The Dissertation Committee for Alejandro Berrío Escobar Certifies that this is the approved version of the following dissertation:

Gene Regulatory Evolution and the Origin of Complex Behaviors in the Prairie Vole, *Microtus ochrogaster*

Committee:

Steven M. Phelps, Supervisor

Johann Hofmann

Mikhail V. Matz

Nigel S. Atkinson

Thomas E. Juenger

Gene Regulatory Evolution and the Origin of Complex Behaviors in the Prairie Vole, *Microtus ochrogaster*

by

Alejandro Berrío Escobar, BSc.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2016

Dedication

I dedicate my dissertation to Sarita, Martín, Juli, el hermanito Andrés,
Hernán Hugo and my mother Irma.

Acknowledgements

Some stars come to Texas and shine by themselves, while many more will vanish in its infinite lands and skies. It would be impossible for my lonely heart to be shining at this moment of life, without the kind support of many people who have contributed to my own scientific and personal growth. First, I must thank my advisor Steve Phelps, not only for financial and educational support, but for all the scientific discussions, experiences, travels, lunches, coffees, dinners, drinks, tears and laughs. I will always be thankful for spending more than 10% of my life under your mentorship. Indeed, I feel very lucky and happy of having you as my mentor and for letting me be a member of your lab; I will always appreciate your patience, care and support. Thanks for letting me watch Leo and Lucy when you and Jason were gone. Likewise, thanks to everybody in my lab: Mariam, Tracy, Gerard, David, Stavana, Erin and Jason I. Next, I need to thank my committee: Misha, Hans, Nigel and Tom. Thank you for offering me your time and experience. Thank you again for your advice, for letting me to work in your labs, to use your equipment, and to learn from you. The more I write this section, the more complex it gets because you brought people who became my friends. Thank you Galina Aglyamova, your meticulousness is amazing, I am grateful of learning to do 2bRAD from you.

Thank you Mike Ryan and Mathew Leibold for your support during my early stages in grad school, I know it was very difficult to me but I am grateful that you empowered the vanishing star in me. I will always be thankful with all the staff members in EEB and IB, especially to Tamra Rogers. I am also grateful with Drs Larry Young and Lisa McGraw who kindly shared their RNAseq data. Special thanks also to my talented friend Rafael Guerrero, who taught and helped me with some population genetic analysis.

My deep acknowledgments to my dear friends, who I admire and respect, those who made the best of being in Patterson and anywhere in Austin, I am very fortunate to have you because we could always talk about anything. Thank you: Becca Tarvin, Marie Strader, Sean Maguire, Patricia Salerno, Sofia Rodriguez and Catalina Cuellar.

It would be impossible to forget all the very special people who made me feel loved and well accepted over my time in Austin. I am deeply thankful to Moises Bernal (our jokes will always be the best!), Vicky Huang, Mariska Brady, Laura Abondano, Amalia Diaz, Juan Diego Palacio, Monica Guerra, Tamara Tabbakh, Ignacio Gallardo, Carlos Guarnizo, Tinisha Hankock, Alejandro Puyana, Liz Milano, Ammon Thomson, Louise and Ben Liebeskind, Maria José La Rota, Nathan Leclear, Sarah Davies, Sarah Sussman, Edgardo Ortiz, Teo Nakov, Lina Maria Valencia, Dimitri Blondel, Bret Pasch, and all-time soccer stars in my Competitive Excluders team, thank you Lynn, Patrick, Gautam, Jose Luis, Max, Spencer, etc.

Osquitar, thanks for being my brother. I truly appreciate for being close to me, in the good and in the most challenging times. I am very grateful with Mendy Black… Despite I will never know if you are the worst, or the best. I know that you are wonderful because you make everything better and easier. Thank you, Melinda and Robert Black. I know I will always have a place to visit in Texas where I will feel like at home.

To finish, I am deeply grateful with all my family in Colombia, you never questioned my interests in art, science and nature. Sarita and Martin, you are my inspiration to keep doing science. I love you!

Gene Regulatory Evolution and the Origin of Complex Behaviors in the Prairie Vole, *Microtus ochrogaster*

Alejandro Berrío Escobar, PhD.

The University of Texas at Austin, 2016

Supervisor: Steven M. Phelps

Understanding variation in form and behavior within and among species requires mapping genotypes to phenotypes. Much of this variation depends on differences in regulatory DNA scattered throughout the genome; in the context of behavior, these regulatory sequences govern gene expression in regions of the brain that shape behavior. Surprisingly few studies have characterized the regulatory changes that underlie the adaptive evolution of brain and behavior. In my PhD dissertation project, I investigated the adaptive role of gene regulation in the evolution of pair-bonding and sexual fidelity in the prairie voles, *Microtus ochrogaster*. Expression of *Avpr1a* in the ventral pallidum plays a critical role in the origin and evolution of pair-bonding in these monogamous voles. In Chapter 1, I have applied phylogenetic and population genetic methods to find signatures of selection in functional elements in the prairie vole genome. I identified a regulatory element of the *Avpr1a* locus that is under positive selection, this sequence coincides with the origins of expression of this gene in a reward region, the ventral pallidum. Then, I tested its causality using transgenic mouse enhancer assays. I found that transgenic mice expressing a reporter under the control of this prairie vole enhancer were able to drive expression in the ventral pallidum, but expression was sensitive to insertion site. Interestingly, this gene also shows profound differences between individuals. In Chapter 2, I applied population genomic tools to demonstrate that this locus shows signatures of balancing selection in a polymorphic enhancer that predicts expression in a spatial memory circuit. I found that alleles that predict aspects of space use and sexual fidelity are strongly linked to each other. Moreover, I show evidence that the evolution of this regulatory element seems to be mediated by a mix of balancing, epistatic and density-dependent selection. In Chapter 3, I performed RNA-sequencing experiments to analyze monogamy-related genomic changes in the brain. I found massive changes in gene expression of prairie voles in contrast to promiscuous meadow voles, despite their gene expression modules are very well preserved. Moreover, neuroplasticity –a neural process involved with learning— was strongly activated in prairie but not it meadow vole brains. Overall, the results of these experiments reveal the potential for gene regulation to drive the adaptive evolution of complex behaviors.

Table of Contents

List of Tables

List of Figures

INTRODUCTION

Gene Regulatory Evolution

Evolution by natural selection occurs in both coding and non-coding sequences that govern the development and function of morphological, physiological and behavioral traits. Identifying the regulatory factors that influence the development and evolution of adaptive traits is becoming increasingly crucial as new genomic methods are becoming available to molecular ecologists. Moreover, empiric and theoretic work suggests that *cis*-regulatory variation contributes largely to morphology, physiology and behavior (Wray 2007). It is remarkably fascinating to think that within all the mammalian gene regulatory sequences, there are many of the instructions by which a small group of sinapsids with nocturnal habits evolved and radiated forms and behaviors that allowed them to colonize all kinds of diurnal, arid, aquatic, and aerial environments. The evolution and ecological social interactions of most animals can be influenced by behavioral traits that are determined by neuronal and genomic processes affecting cognition, for example: perception, learning, memory and decision making (Dukas, 2004; Rittschkof and Robinson, 2014). However, research addressing the genetic mechanisms of behavior and cognition is uncommon; perhaps, this has been a consequence of the popular idea of the "phenotypic gambit". The phenotypic gambit is the assumption by which researchers disregarded the intrinsic genetic mechanisms in the course of the understanding the evolutionary trajectories of behavior (Grafen, 1984). In this dissertation, I integrate the use of various methods from molecular genetics, functional genomics and computational biology to gain better insights into the adaptive and evolutionary origin of monogamy in prairie voles.

The first studies that brought concise explanations on the importance of gene regulation were initially made by François Jacob and Jacques Monod who revealed the model of Lac operon. This lac element regulates the expression of the b-galactosidase enzyme in enteric bacteria as a function of the presence of lactose (Jacob & Monod 1961). This model unified many ideas of how genes are expressed and helped to understand how bacterial DNA sequences coordinate the synthesis of RNA and ultimately proteins. This prokaryotic model of gene regulation advanced the field but fails to explain how eukaryotic cells maintain a precise system of cell differentiation despite having enormous genomes. The regulatory organization of eukaryotic genomes is far more complex than in prokaryotes. This complexity led to the development of a model of eukaryotic gene regulation as proposed by Britten and Davidson (1969). They proposed that eukaryotic cells contained a complex and coordinated program of gene expression where different components at a locus interacted together in order to express a specific gene, this model included an integrator gene, a producer gene, a receptor site, and a sensor site (Britten & Davidson 1969). This model of gene regulation also explained why eukaryotic genomes contain large amounts of non-coding DNA sequences that have usually being considered as 'Junk' DNA (Ohno 1972). Then, King and Wilson observed that coding sequences between humans and chimpanzees were extremely similar despite the big morphological differences. They also observed that highly repetitive DNA from these two apes hybridized at a lower dissociation temperature than from human DNA alone, suggesting that most of the phenotypic differences between humans and chimpanzees were caused by changes in regulatory sequences rather of changes in coding sequences (King & Wilson 1975).

Recent advances in functional analysis of the genome have revolutionized the understanding of the relative importance of gene regulation in evolution. Researchers from the Encode project argued that nearly 80% of the DNA contribute to gene regulation and that most of the 'junk' DNA was actually regulatory (Bernstein et al. 2012). While some scientists have criticized this as a large overestimate (Graur et al. 2013), it is clear that non-coding DNA plays a critical role in gene regulation. Since the encode program was opened to the public, massive amounts of functional data associated with all animal genomes have been released. In addition to the technical and empiric advances, new statistical developments have also allowed the discovery of regions that are evolving slower or faster than the phylogenetic expectation (Pollard et al. 2006; Hubisz et al. 2011). In Encode, each genome can be compared with the human and mouse genomes, which contain extensive and detailed maps or tracks of epigenetic information and measurements of DNA conservation and acceleration that allow researchers to identify the regions were selection may be acting. Indeed, these tools have allowed several studies to look for evidence of selection among regions of the genome that are associated with open and regulatory chromatin. Some of these fast evolving regions in the human lineage have been coined the Human Accelerated Regions (HARs). These elements are associated with many of the morphological traits that differentiate humans from our close relatives, such as: precision grip, brain size expansion, and lactose intolerance (Tishkoff et al. 2007; Prabhakar et al. 2008; Boyd et al. 2015). Many other studies have been able to identify gene regulatory elements that explain adaptive evolution among natural populations. Regulatory changes are predominant in scans of selection among divergent populations of marine and lake three spine sticklebacks (*Gasterosteus acueatus*) (Jones et al. 2012). In fact, a recurrent deletion of an enhancer in the *Pitx1* gene explains the adaptive loss of pelvic spines in the stickleback fish (Chan et al. 2010). Despite all the morphological and physiological evidence claiming the relative importance of gene regulation in adaptive evolution, few studies have addressed the involvement of gene regulation in the adaptive evolution of animal behavior. Some gene modules that are known to regulate the development of the egg in insects have been linked to the regulation of labor division and foraging behavior in honeybees (Toth & Robinson, 2007). The gene *for,* which is known for controlling feeding behavior in the fruit-fly, has been found to regulate the transition from nurses to foraging honey bees. As long as honeybees age, the expression of *for* increases and the transition to forager occurs (Whitfield et al. 2003).

Animals rely on their perception of the environment, experience and social interactions to make adaptive decisions in order to increase their reproductive fitness (Dukas & Ratcliffe 2009). Prairie voles and its allies in the Old and New World have become an excellent model to understand the evolution of the social brain (McGraw & Young 2010). Prairie voles exhibit extraordinary variation in social behavior at both the interspecific and intraspecific level (Getz, McGuire, & Pizzuto, 1993). Prairie voles are socially monogamous rodents that form life-long pair-bonds but not necessarily mutually exclusive, exhibit bi-parental care, and extremely aggressive toward intruding conspecifics (Phelps & Ophir 2009). Interestingly, most of the neurobiological and physiological bases of these complex behaviors are becoming well understood. In fact, the regulation of the arginine-vasopressin receptor in neural circuits has been linked to the differences in behavior within and between species (Young and Hammock 2007). However, little is known about the adaptive value gene regulation at driving behavioral variation in both the genome-wide level and even at the *Avpr1a* locus. In my dissertation, I addressed my interests in making progress in the understanding of the role of gene regulation in adaptive evolution of complex behaviors. In chapter I, I studied the regulatory evolution of pair-bonding by identifying functional elements at the *Avpr1a* locus using ChIP-seq, and then tested for signatures of rapid evolution driven by positive selection at the phylogenetic level. In chapter II, I tested for diversifying or balancing selection maintaining high levels of genetic variation associated with social variation in sexual fidelity. Finally, in chapter III, I evaluated gene expression at the genome-wide level to gain insights on additional candidate genes and their regulatory elements.

CHAPTER 1

Becoming Monogamous: Rapid evolution of a *cis*-regulatory element in the *Avpr1a* locus

ABSTRACT

The evolution of regulatory DNA is thought to play a critical role in the adaptive diversification of complex phenotypes. Although a variety of morphological innovations have been tied to changes in regulatory sequences, we know little about how such changes influence behavior. We explored the evolution of the vasopressin 1a receptor (V1aR, encoded by *Avpr1a*) to ask whether there was evidence of regulatory adaptation among monogamous vole species. Our phylogenetic analysis suggest that the two monogamous species in our analysis are sister taxa compared to other voles within our sample, indicating that they share patterns of *Avpr1a* expression and pairbond formation by common descent. ChIP-seq targeting H3K27ac from the ventral pallidum revealed a pair of putative enhancers. One of these enhancers exhibited significant evidence of adaptation coinciding with the origins of monogamy. A second enhancer showed evidence of purifying selection across voles, and across mammals more generally. Lastly, transgenic mice expressing a reporter under the control of these prairie vole enhancers were able to drive expression in the ventral pallidum, but expression was sensitive to insertion site. The results highlight the tractability of combining functional genomics, evolutionary genetics and behavioral neuroscience to understand the evolution of complex behaviors.

INTRODUCTION

Few questions are more fundamental to the biological sciences than the relationship between genomic and phenotypic diversity. Among the many phenotypes biologists hope to understand, social behavior is among the most compelling and complex, including mating decisions, cultural learning, and parental care to name just a few (Lea & Ryan 2015; Aplin et al. 2014; Dulac et al. 2014). Such behaviors vary not only within and among species, they are also dynamic, changing over time and across contexts. One way to manage this complexity is to focus on the more stable phenotypes that govern behavioral decisions (Phelps 2010; Hamer 2002). For example, in the socially monogamous prairie vole (*Microtus ochrogaster*) neural expression of the vasopressin 1a receptor V1aR, encoded by *Avpr1a* , is a critical regulator of pairbond formation (Winslow et al. 1993). In the current manuscript, we examine the relationship between nucleotide variation and *Avpr1a* expression in the ventral pallidum, a region critical for reward in general (Smith et al. 2009) and for prairie-vole bonding in particular (Pitkow et al. 2001; Lim & Young 2004). By doing so, we explore how selection has shaped DNA sequences that contribute to species differences in gene expression and complex social behavior.

Gene expression is an interesting phenotype not only because it is so intimately tied to genome sequence, but also because changes in gene regulation are increasingly considered central to evolutionary innovation (King & Wilson 1975; Stern 2000; Wray 2007). Mutations that arise within regulatory regions can alter a gene's function in specific tissues, leaving function in other tissues unchanged; this reduces the negative pleiotropic consequences of mutations and presumably increases their capacity to contribute to adaptation (Stern 2000; Wray 2007). Indeed, many studies have demonstrated the importance of gene regulation in the evolution of form, including insect wing patterns (Gompel et al. 2005; Warren et al. 1994; Jeong et al. 2008), stickleback pelvic spines (Shapiro et al. 2004), and the human neocortex (Boyd et al. 2015). Despite this recent progress, few studies have examined how adaptive sequence evolution contributes to behavior.

The capacity for pair-bond formation is a complex phenotype governed by many neuromodulators, including dopamine (Young & Wang 2004), opioids (Resendez et al. 2016), estrogen receptors (Cushing 2016) and corticosteroids (DeVries et al. 1996; Blondel et al. 2016; Lim et al. 2006; Lim et al. 2007). But among these regulators, the vasopressin 1a receptor (V1aR) is particularly well studied. Prairie voles have unusually high levels of V1aR in the ventral pallidum, a region critical to reward and addiction (Insel et al. 1994; Young & Wang 2004). Site-specific injections of V1aR antagonists or *Avpr1a* shRNA vectors reveal that pallidal V1aR is necessary for pair-bond formation but not for normal mating (Lim et al. 2004; Barrett et al. 2013). Species comparisons reveal that another monogamous species, the pine vole (*M. pinetorum*) has a similar elevation in pallidal V1aR abundance, while promiscuous congeners and many other promiscuous rodents lack this expression (Insel et al. 1994). Indeed, over-expressing the receptor in the pallidum of the promiscuous meadow vole produces key aspects of pair-bonding (Lim et al. 2004). Understanding how *Avpr1a* expression came to be elevated in the ventral pallidum of prairie and pine voles offers insights into the origins of a complex social behavior.

There are three receptors for the neuropeptide vasopressin, and V1aR is the predominant form in the brain (Caldwell et al. 2008). V1aR is a G-protein coupled receptor encoded by *Avpr1a*, a gene with two exons separated by a ~2.5kb intron. Initial work on individual and species differences in *Avpr1a* regulation focused on length differences in a complex microsatellite flanking the gene's transcription start site (Hammock et al. 2005; Ophir, Wolff, et al. 2008). The two monogamous species studied both share a long microsatellite, while two promiscuous species, meadow and montane voles, have very short microsatellite lengths. However, phylogenetic data indicate that the two promiscuous species examined are sister taxa, and that the microsatellite has been lost in their lineage but not in other promiscuous voles (Fink et al. 2006). Indeed, transgenic studies demonstrate that the microsatellite sequence is not responsible for pallidal V1aR expression (Donaldson & Young 2013). These findings highlight the need for a more systematic examination of *Avpr1a* regulatory evolution.

In the current study, we ask whether there has been adaptive regulatory evolution at the *Avpr1a* locus associated with origins of monogamy in voles. We begin by selecting 6 microtine species and a non-microtine outgroup species, each with a well characterized mating system. We use the sequences of putatively neutral nuclear genes to generate a well resolved phylogeny of this group. Next we sequence ~8 kb of the *Avpr1a* locus, and use ChIP-seq targeting a marker of active enhancers, acetylation of lysine 27 in histone 3 (H3K27ac), to identify regulatory regions at the locus. We test whether identified enhancers exhibit evidence of adaptation by looking for accelerated sequence evolution associated with the origins of monogamy, and by examining patterns of nucleotide diversity within prairie voles. Together these data offer novel insights into the adaptive regulation of gene expression critical to a complex behavior.

MATERIALS AND METHODS

Animal sampling and DNA extractions

We sampled livers from 6 species of *Microtus* voles (*Microtus arvalis*, *M. richardsoni*, *M. montanus*, *M. pennsylvanicus*, *M. pinetorum* and *M. ochrogaster*) and one outgroup*, Myodes gapperi*. With the exception of *M. ochrogaster*, all samples were kindly provided by The Washington Burke Museum in alcohol preparations. To assess intraspecific variation in prairie voles, 32 individuals collected in the vicinity of Urbana, IL (Champaign County) were also included in these analyses (Okhovat et al. 2015). All genomic DNA extractions were prepared using the QIAGEN DNEasy kit for tissue and blood following the manufacturer's protocol.

Polymerase chain reactions

We amplified three putatively neutral loci (*Lcat,* 920 bp; *Acp5,* 375 bp; and *Fgb,* 725 bp) and 8 kb of the *Avpr1a* locus using the primers listed in Table 1.1. All polymerase reactions (20 µL) were prepared with 1X GoTaq® Hot Start Polymerase buffer (Promega), 1.0 mM of MgCl2, 0.2 mm of each dNTP, 0.5 µm of each primer, 1.25 U HotStart Taq Polymerase and 1.0 µL of diluted DNA. Amplifications were carried out on a Veriti Thermal Cycler (Applied Biosystems) using the following profiles: *Lcat* – initial denaturation at 95 °C for 2 min, 35 cycles of 95 °C for 40 s, annealing at 57 °C for 40 s, extension at 72 °C for 90 s, and a final extension at 72 °C for 7 min. *Acp5* – initial denaturation at 95 °C for 2 min, 35 cycles of 95 °C for 30 s, annealing at 58 °C for 30 s, extension at 72 °C for 60 s, and a final extension at 72 °C for 7 min. *Fgb* – initial denaturation at 95 °C for 2 min, 35 cycles of 94 °C for 40 s, annealing at 53.5 °C for 40 s, extension at 72 °C for 60 s, and a final extension at 72 °C for 7 min. *Avpr1a* – to avoid amplification of a pseudogene present in the prairie voles, and to improve our ability to amplify non-coding sequences across species, we used a semi-nested PCR to target the *Avpr1a* locus. The first amplification consisted of an initial denaturation at 95 °C for 2 min, 35 cycles of 95 °C for 40 s, annealing at 58 °C for 30 s and extension at 72 °C for 240 s, and a final extension at 72 °C for 7 min; 1ul of the resulting reaction was taken as template and amplified using the same forward primer but a second reverse primer, with PCRs conditions including initial denaturation at 95 °C for 2 min; 35 cycles of 95 °C for 40 s, annealing at 58 °C for 30 s, extension at 72 °C for 210 s, and a final extension at 72 °C for 7 min. Each amplicon was first visualized on 1% agarose gels in order to check its band size and specificity, and then cleaned with Qiaquick PCR purification kits (Qiagen) according to the manufacturer's protocol. Sanger sequencing was conducted at the University of Texas at Austin Institute for Cellular and Molecular Biology (ICMB). The resulting ABI Chromatograms were processed and analyzed using 'Map to Reference' parameters in Geneious v6.1 (Biomatters).

Phylogenetic reconstruction

We first assembled a concatenated "neutral sequence" by combining the putatively neutral loci from each species. To align the sequences, we used MAFFT implemented in the Pairwise/Multiple Align tool of Geneious V6.1 using default settings, followed by manual curation to resolve indels or mismatched substitutions. After removing conserved coding sequences, the neutral alignment contained 1279 bp for each species. MrModeltest v2.3 identified "K80", "GTR + Γ", "JC69", and "HKY85 + I" as the best-fit models of nucleotide substitution among 16 model tests (Nylander 2004). All these models provided the same tree topologies and similar branch lengths, suggesting the resulting phylogeny was not sensitive to a specific model. We selected the GTR using a gamma-distribution to infer a tree of three putatively neutral non-coding sequences. MrBayes was run for 2000000 generations with subsampling every 100th generation, discarding 30% of the first generated burn-in trees, and chain temperature was set to 0.2.

Chromatin immunoprecipitation

To identify putative enhancer sites active in the ventral pallidum of prairie voles, we performed ChIP-seq targeting a marker for active enhancers, H3K27ac. Nine lab-reared males from our breeding colony at the University of Texas at Austin were euthanized and their brains were immediately harvested, blocked using a brain slicer matrix, and a 1 mm coronal section was extracted and placed in 1.4% paraformaldehyde. While the section was fixing, we collected four punches from the ventral pallidum of each brain and placed them in a microfuge tube with 1.4% paraformaldehyde at room temperature (RT). The cross-linking reaction was quenched after 15min total by adding glycine (2M). Samples were then washed three times with phosphate-borate solution (PBS) containing proteinase inhibitors (PI and PMSF). The tissue punches were homogenized using disposable grinder pestles and rewashed with PBS supplemented with PI and PMSF. Cell membranes were lysed in buffer (5mM PIPES pH8.0, 85mM KCl, 0.5% NP40) supplemented with proteinase inhibitors (PI). Nuclear membranes were further lysed in buffer (50 mM Tris-HCl pH8.0, 10 mM EDTA, 1% SDS, PI) on ice for 10 min. Chromatin was sonicated on ice with 6 pulses of 10s at 80% of power to generate fragments of a size range of 150-400bp. To pull down chromatin, 150 μL of sonicated solutions were precleared with Dynabeads and a 250μL of dilution buffer (0.01 %SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-Cl pH=8, 150mM NaCl) for 2 hours at 4°C. 15uL were aliquoted to use as control input DNA. The remaining chromatin was incubated with H3K27ac antibody (Abcam, ab4729) with overnight rotation at 4°C. After antibody binding, the magnetic beads were passed thorough sequential washes with fresh RIPA buffer, high-salt buffer, LiCl buffer and two final washes on 1X TE buffer. Chromatin was isolated from beads in fresh elution buffer with a 15 min incubation at 65°C. The chromatin precipitate was incubated with RNAse-A for 4h at 65C to degrade traces of RNA, and then incubated with Proteinase K to degrade protein. Lastly, DNA was purified with a standard phenol-chloroform extraction.

Library preparation and sequencing

DNA from H3K27ac-ChIP and INPUT were combined in pools of 3 individuals with similar concentrations, with 3 pools corresponding to 9 individuals. Libraries were prepared using the KAPA Library Prep Kit following the manufacturer's instructions but excluding size-selection. Briefly, DNA was end-repaired, tailed with adenines and ligated to different NEXTflex DNA barcodes (Bio Scientific). Barcoded libraries were PCR-amplified for 10 cycles. The quality control and fragment distribution was examined before sequencing using the Agilent Bioanalyzer by University of Texas at Austin Genome Sequencing and Analysis Facility (GSAF). All samples were sequenced in the Illumina NextSeq platform at the GSAF facility (>30 million reads (PE75) per pooled sample). The quality of the reads was examined by visualizing the FastQC output for each sample. Experimental and input reads were aligned to the prairie vole draft genome assembly using bwa with settings for 75bp paired-end reads. We used SAMtools to estimate mapping efficiency (Li et al. 2009). A local duplication kept the extended region around the *Avpr1a* locus from assembly into the prairie vole genome (http://www.broadinstitute.org/software/allpaths-lg/blog/?p=618). Therefore, BAC contigs containing the locus and its pseudogene were manually added to the assembly (NCBI accessions: DP001225, HQ156469). To call for significant peaks of H3K27ac, we used Model-based Analysis of ChIP-Seq, MACS2 (Zhang et al. 2008), software that identified peaks of H3K27ac-DNA interactions with a Qvalue-cutoff of 0.05. Fold enrichments were plotted using ggPlot in R. Additional details of the pipeline used can be found at the author's GitHub repositories site (https://github.com/wodanaz/ChIPseq).

Phylogenetic tests of selection

Our phylogenetic data suggest that monogamy evolved once at the common ancestor of prairie and pine voles. We noted that there were only six synapomorphies at this short node: five within a single enhancer, and the sixth at an adjacent H3K27ac peak. To test for the probability of observing this clustering of substitutions by chance, we estimated the probability of six substitutions occurring within a window of 563 bp or less. To do this, we ran 10,000 simulations in which we randomly assigned six substitutions to a sequence of length 5874bp (the length of our non-coding sequence) and calculated the distance between the most distant substitutions.

To test for signatures of selection more explicitly, we used a likelihood ratio test to examine whether ventral pallidum enhancers identified by H3K27ac ChIP-seq were evolving more quickly at the onset of monogamy than in the rest of the phylogeny. Based on our phylogeny, we defined the "foreground" as the branch that corresponds to the origin of monogamy in the common ancestor of pine and prairie voles (dark red in figure 1.3); branches after the split of prairie and pine vole lineages were excluded; all other branches were designated "background" Using a Poisson distribution, we calculated the likelihood of observing a given number of substitutions in the foreground given the maximum likelihood estimate of the rate in the background. We compared this to the likelihood of observing the same number of mutations in the foreground given a model in which the foreground branch was evolving at a different rate. We compared these two models using a likelihood ratio test, where the ratio was assumed to follow a chi-squared distribution with one degree of freedom.

We compared rates of foreground evolution between the putative pallidal enhancers and other non-coding sequences using a similar strategy. Specifically, we asked whether changes in the foreground were significantly faster within an enhancer than in other non-coding sequences. These models were also compared using a chi-squared test with one degree of freedom.

To assess whether putative enhancers had evolved recently or were likely to be ancient, we examined the extent of conservation at the locus across various taxonomic scales using PHylogenetic Analyses with Space/Time, PHAST v1.3 (Hubisz et al. 2011) First, we downloaded mammalian and glire homologues of the sequences for *Lcat, Acp5*, *Fgb* and *Avpr1a* (Table 1.2) using the genome Blast/Blat tools implemented in Ensembl (Yates et al. 2016). Next, these sequences were aligned using the MAFFT algorithm. Sequences were aligned respect to a prairie vole reference, and sequences not present in the prairie vole reference were deleted. *PhyloFit* generated neutral models of evolution for the mammalian, glire and vole clades by fitting the neutral topologies to their respective sequence alignments. To estimate conservation scores and identify conserved elements we used *PhastCons*, fitting each of the neutral models to their respective *Avpr1a* alignments using the General Time-Reversible (GTR) substitution model. We plotted the posterior probability scores for each clade using the library Gviz implemented in R (Hahne & Ivanek 2016).

Population tests of selection

To assess whether positive selection leaves signatures of more recent selection, we used a modification of the *McDonald-Kreitman (MK)* test that has been adapted to non-coding DNA (Bustamante et al. 2002). It compares the ratio of polymorphism to divergence between two species at two types of genetic regions, one of which is putatively neutral. This method assumes that, for neutrally evolving sequences, polymorphism and divergence are proportional and dependent on mutation rate. We estimated polymorphism of 32 individuals from a population of prairie voles from Champaign County, IL, and divergence was estimated by comparing *Avpr1a* sequences from prairie voles and water voles, *M. richardsoni* (Okhovat et al. 2015). First, we used a Fisher's exact test to compare the ratio of fixed differences to polymorphisms in the vicinity of the pallidal peak 1 or pallidal peak 2 to the ratio observed in our putatively neutral loci.

To further characterize localized patterns of selection, we performed a sliding window analysis in which we calculated levels of polymorphism and divergence along the *Avpr1a* sequence in 300 bp windows, comparing the observations in each window to our null expectation based on our three neutral markers. Because these tests are descriptive, we did not correct for multiple testing. Scores for Divergence (K), nucleotide diversity (π) and MK p-values were plotted respect to the first base of each window using the library Gviz implemented in R (Hahne & Ivanek 2016), excluding microsatellite sequences.

To further characterize the null expectations for diversity and divergence based on our neutral data, we used the within and between species alignments in our neutral markers to generate a neutral model of sequence evolution, and simulated a neutral alignment of 100 kb using the program *base_evolve* implemented in PHAST. For each 300 bp window in this simulated sequence, we used DNAsp to calculate the nucleotide diversity (π) and divergence (K). We used these measures to generate a bivariate kernel density estimate of the probability of observing particularly combinations of π and K in our neutral data using the function kde2d implemented in the R package MASS

(Venables & Ripley 2002). We plotted our observed levels of π and K from each 300 bp window in our *Avpr1a* data over the kernel density estimate, color coding each window based on functional features of the gene.

Transgenic mouse reporter assay

To determine whether enhancer sequences that showed evidence of selection were capable of driving expression in the ventral pallidum, we synthesized the region spanning the H3K27ac ChIPseq peaks (875bps, IDT DNA technologies, Coralville, IA). We first cloned this putative enhancer into a Gateway entry vector (pENTR/D-TOPO; Invitrogen, Carlsbad, CA). We then subcloned the putative enhancer into an hsp68-*lacZ* Gateway vector (Addgene plasmid# 37843, kindly delivered by Pennacchio lab) via RS Clonase (Invitrogen, Carlsbad, CA). An individual colony was isolated and cultured, and the final plasmid was isolated using the ZymoPURE™ Plasmid Maxiprep Kit (Zymo, Irvine, CA). We confirmed the identity and sequence of the construct by Sanger sequencing. Next, we digested the vector with Sal1 (NEB, Ipswich, MA) and submitted ~20ug of the transgene to the Mouse Genetic Engineering Facility of The University of Texas at Austin for pronuclear injection.

The transgene fragment was isolated from a 0.8% agarose gel using a silica matrix method (QIAEX II kit). The DNA was resuspended in 10 mM Tris-HCl (pH 7.5) and 0.1 mM EDTA (Injection buffer) at 100 ng/ul. A DNA aliquot was diluted to 1 ng/ul in injection buffer for microinjection (Behringer et al. 2014). To generate embryos for pronuclear injection, we crossed C57BL/6J (JAX) males and B6D2F1/J (JAX) females. Injected embryos were surgically transferred to CD-1 (Charles River) recipient female mice. Seventy two independent lines were generated, of which fourteen were positive for the *lacZ* transgene. These 14 brains were fresh frozen on dry ice and stored at -80C. We sliced twelve *lacZ* positive brains in sections of 20µm thick using a cryostat Microm HM550 cryostat (ThermoFisher Scientific, Walldorf, Germany) and stained for *lacZ* using Xgal and counterstained with neutral red. All animal work was approved by the University of Texas Animal care and use Committee.

RESULTS

Phylogenetic history of New World voles

Despite our limited taxonomic sample, we found that our phylogenetic tree was better resolved than many published trees, and had high levels of support overall. The tree recovered a novel sister-group relationship (Bayesian Posterior Probability, BPP=1.00) between monogamous pine and prairie voles (Fig 1.1A). Similarly, the data strongly indicate that *M. pennsylvanicus* and *M. montanus* are sister taxa (BPP=1.00). These two groups also share a common ancestor (BPP= 0.89) that diverged from *M. richardsoni* (BPP=0.83). We also found strong support for the hypothesis that our New World voles form a monophyletic group respect to the European *M. arvalis* (BPP= 1.00). Lastly, the monophyly of Microtus with respect to Myodes was strongly supported (BPP=1.00). Bayesian trees generated with other nucleotide substitution models had similar topologies and branch lengths.

Functional characterization of the *Avpr1a* locus

To identify putative regulatory elements we performed ChIP-seq targeting H3K27ac from ventral pallidum punches. We sequenced 31,292,819,850 bp from a total of 9 individuals from three pooled batches with high quality scores (Phred>33) across all bases, obtaining a total of 404749832 uniquely mapped reads (75%-91%) for both input and H3K27ac libraries, however, only one ChIP-seq library showed low mapping efficiency (32%). We identified 73121 peaks across the genome (Q-value<0.05), including the promoters of genes known to exhibit high expression in the ventral pallidum, such as *Slc34a3* and *Syde2*. We observed 28 peaks within 1Mb of *Avpr1a*, and 2 peaks within 100kb of *Avpr1a* (GEO accession #). We found that these two peaks of H3K27ac were significant in the 5' region (peak 1: *P*= 1x10^{-12}, and peak 2: *P*= 1x10^{-9}), located around 2000bp from the translation start site of the *Avpr1a* locus and flanking an *Avpr1a* microsatellite (Fig 1.2). The summit of 5' peak occurs at 2022 bp 5' of the reported transcription start site (Young et al. 1999), and 2255 bp 5' of the translation start site. This first pallidal peak is novel and spans 340bp of sequence. The second peak is 437bp in length, and its summit is located at 1333 bp 5' of the reported transcription start site and 1566 bp 5' of the translation start site. This peak coincides with a mouse brain DNase I Hypersensitivity Site from ENCODE and a highly conserved DNA element across mammals. The closest peak outside the *Avpr1a* locus was found 62kb downstream the translation start site. Moreover, we found no significant peaks in the vicinity of the duplicated *Avpr1a* pseudogene, actually, the closest peak was located 25Kb downstream the pseudogene. We used these two peaks as *a priori* boundaries for subsequent tests of selection.

Phylogenetic signals of selection

We observed six shared-derived substitutions clustered within 563 bp of the *avp1a* locus. Five of these substitutions are located inside a 340 bp element characterized by H3K27ac enrichment in the ventral pallidum (a putative pallidal enhancer). The sixth substitution is placed within the second pallidal H3K27ac peak (437 bp). A Fisher's exact test revealed that probability of observing five synapomorphies in the first 340bp pallidal peak a significantly higher rate of substitution than observed in the remainder of the noncoding sequence (5:52, vs 1:625, Fisher's exact P<0.0001). The rate of change within the second peak was not significantly different than in the remainder of the non-coding sequence (1:65, vs 0:565, Fisher's exact P=0.1054). Simulation results demonstrate that these 6 substitutions cluster over a significantly shorter span than expected by chance (P= 0.0001).

We next compared rates of evolution at each putative pallidal enhancer to those in the background. A likelihood ratio test reveals that the first peak of H3K27ac exhibits evidence of acceleration during the evolution of monogamy, at the common ancestor of prairie and pine voles ($\lambda_{\text{Background}}$=0.74, $\lambda_{\text{Foreground}}$ 4.90, LR= 187.84, P= 9.4x10^{-43}). In contrast, we found no significant evidence of acceleration at the second peak of ChIP-seq ($\lambda_{\text{Background}}$=0.84, $\lambda_{\text{Foreground}}$ 0, LR= 1.02, P= 0.31) (Fig 1.3B). This last region coincides with a DNase I Hypersensitivity Site from the mouse brain (ENCODE) and a highly conserved DNA element in mammals (Fig 1.2A).

Population signals of selection

We explored whether positive selection can be detected in the *Avpr1a* locus within a population of prairie voles that were sequenced from Champaign County, IL (Okhovat et al. 2015)**.** Our McDonald-Kreitman test revealed that pallidal peak 1 exhibited a significant increment of divergence vs diversity in contrast to putatively neutral loci (22:6

vs 30:30, *P*=0.019, Fig 1.4A-B, orange dots), a signal of positive selection. We did not obtain an equivalent pattern from pallidal peak 2 (13:12 vs 30:30, *P*=1.0, depicted in yellow dots). The sliding window analysis revealed two additional peak in which divergence was high but diversity was low. One within the transcription and translation start sites (5'UTR) were polymorphism may be expected to be low and divergence high (Andolfatto 2005; Haddrill et al. 2008). The next peak does not correspond to an enhancer observed in the ventral pallidum, or previous reported in the retrosplenial cortex (Okhovat et al. 2015). This element contains no fixed foreground substitutions (Fig 1.4A, violet dots). Lastly, our sliding window analysis suggests a region of the *Avpr1a* locus with high levels of diversity (4B dark green dots).

Causal transgenic manipulations

A total of 189 injected mice embryos were transferred to foster mothers, of which 72 independent lines were born. We sampled tail tissue to verify presence of the transgene using the method reported by Kobayashi and collaborators (2004) at weaning. We examined 12 transgenic mice that resulted positive for the transgene. X-gal blue staining was used to determine the number of individuals that showed positive expression in the ventral pallidum. Of 12 examined lines, 10 independent lines showed expression of *lacZ* in the brain, and 5 showed positive expression in the ventral pallidum (Fig 1.5).

DISCUSSION

Comparisons of monogamous and non-monogamous vole species have been a central model in the investigation of species differences in genes, brains and behavior. The simple two-species comparisons have come under criticism by evolutionary biologists for failing to account for phylogeny (Fink et al. 2006). These concerns are reinforced by our current findings, which provide strong support for a rapid origin of monogamy at the common ancestor of pine and prairie voles. This topology was initially surprising, but a closer examination of published phylogenies revealed that the placement of these monogamous voles were generally poorly resolved (Fink et al. 2006), including a recent study using multiple nuclear markers and a larger taxonomic sampling (Martínková and Moravec 2012). Thus our results are not in conflict with prior phylogenetic studies. Our result is also consistent with classical taxonomic nomenclature that put pine and prairie voles together in the genus *Pitymys* on the basis of a shared tri-cuspid morphology of the first molar (Tamarin 1985).

These results have significant implications for our understanding of monogamy and its evolution in this model clade. For example, recent molecular phylogenetics work suggests that there are many species excluded from our analysis that are more closely related to pine voles than are prairie voles (Robovský et al. 2008; Martínková & Moravec 2012). Taken together, these data imply that monogamy is significantly more common in *Microtus* than has been previously appreciated. The *Pitymys* group has been in North America since the Pleistocene, and today many of its habitats include sandy soil grasslands with cool temperatures and low productivity, including Midwest prairies, and highland pine habitats in Mexico (Tamarin 1985; Harris 1988; Escalante et al. 2004); the natural history of this group suggests that monogamy may have arisen as a response to scarce resources and low densities (Emlen & Oring 1977) – a finding that would be consistent with ecological correlates of mammalian monogamy (Lukas & Clutton-Brock

2013). With respect to the evolution of *Avpr1a* expression more specifically, our results indicate that the many similarities between pine and prairie voles in the neural distributions of neuropeptide receptors (Insel et al. 1994; Young & Hammock 2007) reflect shared descent rather than adaptive convergence.

Functional genomic approaches paired with bioinformatics and sequence analysis are powerful methods to characterize gene regulation in the brain (Visel et al. 2009; Konopka & Geschwind 2010; Landt et al. 2012; Harris & Hofmann 2014; Maze et al. 2014; Okhovat et al. 2015). We used ChIP-seq targeting a well characterized marker of active enhancers, H3K27ac (Creyghton et al. 2010), to identify pallidal regulatory regions in the vicinity of the *Avpr1a* locus. We found two distinct peaks that flanked a 5' microsatellite sequence. Remarkably, we found that all of the substitutions that coincide with the origin of monogamy occur within the two regulatory sequences defined by our ventral pallidum ChIP-seq data. Indeed, a simple Fisher's exact test revealed that this clustering was much stronger than expected by chance ($P<0.0001$), a result confirmed by a more rigorous simulation that takes into account the many ways a cluster could occur across the locus ($P=0.0001$).

To our surprise, we found that the foreground substitutions were almost entirely clustered in the novel 5' enhancer element we refer to "pallidal peak 1," suggesting this sequence may have played a unique role in the derived expression patterns that characterize prairie and pine voles, including enhanced ventral pallidum expression. A likelihood ratio test revealed that this putative pallidal enhancer evolved significantly more rapidly at the origin of monogamy than it did in the rest of the phylogeny (*P*=

9.4×10^{-43}). Similarly, pallidal peak 1 showed a reduced ratio of polymorphism to divergence compared to the rest of the locus (P<0.0001) or compared to neutral markers (P=0.019). Together these data suggest that sequence changes in this putative enhancer contribute to the adaptive evolution of V1aR expression patterns and the origins of monogamy.

In contrast to the accelerated and novel evolution of apparent enhancer function in pallidal peak 1, the putative enhancer element we refer to as "pallidal peak 2" does not differ between foreground and background (LRT, P=0.31). Moreover, this element shows low rates of evolution compared to the rest of the non-coding sequence, suggesting that the element has been subject to purifying selection across mammals. Similarly, this conserved peak contains a DNAse hypersensitive site evident in *Mus* brain, and a highly conserved binding site for the tethering insulator protein known as CTCF (Ong & Corces 2014). Interestingly, this generally conserved region also contained an insertion unique to Cricetidae, which is highly conserved across the vole species we examined. One possible explanation is that open chromatin at this region has enabled the insertion of a transposable element (Gangadharan et al. 2010). Given that this sequence element is highly conserved across voles and seems to be present in other cricetids such as the deer mice, we think that this insertion may play a role in tuning the regulation of the *Avpr1a* in the rodent brain — a finding that would be consistent with a role for transposable elements in regulatory innovations in other contexts (Morgan et al. 1999; Sun et al. 2004; Iida et al. 2004).

Because the two putative enhancers flanked one another and contained all the substitutions coincident with the increases in pallidal V1aR, we decided to test whether this sequence was able to drive expression of a *lacZ* reporter in the ventral pallidum of a lab mouse. We examined *lacZ* expression in 12 lines positive for the *lacZ* transgene. We found that in 10 lines, the prairie vole sequence was able to drive neural expression in the brain, and in 5 lines, the prairie vole sequence was able to drive *lacZ* expression in the ventral pallidum. These data demonstrate that the putative enhancers are indeed capable of driving expression in the ventral pallidum. However, they also demonstrate that this sequence is not sufficient to reliably drive pallidal V1aR expression on its own. This is consistent with the existence of other H3K27ac peaks outside the focal 8 kb of our study. It is also worth noting that our transgenic construct included a highly conserved CTCF binding site. A variety of recent studies implicate CTCF in the formation of chromatin loops, with the orientation of loops resulting from the orientation of pairs of CTCF binding sites (Splinter 2006; Holwerda & de Laat 2013; Oti et al. 2016). Our transgenic approach does not control either for the insertion site or its orientation; such variation in the chromatin context seems important to the function of the targeted sequence. Recent advances in the use of conformation capture methods (4C, Hi-C, ChIA-PET; Rusk 2009; de Wit and de Laat 2012; Vietri Rudan et al. 2015; Mifsud et al. 2015) and genome editing (Esvelt & Wang 2013; Makarova et al. 2015; Graham & Root 2015) suggest it should be possible to tease apart the specific roles of additional enhancers and the more general contributions of chromatin context, though this is beyond the scope of the current study.

Our results highlight the power of integrating neuroscience, functional genomics and evolutionary analysis to explore the adaptive evolution of gene expression and complex behavior. Our phylogenetic analyses suggest a common origin of monogamy in prairie and pine voles, a result that provides new insights into the distribution and origins of pair-bonding in this clade. In contrast to work on the *Avpr1a* coding sequence (Fink et al. 2007; Turner et al. 2010), or on repetitive microsatellite sequences (Fink et al. 2006; Turner & Hoekstra 2008), multiple lines of evidence indicate selection has shaped the function of a pallidal enhancer associated with both the origin of monogamy and with the emergence of the pallidal *avpr1a* expression critical to bond formation. Our transgenic data reveal that this enhancer sequence is indeed able to drive expression in the ventral pallidum, but that it is not sufficient. Although our work focuses on a single gene, we believe this integrative approach generalizes readily across the genome, and provides a model for how to approach adaptive diversity in DNA sequence, neuronal function, and complex behaviors.

TABLES

Amplicon	Size (kb)	Primer sequences	Outer primer sequences (if nested)
5’ non-coding seq.	3.4	F:TGTGGCACCCAGGTAAATGC R:GTAGCAGATGAAGCCATAGCAG	F:GCATGTGATTCTGGAATTTGTAAC R:ATAGTCTTCACGCTGCTGACA
Promoter + 5’ UTR	1.7	F:AATAGACCAACGTTCTTAAG R:GCTCCTCGTTGCGTACATC	*Not nested*
First exon	1.2	F:CGGAAGCGGGAAGGAAGCAGCC R:CTCCCTCAGCCCATGATGCAG	F:GYGGTAGCCTAAACGCAGA R:GTTGGGATGRTTGAGAACCACA
Intron	2.5	F:CTACATCCTCTGCTGGGCTCC R:CATGTATATCCAGGGGTTGC	F:GCCTTGTGTCAGCAGCGTG R:TGTCTGTAGGCACCTTCTGTTCTG
Second exon	1.0	F:GCTGCTCTAACAGTGGTTGGTTTG R:CACATCACATGACTTAAACCAATC	F:GCCTTGTGTCAGCAGCGTG R:TGTCTGTAGGCACCTTCTGTTCTG
3’ UTR	0.6	F:CTACATCCTCTGCTGGGCTCC R:CATGTATATCCAGGGGTTGC	F:GCCTTGTGTCAGCAGCGTG R:TGTCTGTAGGCACCTTCTGTTCTG
3’ flanking	0.6	F:CGGACCATATAGAGATCATAAGAG R: GGGATAGAGGCAGAGACCCA	F: GTCCATTGTCTAAATCCGGACC R: GAACATGAGCAAAGAAGTCGG
Lcat	0.7	F: AGAGGACTTCTTCACCATCTGGCT R:TGTGCCCAATAAGGAAGACAGGCT	*Not nested*
Fgb	0.7	F: GGCAATGATAAGATTAGCCAGCCAGCTCAC R: AACGGCCACCCCAGTAGTATCTG	*Not nested*
acp5	0.5	F: AATGCCCCATTCCACACAGC R: GCAGAGACGTTGCCAAGGTG	*Not nested*

Table 1.1. PCR and sequencing primers (5’ to 3’) for characterizing *Avpr1a* , *Lcat*, *Fgb and acp5* sequence variation across species.

	ENSEMBL Genomic Range			
Species (Assembly)	***Avpr1a***	***Lcat***	***Acp5***	***Fgb***
Homo sapiens (GRCh38)	chr12:63538078-63546832	chr16:67939347-67944491	chr19:11576366-11577280	chr4:154569738-154570400
Pan troglodytes (CHIMP2.1.4)	chr12:26138709-26148033	chr16:67228023-67230496	chr19:11806099-11807062	chr4:157647751-157648413
Macaca mulatta (MMUL_1)	chr11:60382040-60390796	chr20:66278551-66279616	chr19:11410159-11411141	chr5:146770867-146771529
Canis familiaris (CanFam3.1)	chr10:9250685-9259553	chr5:81555683-81558333	chr20:49814430-49815939	chr15:52227070-52227361
Felis catus (Felis_catus_6.2)	chrB4:103421857-103430787	chrE2:45568720-45570933	chrA2:8521753-8523087	chrB1:73945557-73946034
Loxodonta Africana (loxAfr3)	scaffold_2:52877434-52886163	scaffold_48:9349290-9351588	scaffold_26:32239254-32240395	scaffold_51:1110285-1110878
Mus musculus(GRCm38)	chr10:121883859-121892319	chr8:105941286-105943677	chr9:22129385-22130147	chr3:83042653-83043291
Rattus norvegicus (Rnor_6.0)	chr7:67525681-67534134	chr19:37912929-37917224	chr8:23145633-23146123	chr2:182028487-182028589
Cavia porcellus (cavPor3)	scaffold_9: 35040551-35050845	scaffold_22:8726283-8730168	scaffold_226:685226-687548	scaffold_7:31718963-31718852
Ictidomys tridecemlineatus (spetri2)	JH393355.1: 1786023-1800018	JH393285.1:18293836-18297463	JH393469.1:1598276-1600836	JH393386.1:1312376-1313272
Oryctolagus cuniculus (OryCun2.0)	chr4: 42141848-42152606	chr5:23503048-23506887	AAGW02082541:501-750	chr15:10901796-10903300

Table 1.2. Genome locations of sequences downloaded from Ensembl for Mammals and Glires.

FIGURES

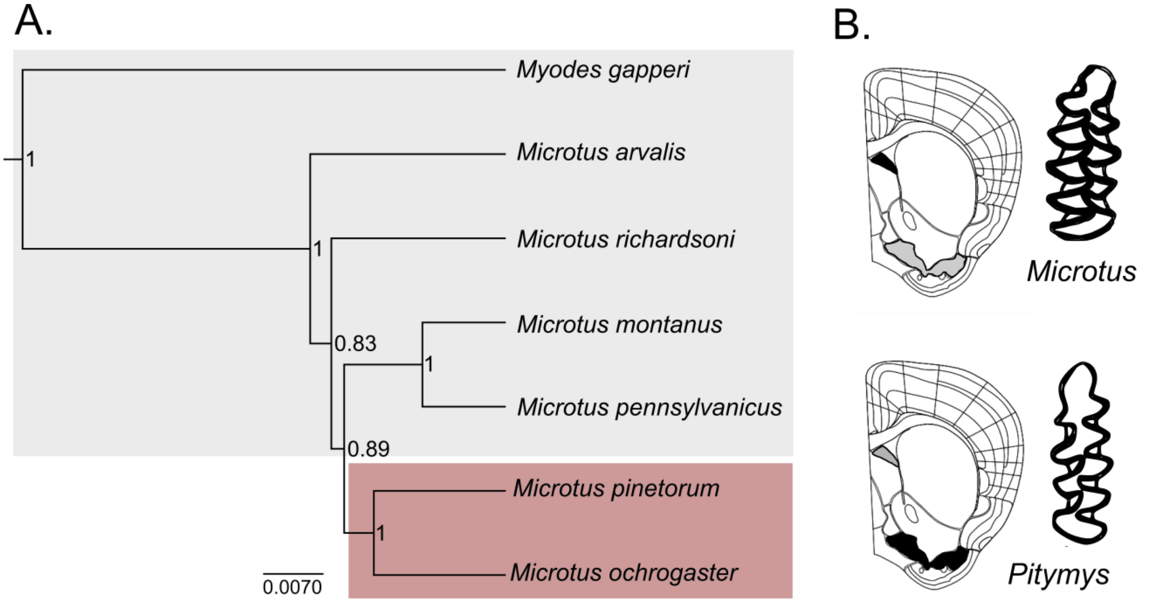


Figure 1.1. Monogamy evolved once in New World Voles. A) Bayesian tree was obtained from three neutral genes: *Acp5*, *LCAT*, and *FGB* (nodes include the posterior probability). Voles from the historic *Pitymys* group are highlighted in dark red and voles from the *Microtus* group are highlighted in grey. B) Pine and prairie voles (*M. pinetorum, M. ochrogaster*) in the historic group Pitymys share derived traits of elevated pallidal V1aR abundance and molar morphology.

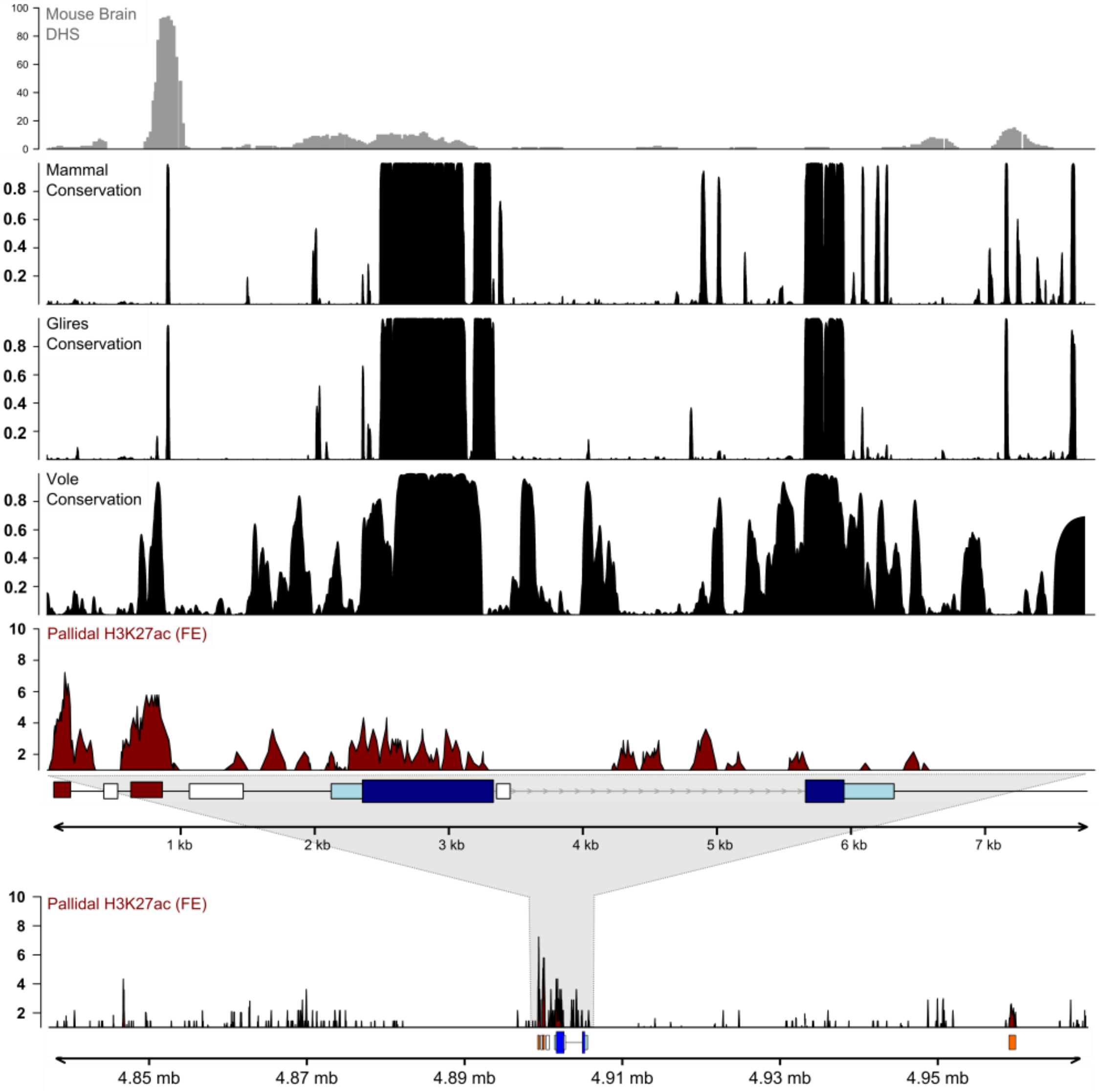


Figure 1.2: Functional characterization of *Avpr1a*. First track (grey) reveals DNAse hypersensitivity (DHS) for mouse whole brain from ENCODE; the x-axis depicts position along the *Avpr1a* locus (bottom panel), the y-axis depicts DNAse I sensitivity as a continuous function using sequencing tag density The next 3 tracks (black) show conservation for mammals including voles, the group glires (rodents and lagomorphs), and new world voles alone. The y-axis corresponds to phastCons conservation score

across the selected species, this score depicts the posterior probability a phastCons's phylogenetic hidden Markov model (HMM) is in its most-conserved state at that nucleotide position (Siepel et al. 2005). Sequences of the species used are provided in Table 1.2; mammal tree included: elephant, cat, dog, rhesus, human, chimp, rat, mouse, and common, water, meadow, pine and prairie voles; glire tree included: rabbit, squirrel, guinea pig, rat, mouse, and common, water, meadow, pine and prairie voles; and the Arvicolini tree included southern red-backed, common, water, meadow, montane, pine and prairie voles. The bottom track (dark red) depicts *Avpr1a* sequences obtained by performing ChIP-seq targeting H3K27ac, a marker of active enhancers, from the ventral pallidum of prairie voles. The y-axis depicts fold-enrichment (FE) of ChIP-seq results compared to input DNA.

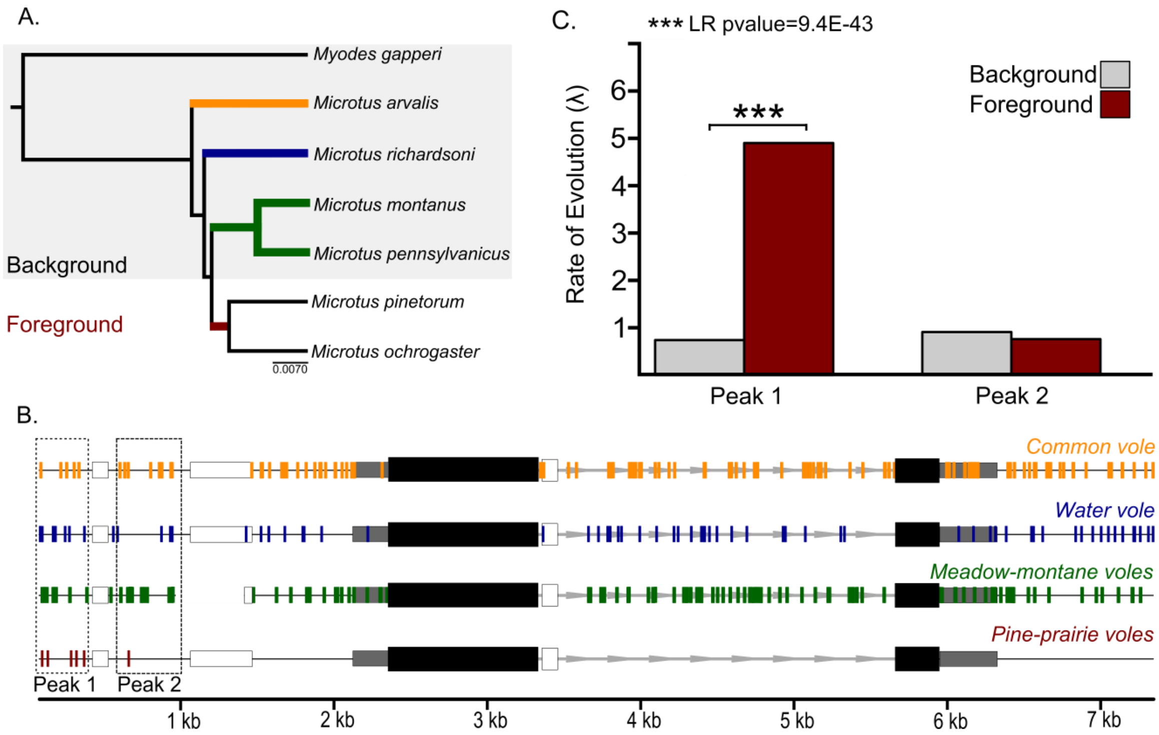


Figure 1.3: Phylogenetic tests of selection. A) Phylogenetic representation of the origin of monogamy. Background branches include all branches of the phylogeny within the shaded gray box (black or colored lines). The branches leading to montane and meadow voles are shown in (green), branch leading to water voles (blue), and branch leading to the European common vole (yellow). The foreground branch, shown in red, corresponds to the common ancestor of the monogamous prairie and pine voles. B) Representative tracks of *Avpr1a* substitutions used in the likelihood ratio test. Changes that occur in specific lineages colored in panel A are shown in panel B. For visual clarity, background changes along uncolored (black) branches are not shown. C) Maximum likelihood estimates of the rates of evolution of putative pallidal enhancers in peak 1 and peak 2, in the foreground and background. Rate heterogeneity identified using the likelihood ratio test.

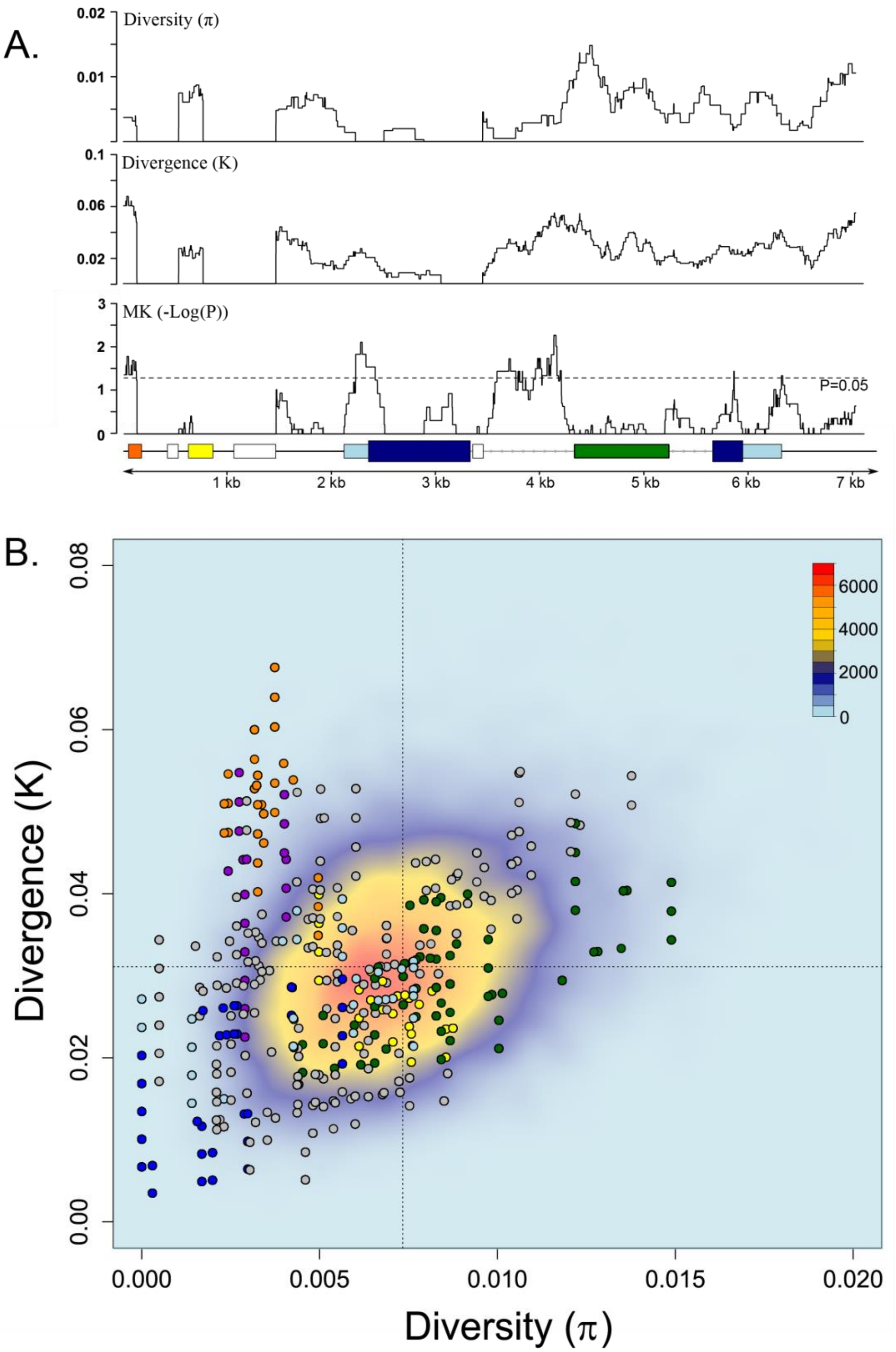


Figure 1.4: Sliding window analyses of selection based on the McDonald-Kreitman test. A) Top panel depicts nucleotide diversity (π) along *Avpr1a* locus. Y-axis

corresponds to the average number of nucleotide differences per site between two sequences. Middle panel depicts divergence (K) between prairie voles (*M. ochrogaster*) and water voles (*M. richardsoni).* Y-axis depicts average proportion of nucleotide differences between populations or species. Bottom panel reports p-values for sliding window comparisons of polymorphism:divergence compared to equivalent data from neutral markers using a Fisher exact test *Avpr1a*. B. Values of polymorphism (π) and divergence (K) for each 300bp window in the avpr1a locus. Data are superimposed on a kernel density estimated from 300bp windows obtained from 100,000bp simulated neutral sequence data. The color legend depicts the number of simulated windows within each grid point after kernel density transformation. Windows overlapping with H3K27ac pallidal peak 1 region (orange) shows low diversity and high divergence; H3K27ac pallidal peak 2 (yellow) shows no evidence of positive selection; coding sequences (blue) display lower polymorphism and divergence; green dots represent a putative enhancer region with more polymorphism than divergence corresponding to an enhancer associated with avpr1a expression in the retrosplenial cortex, an area previously shown to be under balancing selection (Okhovat et al. 2015).

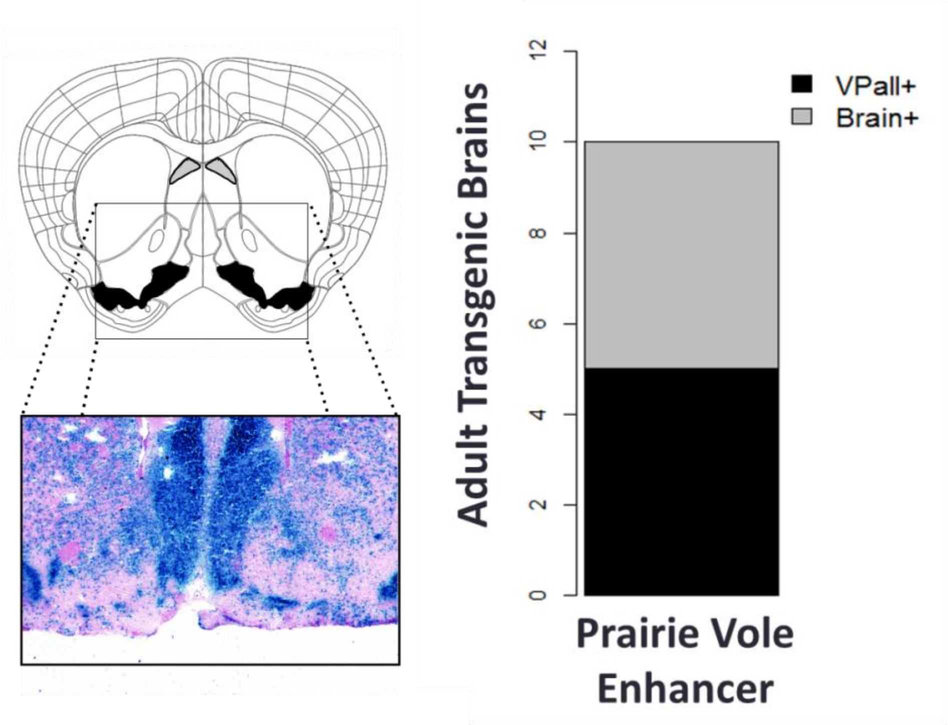


Figure 1.5: Functional characterization of putative enhancer adaptations. Pronuclear injection of a construct spanning H3K27ac pallidal peaks 1 and 2, a mouse minimal promoter (*hsp68*) and the reporter gene *LacZ* reveals that these elements are capable of driving pallidal expression of LacZ in the mouse brain. Top left, diagram showing coronal section of rodent brain, with ventral pallidum shown in black. Lower left, detail of transgenic mouse showing pallidal expression of LacZ. Right, number of 12 transgenic lines expressing LacZ in the brain generally (gray) or the ventral pallidum specifically (black).

CHAPTER 2[*]

Balancing selection on *Avpr1a* : Evidence for epistatic selection and local adaptation

ABSTRACT

Evolutionarily adaptive changes in social behavior are determined by genomic variation, but we know little about how the two are related. In one example, SNPs at the *Avpr1a* locus predict expression of the vasopressin 1a receptor in the retrosplenial cortex (RSC), a brain region that mediates spatial memory; cortical V1aR abundance in turn predicts diversity in space-use and sexual fidelity in the field. To examine the potential contributions of selective and neutral forces to variation at the *Avpr1a* locus, we explore sequence diversity at *Avpr1a* and across the genome in two populations of wild prairie voles. Here, we found strong evidence of balancing selection on the *Avpr1a* locus. Moreover, we also found that the four SNPs that predict high V1aR expression in the RSC are in stronger linkage disequilibrium than expected by chance. Analysis of population structure at two sites revealed that this was unlikely to be due to admixture. Similarly, a haplotype network suggested common origins of major allele classes across populations. Interestingly, we found that the two populations had extremely low levels of genetic differentiation. Despite their similarity, the two populations did seem to differ in the frequency of alternative *Avpr1a* alleles, with measures of differentiation concentrated at the same regions of the locus shown to be under balancing selection. Together, our

[*] Population data and analyses of *Avpr1a* locus in this chapter has previously been published as Okhovat, M; **Berrio, A**; Wallace, G; Ophir, A; Phelps, S. 2015. Sexual fidelity trade-offs promote regulatory variation in the prairie vole brain. Science, 350(6266), pp.1371–1374. Berrio performed experimental design, sequencing experiments, evolutionary analyses, and writing.

data suggest that the balanced polymorphism in *Avpr1a* results from strong local selection at this locus, resulting in allelic frequencies that are associated with unique patterns of spatial cognition and sexual fidelity across populations.

INTRODUCTION

Individual differences in social behavior are common, ranging from the continuous variation that characterizes "personality" (Bell & Sih 2007) to the dramatic variation that defines alternative life-history strategies. Among spadefoot toads, for example, individuals adopt a cannibalistic morphology and behavior that is shaped by both genetic variation and environmental cues (Pfennig 1992; Bazazi et al. 2012). Moreover, the territorial behavior of male side-blotched lizards consists of three discrete phenotypes that vary in aggressiveness, allowing each to predominate for brief periods before being displaced by an alternative, with population dynamics that have been compared to a game of rock-paper-scissors (Sinervo & Lively 1996). Indeed, classic game theoretic approaches (Smith & Price 1973), and more recent models (Slatkin 1979; Sokolowski et al. 1997) suggest frequency- and density-dependent selection should be major contributors to variation in social behavior. Despite these advances, we know little about how selection maintains variation in the nervous system. Here we examine the population genetics of the *Avpr1a* locus, a gene extensively implicated in male social behavior, in the socially monogamous prairie vole, *Microtus ochrogaster* (Phelps & Young 2003; Hammock et al. 2005; Caldwell et al. 2008; Ophir et al. 2008; Barrett et al. 2013).

Prairie voles are well known for forming enduring pair-bonds characterized by shared territory defense, bi-parental care of young, and selective attachment between mates (Getz, Carter, & Gavish, 1981). Although monogamy is clearly the modal mating system, a significant number of individuals mate outside the pair-bond. Multiply sired litters have been detected by multiple labs (Solomon et al. 2004; Ophir et al. 2008), and ~25% of young are sired outside the pair bond (Ophir et al. 2008). Moreover, populations undergo drastic changes in population density annually, ranging from densities that are near zero to as many as 625 animals per hectare (Getz et al. 2001). These changes are accompanied by changes in extra-pair encounter rates and extra-pair fertilizations (Blondel et al. 2016; Solomon et al. 2004; McGuire et al. 1990). This natural history suggests a number of opportunities for frequency- or density-dependent selection to shape the mechanisms of fidelity.

The gene *Avpr1a* encodes for the vasopressin 1a receptor (V1aR), the predominant vasopressin receptor in the central nervous system, a protein critical to male social behavior in many taxa (Donaldson & Young 2008; van Kesteren et al. 1996; Goodson & Bass 2000; Goodson et al. 2009; Bachner-Melman et al. 2005). Among prairie voles, expression of V1aR in regions of the brain that coordinate reward contributes to the ability of male prairie voles to form pairbonds (Lim & Young 2004). Expression of *Avpr1a* in reward centers is uniformly high among prairie voles (Phelps & Young 2003), a finding consistent with the fact that selection seems to favor the capacity to form pairbonds (Ophir et al. 2008; Phelps & Ophir 2009). Interestingly, other brain regions, including the retrosplenial cortex (RSC), vary tremendously among individuals (Insel et al. 1994; Phelps & Young 2003). The RSC is a critical node in a circuit that coordinates spatial memory and navigation (Troy & Whishaw 2004; Todd & Bucci

2015). Moreover, the abundance of V1aR in the RSC is predictive of differences in territorial intrusion rates and extra-pair paternity among males (Ophir et al. 2008, Okhovat et al. 2015).

We recently demonstrated that individual differences in the RSC are well predicted by a set of four highly linked single nucleotide polymorphisms (SNPs) in the *Avpr1a* locus that co-localize with markers of regulatory DNA, including a SNP in a DNAse I hypersensitive site 5' of the *Avpr1a* locus, and two additional SNPs within a putative intron enhancer identified by ChIP-seq (Okhovat et al. 2015). Field experiments suggest that the two alternative allele classes, which we refer to as HI and LO RSC alleles, are under opposing selection when environments favor intra-pair or extra-pair fertilization (Okhovat et al. 2015). Here we use genome-wide polymorphism data to more rigorously test whether the *Avpr1a* locus has indeed been under balancing selection. Next we ask whether the linkage observed between the SNPs of HI and LO *Avpr1a* alleles is greater than expected by chance. We find an excess of linkage between SNPs that suggests selection may have favored specific combinations of alleles to be in phase with one another; we use genome-wide patterns of polymorphism in two different populations to test the alternative hypothesis that this excess of linkage is a by-product of admixture-induced population structure. Next we use haplotype networks to visualize the relatedness of HI and LO RSC alleles, and ask whether HI or LO allele classes have a common origin. Lastly, we use our *Avpr1a* haplotypes and genome-wide sequencing data to ask whether there is evidence of local adaptation of *Avpr1a* allele frequencies between two populations of prairie voles. Our results provide the first detailed examination of population variation in a locus directly linked to differences in brain function and social behavior.

MATERIALS AND METHODS

Population sampling and DNA extractions

To assess intraspecific variation in prairie voles, 32 individuals were collected in Champaign County, IL, and 29 individuals were collected from Jackson County, IL. All genomic DNA extractions were prepared from liver samples using the QIAGEN DNEasy kit for tissue and blood following the manufacturer's protocol.

Polymerase chain reactions

Approximately 7.7kb of the *Avpr1a* locus was amplified and sequenced from individuals of Champaign County, and details of this amplification have been previously published (Okhovat et al. 2015). Here we sequenced a subset of this locus (5.5 Kb) in the population of Jackson County, using the primers listed in Table 1. All polymerase chain reactions (20 μL) were prepared with 1X GoTaq® Hot Start Polymerase buffer (Promega), 1.0 mM of MgCl2, 0.2 mm of each dNTP, 0.5 μl of each primer, 1.25 U HotStart Taq Polymerase and 1.0 μL of diluted DNA. Amplifications were carried out on a Veriti Thermal Cycler (Applied Biosystems) using the following temperatures: initial denaturation at 95 °C for 2 min; 35 cycles of 95 °C for 40 s, annealing at 58 °C for 30 s and extension at 72 °C for 240 s; and a final extension at 72 °C for 7 min. Semi-nested PCRs were necessary to amplify this locus in some individuals; for these we used an initial denaturation at 95 °C for 2 min; 35 cycles of 95 °C for 40 s, annealing at 58 °C for

30 s and extension at 72 °C for 210 s; and a final extension at 72 °C for 7 min. Each amplicon was first visualized on 1% agarose gels in order to check its band size and specificity, and then cleaned with Qiaquick PCR purification kits (Qiagen) according to the manufacturer's protocol. Sanger sequencing was conducted at the University of Texas at Austin Institute for Cellular and Molecular Biology (ICMB). The resulting ABI Chromatograms were processed and analyzed using 'Map to Reference' parameters in Geneious v6.1 (Biomatters). The *Avpr1a* locus contains three highly polymorphic microsatellite sequences (see figure 2.2). Given that microsatellite sequences are difficult to sequence accurately, they were excluded from all analyses.

Cloning

Thirteen individuals were selected for cloning and direct determination of the haplotype phase. Of these, five were heterozygous animals from Champaign County and three from Jackson County. A ~5.5 Kb sequence amplicon was diluted in water and then ligated and cloned using the StrataClone PCR Cloning kit and the recommended protocol from the manufacturer (Agilent Technologies). Briefly, ligation reactions were transformed into chemically competent StrataClone SoloPack cells and plated onto LB agar plates supplemented with kanamycin and spread with X-gal at 2% (FisherScientific). At least three transformed white colonies were transferred to a tube with 5 mL of LB broth to grow overnight at 37C. Plasmid DNA was isolated using the QIAprep spin miniprep kit (Qiagen). The insert size was confirmed by restriction digestion with EcoRI and visualization in a 1% agarose gel. Sanger sequencing was conducted at the University of Texas at Austin Institute for Cellular and Molecular Biology (ICMB).

Haplotype reconstruction of the *Avpr1a* locus

Haplotypes of 151 polymorphic sites along a 7.7 Kb segment of the *Avpr1a* locus were reconstructed in 32 individuals of the Champaign County population; for the Jackson County population, consensus sequences spanning 5.5Kb with 109 polymorphic sites were used to reconstruct haplotypes using default settings in the statistical package PHASE v2.1 (Stephens et al. 2001). Based on data from cloned PCR amplicons, the phase was known for at least 22% of all polymorphic sites. These known phases were also included in the PHASE algorithm.

To illustrate the relationships among haplotypes and examine the origin of the HI alleles, a haplotype network of the *Avpr1a* locus was constructed using the Median Joining network algorithm (Bandelt et al, 1999) implemented in PopART v1.7 (Leigh and Bryant, 2016, http://popart.otago.ac.nz); the resulting network was edited in Inkscape v0.48.

Linkage disequilibrium

To estimate linkage among SNPs across the *Avpr1a* locus, we computed D and its confidence statistic values (R^2) using HAPLOVIEW v4.2, excluding all microsatellite loci. Consequently, these values were used to evaluate a sliding window track of LD across the *Avpr1a* locus by evaluating at least three SNPs within a window of 300bp. Windows with fewer than three SNPs were treated as missing data. To test whether the set of polymorphisms defining HI and LO alleles exhibited higher linkage disequilibrium than expected by chance, we first regressed pairwise LD values (R^2) against distances between sites, and then calculated the average residual R^2 for each of 6 pairwise comparisons of the 4 strongly linked sites. To estimate the null distribution of this statistic, we randomly sampled 4 SNPs (excluding the 4 that define HI/LO alleles) and calculated the average residual R^2 based on their nucleotide distance). For an alternative analysis, we randomly selected a focal SNP and noted the position of other SNPs with an LD as great as that observed in the HI/LO SNPs, we then calculated the distance along the *Avpr1a* locus that this randomly selected linkage group spanned. We compared the observed span to the distribution of spans expected from a linkage group chosen at random.

Calculation of population genetic summary statistics

For our initial samples from Champaign County, we used the package *DnaSP v5.1* (Librado & Rozas 2009) to compute *nucleotide diversity* (π) in 300bp sliding windows. The recombination rate *rho* was computed with the program *RECSLIDER* (http://genapps.uchicago.edu/labweb/index.html) within a window size of 10 variable sites. A preliminary estimate of *rho*=0.0044 was based on the average recombination rate between sites, obtained from *DnaSP*.

RAD library construction and sequencing

Genome-wide 2bRAD-seq is based on the use of type IIB restriction enzymes that target a small fraction of the genome. Type IIB enzymes excise 36bp-fragments alongside the recognition site to allow detection of genetic variants. Vole DNA was treated with the enzyme BcgI (NEB) to produce sufficient genome fragmentation; fragments were ligated to Illumina adaptors, amplified for 10 cycles, and then purified according to the protocol designed by Wang and collaborators (2012). Final library preparations were pooled and sequenced by University of Texas at Austin Genome Sequencing and Analysis Facility (GSAF). Quality control and fragment distribution was examined before sequencing using the Agilent Bioanalyzer. The pooled sample was sequenced using a 50bp single-read strategy on the Illumina HiSeq4000 platform at the GSAF facility, which generated a total of 277,180,409 reads. The quality of the reads was examined and approved by visualizing the FastQC output for each sample, followed by processing with pipelines for genome-guided genotyping developed by Mikhail Matz (https://github.com/z0on/2bRAD_GATK). Reads that lacked the overhanging restriction site were discarded. Individual samples were identified using the barcodes incorporated during ligation and PCR amplification stages. Average read count for seventy samples that passed the trimming and other filters was 373,289 reads. After this initial quality check, one sample was dropped due to low sequencing depth.

Genome-guided RAD-genotyping and variant discovery

Trimmed reads were aligned to the prairie vole draft genome assembly using bowtie2 (Langmead & Salzberg 2012); mapping efficiency was >95% for all the samples according to the flagstat report implemented in SAMtools software (Li et al. 2009). The program GATK (McKenna et al. 2010) was used to identify genetic variants; UnifiedGenotyper tool was run twice followed by base quality score recalibration using the script GetHighQualVcfs.py to score SNPs with a quality percentile of 75 or higher (Kyle Hernandez, https://github.com/kmhernan/tacc-launcher-bio/blob/master/utils/GetHighQualVcfs.py). The last recalibration step was done using vcftools (Danecek et al. 2011), producing a total output of 132,573 SNPs, of which 4955 passed a reproducibility test that was run in four sets of replicate samples. The false discovery rate (FDR) was estimated from each portion of SNPs (i.e. tranche) based on the difference between the estimated and expected transition(Ti)/transversion(Tv) ratio of 2.41. This ratio was estimated from 4955 polymorphic sites that were fully consistent and highly reproducible among replicate samples. This model only considers tranches with a Ti/Tv ratio that is higher than expected due to a deficiency of false positives. The 31,965 SNPs that passed this final filter were recorded in a VCF file.

Population structure

For our two populations of prairie voles, we conducted a principal component analysis (PCA) using the 'snpgdsPCA' tool implemented in the SNPRelate library (Patterson et al. 2006) to examine differences between populations. Next we estimated genetic clusters using ADMIXTURE v1.3 (Alexander et al. 2009), which estimates the ancestry in a model-based manner from SNP datasets. Finally, To estimate global F_{ST}, we used BayeScan to identify F_{ST} outliers among the 31,965 SNPs that passed our quality controls (Foll & Gaggiotti 2008), and measured global F_{ST} using the Weir and Cockerham method implemented in Vcftools (Danecek et al. 2011). We used 4P (Benazzo et al. 2015) to calculate genome-wide diversity per population from the VCF file, and custom scripts in R (Team 2015) to calculate the global genome average of D_{xy} between populations.

Population differentiation at *Avpr1a*

To examine population differences at the *Avpr1a* locus, we used the *Gene Flow and Genetic Differentiation* tool in *DnaSP* to compute F_{ST} and D_{xy}. Gene tracks of F_{ST}

were generated using the program *SLIDER* (http://genapps.uchicago.edu/labweb/index.html).

RESULTS

Evidence of balancing selection in the *Avpr1a* locus

Genetic variation at the *Avpr1a* locus predicts expression of V1aR in prairie vole RSC (Fig 2.1A). We examined the frequency spectrum of genome-wide SNPs and at the *Avpr1a* locus among wild-caught samples from Champaign County, IL. We found that the *Avpr1a* locus was strongly skewed toward an excess of intermediate-frequency alleles compared with the rest of the genome (Fig. 2.1B), a typical signature of balancing selection. Consistent with our prior analysis, the *Avpr1a* frequency spectrum is strongly skewed towards intermediate frequencies (KST, D = 0.40, P < 2.2e-16). The *Avpr1a* region also had a significantly positive Tajima's D (P<0.05; see also Okhovat et al. 2015), which is consistent with balancing selection at this locus. In contrast, we found that the genome-wide average of Tajima's D is not different from 0 (TD= -0.0040).

Patterns of polymorphism and linkage disequilibrium at the *Avpr1a* locus

To characterize local patterns of polymorphism at *Avpr1a*, we used reconstructed haplotypes from the Champaign population of wild-caught voles. From these, we estimated nucleotide diversity (π), linkage disequilibrium (LD, R^2) and recombination (*rho*). A sliding window analysis shows local peaks of nucleotide diversity and LD within a known putative intron enhancer (Fig 2.2a) (Okhovat et al. 2015). We found that average nucleotide diversity for the *Avpr1a* locus (π_{Avpr1a} = 0.0045) was significantly higher than across the rest of the genome (π_{genome}= 3.1x10^{-6}, Mann-Whitney U-Test, P < 2.2x10^{-16}).

Diversity values ranged from 0 to 0.015, with two local peaks (π > 0.01) spanning roughly 200 bp (between positions 4520 and 4730) and 500bp (between 7060 and 7596).

On average, linkage disequilibrium is low—even between adjacent sites—and decays with distance (0.00003*bp + 0.19). There is, however, considerable variation in LD across the region. This heterogeneity in LD patterns is apparent in our recombination estimates, which suggest the presence of two hotspots near the boundaries of the intron enhancer and its RSC-predictive SNPs (Fig 2.2a). Additionally, a set of linked SNPs predictive of RSC variation are more highly linked than 4 SNPs chosen at random, even after correcting for the decay of LD with distance (Fig 2.2b; *P*=0.0001, Fig 2.3). The unusually high levels of association suggest selection favored specific combinations of SNPs that shape the functions of HI and LO RSC alleles.

Genome-wide population structure

We found very weak genome-wide structure between two populations of prairie voles separated by ~200 miles (Fig 2.4A). A principal components analysis suggested that, while it represents the largest source of variance, population structure accounts for only 3.5% of genetic variation in our sample (PC1, vertical axis in Fig 2.4B). Moreover, our admixture model confirms very low population structure, which favors a model with a single population. Cluster one (K1) exhibited less cross-validation error than clusters K2 and K3 (CV error (K=1): 0.52 vs CV error (K=2): 0.55 vs CV error (K=3): 0.60) (Fig 2.4B). In addition, genome-wide values of absolute (D_{xy}) and relative nucleotide divergence (F_{ST}) indicate that these two populations are extremely similar (BayeScan

F_{ST}=0.06, Weir and Cockerham mean F_{ST} estimate: 0.02, Global D_{xy} = 0.0069, assuming 1/3 of the restrictions cut sites were sequenced).

Haplotype Network of the *Avpr1a* locus

We sampled 110 haplotypes from a total of 61 wild-caught prairie voles collected in Champaign and Jackson County. As in our genome-wide data, the general pattern of our median-joining network analysis shows extensive mixing of *Avpr1a* haplotypes across populations (Fig 2.5). However, LO alleles were extremely diverse in both populations, we observed 103 distinct LO haplotypes (average π = 0.005 across all LO alleles). The HI alleles, in contrast, were far less diverse (average π = 0.001) in only 7 haplotypes. All but one HI allele clustered together on the haplotype network. Examination of the sequence of the one HI allele that clustered with LO alleles suggests it was a recombinant containing the intronic HI allele SNPs, but a set of 5' sequences more closely resembling the LO alleles. For both HI and LO alleles, there was no clear segregation by population, suggesting that variants in both populations have a common origin. Lastly, we validated the specificity of our PCR primers given that none of the *Avpr1a* haplotypes clustered near a pseudogene haplotype sequence.

Patterns of population differentiation at the *Avpr1a* locus

The average F_{ST} between pop X and Y at the *Avpr1a* locus was considerably elevated compared to the rest of the genome ~0.25 (Fig 2.6A, Mann-Whitney U-test, $p < 2.2e-16$). This result is consistent with our sliding-window analysis of wild-caught samples from Champaign County (see Fig 2.4B), which found peaks of nucleotide diversity (π), recombination rate (*rho*) and linkage disequilibrium (LD, R^2) in a putative regulatory element of this locus (Fig 2.2A). We found similar patterns of nucleotide diversity when we included wild-caught samples from Jackson County in this analysis (Fig 2.6B, *top*). Moreover, we also found abnormally high values of relative (F_{ST}) (Fig 2.6B, *middle*) and absolute divergence (D_{xy}) (Fig 2.6B, *bottom*) at these same sites. Overall, the values of relative and absolute differentiation at the *Avpr1a* locus depart significantly from the whole genome, our sliding windows analysis revealed a maximum F_{ST} value of 0.49 in the intronic region. Interestingly, this pattern of relative differentiation is similar to the pattern of absolute nucleotide divergence across the locus, which reached a local average of Dxy_{Avpr1a} =0.006 and a maximum value in the putative enhancer of D_{xy}= 0.02. Comparing this peak with genome-wide average Dxy_{global}=0.0069, these values are statistically significant differences (Mann-Whitney U-test, $P < 2.2e-16$).

DISCUSSION

Prairie voles are socially monogamous rodents that exhibit bi-parental care and territory defense, but also display considerable individual differences in territorial behavior, space use and sexual fidelity (Getz, McGuire, and Pizzuto 1993; Carter, Getz, and Cohen-Parsons 1986; Solomon et al. 2004; Phelps and Ophir 2009). Four highly linked SNPs reside in *cis*-regulatory elements at *Avpr1a* and predict expression of V1aR in the retrosplenial cortex (RSC); RSC expression in turn predicts aspects of space use and sexual fidelity (Fig 2.1A). Moreover, balancing selection is thought to maintain nucleotide diversity at this locus (Ophir et al. 2008; Okhovat et al. 2015). Here, we examine population patterns of variation at the *Avpr1a* locus to understand the origin and evolution of this interesting variation in brain and behavior.

We recently compared the frequency spectrum at the *Avpr1a* locus to three putatively neutral loci and found that the *Avpr1a* locus had a significant excess of intermediate frequency polymorphisms (Okhovat et al. 2015). While this was strongly suggestive of balancing selection, the limited sampling of neutral loci leaves open alternative possibilities, such as population admixture, that could lead to heterogeneity in frequency spectra across regions of the genome. We used 2bRADseq to construct a reduced-representation map of genome-wide frequency spectra from wild-caught prairie voles collected in Champaign County, IL. As expected from neutral distribution of allele frequencies near mutation-drift equilibrium (Nei, Chakraborty, and Fuerst 1976; Luikart et al. 1998), we observed a high density peak of alleles at low frequency (<0.1) and a smaller peak that corresponds to fixed differences respect to the reference genome. Moreover, the analysis of population summary statistics at the genome-wide level was consistent with our previous results; both the frequency spectrum and Tajima's D deviated from the genome expectation (Fig 2.1). Genome-wide Tajima's D agreed with the neutral expectation, while the *Avpr1a* locus had a significantly positive Tajima's D (P<0.05; see also Okovhat et al. 2015). The absence of a secondary mode of RADseq sites corresponding to intermediate allele frequencies similar to the *Avpr1a* suggests that simple admixture is unlikely to account for the skewed frequency distribution we observe at the *Avpr1a* locus. Our results strongly support the hypothesis that balancing selection is actively maintaining high levels of variation in a regulatory element associated with *Avpr1a* expression in a spatial memory circuit.

To characterize the evolution of this locus more closely, we used a combination of cloning and phase-estimation to reconstruct the haplotypes from this wild-caught population. From these haplotypes, we estimated nucleotide diversity (π), recombination (*rho*) and linkage disequilibrium (LD, R^2). We found local peaks of both nucleotide diversity and linkage disequilibrium, a classic signature of balancing selection, co-localized with regulatory elements implicated in RSC *Avpr1a* expression (Fig 2.2A). The largest LD block, for example, coincides with a 5' peak in π and overlaps with a DNAse I hypersensitive site that contains a SNP contributing to HI and LO RSC alleles. The second largest linkage block occurs in the intron, where it overlaps another local peak of nucleotide diversity; this block overlaps an enhancer identified in the RSC by H3K4me1 ChIP-seq, and two additional SNPs associated with HI and LO RSC alleles (Okhovat et al. 2015). Interestingly, the pattern of LD within and between these two blocks is somewhat unusual. Although the two blocks are not linked to one another, the SNPs that define HI and LO alleles are strongly linked, defining points of significant LD that are very distinct from the broader pattern of recombination between these blocks (Fig 2.2B). Moreover, a detailed examination of LD patterns reveals these SNPs are poorly linked to other polymorphisms within each of these blocks. Our estimates of the recombination rate *rho*, moreover, suggest a local elevation in recombination within the putative RSC intron enhancer. We speculate that this pattern may reflect not only balancing selection on the locus, but epistatic selection favoring specific combinations of regulatory variants that influence RSC V1aR and male mating strategies.

To test whether patterns of LD observed between SNPs that define HI and LO RSC-expressing alleles are higher than we would expect by chance, we first calculated the LD between possible pairs of polymorphic sites and plotted LD against distance in base pairs (Fig 2.3). Pairwise comparisons for SNPs that define HI and LO alleles are shown in red, all other comparisons are shown in blue. We observed that the background levels of LD are surprisingly low along the *Avpr1a* locus, suggesting that the standing diversity at the locus is extremely old. We found that the residual LD after correcting for distance was indeed significantly higher than expected for four randomly selected SNPs. More conservatively, we also tested for excess LD by choosing 4 SNPs with LD equal to or greater than observed for the HI and LO SNPs and calculated the span of the sites across the locus. We found that the probability of 4 linked SNPs spanning 4897 bp or more was $P<10^{-4}$. These data all suggest that the extent of LD among HI and LO SNPs is substantially greater than expected by chance, findings which suggest selection has actively maintained specific combinations of SNPs at the locus.

One alternative explanation for unusual patterns of LD among HI/LO SNPs is that an adaptive allele arose in another population and spread into our population by admixture. To explore the structure of prairie vole populations more thoroughly, we conducted 2bRADseq on a second population of prairie voles located ~200 miles from our original source, and we examined genome-wide polymorphisms from both populations for evidence of admixture (Fig 2.4A). Interestingly, although a principal components analysis revealed limited differentiation between the two populations (3.5%, Fig 2.4A), an ADMIXTURE analysis suggested that a model with a single population was slightly better than a model with two populations. Similarly, the estimated genome-wide F_{ST} of the two populations was just 0.06, confirming the very low differentiation of the two populations. While an ADMIXTURE model with k=1 was better than a model in which k=2 (corresponding to the two sampled populations), both of these models were better than a model that included three or more clusters. Overall, we found no evidence that admixture shapes the distribution of genetic variation in our two sampled populations. While migration from other (unsampled) populations could still be the source of high LD, admixture seems an unlikely explanation for the origin of the unusual patterns of association we observe at the *Avpr1a* locus.

To evaluate the origin of the HI/LO RSC alleles, we sequenced and cloned additional *Avpr1a* loci from our more southern Jackson County population, limiting our efforts to ~5.5kb that spanned the intron and 5’ HI/LO SNPs. We evaluated the haplotype structure of the *Avpr1a* locus by reconstructing a neighbor-joining haplotype network (Fig 2.5). Our haplotype network shows an abundance of haplotypes that can be clustered in at least four haplogroups, each of which is present in both populations. Additionally, we identify only a handful of mutations separating the origin of HI haplotypes present in Champaign and Jackson Counties. A single HI allele seems more closely related to other LO alleles; a closer examination of the sequence, however, reveals that it is a recombinant form, the only one that decouples the 5’ HI SNP from the intronic HI SNPs.

Together these data suggest that the HI allele is the derived form, and shares a common origin across these populations.

Together our data suggest selection has maintained specific combinations of SNPs that contribute to HI and LO alleles, which in turn shape RSC expression, social cognition and mating fidelity (Okhovat et al. 2015). It remains possible, however, that the HI allele arose in another population and rose to intermediate frequencies through selection. Such a scenario would require selection to be old enough that the resulting signatures of admixture were not visible to our 2bRAD-seq analysis. This is a difficult hypothesis to refute. An alternative approach would be to test for evidence of epistatic selection more directly. For example, it may be possible to examine rare but naturally occurring recombinants among the SNPs that define HI and LO alleles to ask whether recombinants occur significantly less often than predicted by chance. Similarly, one could ask whether such recombinants or artificially induced mutations alter RSC expression of the *Avpr1a* locus.

There are many examples of gene interactions evolving by epistatic selection. Including, the classic epistatic effects in the major complex of histocompatibility (Gregersen et al. 2006; Trowsdale & Knight 2013), and the white color mutations in the MCR1A locus that contributes to coat color in domestic pigs only when specific mutations in the KIT locus that encodes the mast/stem cell growth factor (MGF) are inherited (Marklund et al. 1998). Epistasis has even been detected between coding and regulatory elements (Lappalainen et al. 2011), but few studies have revealed epistatic selection on specific combinations of gene regulatory elements or transcription factor binding sites (Phillips 2008; Anderson et al. 2015).

While balancing selection has maintained variation in the *Avpr1a* locus, frequency- and density-dependent selection may also vary across populations. If frequency- or density-dependent selection on the *Avpr1a* locus differs between our populations, for example, we would expect *Avpr1a* to show stronger differentiation between our populations than expected from our genome-wide RAD-seq data. To investigate this, we estimated relative and absolute values of differentiation (F_{ST} and D_{xy}). Indeed, we found that for both F_{ST} and D_{xy}, the intronic enhancer at the *Avpr1a locus* was much more differentiated between populations than predicted based on whole genome estimates (F_{ST}=0.02, Mann-Whitney U-test, $P< 2.2x10^{-16}$, Fig 2.6A; Dxy_{global}=0.0069, Mann-Whitney U-test, $P< 2.2x10^{-16}$). These results are surprising, given that the *Avpr1a* locus behaves as a balanced polymorphism, which usually suggests reduced F_{ST} (Hohenlohe et al. 2010). A sliding window analysis revealed that elevated values for both relative differentiation (F_{ST}) and absolute divergence (D_{xy}) occur in the vicinity of the intron enhancer (Fig 2.6B). These data suggest that balancing selection operating on the RSC intron enhancer may favor different frequencies of HI and LO alleles in these two populations. It is not clear what the source of such local adaptation might be, but the populations differ in the severity of winter and the relative abundance of suitable grassland habitat, either of which might contribute to fluctuations in population density and resource availability. Population density alters extra-pair encounter rates and observed sexual fidelity (McGuire et al. 1990; Streatfeild et al. 2011). Population-specific fluctuations in the defensibility of females could thus influence the strength and direction of selection on HI and LO alleles.

In conclusion, population genetic data indicate that prairie voles harbor an unusually high level of genetic diversity at the *Avpr1a* locus compared with the rest of the genome, and both nucleotide diversity and linkage patterns are elevated in the vicinity of SNPs that predict cortical V1aR abundance and patterns of space use and sexual fidelity. Moreover, the SNPs that define HI and LO RSC alleles are more highly linked that predicted by chance. This pattern is not accompanied by evidence of admixture in the genome, suggesting selection has actively maintained specific combinations of polymorphisms. Comparisons of populations reveal very low between population differentiation, but substantial differences in the frequencies of variants within the *Avpr1a* locus. This suggests that in addition to balancing selection within a population, there may also be local adaptation in *Avpr1a* allele frequencies. Together the data reveal how individual differences in brain and behavior can be maintained by selection that fluctuates in time and space.

Primers	**Sequence 5’ to 3’**
Avpr1a PCR (For)	GCCACAAATAGACCAACGTTCTTAAG
Avpr1a Seq 1	ATTCCCATAGTAAAGATTGTTTG
Avpr1a Seq 2	GCCTTGTGTCAGCAGCGTG
Avpr1a Seq 3	GACTGGGAAAGGATTCAAGAAGTC
Avpr1a Seq 4	GTCATCTGCGAGACCTAACAC
Avpr1a Seq 5	TCTGTGGTGTGAATAGTTCC
Avpr1a Seq 6	GTTGGGATTGTTGAGAACCACA
Avpr1a Seq 7	CTGTATACTGTGCATAGAAGC
Avpr1a Seq 8	GCTGCTCTAACAGTGGTTGGTTTG
Avpr1a Seq 9	GTGCAGTGTGCAGGGTCTTGCTC
Avpr1a Seq 10	CAGGTGGAAACAGGAATGAATCTG
Avpr1a PCR (Rev)	TGGCATCCCTTGTACAAACT

Table 2.1. PCR and sequencing primers (5’ to 3’) for characterizing *Avpr1a* sequence variation

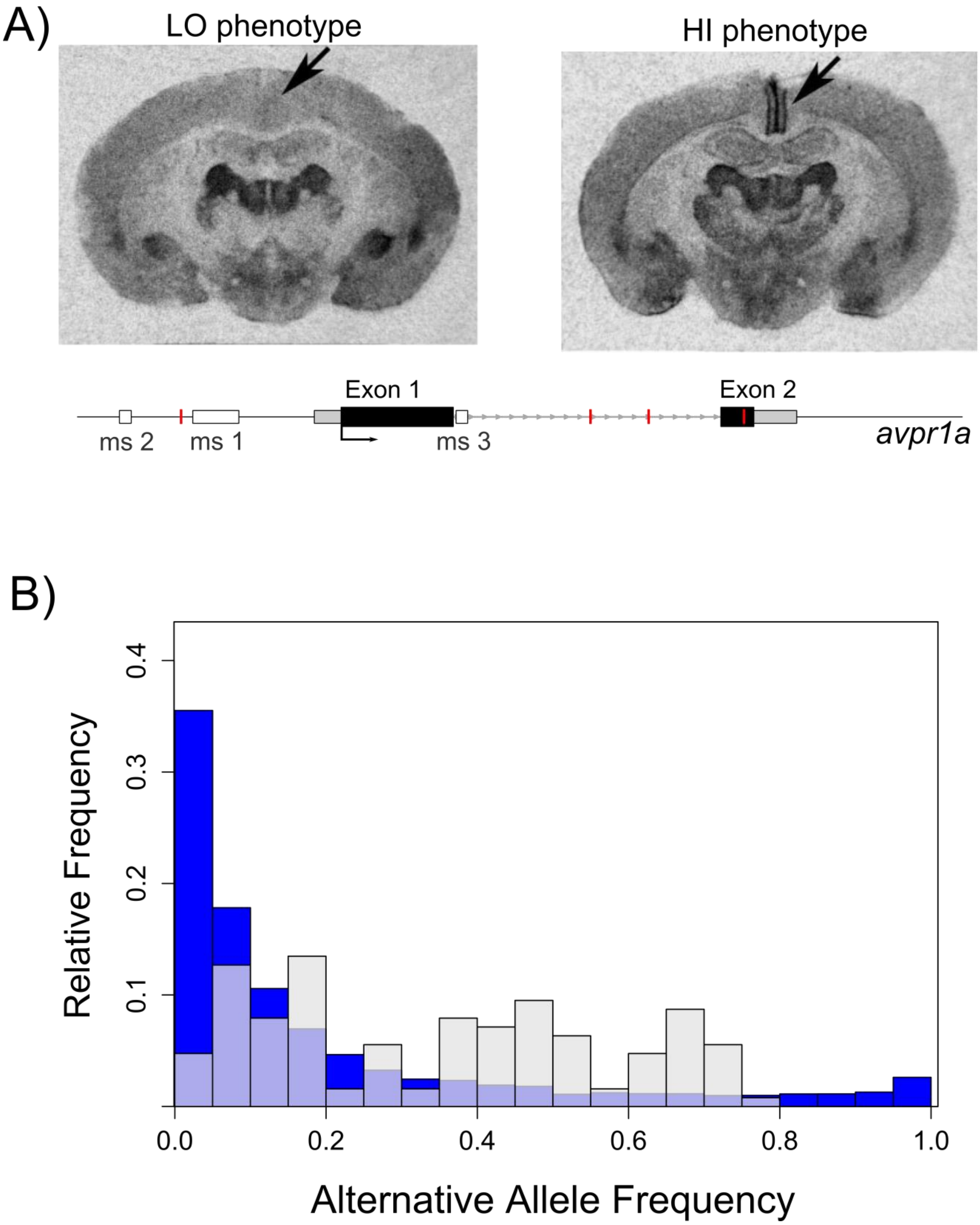


Figure 2.1: **Evidence of balancing selection on the *Avpr1a* locus**. A. Low and high V1aR expressing phenotypes in the retrosplenial cortex (RSC). *Avpr1a* locus contains

two exons and three microsatellite sequences. SNPs associated withRSC-V1aR expressionare represented as red vertical lines. B. Distribution of allele frequencies observed for randomly samples genome-wide loci (blue columns) and at the *Avpr1a* locus (gray columns).

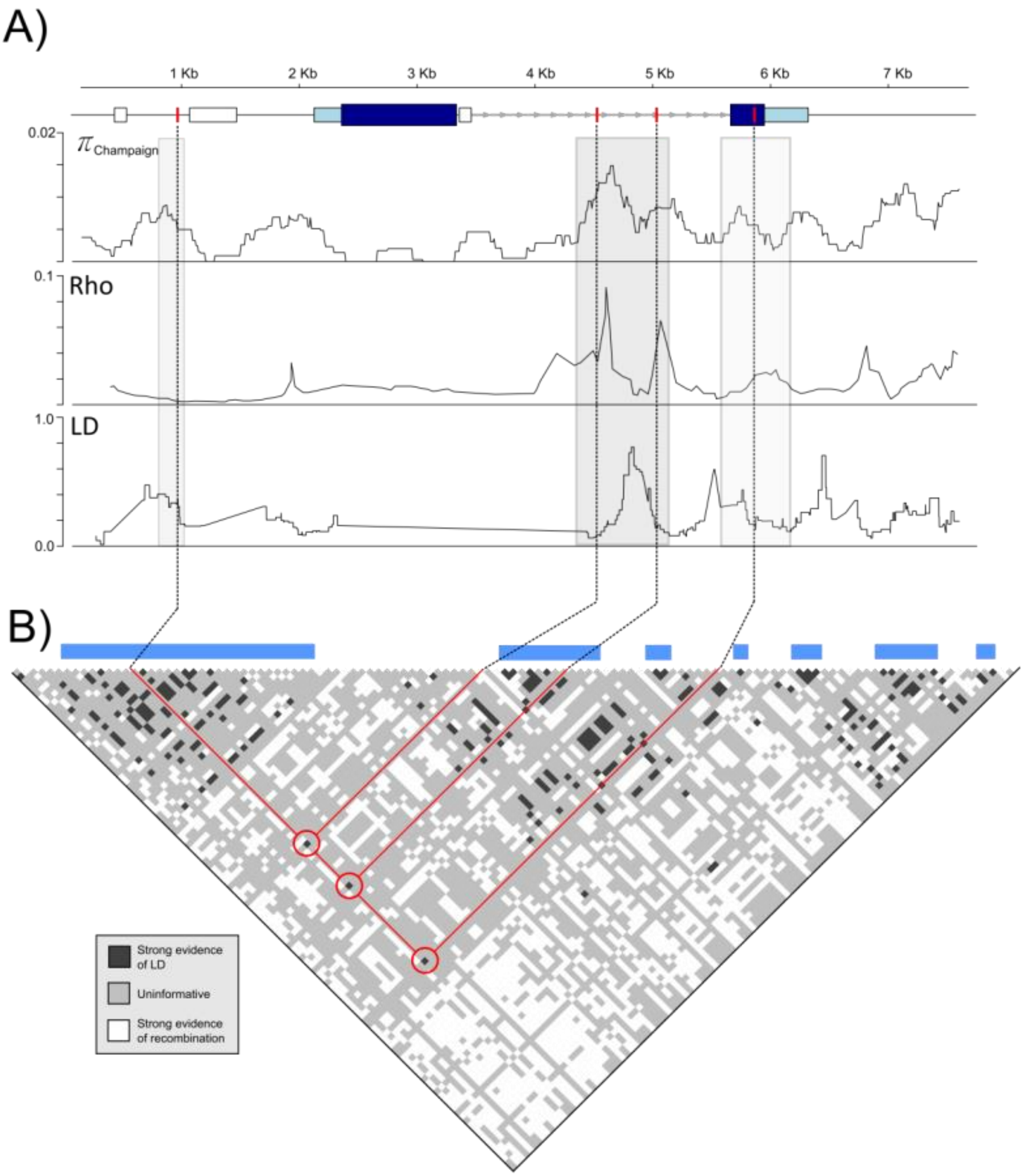


Figure 2.2: Patterns of diversity, recombination and linkage at the *Avpr1a* locus. A) *Top panel*, structure of the *Avpr1a* locus. Red lines correspond to SNPs associated with high V1ar expression in RSC; location of the putative intron enhancer is highlighted in gray (Okhovat et al. 2015). *Second panel*, sliding-window analysis of nucleotide diversity

(π) in animals from Champaign County, IL*; third panel*, recombination rate (Rho); *bottom track*, linkage disequilibrium (LD, R^2). B) Heatmap depicts significant evidence of linkage (black) and recombination (white) between pairs of SNPs. The four SNPs linked to RSC are strongly linked with each other, but poorly linked to surrounding sequences. Seven LD blocks are depicted in blue.

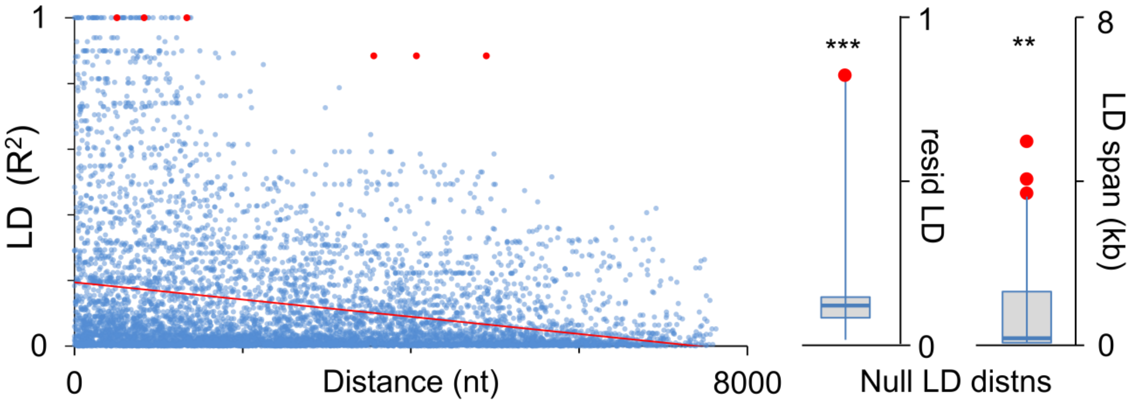


Figure 2.3: Excess of linkage among SNPs defining HI and LO alleles. The four SNPs associated with RSC-V1aR are strongly linked. Left panel depicts LD decay with distance (in nucleotides, nt). Red dots correspond to LD between pairs of SNPs linked to RSC expression, blue dots to comparisons between SNPs not linked to RSC. Right panels demonstrate that the average residual LD for RSC-linked SNPs (red) is much larger than four randomly sampled SNPs at the locus, and that the span of four linked SNPs is significantly shorter than the span of the SNPs that define the HI allele. Box and whisker plot depicts median, quartiles and range of null distribution.

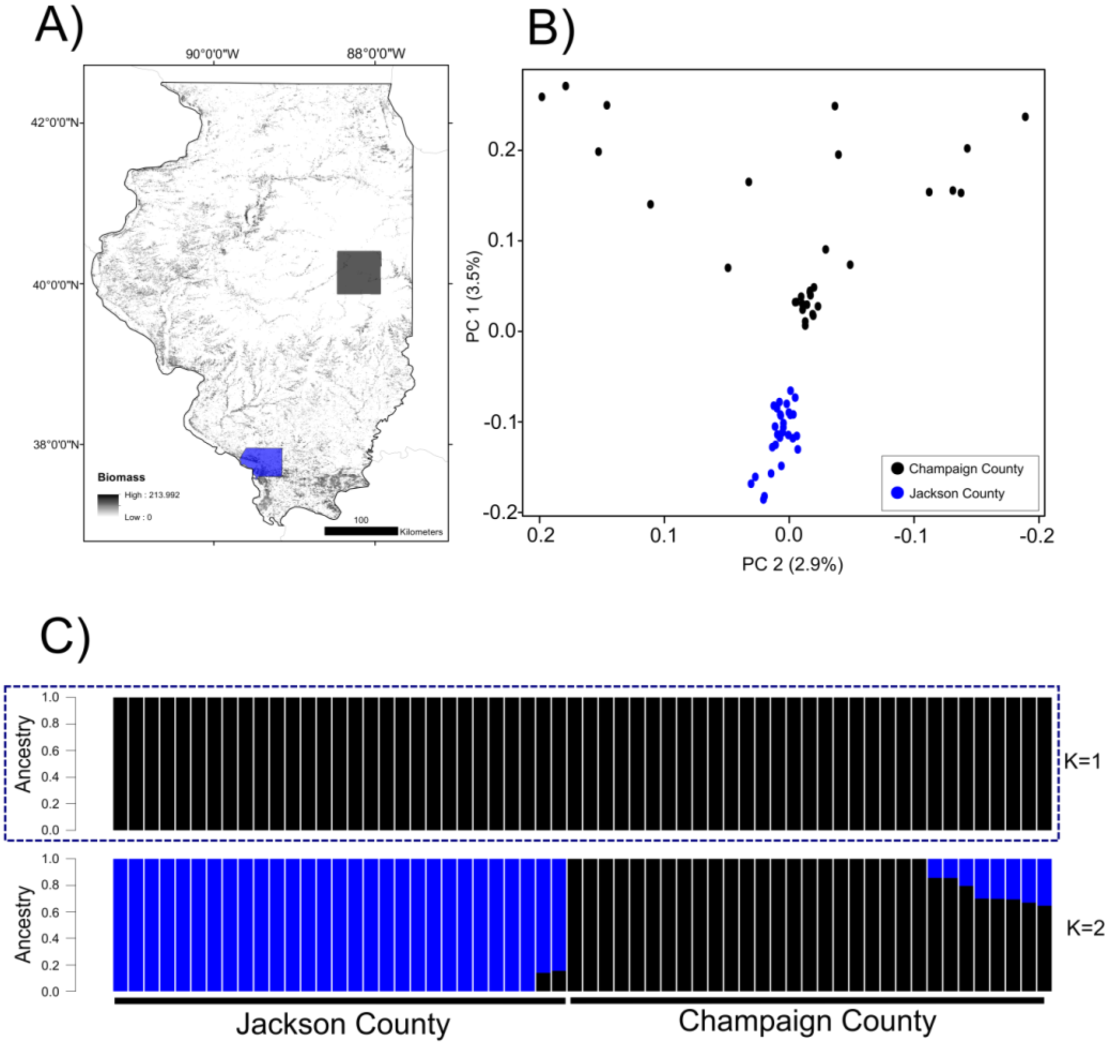


Figure 2.4: Population structure analysis of prairie voles in central and southern Illinois. A) Sampling locations of prairie vole populations, Champaign County, IL (black), Jackson County, IL (blue), shaded biomass is depicted to represent differences in forest vs prairie. B) Principal component analysis (PCA) generated on the basis of individual genotypes from 2bRAD data. PC1 and PC2 represent eigenvectors that accounted for 3.5% and 2.9% of the total genetic variation. C) Admixture cluster analysis

representing the inferred ancestry from K ancestral populations. Blue dashed box highlights the cluster (K) with the smallest error.

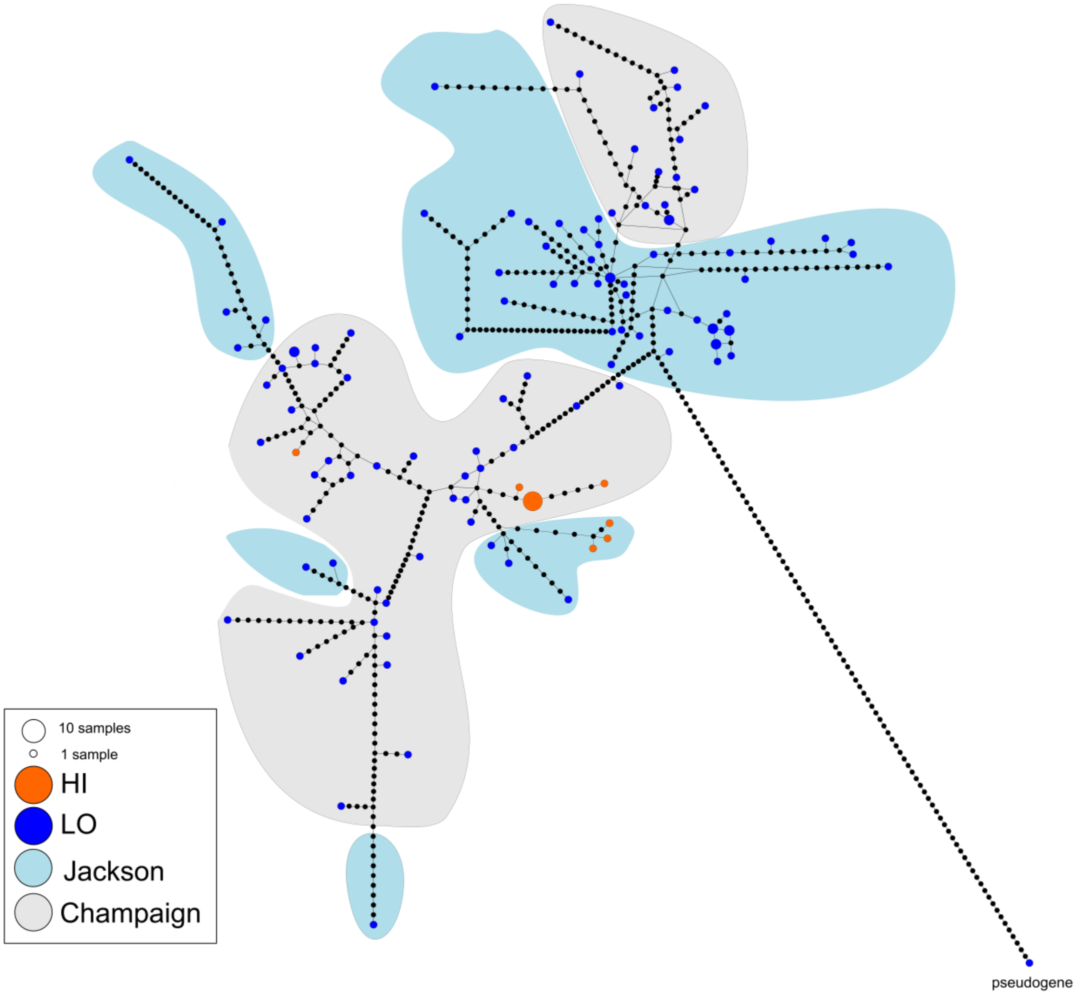


Figure 2.5: *Avpr1a* haplotype network. Median-joining network of the *Avpr1a* locus. Each cluster is color coded to display the original population of the inferred haplotype. Gray represents Champaign County and light blue represents Jackson County. Orange circles represent HI haplotypes and darker blue represents LO haplotypes. Black dots represent single mutations separating observed haplotypes. The pseudogene sequence was used as "outgroup".

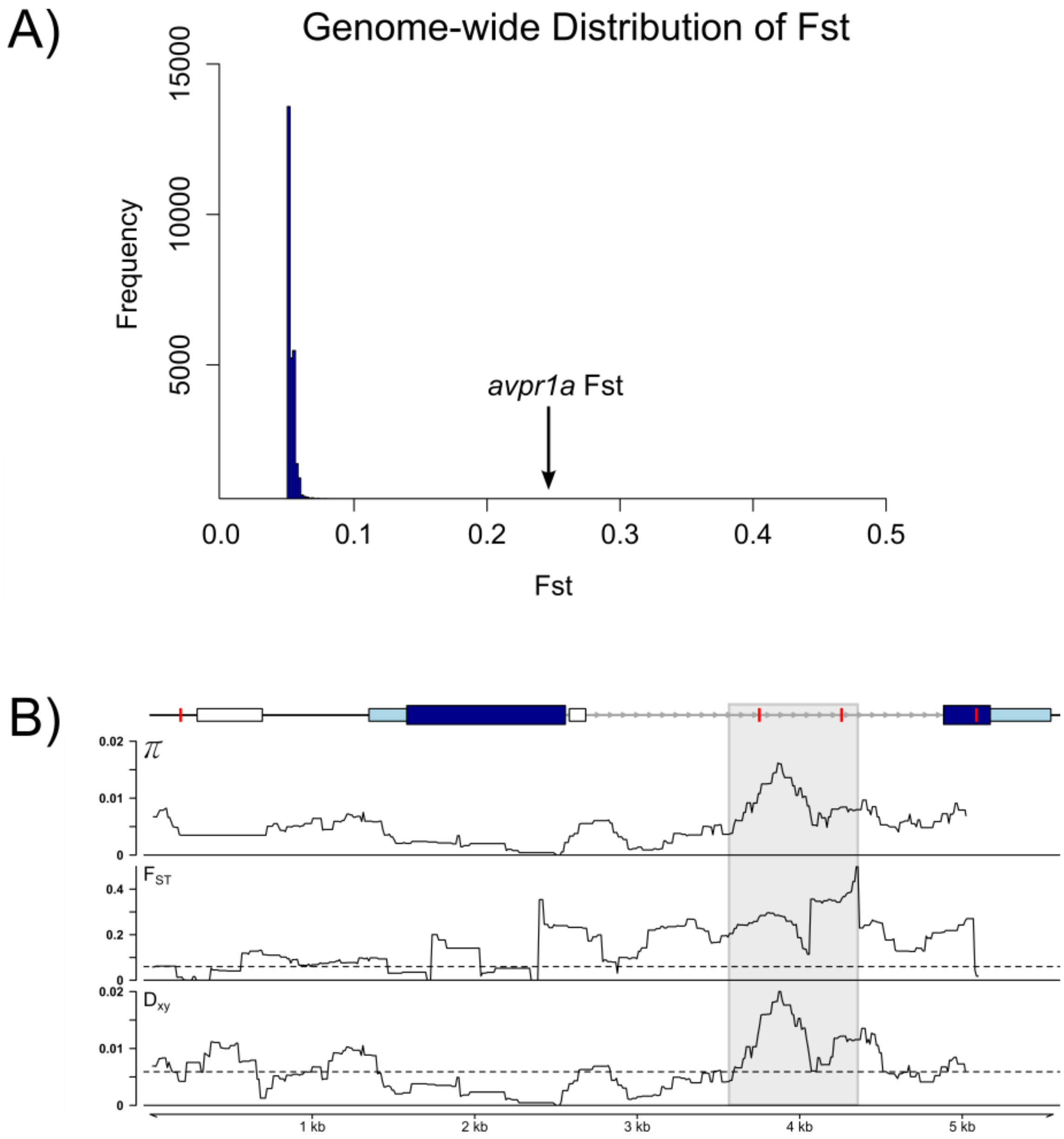


Figure 2.6: Patterns of population differentiation in the *Avpr1a* locus compared to genome-wide. A. Distribution of Fst between Champaign and Jackson County populations per site across the prairie vole genome. B. Sliding-window analysis of nucleotide diversity *top;* relative differentiation (Fst) *middle*; absolute differentiation

(D_{xy}) *bottom*, of the two populations. Location of the putative intron enhancer is highlighted in gray, and RSC-associated SNPs in red (Okhovat et al. 2015).

CHAPTER 3

The transcriptome of male bonding

ABSTRACT

Prairie voles are relatively unique among mammals, in that males contribute to parental care and form enduring pairbonds with their mates. The vasopressin 1a receptor in the ventral pallidum has been the best known modulator of male attachment, but pairbonding requires a much broader set of molecular and cellular mechanisms that are not well understood. Here we employed next-generation sequencing in order to examine the patterns and dynamics of gene expression underlying the onset of male pairbonding in the prairie voles in contrast to meadow voles. We investigated three brain regions involved in the reward and limbic system in response to 30 minutes, 2 hours and 12 hours of mating. We identified massive changes in gene expression between these two species across all three brain regions. Time-sensitive differential gene expression was the highest for genes involved with neuroplasticity in the prairie voles but not in the meadow voles. Comparisons across gene networks in these species indicate high module conservation and preservation, suggesting that most of the cognitive differences between species may have a regulatory basis. This is the first study that provides a wide-ranging list of novel candidate genes for pair bonding. Results of this study may contribute to better understanding of social evolution in mammals but also will provide insights social attachment and its mechanisms in humans.

INTRODUCTION

My dissertation thus far has focused on individual and species differences in the regulation of the vasopressin 1a receptor. However, both intraspecific and interspecific variation in social behavior is likely to rely on a much larger set of genes. Species differences in gene expression are often attributed to changes in gene regulation and transcription factor binding (Garfield & Wray 2010; Wittkopp 2010; Romero et al. 2012). Meanwhile, individual differences are likely to include both genetic differences and individual responses to environment and experience. Changes in social behavior have been linked to extensive shifts in brain gene expression (Robinson et al. 2008; Chandrasekaran et al. 2011). Investigating species differences in gene regulation can inform our understanding of behavioral diversity across species, populations, individuals, and ontogenies.

Progress in genomics and sequencing has revolutionized studies in evolutionary ecology as thousands of genetic markers and genes can be examined thoroughly in non-model organisms. Many recent studies have disentangled the life histories of animals in their natural environments. For example, Dixon *et al* (2015) identified that some corals trigger adaptive responses in gene regulation to changes in heat by combining RAD sequencing with gene expression profiles. Developmental wing patterns of butterflies can be predicted using gene expression profiles (Hines et al. 2012; Connahs et al. 2016). Wang and collaborators identified multiple gene regulatory networks that control limb development in bat wing formation (Wang et al., 2014). In terms of behavior, many studies have explored the influence of gene expression in different behavioral states at the species or individual level. Changes in gene regulation have been linked to behavioral shifts in honey bees (Whitfield et al. 2003), swordtails (Cummings et al. 2008), sticklebacks (Sanogo et al. 2012), and primates (Nowick et al. 2009; Runcie et al. 2013). Despite this progress, however, we know little about the transcriptome-level changes associated with pairbond formation. In the current study, we examine changes in gene expression that accompany pairbond formation in the socially monogamous male prairie vole, *Microtus ochrogaster*, and compare it to patterns of gene expression in the promiscuous male meadow vole, *Microtus pennsylvanicus*.

Prairie voles are well known for their ability to form enduring pairbonds in response to mating. After sexually naïve prairie voles are paired with one another, an extended bout of mating ensures (Getz, Carter, and Gavish 1981). The onset of monogamous behaviors includes pair-bonding and selective aggression (Getz, McGuire, and Pizzuto 1993). The formation of a pairbond requires prolonged mating – 6 hours of mating, for example, is generally insufficient to elicit an enduring bond, but 24 hours elicits its reliably (Insel et al. 1995). This requirement is thought to enable a male to assess whether he can effectively monopolize a female, while it allows a female to assess whether a male is able to effectively displace potentially infanticidal intruders (Wolff et al. 2002; Phelps & Ophir 2009). Yet, we don’t know the neurogenetic mechanisms that occur during pairbond formation that could explain how prairie vole males identify their mating partners. Since mating rewards the rodent brain, it has been hypothesized that paibonding is a consequence of conditioned reward learning where olfactory and sexual signatures form strong associations in the mesolimbic circuit (Insel & Young 2001; Young & Wang 2004).

From a more reductionist perspective, the formation of a pairbond depends on the activity of multiple molecular and neurobiological mechanisms in response to specific social and environmental cues. The formation of pair-bonds in prairie voles seems to be mediated in part by specific reward circuits in the brain (Young & Wang 2004). Dopaminergic neurons projecting from the ventral tegmental area (VTA) are thought to contribute to pair-bond formation by altering reward in response to sexual stimulation (Liu, Curtis, and Wang 2001), and by linking behavioral reward to sensory inputs arriving from the amygdala (Keshavarzi et al. 2015). Furthermore, genital stimulation and the release of pheromonal and olfactory signals during mating initiate an active discharge of vasopressin and oxytocin from the hypothalamus (Gobrogge et al. 2009), two major modulators of social behavior that play a central role in pairbond formation. In addition to the neuropeptides oxytocin and vasopressin, other neuromodulatory systems such as opioids (Resendez et al. 2016), sex steroids (Cushing et al. 2001) and stress hormones (Bales, Kramer, Lewis-Reese, & Carter, 2006; Lim et al., 2007) have all been implicated in the regulation of bonding. These diverse molecular systems have all been shown to act in reward regions, the hypothalamus or the amygdala. Indeed, these brain regions have been implicated in a tremendous variety of reproductive and social behaviors (Newman 1999; Goodson 2005; O’Connell & Hofmann 2012), and have been described as parts of a larger “social decision-making circuit” (O’Connell & Hofmann 2012). A thorough understanding of the role of gene expression in pairbonding will require the investigation of transcriptome profiles across a variety of brain regions and species.

To better understand the genomic mechanisms that govern responses to mating in prairie voles, we first generated a transcriptome reference assembly of prairie voles. Next we used this assembly to analyze gene expression profiles from males of two species (monogamous prairie voles and promiscuous meadow voles), across three groups of brain regions associated with mating and pair-bond formation. Targeted brain regions include three major groups of related nuclei: 1) adjacent reward regions known as the ventral pallidum and nucleus accumbens (VP/NAcc); 2) the various nuclei of the hypothalamus (HYP), including regions known to be involved in sexual behavior and parental care; and 3) the assorted nuclei of the amygdala (AMYG), a suite of adjacent brain regions critical in emotional learning and a variety of social behaviors (Insel & Young 2001; Ferguson et al. 2002; Young & Wang 2004). We examined transcriptomes of prairie and meadow voles in each of these brain regions through a series of time samplings before and immediately following mating (virgins, or 30 minutes, 2 hours and 12 hours post-mating). With these data we asked whether there are significant differences in gene expression between promiscuous and monogamous voles, whether there are coordinated transcriptional responses to mating, and whether there are specific sets of genes that coordinate responses to mating in one or both species.

MATERIALS AND METHODS

Animals

For the transcriptome assembly, adult tissues were dissected from single, virgin adult male prairie voles and transferred immediately to RNAlater. Animals were derived from the prairie vole colony at Yerkes National Primate Research Center and all the procedures were performed as per guidelines that were reviewed and approved by the Emory Institutional Animal Care and Use Committee and were conducted in accordance with the Guide for Care and Use of Laboratory Animals published by the National Research Council.

For the time-series experiment, prairie voles (Microtus ochrogaster) and meadow voles (Microtus pennsylvanicus) were housed in same-sex groups with two or three voles per cage from postnatal day 21. Housing consisted of a ventilated 36 cm × 18 cm × 19 cm plexiglass cage filled with Bed-o'Cobs laboratory animal bedding (The Andersons Inc., Maumee, Ohio) under a 14/10 hour light/dark cycle (lights on 7:00 AM–9:00 PM) at 22°C with access to food (rabbit diet; LabDiet, St. Louis, Missouri) and water ad libitum. These laboratory breeding colonies were originally derived from field captured voles in Illinois. All procedures were approved by the Emory University Institutional Animal Care and Use Committee.

Tissue Collection

All animals were 60-90 days old and sexually naive at the time of the experiment. Control tissue was collected from the brains of sexually naive males. For time course samples, a single male prairie vole or meadow voles was paired with an unrelated female partner primed with estradiol benzoate (Sigma Aldrich, St. Louis, MO, USA, BP958) for two days prior to testing to ensure sexual receptivity. Pairs were observed and the time of first intromission was recorded. Males were euthanized using isofluorane at 30 min, 2 hours and 12 hours following the first intromission. Whole brains from male voles at 30 minutes, 2 hours, and 12 hours, post mating (virgin males were used as controls) were harvested and dissected on a block of dry ice to remove the ventral pallidum, the hypothalamus, and the amygdala, and were stored in RNAlater (Applied Biosystems AM7020).

Extraction of total RNA and library preparations

Total RNA was extracted from 50-80 mg of tissue using TRIzol (Sigma-Aldrich) following the manufacturer's protocols. An additional DNAse (PureLink) treatment was included to eliminate any contaminating DNA. RNA quality and quantity were assessed on a BioAnalizer and quantified by spectrophotometry. cDNA was prepared using Superscript reverse transcriptase (Invitrogen) and purified with QIAquick PCR purification kit. After checking cDNA quantity and quality in the Bioanalyser (Agilent), libraries were prepared using Illumina's TrueSeq Sample Prep Kit with a starting amount of 1.25ug of RNA. Then, libraries were normalized using the Evrogen Trimmer kit (Evrogen).

Transcriptome assembly and gene annotation

In order to assemble a prairie vole transcriptome, we used Illumina short read libraries derived from male gonads (kindly provided by Dr Lisa McGraw and Dr Larry Young from Emory University), whole brains, and multiple SRA files downloaded from

NCBI (SRA Accession codes in Table 3.1). After filtering and eliminating adapters using fastq and cutadapt tools, we mapped the processed reads into the *Microtus ochrogaster* genome (http://www.broadinstitute.org/software/allpaths-lg/blog/?p=618) using TopHat2 (Kim et al. 2013). A pseudogenized duplication kept the extended region around the *Avpr1a* locus from assembly into the prairie vole genome). Therefore, BAC contigs containing the locus and its pseudogene were manually added to the assembly (NCBI accessions: DP001225, HQ156469). The outputs from each library were merged and the resulting BAM file was sorted and indexed. The final transcriptome was assembled with the genome-guided program StringTie (Pertea et al. 2015). To annotate the transcriptome, we aligned the resulting fasta file to SwissProt database using blastx and extracted the uniprot annotation files using custom scripts written by Dr Mikhail Matz. KOG and KEGG annotations were retrieved from online databases by submitting the transcriptome in fasta format to the following websites with default settings: for KOG annotations, (http://weizhong-lab.ucsd.edu/metagenomic-analysis/server/kog/) and KEGG (http://www.genome.jp/kegg/kaas/).

Differential expression sequencing

To identify differences in gene expression of prairie vs meadow vole, RNA was collected following mating for three brain regions – hypothalamus, amygdala and ventral pallidum/nucleus accumbens. RNA was isolated, reverse transcribed with Superscript reverse transcriptase II (Invitrogen), and prepared for Illumina sequencing using Illumina's TrueSeq Sample Prep Kit. The samples were then delivered to the Emory University genomics core facility and sequenced using the Illumina HISEQ 50 bp single ends. All sequence reads of this experiment were kindly provided by our collaborators, Drs Lisa McGraw and Larry Young.

Sequence files were trimmed and cleaned fastq tools and mapped to our prairie vole reference-transcriptome using BWA (Li & Durbin 2009). First, the whole dataset was analyzed using the R library arrayQualityMetrics to identify outliers and sample quality. 6 outliers were removed, resulting in sample sizes of 44 prairie and 45 meadow vole samples. The counts file was then downloaded and analyzed using the library DESeq2 in R (Love et al. 2014). We used the full design [~ Species * Brain-Region * Time] in order to identify the effects of species, brain region, time and all the interactions to explain the variation in gene expression. Principal component analysis (PCA), principal coordinate analysis (PCOA) and permutational multivariate analysis of variance (permanova) tests using the Manhattan method and 10,000 permutations were executed to identify the main factors explaining the variation in gene expression of this experiment.

To facilitate the analysis of time effects, we parsed our dataset into five subsets. More specifically, we split our reads by brain region to focus on species differences within a brain region, and split our reads by species sets to identify tissue differences within a species. Here, outlier analysis excluded two individuals, one for the ventral pallidum/NAcc data set and the other outlier from the hypothalamus data set. We used [~ Species + Time + Species:Time] as a full design for each brain region, and [~ Brain-region + Time + Brain-region:Time] for each species. Comparisons were made between species and across brain regions, having meadow voles, hypothalamus and virgins (time 0) as reference for each contrast. To retrieve differentially expressed genes and their gene ontology (GO) enrichment, we filtered our RNA-seq data with a False Discovery Rate (FDR) q-value<0.1 for each contrast and constrained to include only differentially expressed genes (DEGs) with at least two-fold difference.

Functional annotation analysis

To identify gene ontology enrichment for molecular functions (MF), cellular components (CC) and biological processes (BP), we used the Gene Ontology Mann-Whitney Test (GOMWU) script by Misha Matz.

Gene co-expression network analysis

To identify groups of co-regulated genes that were differentially expressed between prairie and meadow voles in response to mating, we used a weighted gene correlation network analysis (WGCNA) (Langfelder & Horvath 2008). This method permits detecting biologically meaningful groups of co-regulated genes. All the genes identified by the DESeq2 analysis were used in our WGCNA. Three unsupervised networks were constructed for species differences in the ventral pallidum/nucleus accumbens, amygdala and hypothalamus with no outlier samples identified (sample distance PCOA confirms this finding (Fig 3.2)). Soft-power thresholds for ventral pallidum/nucleus accumbens, amygdala and hypothalamus were 22, 20 and 18 respectively. Genes inside a co-expression module tend to have high levels of connectivity; therefore, a mathematical construct of the principal component of gene expression of all the genes in that module can be summarized as an "eigengene" (Langfelder & Horvath 2008). The membership of each gene in the module can be correlated with its gene significance (GS) from a specific trait-category if available. GS can be defined as the absolute value of the correlation between genes and the trait-category of interest (Langfelder & Horvath 2008). Here we correlated groups of co-expressed genes with our categories of interest defined as prairie or meadow voles, and these at specific times (i.e. premating (virgin state), 30 minutes, 2 hours, 12 hours, and post-mating). Groups of co-expressed genes that were strongly correlated with a specific species or with specific times (i.e. before and after mating) were further analyzed for GO enrichment. To confirm that the strongest module eigengene correlations corresponded to real biological effects, we shuffled the time-condition designations among samples to confirm the observed correlations disappeared.

Preservation analysis

We also used preservation analysis as an alternative and independent test to validate the preservation or conservation of specific network modules between species (Langfelder et al. 2011). To assess preservation of prairie vole modules in the meadow vole brain network, a gene co-expression network analysis was performed separately on both species subsets, designating the prairie vole network as reference. The functions overlapTable() and modulePreservation() implemented in the WGCNA package for R (Langfelder & Horvath 2008) were used to identify the number of genes that are shared between specific brain modules in both species; and the summary Z statistics reveal preservation as a function of module size, modules at specific Z thresholds indicate no preservation if Z-summary < 2, weak to moderate evidence of preservation if 2<Z-summary<10), and strong evidence of preservation if Z-summary>10.

RESULTS

Assembly and annotation of the prairie vole reference transcriptome

A genome-guided reference transcriptome was produced for prairie voles (Table 3.2). The transcriptome assembly generated a total of 79609 prairie vole individual contigs, of which ~28K were annotated using Uniprot, SwissProt, KEGG, GO and KOG categories. The average length of the contigs was 1536 bp with an N50 equal to 3083. The combined size of the transcriptome represents 122.3 Mb altogether. 100% of the transcripts are covered in KOG database. The transcriptome contiguity (Martin & Wang 2011), a measure of the proportion of contigs covering the uniprot reference at 0.75-threshold, is equal to 0.53.

Differential gene expression analysis

We generated expression data by mapping RNA-seq reads to our prairie vole reference transcriptome, and found that >85% of the reads from both species aligned to the reference transcriptome. After removing outlier samples, the average number of raw read counts per sample for prairie voles ranged from 21975910 and 41140809, and for meadow voles it ranged between 20934671 to 41851176; indicating little mapping bias. We identified 22002 annotated genes that were differentially expressed between species and across all three brain regions (Fig 3.1A). Of these, 16938 were differentially expressed between vole species within the ventral pallidum/nucleus accumbens, 15892 in the amygdala, and 16714 in the hypothalamus (Fig 3.1B). Of these genes, 8656 were significantly over-expressed genes in the prairie vole ventral pallidum, 8248 in the amygdala, and 8503 in the hypothalamus with an adjusted p-value <0.1 in the species contrast result.

A permutational multivariate analysis of variance (permanova) on normalized and variance stabilized read count data showed that the main factors were species (R^2=0.34, P <10^{-5}), followed by brain region (R^2=0.21, P<10^{-5}), time (R^2=0.04, P =0.004), and species-by-time interaction (R^2=0.03, P= 0.017). The principal component analysis function implemented in the R package DESeq2 provided similar but overestimated patterns, 68% of the variation in normalized gene expression is explained by species, and 14% is explained by brain region (Fig 3.1C). Principal coordinate analysis of the variance-stabilized data confirmed these observations and also revealed that a small fraction of expression differences seem to be explained by time effects, as virgins seemed to be more distant to post-mated samples in the prairie than in the meadow voles (Fig 3.1D).

Gene expression differences over time

We identified a total of 2535 prairie vole genes, and 484 genes in the meadow vole genes, that differed in gene expression during the first 30 minutes of mating among brain regions (Fig. 3.2A). Particularly, the hypothalamus of prairie voles maintains a massive increase of differentially expressed genes after mating respect to only 12 genes that are differentially expressed in meadow voles for the same time contrasts (Fig 3.2A). Interestingly, some of the most significant differences correspond to early immediate genes such as FosB, which increase in expression after 30 minutes of mating in prairie and meadow voles (Fig. 3.2B).

Species differences in gene expression over time

Heatmaps of clustered gene expression revealed several genes that exhibited species by time effects. The most significant (FDR <0.1) genes for the ventral pallidum were 12 (Fig 3.3B), for the amygdala were 25 (Fig 3.4B), and for the hypothalamus were 35 (Fig 3.5B). With a log2 fold change difference >2 and a p-adjusted <0.1, the most strongly up-regulated gene in the ventral pallidum of prairie vole samples that also showed significant difference after 12 hours of mating was SMARCAL1 which is known to stabilize DNA topology (Fig 3.3C). In the amygdala was Discs5 that may play roles in the maintenance of cell structure and the transmission of extracellular signals to the membrane and the cytoskeleton (Fig 3.4C). And, C2CD3 in the hypothalamus, this gene is a component of centriole elongation (Fig 3.5C). Interestingly, the hypothalamus was the only brain region that contained under-regulated genes; the most significant was plasminogen activator inhibitor-2 (PAI2), which is known to inhibit endocellular proteases.

Species differences in gene ontology enrichment

Among all genes that pass the false discovery rate of 0.1, an additional fraction showed more than a twofold difference in expression differences between species in the ventral pallidum (1463 genes, Fig. 3.6B), amygdala (1365 genes, Fig. 3.7B), and hypothalamus (1413 genes, Fig. 3.8B). Moreover, species differences among all brain regions revealed very similar gene ontology (GO) categories of overregulated molecular functions (MF), biological processes (BP), and cellular compartments (CC) across brain regions. Structural constituents of muscle, axoneme, chromatin modification, and cell projection organization are upregulated functions in prairie voles in contrast to meadow voles among all brain regions. In contrast, mitochondrial and proteasome complex genes are the cellular components that are more downregulated in the prairie voles (Fig. 3.6C-3.8C).Similarly, endopeptidases, endnucleases and microtubule motor GO terms are the molecular functions that are more expressed among brain regions in the prairie voles. In the other hand, poly(A)-RNA binding, endoplasmic reticulum, oxidoreductase, and GTPase binding are some of the GO categories for molecular function that are downregulated among prairie vole brain regions.

Network analysis of co-expression

To determine the patterns of co-regulated genes that underlie species differences among three brain regions, the nearly 28K genes normalized genes were entered into WGCNA analysis for co-expression analysis. No additional outlier samples were identified after pre-adjacency analysis. Then, we identified the correlated modules with either prairie voles or post-mating specific modules on each brain network. WGCNA analysis revealed 14 modules in the ventral pallidum/nucleus accumbens. Of these, the brown module was strongly correlated with overexpression in the ventral pallidum of prairie voles (corr=0.97, P=2x10^{-20}), the dark-magenta module was strongly correlated with downregulation in the prairie vole ventral pallidum (corr=-0.99, P=1x10^{-26}), and the bisque4 module was highly correlated and overexpressed in mated individuals (corr=0.65, P=7x10^{-5}). 15 modules were found in the amygdala, of which the salmon module was highly correlated with overexpression in prairie voles (corr=0.95, P=3x10^{-17}), the brown was correlated with downregulation in prairie voles (corr=-0.99, P=2x10^{-}

^{25}), and antiquewhite4 was associated with postmated animals (corr=0.65, P=6x10^{-5}). Finally, of 12 modules found in the hypothalamus, the turquoise module showed high correlation values with prairie voles (corr=0.83, P=2x10^{-8}), the antiquewhite4 module was correlated with under-expressed genes in the prairie vole hypothalamus (corr=-0.89, P=3x10^{-11}), and the darkolivegreen module was strongly correlated with postmated individuals (corr=0.81, P=4x10^{-8}). Interestingly all this significantly correlated modules in the prairie voles showed completely opposite directions in the meadow voles (Fig 3.9A-C). And additional GO analysis of these modules confirmed findings from differences in gene expression in DESeq2 (Fig 3.9D-F).

Preservation analysis

In this section, we compared module assignations for the prairie and meadow voles using overlap analysis. A subset of 8000 genes were filtered by time effects (p-adjusted <0.1) from the prairie voles and meadow voles and imported to WGCNA analysis. We observed that almost all the modules in the prairie vole overlap with meadow vole modules, including the grey module – a module where all the unassigned genes are classified– which overlapped with 32% of unassigned genes in the meadow vole. This implicates that modules are preserved between species; despite they can be split in two or more interspecific modules. We found relatively small grey modules for both species networks, the largest prairie vole module (i.e. darkmagenta) also overlapped the major fraction of grey meadow vole eigengenes (7%). This module also contained 89% of blue module genes, suggesting that these may play similar functions in both species and that there are not really unique modules for each of the species (Fig 3.10).

DISCUSSION

The social brain relies on interactions between the hypothalamus, the amygdala and other brain reward regions. Pair-bond formation is a process that involves sexual stimulation, olfactory input and social interactions. Oxytocin and vasopressin are the best known drivers of these processes but we know little about the genome-wide and cellular level processes driving the formation of pairbonds. Transcriptome analyses of the prairie vole brain are necessary for illustrating the neurogenetic topology of molecular functions underlying social behaviors. This scenario can be achieved by finding candidate genes, molecular pathways and gene co-expression networks in the context of species, brain regions, and time during the onset of pair-bonding. Here, we report an updated and annotated reference transcriptome assembly of prairie voles (Table 3.2), and a functional analysis of the effect of mating in the brain of monogamous and promiscuous male voles.

Our clustering analysis revealed that variation in gene expression among all samples in this experiment are explained mainly by species and then brain region (Fig 3.1). Surprisingly, the differences in gene expression are massive between species, while time effects seem subtle. A plausible explanation of this result may be caused by lower number of premating samples in the experiment, therefore We think this finding suggests a different neurogenomic states in the prairie voles in contrast to meadow voles during the formation of pair bonds and selective aggression (Cirelli et al. 2004; Toth & Robinson 2010; Chandrasekaran et al. 2011). However, the effect of mating and time seems to be only slightly more important in prairie voles (Fig 3.1D).

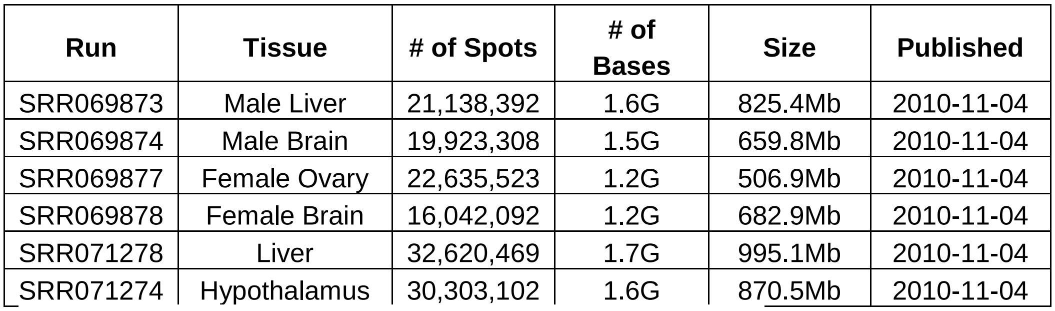
Mating is an adaptive behavior; therefore the mechanisms of sexual reward must be conserved across taxa. But species differences in postmating behavior should be explained by differences in neural gene expression. Interestingly, we found evidence that mating causes the expression of early immediate gene (e.g. *FosB*) in both species during the first two hours of mating (Fig 3.2). Based on the fact that the species factor explained more that 50% of differences in gene expression, we think that these early immediate genes and transcription factors activate differential cascades of downstream genes in these two species.

Consistent with the hypothesis that pairbonding formation is mediated by learning in the prairie vole brain (Insel & Young 2001; Young & Wang 2004), we expected that prairie voles activate genes that are necessary for formation of memories related to a partner while prairie voles during the onset of pairbonding and selective aggression. Contrastingly, meadow voles would activate genes that are essential in the maintenance of homeostasis in brain after the consequences of sex. Indeed, we observed massive gene expression differences between prairie and meadow among all brain regions. Nevertheless, some few differentially expressed genes were indicative of postmating differences between both species The upregulation of *Smarcal1* in the ventral pallidum of prairie voles may suggest higher activity of neuron-glial migration as mating progresses over time (Fig 3.3C). This gene encodes a protein that participates in chromatin remodeling. *Smarcal1* has been identified in cerebrovascular disease; mutations in this gene have also been associated with microcephaly and social, language, motor, or cognitive abnormalities (Deguchi et al. 2008). The overexpression of this gene suggests that structural changes in ventral pallidum and glial cell migration may contribute to pair-bond formation by enhancing aspects of reward and motivation. Disc large homologue 5, (*Dlg5*) is the most upregulated gene in the prairie vole amygdala after 12 hours of mating; this gene encodes a protein that facilitates interactions between the plasma membrane and the cytoskeleton, facilitating cell migration, adhesion and proliferation (Liu et al. 2014). Moreover, other members of the discs-homolog family have been involved with post-synaptic stabilization and regulation in the adult rat brain (Cho et al. 1992). Finally, the C2 calcium dependent domain containing 3 (*C2cd3*) is upregulated in the prairie vole hypothalamus, the protein encoded by this gene regulates centriole elongation and has been linked to microcephaly and facial deformities (Thauvin-Robinet et al. 2014). The upregulation of these genes associated with active regulation of cell shape, synaptic formation and maintenance across all brain regions of the prairie voles, suggests that mating activates reward-dependent neural plasticity. We think this process favors the formation of partner preference memory and pair bond formation in prairie voles but not in the meadow voles.

Contrasts of gene expression differences between species across different regions revealed thousands of genes that exhibit differences in gene expression between prairie and meadow voles (Fig 3.6-3.8). Neuroplasticity at both synaptic and structural level is known to be the primary basis of learning and memory formation. Learning involves a variety of structural changes mediated in part by reorganizing the actin cytoskeleton (Kneussel & Wagner 2013). The myosin II complex is an important regulator of this mechanism (Rex et al. 2010), and these genes have also been involved in the onset of human neuronal disorders such as autism and schizophrenia (Newell-Litwa et al. 2015). Interestingly, prairie voles upregulate genes related to axoneme elongation, cell migration, myosin-related processes, and microtubule dynamics, while the meadow vole brain is upregulating genes that maintain mitochondrial homeostasis. In addition, molegular function GO terms alsp revealed stronger endonuclease and endopeptidase activity in the prairie vole brain that may contribute to both RNA and protein metabolism. More specifically, the upregulation of aspartic-type endopeptidase causes the hydrolysis of internal alpha-peptide bonds. This is, however, a novel finding despite other endonucleases have been linked with learning, memory formation and neurodegenerative disorders (D'Agostino et al. 2013; Schneider et al. 2002; Walther et al. 2009). The activation of these gene ontology categories further reinforces the idea that learning and neural plasticity are possible mechanisms shaping the formation of partner memories in the prairie vole, therefore, genes within these categories are new candidates mediating the onset of pairbonding and selective aggression.

Furthermore, our WGNCA analysis provided an additional validation that our findings on species differences are driven by synaptic plasticity processes in the prairie vole brains rather than stochastic effects due to experimental design. Co-expression modules are thought to be blind to how the experiment of gene expression has been designed. Therefore, if condition-correlated modules are enriched by similar gene ontology categories, we can ascertain the biological meaning of our data. Indeed, the most highly correlated modules in the prairie vole ventral pallidum (brown module), amygdala (salmon module) and hypothalamus (darkolivegreen module) revealed that myosin II complex and endopeptidase genes are enriched and co-regulated among all these brain regions respect to meadow voles (Fig 3.9D-F). Interestingly all these modules exhibited opposite trends between species, while co-regulated genes in the ventral pallidum/NAcc (brown module) are overregulated in the prairie voles they are down-regulated in the meadow voles, for example. This observation suggests that the modules within the brain networks are highly conserved and preserved between these two species, but their regulation is completely different. To further investigate this observation, we analyzed differential gene expression by parsing our data set in two independent WGCNA analyses. Then we overlapped the resulting networks and found that indeed, each module eigengene identity is well conserved in both species despite some modules of the prairie vole are split in the meadow voles and vice-versa (Fig 3.10).

Overall, these data suggest that species differences in gene expression are pervasive across all three of the brain regions we examined. Expression differences seemed to be particularly enriched for genes related to the structural demands of neuroplasticity; this seemed particularly true for the ventral pallidum and nucleus accumbens, a pair of reward-related regions known to be critical to learning, memory and pair-bond formation. Because our experimental design emphasized species differences over the course of pair-bond formation, it is perhaps not surprising that plasticity-related genes would be most over-represented among species differences. Our data suggest that a relatively small number of pre-mating species differences in gene expression may drive massive differences in gene expression in response to mating. This suggests how a relatively small number of loci may have effects that become magnified by their ability to recruit major networks of genes needed to execute a behavior. This may have major consequences for our understanding of species and individual differences in brain and behavior.



Run	Tissue	# of Spots	# of Bases	Size	Published
SRR069873	Male Liver	21,138,392	1.6G	825.4Mb	2010-11-04
SRR069874	Male Brain	19,923,308	1.5G	659.8Mb	2010-11-04
SRR069877	Female Ovary	22,635,523	1.2G	506.9Mb	2010-11-04
SRR069878	Female Brain	16,042,092	1.2G	682.9Mb	2010-11-04
SRR071278	Liver	32,620,469	1.7G	995.1Mb	2010-11-04
SRR071274	Hypothalamus	30,303,102	1.6G	870.5Mb	2010-11-04

Table 3.1. SRA accessions for raw transcriptome libraries

Assembly details	Summary statistics
Total number of reads	443’528.283 reads
Transcriptome total size	122.3 Mb altogether (122269988 bp)
Maximum contig length	41682 bp
Average contig length	1536 bp
N50	3083 bp
Total number of contigs>350bp	79609
Number of ambiguities	0
Number of Ns	0
Contiguity at .75 threshold	0.53
KOG completeness	100%
Number of Uniprot hits	28462
Number of GO annotated contigs	28869
Number of KOG annotated contigs	16600
Number of KEGG annotated contigs	18436

Table 3.2: Summary of genome-guided transcriptome assembly

FIGURES

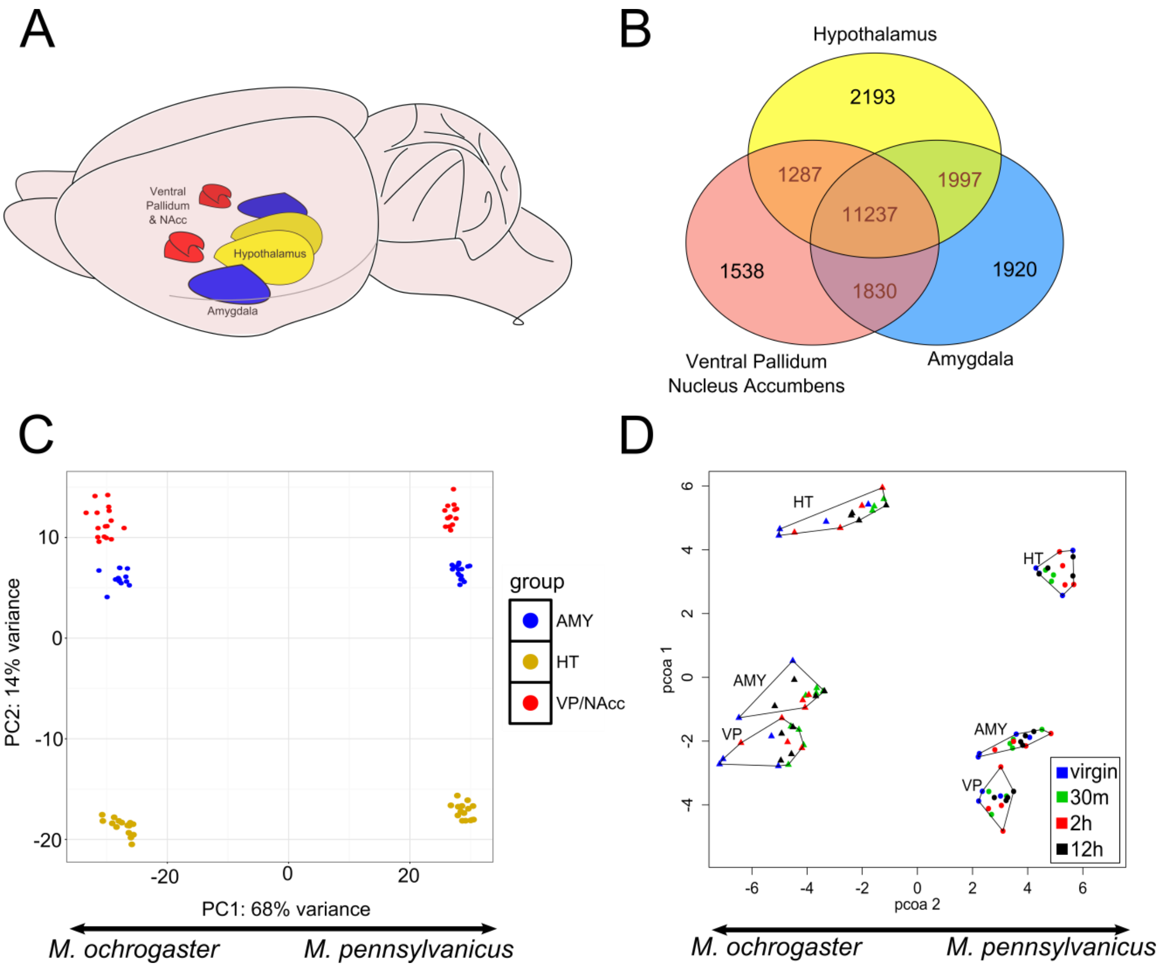


Figure 3.1: Species and brain region differences in gene expression. A. The vole brain representing components of the reward and limbic system: ventral pallidum/nucleus accumbens (red), amygdala (blue) and hypothalamus (yellow). B. Venn diagram representing the number of differentially expressed genes in the species contrast across brain regions. C. Principal component analysis shows clusters by species and brain region. D. PCOA distances between prairie and meadow vole samples.

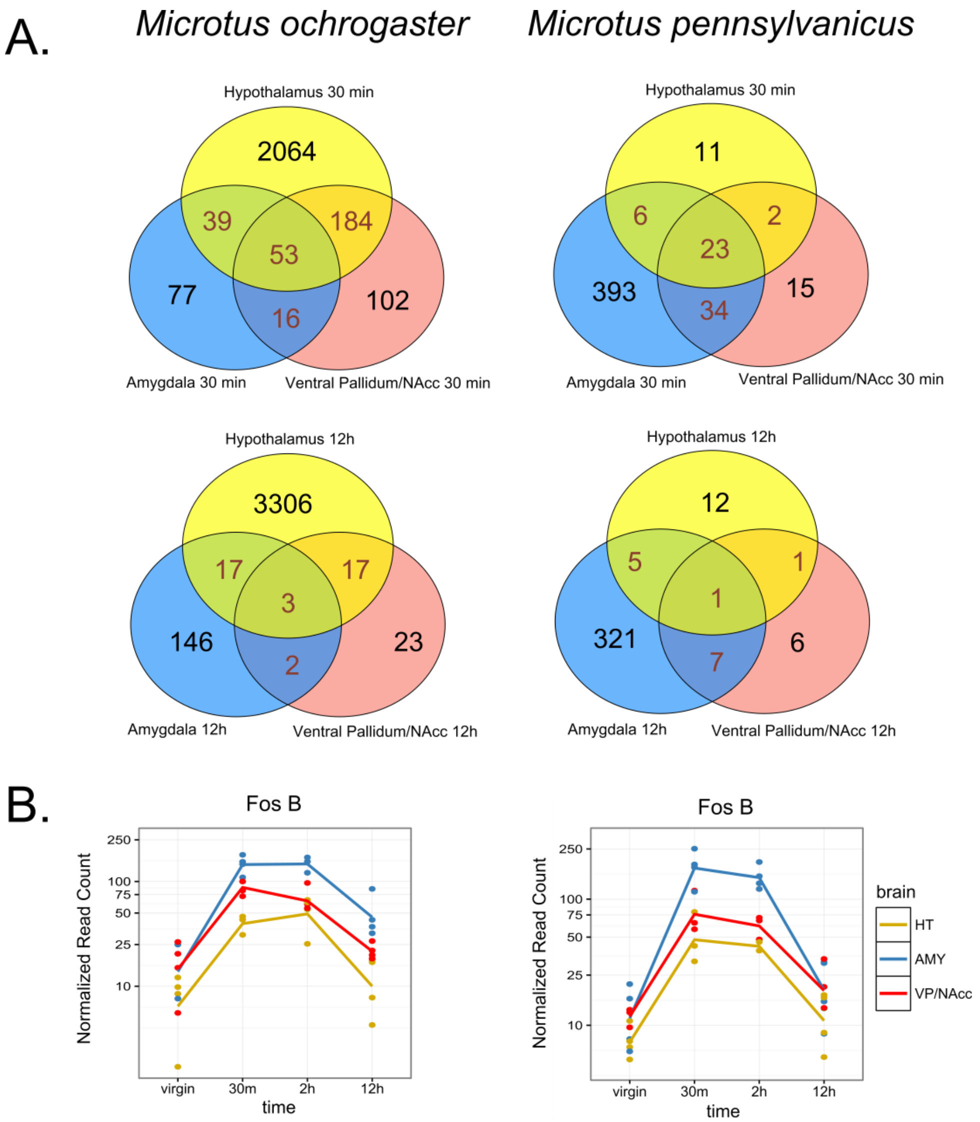


Figure 3.2: Differentially expressed genes over time. A. Venn diagrams representing the number of differentially expressed genes after 30 minutes and 12 hours of mating in contrast to virgins in the prairie voles (left) and meadow voles (right). **B.** Gene expression changes of the most upregulated (Log2FoldChange>2) gene in the ventral pallidum, amygdala and hypothalamus of prairie voles (left) and meadow voles (right).

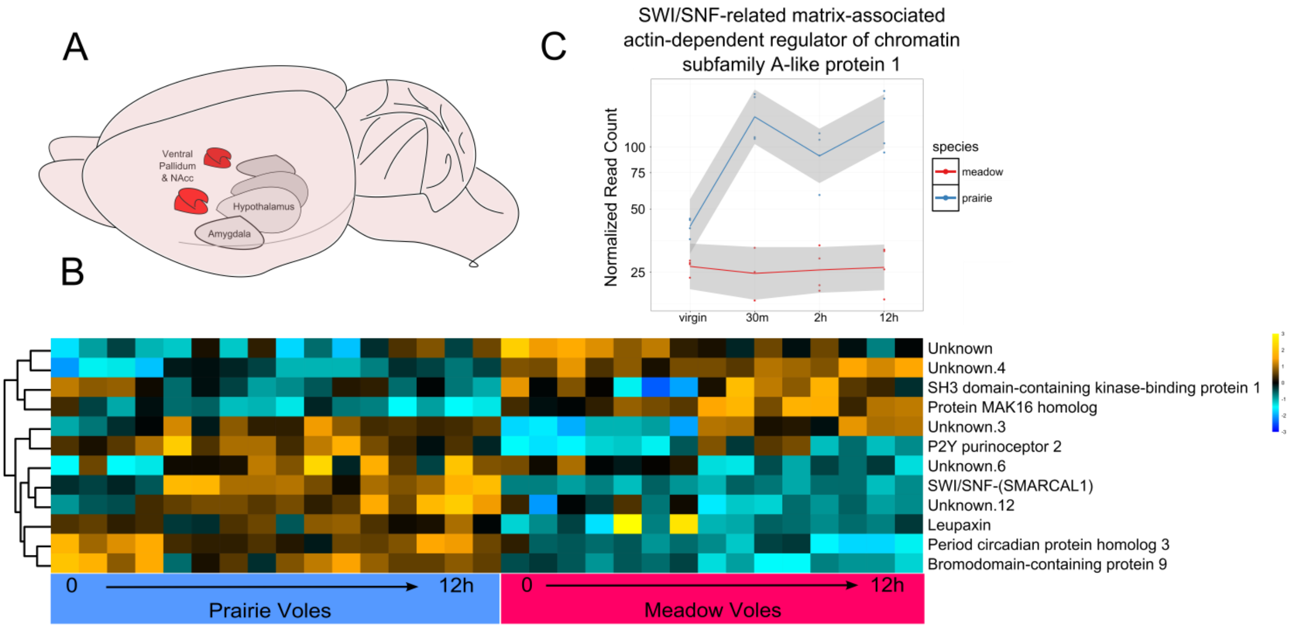


Figure 3.3: Species differences in gene expression over time in the ventral pallidum/nucleus accumbens. A. View of the vole brain illustrating the position of the ventral pallidum and nucleus accumbens in red. B. Heatmap of differentially expressed genes (passing a FDR-adjusted p<0.1,) between species during 12 hours of mating. The rows correspond to clustering between significant genes and the color scale represents the relative change to the mean across all samples. C. Gene expression changes of the most upregulated (Log2FoldChange>2) gene in the prairie vole ventral pallidum/nucleus accumbens of prairie vs meadow voles.

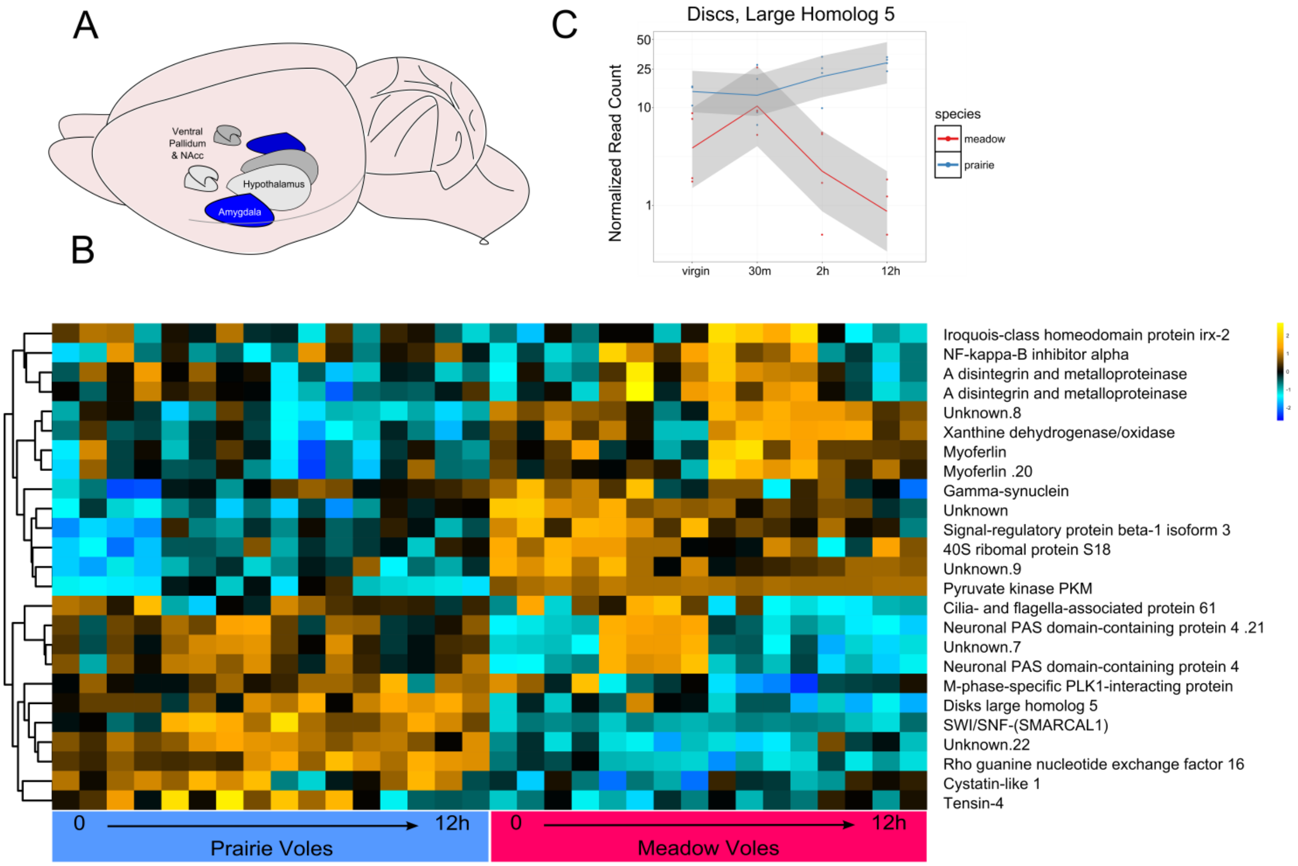


Figure 3.4: Species differences in gene expression over time in the amygdala. A. View of the vole brain illustrating the position of the amygdala in blue. B. Heatmap of differentially expressed genes (passing a FDR-adjusted p<0.1,) between species during 12 hours of mating. The rows correspond to clustering between significant genes and the color scale represents the relative change to the mean across all samples. C. Gene expression changes of the most upregulated (Log2FoldChange>2) gene in the prairie vole amygdala of prairie vs meadow voles.

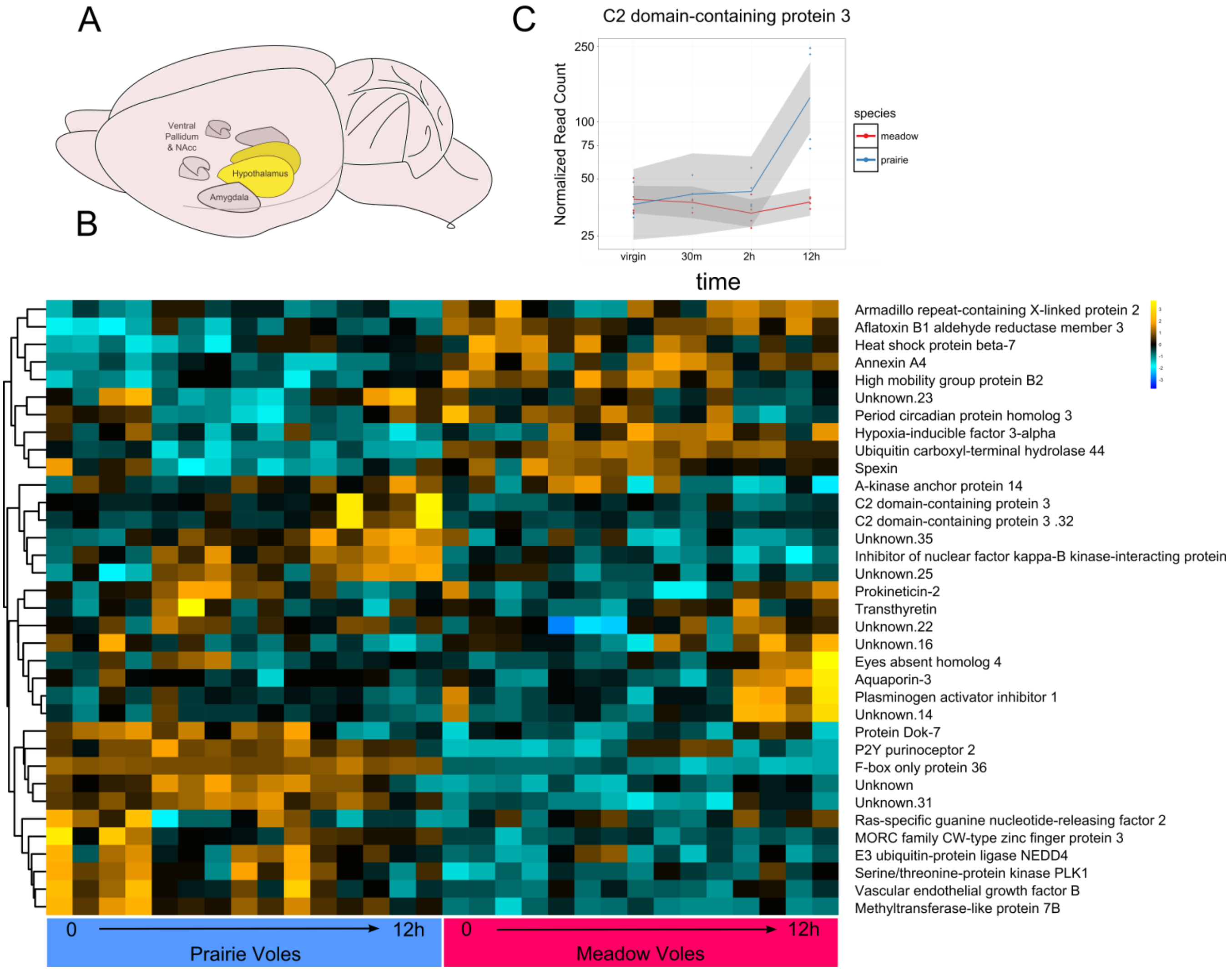


Figure 3.5: Species differences in gene expression over time in the hypothalamus. A. View of the vole brain illustrating the position of the hypothalamus in yellow. B. Heatmap of differentially expressed genes (passing a FDR-adjusted p<0.1,) between species during 12 hours of mating. The rows correspond to clustering between significant genes and the color scale represents the relative change to the mean across all samples. C. Gene expression changes of the most upregulated (Log2FoldChange>2) gene in the prairie vole hypothalamus of prairie vs meadow voles.

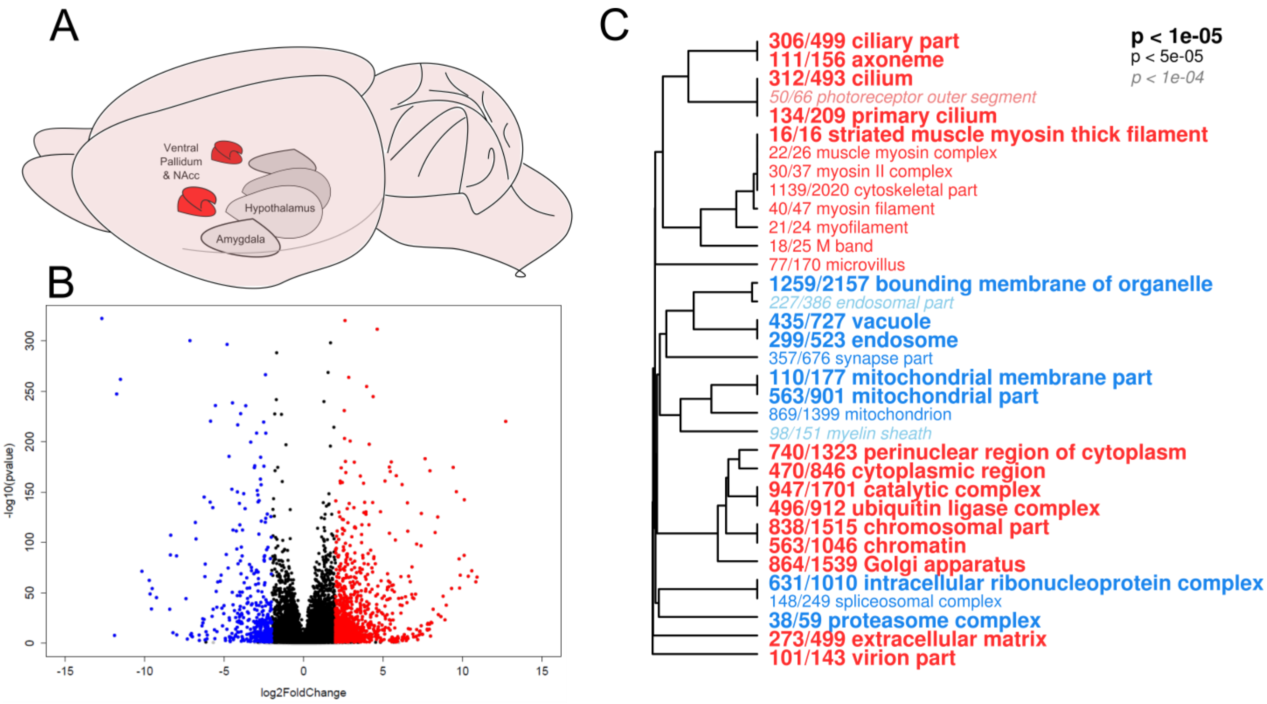


Figure 3.6: Differential gene ontology enrichment between species in the ventral pallidum/nucleus accumbens. A. View of the vole brain illustrating the position of the ventral pallidum and nucleus accumbens in red. B. Volcano plot of differential expression in the ventral pallidum and nucleus accumbens between prairie and meadow voles. The x axis shows normalized read counts log2 fold change between prairie voles and meadow voles. The y axis shows p-value (-log base 10) for differential expression. At a p-adjusted p<0.1, downregulated genes in prairie voles (blue) at a twofold difference lower than -2; and upregulated genes (red) at a twofold difference higher 2. C. Gene ontology enrichment analysis (GOWUA) for cellular component categories, upregulated (red) and downregulated (blue) categories in prairie vole; font size reflects p-value.

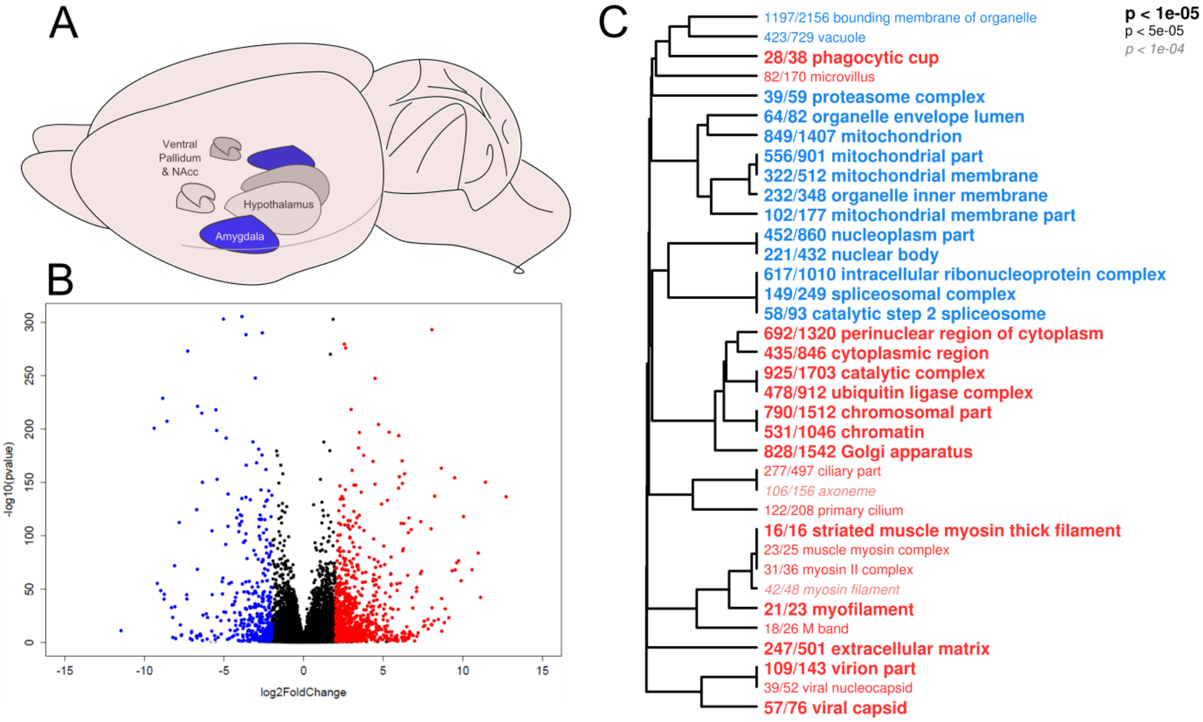


Figure 3.7: Differential gene ontology enrichment between species in the amygdala. A. View of the vole brain illustrating the position of the amygdala (blue). B. Volcano plot of differential expression in the amygdala between prairie and meadow voles. The x axis shows normalized read counts log2 fold change between prairie voles and meadow voles. The y axis shows p-value (-log base 10) for differential expression. At a p-adjusted p<0.1, downregulated genes in prairie voles (blue) at a twofold difference lower than -2; and upregulated genes (red) at a twofold difference higher 2. C. Gene ontology enrichment analysis (GOWUA) for cellular component categories, upregulated (red) and downregulated (blue) categories in prairie vole; font size reflects p-value.

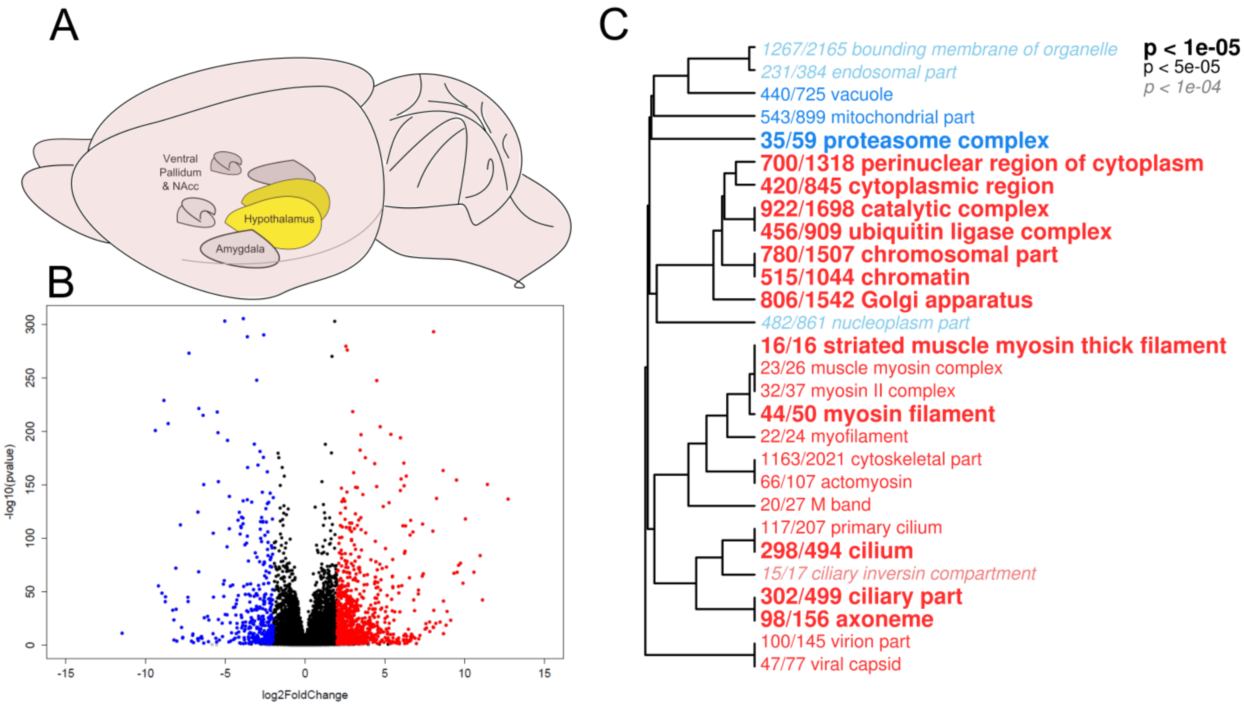


Figure 3.8: Differential gene ontology enrichment between species in the hypothalamus. A. View of the vole brain illustrating the position of the hypothalamus (yellow). B. Volcano plot of differential expression in the hypothalamus between prairie and meadow voles. The x axis shows normalized read counts log2 fold change between prairie voles and meadow voles. The y axis shows p-value (-log base 10) for differential expression. At a p-adjusted p<0.1, downregulated genes in prairie voles (blue) at a twofold difference lower than -2; and upregulated genes (red) at a twofold difference higher 2. C. Gene ontology enrichment analysis (GOWUA) for cellular component categories, upregulated (red) and downregulated (blue) categories in prairie vole; the font size of the categories represents the significance as indicated by the labels on top.

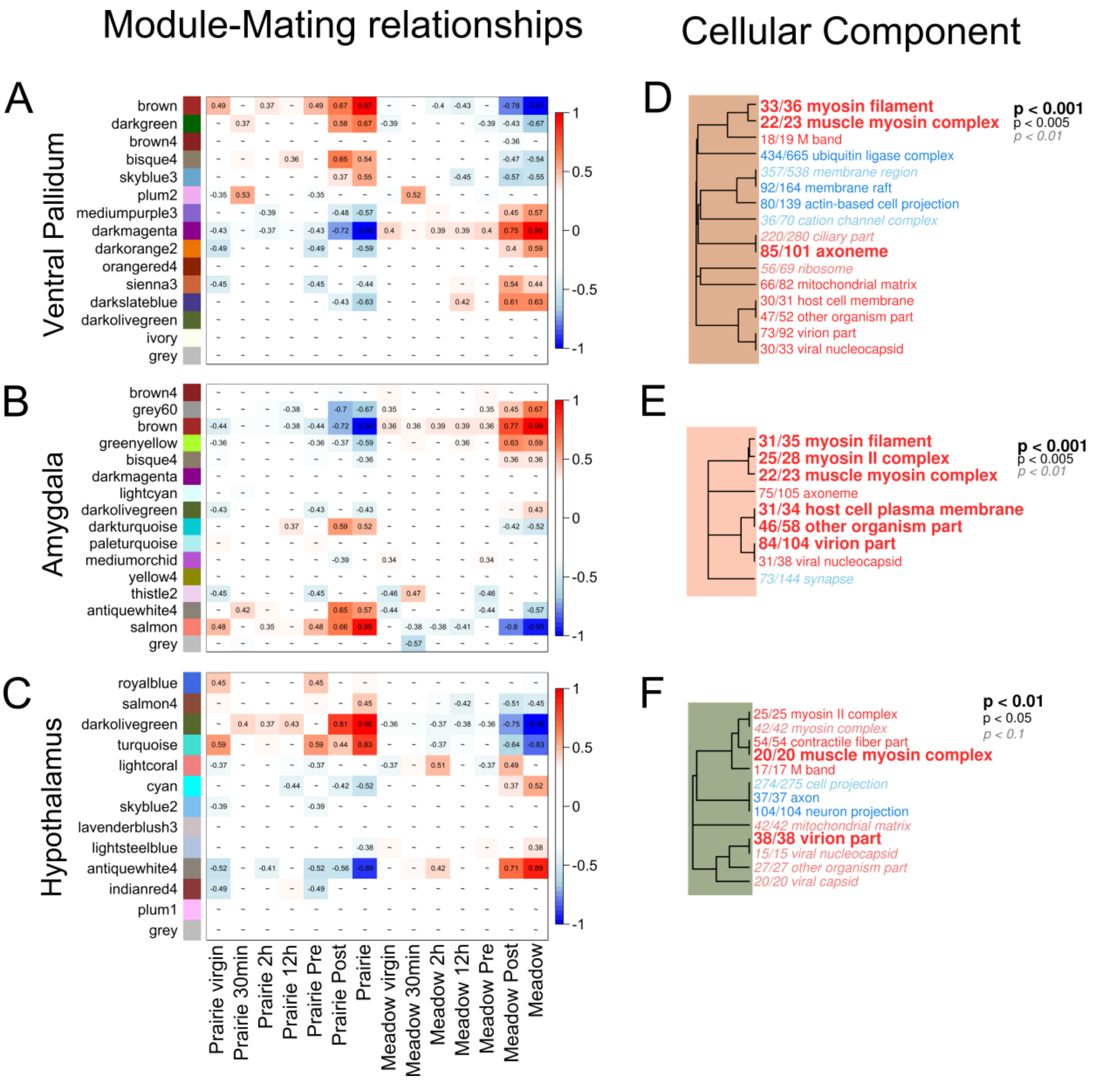


Figure 3.9: Relationships between eigengene co-expression modules and pre- and post-mating times in prairie voles and meadow voles. A-C. Correlation heatmaps between brain region modules eigengenes (rows) and mating times in two species of voles (columns). The numbers within cells are the Pearson's correlation coefficient with a $P_{corr.test}$ <0.01, higher p-values are not listed. D-E. Gene ontology enrichment for the eigengene module with highest correlation with the ventral pallidum, amygdala and hypothalamus.

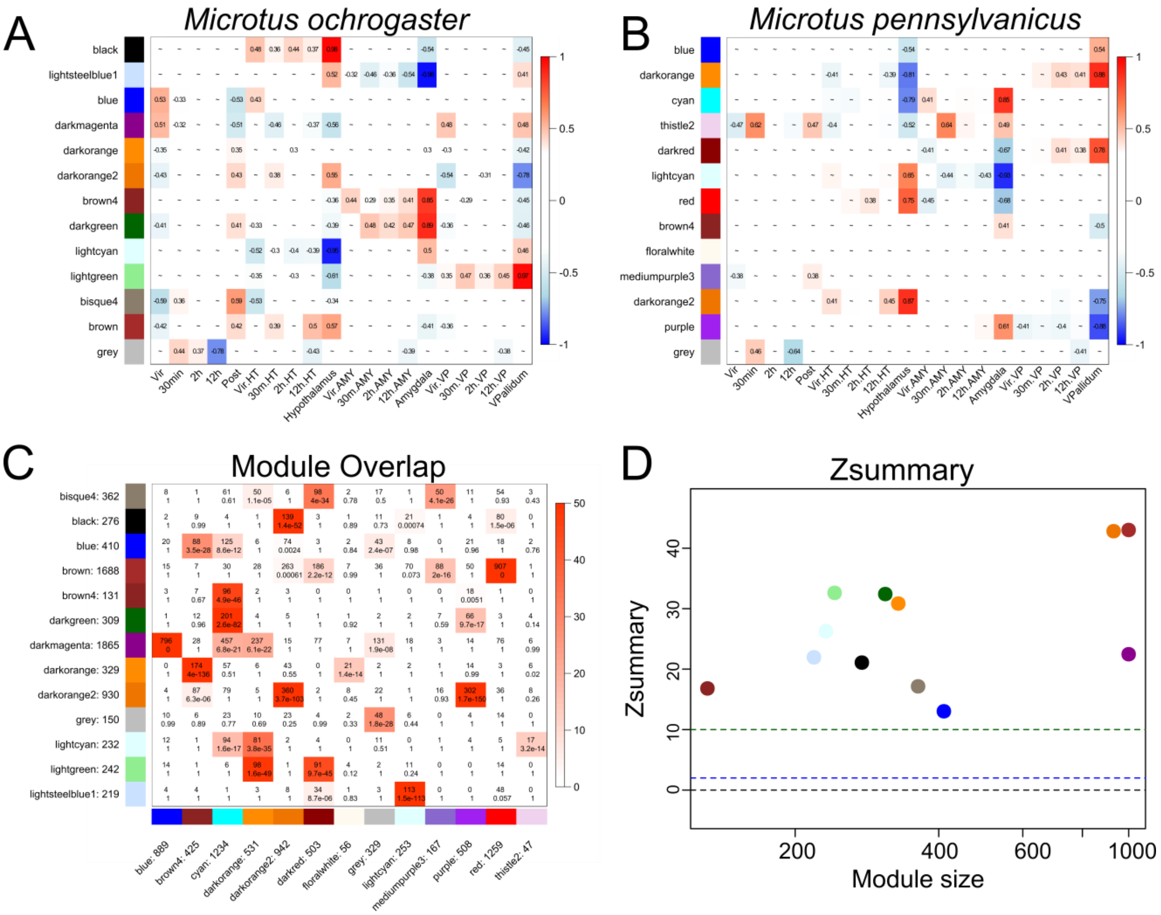


Figure 3.10: Preservation analysis of brain region and time eigengenes modules between species. Module-mating relationship heatmaps, eigengenes (rows) and mating times across brain regions (columns) in prairie voles (A), and in meadow voles (B) . C. Overlap between prairie and meadow vole co-expression modules. D. Summary Z statistic plot as a function of module size, each dot represents a prairie vole module and its degree of preservation in meadow voles. The dashed lines represent thresholds, summary < 2 indicates no preservation, 2<Zsummary<10 weak to moderate evidence of preservation, and Zsummary>10 strong evidence of preservation.

CONCLUDING REMARKS

The particular focus in my dissertation was to study the evolution of underlying cellular and molecular mechanisms involved with the formation of pairbonds and sexual fidelity in social monogamous prairie voles. The long term aim of my research interests is to seek novel explanations for proximate and ultimate questions in behavioral ecology on the subject of adaptive gene regulation. Through the use of traditional molecular biology methods such as PCR and Sanger sequencing, high throughput sequencing methods such as ChIP-seq, 2bRAD and RNAseq, and the use of bioinformatic and statistical tools, I analyzed molecular data from the vole brain, and I was able to validate the importance of gene regulation in the evolution of brain and behavior. The most interesting results I found in my dissertation so far, are: i) Positive selection have driven the evolution of at least one enhancer that increase the expression of arginine vasopressin receptor (V1aR) in the ventral pallidum; ii) A mix of balancing selection, local adaptation and selection on epistasis on a regulatory element may explain the variation in expression of V1aR in the retrosplenial cortex, which regulates spatial memory and sexual fidelity; iii) massive changes in gene expression are necessary for the origin of neurogenetic states in the limbic system of prairie voles, which coincide with the formation of pairbonds and selective aggression.

At least 5% of mammals are thought to be either socially or genetically monogamous; some examples include California mice, titi monkeys, humans, and prairie voles. But, how have prairie voles evolved to recognize their partners and maintain sexual fidelity? This has been a question that cognitive ecologists have tried to answer for long time. Hence, in my dissertation, I built upon the legacy of many years of research on some of the hypotheses addressing aspects of the evolution of monogamy of prairie voles and its relatives. Interestingly, the results of my work suggest that adaptive changes in gene regulation are important for the formation and evolution of monogamous behaviors. More specifically, I found that adaptive evolution occurs at putative regulatory elements that are essential for the expression of vasopressin receptor, a gene product that is involved in the formation of pairbonds and the maintenance of sexual fidelity in prairie voles. Additionally, I sought to understand some genic mechanisms that would explain aspects of the onset of pairbonding in prairie voles. To do this, I compared the transcriptomes from the limbic system of male prairie voles during the first 12 hours of mating, and I contrasted them to promiscuous meadow voles. Interestingly, the results of this analysis showed that prairie voles and meadow voles differ in both gene expression and behavior, revealing the activation of different neurogenetic states between monogamous and promiscuous voles. Moreover, the most upregulated genes in monogamous voles are part of known pathways associated with learning and memory formation.

From the perspective of a vole, or even a human being, attachment requires the linking the complex sensory cues associated with a specific partner to the rewards of sex and affiliation. Monogamy thus requires the formation of strong memories in the limbic system, drawing on the more general and ancient mechanisms of neural plasticity and conditioned reward. In the natural world, being more or less monogamous also requires different strategies for keeping track of other individuals, making demands on circuits important to spatial or contextual cognition. Overall, the studies I presented in this dissertation highlight the importance of social cognition and the diverse mechanisms of learning and memory for the effective execution of a given behavioral strategy. Central to this complex phenotype, as to many others, is the role of gene regulation. Selection can alter the propensity to bond by changing the expression of a gene in a reward center like the ventral pallidum, or it can favor diversity in fidelity by preserving regulatory variation in a memory circuit like the retrosplenial cortex. Ultimately, species and individual differences in bonding and memory must draw on the transcriptional regulation of neuroplasticity. One of the challenges of behavior is that it is inherently plastic, responsive to the environment. Perhaps it should not be a surprise that the evolution of behavior requires modifications of this plasticity. Finally, I would like to encourage the reader to take the advantage of many bioinformatics methods that are available to also investigate gene regulatory processes that may contribute to the evolution and maintenance of behavioral diversity.

BIBLIOGRAPHY

Alexander, D.H., Novembre, J. & Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9), pp.1655–1664.

Anderson, D.W., McKeown, A.N. & Thornton, J.W., 2015. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *eLife*, 4.

Andolfatto, P., 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437(7062), pp.1149–52.

Aplin, L.M. et al., 2014. Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature*, 518(7540), pp.538–541.

Bachner-Melman, R. et al., 2005. AVPR1a and SLC6A4 gene polymorphisms are associated with creative dance performance. *PLoS genetics*, 1(3), p.e42.

Bales, K.L. et al., 2006. Effects of stress on parental care are sexually dimorphic in prairie voles. *Physiology & Behavior*, 87(2), pp.424–429.

Barrett, C.E. et al., 2013. Variation in vasopressin receptor (Avpr1a) expression creates diversity in behaviors related to monogamy in prairie voles. *Hormones and behavior*, 63(3), pp.518–26.

Bazazi, S. et al., 2012. Vortex formation and foraging in polyphenic spadefoot toad tadpoles. *Behavioral Ecology and Sociobiology*, 66(6), pp.879–889.

Behringer, R. et al., 2014. Production of Transgenic Mice by Pronuclear Microinjection. In K. Richard Behringer, Marina Gertsenstein, V. Nagy, & A. Nagy, eds. *Manipulating the Mouse Embryo: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, pp. 237–319.

Bell, A.M. & Sih, A., 2007. Exposure to predation generates personality in threespined sticklebacks (Gasterosteus aculeatus). *Ecology letters*, 10(9), pp.828–34.

Benazzo, A., Panziera, A. & Bertorelle, G., 2015. 4P: fast computing of population genetics statistics from large DNA polymorphism panels. *Ecology and Evolution*, 5(1), pp.172–175.

Bernstein, B.E. et al., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74.

Blondel, D. V. et al., 2016. Effects of population density on corticosterone levels of prairie voles in the field. *General and Comparative Endocrinology*, 225, pp.13–22.

Boyd, J.L. et al., 2015. Human-Chimpanzee Differences in a FZD8 Enhancer Alter Cell-Cycle Dynamics in the Developing Neocortex. *Current Biology*, 25(6), pp.772–779.

Britten, R.J. & Davidson, E.H., 1969. Gene regulation for higher cells: A theory. New facts regarding the organization of the genome provide clues to the nature of gene regulation. *Science*, 25(3891), pp.349–357.

Bustamante, C.D. et al., 2002. The cost of inbreeding in Arabidopsis. *Nature*, 416(6880), pp.531–534.

Caldwell, H.K. et al., 2008. Vasopressin: Behavioral roles of an “original” neuropeptide A. Lajtha & R. Lim, eds. *Progress in Neurobiology*, 84(1), pp.1–24.

Carter, C.S., Getz, L.L. & Cohen-Parsons, M., 1986. Relationships between Social

Organization and Behavioral Endocrinology in a Monogamous Mammal. In pp. 109–145.

Chan, Y.F. et al., 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science (New York, N.Y.)*, 327(5963), pp.302–5.

Chandrasekaran, S. et al., 2011. Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. *Proceedings of the National Academy of Sciences*, 108(44), pp.18020–18025.

Cho, K.-O., Hunt, C.A. & Kennedy, M.B., 1992. The rat brain postsynaptic density fraction contains a homolog of the drosophila discs-large tumor suppressor protein. *Neuron*, 9(5), pp.929–942.

Cirelli, C., Gutierrez, C.M. & Tononi, G., 2004. Extensive and divergent effects of sleep and wakefulness on brain gene expression. *Neuron*, 41(1), pp.35–43.

Connahs, H., Rhen, T. & Simmons, R.B., 2016. Transcriptome analysis of the painted lady butterfly, Vanessa cardui during wing color pattern development. *BMC Genomics*, 17(1), p.270.

Creyghton, M.P. et al., 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50), pp.21931–21936.

Cummings, M.E. et al., 2008. Sexual and social stimuli elicit rapid and contrasting genomic responses. *Proceedings of the Royal Society of London, Series B*, 275(1633), pp.393–402.

Cushing, B.S., 2016. Estrogen Receptor Alpha Distribution and Expression in the Social Neural Network of Monogamous and Polygynous Peromyscus C. S. Rosenfeld, ed. *PLOS ONE*, 11(3), p.e0150373.

Cushing, B.S. et al., 2001. The effects of peptides on partner preference formation are predicted by habitat in prairie voles. *Hormones and behavior*, 39(1), pp.48–58.

D'Agostino, G. et al., 2013. Prolyl Endopeptidase-Deficient Mice Have Reduced Synaptic Spine Density in the CA1 Region of the Hippocampus, Impaired LTP, and Spatial Learning and Memory. *Cerebral Cortex*, 23(8), pp.2007–2014.

Danecek, P. et al., 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15), pp.2156–2158.

Deguchi, K. et al., 2008. Neurologic Phenotype of Schimke Immuno-Osseous Dysplasia and Neurodevelopmental Expression of SMARCAL1. *Journal of Neuropathology & Experimental Neurology*, 67(6), pp.565–577.

DeVries, A.C. et al., 1996. The effects of stress on social preferences are sexually dimorphic in prairie voles. *Proceedings of the National Academy of Sciences of the United States of America*, 93(21), pp.11980–4.

Donaldson, Z.R. & Young, L.J., 2008. Oxytocin, vasopressin, and the neurogenetics of sociality. *Science (New York, N.Y.)*, 322(5903), pp.900–904.

Donaldson, Z.R. & Young, L.J., 2013. The relative contribution of proximal 5' flanking sequence and microsatellite variation on brain vasopressin 1a receptor (Avpr1a) gene expression and behavior. *PLoS genetics*, 9(8), p.e1003729.

Dukas, R. & Ratcliffe, J.M., 2009. Introduction. In R. Dukas & J. M. Ratcliffe, eds. *Cognitive Ecology II*. Chicago: The University of Chicago Press, pp. 1–4.

Dulac, C., O'Connell, L.A. & Wu, Z., 2014. Neural control of maternal and paternal behaviors. *Science*, 345(6198), pp.765–770.

Emlen, S. & Oring, L., 1977. Ecology, sexual selection, and the evolution of mating systems. *Science*, 197(4300), pp.215–223.

Escalante, T., Rodriguez, G. & Morrone, J.J., 2004. The diversification of Nearctic mammals in the Mexican transition zone. *Biological Journal of the Linnean Society*, 83(3), pp.327–339.

Esvelt, K.M. & Wang, H.H., 2013. Genome-scale engineering for systems and synthetic biology. *Molecular systems biology*, 9, p.641.

Ferguson, J.N., Young, L.J. & Insel, T.R., 2002. The neuroendocrine basis of social recognition. *Frontiers in neuroendocrinology*, 23(2), pp.200–24.

Fink, S., Excoffier, L. & Heckel, G., 2007. High variability and non-neutral evolution of the mammalian avpr1a gene. *BMC Evolutionary Biology*, 7(1), p.176.

Fink, S., Excoffier, L. & Heckel, G., 2006. Mammalian monogamy is not controlled by a single gene. *Proceedings of the National Academy of Sciences of the United States of America*, 103(29), pp.10956–60.

Foll, M. & Gaggiotti, O., 2008. A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, 180(2), pp.977–993.

Gangadharan, S. et al., 2010. DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo. *Proceedings of the National Academy of Sciences*, 107(51), pp.21966–21972.

Garfield, D.A. & Wray, G.A., 2010. The Evolution of Gene Regulatory Interactions. *BioScience*, 60(1), pp.15–23.

Getz, L., McGuire, B. & Pizzuto, T., 1993. Social organization of the prairie vole (Microtus ochrogaster). *Journal of Mammalogy*, 74(1), pp.44–58.

Getz, L.L. et al., 2001. Twenty-Five Years of Population Fluctuations of Microtus ochrogaster and M. pennsylvanicus in Three Habitats in East-Central Illinois. *Journal of Mammalogy*, 82(1), pp.22–34.

Getz, L.L., Carter, C.S. & Gavish, L., 1981. The mating system of the prairie vole, Microtus ochrogaster: Field and laboratory evidence for pair-bonding. *Behavioral Ecology and Sociobiology*, 8(3), pp.189–194.

Gobrogge, K.L. et al., 2009. Anterior hypothalamic vasopressin regulates pair-bonding and drug-induced aggression in a monogamous rodent. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45), pp.19144–9.

Gompel, N. et al., 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature*, 433(7025), pp.481–7.

Goodson, J.L. et al., 2009. Mesotocin and nonapeptide receptors promote estrildid flocking behavior. *Science (New York, N.Y.)*, 325(5942), pp.862–6.

Goodson, J.L., 2005. The vertebrate social behavior network: Evolutionary themes and variations. *Hormones and Behavior*, 48(1), pp.11–22.

Goodson, J.L. & Bass, A.H., 2000. Forebrain peptides modulate sexually polymorphic vocal circuitry. *Nature*, 403(6771), pp.769–72.

Graham, D.B. & Root, D.E., 2015. Resources for the design of CRISPR gene editing experiments. *Genome biology*, 16, p.260.

Graur, D. et al., 2013. On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. *Genome biology and evolution*, 5(3), pp.578–90.

Gregersen, J.W. et al., 2006. Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature*.

Haddrill, P.R., Bachtrog, D. & Andolfatto, P., 2008. Positive and Negative Selection on Noncoding DNA in Drosophila simulans. *Molecular Biology and Evolution*, 25(9), pp.1825–1834.

Hahne, F. & Ivanek, R., 2016. Visualizing Genomic Data Using Gviz and Bioconductor. In pp. 335–351.

Hamer, D., 2002. Genetics: Rethinking Behavior Genetics. *Science*, 298(5591), pp.71–72.

Hammock, E.A.D. et al., 2005. Association of vasopressin 1a receptor levels with a regulatory microsatellite and behavior. *Genes, Brain and Behavior*, 4(5), pp.289–301.

Harris, A.H., 1988. Late Pleistocene and Holocene Microtus (pitymys) (Rodentia: Cricetidae) in New Mexico. *Journal of Vertebrate Paleontology*, 8(3), pp.307–313.

Harris, R.M. & Hofmann, H.A., 2014. Neurogenomics of behavioral plasticity. *Advances in experimental medicine and biology*, 781, pp.149–68.

Hines, H.M. et al., 2012. Transcriptome analysis reveals novel patterning and pigmentation genes underlying Heliconius butterfly wing pattern variation. *BMC Genomics*, 13(1), p.288.

Hohenlohe, P.A., Phillips, P.C. & Cresko, W.A., 2010. Using population genomics to detect selection in natural populations: Key concepts and methodological considerations. *International journal of plant sciences*, 171(9), pp.1059–1071.

Holwerda, S.J.B. & de Laat, W., 2013. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620), pp.20120369–20120369.

Hubisz, M.J., Pollard, K.S. & Siepel, A., 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in bioinformatics*, 12(1), pp.41–51.

Iida, S. et al., 2004. Genetics and epigenetics in flower pigmentation associated with transposable elements in morning glories. *Advances in biophysics*, 38, pp.141–59.

Insel, T.R., Preston, S. & Winslow, J.T., 1995. Mating in the monogamous male: behavioral consequences. *Physiology & behavior*, 57(4), pp.615–27.

Insel, T.R., Wang, Z.X. & Ferris, C.F., 1994. Patterns of brain vasopressin receptor distribution associated with social organization in microtine rodents. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 14(9), pp.5381–92.

Insel, T.R. & Young, L.J., 2001. The neurobiology of attachment. *Nature reviews.*

Neuroscience, 2(2), pp.129–36.

Jacob, F. & Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3), pp.318–356.

Jeong, S. et al., 2008. The Evolution of Gene Regulation Underlies a Morphological Difference between Two Drosophila Sister Species. *Cell*, 132(5), pp.783–793.

Jones, F.C. et al., 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, 484(7392), pp.55–61.

Keshavarzi, S. et al., 2015. Dendritic Organization of Olfactory Inputs to Medial Amygdala Neurons. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 35(38), pp.13020–8.

van Kesteren, R.E. et al., 1996. Co-evolution of ligand-receptor pairs in the vasopressin/oxytocin superfamily of bioactive peptides. *The Journal of biological chemistry*, 271(7), pp.3619–26.

Kim, D. et al., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4), p.R36.

King, M. & Wilson, a., 1975. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184), pp.107–116.

Kneussel, M. & Wagner, W., 2013. Myosin motors at neuronal synapses: drivers of membrane transport and actin dynamics. *Nature Reviews Neuroscience*, 14(4), pp.233–247.

Kobayashi, A., 2004. Requirement of Lim1 for female reproductive tract development. *Development*, 131(3), pp.539–549.

Konopka, G. & Geschwind, D.H., 2010. Human Brain Evolution: Harnessing the Genomics (R)evolution to Link Genes, Cognition, and Behavior. *Neuron*, 68(2), pp.231–244.

Landt, S.G. et al., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9), pp.1813–1831.

Langfelder, P. et al., 2011. Is my network module preserved and reproducible? *PLoS Computational Biology*, 7(1).

Langfelder, P. & Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9, p.559.

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), pp.357–359.

Lappalainen, T. et al., 2011. Epistatic Selection between Coding and Regulatory Variation in Human Evolution and Disease. *The American Journal of Human Genetics*, 89(3), pp.459–463.

Lea, A.M. & Ryan, M.J., 2015. Irrationality in mate choice revealed by tungara frogs. *Science*, 349(6251), pp.964–966.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–9.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), pp.1754–60.

Librado, P. & Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(11), pp.1451–1452.

Lim, M.M. et al., 2007. CRF receptors in the nucleus accumbens modulate partner preference in prairie voles. *Hormones and Behavior*, 51(4), pp.508–515.

Lim, M.M. et al., 2006. Distribution of Corticotropin-Releasing Factor and Urocortin 1 in the Vole Brain. *Brain, Behavior and Evolution*, 68(4), pp.229–240.

Lim, M.M., Murphy, A.Z. & Young, L.J., 2004. Ventral striatopallidal oxytocin and vasopressin V1a receptors in the monogamous prairie vole (Microtus ochrogaster). *The Journal of comparative neurology*, 468(4), pp.555–70.

Lim, M.M. & Young, L.L., 2004. Vasopressin-dependent neural circuits underlying pair bond formation in the monogamous prairie vole. *Neuroscience*, 125(1), pp.35–45.

Liu, J. et al., 2014. DLG5 in cell polarity maintenance and cancer development. *International journal of biological sciences*, 10(5), pp.543–9.

Liu, Y., Curtis, J.T. & Wang, Z., 2001. Vasopressin in the lateral septum regulates pair bond formation in male prairie voles (Microtus ochrogaster). *Behavioral Neuroscience*, 115(4), pp.910–919.

Love, M.I., Huber, W. & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12), p.550.

Luikart, G. et al., 1998. Distortion of allele frequency distributions provides a test for recent population bottlenecks. *The Journal of heredity*, 89(3), pp.238–47.

Lukas, D. & Clutton-Brock, T.H., 2013. The Evolution of Social Monogamy in Mammals. *Science*, 341(6145), pp.526–530.

Makarova, K.S. et al., 2015. An updated evolutionary classification of CRISPR–Cas systems. *Nature Reviews Microbiology*, 13(11), pp.722–736.

Marklund, S. et al., 1998. Molecular basis for the dominant white phenotype in the domestic pig. *Genome Research*, 8(8), pp.826–833.

Martin, J.A. & Wang, Z., 2011. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10), pp.671–682.

Martínková, N. & Moravec, J., 2012. Multilocus phylogeny of arvicoline voles (Arvicolini, Rodentia) shows small tree terrace size. *Folia Zoologica*, 61(3-4), pp.254–267.

Maze, I. et al., 2014. Analytical tools and current challenges in the modern era of neuroepigenomics. *Nature Neuroscience*, 17(11), pp.1476–1490.

McGraw, L.A. & Young, L.J., 2010. The prairie vole: an emerging model organism for understanding the social brain. *Trends in neurosciences*, 33(2), pp.103–9.

McGuire, B., Pizzuto, T. & Getz, L.L., 1990. Potential for social interaction in a natural population of prairie voles (Microtus ochrogaster). *Canadian Journal of Zoology*, 68(2), pp.391–398.

McKenna, A. et al., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9), pp.1297–303.

Mifsud, B. et al., 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, 47(6), pp.598–606.

Morgan, H.D. et al., 1999. Epigenetic inheritance at the agouti locus in the mouse. *Nature* *genetics*, 23(3), pp.314–8.

Nei, M., Chakraborty, R. & Fuerst, P.A., 1976. Infinite allele model with varying mutation rate. *Proceedings of the National Academy of Sciences of the United States of America*, 73(11), pp.4164–8.

Newell-Litwa, K.A., Horwitz, R. & Lamers, M.L., 2015. Non-muscle myosin II in disease: mechanisms and therapeutic opportunities. *Disease Models & Mechanisms*, 8(12), pp.1495–1515.

Newman, S.W., 1999. The medial extended amygdala in male reproductive behavior. A node in the mammalian social behavior network. *Annals of the New York Academy of Sciences*, 877, pp.242–57.

Nowick, K. et al., 2009. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proceedings of the National Academy of Sciences of the United States of America*, 106(52), pp.22358–22363.

Nylander, J.A.A., 2004. MrModeltest v2. Program distributed by the author. *Evolutionary Biology Centre Uppsala University*, 2, pp.1–2.

O'Connell, L.A. & Hofmann, H.A., 2012. Evolution of a Vertebrate Social Decision-Making Network. *Science*, 336(6085), pp.1154–1157.

Ohno, S., 1972. So much "junk" DNA in our genome. *Brookhaven symposia in biology*, 23, pp.366–70.

Okhovat, M. et al., 2015. Sexual fidelity trade-offs promote regulatory variation in the prairie vole brain. *Science*, 350(6266), pp.1371–1374.

Ong, C.-T. & Corces, V.G., 2014. CTCF: an architectural protein bridging genome topology and function. *Nature Reviews Genetics*, 15(4), pp.234–246.

Ophir, A.G., Phelps, S.M., et al., 2008. Social but not genetic monogamy is associated with greater breeding success in prairie voles. *Animal Behaviour*, 75(3), pp.1143–1154.

Ophir, A.G., Wolff, J.O. & Phelps, S.M., 2008. Variation in neural V1aR predicts sexual fidelity and space use among male prairie voles in semi-natural settings. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4), pp.1249–54.

Oti, M. et al., 2016. CTCF-mediated chromatin loops enclose inducible gene regulatory domains. *BMC Genomics*, 17(1), p.252.

Patterson, N., Price, A.L. & Reich, D., 2006. Population Structure and Eigenanalysis. , 2(12).

Pertea, M. et al., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3), pp.290–295.

Pfennig, D.W., 1992. Polyphenism in Spadefoot Toad Tadpoles as a Logically Adjusted Evolutionarily Stable Strategy. *Evolution*, 46(5), p.1408.

Phelps, S.M., 2010. From endophenotypes to evolution: social attachment, sexual fidelity and the avpr1a locus. *Current opinion in neurobiology*, 20(6), pp.795–802.

Phelps, S.M. & Ophir, A.G., 2009. Monogamous brains and alternative tactics: Neuronal V1aR, space use and sexual infidelity among male prairie voles. In R. Dukas & J.

M. Ratcliffe, eds. *Cognitive Ecology II*. Chicago: The University of Chicago Press, pp. 156–176.

Phelps, S.M. & Young, L.J., 2003. Extraordinary diversity in vasopressin (V1a) receptor distributions among wild prairie voles (Microtus ochrogaster): patterns of variation and covariation. *The Journal of comparative neurology*, 466(4), pp.564–76.

Phillips, P.C., 2008. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11), pp.855–867.

Pitkow, L.J. et al., 2001. Facilitation of affiliation and pair-bond formation by vasopressin receptor gene transfer into the ventral forebrain of a monogamous vole. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 21(18), pp.7392–6.

Pollard, K.S. et al., 2006. Forces shaping the fastest evolving regions in the human genome. *PLoS genetics*, 2(10), p.e168.

Prabhakar, S. et al., 2008. Human-specific gain of function in a developmental enhancer. *Science (New York, N.Y.)*, 321(5894), pp.1346–50.

Resendez, S.L. et al., 2016. Dopamine and opioid systems interact within the nucleus accumbens to maintain monogamous pair bonds. *eLife*, 5.

Rex, C.S. et al., 2010. Myosin IIb Regulates Actin Dynamics during Synaptic Plasticity and Memory Formation. *Neuron*, 67(4), pp.603–617.

Robinson, G.E., Fernald, R.D. & Clayton, D.F., 2008. Genes and social behavior. *Science (New York, N.Y.)*, 322(5903), pp.896–900.

Robovský, J., Řičánková, V. & Zrzavý, J., 2008. Phylogeny of Arvicolinae (Mammalia, Cricetidae): utility of morphological and molecular data sets in a recently radiating clade. *Zoologica Scripta*, 37(6), pp.571–590.

Romero, I.G., Ruvinsky, I. & Gilad, Y., 2012. Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, 13(7), pp.505–516.

Runcie, D.E. et al., 2013. Social environment influences the relationship between genotype and gene expression in wild baboons. *Philos Trans R Soc Lond B Biol Sci*, 368(1618), p.20120345.

Rusk, N., 2009. When ChIA PETs meet Hi-C. *Nature Methods*, 6(12), pp.863–863.

Sanogo, Y.O. et al., 2012. Transcriptional regulation of brain gene expression in response to a territorial intrusion. *Proceedings. Biological sciences / The Royal Society*, 279(1749), pp.4929–38.

Schneider, J.S., Giardiniere, M. & Morain, P., 2002. Effects of the prolyl endopeptidase inhibitor S 17092 on cognitive deficits in chronic low dose MPTP-treated monkeys. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 26(2), pp.176–82.

Shapiro, M.D. et al., 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, 428(6984), pp.717–23.

Siepel, A. et al., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8), pp.1034–50.

Sinervo, B. & Lively, C.M., 1996. The rock–paper–scissors game and the evolution of alternative male strategies. *Nature*, 380(6571), pp.240–243.

Slatkin, M., 1979. Frequency- and density-dependent selection on a quantitative character. *Genetics*, 93(3), pp.755–71.

Smith, J.M. & Price, G.R., 1973. The Logic of Animal Conflict. *Nature*, 246(5427), pp.15–18.

Smith, K.S. et al., 2009. Ventral pallidum roles in reward and motivation. *Behavioural Brain Research*, 196(2), pp.155–167.

Sokolowski, M.B., Pereira, H.S. & Hughes, K., 1997. Evolution of foraging behavior in Drosophila by density-dependent selection. *Proceedings of the National Academy of Sciences of the United States of America*, 94(14), pp.7373–7.

Solomon, N.G. et al., 2004. Multiple paternity in socially monogamous prairie voles (Microtus ochrogaster). *Canadian Journal of Zoology*, 82(10), p.1667.

Splinter, E., 2006. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & Development*, 20(17), pp.2349–2354.

Stephens, M., Smith, N.J. & Donnelly, P., 2001. A new statistical method for haplotype reconstruction from population data. *American journal of human genetics*, 68(4), pp.978–89.

Stern, D.L., 2000. Perspective: Evolutionary Developmental Biology and the Problem of Variation. *Evolution*, 54(4), p.1079.

Streatfeild, C.A. et al., 2011. Intraspecific variability in the social and genetic mating systems of prairie voles, Microtus ochrogaster. *Animal Behaviour*, 82(6), pp.1387–1398.

Sun, F.-L. et al., 2004. cis-Acting determinants of heterochromatin formation on Drosophila melanogaster chromosome four. *Molecular and cellular biology*, 24(18), pp.8210–20.

Tamarin, R.H., 1985. *Biology of New World Microtus*, American Society of Mammalogists.

Team, R.C., 2015. R: A Language and Environment for Statistical Computing.

Thauvin-Robinet, C. et al., 2014. The oral-facial-digital syndrome gene C2CD3 encodes a positive regulator of centriole elongation. *Nature genetics*, 46(8), pp.905–11.

Tishkoff, S. a et al., 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature genetics*, 39(1), pp.31–40.

Todd, T.P. & Bucci, D.J., 2015. Retrosplenial Cortex and Long-Term Memory: Molecules to Behavior. *Neural Plasticity*, 2015, pp.1–9.

Toth, A.L. & Robinson, G.E., 2007. Evo-devo and the evolution of social behavior. *Trends in genetics : TIG*, 23(7), pp.334–41.

Toth, A.L. & Robinson, G.E., 2010. Evo-Devo and the Evolution of Social Behavior: Brain Gene Expression Analyses in Social Insects. *Cold Spring Harbor Symposia on Quantitative Biology*, 74(0), pp.419–426.

Trowsdale, J. & Knight, J.C., 2013. Major Histocompatibility Complex Genomics and Human Disease. *Annual Review of Genomics and Human Genetics*, 14(1), pp.301–323.

Troy, H.K. & Whishaw, I.Q., 2004. A reaffirmation of the retrosplenial contribution to rodent navigation: Reviewing the influences of lesion, strain, and task. *Neuroscience* *and Biobehavioral Reviews*, 28(5), pp.485–496.

Turner, L.M. et al., 2010. Monogamy Evolves through Multiple Mechanisms: Evidence from V1aR in Deer Mice. *Molecular Biology and Evolution*, 27(6), pp.1269–1278.

Turner, L.M. & Hoekstra, H.E., 2008. Reproductive protein evolution within and between species: Maintenance of divergent ZP3 alleles in Peromyscus. *Molecular Ecology*, 17(11), pp.2616–2628.

Venables, W.N. & Ripley, B.D., 2002. *Modern Applied Statistics with S* Fouth., New York, NY: Springer New York.

Vietri Rudan, M. et al., 2015. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports*, 10(8), pp.1297–1309.

Visel, A. et al., 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231), pp.854–858.

Walther, T. et al., 2009. Improved learning and memory in aged mice deficient in amyloid beta-degrading neutral endopeptidase. *PloS one*, 4(2), p.e4590.

Wang, S. et al., 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature methods*, 9(8), pp.808–810.

Wang, Z. et al., 2014. Unique expression patterns of multiple key genes associated with the evolution of mammalian flight. *Proceedings. Biological sciences / The Royal Society*, 281(1783), p.20133133.

Warren, R.W. et al., 1994. Evolution of homeotic gene regulation and function in flies and butterflies. *Nature*, 372(6505), pp.458–461.

Whitfield, C.W., Cziko, A.-M. & Robinson, G.E., 2003. Gene expression profiles in the brain predict behavior in individual honey bees. *Science (New York, N.Y.)*, 302(5643), pp.296–9.

Winslow, J.T. et al., 1993. A role for central vasopressin in pair bonding in monogamous prairie voles. *Nature*, 365(6446), pp.545–8.

de Wit, E. & de Laat, W., 2012. A decade of 3C technologies: insights into nuclear organization. *Genes & Development*, 26(1), pp.11–24.

Wittkopp, P.J., 2010. Variable Transcription Factor Binding: A Mechanism of Evolutionary Change. *PLoS Biology*, 8(3), p.e1000342.

Wolff, J. et al., 2002. Multi-Male Mating by Paired and Unpaired Female Prairie Voles (Microtus ochrogaster). *Behaviour*, 139(9), pp.1147–1160.

Wray, G.A., 2007. The evolutionary significance of cis-regulatory mutations. *Nature reviews. Genetics*, 8(3), pp.206–16.

Yates, A. et al., 2016. Ensembl 2016. *Nucleic Acids Research*, 44(D1), pp.D710–D716.

Young, L.J. & Hammock, E.A.D., 2007. On switches and knobs, microsatellites and monogamy. *Trends in genetics : TIG*, 23(5), pp.209–12.

Young, L.J. & Wang, Z., 2004. The neurobiology of pair bonding. *Nature Neuroscience*, 7(10), pp.1048–1054.

Zhang, Y. et al., 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9), p.R137.

VITA

Alejandro Berrio was born in Medellin, Colombia. He completed his studies of high-school at the Instituto Jorge Robledo. Then, he joined the Colombian Army for one year as a mandatory enforcement rule of his country; after which, he entered the Biology Department at the Universidad de Antioquia in Medellin, where he received the degree of Bachelor of Science with an awarded research thesis on his studies of DNA repair during spermatogenesis and meiosis of local grasshoppers. During the following years, he worked as a research associate at the Center of Coffee Research (CENICAFE), Chinchina, where he investigated genetic populations and cytogenetics of the biggest coffee pest in the world –the coffee berry borer. In August of 2009, he joined the Phelps lab at the University of Florida, but then he moved with Phelps lab to Austin Texas in August 2010, where he continued his studies in the Department of Ecology, Evolution and Behavior at the University of Texas at Austin. In the Fall of 2016, he will continue his scientific career by joining a postdoctoral position in Wray Lab at Duke University.

Permanent email Address: alebesc@gmail.com

This manuscript was typed by the author.