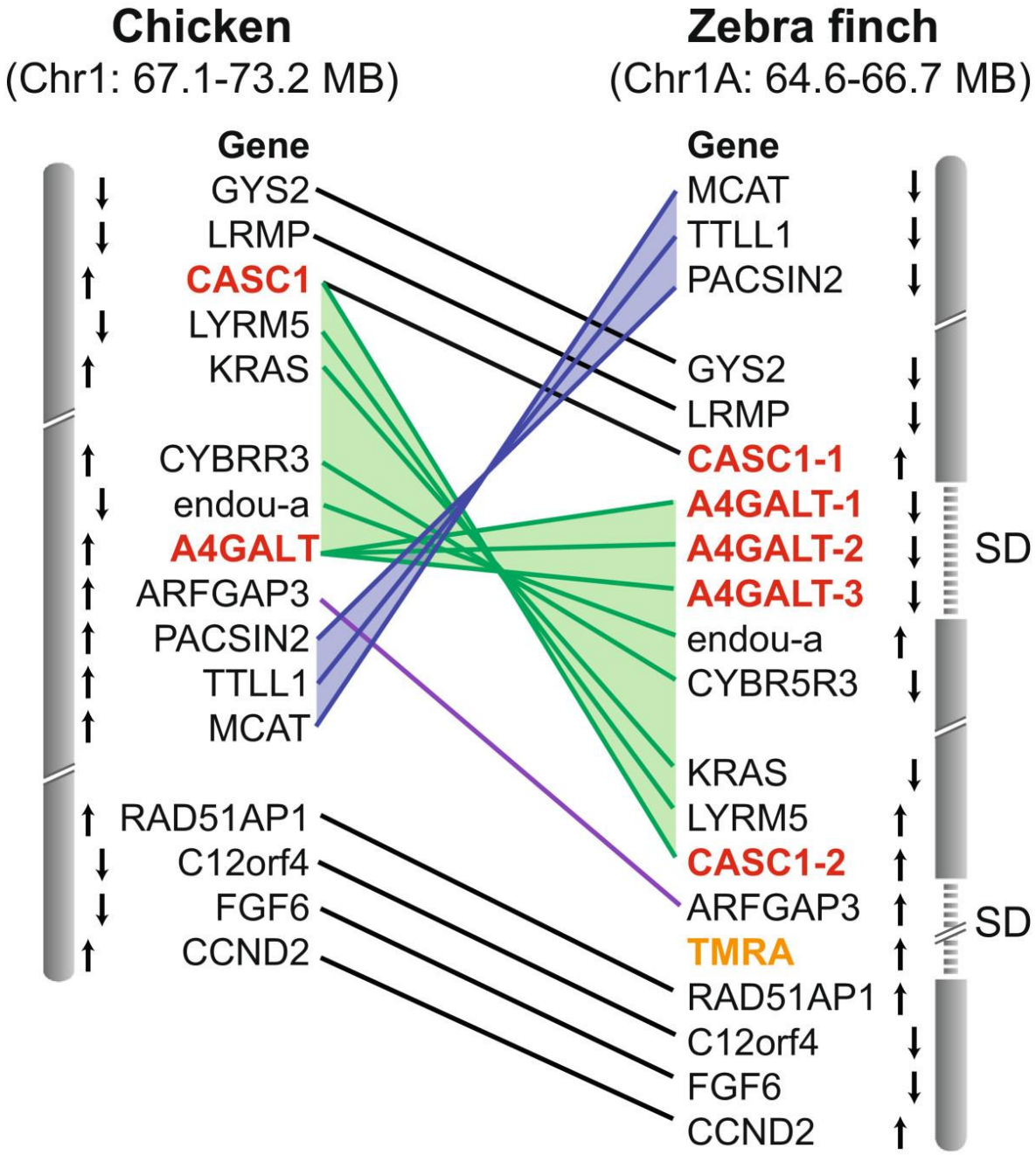

Figure 3. Example of songbird novel and duplicated genes associated with regions of syntenic disruption. Chromosomal maps of the syntenic order of genes in chicken (chr1) compared to Zebra finch (chr1A) reveal that songbird-unique novel (TMRA, in orange) and duplicated/expanded genes (CASC1, A4GALT; in red) are located in chromosomal regions containing syntenic disruptions (SD) that are unique to the songbird lineage (i.e., the syntenic flanking genes in Zebra finch are different from those in other avian and non-avian species). Small black up/down arrows next to each gene indicate orientation on the minus/plus strand of DNA. Line colors denote genes in rearranged syntenic blocks, with shaded regions representing apparent chromosomal inversions.

Next, I examined whether novel songbird genes as well as their parent genes map within or near songbird lineage-specific SD sites. I found that 6 of 10 novel genes are present at the start, end, or within SD regions (position of SDs indicated in Table S1.6). This suggests that a large proportion of the duplication events that gave rise to novel genes seems to have occurred in regions of chromosomal instability. Two regions were associated with multiple duplication events: a chromosomal inversion in chr1A (expansions of *A4GALT* and *CASC1*, and occurrence of *TMRA*; Figure 3), and a chromosomal rearrangement on chrZ (associated with the expansions of *FN3KRP* and *RIOK2*). Interestingly, most of the putative *de novo* novel (*n* = 6 of 8) and ~50% of the duplicated (*n* = 18 of 32) genes that are not unique to songbirds (found in subsets of songbirds, all passerines, or other Neoaves; Table S1.5) are also located within or close to SDs, supporting the association between chromosomal breakpoints and the emergence of novel genomic features in non-songbirds as well.

Predicted protein analysis of songbird-unique genes

We assessed the potential function of novel songbird genes by examining the predicted protein domains of the largest open reading frame. These were obtained either from the Zebra finch Ensembl model, or by mapping the Zebra finch parent gene and its chicken ortholog onto the novel gene locus, in cases where an Ensembl model for the novel duplicate was absent or incomplete (in the latter case noting the possible occurrence of additional unmapped exons/domains in nearby regions). I then compared the predicted domains between parent and duplicate genes to identify potential changes in function. As an outgroup to birds, I verified the domain organization of the parent genes in humans, where gene predictions are often more complete than in other vertebrates.

In one case, I found one gene where the duplicate copies are nearly identical. The *A4GALT* expansion consists of 3 complete tandem copies; all contain a glycosyltransferase DXD-sugar binding motif and an alpha 1,4-glycosyltransferase domain, the latter involved in protein glycosylation. Due to high conservation, it is hard to determine which copy is orthologous to the ancestral *A4GALT*. For the remaining gene families, I found differences in domain annotation between the duplicate and parent genes, with some genes representing partial copies and others gaining additional exons.

For *RIOK2*, the parent gene contains a RIO-like kinase, a RIO2 kinase N-terminal domain, a winged helix domain that confers DNA binding properties, and a coiled-coil domain, indicative of a function in the regulation of gene expression. The model-less *RIOK2L* contains only the coiled-coil domain, and thus is a partial duplicate. Since there are no sequence gaps upstream of the model, the missing N-terminal domains and exons cannot be in a gap. Interestingly, *RIOK2L* overlaps with, and is antisense to, one of the 3'-splice variants of *FN3KRPL2*.

As in the chicken and human orthologs, *FN3KRP* and its duplicate paralog present in all passerines (*FN3KRPL2*) contain six exons, and the predicted large fructosamine-3-kinase domain that covers most of the open reading frame is indicative of a role in deglycation and functional activation of proteins protective against hyperglycemia. *FN3KRPL2* has additional downstream exons, part of several transcripts that are alternatively spliced in the 3'UTR region of this gene. The copy specific to songbirds (*FN3KRPL1*), however, lacks the amino-terminus exon (~45 aa residues) of the parent gene. Since there are no sequence gaps upstream of this locus where that exon might be located, this copy is partial, with a likely disruption of the main functional domain.

Both copies of *CASC1* in songbirds lack a coiled-coil domain close to the amino-terminus present in the chicken and human orthologs. However, since the predicted peptides lack a starting

codon, this domain may be present in the gaps upstream to the models. Furthermore, *CASC1-1* is shorter than *CASC1-2*, consisting only of the 282 amino acid residues at the carboxy-terminus.

As in chicken and human, the parent *URB1* gene in songbirds contains a predicted nucleolar pre-ribosomal-associated protein 1 domain close to the amino-terminus and an armadillo (ARM-) like fold involved in interactions with other proteins and nucleic acids. The expanded set of duplicate *URB1* copies do not contain either of the domains above, although in some cases the missing domains might be hiding in gaps.

In the case of the *YTHDC2* expansion, the expanded set includes several copies that are differently shared across species, and which display marked changes in the predicted structure across the different copies (Figure 4A). In human and chicken, the parent gene contains several domains (R3H, DEAD, Ank_rpt, HELICc, HA2, OB-fold, YTH) involved in functions like binding to and inducing conformational changes in single stranded nucleic acids (RNA or ssDNA). In songbirds it is highly truncated (*YTHDC2*, Figure 4B), lacking most predicted domains. The songbird-unique duplication (*YTHDC2L1*, Figure 4B) also lacks all domains except the HA2 and OB-fold. The alignment of an EST containing a polyA at this locus confirms that the sequence is complete at the 3' end, thus this copy lacks the YTH domain. All 3 songbirds contain relatively complete copies of the gene that lack the amino-terminus (~74 residues) but contain most other domains of the parent gene at different syntenic locations (e.g., *YTHDC2L5* in Zebra finch, Figure 4B). Overall, these observations suggest continued expansion of this gene following divergence of songbirds.

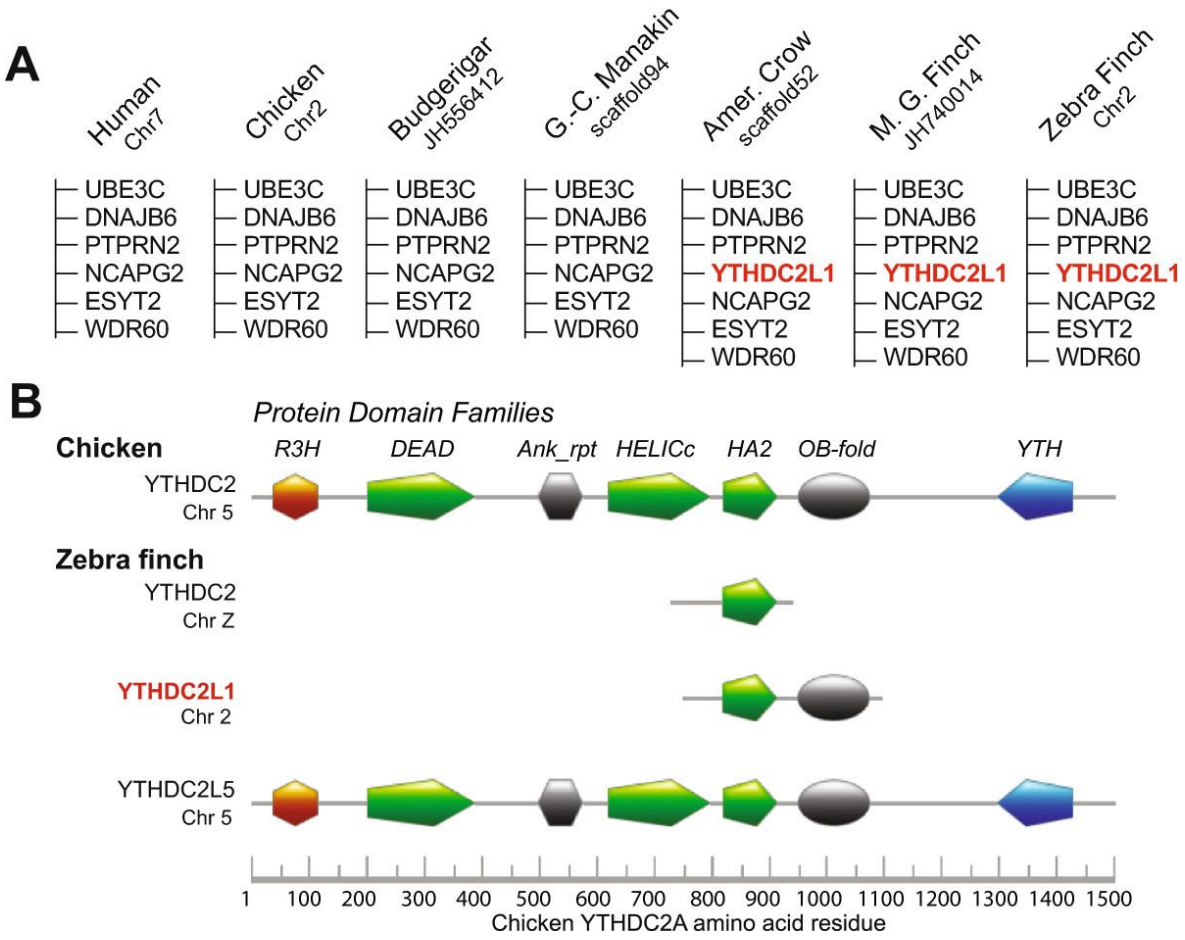

Figure 4. Syntenic and protein functional domain analysis of YTHDC2L1 and YTHDC2L5. (A) Schematic representation of conserved chromosomal loci in avian and non-avian vertebrate species showing the relative position of YTHDC2L1 (in red), a novel expansion of YTHDC2 that is only present in songbirds. Adjacent genes are indicated in black. The chromosome or scaffold number is indicated beneath each species common name. (B) An alignment of the protein family domains predicted for amino acid sequences derived from Chicken YTHDC2, the orthologous YTHDC2 "parent" gene in Zebra finch, as well as the copies of YTHDC2 that are only present in songbirds (YTHDC2L1), and Zebra finch (YTHDC2L5). Specific protein family domains predicted by InterProScan5 are aligned relative to Chicken YTHDC2, and are indicated by the various colored symbols. YTHDC2 and YTHDC2L1 lack nearly all of the major protein family domains that are characteristic of YTHDC2. In contrast, YTHDC2L5 appears to be a nearly complete copy of YTHDC2.

All *HYDIN*-related copies are partial compared to the parent gene in chicken and human, which contains a predicted PapD-like domain related to periplasmic chaperoning function, and a P-loop NTPase domain thought to be involved in modulating conformational changes in other proteins. The Zebra finch *HYDIN* parent gene contains only a short fragment of the parent gene and has no predicted model; all other duplicates contain truncated portions of the 5' part of the parent gene. Because these loci are flanked by gaps and several unassembled sequences (chrUn) contain portions of the gene, some of which close to the 3' end, several assembled loci may be partial due to assembly incompleteness. I also note that the alignment scores of the chicken parent gene and songbird-expanded copies are rather low, reflecting considerable divergence.

Similarly to the orthologous gene in chicken and humans, *RNF4* and its duplicates in passerines each contain a predicted single RING finger motif close to the carboxy-terminus. However, the songbird-specific copy and one of the copies shared by passerines are truncated, lacking most of the 5' half of the gene. Low alignment scores of these duplicates point to significant divergence from the parent locus.

We identified one novel gene (ENSTGUG00000012248) with no identifiable parent gene that appears to have arisen *de novo* in the songbird lineage. Its only trace outside of songbirds is in the form of a short, truncated segment of one of its coding exons in the correct syntenic position in two non-songbird species (Peregrine falcon and Bar-tailed trogon, data not shown). Its predicted protein (330 aa) contains a putative amino-terminus cytoplasmic domain, three transmembrane domains, and a carboxy-terminus extracellular domain, the latter with a putative C-type lectin domain (Figure 5). Thus, it appears to encode a polytopic transmembrane α-helical protein, suggesting a role related to the cell surface, possibly involving carbohydrate binding activity. Due to a gap in the Zebra finch genomic sequence, this analysis required sequencing an ESTIMA cDNA, CK302958, the longer of two brain-derived cDNAs that map specifically to this locus

(complete sequence submitted to GenBank, accession ID:KM520127). Results were confirmed by comparing cloned cDNAs from the Dark-eyed junco, another songbird species [150]. Based on its predicted structure and discrete expression in the song nucleus RA (see below), I annotated this gene *TMRA* (transmembrane protein of the robust nucleus of the arcopallium).

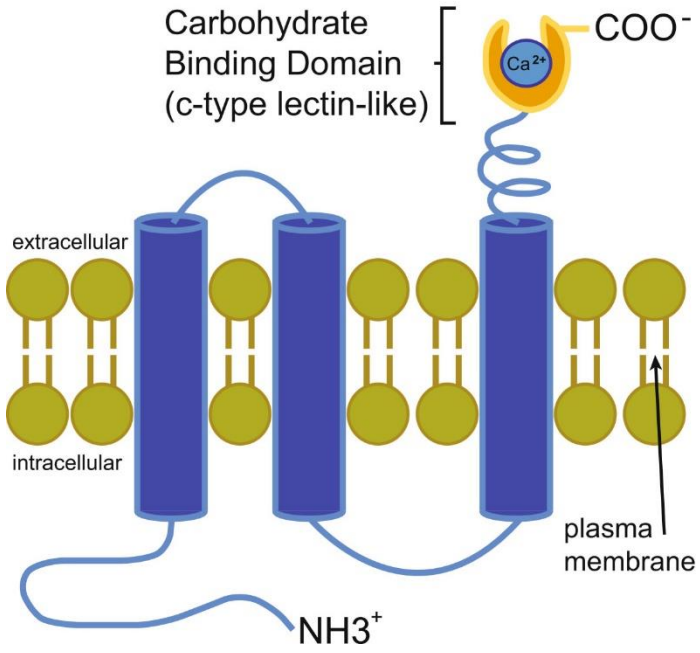

Figure 5. Schematic model depicting the predicted structure of songbird novel gene TMRA in the plasma membrane. Protein functional domain analysis (InterProScan5) predicts the protein coding sequence of TMRA contains three transmembrane spanning domains connected to an extracellular C-type Lectin-like (CLEC) domain that is typically associated with carbohydrate binding.

Expression analysis of songbird-unique genes

To gain insights into the possible function of the identified novel songbird genes, I performed an expression analysis using both publicly available expressed sequence tags (ESTs) and RNA-seq data [123, 150] from brain, liver, muscle, skin, spleen, testes, and whole embryo of Zebra finch and an additional songbird species, the Dark-eyed junco; as well as *in situ* hybridization of Zebra finch brain sections. The results fell into several categories (Table 1): songbird novel genes with no evidence of expression in these tissues (*A4GALT*, *RNF4L1*), expressed in non-brain tissues only

(*FN3KRPL1*, *HYDINL1*, *RIOK2L*), expressed in multiple tissues including brain (*CASC1*, *TMRA*, *URB1L3*), and one gene with expression detected solely in the brain (*YTHDC2L1*). In the case of *A4GALT*, I found no evidence for expression of either the parent gene or the duplicate copy unique to songbirds in these tissues. In contrast, for all other novel genes, I find that transcriptional data reveal differential tissue expression of parent and novel genes, suggesting functional differentiation among loci (Table 1).

In some cases, songbird novel genes show more limited expression than their parent genes. Songbird-unique gene *HYDINL1* has lost expression in liver relative to parent gene *HYDIN*. For the *RIOK2* and *RNF4* gene expansions, I found evidence of expression in brain and other tissues for the parent genes, but limited (*RIOK2L*, expressed only in skin) or no expression (*RNF4L1*) of the songbird duplications (Table 1). In both cases, I was unable to detect parent gene expression in the brain by *in situ* with probes from brain-derived cDNAs, suggesting that expression levels are either very low and/or brain state-dependent.

In other instances, the songbird duplication was expressed in additional tissues beyond those of the parent gene, as in the case of *FN3KRPL1*, which is expressed in liver and skin, as well as sharing expression in embryo with parent gene *FN3KRP* (Table 1). Several brain-derived ESTs map specifically to parent gene *FN3KRP* and to passerine duplicate *FN3KRPL2*, thus both loci are transcriptionally active in the brain. Interestingly, several transcript variants for *FN3KRPL2* differ on the length of the 3'UTR region. *In situ* hybridization revealed that *FN3KRP* and *FN3KRPL2* expression, including several transcript variants for the latter, is uniformly low throughout the brain, including nuclei of the song system (data not shown). In contrast, there was no evidence that songbird-specific *FN3KRPL1* is transcriptionally active in the brain.

With *URB1*, I found further evidence of expressional divergence. RNA-seq showed the parent gene to be expressed widely in embryo, liver, muscle, skin, and testes; whereas expression of songbird duplicate *URB1L3* was only detected in skin (Table 1). *In situ* hybridization revealed uniform brain expression of the parent gene and duplicate copies (Figure 6A). However, due to cross-alignment of probes, I cannot unequivocally assign cDNAs from the duplicate copies to a specific locus. At higher resolution, the labelling of both parent and duplicate genes is cellular, but rather than displaying the cytoplasmic pattern typical of most mRNAs, expression is concentrated within nuclei, consistent with nucleolar localization (Figure 6B).

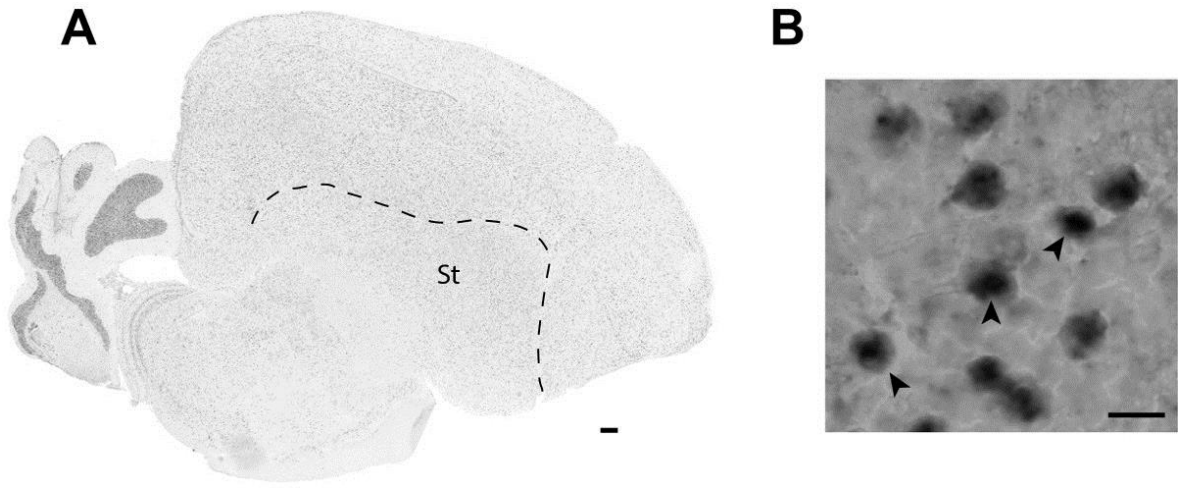

Figure 6. Expression of URB1 in the adult male Zebra finch brain. (A) Photomicrograph of *in situ* hybridization showing uniform expression of URB1 in the pallium. (B) High-power view reveals URB1 enrichment in individual pallial neurons. Note that in several of the cells indicated by the arrowheads the intracellular labelling appears more robust in the nucleus than in the surrounding cytoplasm, forming a pattern reminiscent of a fried-egg. Scale bars: 10 µm.

In the case of *CASC1-1* and *CASC1-2*, RNA-seq revealed shared expression in embryo and testes, with *CASC1-1* also expressed in spleen and *CASC1-2* detected in liver and muscle (Table 1). I identified two Zebra finch brain cDNA clones, FE723736 and DV948439, associated with *CASC1-1* and *CASC1-2*. FE723736 aligns completely (98.1% identity) to exons 2-5 of the 5-exon *CASC1-1*. It also aligns well (98.9% identity) to *CASC1-2*, but this alignment is partial, since the *CASC1-2* locus lacks the 100 bp 4th exon predicted in *CASC1-1* and present in FE723736. Thus, FE723736 is transcribed from *CASC1-1*, indicating that this gene is unequivocally expressed in the brain. In contrast, Zebra finch clone DV948439 aligns completely with high scores to both loci, thus we cannot establish from which locus it is transcribed. Because we cannot exclude the possibility that both clones are transcript variants from *CASC1-1*, we cannot conclusively establish whether *CASC1-2* is transcriptionally active in the brain. Since I used high stringency hybridization conditions, and these clones are relatively small, I predicted that I might detect differences in their brain distributions, if present. Indeed, these clones displayed partially overlapping but distinct patterns. DV948439 revealed strong labelling throughout the brain, including the pallium, thalamus, and both granular and Purkinje cell layers of the cerebellum (Figure 7A). The distribution and density of labelled cells in the pallium was uniform, but the relative level of expression varied from cell to cell (Figure 7B,C, left panels). Fiber tracts and white matter were devoid of signal, suggesting the probe is detecting transcripts that are not expressed in glia, but cells in walls of the lateral ventricles were strongly labelled (Figure 7D, left panel). In contrast, FE723736 revealed specific expression in the cells that define the ventricular wall (Figure 7B,D, right panels), and in large neurons within the globus pallidus (Figure 7C, right panel).

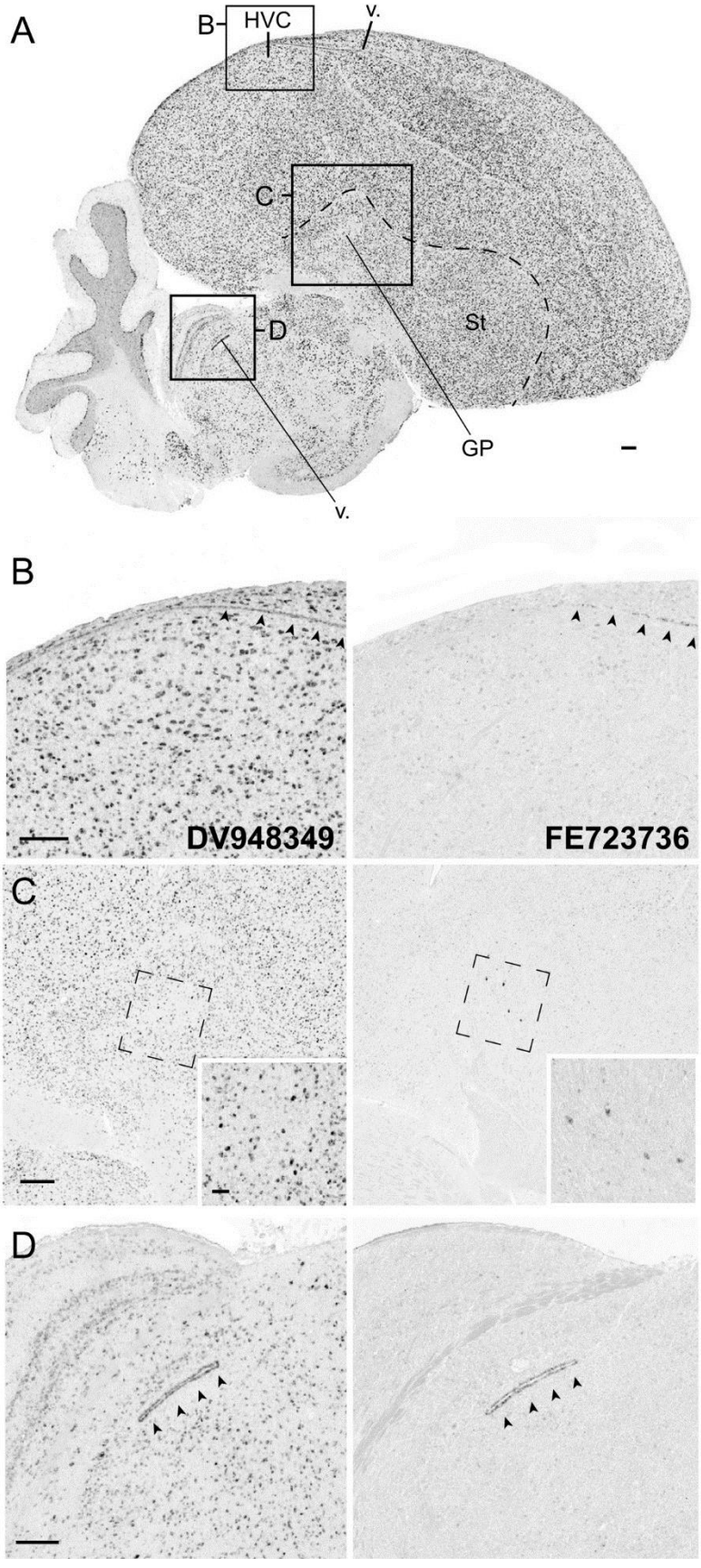

Figure 7. Differential expression of CASC1 genes in song nucleus LMAN of the adult male Zebra finch. (A) Photomicrograph of an in situ hybridization conducted with a probe (DV948439) that is not locus specific reveals expression of CASC1-1 and/or CASC1-2 throughout the brain (See Additional file 1: Figure S3 for details). The approximate positions for the photomicrographs shown in panels B-D are depicted by the black squares. (B-D) CASC1-1/2 mRNA is highly expressed in song nucleus HVC (B, left panel), the globus pallidus (C, left), and ependymal cells of the lateral ventricle (D, left; arrowheads). In contrast, CASC1-1, revealed by a probe that is specific to this locus, is differentially expressed in ependymal cells of the lateral ventricle (B, right panel), large likely GABAergic cells in globus pallidus (C; right), and ependymal cells of the fourth ventricle in the midbrain (D, right). The dashed rectangles in B indicate the approximate positions of the high-power photomicrographs depicting labelled cells in the globus pallidus. Anatomical abbreviations: HVC, proper name; GP, Globus pallidus; St, striatum; v., lateral ventricle. Scale bars: 100 µm.

The truncated *YTHDC2* parent gene lacks any EST or RNA-seq evidence of expression, supporting the conclusion that it is functionally inactive in songbirds (Table 1). I identified several Zebra finch brain cDNA clones that align to varying degrees to the 7 *YTHDC2*-related genes. Based on alignment scores and the presence of unique exons, I was able to unequivocally assign two clones to specific copies of *YTHDC2*: CK309358 aligns with 99.9% identity to songbird-unique *YTHDC2L1*, including a 3' exon that is not present in any of the other *YTHDC2*-related loci, whereas DV946054 aligns with 100% identity to Zebra finch-unique copy *YTHDC2L5*, including a first exon that is not present at any other loci. Thus, both *YTHDC2L1* and *YTHDC2L5* are transcriptionally active. *In situ* hybridization reveals that *YTHDC2L5* is expressed at low levels throughout the brain (Figure 8A,B), with RNA-seq data revealing expression in all other tissues examined (Table 1). In contrast, songbird copy *YTHDC2L1* appears to be exclusively expressed in LMAN (Figure 8C), a brain nucleus within the AFP and critical for song learning and vocal variability [93, 94]. Strongly labelled cells are uniformly distributed in LMAN, displaying cellular labelling that is diffuse in the cytoplasm and strong in the nucleus. The somata of labelled cells are large, consistent with neuronal identity. The nuclear signal appears to label distinct foci, suggesting that I may be detecting sites of active transcription (Figure 8D).

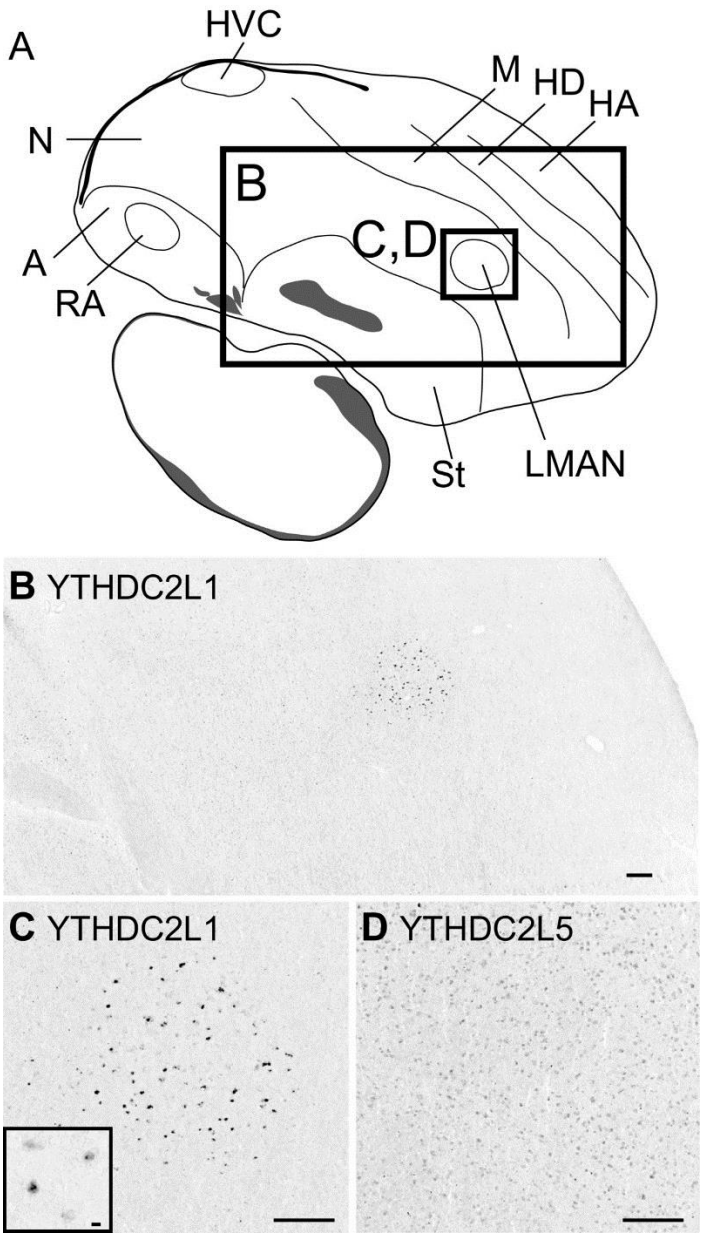

Figure 8. Differential expression of YTHDC2L1 in song nucleus LMAN of the adult male Zebra finch brain. (A) A schematic depicting a sagittal brain section ~2 mm from the midline shows the approximate location of the in situ photomicrographs presented in panel B. (B) Photomicrograph of in situ hybridization of songbird novel gene YTHDC2L1 shows discrete expression in song nucleus LMAN. (C) Detailed view of this section reveals that expression of YTHDC2L1 is restricted to large cells of LMAN, with labelled foci evident in some cells pairs within cellular nuclei (inset). (D) A comparable view of paralogous gene YTHDC2L5 shows low levels of expression, non-differential in LMAN. Anatomical abbreviations: A, arcopallium; H, hyperpallium; HVC, proper name; LMAN, lateral magnocellular nucleus of the nidopallium; M, mesopallium; MD, dorsal mesopallium; MV, ventral mesopallium; N, nidopallium; RA, robust nucleus of the arcopallium; St, striatum. Scale bars: 100 µm in B–D; 20 µm in C inset.

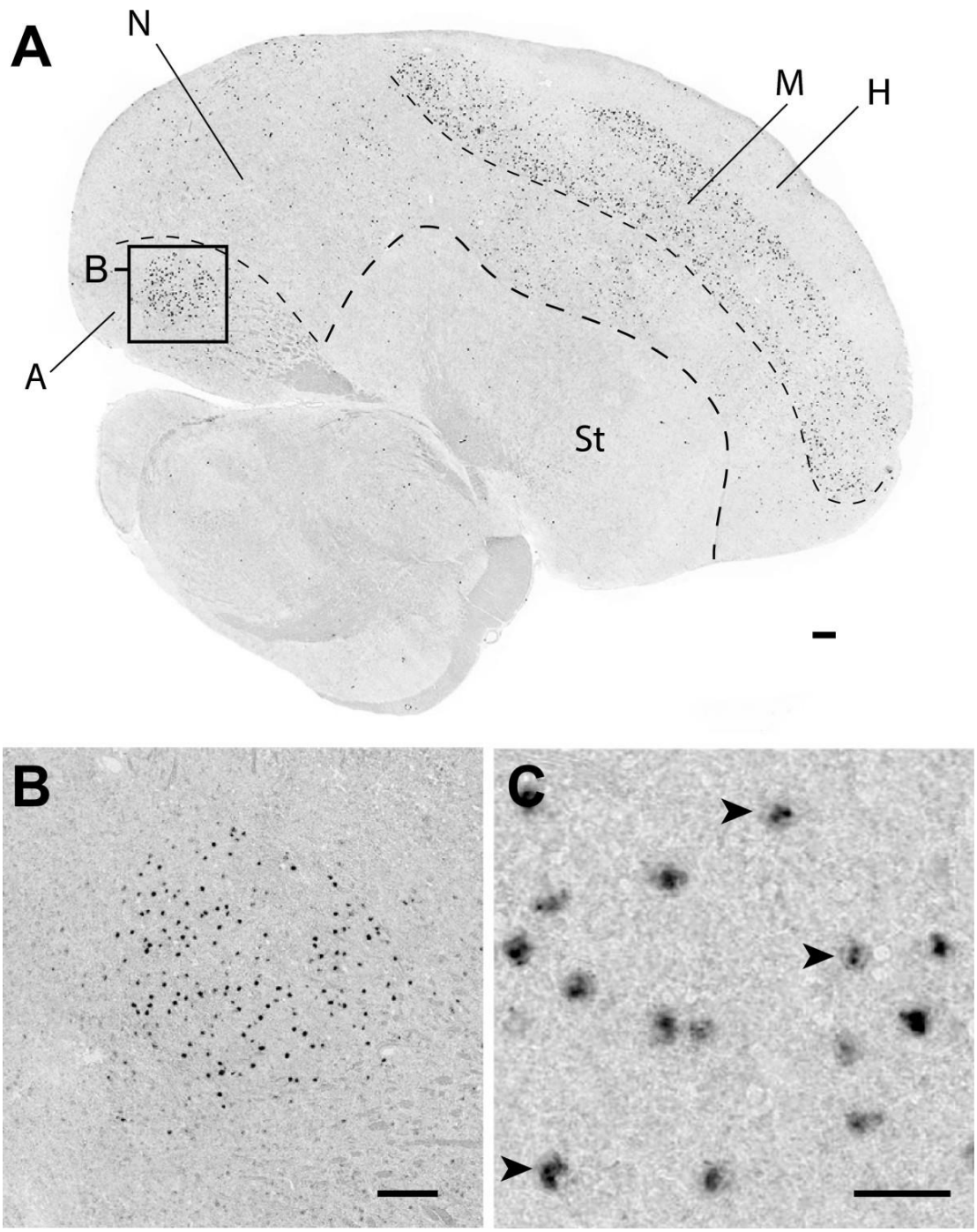

Figure 9. Differential expression of TMRA in song nucleus RA of the adult male Zebra finch. (A) Photomicrograph of *in situ* hybridization showing the distribution of TMRA expressing cells in a parasagittal brain section that includes song nucleus RA (~2.0 mm from the midline, position indicated by inset B). (B) High-power view reveals the enrichment of TMRA in individual cells within RA. (C) Detailed views reveal that in some cells in RA (indicated by the arrowheads), labelling is largely restricted to nuclei, and expression is low in the surrounding cytoplasm. Anatomical abbreviations: A, arcopallium; H, hyperpallium; M, mesopallium; N, nidopallium; St, striatum. Scale bars: 100 µm in A and B; 20 µm in C.

For the *de novo* gene *TMRA*, RNA-seq data detected wide expression in embryo, liver, muscle, spleen, and testes; brain expression was shown by the presence of finch and junco cDNA clones which mapped specifically to this locus (Table 1). *TMRA* is sparsely expressed throughout the nidopallium, but enriched in the mesopallium (Figure 9A; mesopallium as recently defined in [151, 152]). Most strikingly, *TMRA* is a prominent marker of song nucleus RA (Figure 9A,B), a structure required for the production of learned song [59, 153]. Labelled cells have large somata, indicating neuronal identity (Figure 9C). As with *URB1* and *YTHDC2*, expression is enriched in cellular nuclei (Figure 9C; arrowheads), and in many cells labelling appears focal, perhaps indicating that the transcript is concentrated in nucleoli, or that I was detecting independent transcriptional sites.

Curating misannotated novel genes in songbirds

Many genes in the candidate set, including those reported in previous studies (Supplementary Table 3 in [123]), did not pass criteria for songbird-unique novel genes. I provide corrected annotations for these false positives, based on the results of the comparative BLAT alignments and syntenic analysis, subdividing them into several categories (Tables S.1-5).

A large set of apparent duplications consisted of typically large, multi-exonic genes where the Ensembl prediction failed to group all exons under the same model, resulting in two or more partial adjacent models annotated as duplicates or expansions of that ortholog (n = 203; Table S1). Such cases were easily identified by examining the BLAT alignment to the Zebra finch genome of a more complete ortholog from a different species such as chicken, mouse, or human (for example, Figure S3A in Additional file 2).

Another large group consisted of separate gene pairs or sets, often members of the same gene family, where one or more models were misidentified as a gene duplication or expansion. These false positive duplications were found in tandem on the same chromosome (n = 90, Table

S1.2A; for example, Figure S3B in Additional file 2), or on separate chromosomes (n = 59, Table S1.3A). In other cases, the models represent true duplicate pairs, but these expansions are also present in chicken and/or non-avian organisms, and thus they are not unique to songbirds (n = 97 adjacent models, Table S2B; and n = 68 models from separate chromosomes, Table S1.3B). Some of these are also present in lizard, frog, and/or fish, but not in mouse or human, thus they appear to be duplications that occurred in a distant vertebrate ancestor but were lost in mammals (e.g., *MRC1-1* and *MRC1-2*). Other cases appear to have originated in birds, as they also occur in chicken and/or turkey, but not in non-avian species (e.g., *TTR-1* and *TTR-2*).

Among Zebra finch genes misannotated as novel and without identified orthologs in other species, the vast majority consisted of short segments of known genes that were not incorporated into the main Ensembl model, likely due to sequence gaps or regions of low assembly quality (n = 154, Table S1.4A). A smaller subset consisted of previously uncharacterized genes, described as novel by Ensembl, but orthologous loci could be identified in chicken and/or other organisms (n = 64, Table S1.4B). Of note, some of the chicken loci in this subgroup did not have a predictive Ensembl model but could be found at the correct syntenic location through BLAT-alignment.

A small set of candidate novel Ensembl gene models (n = 25) was found to be part of massively expanded gene families (e.g., olfactory receptors, keratins, zinc fingers). As their curation would require extensive further analysis to establish correct orthology, these were not studied further. Finally, a subset of candidate novel genes (n = 28) displayed a loss of intronic regions and had flanking repetitive elements (e.g., LTRs, LINEs, SINEs) characteristic of retrotransposon-mediated duplication. These genes will be further reported elsewhere.

Our phylogenetic searches revealed some gene duplications present in Zebra finches but in no other species (n = 32). While these could represent real Zebra finch-specific features, in most cases the two genes in a duplicated pair occur in tandem, often in regions of poor sequence

quality, flanked by gaps, and tend to have very high (>95) percent identities. They likely represent assembly artifacts, due to improper placement of allelic variants or misassembly of repetitive sequences. Our lab is currently generating further sequence and a new genome assembly for the Zebra finch, and intend to further examine this subset elsewhere.

A distinct subset of novel Zebra finch genes were also present in the Medium ground finch but not in the crow or in any other bird; they thus appear to be specific to the finch lineage (n = 17, Table S1.5A). In other cases, the genes were present in all passerines (3 songbirds, manakin, and Rifleman) but in no other birds; they thus represent features shared among Passeriformes (n = 16, Table S1.5B). Yet other cases could be found in numerous avian species, but not chicken; these likely originated early in the radiation of Neoaves, or represent genetic losses specific to Galliformes (n = 7, Table S1.5C).

Discussion

My analysis identified with high certainty several songbird-unique genes, and drastically reduced the number of genes misannotated as novel by automated detection algorithms [154, 155]. In addition to significantly extending and improving upon previous lists of songbird novel genes [123, 145], this effort demonstrates the necessity of a systematic curation pipeline that incorporated synteny analysis in order to accurately predict gene identity, and establishes a template for using comparative genomics to identify novel genes in any genome. The pipeline is particularly effective in eliminating false positive novel gene annotations by identifying orthologs undetected by automated gene prediction algorithms due to gene model incompleteness. This effort also illustrates how analysis of a large number of genomes can enable the discovery of genomic features unique to specific groups and possibly associated with group-specific traits, a strategy that will become increasingly feasible as larger collections of genomes from other animal groups become available.

Although in this study I have focused on the contribution of novel genes to the evolution of the songbird lineage; other factors are likely to have played a role as well. Such factors include differential gene substitution rates, chromosomal rearrangements, retrotransposon-related events, and modification of regulatory regions. Several of these are being explored in companion papers to this study [33, 147].

Lineage-specific expansions have been reported in Zebra finch versus chicken [123], but the incorporation of 45 newly sequenced, high-coverage (30 – 120X) avian genomes as well as representative non-avian genomes allows us to identify the specific set of duplications that arose following the divergence of ancestral oscine passerines from their closest living relatives (i.e., suboscine passerines, ~32 mya), but before the songbird crown radiation which includes both finches and crows (~20 mya) [33]. I note that this represents a high-confidence set: if these genes were not unique to songbirds, and their absence in non-songbirds a consequence of incomplete sequencing, I would expect them to be randomly distributed across the 45 non-songbird genomes sequenced. Instead, I find them only in songbirds, and in none of the non-songbird species examined. Due to my strict criteria, the list of songbird-unique genes is likely an underestimate, as I have excluded loci not mapped to a known chromosome, due to the possibility that these represent alleles rather than actual paralogs, as well as gene sets annotated by Ensembl as one-to-many orthologs, which require further analysis to establish exact orthology. In addition to identifying genes uniquely present and shared among songbirds, this analysis also reveals further sets of lineage-specific genes which characterize finches, passerines, neoavian birds, or Galliformes. The identification of gene sets common to these avian clades represents a significant advance for identifying genomic innovations whose emergence may be linked to some of the characteristic traits of these groups.

Our identification of SDs that emerged following evolutionary divergence of Galliformes (e.g., chicken) and Neoaves (e.g., Zebra finch) substantially improves on previous studies [123, 149] by refining the location of SD sites, identifying breakpoints on chrs 11-28 and Z, and distinguishing SDs present in Zebra finch only, thus possibly specific to songbirds, from those present in chicken only, and thus possibly specific to Galliformes. The fact that the majority of novel genes, both those unique to songbirds as well as those present in other avian groups, are located within or immediately adjacent to SDs suggests that chromosomal rearrangement is a major mechanism for the emergence of novel genomic features in passerines and other avian groups, as found in other lineages [156, 157]. This corroborates previous reports establishing non-allelic homologous recombination following inter- or intra-chromosomal rearrangement as an essential mechanism for genome evolution [158, 159]. Of note, songbird chromosomes 1 and 1A are known to have undergone significant rearrangement, having split from chicken chromosome 1 [160], which my findings suggest has had repercussions for novel gene evolution: a single rearrangement on Chr1A can be associated with multiple novel genes; another SD on Chr1 harbors the previously described growth hormone gene duplication (*GHL*) [161], which I conclude to be passerine-specific. Related to these rearrangements, avian SDs have been previously associated with a high occurrence of repetitive elements [149], which are thought to provide a substrate for non-allelic homologous recombination and genomic instability [162, 163]. It is also possible that some of the genes reported here, as well as others containing repetitive elements, may have arisen through retrotransposon-mediated duplication mechanisms, such as 5' transduction [164].

A major outcome of this study was the discovery that most of the novel songbird genes are transcriptionally active, with both EST and RNA-seq data supporting differential tissue expression of parent genes and songbird novel loci, indicating a diversification of function following duplication. Compellingly, some novel genes' brain expression indicates an association

with songbirds' neural system for learned vocalizations, suggesting that their evolution could be related to the emergence of this characteristic songbird trait. Although other studies have identified gene enrichments in song nuclei [126, 165, 166], this is the first report that genes found only within the songbird lineage are transcriptionally active in these nuclei. This suggests that certain novel genomic features of songbirds may have evolved to support the function of the circuitry dedicated to vocal learning behavior. For other novel genes, transcriptional evidence suggests that their function is associated with other, non-neuronal tissues, including skin, muscle, liver, and testis. Although more targeted experimental approaches involving gene manipulations will be required to establish the exact functions of songbird novel genes, I discuss potential implications in the context of their predicted protein domain architecture and selective tissue expression.

The multiple copies of *A4GALT* are complete duplications of the parent gene, thus this gene expansion might represent an increase in molecular function, or a diversity of functions if accompanied by divergent expression patterns. Although I found no evidence of expression of these genes, in other organisms *A4GALT* has been implicated in glycosylation of surface antigens related to the P blood group system, indicating that this gene and its expansion are likely related to organs and systems outside of the range of tissues explored in these analyses [167].

Several other gene expansions (*RIOK2*, *RNF4*, *URB1*, *HYDIN*) are predicted to encode much shorter proteins that lack specific domains compared to the parent genes, sometimes even lacking a recognizable ORF. For these genes, there are indications of differential expression in non-brain tissues by RNA-seq, although I cannot unequivocally demonstrate brain expression due to cross-alignment of cDNAs to multiple loci (Table 1). These truncated genes could act as partial competitive inhibitors of the parent gene, as seen with the human specific duplication of the *SRGAP2* gene, which in turn causes slower brain development in humans relative to other

mammals [143]. It is also possible that these might represent pseudogenes resulting from a complete duplication followed by a degradation of the coding sequence and loss of transcriptional activity of one paralog. A notable exception was *FN3KRP*, a gene related to deglycation of proteins and thus possibly protective against hyperglycemia [168]. *FN3KRPL2*, which is shared by all passerines and complete in terms of coding domains, has even gained complexity in the form of multiple 3'UTR variants. I also note that in some cases I found evidence of brain expression of the parent gene only (e.g., *RIOK2* and *RNF4*, both low to undetectable by *in situ* but with associated cDNAs) or of the parent gene and its duplicates (*FN3KRP*, *URB1*). While the patterns were broad and uniform, thus uninformative with regards to regional specializations, they establish a link to basic, non-specialized aspects of brain function. Intriguingly, I note that two songbird novel genes exhibit exclusive expression in skin (*RIOK2L*, *URB1L3*), a finding without clear precedent that points to potential unexplored molecular specializations of songbirds.

There is significant sequence divergence between *CASC1-1* and *CASC1-2*, which, along with the differential expression patterns detected, suggests a divergence of molecular function. The restricted expression of the *CASC1-1* in the ventricular zone is intriguing, given that the parent gene is related to the control of cell proliferation [169], and the subventricular region adjacent to the ventricles is a site of continued proliferation of neuronal precursor cells in adulthood [170, 171]. The other transcript is broadly expressed but cannot be unambiguously linked to either paralog. One possible interpretation is that the two paralogs have very distinct expression patterns. In this regard, the chromosomal rearrangement that gave rise to this duplication (Figure 3) likely disrupted the regulatory promoter of *CASC1-1*, leading to differences in expression patterns. Alternatively, the two transcripts analyzed might be variants of *CASC1-1*, with *CASC1-2* representing a pseudogene. In either scenario, further studies of the *CASC1* duplication are worth

pursuing, and analysis of additional genes in the proximity of genomic rearrangement sites could lead to further insights into evolving patterns of gene regulation in the avian brain.

The *YTHDC2* gene expansion clearly illustrates a divergence of function across paralogs, which differ in both structure and expression. *YTHDC2* is predicted to encode a protein capable of binding to RNA (through its YTH domain [172]) and inducing conformational changes (through its RNA helicase activity [173]). Although the functionally complete copy *YTHDC2L5* is expressed broadly in the brain and other tissues, the songbird-unique copy *YTHDC2L1* is expressed solely in song learning nucleus LMAN in the brain, and is highly truncated relative to the parent gene, retaining only the HA2 and OB-fold domains associated with RNA helicase activity regulation [174]. This indicates a neofunctionalization of songbird paralog *YTHDC2L1*, and suggests that it may play a role in RNA regulation in LMAN. This conclusion that will await targeted experimental confirmation.

It is unclear how the complete *TMRA* arose in songbirds, but given the presence of a short, exonic segment in the correct syntenic position in falcon and trogon, two clades recently shown to be closely related to songbirds [33], *TMRA* appears to represent a *de novo* gene gain in songbirds with ancestral non-coding origins, as shown previously for genes which originated *de novo* in human from non-coding sequences in chimp [175]. *TMRA* is a remarkable marker of song nucleus RA, suggesting a role related to the neural coding of learned vocalizations, as RA represents the cortical output for vocal-motor control and is essential for the production of learned vocalizations [59, 153]. *TMRA* is a member of the *CLEC* family of transmembrane protein genes, with a function likely associated with cell surface recognition processes required for cell-cell and/or cell-substrate interactions [176]. Although one cannot exclude the possibility that *TMRA* may play a role in response to pathogens, as occurs for other members of the lectin family [177], I note that members of some gene superfamilies related to immune system function (e.g., *N-CAM*) also play

major roles in the nervous system, modulating cell-cell adhesion and interactions with extracellular matrix that are critical for neural development and function [178].

The discovery of novel songbird genes expressed specifically in the vocal control system provides evidence that some molecular specializations unique to this group may be associated with vocal learning. This trait evolved in three avian lineages (songbirds, parrots, and hummingbirds), all possessing dedicated circuits for this behavior with marked similarities in their neuroanatomical organization [20, 56, 57]. Given these parallels, one might expect convergent similarities in the molecular organization of these circuits. Recent evidence supports a much closer relationship between parrots and passerines than previously recognized [32, 33], leading to the intriguing possibility that some molecular specializations of their vocal learning circuits may have evolved in a common ancestor. Indeed, recent studies have identified a number of shared molecular specializations in analogous vocal control nuclei across avian vocal learners [166]. My demonstration that songbird novel genes have been incorporated into their unique vocal control nuclei suggests that in addition, these lineages also possess unique molecular specializations related to their particular vocal learning circuits. These specializations could relate to neuronal populations and connections unique to songbirds. For example, the songbird direct pre-vocal motor cortical projection to the basal ganglia, HVC-to-Area X, is absent in parrots [179], and possibly also in hummingbirds (Mello et al., unpublished). Alternatively, the roles played by songbird-unique genes could be subserved by functionally analogous genes in parrots and hummingbirds. Further study of avian vocal learners may reveal further group-specific specializations, as well as shared molecular features that may represent fundamental requirements for vocal learning.

2.3 Methods

Curation of Ensembl models

We manually curated three distinct categories of gene models predicted in the genome of the Zebra finch (taegut3.2.4) by Ensembl's Genebuild pipeline (e59; [155]). The first set consisted of the set of Ensembl models annotated as Uncharacterized Proteins with a status of 'Novel' in Zebra finch. I identified this set by retrieving the complete set of protein-coding models in Zebra finch from Ensembl BioMart (ensembl.org/biomart/martview), and excluding all models annotated as "Known," models with orthologs in other species, and models placed onto 'chromosome Unknown,' which are thought to largely represent allelic variants. The second set consisted of all genes identified by Ensembl as constituting Zebra finch-specific duplications based on the presence of a hyphen and a number following the gene symbol (e.g., *CASC1-1*, *CASC1-2*). I note that many of these genes' status have changed to 'novel' or 'uncharacterized' protein in the latest release (e75). The third set included in this analysis consisted of models previously identified as belonging to expanded gene families in the initial description of the Zebra finch genome (see Supplementary Figure 3 in [123]). Of note, I did not re-analyze the expanded gene sets for *PAK3* and *PIM1* that had been previously characterized [145].

To curate all three sets of genes and thus obtain a set of putative novel, duplicated, or expanded loci, I used the following steps: 1) I retrieved the complete nucleotide and protein sequences for each predicted model from Ensembl Biomart. 2) I BLAT-aligned [180] each model sequence to the genome assemblies of Zebra finch (taeGut1); two galliform genomes, chicken (galGal4) and turkey (melGal1); and five non-avian genomes of each major vertebrate lineage, namely a lizard (anoCar2), frog (xenTro3), Zebrafish (danRer7), mouse (mm10), and human (hg19) utilizing the UCSC genomic browser (genome.ucsc.edu, [181]). 3) I manually examined all high scoring (>50) hits within their genomic context, taking into account sequence quality and genome

assembly gaps as well as comparing alignments to any existing annotations at these loci (e.g., human proteins mapped by chained tBLASTn, refSeqs from other species, and expressed sequence tags (ESTs)). This allowed us to identify some models as being artifactual due to misalignment or redundant alignment to known loci, excluding these from further analysis. In some cases, I also identified previously unrecognized paralogs for which no Ensembl model is currently available. These were added to my candidate novel gene set for further analysis. 4) To confirm the identity of each non-artifactual BLAT alignment of the models, I conducted a syntenic analysis in the UCSC browser, comparing the genes flanking each hit (at least three genes upstream and downstream) in Zebra finch to all other species of interest. This allowed us to exclude models whose syntenic placement revealed them to be known genes, including hits to known paralogs and related gene family members. I also excluded models for which I identified orthologs in species aside from Zebra finch, making note where possible of the "parent" gene orthologous between chicken and Zebra finch which may have been duplicated in songbirds to give rise to the novel genes examined further in this study (for example, see Figures 3,4A). This approach allowed us to correctly annotate models based on a combination of sequence identity and synteny. It is also highly sensitive in discriminating paralogs from related gene family members, and in detecting additional loci not currently predicted by an Ensembl model. I have contacted Ensembl for incorporation of these corrected gene annotations into a future Zebra finch genome annotation release.

Identification of songbird-unique genes

We next used a BLAST resource developed by BGI (phybirds.genomics.org.cn) to search for evidence of the models representing candidate novel songbird genes resulting from my curation effort (see preceding paragraph) in 45 new avian genomes (described in [33, 147], the Budgerigar genome used is further described in [182]). This includes basal ratites, galloanseriformes, and a

range of shorebirds and landbirds, as well as other vocal learning groups (e.g., parrots, hummingbirds) and their sister taxa (falcons, swifts). I determined the number of hits of each model to each of these species, and identified models where passerines possessed additional gene copies not present in other avian species. This resource provides only the BLAST hits themselves with no genomic context, which makes it impossible to separate true hits from alleles, as well as to establish orthology among hits in different species. In order to address these limitations and to precisely identify the set of novel genes which arose in the songbird lineage, I examined the alignment and synteny of these Zebra finch models in genomes of critical comparative relevance to my goals, namely: two songbirds, Medium ground finch and American crow; two non-songbird passerines, Golden-collared manakin, a suboscine, and Rifleman, a New Zealand wren; and the nearest non-passerine relative, Budgerigar. To accomplish this task, I generated provisional annotations of these genomes by BLAT-aligning the complete Zebra finch Ensembl model set using a standalone server-based BLAT implementation with parameters replicating the web-based UCSC browser. I then BLAT-aligned the putative novel gene set with more sensitive parameters (e.g., allowing for more mismatches, returning lower-scoring alignments) in order to ensure that any trace of these models would be detected. I also BLAT-aligned the set of orthologous parent models from Zebra finch and chicken (in which predicted models are often more complete), in order to detect novel genes where the Zebra finch copy may not cross-align well to the other species due to genetic divergence. Finally, I imported these genomes and all described BLAT alignments into the Integrated Genomics Browser (IGV, [183]), a server-side alternative to the web-based UCSC browser. This procedure allowed us to examine the alignments of the candidate songbird-specific novel genes within the syntenic context of these genomes, enabling us to distinguish orthologs of known genes from novel genes. By identifying which novel loci were present in each species, I was also able to pinpoint their phylogenetic origin. I confirmed the

accuracy of this method for establishing phylogenetic placement by replicating the results in the web-based UCSC browser for two species, Medium ground finch (geoFor1) and Budgerigar (melUnd1), which became available on the UCSC site midway through the study. No discrepancies were found between the two methods.

Detection of Zebra finch chromosomal rearrangements by alignment of homologous synteny blocks in chicken and Zebra finch

To identify novel genes and/or duplications that might be associated with chromosomal rearrangements in Zebra finch, I used a previously established genomics approach [149] to first identify any breaks in gene synteny by analyzing the order of appearance of orthologous gene pairs in the genomes of chicken and Zebra finch. To accomplish this, I used Ensembl BioMart [184] to retrieve the complete set of 11,132 genes from Ensembl that have been predicted to be of type "ortholog_one-to-one" in both Zebra finch (taeGut3.2.4) and Chicken (WASHUC2, May 2006). I only included genes with known physical locations in both genomic assemblies. I then used SyntenyTracker [148] in Orthologous Gene Pair mode with default settings (i.e., distance between markers 1 Mb, block size 0 bp, block length 2 Mb, jumping distance 2 Mb, reference genome ''Zebra finch,'' target genome ''chicken'') to identify Homologous Synteny Blocks (HSBs) for each orthologous chromosome pair. Here I define HSBs as a continuous block of two or more adjacent homologous genes that appear without interruption, and on the same chromosomes in the two species being compared. To verify the results of SyntenyTracker, and further refine the breakpoint analysis, I also aligned the entire set of orthologous gene pairs according to their relative position in the Zebra finch genomic assembly, and then scanned the alignments to search for cases where two or more genes appeared with the correct gene order, but in different positions on the chromosome (e.g., translocation), or where the gene order was reversed (e.g., inversion). I

retrieved from Biomart a subset of Zebra finch genes, which included known and novel protein coding genes, as well as non-coding and pseudogenes (e.g., miRNAs, snoRNAs) that have no known orthologs in chicken. Paired orthologous and non-orthologous genes were then sorted according to their relative chromosomal positions within the Zebra finch assembly to identify genes lying within a predicted chromosomal "breakpoint", or gap in genomic sequence between two consecutive HSBs. To determine which of the syntenic rearrangements between Zebra finch and chicken had occurred in the songbird or the chicken lineage, I also compared the syntenic regions flanking the SD sites to mammalian species that have well-assembled and curated genomes (i.e., mouse and humans) as well as to lizard when the corresponding regions were well-assembled.

Protein coding domain motif analysis

To determine whether novel and duplicated genes contain or might be missing specific protein coding domains that specify their possible molecular function, I analyzed each novel gene's protein coding sequences with Interproscan5 [185] using default search parameters and all available protein domain definitions (e.g., SUPERFAMILY, SMART, PfamA, PROSITE). For genes that had no Ensembl model, I analyzed the most complete open reading frame prediction available in the songbird genomes. I paid particularly close attention to the duplicated genes, since any differences between these and their parent genes would suggest a possible divergence in protein function. Schematic representations of predicted protein domain structures were designed using Prosite MyDomains (prosite.expasy.org/mydomains/).

Analysis of songbird novel gene expression

To explore the expression of genes of interest, I searched for Zebra finch brain-derived cDNA clones from ESTIMA [127], SongbirdTranscriptome.net [125], or the Rockefeller database [126], as well as clones derived from several additional songbird tissues [150], that were aligned to novel gene loci in the UCSC genome browser. For cases where the novel gene represented a songbird duplication or expansion I searched for cDNAs for both the expanded copies and suspected ortholog. To maximize the likelihood that the resulting probe would be specific to a given locus, I selected clones containing primarily 3'-untranslated sequence, minimizing the inclusion of protein coding regions that might be conserved among close paralogs and/or related gene family members. In the case of TMRA, I also fully sequenced the cDNA clone to define the 3'-end of each clone and establish the presence of a polyadenylation tag (polyA). To directly test for clone specificity, I BLAT-aligned the complete nucleotide sequence for each clone to the Zebra finch genome, and analyzed the resulting hits. In cases where a clone aligned to multiple loci, I attempted to identify the locus that generated that clone by determining the BLAT query that produced the largest alignment (e.g., number of exons) with the highest percent identity to the aligned region, as well as identifying any locus-specific segments in the clones. Additional confirmation of songbird novel gene expression was obtained by analyzing previously published Zebra finch RNA-seq data from several tissues; spleen, muscle, skin, liver, testes, and whole embryo [123]; available through NCBI with BioProject accession code PRJNA17289. Only reads that mapped to unique locations, and which overlapped with novel gene exons were considered.

Brain preparation for *in situ* hybridization

This study used 10 adult male Zebra finches that were bred and housed at OHSU in accordance with IACUC guidelines. Birds were moved into sound-attenuated chambers the evening before

sacrifice and monitored for 1 hour following lights-on to ensure that they were non-singing to reduce variation due to hearing- or singing-induced gene regulation in song nuclei and adjacent regions [58, 186]. Following decapitation, brains were removed and flash frozen in Tissue-tek embedding medium on a dry ice-isopropanol slurry in under 5 minutes to ensure sufficient RNA quality. Brains were cut into 10 μm sections on a Leica CM1850 cryostat, placed onto glass slides, fixed for 5 minutes in a 3% paraformaldehyde fixative solution, and stored at -80°C until further use.

Analysis of brain expression by *in situ* hybridization

We followed a previously established protocol for optimized detection of gene expression through non-radioactive *in situ* hybridization [128, 187]. Briefly, for each brain-derived cDNA clone, I generated digoxygenin-labelled sense and antisense riboprobes, performed *in situ* hybridization using high-stringency hybridization and wash conditions (~65°C, low salinity, high pH), and detected cellular labelling by immunohistochemical detection with alkaline phosphatase precipitation. Replicates were run for each probe on at least two adjacent sections (n = 3 brains) for all of the major song nuclei. Resulting high-quality sections were then imaged for digital analysis at 0.42 μm/pixel with an Olympus Nanozoomer. For each nucleus I assessed relative level of brain expression based on visual assessment and a scoring scale from low to high. Genes that I found to be expressed in song nuclei were qualitatively analyzed for enrichment (or impoverishment) relative to surrounds (e.g., RA vs. arcopallial shelf) by at least two independent observers.

3 ‘Core sets’ of genes expressed in vocal learning circuits

3.1 Background

At present, of the three lineages of birds known to learn their vocalizations—songbirds, parrots, and hummingbirds—only the songbird vocal learning circuit has been extensively characterized in terms of its anatomy, physiology, and molecular features (Figures 1 and 10). Although parrot and hummingbirds have been demonstrated to have specialized brain nuclei for vocal learning analogous to songbird HVC, RA, and Area X, it is unclear whether the known properties of these songbird nuclei (e.g., interconnections, cell types, and expressed transcripts) are shared more broadly among vocal learning birds. The identification of convergent properties in these independently evolved structures would point to fundamental molecular properties of vocal learning circuits, lending unique insight into the neurobiology and evolution of a complex, learned motor behavior.

Vocal pathways in parrots are comparable to those of songbirds (Figure 10). NLC and AAC constitute a direct vocal-motor pathway from nidopallium to arcopallium to the brainstem vocal nuclei, analogous to songbird HVC and RA, respectively, while the anterior MST and NAOM plus thalamic DMM are organized into a cortico-basal ganglia-thalamo-cortical loop, analogous to the songbird anterior forebrain pathway [188-192]. As in songbirds, singing-induced expression of the activity-dependent gene EGR1 (aka ZENK) has confirmed the involvement of these nuclei in vocalizations [56]. Budgerigars have projections that are absent in songbirds (DMm to direct pathway nuclei, AAc-to-NAom, NAom-to-NLc) and lack an equivalent of the songbird HVC-to-X projection (i.e., their NLC does not appear to contain NLC-to-MSt projection neurons).

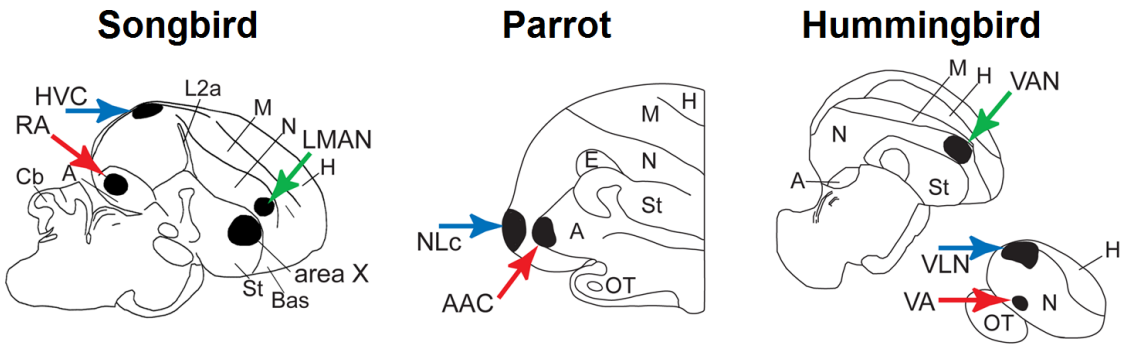

Figure 10. Analogous song-activated brain nuclei in avian vocal learners. Schematic images depicting the neuroanatomical locations of vocal nuclei in songbird (zebra finch, in sagittal), parrot (Budgerigar, in frontal) hummingbird (Anna's hummingbird, in sagittal) brain sections. Colors indicate nuclei proposed to be analogous in function and approximate circuit location. Abbreviations, general brain regions: H, hyperpallium; M, mesopallium; N, nidopallium; A, arcopallium; St, striatum; Cb, cerebellum; Bas, nucleus basalis; OT, optic tectum. Abbreviations, songbird vocal nuclei: Area X, proper name; HVC, proper name, LMAN, lateral magnocellular nucleus of the anterior nidopallium; RA, robust nucleus of the arcopallium. Abbreviations, parrot vocal nuclei: NLC, central nucleus of the lateral nidopallium; AAC, central nucleus of the anterior arcopallium. Abbreviations, hummingbird vocal nuclei: VA, vocal nucleus of the arcopallium; VAN, vocal nucleus of the anterior nidopallium; VLN, vocal nucleus of the lateral nidopallium.

Using singing-induced EGR1 expression, vocal control areas were identified in the Somber hummingbird (*Aphantochroa cirrochloris*) and the Rufous-breasted hermit (*Glaucis hirsuta*), which both possess complex acoustic repertoires indicative of vocal learning [57, 193]. Comparable nuclei are seen in other hummingbirds, indicating their widespread occurrence [57, 64]. They resemble the vocal nuclei of songbirds and parrots in cytoarchitectonics and can be grouped into anterior (VSt, VAN) and posterior-lateral (VLN, VA) sets (Figure 10). Previous work has suggested that VAN and VA (presumed hummingbird LMAN and RA) project respectively to VA and to brainstem areas involved in vocal-motor and respiratory control, indicating the presence of conserved features in connectivity [64]. However, these observations are preliminary (based on DiI in fixed tissue) and connections of VLN and VSt (presumed hummingbird HVC and Area X) were not studied. Thus, the extent to which the circuitry of hummingbird vocal nuclei more closely resembles that of songbird or parrots remains an open question, pending further tract-tracing experiments.

In addition to their anatomical similarities, the identification of shared molecular and electrophysiological traits among avian vocal learning nuclei was suggestive of broader convergence in the neurobiology of vocal learning. *In vivo* electrophysiological recordings in NLC, the premotor vocal nucleus of parrots, revealed the presence of auditory responses, suggesting convergence at the physiological level with songbird HVC, which is also characterized by neurons activated by both auditory perception and vocal production [194, 195]. Early molecular studies identified intriguing similarities and differences in neurotransmitter receptors, neuropeptides, and sex hormone-related genes across avian vocal learners [64, 196-203]. However, the scope of these studies was limited, pending a broader survey of molecular convergence and divergence among avian vocal learners.

To address the possibility of convergence in the molecular properties of vocal learning circuits, I sought to examine whether the analogous vocal control nuclei of songbirds, parrots, and hummingbirds also share molecular markers. I define molecular markers as genes that are differentially transcribed in select brain nuclei relative to their surrounding tissue, and that thus might subserve the unique anatomical and physiological properties of those nuclei. For my choice of model species, I focused on comparison of zebra finch (*Taeniopygia guttata*), Budgerigar (*Melopsittacus undulatus*), and Anna's hummingbird (*Calypte anna*). These species represent the ideal comparative models, as their vocal behavior and song nuclei have been more extensively characterized than any other species in their clades. In zebra finches, vocal learning occurs during an early critical period and results in a stable song that endures throughout life [91, 204-207]. This is in contrast to open-ended learners like canaries and North American sparrows, which modify their song seasonally [208, 209]. In Budgerigars, the best studied parrot, song is a long, complex, multi-syllabic warble that is variable across bouts [210-213]. Budgerigars are also capable of imitating the vocalizations of other species, including humans. When they are deafened or raised

in acoustic isolation, the resulting song lacks the frequency spectrum and breadth of elements found in normal song [214]. Anna's hummingbird is a confirmed vocal learning hummingbird. Its song contains consecutive phrases, each a series of syllables with distinct spectro-temporal elements [215, 216] resembling the song of other hummingbirds in complexity [217, 218], but less stereotyped than in some oscines (e.g., zebra finch). Anna's hummingbirds do not appear capable of the improvisation or adult learning typical of open-ended learners. Regional dialects consist of variants in syllable spectral composition, relative position, and repetition rates [219, 220]. When raised in isolation, Anna's produce abnormal song that lacks species-specific traits in syllable composition, frequency range, and phrase structure [216]. In other hummingbirds, evidence for learning is based on individual and regional variability [221-224], or complexity in spectral and syntactic organization [193].

3.2 Results & Discussion

In order to be comprehensive in assessing the degree of molecular similarity among vocal learners, I chose to focus specifically on the pallial ('cortex-like') nuclei of the vocal motor pathway (HVC and RA and their analogs), as these nuclei are critical for song production and their molecular and electrophysiological characteristics have been well-characterized in zebra finch [100, 225, 226].

Shared gene expression in songbird HVC and hummingbird VLN revealed by microarray

Our lab has previously identified neurochemical specializations of zebra finch HVC by comparing differential expression between HVC and an adjacent nidopallial region, revealing 400+ genes to be downregulated or enriched markers of this nucleus relative to surrounding tissue [165]. These results were obtained using a microarray spotted with 19,213 individual cDNAs corresponding to approximately 17,214 unique genes [127]. Our collective efforts subsequently confirmed

differential HVC expression in songbirds of 239 of these genes through an optimized, high-throughput *in situ* hybridization (ISH) protocol [187]. Results deemed by stringent criteria to be of high quality were subsequently scanned at microscopic, subcellular resolution. As part of our lab's effort to compile a high-resolution zebra finch brain expression atlas resource to further songbird research, I processed these images for clarity, uploaded them to a central database, and helped to develop user-friendly, publicly available website (zebrafinchatlas.org). Final gene image sets (*n* = 447) were exhaustively annotated by myself and others in order to gain a comprehensive insight into differential gene expression in the songbird brain, as well as establish a significance cutoff threshold for the microarray differential expression analysis of HVC. Using this cutoff, I established that approximately 800 genes from the microarray could be expected to represent true differential HVC markers in zebra finch, nearly doubling the number from our previously reported analysis of the dataset [165]. This provided a thoroughly verified reference point for my subsequent comparative analyses.

In parallel, a differential microarray analysis of Anna's hummingbird VLN versus an adjacent nidopallial shelf had been run previously by our lab using an identical cDNA microarray as was used to profile zebra finch HVC. A total of 959 cDNA spots were determined to be differential between VLN and surrounding tissue at a false discovery rate $p < 0.001$. I then cross-referenced this list against the verified HVC versus surrounding tissue microarray (1247 cDNAs) in order to identify markers that showed significant differential expression in both HVC and VLN (209 cDNAs, corresponding to 176 genes). Surprisingly, 148 / 176 (84%) genes were up- or downregulated in the same direction in finch and hummingbird (Figure 11, Table S2).

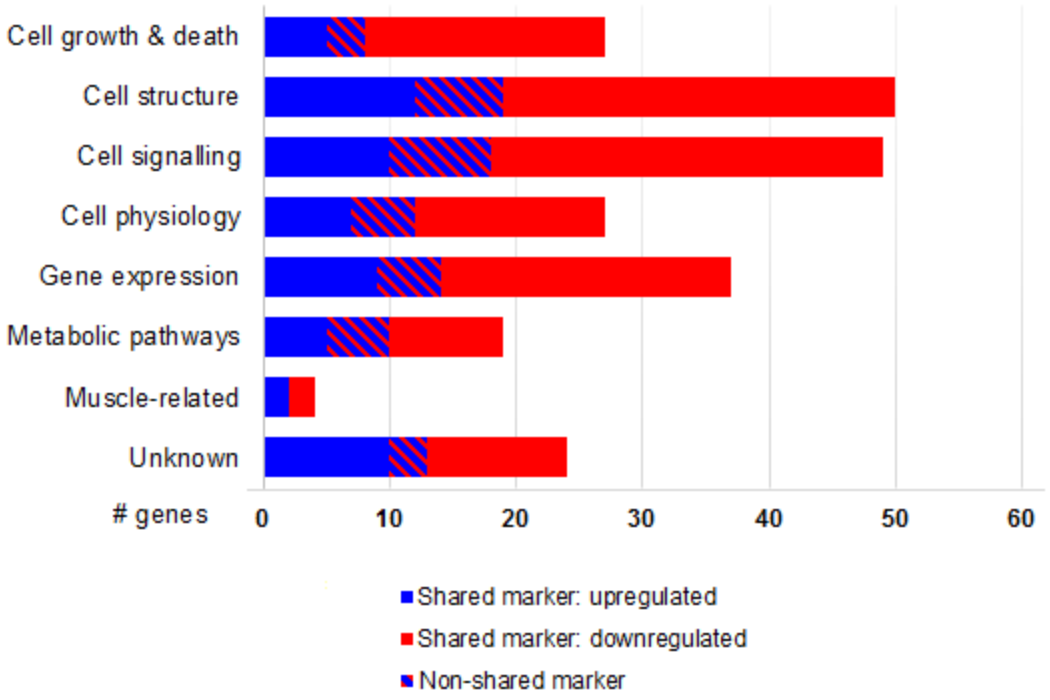

Figure 11. Conserved differential gene markers of zebra finch HVC and Anna's hummingbird VLN, organized by functional category. A total of 148 / 176 (84%) of markers show conservation in the direction of gene regulation in nucleus relative to surrounding tissue (59 shared upregulation, 120 shared downregulation, 30 opposite regulation). Complete results listed in Table S2.

Among the top functional categories for shared differential markers were cell structure (50 genes), cellular signaling (49 genes), and cell electrophysiology (27 genes), consistent with the expected properties of a highly interconnected forebrain nucleus. A large number of genes were associated with gene expression (37 genes), suggesting that the molecular identity of these nuclei could be shaped in part by a critical set of constitutively expressed transcriptional regulators, a possibility explored in more depth in Chapter 4. An additional 27 genes were involved in cell growth, migration, and apoptotic processes, potentially providing a molecular substrate for the adult neurogenesis and replacement observed in HVC-RA projection neurons [98], and suggesting that hummingbird VLN could contain a cell population sharing similar properties.

Shared gene expression in songbird, parrot, and hummingbird vocal nuclei revealed through ISH

In order to verify the shared HVC / VLN expression data identified in the microarray, assess whether this shared expression also extended to Budgerigar, and to identify the level of conserved expression between RA and its analogs in parrot and hummingbird, I performed extensive comparative ISH analyses.

Beyond those markers identified in the HVC / VLN microarray study, additional markers were selected in conjunction with a larger collaboration I participated in, with the goal of identifying transcriptional specializations of forebrain vocal nuclei that were not only shared by avian vocal learners, but also humans. For this analysis, additional microarray data were collected from a large number of brain regions (all major vocal nuclei and their adjacent tissues, all major forebrain divisions, as well as further midbrain hindbrain structures), from zebra finch, budgerigar, and Anna's hummingbird, as well as from two vocal non-learning birds (dove and quail), in addition to a large transcriptional dataset from different cortical layers of 12 regions in a non-human primate (Rhesus macaque). The zebra finch microarray data were then mapped onto the microarray results obtained from 4 human subjects for 231 discrete brain subdivisions, performing an 'all-against-all' comparison of region-specific transcriptional profiles, in order to determine which songbird and human brain regions shared the largest degree of transcriptional similarity. In addition to broadly conserved regions in the brainstem, these results confirmed the exceptional shared molecular specializations between mammalian cortex and avian pallium. Remarkably, these results also identified that the region most similar to songbird RA was the analogous structure in humans, primary laryngeal motor cortex (LMC). Further exceptional similarity was established between Area X and a subregion within human putamen (dorsal striatum). HVC and LMAN did not align well to any human brain regions sampled. Finally, shared specializations of RA / LMC and Area X / putamen were compared to the differential expression patterns observed in

analogous regions in the array data from vocal learning and non-learning birds and the macaque, in order to identify transcriptional specializations of vocal nuclei shared by all four human and avian vocal learning species, but not in the control regions of the three vocal non-learning species (for additional details, see [166]). However, this represented an unprecedented analysis of convergent gene expression both in scope and design, and as such it was unclear how to establish an acceptable cutoff for significance. To this end, I attempted to confirm several of these predicted expression specializations through an extensive comparative ISH analysis in zebra finch, budgerigar, and Anna's hummingbird.

I selected probes for ISH based on the quality of their sequence alignment to the Budgerigar and Anna's hummingbird genomes, which became available early in the study through our participation in a large effort to sequence, assemble, and annotate the genomes of 45 avian species [33, 147]. ISH was performed on brain sections containing vocal nuclei as well as major brain subdivisions and nuclei of all 3 species. For each gene, ISH conditions were optimized per species by adjusting the hybridization stringency through successive rounds of ISH (see Methods).

In sum, I was able to obtain multi-species ISH data for 18 genes (Tables 2 and 3). For the HVC / VLN gene set (Table 2), I was able to obtain strong ISH signal for four genes in hummingbird, CRHBP, NEFM, PVALB, and ZEB2. In all four cases, the gene was confirmed to be a shared marker upregulated in both HVC and VLN. For two further genes from this set, CADPS2 and HPCAL1, the probe failed to cross-hybridize to hummingbird. ISHs for these genes were subsequently run against budgerigar, confirming shared HVC / NLC regulation in 4 / 4 genes tested (HPCAL1, NEFM, PVALB, ZEB2 – NLC was not assessed for CADPS2 or CRHBP). These results suggest that the array data accurately reflect the level of shared expression in these nuclei, and that these markers appear to also be shared in the analogous nucleus in budgerigar. Thus, it is possible that these genes represent fundamental molecular specializations of HVC-like, premotor vocal nuclei in birds.

Cross-species ISH were attempted for 3 additional markers of zebra finch HVC—ALDH1A2, NTS, and RGS4. In the case of ALDH1A2 and NTS, strong signal was obtained from all three species, revealing a lack of expression in parrot NLC and hummingbird VLN – thus, these genes represent exclusive markers of songbird HVC. RGS4 was a shared marker of HVC and NLC, but failed to hybridize to hummingbird. Thus, for this set, zebra finch HVC marker status alone was less successful in predicting shared regulation than the HVC / VLN microarray results.

Gene	Probe EST ID	Zebra finch	Budgerigar	Anna's hummingbird
		HVC	NLC (HVC-like)	VLN (HVC-like)
ALDH1A2	CK234972	yes (+)	no (non-differential)	no (non-differential)
CADPS2*	DV955943	yes (+)	no signal	no signal
CRHBP*	DV955207	yes (+)	no signal	yes (+)
HPCAL1*	CK306795	yes (–)	yes (–)	no signal
NEFM*	CK313443	yes (+)	yes (+)	yes (+)
NTS	CK302282	yes (+)	no (non-differential)	no (non-differential)
PVALB*	CK305573	yes (+)	yes (+)	yes (+)
RGS4	CK314393	yes (+)	yes (+)	no signal
TAC1	DV958953	no (non-differential)	yes (+)	no (non-differential)
ZEB2*	DV952827	yes (+)	yes (+)	yes (+)

Table 2. Markers of HVC-like nuclei revealed through comparative ISH analysis. Table cells indicate whether the gene is a shared marker of the nucleus or not, included the direction of differential regulation, up (+) or down (-), relative to surrounding tissue. (*) indicate genes identified as shared HVC / VLN markers by cDNA microarray analysis.

I found that markers identified through the comparative microarray analysis of zebra finch RA and human LMC were highly consistent with comparative ISH data (Table 3). I was able to collect and analyze high-quality zebra finch ISH data for 11 of these genes (C1QL3, GABRB3, GPM6A, LINGO1, NECAB2, NEUROD6, NOL4, PLXNC1, PVALB, RTN4R, SNCA; patterns available on zebrafinchatlas.org). I found that the finch pattern in RA confirmed the pattern from the comparative array analysis in all but one case, NOL4. In this case, the gene was predicted to be a downregulated marker of RA, but appeared to be downregulated in a larger portion of arcopallium, including but extending beyond RA. I was able to obtain comparative ISH data for 5 of

these genes (C1QL3, GPM6A, PVALB, RTN4R, SNCA). In all 5 cases, ISH data revealed the RA/LMC marker to be a marker of budgerigar AAC as well. In Anna's hummingbird, ISH data revealed shared differential regulation in VA for 2 genes (PVALB, SNCA). In all other ISHs attempted in hummingbird, the probe failed to cross hybridize. An additional four markers (GDA, NEFM, RGS4, UNC5A) were run based on early results from the comparative RA/LMC analysis. In all four cases the pattern in budgerigar AAC was consistent with the pattern in finch RA. In one case (NEFM), the pattern in Anna's hummingbird VA was consistent with RA; in all other cases the probe failed to hybridize.

The majority of the shared RA / LMC markers identified in the array study had decreased expression relative to the surrounding regions (n = 50, 91%). Further, several shared markers, including PLXNC1 and SLIT1, have known roles in axon guidance. This suggests that one of the critical properties shared by LMC/RA-like structures is their specialized connectivity, perhaps reflective of the fact that one of the defining characteristics of these nuclei is their direct corticospinal projection to hindbrain motor nuclei for vocalization and respiration, a trait not shared with non-learning species (see Introduction).

Gene	Probe EST ID	Zebra finch	Budgerigar	Anna's hummingbird
		RA	AAC (RA–like)	VA (RA–like)
C1QL3*	FE726945	yes (–)	yes (–)	no signal
GDA	FE735906	yes (–)	yes (–)	no signal
GPM6A*	CK307194	yes (–)	yes (–)	no signal
NEFM	CK313443	yes (+)	yes (+)	yes (+)
NOL4*	CK310402	no (non-differential)	not attempted	not attempted
PVALB*	CK305573	yes (+)	yes (+)	yes (+)
RGS4	CK314393	yes (+)	yes (+)	no signal
RTN4R*	CK314393	yes (–)	yes (–)	no signal
SNCA*	DV950418	yes (–)	yes (–)	yes (–)
TAC1	DV958953	no (non-differential)	yes (+)	no (non-differential)
UNC5A	DV956337	yes (–)	yes (–)	no signal

Table 3. Markers of RA-like nuclei revealed through comparative ISH analysis. Table cells indicate whether the gene is a shared marker of the nucleus or not, included the direction of differential regulation, up (+) or down (-) relative to surrounding tissue. (*) indicate genes identified as shared specializations of avian RA-like nuclei and human LMC through a large-scale comparative microarray analysis [166].

One gene in particular, PVALB, represented an exceptional marker of both HVC-like and RA-like nuclei in all three species (Figure 12). Its broad conservation as a marker has subsequently been confirmed by other groups [227, 228]. PVALB encodes the calcium binding protein parvalbumin, which is known to be associated with fast-spiking neurons and muscle cells, where it serves as a calcium buffer to protect fast-spiking cells from cytotoxicity [227]. Its broad conservation in multiple song nuclei in all vocal learners studied thus far suggests that fast spiking may be a general property of vocal learning circuits, and that PVALB might thus be an excellent first-choice marker for comparative molecular analysis of known and suspected vocal learners whose vocal control circuits have not thus far been identified, such as bats, cetaceans, and pinnipeds.

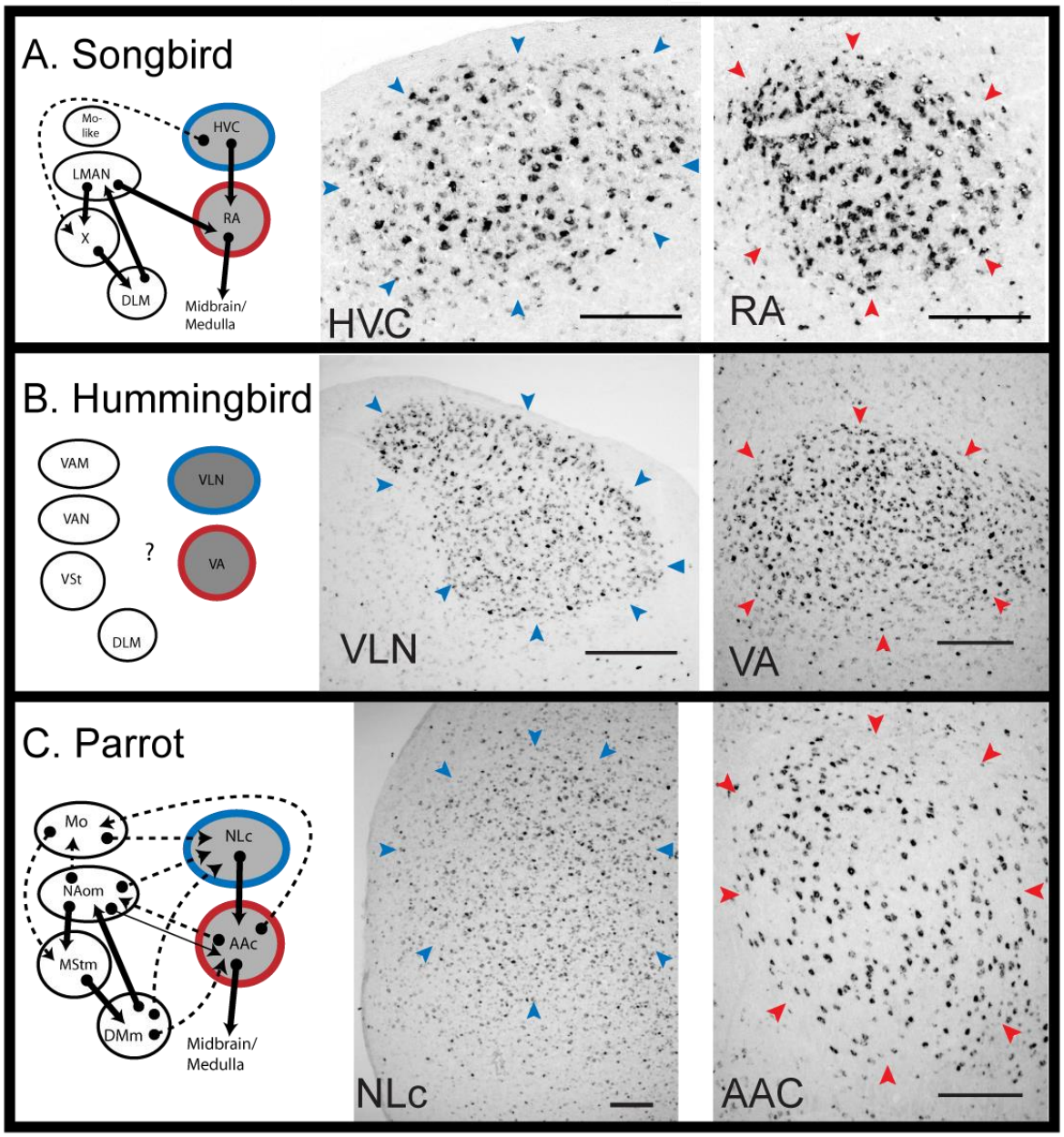

Figure 12. Comparative ISH of PVALB in HVC-like nuclei. In connectivity diagrams at left, schematics show major connections of vocal nuclei in zebra finch (A), Anna's hummingbird (B), and Budgerigar (C). Nuclei in the direct vocal-motor pathway are in gray and those in the anterior forebrain pathway are in white. For songbird and parrot, arrowheads indicate projections that are confirmed to be shared (thick), confirmed to be unique to the lineage (dashed lines), or suspected but unconfirmed (thin). In hummingbird, little is known of song system connectivity. Cross-species ISH, in center and right panels, show a convergent pattern of PVALB upregulation in HVC-like nuclei (center, blue arrowheads) and RA-like nuclei (right, red arrowheads). Scale bars: 250 µm.

The case for suboscines, the outgroup of songbirds, as potential vocal learners

The discovery of core sets of markers in HVC- and RA-like nuclei across vocal learners suggests that these gene specializations may represent fundamental molecular requirements for all vocal learning circuits. Based on phylogeny, it is clear that in humans and hummingbirds these represent unique evolutionary events, supporting the conclusion that they are convergent molecular specializations (Figure 1). However, the conclusion is less clear in the comparison of parrots and songbirds, as molecular and genomic data have demonstrated the groups to be closer relatives than previously suspected (Figures 1, 13). This is complicated by the fact that one lineage lies between them: the suboscine passerines, the nearest sister taxon of songbird (oscine passerines). Suboscine passerines had previously been shown to be vocal nonlearners, based on several results: 1) they fail to imitate the songs of closely related species in cross-fostering experiments, developing species typical song [229]; 2) they develop species-typical song even when deafened by total cochlear removal [45]; and 3) song nuclei are not revealed in the brains of suboscines by sex steroid markers of songbird nuclei [230, 231]. Concluding that suboscines are non-vocal learners, two possibilities emerge: either vocal learning evolved once in the psittacopasserae lineage —before the last common ancestor of songbirds, suboscines, and parrots—and was subsequently lost in suboscines, or vocal learning has evolved twice, independently in songbirds and parrots.

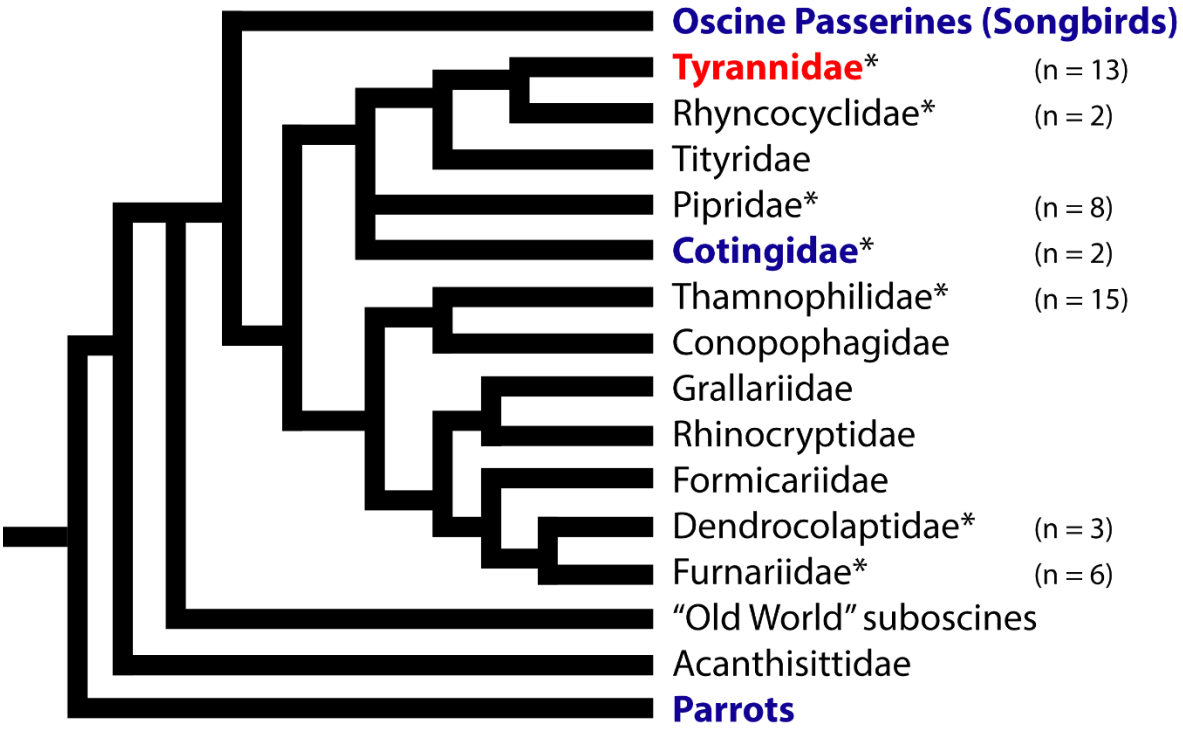

Figure 13. Schematic phylogeny of songbirds, parrots, and suboscine families. Phylogenetic relationships after Ohlson, et al. 2013 [232]. Lineages in blue possess learned vocal behavior, including suboscine bellbirds (Cotingidae). Tyrannid suboscines, in red, are demonstrated vocal nonlearners. All other suboscines, in black, have not been thoroughly investigated for learned vocal behavior. (*) indicates families from which Brazilian samples were collected, with numbers of species collected at right.

However, several important facts suggest that the conclusion that all suboscines are vocal nonlearners has been made prematurely. First, all previously mentioned studies of suboscines were performed on members of a single family, Tyrannidae (Figure 13, in red). New World suboscines (Tyrannides) alone are comprised of at least 12 separate families, in addition to 4 families of Old World suboscines (Eurylaimides), and one basal passeriform family (Acanthisittidae), altogether comprising approximately 1,000 species, one of the largest avian suborders [232]. Considering this, the conclusion that all suboscines are non-learning based on the study of three closely related species from just one of 17 global suboscine families, appears largely unsubstantiated.

Casting further doubt, recent behavioral work has demonstrated that bellbirds (*Procnias spp*.), suboscines of the family Cotingidae, appear to exhibit regional vocal dialects [233] as well as imitating the calls of other species [234]. It was further demonstrated that bellbird songs fall into three regional dialects, and that the dialect learned by any particular bellbird does not correlate with its genetic relatedness, suggesting further that bellbird song variants are not inherited genetically, but rather transmitted culturally, through learning [233]. Additionally, it has been noted that male manakins, of the family Pipridae, appear to develop coordinated duet songs over time for cooperative lekking displays [235]. Finally, in two further suboscine families, Thamnophilidae and Furnariidae, males and females perform complexly timed duets, requiring a high degree of coordination between partners [236-238]. Similar duetting behavior in songbirds has been shown to require heavy input from the song system [239]. Thus, it stands that one suboscine family is a confirmed vocal nonlearner, one family has been reported as a putative vocal learner, and three further families display complex vocal behavior suggestive of a song system.

As I am exceptionally interested in the evolution of the song system, I sought to perform an extensive survey of suboscine families to specifically address whether conserved markers of the song system could reveal the presence of vocal nuclei in any members of this group. Identifying which of the shared markers of songbird and parrot song nuclei also exhibited shared expression in suboscines would inform the question of whether these represent conserved specializations from a common ancestor (if shared), or convergent specializations evolved independent in songbirds and parrots (if not present in suboscines).

In order to survey as many suboscine families as possible, I developed a mobile brain dissection unit to collect fresh tissue samples in the field (see Methods). In sum, I collected 49 brain samples, representing 26 species from 7 suboscine families, comprising the largest collection of suboscine brain material in the world (Figure 13). I then selected genetic markers to test across

species, prioritizing genes that: i) were identified by my earlier analyses as shared song system markers in multiple vocal learners, ii) were markers of multiple song nuclei, iii) were markers of non-vocal neuroanatomical features (in order to discriminate potential song nuclei from these other structures), and iii) that had coding sequence-biased riboprobes available. For my initial study of the brain collection, I selected three species where tissue quality was high, covering the major branches of South American suboscine diversity: the Amazonian antshrike (Thamnophilidae: *Thamnophilus amazonicus*), the Helmeted pygmy tyrant (Tyrannidae: *Lophotriccus galeatus*), and the Screaming piha (Cotingidae: *Lipaugus vociferans*). This set was also selected because it both contained a member of the suboscine family that had been shown to be vocal non-learning (Tyrannidae) and a member of the family reported as vocal learning (Cotingidae).

In total, I was able to determine suboscine brain gene expression patterns for 13 genes via ISH analysis (Table 4). Of these, 7 were markers of songbird HVC (ALDH1A2, NEFM*, NTS, PVALB*, RGS4*, RHOB, SNCA), with three(*) of these representing genes identified as shared markers with parrot HVC-like nucleus, NLC. I did not detect differential expression consistent with an HVC-like nucleus in any of the three suboscine species.

I then tested 6 of the genes that I had shown to be shared markers of RA and its parrot equivalent, AAC (Table 4, genes marked by *). In all 6 cases, these genes were differentially expressed in an arcopallial nucleus in the suboscine tissues. Darkfield microscopy revealed the nucleus to be localized in a similar arcopallial location as RA, at the caudal extent of a densely myelinated fiber tract, the occipitomesencephalic tract, OM, although the nucleus appeared to be less densely myelinated than RA (Figure 14). Nissl staining revealed the nucleus to be comprised of a ball-like cluster of large cells, similar in size although more cytoarchitectonically diffuse than RA (Figure 14, bottom panels).

In the course of examining these expression patterns, I noticed shared markers of RA-like nuclei in parrots and songbirds were also correlated with differential expression in an additional arcopallial zone, lateral to RA, which I term 'Alat.' In songbirds, markers of Alat exhibit somewhat more diffuse expression borders than RA, which has starkly contrasted borders reminiscent of a bullet-hole in ISH images (Figure 15, right panels). Markers of the suboscine RA-like arcopallial nucleus also exhibited diffuse borders (Figure 15, left panels). In order to determine whether the suboscine nucleus more closely resembled RA or Alat, I performed several additional ISH experiments, including markers that were differentially expressed in RA only or Alat only (Table 4). In 14 of 16 (88%) comparative ISH experiments performed, the differential pattern of suboscine arcopallial domain was consistent with songbird Alat marker status, whereas it was consistent with songbird RA marker status in only 9 / 16 (56%) of experiments. Thus, the nucleus appears to share a higher molecular affinity with Alat than RA. However, the high degree of overlap between RA markers and Alat markers—as well as the high correlation of these RA+Alat markers with shared markers of songbird RA, parrot AAC, and humming VA—suggest that all of these nuclei could have derived from a common, ancestral nucleus, a hypothesis that will be expanded on more fully in Chapter 5.

Since these data were collected, some evidence has been published suggesting that the Eastern phoebe possess an RA-like forebrain nucleus that projects to the hindbrain vocal motor nuclei, and that bilateral lesion of this structure leads to subtle changes in song [240]. However, the equivalency of this structure with songbird RA was based on the shared differential expression of a single molecular marker, GRIK1. Importantly, no differential expression was detected in this 'RA-like' nucleus for PVALB, a marker I have identified as being widely shared in RA-like nuclei across avian vocal learners, as well as a marker of the suboscine Alat that I have identified (Figure

15). Thus, the resolution of the exact molecular identity of suboscine RA-like nuclei will depend on the further characterization of molecular markers in additional families.

		Zebra finch		Amazonian antshrike	Helmeted pygmy tyrant	Screaming piha
Gene	**Probe ID**	**RA**	**Alat**	**Arco nucleus**	**Arco nucleus**	**Arco nucleus**
SHARED: RA (-), ALAT (-)						
GDA*	FE735906	(–)	(–)	(–)		
GPM6A*	CK307194	(–)	(–)	(–)		
SNCA*	DV950418	(–)	(–)	(–)		
SHARED: RA (+), ALAT (+)						
CNTNAP2	DV948190	(+)	(+)	n.-d.	n.-d.	
NEFM*	CK313443	(+)	(+)	(+)		(+)
PVALB*	CK305573	(+)	(+)	(+)	(+)	
RGS4*	CK314393	(+)	(+)	(+)	(+)	
ALAT ONLY						
NTS	CK302282	n.-d.	(–)	(–)	(–)	(–)
RA ONLY						
LINGO1	CK315247	(–)	n.-d.	n.-d.		
RHOB	CK308435	(+)	n.-d.		n.-d.	
CONTROL (n.-d. in RA, Alat)						
ALDH1A2	CK234972	n.-d.	n.-d.	n.-d.	n.-d.	n.-d.
GAD2	CK311753	n.-d.	n.-d.	n.-d.	n.-d.	
TAC1	DV958953	n.-d.	n.-d.	n.-d.	n.-d.	n.-d.
TOTALS						
Suboscine ISH consistent with zebra finch **Alat** marker status:				8 / 9	4 / 5	2 / 2
Suboscine ISH consistent with zebra finch **RA** marker status:				6 / 9	2 / 5	1 / 2

Table 4. Comparative ISH of genes differentially expressed in arcopallial brain regions between zebra finch and three suboscine passerines; Amazonian antshrike (*Thamnophilus amazonicus*), Helmeted pygmy tyrant (*Lophotriccus galeatus*), and Screaming piha (*Lipaugus vociferans*). Table cells indicate whether gene is upregulated (+) or downregulated (–) in specified arcopallial nucleus relative to surrounding arcopallium, or non-differential. Blank cells indicate cases where ISH data are not yet available. Totals present number of experiments where differential expression in a suboscine arcopallial domain was consistent with expression in zebra finch RA or Alat out of the total number of gene markers tested in given species. (*) indicate genes identified as a shared marker of song nuclei across multiple vocal learners. Abbreviations: Alat, lateral domain of the arcopallium; Arco, arcopallium; n.-d., non-differential expression in nucleus; RA, robust nucleus of the arcopallium.

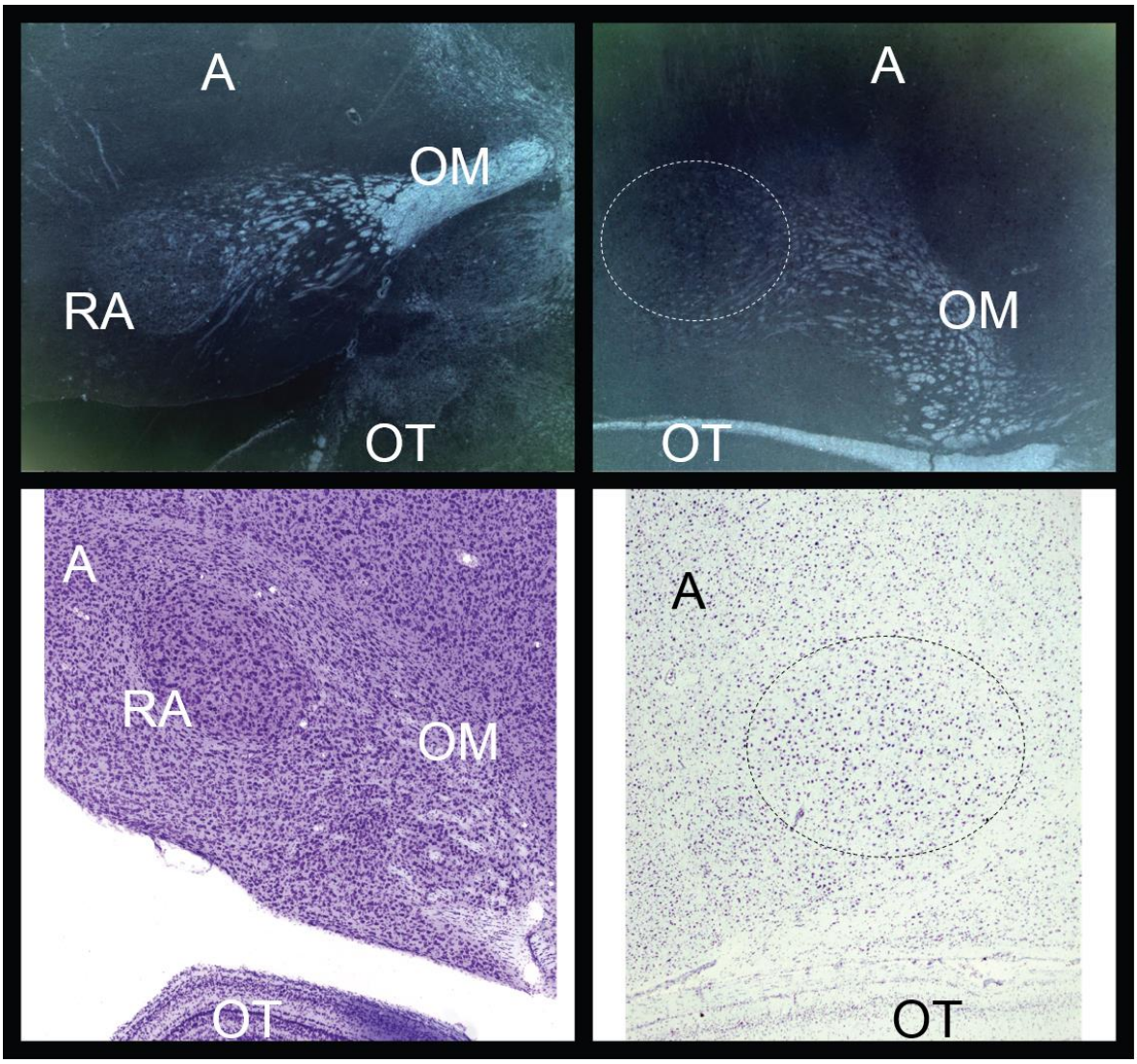

Figure 14. Comparative neuroanatomy of songbird RA (left panels) in zebra finch and a similarly located arcopallial nucleus (right panels, proposed boundaries of nucleus indicated by a dashed circle) in a suboscine, Amazonian antshrike (*Thamnophilus amazonicus*). Darkfield microscopy (top panels) reveals myelin patterns in arcopallium, including shared connection to OM. Nissl staining (bottom panels) reveal similar cytoarchitectonics between these nuclei, including larger cells than surround as well as a 'ball-like' organization. Abbreviations: A, arcopallium; OM, occipitomesencephalic tract, OT: optic tectum, RA: robust nucleus of the arcopallium.

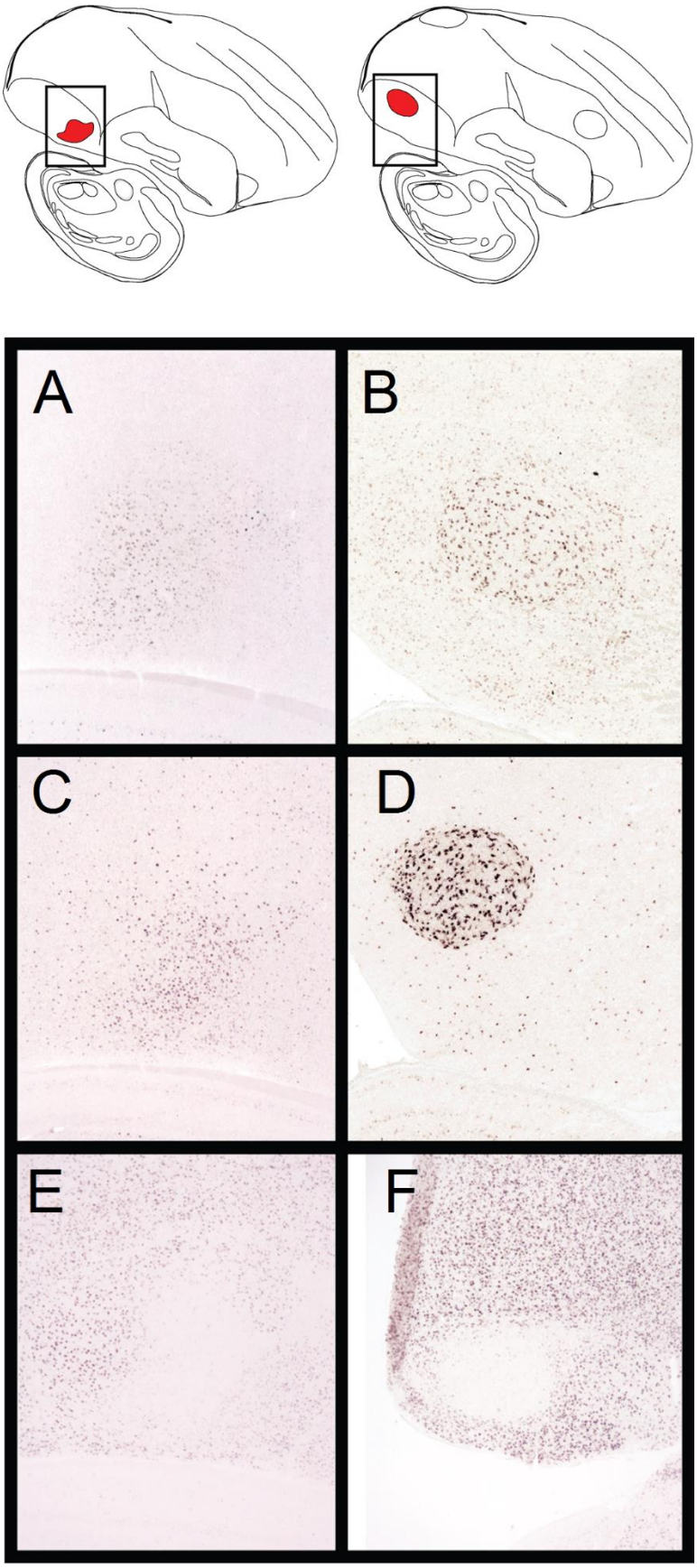

Figure 15. Schematics of sagittal brain sections and in situ hybridization data in Amazonian antshrike (Thamnophilus amazonicus), left column, and zebra finch, right column. Red indicates relative position of an arcopallial nucleus in the antshrike and RA in the finch. ISH images correspond to the square window. Two positive marker genes, NEFM (A,B) and PVALB (C,D) share upregulated expression in this nucleus, while a negative marker gene, SNCA (E-F), shares downregulated expression.

3.3 Methods

Tissue sample collection

For budgerigar and Anna's hummingbird samples, only male animals were selected, due to the robust sexual dimorphism in the zebra finch song system, and since vocal nuclei have not been verified in female hummingbirds. Budgerigars were purchased from a commercial supplier. Anna's hummingbirds were caught on the day of the experiment using playback of Anna's hummingbird song to incite territorial defense behavior in males, who were subsequently caught by a drop-net trap. Animals were humanely sacrificed through quick decapitation. All procedures were performed in accordance with approved IACUC protocol.

Collecting suboscine brain tissue samples of suitable quality for *in situ* hybridization presented a challenge. North American suboscines have been conclusively demonstrated to be vocal non-learning [43, 45], and South American suboscines have proven difficult to maintain in laboratory environments (Ana Amador, personal communication). To circumvent this obstacle, I developed a mobile dissection protocol, adjusting all necessary tools and reagents to fit in the space of a backpack. Species were confidently identified through the assistance of a local ornithological expert, and collected via small caliber rifle. Traditional mist-net capture was an inappropriate alternative, as many of the birds of interest, including the Screaming Piha, live almost exclusively in the high canopy, and other species of interest, including the Amazonian Antshrike, seldom emerge from dense undergrowth. Further, stress from capture could result in significant brain expression changes that would alter the results obtained via ISH. Brains were quickly dissected (< 5 min) using sterile tools and nitrile gloves and flash frozen in a mix of dry ice—which was contained in a small styrofoam box to limit sublimation—and isoflurane, in order to preserve RNA from degradation. Field collection was performed preferentially in early morning hours, comparable to our collection time for zebra finches.

Selection of clones

For all *in situ* hybridizations on avian brains, we used zebra finch clones. The probes are either derived from the Zebra finch ESTIMA cDNA clone collection (Songbird Neurogenomics Consortium), or through RT-PCR cloning (for EGR1). Unlike in same-species hybridizations, where 3'-biased clones are preferable for their specificity, in cross-species hybridizations I favored coding sequence clones, as gene coding regions are more likely to be conserved, resulting in stronger probe hybridization. Selection of clones was based on the quality and specificity of the BLAT alignment of the sequence to budgerigar and Anna's hummingbird genomes.

Cross-species *in situ* hybridizations

ISH were performed as in Chapter 2, with modifications to the protocol to increase the strength of cross-species probe hybridization. In brief, for each gene, digoxygenin-labeled antisense riboprobes were hybridized to 10 µm brain sections containing vocal nuclei as well as major brain subdivisions and nuclei of all 3 species. Zebra finch control slides were run in parallel with all cross-species hybridizations to ensure that failure of cross-hybridization could be discriminated from experimental failure for technical reasons. Probes were hybridized to sections at lower stringency (63°C) to increase the strength of probe cross-hybridization, followed by high stringency washes and probe immunohistochemical detection. Optimized conditions were established for each probe to preserve cellular signal while minimizing non-specific tissue staining, decreasing stringency (60°C) for probes that failed to cross-hybridize, or increasing stringency (65 - 67°C) for probes where non-specific background signal precluded clear interpretation of the brain expression pattern. As further specificity controls, I included hybridizations with no probe in all experiments, as well as positive control markers with known patterns (e.g., GAD2).

4 The role of genomic regulatory elements in the evolution of vocal learning

4.1 Background

Genes that are co-expressed in select tissues and cell types, or that are co-expressed under specific cellular or behavioral conditions, have been shown to be associated with genomic regulatory mechanisms that modulate their expression at the level of the genome [241]. This context-specific expression is largely controlled by the actions of transcription factors, which accomplish this largely through their binding to short (6 – 12 bp), specific DNA regions, known as transcription factor binding motifs, in order to increase or decrease the rate of target gene expression [242]. These motifs tend to be concentrated into clusters known as promoters, located immediately upstream of a gene's 5' transcription start site (TSS), or sites, as well as enhancers, located in intronic and intergenic regions. Identifying suites of genes whose promoters share enriched DNA binding motifs represents a highly informative method for finding genes under shared regulatory control, and which may be activated by and involved in similar biological processes [241].

The songbird song system, in addition to being well-characterized in terms of its anatomical and physiological properties, has been extensively profiled for differential expression of molecular markers [125-127, 165, 243, 244], thus representing a neurobiological system highly amenable to study through computational cis-regulatory analysis. However, few analyses of songbird transcriptional regulation have taken place. With the exception of one recent study that examined transcription factor binding in the context of singing-induced gene regulation [139], all studies of songbird cis-regulatory elements have examined relatively small numbers ($n < 20$) of promoters [245-249].

In this work, my approach was to perform a computational promoter analysis to identify transcription factor binding motifs associated with differential expression in the song system. By identifying such song system-enriched motifs, I sought to also identify sets of predicted target genes associated with each motif, as these could represent co-regulated functional pathways that underlie the unique properties of song system nuclei. By integrating innovative molecular and genomic techniques, I sought to critically assess the regulatory dynamics of gene expression in a circuit devoted to learned vocal behavior. This work improves our understanding of the genetic and neurobiological basis of vocal learning, the behavioral basis for human speech and language.

4.2 Results & Discussion

I sought to identify cis-regulatory motifs associated with the promoters of gene sets differentially expressed in the song system that could thus subserve the unique molecular properties of song system nuclei and cell types. For this purpose, my overall strategy was to: 1) develop an evidence-based promoter identification pipeline and database of brain-expressed promoters in zebra finch; 2) develop a promoter analysis pipeline in order to identify motifs enriched in the promoters of genes differentially regulated in RA versus surrounding arcopallium, or differentially regulated between HVC-RA and HVC-X projection neurons; 3) for each enriched motif, identify target genes on the basis of significant motif–promoter binding sites, analyzing target gene sets as potentially co-regulated gene networks that could subserve the functional specializations of the song system; and 4) identify whether comparable enrichments exist in orthologous gene promoters of chicken, in order to identify whether gene regulatory pathways used by the song system are uniquely present in songbirds or represent songbird-specific usage of broadly conserved regulatory pathways.

Promoters of brain-expressed genes in zebra finch

In order to identify the changes in regulatory sequence that may be associated with the co-expressed gene networks in vocal learning circuits, I developed a computational pipeline for promoter identification and analysis in the zebra finch. It has been demonstrated that the accuracy of *de novo* promoter predictions can be greatly improved by incorporating 5'-biased cDNA to accurately determine transcription start sites (TSSs) [250-252]. It is also well established that different tissues differentially use distinct TSSs, which can be several hundred kb apart [253]. Thus, the study of tissue-specific gene regulation depends not only on the correct identification of a TSS, but also that the TSS is active in the tissue of interest. With these issues in mind, I developed a robust pipeline for promoter identification that is designed to integrate expression data from a tissue of interest to identify true TSSs active in any specified tissue, condition, or behavioral state of interest.

This pipeline was applied to the zebra finch genome using all available brain-derived sequenced gene transcripts that would be expected to contain the complete 5' end. This included 6,153 published full-length zebra finch mRNA sequences, derived from NCBI, and 9,580 5' ESTs derived from a study that applied RIKEN 5'-cap-trapper methodology to brain-derived zebra finch cDNAs [125]. Using restrictive criteria, these were mapped onto the zebra finch genome, filtered to remove artifactual alignments and alignments where >100 bp of the 5' end of the transcript failed to align to the genome, and assigned to genes based on my previous annotation of the zebra finch genome (see Chapter 2). This analysis produced a set of 4,451 high-confidence TSSs with evidence of expression in the brain. To further support that these reflected true TSSs, I also performed an alignment of the complete chicken mRNA and RefSeq set from NCBI, where gene annotations are generally more complete than in zebra finch, in order to confirm that brain expression data accurately predicted 5'-most exons.

For all subsequent promoter analyses, I defined promoters as -500 bp upstream and +50 bp downstream of the TSS, except where explicitly stated otherwise. The study of a relatively short proximal promoter region was selected to maximize the potential for finding significant enrichments by reducing the motif signal-to-noise ratio. Hence, although proximal promoters may not contain all of the information required to precisely control transcription of individual genes in time and space during development (this level of regulatory control is known to be largely mediated through promoter–enhancer interactions [241]), analysis of promoters alone can generate meaningful models of transcriptional regulatory networks.

A portion (7.4% of cDNAs and 5.4% of 5' ESTs) of the expressed transcript sequences failed to fully align to the current finch assembly, or aligned incorrectly to a paralogous gene. Of the sequences that produced a high-scoring genomic alignment, a further 30% had to be removed as >100 bp of their 5' sequence ends failed to align, removing confidence in their ability to predict a true TSS. In most cases this was due to the presence of sequence gaps in the current zebra finch assembly (taeGut1). Thus, the number of promoters obtained stands to drastically improve as the pipeline is reapplied to a forthcoming zebra finch assembly derived from PacBio long-read sequence. Still, this conservative promoter set stands in contrast to the typical approach for promoter collection, which attempts to gather the largest promoter set possible by using all *in silico*–predicted gene models to generate TSS predictions. Although greatly improving the total number of promoters available, such an approach stands to add substantial noise to any analysis of promoters, as I found through my extensive genome annotation efforts that *in silico* gene model prediction often failed to predict noncoding 1st exons (and thus TSSs), which for some genes can be up to 1 Mb upstream of the first coding exon[254]. Further, the pipeline employed here takes into account only promoters with evidence of expression in the tissue of interest (brain), thus further improving the biological meaningfulness of the final output of analysis.

Generation and *in situ* validation of song nucleus gene expression sets

To study regulatory networks active in song nuclei, I started by compiling large sets of song nucleus marker genes, which were initially determined through analysis of ISH data generated from our extensive efforts to characterize brain gene expression patterns in zebra finch (the ZEBrA gene expression atlas, described in Chapter 3). The number of genes represented in ZEBrA is large (*n* = 500 as of writing) and, through our collective efforts, extensively annotated. I assessed the overlap between regional markers annotated in the ZEBrA gene atlas and my brain-expressed promoter set, and on average had promoters for approximately 50% of gene markers of the four major song nuclei, Area X, HVC, LMAN, and RA (see Table 5).

Song nucleus	Upregulated	Downregulated
Area X	23 / 47 (49%)	2 / 7 (29%)
HVC	62 / 103 (60%)	27 / 47 (57%)
LMAN	38 / 59 (64%)	25 / 42 (60%)
RA	28 / 44 (64%)	28 / 60 (47%)

Table 5. Promoters of song system marker genes with high quality *in situ* expression data. Number of available promoters are expressed as a percentage out of the total number of annotated markers in ZEBrA (zebrafinchatlas.org). Genes are upregulated or downregulated relative to surrounding pallial tissue.

In order to increase the number of song nucleus marker gene promoters, I reanalyzed the results of two large microarray studies profiling differential gene expression in HVC and RA. The first of these was an attempt to profile the cell type-specific expression of the two long-range projection neuron classes of HVC [243]. In this experiment, the authors injected retrograde tracer cholera toxin B (CTB) conjugated to AlexaFluor dyes into RA and Area X into three silent and four singing birds. Approximately 2,000 retrogradely labeled HVC-RA neurons and 1,500 HVC-X neurons were then collected through laser capture microdissection. The pooled cell type homogenates were run on custom microarray previously generated by printing 5,000 unique, sequence-verified cDNA clones onto glass slides [125]. A differential microarray screen was performed, contrasting

gene expression in one cell type versus the other. After extensive annotation and validation through qPCR, 150,350 spots successfully hybridized and a total of 1,318 spots (0.88% of all the total) met a two-fold cutoff for differential expression. From this dataset, a small number of differentially expressed genes were described [243, 255].

Our lab obtained the raw data from the array screen for differential HVC projection neuron transcription, applying a strict false discovery rate cutoff of $q < 0.05$, resulting in a final count of 167 cDNAs significantly upregulated in HVC-X relative to HVC-RA neurons, and 322 cDNAs upregulated in HVC-RA relative to HVC-X. Although this provided an important source of molecular data on these two important cell types, I discovered that many of the cDNA clones' annotated gene identities were inaccurate. I therefore performed a full reannotation of all differentially expressed transcripts, using genome-wide alignment of cDNAs followed by manual synteny verification, performed as described Chapter 2. This resulted in a curated set of 210 genes, corresponding to 139 putative HVC-RA markers and 71 putative HVC-X markers. By cross-listing these gene sets with my brain-expressed promoter set, I determined that promoters were available for 108 genes from the HVC-RA set, and 61 genes from the HVC-X set (Table S3). To further verify my brain-expressed promoter set, I performed a detailed inspection of all 210 gene loci, verifying the presence of a true TSS in the zebra finch genome by examining the alignment of the mRNA and 5'-sequenced EST set described previously, as well as a suite of RNA-seq data previously generated from zebra finch brains, testes, and whole embryos[125], and a set of h3k27ac chromatin immunoprecipitation sequencing (ChIP-seq) data, a histone mark indicating the presence of active cis-regulatory regions [139]. These manual annotation efforts confirmed all 170 predicted promoters, as well as identified an additional 4 HVC-RA and 1 HVC-X genes with evidence of a full promoter, establishing a 0% false positive rate and a 3% false negative rate for the automatic promoter retrieval pipeline (Table S3).

The clean separation of individual cell types from heterogeneous tissue is a technically challenging process, and it is known that laser capture microdissection can sometimes result in contaminated samples. Given the importance of having a highly accurate cell type-specific gene promoter set, I sought to confirm the conclusions of the HVC cell type microarray by assessing the brain expression patterns experimentally through ISH. I first assessed whether the differential expression patterns could be confirmed through fluorescently labeled ISH onto brain sections where cells had been retrogradely labeled from Area X or RA. Through our previous efforts, our lab had thus analyzed 8 of the putative HVC-RA markers and 5 of the HVC-X markers. In 100% of cases, these fluorescent ISH experiments were consistent with the patterns identified from the microarray analysis (Figure 16, Table S3). I then sought to analyze a further set through digoxygenin (DiG) ISH. Although this is insufficient to conclusively determine cell type-specific expression, it can be used to distinguish neuronal expression from expression in glial or endothelial cells on the basis of cytoarchitectonics. It also allows for an initial assessment of cellular specificity of labeling, as HVC-X markers tend to show a pattern of expression in large, sparse cells; whereas HVC-RA markers tend to be expressed in smaller, clustered cells (Figure 17). In total, 74 additional brain gene ISH patterns were obtained and analyzed to determine whether the pattern of expression in HVC was consistent with a neuronal identity, as well as with the DiG ISH pattern of markers confirmed through retrograde labeling. In 65 cases (88%), the ISH pattern confirmed neuronal gene expression in HVC (for example, see Figure 17). In 8 cases (11%), the ISH pattern revealed expression in glial cells. In one case, the ISH pattern revealed an absence of gene expression in HVC. This high rate of array–ISH consistency supports the accuracy of the results obtained in the microarray analysis, as well as contributes additional brain expression pattern information to support future analyses via their deposition in a public repository (ZEBrA).

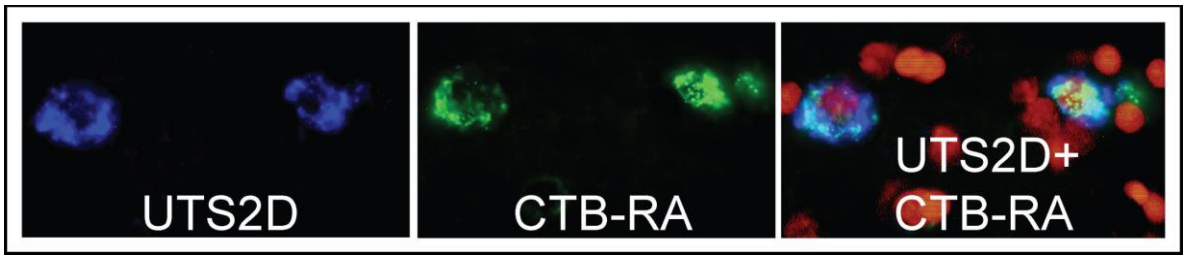

Figure 16. FISH of molecular specializations specifically expressed in HVC projection neurons. Fluorescence ISH of gene marker of HVC-X projection neurons (UTS2D, 1st panel, blue) on sections injected with with retrograde tracer (CTB) into RA (2nd panel, green) confirms predicted cell-type specific expression of this gene (3rd panel: merge of 1st and 2nd panels, plus propidium iodide stain in red to reveal nuclei). Abbrevs: CTB, choleratoxin B; HVC, proper name. UTS2D, Urotensin 2B

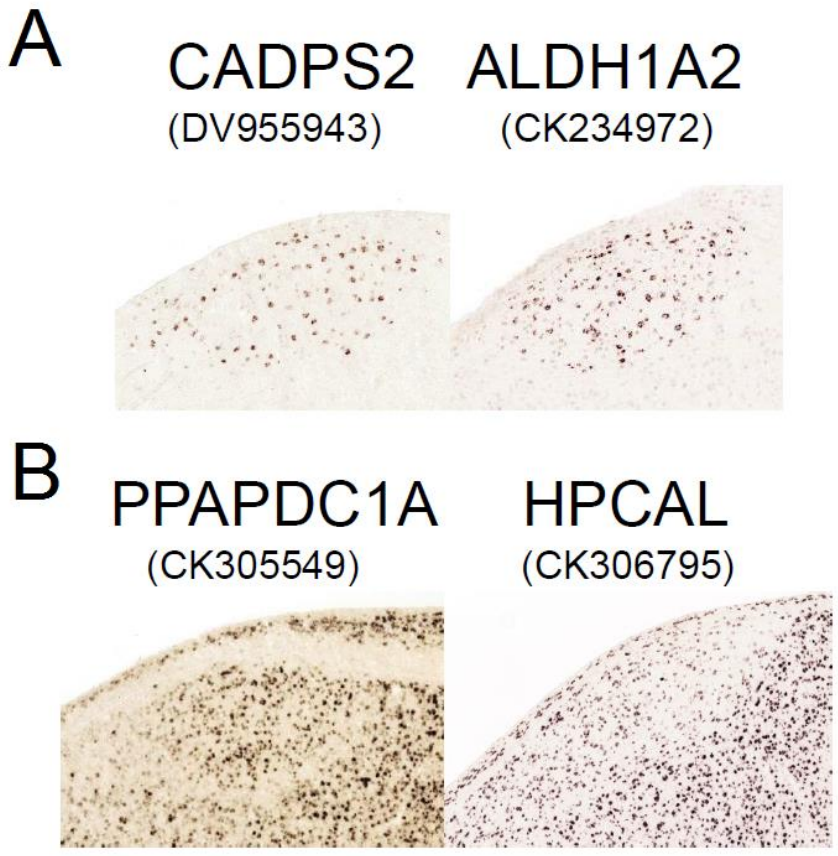

Figure 17. *In situ* hybridization of markers of HVC projection neurons. A) digoxigenin (DiG) ISH patterns of HVC-X neuron markers confirmed through FISH on sections retrogradely labeled from Area X and RA (ALDH1A2) are consistent with the DiG ISH patterns of additional differential HVC-X markers from the microarray study (CADPS2), showing a pattern of upregulation in large, sparse cells in HVC. In these cases expression in HVC as a whole is upregulated relative to surrounding nidopallial shelf. B) DiG ISH patterns of HVC-RA neuron markers confirmed through FISH on sections retrogradely labeled from Area X and RA (HPCAL1) are consistent with the DiG ISH patterns of additional differential HVC-RA markers from the microarray study (PPAPDC1A), showing a pattern of upregulation in small, clustered cells in HVC. In these cases expression in HVC as a whole is not upregulated relative to surrounding nidopallial shelf.

In addition to HVC, I sought to analyze promoters of genes differentially regulated in RA. For this purpose, I analyzed the results of a microarray study [139] assessing differential expression in RA relative to surrounding ventral arcopallial tissue, which had been run on an Agilent oligoarray containing 43,488 60-mer oligonucleotides sourced from zebra finch brain transcripts collected by three major research efforts [125-127], representing approximately 10,000 zebra finch genes. In order to confirm the gene identity of reported differentially expressed oligo probes, I aligned all probe sequences to the zebra finch assembly, removing those probes which failed to map or which could not be conclusively assigned to a single gene locus. This resulted in a final set of 521 genes downregulated in RA, and 349 genes upregulated in RA. Taking the intersection of this set and the brain-expressed promoter collection resulted in a final set of 138 positive RA markers (RA (+)), and 212 negative RA markers (RA (-)). ISH data corresponding to these sets were retrieved from ZEBrA, confirming the appropriate differential regulation 100% of RA (+) markers tested (*n* = 11) and 83% of RA (-) markers (*n* =10 / 12), with two genes (MPZL1 and KCNAB2) showing upregulation in RA rather than the expected downregulation (Table S3), ZEBrA).

Analysis of zebra finch promoters

I analyzed several general features of brain-expressed promoters in zebra finch. The results of an analysis of the repeat content of promoter sets using Repeatmasker revealed that brain-expressed promoters in zebra finch have high GC content (53.77% for promoters -1,000 and +500 bp relative to the TSS) relative to the background GC content of the full genome (41.45%). This is slightly higher than the GC content reported for the full chicken genome promoter set (51.5% [256]). Additionally, promoters of genes differentially expressed in HVC and RA are exceptionally GC-rich relative to the full brain-expressed promoter set (Table 6).

Analyses of vertebrate promoters, including chicken, have demonstrated that high- and low-GC promoters, known as broad and sharp promoters respectively, are distinct promoter types

that differ in their primary mechanism of regulatory control [253, 257, 258]. So-called GC-rich 'broad' promoters are known to be more widely expressed, feature more variable TSS locations, and deprived of TATA boxes relative to 'sharp' class promoters, which are comparatively enriched for TATA boxes and typically have a single TSS peak [258]. I attempted to further assess whether any of the song system marker gene promoter sets were biased for broad versus sharp promoters. To this end, I performed an analysis to identify localized motif enrichment for TATA box motifs (Figure 18), which implemented an algorithm designed to identify motifs that tend to be overrepresented in spatially conserved regions of a promoter set (CentriMo [259]). Analysis of the full zebra finch promoter set identified a spatially conserved TATA box binding element motif peak to be 25 bp upstream of the TSS (local enrichment p=2.2e-400, Figure 18), in keeping with the canonical location of TATA boxes in chicken (30 bp upstream of TSS [256]) and mammals (40 bp upstream of TSS [252]). The finding of a canonical TATA box motif peak spatially enriched in the biologically expected location serves to further support the accuracy of the brain-expressed promoter identification pipelines.

Approximately 16% of mammalian promoters [252, 253] and 10% of chicken promoters [256] contain TATA boxes. I found that 10% (n = 14/138) of RA (+) gene set promoters contained TATA boxes, in keeping with these figures. However, just 7% of all HVC-RA (n = 8/102), HVC-X (n = 4/61), and RA (-) (n = 16/212) gene set promoters contained significant TATA binding motif sites. These data suggest that promoters of genes differentially regulated in the HVC and RA subsets share canonical features of genes that are more widely transcribed than those of the full brain-expressed promoter set, potentially subserving a variety of different functions, and thus requiring complex regulatory control. Further assays of these genes' brain-wide transcriptional patterns, as well as a genome-wide screening assay for histone H3 lysine 4 trimethylation (H3K4me3), a mark

associated with unmethylated CpG islands, could be used to experimentally support this conclusion.

Promoter length (relative to TSS)	Full promoter set		HVC-RA		HVC-X		RA (+)		RA (-)	
	GC level (%)	Simple repeats (%)	GC level (%)	Simple repeats (%)	GC level (%)	Simple repeats (%)	GC level (%)	Simple repeats (%)	GC level (%)	Simple repeats (%)
-100, +50	63.66	9.40	66.39	11.41	70.43	11.96	62.37	9.45	63.14	10.03
-450, +100	57.58	5.78	60.19	5.84	63.11	5.03	56.34	5.78	56.76	5.51
-1,000, +500	53.77	4.27	55.57	4.51	56.81	3.91	52.91	4.45	52.86	4.03

Table 6. GC content and repetitive sequence in zebra finch brain-expressed gene promoters, and gene subsets differentially expressed in the song system.

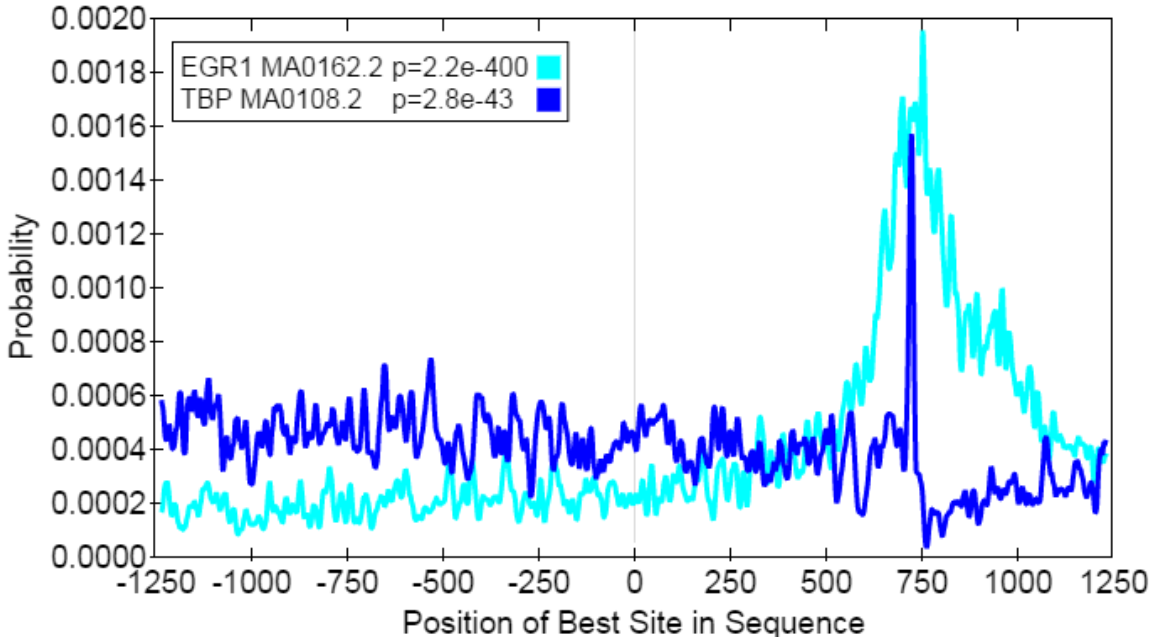

Figure 18. Position of best location of TATA-binding protein (TBP) and immediate early gene EGR1 consensus motifs in brain-expressed zebra finch promoters. TSS is located at position 750. TBP peak is located at position 725, EGR1 peak is located at position 755.

Cis-regulatory motifs enriched in promoters of song system-regulated genes

In order to identify motifs that could serve to regulate transcription in the song system, I performed a computational analysis to identify significant associations with known transcription factor binding motifs in song system-regulated gene sets. For this purpose, I developed a non-

redundant database of known motifs, employed an enrichment algorithm (AME [260]) to identify motifs significantly associated with promoter sets of interest, and finally performed a motif target search to predict which promoters were targeted by each motif, thus building sets of gene networks likely to be co-regulated in song system nuclei (for details, see Methods). The use of statistical enrichment algorithms to measure transcription factor binding is considered preferable to approaches based on counting of motif hits identified using an arbitrary cutoff for hit significance, as it has been experimentally demonstrated that low-affinity binding sites represent a critical regulatory mechanism in eukaryotic genomes [261, 262], and that a promoter with multiple low-affinity motif binding sites is in many cases more likely to be bound by a transcription factor than a promoter with a single high-affinity motif binding site [263].

For the song nucleus gene sets derived from ZEBrA (Table 5), the promoter enrichment pipeline failed to detect significant enrichment of known motifs, likely due to the relatively small number of these gene sets (between 2 and 62 genes each). However, the ISH data from these gene sets represented an excellent means of validating the larger microarray-derived gene sets analyzed subsequently.

As a result of the motif enrichment analysis of the HVC cell type and RA differential array data, I identified 9 significant gene set–binding motif associations in HVC-RA promoters, 4 in HVC-X promoters, 3 in RA (+) promoters, and 9 in RA (-) promoters (Tables 6, 7, S3, S4, S5). These final numbers were the result of a motif cluster analysis. Many of the motifs initially identified as being enriched both within and across gene promoter sets were found to be highly redundant, thus I prepared a pairwise correlation matrix of motif similarity for all motifs enriched in a given gene target set, in order to condense motifs enriched in redundant gene target sets (Table S4). Collectively, these associations represent 15,160 gene-motif associations and over 48,000 individual predicted binding sites (Table S5). I describe here the overall findings, predicted song

nucleus-specific motif–gene target pathways, and discuss the known functional associations of these predicted co-regulated gene networks.

TF MOTIF	**HVC-RA** (n = 102)	**HVC-X** (n = 61)	**RA (+)** (n = 138)	**RA (-)** (n = 212)	**Brain-expressed gene promoters** (n = 4,451)
EGR1	69 1.19E-02	47 6.65E-03	84 n.s.	139 1.91E-03	2,899
FOXJ3	9 1.52E-02	6 n.s.	10 n.s.	12 8.91E-05	293
FOXP1	27 2.32E-03	13 n.s.	36 n.s.	49 3.69E-04	1,062
NFATC2	9 n.s.	12 n.s.	29 2.70E-03	35 n.s.	670
NHLH1	25 1.39E-02	17 2.07E-02	24 n.s.	37 9.14E-05	888
SRF	21 4.26E-03	9 n.s.	36 n.s.	58 4.92E-05	1,124
TBP	8 3.10E-04	4 n.s.	14 9.36E-06	16 4.70E-05	325
TCF4	28 4.20E-02	22 5.64E-04	23 n.s.	35 n.s.	910
TFAP2E	21 n.s	13 n.s.	37 n.s.	59 6.59E-04	1,111
ZFP128	5 n.s.	1 n.s.	7 n.s.	9 9.46E-03	174
ZNF410	75 2.76E-03	46 n.s.	84 n.s.	139 n.s.	2,829
ZNF263	72 2.87E-06	41 1.85E-02	88 1.53E-03	142 1.87E-05	2,867

Table 7. Song system gene set-associated motifs. The top number in each cell corresponds to predicted number of significant motif–gene promoter associations in each target gene set, determined using a false discovery rate cutoff of $p < 0.0001$. The bottom number in each cell indicates motif enrichment scores for the target gene set, reported as Bonferroni adjusted p-values. Significance of enrichment takes into account strength of match to motif and number of hits to promoter. n.s.: non-significant, TF: transcription factor.

Immediate early gene EGR1 (aka ZENK) was significantly associated with both HVC projection neuron promoter sets, in addition to the RA (-) set (Table 7). This represented an expected result, given the importance of this immediate early gene in all neuronal tissues, as well as the fact that enrichment of this motif had previously been reported for activity-dependent gene

regulation in the song system[139]. In keeping with its known role as a core promoter-binding element, I found EGR1 binding sites to be preferentially located in close proximity to the predicted TSS location (Figure 18). As a key immediate early gene regulated by behavior and neuronal activity, EGR1 has been extensively studied in songbirds, where it has been demonstrated to be expressed in the song system in response to hearing or producing song [58, 264, 265]. Its predicted HVC and RA target gene networks are large (Table 7, Table S5), and showed significant functional association with biological processes involving nervous system development, axonal projections, and synaptic activity (Table S6). EGR1 has been shown to bind to its own promoter, resulting in a downregulation of its own activity [266]. This could serve as a potential mechanism behind the finding that repeated exposure to song leads to a decreased induction of EGR1 over time in zebra finch auditory cortex [267]. Supporting this hypothesis, I found a significant EGR1 binding site in its own promoter (Table S5). These confirmatory findings added validation to the promoter motif enrichment methodologies used in this study, as well as providing song nucleus-specific EGR1 target gene predictions that could subserve the immediate early gene response in the zebra finch.

Gene	Zebra finch gene locus	Brain expression evidence (ESTs)	Brain expression evidence (mRNAs)	Brain expression pattern (ISH)
EGR1	chr13:18142-21075	CK234258, CK308891, CK313422, DV947454, DV953620, EE061142, FE734327, FE732368	EF052676, JX296528, EF063681	activity-induced, all song nuclei [264]
FOXJ3	chr1:88321169-88329070	no	no	-
FOXP1	chr12:17203773-17366141	FE728400, CK316597, DV947108, FE19884, DV957717, DV948713, FE715201, DV950234, DV954317, DV957839, FE736659, FE729356, DV955340	AY549152, JN800730, JN800727	Area X (+), HVC (+), LMAN (-), RA (+) [132]
NFATC2	chrUn:19118196-19121362	no	no	-
NHLH1	could not be localized in taeGut1	-	-	-
SRF	chr3:19786724-19796586	DV580161, EE052995, FE727176, FE731342, CK302994, DV952757, FE723648, FE721454, FE736908, FE736747	no	-
TBP	chr3:42034996-42043036	FK818232, FE736399, CK315004, FK818231	no	-
TCF4	could not be localized in taeGut1	-	-	-
TFAP2E	chr23:5575660-5595827	CK315244	no	-
ZFP128	could not be localized in taeGut1	-	-	-
ZNF263	could not be localized in taeGut1	-	-	-
ZNF410	chr5:37744483-37771159	FE716191, CK308593, CK315055, FE736371	no	-

Table 8. Evidence for brain expression of transcription factors associated with motifs enriched in HVC projection neuron and differential RA gene sets. Genomic locations refer to zebra finch assembly taeGut1.

FOXP1 mutations in humans have been linked to intellectual disability, autism, epilepsy, and motor delays in human, with one of its most consistent features being significant speech and language deficits [268-273]. FOXP1 has received considerably less attention as a speech-disorder linked gene than FOXP2, in large part due to the confound that FOXP1 variants in humans have so far not been found that result solely in language deficits, without the presence of additional cognitive phenotypes, supporting the conclusion that it is a key player in many neurological

processes beyond vocal production (reviewed in [274]). Although FOXP1 is known to be upregulated in RA and both HVC projection neurons in songbirds, FOXP2 shows no such enrichments [132, 275]. This pattern of FOXP1 and FOXP2 brain expression has been shown to parallel its expression pattern in humans [132]. This suggests that these two transcription factors may play separate roles in the telencephalic circuits for vocal learning and vocal motor production in humans and avian vocal learners, and motivates further interest in the distinct role of FOXP1 in the production of learned vocalizations.

FOXP1 has been shown to play a role in the development of axonal projections and circuit formation [276], and brain-specific FOXP1 deletion in inbred mice has been linked to morphological and physiological deficits in multiple neuron types, leading to a range of behavioral abnormalities [277]. FOXP1's predicted target gene sets in HVC were found to have significant functional associations with axons and vesicles, as well as GABAergic synaptic pathways (Table S6), providing a potential molecular substrate for the morphological and physiological deficits observed in rodent FOXP1 knockout. $GABA_A$ receptors have been shown to be necessary for mediating the fast IPSPs evoked in HVC-X cells by both HVC interneurons and HVC-RA, which are connected through a well-studied HVC microcircuit [278]. I found significant FOXP1 motif sites in the promoters of $GABA_A$ receptor subunit GABRA5 and GABARAPL1, a $GABA_A$ receptor-associated gene that is a differential marker of HVC-X neurons (Table S5). This suggests a specific prediction, that FOXP1's deficits could in part be due to its effects on GABRA5 and GABARAPL1 expression, which would be expected to have a disruptive effect on vocal behavior, based on its demonstrated necessity in the physiological function of the HVC microcircuit. Based on these findings, manipulation of FOXP1 levels in HVC represent a highly promising avenue of research, with specific predictions I have presented as to the expected molecular and physiological outcomes.

Predicted target genes of FOXP1 in RA were similarly associated with nervous system function, axons, dendrites, and synapses, including GABA receptor gene GABRE (Table S6), which was shown to be upregulated in RA in the differential microarray analysis (Table S3). Although $GABA_E$ has not been specifically investigated for its role in RA, GABAergic responses have been suggested to underlie its characteristic bursting behavior, which characterize RA's time-locked output during song production [279].

A motif associated with SRF was significantly enriched in the HVC-RA and RA (-) promoter sets (Table 7). SRF is expressed in the songbird brain throughout development, based on the genomic alignment of brain-derived ESTs obtained from embryonic, juvenile, and adult brains (Table 8). In humans, SRF is a critical transcription factor expressed throughout the body and throughout development, with noted roles in neurons and muscle cells [280]. In addition to its primary roles in regulating cellular growth and the actin cytoskeleton [281, 282], it is an essential regulator in the activity-dependent immediate early gene response [283], the regulation of miRNAs [284], the cellular response to androgen [285], and in the establishment of cortical axonal projections [286].

SRF's role in androgen sensitivity is of especial interest in the context of the song system. Androgens have been demonstrated to play an essential role in the sexual differentiation, growth, and development of the song system [287-289]; and are sufficient to induce the development of a song system and singing behavior in female canaries, who do not display singing behavior under normal conditions [290]. Specifically, androgens are responsible for the seasonal neurogenesis and adult replacement of HVC-RA cells [291]. Given the established importance of androgens in the function of the song system, it was indeed surprising that I did not detect evidence of enrichment of known androgen receptor binding motifs in song system regulated gene promoters. Interestingly, however, evidence suggests that androgen receptor is capable of gene activation in

the absence of a DNA binding site, through its ability to recruit SRF to promoter SRF binding sites [285]. Thus, it may be possible that androgen partially exerts its effects in the song system through its interactions with SRF. Supporting this possibility, predicted target genes of SRF in HVC promoter sets were shown to be functionally associated with the response to hormones, including androgen (Table S6). A specific prediction suggested by this hypothesis would be that inhibition of SRF in HVC, either through genetic manipulation or pharmacological inactivation, could be sufficient to block the adult replacement of HVC-RA neurons in response to androgen.

In a previous analysis of gene regulatory mechanisms in the zebra finch estradiol-synthetic pathway, SRF was predicted to be present in the promoter of aromatase [248], an association which is confirmed in the present study (Table S5). Interestingly, in the largest published study of zebra finch regulatory networks to date [139], SRF was strongly associated with enhancers of genes that undergo gradual downregulation in RA in response to singing. The associations shown here, of SRF's binding motif with genes constitutively downregulated in adult RA and differentially enriched in HVC-RA-projecting neurons, extend these findings, indicating that SRF may perform a variety of regulatory duties in the song system, mirroring its previously discussed involvement in multiple mammalian regulatory processes.

The NHLH1 binding motif was found to have a significant association with both HVC projection neuron population sets, as well as the RA (+) gene promoter sets. NHLH1 expression is exclusive to the nervous system, and has been shown to be required for basic neuronal function in inbred mice [292]. NHLH1 has been shown to play important, conserved roles in the neuronal differentiation and migration across vertebrates [293-295].

Although its HVC targets were not associated with any functional enrichments, its predicted RA target network was enriched for genes involved in neural projections, including CNTNAP2 (Table S5). Although best known for harboring one of the most common genetic

susceptibility variants for autism, CNTNAP2 variants are perhaps most strongly associated with language deficits in humans [118], evidence suggests that it may be a significant downstream effector of the language impairments associated with mutations in FOXP2 [118, 137, 138]. Songbird researchers have noted its differential expression and protein abundance in song nuclei, including RA [135, 136]. The known importance of FOXP2, which is not differentially expressed in RA, in the regulation of CNTNAP2 activity in speech and vocal learning strongly motivates further study of the role NHLH1 regulation could play in modifying CNTNAP2's activities, in RA and elsewhere in the song system.

The NHLH1 gene could not be localized in the current zebra finch assembly, which would be required to distinguish alignment of potential brain-expressed sequences to either NHLH1 or its related family member NHLH2. However, NHLH1 has previously been reported to be expressed in several sites in the developing chick forebrain, hindbrain, and spinal cord [296]. A search of NCBI's gene sequence databases revealed predicted NHLH1 orthologs in the conserved chicken syntenic locus for three other songbirds (*Geospiza fortis*, XM_005431031; *Pseudopodoces humilis*, XM_005533400; *Sturnus vulgaris*, XM_014892715), indicating that future zebra finch assemblies will likely provide localization of the gene.

The NFATC2 motif was found to be enriched solely in the RA (+) set. NFAT family genes serve as transcriptional co-factors that participate in numerous developmental processes, most significantly in axon growth and guidance [297]. NFATC2 mutant inbred mice have been shown to present significant axon guidance defects, as well as an inability to respond to neurotrophins or netrin-1 [298]. Supporting the association of this motif with axonal function, its gene targets in the differential RA set were functionally associated with axons, cell-cell signaling, and nervous system development (Table S6). Although expression of NFATC2 was not detected in the zebra finch brain (Table 8), NFAT genes share overlapping functions and binding motifs across vertebrates [299],

and it is possible that the enriched motif could correspond to a related NFAT transcription factor. Interestingly, NFAT transcription factors have been shown to form cooperative protein complexes with FOXP family members, altering their regulatory target binding preferences [300]. In a case of NFATC2 gene disruption due to interstitial duplication, a patient presented delayed onset of speech, although this was noted in conjunction with moderate, general neurodevelopmental delay, in addition to cardiac and skeletal abnormalities [301]. However, the delay in speech acquisition long outlasted the improvement in the patient's other behavioral and motor delays, requiring speech therapy [302].

Another motif, corresponding to TCF4, was significantly associated exclusively with promoters of the HVC projection neuron gene sets. Human mutations in TCF4 are associated with severe intellectual disability, psychomotor delay, epilepsy, and various growth abnormalities, in addition to a total absence of language facility [303, 304]. In inbred mice, TCF4 has been shown to act in concert with other transcription factors in the brain, playing a significant role in fear conditioning and sensorimotor gating [305]. Disruption of TCF4 in cultured neurons has been shown to disregulate axon growth and branching by abolishing the effects of hepatocyte growth factor (HGF) [306]. Supporting a similar function in songbirds, I found TCF4 targets in HVC to be functionally associated with axonal projections and the growth cone (Table S6), suggesting that TCF4 could be a critical factor in the development and maintenance of long-range projections of HVC.

The TCF4 gene is not present in the current zebra finch assembly, and could not be associated with any data suggestive of brain expression. The TCF4 binding motif, derived from SELEX-sequenced binding sites of human TCF4 [307], has been shown to share a similar binding profile to several other vertebrate basic helix-loop-helix transcription factors [308]. Thus, it is

possible that another transcription factor with a DNA recognition motif similar to human TCF4 could bind to the enriched motif identified in this study.

One motif, corresponding to ZNF263, was significantly associated with all promoter sets, suggesting that this motif could be associated with a transcription factor that performs a more general role in the zebra finch song system. ZNF263 is widely expressed, and has been experimentally demonstrated to bind to more than 5,000 binding sites in human cell cultures [309], including the CNTNAP2 promoter, an association confirmed by my data (Table S5). Interestingly, an allele that disrupts the ZNF263 core binding site in the CNTNAP2 promoter has been associated with autism and delayed speech onset [310]. Despite its numerous and diverse predicted target genes in zebra finch, its predicted HVC and RA target networks were associated with only general, non-neuron-specific cellular functions, including translation, RNA binding, and ribosomal function (Table S6). Further, the gene itself could not be localized in the current zebra finch assembly, precluding identification of brain-associated expressed transcripts. Zinc finger transcription factors represent a highly expanded gene family in vertebrates [311], which presents an impediment to confident assessment of orthology and all subsequent genomic analysis.

ZNF410 was enriched in HVC-RA gene set (Table 7), with nearly 3,000 predicted motif-promoter binding associations across the brain-expressed promoter set (Table S5). ZNF410 has been demonstrated in humans to be a reference housekeeping gene expressed ubiquitously across several tissue types [312]. Supporting a similar role in zebra finches, it was found to be transcribed in brain tissue (Table 8), and its predicted HVC and RA target gene networks were significantly enriched for numerous general as well as neuronal functional processes, largely overlapping with functional associations enriched in the full HVC and RA gene target sets (Table S6).

ZFP128, a motif with few predicted gene targets and solely enriched in the RA (-) set, corresponds to a C2H2 zinc finger transcription factor unique to the mouse genome. Although poorly known, the gene has been hypothesized to participate in the bone morphogenetic protein (BMP) signaling pathway [313]. The gene could not be localized in the zebra finch assembly, and its predicted HVC and RA targets were not significantly associated with any known functions.

Finally, a motif derived from FOXJ3 was found to be enriched in the HVC-RA and RA (-) sets (Table 7). Although not well described, FOXJ3 has been suggested to participate in neural development, with expression in neural crest tissues described in inbred mice [314]. Its predicted song system gene targets were not significantly associated with any known functions, and no brain expression was identified at its gene locus (Table 8).

Brain expression could not be detected—or assessed, in the case of genes that did not align to the taeGut1 assembly—for a number of the transcription factors whose motifs were found to be enriched in HVC and RA (Table 8). However, it should be noted that it is possible that these transcription factors are expressed only under developmental or behavioral conditions not directly assessed by available transcriptomic resources. Thus, rather than being constitutively expressed or repressed by constitutively active transcription factors, it is possible that predicted motif target genes are instead primed to respond only under select conditions and states. Further, functional overlap with related gene family members has been noted for many of the transcription factors discussed, and it is possible that alternate transcription factors whose DNA binding motifs have shifted in songbirds could be regulating the gene sets identified in this study. Careful experimental interrogation, ideally through the use of ChIP-seq to specifically identify sites of TF-DNA interaction, will be necessary to conclusively demonstrate the motif-gene associations predicted in this work.

Comparative promoter analysis between zebra finch and chicken

The finding that gene sets regulated in brain nuclei for vocal learning are associated with select binding motif enrichments motivates an evolutionary question: are these motif–gene set associations also present in species that do not learn their vocalizations, that do not have a song circuit, and that thus do not display differential gene regulation in these regions? To begin an exploration of this new avenue of research, I attempted to test two simple models for the evolution of region-specific expression in a lineage-specific brain region, such as a songbird vocal control nucleus. In the first model, cis-regulatory element–gene associations are relatively fixed over evolutionary time, and newly evolved structures (such as song system nuclei) take advantage of preexisting transcription factor regulatory pathways. In this case, the song system gene sets described earlier would be expected to show similar enrichments in gene sets derived from orthologous promoters of a non-songbird avian species. In the second model, new motif–promoter associations are established, in order for transcription factors already expressed in a given brain region to modify the expression of genes not previously under their regulatory control. In this case, the song system gene sets described earlier could show unique enrichments, not observed in gene sets derived from orthologous promoters of a non-songbird. Clearly, these simple models are not mutually exclusive, but to my knowledge, the degree to which either model is supported by the analysis of regulatory motif conservation in brain co-expressed gene sets has not been specifically tested. I sought to assess whether the motif enrichments linked to genes expressed in zebra finch song nuclei also occurred in orthologous gene promoter sets in chicken, a non vocal learner. In doing so, my goal was to assess the level of support for these two models of the evolution of gene regulatory regions, which could have larger implications for the evolution of the brain and behavior.

To test the level of support for these two models, I reapplied the motif enrichment pipeline performed previously in zebra finch to the orthologous promoter sequences of chicken, to test whether the motif–gene set enrichments observed in zebra finch were also present in chicken. I was able to identify orthologous chicken promoters for 4,351 (98%) of the zebra finch brain-expressed promoter set, specifically identifying chicken orthologs for 86 of the 102 genes (84%) in the HVC-RA set, 56 / 61 (92%) for the HVC-X set, 135 / 138 (98%) for the RA(+) set, and 200 / 212 (94%) for the RA(–) set. Overall, I found that 10 / 25 (40%) motif–gene set enrichments were conserved between finch and chicken (Table 9). 15 / 25 (60%) of enrichments were found to be uniquely enriched in zebra finch gene sets. In 6 cases, chicken gene sets displayed motif enrichment that was not observed in the orthologous zebra finch gene set. In 17 cases, motifs shared non-enrichment in both chicken and zebra finch.

Thus, the data seem to suggest that gene sets expressed differentially in the song system display slightly more unique motif enrichments (60%) than broadly conserved enrichments shared in chicken (40%). However, I note that the overall level of motif enrichment conservation between finch and chicken roughly correlated with the percentage of orthologous chicken promoters obtained and analyzed, and that in some cases the enrichment scores were near the threshold for significance ($p < 0.05$), suggesting that some of the differences in gene set enrichment could potentially be explained by lower numbers of promoters in the chicken sets relative to the finch sets. In consideration of this, I draw the more conservative conclusion that motifs enriched in the promoters of zebra finch song system gene sets are neither more nor less likely to be uniquely enriched or conserved in orthologous promoters of a non-songbird. Thus, these data provide roughly equal support for models 1 and 2, suggesting that the evolution of differential regulation in the song system takes advantage of preexisting, broadly conserved regulatory pathways, as well as evolving new transcription factor–gene target associations.

MOTIF	HVC-RA	HVC-X	RA (+)	RA (–)
EGR1	Chk: 5.12e-03 Zf: 1.19E-2	Chk: n.s. ZF: 6.65E-3	Chk: 1.25e-3 Zf: n.s.	Chk: 7.22e-5 Zf: 1.91E-03
FOXJ3	Chk: n.s. Zf: 1.52E-02	Chk: n.s. Zf: n.s.	Chk: 9.01e-3 Zf: n.s.	Chk: 1.35e-2 Zf: 8.91E-05
FOXP1	Chk: n.s. Zf: 2.32E-03	Chk: n.s. Zf: n.s.	Chk: 3.13e-3 Zf: n.s.	Chk: 1.64e-3 Zf: 3.69E-04
NFATC2	Chk: n.s. Zf: n.s.	Chk: n.s. Zf: n.s.	Chk: n.s. Zf: 2.70E-03	Chk: 3.85e-3 Zf: n.s.
NHLH1	Chk: n.s. Zf: 1.39E-02	Chk: n.s. Zf: 2.07E-02	Chk: 2.69e-3 Zf: n.s.	Chk: n.s. Zf: 9.14E-05
SRF	Chk: n.s. Zf: 4.26E-03	Chk: n.s. Zf: n.s.	Chk: 5.73e-4 Zf: n.s.	Chk: 4.57e-2 Zf: 4.92E-05
TBP	Chk: n.s. Zf: 3.10E-04	Chk: n.s. Zf: n.s.	Chk: 3.47e-3 Zf: 9.36E-06	Chk: n.s. Zf: 4.70E-05
TCF4	Chk: n.s. Zf: 4.20E-02	Chk: n.s. Zf: 5.64E-04	Chk: n.s. Zf: n.s.	Chk: n.s. Zf: n.s.
TFAP2E	Chk: n.s. Zf: n.s	Chk: n.s. Zf: n.s.	Chk: n.s. Zf: n.s.	Chk: n.s. Zf: 6.59E-04
ZFP128	Chk: n.s. Zf: n.s.	Chk: n.s. Zf: n.s.	Chk: n.s. Zf: n.s.	Chk: n.s. Zf: 9.46E-03
ZNF410	Chk: n.s. Zf: 2.76E-03	Chk: n.s. Zf: n.s.	Chk: n.s. Zf: n.s.	Chk: n.s. Zf: n.s.
ZNF263	Chk: 3.15e-07 Zf: 2.87e-06	Chk: 9.57e-04 Zf: 1.85e-02	Chk: 1.32e-9 Zf: 1.53e-03	Chk: 1.25e-15 Zf: 1.87E-05

Table 9. Motifs enriched in promoters of song system marker gene sets in zebra finch and chicken. Motifs displaying shared gene set enrichment in both species indicated in yellow, motifs uniquely enriched in zebra finch gene sets indicated in blue, motifs uniquely enriched in chicken indicated in red. Abbreviations: Chk, chicken; Zf, zebra finch.

Previous studies provide additional support for both models. Supporting the first model, it has been extensively demonstrated that some regulatory regions are highly conserved across vertebrates [315], and comparative analyses have demonstrated that some predicted regulatory motifs are broadly conserved across mammalian promoters [316]. Thus, the evolution of a novel brain nucleus could be accomplished by altering the expression of specific transcription factors within a brain region, in order to alter that region's functional properties, taking advantage of pre-existing motif–promoter associations.

Supporting the second model, it has been shown that although gene expression across tissues is widely conserved across vertebrates, there is little correlation in conservation of the cis-regulatory regions of co-expressed genes, suggesting that conserved gene co-expression can be accomplished through alternate regulatory mechanisms [317]. It is also known that ultra-conserved noncoding regulatory regions are relatively few in number and largely associated with enhancers rather than promoters [315]. Outside of these conserved regulatory regions, non-coding genomic regions appear to be under low selective pressure, evolving rapidly from species to species [123].

Further, extensive work has demonstrated the importance of endogenous retroviral elements, or retrotransposons, in shaping the course of genomic evolution [318, 319], in some cases through their ability to alter cis-regulatory sequences and gene expression. These elements translocate across genomic locations in an inexact manner, and have been demonstrated to interrupt and alter gene regulatory regions through insertion, in some cases bringing with them genomic regulatory motifs from their original locus, in some cases with significant consequences for gene expression [320, 321]. In my own work, I have discovered a gene duplication where one copy exhibited altered expression relative to the other copy, and also showed evidence of retroviral insertion into its promoter region (see discussion of CASC1 duplication, Chapter 2). Thus one may conclude that this shuffling of regulatory regions may be a mechanism that allows new motif–gene target promoter associations to form, potentially resulting in the large-scale alterations in transcription that allow a new brain nucleus with specialized properties to emerge.

2.1 Methods

Sequence Data

The genomes of zebra finch and chicken were downloaded from public databases (taeGut1, NCBI: PRJNA122047; galGal4, NCBI: PRJNA13342). The complete set of publicly available zebra finch full-length mRNAs, including all RefSeqs, was retrieved from NCBI. All 5'-cap ESTs, which were derived from a comprehensive set of diverse brain tissues under a range of behavioral conditions [125-127], were also retrieved from NCBI.

Identification of brain-expressed gene promoters

In order to derive a database of all zebra finch promoters with experimental support from brain expression studies, I first aligned the complete set of brain-derived mRNA and EST sequences to the zebra finch genome assembly using BLAT [180], with parameters optimized for same-species alignment as described in Chapter 2. These were then filtered so as to include only those hits where the alignment to the genome started within the first 50 bp of the query sequence, and represented the highest scoring alignment of each query sequence to the assembly. BLAT outputs were then combined, converted to BED format, and curated to remove artifactual alignments using custom scripts. Bedtools [322] was used to merge all overlapping alignments, which were then assigned names based on the most upstream overlapping Ensembl model. A total of 525 loci did not overlap with any Ensembl model. Ad hoc manual inspection of a sample of non-gene overlapping alignments revealed them to largely represent lncRNAs and retroviral elements, unsuitable for gene promoter analysis. All 153 alignments with duplicate Ensembl model names were also manually inspected. These misannotations were found to derive from artifactual Ensembl models, and were reannotated on the basis of sequence similarity to known RefSeqs from other species aligned to the locus in the UCSC genome browser. This resulted in a high-

confidence set of 4,451 experimentally supported first exon-containing loci. Finally, I used custom scripts and implementations of the bedtools 'slop' and 'getfasta' functions [323] to generate promoter sets according to the length specifications described in the text.

Genomic repeat content analysis

Repetitive sequence content was assessed using the RepeatMasker application (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker), using the more sensitive cross-species search algorithm (cross_match), with chicken as the comparative contrast species, and otherwise default parameters.

Microarray datasets

In order to analyze promoters of genes with HVC projection neuron-selective expression, I reanalyzed the results of a microarray study which contrasted expression in HVC-RA and -X neurons identified through fluorescent retrograde labeling and obtained through laser capture microdissection [243, 255]. The results provided 324 clones from HVC-RA and 175 from HVC-X neurons that were significantly different at a false discovery rate of $p < 0.05$. These were then manually annotated using extensive cross-species alignment & synteny approach, described previously. This condensed the set to 144 putative HVC-RA marker genes and 74 putative HVC-X marker genes. An analysis of archival double labeling data from our lab contributed an additional 4 HVC-RA and 5 HVC-X marker genes. I then performed a manual analysis of the entire set, considering alignments of chicken and human genes as well as 5' ESTs, to identify genes for which we have true first exons. I then identified the intersection of the array set and the full brain-expressed promoter set in order to generate promoter sets for song system markers.

In situ hybridization analysis

I performed an analysis of FISH data collected previously by our lab, in order to confirm cell type-selective expression of HVC-RA and HVC-X markers identified by microarray. Briefly, these data were collected by first bilaterally injecting fluorescence-conjugated cholera toxin B into Area X and RA to retrogradely label HVC projections into these nuclei, and subsequently performing FISH of these sections with Alexa-conjugated riboprobes to test gene markers in order to quantify the number of cells where probe labeling co-localized with retrograde labeling from Area X and RA, in order to identify selective markers of each projection neuron subtybe (Figure 16). DiG ISH was performed as in previous chapters.

Transcription factor motif collection

To date, no transcription factor binding sites have been directly determined in a songbird. To develop a suitable set of positional weight matrices (PWMs) for use in the analysis of zebra finch promoters, I first investigated the use of a library of inferred sequence binding motifs for zebra finch, among >300 other species (CisBP [308]). To reduce redundancy and improve species-specific motif accuracy, the authors of CisBP attempt to link each TF present in a given genome to directly determined PWMs from the nearest available species. However, a cutoff is imposed to limit the allowable amount of sequence divergence in predicted DNA binding motifs between species. Because of this, many TFs with known activity in the zebra finch song system (e.g., FOXP1) are not represented in the database. As an alternative, I assembled a set of motifs from several vertebrate PWM databases to cover the fullest possible range of TFs present in the zebra finch genome. This set comprised PWMs derived from protein-binding microarrays in mouse (UniProbe[324]), high-throughput SELEX and ChIP sequencing in human and mouse [307], and JASPAR, an extensively curated, non-redundant motif database integrating experimental binding data from numerous

sources [325]. In total, this PWM set consisted of 1,748 motifs, between 6 and 23 bp in width (average width 13.2). PWMs were converted to the MEME motif format, adjusting positional dependencies based on the background nucleotide sequence composition of the full zebra finch promoter set.

Motif enrichment analysis

To identify binding site motifs that were associated with specific gene promoter sets of interest, I performed a motif enrichment analysis to identify overrepresented motif–target set associations using a threshold-free linear regression algorithm that has been demonstrated to outperform other available algorithms for detection of motif enrichment [260]. I calculated the enrichment of motifs from the previously constructed PWM set within each gene promoter set of interest relative to its background promoter set. For each target gene set, a background set consisting of all other brain-expressed promoters was first used to calculate position-specific priors, in order to account for biased likelihood of motifs generally present in the full promoter set. Enrichment was calculated using average odds scoring, which calculates PWM scores as odds scores for overall probability of motif-promoter binding and averages these over the set of target promoters, calculating the enrichment relative to the probability of finding the motifs in shuffled input sequences, in order to prevent bias due to uneven positional dependence of background dinucleotide frequencies.

Motif target search

As increasing the number of PWMs used to search a sequence set also increases the false discovery rate, I sought to reduce the number of redundant motifs before performing the final motif target searches. I used a method to generate a pairwise correlation matrix of motif similarity for all motifs enriched in a given gene target set (MAST [326]). Motifs with pairwise correlations

above 0.60 were considered redundant, and combined. To obtain sets of predicted gene targets of enriched transcription factors identified through the motif enrichment and discovery analyses, the FIMO algorithm was used [327], using a strict false discovery rate cutoff of $p < 1E-4$.

Functional annotation of motif target gene sets

To identify the potential function of predicted transcription factor target gene sets, I ran a gene ontology analysis using StringDB v10.0 [328] to identify enrichments for functional associations within gene sets using annotations from the Gene Ontology database (GO [329]) and KEGG PATHWAY database [330].

Comparative promoter analysis

Ensembl gene IDs associated with the complete zebra finch promoter set were used to obtain the set of orthologous chicken promoters from Ensembl BioMart [184]. To identify the level of base conservation between zebra finch and chicken promoters, BLAT was used, with parameters sensitized for cross-species alignment, as described in previous chapters. Chicken gene promoter sets corresponding to those used in the zebra finch promoter analyses were constructed by matching gene orthologs between finch and chicken. The motif discovery and enrichment pipelines were run as before, using chicken promoter sets, to identify differences and similarities in motif enrichment in gene sets of interest between finch and chicken.

5 New models for understanding the evolution of a complex behavior: a general discussion

5.1 Current models of the evolution of learned vocal behavior

Current and past models for the evolution of vocal learning have largely treated it as a binary trait, where some species are vocal learners and the rest are vocal nonlearners, with no intermediary species [20, 64, 331]. One of the basic tenets of Darwin's Theory of Evolution is that evolution results from the gradual, successive accumulation of mutations across generations [332]. It seems improbable that the substrates for vocal learning—including a modified vocal apparatus and multiple brain structures devoted to vocal control, which are interconnected throughout and beyond the brain and involve the coordinated expression of thousands of genes—could have evolved overnight, through a single or even several mutation events. Such a mechanism of origin has been proposed for the song system by Erich Jarvis and colleagues, who suggested that vocal learning circuits evolved at once, through the duplication of a preexisting motor pathway exhibiting similar connectivity, as a result of gene duplication [333, 334]. Although this would seem to contradict the generally accepted notion that the mutations that confer fitness are rare, of small effect, and accumulate slowly over evolutionary timescales (i.e., 'gradualism'), subsequent evolutionary theorists have suggested the possibility of abrupt mutation events that result in dramatic phenotypic changes, which confer extreme fitness benefits, termed 'hopeful monsters' [335]. However, convincing examples of such 'hopeful monsters' have so far been identified only in unicellular organisms [336].

This hypothesis has support from linguists such as Noam Chomsky and primatologists such as Marc Hauser, who have argued that human language facility must have arisen from a single gene event, given that many aspects of language, including certain properties of grammar, are shared amongst humans universally and not found elsewhere in the animal kingdom [337]. This

appears further supported by the finding that parrots seem to possess a second song system within their song system which exhibits differential connectivity and gene expression, interpreting this pathway as a second motor pathway duplication [228]. However, in no case has a gene or set of genes that could result in the duplication of an entire set of brain nuclei and attendant circuitry been identified. For now, this hypothesis rests on a 'hopeful monster' of unprecedented scale. It thus seems necessary to consider hypotheses that implicate intermediate forms between vocal learners and nonlearners.

Jarvis has proposed such a hypothesis, the "continuum hypothesis" for the evolution of vocal learning [338]. In support of this hypothesis, Jarvis and colleagues have reported the identification of a direct corticospinal projection to the hindbrain vocal nucleus in inbred mice, which they propose could serve to control their ultrasonic 'vocalizations' [339]. Further, they interpret their finding in a vocal non-learning suboscine, the Eastern phoebe, of a similar direct projection from an 'RA-like' nucleus to the hindbrain vocal motor nucleus, as additional support [240]. On the basis of these findings, they concluded that vocal learning behavior exists as a continuum, where non-human primates, which lack such a direct projection [79-82], are considered "limited vocal learners;" mice are considered "moderate vocal learners," songbirds and parrots are considered "complex vocal learners," and humans alone are considered "high vocal learners" [338]. Thus, they predict, vocal learning evolves by gradualism, with intermediate species only partially learning their vocalizations.

The first issue with this hypothesis is of a more philosophical nature, as it implies that the supposed 'limited' or 'moderate' vocal learning of mice, monkeys, and songbirds are incomplete behaviors, imperfect versions of the implied behavioral ideal, the 'high vocal learning' of humans. This thinking is not unlike the *scala naturae*, the ancient reasoning that placed all of creation along a ladder, with the 'lower' animals firmly set below the 'higher' forms of Men and angels, with all of

Creation climbing up towards the heavens. Within the realm of science, this line of thinking was long ago superseded by the Theory of Evolution [332], and subsequent advent of modern phylogenetic systematics, or cladistics, in which all living organisms are descended from a common origin, constituting equal branches in a tree (or, more accurately, bush) of life [340]. Under this more nuanced understanding of evolution, where all living organisms are equally advanced from this common origin, language referring to 'higher' and 'lower' life forms is avoided, as it reflects misunderstanding of evolutionary theory stemming from anthropocentric bias. Finally, it is nonsensical to consider any behavior as an incomplete behavior – all behaviors are simply behaviors.

This vocal continuum hypothesis is more seriously contradicted by the existing data. First, in both their demonstrations of a direct corticospinal projection in inbred mouse and Eastern phoebe, the projection was identified on the basis of relatively few labeled fibers, was reproducibly detected in relatively few animals with a high failure rate, and appeared highly anatomically variable from animal to animal [240, 339]. Given that the injection target in both cases was a very small hindbrain nucleus, and that several hindbrain nuclei receive input from adjacent cortical motor regions to the proposed RA-like nuclei in these species, a very strong possibility exists that adjacent nuclei were retrogradely labeled through leakage at the injection site or through upwelling through the pipette injection column, or that axonal projection passing through the injection site were labeled inadvertently, a common technical confound known as the 'fibers of passage' issue. More problematic is the fact that the vocal behavior of these particular species has been thoroughly demonstrated as unlearned, perhaps more conclusively than in any other species. In inbred mice, the ultrasonic 'vocalizations' now widely studied as a model of human speech are not produced through the larynx at all, and thus are more accurately termed phonations, more similar to human wheezing and coughing [341, 342]. It was further

demonstrated that the two forms of rodent ultrasonic phonation represent by-products of either decreased body temperature, in the case of pup 'isolation calls' [343], or physical activity, in the case of the additional adult 'call' types [344]. The social aspect ascribed to these 'calls' is thus confounded by the fact that pup core body temperature decreases with maternal isolation [343], and that adult 'calls' are associated with social behaviors with a strong physical component such as tickling, play, mating, and general locomotion [345, 346]. Most damning of all, inbred mice with engineered deafness develop normal 'calls' indistinguishable in acoustic features and social production from undeafened mice [28, 347]. It is indeed possible that mouse phonatory behavior contains social cues, just as a cough or a wheeze could indicate respiratory illness. But the supposition that these acoustic emissions show evidence of learning, even 'limited learning,' has been thoroughly disproven. Given the lack of support for the 'continuum hypothesis,' I propose an alternate model for the evolution of vocal behavior.

5.2 The 'Stepwise Hypothesis,' a new model for the evolution of vocal learning

The Stepwise Hypothesis for the origins of vocal learning posits that the complex vocal learning behavior of humans and avian vocal learners can be seen as one step along a pathway. In contrast to the motor hypothesis, which treats vocal learning as a binary trait, in the stepwise hypothesis attaining the phenotype of complex vocal learning is dependent on several intermediate steps, each conferring a unique fitness advantage. However, in contrast to the continuum hypothesis, there is no implied elevation or directionality: each step is neither higher nor lower, and species may 'walk' in any direction.

Step 1. Innate motor variability and learned contextual usage

Most animals that vocalize are social, and this vocal behavior changes bioacoustically over a protracted period of development, largely a by-product of the physical development of the vocal apparatus. Domestic dogs, cats, and chimps exhibit a degree of individual call variation that is more variable in juveniles, where the development of the vocal apparatus is still in flux. Following this, animals can be trained to produce slightly different calls, selecting from an innate set of developmental outcomes of the larynx or syrinx. Anecdotally, my own cat, Rhiannon, when less than a year old, produced a variable set of 'meow' calls, including an infrequently produced, high-frequency 'coo' call, which I found to be endearing. Through selective provisioning of treats, I rewarded instances of 'coo' call production. Consequently, her rate of 'coo' calls increased, and all other 'meow'-like calls disappeared as she attained maturity. I did not myself produce the 'coo' call for her to imitate; thus it represented an innate, unlearned vocalization.

I propose that all current data on vocal individuality and context-specific usage in the calls of non-human primates can be most parsimoniously explained as similar phenomena, given the neuroanatomical and behavioral data discussed in Chapter 1. In both the example of my cat and of social call production in Apes, this context-specific associational learning confers clear fitness benefits, including social cohesion, alert to the presence of a predator, and procurement of a treat. I adopt Janik and Slater's terminology here, designating Step 1 'learned contextual usage' [19], with a requirement for innate vocal variability that is generally shared among vocalizing organisms.

Step 2. Selection on actively injected variability

One of the most fundamental insights of Darwin's Theory of Evolution was that variability in phenotype is a necessary prerequisite for natural selection [332]. Indeed, as populations lose their variation through inbreeding or strong stabilizing selection, so, too, do they lose their ability to evolutionarily adapt to changes in their environment.

More recently, it has been demonstrated that the neurobiology of learning mirrors this key aspect of evolution. During juvenile song learning, LMAN acts as a driver of variability, injecting a highly variable spiking signal into the motor output nucleus RA [348]. This variability signal has been shown to be a critical requirement for song learning [349]. Following this observation, studies of human motor behavior have confirmed that initial motor variability predicts subsequent ability to learn the performance of a novel motor task [350]. Integrating these findings, I now suggest that the evolution of the ability to actively inject variability into the motor signal represents the key evolutionary event that precedes the evolution of complex vocal learning.

A fundamental insight in support of this step came with the discovery that the variable food begging calls of infant songbirds share many overlapping features with adult song [351-354]. In a key study, it was shown that nestling food begging calls: 1) were more variable in males, resemble later subsong by several spectral parameters, 2) were altered by deafening (unlearned 'chip' contact calls are not and female nest begging calls were not), 3) induced differential expression of *c-fos* in RA (this was not the case in females, nor in the production of unlearned contact calls), and 4) electrolytic lesion of the developing RA significantly altered the parameters of several acoustic features of male food begging calls (lesions did not alter unlearned contact calls, and lesions of RA in females did not alter begging calls) [351]. On the basis of these

observations, Liu and colleagues suggested that this variability in infant food begging calls may be a developmental prerequisite for eventual song learning [351].

In certain species, this begging has become more elaborated, where begging calls that more closely resemble calls made by the mother result in higher food provisioning to the nestling [352-354]. Thus there is a clear selective advantage for infants to actively inject variability into their calls, and actively select calls that most closely resemble a tutored vocalization. Further, this ability is mediated by a forebrain motor control nucleus with a direct corticospinal projection to the hindbrain motor nuclei, RA. It is important to note that these vocalizations are calls, which would emerge in some form without tutoring, but that their acoustic character is shaped through auditory input from a tutor. I propose that this step, 'Selection on actively injected variability,' is a discrete phenotypic unit with a select fitness advantage distinct from that of learned song, which is more fully dependent on tutoring, and predict that the closer examination of variable infant nest begging across avian phylogeny will reveal further species that possess variable, tutored infant nest begging behavior but not learned song.

Temporal vocal learning (Step 3) versus Spectral vocal learning (Step 4)

It has been demonstrated that the learning of the temporal and spectral features of song require separate pathways and connections [355]. As reviewed in Chapter 1, the production of a learned vocalization is dependent on the vocal-motor pathway (HVC to RA to nXIIts). Lesioning of HVC or RA will lead to a selective loss of learned song production, although calls will still be produced. Although lesioning of the AFP pathway for learning in juveniles will result in an inability to learn song, lesioning of these nuclei in adults produces little degradation in song, suggesting that once a song has been learned, the vocal motor pathway is sufficient for its production. Further work has demonstrated that the pathways for learning in the spectral and temporal domains of birdsong are dissociated, and that learning in these domains can be modified selectively [355].

Given these observations, I suggest that suboscines may represent an example of an organism with vocal learning in the temporal, but not spectral domain. This is supported by: 1) the fact that many suboscines exhibit complex duetting behavior, requiring temporally precise coordination of song between males and females, and 2) my findings that suboscines share expression in an RA-like nucleus, but do not display evidence of an HVC nor any other AFP nuclei. All demonstrations of a lack of vocal learning in suboscines performed thus far have focused on a lack of difference in the spectral features of song, in species that do not exhibit duetting behavior. I predict that lesioning of the RA-like nucleus in a duetting suboscine, while not altering the spectral features of song, would detrimentally affect the bird's ability to modify the temporal features of song in response to its duet partner. In this case, temporal vocal learning could confer a unique fitness advantage—evaluation of performance by a mate—maintained by the sexual selection pressure.

Under this interpretation, the last common ancestor of parrots, songbirds, and suboscines was likely a temporal vocal learner, and songbirds and parrots have convergently evolved spectral vocal learning and its underlying forebrain circuitry. This is supported by: 1) the fact that although the position of RA is well conserved in songbirds and parrots, the position of HVC/NLC and the AFP nuclei is drastically different (Figure 10); 2) although direct motor pathway for temporal learning is conserved in parrots and songbirds (HVC/NLC to RA/AAC to nXIIts), the other pathways of the song system show alternate connectivity between parrots and songbirds (Figure 12); 3) my own findings that markers of songbird RA were more broadly shared in parrot AAC than markers of HVC / NLC (Tables 2,3); and 4) shared markers of RA / AAC are also remarkably well-conserved in an RA-like nucleus in suboscines (Table 4, Figure 15), similar in position and cytoarchitectonics to songbird RA (Figure 14).

I believe these results are more parsimoniously explained by the Stepwise Hypothesis. In this case, the evolution of an RA-like brain nucleus—which is defined by a direct corticospinal projection onto brainstem vocal and respiratory centers and a 'core' set of molecular specializations broadly shared by all vocal learning birds and humans—is a separate event preceding the evolution of an interconnected cortico-basal ganglia-striato-cortico loop. Emerging evidence shows that parrots, the premier vocal learners of the animal kingdom, may have an additional AFP pathway to subserve their advanced imitative abilities [228]. This could perhaps be interpreted as a fifth step, reflecting a level of inter-species vocal imitation that exceeds even our own (think about how well you can imitate a parrot). However, these imitative abilities are largely confined to the spectral domain, and reflect increasing degrees of spectral imitative ability rather than a discrete behavioral phenotype. Thus, I classify parrots and humans both as spectral vocal learners.

In one of the most celebrated papers on evolutionary theory since Darwin, "The Spandrels of San Marco and the Panglossian Paradigm," Gould & Lewontin describe the history of St. Mark's Cathedral in Venice. Following its initial construction, structures that had been built to serve one purpose—such as its spandrels, which were initially built for the practical purpose of supporting the Cathedral's massive dome—took on new and completely different purposes over time. Later changes to the Cathedral were not accomplished through the addition of new structures, but the extensive elaboration of preexisting structures, such as the spandrels. Similarly, in Chapter 2, I was able to identify 10 novel genes unique to songbirds, 2 of which exhibit differential expression in the song system. However, in Chapters 2 and 3, I also found that hundreds of genes are differentially expressed in the song system, and that these expression patterns may be regulated in part by taking advantage of transcription factor–gene target associations broadly conserved in vocal learning and non-learning birds alike. It follows that although novel gene evolution appears

to contribute to the properties of brain nuclei for vocal learning, the primary molecular identity of these circuits emerges through incorporation of preexisting genes, as in the example of the spandrels.

Given that the specialized properties of brain circuits evolved for vocal learning take advantage of preexisting genes and regulatory pathways, I suggest that vocal behavior could evolve similarly, with each 'Step' representing an elaboration upon preexisting behaviors and neural circuits: spectral vocal learning can be thought of as an elaboration (the AFP) in the spectral domain upon structures initially evolved for temporal vocal learning (RA and its corticospinal projection), which may be a sexual selection-mediated elaboration in adults upon the tutored, variable infant nest begging behavior of juveniles, which may be an elaboration on the developmental plasticity and associational learning widely shared by vocalizing animals.

Viewing these data through the lens of the Stepwise Hypothesis, no step is classified as 'incomplete.' Rather each step is associated with its own complete behavioral phenotype, conferring unique fitness advantages. Further, the underlying brain circuitry and molecular data suggest that the circuit does not emerge all at once, but rather through successive steps, where preexisting genes and structures take on additional duties to support novel structures. Thus, duetting suboscines are no more 'on their way' towards becoming 'more evolved' spectral learners like their oscine and psittacine kin than apes are waiting for the day they will become 'complete,' talking humans. In the unceasing evolutionary race, where there exists no endpoint so much as a constant pressure to outrun life's myriad challenges, every step is as good as the next.

6 Conclusions

6.1 Conclusions

I performed a comparative analysis of 48 avian genomes to identify genomic features that are unique to songbirds, as well as an initial assessment of function by investigating their tissue distribution and predicted protein domain structure. In order to facilitate these analyses, I developed original computational tools for lineage-specific novel gene discovery that overcome long-standing obstacles in this field by incorporating knowledge from evolutionary systematics. I also developed extensive resources for high-throughput analyses of avian genomes. I described and characterized 10 novel genes found uniquely in songbirds, including two with differential activity in brain regions necessary for vocal learning. These findings support the hypothesis that lineage-specific novel genes subserve some of the properties of vocal learning circuits.

In order to address the degree of molecular similarity among independently evolved vocal learning circuits, I identified molecular markers—transcripts uniquely expressed in specific cell-type populations—of vocal nuclei in a songbird, and subsequently demonstrated that a set of these represent shared specializations of analogous vocal nuclei in a parrot and hummingbird. In a nucleus critical for the fine motor control of learned vocalizations, these molecular specializations are further shared with humans. The finding that core sets of genes are convergently expressed in HVC- and RA-like nuclei across vocal learners suggests that these gene specializations may represent fundamental molecular requirements for all vocal learning circuits

In a complementary study, I explored whether expression of song nucleus gene markers was also conserved in suboscines, a group closely related to songbirds thought to be non vocal learners lacking a song circuit, although limited evidence suggests some species may in fact display learned vocal behavior. In order to complete this study, I collected high quality brain tissue in the field (Brazil) from 49 brains from 26 species representing 7 of the 12 families of New World

suboscines. This represents the largest collection of suboscine brain material in the world. ISH gene expression analysis of this material largely supports the lack of song nuclei in this group, although intriguing conserved expression is noted of several RA markers in a similarly located arcopallial domain, which may represent evidence for molecular 'preadaptation,' paving the way for the later emergence of true vocal learning nuclei, through a process that I describe as the 'Stepwise Hypothesis' of vocal learning evolution. Together, these findings support the hypothesis that unrelated vocal learning groups have convergently evolved shared molecular mechanisms within their vocal learning circuits, but also suggest that there may be specific differences that reflect multiple evolutionary solutions to the problem of building a song circuit in the evolutionary histories of these groups.

Finally, in order to identify the changes in regulatory sequence that may be associated with the co-expressed gene networks in vocal learning circuits, I developed a computational pipeline for avian promoter identification and analysis. I identified candidate vocal learning regulatory motifs associated with differentially expressed gene sets in RA, as well as the projection neurons of HVC. I propose that the predicted target genes of these motifs could subserve the known molecular and physiological properties of brain nuclei critical for learned vocal behavior, including neurogenesis, axon projection and guidance, and androgen sensitivity. These findings support the hypothesis that genomic regulatory elements underlie regional and cell type-specific gene expression networks in vocal learning circuits, and present a framework for understanding how shifts in gene regulatory sequence composition could lead to the emergence of vocal learning circuits. In sum, this work provides a roadmap for integrating knowledge from genomics, gene expression analysis, neuroanatomy, physiology, and phylogenetics to provide novel insights into the evolution of behavior.

6.2 Future Directions

This work demonstrated that incorporating synteny-based methodologies into gene annotation and novel gene finding pipelines improves the quality and reliability over purely alignment-based methodologies. New genomic sequencing technologies substantially improve on the read length and sequence contiguity over previous methods, and ambitious plans are underway to sequence the genomes of tens of thousands of species, forming a more complete picture of the molecular blueprints for life on earth [356]. The novel gene pipeline described here could be applied to identify genomic innovations that characterize species and clades at multiple levels. By tracking individual loci across phylogenetic trees, we might improve our understanding of general principles that govern the expansion and contraction of gene families. As it focuses on gene loci rather than quality of alignment, this pipeline is robust to the sometimes drastic changes in gene exon structure and sequence composition that can occur in the course of genome evolution. This motivates deeper integration of synteny, in addition to sequence alignment, into even routine genome annotation pipelines.

Although I identified several novel genes unique to the songbird lineage, including two uniquely expressed in song system nuclei, my analyses did not extend to experimental testing, to identify the role these genes could play in vocal learning. At ~3 and 2 kb, respectively, TMRA and YTHDC2L1 are both well under the packaging capacity limits of AAV and lentiviral vectors. Future experiments could seek to introduce custom short hairpin RNAs to knock down expression of these genes with high precision, and subsequently measure the effect on song learning.

The finding that unrelated vocal learners share a core set of vocal learning circuit markers opens up several exciting avenues for future research. The first task will be to gain deeper understanding into the molecular pathways and functions supported by these critical gene sets. Already, others have begun investigating some of these genes, establishing their importance in the

development of the critical cortico-spinal projections that characterize vocal learning circuits [357]. Future work should focus on analysis of the most important functional categories represented by these gene sets (Figure 11, Table S2), with special focus on ion channels, which could eventually lead to molecular substrates for these circuits' unique electrophysiological properties. An important caveat worth noting is that all of these studies were carried out in adult animals. Recent work has demonstrated that marker patterns in zebra finches can shift significantly over the course of development [358]. This raises interesting questions about whether these developmental patterns are also seen in non-songbird vocal learners. A concerted effort to study comparative development of specialized expression in convergent song systems could lend insight into whether there exist molecular correlates for the behavioral differences in song learning ontogeny seen in these groups. A better understanding of the most critical mechanisms that regulate neuroplasticity in open vs closed learners (represented by budgerigars and parrots, respectively) could have implications for how critical periods are closed, and how they may be reopened.

The finding that markers of song RA that are shared in its parrot analog AAC are largely also markers of an arcopallial domain in suboscines suggests that the common origin for this nucleus was likely present in the last common ancestor of these avian lineages. Most immediately, there is a need to run additional ISHs in suboscines to determine whether this domain shares more markers with songbird RA or Alat, as the affiliation with Alat described in this work was based on a relatively small sample size. It will also be necessary to define the projection targets of this region, through the use of anterograde tract tracing. Retrograde tract tracing from nXIIts could also be used to determine whether this or any other forebrain region contains a direct corticospinal projection to this syringeal control nucleus. Further, it will be necessary to determine whether this domain supports any component of vocal behavior in suboscines, especially with

respect to temporal learning, which could be critical to the complexly timed duets of some suboscine species [238, 359]. The possibility that vocal control pathways in suboscines represent a sort of intermediate stage in the evolution of complex vocal learning presents a prime opportunity to combine the work of field ecologists with molecular neuroscientists and genomicists to reveal novel insights into the evolution of complex behavior.

The advent of high-throughput sequencing has enabled the study of large scale patterns of gene activation in neural circuits, as well as the regulatory mechanisms that govern them. Nevertheless, sequence gaps interrupt many promoter regions, which are notoriously difficult to sequence due to high GC content [360], as well as artifacts of misassembly that generally plague alignment-based analysis of genome assemblies [361]. To ameliorate these deficits, our lab and others are in the process of preparing a new zebra finch assembly, constructed using PacBio long-read sequencing. Preliminary analyses indicate that this assembly contains substantial new genomic sequence, and is more contiguous than previous assemblies, filling in many of the sequence gaps present in the current zebra finch assembly, including those that interrupt promoters. The inclusion of additional promoters of song system markers, excluded from this study due to poor sequence quality, would likely bolster the promoter–motif associations presented here, as well as presenting new ones.

A primary goal for future comparative work is to apply the promoter identification and motif enrichment pipelines described here to additional songbird species for which brain expression data are available, such as Great Tit [362] and Canary [363], in order to identify whether these species share the promoter–motif associations and gene target network functional associations with those identified in zebra finch. The development of multi-species phylogenetic footprinting approaches, similar to recent models tracing the evolution of some human-specific regulatory elements [364], stands as an especially promising goal to distinguish zebra finch-

specific motif–promoter associations from those shared by all songbirds. For those species without sufficient brain transcriptome data, the highly curated gene promoter databases developed here could be aligned to comparable sequences upstream of the orthologous genes from the large set of avian genomes in NCBI (n=68, 9 songbirds and 59 non-songbird species). The promoter analysis pipeline developed here could be used to identify domains that are conserved throughout avian phylogeny, as well as determining which of the song system-enriched regulatory motifs identified in this work represent unique, conserved enrichment in song system marker genes across all songbirds.

This work delivers a set of highly promising candidate drivers of expression in vocal control circuits, which could be used to design artificial promoters to confer brain region and even cell type-selective expression to viral gene constructs. The most promising motifs, including FOXP1 and SRF, should first be validated through the use of chromatin immunoprecipitation sequencing (ChIP-seq) studies, which serve as an ideal means by which to experimentally assess transcription factor binding sites, confirming TF-DNA associations presented here as well as presenting new ones. Promoters strongly associated with song circuit-enriched motifs could then be tested for the ability of these promoters to drive song circuit-selective expression *in vivo* through the use of viral constructs. Thus, the present computational work presents preliminary findings that serve to prioritize which specific transcription factors and tissues may represent the most promising starting points, as well as making specific predictions about which functional associations might be expected.

Bibliography

1. Leiner HC, Leiner AL, Dow RS: **Cognitive and language functions of the human cerebellum.** *Trends in Neurosciences* 1993, **16:**444-447.
2. Stout D, Toth N, Schick K, Chaminade T: **Neural correlates of Early Stone Age toolmaking: technology, language and cognition in human evolution.** *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 2008, **363:**1939-1949.
3. Fisher SE, Marcus GF: **The eloquent ape: genes, brains and the evolution of language.** *Nat Rev Genet* 2006, **7:**9-20.
4. Bolhuis JJ: *Birdsong, speech, and language: exploring the evolution of mind and brain.* MIT press; 2013.
5. Seidenberg MS, Petitto LA: **Signing behavior in apes: A critical review.** *Cognition* 1979, **7:**177-215.
6. Seyfarth R, Cheney D, Marler P: **Monkey responses to three different alarm calls: evidence of predator classification and semantic communication.** *Science* 1980, **210:**801-803.
7. Arnold K, Zuberbühler K: **Language evolution: Semantic combinations in primate calls.** *Nature* 2006, **441:**303-303.
8. Hewes GW, Andrew RJ, Carini L, Choe H, Gardner RA, Kortlandt A, Krantz GS, McBride G, Nottebohm F, Pfeiffer J, et al: **Primate Communication and the Gestural Origin of Language [and Comments and Reply].** *Current Anthropology* 1973, **14:**5-24.
9. Savage-Rumbaugh ES, Wilkerson BJ, Bakeman R: **Spontaneous gestural communication among conspecifics in the pygmy chimpanzee (*Pan paniscus*).** *Progress in ape research* 1977**:**97-116.
10. Tomasello M, Gust D, Frost GT: **A longitudinal investigation of gestural communication in young chimpanzees.** *Primates* 1989, **30:**35-50.
11. Pika S, Liebal K, Tomasello M: **Gestural communication in young gorillas (*Gorilla gorilla*): Gestural repertoire, learning, and use.** *American Journal of Primatology* 2003, **60:**95-111.
12. Janik VM, Slater PJB: **Vocal Learning in Mammals.** *Advances in the Study of Behavior* 1997, **26:**59-100.
13. Zuberbühler K: **Referential Signaling in Non-Human Primates: Cognitive Precursors and Limitations for the Evolution of Language.** In *Advances in the Study of Behavior. Volume* Volume 33: Academic Press; 2003: 265-307
14. Slocombe KE, Zuberbühler K: **Chimpanzees modify recruitment screams as a function of audience composition.** *Proceedings of the National Academy of Sciences* 2007, **104:**17228-17233.
15. Slocombe KE, Zuberbühler K: **Functionally Referential Communication in a Chimpanzee.** *Current Biology* 2005, **15:**1779-1784.
16. Schel AM, Townsend SW, Machanda Z, Zuberbühler K, Slocombe KE: **Chimpanzee Alarm Call Production Meets Key Criteria for Intentionality.** *PLoS ONE* 2013, **8:**e76674.
17. Watson Stuart K, Townsend Simon W, Schel Anne M, Wilke C, Wallace Emma K, Cheng L, West V, Slocombe Katie E: **Vocal Learning in the Functionally Referential Food Grunts of Chimpanzees.** *Current Biology* 2015, **25:**495-499.
18. Nottebohm F: **The Origins of Vocal Learning.** *The American Naturalist* 1972, **106:**116-140.
19. Janik VM, Slater PJB: **The different roles of social learning in vocal communication.** *Animal Behaviour* 2000, **60:**1-11.

20. Jarvis ED: **Learned Birdsong and the Neurobiology of Human Language.** *Annals of the New York Academy of Sciences* 2004, **1016:**749-777.
21. Doupe AJ, Kuhl PK: **Birdsong and Human Speech: Common Themes and Mechanisms.** *Annual Review of Neuroscience* 1999, **22:**567-631.
22. Lilly JC: **Vocal Mimicry in Tursiops: Ability to Match Numbers and Durations of Human Vocal Bursts.** *Science* 1965, **147:**300-301.
23. Knörnschild M, Nagy M, Metz M, Mayer F, von Helversen O: **Complex vocal imitation during ontogeny in a bat.** *Biology Letters* 2010, **6:**156-159.
24. Boughman JW: **Vocal learning by greater spear-nosed bats.** *Proceedings of the Royal Society of London Series B: Biological Sciences* 1998, **265:**227-233.
25. Poole JH, Tyack PL, Stoeger-Horwath AS, Watwood S: **Animal behaviour: Elephants are capable of vocal learning.** *Nature* 2005, **434:**455-456.
26. Sanvito S, Galimberti F, Miller EH: **Observational Evidences of Vocal Learning in Southern Elephant Seals: a Longitudinal Study.** *Ethology* 2007, **113:**137-146.
27. Reichmuth C, Casey C: **Vocal learning in seals, sea lions, and walruses.** *Current Opinion in Neurobiology* 2014, **28:**66-71.
28. Mahrt EJ, Perkel DJ, Tong L, Rubel EW, Portfors CV: **Engineered Deafness Reveals That Mouse Courtship Vocalizations Do Not Require Auditory Experience.** *The Journal of Neuroscience* 2013, **33:**5573-5583.
29. Baptista LF, Schuchmann K-L: **Song learning in the Anna hummingbird (*Calypte anna*).** *Ethology* 1990, **84:**15-26.
30. Sibley CG, Ahlquist JE: *Phylogeny and classification of birds: a study in molecular evolution.* New Haven, CT: Yale University Press; 1991.
31. Hackett SJ, Kimball RT, Reddy S, Bowie RCK, Braun EL, Braun MJ, Chojnowski JL, Cox WA, Han K-L, Harshman J, et al: **A Phylogenomic Study of Birds Reveals Their Evolutionary History.** *Science* 2008, **320:**1763-1768.
32. Suh A, Paus M, Kiefmann M, Churakov G, Franke FA, Brosius J, Kriegs JO, Schmitz J: **Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds.** *Nat Commun* 2011, **2:**443.
33. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al: **Whole-genome analyses resolve early branches in the tree of life of modern birds.** *Science* 2014, **346:**1320-1331.
34. Slater PJB: **Fifty years of bird song research: a case study in animal behaviour.** *Animal Behaviour* 2003, **65:**633-639.
35. Zeigler HP, Marler P: *Neuroscience of Birdsong.* Cambridge: Cambridge University Press; 2008.
36. Ames PL: **The morphology of the syrinx in passerine birds.** *Peabody Mus Bull* 1971, **37:**1-194.
37. Gaunt AS: **An Hypothesis concerning the Relationship of Syringeal Structure to Vocal Abilities.** *The Auk* 1983, **100:**853-862.
38. Suthers R, Goller F, Pytte C: **The neuromuscular control of birdsong.** *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* 1999, **354:**927-939.
39. Kay RF, Cartmill M, Balow M: **The hypoglossal canal and the origin of human vocal behavior.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95:**5417-5419.
40. Brenowitz EA: **Comparative approaches to the avian song system.** *Journal of Neurobiology* 1997, **33:**517-531.

41. Nottebohm F, Stokes TM, Leonard CM: **Central control of song in the canary, *Serinus canarius*.** *The Journal of Comparative Neurology* 1976, **165:**457-486.
42. Takahasi M, Okanoya K: **Song Learning in Wild and Domesticated Strains of White-Rumped Munia, *Lonchura striata*, Compared by Cross-Fostering Procedures: Domestication Increases Song Variability by Decreasing Strain-Specific Bias.** *Ethology* 2010, **116:**396-405.
43. Kroodsma DE: **Songs of the Alder Flycatcher (*Empidonax alnorum*) and Willow Flycatcher (*Empidonax traillii*) Are Innate.** *The Auk* 1984, **101:**13-24.
44. Nottebohm F, Nottebohm ME: **Vocalizations and breeding behaviour of surgically deafened ring doves (Streptopelia risoria).** *Animal Behaviour* 1971, **19:**313-327.
45. Kroodsma DE, Konishi M: **A suboscine bird (eastern phoebe, Sayornis phoebe) develops normal song without auditory feedback.** *Animal Behaviour* 1991, **42:**477-487.
46. Marler P, Tamura M: **Culturally Transmitted Patterns of Vocal Behavior in Sparrows.** *Science* 1964, **146:**1483-1486.
47. Olson CR, Wirthlin M, Lovell PV, Mello CV: **Proper Care, Husbandry, and Breeding Guidelines for the Zebra Finch, Taeniopygia guttata.** *Cold Spring Harbor Protocols* 2014, **2014:**pdb.prot084780.
48. Eens M, Pinxten R, Verheyen RF: **Song learning in captive European starlings, Sturnus vulgaris.** *Animal Behaviour* 1992, **44:**1131-1143.
49. Adret P: **Vocal Learning Induced With Operant Techniques: an Overview.** *Netherlands Journal of Zoology* 1993, **43:**125-142.
50. Tchernichovski O, Mitra PP, Lints T, Nottebohm F: **Dynamics of the Vocal Imitation Process: How a Zebra Finch Learns Its Song.** *Science* 2001, **291:**2564-2569.
51. Mello CV: **The Zebra Finch, *Taeniopygia guttata*: An Avian Model for Investigating the Neurobiological Basis of Vocal Learning.** *Cold Spring Harbor Protocols* 2014, **2014:**pdb.emo084574.
52. Seller TJ: **Midbrain vocalization centres in birds.** *Trends in Neurosciences* 1981, **5:**301-303.
53. Jurgens U: **The role of the periaqueductal grey in vocal behaviour.** *Behav Brain Res* 1994, **62:**107-117.
54. Wild JM, Li D, Eagleton C: **Projections of the dorsomedial nucleus of the intercollicular complex (DM) in relation to respiratory-vocal nuclei in the brainstem of pigeon (*Columba livia*) and zebra finch (*Taeniopygia guttata*).** *Journal of Comparative Neurology* 1997, **377:**392-413.
55. Goodson JL, Bass AH: **Vocal-acoustic circuitry and descending vocal pathways in teleost fish: convergence with terrestrial vertebrates reveals conserved traits.** *J Comp Neurol* 2002, **448:**298-322.
56. Jarvis ED, Mello CV: **Molecular mapping of brain areas involved in parrot vocal communication.** *The Journal of Comparative Neurology* 2000, **419:**1-31.
57. Jarvis ED, Ribeiro S, da Silva ML, Ventura D, Vielliard J, Mello CV: **Behaviourally driven gene expression reveals song nuclei in hummingbird brain.** *Nature* 2000, **406:**628-632.
58. Mello CV, Vicario DS, Clayton DF: **Song presentation induces gene expression in the songbird forebrain.** *Proceedings of the National Academy of Sciences of the United States of America* 1992, **89:**6818-6822.
59. Nottebohm F, Paton JA, Kelley DB: **Connections of vocal control nuclei in the canary telencephalon.** *The Journal of Comparative Neurology* 1982, **207:**344-357.
60. Moore JM, Székely T, Büki J, DeVoogd TJ: **Motor pathway convergence predicts syllable repertoire size in oscine birds.** *Proceedings of the National Academy of Sciences* 2011, **108:**16440-16445.

61. Devoogd TJ, Krebs JR, Healy SD, Purvis A: **Relations between Song Repertoire Size and the Volume of Brain Nuclei Related to Song: Comparative Evolutionary Analyses amongst Oscine Birds.** *Proceedings: Biological Sciences* 1993, **254:**75-82.
62. Karten HJ: **The organization of the ascending auditory pathway in the pigeon (*Columba livia*). I. Diencephalic projections of the inferior colliculus (nucleus mesencephali lateralis, pars dorsalis).** *Brain Research* 1967, **6:**409-427.
63. Kuenzel WJ, Masson M: *A stereotaxic atlas of the brain of the chick (Gallus domesticus).* Baltimore, MD: Johns Hopkins University Press; 1988.
64. Gahr M: **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* 2000, **426:**182-196.
65. Bottjer SW, Halsema KA, Brown SA, Miesner EA: **Axonal connections of a forebrain nucleus involved with vocal learning in zebra finches.** *The Journal of Comparative Neurology* 1989, **279:**312-326.
66. Paton J, Manogue K, Nottebohm F: **Bilateral organization of the vocal control pathway in the budgerigar, Melopsittacus undulatus.** *J Neurosci* 1981, **1:**1279-1288.
67. Durand SE, Heaton JT, Amateau SK, Brauth SE: **Vocal control pathways through the anterior forebrain of a parrot (*Melopsittacus undulatus*).** *The Journal of Comparative Neurology* 1997, **377:**179-206.
68. Brauth SE, Heaton JT, Durand SE, Liang W, Hall WS: **Functional Anatomy of Forebrain Auditory Pathways in the Budgerigar (Melopsittacus undulatus).** *Brain, Behavior and Evolution* 1994, **44:**210-233.
69. Vates GE, Broome BM, Mello CV, Nottebohm F: **Auditory pathways of caudal telencephalon and their relation to the song system of adult male zebra finches (*Taenopygia guttata*).** *The Journal of Comparative Neurology* 1996, **366:**613-642.
70. Reiner A, Perkel DJ, Bruce LL, Butler AB, Csillag A, Kuenzel W, Medina L, Paxinos G, Shimizu T, Striedter G, et al: **Revised nomenclature for avian telencephalon and some related brainstem nuclei.** *The Journal of Comparative Neurology* 2004, **473:**377-414.
71. Jarvis ED, Gunturkun O, Bruce L, Csillag A, Karten H, Kuenzel W, Medina L, Paxinos G, Perkel DJ, Shimizu T, et al: **Avian brains and a new understanding of vertebrate brain evolution.** *Nat Rev Neurosci* 2005, **6:**151-159.
72. Nottebohm F, Arnold AP: **Sexual dimorphism in vocal control areas of the songbird brain.** *Science* 1976, **194:**211-213.
73. Nottebohm F, Kelley DB, Paton JA: **Connections of vocal control nuclei in the canary telencephalon.** *J Comp Neurol* 1982, **207:**344-357.
74. Bottjer SW, Miesner EA, Arnold AP: **Forebrain lesions disrupt development but not maintenance of song in passerine birds.** *Science* 1984, **224:**901-903.
75. Sohrabji F, Nordeen EJ, Nordeen KW: **Selective impairment of song learning following lesions of a forebrain nucleus in the juvenile zebra finch.** *Behav Neural Biol* 1990, **53:**51-63.
76. Scharff C, Nottebohm F: **A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: implications for vocal learning.** *J Neurosci* 1991, **11:**2896-2913.
77. Dum R, Strick P: **The origin of corticospinal projections from the premotor areas in the frontal lobe.** *The Journal of Neuroscience* 1991, **11:**667-689.
78. Martin JH: **The corticospinal system: from development to motor control.** *The Neuroscientist* 2005, **11:**161-173.

79. Jürgens U: **The role of the periaqueductal grey in vocal behaviour.** *Behavioural Brain Research* 1994, **62:**107-117.
80. Jürgens U: **Neuronal Control of Mammalian Vocalization, with Special Reference to the Squirrel Monkey.** *Naturwissenschaften* 1998, **85:**376-388.
81. Jürgens U: **Neural pathways underlying vocal control.** *Neuroscience & Biobehavioral Reviews* 2002, **26:**235-258.
82. Jürgens U: **The Neural Control of Vocalization in Mammals: A Review.** *Journal of Voice* 2009, **23:**1-10.
83. Leonardo A, Fee MS: **Ensemble Coding of Vocal Control in Birdsong.** *J Neurosci* 2005, **25:**652-661.
84. Dave AS, Margoliash D: **Song Replay During Sleep and Computational Rules for Sensorimotor Vocal Learning.** *Science* 2000, **290:**812-816.
85. Brawn TP, Fenn KM, Nusbaum HC, Margoliash D: **Consolidation of sensorimotor learning during sleep.** *Learning & Memory* 2008, **15:**815-819.
86. Shank SS, Margoliash D: **Sleep and sensorimotor integration during early vocal learning in a songbird.** *Nature* 2009, **458:**73-77.
87. Margoliash D: **Sleep, learning, and birdsong.** *ILAR J* 2010, **51:**378-386.
88. Johnson F, Sablan MM, Bottjer SW: **Topographic organization of a forebrain pathway involved with vocal learning in zebra finches.** *The Journal of Comparative Neurology* 1995, **358:**260-278.
89. Luo M, Ding L, Perkel DJ: **An Avian Basal Ganglia Pathway Essential for Vocal Learning Forms a Closed Topographic Loop.** *J Neurosci* 2001, **21:**6836-6845.
90. Lidsky TI, Manetto C, Schneider JS: **A consideration of sensory factors involved in motor functions of the basal ganglia.** *Brain Research* 1985, **356:**133-146.
91. Bottjer SW: **Developmental regulation of basal ganglia circuitry during the sensitive period for vocal learning in songbirds.** *Ann N Y Acad Sci* 2004, **1016:**395-415.
92. Perkel DJ: **Origin of the anterior forebrain pathway.** *Ann N Y Acad Sci* 2004, **1016:**736-748.
93. Bottjer S, Miesner E, Arnold A: **Forebrain lesions disrupt development but not maintenance of song in passerine birds.** *Science* 1984, **224:**901-903.
94. Brainard MS, Doupe AJ: **Interruption of a basal ganglia-forebrain circuit prevents plasticity of learned vocalizations.** *Nature* 2000, **404:**762-766.
95. Hahnloser RHR, Kozhevnikov AA, Fee MS: **An ultra-sparse code underlies the generation of neural sequences in a songbird.** *Nature* 2002, **419:**65-70.
96. Yu AC, Margoliash D: **Temporal Hierarchical Control of Singing in Birds.** *Science* 1996, **273:**1871-1875.
97. Benton S, Cardin JA, DeVoogd TJ: **Lucifer Yellow filling of area X-projecting neurons in the high vocal center of female canaries.** *Brain Research* 1998, **799:**138-147.
98. Nottebohm F: **A brain for all seasons: cyclical anatomical changes in song control nuclei of the canary brain.** *Science* 1981, **214:**1368-1370.
99. Wang N, Hurley P, Pytte C, Kirn JR: **Vocal control neuron incorporation decreases with age in the adult zebra finch.** *The Journal of neuroscience : the official journal of the Society for Neuroscience* 2002, **22:**10864-10870.
100. Dutar P, Vu HM, Perkel DJ: **Multiple Cell Types Distinguished by Physiological, Pharmacological, and Anatomic Properties in Nucleus HVc of the Adult Zebra Finch.** *Journal of Neurophysiology* 1998, **80:**1828-1838.
101. Luo M, Perkel DJ: **Long-range GABAergic projection in a circuit essential for vocal learning.** *J Comp Neurol* 1999, **403:**68-84.

102. Denisenko-Nehrbass NI, Jarvis E, Scharff C, Nottebohm F, Mello CV: **Site-Specific Retinoic Acid Production in the Brain of Adult Songbirds.** *Neuron* 2000, **27:**359-370.
103. Agate RJ, Hertel M, Nottebohm F: **FnTm2, a novel brain-specific transcript, is dynamically expressed in the song learning circuit of the zebra finch.** *The Journal of Comparative Neurology* 2007, **504:**127-148.
104. Fiete IR, Hahnloser RHR, Fee MS, Seung HS: **Temporal Sparseness of the Premotor Drive Is Important for Rapid Learning in a Neural Network Model of Birdsong.** *J Neurophysiol* 2004, **92:**2274-2282.
105. Long MA, Jin DZ, Fee MS: **Support for a synaptic chain model of neuronal sequence generation.** *Nature* 2010, **468:**394-399.
106. Fee MS, Kozhevnikov AA, Hahnloser RHR: **Neural Mechanisms of Vocal Sequence Generation in the Songbird.** *Annals of the New York Academy of Sciences* 2004, **1016:**153-170.
107. Amador A, Perl YS, Mindlin GB, Margoliash D: **Elemental gesture dynamics are encoded by song premotor cortical neurons.** *Nature* 2013, **495:**59-64.
108. Roberts TF, Gobes SMH, Murugan M, Olveczky BP, Mooney R: **Motor circuits are required to encode a sensory model for imitative learning.** *Nat Neurosci* 2012, **15:**1454-1459.
109. Doupe AJ: **A neural circuit specialized for vocal learning.** *Current opinion in neurobiology* 1993, **3:**104-111.
110. Bottjer SW, Miesner EA, Arnold AP: **Changes in neuronal number, density and size account for increases in volume of song-control nuclei during song development in zebra finches.** *Neuroscience letters* 1986, **67:**263-268.
111. DeWulf V, Bottjer SW: **Neurogenesis within the juvenile zebra finch telencephalic ventricular zone: A map of proliferative activity.** *The Journal of Comparative Neurology* 2005, **481:**70-83.
112. Walton C, Pariser E, Nottebohm F: **The Zebra Finch Paradox: Song Is Little Changed, But Number of Neurons Doubles.** *The Journal of Neuroscience* 2012, **32:**761-774.
113. Biddle A, Watson L, Hooper C, Lohr K, Sutton S: **Criteria for Determining Disability in Speech-Language Disorders.** In *Agency for Healtcare Research and Quality (AHRQ Evidence Reports/Technology Assessments).* Rockville, MD; 2002
114. Barry JG, Yasin I, Bishop DVM: **Heritable risk factors associated with language impairments.** *Genes, Brain and Behavior* 2007, **6:**66-76.
115. Airey DC, Castillo-Juarez H, Casella G, Pollak EJ, DeVoogd TJ: **Variation in the volume of zebra finch song control nuclei is heritable: developmental and evolutionary implications.** *Proceedings of the Royal Society of London Series B: Biological Sciences* 2000, **267:**2099-2104.
116. Woodgate JL, Buchanan KL, Bennett ATD, Catchpole CK, Brighton R, Leitner S: **Environmental and Genetic Control of Brain and Song Structure in the Zebra Finch.** *Evolution* 2014, **68:**230-240.
117. Lai CSL, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP: **A forkhead-domain gene is mutated in a severe speech and language disorder.** *Nature* 2001, **413:**519-523.
118. Whitehouse AJO, Bishop DVM, Ang QW, Pennell CE, Fisher SE: **CNTNAP2 variants affect early language development in the general population.** *Genes, Brain and Behavior* 2011, **10:**451-456.
119. Berkel S, Marshall CR, Weiss B, Howe J, Roeth R, Moog U, Endris V, Roberts W, Szatmari P, Pinto D, et al: **Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation.** *Nat Genet* 2010, **42:**489-491.

120. Kang C, Riazuddin S, Mundorff J, Krasnewich D, Friedman P, Mullikin JC, Drayna D: **Mutations in the Lysosomal Enzyme–Targeting Pathway and Persistent Stuttering.** *New England Journal of Medicine* 2010, **362:**677-685.
121. Fisher SE, Lai CSL, Monaco AP: **Deciphering the Genetic Basis of Speech and Language Disorders.** *Annual Review of Neuroscience* 2003, **26:**57-80.
122. Dong S, Replogle KL, Hasadsri L, Imai BS, Yau PM, Rodriguez-Zas S, Southey BR, Sweedler JV, Clayton DF: **Discrete molecular states in the brain accompany changing responses to a vocal signal.** *Proceedings of the National Academy of Sciences* 2009, **106:**11364-11369.
123. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S, et al: **The genome of a songbird.** *Nature* 2010, **464:**757-762.
124. Lovell PV, Clayton DF, Replogle KL, Mello CV: **Birdsong "transcriptomics": neurochemical specializations of the oscine song system.** *PLoS ONE* 2008, **3:**e3440.
125. Wada K, Howard JT, McConnell P, Whitney O, Lints T, Rivas MV, Horita H, Patterson MA, White SA, Scharff C, et al: **A molecular neuroethological approach for identifying and characterizing a cascade of behaviorally regulated genes.** *Proceedings of the National Academy of Sciences* 2006, **103:**15212-15217.
126. Li X, Wang X-J, Tannenhauser J, Podell S, Mukherjee P, Hertel M, Biane J, Masuda S, Nottebohm F, Gaasterland T: **Genomic resources for songbird research and their use in characterizing gene expression during brain development.** *Proceedings of the National Academy of Sciences* 2007, **104:**6834-6839.
127. Replogle K, Arnold A, Ball G, Band M, Bensch S, Brenowitz E, Dong S, Drnevich J, Ferris M, George J, et al: **The Songbird Neurogenomics (SoNG) Initiative: Community-based tools and strategies for study of brain gene function and evolution.** *BMC Genomics* 2008, **9:**131.
128. Lovell PV, Carleton JB, Mello CV: **Genomics analysis of potassium channel genes in songbirds reveals molecular specializations of brain circuits for the maintenance and production of learned vocalizations.** *BMC Genomics* 2013, **14:**470.
129. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S, et al: **The genome of a songbird.** *Nature* 2010, **464:**757-762.
130. Nam K, Mugal C, Nabholz B, Schielzeth H, Wolf JB, Backstrom N, Kunstner A, Balakrishnan CN, Heger A, Ponting CP, et al: **Molecular evolution of genes in avian genomes.** *Genome Biol* 2010, **11:**R68.
131. Haesler S, Wada K, Nshdejan A, Morrisey EE, Lints T, Jarvis ED, Scharff C: **FoxP2 Expression in Avian Vocal Learners and Non-Learners.** *J Neurosci* 2004, **24:**3164-3175.
132. Teramitsu I, Kudo LC, London SE, Geschwind DH, White SA: **Parallel FoxP1 and FoxP2 expression in songbird and human brain predicts functional interaction.** *J Neurosci* 2004, **24:**3152-3163.
133. White SA, Fisher SE, Geschwind DH, Scharff C, Holy TE: **Singing Mice, Songbirds, and More: Models for FOXP2 Function and Dysfunction in Human Speech and Language.** *J Neurosci* 2006, **26:**10376-10379.
134. Haesler S, Rochefort C, Georgi B, Licznerski P, Osten P, Scharff C: **Incomplete and Inaccurate Vocal Imitation after Knockdown of *FoxP2* in Songbird Basal Ganglia Nucleus Area X.** *PLoS Biol* 2007, **5:**e321.
135. Panaitof SC, Abrahams BS, Dong H, Geschwind DH, White SA: **Language-related *Cntnap2* gene is differentially expressed in sexually dimorphic song nuclei essential for vocal learning in songbirds.** *The Journal of Comparative Neurology* 2010, **518:**1995-2018.
136. Condro MC, White SA: **Distribution of language-related Cntnap2 protein in neural circuits critical for vocal learning.** *Journal of Comparative Neurology* 2014, **522:**169-185.

137. Vernes SC, Newbury DF, Abrahams BS, Winchester L, Nicod J, Groszer M, Alarcón M, Oliver PL, Davies KE, Geschwind DH, et al: **A Functional Genetic Link between Distinct Developmental Language Disorders.** *New England Journal of Medicine* 2008, **359:**2337-2345.
138. Kos M, van den Brink D, Snijders TM, Rijpkema M, Franke B, Fernandez G, Hagoort P: **CNTNAP2 and Language Processing in Healthy Individuals as Measured with ERPs.** *PLoS ONE* 2012, **7:**e46995.
139. Whitney O, Pfenning AR, Howard JT, Blatti CA, Liu F, Ward JM, Wang R, Audet J-N, Kellis M, Mukherjee S, et al: **Core and region-enriched networks of behaviorally regulated genes and the singing genome.** *Science* 2014, **346**.
140. Feduccia A: *The origin and evolution of birds.* Yale University Press; 1999.
141. Fowler ME, Miller RE: *Zoo and wild animal medicine: current therapy.* Elsevier Health Sciences; 2008.
142. Lespinet O, Wolf YI, Koonin EV, Aravind L: **The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes.** *Genome Research* 2002, **12:**1048-1059.
143. Dennis Megan Y, Nuttle X, Sudmant Peter H, Antonacci F, Graves Tina A, Nefedov M, Rosenfeld Jill A, Sajjadian S, Malig M, Kotkiewicz H, et al: **Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication.** *Cell* 2012, **149:**912-922.
144. Hillier LW, Miller W, Ewan Birney , Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME, et al: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432:**695-716.
145. Kong L, Lovell PV, Heger A, Mello CV, Ponting CP: **Accelerated Evolution of PAK3- and PIM1-like Kinase Gene Families in the Zebra Finch, *Taeniopygia guttata*.** *Molecular Biology and Evolution* 2010, **27:**1923-1934.
146. Lomsadze A, Burns PD, Borodovsky M: **Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm.** *Nucleic Acids Research* 2014, **42:**e119.
147. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al: **Comparative genomics reveals insights into avian genome evolution and adaptation.** *Science* 2014, **346:**1311-1320.
148. Donthu R, Lewin H, Larkin D: **SyntenyTracker: a tool for defining homologous synteny blocks using radiation hybrid maps and whole-genome sequence.** *BMC Research Notes* 2009, **2:**148.
149. Völker M, Backström N, Skinner BM, Langley EJ, Bunzey SK, Ellegren H, Griffin DK: **Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution.** *Genome Research* 2010, **20:**503-511.
150. Peterson M, Whittaker D, Ambreth S, Sureshchandra S, Buechlein A, Podicheti R, Choi J-H, Lai Z, Mockatis K, Colbourne J, et al: **De novo transcriptome sequencing in a songbird, the dark-eyed junco (Junco hyemalis): genomic tools for an ecological model system.** *BMC Genomics* 2012, **13:**305.
151. Jarvis ED, Yu J, Rivas MV, Horita H, Feenders G, Whitney O, Jarvis SC, Jarvis ER, Kubikova L, Puck AEP, et al: **Global view of the functional molecular organization of the avian cerebrum: Mirror images and functional columns.** *Journal of Comparative Neurology* 2013, **521:**3614-3665.
152. Chen C-C, Winkler CM, Pfenning AR, Jarvis ED: **Molecular profiling of the developing avian telencephalon: Regional timing and brain subdivision continuities.** *Journal of Comparative Neurology* 2013, **521:**3666-3701.

153. Simpson H, Vicario D: **Brain pathways for learned and unlearned vocalizations differ in zebra finches.** *J Neurosci* 1990, **10:**1541-1556.
154. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, et al: **The Ensembl genome database project.** *Nucleic Acids Research* 2002, **30:**38-41.
155. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M: **The Ensembl Automatic Gene Annotation System.** *Genome Research* 2004, **14:**942-950.
156. Lynch M, Conery JS: **The Evolutionary Fate and Consequences of Duplicate Genes.** *Science* 2000, **290:**1151-1155.
157. Lynch M: **Gene Duplication and Evolution.** *Science* 2002, **297:**945-947.
158. Skinner BM, Griffin DK: **Intrachromosomal rearrangements in avian genome evolution: evidence for regions prone to breakpoints.** *Heredity* 2012, **108:**37-41.
159. Lupski JR, Stankiewicz P: **Genomic Disorders: Molecular Mechanisms for Rearrangements and Conveyed Phenotypes.** *PLoS Genet* 2005, **1:**e49.
160. Itoh Y, Arnold A: **Chromosomal polymorphism and comparative painting analysis in the zebra finch.** *Chromosome Research* 2005, **13:**47-56.
161. Yuri T, Kimball RT, Braun EL, Braun MJ: **Duplication of Accelerated Evolution and Growth Hormone Gene in Passerine Birds.** *Molecular Biology and Evolution* 2008, **25:**352-361.
162. Han K, Lee J, Meyer TJ, Remedios P, Goodwin L, Batzer MA: **L1 recombination-associated deletions generate human genomic variation.** *Proceedings of the National Academy of Sciences* 2008, **105:**19366-19371.
163. Stankiewicz P, Lupski JR: **Genome architecture, rearrangements and genomic disorders.** *Trends in Genetics* 2002, **18:**74-82.
164. Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA: **Emergence of primate genes by retrotransposon-mediated sequence transduction.** *Proceedings of the National Academy of Sciences* 2006, **103:**17608-17613.
165. Lovell PV, Clayton DF, Replogle KL, Mello CV: **Birdsong "Transcriptomics": Neurochemical Specializations of the Oscine Song System.** *PLoS ONE* 2008, **3:**e3440.
166. Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, Roulhac PL, Howard JT, Wirthlin M, Lovell PV, Ganapathy G, et al: **Convergent transcriptional specializations in the brains of humans and song-learning birds.** *Science* 2014, **346:**1333.
167. Thuresson B, Westman JS, Olsson ML: **Identification of a novel A4GALT exon reveals the genetic basis of the P1/P2 histo-blood groups.** *Blood* 2011, **117:**678-687.
168. Szwergold BS: **Fructosamine-6-phosphates are deglycated by phosphorylation to fructosamine-3,6-bisphosphates catalyzed by fructosamine-3-kinase (FN3K) and/or fructosamine-3-kinase-related-protein (FN3KRP).** *Medical Hypotheses* 2007, **68:**37-45.
169. Zhang Z, Futamura M, Vikis HG, Wang M, Li J, Wang Y, Guan K-L, You M: **Positional cloning of the major quantitative trait locus underlying lung tumor susceptibility in mice.** *Proceedings of the National Academy of Sciences* 2003, **100:**12642-12647.
170. Goldman SA, Nottebohm F: **Neuronal production, migration, and differentiation in a vocal control nucleus of the adult female canary brain.** *Proceedings of the National Academy of Sciences* 1983, **80:**2390-2394.
171. Alvarez-Buylla A, Nottebohm F: **Migration of young neurons in adult avian brain.** *Nature* 1988, **335:**353-354.
172. Zhang Z, Theler D, Kaminska KH, Hiller M, de la Grange P, Pudimat R, Rafalska I, Heinrich B, Bujnicki JM, Allain FH-T, Stamm S: **The YTH Domain Is a Novel RNA Binding Domain.** *Journal of Biological Chemistry* 2010, **285:**14701-14710.
173. Morohashi K, Sahara H, Watashi K, Iwabata K, Sunoki T, Kuramochi K, Takakusagi K, Miyashita H, Sato N, Tanabe A, et al: **Cyclosporin A Associated Helicase-Like Protein**

Facilitates the Association of Hepatitis C Virus RNA Polymerase with Its Cellular Cyclophilin B. *PLoS ONE* 2011, **6:**e18285.
174. Walbott H, Mouffok S, Capeyrou R, Lebaron S, Humbert O, van Tilbeurgh H, Henry Y, Leulliot N: **Prp43p contains a processive helicase structural architecture with a specific regulatory domain.** *The EMBO Journal* 2010, **29:**2194-2204.
175. Knowles DG, McLysaght A: **Recent de novo origin of human protein-coding genes.** *Genome Research* 2009, **19:**1752-1759.
176. Drickamer K: **C-type lectin-like domains.** *Current Opinion in Structural Biology* 1999, **9:**585-590.
177. Cambi A, Figdor CG: **Dual function of C-type lectin-like receptors in the immune system.** *Current Opinion in Cell Biology* 2003, **15:**539-546.
178. Jessell TM: **Adhesion molecules and the hierarchy of neural development.** *Neuron* 1988, **1:**3-13.
179. Striedter GF: **The vocal control pathways in budgerigars differ from those in songbirds.** *The Journal of Comparative Neurology* 1994, **343:**35-56.
180. Kent WJ: **BLAT—The BLAST-Like Alignment Tool.** *Genome Research* 2002, **12:**656-664.
181. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006.
182. Ganapathy G, Howard JT, Ward JM, Li J, Li B, Li Y, Xiong Y, Zhang Y, Zhou S, Schwarz D, et al: **High-coverage sequencing and annotated assemblies of the budgerigar genome.** *Gigascience*, **in press**.
183. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer.** *Nat Biotech* 2011, **29:**24-26.
184. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al: **Ensembl BioMarts: a hub for data retrieval across taxonomic space.** *Database (Oxford)* 2011, **2011:**bar030.
185. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30:**1236-1240.
186. Jarvis ED, Nottebohm F: **Motor-driven gene expression.** *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94:**4097-4102.
187. Carleton JB, Lovell PV, McHugh A, Marzulla T, Horback KL, Mello CV: **An Optimized Protocol for High-Throughput In Situ Hybridization of Zebra Finch Brain.** *Cold Spring Harbor Protocols* 2014.
188. Paton JA, Manogue KR, Nottebohm F: **Bilateral organization of the vocal control pathway in the budgerigar, Melopsittacus undulatus.** *J Neurosci* 1981, **1:**1279-1288.
189. Striedter GF: **The vocal control pathways in budgerigars differ from those in songbirds.** *Journal of Comparative Neurology* 1994, **343:**35-56.
190. Durand SE, Heaton JT, Amateau SK, Brauth SE: **Vocal control pathways through the anterior forebrain of a parrot (*Melopsittacus undulatus*).** *Journal of Comparative Neurology* 1997, **377:**179-206.
191. Brauth SE, Heaton JT, Shea SD, Durand SE, Hall WS: **Functional anatomy of forebrain vocal control pathways in the budgerigar (*Melopsittacus undulatus*).** *Annals of the New York Academy of Science* 1997, **807:**368-385.
192. Brauth SE, Liang W, Roberts TF: **Projections of the oval nucleus of the hyperstriatum ventrale in the budgerigar: relationships with the auditory system.** *J Comp Neurol* 2001, **432:**481-511.

193. Ferreira ARJ, Smulders TV, Sameshima K, Mello CV, Jarvis ED: **Vocalizations and associated behaviors of the Sombre Hummingbird (*Aphantochroa cirrhochloris*) and the Rufous-breasted Hermit (*Glaucis hirsutus*).** *The Auk* 2006, **123:**1129-1148.
194. Plummer TK, Striedter GF: **Auditory responses in the vocal motor system of budgerigars.** *Journal of Neurobiology* 2000, **42:**79-94.
195. Prather JF, Peters S, Nowicki S, Mooney R: **Precise auditory-vocal mirroring in neurons for learned vocal communication.** *Nature* 2008, **451:**305-310.
196. Clayton DF: **Role of gene regulation in song circuit development and song learning.** *Journal of Neurobiology* 1997, **33:**549-571.
197. Wada K, Sakaguchi H, Jarvis ED, Hagiwara M: **Differential expression of glutamate receptors in avian neural pathways for learned vocalization.** *The Journal of Comparative Neurology* 2004, **476:**44-64.
198. Ball GF, Faris PL, Hartman BK, Wingfield JC: **Immunohistochemical localization of neuropeptides in the vocal control regions of two songbird species.** *Journal of Comparative Neurology* 1988, **268:**171-180.
199. Ball GF, Nock B, McEwen BS, Balthazart J: **Distribution of alpha 2-adrenergic receptors in the brain of the Japanese quail as determined by quantitative autoradiography: implications for the control of sexually dimorphic reproductive processes.** *Brain Research* 1989, **491:**68-79.
200. Ball GF, Nock B, Wingfield JC, McEwen BS, Balthazart J: **Muscarinic cholinergic receptors in the songbird and quail brain: a quantitative autoradiographic study.** *Journal of Comparative Neurology* 1990, **298:**431-442.
201. Cookson KK, Hall WS, Heaton JT, Brauth SE: **Distribution of choline acetyltransferase and acetylcholinesterase in vocal control nuclei of the budgerigar (Melopsittacus undulatus).** *J Comp Neurol* 1996, **369:**220-235.
202. Durand SE, Brauth SE, Liang W: **Calcitonin gene-related peptide immunoreactive cells and fibers in forebrain vocal and auditory nuclei of the budgerigar (Melopsittacus undulatus).** *Brain Behav Evol* 2001, **58:**61-79.
203. Roberts TF, Cookson KK, Heaton KJ, Hall WS, Brauth SE: **Distribution of tyrosine hydroxylase-containing neurons and fibers in the brain of the budgerigar (Melopsittacus undulatus): general patterns and labeling in vocal control nuclei.** *J Comp Neurol* 2001, **429:**436-454.
204. Eales LA: **Song learning in zebra finches: some effects of song model availability on what is learnt and when.** *Animal Behavior* 1985, **33:**1293-1300.
205. Bohner J, Chaiken ML, Ball GF, Marler P: **Song acquisition in photosensitive and photorefractory male European starlings.** *Hormones and Behavior* 1990, **24:**582-594.
206. Clayton DF: **Role of gene regulation in song circuit development and song learning.** *Journal of Neurobiology* 1997, **33:**549-571.
207. Nordeen KW, Nordeen EJ: **Synaptic and molecular mechanisms regulating plasticity during early learning.** *Ann N Y Acad Sci* 2004, **1016:**416-437.
208. Nottebohm F, Nottebohm ME, Crane L: **Developmental and seasonal changes in canary song and their relation to changes in the anatomy of song-control nuclei.** *Behavioral and Neural Biology* 1986, **46:**445-471.
209. Smith GT, Brenowitz EA, Wingfield JC, Baptista LF: **Seasonal changes in song nuclei and song behavior in Gambel's white-crowned sparrows.** *Journal of Neurobiology* 1995, **28:**114-125.

210. Farabaugh SM, Linzenbold A, Dooling RJ: **Vocal plasticity in budgerigars (Melopsittacus undulatus): evidence for social factors in the learning of contact calls.** *J Comp Psychol* 1994, **108:**81-92.
211. Farabaugh SM BE, and Dooling RJ: **Analysis of warble song of the budgerigar Melopsittacus undulatus.** *Bioacoustics* 1992, **4:**111-130.
212. Tu HW, Smith EW, Dooling RJ: **Acoustic and perceptual categories of vocal elements in the warble song of budgerigars (Melopsittacus undulatus).** *J Comp Psychol* 2011, **125:**420-430.
213. Tu HW, Osmanski MS, Dooling RJ: **Learned vocalizations in budgerigars (Melopsittacus undulatus): the relationship between contact calls and warble song.** *J Acoust Soc Am* 2011, **129:**2289-2297.
214. Eda-Fujiwara H WA, Okumura H: **Effects of deafening on temporal pattern of vocalizations in the budgerigar Melopsittacus undulatus.** *J Ethol* 1995, **13:**145-152.
215. Stiles F: **Aggressive and courtship displays of the male Anna's Hummingbird.** *Condor* 1982, **84:**208-225.
216. Baptista LF, Schuchmann KL: **Song learning in the anna hummingbird (*Calypte anna*).** *Ethology* 1990, **84:**15-26.
217. Ferreira AR, Smulders TV, Sameshima K, Mello CV, Jarvis ED: **Vocalizations and Associated Behaviors of the Sombre Hummingbird (Aphantochroa Cirrhochloris) and the Rufous-Breasted Hermit (Glaucis Hirsutus).** *Auk* 2006, **123:**1129-1148.
218. Ficken MS, Rusch KM, Taylor SJ, Powers DR: **Reproductive behaviora and communication in blue-throated hummingbirds.** *The Wilson Bulletin* 2002, **114:**197-209.
219. Mirsky EN: **Song divergence in hummingbird and junco populations on guadalupe island.** *Condor* 1976, **78:**230-235.
220. Baptista LFM, M.: **The source of the dive-noise of the Anna's Hummingbir.** *Condor* 1979, **81:**87-89.
221. Snow DW: **The singing assemblies of little hermits phaethornis-longuemareus.** *The Wilson Bulletin* 1968, **85:**163-177.
222. Snow DW: **Comparison of the leks of Guy's hermit hummmingbird Phaethornis guy in Costa Rica and Trinidad.** *Ibis* 1977, **119:**211-214.
223. Wiley RH: **Song Groups in a Singing Assembly of Little Hermits.** *The Condor* 1971, **73:**28-35.
224. Gaunt SLL, Baptista LF, Sanchez JE, Hernandez D: **Song learning as evidence from song sharing in two hummingbird species (*Colibri coruscans* and *C. thalassinus*).** *The Auk* 1994, **111:**87-103.
225. Kubota M, Taniguchi I: **Electrophysiological Characteristics of Classes of Neuron in the HVc of the Zebra Finch.** *Journal of Neurophysiology* 1998, **80:**914-923.
226. Mooney R: **Different Subthreshold Mechanisms Underlie Song Selectivity in Identified HVc Neurons of the Zebra Finch.** *J Neurosci* 2000, **20:**5420-5436.
227. Hara E, Rivas MV, Ward JM, Okanoya K, Jarvis ED: **Convergent Differential Regulation of Parvalbumin in the Brains of Vocal Learners.** *PLoS ONE* 2012, **7:**e29457.
228. Chakraborty M, Walløe S, Nedergaard S, Fridel EE, Dabelsteen T, Pakkenberg B, Bertelsen MF, Dorrestein GM, Brauth SE, Durand SE, Jarvis ED: **Core and Shell Song Systems Unique to the Parrot Brain.** *PLoS ONE* 2015, **10:**e0118496.
229. Kroodsma DE: **Male eastern phoebes (Sayornis phoebe; Tyrannidae, Passeriformes) fail to imitate songs.** *Journal of Comparative Psychology* 1989, **103:**227-232.

230. Gahr M, Güttinger H-R, Kroodsma DE: **Estrogen receptors in the avian brain: Survey reveals general distribution and forebrain areas unique to songbirds.** *The Journal of Comparative Neurology* 1993, **327:**112-122.
231. Saldanha CJ, Schultz JD, London SE, Schlinger BA: **Telencephalic Aromatase but Not a Song Circuit in a Sub-Oscine Passerine, the Golden Collared Manakin (Manacus vitellinus).** *Brain, Behavior and Evolution* 2000, **56:**29-37.
232. Ohlson JI, Irestedt M, Ericson PG, Fjeldså J: **Phylogeny and classification of the New World suboscines (Aves, Passeriformes).** *Zootaxa* 2013, **3613:**1-35.
233. Saranathan V, Hamilton D, Powell GVN, Kroodsma D, Prum RO: **Genetic evidence supports song learning in the three-wattled bellbird *Procnias tricarunculata* (Cotingidae).** *Molecular Ecology* 2007, **16:**3689-3702.
234. Kroodsma D, Hamilton D, Sánchez JE, Byers BE, Fandiño-Mariño H, Stemple DW, Trainer JM, Powell GVN: **Behavioral Evidence for Song Learning in the Suboscine Bellbirds (Procnias spp.; Cotingidae).** *The Wilson Journal of Ornithology* 2013, **125:**1-14.
235. Trainer JM, McDonald DB, Learn WA: **The development of coordinated singing in cooperatively displaying long-tailed manakins.** *Behav Ecol* 2002, **13:**65-69.
236. Tobias JA, Seddon N: **Signal Jamming Mediates Sexual Conflict in a Duetting Bird.** *Current Biology* 2009, **19:**577-582.
237. Amador A, Trevisan MA, Mindlin GB: **Simple neural substrate predicts complex rhythmic structure in duetting birds.** *Physical Review E* 2005, **72:**031905.
238. Laje R, Mindlin GB: **Highly Structured Duets in the Song of the South American Hornero.** *Physical Review Letters* 2003, **91:**258104.
239. Fortune ES, Rodríguez C, Li D, Ball GF, Coleman MJ: **Neural Mechanisms for the Coordination of Duet Singing in Wrens.** *Science* 2011, **334:**666-670.
240. Liu W-c, Wada K, Jarvis ED, Nottebohm F: **Rudimentary substrates for vocal learning in a suboscine.** *Nat Commun* 2013, **4**.
241. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements.** *Nat Rev Genet* 2004, **5:**276-287.
242. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, et al: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotech* 2005, **23:**137-144.
243. Lombardino AJ, Hertel M, Li X-C, Haripal B, Martin-Harris L, Pariser E, Nottebohm F: **Expression profiling of intermingled long-range projection neurons harvested by laser capture microdissection.** *Journal of Neuroscience Methods* 2006, **157:**195-207.
244. Drnevich J, Replogle KL, Lovell P, Hahn TP, Johnson F, Mast TG, Nordeen E, Nordeen K, Strand C, London SE, et al: **Impact of experience-dependent and -independent factors on gene expression in songbird brain.** *Proceedings of the National Academy of Sciences* 2012, **109:**17245-17252.
245. Ramachandran B, Schlinger BA, Arnold AP, Campagnoni AT: **Zebra finch aromatase gene expression is regulated in the brain through an alternate promoter.** *Gene* 1999, **240:**209-216.
246. Gasper JS, Shiina T, Inoko H, Edwards SV: **Songbird Genomics: Analysis of 45 kb Upstream of a Polymorphic Mhc Class II Gene in Red-Winged Blackbirds (Agelaius phoeniceus).** *Genomics* 2001, **75:**26-34.
247. Saab SS, Lange HS, Maney DL: **Gonadotrophin-Releasing Hormone Neurones in a Photoperiodic Songbird Express Fos and Egr-1 Protein After a Single Long Day.** *Journal of Neuroendocrinology* 2010, **22:**196-207.

248. London SE, Clayton DF: **Genomic and neural analysis of the estradiol-synthetic pathway in the zebra finch.** *BMC Neuroscience* 2010, **11:**1-16.
249. Lovell PV, Kasimi B, Carleton J, Velho TA, Mello CV: **Living without DAT: Loss and compensation of the dopamine transporter gene in sauropsids (birds and reptiles).** *Scientific Reports* 2015, **5:**14093.
250. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, et al: **CAGE: cap analysis of gene expression.** *Nat Meth* 2006, **3:**211-222.
251. Ohler U, Liao G-c, Niemann H, Rubin GM: **Computational analysis of core promoters in the Drosophila genome.** *Genome Biology* 2002, **3:**1-12.
252. Jin VX, Singer GA, Agosto-Pérez FJ, Liyanarachchi S, Davuluri RV: **Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs.** *BMC Bioinformatics* 2006, **7:**1-13.
253. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engstrom PG, Frith MC, et al: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38:**626-635.
254. Davuluri RV, Grosse I, Zhang MQ: **Computational identification of promoters and first exons in the human genome.** *Nat Genet* 2001, **29:**412-417.
255. Lombardino AJ, Li X-C, Hertel M, Nottebohm F: **Replaceable neurons and neurodegenerative disease share depressed UCHL1 levels.** *Proceedings of the National Academy of Sciences* 2005, **102:**8036-8041.
256. Abe H, Gemmell NJ: **Abundance, arrangement, and function of sequence motifs in the chicken promoters.** *BMC Genomics* 2014, **15:**1-12.
257. Gardiner-Garden M, Frommer M: **CpG islands in vertebrate genomes.** *J Mol Biol* 1987, **196**.
258. Valen E, Sandelin A: **Genomic and chromatin signals underlying transcription start-site selection.** *Trends in Genetics* 2011, **27:**475-485.
259. Bailey TL, Machanick P: **Inferring direct DNA binding from ChIP-seq.** *Nucleic Acids Research* 2012, **40:**e128.
260. McLeay RC, Bailey TL: **Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data.** *BMC Bioinformatics* 2010, **11:**1-11.
261. Nichols M, Weih F, Schmid W, DeVack C, Kowenz-Leutz E, Luckow B, Boshart M, Schütz G: **Phosphorylation of CREB affects its binding to high and low affinity sites: implications for cAMP induced gene transcription.** *The EMBO Journal* 1992, **11:**3337-3346.
262. Jiang J, Levine M: **Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen.** *Cell* 1993, **72:**741-752.
263. Tanay A: **Extensive low-affinity transcriptional interactions in the yeast genome.** *Genome Research* 2006, **16:**962-972.
264. Mello C, Clayton D: **Song-induced ZENK gene expression in auditory pathways of songbird brain and its relation to the song control system.** *J Neurosci* 1994, **14:**6652-6666.
265. Mello CV, Velho TAF, Pinaud R: **Song-induced gene expression: A window on song auditory processing and perception.** *Ann N Y Acad Sci* 2004, **1016**.
266. Cao X, Mahendran R, Guy GR, Tan YH: **Detection and characterization of cellular EGR-1 binding to its recognition site.** *J Biol Chem* 1993, **268:**16949-16957.
267. Mello C, Nottebohm F, Clayton D: **Repeated exposure to one song leads to a rapid and persistent decline in an immediate early gene's response to that song in zebra finch telencephalon.** *J Neurosci* 1995, **15:**6919-6925.

268. Vernes SC, MacDermot KD, Monaco AP, Fisher SE: **Assessing the impact of FOXP1 mutations on developmental verbal dyspraxia.** *Eur J Hum Genet* 2009, **17:**1354-1358.
269. Horn D, Kapeller J, Rivera-Brugués N, Moog U, Lorenz-Depiereux B, Eck S, Hempel M, Wagenstaller J, Gawthrope A, Monaco AP, et al: **Identification of FOXP1 deletions in three unrelated patients with mental retardation and significant speech and language deficits.** *Human Mutation* 2010, **31:**E1851-E1860.
270. Hamdan FF, Daoud H, Rochefort D, Piton A, Gauthier J, Langlois M, Foomani G, Dobrzeniecka S, Krebs M-O, Joober R, et al: **De Novo Mutations in *FOXP1* in Cases with Intellectual Disability, Autism, and Language Impairment.** *The American Journal of Human Genetics* 2010, **87:**671-678.
271. Carr CW, Moreno-De-Luca D, Parker C, Zimmerman HH, Ledbetter N, Martin CL, Dobyns WB, Abdul-Rahman OA: **Chiari I malformation, delayed gross motor skills, severe speech delay, and epileptiform discharges in a child with FOXP1 haploinsufficiency.** *Eur J Hum Genet* 2010, **18:**1216-1220.
272. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, MacKenzie AP, Ng SB, Baker C, et al: **Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations.** *Nat Genet* 2011, **43:**585-589.
273. Talkowski Michael E, Rosenfeld Jill A, Blumenthal I, Pillalamarri V, Chiang C, Heilbut A, Ernst C, Hanscom C, Rossin E, Lindgren AM, et al: **Sequencing Chromosomal Abnormalities Reveals Neurodevelopmental Loci that Confer Risk across Diagnostic Boundaries.** *Cell* 2012, **149:**525-537.
274. Bacon C, Rappold GA: **The distinct and overlapping phenotypic spectra of FOXP1 and FOXP2 in cognitive disorders.** *Human Genetics* 2012, **131:**1687-1698.
275. Chen Q, Heston JB, Burkett ZD, White SA: **Expression analysis of the speech-related genes FoxP1 and FoxP2 and their relation to singing behavior in two songbird species.** *Journal of Experimental Biology* 2013, **216:**3682-3692.
276. Tamura S, Morikawa Y, Iwanishi H, Hisaoka T, Senba E: **Foxp1 gene expression in projection neurons of the mouse striatum.** *Neuroscience* 2004, **124:**261-267.
277. Bacon C, Schneider M, Le Magueresse C, Froehlich H, Sticht C, Gluch C, Monyer H, Rappold GA: **Brain-specific Foxp1 deletion impairs neuronal development and causes autistic-like behaviour.** *Mol Psychiatry* 2015, **20:**632-639.
278. Mooney R, Prather JF: **The HVC Microcircuit: The Synaptic Basis for Interactions between Song Motor and Vocal Plasticity Pathways.** *J Neurosci* 2005, **25:**1952-1964.
279. Spiro JE, Dalva MB, Mooney R: **Long-Range Inhibition Within the Zebra Finch Song Nucleus RA Can Coordinate the Firing of Multiple Projection Neurons.** *Journal of Neurophysiology* 1999, **81:**3007-3020.
280. Cooper SJ, Trinklein ND, Nguyen L, Myers RM: **Serum response factor binding sites differ in three human cell types.** *Genome Research* 2007, **17:**136-144.
281. Miano JM: **Serum response factor: toggling between disparate programs of gene expression.** *Journal of Molecular and Cellular Cardiology* 2003, **35:**577-593.
282. Miano JM, Long X, Fujiwara K: **Serum response factor: master regulator of the actin cytoskeleton and contractile apparatus.** *American Journal of Physiology - Cell Physiology* 2007, **292:**C70-C81.
283. Johansen F-E, Prywes R: **Serum response factor: transcriptional regulation of genes induced by growth factors and differentiation.** *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1995, **1242:**1-10.
284. Niu Z, Li A, Zhang SX, Schwartz RJ: **Serum response factor micromanaging cardiogenesis.** *Current Opinion in Cell Biology* 2007, **19:**618-627.

285. Vlahopoulos S, Zimmer WE, Jenster G, Belaguli NS, Balk SP, Brinkmann AO, Lanz RB, Zoumpourlis VC, Schwartz RJ: **Recruitment of the Androgen Receptor via Serum Response Factor Facilitates Expression of a Myogenic Gene.** *Journal of Biological Chemistry* 2005, **280:**7786-7792.
286. Lu PPY, Ramanan N: **Serum Response Factor Is Required for Cortical Axon Growth But Is Dispensable for Neurogenesis and Neocortical Lamination.** *The Journal of Neuroscience* 2011, **31:**16651-16664.
287. Arnold AP, Nottebohm F, Pfaff DW: **Hormone concentrating cells in vocal control and other areas of the brain of the zebra finch (Poephila guttata).** *The Journal of Comparative Neurology* 1976, **165:**487-511.
288. Schlinger BA, Arnold AP: **Androgen effects on the development of the zebra finch song system.** *Brain Research* 1991, **561:**99-105.
289. Wade J, Arnold AP: **Sexual Differentiation of the Zebra Finch Song System.** vol. 1016. pp. 540-559: Blackwell Publishing Ltd; 2004:540-559.
290. Nottebohm F, Arnold A: **Sexual dimorphism in vocal control areas of the songbird brain.** *Science* 1976, **194:**211-213.
291. Alvarez-Buylla A, Kirn JR: **Birth, migration, incorporation, and death of vocal control neurons in adult songbirds.** *J Neurobiol* 1997, **33:**585-601.
292. Cogliati T, Good DJ, Haigney M, Delgado-Romero P, Eckhaus MA, Koch WJ, Kirsch IR: **Predisposition to Arrhythmia and Autonomic Dysfunction in Nhlh1-Deficient Mice.** *Molecular and Cellular Biology* 2002, **22:**4977-4983.
293. Murdoch JN, Eddleston J, Leblond-Bourget N, Stanier P, Copp AJ: **Sequence and expression analysis of Nhlh1: a basic helix-loop-helix gene implicated in neurogenesis.** *Dev Genet* 1999, **24:**165-177.
294. Fritzsch B, Eberl DF, Beisel KW: **The role of bHLH genes in ear development and evolution: revisiting a 10-year-old hypothesis.** *Cellular and Molecular Life Sciences* 2010, **67:**3089-3099.
295. Duncan MK, Bordas L, Dicicco-Bloom E, Chada KK: **Expression of the helix-loop-helix genes Id-1 and NSCL-1 during cerebellar development.** *Developmental Dynamics* 1997, **208:**107-114.
296. Theodorakis K, Kyriakopoulou K, Wassef M, Karagogeos D: **Novel sites of expression of the bHLH gene NSCL1 in the developing nervous system.** *Mechanisms of Development* 2002, **119, Supplement:**S103-S106.
297. Nguyen T, Di Giovanni S: **NFAT signaling in neural development and axon growth.** *International Journal of Developmental Neuroscience* 2008, **26:**141-145.
298. Graef IA, Wang F, Charron F, Chen L, Neilson J, Tessier-Lavigne M, Crabtree GR: **Neurotrophins and Netrins Require Calcineurin/NFAT Signaling to Stimulate Outgrowth of Embryonic Axons.** *Cell* 2003, **113:**657-670.
299. Graef IA, Chen F, Crabtree GR: **NFAT signaling in vertebrate development.** *Current Opinion in Genetics & Development* 2001, **11:**505-512.
300. Wu Y, Borde M, Heissmeyer V, Feuerer M, Lapan AD, Stroud JC, Bates DL, Guo L, Han A, Ziegler SF, et al: **FOXP3 controls regulatory T cell function through cooperation with NFAT.** *Cell* 2006, **126:**375-387.
301. Briand-Suleau A, Martinovic J, Tosca L, Tou B, Brisset S, Bouligand J, Delattre V, Giurgea I, Bachir J, Folliot P, et al: **SALL4 and NFATC2: Two major actors of interstitial 20q13.2 duplication.** *European Journal of Medical Genetics* 2014, **57:**174-180.

302. Blanc P, Gouas L, Francannet C, Giollant M, Vago P, Goumy C: **Trisomy 20q caused by interstitial duplication 20q13.2: Clinical report and literature review.** *American Journal of Medical Genetics Part A* 2008, **146A:**1307-1311.
303. Amiel J, Rio M, Pontual Ld, Redon R, Malan V, Boddaert N, Plouin P, Carter NP, Lyonnet S, Munnich A, Colleaux L: **Mutations in *TCF4*, Encoding a Class I Basic Helix-Loop-Helix Transcription Factor, Are Responsible for Pitt-Hopkins Syndrome, a Severe Epileptic Encephalopathy Associated with Autonomic Dysfunction.** *The American Journal of Human Genetics* 2007, **80:**988-993.
304. Zweier C, Peippo MM, Hoyer J, Sousa S, Bottani A, Clayton-Smith J, Reardon W, Saraiva J, Cabral A, Göhring I, et al: **Haploinsufficiency of *TCF4* Causes Syndromal Mental Retardation with Intermittent Hyperventilation (Pitt-Hopkins Syndrome).** *The American Journal of Human Genetics*, **80:**994-1001.
305. Brzózka MM, Radyushkin K, Wichert SP, Ehrenreich H, Rossner MJ: **Cognitive and Sensorimotor Gating Impairments in Transgenic Mice Overexpressing the Schizophrenia Susceptibility Gene *Tcf4* in the Brain.** *Biological Psychiatry* 2010, **68:**33-40.
306. David MD, Yeramian A, Duñach M, Llovera M, Cantí C, de Herreros AG, Comella JX, Herreros J: **Signalling by neurotrophins and hepatocyte growth factor regulates axon morphogenesis by differential β-catenin phosphorylation.** *Journal of Cell Science* 2008, **121:**2718-2730.
307. Jolma A, Yan J, Whitington T, Toivonen J, Nitta Kazuhiro R, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al: **DNA-Binding Specificities of Human Transcription Factors.** *Cell* 2013, **152:**327-339.
308. Weirauch Matthew T, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi Hamed S, Lambert Samuel A, Mann I, Cook K, et al: **Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity.** *Cell* 2014, **158:**1431-1443.
309. Frietze S, Lan X, Jin VX, Farnham PJ: **Genomic Targets of the KRAB and SCAN Domain-containing Zinc Finger Protein 263.** *Journal of Biological Chemistry* 2010, **285:**1393-1403.
310. Chiocchetti AG, Kopp M, Waltes R, Haslinger D, Duketis E, Jarczok TA, Poustka F, Voran A, Graab U, Meyer J, et al: **Variants of the CNTNAP2 5[prime] promoter as risk factors for autism spectrum disorders: a genetic and functional approach.** *Mol Psychiatry* 2015, **20:**839-849.
311. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: **A census of human transcription factors: function, expression and evolution.** *Nat Rev Genet* 2009, **10:**252-263.
312. Kidd M, Nadler B, Mane S, Eick G, Malfertheiner M, Champaneria M, Pfragner R, Modlin I: **GeneChip, geNorm, and gastrointestinal tumors: novel reference genes for real-time PCR.** *Physiological Genomics* 2007, **30:**363-370.
313. Jiao K, Zhou Y, Hogan BLM: **Identification of mZnf8, a Mouse Krüppel-Like Transcriptional Repressor, as a Novel Nuclear Interaction Partner of Smad1.** *Molecular and Cellular Biology* 2002, **22:**7633-7644.
314. Landgren H, Carlsson P: **Foxj3, a novel mammalian forkhead gene expressed in neuroectoderm, neural crest, and myotome.** *Developmental Dynamics* 2004, **231:**396-401.
315. Dimitrieva S, Bucher P: **UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks.** *Nucleic Acids Research* 2013, **41:**D101-D109.

316. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3[prime] UTRs by comparison of several mammals.** *Nature* 2005, **434:**338-345.
317. Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJ, Wilde A, Brudno M, et al: **Conservation of core gene expression in vertebrate tissues.** *Journal of Biology* 2009, **8:**1-17.
318. Meyer TJ, McLain AT, Oldenburg JM, Faulk C, Bourgeois MG, Conlin EM, Mootnick AR, de Jong PJ, Roos C, Carbone L, Batzer MA: **An Alu-Based Phylogeny of Gibbons (Hylobatidae).** *Molecular Biology and Evolution* 2012, **29:**3441-3450.
319. Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ, Herrero J, Roos C, Aken B, et al: **Gibbon genome and the fast karyotype evolution of small apes.** *Nature* 2014, **513:**195-201.
320. Bekpen C, Tastekin I, Siswara P, Akdis CA, Eichler EE: **Primate segmental duplication creates novel promoters for the LRRC37 gene family within the 17q21.31 inversion polymorphism region.** *Genome Research* 2012, **22:**1050-1058.
321. Giannuzzi G, Siswara P, Malig M, Marques-Bonet T, Program NCS, Mullikin JC, Ventura M, Eichler EE: **Evolutionary dynamism of the primate LRRC37 gene family.** *Genome Research* 2013, **23:**46-59.
322. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.
323. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** In *Current Protocols in Bioinformatics.* John Wiley & Sons, Inc.; 2002
324. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML: **UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions.** *Nucleic Acids Research* 2015, **43:**D117-D122.
325. Mathelier A, Fornes O, Arenillas DJ, Chen C-y, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R, et al: **JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles.** *Nucleic Acids Research* 2016, **44:**D110-D115.
326. Bailey TL, Gribskov M: **Combining evidence using p-values: application to sequence homology searches.** *Bioinformatics* 1998, **14:**48-54.
327. Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif.** *Bioinformatics* 2011, **27:**1017-1018.
328. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al: **STRING v10: protein–protein interaction networks, integrated over the tree of life.** *Nucleic Acids Research* 2015, **43:**D447-D452.
329. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The gene ontology consortium.** *Nat Genet* 2000, **25**.
330. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Research* 2000, **28:**27-30.
331. Nottebohm F: **Brain Pathways for Vocal Learning in Birds: A Review of the First 10 Years.** In *Progress in Psychobiology and Physiological Psychology. Volume* 9. Edited by Sprague JMS, Epstein ANE. New York: Academic Press; 1980: 85-125
332. Darwin C: *The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* 1st edn. London: John Murray; 1859.

333. Feenders G, Liedvogel M, Rivas M, Zapka M, Horita H, Hara E, Wada K, Mouritsen H, Jarvis ED: **Molecular Mapping of Movement-Associated Areas in the Avian Brain: A Motor Theory for Vocal Learning Origin.** *PLoS ONE* 2008, **3:**e1768.
334. Jarvis ED: **Evolution of brain pathways for vocal learning in birds and humans.** *Birdsong, speech, and language: exploring the evolution of mind and brain* 2013**:**63-107.
335. Gould SJ: **Punctuated Equilibrium's Threefold History.** 2002.
336. Kutschera U, Niklas KJ: **Macroevolution via secondary endosymbiosis: a Neo-Goldschmidtian view of unicellular hopeful monsters and Darwin's primordial intermediate form.** *Theory in Biosciences* 2008, **127:**277-289.
337. Hauser MD, Chomsky N, Fitch WT: **The faculty of language: what is it, who has it, and how did it evolve?** *Science* 2002, **298:**1569-1579.
338. Petkov CI, Jarvis ED: **Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates.** In *Frontiers in evolutionary neuroscience*, vol. 4. pp. 12; 2012:12.
339. Arriaga G, Zhou EP, Jarvis ED: **Of Mice, Birds, and Men: The Mouse Ultrasonic Song System Has Some Features Similar to Humans and Song-Learning Birds.** *PLoS ONE* 2012, **7:**e46610.
340. Mayr E: *The growth of biological thought: Diversity, evolution, and inheritance.* Harvard University Press; 1982.
341. Roberts LH: **The rodent ultrasound production mechanism.** *Ultrasonics* 1975, **13:**83-88.
342. Roberts LH: **Evidence for the laryngeal source of ultrasonic and audible cries of rodents.** *Journal of Zoology* 1975, **175:**243-257.
343. Blumberg MS, Alberts JR: **Ultrasonic vocalizations by rat pups in the cold: An acoustic by-product of laryngeal braking?** *Behavioral Neuroscience* 1990, **104:**808-817.
344. Blumberg MS: **Rodent ultrasonic short calls: Locomotion, biomechanics, and communication.** *Journal of Comparative Psychology* 1992, **106:**360-365.
345. Burgdorf J, Kroes RA, Moskal JR, Pfaus JG, Brudzynski SM, Panksepp J: **Ultrasonic vocalizations of rats (Rattus norvegicus) during mating, play, and aggression: Behavioral concomitants, relationship to reward, and self-administration of playback.** *Journal of Comparative Psychology* 2008, **122:**357-367.
346. Mällo T, Matrov D, Herm L, Kõiv K, Eller M, Rinken A, Harro J: **Tickling-induced 50-kHz ultrasonic vocalization is individually stable and predicts behaviour in tests of anxiety and depression in rats.** *Behavioural Brain Research* 2007, **184:**57-71.
347. Hammerschmidt K, Reisinger E, Westekemper K, Ehrenreich L, Strenzke N, Fischer J: **Mice do not require auditory input for the normal development of their ultrasonic vocalizations.** *BMC Neuroscience* 2012, **13:**40.
348. Kao MH, Doupe AJ, Brainard MS: **Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song.** *Nature* 2005, **433:**638-643.
349. Ölveczky BP, Andalman AS, Fee MS: **Vocal Experimentation in the Juvenile Songbird Requires a Basal Ganglia Circuit.** *PLoS Biol* 2005, **3:**e153.
350. Wu HG, Miyamoto YR, Castro LNG, Olveczky BP, Smith MA: **Temporal structure of motor variability is dynamically regulated and predicts motor learning ability.** *Nat Neurosci* 2014, **17:**312-321.
351. Liu W-c, Wada K, Nottebohm F: **Variable Food Begging Calls Are Harbingers of Vocal Learning.** *PLoS ONE* 2009, **4:**e5929.
352. Colombelli-Négrel D, Hauber Mark E, Robertson J, Sulloway Frank J, Hoi H, Griggio M, Kleindorfer S: **Embryonic Learning of Vocal Passwords in Superb Fairy-Wrens Reveals Intruder Cuckoo Nestlings.** *Current Biology* 2012, **22:**2155-2160.

353. Colombelli-Négrel D, Hauber ME, Kleindorfer S: **Prenatal learning in an Australian songbird: habituation and individual discrimination in superb fairy-wren embryos.** *Proceedings of the Royal Society of London B: Biological Sciences* 2014, **281**.
354. Colombelli-Négrel D, Webster MS, Dowling JL, Hauber ME, Kleindorfer S: **Vocal imitation of mother's calls by begging Red-backed Fairywren nestlings increases parental provisioning.** *The Auk* 2016, **133:**273-285.
355. Ali F, Otchy Timothy M, Pehlevan C, Fantana Antoniu L, Burak Y, Ölveczky Bence P: **The Basal Ganglia Is Necessary for Learning Spectral, but Not Temporal, Features of Birdsong.** *Neuron* 2013, **80:**494-506.
356. Genome 10K Community of Scientists: **Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species.** *Journal of Heredity* 2009, **100:**659-674.
357. Wang R, Chen C-C, Hara E, Rivas MV, Roulhac PL, Howard JT, Chakraborty M, Audet J-N, Jarvis ED: **Convergent differential regulation of SLIT-ROBO axon guidance genes in the brains of vocal learners.** *Journal of Comparative Neurology* 2015, **523:**892-906.
358. Olson CR, Hodges LK, Mello CV: **Dynamic gene expression in the song system of zebra finches during the song learning period.** *Developmental Neurobiology* 2015, **75:**1315-1338.
359. Seddon N, Tobias JA: **Duets defend mates in a suboscine passerine, the warbling antbird (Hypocnemis cantator).** *Behav Ecol* 2006, **17:**73-83.
360. Benjamini Y, Speed TP: **Summarizing and correcting the GC content bias in high-throughput sequencing.** *Nucleic Acids Research* 2012, **40:**e72.
361. Salzberg SL, Yorke JA: **Beware of mis-assembled genomes.** *Bioinformatics* 2005, **21:**4320-4321.
362. Laine VN, Gossmann TI, Schachtschneider KM, Garroway CJ, Madsen O, Verhoeven KJF, de Jager V, Megens H-J, Warren WC, Minx P, et al: **Evolutionary signals of selection on cognition from the great tit genome and methylome.** *Nat Commun* 2016, **7**.
363. Frankl-Vilches C, Kuhl H, Werber M, Klages S, Kerick M, Bakker A, de Oliveira EH, Reusch C, Capuano F, Vowinckel J, et al: **Using the canary genome to decipher the evolution of hormone-sensitive gene regulation in seasonal singing birds.** *Genome Biology* 2015, **16:**1-25.
364. Yokoyama KD, Zhang Y, Ma J: **Tracing the Evolution of Lineage-Specific Transcription Factor Binding Sites in a Birth-Death Framework.** *PLoS Comput Biol* 2014, **10:**e1003771.