2021

The Role of DNA Methylation in Defining the Vocal Learning Transcriptome of the Zebra Finch

Caitlin Sun Gilbert

The Role of DNA Methylation in Defining the Vocal Learning Transcriptome of the Zebra Finch

A Thesis Presented to the Faculty of
The Rockefeller University
in Partial Fulfillment of the Requirements for
the degree of Doctor of Philosophy

by
Caitlin Sun Gilbert
June 2021

THE ROLE OF DNA METHYLATION IN DEFINING THE VOCAL LEARNING TRANSCRIPTOME OF THE ZEBRA FINCH

Caitlin Gilbert, Ph.D.

The Rockefeller University 2021

Vocal learning is a rare, complex behavior which is a critical component of human spoken language acquisition. It is convergent across several independent lineages of birds and mammals, including songbirds and humans. The development of speech and song production in humans and songbirds is strikingly similar, though the molecular mechanisms underlying these similarities are not yet understood. Our lab has previously found convergently differentially expressed genes in the vocal learning circuitry of humans and song-learning birds relative to adjacent non-vocal motor circuits. Most notably, the RA song nucleus in the songbird is molecularly convergent with the human laryngeal motor cortex. What remains unknown, however, is how such striking molecular convergences came to be. One likely mechanism by which differential expression can be established in the brain is DNA methylation, an epigenomic signature that is canonically associated with transcriptional repression. In my thesis work, I examined whether this differential specialized gene expression in the zebra finch, one of the most well-studied songbirds, is associated with differential DNA methylation. To do so, I performed whole-genome bisulfite sequencing in the RA song nucleus and its adjacent non-vocal motor region. I found that the convergent specialized transcriptome in this nucleus is in part defined by DNA methylation in a subset of these genes. A significant proportion of downregulated genes in RA exhibit increased gene body methylation relative to the adjacent non-vocal motor region. I additionally profiled the potential writers and readers of DNA methylation in the zebra finch brain to explore mechanisms for establishing the

methylome over development. Strikingly, I found that the de novo writer of DNA methylation, *DNMT3A*, is specifically upregulated in the RA relative to the surrounding arcopallium only at post-hatch day ~60, not earlier or in adulthood. These findings indicate that differential DNA methylation could be contributing to specialization of gene down regulation, and thereby influence circuit specializations of vocal learning brain circuits. This study represents the first unbiased, basepair-resolution, genome-wide analysis of DNA methylation in the songbird as well as one of the first brain methylome studies in a non-mammalian vertebrate. Understanding the genomic mechanisms for vocal learning in the songbird will expand our understanding of speech and language disorders in humans, especially those that are congenital.

For Mama and Baba

ACKNOWLEDGMENTS

The work that has culminated in this thesis would not have been possible without the incredible scientific mentors I have had both before and during my time in graduate school. First, to my PhD advisor, Dr. Erich Jarvis, thank you for allowing your students to take on novel and ambitious projects and ask bold and paradigm-shifting questions; I know my thesis work could not have happened in any other lab. My brilliant committee members, Dr. David Allis and Dr. Daniel Kronauer, have guided my work with their insightful feedback and curiosity. I will forever appreciate how their kindness and support have shaped my time as a scientist at Rockefeller. I am also indebted to my mentor in the Greenberg lab, Dr. Harrison Gabel, for being a role model for the kind of scientist I wanted to be during my PhD: tenacious and deeply thoughtful about the science while always supporting the work and growth of others.

The wildly diverse projects within the Jarvis lab fueled both direct and indirect collaborations without which this work would not have been possible. To Greg, Ha Na, Lindsey, Matt, Matthew, and James, thank you for your direct contributions and, in several cases, training to make my work possible. A special thanks to Dr. Samara Brown for being a constant source of support, both scientifically and emotionally, for me and many others in the lab. There is no way I would have survived the past six years without colleagues who I've also been so privileged to call my friends: César, Steph, Brigid, Lindsey, Greg, and Matt—thank you for keeping the lab weird and fun; I'm going to deeply miss having you all as labmates.

I have also been so privileged to be a part of the Rockefeller community. The single best part of my time here has been my outreach work as a co-director of the Summer Neuroscience Program—to my fellow co-directors past and present as well as the entire RockEdu team that we so loved collaborating with, you are doing the most important work on campus, and I can't wait to

see all you will accomplish. Much of my motivation in lab stemmed from the time I spent learning from and teaching brilliant SNP high schoolers every summer. I also owe so much of my success in graduate school to the constant support and kindness from the Dean's Office, without which I would be extremely lost and likely barely functioning—to Cris, Stephanie, Marta, Emily, Kristen, and Sid, you are all champions for students, and I have been so lucky to have you all on my team.

My time as a PhD student has been the most challenging period of my life, and I would not have been able to accomplish all that I have let alone overcome the most difficult parts without the constant presence and support of my family and friends. Finally, to Jeremy, your love and steadfast encouragement are the reason I have been able to make it through the past few years of my PhD—thank you for being the best dog dad to Henry, too.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

PFP (posterior forebrain pathway)

AFP (anterior forebrain pathway)

HVC (nidopallial song nucleus)

RA (robust nucleus of the arcopallium)

LMAN/MAN (lateral magnocellular nucleus of the anterior nidopallium)

Area X (the medial striatal song nucleus)

DLM (the thalamic nucleus)

LMC (laryngeal motor cortex)

ASt (anterior striatum)

aT (anterior thalamus)

NIf (nucleus interfacialis of the nidopallium)

Av (avalanche)

MO (mesopallium oval)

PLN (posterior lateral nidopallium)

PLMV (posterior-lateral ventral mesopallium)

AN (anterior nidopallium)

AMV (anterior ventral mesopallium)

LAI (lateral interior arcopallium)

VSt (ventral striatum)

mC (methylated cytosine)

mCH (methylated CA, CC, or CT)

mCG (methylated CG)

hmC (hydroxymethylated cytosine)

MBD (methyl-binding domain)

WGBS (whole-genome bisulfite sequencing)

LCM (laser capture microscopy)

PHD (post-hatch day)

VGP (Vertebrate Genomes Project)

DMR (differentially methylated region)

DMS (differentially methylated site)

ISH (in situ hybridization)

fISH (fluorescent in situ hybridization)

RRBS (reduced-representation bisulfite sequencing)

FDR (false discovery rate)

TPM (transcripts per million)

CHAPTER I: BACKGROUND AND INTRODUCTION

1.1 Vocal learning and speech acquisition

1.1.1 What is vocal learning?

Vocal production learning is the ability to produce novel vocalizations based on auditory input. It is a complex learned behavior involving the ability to imitate sounds and modify the acoustics of these sounds through a vocal organ such as the larynx or syrinx. It is a rare trait in the animal kingdom, having only been documented in three of over forty orders of birds (parrots, hummingbirds, and songbirds)[1,2]; and four of over thirty orders of mammals (cetaceans [dolphins, whales][3], bats[4], elephants[5], and humans). Notably, among primates, only humans have vocal production learning. Other types of acoustic learning include vocal usage learning and auditory learning, which are much more common[6,7]. Vocal usage learning involves learning when to use specific vocalizations (innate or learned) in certain contexts (e.g. knowing when to use an alarm call versus a mating call). Auditory learning is the ability to perceptually learn sounds and then associate these sounds with a given stimulus (e.g. a dog knowing to sit when hearing the command "Sit"). Many vertebrates are capable of either vocal usage learning or auditory learning, including the five groups of mammals and three groups of birds that have vocal production learning. The significant evolutionary distance across these different vocal learning groups of animals suggests that vocal learning is a product of convergent evolution[8,9]. For simplicity, I will use "vocal learning" to refer to vocal production learning and "usage learning" to refer to vocal usage learning.

Early in life, human language acquisition rests on various auditory inputs, mirroring the development of vocal learning in other vocal learning-capable species. In the zebra finch, a

songbird that is the most studied model for vocal learning[1,10], vocal learning manifests as song production, while in humans, it is speech production. With this analogous behavior and behavioral development between humans and songbirds, understanding the vocal learning mechanisms in the songbird brain could elucidate a better understanding of the genetic and molecular underpinnings of human language acquisition. Moreover, such studies could provide insight into speech and language disorders, including those that exist in the context of autism-spectrum disorders.

1.1.2 Analogous development of vocal learning in songbirds and humans

In both songbirds and humans, the ability to imitate vocalizations matures over post-hatch/postnatal development. Song or speech learning occurs in two overlapping stages, called the sensory and sensorimotor phases (Fig 1.1)[11]. The sensory phase involves song or speech perception and auditory learning, while the sensorimotor phase involves the beginnings of song or speech production, i.e. learned vocalizations. In the sensory phase for zebra finches, a young juvenile bird listens to and memorizes the song of an adult tutor, often the bird's father, to form an auditory template for the song[11]. Then, during sensorimotor learning, the juvenile begins to sing and gradually learns to produce the correct song by refining its own vocalizations to match the tutor's template. Throughout this phase, the juvenile bird uses auditory feedback to learn to match its immature vocalizations to the template song[11]. Towards the beginning of the sensorimotor phase, at around 30 days post-hatch, the zebra finch begins to produce an immature song known as subsong that is similar to babbling sounds produced by 6-8-month old babies[1,10]. During the sensory phase, both juvenile birds and human babies listen to various features of the song or speech sounds in their environment like prosodic cues (e.g. pitch, duration, loudness changes), transitional probabilities (i.e. the probability that one syllable will follow another), and the environment of the

critical period (e.g. presence or absence of a parent's vocalizations). Both birds and humans also have a critical period marked by the end of the sensory phase—at this point, it becomes difficult for humans to easily learn new languages (and accents) and for birds to learn a new song. In humans, the sensorimotor phase starts at around 10 months, at which point babies begin to produce language-specific speech. By around 1-2 years, the babies will be able to produce meaningful speech (Fig 1.1).

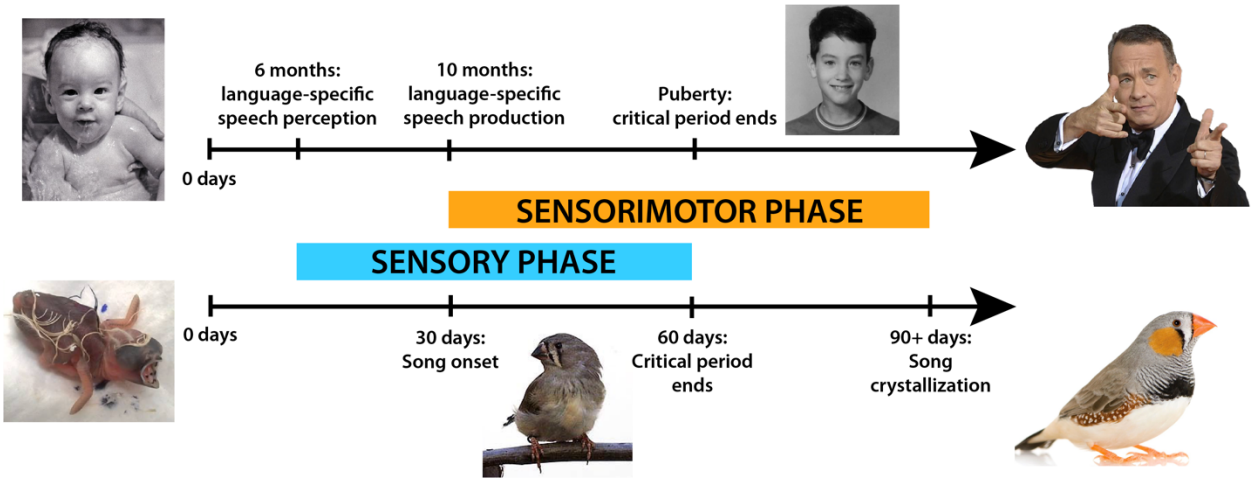


Figure 1.1. Analogous development of vocal learning in humans and songbirds

Timeline for development of vocal learning in the zebra finch with sensory and sensorimotor phases mapped out is adapted from Solis et al[11]. Timepoints listed for both human and finch development mark the key milestones during speech or song learning. In both the top and bottom timelines, speech and song production critical period end points are listed with pictures indicating development from newborn to adolescent to adult stages, from left to right.

Of course, there are several key differences in the development of speech and song. Songbirds have another critical period at the end of the sensorimotor phase, between 60-90 days post-hatch, where they crystallize their song; after this point, the zebra finch can only produce this one song with limited variability. This song crystallization marks a key difference between humans

and songbirds in their respective capacities for vocal learning: humans are open-ended vocal learners in that they can learn new languages, albeit with significantly greater difficulty, after this sensorimotor period, while zebra finches are closed-ended vocal learners, unable to learn new songs after this critical period. Moreover, only male, not female, zebra finches are vocal learners, while speech is much less sexually dimorphic in humans[12].

1.2 Convergence of speech and song neural pathways in the avian and human brains

1.2.1 Circuitry underlying both speech and song production

The parallels between the development of human and zebra finch vocal learning behavior are accompanied by parallels in brain circuitry: in the brains of both vocal learners, there are distinct regions that specifically control the production of either song or speech. In the zebra finch, these regions that underlie song production are called nuclei and are divided into two key subpathways, the posterior forebrain pathway (PFP) and the anterior forebrain pathway (AFP). The PFP consists of the nidopallial song nucleus (HVC) projecting to the robust nucleus of the arcopallium (RA), which in turn projects directly to the XII motor nucleus in the brainstem that controls the muscles of the syrinx to produce song (Fig 1.2, solid black and red arrows on left)[2,13]. The AFP is a pathway connecting the lateral magnocellular nucleus of the anterior nidopallium (LMAN), the medial striatal song nucleus (Area X), and the thalamic nucleus DLM (Fig.1.2, solid white arrows on left)[2,13]. While the PFP is important for the direct motor control of song production, the AFP is involved in song acquisition and modification of song. The AFP interfaces with the PFP via projections from LMAN to RA and HVC to Area X (Fig 1.2, dotted arrows on left)[2,13]. In the human brain, the proposed PFP consists of direct projections from the laryngeal motor cortex (LMC) to the nucleus ambiguous in the brainstem which controls the muscles of the

larynx (Fig 1.2, red arrow on right), while the AFP consists of Broca's area, the anterior striatum (ASt), and anterior thalamus (aT) (Fig 1.2, white arrows on right)[2,13].

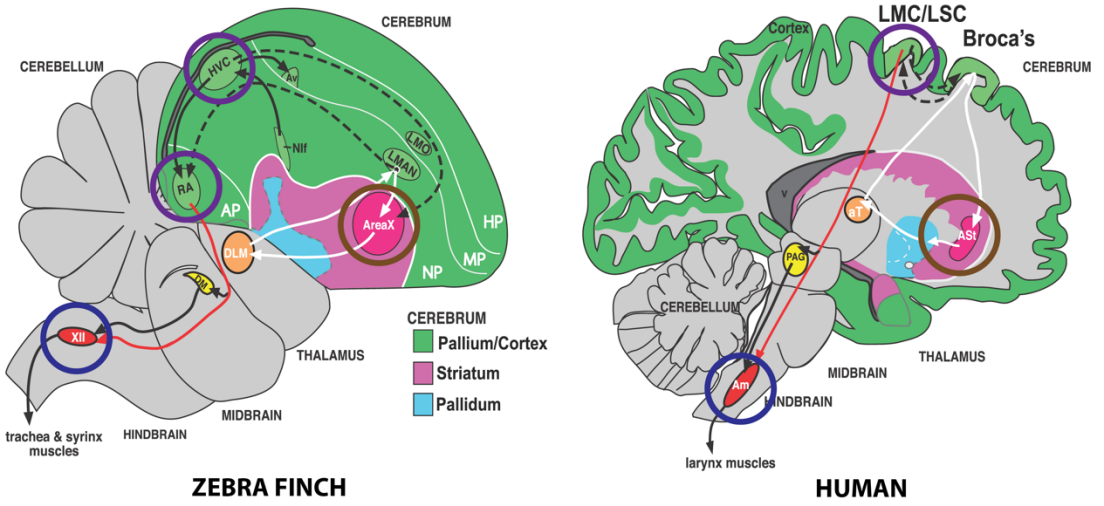


Figure 1.2. Convergent molecular specializations of vocal learning brain regions

Figure adapted from Pfenning et al[13] to show analogous vocal learning circuitry in the zebra finch and human brain. Black arrows indicate the PFP, red arrows indicate the direct projection within the PFP from vocal motor cortex regions to the brainstem, white arrows indicate the AFP, and dotted arrows indicate projections between the PFP and AFP. In the human brain, PFP and AFP are proposed pathways. On left, thin white lines denote pallial subdivisions of the finch brain: AP = arcopallium, NP = nidopallium, MP = mesopallium, HP = hyperpallium. Thick purple, blue, and brown circles indicate regions that have been shown to be molecularly convergent via microarray and RNA-seq data[13,14]—matching colors signify regions that are molecularly convergent between zebra finch and human, e.g. RA and HVC in the finch are convergent with LMC in the human.

To hypothesize how these specialized brain circuits developed independently in songbirds and humans, the motor theory of vocal learning was proposed: it posits that the vocal learning pathways evolved out of a pre-existing motor learning pathway[15] such that broader motor regions

subspecialize for song and speech functions[15–17]. This theory was in part based on activity-dependent gene expression found in the zebra finch exclusively during singing but not other complex motor behaviors like hopping[15]: during singing, the song nuclei are active, while during hopping, the surrounding motor regions are active (Fig 1.3).

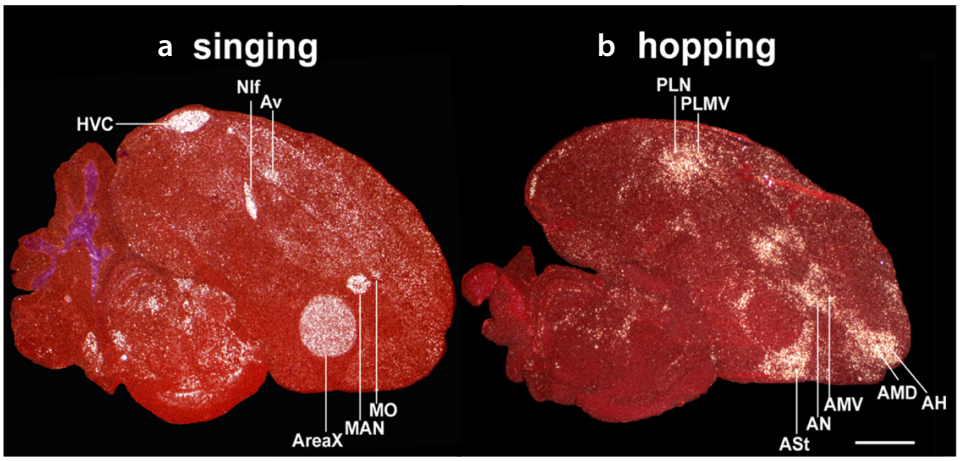


Figure 1.3. Specialized vocal learning circuitry in the songbird brain[15]

Figure from Feenders et al.[15] shows in situ hybridization images for the expression of immediate early genes (c-fos in a and ZENK in b), of zebra finch brain (adult male, midsagittal sections) while the animal is singing (a) or moving (b). In (a), the section is from a bird singing while not moving, while in (b), the section is from a deafened bird hopping in the dark (birds are silent in the dark). Labeled regions in (a) are six of the seven vocal nuclei (HVC, NIf = Nucleus Interfacialis of the Nidopallium, Av = Avalanche, AreaX, MAN, MO = Mesopallium Oval; RA is the seventh nucleus not shown here, as it is lateral to this sagittal plane), while labeled regions in (b) are five adjacent movement-associated areas (PLN = Posterior Lateral Nidopallium, PLMV = Posterior-Lateral Ventral Mesopallium, ASt = Anterior Striatum, AN = Anterior Nidopallium, AMV = Anterior Ventral Mesopallium) for each of these vocal nuclei, respectively (ie. ASt is the movement associated region adjacent to AreaX in the striatum).

This regional specialization also supports the use of the surrounding motor regions as controls to define the molecular specializations of the speech and song circuits, as the key biological difference between the nuclei and surround is the vocal learning functionality in the former. The respective surrounding motor regions for the four central vocal nuclei in the zebra finch (RA, HVC, Area X, and LMAN) are the lateral intermediate arcopallium (LAI), the postlateral nidopallium (PLN), ventral striatum (VSt), and the anterior nidopallium (AN).

1.2.2 Specialized transcriptomes underlying speech and song production pathways

Using the framework of specialized vocal learning regions as defined by their surrounding motor regions, our lab has shown over the past several years that not only are these analogous structures in the zebra finch and human brains functionally similar, but that they also share transcriptomic specializations. Microarray-based and, more recently, RNA-seq-based studies have found that the RA song production nucleus in the zebra finch is most transcriptomically similar to the LMC in humans (Fig 1.2, purple circles): the specialized gene program underlying this convergence consists of ~100 genes[13,14] that are differentially expressed relative to the adjacent motor regions. Moreover, the latest findings from the lab indicate that the HVC nucleus also shares this molecular convergence with the LMC[14] (Fig 1.2, purple circles). These gene sets are enriched for specialized functions like axon guidance, synaptic transmission, and neural protection, which are correlated with the specific phenotypic functions of these regions in the songbird[13,14]. Importantly, these gene expression specializations are not found in the brains of non-human primates or vocal non-learning birds[13].

1.3 DNA methylation in the brain

1.3.1 Why study DNA methylation in the songbird brain?

Unraveling the transcriptomic specializations required to build vocal learning circuitry in the brain requires the exploration of the regulatory mechanisms that establish these gene expression changes. Gene expression itself is not merely the presence or absence of a gene, but the dynamic result of cell intrinsic (e.g. chromatin changes) and extrinsic factors (e.g. secreted proteins, oxygen, and small molecules) that act as a network of dimmer switches to control the amount of transcripts produced for any given gene. This vast and complex network of interactions between the genome, epigenome, proteome, and other regulatory components determine when and where specific genes are activated and the amount of protein product produced. Especially for cells in the brain, many of these changes occur over critical developmental windows to set up more stable transcriptomes that are perhaps less in flux, yet still retain a historical footprint of the changes that have occurred via specific epigenomic marks. A complete transcriptome can be extremely powerful in understanding gene expression, but it's in looking at this epigenomic footprint that the mechanisms underlying this transcriptome are best understood.

One of the many features of this epigenomic footprint is a form of cell intrinsic gene regulation, DNA methylation, which interfaces directly with the process of transcription to slow the rate of transcript production. DNA methylation, a repressive, covalent modification of genomic cytosine residues, provides a level of regulatory information outside of the genomic sequence (i.e. the genomic sequence itself is unchanged). Recent work has shown DNA methylation to be a critical epigenomic signature of cellular diversity in the developing and adult mammalian brain, as it defines differential expression in cell types and is required for proper brain development[18–20]. Importantly, one of the best predictors of cell type in the mammalian brain is gene body DNA

methylation[19,21,22]. Indeed, the specialized transcriptome underlying the unique cell type heterogeneity in the song production regions in the zebra finch brain[23] highlights the importance of establishing the interacting dynamics between the transcriptome and methylome in the finch.

1.3.2 DNA methylation in the mammalian brain

Cytosine methylation is the most well-studied form of DNA methylation. In mammals, the modification primarily exists in the CG context (mCG) (Fig 1.4a): the mCG mark has been shown to play a role in transcriptional regulation[24], cellular differentiation[24,25], and the development of various diseases, including neurological disorders[26] and cancer[27,28]. Despite the fact that cytosine methylation in other dinucleotide contexts (mCH, where H is A, C, or T) (Fig 1.4a) was established in the 1980s[29], mCH was dismissed as a technological artifact until its biological importance in mice and humans[30–32] was conclusively shown. The coincident development of higher-resolution methylome sequencing technologies like whole-genome bisulfite sequencing paved the way for these and more recent discoveries. In the past few years, work has shown that other DNA residues besides cytosine, like adenine, can also be methylated (i.e. mA instead of mC), including in mammalian embryonic stem cells[33] and brains, where it is required for fear extinction in mice[34].

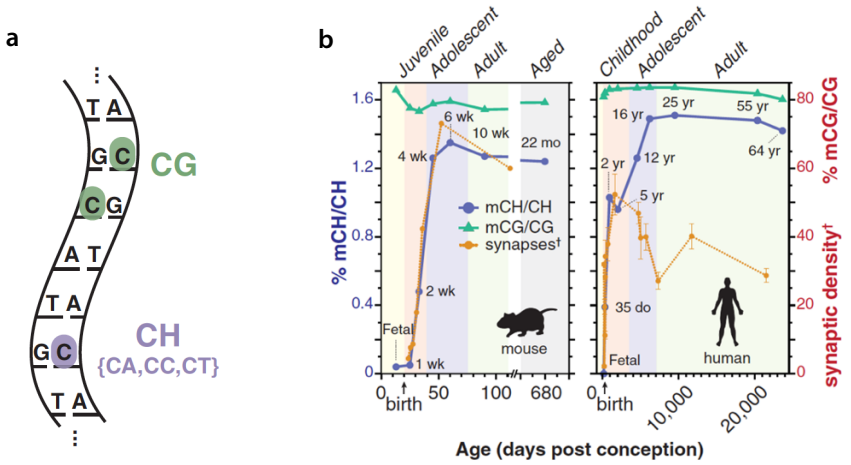


Figure 1.4. Development of mammalian neuronal CG and CH methylomes[32]

Figure 1.4 from Lister et al.[32] and shows in (a) the two different dinucleotide contexts of cytosine methylation, mCH (purple) and mCG (green), and in (b) the accumulation of mCH (blue line), mCG (green line), and synapses (orange line) over brain development in the mouse and human frontal cortex from birth to adulthood. Amount of methylation mC/C calculated as the fraction of methylated cytosines in all cytosines for each dinucleotide context. Synaptic density quantitated from De Felipe et al.[35] and Huttenlocher et al.[36].

While DNA methylation at CG dinucleotides is known to be an important epigenomic regulatory mechanism common to virtually all mammalian cell types, evidence from the last decade indicates that, during early postnatal development, neuronal genomes also contain uniquely high levels of the alternative form of methylation, mCH. In human adult neurons, approximately half of the cytosine methylation is in the CH context, while, in mouse adult neurons, ~34% of cytosine methylation is mCH[32]. In fact, mCH occurs in 5% of CH and 10% of CA nucleotide sites[32]. mCH moreover accumulates significantly in mammalian post-mitotic neurons through early childhood and adolescence: the maturation and development of neural circuitry (noted by a burst of synaptogenesis) dovetails with the rapid rise in mCH (Fig 1.4b)[32]. mCG accumulation, on

the other hand, is much more dynamic during mammalian embryogenesis[37] before plateauing at a high level during postnatal development[32] (Fig 1.4b).

1.3.3 Writers, erasers, and readers of DNA methylation

There are several proteins that establish the patterning of the dynamic DNA methylome, by depositing or removing methyl groups; these "writers" and "erasers" of DNA methylation are particularly important during the development of the brain. The DNA methyltransferases *DNMT1*, *DNMT3A*, and *DNMT3B* methylate the five-position carbon of cytosine (Fig 1.5). mCG is established by the *de novo* methyltransferases, *DNMT3A* and *DNMT3B*, and is then maintained by *DNMT1* which acts on hemi-methylated (i.e. one strand of the DNA has a methylated cytosine that provides a template for the other strand) sites. This maintenance methylation is limited to asymmetric (i.e. hemi-methylated) mCG[38], leaving no means for CH methylation to be maintained in replicating cells[20]. The rapid accrual of mCH in postmitotic neurons stems largely from the rise in *DNMT3A* in the juvenile brain[18,32,39,40]. Neuronal *DNMT3A* expression then decreases, but is maintained at measurable levels throughout adulthood[32,40].

Cytosine methylation is "erased" when methylcytosine is oxidized to hydroxymethylcytosine (hmC) via the Tet family of dioxygenases, *TET1*, *TET2*, and *TET3*[41,42] (Fig 1.5). The majority of hydroxymethylation exists in the hmCG context, as the Tet enzymes preferentially act on CG dinucleotides[42]. The Tet enzymes can further serially oxidize hmC to formylcytosine and carboxylcytosine which can then be excised to yield a nonmethylated cytosine[43,44].

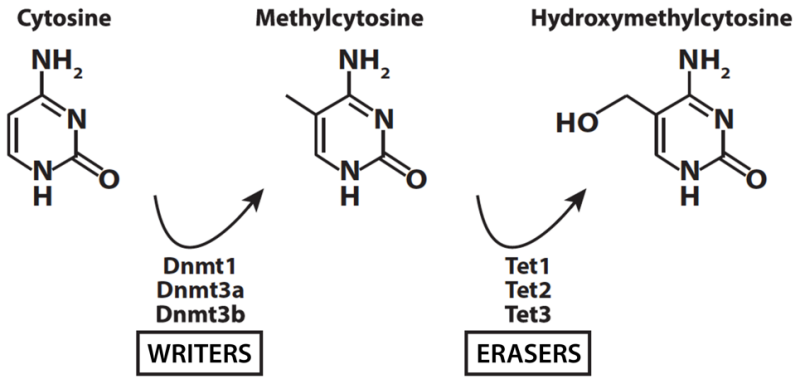


Figure 1.5. The writers and erasers of DNA methylation[45]

Figure adapted from Kinde et al[45] to show the chemical structures of the different forms of cytosine as cytosine undergoes "writing" to methylcytosine via *DNMT1*, *DNMT3A*, *DNMT3B* and the subsequent "erasing" from methylcytosine to hydroxymethylcytosine via *TET1*, *TET2*, and *TET3*.

Specific proteins, known as readers, bind directly to methylated DNA to confer changes in gene expression, often by modulating transcription itself. *MECP2*, one of the most well-studied readers of methylation and part of the family of methyl-binding-domain (MBD) proteins, was initially found to repress transcription via the recruitment of histone deacetylase complexes[46–48]; more recently, it has been shown to downregulate genes via recruitment of corepressor complexes[39,49–51]. In the mouse and human brain, *MECP2* binds strongly with both mCG and mCH, with mCAC being the most preferential binding context out of all mCH[39,50,52–54]. This binding is critically important for brain development, as loss-of-function mutations in *MECP2* are associated with severe neurological deficits, most notably seen in Rett Syndrome[26]. Other MBD proteins bind to different methylated cytosine contexts with varying affinities. In both *MBD1/4*, mCAC is preferentially bound across all mCH, with *MBD2/3* only binding to mCG[53]. Other

methyl-binding proteins include the zinc-finger protein Kaiso, which specifically binds mCG[55], and SRA-domain proteins, like *KYP*, which binds to both mCG and mCH[56].

1.3.4 Relationship between DNA methylation and transcription

In mammalian neurons, the presence of cytosine methylation has been largely correlated with repression of gene expression. Both mCG and mCH have been shown to be anticorrelated with gene expression in gene bodies, upstream, and downstream regions, with hypomethylation observed around the TSS[18,19]. However, recent work has demonstrated that gene body mCH is a clear epigenomic predictor for gene expression in that specific patterns of mCH define different neuronal cell types, e.g. excitatory versus inhibitory cell populations[19]. Neurological functions require the proper deposition and interpretation of mCH (specifically mCA)[39,40]. In the brain, *MECP2* acts to stabilize gene expression by binding to gene body mCA; in the absence of *MECP2*, longer genes tend to be significantly misregulated[39].

The specific deposition patterns of mCG and mCH contribute to their distinct functions, especially in how their respective abundances affects where readers can bind to mediate transcriptional changes. From early development, the relatively fewer CG dinucleotides in the genome are highly methylated, making the ratio of mCG/CG quite high. While the number of genomic CH dinucleotides is much higher, these sites are sparsely methylated over the course of brain development, and the relative mCH/CH levels are much lower in adult postmitotic neurons. Moreover, *DNMT3A* targets genes for mCA methylation that are more lowly expressed during the early postnatal period[40], creating a rheostat-like tuning system on neuronal gene expression.

It is not known how the presence or absence of DNA methylation itself directly impacts the rate of transcription. Recently, however, Adrian Bird's lab mathematically tested several

models of how a reader like *MECP2* might affect the progression of RNA polymerase Pol II along the gene body[51]. In their work, they find that DNA methylation marks act as "slow sites" causing "congestion" of Pol II along the transcript: *MECP2* dynamically binds to methylated sites, causing Pol II to proceed more slowly along the gene body, and ultimately generating fewer transcripts (ie. dampened gene expression) (Fig 1.6)[51]. This model suggests how gene body methylation might function to dampen gene expression while promoter methylation would prevent the initiation of transcription itself. Moreover, the larger proportion of mCH sites in the genome would allow for greater opportunities for a reader to bind to affect transcriptional changes.

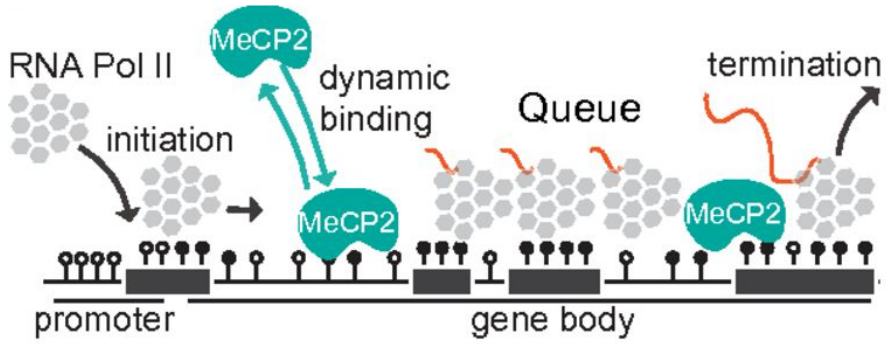


Figure 1.6. Congestion model of methylation and transcription[51]

Figure from Cholewa-Waclaw et al.[51] that shows the schematic for a model by which *MECP2* (teal protein) creates "dynamical obstacles" when it binds to methylated DNA (filled-in black circles) that cause RNA Pol II (grey dots) to pause along the gene body and thereby slow the rate of transcription.

In this thesis, I outline my work to characterize the relationship between the specialized methylome and transcriptome in vocal learning regions of the zebra finch brain. Chapter II presents the methods I used and developed for the work underlying this thesis. Chapter III focuses on the integrative analysis connecting the adult transcriptome with the adult methylome in the zebra finch

brain. Chapter IV explores the potential mechanisms required to build this specialized methylome as well as the coincident transcriptomic changes occurring over the development of the zebra finch brain. Together, the goal of this work is to characterize the unique zebra finch methylome and begin to explore how DNA methylation can define the specialized transcriptome underlying vocal learning in a songbird.

CHAPTER II: MATERIALS AND METHODS

2.1 Animals and Tissue Collection

2.1.1 Adult zebra finch samples

Tissue samples for whole-genome bisulfite sequencing (WGBS) were collected from adult (≥ 90 days old post-hatch, ranging from ~6 months to 3 years old) male zebra finches. Male finch brain regions were dissected as females are not vocal learners; while female finches have song nuclei at birth, these regions atrophy over development. All birds were individually housed overnight in sound isolation chambers and subsequently sacrificed in the dark before the lights came on in the morning. Rapid decapitation was used to avoid the use of isoflurane or other anesthesia known to have an effect on activity-dependent gene expression[15,57]. Brains were extracted and regions were microdissected within ~2 minutes of sacrifice. For the microdissections, 250-300µm coronal sections were collected using a Stoelting tissue slicer, after which sections were transferred to a petri dish on ice filled with pre-chilled 1x HBSS (Thermo Fisher 14175095) under a dissection microscope. The RA region is easily identifiable due to the higher myelination of the nucleus, which makes it appear as a dark spot under light microscopy; the medially adjacent lateral interior arcopallium (LAI) was collected as the non-vocal motor region (Fig. 2.1). The RA and LAI were collected from each hemisphere of the brain, such that each biological replicate contained each region pooled across both hemispheres of each animal. Replicate samples were immediately flash frozen at -80°C following collection.

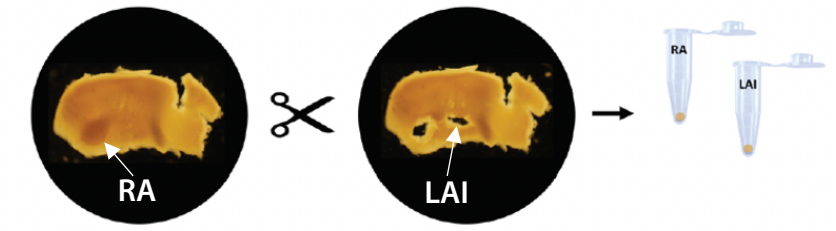


Figure 2.1. Coronal microdissection of RA and LAI in adult zebra finch brain

Figure shows the RA (dark spot toward left edge of section) and LAI regions collected from one hemisphere of a coronally-sectioned zebra finch brain. Pictures were taken under light microscope before and after microdissection of region.

Samples for RNA-sequencing were collected in another study in the lab[14] under the same behavioral conditions (ie. overnight, singly-housed males) but different mode of dissection: regions were collected via laser-capture microscopy (LCM) using serial sagittal sections of the brain. For both WGBS and RNA-seq, four biological replicates per region were collected. For brain sections for in situ hybridizations, samples were also collected from male birds that were singly-housed in the dark for at least 3 hours. All birds were from the Jarvis lab zebra finch colony at the Rockefeller University Comparative Biosciences Center, in which non-breeding finches are housed in single-sex cages with 5-8 birds per cage. Only male birds were used for experimental samples as vocal learning is limited to male zebra finches, and there was the potential for genomic confounds stemming from the W chromosome (the female is heterogametic with ZW, while the male is homogametic with ZZ chromosomes).

2.1.2 Juvenile zebra finch samples

For developmental in situ hybridization experiments, brains were collected from juvenile zebra finches at post-hatch day (PHD) 6, 35, 58, and 100. Sextyping is necessary for birds younger than ~40 PHD, so samples were collected regardless of sex for younger timepoints and subsequent genotyping determined whether or not samples were processed for experimentation. Only male samples were used for in situ hybridization. For juvenile RNA-Seq experiments, samples were collected from PHD30 male birds (n=3) using LCM-based techniques in another study in our lab[58]; three biological replicates were collected for every song nucleus and respective motor surround region.

2.1.3 Mouse samples

Pilot experiments of WGBS included tissue samples dissected from the frontal cortices and livers of 8-week-old male C57/B6 mice (n=2) at Duke University. Brains were extracted and bulk frontal cortex was dissected immediately followed sacrifice by cervical dislocation. ~0.5 cm sections were dissected from each liver. Samples were immediately flash frozen at -80°C post-collection.

2.2 Library Preparations and Sequencing Protocols

2.2.1 Whole-genome bisulfite sequencing

For all adult zebra finch RA and LAI and mouse brain and liver samples, genomic DNA was extracted using the QIAamp DNA Micro Kit (Qiagen 56304), with a minimum yield of 200 ng at a concentration of 4 ng/µL. For DNA fragmentation, the 200 ng of DNA in 50 µL of 1X TE Buffer (low EDTA, pH = 8.0) was then sonicated in a Covaris S220 Ultrasonicator (10% duty

factor, 200 cycles/burst, 175 peak incident power, 7°C water temperature, 3 minutes). Following sonication, tubes were transferred to fresh tubes and briefly spun down, and 3 µL of 1 ng/µL lambda DNA (Fisher/Promega D1501) was added to each sample to serve as an in-sample unmethylated control.

For the remainder of the library preparation, NuGen Ovation Ultralow Methyl-Seq DR Multiplex System 1-8 (original: NuGen 0335-32, now: Tecan 0541-32) was used with the EpiTect Fast DNA Bisulfite Kit (Qiagen 59824) for the bisulfite conversion portion of the protocol. For bead purification steps, a DynaMag-96 Side (Thermo Fisher 12331D) was used. Protocol break point occurred between post-ligation purification (at which point samples were eluted in low EDTA 1x TE Buffer and stored at 4°C overnight) and final end repair. For bisulfite conversion, PCR cycling conditions were modified as follows to facilitate detection of mCH: (95°C – 5 min, 60°C – 20 min) x 4 cycles, hold at 20°C. For library amplification, all reagents were pre-chilled and/or kept on ice and 11 cycles were used for the thermal cycling program. A Qubit 2.0 Fluorometer and dsDNA HS Assay Kit (Qubit Q32851) were used to quantify library DNA concentrations and an Agilent 2100 Bioanalyzer was used to analyze library quality prior to sequencing. Protocol was adapted from the Gabel lab (Washington University in St. Louis).

Sequencing of 100 bp, paired-end reads was done on the Hiseq 2500 platform from Illumina at the Weill Cornell Epigenomics Core Facility.

2.2.2 Bulk RNA-seq

RNA sample and library preparations for adult zebra finch samples followed methods described in Gedman et al[14]. Briefly, RNA was isolated from each sample using the Picopure RNA Isolation kit (Applied Biosystems KIT0204), and cDNA was generated using the SMART-Seq

Ultra Low-Input RNA kit (Takara 634892). Sequencing libraries were prepped using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs E7645L) and RNA sequencing of 150bp, paired-end reads was conducted on the NextSeq 500 platform from Illumina.

RNA sample and library preparations for PHD30 zebra finch samples followed methods described in Choe et al[58]. Briefly, RNA was isolated from each sample using the Picopure RNA Isolation kit (Applied Biosystems KIT0204), and cDNA was generated using the SMART-Seq Ultra Low-Input RNA kit (Takara 634892). Sequencing libraries were prepped using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs E7645L) and RNA sequencing of 150 bp, paired-end reads was conducted on the Novaseq 6000 platform from Illumina.

2.3 Bioinformatic Analyses

2.3.1 Transcriptomic analysis and differential gene expression

For all transcriptomic analysis, all adult and juvenile raw RNA-Seq reads were processed through a custom NGS pipeline, CountMatrix (https://github.com/mattisabrat/CountMatrix) that performs quality control, trimming, alignment, and mapping to genes from raw read files—it acts as an automated wrapper for several well-known packages including FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc), Trimmomatic[59], STAR[60], and FeatureCounts[61]. For this work, the CountMatrix pipeline included trimming with Trimmomatic[59] to remove adapter sequences and low-quality reads, and alignment with STAR[60]. All reads were mapped to the newly assembled and annotated high-quality, long-read based, Vertebrate Project Genome (VGP) zebra finch genome (bTaeGut1_v1, RefSeq Accession: GCF_003957565.1).

The count matrix generated from this pipeline was used to run DESeq2[62] (v1.28.1) for all downstream differential expression testing. Within DESeq2, a linear model was constructed to

model brain region (e.g. RA or LAI) and whether the region was a song region or surrounding control region. As juvenile and adult samples were prepared and sequenced separately, the only differential expression established for integrated analysis with methylome data was between RA and LAI within each age (i.e. adult RA v adult LAI). A variance stabilizing transformation was used on the DESeq object to run the differential expression pipeline. Following multiple test correction, genes with $q < 0.05$ were called as differentially expressed. 20,905 genes were found to be expressed in the zebra finch arcopallium using the bTaeGut1_v1 genome.

2.3.2 Methyl-seq processing and methylation calling

All Methyl-seq raw reads were first trimmed using TrimGalore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) a wrapper for FastQC and cutadapt[63] using the following specifications: --stringency 3 --clip_R1 3 --clip_R2 3. With multiple aligners available for methylation sequencing data, alignment via BS-Seeker2[64] (4 mismatches allowed, XS=0.3,2) using bowtie2[65] yielded the highest percent of genome mapped (82%)—paired-end reads from zebra finch samples were mapped to the VGP zebra finch genome (bTaeGut1_v1, RefSeq Accession: GCF_003957565.1), while paired-end reads from mouse samples were mapped to the mm10 genome (RefSeq Accession: GCF_000001635.20). Duplicate reads were removed and only uniquely mapping reads were kept.

For all-cytosine methylation (mC) calling, samtools[66] was first used to sort alignment files and CGmaptools[67] was used to produce CGmap and wig files. CGmap files were further filtered for at least 10x coverage. For base-specific-cystosine methylation (mCG v mCH), methylpy[68] (https://github.com/yupenghe/methylpy) was used to generate allc ("all-cytosine") files (Fig. 2.2). Allc files were further filtered for at least 10x coverage. Methylation levels (ie. mC/C, mCG/CG,

or mCH/CH) were estimated as the fraction of cytosines in the sequenced population which are methylated—to estimate mC/C, the fraction of all MethylC-Seq base-calls at cytosine reference positions that were cytosine (protected from bisulfite conversion) was calculated. This estimate was then corrected to account for the low error rate stemming from the failure of the chemical conversion of unmethylated cytosine to uracil. To calculate this bisulfite non-conversion frequency, the frequency of all cytosine base-calls at reference cytosine positions in the lambda genome (unmethylated spike-in control) was normalized by the total number of base-calls at reference positions in the lambda genome.

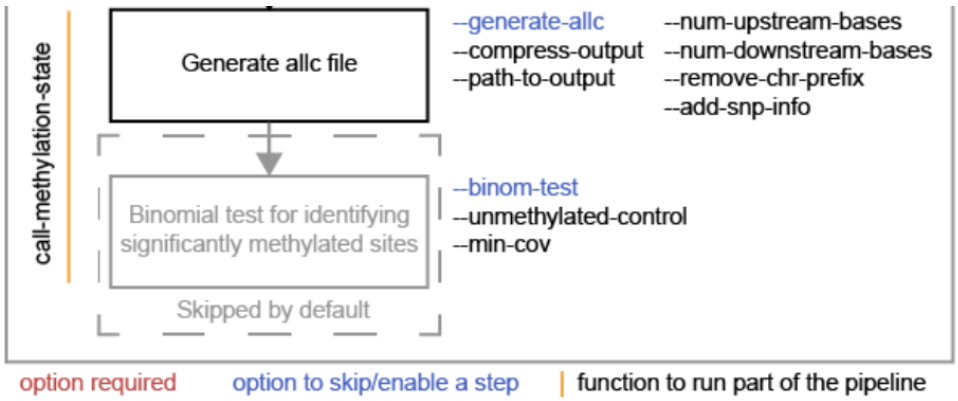


Figure 2.2. MethylPy pipeline for calling methylation state[68]

Figure from github.com/yupenghe/methylpy shows the bioinformatic parameters for calling methylation using an unmethylated control within the MethylPy package.

2.3.3 DMR Identification

Differentially methylated regions (DMRs) between RA and LAI were defined by first identifying differentially methylated sites (DMSs) between the two tissues using a modified version of methylpy's DMRfind (Fig. 2.3). A beta-binomial distribution was used to model the methylation level of every single cytosine site in each of the tissues. Then, differentially

methylated sites were identified if the methylation levels of certain sites were significantly different between tissues (p-value ≤ 0.01) and the minimum methylation difference was greater than or equal to 0.3. DMSs within 500 bp of each other were merged into DMRs. For each DMR, the methylation difference between each of the tissue pairs (i.e. pairwise comparisons) was computed and only DMRs that had a significant methylation difference (p-value ≤ 0.01) and methylation difference greater than or equal to 0.3 in at least one of the pairwise comparisons were retained.

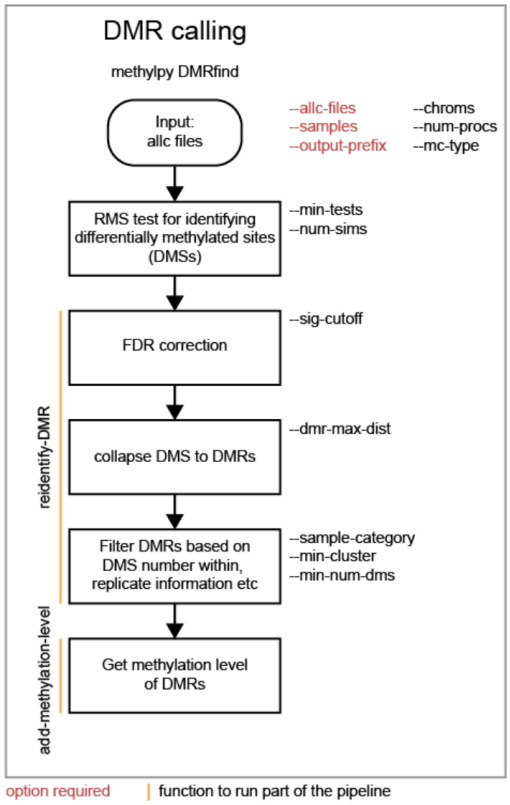


Figure 2.3. MethylPy pipeline for finding DMRs[68]

Figure from github.com/yupenghe/methylpy shows the bioinformatic pipeline and parameters to identify statistically significant DMRs between methylation files within the Methylpy package.

2.3.4 Metagene analysis

Metagene plots for mCG and mCH along gene bodies were generated using deeptools[69]. First, computeMatrix's "scale-regions" function was used to normalize the gene body region (from TSS to TES) to 1kb for all genes and average the specific methylation signal (parameters: --m 1000 --sortUsing mean --binSize 50) across this window. Then, plotHeatmap was used to run k-means clustering on mCG and mCH matrices independently. The number of clusters for each plot was selected to achieve a maximal silhouette score (0 = clusters poorly formed, 1 = clusters perfectly formed), resulting in k=5 for mCG and k=14 for mCH. Genes in each cluster were then integrated with transcriptomic data to establish proportions of differential expression in each cluster. Pattern differences between RA and LAI clusters were established by comparing the distribution of methylation scores across the gene body for RA versus LAI, using a two-sample Kolmogorov-Smirnov test for significance.

2.3.5 Integration of methylation and gene expression data

Methylome and transcriptome data were overlayed using the R packages GenomicRanges[70] (v1.40.0), GenomicFeatures[70] (v1.40.1), rtracklayer[71] (v1.48.0), and bumphunter[72] (v1.30.0). For matching specific DMRs or other methylation Granges objects to genes, the bumphunter function matchGenes() was used with promoters defined as being within 2kb of the TSS. For all correlations calculated between gene expression levels and methylation levels, Spearman correlations were used. For permutation testing of significant negative Spearman correlations, the perm.relation() function in the wPerm (v1.0.1) package was used; in each iteration of this function, the alternative hypothesis was tested 9999 times. All integrative analyses and visualization were done using R (v4.0.2) with tidyverse (v1.3.0).

2.4 Examining methylation readers and writers

2.4.1 *DNMT3A* in situ hybridization

For all in situ hybridization (ISH) experiments, brains were embedded in TissueTek OCT (Sakura 4583) within 5 minutes of sacrifice and immediately frozen at -80°C. Brains were subsequently cryosectioned (either sagitally or coronally) at a thickness of 12 μm. Two sets of anti-sense FITC labeled RNA probes were generated for *DNMT3A* for testing and sense RNA probes were used to validate ISH results. All probes were generated from zebra finch brain cDNA. In brief, sections were first fixed in 4% PFA/PBS, permeabilized, dehydrated, and then hybridized with antisense *DNMT3A* probes (1:100) at 65°C overnight in a HybEZ II Oven (ACD PN 321710/321720). Next, sections were washed in room-temperature chloroform and processed through a series of stringency washes in SSC (0.2x SSC 4 times for 30 minutes each, followed by cooling to room-temperature and one 5-minute 1x SSC wash) and a series of buffer washes (3x 5-minute wash in a 0.1M Tris/0.15M NaCl/0.05% Tween-20/water buffer at room-temperature), incubated with 0.5% Blocking Solution (Roche 11 096 176 001) for one hour, and then incubated with POD-anti-FITC antibody (1:1000) (Roche 11 426 346 910) for at least one hour. Following a series of wash steps (2x 10-minute washes in 1x PBS at room-temperature, one 10-minute wash in 0.1% BSA-PBS), sections were incubated in FITC-TSA in Amplification Plus Diluent (1:100) (Perkin Elmer NEL753001KT) before a final 5-minute 1x PBS wash and slide mounting with Vectashield fluorescent mounting media with DAPI (Vector Labs H-1500).

2.4.2 *MECP2* and MBDs transcriptomic analysis

To determine whether *MECP2* was present in the zebra finch brain, raw reads from sequenced finch somatic tissues (ovaries, muscle, testes) and from sequenced finch brain tissues

(juvenile and adult song nuclei and surrounding regions) were first mapped to the VGP bTaeGut1_v1 genome and sample coverages were visualized using the IGV genomic browser[73]. For assessing the brain expression of *MBD1-5*, *TET1-3*, *DNMT1*, *DNMT3A/B*, and *MECP2*, a full matrix of raw counts for all song nuclei (RA, HVC, Area X, LMAN) and their respective surrounds (LAI, PLN, VSt, and AN) was generated for both adult and juvenile samples. Variance stabilizing transformation within the DESeq2 pipeline was used to generate this DESeq object. Genes with very low counts were filtered out via minimal pre-filtering to keep only rows that have at least 30 reads total for each sample; genes remaining in the raw counts matrix were deemed to be "expressed" in the given region, and subsequently categorized into "low" or "high" expression if the normalized read count was on average < 100 or > 2000, respectively.

CHAPTER III: INTEGRATING THE SPECIALIZED METHYLOME AND TRANSCRIPTOME OF THE ADULT ZEBRA FINCH ARCOPALLIUM

3.1 Characterization of the zebra finch methylome

3.1.1 Similarities and differences between the zebra finch and mammalian brain methylomes

Before examining the connection between the methylome and transcriptome in the zebra finch brain, it was important to establish how the zebra finch DNA methylome in the brain compares to those of well-studied mammalian methylomes like in mice and humans, as this study was one of the first to examine non-mammalian vertebrate brain methylomes at basepair resolution. Other groups have explored aspects of DNA methylation in the finch brain, including using reduced-representation bisulfite sequencing (RRBS) and profiling the methylome of a finch cell line[74,75]. However, the use of RRBS is biased towards CpG sites[76], preventing a complete profiling of the critically important mCH mark in the genome. In this work, WGBS, an unbiased sequencing method, was performed on RA and its adjacent motor surround, LAI, from four adult male zebra finches. WGBS profiles the entire methylome at basepair resolution via differentiation of methylated from unmethylated cytosines following sodium bisulfite treatment, which converts unmethylated cytosines to uracil. The zebra finches were behaviorally controlled to limit light and activity-induced gene expression and potential methylome changes in these regions. To ensure technical accuracy of the WGBS experiment, mouse frontal cortex samples were sequenced along with samples from the finch arcopallium (RA and LAI) as a positive control. After sequencing, the epigenome-wide proportions and distributions of mCG and mCH were calculated across zebra finch and mouse tissues.

From published work on mammalian DNA methylomes, the breakdown of methylated cytosines in human frontal cortex NeuN+ tissue is 48% mCH and 52% mCG[32] (Fig. 3.1a). In bulk mouse frontal cortex, the breakdown of methylated cytosines is 38% mCH and 62% mCG (Fig. 3.1a), comparable to what has been shown in NeuN+ mouse frontal cortex[32]. Interestingly, in the zebra finch arcopallium, the proportion of mCH (60%) is much higher than mCG (40%), mirroring an opposite distribution to that seen in the mouse brain, with about half of the arcopallium mCH occurring in the mCA context (Fig. 3.1b).

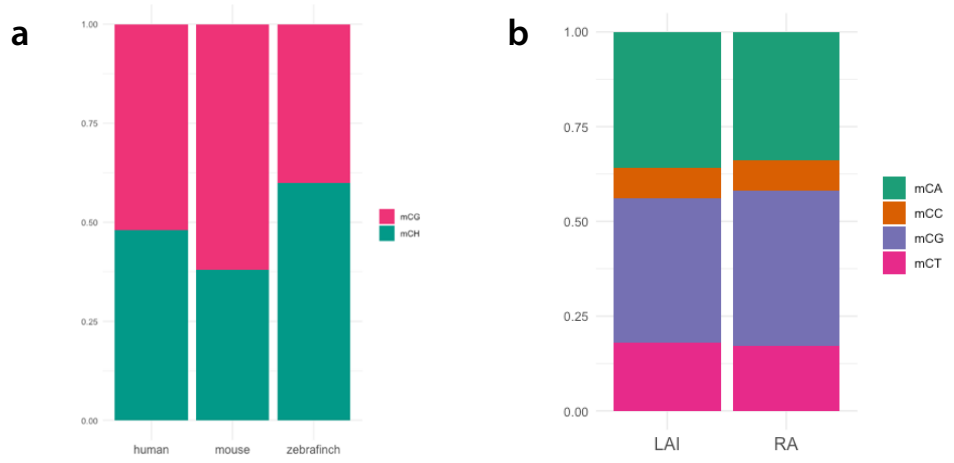


Figure 3.1. DNA methylation context proportions in human, mouse, and zebra finch brains

Barplots of **(a)** cross-species comparisons of methylated cytosines in CH and CG dinucleotide contexts and **(b)** dinucleotide proportions of methylated cytosines within zebra finch arcopallium. Human data adapted from Lister et al, 2013[32]. Human samples were collected from NeuN+ sorted frontal cortex tissue[32], mouse samples were collected from bulk frontal cortex tissue, and zebra finch samples are from bulk arcopallial tissue (RA and LAI).

3.1.2 Characterizing the adult arcopallial methylome

This unique breakdown is partially due to the fact that the overall methylation level of zebra finch mCG (i.e. how many CG sites are methylated) is ~60%, which is much lower than the ~80% mCG/CG in mammalian brains—in general, mCG sites tend to show greater intermediate methylation (between 25-80%) in the finch (Fig. 3.2) than in mammals, in which mCG sites tend to be highly methylated (>80%)[32]. In Fig. 3.2.a, the distribution of mC/C at every base (≥10x coverage) is shown for mCA and mCG, while in Fig. 3.2.b, this distribution is based on mC/C over 500bp non-overlapping windows across the genome. The distributions of mCA, mCT, and mCC are nearly identical, so only mCA is shown here for readability.

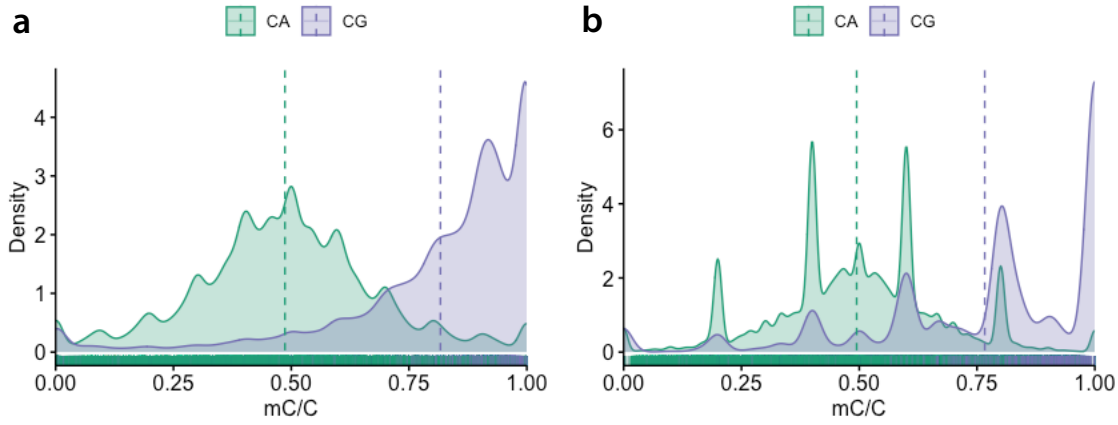


Figure 3.2. Density plot of RA mCA v mCG methylation across the adult finch genome

In **(a)** the distributions of mCA/CA (green) and mCG/CG (purple) are shown for all bases (>10x coverage), while in **(b)** the distributions of methylation levels are shown for 500bp non-overlapping windows across the genome. Dotted lines indicate means for each mC/C distribution.

3.1.3 Discussion

We show that the DNA methylome of the zebra finch brain (arcopallium) is both similar to and unique from those of mammalian models. Recent work profiling gross methylome patterns

across vertebrate and non-vertebrate species indicates that great tit (in the same order, Passeriformes, as the zebra finch) and chicken genomes display significantly reduced hypermethylation (mCG/CG > 70%) compared to mammalian genomes[77], supporting our findings on the distribution of mC/C for each dinucleotide. Neither genome size nor basepair composition appear to play a role in defining this unique feature of bird genomes[77]. The unique levels of mCG and mCH potentially stem from the altered levels of canonical methylation writers, like *Dnmt3a*, and erasers, like *Tet1*. As only arcopallial methylomes were sequenced for the zebra finch here, further brain-region-specific sequencing would elucidate potential subdivision or zebra finch-specific (i.e. distinct from other birds) methylome features. With respect to similarities across vertebrate methylomes, the overall mCH level in the finch and mammals is around ~2-4%[32,77], with the highest proportion of mCH being mCA (Fig 3.1.b). Moreover, the most common trinucleotide context in the finch brain for mCA is mCAC[77], as it is in the mammalian brain[39,40,50,54].

3.2 Differential methylation and gene expression between vocal learning and surrounding brain regions

3.2.1 Integrating DMRs and differential gene expression in RA versus LAI

To begin to explore how methylation might underlie differential expression in the finch arcopallium, regions of significant differential methylation between the RA and LAI methylomes were identified. These DMRs between RA and LAI were defined by first identifying significantly differentially methylated sites (DMSs) between the two tissues (p-value ≤ 0.01), and then DMSs within 500 bp were merged—only DMRs with ≥ 0.3 difference in methylation in at least one pairwise comparison were retained. Between RA and LAI, 4649 DMRs were found. Overall,

DMRs between RA and LAI were typically found in gene bodies as opposed to intergenic regions (i.e. promoter, upstream, and downstream) (Fig 3.3). Promoters (within 2kb of the TSS) had the lowest number of DMRs (Fig 3.3) DMSs were also typically in the mCH context, with only ~6% of DMSs in the mCG context (Fig 3.4).

To connect transcriptome data to this differential methylation data, we utilized RNA-Seq data[14] from RA and LAI regions under the same behavioral conditions (n=4, silent, single-housed adult males). A total of 2134 genes expressed in the arcopallium (across RA and LAI) had DMRs, yielding a ~10% rate of arcopallial genes with DMRs (i.e. 2134/20905 genes). Genes and DMRs were matched using the R packages, bumphunter and GenomicRanges. Often, more than one DMR was found on any given gene, for an average of ~2 DMRs per gene.

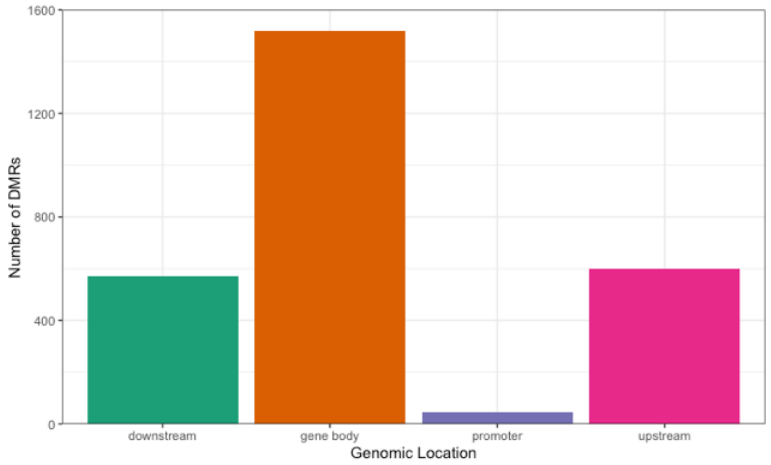


Figure 3.3. Genomic locations of DMRs between RA and LAI

Figure shows the breakdown of genomic locations (downstream = downstream of the gene, upstream = upstream of the gene, promoters = ≤ 2kb upstream of the TSS, gene body = between the TSS and TES of the gene) for all 4649 DMRs between RA and LAI. Regions were defined using the "matchGenes" function of the bumphunter package with annotated VGP bTaeGut1_v1 genome (note: the annotation for this genome is constantly being improved, so future work would elucidate more named genes—the number of genes should remain the same for the same genome).

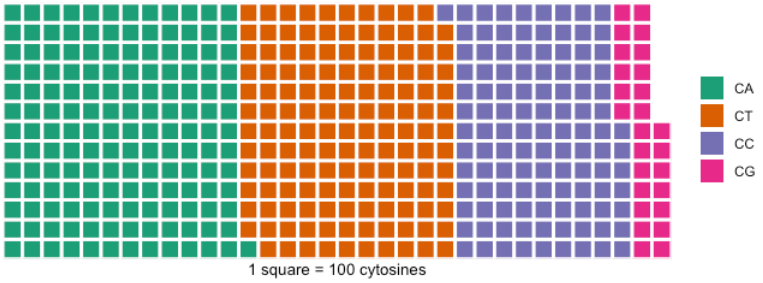


Figure 3.4. Dinucleotide methylation contexts of DMSs between RA and LAI

Figure is a waffle plot visualization representing the breakdown of all DMSs between RA and LAI, which are used to determine DMRs. Each square represents 100 DMSs and the colors represent the four dinucleotide contexts of cytosine methylation—green, orange, and purple colors together represent mCH DMSs between RA and LAI.

After establishing this baseline intersection of DMRs with the arcopallial transcriptome, it was critical to identify and examine the subset of differentially expressed genes (DEG) between RA and LAI as this geneset defines the specialized transcriptome of the RA song nucleus. 781 genes out of 20,905 expressed genes were found to be significantly differentially expressed between RA and LAI (Benjamini-Hochberg, adjusted $p \leq 0.05$). DMRs are found on a significant fraction of DEGs, with 20% of DEGs containing DMRs (151/781 genes), approximately twice the rate of arcopallial genes with DMRs. Moreover, 80% of these DMRs were found in downregulated genes in RA vs LAI (221/276 DMRs). In examining the direction of the differential methylation for each DMR, i.e. increased or decreased methylation in RA relative to LAI, there was no clear correlation between the extent of methylation difference or DMR location and the fold change of expression of the gene (Fig 3.5a). Yet, several critical genes involved in brain development and signaling like *NEUROD6*, *GRIK1*, and *CNTNAP5* had DMRs with anticorrelated direction of expression change and methylation difference, e.g. hypermethylated DMRs on downregulated genes (Fig 3.5a). One example of this type of DEG with DMRs was *MEF2C*, a critical downregulated gene in RA with two hypermethylated DMRs on its gene body (Fig 3.5b).

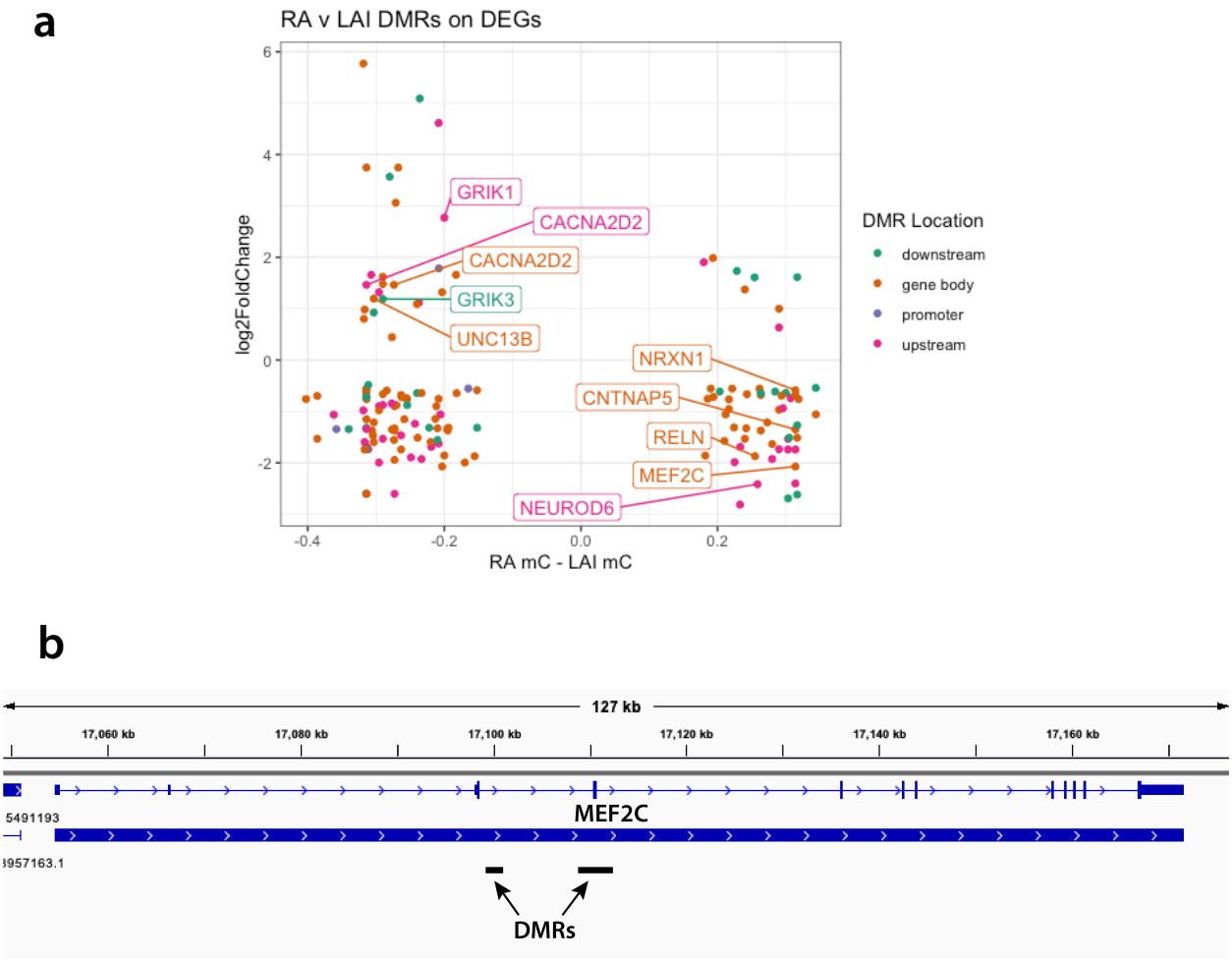


Figure 3.5. DMRs on upregulated and downregulated differentially expressed genes

Figure in (a) shows all DMRs on DEGs plotted such that the difference in methylation (positive = hypermethylation in RA, negative = hypomethylation in RA) is on the x-axis and the normalized fold change in expression is on the y-axis (positive = upregulated DEG, negative = downregulated DEG). DMR points are colored according to their genomic location. A few DMRs which are anticorrelated to direction of gene expression (i.e. upper left and lower right quadrants of the graph) are labeled for the gene that they occur on, as indicated. Sometimes, multiple DMRs exist on one gene, e.g. upregulated *CACNA2D2* has two DMRs that are hypomethylated in RA. In (b), a gene browser representation of one of the differentially downregulated genes in RA, *MEF2C*, is shown with two hypermethylated DMRs (higher RA methylation) located on the gene body.

3.2.2 Differential mCH gene body methylation defines a subset of differentially expressed genes in the arcopallium

With the uniquely high proportion of mCH in the finch brain and the established functional role of mCH in defining cell types in the brain, it was important to next separate mCH and mCG while integrating methylome and transcriptome data. As the bulk of DMRs were found in gene bodies, we next focused on defining the patterns of gene body methylation in RA versus LAI for both mCH and mCG. All gene bodies were first scaled to 1kb in length and metagene plots were generated from mean mCH or mCG across non-overlapping 50 bp windows across the genome. Gene bodies in RA and LAI for mCH or mCG were subsequently sorted by decreasing methylation level and clustered using the k-means algorithm, with the number of clusters selected to achieve a maximal silhouette score (Fig 3.6). As mCG and mCH are distinct signals, RA versus LAI clustering was done independently for each context: 14 clusters were calculated for mCH RA and LAI, while 5 clusters were calculated for mCG RA and LAI (Fig 3.6). Across all clusters, only two clusters in mCH had significantly different patterns between RA and LAI (two-sample Kolmogorov-Smirnov test performed on comparison of distribution of methylation scores between RA and LAI for each cluster, k7: $p = 4.1e^{-12}$, k9: $p = 1.7e^{-13}$): cluster k7, where the RA gene body was hypermethylated relative to LAI, and cluster k9, where the RA gene body was hypomethylated relative to LAI (Fig 3.6).

Figure 3.6. Gene body mCG and mCH pattern differences between RA and LAI

Figure on following pages shows metagene plots for gene body methylation in RA versus LAI in (a) the mCG context and in (b) the mCH context. At the top of both (a) and (b) are line graphs showing the average methylation signal in each of the clusters of the heatmaps below; in (b) cluster 7 (teal line) and cluster 9 (lime line) are labeled next to the line graphs to highlight their opposing signals between RA and LAI. Heatmaps are ordered by mean methylation level across each pair of clusters in descending order, with dark red corresponding to high methylation and dark blue corresponding to low methylation. For all annotated genes in the genome, every gene body, from TSS to TES, is scaled to 1kb across the heatmaps. On the heatmaps, clusters k7, k8, and k9 are labeled with k7 boxed in golden yellow and k9 boxed in purple. (c) is a zoomed-in view of clusters k7, k8, and k9 from the mCH metagene plots.

Figure 3.6. Gene body mCG and mCH pattern differences between RA and LAI

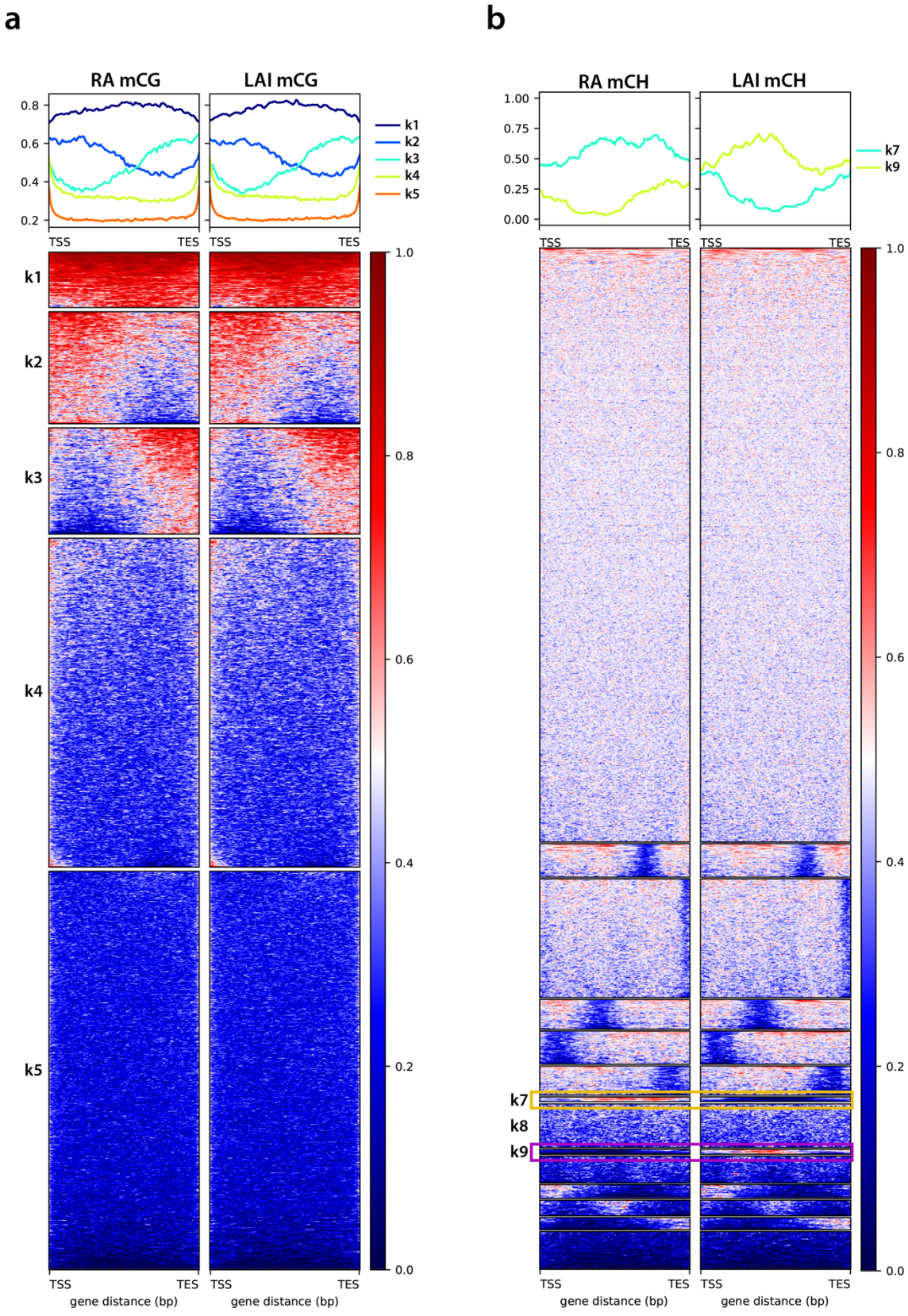
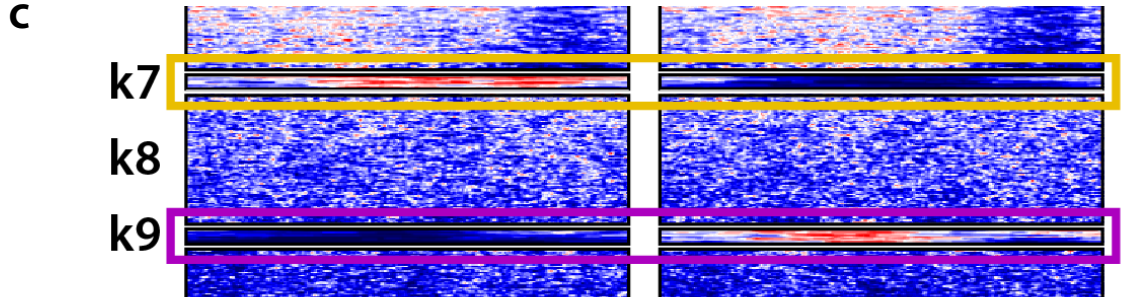


Figure 3.6. Gene body mCG and mCH pattern differences between RA and LAI



In the cluster with RA gene body hypermethylation, k7, ~40% of genes were DEGs (44/118 genes in the cluster), whereas in the cluster with RA gene body hypomethylation, k9, only a handful of genes were DEGs (7/124) (Fig 3.7). Moreover, in examining other clusters where there was no difference between gene body RA and LAI methylation, for either mCH or mCG, we found comparably low proportions of DEG, ranging from 2-5% (Fig 3.7). Examining the genes in k7 further revealed that 40 out of the 44 DEG were differentially downregulated genes, including several of the genes that had been identified as having hypermethylated DMRs like *MEF2C*, *CNTNAP5*, and *ADCY1* (Fig 3.8). The genes in the k7 cluster were also enriched for several neurobiological functions including synaptic signaling, neurodevelopment, and synapse structure, among several other neural functions (Table 3.1).

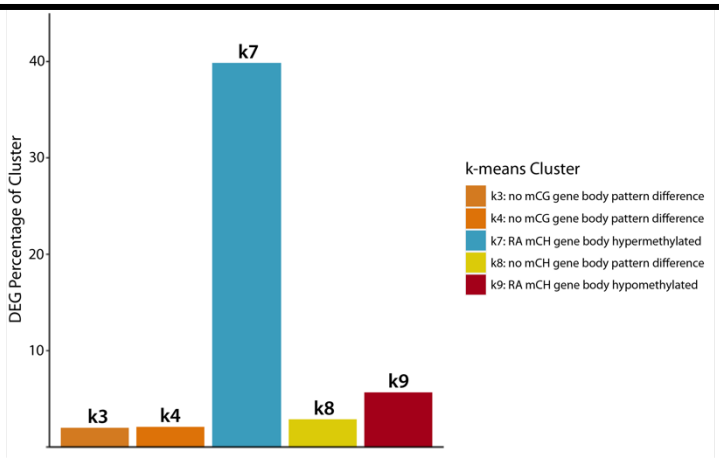


Figure 3.7. Proportion of DEGs in differentially methylated clusters between RA and LAI

Bar plot shows the proportion of DEGs in each of five clusters from the heatmaps in Fig 3.6. k3 and k4 were two clusters in which there was no mCG gene body pattern difference ($p = 0.22, 0.39$), while k8 showed no mCH gene body pattern difference ($p = 0.13$). k9 showed RA gene body hypomethylation and k7 showed RA gene body hypermethylation relative to LAI.

Figure 3.8. Proportion of differentially expressed genes in k7 cluster of RA gene body hypermethylation

Figure on following page shows all 118 genes in the k7 cluster from the mCH heatmap in Fig 3.6.b sorted by normalized change in expression (log2FoldChange) in descending order, with dark red representing highly upregulated genes and dark blue representing highly downregulated genes. These genes are further annotated to indicate whether or not they are a DEG (yes = purple, no = golden yellow).

Figure 3.8. Proportion of differential genes in k7 cluster of RA gene body hypermethylation

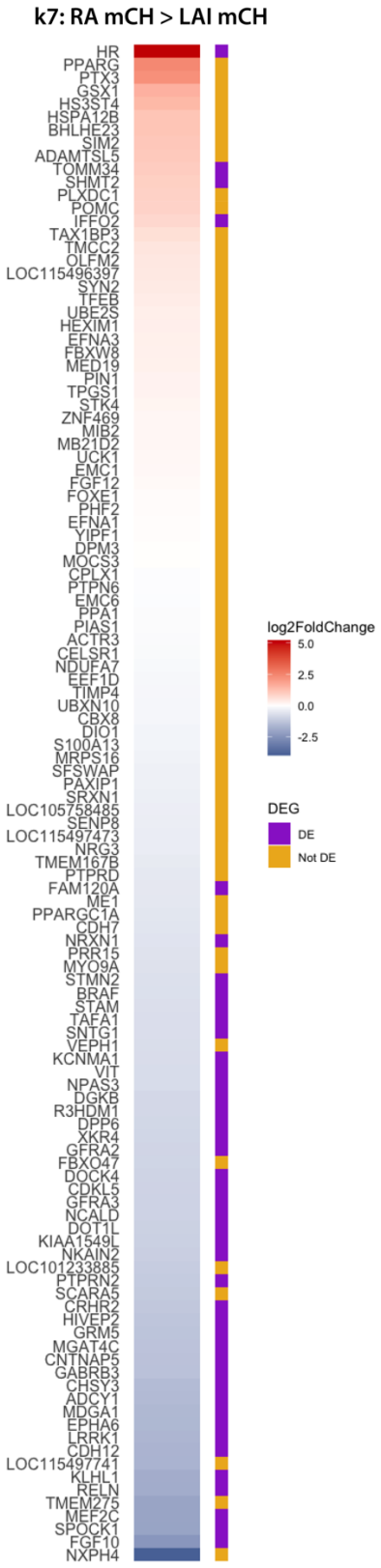


Table 3.1. Gene ontology of k7 cluster of genes with hypermethylated gene body mCH

Enrichment FDR	Genes in list	Total genes	Functional Category
1.3E-08	20	434	Synapse organization
7.3E-08	45	2474	Nervous system development
2.5E-07	23	746	Anterograde trans-synaptic signaling
2.5E-07	23	746	Chemical synaptic transmission
2.5E-07	23	762	Synaptic signaling
2.5E-07	23	756	Trans-synaptic signaling
3.2E-07	10	103	Synaptic transmission, glutamatergic
1.2E-06	13	239	Regulation of synapse organization
1.6E-06	17	468	Regulation of trans-synaptic signaling
1.6E-06	13	248	Regulation of synapse structure or activity
1.6E-06	17	467	Modulation of chemical synaptic transmission
1.6E-06	58	4319	Regulation of biological quality
3.7E-06	33	1774	Cell-cell signaling
3.7E-06	11	182	Synapse assembly
4.6E-06	9	111	Regulation of synapse assembly
4.6E-06	8	79	Regulation of synaptic transmission, glutamatergic
8.3E-06	24	1054	Central nervous system development

Table shows the top functional categories, sorted by significance (FDR) that were enriched in the k7 cluster genes against a background of all genes expressed in the zebra finch arcopallium. Functional category tags were defined via the PANTHER classification system[78].

3.2.3 Discussion

We find that mCH gene body hypermethylation, not promoter or mCH gene body methylation, underlies a proportion of differential downregulation that defines the RA song production nucleus. Moreover, these downregulated genes are involved in synaptic signaling and neurodevelopment. One potential way in which this candidate geneset could be expanded is with improved genome annotation; many of the genes in the clusters in Fig 3.6, for example, are named with the “LOC” prefix and could eventually be named as this annotation is updated. It will also be interesting to establish how other epigenomic mechanisms like chromatin conformational changes

could define these differential categories of genes, e.g. hypermethylated gene bodies in RA relative to LAI. The metagene analysis here proved essential in understanding the complexity in the DMR findings, which could be a result of the bulk-sequencing of the cell type heterogeneity of RA.

Additionally, due to the constraints of RA and LAI tissue size, samples could not be split for both RNA-Seq and WGBS, resulting in the collection of transcriptome and methylome data from separate animals and distinct dissection protocols. RNA-Seq samples were collected using laser-capture microscopy, while WGBS samples were collected via rapid microdissection; the former were collected in sagittal sections, while the latter were collected in coronal sections. The results of these differences is that the LAI region was collected differently, but RA, due to its well-defined morphology was collected comparably in both. We do not expect this distinction in dissection to yield meaningful differences in the interpretation of the sequencing of the bulk tissue, especially over four individual biological replicates for each sequencing pipeline. However, to reduce any potential confounds, future experiments should use the same dissection methods and, ideally, use the same tissue to be split across multiple types of library preparations. The latter experiment could now be possible with recently-developed single-nuclei protocols[21,79].

3.3 Correlation of methylation and gene expression in the arcopallium

3.3.1 Gene body, not promoter, DNA methylation is negatively correlated to gene expression

Thus far, we have shown that gene body methylation, specifically mCH, is tied to downregulation of gene expression in the RA relative to LAI. However, it was important to establish a better proxy for gene expression than normalized fold change in expression, which, while informative for direction of regulation and extent of difference between samples, does not capture a unit of expression for individual genes in each sample. Therefore, we next examined the

correlation between methylation level and the transcript count per million (TPM). An added benefit of using TPM to describe gene expression is that it normalizes for the length of the gene; indeed, many neuronal genes are long genes (>100kb)[39] that can be differentially repressed via mCA sites[39]. Average methylation levels (%mC) were then calculated for all gene bodies and promoter regions and mapped against TPM (Fig 3.9) for RA and LAI separately. Higher methylation levels in gene bodies were found to be negatively correlated (Spearman R = -0.35, p < $2.2e^{-16}$) with gene expression, with no significant correlation (Spearman R = 0.025, p = 0.78) found between promoter methylation and TPM (Fig 3.9). Despite the correlation with the level of methylation, there was significant variation, indicating the potential presence of other factors influencing gene expression levels.

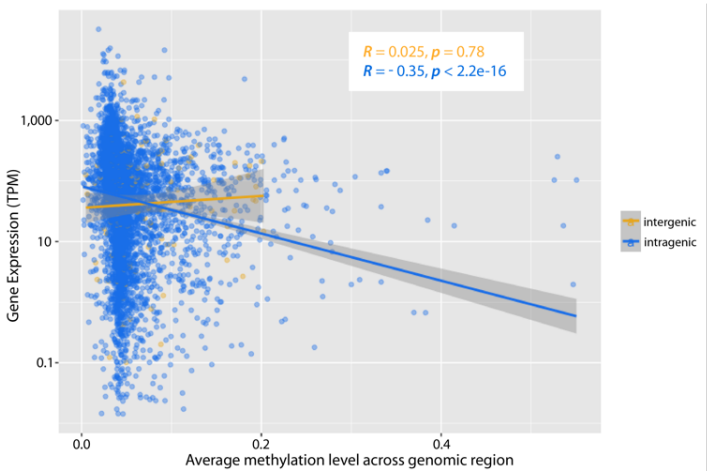


Figure 3.9. Correlation between gene body methylation and gene expression

Scatterplot for all gene bodies (blue, intragenic) and promoters (yellow, intergenic) in the finch arcopallium, with average level (calculated across the genomic region) of cytosine methylation in RA on the x-axis and TPM on the y-axis. Methylation values were calculated for all genomic regions by averaging across each gene body region (from TSS to TES) and promoter region (within 2kb of TSS). Spearman correlations and p-values are shown on the plot.

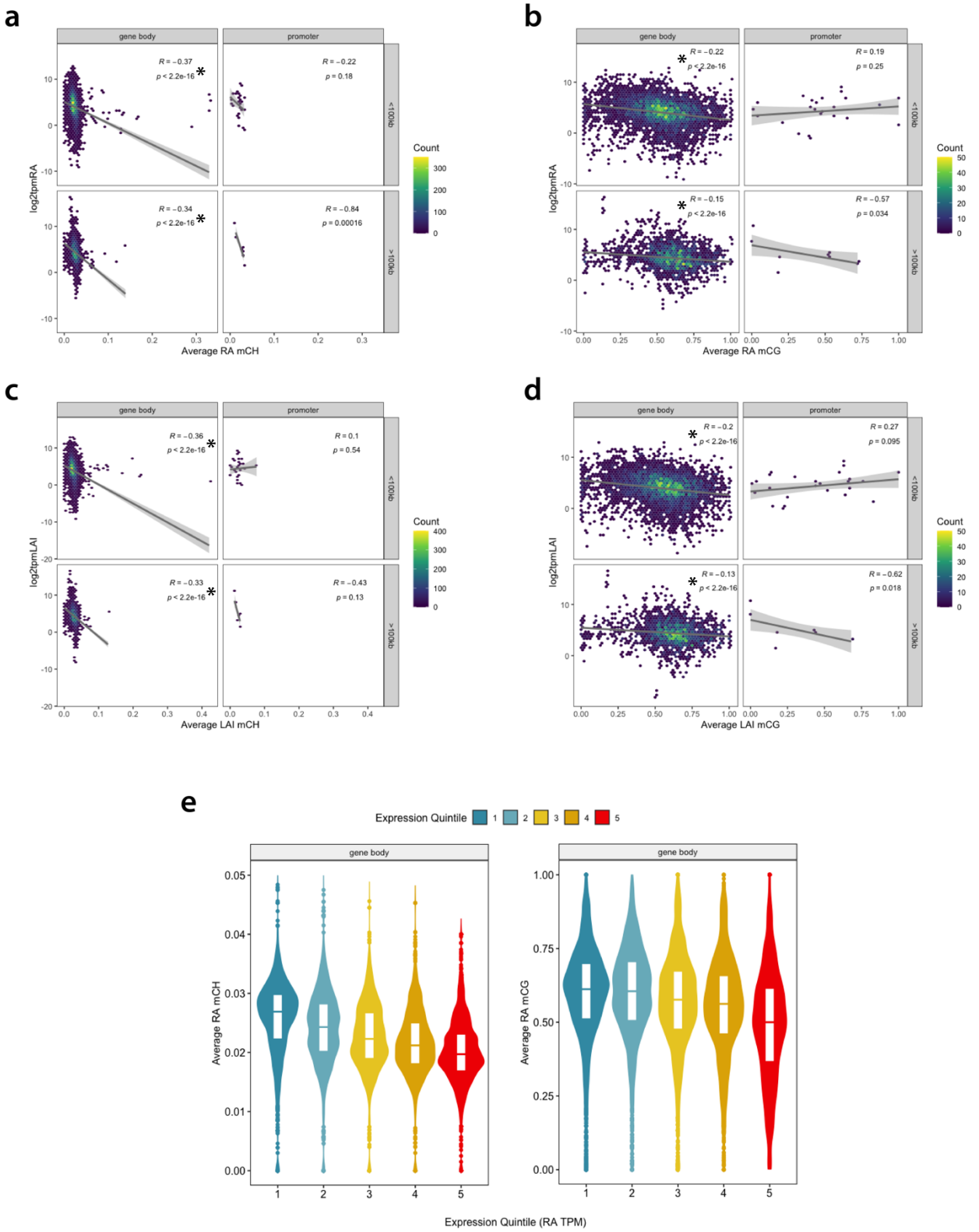
3.3.2 Unique context-specific methylation correlation patterns with gene expression

To further parse this negative correlation, we then separated the potential effects of mCG and mCH on gene expression. To do so, average methylation levels in each context were calculated for all gene body and promoter regions. For gene bodies in both mCG and mCH, there was a significant negative correlation between the level of methylation and the expression of the gene (log2TPM), with the size of the correlation being larger for mCH (Spearman $R \approx -0.3$) than for mCG (Spearman $R \approx -0.15$) (Fig 3.10a-d). This correlation remained significant irrespective of whether or not the gene was a long gene (>100kb) and whether the methylation was calculated for RA or LAI (Fig 3.10a-d). All significant p-values for correlations across Fig 3.10 were then subjected to permutation testing in which each Spearman correlation was tested over 9999 randomized simulations: the only significant correlations that were not recapitulated (therefore, significant) during this testing were gene body correlations (Fig 3.11). To analyze the gene body-TPM negative correlation further, we ranked and binned all expressed genes into five quintiles of expression and saw the same negative correlation in gene bodies (Fig 3.10 e): genes with higher TPM had lower average mCH or mCG levels. Finally, this correlation was established for all four dinucleotide contexts of cytosine methylation by calculating the mC/C level for gene bodies: mCA and mCG were the only contexts to have significant correlations (Spearman $R = -0.043, -0.12$), though both were smaller than the correlations seen in Figures 3.9-3.10 (Fig 3.12).

Figure 3.10. Correlation between gene body mCG or mCH and gene expression

On following page, **a-d**, the scatterplots for genes are visualized as hex density plots (bins = 50), with the count indicating the density of points in each hex tile (yellow = high density, dark blue = low density). The y-axis for these plots represents gene expression (calculated as log2TPM) and the x-axis represents average levels of methylation (across every gene body or promoter region). Within each of the four panels, the plot is faceted vertically to distinguish gene bodies (left) from promoters (right) and horizontally to separate genes <100kb (top) from genes >100kb (bottom). Spearman correlations and p-values are shown in the upper right corner of each sub-panel. In **a** and **c**, mCH levels are shown on the x-axis, and in **b** and **d**, mCG levels are shown. In **a** and **b**, gene expression and methylation levels are shown for RA, and, in **c** and **d**, they are shown for LAI. Significant correlations confirmed by permutation testing ($p < 0.01$) are indicated by asterisks. In **e**, the distribution of gene body mCH (left) or mCG (right) is shown for all genes ranked by RA TPM and binned into quintiles (1 = lowest gene expression, 5 = highest gene expression) in a violin plot. The distributions for each of the five quintiles (mCH and mCG separately) were statistically compared with Bonferroni-corrected t-tests between each possible pair (e.g. mCH quintile 1 and quintile 5) of distributions. All possible distribution comparisons were significant except for those between mCG quintiles 1 and 2 and 3 and 4.

Figure 3.10. Correlation between gene body mCG or mCH and gene expression



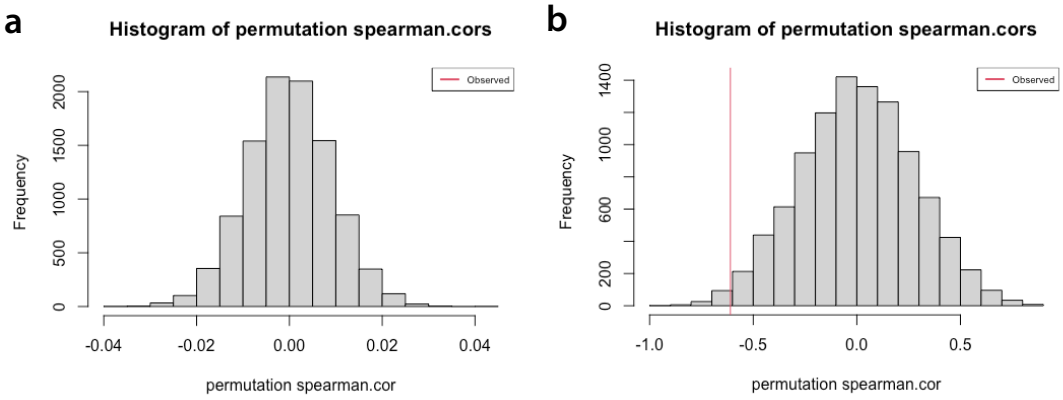


Figure 3.11. Examples of permutation testing for significant correlations

Example histograms of permutation test results for two different significant negative correlations established in Figure 3.10. On the x-axis is the spearman correlation generated from each randomized simulation. In both **a** and **b**, the spearman correlation is used as the test statistic and if the observed correlation is not recapitulated from 9999 randomized simulations, it's significant with a $p < 0.001$ (as it is in **a**). In **a**, the negative correlation being tested is for RA gene body mCH for genes >100kb, while in **b**, the negative correlation tested is for LAI promoter mCG for genes >100kb; in the former, the correlation is highly significant, as the observed correlation line doesn't appear in the range of simulations, while in the latter case, the correlation is borderline significant, as the observed correlation was found in ~150 (left of the red line) out of 9999 simulations.

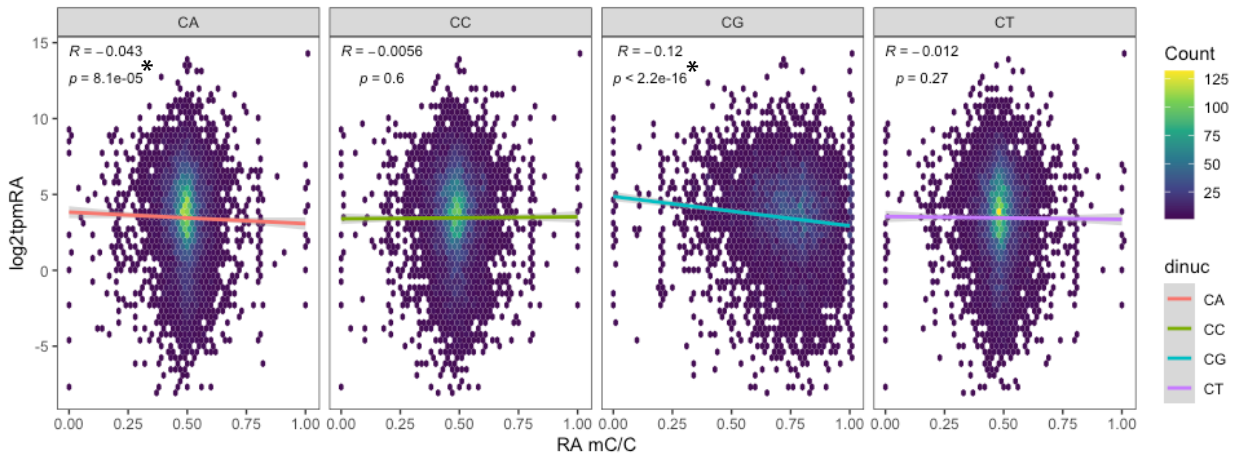


Figure 3.12. Correlation between gene body mCG or mCA and gene expression

Hex density plots as in Figure 3.10, but, here, methylation levels are calculated as the fraction of methylated cytosines for each context (mC/C) in 500bp non-overlapping windows for all gene bodies. Spearman correlations and p-values are shown for each panel (corresponding to each dinucleotide context) with asterisks indicating $p < 0.001$ from subsequent permutation testing, as shown in Figure 3.11.

3.3.3 Discussion

From these combined analyses, we find that gene body, not promoter, methylation is negatively correlated with gene expression, and that this relationship is specific to mCA and mCG. The size of the negative correlation varies from approximately -0.3 to -0.01, which likely stems from the mode of methylation calculation, i.e. averaging methylation levels across regions or calculating fractions of methylated cytosines. Despite this discrepancy, the significance of the mCA and mCG correlations is extremely robust, holding up to rigorous permutation testing. Moreover, the correlations exist regardless of gene length or region (i.e. RA or LAI), suggesting that the mechanisms creating this repressive relationship between transcriptome and methylome are consistent across the finch arcopallial genome.

CHAPTER IV: ESTABLISHING THE VOCAL LEARNING METHYLOME AND TRANSCRIPTOME OVER DEVELOPMENT

4.1 Writers and readers of DNA methylation over development in the zebra finch

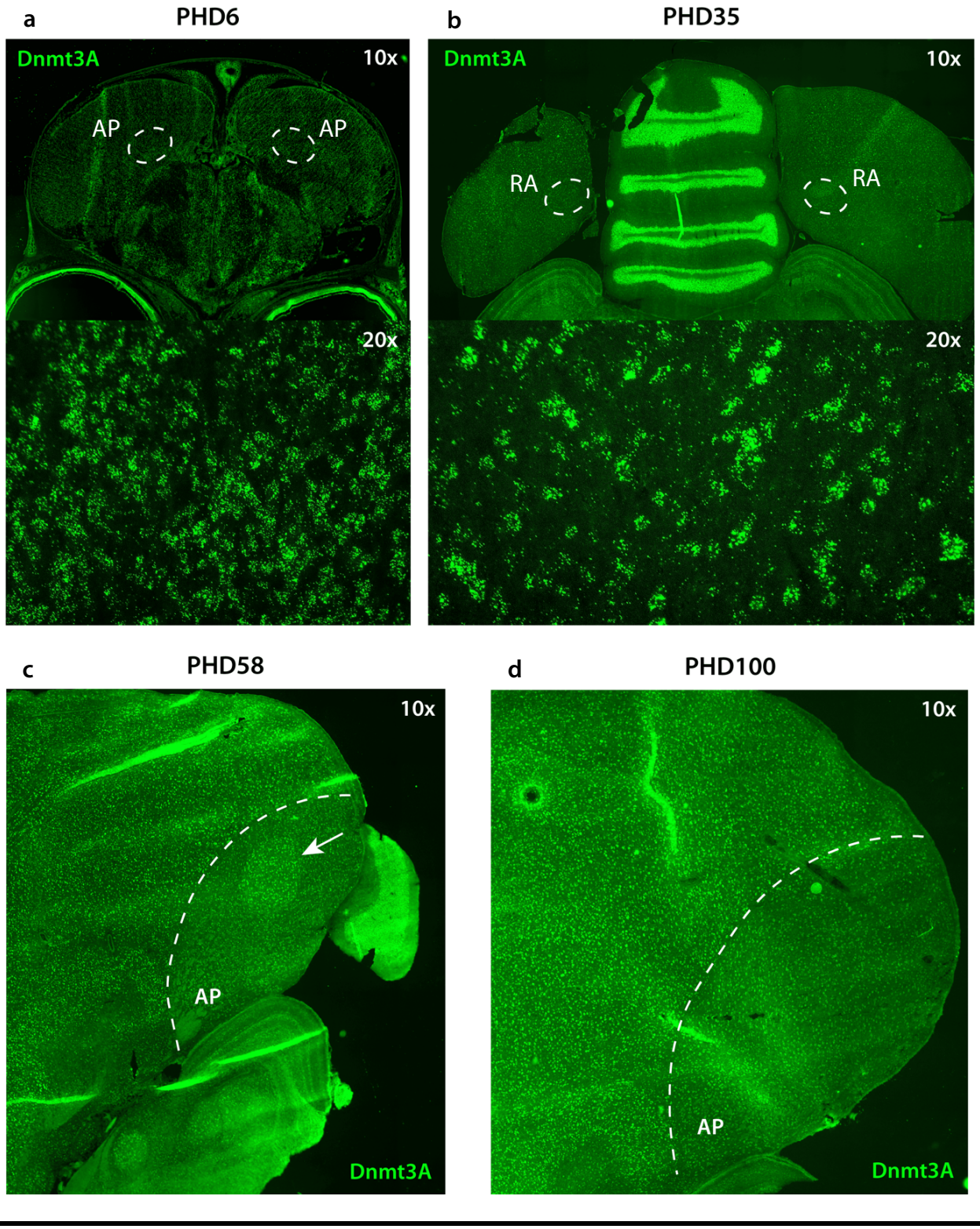
4.1.1 Developmentally-specific upregulation of Dnmt3A in the RA song nucleus

Why are certain differentially expressed genes differentially methylated between RA and LAI, while others are not? To begin to answer this question and understand how the methylome is built, the writers and readers of DNA methylation must be profiled. Based on both RNA-Seq and in situ data, we found that the canonical *de novo* methylation writer, *DNMT3A*, is expressed in the zebra finch brain. As in the mammalian brain[32], mRNA levels for *DNMT3A* are high in the juvenile finch brain (PHD6-35) and subsequently drop down in adulthood (>PHD90) (Figure 4.1, Table 4.1). Profiling the mRNA expression of DNMT3A using fluorescent in situ hybridization (fISH) at multiple developmental timepoints provided a more complete picture beyond the two developmental timepoints from the RNA-Seq data. Most strikingly, we find that *DNMT3A* is specifically upregulated in RA at a late juvenile age (58 PHD, n=2), but not at PHD6 (n=2), PHD35 (n=2) or in adulthood (n=3) (Fig 4.1a-d). fISH images at 20x reveal the high levels of cellular *DNMT3A* at PHD6 and PHD35 (Fig 4.1a-b).

Figure 4.1. Expression of *Dnmt3A* over zebra finch brain development

On the following page, fISH images for *DNMT3A* expression (green) in male finch brains over multiple timepoints: (**a**) PHD6, (**b**) PHD35, (**c**) PHD58, and (**d**) PHD100. For (**a**) PHD6 and (**b**) PHD35, both 10x and 20x images are shown and brains are sectioned coronally; white dotted circles denote the location of arcopallium (AP) at PHD6 or RA at PHD35 in each hemisphere, as the RA nucleus does not develop until ~PHD15, as well as represent the inset for the location of the 20x images below. For both (**c**) PHD58 and (**d**) PHD100, brains are sectioned sagitally and imaged at 10x, with the white dotted line indicating the boundary for the arcopallium (AP); the white arrow is pointing at RA.

Figure 4.1. Expression of *Dnmt3A* over zebra finch brain development



4.1.2 Other methylation writers, erasers, and readers in the finch arcopallium

To profile the other potential writers, erasers, and readers that could help to pattern the methylome, we jointly processed the RNA-Seq data from juvenile (PHD30) and adult (>PHD90) finch brains and examined the normalized raw read counts across all song nuclei (RA, HVC, LMAN, Area X) and their respective surrounding regions (LAI, PLN, AN, VSt). *DNMT3A* was higher in all brain regions in juveniles relative to adults (Table 4.1). The other *de novo* writer of DNA methylation, *DNMT3B*, was not expressed in any of the brain regions in juveniles or adults (Table 4.1). The maintenance writer of DNA methylation, *DNMT1*, was lowly expressed in the brain across development. Interestingly, of the erasers of DNA methylation, *TET1* was more highly expressed in the juvenile finch brain than in the adult brain, while *TET2* was relatively consistent over development, and *TET3* was not expressed at all. Most surprising was the absence of the canonical methylation reader, *MECP2*, and its family member, *MBD2*, from the finch brain. Of the other MBD family of proteins found in the finch genome, only *MBD3-5* were expressed across the brain (Table 4.1). High and low levels of expression were defined based on whether the normalized read count was on average <100 or >2000, respectively. Importantly, based on DESeq2 differential expression analysis for these readers, none of these genes were significantly differentially expressed between song nucleus and surround. The specific differences in expression across nuclei do indicate that the MBD proteins could be responsible for the developmental differences and specialized expression in pallial song nuclei.

Table 4.1. Methylation writers, erasers, and readers in the juvenile and adult finch brain

	Adult (PHD > 90)								Juvenile (PHD 30)							
	Arcopallium		Posterior Nidopallium		Anterior Nidopallium		Striatum		Arcopallium		Posterior Nidopallium		Anterior Nidopallium		Striatum	
	RA	LAI	HVC	PLN	LMAN	AN	AreaX	VSt	RA	LAI	HVC	PLN	LMAN	AN	AreaX	VSt
Readers																
MECP2	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
MBD2	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
MBD3	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	high	expressed	expressed	high	expressed	expressed
MBD4	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed
MBD5	expressed	low	expressed	expressed	low	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed
Writers																
DNMT3A	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	high	high	high	high	high	high	high	high
DNMT3B	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
DNMT1	low	low	low	low	low	low	low	low	low	low	low	low	low	low	low	low
Erasers																
TET1	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	high	high	high	high	high	high	high	high
TET2	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed	expressed
TET3	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent

Table shows the relative expression or absence of each of the methylation readers, writers, and erasers across all song nuclei and their surrounding regions in the adult and juvenile zebra finch.

4.1.3 The absence of MeCP2 in the zebra finch brain

To solidify the finding from the song nuclei RNA-Seq data that *MECP2* is not expressed in the brain, RNA-Seq data from finch non-neuronal tissues were also aligned to the same bTaeGut1_v1 genome and raw reads were visualized in a genome browser. Indeed, *MECP2* was expressed, albeit at low levels, in the three non-brain tissues processed (ovaries, testes, and muscle) but was completely absent in any of the brain samples (Fig 4.2).

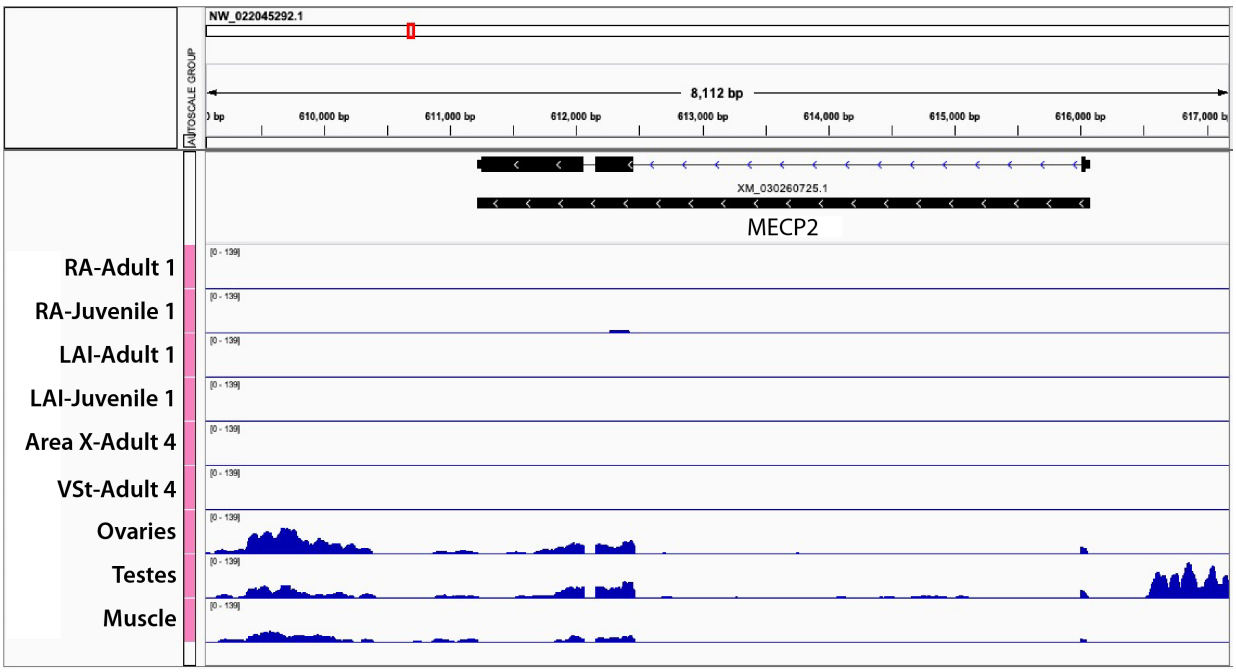


Figure 4.2. Absence of *MECP2* in the juvenile and adult zebra finch brain

IGV browser snapshot showing the read pileups for different zebra finch adult and juvenile brain samples as well as non-brain tissue samples for *MECP2*. All tracks are autoscaled for comparability. Non-brain tissues shown are zebra finch ovaries, testes, and muscles samples sequenced by the Vertebrate Genome Laboratory at the Rockefeller University. Replicate samples from adult (n = 4) and juvenile (n = 3) brain tissue cover arcopallium (RA, LAI) and the striatum (Area X, VSt).

4.1.4 Discussion

We find that of all the potential mechanistic players in establishing a methylome, only *DNMT3A* is differentially expressed in a song nucleus relative to its surrounding region, and this upregulation only occurs in RA at a specific developmental timepoint (~PHD60). The specificity of the upregulation to RA suggests greater deposition of methyl marks in RA relative to LAI,

highlighting the importance of RA hypermethylation in defining the vocal learning transcriptome for the region. One potential explanation for the specific upregulation of *DNMT3A* at ~ PHD60 is that the levels of *DNMT3A* in the RA decrease at a slightly later timepoint than they do in the surrounding arcopallium. To establish the specific window in which this RA *DNMT3A* upregulation occurs would require additional RNA-sequencing or fISH at ~PHD45-50 and ~PHD70-90. This time period around ~PHD60 is particularly important for the juvenile finch as it represents the closing of the sensory period of vocal learning, during which the young bird is taking in auditory cues and feedback from his tutor and environment.

The absence of *MECP2* in the finch brain is incredibly striking, as the only studied mechanism by which DNA methylation confers changes in transcription is via this canonical reader. Furthermore, *MECP2* is well-studied as a mammalian reader of DNA methylation in the brain. All the analyses up to this point suggest a clear relationship between gene expression and methylation in the finch genome; yet, the mechanism for this interaction requires a reader to bind to methylated DNA to then recruit other co-repressor proteins or induce Pol II pausing to impart transcriptional change. Of the three MBD proteins expressed in the juvenile and adult zebra finch brain (*MBD3-5*), only *MBD4* has been shown to bind to mCH (specifically mCAC) sites[53], making it the most likely candidate for the novel neuronal methylation reader in the zebra finch. However, further in vitro binding assays are necessary to test this hypothesis in the zebra finch.

4.2 The dynamic transcriptome across the development of vocal learning

4.2.1 Comparing transcriptomes of juvenile and adult arcopallium

The adult brain methylome represents an epigenomic footprint for historical transcriptomic changes over development. Indeed, the highly specialized RA song nucleus in the adult finch was

at one point unspecialized arcopallium in the very juvenile finch. The transformative changes in the transcriptome, facilitated in part by patterning in the methylome, pave the way for the adult RA and LAI to be such specialized brain regions. In comparing the juvenile (PHD30) and adult RNA-Seq datasets, we are able to establish two broad categories of genes: genes that become differentially expressed in the adult and genes that are already differential by PHD30.

Since the juvenile and adult samples were dissected, prepared, and sequenced completely independently, direct differential comparisons between adult RA and juvenile RA or adult LAI and juvenile LAI would be entirely confounded by batch effects. To avoid these comparisons, RA and LAI were compared within each age's samples. Using the same DESeq2 pipeline we used for the adult RA versus LAI (n = 4) DEG analyses, we further identified juvenile DEGs between RA and LAI (n = 3) (Fig 4.3). Overall, RA and LAI are significantly less distinct from one another at PHD30 than in adulthood, as only 212 genes are differentially expressed between RA and LAI in the juvenile brain, while there are 781 DEGs in adult arcopallium (Fig 4.3). Moreover, only 86 genes have shared DEG status, i.e. are differentially expressed in both, between the juvenile and adult arcopallium, with a majority (55/86) of these genes being differentially downregulated at both ages (Fig 4.4).

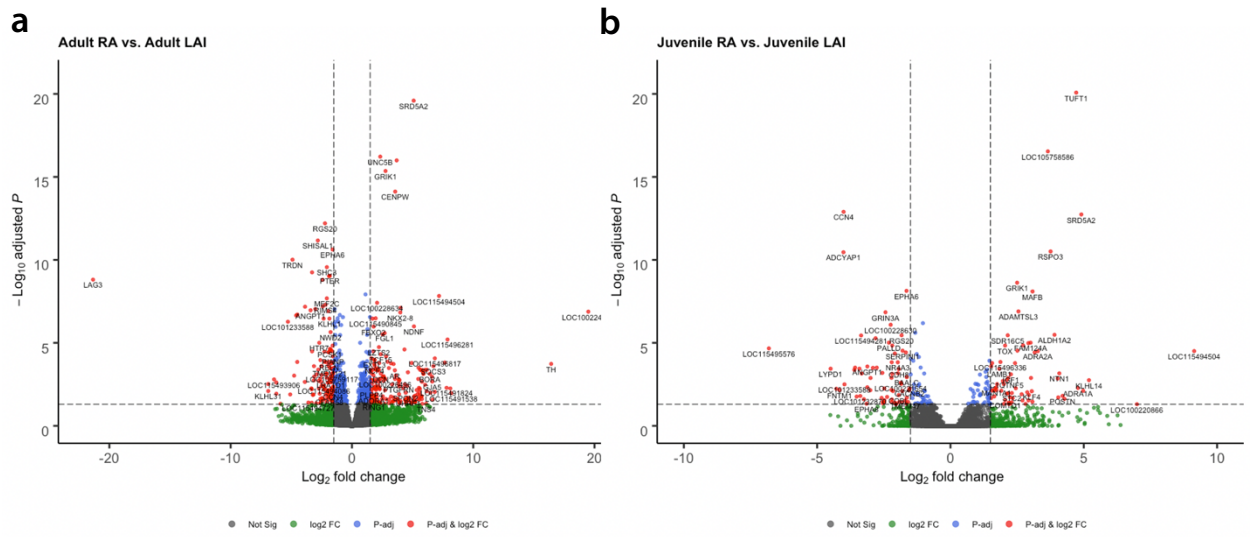


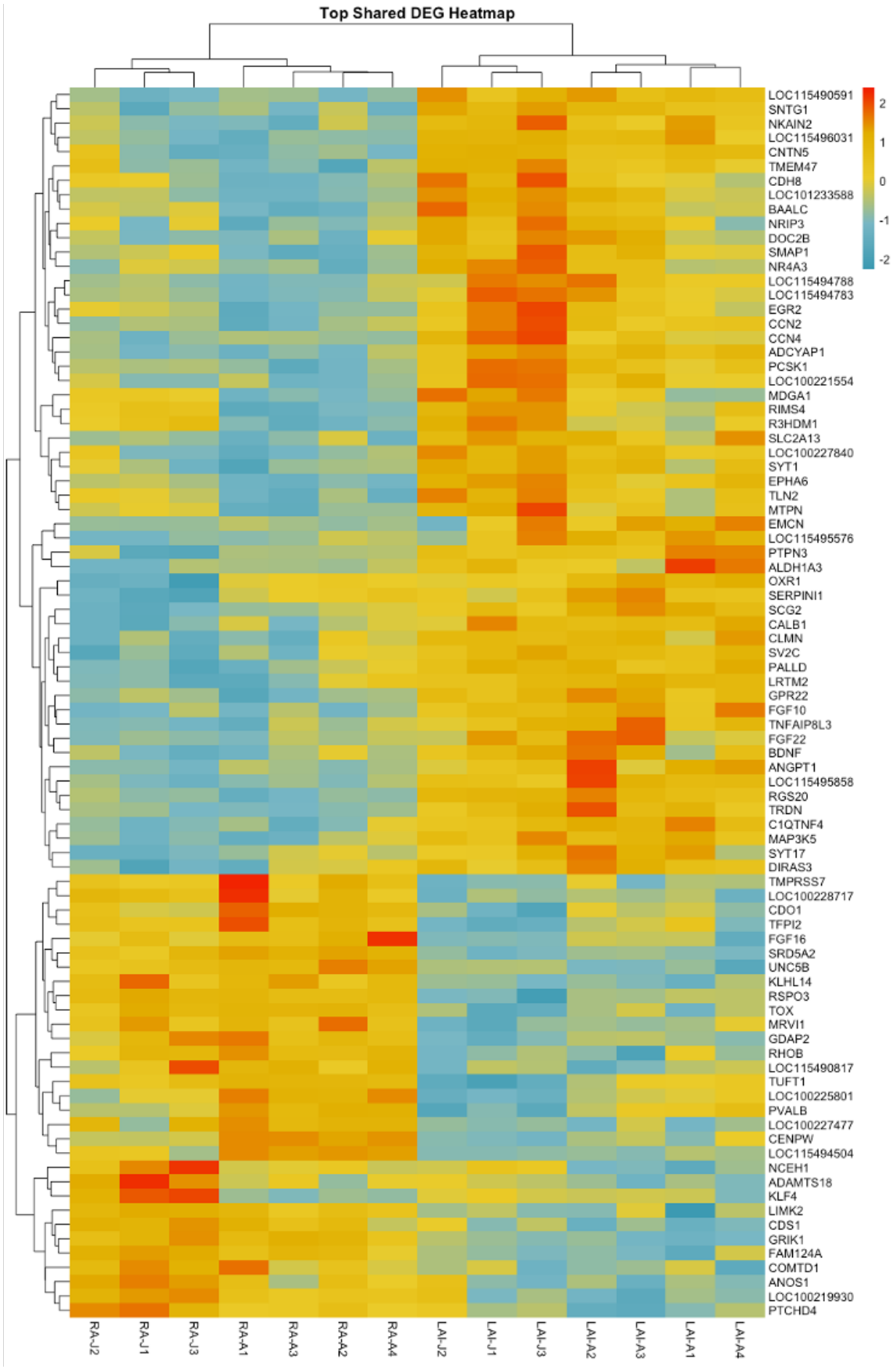
Figure 4.3. Differentially expressed genes between adult or juvenile RA and LAI

Volcano plots for differentially expressed genes between RA and LAI in the adult brain (**a**) and in the juvenile brain (**b**). In both **a** and **b**, the relative change in expression (log_2 fold change) is plotted against the normalized adjusted p-value ($-log_{10}$ adjusted p) for each gene. Colors are associated with different thresholds: green = genes only with log_2foldchange > |1|, blue = genes only with padj ≤ 0.01, red = genes that meet both the green and blue thresholds for fold change and adjusted p-value, grey = not significantly differential by either fold change or adjusted p-value.

Figure 4.4. Expression of shared arcopallial DEG between adult and juvenile

On next page, heatmap of the top shared differentially expressed genes between RA and LAI in both adult and juvenile arcopallium (i.e. genes are DEG in adult and in juvenile). Columns represent all arcopallial samples (adult: A1-4, juvenile: J1-3), where each row shows the z-score (normalized expression) for each gene, where red represents upregulation and blue represents downregulation.

Figure 4.4. Expression of shared arcopallial DEG between adult and juvenile



4.2.2 The adult methylome as an epigenomic footprint for transcriptomic changes

Having established the different gene sets across the developing arcopallial transcriptome, methylome data could then be overlaid onto these gene sets. Since the methylome data is from the adult arcopallium, it was important to comparably match this data to the adult DEG set, broken down into the two broad groups of genes: genes that become differentially expressed between PHD30 and PHD90 (i.e. not DE at PHD30 and DE in the adult) and genes that are already differentially expressed by PHD30 (i.e. DE at PHD30 and in the adult). In splitting the genes in this way, two developmental timelines are presented for how the methylome could be shaping the transcriptome in the arcopallium: differential methylation observed in the adult was deposited before or after PHD30, the beginning of the key vocal learning critical period in the finch.

Of the 781 adult DEGs, 695 become differentially expressed, while 86 are already differentially expressed by PHD30 (Fig 4.5). In the former group, approximately half the genes are upregulated and half are downregulated (in RA relative to LAI); in the latter group of shared DEGs, 30 genes are upregulated, 56 are downregulated, and 1 changes the direction of its differential expression from upregulated to downregulated (*KLF4*) between PHD30 and adulthood. Within genes that become differentially expressed, a much higher proportion of downregulated genes have DMRs (30%) than do upregulated genes (9%), while the DMR proportion is about the same for up- and down-regulated genes in the shared DEGs set (33%, 27%) (Fig 4.5). Furthermore, the number of DMRs per gene for upregulated DEGs is ~1 DMRs/gene, but ~2 DMRs/gene for downregulated genes: there are 189 DMRs across the 93 genes with DMRs in the adult-only downregulated geneset, but just 45 DMRs across the 34 genes with DMRs in the adult-only upregulated geneset. Several of the previously defined candidate genes also appear in the 93 gene set (Fig 4.5), including *CNTNAP5*, *GRIK2*, *NEUROD6*, *RELN*, and *NRXN1*.

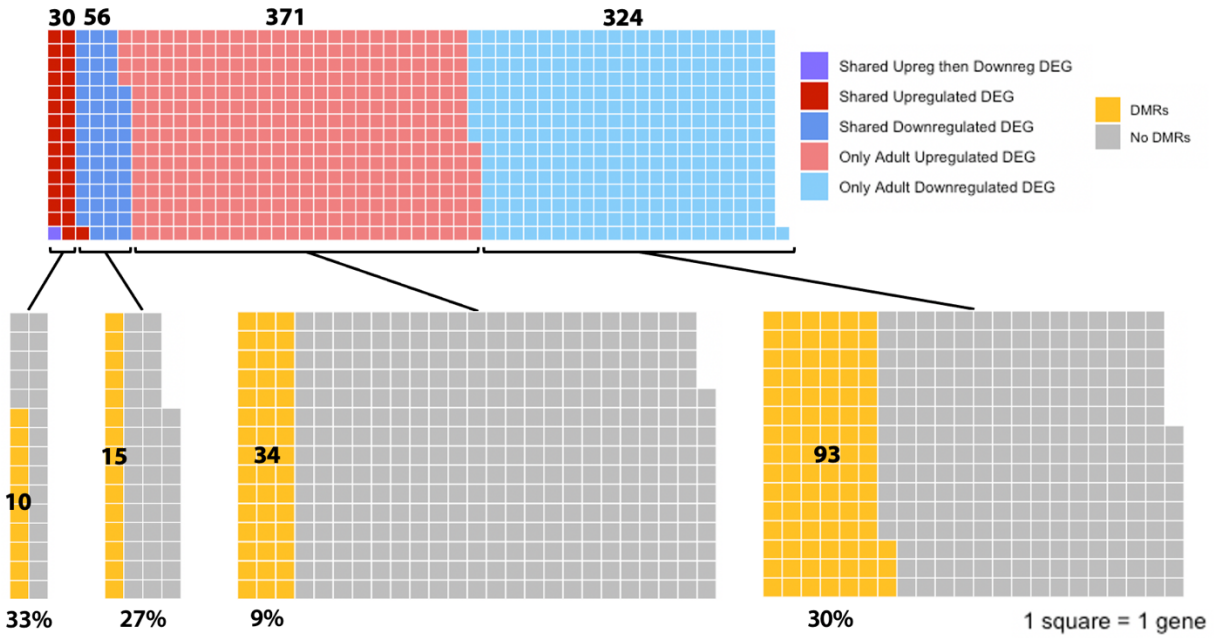


Figure 4.5. Proportions of DEGs with DMRs across development

Waffle plots for all genes differentially expressed between RA and LAI in the adult zebra finch brain, where each square is one gene. The top waffle plot represents all adult DEGs colored according to DEG status: red = upregulated in both juvenile and adult, blue = downregulated in both juvenile and adult, coral = not DE in juvenile and upregulated in adult, light blue = not DE in juvenile and downregulated in adult, purple = upregulated in juvenile and downregulated in adult. The four lower waffle plots show the proportion of genes with DMRs (golden yellow) in each of these genesets (except for purple). For each section of DMRs, both the number of genes and the DMR proportion are shown.

4.2.3 Discussion

While there were limitations for the extent of comparison between the juvenile and adult RNA-Seq data, the differential expression analysis revealed that the arcopallium is much less transcriptomically specialized into RA and LAI in the juvenile as it is in the adult. Future RNA-seq experiments that combine developmental and adult finch arcopallial samples under well-controlled conditions (e.g. same sequencing runs and library preparations) would allow for RA and LAI to be separately studied in how their respective transcriptomes develop. We also found that the subset of genes that becomes differentially downregulated in adulthood has a significant number of DMRs. The specific upregulation of *DNMT3A* in RA in the middle of the critical PHD30 to PHD90 window provides a hypothetical mechanism for how this geneset could become differentially downregulated. The higher number of DMRs per gene seen in downregulated genes is further supported by the significant negative correlation we found between hypermethylation and reduced gene expression.

CHAPTER V: CONCLUSIONS AND FUTURE DIRECTIONS

5.1 Conclusions

5.1.1 Gene body mCH and downregulation of gene expression

In summary, these results reveal the importance of gene body mCH not only as a broad epigenomic mechanism for defining gene expression in a non-mammalian vertebrate brain, but also as a key feature in a subset of the specialized vocal learning transcriptome. DMR analyses, metagene analyses, and correlational testing all point to RA gene body hypermethylation as a mechanism by which genes can be differentially downregulated in RA. Future work that profiles individual neuronal cell types will likely reveal cells with specific sets of genes within the specialized transcriptome that exhibit unique methylome patterns, though we would expect the inverse relationship between methylation and expression to be conserved across cell types.

We further show that the vast majority of transcriptomic specialization in the arcopallium occurs during the developmental window between song onset (PHD30) and the crystallization of song (PHD90). As the adult methylome serves as an epigenomic footprint for the history of transcriptomic changes over zebra finch brain development, we find that differential methylation defines a subset of genes that become differentially downregulated in RA. Methylome profiling at PHD30-45 would provide a more complete picture of these epigenomic dynamics, though single-nuclei approaches combined with pooling of many samples is likely required to generate sufficient replicates for a reasonable cost across development. Additionally, while it is still unknown how RA and LAI individually change over development, the upregulation of *DNMT3A* in RA at ~PHD60 suggests that RA is hypermethylated, thereby downregulating expression of genes within

RA during development. Additional epigenomic mechanisms like conformational changes in chromatin likely also play a role in defining gene expression in the arcopallium over development.

5.1.2 Potential mechanisms for maintaining vertebrate DNA methylomes

Many of the methylome features seen in the finch brain mirror what has been shown in the mammalian brain. However, the distributions of mCG and mCH as well as the absence of *MECP2* in the finch brain strongly suggest a novel mechanism for reading and writing non-mammalian vertebrate brain (or potentially, a uniquely bird-specific mechanism) genomes. Our survey of potential readers in the finch brain points to *MBD4* as a likely alternative reader, as it is expressed throughout the finch brain and has been shown to bind to mCH (mCA, notably), thereby conferring transcriptional changes. However, detailed developmental profiling, binding assays, and loss-of-function experiments would be necessary to show how *MBD4* functions in zebra finch neurons.

From the finding that many DEGs had more than one DMR (on average, ~2), it was critical to see if gene length potentially played a role in the extent of differential methylation. In the mammalian brain, it has been shown that in the absence of *MECP2*, longer genes tend to be significantly misregulated, as the length of the gene corresponds with the amount of mCA for *MECP2* to bind with[39]. We find that genes with DMRs are significantly longer than DEGs and a random set of genes expressed in the arcopallium (Fig 5.1). DEGs also tend to be longer than the random set of expressed genes, but not as long as genes with DMRs. It is unclear if longer genes tend to be more methylated than are shorter genes, but it is possible that genes with DMRs might have more mCH sites for which a reader like *MBD4* could bind to mediate transcriptional repression. It will be especially critical to profile neurons rather than bulk tissue to clarify this

potential mechanism, as the methylomes of neuronal and non-neuronal populations are functionally distinct[19,32].

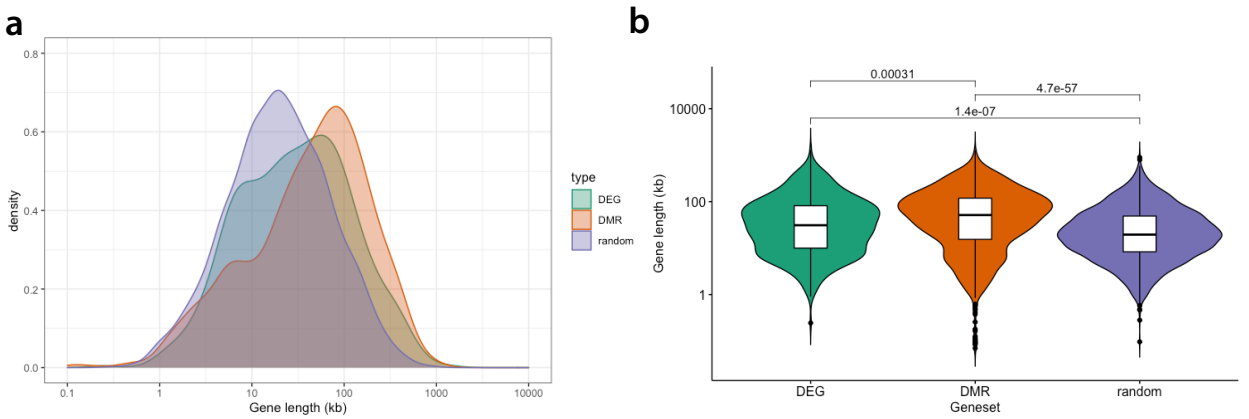


Figure 5.1. Distribution of gene lengths for different genesets

Plots showing the gene length distributions for different sets of genes: green = DEG, orange = genes with DMRs, and purple = a random set of 2,000 genes expressed in the arcopallium. **a**, density plot of the gene lengths (in kb) for the three sets of genes. Two-sample Kolmogorov-Smirnov tests between genes with DMRs and DEGs and between genes with DMRs and the random set of genes are highly significant ($p < 2.2e^{-16}$ for both comparisons). **b,** a violin plot comparing the lengths of the genesets, with Bonferroni-corrected t-test p-values shown for each comparison.

5.2 Future Directions

5.2.1 Cell-type-specific methylomes and transcriptomes in the finch arcopallium

While we have learned much about the brain epigenome with bulk-tissue methylome data, neuronal cell-types do exhibit specific methylation patterns, which this study could not capture. Comparably, all RNA-Seq data analyzed in this study came from bulk tissue, though recent work

has profiled some song nuclei (no surround regions) at the single-cell level[23]. A potential future experiment would involve single-nuclei methods to assay methylome patterns of the extremely heterogeneous arcopallial tissue. The RA nucleus, specifically, has a significant proportion of non-neuronal cells, and it is well-established that neuronal versus non-neuronal cell populations are marked by distinct methylome (especially mCH) patterns[19,32]. Beyond the arcopallium, it would be extremely interesting to see if neurons in song nuclei in the PFP versus the AFP exhibit unique methylomes. Future work that sorts NeuN+/- populations could better clarify the interaction between methylation and transcription in specific cell types, though would likely require pooling tissue across many animals. Single-nuclei protocols could potentially also allow for RNA and DNA to be collected from the same samples, though pooling might also be required across animals.

5.2.2 In vivo functional studies of DNA methylation in the songbird brain

Our findings identifying both *DNMT3A* as a critical writer of methylation in the arcopallium and a set of candidate genes with RA gene body hypermethylation have laid the groundwork for exciting functional studies in the developing juvenile zebra finch. One potential experiment could involve the localized knockdown of *DNMT3A* in the developing RA: a *DNMT3A* shRNA injection into a PHD15 RA with methylome and transcriptome profiling before and after treatment would show whether *DNMT3A* is necessary for generating the specialized methylome in the arcopallium. Moreover, great care was taken in this study to control for the effects of behavior on both the transcriptome and methylome, with the hope that these results could serve as an epigenomic baseline for future behavioral experiments that manipulate the conditions of the vocal learning critical period, e.g. prevention of singing in the juvenile or no tutor exposure. For

each behavioral perturbation, the transcriptome and methylome would be profiled before and after the specific behavioral treatment.

REFERENCES

1. Doupe, A. J. & Kuhl, P. K. BIRDSONG AND HUMAN SPEECH: Common Themes and Mechanisms. (2003) doi:10.1146/annurev.neuro.22.1.567.
2. Petkov, C. I. & Jarvis, E. D. Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Front. Evol. Neurosci.* **4**, 12 (2012).
3. Deecke, V. B., Ford, J. K. & Spong, P. Dialect change in resident killer whales: implications for vocal learning and cultural transmission. *Anim. Behav.* **60**, 629–638 (2000).
4. Boughman, J. W. Vocal learning by greater spear-nosed bats. *Proc. Biol. Sci.* **265**, 227–233 (1998).
5. Poole, J. H., Tyack, P. L., Stoeger-Horwath, A. S. & Watwood, S. Animal behaviour: elephants are capable of vocal learning. *Nature* **434**, 455–456 (2005).
6. Janik, V. M. & Slater, P. J. The different roles of social learning in vocal communication. *Anim. Behav.* **60**, 1–11 (2000).
7. Wirthlin, M. *et al.* A Modular Approach to Vocal Learning: Disentangling the Diversity of a Complex Behavioral Trait. *Neuron* **104**, 87–99 (2019).
8. Jarvis, E. D. Learned birdsong and the neurobiology of human language. *Ann. N. Y. Acad. Sci.* **1016**, 749–777 (2004).
9. Bolhuis, J. J., Okanoya, K. & Scharff, C. Twitter evolution: converging mechanisms in birdsong and human speech. *Nat. Rev. Neurosci.* **11**, 747–759 (2010).
10. Gervain, J. Plasticity in early language acquisition: the effects of prenatal and early childhood experience. *Curr. Opin. Neurobiol.* **35**, 13–20 (2015).

11. Solis, M. M., Brainard, M. S., Hessler, N. A. & Doupe, A. J. Song selectivity and sensorimotor signals in vocal learning and production. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 11836–11842 (2000).
12. Choe, H. & Jarvis, E. D. The role of sex chromosomes and sex hormones in vocal learning systems. *Horm. Behav.* (2021).
13. Pfenning, A. R. *et al.* Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* **346**, 1256846 (2014).
14. Gedman, G. *et al.* As above, so below: Whole transcriptome profiling supports the continuum hypothesis of avian dorsal and ventral pallium organization. *Cold Spring Harbor Laboratory* 2020.11.13.375055 (2020) doi:10.1101/2020.11.13.375055.
15. Feenders, G. *et al.* Molecular mapping of movement-associated areas in the avian brain: a motor theory for vocal learning origin. *PLoS One* **3**, e1768 (2008).
16. Mello, C. V., Vates, G. E., Okuhata, S. & Nottebohm, F. Descending auditory pathways in the adult male zebra finch (Taeniopygia guttata). *J. Comp. Neurol.* **395**, 137–160 (1998).
17. Margoliash, D. *et al.* Distributed representation in the song system of oscines: evolutionary implications and functional consequences. *Brain Behav. Evol.* **44**, 247–264 (1994).
18. Guo, J. U. *et al.* Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.* **17**, 215–222 (2014).
19. Mo, A. *et al.* Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron* **86**, 1369–1384 (2015).
20. He, Y. & Ecker, J. R. Non-CG Methylation in the Human Genome. *Annu. Rev. Genomics Hum. Genet.* **16**, 55–77 (2015).

21. Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).

22. Liu, H. *et al.* DNA Methylation Atlas of the Mouse Brain at Single-Cell Resolution. *bioRxiv* 2020.04.30.069377 (2020) doi:10.1101/2020.04.30.069377.

23. Colquitt, B. M., Merullo, D. P., Konopka, G., Roberts, T. F. & Brainard, M. S. Cellular transcriptomics reveals evolutionary identities of songbird vocal circuits. *Science* **371**, (2021).

24. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).

25. Huang, K. & Fan, G. DNA methylation in cell differentiation and reprogramming: an emerging systematic view. *Regen. Med.* **5**, 531–544 (2010).

26. Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23**, 185–188 (1999).

27. Kulis, M. & Esteller, M. 2 - DNA Methylation and Cancer. in *Advances in Genetics* (eds. Herceg, Z. & Ushijima, T.) vol. 70 27–56 (Academic Press, 2010).

28. Huang, Y. & Rao, A. Connections between TET proteins and aberrant DNA modification in cancer. *Trends Genet.* **30**, 464–474 (2014).

29. Woodcock, D. M., Crowther, P. J. & Diver, W. P. The majority of methylated deoxycytidines in human DNA are not in the CpG dinucleotide. *Biochem. Biophys. Res. Commun.* **145**, 888–894 (1987).

30. Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5237–5242 (2000).

31. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).

32. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).

33. Wu, T. P. *et al.* DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature* **532**, 329–333 (2016).

34. Li, X. *et al.* The DNA modification N6-methyl-2'-deoxyadenosine (m6dA) drives activity-induced gene expression and is required for fear extinction. *Nat. Neurosci.* **22**, 534–544 (2019).

35. De Felipe, J., Marco, P., Fairén, A. & Jones, E. G. Inhibitory synaptogenesis in mouse somatosensory cortex. *Cereb. Cortex* **7**, 619–634 (1997).

36. Huttenlocher, P. R. & Dabholkar, A. S. Regional differences in synaptogenesis in human cerebral cortex. *J. Comp. Neurol.* **387**, 167–178 (1997).

37. Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**, 91–99 (1985).

38. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**, 425–432 (2007).

39. Gabel, H. W. *et al.* Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature* **522**, 89–93 (2015).

40. Stroud, H. *et al.* Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. *Cell* **171**, 1151-1164.e16 (2017).

41. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).

42. Hu, L. *et al.* Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation. *Cell* **155**, 1545–1555 (2013).

43. Maiti, A. & Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J. Biol. Chem.* **286**, 35334–35338 (2011).

44. He, Y.-F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).

45. Kinde, B., Gabel, H. W., Gilbert, C. S., Griffith, E. C. & Greenberg, M. E. Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6800–6806 (2015).

46. Lewis, J. D. *et al.* Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* **69**, 905–914 (1992).

47. Nan, X., Campoy, F. J. & Bird, A. MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell* **88**, 471–481 (1997).

48. Nan, X. *et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386–389 (1998).

49. Lyst, M. J. *et al.* Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor. *Nat. Neurosci.* **16**, 898–902 (2013).

50. Lagger, S. *et al.* MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. *PLoS Genet.* **13**, e1006793 (2017).

51. Cholewa-Waclaw, J. *et al.* Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 14995–15000 (2019).

52. Sperlazza, M. J., Bilinovich, S. M., Sinanan, L. M., Javier, F. R. & Williams, D. C., Jr. Structural Basis of MeCP2 Distribution on Non-CpG Methylated and Hydroxymethylated DNA. *J. Mol. Biol.* **429**, 1581–1594 (2017).

53. Liu, K. *et al.* Structural basis for the ability of MBD domains to bind methyl-CG and TG sites in DNA. *J. Biol. Chem.* **293**, 7344–7354 (2018).

54. Tillotson, R. *et al.* Neuronal non-CG methylation is an essential target for MeCP2 function. *Mol. Cell* (2021) doi:10.1016/j.molcel.2021.01.011.

55. Nikolova, E. N., Stanfield, R. L., Dyson, H. J. & Wright, P. E. CH···O Hydrogen Bonds Mediate Highly Specific Recognition of Methylated CpG Sites by the Zinc Finger Protein Kaiso. *Biochemistry* **57**, 2109–2120 (2018).

56. Johnson, L. M. *et al.* The SRA methyl-cytosine-binding domain links DNA and histone methylation. *Curr. Biol.* **17**, 379–384 (2007).

57. Wada, K. *et al.* A molecular neuroethological approach for identifying and characterizing a cascade of behaviorally regulated genes. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 15212–15217 (2006).

58. Choe, H. N. *et al.* Estrogen and sex-dependent loss of the vocal learning system in female zebra finches. *Horm. Behav.* **129**, 104911 (2021).

59. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

60. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

61. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

62. Michael I Love, W. H. &. S. A. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* (2014).

63. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

64. Weilong Guo, Petko Fiziev, Weihong Yan, Shawn Cokus, Xueguang Sun, Michael Q Zhang, Pao-Yang Chen & Matteo Pellegrini. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics* (2013).

65. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

66. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

67. Guo, W. *et al.* CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. *Bioinformatics* **34**, 381–387 (2018).

68. Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212–216 (2015).

69. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-5 (2016).

70. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).

71. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).

72. Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* **41**, 200–209 (2012).

73. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

74. Steyaert, S. *et al.* A genome-wide search for epigenetically regulated genes in zebra finch using MethylCap-seq and RNA-seq. *Sci. Rep.* **6**, 20957 (2016).

75. George, J. M. *et al.* Acute social isolation alters neurogenomic state in songbird forebrain. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 23311–23316 (2020).

76. Sun, Z., Cunningham, J., Slager, S. & Kocher, J.-P. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics* **7**, 813–828 (2015).

77. de Mendoza, A. *et al.* The emergence of the brain non-CpG methylation system in vertebrates. *Nat Ecol Evol* **5**, 369–378 (2021).

78. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **8**, 1551–1566 (2013).

79. Hu, Y. *et al.* Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.* **17**, 88 (2016).