THE UNIVERSITY OF CHICAGO

GENETIC ANALYSES OF PAVLOVIAN CONDITIONING IN OUTBRED MODELS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY
ALEXANDER F. GILETA

CHICAGO, ILLINOIS
DECEMBER 2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY FILES AVAILABLE ONLINE

Supplemental File 3.1 - SD Rat Sample Info
Supplemental File 3.2 - PCA Loadings
Supplemental File 3.3 - Heritabilities
Supplemental File 3.4 - SD Manhattan Plots
Supplemental File 3.5 - SD QQ Plots
Supplemental File 4.1 - Genetic Correlations
Supplemental File 4.2 - HS Manhattan Plots
Supplemental File 4.3 - HS QQ Plots
Supplemental File 4.4 - LocusZoom Plots
Supplemental File 4.5 - METAL Manhattan Plots
Supplemental File 4.6 - HS and METAL GWAS Hits Unfiltered Table
Supplemental Text 2.1 - ddGBS Protocol
Supplemental Text 2.2 - GBS and ddGBS Primer and Adapter Sequences

ACKNOWLEDGEMENTS

Above all, I need to thank my advisor Abraham Palmer. From the day I joined his lab, he has offered unwavering support and guidance, taking the time to meet whenever necessary to discuss my project. Abe's seemingly encyclopedic knowledge of the field always provided great insight into where I may be encountering issues, what questions could arise from results, and the next steps I needed to take. His edits to my papers, grants, and presentations were invaluable and taught me a great deal about what people want and need to hear when I'm presenting my work. Abe allowed me to work independently, giving just the right amount of push, balanced by an unbelievable amount of understanding for the time it took to work through obstacles I encountered in my research. He also tolerated my seemingly endless maladies and injuries throughout the years. Overall, an incredible advisor both academically and personally, and I am very grateful to have had his mentorship.

I would also like to thank the members of my committee: John Novembre, Terry Robinson, Mark Abney, and Dan Nicolae. John took on the role of my internal advisor and welcomed me into his lab during a transitory period after Abe moved to UCSD. Terry offered phenomenal mentorship when I was learning about neurobiology and behavior in rodents, topics with which I had no prior experience. Mark and Dan provided advice on statistical quandaries on numerous occasions throughout my projects. I would also like to thank Shelly Flagel and her lab at the University of Michigan for the advice over the years and collaboration on all aspects of my project, and Shelly in particular for co-advising me on my F31 grant and always being responsive. Even outside of my thesis research, I have been surrounded by a remarkable group of faculty: Anna DiRienzo, Jocelyn Malamy, Marcelo Nobrega, Carole Ober, Matthew Stephens, and others. It has been a privilege to work with these scientists over the years.

Of course, I would be no where without the help and support of my fellow Palmer lab members, new and old. Natalia Gonzales and Shyam Gopalakrishnan taught me everything I needed to know to get started on my journey, and Natalia continued to be there for questions, idea exchanges, and venting frustrations through the thick of it. I couldn't have accomplished all the wet lab work without Celine St. Pierre and my undergraduate team of Liz Joyce, Africa McLeod, and Rachael Maguire. Then on the computational side, Apurva Chitre has been such an amazing asset for the lab, always generous with her time and willing to help with issues any other members were encountering. Additionally, I would like to thank Jianjun Gao, April Williams, Peter Carbonetto, and Yu-yu Ren for their assistance with the early stages of data analysis for my project, as well as Oksana Polesskaya for running the lab like a boss all these years. Lastly, all the other members over the years: Xinzhu Zhou, Amelie Baud, Sandra Sanchez-Roige, Amanda Barkley-Levenson, Kat McMurray, Clarissa Parker, Amy Hart, Emily Leung, Riyan Cheng, and Hannah Bimschleger. You all have been awesome.

On the more personal note, I have to thank Sue Levison, the unofficial mom for all students in HG and GGSB. She puts in 120% effort all the time and cares immensely about all of us. Of course, Candice Lewis for her tireless efforts as HG admin for a few years. Then there are my fellow nerds Alex Advani, Katie Mika, Aarti Venkat, Bill Richter, Diedre Reitz, Andrei Anghel, Sahar Mozafarri, and numerous others who were as good a support network as you could possibly ask for. Outside of the academic world, I really have to thank Dan Ciesla for sticking by my side the majority of my journey through grad school, listening to all my woes and celebrating the successes. I'm grateful to all the friends who kept offering words of encouragement when I felt like I was never going to finish and that my project was futile. To the

members of the various sports leagues I participated in through CMSA and AAC, you all kept me sane and grounded.

Finally, there's my family. My extended family for their encouragement, love, and genuine interest in my work. My brother for comic relief and real talk. And my parents, Chrystine Raheb and Frank Gileta, who have been there for me since day one, literally. They never pushed me in any specific direction, let me follow my passion, and supported me intellectually, emotionally, and monetarily, allowing me to pursue this path. My dad helped me move out to Chicago and then again to San Diego when I decided to follow Abe to UCSD. He always has my best interests in mind and takes the time to research any things I tell him and confer with me, whether it be choice of schools, next career steps, random purchases, really anything. I have only made it this far because of the security, support, and motivation he provided. Then there is my mom, the village socialite and my emotional rock. She taught me how to navigate people, stay strong in tough situations, plan ahead, be compassionate, and most other things under the sun. She is always there when I need her, listened on the phone for hours to help me through some hard times, and is always, always supportive of my choices. They made me who I am today, and I am eternally grateful. This one was for them.

ABSTRACT

Addiction is a heritable trait. There is substantial inter-individual variability in the susceptibility to the development of addiction. Environmental cues that have been repeatedly paired with rewards are believed to be major contributors to the progression to and maintenance of addiction. These reward-associated cues are attributed with incentive salience, which makes them attractive, desirable, and capable of prompting motivated, reward-seeking behaviors. It is thought that the variability in individuals' susceptibilities to addiction is due in part to differences in the degree to which individuals attribute incentive salience to reward cues. In my dissertation, I performed the first mapping studies aimed at identifying genetic loci influencing the propensity to attribute incentive salience, estimating the heritability of this trait in the process. These genome-wide association studies (GWAS) were carried out in independent, outbred rat populations to provide replication, and the results of the studies were meta-analyzed. To successfully accomplish the GWAS, I optimized a reduced-representation sequencing approach called genotyping-by-sequencing (GBS) for use in rats and designed a variant calling workflow to obtain dense, high-quality genotypes from the GBS data. In the process of performing the mapping studies, I discovered substantial divergence between different vendor populations (F_{ST} > 0.4) of a commonly used laboratory rat strain, the Sprague Dawley (SD). Ultimately, I uncovered 21 genome-wide significant loci in the SD and 22 in the heterogeneous stock associated with various quantitative metrics that capture different aspects of this complex behavior in rats. Within these loci were a handful of candidate genes that warrant *in vivo* follow up experiments to test their effects on this important addiction-related behavior. Notably, the candidate gene TAAR1 has significant evidence linking it to addiction and potential therapeutic uses.

CHAPTER 1

INTRODUCTION

1.1 Mental illness and the genetics of complex traits

According to a recent study, there is a 29.2% lifetime prevalence of psychiatric disorders (mood disorders, anxiety disorders, substance abuse disorders, etc.) in adults world-wide [1]. It has been estimated that the global burden of mental illness accounts for the largest portion of years lived with disability (32.4%) of all human diseases and disabilities, and that it is on par with cardiovascular and circulatory disease for disability-adjusted life-years (13%) [2,3]. Unfortunately, because the biology of mental disorders is poorly understood, development of effective pharmacological treatments has proven difficult [4,5]. The molecular targets utilized by modern treatments have remained largely constant since the advent of psychiatric medications in the 1950s [5]. Due to the largely debilitating nature of these disorders and the high economic cost ($2.5 trillion a year globally; [6]) and societal impact, it is imperative that researchers develop innovative methods for dissecting the neural pathways contributing to these disorders in order to help identify novel molecular targets for therapeutic intervention.

While a portion of the incidence rate of mental illnesses can be attributed to life history, there is also a substantial contribution of genetics in predisposing individuals to development of these disorders [7,8]. Twin and family-based studies have estimated that psychiatric diseases have heritabilities anywhere in the range of 30-85% [7], with substance abuse disorders being among the most heritable (39-72%) [9]. Historically, linkage, candidate gene, and targeted resequencing studies have been used to investigate the genetic architecture of these mental disorders, under the notion that the observed genetic signal would be concentrated in a few, highly penetrant genes. However, the vast majority of linkage studies failed, and candidate gene studies often failed to

produce replicable findings [10]. With the advent of next-generation sequencing and microarray technologies for discovering and genotyping hundreds of thousands of single-nucleotide polymorphisms (SNPs) in the human genome, it became clear that the aggregate of these SNPs could explain a substantial fraction of the heritable burden of many common diseases and other traits [11]. These studies have shown that psychiatric disorders are polygenic and arise from multiple alleles of small effect that contribute to an individual's susceptibility [7].

In the past 13 years, genome-wide association studies (GWAS) have become a staple for statistically identifying genes influencing numerous mental and physiological traits [12–23]. Despite the relative success of human GWAS, a large portion of the estimated heritability of these traits remains unaccounted for by the identified loci [24,25]. It has been suggested that the this "missing" heritability could be due to multiple reasons, including: (1) the power to detect common variants of marginal effect or rare variants of moderate to large effect is low [26], or (2) genetic interactions occur between loci [27]. While it remains intractable to assay the contribution of genetic interactions at a genome-wide scale, we can tackle the issue of power in a few ways. The first is to vastly increase sample sizes [28], which numerous recent GWAS have accomplished through large consortiums [22,29,30]. An alternative approach would be to increase power by reducing the influences of diagnostic, environmental, and genetic heterogeneity on the phenotype of interest. Alas, constraining for these factors in human populations is impossible, and recording them is equally problematic. For these reasons among others, a large contingent of psychiatric researchers have turned focus to utilizing model systems (rats, mice, zebrafish, etc.) for genetic mapping, as well as searching for endophenotypes: heritable traits believed to be closer to the underlying genes than the overarching clinical disease diagnosis [31].

1.2 Animal models of psychiatric disorders

Chief among the issues facing neuropsychiatric research are the difficulty of evaluating human subjects over long periods and the ethical implications of experimentally controlling aspects of the human experience for the benefit of science. While animal models will never fully recapitulate the suite of symptoms experienced by affected humans, they still provide great utility for piecing apart the neural pathways underlying disease-relevant phenotypes. The use of model systems allows for studying behaviors in carefully controlled environments to minimize non-genetic phenotypic variation, experimental manipulations, and direct measurement of an array of physiological traits. The strength of animal models comes less from the ability to model the complex psychiatric traits as a whole so much as the ability to study specific phenotypes with either hypothetical or proven connections to human disease. Examples of such behavioral endophenotypes may include anhedonia in depression, increased locomotion in ADHD, or decreased sociability in autism [32]. These traits can be quantitatively measured and are thought to have simpler genetic etiologies that are more amenable to genetic mapping. They also tend to avoid the subjective nature of many of the diagnostic criteria for psychiatric disorders.

Though psychiatric research in non-human primates [33], *Drosophila* [34], and zebrafish [35] has been abundant and indispensable, rodent models strike a balance of exhibiting a rich behavioral repertoire while remaining cost-efficient for studying in large quantities. For years, rats lagged behind mice in terms of available genomic tools and computational resources; however, this gap has recently been considerably reduced [36]. Tools for forward and reverse genetics in both mice and rats have now been refined, allowing for discovery and validation of gene function through direct manipulation of the genes and their products [36,37]. Rats specifically are an advantageous model due to their exceptionally complex behavioral range compared to most other

models; learning and reliably performing sophisticated tasks [36]. Their larger size also allows for greater ease in making detailed physiological measurements and isolating specific brain regions for stimulation, lesion, or expression profiling. Despite their superiority in many facets of research, few genetic mapping studies have been published using rats beyond an F2 cross [38–40].

1.3 Addiction and utility of Pavlovian conditioned approach

Drug abuse and addiction are widespread issues facing the global community [41]. In 2011, the UN Office of Drugs and Crime estimated that up to 6.1% of the world population had used an illicit drug in the past year. The 2012 National Survey on Drug Use and Health showed that 9.2% of the adult US population had used an illicit drug within the past month, a 10 year high [42]. Furthermore, 20.2% of illicit drug users will experience dependency throughout their lifetime [43]. Substance abuse is an incredible economic and social burden, costing the US over $600 billion annually [44]. It is this burden that necessitates the discovery of the factors underlying risk for drug abuse, so that addiction may be both prevented and treated effectively.

Drug addiction is a chronic relapsing mental disorder characterized by compulsive drug "wanting" and seeking despite adverse consequence, and often, lack of pleasure during use [45]. Drug-associated cues are key triggers of reinstatement of drug craving and drug-seeking behaviors [46], which are predictive of relapse risk [47]. Several theories based on the reinstatement model [48] propose that when an addict encounters a stimulus previously associated with a drug reward, the cue disproportionately attracts their attention and is able to arouse emotional and motivational states that prompt and sustain drug-seeking behavior [49–52]. Robinson and Berridge proposed the 'Incentive-Sensitization Theory' to explain this behavior [51]. According to their model, repeated drug use causes persistent neuroadaptations in the brain's reward circuitry, rendering the

neural system hypersensitive to the attribution of motivational value to a reward-predictive cue through Pavlovian learning. The property acquired by the cue, which allows it to attract attention and reinforce appetitive behaviors, is known as "incentive salience" [53].

In human subjects, it has been shown that both drug and alcohol associated cues that have acquired salience can evoke involuntary attentional, emotional, and behavioral responses [54–60]. Elevated levels of attentional bias are found in substance abusers, and are correlated with increased craving, poorer treatment outcome, and increased likelihood of relapse [61–64]. This evidence emphasizes the importance of salient drug cues in the development of maladaptive drug use through their prompting of drug craving, approach, and engagement. Additionally, there is significant inter-individual variability in the degree to which salient stimuli command attentional resources in humans [63,65–68]. It is also known that certain genetic factors can increase an individual's risk for developing drug abuse [69,70]. It is hypothesized that individuals who are genetically predisposed to attribute greater incentive salience to drug-associated cues are experiencing amplified craving and cue-approach, increasing their risk for developing addictive behaviors. This would also help to explain the comorbidity of many addictions.

Our collaborators in this work, Drs. Terry Robinson and Shelly Flagel, have an established history of utilizing rats to study the attribution of incentive salience with respect to behaviors modeling compulsive drug seeking and taking in humans. It is believed there is a shared neurobiological mechanism between species, and that rats (not mice) can reliably reproduce this complex behavioral phenotype. Their lab has successfully studied this trait using the Pavlovian conditioned approach (PavCA) paradigm, in which a lever stimulus is non-contingently paired with a subsequent food reward, and rodents are scored on their approach and interaction with either the stimulus or reward receptacle [53]. Their work has shown significant individual differences in

the tendency of rats to attribute incentive value to reward-predictive stimuli [53]. The variation has been linked to neurobiological differences in dopamine systems and hypothalamic pituitary adrenal axis activity [71–74], as well as differential engagement of the cortico-striatal-thalamic "motive circuit" [74].

There are multiple lines of evidence showing that the observed variation is heritable [53,75,76]. The implication of this heritable variation is that only a subset of individuals will robustly attribute incentive salience to a discrete reward cue. These individuals are designated 'sign-trackers' due to their approach (i.e. tracking) and interaction with cues (i.e. signs) associated with food or drug reward, as opposed to 'goal-trackers', who will approach the reward (i.e. goal) upon cue presentation. Importantly, rats' attribution of incentive salience to a cue associated with a food reward has been successfully used as a proxy for drug (i.e. cocaine, nicotine, & opioid) paradigms [75,77,78]. Performance with food reward is predictive of the degree to which drug associated cues will motivate drug approach, self-administration, and reinstate drug seeking after extinction [79,80]. The current evidence in the field supports the use of the attribution of incentive salience to reward cues as an endophenotype for the fundamentally human disorder of addiction, and the theory that aberrant and inordinate attribution of incentive salience to reward cues is a key underlying factor in the genetic risk of addiction. Through investigation of this trait, I hoped to provide insights into the genetic architecture underlying the interindividual variability in the attribution of incentive salience to reward cues, specifically with respect to . In turn, this would help to elucidate the emotional, motivational, and cognitive features of conditioned drug cues [81]. This work also had the potential lead to pharmacotherapies, as GWAS hits have previously been successful in identifying important drug targets [82–84].

1.4 Rodent mapping populations

Rodent populations used in genetic mapping studies can be broadly divided into four categories, structured mouse populations, outbred mouse populations, structured rat populations, and outbred rat populations. Structured mouse populations include strains such as advanced intercross lines (AILs) [85–87], recombinant inbred lines (RILs) [88,89], and the diversity outcross (DO) mice [90,91]. There are then structured rat population, including AILs [38] as well as the heterogeneous stock (HS) [92,93,40,94], an analogous population to DO mice. Alternative to structured populations are commercially available strains that have been outbred for decades [95]. This includes mouse strains such as Swiss Webster mice (CFW) [96], and rat strains, such as Sprague Dawley (SD) [97]. The primary advantage for using commercially available outbred stocks for QTL mapping studies is the significant decay of linkage disequilibrium (LD) that has occurred between adjacent genomic loci [98]. The more LD is broken down, the higher the mapping resolution (assuming a sufficient sample size) and the more closely the linkage landscape emulates that observed in humans. Using these commercially available populations has proven successful in mice [95,96,99]; however, no mapping studies have been performed in analogous rat populations (the Sprague Dawley) prior to the work presented in this dissertation. Since my work only concerns genetic analyses in rats, I will focus the following discussion on the SD and HS.

The population history of Sprague Dawley rats is poorly understood. The SD originated in 1925 from a cross between a hooded male hybrid of unknown origin and an albino Wistar female [100]. The small number of breeders and the significant inbreeding and selection the early generations limited the amount of genetic variation present in the population [92]. However, this also putatively reduced the level of rare variation, which is beneficial for mapping efforts [36]. In the following decades, the strain was acquired by multiple commercial vendors (Charles River

Inc., Harlan [now Envigo], Taconic, etc.), all of which have multiple breeding facilities for the SD and none of which exchange breeding pairs. Despite being maintained as an outbred population to maximize genetic variation, this physical separation of the rats is likely to have caused a significant amount of drift and differentiation between the various vendors and colonies. This hypothesis is supported by numerous studies of physiological differences between these vendor populations and colonies [101–107]. However, as previously mentioned, a major potential benefit of the SD is the extensive decay of LD from nearly a century of outbreeding. Additionally, commercial availability is highly advantageous when considering the cost of maintaining a large colony of rodents for a genetic study.

The structured alternative to the SD is the HS. The National Institutes of Health's HS was created in 1984 by interbreeding eight inbred laboratory rat strains of independent origin and then outcrossing their progeny for numerous generations. The original purpose of this line was establishing a strain with high levels of genetic diversity for use in selection studies. The founder strains were therefore initially chosen to represent a broad spectrum of genetic and phenotypic diversity [92,93,108], which fortuitously benefitted genetic mapping studies as well. The outbreeding scheme ensured maximal recombination and breakdown of LD while minimizing inbreeding and genetic drift. Though the HS has now been maintained for 80 generations, the LD still remains far more extensive than is observed in human populations, which improved power for identifying QTL, though limits the mapping resolution. The primary advantage of the HS is that its genome is now a mixture of the original eight known inbred genetic backgrounds. Therefore, it is possible to use available deep sequencing data on these inbred lines to perform imputation and call genotypes at millions of SNPs using just the information from a few thousand. Additionally, since the stock has only bred for 80 generations, the majority of the variation in the population

remains common. These benefits has been leveraged previously to successfully identify numerous QTL for complex physiological and behavior traits [40,109,110].

Each of these outbred rat populations has their benefits and limitations. We pursued GWAS in both populations in parallel, taking advantage of opportunities that were part of large collaborative efforts. Since it is not guaranteed that the SD and the HS each contain the optimal set of functional variation to study the phenotype(s) of interest, using both also provided a complementary approach to investigate the genetic basis of a novel behavioral trait associated with addiction. Though, to do so at a scale large enough to be useful for QTL mapping required identifying a cost-efficient method of obtaining genotype data on thousands of samples.

1.5 Reduced-representation sequencing approaches

Unlike human, mice, and a few other model species, no affordable microarray is available for high-density genotyping in rats. Previous arrays included the RAT-DIV developed for use in the HS [40] and a custom Affymetrix chip design at the University of Michigan; however, both have gone out of production and would have cost over $300 per sample to use, severely limiting our sample size. Commercially available whole-genome sequencing (WGS) library preparation kits are also cost prohibitive, running around $80 per sample, excluding the actual cost of sequencing reagents. For these reasons, the lab developed a more cost-efficient technologies known as reduced-representation sequencing approaches, which includes protocols such as CRoPS [111], genotype-by-sequencing (GBS) [112,113], restriction site-associated DNA sequencing (RADseq) [114,115]. We have successfully utilized GBS in both mice [86,87,96] and rats [76,97,110]. Through extensive optimization that will be discussed in Chapter 2, I was able to reduce the cost of GBS to

only approximately $50 per sample, which covered all steps from DNA extraction to sequencing (the labor costs associated with the analysis are not included in this cost estimate).

The reduced-representation protocols listed above all share similar features, the foremost being that they all digest the genome with restriction enzymes with the intention of sequencing fragments of DNA adjacent to restriction cut sites. These technologies rely on the LD structure of the populations on which they are applied. Since only a small portion of the genome has sufficient sequencing data to make genotype calls (0.1-5%), the hope is that the captured SNPs will effectively 'tag' local haplotypes. Thus, only a fraction of the genome is sequenced, so that each sample only requires a few million reads for accurate genotyping, allowing for dense multiplexing of sample libraries by using barcoded adapters. In model systems with greater LD, fewer SNPs are necessary to capture the majority of the variation in the genome, and therefore more samples can be multiplexed. The portion of the genome captured can also be titrated through selection of appropriate enzyme(s) and size-selection steps for the final sequencing library. This allow researchers to choose broad coverage at low depth or narrow coverage at high depth, depending on their genotyping needs. The approaches also rely heavily on within-sample imputation, since on average only 15-20% of sequenced samples will have sufficient read depth to make a confident genotype call. With GBS, I was able to obtain genotypes at over 200,000 SNPs in SD rats [97], more than sufficient to capture the majority of the variation present in the genome. With the HS, this was greatly increased [110] due to the availability of appropriate reference panels for imputation of SNPS that were not sequenced in any of the samples.

1.6 Accounting for population structure in mapping studies

The function of a genome-wide association study is to correlate the genetic variation that exists in a sample of individuals with the phenotypic variation observed in a trait of interest. Standard GWAS approaches utilize regression techniques that rely on the assumption that the variables in the model are all independent and identically distributed. However, this assumption rarely holds due to population structure and cryptic relatedness of individuals in a sample [116,117]. Violations of this assumption can result in numerous spurious associations, where loci appear to be associated with the trait of interest, when in reality, genetic variation and the trait distribution only coincidentally covary across different groups of related individuals. The issue of relatedness is especially salient in commercially available or structured rodent populations, where there are likely to be several sets of closely related individuals in any given sample.

Initial approaches for controlling for population structure in a sample involved methods such as the use of a panel of ancestry information markers [118,119] or principal component analysis (PCA) [120]. However, while these techniques adequately control for large-scale structure due to different ethnicities or breeding populations, they failed to control for the more subtle, and often unknown, family-based structure in samples [121]. This motivated the development of variance component models, which utilize pedigree-based or empirical pairwise relatedness estimates to model the phenotypic correlations between samples [122,123]. With the advent of high-density genotyping methods (i.e. microarrays & NGS), it is possible to obtain precise estimates of the genetic relatedness of all samples in a study. Linear mixed models (LMMs) including s genetic relationship matrix (GRM) as a random effect term are able to effectively control for relatedness in structured populations and attenuate the false positive rate [124,125].

Although LMMs including a GRM can effectively control for populations structure and ethnicity, it is often advantageous to perform GWAS separately on genetically distinct populations.

In humans, heterogenous effects of SNPs are often observed across different ethnic groups [126–128]. This heterogeneity could reflect allelic heterogeneity, differences in LD structure, or differences in genetic background that affect gene-by-gene or gene-by-environment interactions [126,129–131]. In rodents particularly, variation in LD structure is prominent across populations and has the implication that SNPs identified in one population do not necessarily tag the same set of SNPs in an alternate population because they are on different LD blocks. Further, if different modifier loci are present in these populations, the penetrance of alleles may vary greatly [132], which cannot yet be dissected by current methods. Regardless, these observations suggest that unless it is known that two populations are genetically homogenous, it is prudent to perform GWAS on the samples separately to search for association unique to a given population, and then meta-analyze the results.

1.7 Replication and meta-analysis of GWAS results

With highly polygenic quantitative and behavioral traits, the true distribution of effect sizes is unknown and are likely to vary across traits and populations [133]. The probability of an observed association being true is dependent on the power to detect it, which is a function of minor allele frequency, effect size, and sample size [133,134]. GWAS typically rely on a significance threshold, which are susceptible to type 1 errors. Therefore, replication studies are considered essential to further minimize false positive errors. This need for replication is underlined further by the Winner's Curse phenomenon, whereby the effect size of associations near the genome-wide significance threshold tend to be systematically overestimated [135]. Ideally, these replication studies

would use the same statistical methods, phenotypes, and genetic variants; however, this is not always possible.

In addition to replication of positive results, performing GWAS for the same phenotype in independent samples can increase power through meta-analysis of the datasets [136]. Increasing the sample size has the effect of propelling true genetic associations of modest effect well over the significance threshold [137]. Approaches for meta-analysis involve using summary statistics to estimate z-scores and weighting the studies proportionally to their sample size [138]. Alternate methods have improved on the precision of the weighting schema by including the allele frequency and imputation quality for the SNP under consideration [139]. Though beneficial to GWAS discovery in human psychiatric traits [22,140,141], meta-analysis has rarely been applied to rodent mapping studies. In large part, this is due to a lack of consistent genotyping methodologies between studies and the lack of strain-specific variant panels and linkage maps to accurately impute. Fortunately, with the use of structured populations with known founders such as the HS, genome-wide imputation is achievable, allowing us to combine GWAS results with those from populations like the SD.

1.8 Dissertation overview

The primary aim of my dissertation project was to identify genetic loci associated with the propensity to attribute incentive salience to reward-associated cues. I accomplished this goal using the methods described in this chapter to perform genotyping and QTL mapping in two independent samples of outbred rats and then meta-analyzing the results. In Chapter 2, I discuss the steps I took to adapt and optimize genotyping-by-sequencing for use in rats, as well as outline the

computational pipeline our lab collaboratively developed to reliably call and impute genotypes in HS rats. In Chapter 3, I show that through the use of the double digest genotyping-by-sequencing (ddGBS), I was able discover over 200,000 SNPs in 4,061 SD rats and use them to characterize the genetic differences that exist between SD rats from two major commercial vendor, Charles River and Harlan. I then went on to estimate the heritability of various metrics from the Pavlovian condition approach behavioral paradigm and use linear mixed models to identify a number of associated genetic loci in each population and meta-analyzed the results. In Chapter 4, I performed a replication GWAS for the same set of PavCA metrics in an independent sample of 2,449 HS rats phenotyped at two testing centers that were part of a P50 grant. I then used the ~54,000 SNPs that overlapped between the SD and HS studies to perform a meta-analysis for PavCA in a combined sample of over 6,000 rats; by far the largest rodent mapping effort to date. I conclude the work in Chapter 5, discussing the successes and pitfalls of these studies and their implications for future research in the fields of rodent QTL mapping and addiction.

CHAPTER 2

OPTIMIZED DOUBLE-DIGEST GENOTYPING-BY-SEQUENCING METHOD AND VARIANT CALLING PIPELINE IN HETEROGENEOUS STOCK RATS

2.1 Abstract

The heterogeneous stock (HS) is an outbred rat population derived from eight inbred rat strains. The population is maintained with the goal of minimizing inbreeding and maximizing the genetic diversity of the stock. Only a few genotyping microarrays have been created for rats; they were expensive and are no longer in production. To obtain high-density genome-wide marker data for genetic mapping, we have adapted genotype-by-sequencing (**GBS**) for use in rats. In this chapter, I outline the steps we took to optimize an efficient double digest genotype-by-sequencing (**ddGBS**) protocol for rats. To analyze the ddGBS sequencing data, we evaluated multiple existing computational tools and designed a workflow that allowed us to call and impute over 3.7 million SNPs genome-wide in the HS. We also compared various rat genetic maps for use in imputation, including a recently developed map specific to the HS. Using the pipeline we established, we obtained concordance rates of 99% with data from a rat genotyping array. The computational pipeline that we have developed could be easily adapted for use in other species.

2.2 Introduction

Advances in next-generation sequencing technology over the past decade have enabled the discovery of high-density, genome-wide single nucleotide polymorphisms (**SNPs**) in model systems. Comprehensive assays of the standing genetic variation in these organisms has allowed for the identification of quantitative trait loci and the application of numerous population genetic and phylogenetic analyses. However, due to the linkage structure in many structured breeding

populations, it is often not necessary to sequence the whole genome in order to capture the majority of extant genetic variation. SNPs are frequently in strong linkage disequilibrium (**LD**) with adjacent loci, effectively 'tagging' the variation within a certain interval, and thereby reducing the number of sites that need to be genotyped. Several reduced-representation sequencing approaches that take advantage of LD structure have been previously described [111–115,142–147]. Thousands of SNPs can be identified in large numbers of samples for a fraction of the price of other genotyping methods [148,149]. The advantages of these methods are especially attractive when considering genotyping less commonly utilized species or strains for which genotyping microarrays are not available.

Of the existing reduced-representation protocols, the genotyping-by-sequencing (**GBS**) approach developed by Elshire et al. [112] has been frequently modified to accommodate new species, such as: soybean [150], rice [151], oat [152], chicken [153,154], mouse [96], fox [155], and cattle [156], among others. The greatly varying genomic composition among organisms necessitates a diverse and customized set of approaches for obtaining high-quality genotypes. Furthermore, advances in next-generation sequencing technology have led to some changes. As such, both the GBS protocol and computational pipeline require modification when being applied in a new system. Recent work from our group showed that GBS could be effectively applied to outbred mouse [87,96,157] and rat [76] populations. However, the rodent protocol and pipeline utilized had not been extensively optimized prior to employment, leaving significant room for improvement in genotype quality and marker density for use in high-resolution genetic mapping. Additionally, though several tools and workflows exist for the analysis of GBS data, including: Stacks [158], IGST-GBS [150], TASSEL-GBS [159], Fast-GBS [160], and GB-eaSy [161], the

majority were developed and optimized for use in plant species and lack an imputation step. Therefore, we saw the need to create a workflow tailored to HS rats.

The HS is an outbred rat population created in 1984 using eight inbred strains and has been maintained since then with the goal of minimizing inbreeding and maximizing the genetic diversity of the colony [93,94]. After more than 80 generations of accumulated recombination events, their genome has become a fine-scale mosaic of the inbred founders' haplotypes. The breeding scheme has made the HS colony the most diverse and complex of available structured strains of rats, ideal for optimizing our GBS protocol. Additionally, extensive deep sequencing data exists for the eight progenitor strains, allowing for accurate imputation from sites directly captured by GBS to millions of additional SNPs.

Detailed here are the steps taken to optimize a rodent GBS protocol and computational pipeline. Drawing on existing protocols [96,112,113,115] and methodologies as models, we redesigned our GBS approach and have developed a novel, reference-based, high-throughput workflow to accurately and cost-effectively call and impute variants from low-coverage GBS data in rats. This publication is intended as a resource when considering utilizing GBS in a new organism, and the methods we have employed and described herein can be feasibly modified for use in various models. The redesigned double-digest genotype-by-sequencing (**ddGBS**) has broad utility for rodent populations. We demonstrate that with a suitable reference panel, applying reduced representation approaches and imputation in model systems can provide high-confidence genotypes on millions of genome-wide markers.

2.3 Methods

2.3.1 Tissue samples and DNA extraction

Samples for this study originated from three sources: an inhouse advanced intercross line (**AIL**) derived from LG/J and SM/J mice [157], Sprague Dawley (**SD**) rats from Charles River Laboratories and Harlan Sprague Dawley, Inc. [97], and an HS rat colony [94,110]. Early ddGBS optimization utilized AIL genomic DNA extracted from spleen by a standard salting-out protocol. Later optimization steps were performed using genomic DNA from SD rats extracted from tail clippings with the PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific, Waltham, MA). Optimization of the ddGBS sequencing data analysis pipeline was performed on various subsets of our final sample of 4,973 HS rats. HS rat DNA was extracted from spleen tissue using the Agencourt DNAdvance Kit (Beckman Coulter Life Sciences, Indianapolis, IN). All genomic DNA quality and purity was assessed by NanoDrop 8000 (Thermo Fisher Scientific, Waltham, MA).

2.3.2 *In silico* digest of rat genome

We used *in silico* digests to aid in the selection of restriction enzymes, with the goal of maximizing the portion of the genome captured at sufficient depth to make confident genotype calls. Increasing the proportion of the genome covered assists with fine-mapping. The rn6 genome assembly was downloaded in FASTA format from UCSC [162]. We used the *restrict* function within EMBOSS (version 6.6.0) [163] in conjunction with the REBASE database published by New England BioLabs (NEB; version 808) [164] to perform *in silico* digest of rn6. We performed the digest with PstI alone and then with PstI paired with each of 7 secondary enzymes: AluI, BfaI, DpnI, HaeIII, MluCI, MspI, and NlaIII. We only considered fragments with one PstI cut site and one cut site from the secondary enzyme because the adapter and primer sets are designed to only allow these fragments to amplify. The number of fragments with one of each of the cut sites were summed for all observed lengths and the results summarized in Figure 2.1 and Table 2.1.

Table 2.1. Restriction enzyme options for double digest.

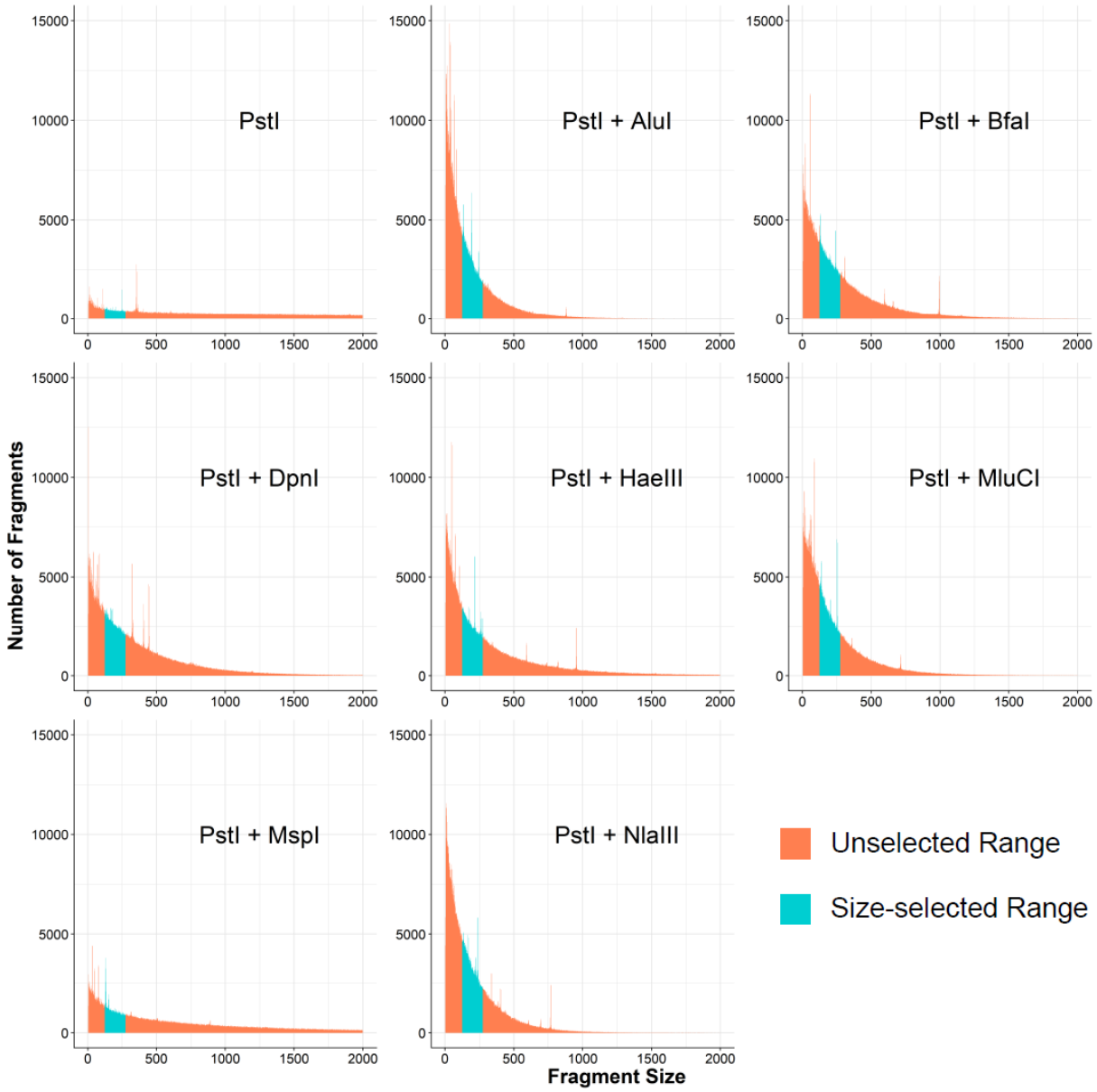
Restriction Enzyme(s)	Recognition sequence	Length of Overhang (bp)	% Genome in 250-400bp Region+	% Genome in 300-450bp Region+
PstI	CTGCA^G	4	0.48%	0.56%
PstI + AluI	AG^CT	0	3.06%	2.88%
PstI + BfaI	C^TAG	2	3.10%	3.25%
PstI + DpnI*	GA^TC	0	2.69%	3.00%
PstI + HaeIII	GG^CC	0	2.71%	2.79%
PstI + MluCI	^AATT	4	3.32%	3.21%
PstI + MspI	C^CGG	2	1.16%	1.24%
PstI + NlaIII	CATG^	4	3.45%	3.31%

* Restriction enzyme is methylation sensitive.

+ Calculated using rn6 genome length of 2,870,182,909bps.

Figure 2.1. *In silico* digest fragment distributions for PstI and potential secondary restriction enzymes.

Each panel represents an independent digest of rn6 with the listed enzyme(s). Regions highlighted in blue are fragments that would be selected by the Pippin Prep (125-275bp) after annealing adapters and primers. These regions are quantified in Table 1 by multiplying the length of the fragments by the number of fragments to estimate the portion of the genome captured.



2.3.3 Restriction enzyme selection

Initial criteria for selecting a secondary restriction enzyme were: a 4bp recognition sequence, no ambiguity in the recognition sequence (i.e. N's), compatible with the NEB CutSmart Buffer, and an incubation temperature of 37ºC. The list of enzymes meeting these criteria at the time included: AluI, BfaI, DpnI, HaeIII, MluCI, MspI, and NlaIII. Using the *in silico* digest data, we looked to maximize the portion of the genome contained within a fragment size range of 125-275bp (250-400bp with annealed adapters and primers). We excluded enzymes that produced blunt ends. We also excluded methylation-sensitive enzymes, as we did not want to limit the breadth of our sequencing efforts, accepting the possibility of read pileup in repeat regions. Based on these criteria, NlaIII, BfaI, and MluCI were selected for further testing.

2.3.4 ddGBS library preparation and sequencing

The full ddGBS protocol is available in Supplemental Text 2.1. In brief, approximately 1µg of DNA is used per sample. Sample DNA, PstI barcoded adapters, and NlaIII Y-adapter are combined in a 96-well plate and allowed to evaporate at 37ºC overnight. Sample DNA and adapters are re-eluted on day two with a PstI/NlaIII digestion mix and incubated at 37ºC for two hours to allow for complete digestion. Ligation reagents are then added and incubated at 16ºC for one hour to anneal the adapters to the DNA fragments, followed by a 30-minute incubation at 80ºC to inactivate the restriction enzymes. Sample libraries are purified using a plate from a MinElute 96 UF PCR Purification Kit (QIAGEN Inc., Hilden, Germany), vacuum manifold, and ddH_2O. Once re-eluted, libraries are quantified in duplicate with Quanit-iT PicoGreen (Thermo Fisher Scientific, Waltham, MA) and pooled to the desired level of multiplexing (i.e. 12, 24, or 48 samples per

library). Pooled libraries are concentrated to obtain the desired volume for use in the Pippin Prep. The concentrated pool is quantified to ensure the gel cassette will not be overloaded with DNA (>5µg). The pool is then loaded into the Pippin Prep for size selection (300-450bps) using a 2% agarose gel cassette on a Pippin Prep (Sage Science, Beverly, MA). Size-selected libraries were then PCR amplified for 12 cycles to increase the quantity of DNA, concentrated to attain the correct concentration, and checked for quality on an Agilent 2100 Bioanalyzer with a DNA 1000 Series II chip (Agilent Technologies, Santa Clara, CA), specifically looking for sufficient quantity and lack of adapter and primer dimer peaks.

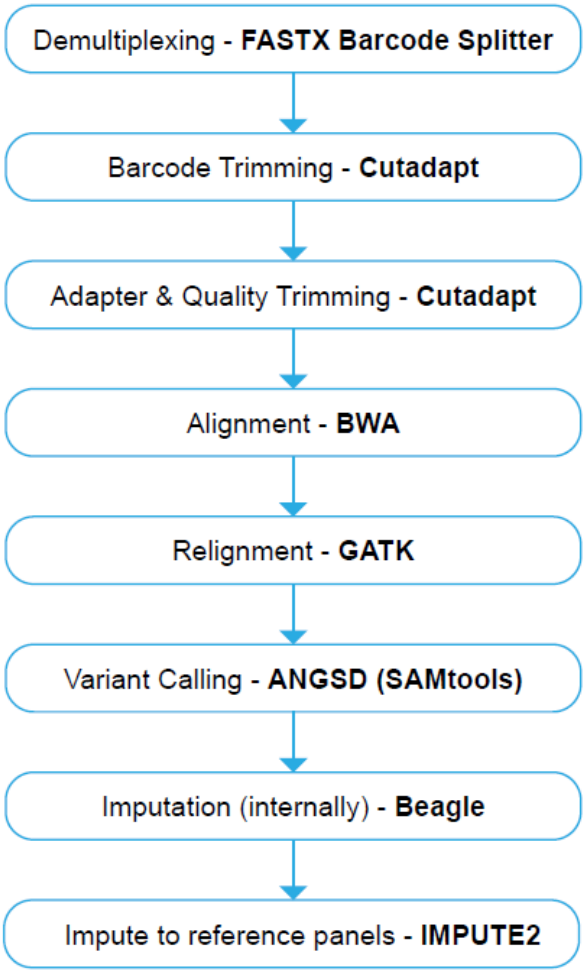
An initial 96 HS samples were sequenced 12 samples per library at Beckman Coulter Genomics (now GENEWIZ) on an Illumina HiSeq 2500 with v4 chemistry and 125bp single-end reads. Subsequently, we began using a set of 48 unique barcoded adapters (Supplemental Text 2.2) to multiplex HS samples 48 per ddGBS library. Each library was run on a single flow cell lane on an Illumina HiSeq 4000 with 100bp single-end reads at the IGM Genomics Center (University of California, San Diego, La Jolla, CA). We chose single-end reads over paired end because we were interested in genotyping SNPs, rather than indels and other structural variants.

2.3.5 Evaluation of ddGBS pipeline performance

The full sequence of steps required to call and impute genotypes from ddGBS data is presented in Figure 2.2 During optimization of the pipeline, performance was assessed by two primary metrics: (1) the number of variants called and (2) genotype concordance rates for calls made in 96 HS animals that had both ddGBS genotypes and array genotypes from the Axiom MiRat microarray. After each modification, the variant count and concordance rates were reevaluated. There were two checkpoints in the GBS pipeline where genotype quality (as a function of concordance rate) was assessed: after initial variant calling and imputation with ANGSD/Beagle

[165–167] and again after imputation to the reference panel with IMPUTE2 [168,169]. Additionally, we checked the transition to transversion ratio ($T_S T_V$), which is expected to be ~2. A final concern was the efficient utilization of computational resources, which was assessed primarily by run time.

Figure 2.2. ddGBS sequencing data analysis workflow.



2.3.6 Demultiplexing

The PstI adapter barcodes were used to demultiplex FASTQ files into individual sample files. Three demultiplexing programs were tested: FASTX Barcode Splitter [170], GBSX [171], and an

in-house Python script. Input included the FASTQ files along with a plain text file containing samples identifiers, associated barcode sequences, and (if applicable) the name of the restriction enzymes used during library preparation. Reads that could not be matched with any barcode (1 mismatch was allowed), or that lacked the appropriate enzyme cut site, were discarded. Samples with less than two million reads after demultiplexing were removed. Data concerning demultiplexing shown in Supplemental Table 2.1 are from a single HS rat sequenced in a 12-sample library on one lane after demultiplexing and adapter/quality trimming.

2.3.7 Adapter and quality trimming

FASTQ read quality was visualized by FastQC v0.11.6 [172]. We compared the efficacy of two rapid, lightweight software options for trimming barcodes, adapters, and low-quality bases from the NGS reads: Cutadapt [173] and the FASTX Clipper/Trimmer/Quality Trimmer tools (Supplemental Table 2.2). [170]. A base quality threshold of 20 was used and reads trimmed to below 25bps were discarded. Variant data was summarized using PLINK/SEQ [174].

2.3.8 Read alignment and indel realignment

Rattus norvegicus genome assembly Rnor_6.0 was used as the reference genome for read alignment with the Burrows-Wheeler Aligner (BWA) [175] using the *mem* algorithm. Using ANGSD to call variants from the aligned reads, we were able to compare variant counts at multiple mapping quality thresholds: 20, 30, 45, 60 and 90 (Supplemental Table 2.3). We then used GATK IndelRealginer [176] to improve alignment quality by locally realigning reads around a reference

set of known indels in 42 whole-genome sequenced inbred rat strains, including the eight HS progenitor strains [177].

2.3.9 Variant calling and imputation

Variants were called using a combination of the SAMtools model in ANGSD v0.911 [165,178] and imputed using Beagle v4.1 [166,167]. Prior to settling on ANGSD/Beagle, GATK's UnifiedGenotyper and HaplotypeCaller [176] were tested with various parameter settings, but we did not obtain satisfactory performance. Using ANGSD, we inferred the major and minor alleles (*-domajorminor* 1) from the genotype likelihoods, kept only high confidence polymorphic sites (*-snp_pval* 1e-6), and estimated the allele frequencies based on the inferred alleles *(-domaf* 1). We discarded sites missing read data in more than 4% of samples (*–minInd* 4774). Additionally, we tested altering thresholds for minimum base quality score (*-minQ*) and mapping quality (*-minMapQ*). We then used Beagle to impute genotypes at variant sites missing calls for subsets of samples. Beagle converted the ANGSD VCF file format to the more conventional VCF 4.1 format and phases genotype data.

2.3.10 HS QC and pre-phasing for reference panel imputation

Prior to imputing SNPs from the 42 inbred strain reference panel, we checked concordance rates for the 96 HS animals with array genotypes, identified Mendelization errors based on known pedigree information, examined the $T_S T_V$ ratio, and assessed whether the sex as recorded in the pedigree records agreed with the sex empirically determined by the proportion of reads on the X-chromosome out of the total number of reads (Supplemental Figure 2.1). Lastly, we only retained

variants previously identified in the 8 HS founders [179] for use in imputation to the 42 genomes [177].

To improve imputation accuracy and computational efficiency, we employed a pre-phasing step prior to reference imputation. A flowchart outlining the pre-phasing protocol is presented in Supplemental Figure 2.2. We tested three methods of converting genotype data to phased .hap files: VCFtools (*--impute*) [180], SHAPEIT [181], and IMPUTE2 (*-prephase_g*) [168]. We assessed performance by checking genotype concordance rates for the 96 HS samples with array data.

2.3.11 Genetic maps

Genetic maps are required for proper phasing and imputation with IMPUTE2. When we began this project, no strain-specific recombination map was available for HS rats. Thus, we considered a sparse genetic map for SHRSPxBN [182]. We also tested two types of linear (with respect to physical distance) genetic "maps", which have recombination rates set at 1cM/Mb or the chromosome specific averages for rats, as reported by Jensen-Seaman et al. [183]. Lastly, late in the evolution of this project, we experimented with an HS-specific genetic map developed by Littrell et al. the Medical College of Wisconsin [184].

2.3.12 HS imputation to reference panel

We used existing sequencing and array data from the HS rat founder and other inbred laboratory rat strains [177] as reference panels for imputation. Genotype data underwent QC and were phased

by Beagle into single chromosome haplotype files. Haplotype files were then created using the workflow detailed in Supplemental Figure 2.2. Imputation by IMPUTE2 was performed in 5Mb chunks of the genome using the aforementioned reference panels and genetic maps. The chunks were subsequently concatenated by order of physical positions for each chromosome individually.

2.4 Results

2.4.1 ddGBS optimization

Previous projects utilizing GBS in mice and rats [76,96,157] often encountered an issue where certain regions of the genome experienced high pileups of reads per sample (>100x), while other regions were covered by just 1-2 reads. This read distribution imbalance can be caused in part by PCR amplification bias, where a subset of fragments are preferentially amplified until they dominate the final library [185,186]. Our original protocol utilized 18 cycles of amplification. We tested reducing this to 8, 10, 12, or 14 cycles and found that only 12 cycles were necessary to yield sufficient PCR product for sequencing; the reduction in the number of PCR cycles was expected to reduce PCR bias.

Another concern regarding previous sequencing results was an excess of long fragments (>700bps as determined by *in silico* digest) with insufficient reads to make confident genotype calls (< 5 reads per sample). These reads ultimately go unused and are therefore wasteful of sequencing reagents. We tested multiple methods of combatting this issue, including: reducing the PCR extension time, increasing the digestion length or enzyme concentration, increasing the selectivity of the primers, performing size selection on the libraries, or using a two-enzyme restriction digest.

We began by testing shorter extension times (15s/20s vs. 30s) to try and reduce the average fragment size of our sequencing libraries by preventing fragments with larger inserts from completing amplification. However, we did not see noticeable improvement in the final library composition at shorter extension times and determined the original length of 30 seconds was appropriate. We additionally evaluated the effects of increasing the length of the restriction digest from 2 hours to 3 or 4 hours, as well as increasing the number of units of PstI enzyme added, to ensure complete digest. Neither of these modifications impacted the final fragment length distribution of the library, confirming the digest was reaching completion by 2 hours with the original PstI concentration.

As an alternative approach, we attempted reducing the complexity of our library to concentrate our reads on a smaller fraction of the genome. We accomplished this using the method employed by Sonah et al. 2013 [150]; adding an additional 1-2 nucleotides to the sequencing PCR primers (A or AG). The addition of these nucleotides increases the selectivity of the primers and reduces the number of amplifiable fragments, sacrificing sequencing breadth for read depth. However, this did not ameliorate the issue of reads being wasted on long fragments and increased the PCR bias (data not shown). Therefore, we chose to retain our original primer design.

Our previous GBS protocol did not have an explicit library fragment size selection step. The final library was purified using a MinElute PCR Purification Kit (QIAGEN Inc., Hilden, Germany), which isolates PCR products 70bp-4kb in length, leaving a wide range of fragment sizes in the final library and trusting that only shorter fragments would bridge amplify on the flow cell. This method was imprecise and had low reproducibility, negatively impacting our ability obtain reads at consistent sites across libraries. Rather than attempt size selection by gel extraction, we chose to utilize a Pippin Prep, which automates the elution of DNA libraries of desired fragment

size ranges. By using this automated size selection, we reduce the proportion of the genome targeted for sequencing, and because restriction enzymes make the consistent cuts across samples, ensure the same fragments are sequenced in the majority of libraries. Since the clustering process involves a bridge amplification step that preferentially amplifies library fragments with shorter insert sizes [187], we kept the size selection window narrow (250-400bps) to avoid introducing a bias in which fragments were sequenced.

PstI has a 6bp recognition sequence for cleaving DNA. When used alone, *in silico* digest of the rn6 reference genome (Fig 2.1; Table 2.1) showed that only ~0.5% of the genome would have fallen within a 150bp fragment size window selected on the Pippin Prep. Given the LD structure of other outbred rodent lines such as DO and CFW mice [96], we speculated this small fraction would not yield sufficient SNPs to tag the majority of variation in the genome. Additionally, we were concerned about potential biases in coverage, heterozygosity, and the minor allele frequency (**MAF**) spectrum that may be introduced by incomplete capture of the genome in our libraries [188]. To increase the proportion of the genome captured within the fragment size window, we pursued a double digest of the genome using a secondary enzyme with a more frequently occurring recognition sequence (Figure 2.1; Table 2.1). BfaI, MluCI, and NlaIII were chosen for further testing due to their compatibility with PstI digestion reagents and temperatures, sticky ends, and the proportion of the genome falling in the size selection window. We ruled out BfaI because it only had a 2bp overhang after cleavage, which led to a high concentration of adapter dimer in the sequencing libraries. Ultimately, we proceeded with NlaIII over MluCI, as it contained the greatest portion of the genome in our size selection window.

In our previous GBS protocol, all fragments were cut on both ends by PstI. By using a substantially lower concentration of the barcoded PstI adapter than the common PstI adapter, we

ensured the barcoded adapter would be the limiting reagent and the majority of fragments with an annealed barcoded adapter would have a common adapter on the other end. This is crucial, as having one of each of the adapters is required for proper amplification of the fragments on the flow cell. However, when using both PstI and NlaIII, the library is predominantly composed of fragments cut on both sides by NlaIII, which will amplify during PCR with a common adapter, but not on the flow cell. Therefore, we employed a Y-adapter [113] to control the direction of the first round of PCR and prevent two-sided NlaIII fragments from dominating the final sequencing library (Supplemental Text 2.2).

We tested numerous quantities of PstI and NlaIII adapters in an attempt minimize the amount used and avoid adapter dimers in the final libraries. For the barcoded PstI adapters, we tested 120pmol, 40pmol, 20pmol, 4pmol, 3pmol, 1.5pmol, 0.6pmol, and 0.2pmol and for the NlaIII Y-adapter, 16pmol, 8pmol, 4pmol, 2pmol, 1pmol, and 0.5pmol. We found that 0.2pmol of PstI adapter and 4pmol of NlaIII Y-adapter yielded sufficient library and avoided adapter dimers.

We sequenced a trial flow cell with 8 pooled ddGBS libraries of 12 SD rat samples each (96 total) on a HiSeq 2500 (Illumina, San Diego, CA) with 125bp reads and v3 chemistry, obtaining an average of 15.3 million reads per sample. Given the NlaIII *in silico* digest results suggested we were capturing ~3.4% of the genome and that we were using 125bp reads, this was approximately 20x coverage of captured sites. We subsequently increased the number of samples to 48 per library for the HS rats because we hypothesized 5x would be sufficient coverage per sample when utilizing imputation to a reference panel. We also discovered that a portion of the reads contained sequence fragments of the NlaIII adapter sequence, indicating there were fragments with insert sizes smaller than 125bps in the final library. To avoid this, we increased the fragment size range to 300-450bps (Table 1), roughly 175-325bp insert sizes. Due to the high

concentrations of our libraries after pooling, the library size distribution obtained from the Pippin Prep was uniformly shifted towards higher fragment lengths (Supplemental Figure 2.3).

The final ddGBS protocol can be found in Supplemental Text 2.1 and the necessary primer and adapter sequences in Supplemental Text 2.2. This protocol was used for the sequencing of all HS rats included in the following computational optimization steps.

2.4.2 Demultiplexing

The number of base pairs of sequencing data retained after demultiplexing was fairly consistent across demultiplexing software (Supplemental Table 2.1). We ultimately decided to use FASTX Barcode Splitter because it yielded the greatest number of reads after quality/adapter trimming and had faster run times, though this may be predominantly attributable to outputting uncompressed FASTQ files. An average of 330 million 100bp reads were obtained per library, resulting in ~7 million reads per sample. Supplemental Figure 2.4 shows the distribution of reads counts for all samples after demultiplexing, but prior to sample filtering based on read count.

2.4.3 Adapter and quality trimming

Read quality was substantially elevated after trimming the barcode and adapter sequences and low-quality base pairs at the ends of reads (Supplemental Figure 2.5). Overall read counts were only marginally reduced by quality trimming (Supplemental Table 2.1). We observed that the number of called variant sites and the genotyping rate were both greater when using reads initially processed by Cutadapt (Martin, 2011) than reads processed by the FASTX_Toolkit (Supplemental

Table 2.2). Importantly, a large portion of the additional identified variants were known variant sites from the 42 inbred strains reference set (Supplemental Figure 2.6), indicating the elevated call rate was at least in part due to capturing more true variant sites. We viewed this as sufficient support for proceeding with Cutadapt for quality trimming.

2.4.4 Read alignment quality

The number of called variants and genotype call rates were identical at read mapping quality (mapQ) thresholds of either 20 or 30 (Supplementary Table 2.3) within ANGSD. As the ANGSD mapQ threshold was raised to 45, there was a small reduction in the number of called variants, and then much greater losses at thresholds of 60 or 90. Fortunately, genotype concordance rates at both low and high mapQ thresholds were stable, despite the putatively higher quality of the alignments (data not shown). This permitted us to select a lower mapQ threshold (mapQ = 20), maximizing the number of variants called without sacrificing genotype quality.

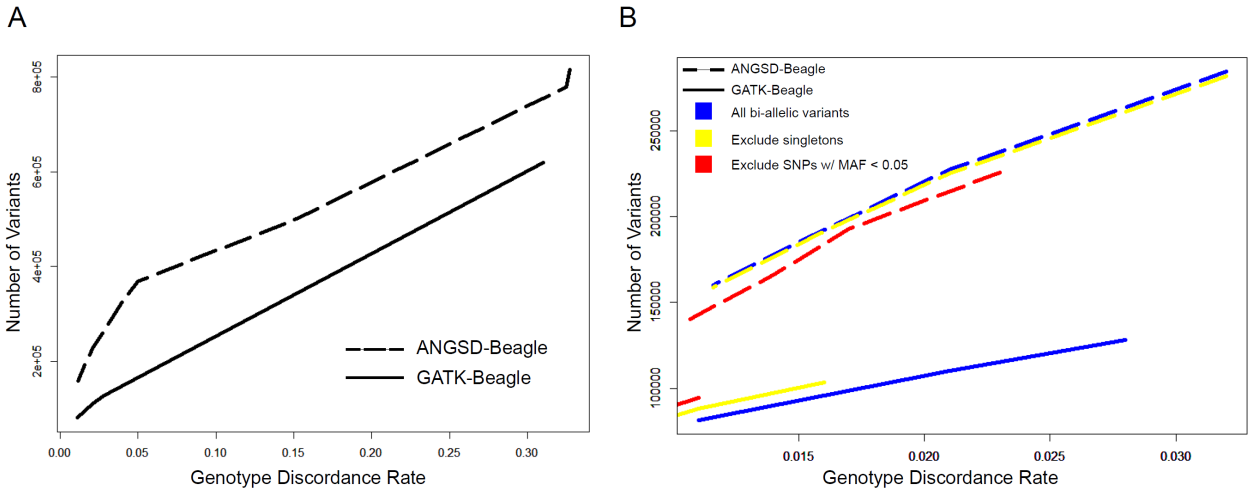
2.4.5 Variant calling

Neither GATK's UnifiedGenotyper nor HaplotypeCaller [176,189] called an adequate number of SNPs for fine-mapping in the HS. Panel A of Figure 2.3 shows that the combination of HaplotypeCaller and Beagle had lower rates of genotype concordance with array data than calls made by ANGSD and Beagle. We tried several adjustments to the GATK default thresholds for variant quality, heterozygosity, graph pruning, dangling branch length, and number of reads sharing an alignment start to declare an active region; however, no alterations significantly improved the results (data not shown; available by request). As shown in Panel B of Figure 2.3,

ANGSD/Beagle yielded a substantially greater number of SNPs at all genotyping error rate thresholds. This observation held when variants were limited only to biallelic sites and SNPs with an MAF > 0.05.

Figure 2.3. Genotype discordance rates between array data and variants called by GATK or ANGSD.

The two panels compare the number variants called by combination of ANGSD and Beagle or GATK HaplotypeCaller and Beagle at various thresholds of genotype discordance with array data. Calls were made using the 96 HS rats with array data. (A) The x-axis represents the genotype discordance rate thresholds and the y-axis is the number of variants that surpass that threshold for each genotype calling method. (B) Additional filters were applied to the original SNP sets and the plot zooms in on a smaller range of acceptable discordance rates. Blue lines represent the unfiltered SNP set. Yellow lines have been filtered for singletons. Red lines have further excluded SNPs with an MAF < 0.05. Each line contains the same number of points.



Within ANGSD, we compared the 4 possible models for estimating genotype likelihoods: SAMtools, GATK, SOAPsnp and SYK. The SOAPsnp model demonstrated an advantage in genotype accuracy and number of variants calls post-imputation with Beagle (Supplemental Figure 7). However, SOAPsnp requires considerably more computing time and resources, outweighing its marginal benefits when applied to a large sample set (>2,000). The GATK model showed a

greater number of variants for more lenient discordance rate threshold, but as stringency increased, the number of variants converged across the remaining 3 models. We proceeded with the SAMtools model for genotype likelihood estimation due to its previous support in the GBS literature [160], accepting a nominal decrease in highly concordant variants for a large gain in efficiency.

2.4.6 Imputation to reference panel

We observed a highly upward skewed transition/transversion ratio compared to the expectation of ~2 (Supplemental Table 2.4) in our ANGSD/Beagle SNP data prior to imputation with IMPUTE2. This issue was ameliorated if we filtered the SNP set for only "known" variants that were previously identified in either the 42 inbred strains [177] or the 8 deep-sequenced HS founders [179]. For imputation, we therefore only provided IMPUTE2 with previously identified variant sites from our ANGSD/Beagle output. Prior to running IMPUTE2, we also filtered the variants for biallelic sites with a genotype call in more than one individual. Using pedigree data for the HS rats, we were further removed sites and samples showing high levels of Mendelization errors. Lastly, we removed any samples where the sex chromosome read ratio was contrary to their reported sex (Supplemental Figure 2.1).

To determine which reference set to impute outwards to, we tested six different possible combinations of available reference data (Table 2.2). The most accurate imputation was observed for the reference set containing only the 8 deep-sequenced HS founder strains from UMich; however, imputation to this set had the lowest genotyping rate of all panels. In contrast, using the 42 rat inbred strains displayed a balance of high accuracy and low missingness, leading us to

choose this as our reference set. To justify our choice, we also compared the 8 founders' data within the 42 inbred strains to other 34 rat strains in the reference set. As expected, discordance rates were much higher when only considering non-founders. Interesting though, genotyping rates were greater for the 34 than the 8 founders, suggesting a combination of the two was the optimal choice.

Table 2.2. Imputation accuracy based on different variant reference panels for IMPUTE2.

		Chr1	Chr2
42 Inbred Strains	**Discordance rate**	0.011	0.010
	# Variants	790,659	882,993
	Genotyping Rate	0.85	0.81
UMich 8 founders + 42 Inbred Strains	**Discordance rate**	0.007	0.011
	# Variants	864,670	898,621
	Genotyping Rate	0.63	0.64
All 34 non-founder inbred strains	**Discordance rate**	0.035	0.030
	# Variants	812,550	912,749
	Genotyping Rate	0.84	0.80
Only the 8 HS founders from the 42 inbred strains	**Discordance rate**	0.012	0.011
	# Variants	805,424	902,061
	Genotyping Rate	0.57	0.53
Only the UMich 8 founders	**Discordance rate**	0.0059	0.008
	# Variants	865,514	898,621
	Genotyping Rate	0.42	0.41
Old 8 HS only	**Discordance rate**	0.0095	0.0096
	# Variants	507,909	540,844
	Genotyping Rate	0.43	0.40

All genetic map options produced promising results from imputation (Supplemental Table 2.5). Surprisingly, the genetic maps based on a constant rate of recombination across the chromosome yielded similarly high quality imputed genotypes as the HS-specific map with variable rates tailored to the HS LD landscape [184]. Due to the marginal differences, we chose to use the chromosome-specific values initially published by Jensen-Seaman [183] for simplicity without compromising our results. We also found no large difference in resultant imputation accuracy between pre-phasing with VCFtools and IMPUTE2; however, SHAPEIT showed a higher rate of genotype discordance than either (data not shown).

To obtain our final set of ~3.7 million variants, a final round of variant filtering is performed after imputation to the 42 genomes data with IMPUTE2 to remove any remaining SNPs with MAF < 0.005, a post-imputation genotyping rate < 90%, and or that violate HWE at a threshold of $1x10^{-10}$.

2.5 Discussion

The use of microarrays and WGS for genotyping large samples in emergent model species remains cost-prohibitive. There is therefore an urgent and wide-spread need for a high-performance and economical methods of obtaining genome-wide genotype data. While reduced-representation approaches have been utilized in numerous species of plants and animals, including rodents [76,87,96,115,157], there has yet to be a published protocol optimized specifically for rats. Prior to sequencing thousands of HS samples with GBS for our mapping efforts, we wanted to ensure we were capturing the greatest possible number of high-quality variants. The protocol we present here in the culmination of thorough testing and optimization of each step of the GBS protocol for rats. We have now applied the approach to 4,973 HS rats, as well as 4,608 Sprague Dawley rats [97].

Our previous GBS computational pipeline (Parker *et al*, 2016), designed for use with CFW mice, was unsuitable for our current genotyping efforts in HS rats, due in part to the much higher levels of genetic diversity of HS population. There are multiple reasons we chose to develop our own computational pipeline for GBS rather than using existing workflows. Foremost, the prominent GBS analysis pipelines were developed and optimized for use in crop species [150,158–161], which are polyploid and have differing levels of variation and LD than outbred rodent populations. Additionally, there were elements of each pipeline that did not meet our needs or lacked customizability. For instance, TASSEL-GBS v2 [159] trims all reads to 92 base pairs; however, other projects underway in our lab utilized up to 125bp reads, leading to a ~20% reduction in data. TASSEL-GBS also ignores read base quality scores, which are informative in probabilistic frameworks for estimating uncertainty in alignments and variant calls [189–191], and uses a naïve binomial likelihood ratio method for calling SNPs. Stacks has previously shown poor demultiplexing [160,171] and does not make use of the reference genome for priors when calling SNPs [158]. Fast-GBS relies on Platypus for variant calling [160,192], which employs a Bayesian method of constructing candidate haplotypes that works poorly with low-pass sequencing data and does not scale well [193]. Lastly, none of these pipelines included an imputation step, which is crucial for filling in missing genotypes in GBS data and can provide hundreds of thousands of additional SNPs given an appropriate composite reference panel [194,195].

Though we have not explicitly tested each alternate GBS pipeline for the purposes of this publication, this has been recently done by Wickland et al. [161]. Their pipeline GB-eaSy, which ours most closely resembles, was found to be superior by a number of metrics to Stacks, TASSEL-GBS, IGST, and Fast-GBS. Similar to GB-eaSy, our pipeline utilizes a double-digest GBS protocol, aligns reads to the reference genome with *bwa-mem*, and uses the SAMtools genotype

likelihood model for calling SNPs [196]. The combination of bwa-mem and SAMtools algorithm was independently shown to have the best performance for calling SNPs from Illumina data [197], further supporting our choice of these programs for read alignment and variant calling. Additionally, using the ANGSD wrapper provided us with the ability to convert the posterior genotype probabilities into dosages for mapping studies [165], useful for conveying important uncertainty in sample genotype calls.

A minor difference between GB-eaSy and the proposed pipeline is the use of Cutadapt [173] rather than GBSX [171] for demultiplexing, though both performed equally well (Table S1). The primary improvement is our extension of the pipeline with the implementation of effective internal and reference-based imputation steps using the 42 inbred rat genomes [177] and 8 deep-sequenced HS founders from UMich [179]. There are two stages of imputation in the pipeline: the first one is accomplished by Beagle, which aims to fill in missing genotypes at called variants using information from other samples. This round of imputation raises the genotype call rate to 100% but may also introduce errors if there is low confidence in the existing genotype calls, emphasizing the need for careful filtering steps. The second stage of imputation made use of IMPTUE2 and an external reference panels of variants called from WGS data on the 8 inbred HS founders, as well as 34 additional inbred rat strains. We decided to include the 34 additional strains because of the elevated genotyping rate we observed upon their inclusion in the IMPUTE2 reference panel. We believe they may help make calls in small chunks of the genome that do not compellingly match the 8 known founders.

In summary, we have redesigned a GBS protocol and genotyping and imputation pipeline to obtain dense genotypes on genome-wide markers in highly-multiplexed HS rats. After quality filtering on the level of SNP and sample, over 3.7 million were called. The ddGBS protocol and

bioinformatic methods used to produce this data are publicly available, easy to handle, and cost-effective. The presented workflow could be feasibly followed with marginal modifications for application in other species.

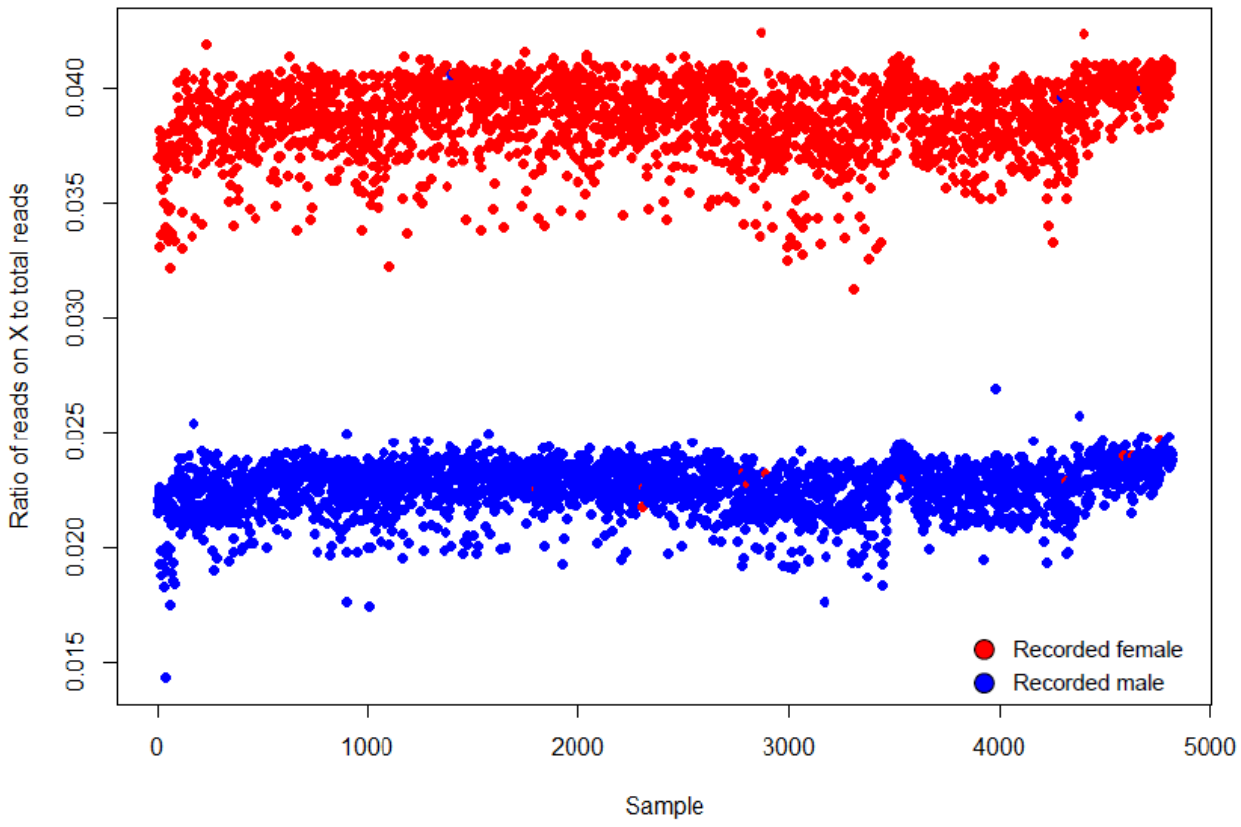
2.6 Contributions

This work was funded primarily by Dr. Abraham Palmer's P50 grant (**P50 DA037844)**. Additional funding came from the Sprague Dawley PavCA grant **(R21 DA036672)** and my individual support through my F31 (**DA039638-02**) and GRTG T32 (**GM007197**) training grant through the University of Chicago. I performed all optimization of the ddGBS sequencing library preparation protocol for use in rats. I also prepared the original HS library with 96 samples to test the efficacy of ddGBS in this population. Later HS libraries were constructed by Celine St. Pierre and Hannah Wladecki. Jianjun Gao performed the majority of the computational analyses presented within this chapter, including all variant calling and concordance checks with array data. However, we worked collaboratively to select the computational tools and parameters that ultimately went into analyzing the raw sequencing data and calling variants. Using our joint data, I composed the written transcript for this chapter.

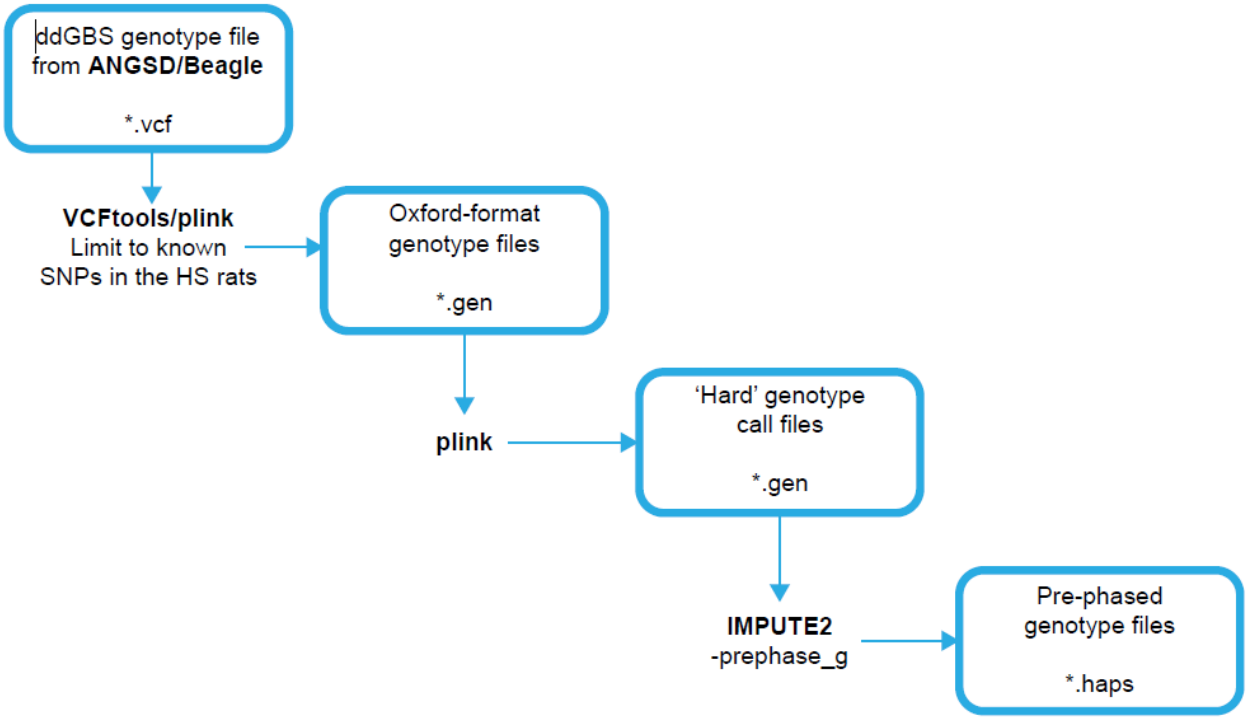
2.7 Appendix A: Supplemental Figures

Supplemental Figure 2.1. Ratio of reads on X-chromosome to total sequencing reads.

The color of the points indicates the pedigree-recorded sex of the samples. Females are expected to have approximately twice as many reads for the X-chromosome. Samples that did not cluster with their pedigree-recorded sex were removed from the study for possible sample mix-up.

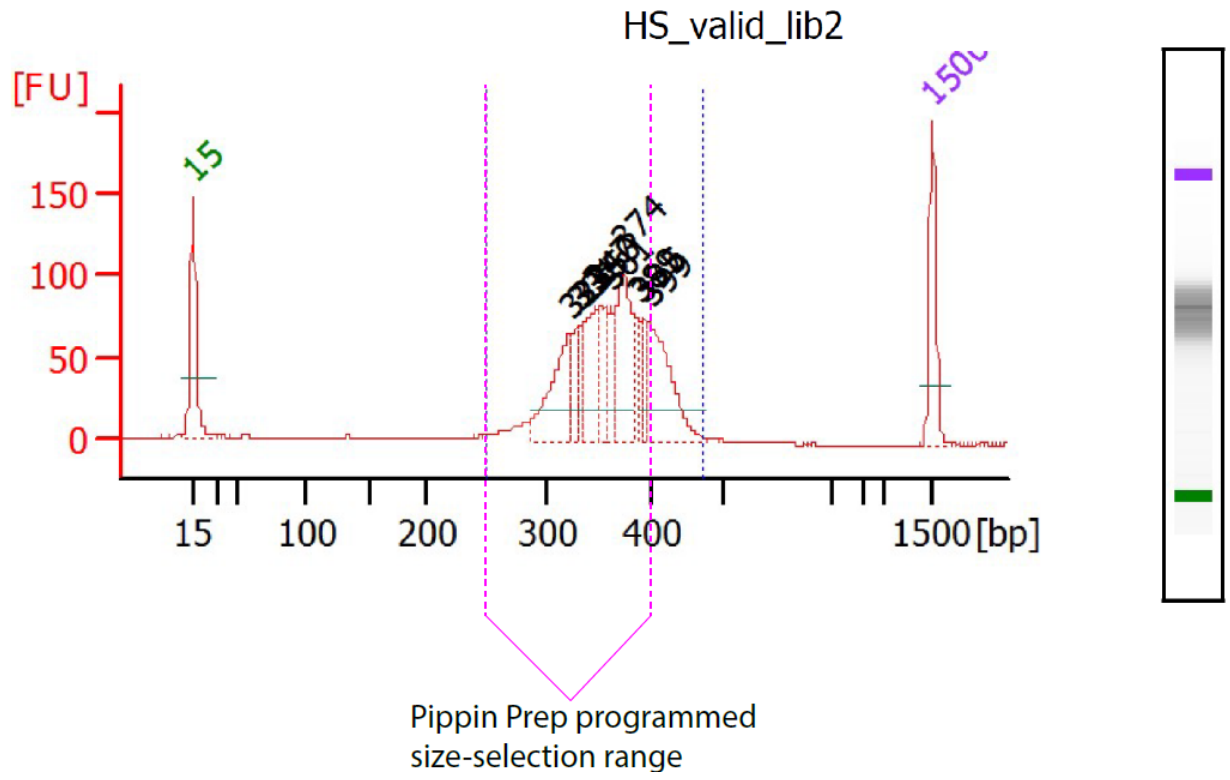


Supplemental Figure 2.2. Data preparation workflow for imputation with IMPUTE2.



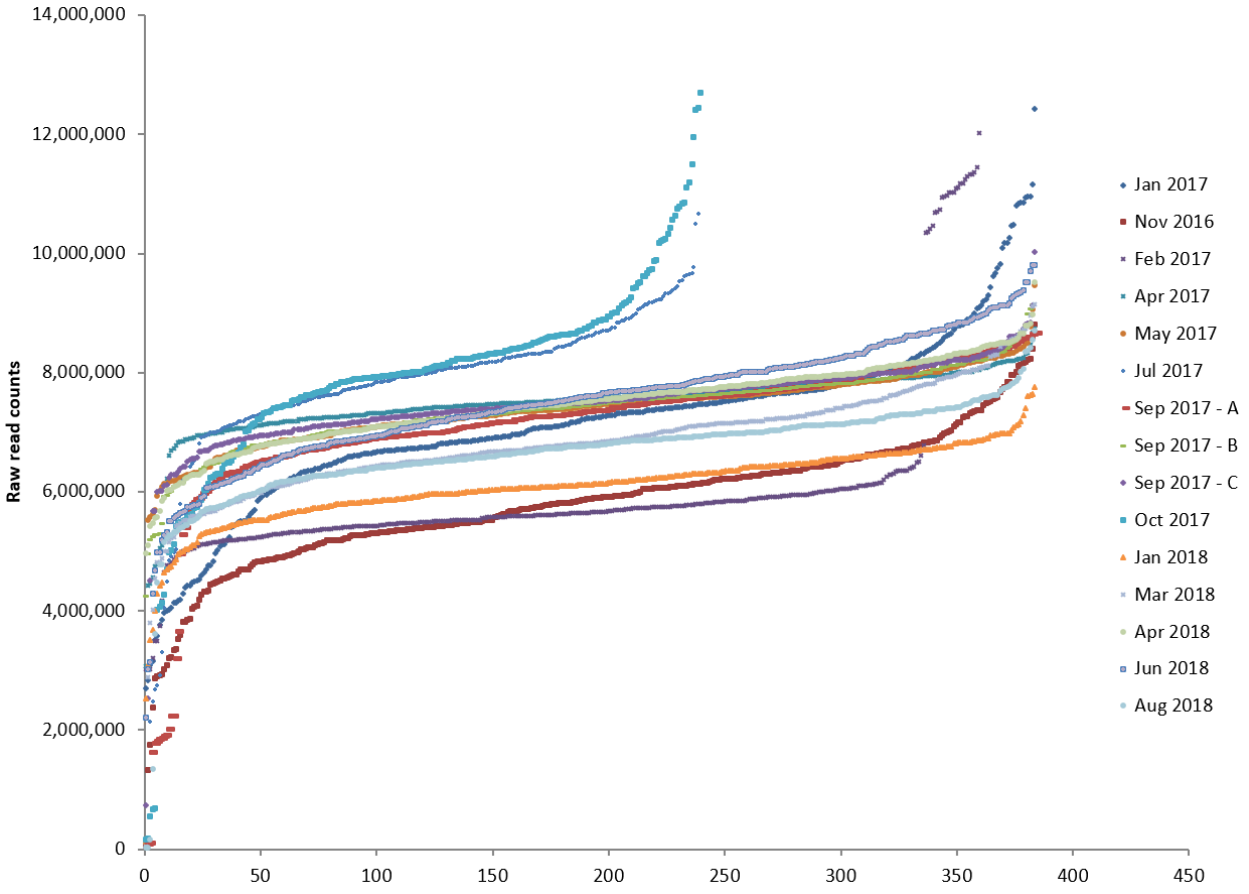
Supplemental Figure 2.3. Programmed vs. empirical Pippin Prep fragment size range.

This plot comes from the Bioanalyzer output for a pooled HS library. The x-axis shows the library fragment sizes in base pairs, and the y-axis is in fluorescent units, which represent the quantity of the fragments on the gel chip. There is approximately a 50-75bp shift in the empirical library distribution compared to expectation due to the high quantity of fragments loaded into the Pippin Prep gel cassette.



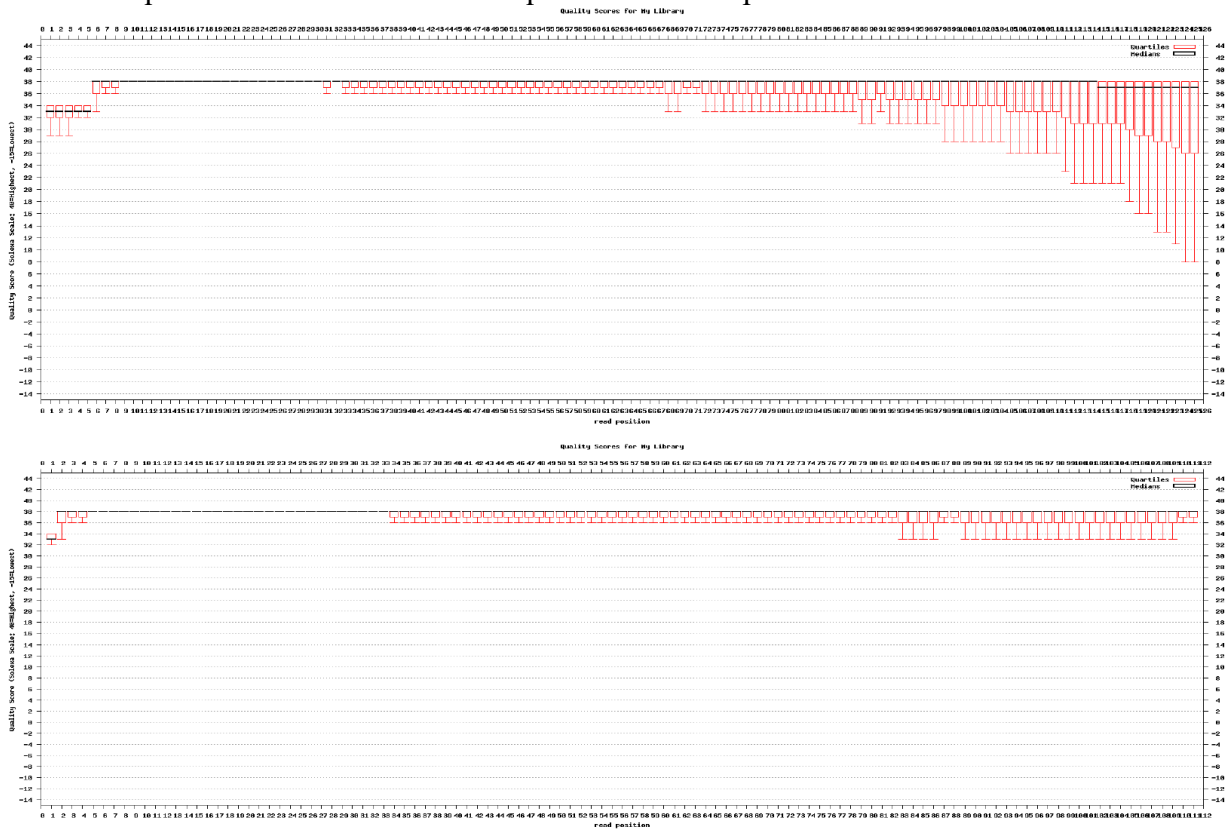
Supplemental Figure 2.4. Raw read counts grouped by shipment batch.

Raw read counts are on a per-sample basis after demultiplexing FASTQ files with FASTX Barcode Splitter.

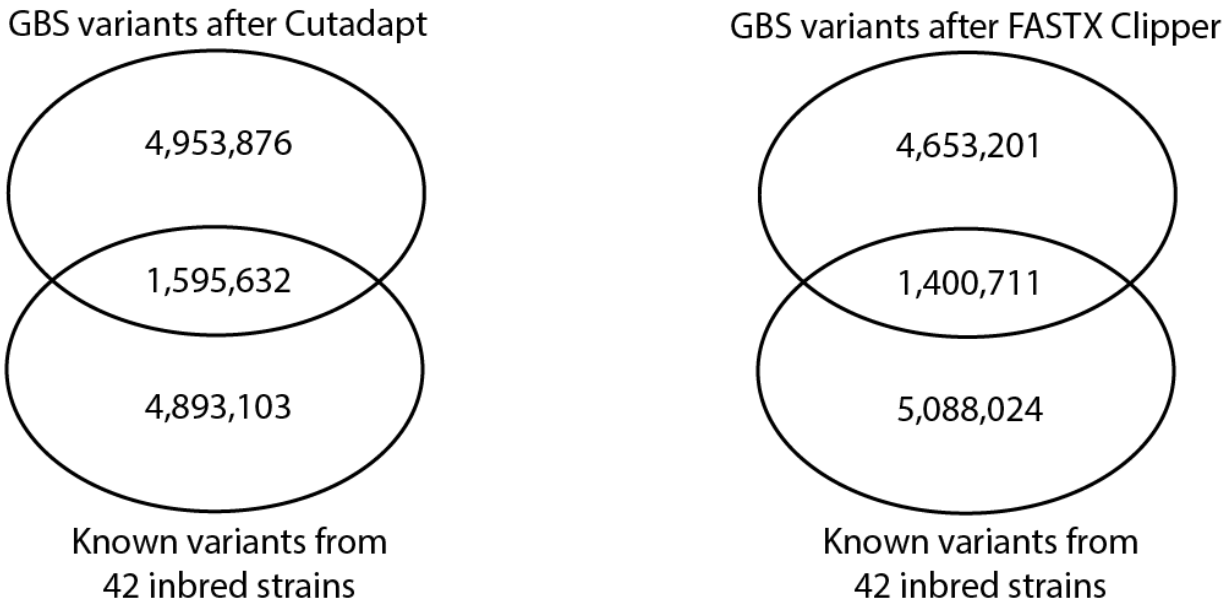


Supplemental Figure 2.5. FASTQC results pre- and post-filtering with Cutadapt.

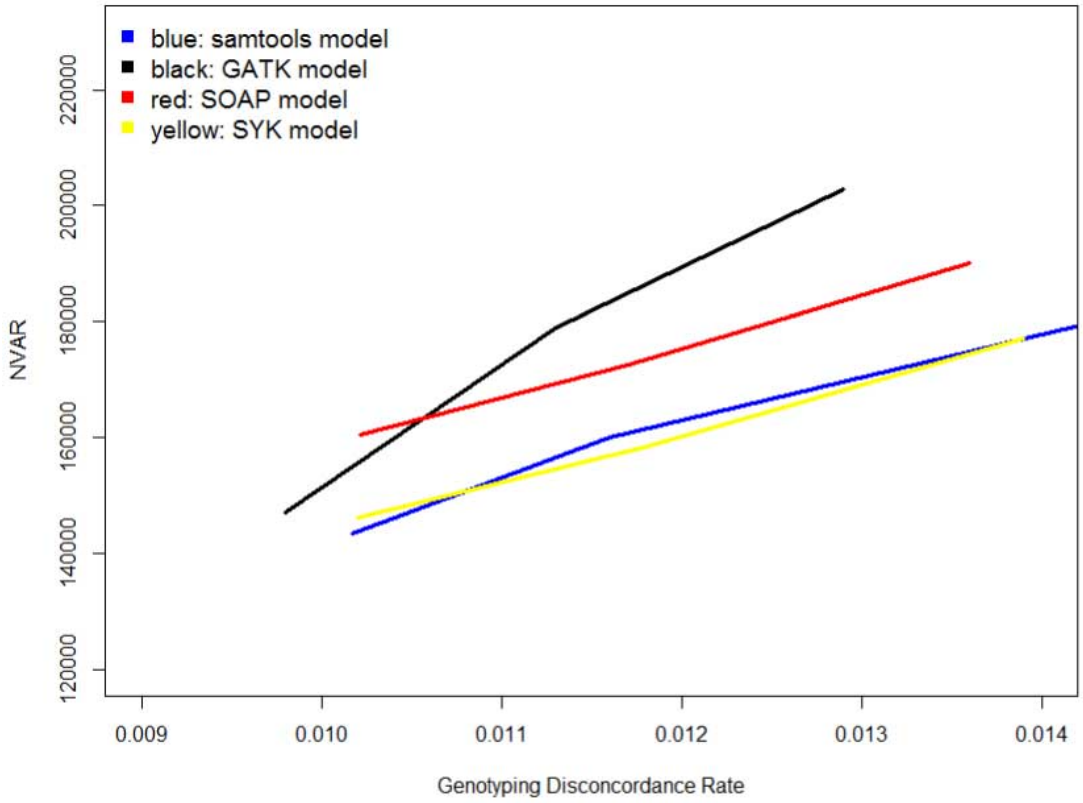
FASTQC results are from a single sample from the original set of 96 HS samples prepared in 12-plex and sequenced on the Illumina HiSeq 2500 with 125bp reads.



Supplemental Figure 2.6. Overlap of called SNPs with known variants after read trimming with FASTX or Cutadapt.

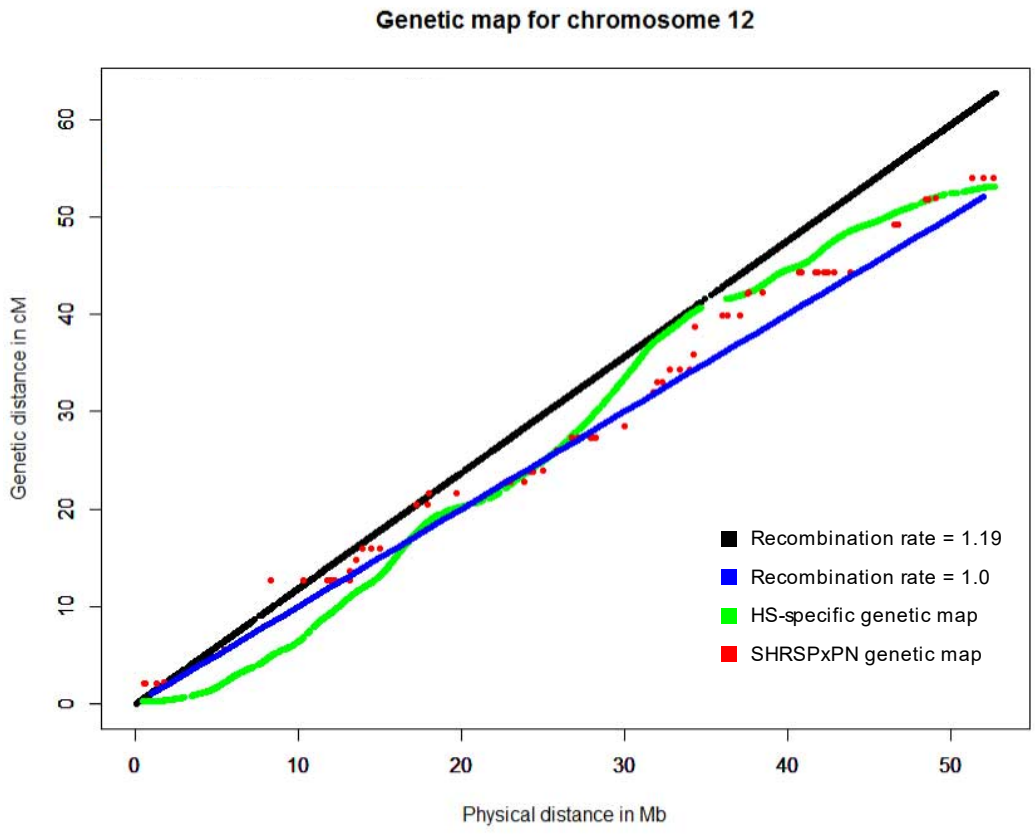


Supplemental Figure 2.7. Number of variants by genotype discordance rates for 4 ANGSD genotype likelihood models.



Supplemental Figure 2.8. Available rat genetic maps.

Plotted physical and genetic distances are for chromosome 12.



2.8 Appendix B: Supplemental Tables

Supplemental Table 2.1. Demultiplexing performance.

All methods began with the same number of reads from the original FASTQ. Final read and base pair counts are from after the reads have been trimmed of adapter, barcode, and restriction site sequences, as well as low-quality base pairs (< Q20).

	In-house Python Script	GBSX	FASTX Barcode Splitter
Reads with NlaIII adapter sequence	545,177 (3.07%)	475,581 (2.67%)	547,697 (3.07%)
Total bps processed	2,061,523,464	2,116,436,361	2,227,542,500
Total bps written to file	2,059,714,312	2,114,841,934	2,225,724,833
Proportion of bps retained	99.91%	99.92%	99.92%
Reads post-processing	17,771,754	17,786,280	17,820,340

Supplemental Table 2.2. Comparison of variants calls after filtering with FASTX vs Cutadapt.

Data shown comes from the original set of 96 HS samples prepared in 12-plex and sequenced on the Illumina HiSeq 2500. At this step of pipeline optimization, variants were called utilizing GATK UnifiedGenotyper. Calls were unfiltered.

	FASTX Clipper	Cutadapt
Number of variants	6,075,821	6,581,115
Genotyping call rate	0.17	0.19
Mean minor allele count	3.96	4.25
Mean minor allele frequency	0.15	0.15
Number of singletons	433,960	548,975
Number monomorphic sites	807,453	773,074
Transition/transversion ratio	2.32	2.40
$T_I T_V$ ratio for singletons	3.23	3.40
Mean variant read depth	109.56	126.35
Mean quality score	601.79	715.56

Supplemental Table 2.3. Variant metrics resulting from reads filtered at different mapping quality thresholds.

Data shown comes from the original set of 96 HS samples prepared in 12-plex and sequenced on the Illumina HiSeq 2500. Variants were called utilizing the SAMtools model and the -minMapQ filter in ANGSD. Calls were unfiltered.

	MAPQ = 20	MAPQ = 30	MAPQ = 45	MAPQ = 60	MAPQ = 90
Number of variants	372,860	372,330	363,790	316,949	233,322
Genotyping call rate	0.64	0.64	0.64	0.61	0.75
Mean minor allele count	5.96	5.96	6.06	5.86	7.36
Mean minor allele frequency	0.18	0.18	0.18	0.18	0.19
Number of singletons	16,781 (4.50%)	16,732 (4.49%)	16,550 (4.55%)	17,352 (5.47%)	11,773 (5.05%)
Number of monomorphic sites	122,478 (32.85%)	122,188 (32.82%)	116,738 (32.09%)	100,074 (31.57%)	56,179 (24.08%)
Transition/transversion ratio	1.23	1.24	1.26	1.31	1.41
$T_I T_V$ ratio for singletons	1.27	1.28	1.28	1.31	1.38
Mean variant read depth	157.78	157.73	159.25	152.48	188.80
Mean quality score	2,547	2,548	2,556	2,461	2,954

Supplemental Table 2.4. Transition/transversion ratio before and after known sites filtering.

The presented data comes from ANGSD/Beagle variant calls for 3,601 HS samples, prior to imputation with IMPUTE2. Known SNPs came from both the 42 inbred genomes from Hermsen et. al 2015 [177] and the 8 inbred HS founder strains sequenced by the University of Michigan [179].

	Unfiltered SNPs	Filtered for known SNPs
AC	15,157	9,166
AG	888,657	42,275
AT	15,432	7,610
CG	18,043	8,061
CT	893,653	41,938
GT	15,118	9,177
T_S	1,782,310	84,213
T_V	63,750	34,014
$T_S T_V$	27.96	2.48
Total # SNPs	1,846,060	118,227

Supplemental Table 2.5. Imputation accuracy for chromosome 12 across different genetic maps.

The number of variants used for the concordance check is dependent on the overlap of the imputed variants with array data for the 96 HS rats with array genotypes. The MAF filter only removes monomorphic sites within the 96 HS rat sample used for the concordance check.

	cM/Mb = 1.00	cM/Mb = 1.16	SHRSPxPN	HS-specific
Number of variants before QC	158,452	158,452	158,452	158,452
Genotyping rate before QC	0.94	0.92	0.92	0.92
Variant removed for missingness > 10%	22,217	28,959	28,356	28,858
Variants removed for MAF < 0.005	50,380	61,270	61,592	59,812
Variants removed for HWE < $1x10^{-10}$	53	56	57	56
Number of variants after QC	85,802	68,167	68,447	69,726
Genotyping rate after QC	0.93	0.91	0.92	0.91
Number of variants in concordance check	5,912	5,590	5,594	5,646
Discordance rate	0.095	0.011	0.011	0.010

CHAPTER 3

GENETIC CHARACTERIZATION OF OUTBRED SPRAGUE DAWLEY RATS AND UTILITY FOR GENOME-WIDE ASSOCIATION STUDIES

3.1 Abstract

Sprague Dawley (**SD**) rats are one of the most commonly used outbred rat strains. Despite this, the genetic characteristics of SD are poorly understood. We collected behavioral data from 4,625 SD rats acquired predominantly from two commercial vendors, Charles River Laboratories and Harlan Sprague Dawley Inc. Using double-digest genotyping-by-sequencing (**ddGBS**), we obtained dense, high-quality genotypes at 234,887 SNPs across 4,061 rats. This genetic data allowed us to characterize the variation present in Charles River vs. Harlan SD rats. We found that the two populations are highly diverged ($F_{ST} > 0.4$). We also used these data to perform a genome-wide association study (**GWAS**) of Pavlovian conditioned approach (**PavCA**), which assesses the propensity for rats to attribute motivational value to discrete, reward-associated cues. Due to the genetic divergence between rats from Charles River and Harlan, we performed two separate GWAS by fitting a linear mixed model that accounted for within vendor population structure and using meta-analysis to jointly analyze the two studies. We identified 18 independent loci that were significantly associated with one or more metrics used to describe PavCA; we also identified 3 loci that were body weight, which was only measured in a subset of the rats. The genetic characterization of SD rats is a valuable resource for the rat community that can be used to inform future study design.

3.2 Introduction

Rats are among the most commonly used organisms for experimental psychology and biomedical research. Whereas research using mice makes extensive use of inbred strains, in rats, it is more common to use commercially available *outbred* populations. Among the commercially available outbred rat populations, the Sprague Dawley strain (SD) is one of the most widely used. SD rats are distributed by several vendors. Each vendor has multiple breeding locations, and each breeding location has one or more barriers in which the rats are housed. Prior studies have identified numerous physiological differences between SD rats obtained from different vendors [101,103,105–107]. Despite these observations, many researchers appear to assume that SD rats obtained from different vendors or barrier facilities are largely interchangeable. There has been little research into the genetic diversity and population structure that exists among commercially available outbred rats [76]. Prior rat genetic studies have used F_2 and more complex, multi-parental crosses of inbred strains for QTL mapping and GWAS [36,40,94]; however, we are not aware of any such studies that have employed commercially available outbred rats. Recently, we and others have demonstrated the potential benefits and challenges associated with the use of commercially available outbred mice for GWAS [96,198], suggesting that similar studies in rats might also be of value.

SD rats originated in 1925 at the Sprague-Dawley Animal Company (Madison, WI), where they were created by a cross between a hooded male hybrid of unknown origin and an albino Wistar female [100]. In 1950, Charles River Inc. began to distribute SD rats. In 1980, Harlan Inc. (now Envigo, Inc.) began to distribute SD rats after their acquisition of Sprague-Dawley, Inc. [199]. In 1992, Charles River reestablished a foundation colony of SD rats, using 100 breeder pairs from various existing colonies [200]. The resulting litters were used to populate SD colonies globally and have since been bred using a mating system that minimizes inbreeding. Every 3 years,

Charles River replaces 25% of their male breeders in each production colony with rats from a single foundation colony. Charles River also replaces 5% of their foundation colony breeding pairs with rats from the production colonies on a yearly basis. These practices are designed to reduce genetic drift between the production colonies [201]. Harlan also reports using a rotational breeding system to limit inbreeding; however, more detailed information is not publicly available. Since Harlan's acquisition by Envigo, the process has become more transparent [202]. Envigo follows a Poiley rotational breeding scheme [203], whereby animals are cycled through different sections of the colony with each generation, reducing genetic drift and inbreeding.

Here we used SD rats from multiple vendors, breeding locations, and barrier facilities to elucidate the genetic background of SD and to perform a GWAS of components of a complex behavior. DNA samples were obtained from rats used in multiple studies as part of an unrelated Program Project grant (P01DA031656) concerned with individual variation in the propensity to attribute incentive value to food and drug cues [53,204]. All rats were first screened for Pavlovian conditioned approach behavior (PavCA) [205], which provides one index of the degree to which a reward cue has been attributed with incentive salience. Although the genetic analyses reported here were not part of the original design, we took advantage of the opportunities afforded by that large, behavioral study. We extracted genomic DNA from available tissue samples and then used double digest genotype-by-sequencing (ddGBS) to obtain dense genotypes for 4,625 SD rats. We used these genotypes to first genetically characterize different populations of SD, and then in conjunction with behavioral data from PavCA, to perform the largest rodent GWAS to date. Because most of the rats were obtained from two vendors (Harlan and Charles River) we performed two separate GWAS and combined the results using meta-analysis. Our results provide

insights about the population structure and suitability of SD for GWAS, and also explore specific loci associated with Pavlovian conditioned approach.

3.3 Methods

3.3.1 Sprague Dawley samples

Tissue samples from 5,206 male Sprague Dawley rats were obtained, predominantly from Charles River and Harlan, with a few samples from Taconic. A subset of 4,625 of these rats went on to be genotyped by ddGBS and/or WGS. After sample filtering, a final set of 4,061 genotyped SD rats were used for the population genetic and association analyses. Supplemental Table 3.1 lists the number of samples that came from each vendor, breeding location, and barrier facility. Detailed information about these 4,061 rats is available in Supplemental File 3.1 File. Behavioral testing was performed between February 2012 and August 2015 as part of work for multiple studies [78,204,206–212,212–217]. All experiments were approved by the University of Michigan IACUC. Housing, feeding, lighting and other relevant environmental conditions have been previously described. Following sacrifice at the University of Michigan, tissue samples were shipped to the University of Chicago; subsequent processing of those samples is described in the following sections.

3.3.2 Pavlovian conditioned approach

Pavlovian conditioned approach procedures have been thoroughly described previously [218,219] as a means to assess the tendency to attribute incentive motivational value or incentive salience to a cue that has been repeatedly paired with a noncontingent reward. Briefly, rats are placed into a testing chamber in which an illuminated lever (conditioned stimulus; CS) enters the chamber and after 8 seconds the lever-CS retracts and a food pellet (unconditioned stimulus; US) is immediately

delivered into an adjacent food cup. Rats are scored for their three possible responses to the lever-CS entering the cage: approach and interact with the lever, approach and interact with the food receptacle (magazine), or make neither approach. Conditioned responses are captured during the 8-second period during which the lever-CS has entered the chamber, but before the food reward enters the magazine. The following measures are obtained in automated fashion: the number of lever contacts as measured by lever depressions, number of magazine entries as measured by infrared sensor in the food receptacle, and the latency to both during the 8-second lever-CS presentation. The rats are tested in this manner with 25 trials per session and one session is conducted per day for 5 consecutive days. For the purposes of this project, the number of lever contacts and magazine entries are summed across all 25 trials within a given session, and the latencies are averaged across 25 trials within a session.

Along with response counts and latencies, three additional measurements are recorded: 1) the proportion of trials in a session during which a rat made a lever contact ("probability" of lever press), 2) the proportion of trials during which they made a magazine entry ("probability" of magazine entry), and 3) the number of non-CS (NCS) magazine entries that occurred outside of the 8 second trials (when the cue was not present during the intertrial interval). We also calculated composite scores to categorize rats as sign-trackers (ST; defined as rats that preferentially interacted with the lever-CS), goal-trackers (GT; defined as rats that preferentially interacted with the food magazine), and intermediate responders (IR; rats that vacillated between sign- and goal-tracking behavior) [205]. These scores include: response bias ([lever presses – magazine entries]/[lever presses + magazine entries]), latency score ([average magazine entry latency – average lever press latency]/8), and probability difference ([lever press probability – magazine entry probability]). The PavCA index score is the average of the response bias, latency score, and

probability difference. A value of [-1, -0.5] for the PavCA index score indicates a GT, (-0.5, 0.5) indicates an IR, and [0.5, 1] indicates a ST. We performed a Welch's 2-sample t-test to show that the PavCA index score distributions differed significantly between Charles River and Harlan SD rats ($t = 20.161$, $df = 3908.1$, p-value $< 2.2 \times 10^{-16}$). In summary, 11 PavCA metrics were available for analysis, each of which we measured on days 1, 2, 3, 4, and 5 (Supplemental Table 3.7).

3.3.3 Double digest genotyping-by-sequencing (ddGBS)

To obtain genotypes, we used ddGBS, a genotyping method that reduces the complexity of the genome by only sequencing regions proximal to restriction enzyme cut sites [76,112]. We have recently described the technical aspects of this protocol in detail [220]. The ddGBS protocol used in this paper is a synthesis of the GBS approach described in Graboski et al. [221] and used more recently by Parker et al. [96] and Gonzales et al. [157], and an analogous approach known as double digest restriction associated DNA sequencing (ddRADseq) [115].

DNA was extracted from rat tails using the PureLink® Genomic DNA kit. DNA purity was assayed using a Nanodrop 8000 ($260/280 \geq 1.8$) and DNA integrity by gel electrophoresis (minimal smearing). Genomic DNA was then digested using two restriction enzymes: *PstI* (6-bp recognition site) and *NlaIII* (4-bp recognition site). Adapter oligos were annealed to overhangs left by *Pst1* and *NlaIII*. The *PstI* adapters contained 48 unique 4-8 bp in-line indexes [96,157,221]. A Y-adapter was annealed to the *NlaIII* cut sites, which controlled the direction of the first round of PCR amplification and thus ensured that the library was primarily composed of fragments with one of each of the adapters. Post-annealing, sets of 24 individual sample libraries were quantified and pooled. Pooled libraries were PCR amplified for 12 cycles, size-selected for 300-450bp using the Pippin Prep, and quality checked by Agilent Bioanalyzer (peak range ~ 300-500bp and conc.

≥ 20nM). Sequences for the 48 barcoded adapters, Y-adapter, and PCR primers are provided in Supplemental Text 2.2.

Sequencing of pooled libraries was performed by Beckman Coulter Genomics (now GENEWIZ). Sequencing was carried out on the Illumina HiSeq 2500 using v4 chemistry and 125-bp single-end reads. Each lane consisted of a pool of 24 samples, resulting in an average of 8.9 million reads per sample. A total of 4,608 unique ddGBS sample libraries were sequenced. Of these samples, 384 were sequenced twice, resulting in two sets of sequencing data for each sample from the same library prep that were used for to check genotype concordance.

3.3.4 Light whole-genome sequencing

To discover new variants and support imputation, we performed low-coverage whole-genome sequencing (WGS) of 80 SD rats from this same cohort. The rats were selected to represent sign-trackers, goal-trackers, and intermediate responders from each of the major barriers within the 6 major subpopulations of Harlan and Charles River. Sample libraries were prepared using the Illumina TruSeq® PCR-Free Library Prep kit and quality checked using an Agilent Bioanalyzer and qPCR on an Applied Biosystems StepOne Real-Time PCR System to ensure they met Illumina quality standards. Sample pooling (10 samples per pool) was performed by Beckman Coulter Genomics. Each pooled library was sequenced on two lanes of an Illumina 2500 flow cell with 125-bp single-end reads, resulting in an average of 51.6 million reads per sample per two lanes. Assuming that the rat genome (rn6) is ~2.87Gb, this provided about 180x coverage of the rat genome, or about 2.25x coverage per rat.

3.3.5 ddGBS sequence data processing

We have recently described the bioinformatic steps that we use for ddGBS in detail [220]. We follow an analogous approach in this paper, though we deviate at the imputation step due to our use of SD instead of HS rats. Briefly, raw reads from ddGBS were demultiplexed using FASTX Barcode Splitter [170], allowing for 1 mismatch. After demultiplexing, barcodes were trimmed by *cutadapt* [173]. Any reads not matching a sample's barcode within 1-bp were filtered out. We removed 316 samples for which there were less than 4 million reads, leaving 4,292 samples with ddGBS data. We also used *cutadapt* to trim low-quality base pairs (phred quality score < 20) at the ends of the reads, and to remove 3' adapter sequences. Reads trimmed to less than 25-bp were discarded. Next, all reads were aligned to the rat reference genome assembly (rn6) using *bwa* [175]. All ddGBS reads were realigned at known indel sites by GATK's IndelRealigner [176]. Because of a lack of SD-specific variant data, we used variant data from 42 whole-genome sequenced rat strains and substrains [177] as the reference set for indels. We then used GATK to perform base quality score recalibration (BQSR) on the BAM files, using data (SNP & indel) from the 42 rat genomes as the "known" set of variants. For the ddGBS samples that were sequenced twice (381 remaining after filtering for read count), we performed all quality control and variant calling steps in parallel, since our goal was to compare calls made in these samples as a means of estimating the genotyping error rate.

3.3.6 Light WGS data processing

Raw reads from WGS were processed in an analogous manner to the ddGBS data (detailed above) through the alignment step. After alignment, duplicates reads were removed using *picard* [222]. Reads were then realigned and underwent BQSR. The final WGS BAM files from each lane (2

files per sample) were merged. The WGS BAM files for the 63 samples that had undergone both ddGBS and WGS were then merged.

3.3.7 Variant discovery and imputation with ANGSD/Beagle

We found that GATK's HaplotypeCaller tool [176] was ineffectual at making high-confidence SNP calls in our dataset, likely due to the unusual distribution of reads produced by ddGBS. Instead, we used the Samtools genotype likelihood model [196] as implemented by ANGSD [165] to estimate genotype likelihoods from the mapped ddGBS reads. Likelihoods were obtained in 10Mb chunks of the genome, which were subsequently concatenated. Major and minor alleles were inferred from the data based on allele frequency estimates made from the genotype likelihoods. The likelihoods were only estimated at sites where at least 100 samples had reads. ddGBS data results in low call rates at many loci. However, we retained these loci because we anticipated they would be useful for imputation. Next, we used the ANGSD genotype likelihoods to impute missing genotypes (that is for SNPs where only a portion of the rats had genotyping information) using Beagle [166,167], which produced a VCF file containing hard genotype calls (0,1, or 2), dosages ([0,2]), and posterior probabilities for each genotype ([0,1]) for 2,274,118 biallelic SNPs in 4,309 rats (ddGBS+WGS). This is the unfiltered set of SNPs and samples we moved forward with for all subsequent steps. We elected not to pursue variant quality score recalibration using the GATK VariantRecalibrator algorithm [189], because we did not have the required "truth" SNP set. Due to the poorly understood population history of SD rats, it was unclear whether the variation present in the 42 rat genomes would be representative of the variation present in our sample. Using the 42 genomes as a reference for the VariantRecalibrator may also have negatively impacted the calling of novel SD variants.

3.3.8 STITCH (Sequencing To Imputation Through Constructing Haplotypes)

In addition to variant calling and imputation using ANGSD/Beagle, we also explored the use of STITCH [223], since it is specifically designed for low-coverage WGS data lacking haplotype reference panels. However, ddGBS data is higher coverage and sparser than the input for which STITCH was designed. We used the set of alignment files and known variant sites from the 42 genomes [177], as described above. STITCH queries the user for the number of ancestral haplotypes that exist in the population (K). Due to our lack of knowledge about the founder population, we ran STITCH on a single chromosome using 5 different values of K: 2, 3, 4, 5, and 6. We found that K values of 3, 4, and 5 worked similarly well and chose K=4 to maximize the number of SNPs called and minimize error rate as ascertained by comparison of genotype calls between duplicate samples. STITCH yielded 8,691,886 biallelic SNPs, vastly more than were called with ANGSD/Beagle. However, after applying filters for dosage r^2/INFO score ≥ 0.9, MAF ≥ 0.01, HWE p-value $\leq 1x10^{-7}$, as well as removing sites in near perfect pairwise LD > 0.95, we found that the genotypes from STITCH contained fewer SNPs compared to the comparably filtered output from ANGSD/BEAGLE (see Supplemental Table 3.2). For this reason, we did not use the SNP calls made by STITCH in any of the analyses presented in this paper.

3.3.9 Genotype concordance check

Whereas some of our past projects that used GBS were able to determine the accuracy of GBS genotypes by comparing them to genotypes obtained from SNP microarrays, we did not have microarray-based genotypes for this cohort. Instead, we relied on the remaining 381 duplicate samples whose genotypes were called in parallel. To estimate genotyping accuracy, we compared

the rate of concordance of hard genotype calls among the duplicate samples. We first filtered variants by dosage r^2 (DR^2), a measure of the accuracy of the genotype imputation performed by Beagle. We tested three different DR^2 thresholds (≥0.7, ≥0.8, ≥0.9). We then removed variants with MAFs < 1% or that violated Hardy-Weinberg equilibrium at a threshold of $1x10^{-7}$ in either vendor population. Concordance rates were checked by two methods: 1) by using hard calls in the RAW format from *plink 1.9* and dividing the number of times a genotype call matched between duplicate samples by the total number of SNPs and 2) by taking the mean Pearson correlation of the dosages of the duplicate samples. The results are presented in Supplemental Table 3.3. Similar error estimates were obtained by the hard call and dosage approaches. We chose to move forward with the DR^2 threshold of 0.9 for all subsequent analyses, which yielded an error rate of ~0.85%.

3.3.10 Post-genotyping sample filtering

We removed female rats (n=77) and rats from Taconic Farms (n=4) since they made up a very small fraction of the total sample. We also removed rats that showed poor clustering in the PCA analysis, described below. We filtered out individuals with unusually high or low rates of heterozygosity and high degrees of relatedness as detailed below. Lastly, we excluded a small set of duplicate samples (n=7) and samples missing phenotype data for mapping (n=10). All filters and sample numbers are listed in Supplemental Table 3.5. After these steps, 4,061 unique male SD rats from Charles River (n=1,780) and Harlan (n=2,281) remained.

3.3.11 Principal component analysis, identity-by-descent, and heterozygosity

We performed principal component analysis (PCA) on the cohort of 4,228 samples filtered for low read count, Taconic, and females. PCA was performed on hard genotype calls in R using the

prcomp function in R [224] on a set of variants pruned for SNPs with MAF ≤ 0.05 in the combined sample set, SNPs in pairwise LD > 0.5, and SNPs violating HWE at a p-value < $1x10^{-7}$ in either Charles River or Harlan. The first PC clearly separated animals from Harlan and Charles River; however, there was a set of 54 rats that did not visually cluster as expected at the level of vendor or subpopulation (data not shown). These animals were removed from all subsequent analyses (we assumed they reflected inaccurate records, sample mix-ups, or some other technical problems).

With this further reduced set of 4,174 rats, we continued on to assess the genetic relationships among the rats in our sample. SD rats were ordered in multiple batches over several years, and we suspected some of these rats would be closely related (siblings, cousins, etc.). We reapplied the variant filters used for PCA and utilized the --*genome* function in *plink 1.9* on the resulting SNP set to estimate $\hat{\pi}$ (the proportion of genotypes predicted to be identical by descent), for all pairs of samples. Panels A and C in Supplemental Figure 3.4 show that while most of the animals were unrelated, there were a significant subset of closely related pairs, as well as some pairs with exceedingly high IBD1 rates in Harlan.

We used the *plink 1.9* function --*het* to examine possible inflation or deflation of heterozygosity rates in our samples. Panels A and C in Supplemental Figure 3.8 show that a handful of samples in both populations with uncharacteristically high (> 0.25) or low (< 0.25) rates of heterozygosity. We filtered out 34 such samples, as we found they drove the majority of the anomalous signal in our pre-filtering plots in Supplemental Figure 3.4. We also removed 12 samples with more than 30 close relations (defined as $\hat{\pi} \geq 0.1875$) and 32 samples that had a $\hat{\pi} \geq 0.6$ with another sample (likely sample contamination/mix-up). After applying these sample filters, we were left with Panels B and D in Supplemental Figures 3.4 and 3.8.

After reaching our final set of 4,061 rats, we reapplied the SNP filters on the reduced sample set, resulting in 4,502 SNPs that were polymorphic in both populations. We then ran PCA as described above. We also repeated PCA on the samples from Charles River and Harlan separately to examine substructure within each population.

3.3.12 Final variant filtering and minor allele frequency spectrum

After establishing the final cohort of 4,061 rats, we sought to establish a final set of SNPs to be used for the association analyses. First, we removed sites with $DR^2 < 0.9$. We then separated the Charles River and Harlan samples for all subsequent variant filtering performed in *plink 1.9*, which converts the posterior probabilities we received from Beagle to hard genotype calls, so long as the probability is ≥ 0.9. In each population, we removed SNPs with MAF < 0.01 using the *–maf* option in *plink 1.9*. Lastly, we used the *--hardy* function in *plink 1.9* [225] to perform tests for Hardy-Weinberg equilibrium. We did this for samples from each major subpopulation in each vendor (for Harlan: Frederick, Haslett, and Indianapolis; for Charles River: Portage, Raleigh, and St. Constant) because the population structure would have inflated the HWE statistics. We removed all SNPs with an HWE p-value < $1x10^{-7}$ in any of the 6 subpopulations. At the end of this process, we retained a total of 214,309 SNPs in Charles River and 114,568 SNPs in Harlan. A summary of the filtering process is shown in Supplementary Table 3.2. We used the *--freq* option in *plink 1.9* to estimate the MAFs for these final filtered sets SNPs. The distributions are show in Figure 3.1D.

Taking the union of these two sets of SNPs, there were a total of 234,887 polymorphic sites passing our filters between the two populations. Our final set of filtered SNPs contained 105,638 novel SNPs that had not be previously reported by Hermsen et al. [177]. We also compared our final set of SNPs to the most recent dbSNP release for rats (Build 149, November 7, 2016) and

found that 195,299 of the 234,887 SNPs we discovered were not present in the current dbSNP build.

3.3.13 Fixation Index (F_{ST})

To quantify the population divergence between SD rats from Harlan and Charles River, we computed the fixation index (F_{ST}) for each SNP in the union set of 234,887 using *smartpca* within the *EIGENSOFT* package [120,226,227]. Due to the substructure within vendor and vendor subpopulation that we saw in our PCA analyses, we chose to approach the F_{ST} estimation in three different ways. We first grouped the rats solely on vendor to calculate F_{ST} between Harlan and Charles River rats. Then, we broke these populations down further into the 3 major breeding facilities in each of the two vendors (listed previously) and calculated the pairwise F_{ST} between each sample set. Finally, we split the breeding facilities into the barriers that composed them, treating each barrier as a separate population. In each case, we computed all possible pairwise F_{ST} values. Samples from poorly represented barrier facilities were removed from the latter 2 analyses.

3.3.14 Linkage Disequilibrium

We plotted the decay of linkage disequilibrium (LD) using the *--r²* utility in *plink 1.9* and the procedures described in Parker et. al 2016 [96]. Briefly, each population's curve included all SNPs with MAF > 20% and pairwise LD comparisons were restricted to SNPs with allele frequencies within 5% of each other. An average r^2 estimate was obtained using 10,000 randomly selected SNP pairs from each 100kb interval for the distance between two SNPs, starting with 0-100kb and end at 9.9-10Mb. Harlan's curve used 35,770 autosomal SNPs in 2,281 rats and Charles River's used 112,678 autosomal SNPs in 1,780 rats. The CFW mouse curve was downloaded from a

repository for Parker et al. (http://datadryad.org/resource/doi:10.5061/dryad.2rs41) [96] and used for comparison as another commercially available, outbred rodent stock.

3.3.15 LMM covariates and phenotype data pre-processing

To select covariates for the GWAS, we performed univariate linear regression for each potential covariate for each PavCA metric. This was done separately for rats from Charles River and Harlan. Any covariates that accounted for 1% or more of the variance for at least one PavCA metric were passed into multivariate model selection with the R package *leaps* [228]. Within *leaps*, we performed exhaustive search for the best subset of variables using the branch-and-bound algorithm. Model selection with *leaps* was performed for all metrics for all days of testing, as well as the average of days 4 and 5. Out of the 66 models, all covariates that had surpassed the 1% threshold to reach this step were ultimately selected in at least 40% of the *leaps* models. For consistency, this led us to simply use the full set of covariates for all downstream GWAS. These covariates were included: age at testing, housing (binary – single or multiple), light cycle (binary – standard or reverse), and a set of binary 'indicator' variables to model the effects of different experimenters/technicians (10 variables were used for Charles River and 7 were used for Harlan). All covariates were included in the LMMs for association testing by GEMMA, rather than being regressed out prior to GWAS. We excluded an additional 88 rats from the final association analyses because of missing covariate data.

Many of the PavCA metrics were exceedingly non-normally distributed. In most cases this was expected due to how the behaviors were measured and defined. For example, the rats only had a window of 8 seconds in each trial during which to contact the lever and/or make a nose poke into the food magazine. All values for "average latency to lever press" or "average latency to magazine

entry" were therefore necessarily between 0 and 8 seconds. As is typical for latency traits, many of the values were near 0 or exactly 8. Similarly, the "probability" of lever press or magazine entry were very skewed towards the limits 0 and 1, especially after conditioned responding had been established on the later test days of training. Given these unusual distributions, we chose to quantile normalize all metrics prior to association testing, accepting a possible loss in power since samples with identical values are ranked randomly during the quantile normalization procedure.

In prior studies, sign trackers, intermediate responders, and goal trackers are categorized based on their behavior (i.e. PavCA) on days 4 and 5 of training [205]. In an attempt to emulate this approach, we created a binary trait (essentially case/control: ST/GT) and treated the intermediate responders as unknown. The threshold of PavCA index score typically used to define ST and GT are 0.5 and -0.5, respectively. We considered three options: 0, +/-0.25, & +/-0.5; for the 0 threshold all rats were either ST or GT, whereas for the other two thresholds some rats were intermediate, so they were excluded from the analysis.

Lastly, we explored the use of principal component analysis (PCA) to summarize the 55 metrics (Supplementary Table 3.7). We regressed out all covariates mentioned above (age, housing, light cycle, & experimenters) and mean-centered and quantile normalized the residuals. We ran PCA on the normalized residuals in 3 different ways: all 55 metrics simultaneously, all 11 metrics for each of the 5 days, and each of the 11 metrics for all 5 days. We choose the latter two approaches because we reasoned that there may be a specific day, or a metric by itself across days, that was uniquely of interest to the behavior. We only mapped the first PC for each, as it accounted for the majority of the variation for all metrics/days. Raw factor loadings and the percentage of the variance in the PCs explained by each metric for the first 5 PCs in each population are summarized in Supplemental File 3.2. The percent of behavioral variation explained by each of the first 10 PCs

in each population is shown by Supplemental Figure 3.5. Figure 3.5 graphically shows the correlations between each factor and the first two PCs in each population using the *factoextra* package in R [229]. Since the first 3 PCs cumulatively accounted for more than 70% of the variance in both populations, we only considered these for mapping.

A subset of 1,858 rats (Charles River n=957; Harlan n=901) also had body weight measurements, taken when the shipments were received from the breeding facilities, but prior to the onset of behavioral testing. We used this data for mapping.

3.3.16 SNP-based heritabilities

Heritabilities were estimated separately for Charles River and Harlan with the union set of 234,887 SNPs. We used this SNP set to construct genetic relationship matrices (GRMs) for each population using GCTA [230]. We then used the restricted maximum likelihood (REML) approach within GCTA on the GRMs, covariates, and quantile normalized Pavlovian conditioned approach data to calculate the SNP-based heritabilities for each metric, as well the previously mentioned PCs.

3.3.17 GWAS

We used GEMMA [123], which implements an LMM for GWAS analysis. We included a GRM as a random term to account for relatedness and population structure. Though beneficial for preventing false positive associations, GRMs can also reduce power to detect QTLs in populations with greater levels of LD; this is due to proximal contamination [231,232]. To avoid this reduction in power, we used the leave-one-chromosome-out (LOCO) approach [157,233,234]. As described above, we selected covariates that were included as fixed effects in our model (listed in

Supplemental Table 6). When mapping PCs 1, 2, and 3 for all 55 metrics, the fixed covariates were excluded from the LMM, as they had already been regressed out prior to calculation of the PCs. For the body weight GWAS, only age was used as a fixed covariate in the model. Any samples missing measurements for a given metric were removed from that analysis. For all GWAS, genotypes were represented as dosages (continuous [0,2]) in lieu of 'hard' genotype calls [0, 1, 2] to account for uncertainty in the genotype calls. Reported p-values come from the likelihood-ratio test (LRT) performed by GEMMA. Results were plotted using a custom R script and LocusZoom plots were created using the stand-alone software [235], custom SD SNP databases and LD calculations, and a 1.5Mb flanking region.

3.3.18 GWAS meta-analysis

We used the beta estimates and LRT p-values to perform a meta-analysis on the sample sets from Harlan and Charles River using the method outlined by Myers et al 2014 [139]. Analysis was limited to the set of 93,990 SNPs that existed at a frequency of at least 1% in both Harlan and Charles River. The allele frequencies, imputation accuracy (DR^2), and sample sizes were factored into the weighting for the z-statistics, which were then summed across the two GWAs. All analyses were completed in R [224].

3.3.19 Significance thresholds

In human GWAS, $5x10^{-8}$ is widely used as a significance threshold [236]. However, model organisms have widely different levels of LD, meaning that the effective number of independent tests differs between studies. Therefore, many prior studies have used permutation testing [124,237], where phenotype data is shuffled with respect to fixed genotypes. The computational

load of such methods becomes intractable for studies with large sample sizes and several traits being mapped [231]. Thus, we used the sequence of SLIDE [238,239] and MultiTrans [240] to obtain significance thresholds. We used separate thresholds for Charles River and Harlan because the number of SNPs, the LD structure, and the marker-based heritabilities were different. An advantage of this approach is that only one threshold was needed. We used a sliding window of 1000 SNPs and sampled from the multivariate normal 10 million times to obtain a 0.05 significance threshold of 8.96×10^{-7} ($-\log_{10}(p) = 6.05$) for Charles River and 2.19×10^{-6} ($-\log_{10}(p) = 5.66$) for Harlan. We calculated the thresholds for body weight separately due to its substantially greater heritability; however, the differences in thresholds were negligible (8.82×10^{-7} for Charles River and 2.13×10^{-6} for Harlan). We had difficulty obtaining a significance threshold for the meta-analysis because MultiTrans cannot accommodate such a situation and permutation would have been computationally expensive. Therefore, we used the threshold for Harlan for all meta-analyses, even though the meta-analyses utilized ~20k fewer SNPs, meaning that this threshold is likely to be overly conservative.

3.3.20 Power analysis

Power analysis was performed in Quanto v 1.2.4 [241,242] using the model for testing an additive genetic effect for a quantitative trait in independent individuals. We estimated our power with a fixed sample size of 2,000, which was approximately the midpoint for the Harlan and Charles River sample sets. We regressed out the covariates in our LMM from the quantile normalized metrics and calculated the mean and standard deviation from the residuals ($\mu \approx 0$, $\sigma \approx 0.98$). The resulting curve is show in Supplemental Figure 3.7.

3.4 Results

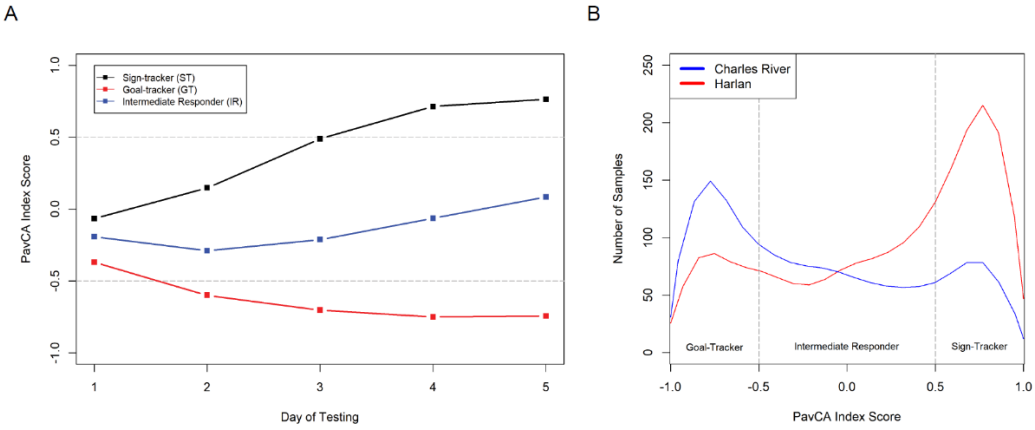
3.4.1 Phenotype

Our final dataset consisted of 4,061 genotyped individuals that were also phenotyped for Pavlovian conditioned approach; 2,281 from Harlan and 1,780 from Charles River, from 5 and 4 different breeding locations, respectively. As noted previously [205], we found that the metrics used to describe performance in PavCA are highly correlated (Supplemental Figure 3.1). Additionally, several of the base and composite PavCA metrics had tail-heavy distributions due to biased responding in sign- and goal-tracking from the animals during the testing periods (Supplemental Figure 3.2). For this reason, we chose to quantile normalize all measurements prior to mapping. The PavCA index score [205], which has been used previously to categorize rats into sign-trackers (**STs**), goal-trackers (**GTs**), and intermediate responders (**IRs**), showed the expected divergence and stabilization for STs and GTs over the five days of testing (Figure 3.1A). We therefore focused initial analyses on days 4 and 5 of testing and the average of days 4 and 5 as previously reported [205].

Charles River and Harlan rats had significantly different average PavCA index score distributions (Figure 3.1B; Welch's 2 -sample t-test, p-value < 2.2×10^{-16}), with Charles River rats biased towards goal-tracking and Harlan rats biased towards sign-tracking. We divided the samples further into the breeding locations that the rats originated from (Supplemental Table 3.1). The differences in PavCA index score between breeding locations within vendor were smaller, but still significant (Supplemental Figure 3.3) [76].

Figure 3.1. PavCA index score progression across days and distribution between Charles River and Harlan.

(A) Samples were classified as ST, GT, or IR based on their average day 4 and 5 PavCA index score. PavCA index scores for each day of training (1-5) were averaged across all STs, GTs, and IRs. (B) Density curves of average day 4 and 5 PavCA index score in Harlan (n=2,281) vs. Charles River (n=1,780) SD rats.



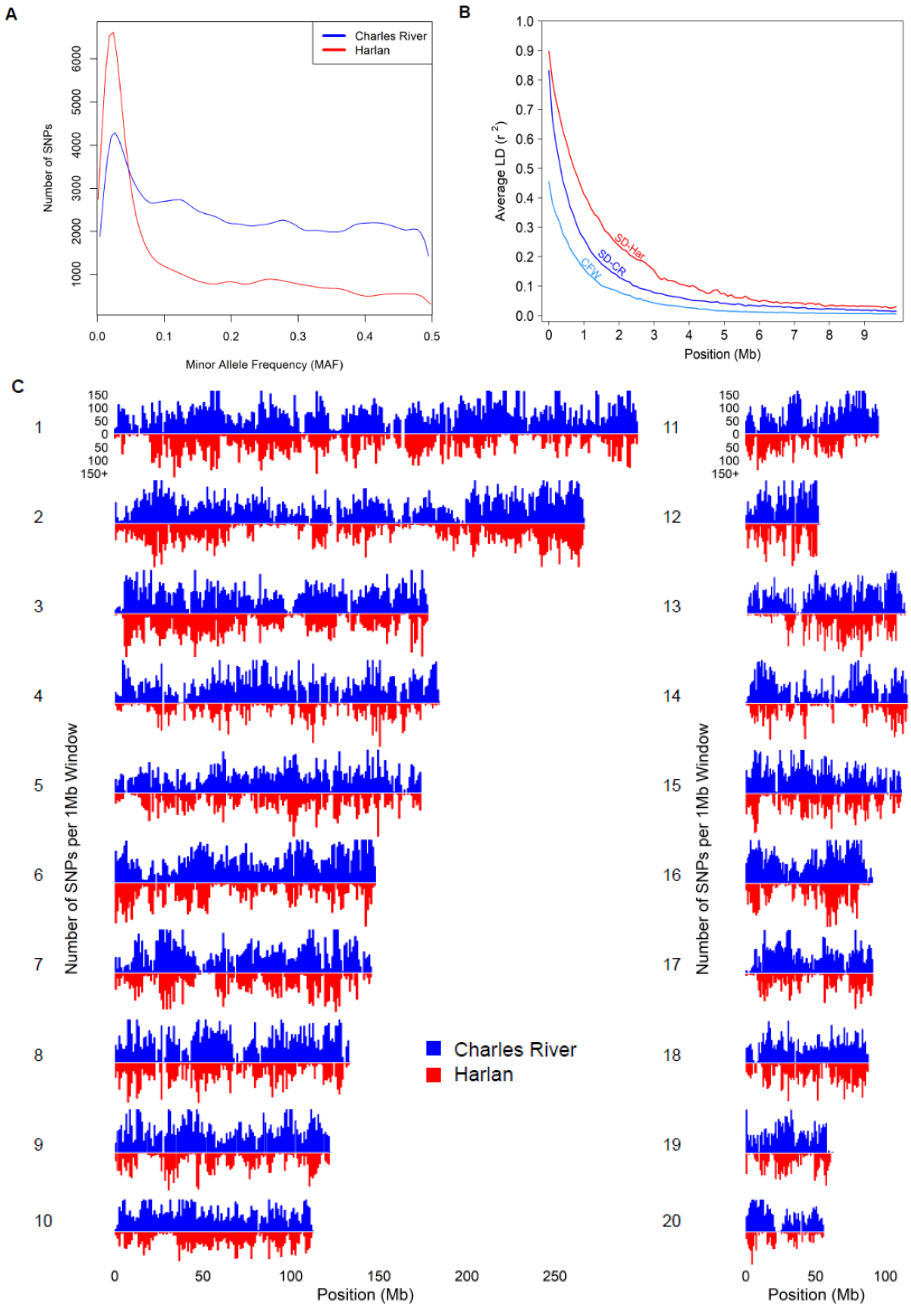
3.4.2 Genotyping and genetic characterization of SD rats

We identified more single nucleotide polymorphisms (SNPs) for the rats from Charles River compared to Harlan (214,309 vs 114,568; Supplemental Table 3.2). Figure 3.2C compares the distribution of SNPs across each chromosome for each vendor. There were some regions where Harlan had few SNPs, but Charles River had many, and other regions where both Harlan and Charles River had few SNPs. We also observed a large difference in the minor allele frequency (MAF) distributions for Charles River and Harlan (Figure 3.2A), with Charles River having a far greater proportion of SNPs with high MAF (>0.05). This observation could reflect the fact that Charles River adhered to their International Genetics Standardization Protocol for 25+ years, whereas Harlan appears to have focused on maintaining diversity within breeding colonies and allowed for a moderate degree of drift between them. After combining the two SNP sets, we

identified a total of 234,887 unique, bi-allelic SNPs. Using the 381 duplicate samples to evaluate genotyping accuracy, we calculated the discordance rate to be 0.85% (Supplemental Table 3.3).

Figure 3.2. Genetic architecture of SD rats from Charles River vs. Harlan.

(A) Density curves of minor allele frequencies for 214,309 SNPs in Charles River and 114,568 SNPs in Harlan, after removing SNPs with MAF < 0.01. (B) Linkage disequilibrium decay rates in SD rats from both vendors and outbred Swiss Webster (CFW) mice. (C) Filtered SNP density per 1Mb window in Charles River vs. Harlan samples for all 20 autosomes.



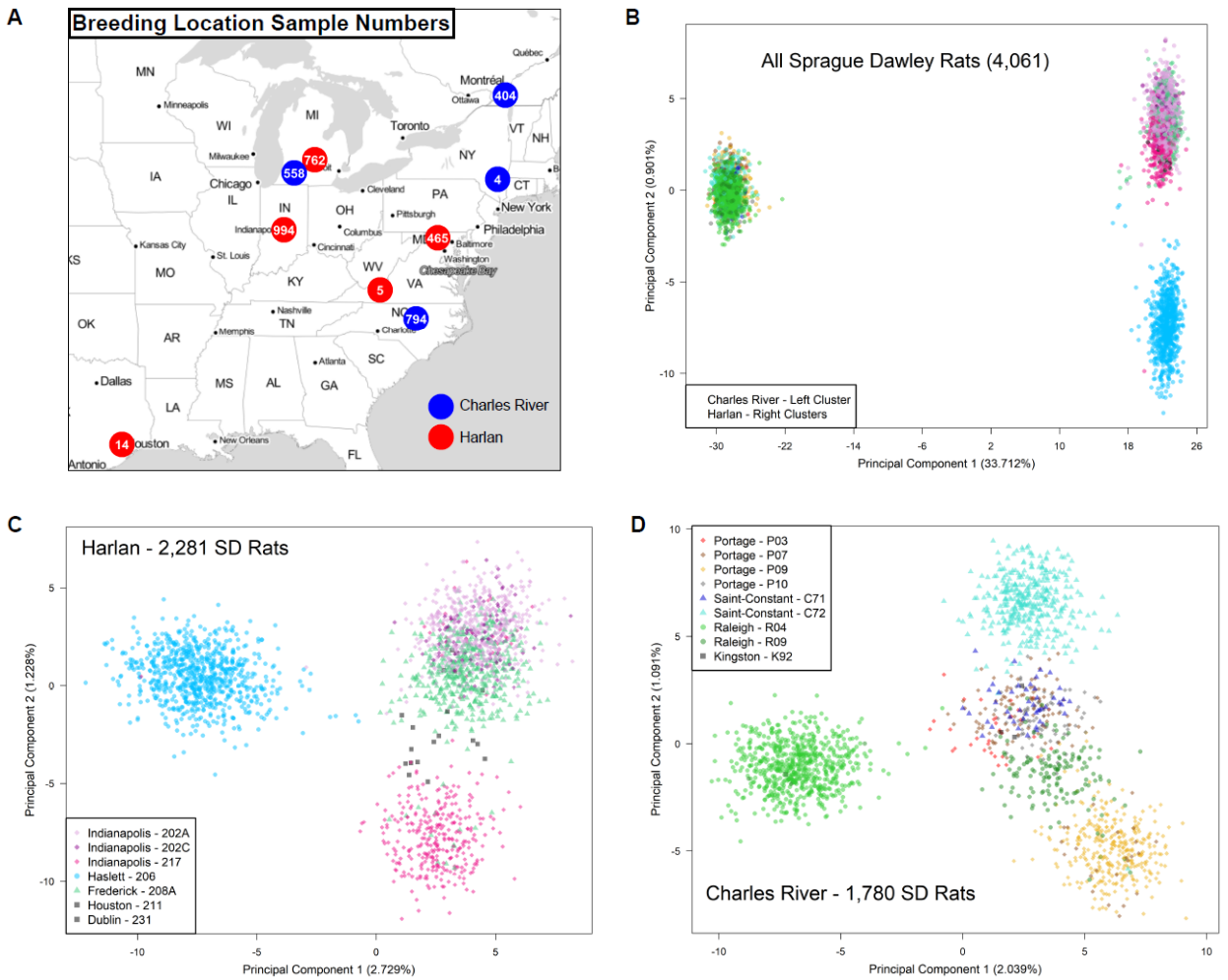
In addition to the variants described above, which were obtained using ANGSD/Beagle, we also used STITCH as an alternative genotyping pipeline [223]. This approach identified a larger total number of variants (Supplemental Table 3.2); however, after pruning SNPs with high linkage disequilibrium (LD; $r^2 \geq 0.95$), STITCH produced fewer SNPs compared to ANGSD/Beagle. Preliminary GWAS suggested the two SNP sets produced broadly similar results. Ultimately, we chose to use the ANGSD/Beagle variants for all subsequent analyses.

To examine the levels of linkage disequilibrium (LD) in Charles River and Harlan, we constructed LD decay curves (Figure 3.2B). These curves show the rate at which LD between two genetic loci dissipates as a function of the distance between the loci. Harlan rats had more extensive LD compared to Charles River. For contrast, we included the decay curve for the Swiss Webster (CFW) line, a commercially available outbred mouse population that has been successfully used for GWAS in the past [96,198].

To further investigate the observed genetic divergence between Charles River and Harlan, we performed principal component analysis (PCA) on a set of 4,502 LD-pruned ($r^2 < 0.5$) SNPs with MAF > 0.05 across both populations (Figure 3.3B). The first PC corresponded to vendor (Charles River or Harlan) and accounted for ~33.7% of the variance. The second PC accounted for just ~0.9% of the variance and reflected population structure within Harlan SD rats. To investigate within vendor structure further (Figure 3.3A), we performed PCA on the samples from each vendor separately using the same set of SNPs. Panels C and D of Figure 3.3 show evidence of substructure at both the level of breeding location (i.e. the city) and barrier facility (i.e., the segregated breeding areas within the building). Interestingly, rats from some barrier facilities showed greater differentiation from barrier facilities within the same breeding location than barriers in other locations.

Figure 3.3. SD population structure and comparison of linkage disequilibrium decay rates.

(A) Map of the nine vendor breeding locations and the number of SD rats obtained from each location. (B) A summary of the genetic data from all 4,061 SD rats based on principal components 1 and 2 from PCA. Each point represents a sample. The left cluster is composed of samples from Charles River and the right clusters are composed of samples from Harlan. (C-D) Repeated PCA analyses on subsets of the samples from Harlan and Charles River, colored by barrier facility of origin.



The fixation index (F_{ST}) is a statistic widely employed by population geneticists to measure the level of structure in populations [243]. It is calculated using the variance in allele frequencies among populations; values closer to 0 indicate genetic homogeneity, and values closer to 1 indicate

genetic differentiation. We calculated the F_{ST} between all Harlan and Charles River breeding locations (Table 1) and barrier facilities (Supplemental Table 3.4) with a sufficiently large number of samples (N > 30). F_{ST} values between vendors were ~0.435, which is much higher than corresponding values for major human lineages [227], whereas the values for different breeding locations within a vendor were substantially lower (Supplemental Table 3.4).

Table 3.1. Pairwise F_{ST} statistics for Harlan and Charles River breeding locations.

	Harlan			Charles River		
Breeding Locations	Frederick, MD	Haslett, MI	Indianapolis, IN	Portage, MI	Raleigh, NC	St. Constant, CAN
Frederick, MD	0	0.028	0.026	0.437	0.434	0.440
Haslett, MI		0	0.009	0.435	0.432	0.438
Indianapolis, IN			0	0.432	0.429	0.434
Portage, MI				0	0.016	0.014
Raleigh, NC					0	0.018
St. Constant, CAN						0

We speculated that some of the rats might share close genetic relationships with one another. We used *plink 1.9* [244,245] to estimate the pairwise proportions of identity-by-descent (IBD; Panels A and C in Supplemental Figure 3.4), which showed that while most rats were only distantly related, a subset shared closer familial relationships. We removed several rats that showed high levels of relatedness with many other samples (presumably due to technical error), as well as any with unreasonably high levels of IBD (S5 Table; Panels B and D in Supplemental Figure 3.4).

3.4.3 SNP heritability and genome-wide association analyses

Although evidence from selective breeding studies has suggested that behavior in the Pavlovian conditioned approach procedure is heritable [75], we are not aware of any specific heritability

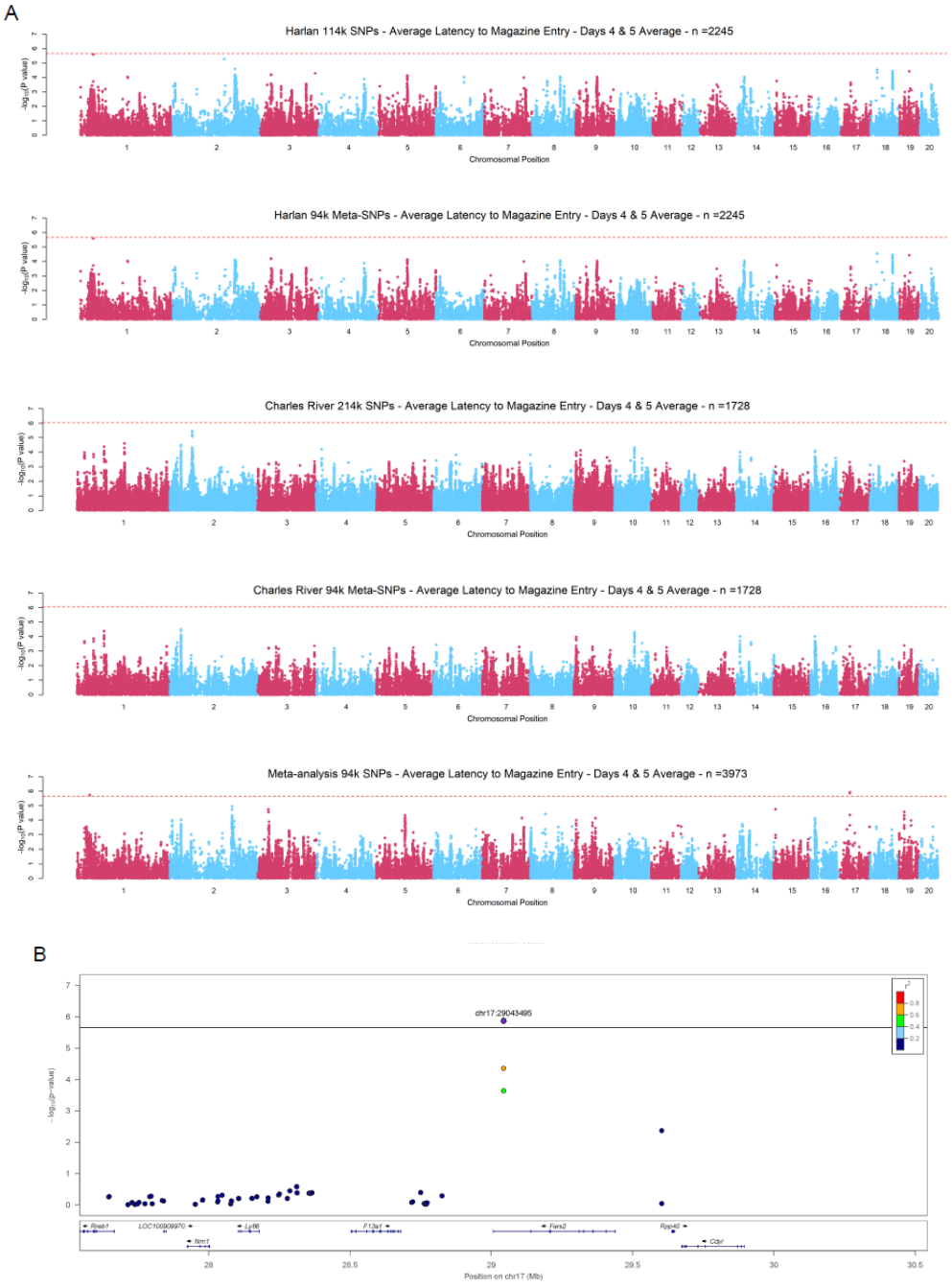
estimates using inbred strains or outbred populations. We used GCTA to calculate the proportion of the variance in the base and composite PavCA metrics that could be explained by the union set of variants from Harlan and Charles River (Supplemental File 3.3). The SNP heritability estimates for all PavCA metrics in Harlan ranged from ~4-11%, whereas they were ~4-21% for Charles River; on average the estimated heritability was about ~1.9-fold greater in Charles River. Importantly, some of the highest heritabilities were for metrics used to designate sign-trackers vs. goal-trackers, such as average of day 4 and day 5 response bias, probability difference, and PavCA index score. However, even the heritability estimates from Charles River were lower than SNP heritability estimates for many other behavioral traits [40,96,157,198]. We also estimated the heritability of body weight, for which we had data in 957 rats from Charles River and 901 from Harlan. The estimates were much higher than for the behavioral metrics: 42.7% (s.e. = 0.070) for Charles River and ~63.2% (s.e. = 0.056) for Harlan.

Next, we performed GWAS for various metrics derived from Pavlovian conditioned approach separately for rats from Harlan and Charles River using GEMMA to fit a linear mixed model that allowed us to account for population structure. We also performed meta-analyses of the two populations using the subset of 93,990 SNPs that overlapped between our two filtered sets. Supplemental Table 3.8 contains information on all loci that obtained permutation-derived genome-wide significance in any of these analyses. Supplemental File 3.4 contains Manhattan plots for all analyses. Because the meta-analysis only contained the overlapping SNPs, we present 5 stacked Manhattan plots for each PavCA metric: Harlan and Charles River with all SNPs, Harlan and Charles River with only the ~94k overlapping SNPs, and the meta-analysis, which only uses the overlapping SNPs.

Previous work on Pavlovian conditioned approach in rats has focused on the average PavCA index score of days 4 and 5 to phenotype rats [205]. When analyzed as a quantitative trait, we did not identify any genome-wide significant results for this metric at a threshold of $-\log_{10}(p) = 5.66$ for Harlan and $-\log_{10}(p) = 6.05$ for Charles River. We also coded PavCA index score as a binary trait using various thresholds to define cases (goal-trackers) and controls (sign-trackers); this approach identified two genome-wide significant associations. The first locus was unique to Harlan and present on chromosome 4 in an intron of *Cntn4* (Page 67 in Supplemental File 3.4; Supplemental Figure 3.5), which encodes a cell-adhesion molecule involved in synaptic signaling, neuronal network formation, and neuropsychiatric disorders such as addiction [246–248]. The second locus only reached genome-wide significance in the meta-analysis and resided on chromosome 17 in the intronic region of *Fars2* (Figure 3.4B), a mitochondrial phenylalanyl-tRNA synthetase involved in oxidative phosphorylation and neuronal functioning [249–251].

Figure 3.4. Stacked Manhattan plots and a LocusZoom plot for day 4 and 5 average latency to magazine entry.

(A) The five Manhattan plots in descending order are: (1) GWAS in Harlan with the full set of 114k SNPs, (2) GWAS in Harlan with the overlapping set of 94k SNPs, (3) GWAS in Charles River with the full set of 214k SNPs, (4) GWAS in Charles River with the overlapping set of 94k SNPs, and (5) meta-analysis of the 94k overlapping SNPs. (B) LocusZoom plot of the genome-wide significant loci on chromosome 17 identified for day 4 and 5 average latency to magazine entry.



We then expanded our search to the average of days 4 and 5 for each of the 10 remaining metrics individually. We identified a genome-wide significant association for the number of head entries into the food magazine during the intertrial interval (i.e. in the absence of a conditioned stimulus; CS) in Harlan, but none of the other 9 metrics analyzed in Harlan and Charles River produced significant associations. However, the meta-analysis of the average day 4 and 5 metrics produced several additional genome-wide significant loci. There was a strong peak for the probability of a magazine entry during the CS interval on chromosome 7 (Page 42 in Supplemental File 3.4) that spanned ~1.2Mb and included the genes *Atxn10*, *Ppara*, and *Wnt7b*. Additionally, we replicated the *Fars2* association seen for the binary coding of the PavCA index score in the meta-analysis of the day 4 and 5 average latency to magazine entry during CS presentation (Fig 3.4A).

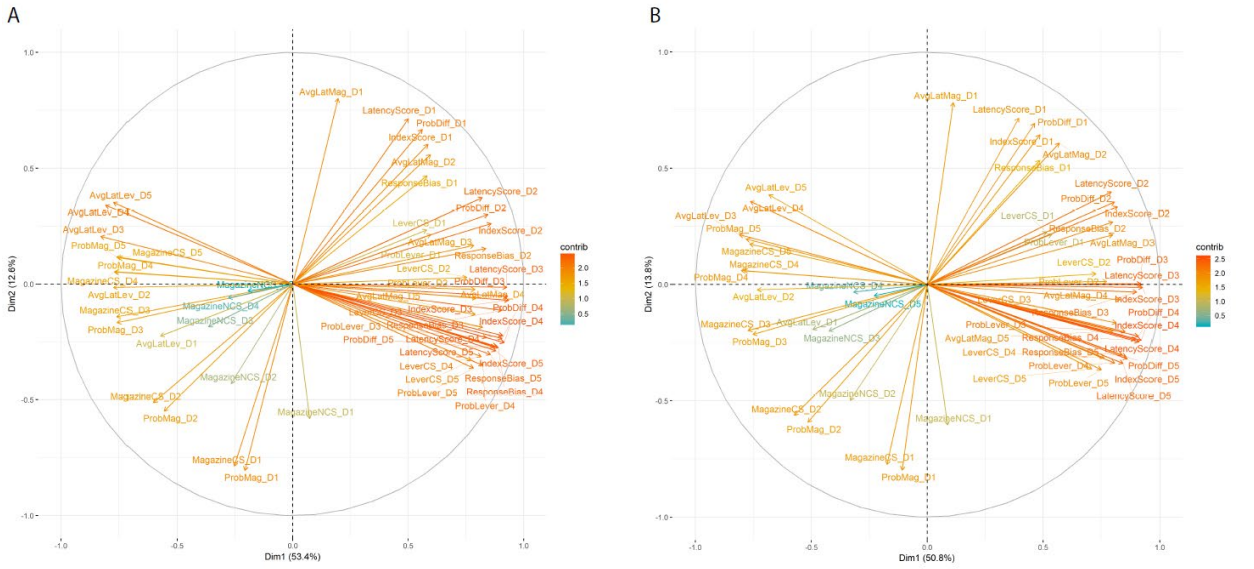
In an effort to further examine this large dataset, we also performed exploratory GWAS for all 11 metrics on days 1-5 for both Harlan, Charles River and the meta-analysis of the two. This large number of additional analyses (110 GWAS and 55 meta-analyses) produced only 5 additional genome-wide significant hits for Harlan, 1 for Charles River, and 5 from the meta-analyses. Because many of the metrics being tested are correlated, a Bonferroni correction would not be appropriate; however, the modest number of significant associations relative to the number of GWAS was not different from expectations under the null hypothesis (no true associations). None of the meta-analyses results replicated results from the individual populations. The meta-analyses did show that there were associations that occurred for the same metrics across multiple days of testing. For example, magazine entries with the CS showed an association with the same

SNP on chromosome 1 on days 3, 4, and 5 in the meta-analysis (Pages 21-24 in Supplemental File 3.4).

We sought to use PCA to analyze all 11 metrics across all 5 days separately for both Harlan and Charles River. A summary of the population-specific factor loadings and the percent variance explained by each factor for the first 5 PCs is presented in Supplemental File 3.2. The percent of variation explained by each of these PCs in each population is shown in Supplemental Figure 3.6. Figure 3.5 shows the correlations between each metric and the first two PCs in each population, helping to visualize which of the 55 metrics loaded most strongly on PC1 and PC2. PC1 accounted for slightly over 50% of the explained variance and loaded metrics from days 4 and 5, supporting our overall approach to Pavlovian conditioned approach. PC2 predominantly loaded metrics from day 1 of training. The factor structure and percent variance explained were very similar between Harlan and Charles River, suggesting that PCA was an effective way to summarize these data. Since the first 3 PCs cumulatively accounted for more than 70% of the variance in both populations, we only considered these for mapping and performed parallel GWAS for Harlan and Charles River. We identified 3 genome-wide significant associations; 2 of them were for PC2 on chromosomes 1 and 17 and were seen exclusively in Harlan. The third came from the meta-analysis of PC1 and was located on chromosome 17 at the same locus as associations for 5 other metrics.

Figure 3.5. Correlation circle plots for 55 metric PCA in Charles River and Harlan.

Correlation circles representing the strength and direction of each metrics' loading onto the first two PCs from PCA for (A) Charles River and (B) Harlan.



Finally, we mapped the first PC from running PCA on each metric across all 5 days and on all 11 metrics for each day (Supplemental File 3.2), which yielded 3 additional genome-wide significant associations. Two associations with PC1 of magazine entries during CS presentation on days 1-5 were discovered on chromosomes 1 and 2. The last associated locus was on chromosome 3 for all PCA metrics on day 4 of training.

For each metric GWAS, we made quantile-quantile (Q-Q) plots (Supplemental File 3.5) to examine the distribution of the test statistics in comparison to the null expectation. We observed moderate levels of inflation for several of the GWAS, which we interpreted as evidence for polygenic signal rather than population structure, since the LMM should have accounted for population structure.

Given the large sample size for a rodent GWAS, the modest number of genome-wide significant associations was surprising. To determine if this was a phenomenon unique to our behavioral metrics, which had relatively low SNP heritabilities (mean 7.8% in Harlan and 14.8% in Charles River), we also performed a GWAS for body weight, for which we had observed much higher heritability (63.2% and 42.7%). One limitation of this analysis is that we only had body weight data a subset of 1,858 rats (Charles River n=957; Harlan n=901). We found three genome wide significant associations for body weight, two in Harlan and one in Charles River. The meta-analysis supported the association from Charles River, but did not yield any additional associations. Though only 3 loci were identified, perhaps due to our limited sample size, the Q-Q plots (Page 89 of Supplemental File 3.5) show that there is an increased polygenic signal for body weight compared to the PavCA metrics, supporting our hypothesis that a more heritable trait would show stronger association in these populations.

3.5 Discussion

Although both rats and mice are widely used in the biomedical sciences, most studies in mice utilize inbred strains whereas studies in rats more commonly use outbred strains. One of the most extensively used outbred rat strains is Sprague Dawley [252]. While SD and other outbred rats have been used for selective breeding studies [253–255], this study was the first to use SD for GWAS. We densely genotyped more than 4,000 SD rats and used the data to characterize the genetic background of SD rats and to perform GWAS for the behaviors that comprise Pavlovian conditioned approach. This represents the largest rodent GWAS ever undertaken, and the first performed using a commercially available outbred rat population. We found dramatic genetic differences between SD rats obtained from Harlan versus Charles River. F_{ST} estimates show that SD rats from Harlan and Charles River are more differentiated than the major human

subpopulations [227] and nearly as diverged as some subspecies of mice [256]. We also found evidence of population structure among the various breeding locations and barrier facilities for each vendor. We found that SD rats from both Harlan and Charles River showed a rapid decay in LD; however, SD from Charles River have more polymorphisms and more favorable MAF and LD profiles, suggesting that future GWAS in SD are best done with rats obtained from Charles River.

We estimated that our meta-analysis was well powered for identify loci that accounted for only 1%-2% of total trait variance (Supplemental Figure 3.7). When we began this study, we were unaware of the large genetic differences between Sprague Dawley rats from Charles River and Harlan. To cope with the observed differentiation, we emulated the approach often taken in human genetics in which multiple groups are analyzed separately before being combined in a meta-analysis (e.g. [257]); however, it is well known that different human populations (e.g. East Asian and European) often do not share the same causal variants. Similarly, our power may have been reduced because causal variants were not shared between the Charles River and Harlan. Modifiers of causal variants might also be dissimilar between Charles River and Harlan, further hindering our meta-analyses. [132]

A notable success of this study was a locus we detected on chromosome 17, which was identified by the meta-analysis of Harlan and Charles River. This locus was identified for multiple metrics on days 4 and 5, as well as in the univariate analysis of the first PC from PCA on the comprehensive set of all 55 metrics. All genome-wide significant SNPs were located in an intron of *Fars2* (Figure 3.3B), which is highly expressed in the Purkinje cells of the cerebellum [249] but has no previously known ties to behavior. Another promising locus for magazine entries and latency to magazine entry during CS presentation was on chromosome 1. That locus was consistent

across multiple days of testing for both of these measures. However, the locus is in an intergenic region with no known genes, indicating either the presence of an unannotated gene, or a regulatory site that influences a nearby gene.

Since PavCA metrics had never been used for a GWAS before, it was unclear which measurements and transformations would yield the best results. Therefore, we used several approaches, which led to the identification of over 20 genome-wide significant loci but involved 267 total tests with a significance threshold of alpha = 0.05. Despite high correlations between many of the metrics, some of the loci we detected were unique to a single PavCA metric, raising the possibility that they could be false positives. Furthermore, nearly half of the QTL identified occurred on earlier training days (days 1-3), which are harder to interpret since the literature surrounding Pavlovian conditioned approach focuses on behavior after training (4/5).

Using genome-wide genotype data, we have provided the first quantitative estimates of the SNP heritability of the component measurements of PavCA. The heritability estimates were lower than we had anticipated, but perhaps reasonable given the complexity of the behavior and new insights into the heterogeneity of neural networks involved in sign-tracking behaviors [258–260]. The highest heritabilities (16-20% in Charles River) were seen for measures typically used to assess the propensity to attribute incentive salience to reward cues, including the averages of the day 4 and day 5 PavCA index score, response bias, and latency score. Previous work had shown that SD rats selectively bred for ~15 generations for high or low responses to a novel environment were also highly divergent for behaviors in PavCA [75]. Those selection studies demonstrated that SD rats had alleles that could exert a strong influence on PavCA performance. The present results are consistent with this conclusion but highlight the extent to which alleles can be concentrated over many generations of selective breeding. We obtained lower heritability estimates for SD rats

from Harlan compared to Charles River, further emphasizing the genetic differences between SD rats from the two vendors. With more highly heritable behavioral traits, we suspect the SD rats would have yielded better results.

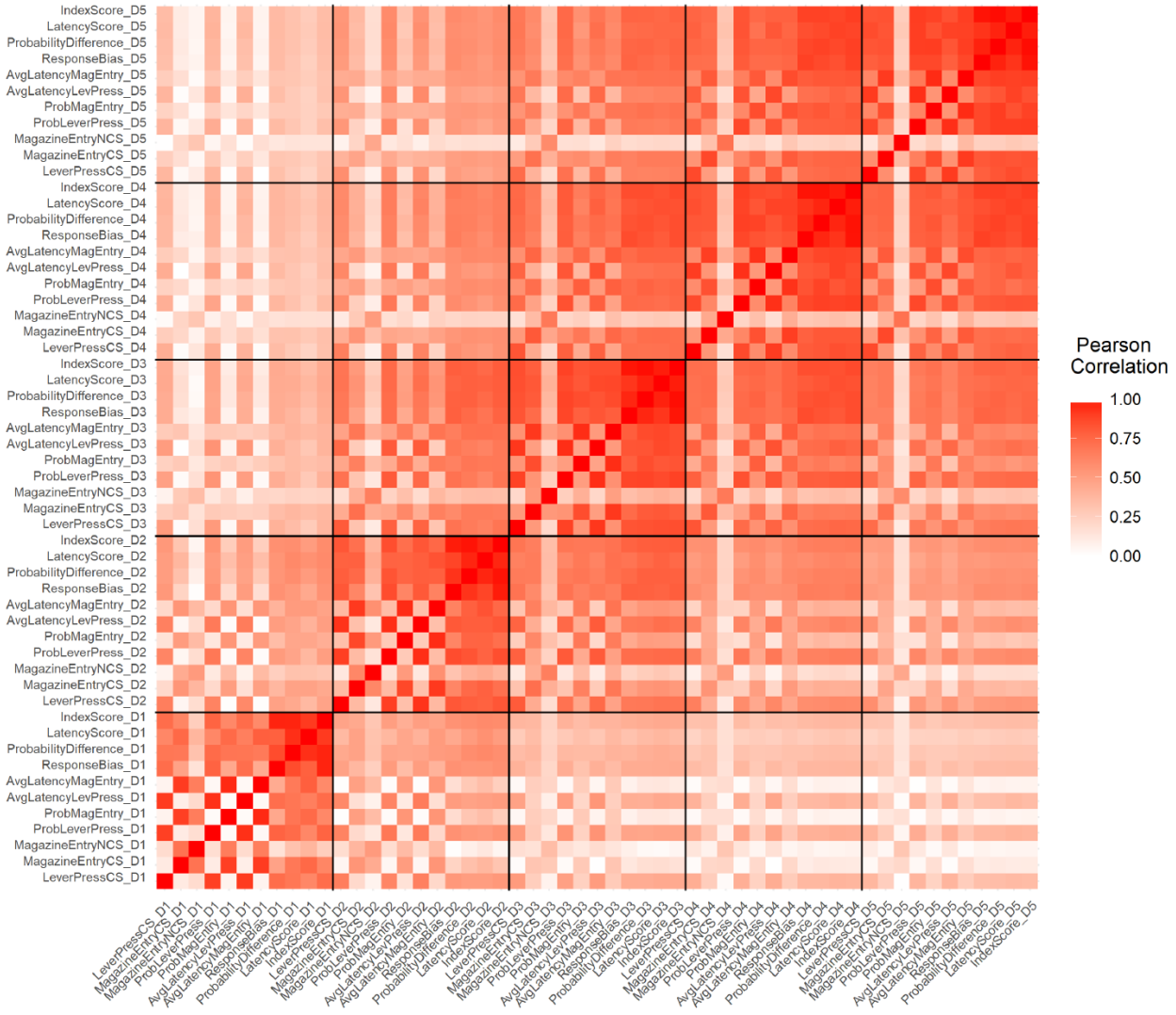
Although our study is the first to carefully document population structure within SD rats, ours is not the first to highlight phenotypic differences among SD from different vendors. In 1973, Prejean et al. reported that the incidence of endocrine tumors varied among SD rats from different vendors [101]. Then Clark et al. (1991) [102] reported differences in noradrenergic neural projections among SD rats from different vendors. Subsequently, Turnbull & Rivier (1999) [103] reported vendor-specific differences in the response to inflammatory stimuli. Then Fuller et al. (2001) [104] reported vendor-specific differences in hypoxic response among SD rats. Even more recently, there have been additional publications reporting differences for a variety of traits among SD rats obtained from different vendors [105,107,261,262], and even suggesting that these phenotypic differences may extend to differences among a single vendor's breeding facility [106]. Our own studies have previously reported both behavioral and genetic differences among SD rats [76], conclusions that are much more comprehensively addressed with the present dataset. Specifically, in addition to wide spread genetic differences, we have also shown that SD rats obtained from Harlan show a much higher proclivity to become sign-trackers compared to SD rats from Charles River. However, neither these prior publications nor the current one can differentiate between two possibilities: that the observed behavioral differences are the result of the different environment in which these animals are raised versus the genetic difference that we have clearly demonstrated. This question could be addressed by future studies in which SD rats are bred in the same facility and the offspring tested in the same manner.

Regardless of whether differences in SD rats obtained from different vendors are due to genetic or environmental differences, our results demonstrate the need for much greater care in the use of SD rats. Future studies may only want to use rats from a single vendor and from a single breeding facility to maintain a consistent genetic background. The differences among vendors can be a source of unwanted phenotypic variability, which alone might be a reason to avoid heterogeneous samples. Vendor differences, especially when not reported in the methods, can also lead problems with replication, since observations made in SD rats from Harlan may not be valid in SD rats from Charles River. Problems of replication and inadequate reporting also extend to differences between breeding facilities within a given vendor. However, a more subtle consequence of using SD rats from multiple vendors or breeding facilities is that spurious correlations can occur. For example, if SD rats from Charles River were higher for traits A and B, compared to SD rats from Harlan, a heterogeneous cohort of SD rats from both vendors would show a significant positive correlation between traits A and B. Such a correlation is unlikely to be due to a shared biological mechanism; it may instead be the result of either genetic population structure or environmental differences between SD rats from the two vendors.

3.6 Appendix C: Supplemental Figures

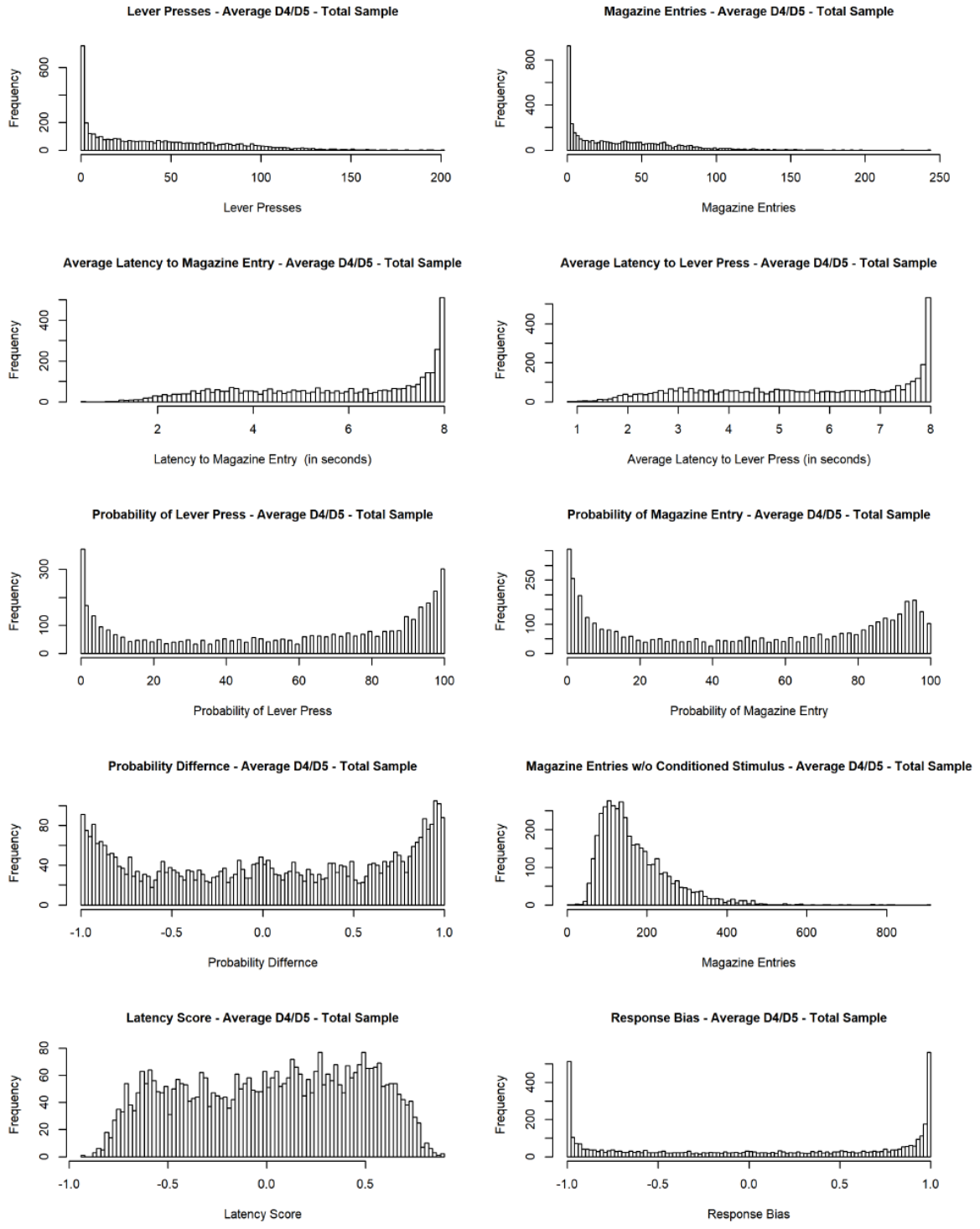
Supplemental Figure 3.1. Correlation heatmap of PavCA metrics across days 1-5.

The heatmap displays the absolute value of the Pearson correlation coefficient between each pair of metrics across all 5 days of testing.

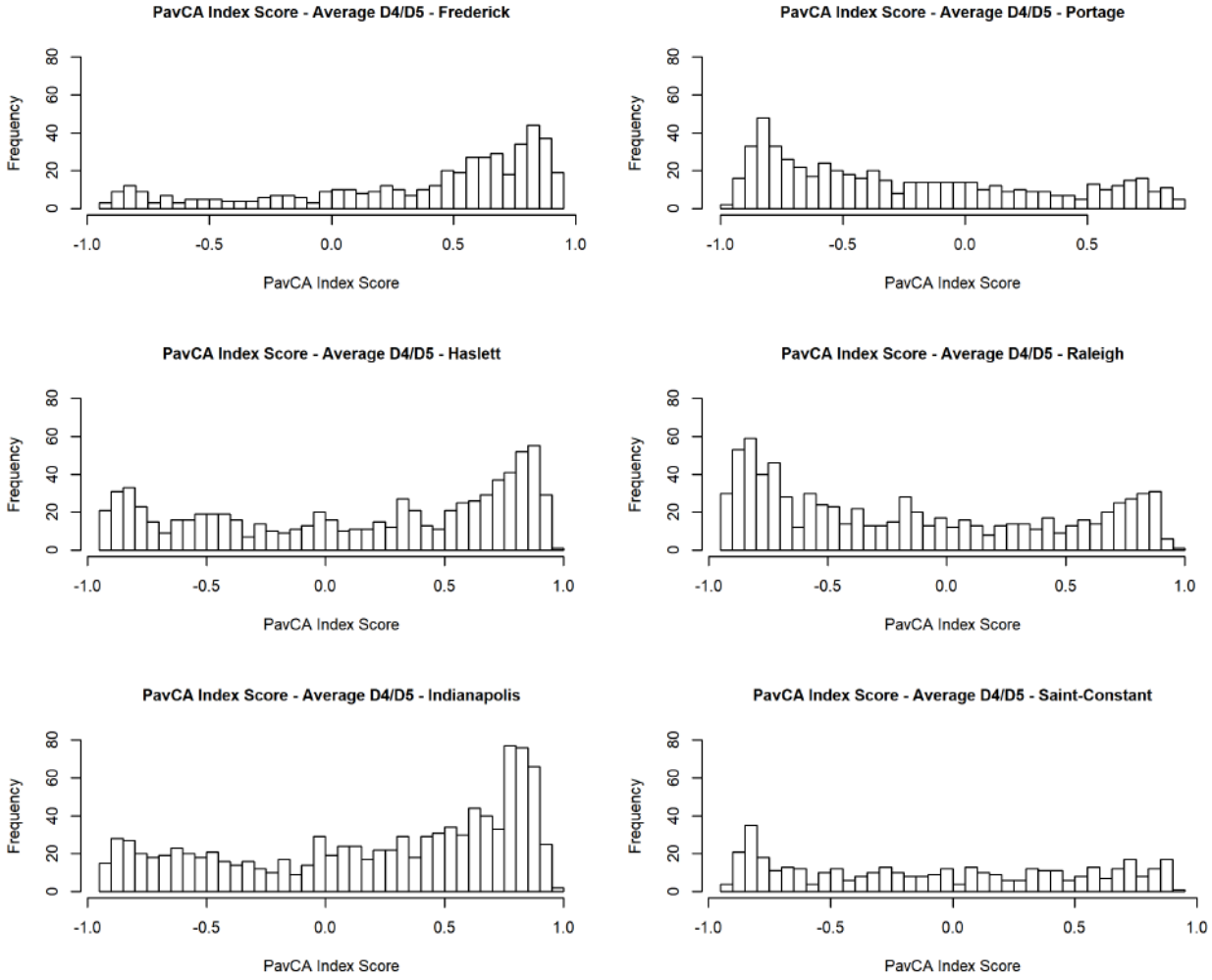


Supplemental Figure 3.2. Distributions of the average of day 4 and day 5 measurements for 10 PavCA metrics.

Histograms were constructed using measurements from the combined Harlan and Charles River sample set.



Supplemental Figure 3.3. Distributions of the average of day 4 and day 5 PavCA index scores for 6 major breeding locations.

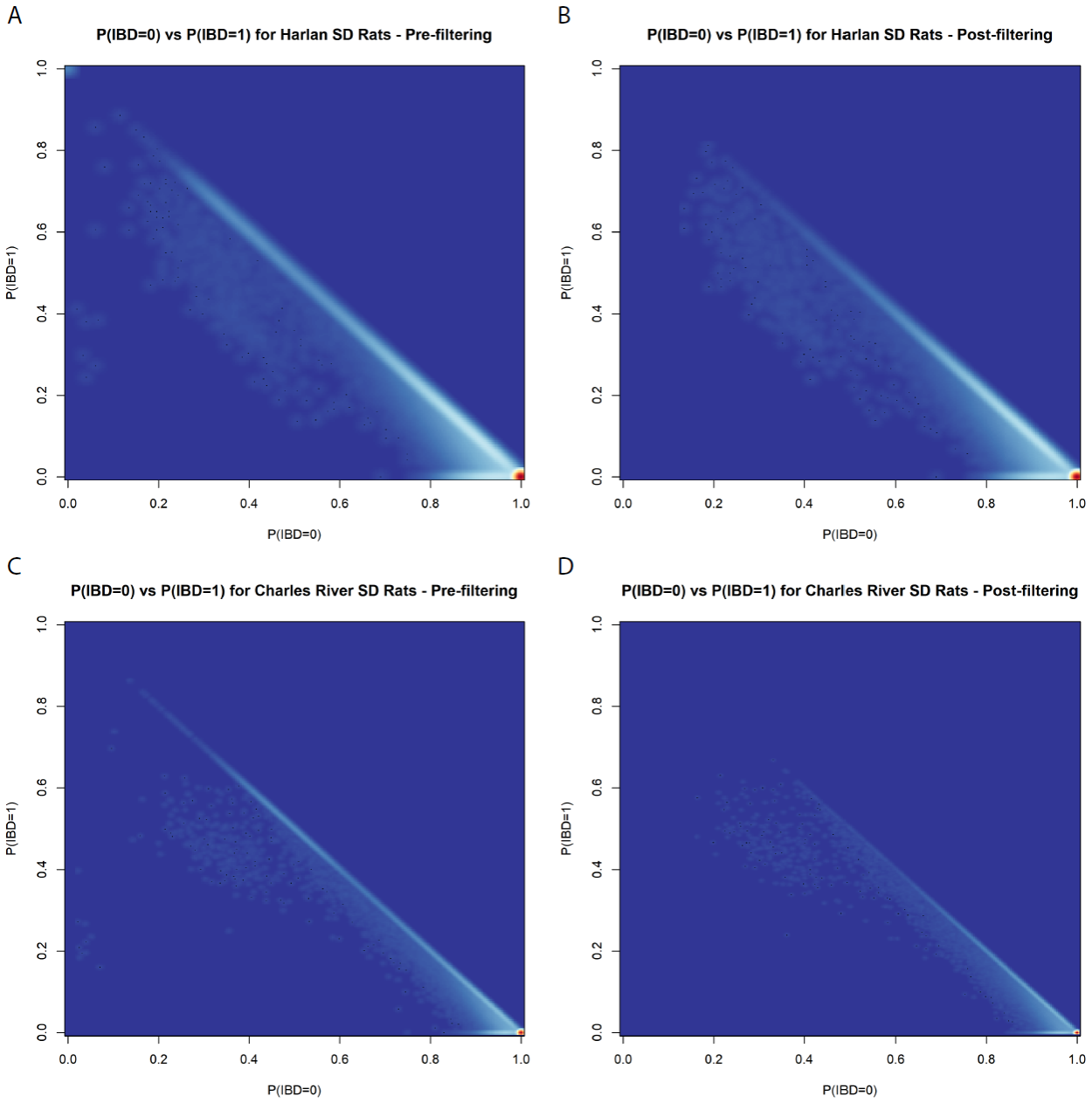


Breeding Location	Frederick	Haslett	Indianapolis	Portage	Raleigh	St. Constant
Frederick	1	1.955e-10	4.645e-06	<2.2e-16	<2.2e-16	<2.2e-16
Haslett		1	0.0131	<2.2e-16	<2.2e-16	<2.2e-16
Indianapolis			1	<2.2e-16	<2.2e-16	<2.2e-16
Portage				1	0.004155	6.657e-06
Raleigh					1	0.03225
St. Constant						1

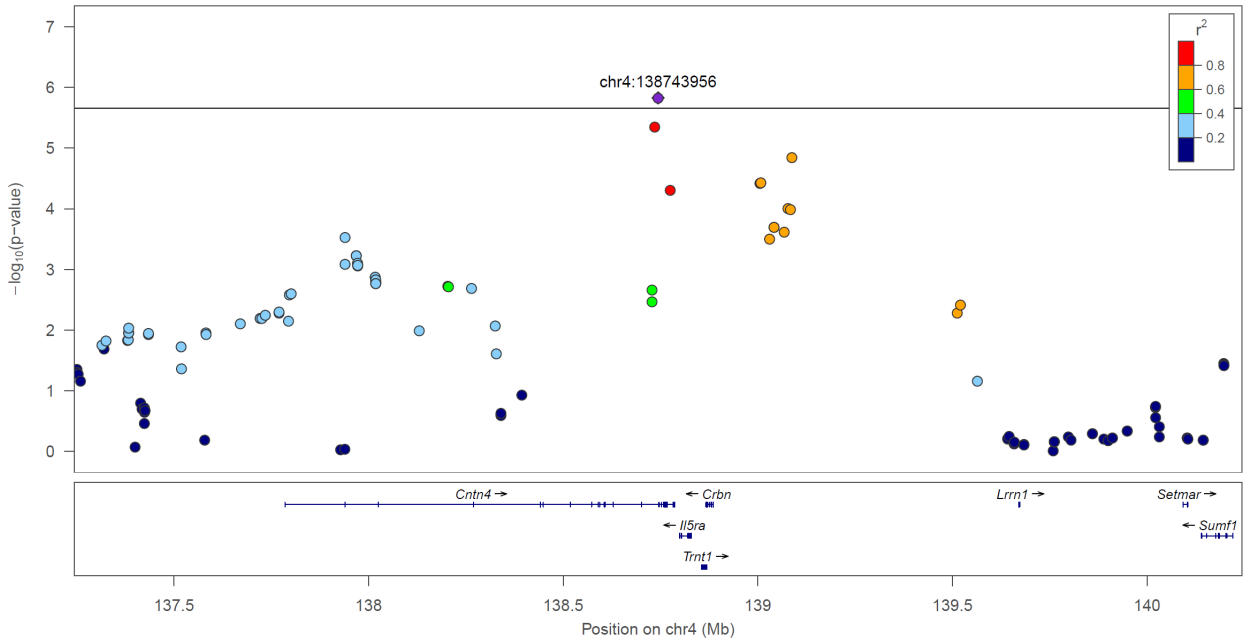
Welch's 2-sample t-test p-values from pairwaise comparisons of day4 and day 5 average PavCA index score distributions between different breeding locations.

Supplemental Figure 3.4. Heatmaps of pairwise identity-by-descent pre- and post-filtering.

Panels A and C show pre-filtering values of P(IBD) = 0 plotted against P(IBD) = 1 for Harlan and Charles River. Unrelated samples cluster in the lower right corner. Samples along the diagonal have 2nd and 3rd degree levels of relatedness, while those clustering around (0.5,0.25) are full-siblings. Panels B and D demonstrate that the sample filtering steps removed several spurious relations from the sample.

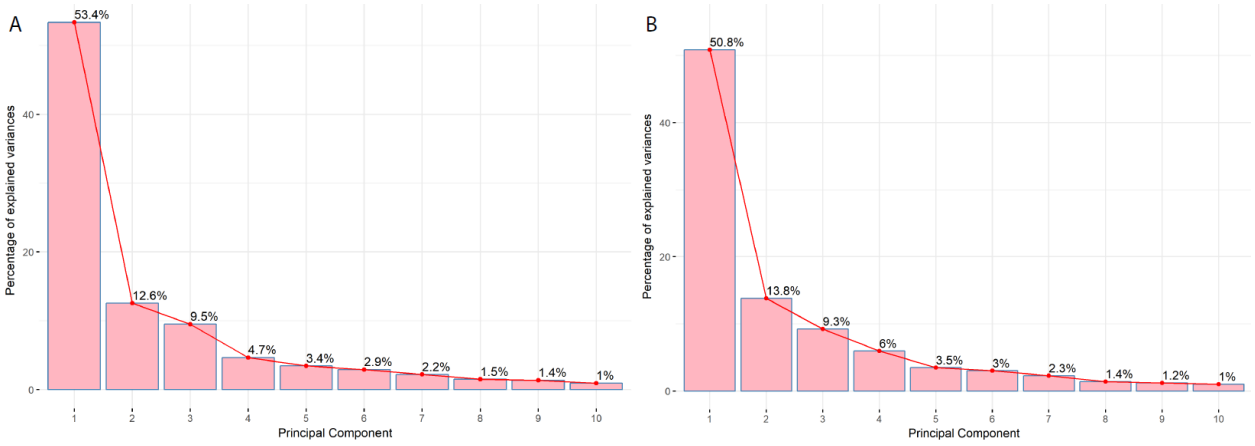


Supplemental Figure 3.5. LocusZoom plot of genome-wide associated region containing *Cntn4*.

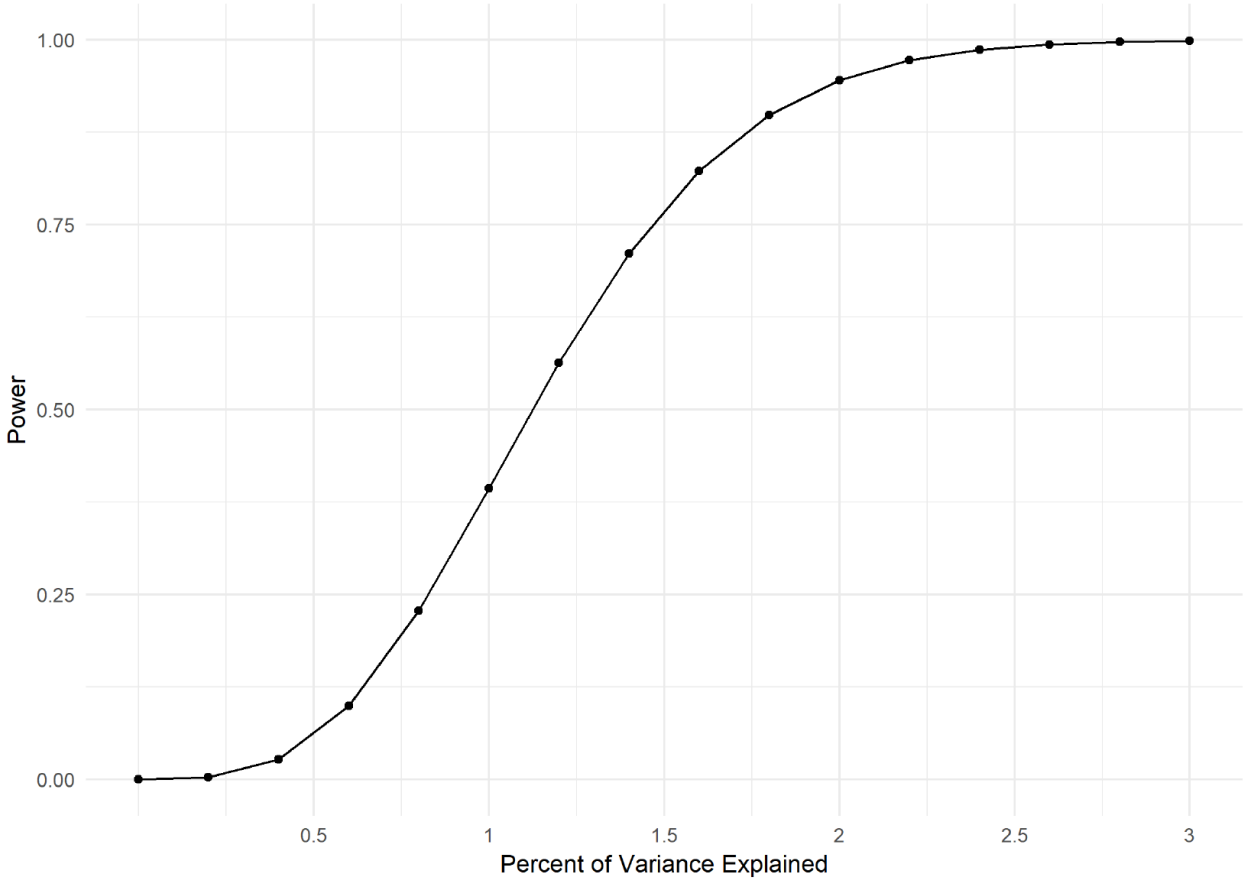


Supplemental Figure 3.6. Scree plots of the PVE for each of the top 10 PCs in the 55 metric PCA analysis.

Panel A shows the PVE for Charles River and Panel B shows the PVE for Harlan.

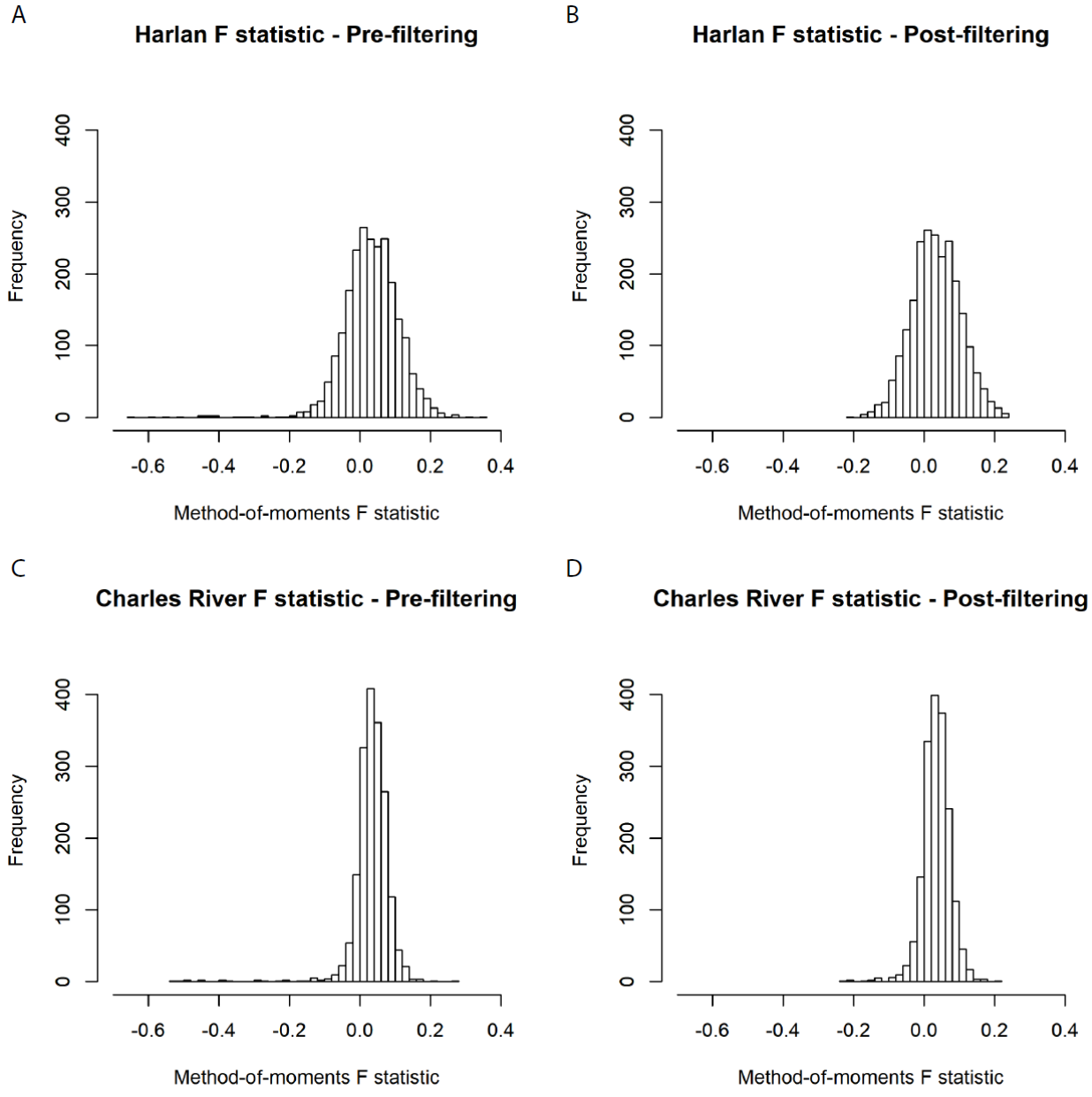


Supplemental Figure 3.7. Power analysis curve for n=2,000 using Quanto.



Supplemental Figure 3.8. Pre- and post-filtering distributions of heterozygosity for Harlan and Charles River.

Panels A and C show pre-filtering distributions of heterozygosity in Harlan and Charles River, as measured by the method-of-moments F coefficient. Panels B and D show the same distributions post-filtering. A value above 0 indicates a deflation of heterozygosity, whereas a value below 0 would be an inflation.



3.7 Appendix D: Supplemental Tables

Supplementary Table 3.1. Sample origins for all 4,061 SD rats in final filtered set.

	# Samples (Barrier Facility)	# Samples (Breeding Location)
Total samples post-filtering	4,061	4,061
Harlan – Dublin, VA – Barrier 231	5	5
Harlan – Frederick, MD – Barrier 208A	465	465
Harlan – Haslett, MI – Barrier 206	762	762
Harlan – Houston, TX – Barrier 211	14	14
Harlan – Indianapolis, IN – Barrier 202A	546	994
Harlan – Indianapolis, IN – Barrier 202C	97	
Harlan – Indianapolis, IN – Barrier 217	351	
Charles River – Kingston, NY – Barrier K92	4	4
Charles River – Portage, MI – Barrier P03	32	558
Charles River – Portage, MI – Barrier P07	121	
Charles River – Portage, MI – Barrier P09	327	
Charles River – Portage, MI – Barrier P10	78	
Charles River – Raleigh, NC – Barrier R04	650	794
Charles River – Raleigh, NC – Barrier R09	144	
Charles River – Saint Constant, CAN – Barrier C71	46	404
Charles River – Saint Constant, CAN – Barrier C72	358	
Unknown Barrier	61	61

Supplementary Table 3.2. List of variant filtering steps and the numbers of SNPs remaining after each step for both ANGSD/Beagle and STITCH.

The values highlighted in green are the final SNP totals used for GWAS in Charles River and Harlan. The values highlighted in orange are the counts we used as criteria for choosing ANGSD/Beagle over STITCH.

	ANGSD/Beagle		STITCH	
Sequential Filtering Steps	**Number of SNPs Harlan**	**Number of SNPs Charles River**	**Number of SNPs Harlan**	**Number of SNPs Charles River**
Total called/imputed	2,274,118	2,274,118	8,691,886	8,691,886
SNPs with dosage $r^2 \geq 0.9$	305,779	305,779	1,992,919	1,992,919
SNPs with MAF ≥ 0.01	235,583	128,572	1,058,812	1,843,240
SNPs with HWE p-value $< 1x10^{-7}$	214,309	114,568	930,685	1,718,475
SNPs after LD pruning $r^2 < 0.95$	44,721	93,692	25,009	75,503
Union set of SNPs	234,887			
Novel SNPs compared to dbSNP (not found in Build 149 - Nov 7, 2016)	195,299			
Novel SNPs compared to 42 genomes ref (not found in Hermsen et al. 2015)	105,638			

Supplemental Table 3.3. Concordance and error rates for ANGSD/Beagle genotypes at different dosage r2 thresholds.

Rates of concordance and genotyping error were calculated by comparing genotypes for 381 duplicate samples called in parallel. Each of the two replicates of the sample was assumed to contribute half the discordant genotypes. Therefore, the per sample genotyping error rate was calculated as half of the observed rate of discordance.

	DR^2>=0.9	DR^2>=0.8	DR^2>=0.7
Total called and imputed by ANGSD/Beagle	2,274,118	2,274,118	2,274,118
SNPs passing dosage r^2 filter	305,779	427,123	541,622
SNPs passing MAF >= 0.01 filter	244,586	326,550	393,569
SNPs passing HWE 10^{-7} filter	204,104	259,111	305,660
Mean Pearson **correlation of dosages** (n=381)	0.983	0.969	0.957
Rate of concordance of hard calls (n=381)	0.983	0.972	0.963
Rate of discordance (1 – concordance)	0.017	0.028	0.037
Error rate (Rate of discordance/2)*100	0.85%	1.4%	1.85%

Supplemental Table 3.4. Pairwise F_{ST} estimates between vendor barrier facilities.

Charles River – Kingston, NY and Harlan – Dublin, VA and Houston, TX were excluded due to low sample numbers.

	Harlan					Charles River							
	Haslett, MI	Indianapolis, IN			Frederick, MD	St. Constant, CAN		Portage, MI				Raleigh, NC	
	206	217	202A	202C	208A	C71	C72	P03	P07	P09	P10	R04	R09
206	0	0.031	0.017	0.020	0.016	0.438	0.435	0.434	0.437	0.433	0.449	0.429	0.438
217		0	0.03	0.033	0.028	0.443	0.441	0.440	0.443	0.439	0.455	0.436	0.444
202A			0	0.066	0.012	0.441	0.439	0.437	0.441	0.437	0.453	0.433	0.442
202C				0	0.015	0.442	0.439	0.438	0.441	0.437	0.453	0.433	0.442
208A					0	0.442	0.439	0.437	0.441	0.437	0.453	0.434	0.442
C71						0	0.024	0.025	0.019	0.027	0.038	0.030	0.026
C72							0	0.017	0.018	0.022	0.036	0.024	0.023
P03								0	0.017	0.019	0.033	0.022	0.021
P07									0	0.015	0.034	0.023	0.019
P09										0	0.037	0.028	0.018
P10											0	0.044	0.037
R04												0	0.027
R09													0

Supplemental Table 3.5. List of sample filtering criteria and number of samples removed.

	Number of Samples	
Total sequenced by ddGBS + WGS	4,625 (17 WGS only)	
Low read count in FASTQ (< 4 million reads) **performed prior to variant calling	316	
Females samples (only females in sample set)	77	
Purchased from Taconic	4	
Lack of phenotype information	10	
Poor clustering in principal component analysis	54	
Putative sample mix-up	18	
Duplicate sample	7	
Inflated/deflated heterozygosity	34	
>= 30 samples with $\hat{\pi}$ >= 0.1875 in sample	12	
$\hat{\pi}$ > 0.6 with another sample	32	
Final Sample Set (Harlan + Charles River)	**4,061**	
	CR – 1,780	**Har – 2,281**

Supplemental Table 3.6. List of covariates used for the GWAS LMMs for Harlan and Charles River.

The first three covariates were used in both Charles River and Harlan analyses. The remaining covariates were unique to each population.

Harlan Covariates	Charles River Covariates
Age of rat in days at start of training (continuous integer)	
Housing condition (binary – single or multiple rats per chamber)	
Light cycle (binary – standard or reverse 12h lighting)	
Experimenter(s) BFS (binary indicator)	Experimenter(s) AAK (binary indicator)
Experimenter(s) BFS/AA (binary indicator)	Experimenter(s) AK (binary indicator)
Experimenter(s) BTS (binary indicator)	Experimenter(s) BFS (binary indicator)
Experimenter(s) CJF (binary indicator)	Experimenter(s) BFS/AA (binary indicator)
Experimenter(s) CJF/JDM (binary indicator)	Experimenter(s) CJF (binary indicator)
Experimenter(s) ESC (binary indicator)	Experimenter(s) CJF/JDM (binary indicator)
Experimenter(s) KP (binary indicator)	Experimenter(s) EGO (binary indicator)
	Experimenter(s) ESC (binary indicator)
	Experimenter(s) LMY (binary indicator)
	Experimenter(s) MLR.LMF (binary indicator)

Supplemental Table 3.7. List of all PavCA metrics collected on SD rats.

There are 11 total metrics across 5 days of training. The first 7 metrics are direct measurements made during the training periods. The following 3 metrics are calculated from the base measurements. The final metric, PavCA index score, is a composite score from the previous 3 metrics.

Day 1 25 8-second trials	**Day 2** 25 8-second trials	**Day 3** 25 8-second trials	**Day 4** 25 8-second trials	**Day 5** 25 8-second trials	**Data Type**
Lever Presses	Lever Presses	Lever Presses	Lever Presses	Lever Presses	Count (summed over trials)
Magazine Entries CS	Magazine Entries CS	Magazine Entries CS	Magazine Entries CS	Magazine Entries CS	Count (summed over trials)
Magazine Entries NCS	Magazine Entries NCS	Magazine Entries NCS	Magazine Entries NCS	Magazine Entries NCS	Count (summed over trials)
Latency to Lever Press	Latency to Lever Press	Latency to Lever Press	Latency to Lever Press	Latency to Lever Press	Continuous [0,8] secs (averaged over trials)
Latency to Magazine Entry	Latency to Magazine Entry	Latency to Magazine Entry	Latency to Magazine Entry	Latency to Magazine Entry	Continuous [0,8] secs (averaged over trials)
Probability of Lever Press	Probability of Lever Press	Probability of Lever Press	Probability of Lever Press	Probability of Lever Press	Proportion [0-1] (0/25 to 25/25 trials)
Probability of Mag Entry	Probability of Mag Entry	Probability of Mag Entry	Probability of Mag Entry	Probability of Mag Entry	Proportion [0-1] (0/25 to 25/25 trials)
Response Bias	Response Bias	Response Bias	Response Bias	Response Bias	Continuous [-1,1]
Latency Score	Latency Score	Latency Score	Latency Score	Latency Score	Continuous [-1,1]
Probability Difference	Probability Difference	Probability Difference	Probability Difference	Probability Difference	Continuous [-1,1]
PavCA Index Score	PavCA Index Score	PavCA Index Score	PavCA Index Score	PavCA Index Score	Continuous [-1,1]

Supplemental Table 3.8. Summary of genome-wide significant associations.

Top SNP	Alleles	EAF - Harlan	EAF - C.R.	Phenotype(s)	Day(s)	Beta - Harlan	Beta - C.R.	Pval - Harlan	Pval - C.R.	Pval - Meta	PVE - Harlan	PVE - C.R.	Nearby Genes (500kb)
chr1:38,824,984	C/T	0.013	0.015	Average Latency to Magazine Entry	4	-0.676	-0.242	**1.65E-06**	9.69E-02	3.77E-06	1.01%	0.12%	N/A
					5	-0.652	-0.294	3.85E-06	4.33E-02	**2.06E-06**	0.94%	0.18%	
					Avg4/5	-0.663	-0.288	2.67E-06	4.86E-02	**1.88E-06**	0.97%	0.17%	
				Magazine Entries - Conditioned Stimulus	3	0.636	0.415	5.44E-06	4.29E-03	**1.42E-07**	0.91%	0.36%	
					4	0.655	0.300	3.33E-06	4.02E-02	**1.70E-06**	0.95%	0.19%	
					5	0.616	0.412	1.28E-05	4.75E-03	**3.25E-07**	0.84%	0.35%	
					Avg4/5	0.661	0.375	2.90E-06	1.03E-02	**2.48E-07**	0.97%	0.29%	
				PC1 - Magazine Entries w/ CS - All Days	All	-0.703	0.423	**8.50E-07**	3.78E-03	1.11E-01	1.11%	0.38%	
chr1:54,194,237	A/G	0.022	0.068	PC2 - All Phenotypes / All Days	All	0.525	0.039	**1.55E-06**	5.93E-01	1.95E-03	1.05%	0.01%	*Afdn, Vom2r7*
chr1:55,674,510	T/A	0.012	0.01	Magazine Entries - Conditioned Stimulus	3	0.446	0.652	1.72E-03	1.95E-04	**1.85E-06**	0.44%	0.61%	*Vom2r9, Dact2*
chr1:55,736,249	G/T	0.014	0.014	Magazine Entries - No Conditioned Stimulus	3	0.635	-0.037	**6.55E-07**	7.91E-01	3.52E-04	1.10%	0.00%	*Vom2r9, Dact2*
chr1:56,030,596	T/C	0.064	0.028	Body Weight	N/A	-6.823	-27.836	1.02E-02	**1.28E-07**	**2.40E-07**	0.74%	3.06%	*Vom2r9, Dact2, Smoc2*
chr1:165,220,406	G/A	0.301	0.58	Latency Score	5	0.164	-0.038	**1.00E-06**	2.86E-01	4.66E-03	1.06%	0.05%	*Pgm2l1, P4ha3, Ppme1, C2cd3, Kcne2, Lipt2, Pold3, Chrd12*
chr1:180,805,571	G/C	0.012	0.013	Magazine Entries - No Conditioned Stimulus	Avg4/5	0.631	-0.059	**1.99E-06**	6.75E-01	1.22E-03	1.00%	0.01%	N/A
chr2:318,643	T/G	0.011	0.046	PC1 - Magazine Entries w/ CS - All Days	All	-0.745	0.138	**1.11E-06**	1.15E-01	3.05E-01	1.08%	0.11%	N/A
chr2:119,024,078	A/G	0.145	0.355	BodyWeight	N/A	8.388	-3.018	**1.30E-06**	8.63E-02	1.53E-01	2.56%	0.33%	N/A
chr3:35,099,667	C/T	0.471	0.285	Average Latency to Lever Press	3	-0.145	0.025	**1.07E-06**	5.07E-01	6.52E-04	1.05%	0.02%	*Kif5c, Lypd6b*
				PC1 - All Day 4 Phenotypes	4	0.139	0.076	5.00E-06	4.91E-02	**1.57E-06**	0.95%	0.18%	
chr4:138,743,956	A/G	0.024	0.712	PavCA Index Score - Binary 0	Avg4/5	-0.223	-0.007	**1.49E-06**	6.99E-01	3.66E-02	1.02%	0.01%	*Cntn4, Il5ra, Trnt1, Crbn*
chr4:142,668,766	C/G	0.504	0.477	Average Latency to Magazine Entry	3	0.144	-0.007	**7.92E-07**	8.23E-01	3.61E-04	1.08%	0.00%	*Grm7*
chr5:115,061,198	C/T	0.011	0.664	Magazine Entries - Conditioned Stimulus	3	0.083	-0.178	5.79E-01	**7.86E-07**	3.23E-06	0.01%	1.08%	*Hook1, Cyp2j10, Cyp2j3, Fggy*
chr7:120,798,791	A/G	0.013	0.012	Average Latency to Magazine Entry	3	-0.524	-0.417	5.69E-05	7.98E-03	**1.73E-06**	0.72%	0.31%	*Dmc1, Kcnj14, Fam227a, Cby1, Tomm22, Josd1, Gtpbp1, Dna14, Nptxr, Ddx17, Kdelr3, Csnk1e, Tmem184b, Moff, Pla2g6*
				Magazine Entries - Conditioned Stimulus	3	0.563	0.429	1.40E-05	6.38E-03	**3.87E-07**	0.84%	0.33%	
chr7:125,865,679	G/T	0.23	0.077	Magazine Entries - No Conditioned Stimulus	2	-0.141	-0.156	4.28E-05	9.89E-03	**1.40E-06**	0.74%	0.29%	*Npap60, Arhgap8, Phf21b, Upk3a, Fam118a, Smc1b, Ribc2, Fblin1*
chr7:126,430,491	C/T	0.732	0.083	Probability of Magazine Entry	Avg4/5	0.142	0.144	2.53E-05	2.30E-02	**1.74E-06**	0.78%	0.23%	*Wnt7b, Atxn10, Ppara, Mirlet7b*
chr8:51,658,262	T/C	0.127	0.744	Latency Score	1	-0.120	-0.157	9.39E-03	5.06E-05	**1.82E-06**	0.30%	0.74%	*Cadm1*
chr10:94,216,916	A/C	0.071	0.926	Body Weight	N/A	11.324	0.075	**8.41E-07**	9.81E-01	6.02E-04	2.65%	0.00%	*Ace, Ace3, Cyb561, Tanc2, Kcnh6, Dcaf7, Taco1, Map3k3, Limd2, Strada, Ccdc47, Ddx42, Ftsj3, Psmc5, Smarcd2*
chr11:14,742,623	A/G	0.789	0.101	Average Latency to Magazine Entry	2	0.141	0.178	1.22E-04	1.79E-03	**8.84E-07**	0.65%	0.43%	*Nrip1*
				Magazine Entries - No Conditioned Stimulus	2	-0.163	-0.113	4.93E-06	3.80E-02	**6.92E-07**	0.92%	0.19%	
chr17:29,043,495	A/C	0.035	0.033	Average Latency to Magazine Entry	Avg4/5	0.304	0.302	2.39E-04	1.63E-03	**1.36E-06**	0.60%	0.44%	*Fars2*
				PavCA Index Score - Binary 0	Avg4/5	0.118	0.182	2.78E-03	8.08E-05	**1.39E-06**	0.40%	0.68%	
				PavCA Index Score - Binary +-0.25	Avg4/5	0.132	0.192	1.80E-03	1.63E-04	**1.57E-06**	0.52%	0.75%	
				PC1 - All Phenotypes / All Days	All	0.323	0.287	1.62E-04	2.99E-03	**1.59E-06**	0.65%	0.40%	
				Probability Difference	4	0.292	0.309	4.23E-04	1.23E-03	**1.84E-06**	0.55%	0.46%	
				Probability of Magazine Entry	5	-0.309	-0.300	2.15E-04	1.88E-03	**1.41E-06**	0.60%	0.43%	
chr17:57,387,756	G/A	0.194	0.668	PC2 - All Phenotypes / All Days	N/A	0.202	-0.014	**4.20E-07**	7.06E-01	1.23E-03	1.17%	0.01%	*Epc1*

CHAPTER 4

REPLICATION GWAS AND META-ANALYSIS FOR PAVLOVIAN CONDITIONED APPROACH IN HETEROGENOUS STOCK RATS

4.1 Abstract

Previously, we performed a genome-wide association study (GWAS) in a large sample of 4,061 Sprague Dawley (SD) rats for the propensity of to attribute incentive salience to reward-associated cues as measured by Pavlovian conditioned approach (PavCA). Here we perform a replication of this GWAS in an independent sample of 2,449 heterogenous stock (HS) rats also measured for performance in PavCA. The samples was split between two centers for testing, which differed in the age of the rats they tested and the previous experience of the rats to behavioral training. In this novel HS sample, we confirm our previous observations in SD that heritability estimates for PavCA are relatively low. We go on to estimate the genetic correlations for PavCA metrics both between and within testing centers. We then perform center-specific GWAS and mega-analyses of the HS rats across 56 different PavCA metrics using a set of over 3.7 million SNPs, leading to the discovery of 22 loci associated with various PavCA metrics across the five days of behavioral training. Using the overlapping set of ~54,000 SNPs between the independent GWAS, we performed a meta-analysis of Harlan SD, Charles River SD, and HS rats. We identified an additional 9 loci and replicated 4 previous associations. Notably, the meta-analysis identified a locus associated with multiple day 5 PavCA metrics that contained the gene *Taar1*, a trace amine receptor with strong ties to addiction.

4.2 Introduction

Pavlovian conditioning (classical conditioning) refers to a learning process in which a rewarding stimulus (unconditioned stimulus; US) such as food or drug is repeatedly paired with a neutral stimulus, creating a strong association between the two. Through this repeated association, the neutral stimulus (conditioned stimulus; CS) becomes a predictive cue for the reward's presence and may also acquire incentive salience. If attributed with incentive salience, this cue transforms from a once neutral stimulus into an incentive stimulus that is "wanted" and capable of eliciting motivated behavior (conditional response; CR) [53,263–265]. While evolutionarily beneficial, these environmental cues have the potential to acquire inordinate amounts of incentive salience due to sensitization of the neural networks that confer motivational value to reward cues, resulting in potentially negative behavioral outcomes [53,266,267]. This "incentive sensitization" of the brain's reward circuitry is believed to play a large role in susceptibility to developing addiction [265,266,268]. Specifically, these salient reward cues have the potential to trigger compulsive craving and reward-seeking behaviors, frequently leading to relapse [53,269].

Studying the attribution of incentive salience to reward-paired cues necessitates an approach that can separate predictive vs. incentive motivational learning. Pavlovian conditioned approach (PavCA) is a unique behavioral paradigm used in rodents that allows for isolation of the degree to which incentive salience is being attributed to a discrete environmental cue associated with reward [218]. In short, PavCA involves presenting a lever CS immediately prior to a non-contingent food or drug reward US over the course of five training sessions. While all rats learn the association between the CS and US, as shown by the development of a CR, the target of the CR varies considerably across individuals. A portion of animals will preferentially approach and engage the CS in the interval prior to reward delivery (sign-trackers; STs), whereas others will

preferentially approach the US location during this period (goal-trackers; GTs). Alternatively, animals may show a mixture of both CRs (intermediate responders; IRs). Only in STs has the lever CS truly gained incentive salience and become a desirable incentive stimulus, proven by the fact that they will work for the CS [74,219]. This has strong implications for addiction because STs have also shown greater cue-induced reinstatement of cocaine-seeking behavior [270,271] and cocaine self-administration [271]. Results from previous studies also suggested that the attribution of incentive salience to reward cues may be a heritable trait, as selected lines showed divergent ST/GT behavior [74,75]. Together, these findings make PavCA an appealing endophenotype for investigating the genetic risk factors that predispose individuals to be more vulnerable to become addicted to rewards.

Recently, we performed the first genome-wide association study (GWAS) for PavCA in a large sample of 4,061 Sprague Dawley (SD) rats, originating from either Charles River or Harlan [97]. Though the heritability estimates for PavCA were lower on average than have been observed for other psychiatric traits [7–9], we uncovered a number of loci linked to various metrics of performance in the PavCA paradigm. However, replication is crucial in GWAS studies [133,272]. Therefore, we simultaneously collected PavCA phenotype data on a sample of 2,449 heterogenous stock (HS) rats as part of a large, NIDA-funded center for GWAS in outbred rats (P50DA037844; www.ratgenes.org), focused on identifying QTL for various behavioral phenotypes relevant to drug addiction. HS rats have been successfully used for mapping numerous quantitative traits [40,110,273]. The HS rats were tested at two geographically distinct centers. At one center (University of Michigan; MI), rats were young and experimentally naïve when beginning PavCA training, whereas in the other center (University of Rochester; NY), the rats were older and had been exposed to multiple behavioral procedures prior to PavCA. Here we perform a replication of

our original SD PavCA GWAS. We first analyze the PavCA metrics separately in the MI-tested and NY-tested HS rats and then together in a mega-analysis of rats from both centers. We then meta-analyze the results of the Charles River and Harlan SD PavCA GWAS with the HS mega-analysis results; a cumulative sample size of over 6,400 rats, and the largest rodent GWAS to date. To our knowledge, this is also the first ever meta-analysis of GWAS results in rodents from different strains.

4.3 Methods

4.3.1 Heterogeneous stock rats

Samples phenotyped and genotyped for this study were part of a large, NIDA-funded center for GWAS in outbred rats (P50DA037844; www.ratgenes.org), focused on identifying QTL for various behavioral phenotypes relevant to drug addiction in heterogeneous stock (HS) rats. The collaboration involved labs at the University of Buffalo, University of California – San Diego, University of Michigan, and University of Tennessee Health Science Center. The NMcwi:HS breeding colony originated in 1984 and was acquired by the Medical College of Wisconsin (MCW) in 2006. A set of 64 breeder pairs is used to maintain the colony. Their breeding scheme takes into account pairwise kinship coefficients between animals to avoid inbreeding and maximize genetic diversity. Rats from generations 73-80 were shipped at 3-6 weeks of age to the phenotype testing locations in 22 batches over the course of 2.5 years (October 2014 – March 2017).

There were 2,449 HS rats tested for PavCA. Of these animals, 1,359 (682 male, 677 female) were tested at the University of Michigan (MI) and 1,090 (550 male, 540 female) were tested at University at Buffalo in Buffalo, New York, USA (NY). Animals were on average 63.4 weeks old (range [49,80]) when tested at the MI center versus an average of 165.1 weeks old (range

[141,205]) at the NY center. Animals tested in NY had also been exposed to 4 additional behavioral testing paradigms prior to undergoing PavCA training, whereas at those tested in MI were naïve.

4.3.2 Pavlovian conditioned approach

University at Buffalo (NY): Rats were housed in same-sex pairs and maintained on a reverse 12h:12h light-dark cycle. Rats were split into testing batches balanced by coat color, age and sex. Prior to PavCA, rats were tested for social behavior, locomotor activity, light reinforcement, reaction time, and delay discounting. Rats were then transferred to a new facility, and testing began after a week of habituation. For two days prior to testing, the rats were given ~25 banana-flavored pellets in their home cage to familiarize them with the reward. Prior to undergoing the PavCA procedure, there was one day of magazine training where 25 pellets are delivered on a 30-second variable time interval (0-60s) with no lever present to ensure they are reliably retrieving food pellets. Rats then underwent 5 days of a standard PavCA procedure [218]. Briefly, there were 25 trials a day during which an illuminated lever (conditioned stimulus; CS) is inserted into the testing chamber for 8 sec, immediately followed by delivery of a food pellet (unconditioned stimulus; US) into the magazine. CS-US pairings occurred on a VT 90-sec schedule (30-150s). All rats were tested during the dark phase at the same time of day.

University of Michigan (MI): Rats were housed three same-sex animals per cage, maintained at a reverse 12:12 light-dark cycle, and were handled on a daily basis prior to testing. Rats were split into seven testing batches balanced by coat color, age and sex. For two days prior to testing, rats were given ~20 banana flavored pellets in their home cage to familiarize them with

the reward. Testing began when rats were 55-75 days old. Pre-training and PavCA testing was carried out analogously to the NY center.

PavCA metrics: There are 11 PavCA metrics recorded or calculated for each day of testing. The primary recorded measures are the number of lever deflections and food magazine nose pokes during presentation of the CS summed across the 25 trials per training session (lever CS & magazine CS). In addition, the sum of the number of nose pokes into the food magazine during the inter-trial interval was recorded (magazine NCS). Calculated measures include: proportion of trials where the lever or magazine were interacted with at least once (probability of lever press or magazine entry, [0-1]), latency to approach of the lever or magazine averaged across trials (latency to lever press or magazine entry; [0-8s]). Additional summary metrics included: tendency to interact with lever or magazine (response bias; [Lever CS – Magazine CS]/[Lever CS + Magazine CS], difference between the proportion of trials with a lever interaction and proportion with a magazine during CS presentation (probability difference; [Prob Lev – Prob Mag]), and the difference between average latency to approach the magazine and the average latency to approach the lever divided by the duration of a single trial (latency score; [Avg Mag Lat – Avg Lev Lat]/8). Lastly, there is the PavCA index score, which is calculated as the average of the response bias, probability difference, and latency score for each day. The average of this score for days 4 and 5 is used to categorize animals' behavior into sign-tracking (PavCA index [0.5-1]), goal-tracking (PavCA index [-1-(-0.5)]), or intermediate responding (PavCA index [-0.5-0.5]).

4.3.3 Covariate selection and phenotype pre-processing

PavCA metrics tend towards highly non-normal distributions, in part due to their artificial bounds and the method by which the summary measures are constructed [97]. Therefore,

normalization was necessary prior to mapping the traits. A significant difference was observed in the distributions of average day 4 and day 5 PavCA index scores between the NY and MI testing centers (Kolmogorov-Smirnov Test; D = 0.18946, p-value < $2.2x10^{-16}$; Supplemental Figure 4.1). Significant differences were also seen between the distributions for males and females within both NY (D = 0.27773, p-value < $2.2x10^{-16-}$) and MI (D = 0.21219, p-value < $1.156x10^{-13}$). Due to these differences, we chose to subset the full sample into NY males, NY females, MI males, and MI females for normalization of the PavCA metrics. All metrics were quantile normalized within testing center and sex groupings, and then males and females were merged within center. Our quantile normalization procedure randomly assigns a rank to identical phenotypic values, potentially reducing our power to detect an association in traits with tail-heavy distributions.

Using the quantile normalized PavCA metrics for each center, we performed univariate linear regression for all remaining potential covariates (age at testing, testing chamber, and shipment batch number) on each PavCA metric. This was done separately for rats tested in NY and MI. We retained all covariates accounting for at least 1% of the variance in one or more PavCA metrics for model selection with the R package *leaps* [228]. Model selection with *leaps* was performed individually on each of the 55 metrics, and the model that explained the greatest portion of the variance for each metric was selected. For both testing centers, all covariates that had surpassed the 1% threshold were also included in at least 25% of the *leaps* selected models. Therefore, we simplified our analyses by regressing out the full set of covariates from each metric to obtain residuals. The 11 covariates selected for NY included: age at the onset of PavCA testing, and binary 'indicator' variables for shipment batches 1/2/6/7/8/9/12 (7 total) and testing boxes 2/6/8. The 11 covariates selected for MI included: age at the onset of PavCA testing and binary 'indicator' variables for shipment batches 1/2/3/6/7/11/12/18/19/20 (10 total). Residuals

were then quantile normalized prior to being used for GWAS, either in both testing centers individually, or across both centers in the case of the mega-analysis. Full covariate data was available for all 2,499 rats.

4.3.4 Genotyping and imputation

After completing testing at the NY and MI centers, spleen samples were collected from the HS rats and shipped to the University of California, San Diego. DNA was extracted from spleen tissue using the Agencourt DNAdvance Kit (Beckman Coulter Life Sciences, Indianapolis, IN) and genomic DNA quality and purity assessed by NanoDrop 8000 (Thermo Fisher Scientific, Waltham, MA). Sequencing libraries were prepared from DNA samples using a double digest genotyping-by-sequencing (ddGBS) approach previously described in Chapter 2. In brief, each library was composed of a set of 48 HS samples. DNA was digested by a combination of PstI and NlaIII. A sequencing adapter with a unique barcode of 4-8bp was ligated to the PstI cut site, and a Y-adapter was ligated to the NlaIII cut site. Sets of 48 sample libraries were pooled in approximately equal quantities and size selected on a Pippin Prep (Sage Science, Beverly, MA). Pooled and size-selected libraries were PCR amplified and quality checked on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Libraries were sequenced at the IGM Genomics Center (University of California, San Diego, La Jolla, CA), each using a single flow cell lane on an Illumina HiSeq 4000 with 100bp single-end reads.

The computational pipeline we used to call high-confidence genotypes from ddGBS data has also been detailed in depth previously in Chapter 2 as well. In short, sequencing data was demultiplexed using FASTX Barcode Splitter [170]. Reads were trimmed for barcodes, adapters,

and low-quality base pairs with Cutadapt [173]. Reads were aligned to reference genome build Rnor_6.0 using *-mem* Burrows-Wheeler Aligner (BWA) [175] and realigned around known indels from 42 whole-genome sequenced inbred laboratory rat strains [177], which includes the eight HS progenitor strains. Variants were call using the SAMtools genotype likelihood model in ANGSD v0.911 [165], and missing genotype calls were imputed with Beagle [166,167]. Lastly, using the Beagle imputed data, we imputed outwards to known sites from reference variant sets, resulting in 3,712,342 SNPs after filtering for MAF (>0.005), genotype missingness (<10%), HWE violations ($<1x10^{-10}$), and Mendelization errors.

4.3.5 SNP-based heritability estimates

Heritabilities estimates were obtained for each testing center, as well as the combined sample of 2,449 rats using the full set of 3,712,342 SNPs. All SNPs and samples were used to construct a single genetic relationship matrix (GRM) in GCTA [230]. We then applied the restricted maximum likelihood (REML) method within GCTA on the GRM and the 3 separate PavCA metric phenotype files (quantile normalized within NY or MI, or across NY+MI), containing NAs for the metric values from the alternate testing center in the single center analyses. Results are reported in Supplemental Tables 4.1, 4.2, and 4.3.

4.3.6 Genetic and phenotypic correlations

Genetic correlations between PavCA metrics were estimated using bivariate genome-based restricted maximum likelihood (GREML) analysis in GCTA [230,274]. Pairwise correlations were calculated between all quantile normalized metrics within each testing center, as well as for each

individual metrics between the two testing centers. The GRM and within-center quantile normalized metrics were the same used for heritability estimation. For the between-center analyses where PavCA metrics were assessed in different animals, the residual covariance between the two traits was dropped from the model. A log likelihood test was performed to obtain a p-value for the genetic correlation coefficient. The null hypothesis for correlation between within-center metrics was $rG \neq 0$ (two-tailed test), whereas for between center correlations of the same metric, the hypothesis was $rG > 0$ (one-tailed test). Reported values are the genetic correlation coefficient and corresponding p-value from either the one- or two-tailed test. Genetic correlations between PavCA metrics measured in NY vs MI are provided in Supplemental Table 4.4. Pairwise within-center genetic correlations are presented in Supplemental File 4.1 and plotted in Supplemental Figures 4.2 and 4.3. Phenotypic correlations were calculated as Pearson correlation coefficients between the quantile normalized metrics.

4.3.7 Linkage disequilibrium

We plotted the decay of linkage disequilibrium (LD) using the *--r²* utility in *plink 1.9* and the procedures described in Parker et. al 2016 [96]. The plot presented in Figure 4.1 is adapted from Gileta et al. 2018 [97] to compare the LD decay curve from the HS to that of the SD and CFW mice. Briefly, the curves include only SNPs with MAF > 20% and pairwise LD comparisons were limited to SNPs within 5% MAF of each other. A sample of 10,000 SNP pairs per 100kb interval between SNPs was used to estimate the average r^2 for SNPs up to 10Mb apart. The HS curve used a set of 102,725 SNPs, LD pruned for pairs with r^2 above 0.999.

4.3.8 GWAS

GWAS analyses were run using a linear mixed model in GEMMA [123]. The fixed effect covariates were regressed out prior, and the residuals were used for mapping. A GRM was included as a random effect term to control for population structure and relatedness with the goal of avoiding false positive associations. We chose to increase the accuracy of our modeling by estimating a GRM from SNP data rather than using the known pedigree data for the HS animals. Due to the pronounced effects of proximal contamination in structure populations with higher levels of LD [232,231], we also applied the leave-one-chromosome-out (LOCO) approach [157,233,234] when constructing the GRMs. By creating 20 individual GRMs, each lacking SNPs on the chromosome being tested, we avoid decreased power from including SNPs in the GRM that are in LD with target SNP. Samples missing a PavCA metric on a given day were removed from that analysis, hence slightly varying sample sizes.

Genotype data was in uncompressed Oxford format (.gen). All genotypes originally called using ANGSD and Beagle were coded as dosages on a continuous [0,2] scale, allowing for uncertainty in the call to be taken into account in the GWAS. The remaining genotypes that were imputed from the reference set by IMPUTE2 were strictly coded as (0, 1, or 2). Separate analyses were run for Michigan and New York by substituting all phenotype values for the alternate testing center to NAs. The combined mega-analysis of all NY and MI data used the phenotype data quantile normalized across all samples from both vendors. We report p-values from the likelihood-ratio test (LRT) performed by GEMMA. All Manhattan plots were created with a custom script in R and were stacked vertically to assist with comparison across analyses (Supplemental File 4.2). The supplemental file also contains an additional meta-analysis of the NY and MI results to compare with the mega-analysis. Samples numbers per analysis varied by phenotype and can be found in the plot titles. Q-Q plots were made with *qqman* in R [275] and can be found in

(Supplemental File 4.3). LocusZoom plots were created using the stand-alone software [235], custom SD SNP databases and LD calculations from the HS data, and a 3Mb flanking region. We additionally performed credible set analysis [276] to calculate the posterior probability of causality for each SNP within the associated interval. This method identifies the smallest set of SNPs that accounts for 99% of the posterior probability. The credible sets of SNPs have been included as tracks in the LocusZoom plots in Supplemental File 4.4 to further narrow down our QTL intervals.

4.3.9 GWAS meta-analysis

In addition to mega-analysis of rats tested in NY and MI, we also performed a meta-analysis of their summary statistics to investigate the concordance of the identified QTL. Meta-analysis was carried out utilizing a custom R script based on an allele frequency and sample size aware weighting of the z-statistics [139] derived from the beta and standard error estimate output of the GEMMA LRT. We then used the program METAL [138] to perform a sample-weighted meta-analysis across three populations: 2,281 Sprague Dawley (SD) rats from Harlan, 1,780 SD rats from Charles River, and the mega-analyzed sample of 2,449 HS rats, yielding a maximum possible sample size of 6,510. Final meta-analysis sample sizes varied based on available covariate and phenotype data and are reported on the plots presented in Supplemental File 4.5. Only the 54,116 SNPs that existed in all three datasets were included in the meta-analysis.

4.3.10 Significance thresholds

Commercially available and structured rodent population have far more extensive stretches of LD than are observed in humans, either by design, or due to breeding stock size or historic bottlenecks. Due to these large LD blocks, tests performed on adjacent SNPs for significance are

often not independent, meaning the effective number of tests performed is far fewer than the number of SNPs tested. Human GWAS uses a genome-wide significance threshold of $5x10^{-8}$ [236]; this would be overly conservative in rodent populations. As an alternative, several studies have used permutation testing [124,237] to establish a significance threshold, whereby phenotype and genotype data are iteratively reordered to sample from a null distribution. While this method can be computationally expensive for a large number of phenotypes [231], we quantile normalized all metrics prior to mapping, so we were able to perform a single round of 1000 permutations of the quantile normalized data to obtain a threshold 0.05 significance threshold of $2.51x10^{-6}$ ($-\log_{10}(p) = 5.6$). This threshold determined by naïve permutation was similar to that obtained from multiTrans [240], which takes into account the phenotypic covariance due to sample related and the heritability of the trait(s) (data not shown). We continued to use the threshold of $-\log_{10}(p) = 5.6$ as a conservative threshold for the HS and SD meta-analyses, which used only 54,116 SNPs.

4.4 Results

4.4.1 Pavlovian conditioned approach

A total of 2,449 HS rats were phenotyped for PavCA; 1,090 at the NY center and 1,359 at the MI center, split equally between males and females. The average of days 4 and 5 (D4D5) PavCA index score is typically used to classify rats as sign-trackers, goal-trackers, or intermediate responders. Therefore, we were interested in differences in the D4D5 PavCA index score distributions between both testing centers and sexes. We observed significant differences in this metric between the NY and MI centers (Supplemental Figure 4.1; Kolmogorov-Smirnov Test - D = 0.18946, p-value < $2.2x10^{-16}$). However, it is not possible to say whether the observed difference was due to difference in testing center environment and experimentation as this difference is confounded with both the

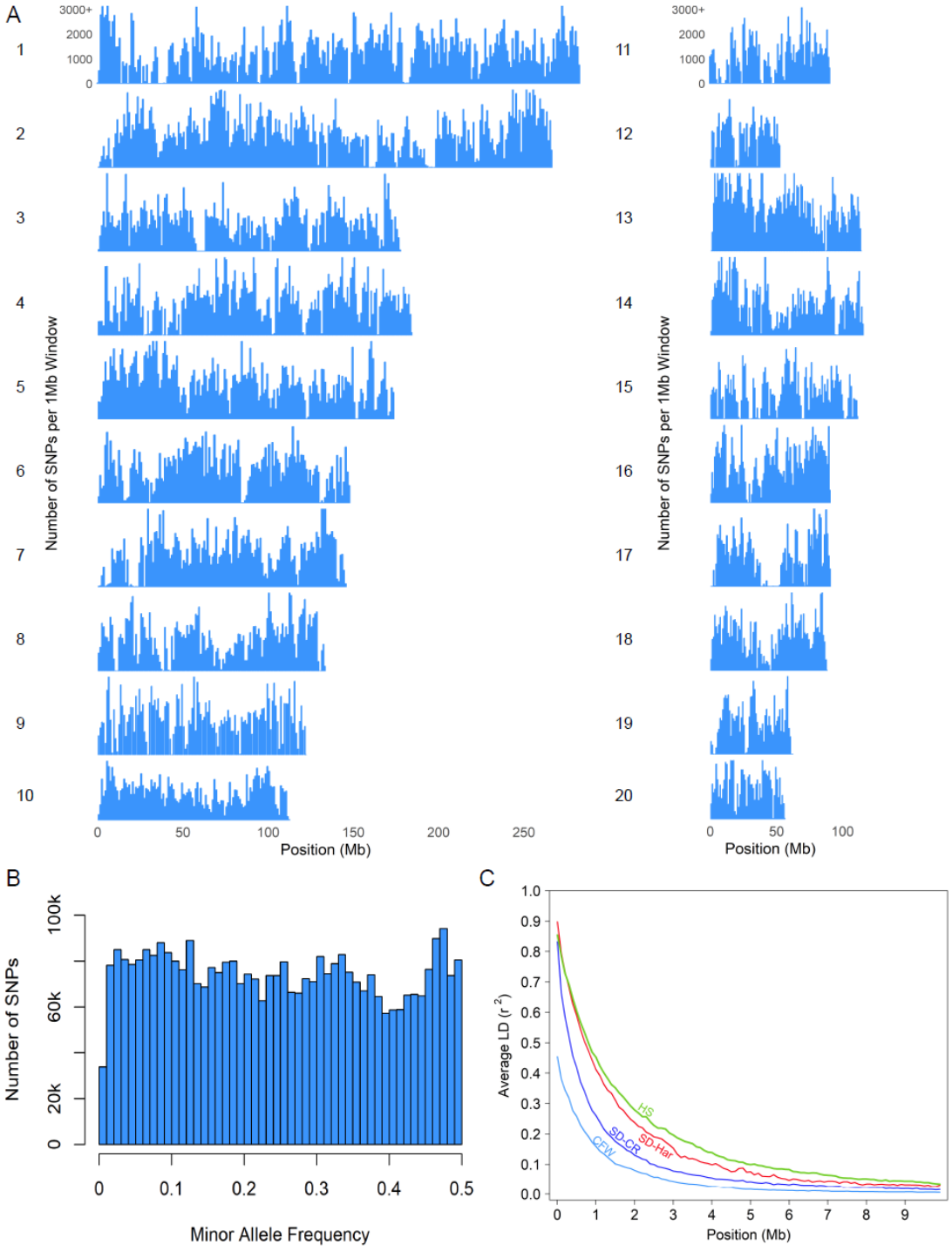
age of the rats (MI ~63 weeks; NY ~165 weeks) and exposure to previous behavioral testing. Interestingly, there was a significant difference between males and females within each center, with females displaying stronger sign-tracking behaviors than males. These results corroborate previous analyses on the HS rats tested at the MI center [267].

4.4.2 Genotype data

We obtained a set of 3,712,342 SNPs after variant calling, imputation to reference panels of known variant sites in the HS founder strains, and quality filtering. Figure 4.1A exhibits that our coverage of the genome is rather comprehensive, with approximately 1,400 SNPs per 1Mb interval on average. The minor allele frequency distribution of these variants (filtered for MAF > 0.005) is strikingly uniform (Figure 4.1B), providing us with improved power to detect associations compared to the SD populations. The decay of linkage disequilibrium (LD) in the HS is slower than observed in commercially available outbred populations of rats and mice (Figure 4.1C). This is expected because there have only been at most 80 generations for recombination to break down the original founder haplotypes in a relatively small breeding population. However, the greater extent of LD further increases the power of our GWAS, with the tradeoff of marginally decreased mapping resolution.

Figure 4.1. SNP coverage and allele frequency distribution and LD decay in the HS.

(A) SNP density per 1Mb window for the full set of 3.7 million imputed SNPs in the HS across all 20 autosomes. (B) Distribution of SNP minor allele frequencies after applying a filter for MAF < 0.005. (C) Linkage disequilibrium decay rates in HS rats as compared to SD rats from Harlan and Charles River and outbred Swiss Webster (CFW) mice.



4.4.3 SNP-based heritabilities and genetic correlations

Previously, we observed heritability estimates for the various PavCA metrics of 4-11% in the Harlan SD and 4-21% in the Charles River SD. Using GCTA, we constructed a GRM with all available SNP data and used it to estimate the proportion of variance explained by additive genetic effects in the HS rats. Consistent with our findings in SD, the SNP-based narrow-sense heritability estimates were in the range of 4.7-16.3% for the combined HS sample (Supplemental Table 4.3). However, due to the experiential and environmental differences between the testing centers, we also calculated the center-specific heritabilities. We observed significant estimates in the ranges of 4.4-24.2% for NY (Supplemental Table 4.1) and 4.6-22.1% for MI (Supplemental Table 4.2). The NY sample showed higher heritability estimated than MI by ~2% on average. The largest heritability estimates were for the first day of PavCA training, while days 2 and 3 typically had the lowest. Among the day 1 metrics, heritabilities were greater for the magazine-associated traits in NY, whereas in MI, the lever-associated traits were the highest. Interestingly, this pattern was reversed for the estimates for days 4 and 5 of training in NY, with lever-associated traits showing greater heritability.

Given the observed heritability estimates, we calculated genetic correlations for all PavCA metrics to investigate whether the additive genetic effects influencing the PavCA metrics were shared between the rats tested at the two centers (Supplementary Table 4.4). All correlations were estimated using the GCTA bivariate GREML analysis. Interestingly, although day 1 of training showed some of the highest heritabilities across the PavCA metrics, the genetic correlations between NY and MI were the lowest on this day. Numerous day 3, 4, and 5 PavCA metrics had genetic correlation estimates of 1, though these came with very large standard errors. We had

previously used this same method to look at the correlations between metrics in Harlan and Charles River SD rats, but few estimates were significantly different than zero (data not shown).

Several of the PavCA metrics are highly correlated with each other on a phenotypic level [97]. To see whether these high phenotypic correlations translated into genetic correlations, we estimated the pairwise genetic correlations between all metrics within either NY or MI (Supplemental Figures 4.2 and 4.3). For a portion of the PavCA metric pairs, an estimate could not be obtained due to limitations in the GCTA algorithm. While the signs of the genetic correlations were predominantly concordant, NY generally showed higher correlations between metrics than MI, despite similar levels of phenotypic correlation. Again, the correlation estimates showed large standard errors, making it difficult to draw definitive conclusions regarding the differences between the centers.

4.4.4 Association analyses in the HS

Due to the life history and age difference of the HS rats tested in NY vs. MI, we could not be certain identical phenotypes were being measured in each center. Therefore, we performed three GWAS for each of the PavCA metrics: one for the 1,090 rats tested in NY, one for the 1,359 rats tested in MI, and a mega-analysis of all 2,449 samples. Using GEMMA, we fit linear mixed models that included a GRM as a random effect term to control for allowed us to account for the familial structure of the sample. We tested all 3.7 million SNPs for association after regressing out significant covariates and normalizing the PavCA metrics appropriately. We discovered a total of 22 loci significantly associated with various PavCA metrics across the five days of training; seven loci unique to Michigan, five loci unique to NY, nine loci unique to the mega-analyses, and one

of which was shared between the NY and mega-analyses. Tables 4.1, 4.2, and 4.3 summarize these findings, with extended information available in Supplementary File 4.6. Manhattan plots and Q-Q plots for each GWAS are available in Supplemental Files 4.2 and 4.3, respectively. For each locus, we also performed credible set analysis [276] using LD information and the association p-values to identify a set of SNPs that was 99% likely to contain the "causal" SNP. We have reported the region over which these credible SNPs spanned, and the genes located within it. LocusZooom plots containing a track for the credible SNP set can be found in Supplemental File 4.4.

Table 4.1. Significantly associated loci for the NY-specific GWAS

Analysis	Top SNP	Ref/Alt Alleles	Alt AF	Phenotype	Training Day	MI Beta	NY Beta	Mega Beta	MI Pval	NY Pval	Mega Pval	Credible Set Start	Credible Set End	Known Genes within Credible Set
New York	chr4:133323878	G/A	0.62	Average Latency to Lever Press	Day 4	0.036	-0.216	-0.081	3.74E-01	2.04E-06	9.90E-03	chr4:132651681	chr4:135593823	Pdzrn3, Cntn3, Ppp4r2, Shq1, Rybp
	chr4:157791902	T/A	0.05	Magazine Entries w/ CS	Day 1	-0.067	0.449	0.187	4.69E-01	2.32E-06	5.72E-03	chr4:153846585	chr4:158504832	Vwf, Cd9, Tnfrsf1a, Ltbr, Scnn1a, Cd27, Tapbp1, Vamp1, Tuba3a, Nop2, Iffp1, Gapdh, Mrp151, Ing4, Pianp, Zfp384, Acrbp, Cops7a, Ptms, Lag3, Mlf2, Cd4, Gpr162, P3h3, Gnb3, Tpi1, Cdca3, Usp5, Spsb2, Atn1, Eno2, Lrrc23, Grcc10, Ptpn6, Phb2, Emg1, Lpcat3, C1s, C1r, C1r1, Clstn3, Pex5
	chr4:18725353	A/G	0.34	Latency Score	Day 2	0.017	-0.229	-0.082	6.92E-01	1.10E-06	1.08E-02	chr4:18347293	chr4:19475049	Sema3a, Sema3d
	chr4:19289093	A/C	0.39	Magazine Entries w/ CS		0.047	-0.234	-0.061	2.56E-01	1.70E-06	5.95E-02			
	chr4:19319672	C/A	0.35	Average Latency to Magazine Entry		0.026	-0.242	-0.085	5.37E-01	6.61E-07	9.78E-03			
				Proportion of Trials with Magazine Entry		-0.039	0.232	0.074	3.62E-01	1.82E-06	2.32E-02			
	chr6:69738076	A/T	0.37	Proportion of Trials w/ Lever Press	Day 5	-0.027	-0.236	-0.113	5.23E-01	1.19E-06	4.84E-04	chr6:69650202	chr6:70691928	Foxg1
	chr7:26724250	A/G	0.35	PavCA Index Score	Day 1	-0.024	0.226	0.079	5.50E-01	1.43E-06	1.22E-02	chr7:26724250	chr7:26783479	Chst11
	chr18:86028462	G/A	0.09	Probability Difference	Day 5	0.263	0.422	0.334	8.17E-04	6.78E-07	1.62E-08	chr18:85838595	chr18:86582998	Rttn, Socs6, Cd226, Dok6

Table 4.2. Significantly associated loci for the MI-specific GWAS

Analysis	Top SNP	Ref/Alt Alleles	Alt AF	Phenotype	Training Day	MI Beta	NY Beta	Mega Beta	MI Pval	NY Pval	Mega Pval	Credible Set Start	Credible Set End	Known Genes within Credible Set
Michigan	chr2:86705559	A/G	0.96	Average Latency to Lever Press	Day 1	0.632	-0.094	0.227	7.01E-07	4.49E-01	1.20E-02	chr2:86,644,565	chr2:86,859,622	Zfp458
	chr2:86790093	T/G	0.96	Proportion of Trials w/ Lever Press		-0.603	0.081	-0.232	2.30E-06	5.16E-01	1.05E-02			
	chr2:86810005	C/T	0.96	Response Bias		-0.602	0.029	-0.277	2.29E-06	8.17E-01	2.23E-03			
	chr2:177832772	C/T	0.95	Average Latency to Lever Press	Day 1	0.575	-0.058	0.240	2.23E-07	6.05E-01	2.94E-03	chr2:176558806	chr2:179723908	Rapgef2, Fnip2, Ppid, Etfdh, Fam198b, Tmem144, Gria2
				Lever Presses w/ CS		-0.526	0.056	-0.223	2.32E-06	6.14E-01	5.56E-03			
	chr2:177539122	A/G	0.94	Proportion of Trials w/ Lever Press		-0.624	0.084	-0.244	1.66E-06	5.08E-01	8.24E-03			
				Response Bias		-0.644	0.029	-0.294	7.20E-07	8.20E-01	1.39E-03			
	chr3:21424663	A/G	0.10	Lever Presses w/ CS	Day 5	0.294	-0.081	0.141	1.51E-06	2.89E-01	3.64E-03	chr3:13233164	chr3:24253315	Olr####, Pdc1, Rc3h2, Zbtb6, Gpr21, Zbtb26, Rabgap1, Strbp, Crb2, Dennd1a, Lhx2, Nr5a1, Nek6, Psmb7, Golga1, Arpc51, Rp135, Ppp6c, Olfm12a, Morn5, Ndufa8, Lhx6, Rbm18, Mrrf, Ptgs1, Dab2ip, Rab14, Gsn, Stom, Ggta1l1, Ggta1, Phf19, Psmd5, Fbxw2, Rabepk, Hspa5, Traf1, Pbx3
	chr7:144653965	G/A	0.99	Average Latency to Magazine Entry	Day 5	-1.195	-0.362	-0.792	2.50E-06	1.65E-01	1.69E-05	chr7:144513164	chr7:144680881	Hoxc4, Hoxc5, Hoxc8, Hoxc9, Hoxc10, Hoxc11, Hoxc12
	chr9:97360745	C/T	0.61	Magazine Entries w/ CS	Day 4	0.202	-0.040	0.096	6.75E-07	3.69E-01	1.78E-03	chr9:96866752	chr9:97681748	Ackr3, Asb18, Gbx2, Agap1
	chr11:83601840	C/A	0.26	Lever Presses w/ CS	Day 2	0.215	0.040	0.146	2.36E-06	4.29E-01	2.16E-05	chr11:82179958	chr11:83606270	Ephb3, Ehhadh, Thpo, Clcn2, Map3k13, Tmem41a, Liph, Senp2, Igf2bp2, Tra2b, Etv5
	chr17:57398753	A/G	0.72	Average Latency to Magazine Entry	Day 3	0.220	-0.044	0.105	1.44E-06	3.72E-01	2.27E-03	chr17:56433977	chr17:60543359	Epc1, Crem, Bambi, Cul2, Adarb2, Idi1, Wdr37, Rab18, Mpp7
				Magazine Entries w/ CS		-0.222	0.043	-0.107	1.17E-06	3.81E-01	1.90E-03			
				Proportion of Trials with Magazine Entry		-0.263	0.031	-0.132	8.77E-09	5.26E-01	1.15E-04			

Table 4.3. Significantly associated loci for the HS mega-analysis.

Analysis	Top SNP	Ref/Alt Alleles	Alt AF	Phenotype	Training Day	MI Beta	NY Beta	Mega Beta	MI Pval	NY Pval	Mega Pval	Credible Set Start	Credible Set End	Known Genes within Credible Set
Mega-Analysis	chr2:138401477	G/A	0.40	Average Latency to Lever Press	Day 4	0.122	0.166	0.148	1.87E-03	2.29E-04	1.02E-06	chr2:137,204,443	chr2:138,830,334	Pcdh18
				Lever Presses w/ CS	Day 5	-0.157	-0.147	-0.154	6.80E-05	1.15E-03	3.76E-07			
	chr2:138402214	C/T	0.41	Latency Score	Day 5	-0.125	-0.162	-0.143	1.51E-03	3.80E-04	2.51E-06			
	chr2:166308959	G/A	0.62	Proportion of Trials w/ Lever Press	Day 4	0.118	0.171	0.149	3.46E-03	1.78E-04	1.60E-06	chr2:165086742	chr2:167019450	Nmd3, Sptssb, B3galnt1, Ppm1l, Arl14, Kpna4, Smc4, Trim59, Ift80
	chr5:42418647	G/A	0.22	Average Latency to Lever Press	Day 5	0.185	0.171	0.177	1.90E-04	1.65E-03	2.49E-06	chr5:41582655	chr5:43059005	N/A
	chr5:42724248	C/G	0.21	Proportion of Trials w/ Lever Press	Day 4	-0.159	-0.229	-0.192	1.40E-03	2.63E-05	3.59E-07			
				Response Bias		-0.118	-0.244	-0.183	1.88E-02	7.88E-06	1.29E-06			
	chr5:42749488	G/C	0.22	Lever Presses w/ CS		-0.179	-0.210	-0.196	2.30E-04	7.88E-05	1.22E-07			
	chr5:42786309	A/T	0.23	Lever Presses w/ CS	Day 5	-0.207	-0.211	-0.208	3.03E-05	9.39E-05	3.22E-08			
	chr6:9775598	T/C	0.09	Average Latency to Magazine Entry	Day 1	0.238	0.302	0.265	6.01E-04	1.98E-04	9.09E-07	chr6:8289654	chr6:10451004	Prkce, Eif3h, Epas1, Six2, Six3, Camkmt, Prep1, Slc3a1
	chr8:59947765	C/T	0.55	Probability Difference	Day 1	0.166	0.129	0.153	3.76E-05	3.90E-03	4.21E-07	chr8:57227186	chr8:60026338	Tmem266, Nrg4, Fbxo22, Chrna3, Chrnb4, Chrna5, Psma4, Hykk, Ireb2, Crabp1, Wdr61, Dnaja4, Acsbg1, Idh3a, Cib2, Gldn, Cyp19a1, Elmod1, Sln, Slc35f2, Rab391, Cu15, Acat1,
	chr8:59951663	C/T	0.56	PavCA Index Score		0.165	0.139	0.156	3.80E-05	1.72E-03	2.00E-07			
				Latency Score		0.167	0.152	0.162	2.51E-05	6.02E-04	5.93E-08			
	chr8:100969613	G/A	0.49	Magazine Entries w/o CS	Day 1	-0.169	-0.130	-0.150	1.67E-05	3.12E-03	5.18E-07	chr8:100276674	chr8:101166439	N/A
				Average Latency to Magazine Entry		0.146	0.122	0.142	1.83E-04	5.55E-03	2.01E-06			
	chr11:30824894	A/G	0.10	Magazine Entries w/o CS	Day 1	0.241	0.278	0.258	9.36E-04	2.22E-04	1.68E-06	chr11:30148919	chr11:30975396	Hunk, Mis18a, Mrap, Urb1, Scaf4, Sod1
	chr13:57185378	A/C	0.18	Magazine Entries w/o CS	Day 1	-0.170	-0.210	-0.189	1.27E-03	2.85E-04	2.29E-06	chr13:56584377	chr13:57185378	Kcnt2, Cfh, Cfhr1, F13b, Aspm
	chr18:85861741	A/T	0.08	Magazine Entries w/ CS	Day 4	-0.243	-0.343	-0.278	2.24E-03	4.16E-05	2.34E-06	chr18:85838595	chr18:86582998	Rttn, Socs6, Cd226, Dok6
				Response Bias		0.265	0.392	0.313	7.92E-04	3.95E-06	1.37E-07			
	chr18:85880215	A/G	0.07	Magazine Entries w/ CS	Day 5	-0.225	-0.416	-0.306	5.21E-03	1.33E-06	3.03E-07			
	chr18:86028462	G/A	0.09	Average Latency to Magazine Entry	Day 5	0.254	0.373	0.305	1.23E-03	9.36E-06	2.21E-07			
				PavCA Index Score		0.225	0.400	0.303	4.35E-03	2.85E-06	3.30E-07			
				PavCA Index Score	Day 45	0.254	0.393	0.315	1.36E-03	4.00E-06	1.19E-07			
				PavCA Index Score	Day 4	0.228	0.391	0.297	4.07E-03	3.69E-06	4.95E-07			
				Probability Difference	Day 5	0.263	0.422	0.334	8.17E-04	6.78E-07	1.62E-08			
	chr18:86256424	A/G	0.05	Latency Score	Day 5	0.296	0.519	0.404	5.56E-03	5.97E-06	4.27E-07			
				Proportion of Trials w/ Lever Press		0.287	0.530	0.407	7.21E-03	3.69E-06	3.45E-07			
	chr18:86352576	G/A	0.05	Lever Presses w/ CS	Day 5	0.282	0.566	0.414	9.51E-03	1.33E-06	3.51E-07			
	chr20:2518031	C/T	0.45	Magazine Entries w/ CS	Day 1	0.167	0.156	0.159	8.81E-05	1.07E-03	9.91E-07	chr20:772414	chr20:3745667	Olr####, Rt1-M#-#, Ubd, Gabbr1, Mog, Zfp57, Znrd1as1, Znrd1, Ppp1r11, Rnf39, C4a, Btn15, Btn17, Btn18, Gnl1, Prr3, Abcf1, Ppp1r10, Mrps18b, Dhx16, Ppp1r18, Nrm, Mdc1, Tubb5, Flot1, Cb707485, Ddr1, Gtf2h4, Vars2, Sfta2, Dpcr1, Cchcr1, Tcf1

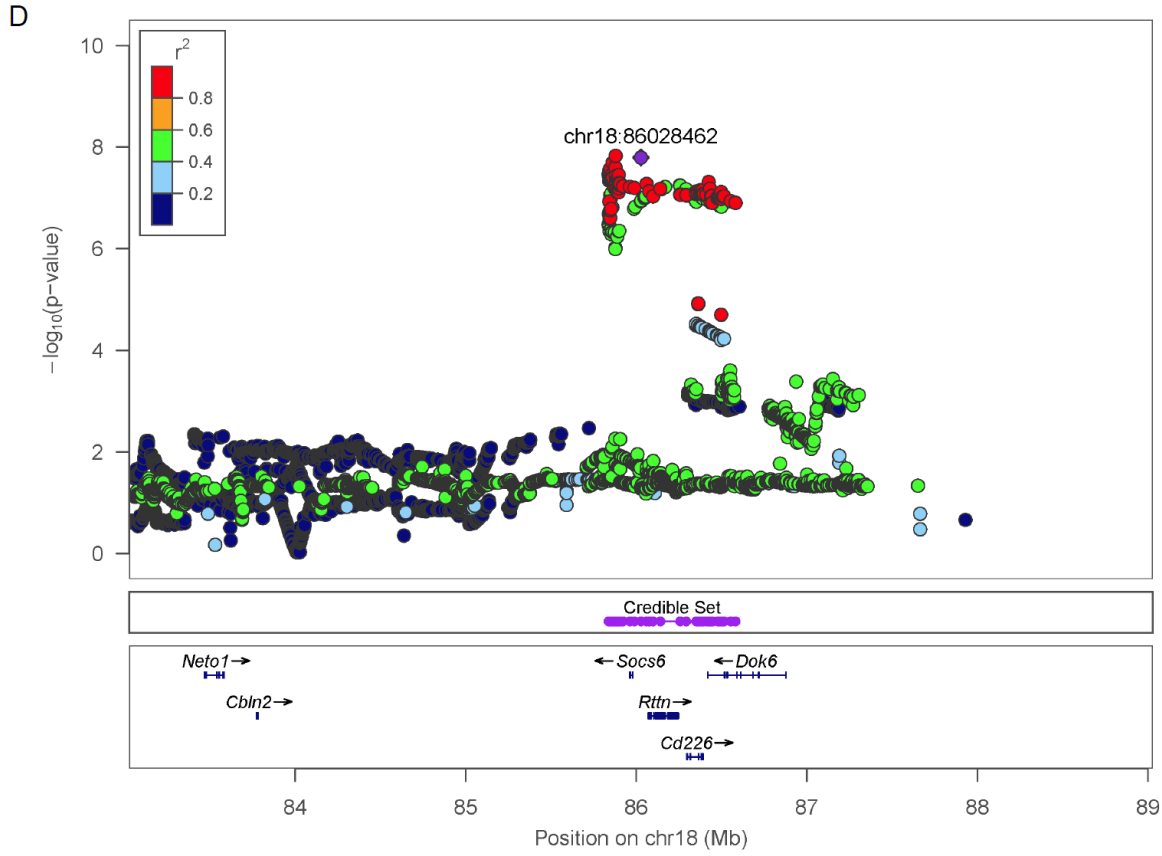
The most promising locus we identified in the mega-analysis was located on chromosome 18 (Figure 4.2). This locus was significantly associated with 11 metrics across days 4 and 5 of training, including the PavCA index score used to categorize rats as sign- or goal-trackers, as well as one metric in the NY-specific GWAS. The credible set for this locus spans 0.75Mb and contains only four known genes. Two of these genes show links to neural development. Little information exists on *Dok6*; however, it has been shown to promote neurite growth *in vitro* and is expressed in the adult brain [277]. The other gene, *Rttn* (Rotatin), has known links to development of the cerebral cortex in mice and humans. Mutations in *Rttn* can cause severe brain malformations and microcephaly [278,279]. In human fibroblast lines, *Rttn* mutants show abnormal ciliary structure and downregulation of key regulators of cortical patterning expressed in the cortical hem, which gives rise to Cajal-Retzius neurons [280]. Rotatin has been shown to colocalize with the Cajal-Retziuz neurons [280]. These neurons are responsible for producing reelin, a key glycoprotein that

regulates neuronal migration in the developing brain [281] and helps control synaptic plasticity [282] and dendrite development [283] in adults. Cajal-Retzius neurons, through their production of reelin, have been implicated in several neuropsychiatric disorders including: Alzheimer's disease [284], schizophrenia [285], autism [286], and bipolar disorder [287].

Figure 4.2. Manhattan and LocusZoom plots for the GWAS for probability difference on day 5 of PavCA training in MI, NY, and the HS mega-analysis.

All Manhattan plots are for probability difference on day 5 of training and contain the full set of 3.7 million SNPs. (A) MI-specific GWAS, (B) NY-specific, and (C) HS mega-analysis. (D) A LocusZoom plot of the 3Mb flanking region of the associated chr18 locus for the mega-analysis. The credible set is included as a track in purple below the points.

Figure 4.2., Continued

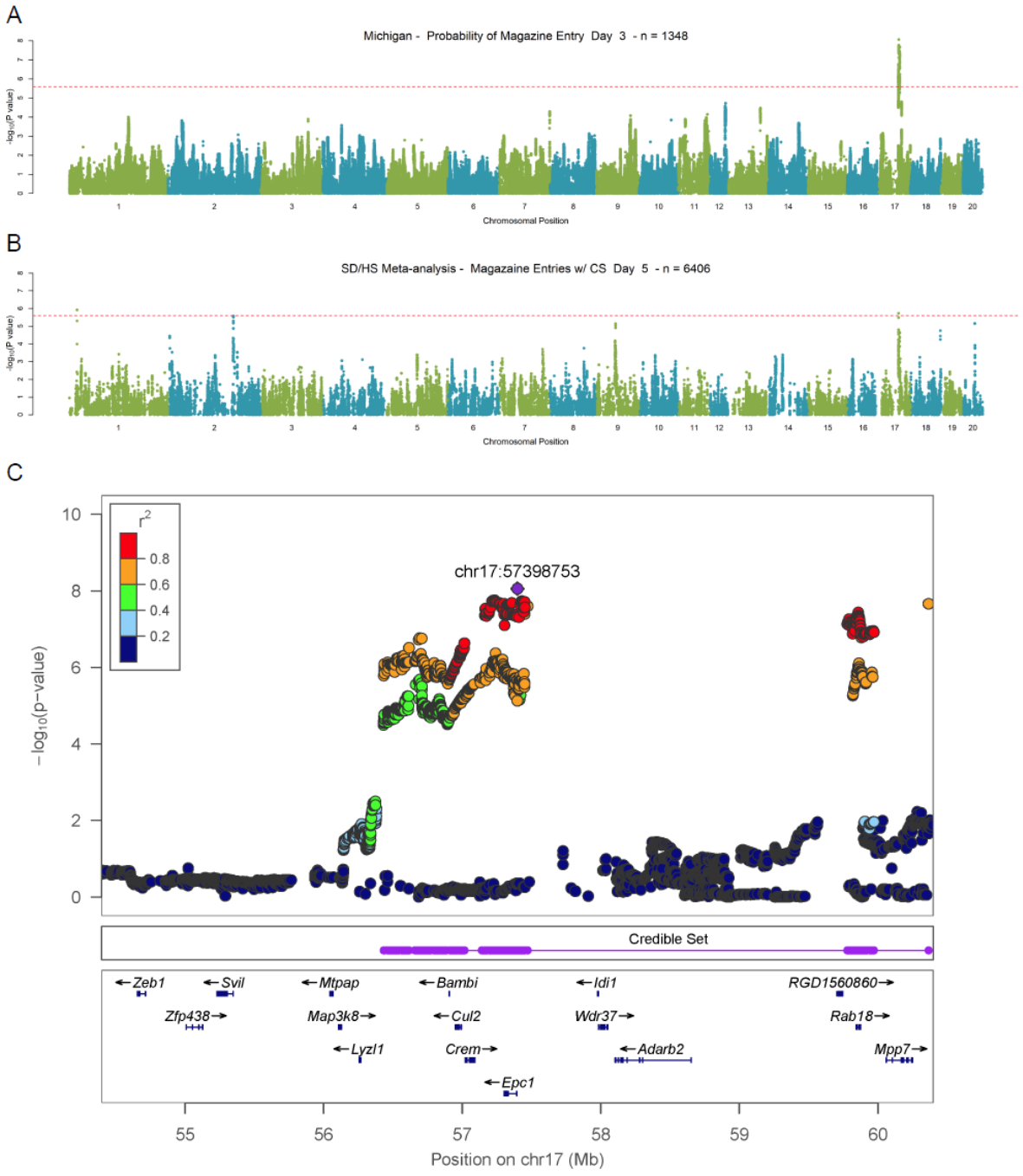


In addition to the chromosome 18 locus, we identified a locus on chromosome 17 strongly associated with three correlated day 3 magazine entry metrics in the Michigan center. This locus had been previously identified in our Harlan SD GWAS as associated with the second principal component summarizing all metrics across all days of training. Proximal to the top identified SNP in this region is the gene *Crem* (CAMP Responsive Element Modulator). Studies have shown that mice deficient in *Crem* have abnormal locomotor activity and lower levels of anxiety [288] and that disruption of CREM and CREB postnatally can lead to neurodegeneration of the striatum [289], which contains crucial components of the brain's dopamine-mediated reward system. In humans, *Crem* was found to be expressed widely in neurons of post-mortem adult brain regions important for learning and memory, such as the hippocampus and prefrontal and temporal cortex

[290]. Also within the credible interval for the association was *Rab18*, known to influence neuronal migration during cortical development through modulation of N-cadherin levels [291].

Figure 4.3. Manhattan and LocusZoom plots for GWAS-identified locus on chromosome 17.

(A) Manhattan plot of the MI-specific GWAS for proportion of trials with a magazine entry on day 3 of training. (B) Manhattan plot of the SD and HS meta-analysis for magazine entries in the presence of the CS on day 5 of training. (C) LocusZoom plot associated with Panel A.



4.4.5 Meta-analysis of HS and SD

Table 4.4. Significantly associated loci for the HS and SD meta-analysis.

Top SNP	Effect/Non-Effect Alleles	Effect AF	N	Phenotype	Day	Direction	P-value	Credible Set Start	Credible Set End	Credible Set Gene List
chr1:22744087	T/C	0.2144	6406	Average Latency to Magazine Entry	Day 5	+++	1.13E-07	chr1:22584681	chr1:22746462	Vnn1, Vnn3, Taar1
		0.2145	6395	PavCA Index Score		+++	1.01E-06			
		0.2144	6406	Magazine Entries w/ CS		---	1.22E-06			
		0.2147	6405	Probability Difference		+++	4.09E-07			
chr2:188162987	T/C	0.491	6406	Average Latency to Magazine Entry	Day 5	---	3.53E-07	chr2:187528608	chr2:188893673	Gon4l, Syt11, Rit1, Msto1, Dap3, Ash1l, Fdps, Pklr, Hcn3, Clk2, Gba, Mtx1, Scamp3, Fam189b, Ssr2, Arhgef2, Lamtor2, Ubqln4, Rab25, Lmna, Sema4a, Pmf1, Bglap, Paqr6, Slc25a44, Glmp, Tmem79, Cct3, Tsacc, Rhbg, Mef2d, Thbs3, Muc1, Trim46, Krtcap2, Dpm3, Slc50a1, Efna1, Efna4, Adam15, Zbtb7b, Shc1, Flad1, Lenep, Cksb1, Shc1, Pygo2, Pbxip1, Kcnn3, Pmvk
		0.4915	6394	PavCA Index Score	Day 45	---	2.47E-06			
		0.4911	6405	Probability Difference	Day 5	---	1.52E-06			
		0.4911	6405	Proportion of Trials with Magazine Entry	Day 5	+++	2.35E-06			
		0.4915	6384	Response Bias	Day 3	---	4.66E-07			
		0.4913	6395	Response Bias	Day 5	---	7.36E-08			
chr2:199231752	A/G	0.4048	6394	PavCA Index Score	Day 2	---	5.96E-07	chr2:199225113	chr2:199231752	N/A
		0.4049	6416	Lever Presses w/ CS		---	1.35E-06			
		0.4047	6419	Probability Difference		---	1.41E-06			
		0.4048	6394	Response Bias		---	2.39E-06			
chr3:13294366	A/G	0.8454	6389	Average Latency to Lever Press	Day 1	---	9.76E-07	chr3:13262120	chr3:13294746	Pbx3
		0.8455	6419	Lever Presses w/ CS		+++	5.10E-07			
		0.8454	6389	Probability Difference		+++	2.10E-06			
		0.8454	6389	Proportion of Trials with Lever Press		+++	8.10E-07			
chr3:49190567	T/C	0.7411	6415	Magazine Entries w/o CS	Day 1	+++	2.35E-06	chr3:47435550	chr3:49955823	Kcnh7, Gcg, Ifih1, Fap, Dpp4, Slc4a10, Tbr1, Psm14, Tank
chr7:125869426	T/C	0.3791	6414	Magazine Entries w/o CS	Day 2	---	1.48E-06	chr7:125865679	chr7:125869426	RGD1204694/Kiaa0930
chr9:54520074	T/C	0.2131	6405	Latency Score	Day 5	---	1.37E-06	chr9:53145306	chr9:55504490	Gls, Stat1, Stat4, Myo1b, Nabp1, Sdpr, Tmeff2, Nab1, Nemp2, Mfsd6, Inpp1, Hibch, Mstn, Pms1
		0.2131	6406	Average Latency to Magazine Entry		---	1.94E-06			
chr10:2617832	T/C	0.4222	6389	Proportion of Trials with Lever Press	Day 1	+++	9.16E-07	chr10:1892581	chr10:2830334	N/A
chr10:95877686	T/C	0.6018	6416	Lever Presses w/ CS	Day 2	+++	2.09E-06	chr10:95772845	chr10:95940367	Helz, Cacng1
		0.602	6419	Average Latency to Lever Press		---	1.92E-06			
chr11:14746031	A/G	0.6921	6419	Average Latency to Magazine Entry	Day 2	---	2.53E-09	chr11:14744594	chr11:14746031	N/A
		0.6922	6418	Magazine Entries w/ CS		+++	2.86E-09			
		0.692	6414	Magazine Entries w/o CS		+++	5.65E-09			
		0.6921	6419	Proportion of Trials with Magazine Entry		+++	8.75E-09			
chr17:56527126	A/G	0.5192	6406	Magazine Entries w/ CS	Day 5	---	1.91E-06	chr17:56433977	chr17:57320503	Crem, Bambi, Cul2, Epc1
chr18:85880215	A/G	0.5907	6405	Probability Difference	Day 5	---	8.47E-07	chr18:85880215	chr18:85893838	N/A
chr20:34118749	A/G	0.1263	6405	Probability Difference	Day 5	+++	6.03E-07	chr20:34118749	chr20:34287647	Slc35f1
		0.1263	6405	Proportion of Trials with Magazine Entry		---	2.99E-07			
		0.1263	6395	Response Bias		+++	1.25E-06			

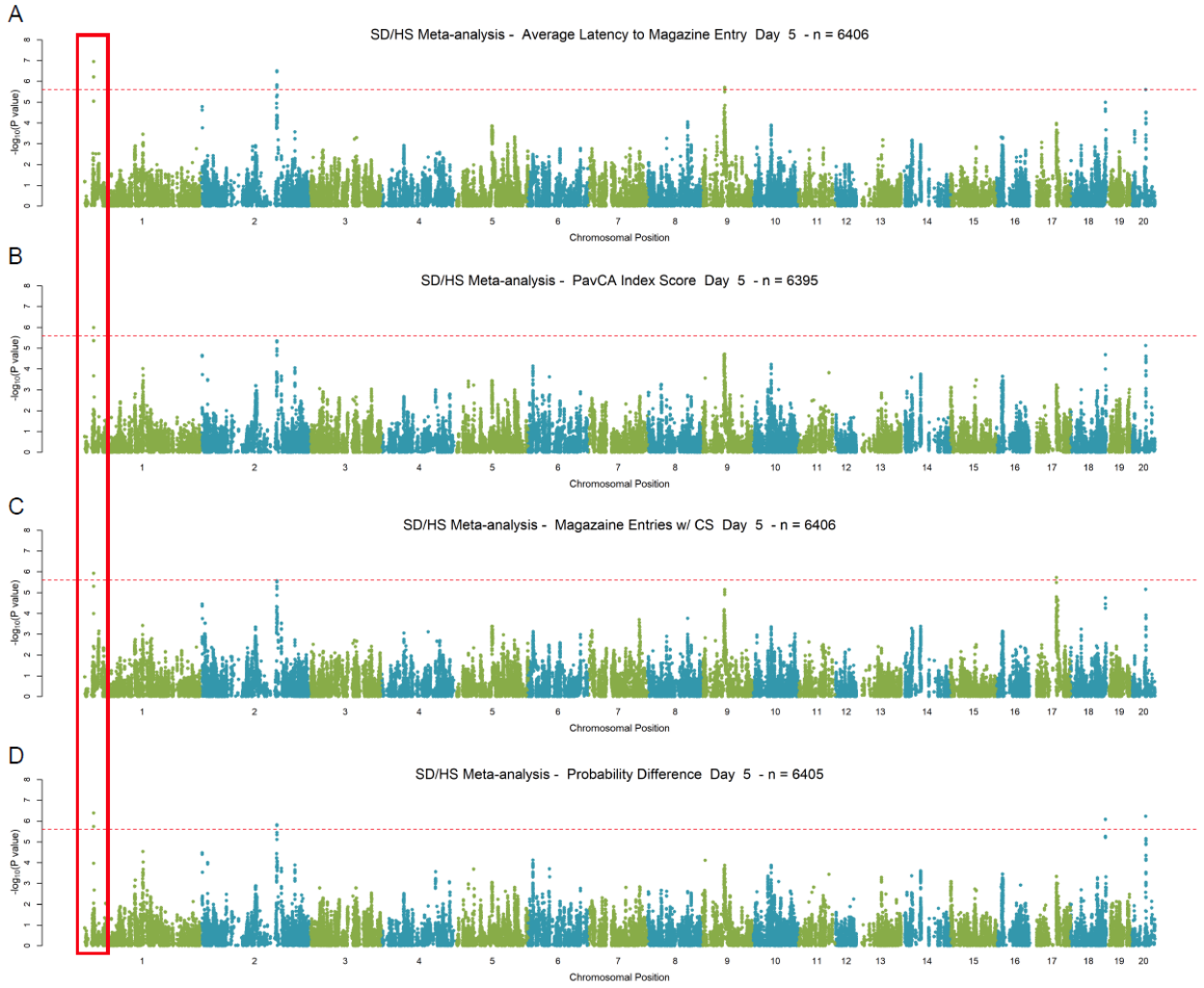
With multiple large, independent samples of rats from different populations tested for PavCA, we decided to perform meta-analysis of the summary statistics from the Harlan SD, Charles River SD, and HS using the standard sample-weighting approach in METAL [138]. We limited the meta-analysis to variants present in all three GWAS, leaving 54,116 SNPs. Though sample size varied by PavCA metric, the total sample of rats was approximately 6,400. We identified 13 loci in the meta-analysis, more than half of which were associated with multiple PavCA metrics (Table 4.4). Directions of effect for all loci were consistent across all three populations and showed the correct sign for the PavCA metrics (i.e. a locus that had a negative effect on magazine entries, had positive

effect on the PavCA index score). Additionally, the p-values in each population were generally trending towards significance (Supplemental File 4.6), indicating the large sample size afforded by meta-analysis of the three studies was necessary to provide sufficient power to detect these loci.

Only four of the 13 loci had been previously identified by our SD or HS GWAS, including the hits on chromosomes 17 and 18 discussed above, as well as two additional loci on chromosomes 7 and 11 associated with metrics on day 2 of training in the meta-analysis of Harlan and Charles River SD rats. Interestingly, the chromosome 17 locus containing *Crem* was associated with a different PavCA metric than in either MI or SD Harlan: magazine entries in the presence of the CS on day 5 of training. The chromosome 18 locus containing *Rttn*, though identified for the same trait in the NY-specific and HS mega-analysis, showed a marginally weaker association in the meta-analysis. In addition to these four recapitulated loci, the meta-analysis identified a novel locus on chromosome 1 associated with four day 5 PavCA metrics (Figure 4.4). Within the credible set for this locus were three genes. *Vnn1/3* are amidohydrolases that are unlikely to influence neural networks. The other gene, *Taar1*, codes for a trace amine-associated receptor that plays a significant role in the regulation of dopamine [292], the key neurotransmitter underlying the brain's reward system. Mice lacking TAAR1 showed greatly increased sensitivity of postsynaptic D2 dopamine receptors in the striatum [293]. TAAR1 agonists also include molecules such as norepinephrine, serotonin, amphetamine, methamphetamine, and MDMA [294]. A wealth of research has been accumulated on TAAR1 and its links to stimulant addiction [295]. This work, in conjunction with our GWAS results, make *Taar1* a prime candidate for further study in relation to PavCA behaviors and addiction.

Figure 4.4. Manhattan and LocusZoom plots for GWAS-identified locus on chromosome 1.

Manhattan plots of the SD and HS meta-analysis for PavCA metrics on day 5 of training, including: (A) average latency to magazine entry, (B) PavCA index score, (C) magazine entries in the presence of the CS, and 9D) probability difference.



We selected the three loci on chromosomes 1, 17, and 18 to highlight due to the limited number of candidate genes in the credible sets and their putative ties to neural networks associated with PavCA-related behaviors. Numerous additional hits were discovered, predominantly for days 1, 4, and 5 of training.

4.5 Discussion

In this study, we used 2,449 HS rats to perform a replication of our heritability estimates and GWAS for PavCA in SD rats. Though PavCA metric heritability estimates differed from those calculated in the SD rats, the new estimates were in the range of 4.4-24.2%, confirming the low additive genetic heritability of this behavioral phenotype. However, considering previous observations about the "selectablility" of sign- vs. goal-tracking behaviors [75], it is possible the broad-sense heritability of PavCA is much greater and that gene-gene interactions play a substantial role in determining the phenotype. Interestingly, the highest heritabilities were seen for day 1 PavCA metrics, when the previous rats' previous experiences likely had the strongest influence. Behaviors on this day likely reflect novelty-seeking or locomotor activity and only show light phenotypic and genetic correlation with PavCA metrics on later days of training. Heritability estimates typically dipped on days 2 and 3 of training and rose again on days 4 and 5 as the behavior became more stably learned.

The genetic correlation estimates for PavCA metrics between NY and MI were (aside from one metric) significantly different than zero. However, the standard errors for the estimates were quite large, leaving uncertainty as to the true degree of shared additive genetics effects between the two centers. Previously, we had also estimated the genetic correlations between Harlan and Charles River and saw that very few metrics had significant correlations. A potential explanation for this could be a high level of polygenicity of the behaviors involved in PavCA. Though behavioral phenotypes may resemble each other across rat populations, they are likely subject to genetic heterogeneity, where several different possible combinations of alleles at a large number of loci are able to produce the same behavioral outcome. This is supported further by the high level

of genetic differentiation that has occurred between Harlan and Charles River, indicating that different alleles are likely at play in each population.

We also estimated within-center genetic correlations for all pairs of metrics. We observed higher genetic correlations between metrics within NY versus within MI, possibly due to the more advanced age of the rats at testing in NY or their previous exposure to behavioral training requiring interactions with a lever. Notably, we had observed a flip in the NY heritability estimates, where lever associated traits gradually showed increased heritability while progressing from day 1 to day 5. This shift was reflected in the within population genetic correlations as well. Initially, lever deflections in the presence of the CS and proportion of trials with a lever interaction were significantly, positively genetically correlated with magazine entries with CS and the proportion of trials with a magazine entry. However, by days 4 and 5 of training, these traits were significantly inversely correlated, suggesting the animals had learned the US-CS association and the CR behaviors had dichotomized into sign-tracking and goal-tracking.

Despite low heritabilities, our large sample size provided us with sufficient power to discover several genome-wide significant loci associated with various PavCA metrics across the five days of training. The NY and MI center-specific GWAS and HS mega-analysis were fruitful, identifying a combined 22 genetic loci associated with 35 out of 56 of the PavCA metrics. We have not performed multiple testing correction for the number of metrics we tested; however, since these metrics are highly phenotypically correlated, we suspect the true number of independent tests is far lower than 56. This suggests that our finding of 22 distinct loci is substantially greater than would be expected with a false discovery rate of 5%. Given the low heritabilities of the PavCA metrics, it is perhaps unsurprising that there was little overlap between the results from the center-specific GWAS and the mega-analysis. The majority of the center-specific associations are for

days 1 and 2 of training, which was also when genetic correlations were the lowest and previous experiences likely had the largest influence on behaviors. The center-specific analyses may also have identified loci of high effect that are capturing differences in behavior unique to young, naïve rats (MI) or older, experienced rats (NY), whereas the mega-analysis identified shared PavCA loci of smaller effect. This is supported by the observation that all loci identified in the mega-analysis were trending towards significance in the center-specific analyses.

The mega-analysis of the HS rats exclusively identified loci associated with days 1, 4, and 5 of training. Interestingly, only 1 of these loci overlapped with those identified in the center-specific GWAS, again suggesting the regions identified in the mega-analysis are associated with behaviors with shared genetic underpinnings between centers, rather than aspects unique to younger, naïve or older, experienced rats. On day 1 of training, it is likely that the identified loci are linked to novelty-seeking behaviors or locomotor activity. Three of the identified day 1 loci were associated with the number of magazine entries in the absence of the CS. This metric was not significantly genetically correlated with any other metric during later training days, indicating these loci likely capture a distinct behavior from PavCA. By days 4 and 5 of training, the rats of reliably learned the CS-US association and show a consistent CR upon presentation of the CS, leading us to believe these loci are more likely directly linked to PavCA behaviors.

Our mega-analysis discovered a locus on chromosome 18 associated with numerous PavCA metrics on days 4 and 5 of training. Within this region were two genes with ties to neuronal growth/development, *Rttn* and *Dok6*. In particular, *Rttn* modulates the production of a protein called reelin that has been previously linked to numerous other psychiatric disorders [296]. It is possible there are mutations in the regulatory elements controlling *Rttn* that have a downstream effect on brain development, and therefore psychiatric outcomes. Further, the reelin produced by

the CR neurons that rotatin putatively controls also supports long-term potentiation, which is important to synaptic plasticity and learning [282,296]. In addition to *Rttn*, another locus on chromosome 17 was identified in both our MI-specific and SD GWAS that contained two genes with known links to brain morphology, *Crem* and *Rab18*. *Crem* stood out due to its role in postnatal degeneration of the striatum and expression pattern in the brain. Both *Rttn* and *Crem* may warrant further investigation into the downstream effects of modulating their expression on behavior.

Lastly, we performed, to our knowledge, the first large-scale meta-analysis of outbred rodent populations with a total sample size of ~6,400 rats. Of the 13 loci identified in the meta-analysis of the HS and SD populations, only four had been identified in the population-specific GWAS. Notably, we replicated the association with the chromosome 17 locus containing *Crem* and *Rab18*. It is not clear why this locus would be associated with various PavCA metrics in the SD GWAS, MI-specific GWAS and meta-analysis, but not NY GWAS. However, the SD rats were closer in age to the HS tested in MI than the HS tested in NY, and therefore the locus could be associated with behavior unique to a certain developmental time period. Failure to replicate the remaining majority of associations from the population-specific GWAS was due in part to the limited number of SNPs that overlapped between the SD and HS datasets (54,116). Additionally, the allele frequencies for the overlapping SNPs varied greatly between population. Several of the loci identified in the HS and SD meta-analysis has a low minor allele frequency in at least one of the three populations. Importantly, all of the nine newly uncovered loci showed elevated p-values in the population-specific analyses, indicating that the large sample size in the meta-analysis is what allowed for their discovery. This is a trend often seen in human GWAS for psychiatric traits [20,22,29].

Surprisingly, our strongest association in the HS and SD meta-analyses was for a set of metrics on day 2 of training. The locus was on chromosome 11 and contained no genes within its credible set. On day 2 of training, it is difficult to say exactly what behavioral processes or learning are occurring. This day of training also often showed the lowest heritability across metrics. However, the identified locus appears to be important as it was consistently identified across studies for the same metrics and day of training. The most exciting finding from the HS and SD meta-analysis was a locus on chromosome 1 containing *Taar1*, a well-studied trace-amine receptor with known links to amphetamine addiction and relapse [295]. Mouse knockout models of TAAR1 showed increased levels of extracellular dopamine [297], increased place preference on treatment with methamphetamine, and a slower rate of extinction [298]. Partial agonists for this receptor have also been shown to lessen cue-induced reinstatement of drug-seeking behaviors for cocaine [299] and methamphetamine [300]. We plan on testing our findings *in vivo* by measuring the effects of treatment with TAAR1 agonists on PavCA behaviors in rats, with treatment both prior to and after PavCA training.

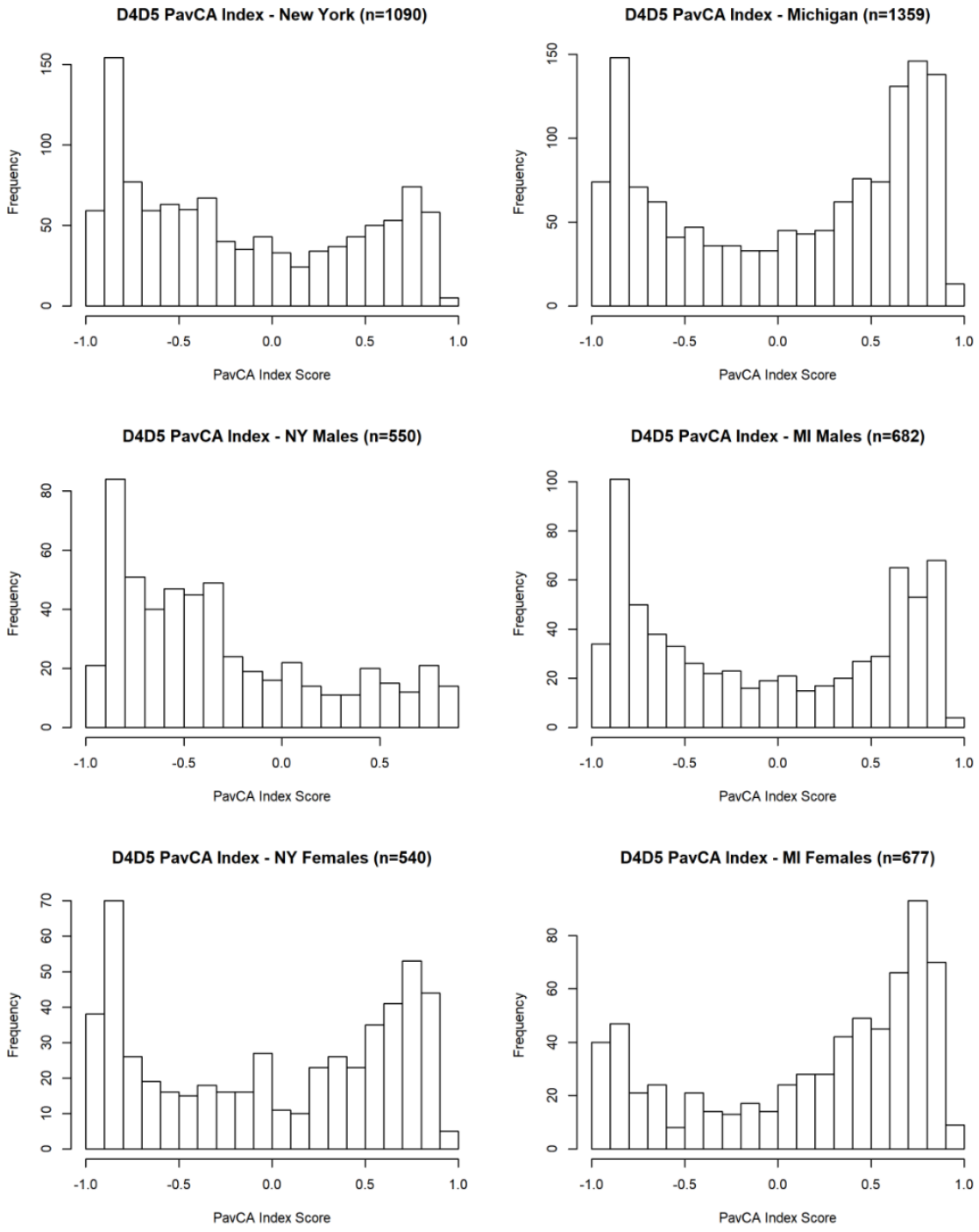
Unexpectedly, only one region clearly replicated between the SD and HS analyses prior to meta-analysis. There are several possible contributing factors to this result. First, the heritabilities of the PavCA metrics are low, requiring exceptionally large sample sizes to discover genes influencing the numerous component behaviors. Differences in genetic background may impact how certain variants effect of the phenotype of interest in ways that are beyond detection under the assumption of an additive genetic model. Allelic heterogeneity is also a possibility with alleles at multiple different loci being capable of producing the same behavioral outcome. Though we had an exceptionally large sample, the differences in the allele frequency spectrums between Harlan SD, Charles River SD, and the HS further hindered our power for discovery. Differences in animal

care between SD vendors and the HS facility may have had a substantial effect on the variation observed in the PavCA behaviors. This issue was this further compounded with difference in age and training center, with SD and MI rats being younger, naïve, and trained at the MI center, and NY rats older, experienced and trained at the NY-center. Since we observed a significant sex difference between males and females in sign- and goal-tracking behaviors, and SD rats were all males, it is further possible that sex played a role in the disparate results of the population-specific analyses.

Our results support the theory that the attribution of incentive salience is a highly polygenic trait. Various associated loci were identified in the three populations of rats, however there was little overlap between them, suggesting there may be a large role of genetic heterogeneity. Recent work has also shown that sign-tracking and goal-tracking behaviors are likely to be more broadly due to bottom-up vs top-down processing, respectively [260]. PavCA behaviors are now recognized to not purely be mediated by ventral striatum phasic dopamine responses controlling the attribution of incentive salience to stimuli, but also other systems such as acetylcholine signaling in the prefrontal cortex controlling attention [259]. This increasingly complex understanding of the biological basis of sign- and goal-tracking behaviors further supports our polygenicity hypothesis. Regardless, our GWAS in the HS and meta-analysis of the SD and HS have provided us with strong candidates for further testing. Additionally, the majority of identified top SNPs occur in intergenic regions, indicating that regulatory elements could be important in modulating these behaviors.

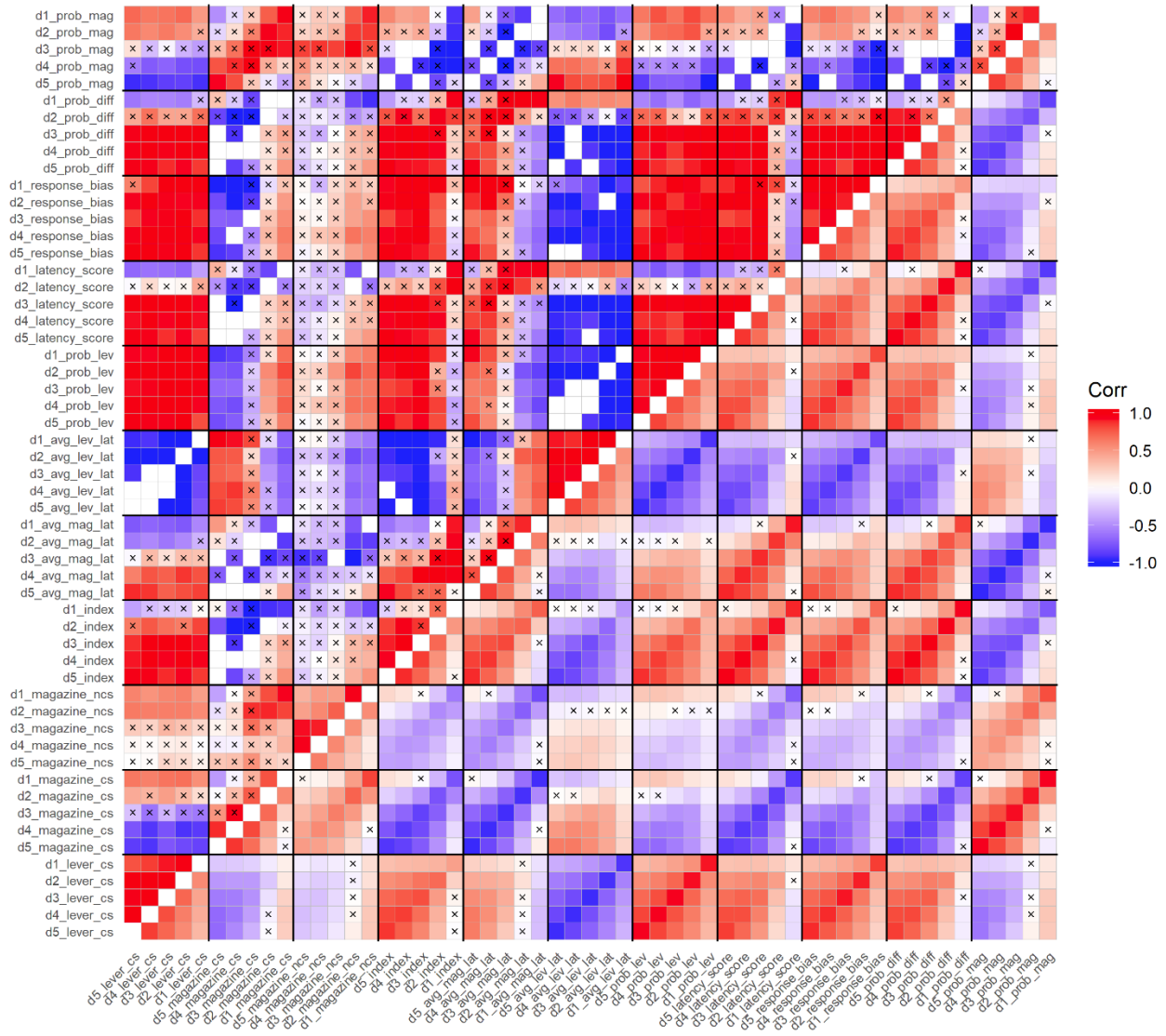
4.6 Appendix E: Supplemental Figures

Supplemental Figure 4.1. Distributions of PavCA index scores across sex and testing centers.



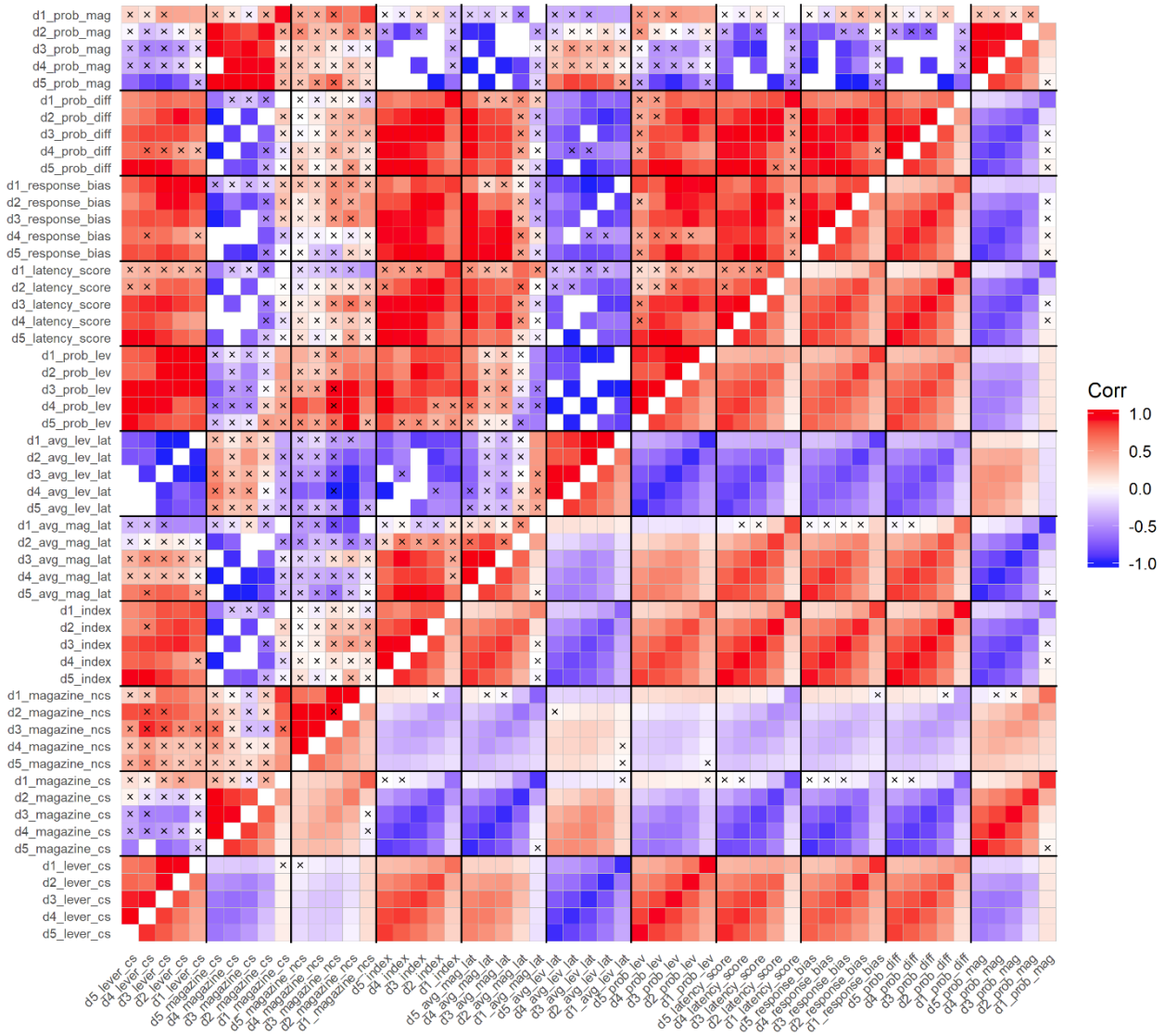
Supplemental Figure 4.2. Pairwise genetic and phenotypic correlations for PavCA metric within HS rats tested in NY.

The upper-left triangle displays the genetic correlations, and the lower-right triangle displays the corresponding phenotypic correlations. Non-significant correlations are indicated by an 'x' in the corresponding tile. White squares indicate genetic correlations that could not be computed by GCTA due to an error.



Supplemental Figure 4.3. Pairwise genetic and phenotypic correlations for PavCA metric within HS rats tested in MI.

The upper-left triangle displays the genetic correlations, and the lower-right triangle displays the corresponding phenotypic correlations. Non-significant correlations are indicated by an 'x' in the corresponding tile. White squares indicate genetic correlations that could not be computed by GCTA due to an error.



4.7 Appendix F: Supplementary Tables

Supplemental Table 4.1. Heritability estimates for 56 PavCA metrics in HS samples tested at the NY center.

Phenotype	V(G)/Vp	Std. Error	P-value	Vg - Percent
Average Latency to Lever Press - Day 1	0.077806	0.034716	2.76E-03	7.8%
Average Latency to Lever Press - Day 2	0.098205	0.035767	2.72E-04	9.8%
Average Latency to Lever Press - Day 3	0.166002	0.041028	3.60E-08	16.6%
Average Latency to Lever Press - Day 4	0.16856	0.042023	1.48E-07	16.9%
Average Latency to Lever Press - Day 5	0.215242	0.044905	1.38E-10	21.5%
Average Latency to Magazine Entry - Day 1	0.233313	0.043988	1.96E-13	23.3%
Average Latency to Magazine Entry - Day 2	0.104274	0.036149	6.87E-05	10.4%
Average Latency to Magazine Entry - Day 3	0.014233	0.027862	2.93E-01	1.4%
Average Latency to Magazine Entry - Day 4	0.043666	0.030582	4.47E-02	4.4%
Average Latency to Magazine Entry - Day 5	0.087404	0.03698	2.14E-03	8.7%
PavCA Index Score - Day 1	0.100365	0.037443	3.15E-04	10.0%
PavCA Index Score - Day 2	0.039089	0.030674	6.99E-02	3.9%
PavCA Index Score - Day 3	0.061694	0.033836	1.65E-02	6.2%
PavCA Index Score - Day 4	0.102177	0.0366	2.19E-04	10.2%
PavCA Index Score - Day 5	0.161078	0.041803	2.88E-07	16.1%
PavCA Index Score - Days 4/5	0.140604	0.040266	3.19E-06	14.1%
Latency Score - Day 1	0.151211	0.040037	2.40E-07	15.1%
Latency Score - Day 2	0.025906	0.028575	1.53E-01	2.6%
Latency Score - Day 3	0.033419	0.03114	1.24E-01	3.3%
Latency Score - Day 4	0.086538	0.035292	1.23E-03	8.7%
Latency Score - Day 5	0.150781	0.041337	1.56E-06	15.1%
Lever Presses CS - Day 1	0.107072	0.036252	4.53E-05	10.7%
Lever Presses CS - Day 2	0.127868	0.038629	1.20E-05	12.8%
Lever Presses CS - Day 3	0.169573	0.042388	1.71E-07	17.0%
Lever Presses CS - Day 4	0.173251	0.042105	6.26E-08	17.3%
Lever Presses CS - Day 5	0.186454	0.04317	8.38E-09	18.6%
Magazine Entries CS - Day 1	0.230224	0.04355	1.60E-13	23.0%
Magazine Entries CS - Day 2	0.088636	0.034998	4.95E-04	8.9%
Magazine Entries CS - Day 3	0.016263	0.028418	2.72E-01	1.6%
Magazine Entries CS - Day 4	0.049213	0.031149	2.86E-02	4.9%
Magazine Entries CS - Day 5	0.111584	0.038314	1.18E-04	11.2%
Magazine Entries NCS - Day 1	0.242195	0.043182	8.33E-16	24.2%
Magazine Entries NCS - Day 2	0.150457	0.041318	1.97E-06	15.0%
Magazine Entries NCS - Day 3	0.178112	0.044039	6.36E-07	17.8%
Magazine Entries NCS - Day 4	0.113557	0.038506	1.14E-04	11.4%
Magazine Entries NCS - Day 5	0.161328	0.042384	7.62E-07	16.1%
Probability Difference - Day 1	0.111967	0.036627	2.51E-05	11.2%
Probability Difference - Day 2	0.021933	0.028319	1.95E-01	2.2%
Probability Difference - Day 3	0.035128	0.031317	1.12E-01	3.5%
Probability Difference - Day 4	0.088375	0.035486	1.03E-03	8.8%
Probability Difference - Day 5	0.169356	0.042543	1.20E-07	16.9%
Probability of Lever Press - Day 1	0.106015	0.036768	9.43E-05	10.6%
Probability of Lever Press - Day 2	0.100299	0.036138	2.41E-04	10.0%
Probability of Lever Press - Day 3	0.155878	0.041228	4.86E-07	15.6%
Probability of Lever Press - Day 4	0.139541	0.039775	4.71E-06	14.0%
Probability of Lever Press - Day 5	0.187685	0.042559	2.37E-09	18.8%
Probability of Magazine Entry - Day 1	0.234746	0.04361	4.16E-14	23.5%
Probability of Magazine Entry - Day 2	0.096153	0.035935	2.74E-04	9.6%
Probability of Magazine Entry - Day 3	0.017861	0.028316	2.48E-01	1.8%
Probability of Magazine Entry - Day 4	0.040229	0.03026	5.81E-02	4.0%
Probability of Magazine Entry - Day 5	0.086075	0.037214	2.86E-03	8.6%
Response Bias - Day 1	0.037296	0.030621	7.82E-02	3.7%
Response Bias - Day 2	0.107355	0.037769	2.31E-04	10.7%
Response Bias - Day 3	0.151614	0.041242	1.34E-06	15.2%
Response Bias - Day 4	0.14775	0.039645	7.13E-07	14.8%
Response Bias - Day 5	0.159154	0.041449	2.21E-07	15.9%

Supplemental Table 4.2. Heritability estimates for 56 PavCA metrics in HS samples tested at the MI center.

Phenotype	V(G)/Vp	Std. Error	P-value	Vg - Percent
Average Latency to Lever Press - Day 1	0.221299	0.038891	3.64E-14	22.1%
Average Latency to Lever Press - Day 2	0.115435	0.032875	3.09E-06	11.5%
Average Latency to Lever Press - Day 3	0.068124	0.030856	6.41E-03	6.8%
Average Latency to Lever Press - Day 4	0.066751	0.030723	7.84E-03	6.7%
Average Latency to Lever Press - Day 5	0.102911	0.03303	9.73E-05	10.3%
Average Latency to Magazine Entry - Day 1	0.092932	0.032526	1.99E-04	9.3%
Average Latency to Magazine Entry - Day 2	0.048872	0.027603	1.83E-02	4.9%
Average Latency to Magazine Entry - Day 3	0.082959	0.031061	5.49E-04	8.3%
Average Latency to Magazine Entry - Day 4	0.104591	0.032279	1.52E-05	10.5%
Average Latency to Magazine Entry - Day 5	0.058213	0.029246	1.04E-02	5.8%
PavCA Index Score - Day 1	0.096554	0.031579	3.80E-05	9.7%
PavCA Index Score - Day 2	0.072055	0.028921	8.11E-04	7.2%
PavCA Index Score - Day 3	0.064584	0.029272	4.30E-03	6.5%
PavCA Index Score - Day 4	0.078733	0.031148	1.47E-03	7.9%
PavCA Index Score - Day 5	0.070977	0.030664	3.81E-03	7.1%
PavCA Index Score - Days 4/5	0.087334	0.032053	6.22E-04	8.7%
Latency Score - Day 1	0.076953	0.030162	6.72E-04	7.7%
Latency Score - Day 2	0.064265	0.028302	2.25E-03	6.4%
Latency Score - Day 3	0.056361	0.028862	1.24E-02	5.6%
Latency Score - Day 4	0.079136	0.031619	2.06E-03	7.9%
Latency Score - Day 5	0.076542	0.031225	2.41E-03	7.7%
Lever Presses CS - Day 1	0.190203	0.037412	5.21E-12	19.0%
Lever Presses CS - Day 2	0.109707	0.032183	5.30E-06	11.0%
Lever Presses CS - Day 3	0.069467	0.030888	5.84E-03	6.9%
Lever Presses CS - Day 4	0.062369	0.030647	1.38E-02	6.2%
Lever Presses CS - Day 5	0.108125	0.033964	8.90E-05	10.8%
Magazine Entries CS - Day 1	0.0906	0.032126	2.32E-04	9.1%
Magazine Entries CS - Day 2	0.048373	0.027195	1.70E-02	4.8%
Magazine Entries CS - Day 3	0.104399	0.031953	1.17E-05	10.4%
Magazine Entries CS - Day 4	0.132513	0.034122	1.90E-07	13.3%
Magazine Entries CS - Day 5	0.078071	0.031081	1.35E-03	7.8%
Magazine Entries NCS - Day 1	0.186044	0.040152	1.73E-08	18.6%
Magazine Entries NCS - Day 2	0.046473	0.029005	4.15E-02	4.6%
Magazine Entries NCS - Day 3	0.020453	0.025979	2.10E-01	2.0%
Magazine Entries NCS - Day 4	0.115423	0.035188	5.98E-05	11.5%
Magazine Entries NCS - Day 5	0.092997	0.033226	5.96E-04	9.3%
Probability Difference - Day 1	0.09352	0.031057	4.18E-05	9.4%
Probability Difference - Day 2	0.071179	0.028796	8.58E-04	7.1%
Probability Difference - Day 3	0.070176	0.029783	2.19E-03	7.0%
Probability Difference - Day 4	0.072352	0.030571	3.00E-03	7.2%
Probability Difference - Day 5	0.058548	0.02926	1.10E-02	5.9%
Probability of Lever Press - Day 1	0.21225	0.039051	5.18E-13	21.2%
Probability of Lever Press - Day 2	0.111675	0.032265	3.59E-06	11.2%
Probability of Lever Press - Day 3	0.053726	0.02896	1.81E-02	5.4%
Probability of Lever Press - Day 4	0.054958	0.029396	2.00E-02	5.5%
Probability of Lever Press - Day 5	0.070003	0.031	5.55E-03	7.0%
Probability of Magazine Entry - Day 1	0.109107	0.033284	1.41E-05	10.9%
Probability of Magazine Entry - Day 2	0.045587	0.027486	2.79E-02	4.6%
Probability of Magazine Entry - Day 3	0.099816	0.031842	2.89E-05	10.0%
Probability of Magazine Entry - Day 4	0.11271	0.033168	7.17E-06	11.3%
Probability of Magazine Entry - Day 5	0.05844	0.03009	1.56E-02	5.8%
Response Bias - Day 1	0.125699	0.034089	1.17E-06	12.6%
Response Bias - Day 2	0.098111	0.031703	3.97E-05	9.8%
Response Bias - Day 3	0.088314	0.031185	1.87E-04	8.8%
Response Bias - Day 4	0.072791	0.029787	1.35E-03	7.3%
Response Bias - Day 5	0.065816	0.030352	5.64E-03	6.6%

Supplemental Table 4.3. Heritability estimates for 56 PavCA metrics in all HS rats combined.

Phenotype	V(G)/Vp	Std. Error	P-value	Vg - Percent
Average Latency to Lever Press - Day 1	0.118976	0.022627	6.16E-15	11.9%
Average Latency to Lever Press - Day 2	0.092842	0.020898	2.12E-10	9.3%
Average Latency to Lever Press - Day 3	0.107673	0.022456	1.14E-11	10.8%
Average Latency to Lever Press - Day 4	0.123661	0.023743	8.35E-13	12.4%
Average Latency to Lever Press - Day 5	0.136695	0.024075	2.17E-15	13.7%
Average Latency to Magazine Entry - Day 1	0.124305	0.023323	8.60E-15	12.4%
Average Latency to Magazine Entry - Day 2	0.056408	0.01778	6.14E-06	5.6%
Average Latency to Magazine Entry - Day 3	0.053626	0.018029	5.65E-05	5.4%
Average Latency to Magazine Entry - Day 4	0.072786	0.019418	6.83E-08	7.3%
Average Latency to Magazine Entry - Day 5	0.080569	0.020691	9.19E-08	8.1%
PavCA Index Score - Day 1	0.076238	0.019424	5.88E-09	7.6%
PavCA Index Score - Day 2	0.061129	0.018231	1.70E-06	6.1%
PavCA Index Score - Day 3	0.074995	0.019852	1.47E-07	7.5%
PavCA Index Score - Day 4	0.09281	0.02136	1.18E-09	9.3%
PavCA Index Score - Day 5	0.11095	0.02271	1.62E-11	11.1%
PavCA Index Score - Days 4/5	0.116284	0.023184	2.15E-12	11.6%
Latency Score - Day 1	0.078297	0.019486	3.99E-09	7.8%
Latency Score - Day 2	0.046934	0.016559	5.89E-05	4.7%
Latency Score - Day 3	0.058833	0.018529	2.15E-05	5.9%
Latency Score - Day 4	0.08657	0.020944	1.39E-08	8.7%
Latency Score - Day 5	0.112164	0.022912	2.98E-11	11.2%
Lever Presses CS - Day 1	0.13056	0.023227	0.00E+00	13.1%
Lever Presses CS - Day 2	0.089402	0.020651	7.02E-10	8.9%
Lever Presses CS - Day 3	0.106975	0.022563	5.01E-11	10.7%
Lever Presses CS - Day 4	0.123737	0.023797	1.38E-12	12.4%
Lever Presses CS - Day 5	0.125009	0.023448	9.76E-14	12.5%
Magazine Entries CS - Day 1	0.109556	0.02197	2.17E-13	11.0%
Magazine Entries CS - Day 2	0.053327	0.017281	1.02E-05	5.3%
Magazine Entries CS - Day 3	0.064389	0.018578	1.06E-06	6.4%
Magazine Entries CS - Day 4	0.08142	0.019732	1.23E-09	8.1%
Magazine Entries CS - Day 5	0.090463	0.021028	9.98E-10	9.0%
Magazine Entries NCS - Day 1	0.163191	0.024996	0.00E+00	16.3%
Magazine Entries NCS - Day 2	0.087308	0.021033	1.36E-08	8.7%
Magazine Entries NCS - Day 3	0.078849	0.020826	6.56E-07	7.9%
Magazine Entries NCS - Day 4	0.097327	0.021836	1.06E-09	9.7%
Magazine Entries NCS - Day 5	0.108491	0.022474	1.77E-11	10.8%
Probability Difference - Day 1	0.069713	0.018406	1.60E-08	7.0%
Probability Difference - Day 2	0.051772	0.017122	1.88E-05	5.2%
Probability Difference - Day 3	0.06876	0.019491	1.47E-06	6.9%
Probability Difference - Day 4	0.088105	0.020888	4.20E-09	8.8%
Probability Difference - Day 5	0.100306	0.021803	1.42E-10	10.0%
Probability of Lever Press - Day 1	0.13102	0.02341	5.55E-17	13.1%
Probability of Lever Press - Day 2	0.094047	0.020889	7.00E-11	9.4%
Probability of Lever Press - Day 3	0.088892	0.020942	2.61E-09	8.9%
Probability of Lever Press - Day 4	0.100127	0.021915	4.86E-10	10.0%
Probability of Lever Press - Day 5	0.114833	0.022622	8.29E-13	11.5%
Probability of Magazine Entry - Day 1	0.127706	0.02308	1.11E-16	12.8%
Probability of Magazine Entry - Day 2	0.047549	0.017168	1.41E-04	4.8%
Probability of Magazine Entry - Day 3	0.058767	0.018368	1.01E-05	5.9%
Probability of Magazine Entry - Day 4	0.074145	0.019545	6.83E-08	7.4%
Probability of Magazine Entry - Day 5	0.072665	0.020208	1.70E-06	7.3%
Response Bias - Day 1	0.077714	0.019858	1.57E-08	7.8%
Response Bias - Day 2	0.083727	0.020231	2.41E-09	8.4%
Response Bias - Day 3	0.106436	0.022486	1.64E-11	10.6%
Response Bias - Day 4	0.111011	0.022383	4.27E-13	11.1%
Response Bias - Day 5	0.087392	0.02092	3.85E-09	8.7%

Supplemental Table 4.4. Between-center genetic correlations for 56 PavCA metrics.

PavCA Metric	Genetic Correlation	S.E.	P-value
Day 1 Average Latency to Lever Press	0.65817	0.233088	2.21E-03
Day 1 Average Latency to Magazine Entry	0.572203	0.188226	2.05E-03
Day 1 PavCA Index Score	0.667065	0.27935	7.60E-03
Day 1 Latency Score	0.486711	0.248907	2.51E-02
Day 1 Lever Presses w/ CS	0.752042	0.182141	5.68E-05
Day 1 Magazine Entries w/ CS	0.438089	0.200683	1.69E-02
Day 1 Magazine Entries w/o CS	0.582208	0.145116	1.11E-04
Day 1 Probability Difference	0.438369	0.255626	4.01E-02
Day 1 Proportion of Trials with Lever Press	0.654373	0.18481	3.70E-04
Day 1 Proportion of Trials with Magazine Entry	0.553284	0.175815	1.38E-03
Day 1 Response Bias	1	0.492624	1.29E-03
Day 2 Average Latency to Lever Press	0.687607	0.225711	1.47E-03
Day 2 Average Latency to Magazine Entry	0.627688	0.344221	2.51E-02
Day 2 PavCA Index Score	1	0.54397	1.03E-03
Day 2 Latency Score	1	0.799982	6.10E-03
Day 2 Lever Presses w/ CS	0.485124	0.210766	1.16E-02
Day 2 Magazine Entries w/ CS	0.711216	0.380207	2.09E-02
Day 2 Magazine Entries w/o CS	1	0.366329	5.29E-04
Day 2 Probability Difference	1	0.786177	3.95E-03
Day 2 Proportion of Trials with Lever Press	0.745609	0.226044	6.12E-04
Day 2 Proportion of Trials with Magazine Entry	0.410715	0.347358	1.09E-01
Day 2 Response Bias	0.700383	0.254438	3.15E-03
Day 3 Average Latency to Lever Press	0.966613	0.253913	8.01E-05
Day 3 Average Latency to Magazine Entry	1	0.854513	9.06E-03
Day 3 PavCA Index Score	1	0.482886	1.12E-04
Day 3 Latency Score	1	0.793204	1.71E-03
Day 3 Lever Presses w/ CS	0.893453	0.24235	1.97E-04
Day 3 Magazine Entries w/ CS	1	0.839521	4.14E-03
Day 3 Magazine Entries w/o CS	1	0.533643	1.18E-03
Day 3 Probability Difference	1	0.698831	5.13E-04
Day 3 Proportion of Trials with Lever Press	1	0.335859	5.89E-04
Day 3 Proportion of Trials with Magazine Entry	1	0.77428	1.07E-02
Day 3 Response Bias	0.781676	0.208944	2.85E-04
Day 4 Average Latency to Lever Press	1	0.221374	1.48E-06
Day 4 Average Latency to Magazine Entry	1	0.455425	8.19E-04
Day 4 PavCA Index Score	1	0.270638	7.13E-05
Day 4 Latency Score	1	0.303483	1.21E-04
Day 4 Lever Presses w/ CS	1	0.217673	2.39E-06
Day 4 Magazine Entries w/ CS	1	0.421777	6.25E-04
Day 4 Magazine Entries w/o CS	0.717413	0.233466	2.27E-03
Day 4 Probability Difference	1	0.332567	2.82E-05
Day 4 Proportion of Trials with Lever Press	1	0.283203	3.05E-05
Day 4 Proportion of Trials with Magazine Entry	1	0.446379	1.00E-03
Day 4 Response Bias	1	0.236159	2.02E-06
Day 5 Average Latency to Lever Press	0.717116	0.171547	8.89E-05
Day 5 Average Latency to Magazine Entry	1	0.353119	1.08E-04
Day 5 PavCA Index Score	0.875965	0.212401	5.73E-05
Day 5 Latency Score	0.874244	0.206515	5.78E-05
Day 5 Lever Presses w/ CS	0.706469	0.186946	2.99E-04
Day 5 Magazine Entries w/ CS	0.989111	0.265934	1.13E-04
Day 5 Magazine Entries w/o CS	0.830963	0.218543	1.71E-04
Day 5 Probability Difference	0.866908	0.243815	2.04E-04
Day 5 Proportion of Trials with Lever Press	0.901498	0.23206	5.60E-05
Day 5 Proportion of Trials with Magazine Entry	1	0.345425	1.04E-03
Day 5 Response Bias	0.686792	0.261377	4.68E-03
Day 4/5 PavCA Index Score	0.93732	0.2014	1.24E-05

CHAPTER 5

CONCLUSIONS

5.1 Summary and significance

Addiction is a complex neuropsychiatric disorder with numerous genetic and environmental factors that contribute to susceptibility. Little is understood about the progression and pathophysiology of addiction; however, we know that life experiences heavily influence an individual's likelihood of developing an addiction. Since it is near impossible to study humans in a controlled manner for extended periods, as well as ethically unacceptable to induce human addiction, researchers need alternate methods of dissecting the disease's etiology. Animal models have long been used to overcome these obstacles. While they will never entirely replicate human circumstances leading to the use and abuse of substances, they allow for isolating certain facets of addiction in a pre-determined environmental context. Animal research also lends itself to the study of behavioral endophenotypes; heritable, often unobserved traits believed to more stably reflect the function of a discrete biological system. Due to the complexity of addiction and plethora of behaviors involved, the discovery of effective endophenotypes is vital.

Chapters 3 and 4 detail a pioneering attempt to identify the genetic underpinnings of an endophenotype long studied in the field of addiction. I aimed to map genes associated with the propensity to attribute incentive salience to reward-associated cues, as studied in rats. For decades, it has been hypothesized that this hypersensitivity to reward cues is critical in the development and maintenance of addiction [51,301]. Environmental cues attributed with incentive salience are capable of eliciting approach, triggering craving, and instigating reward-seeking behaviors that can lead to eventual relapse[218,265,267,271]. Our collaborators and

others developed and refined the Pavlovian conditioned approach behavioral paradigm to measure the degree to which rodents place motivational value on cues associated with a rewarding substance. It was discovered that some animals preferentially approached reward-associated cues after Pavlovian conditioning and that this predisposition could be subjected to selection [75]. Though this trait is very well-studied, no previous attempt has been made to identify genes influencing it in an unbiased manner.

In order to perform the work described in Chapters 3 and 4, I required a method of obtaining dense genotype data on thousands of rat samples. Affordable microarray options are not available in rats, and neither the methodology nor the price-point for light whole-genome sequencing met our needs at the advent of these projects. Therefore, in Chapter 2, I describe the optimization and a protocol and variant calling pipeline for an alternative option known as genotype-by-sequencing. Though this reduced-representation approach had been previously used in rodent species [76,86,96,115], it had never been refined to ensure the highest quality results. Further, it was unclear how best to analyze rodent GBS data, since all existing computational pipelines were designed for plants. I meticulously deconstructed the original protocol, reviewed the available literature for each step, and tested modifications that improved library and sequence data quality. Then, in collaboration with other members of the lab, I worked on designing a workflow using existing tools for turning raw ddGBS sequencing data into genotypes with high concordance with calls from array data. We demonstrate that with the correct set of tools and parameters, ddGBS data could be used to call over 100,000 SNPs, and then imputation to available reference panels could elevate this number to over 3 million. This work has great utility to the rodent genetics community and acts as a resource for others to develop their own versions of GBS for their model system of interest.

The Sprague Dawley project discussed in Chapter 3 was opportunistic in nature. We took advantage of a tremendous sample of rats already phenotyped by our collaborators for various other projects and used their DNA to perform the first ever GWAS in commercially-available outbred rats. My original goal for this project was to determine the heritability of various metrics from Pavlovian conditioned approach and use these trait components to uncover genes influencing the attribution of incentive salience to reward cues. A major complicating factor was that after beginning to genotype SD rats, I quickly realized that the two major commercial vendor showed highly disparate allele frequency spectrums. Using principal component analysis and F_{ST} estimation, I showed that despite both being called Sprague Dawley, rats from Harlan and Charles River were highly genetically differentiated and even had modest within-vendor population substructure. This incidental finding has a substantial impact on the rat community, as the SD are one of the most frequently used strains for psychiatric and physiological research, and results of multiple pre-existing studies may actually be largely driven by differences in the genetic background between vendors. I additionally discovered that the PavCA paradigm had low levels of heritability. The differentiation between the two vendors and low heritability both negatively influenced the power of the GWAS. I had to split the sample and perform the analyses on each population separately and meta-analyze the results, similar to how human GWAS are performed. Ultimately, I discovered 21 genome-wide significant loci associated with various PavCA and summary metrics over the five days of training. Of these loci, a handful showed clear ties to neurobiological function.

Due to the nature of GWAS, results must be replicated to show that the associations weren't spurious, nor a product of winner's curse. To accomplish this, in Chapter 4 I performed

another GWAS for PavCA in an independent sample of HS rats. Though the sample was homogenous in origin, rats were split between two centers, where they were tested as different ages and levels of experience with behavioral testing. The confounding of environment, age, and previous testing necessitated analyzing these sample sets both separately and jointly. I discovered 22 independent genetic loci across these 3 analyses. Interestingly, there was little overlap between the findings in the center-specific and mega-analyses, suggesting different aspects of the genetic underpinnings of PavCA were being uncovered. However, one locus stood out on chromosome 18 that was associated with numerous PavCA metrics across later days of training. This locus contained the gene *Rttn*, believed to play an important role in brain development and the etiology of several psychiatric disorders [278–280]. I also meta-analyzed the results of the HS mega-analysis and Harlan and Charles River GWAS, leading to the identification of 13 loci, 9 of which were previously undiscovered in the population-specific analyses. Notably, the gene *Taar1* was identified as a positional candidate in the HS/SD meta-analysis and is a well-known mediator of dopamine neurotransmission with strong ties to addiction [293,294].

The LD landscape of the HS and SD acted as both an advantage and disadvantage to this study. Although numerous loci were discovered in the various SD and HS GWAS I ran, several have large QTL intervals, making it difficult to pinpoint the "causal" locus. While credible set analysis provides one method of narrowing down the potential SNPs to a more reasonable subset, there is still great ambiguity as to where the signal is originating. There was also a very low rate of replication of genome-wide significant hits between the studies. I hypothesize that there are a few reasons for this. The first is the low trait heritability and non-normality of the PavCA metrics. The distributions were highly skewed with hundreds of animals grouped at the

minimum or maximum values. It was not clear how to properly break ties at extreme values, leading to random sorting of ~30% of the samples with respect to each other during the quantile normalization procedure. This certainly negatively impacted our power within each study, making it less likely to replicate findings from each. Secondly, PavCA is a highly polygenic behavior. The Q-Q plots show low levels of inflation despite adequate control for population structure with LOCO-GRMs, which can be caused by polygenic signal occurring in regions of extended LD. Moreover, with a complex behavior like PavCA, different variants are likely influencing it in each population. Additionally, in the case of the Sprague Dawley rats (especially Harlan), there were a large number of rare alleles, negatively impacting our power. Allele frequencies also varied greatly across the three populations, influencing which loci would have been discoverable in each with the available sample sizes. Lastly, because of the differences in the age of the rats at testing and the environment and experiences, the phenotypic data from NY, MI, and the SD may not have been perfectly comparable. All in all, there were many impediments to the success of these studies, most of which could not be foreseen at their advent.

5.2 Future directions

The results of Chapters 3 and 4 have provided us with multiple positional candidate genes that we can investigate in follow-up studies. Specifically, there is already a wealth of research on TAAR1 in relation to addiction and its influence on the dopamine-mediated reward system, and TAAR1 agonists and partial agonists are available [302], as are TAAR1 knockout rat lines [303]. In conjunction with our collaborators at the Universities of Buffalo and Michigan, we plan on treating HS animals with the a TAAR1 agonist either before or after PavCA training to investigate its effects on initial learning of a Pavlovian conditioned response or later expression of the response after the CS-US association has been made, respectively. We may also obtain a

knockout of TAAR1 and observe its impact on the acquisition of sign-tracking behaviors in the PavCA paradigm. Other identified targets from our studied, such as *Rttn*, have more severe neurological consequences when knocked out [279], necessitating alternate, more elegant approaches. We plan on looking for known eQTL in the regions of these genes to identify possible regulatory elements they can be modified to alter expression in a more attenuated manner.

In addition to functional testing of our current set of candidates, In the near future, I plan on repeating the meta-analyses presented in Chapters 3 and 4 using methods that were designed for trans-ethnic human studies. These methods take into account the differences in the LD landscape between populations and heterogeneous effect sizes, including packages such as MANTRA and RE2C [304,305]. Given the high levels of differentiation between the two SD population and the SD and HS, I believe these programs may be more appropriate than a simple sample-weighting approach. I will also be reevaluating the significance threshold used for the meta-analyses, attempting a traditional permutation based approach as I am concerned that the current threshold may be overly stringent, reducing the discovery rate. Lastly, I would like to construct a phylogenetic tree using common inbred and outbred rat strains in conjunction with SD rats from Harlan and Charles River to determine how closely related they are to modern day strains.

5.3 Concluding remarks

The studies detailed in Chapter 2-4 represent major contributions to the rat genetics and addiction community. In Chapter 2, I developed a ddGBS protocol tailored to rats and showed that sequencing data obtained by this method could be used to call 200,000+ SNPs in SD rats and over 3.7 million in the HS, where reference panels were available. In Chapter 3, I provide

the first genetic characterization of SD rats and have shown that the vendor populations are highly diverged. I went on to make the first heritability estimates for PavCA, a widely studied addiction-related trait. I then performed the first GWAS in commercially-available outbred rats, and the largest rodent GWAS to date. In Chapter 4, I performed a replication of the PavCA GWAS and the first ever meta-analysis of different rat strains. Between the SD and HS GWAS and the meta-analyses, I have identified a handful of exciting candidate loci that warrant further investigation. This work has proven that GBS data is able to provide more than sufficient genome-wide coverage for genetic mapping and that commercially-available outbred rat populations can be successfully use for GWAS.

BIBLIOGRAPHY

1. Steel Z, Marnane C, Iranpour C, Chey T, Jackson JW, Patel V, et al. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. International Journal of Epidemiology. 2014;43: 476–493. doi:10.1093/ije/dyu038

2. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. The Lancet Psychiatry. 2016;3: 171–178. doi:10.1016/S2215-0366(15)00505-2

3. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. The Lancet. 2013;382: 1575–1586. doi:10.1016/S0140-6736(13)61611-6

4. McCammon JM, Sive H. Challenges in understanding psychiatric disorders and developing therapeutics: a role for zebrafish. Disease Models & Mechanisms. 2015;8: 647–656. doi:10.1242/dmm.019620

5. Hyman SE. Revitalizing Psychiatric Therapeutics. Neuropsychopharmacology. 2014;39: 220–229. doi:10.1038/npp.2013.181

6. Trautmann S, Rehm J, Wittchen H. The economic costs of mental disorders: Do our societies react appropriately to the burden of mental disorders? EMBO reports. 2016;17: 1245–1249. doi:10.15252/embr.201642951

7. Geschwind DH, Flint J. Genetics and genomics of psychiatric disease. Science. 2015;349: 1489–1494. doi:10.1126/science.aaa8954

8. Ducci F, Goldman D. The Genetic Basis of Addictive Disorders. Psychiatric Clinics of North America. 2012;35: 495–519. doi:10.1016/j.psc.2012.03.010

9. Goldman D, Oroszi G, Ducci F. The genetics of addictions: uncovering the genes. Nature Reviews Genetics. 2005;6: 521.

10. Kendler KS. What psychiatric genetics has taught us about the nature of psychiatric illness and what is left to learn. Molecular Psychiatry. 2013;18: 1058.

11. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics. 2011;43: 519–525. doi:10.1038/ng.823

12. Klein RJ. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science. 2005;308: 385–389. doi:10.1126/science.1109557

13. DeWan A, Liu M, Hartman S, Zhang SS-M, Liu DTL, Zhao C, et al. HTRA1 Promoter Polymorphism in Wet Age-Related Macular Degeneration. Science. 2006;314: 989–992. doi:10.1126/science.1133807

14. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447: 661–678. doi:10.1038/nature05911

15. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. Journal of Clinical Investigation. 2008;118: 1590–1605. doi:10.1172/JCI34772

16. MAGIC, on behalf of Procardis Consortium, Speliotes EK, Willer CJ, Berndt SI, Monda KL, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature Genetics. 2010;42: 937–948. doi:10.1038/ng.686

17. Gelernter J, Kranzler HR, Sherva R, Koesterer R, Almasy L, Zhao H, et al. Genome-wide association study of opioid dependence: multiple associations mapped to calcium and potassium pathways. Biol Psychiatry. 2014;76: 66–74. doi:10.1016/j.biopsych.2013.08.034

18. Gelernter J, Sherva R, Koesterer R, Almasy L, Zhao H, Kranzler HR, et al. Genome-wide association study of cocaine dependence and related traits: FAM53B identified as a risk gene. Mol Psychiatry. 2014;19: 717–723. doi:10.1038/mp.2013.99

19. The Electronic Medical Records and Genomics (eMERGE) Consortium, The MIGen Consortium, The PAGE Consortium, The LifeLines Cohort Study, Wood AR, Esko T, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics. 2014;46: 1173–1186. doi:10.1038/ng.3097

20. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511: 421–427. doi:10.1038/nature13595

21. Hou L, Bergen SE, Akula N, Song J, Hultman CM, Landén M, et al. Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. Human Molecular Genetics. 2016;25: 3383–3394. doi:10.1093/hmg/ddw181

22. The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. Molecular Autism. 2017;8. doi:10.1186/s13229-017-0137-9

23. Clarke T-K, Adams MJ, Davies G, Howard DM, Hall LS, Padmanabhan S, et al. Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N=112 117). Molecular Psychiatry. 2017;22: 1376.

24. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461: 747–753. doi:10.1038/nature08494

25. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics. 2010;11: 446–450. doi:10.1038/nrg2809

26. Gibson G. Hints of hidden heritability in GWAS. Nature Genetics. 2010;42: 558–560. doi:10.1038/ng0710-558

27. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences. 2012;109: 1193–1198. doi:10.1073/pnas.1119675109

28. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. The American Journal of Human Genetics. 2012;90: 7–24. doi:10.1016/j.ajhg.2011.11.029

29. Li Z, Chen J, Yu H, He L, Xu Y, Zhang D, et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. Nature Genetics. 2017;49: 1576–1583. doi:10.1038/ng.3973

30. eQTLGen, 23andMe, the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Wray NR, Ripke S, Mattheisen M, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nature Genetics. 2018;50: 668–681. doi:10.1038/s41588-018-0090-3

31. Iacono WG. Endophenotypes in psychiatric disease: prospects and challenges. Genome Medicine. 2018;10. doi:10.1186/s13073-018-0526-5

32. Gass P, Wotjak C. Rodent models of psychiatric disorders—practical considerations. Cell and Tissue Research. 2013;354: 1–7. doi:10.1007/s00441-013-1706-7

33. Dettmer AM, Suomi SJ. Nonhuman Primate Models of Neuropsychiatric Disorders: Influences of Early Rearing, Genetics, and Epigenetics. ILAR Journal. 2014;55: 361–370. doi:10.1093/ilar/ilu025

34. van Alphen B, van Swinderen B. Drosophila strategies to study psychiatric disorders. Brain Research Bulletin. 2013;92: 1–11. doi:10.1016/j.brainresbull.2011.09.007

35. Haesemeyer M, Schier AF. The study of psychiatric disease genes and drugs in zebrafish. Current Opinion in Neurobiology. 2015;30: 122–130. doi:10.1016/j.conb.2014.12.002

36. Parker CC, Chen H, Flagel SB, Geurts AM, Richards JB, Robinson TE, et al. Rats are the smart choice: Rationale for a renewed focus on rats in behavioral genetics. Neuropharmacology. 2014;76: 250–258. doi:10.1016/j.neuropharm.2013.05.047

37. Susaki EA, Ukai H, Ueda HR. Next-generation mammalian genetics toward organism-level systems biology. npj Systems Biology and Applications. 2017;3: 15. doi:10.1038/s41540-017-0015-2

38. Becanovic K, Jagodic M, Sheng JR, Dahlman I, Aboul-Enein F, Wallstrom E, et al. Advanced Intercross Line Mapping of Eae5 Reveals Ncf-1 and CLDN4 as Candidate Genes for Experimental Autoimmune Encephalomyelitis. The Journal of Immunology. 2006;176: 6055–6064. doi:10.4049/jimmunol.176.10.6055

39. Tsaih S-W, Holl K, Jia S, Kaldunski M, Tschannen M, He H, et al. Identification of a Novel Gene for Diabetic Traits in Rats, Mice, and Humans. Genetics. 2014;198: 17–29. doi:10.1534/genetics.114.162982

40. Rat Genome Sequencing and Mapping Consortium, Baud A, Hermsen R, Guryev V, Stridh P, Graham D, et al. Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nature Genetics. 2013;45: 767–775. doi:10.1038/ng.2644

41. Degenhardt L, Hall W. Extent of illicit drug use and dependence, and their contribution to the global burden of disease. Lancet. 2012;379: 55–70. doi:10.1016/S0140-6736(11)61138-0

42. Substance Abuse and Mental Health Services Administration. Results from the 2012 National Survey on Drug Use and Health: Summary of National Findings. Rockville, MD; 2013.

43. Glantz MD, Anthony JC, Berglund PA, Degenhardt L, Dierker L, Kalaydjian A, et al. Mental disorders as risk factors for later substance dependence: estimates of optimal prevention and treatment benefits. Psychological Medicine. 2009;39: 1365. doi:10.1017/S0033291708004510

44. National Institute of Drug Abuse. Principles of Drug Addiction Treatment: A Research-Based Guide (Third Edition). 2018; Available: https://www.drugabuse.gov/publications/principles-drug-addiction-treatment-research-based-guide-third-edition

45. Henden E, Melberg HO, Røgeberg OJ. Addiction: Choice or Compulsion? Frontiers in Psychiatry. 2013;4. doi:10.3389/fpsyt.2013.00077

46. Shaham Y, Shalev U, Lu L, de Wit H, Stewart J. The reinstatement model of drug relapse: history, methodology and major findings. Psychopharmacology (Berl). 2003;168: 3–20. doi:10.1007/s00213-002-1224-x

47. Paliwal P, Hyman SM, Sinha R. Craving predicts time to cocaine relapse: Further validation of the Now and Brief versions of the cocaine craving questionnaire. Drug and Alcohol Dependence. 2008;93: 252–259. doi:10.1016/j.drugalcdep.2007.10.002

48. Stretch R, Gerber GJ, Wood SM. Factors affecting behavior maintained by response-contingent intravenous infusions of amphetamine in squirrel monkeys. Can J Physiol Pharmacol. 1971;49: 581–589.

49. Stewart J, de Wit H. Reinstatement of Drug-Taking Behavior as a Method of Assessing Incentive Motivational Properties of Drugs. In: Bozarth MA, editor. Methods of Assessing the Reinforcing Properties of Abused Drugs. New York, NY: Springer New York; 1987. pp. 211–227. doi:10.1007/978-1-4612-4812-5_12

50. Weiss F, Maldonado-Vlaar CS, Parsons LH, Kerr TM, Smith DL, Ben-Shahar O. Control of cocaine-seeking behavior by drug-associated stimuli in rats: effects on recovery of extinguished operant-responding and extracellular dopamine levels in amygdala and nucleus accumbens. Proc Natl Acad Sci USA. 2000;97: 4321–4326.

51. Robinson T. The neural basis of drug craving: An incentive-sensitization theory of addiction. Brain Research Reviews. 1993;18: 247–291. doi:10.1016/0165-0173(93)90013-P

52. DeJong W. Relapse prevention: an emerging technology for promoting long-term drug abstinence. Int J Addict. 1994;29: 681–705.

53. Flagel SB, Akil H, Robinson TE. Individual differences in the attribution of incentive salience to reward-related cues: Implications for addiction. Neuropharmacology. 2009;56: 139–148. doi:10.1016/j.neuropharm.2008.06.027

54. Stormark KM, Laberg JC, Nordby H, Hugdahl K. Alcoholics' selective attention to alcohol stimuli: automated processing? J Stud Alcohol. 2000;61: 18–23.

55. Copersino ML, Serper MR, Vadhan N, Goldberg BR, Richarme D, Chou JC-Y, et al. Cocaine craving and attentional bias in cocaine-dependent schizophrenic patients. Psychiatry Res. 2004;128: 209–218. doi:10.1016/j.psychres.2004.07.006

56. Cox WM, Fadardi JS, Pothos EM. The addiction-stroop test: Theoretical considerations and procedural recommendations. Psychol Bull. 2006;132: 443–476. doi:10.1037/0033-2909.132.3.443

57. Rosse RB, Johri S, Kendrick K, Hess AL, Alim TN, Miller M, et al. Preattentive and attentive eye movements during visual scanning of a cocaine cue: correlation with intensity of cocaine cravings. J Neuropsychiatry Clin Neurosci. 1997;9: 91–93. doi:10.1176/jnp.9.1.91

58. Lubman DI, Peters LA, Mogg K, Bradley BP, Deakin JF. Attentional bias for drug cues in opiate dependence. Psychol Med. 2000;30: 169–175.

59. Franken IH, Kroon LY, Wiers RW, Jansen A. Selective cognitive processing of drug cues in heroin dependence. J Psychopharmacol (Oxford). 2000;14: 395–400. doi:10.1177/026988110001400408

60. Mahler SV, de Wit H. Cue-reactors: individual differences in cue-induced craving after food or smoking abstinence. PLoS ONE. 2010;5: e15475. doi:10.1371/journal.pone.0015475

61. Field M, Mogg K, Zetteler J, Bradley BP. Attentional biases for alcohol cues in heavy and light social drinkers: the roles of initial orienting and maintained attention. Psychopharmacology (Berl). 2004;176: 88–93. doi:10.1007/s00213-004-1855-1

62. Carpenter KM, Martinez D, Vadhan NP, Barnes-Holmes D, Nunes EV. Measures of attentional bias and relational responding are associated with behavioral treatment outcome for cocaine dependence. Am J Drug Alcohol Abuse. 2012;38: 146–154. doi:10.3109/00952990.2011.643986

63. Cox WM, Hogan LM, Kristian MR, Race JH. Alcohol attentional bias as a predictor of alcohol abusers' treatment outcome. Drug Alcohol Depend. 2002;68: 237–243.

64. Marissen MAE, Franken IHA, Waters AJ, Blanken P, van den Brink W, Hendriks VM. Attentional bias predicts heroin relapse following treatment. Addiction. 2006;101: 1306–1312. doi:10.1111/j.1360-0443.2006.01498.x

65. Tomie A, Grimes KL, Pohorecky LA. Behavioral characteristics and neurobiological substrates shared by Pavlovian sign-tracking and drug abuse. Brain Res Rev. 2008;58: 121–135. doi:10.1016/j.brainresrev.2007.12.003

66. Bauer D, Cox WM. Alcohol-related words are distracting to both alcohol abusers and non-abusers in the Stroop colour-naming task. Addiction. 1998;93: 1539–1542.

67. Kilts CD, Kennedy A, Elton AL, Tripathi SP, Young J, Cisler JM, et al. Individual differences in attentional bias associated with cocaine dependence are related to varying engagement of neural processing networks. Neuropsychopharmacology. 2014;39: 1135–1147. doi:10.1038/npp.2013.314

68. Carpenter KM, Schreiber E, Church S, McDowell D. Drug Stroop performance: relationships with primary substance of use and treatment outcome in a drug-dependent outpatient sample. Addict Behav. 2006;31: 174–181. doi:10.1016/j.addbeh.2005.04.012

69. Ho MK, Goldman D, Heinz A, Kaprio J, Kreek MJ, Li MD, et al. Breaking barriers in the genomics and pharmacogenetics of drug addiction. Clin Pharmacol Ther. 2010;88: 779–791. doi:10.1038/clpt.2010.175

70. Hart AB, Engelhardt BE, Wardle MC, Sokoloff G, Stephens M, de Wit H, et al. Genome-Wide Association Study of d-Amphetamine Response in Healthy Volunteers Identifies Putative Associations, Including Cadherin 13 (CDH13). Arking DE, editor. PLoS ONE. 2012;7: e42646. doi:10.1371/journal.pone.0042646

71. Danna CL, Shepard PD, Elmer GI. The habenula governs the attribution of incentive salience to reward predictive cues. Front Hum Neurosci. 2013;7: 781. doi:10.3389/fnhum.2013.00781

72. Darvas M, Wunsch AM, Gibbs JT, Palmiter RD. Dopamine dependency for acquisition and performance of Pavlovian conditioned response. Proc Natl Acad Sci USA. 2014;111: 2764–2769. doi:10.1073/pnas.1400332111

73. Flagel SB, Watson SJ, Robinson TE, Akil H. Individual differences in the propensity to approach signals vs goals promote different adaptations in the dopamine system of rats. Psychopharmacology. 2007;191: 599–607. doi:10.1007/s00213-006-0535-8

74. Flagel SB, Clark JJ, Robinson TE, Mayo L, Czuj A, Willuhn I, et al. A selective role for dopamine in stimulus-reward learning. Nature. 2011;469: 53–57. doi:10.1038/nature09588

75. Flagel SB, Robinson TE, Clark JJ, Clinton SM, Watson SJ, Seeman P, et al. An Animal Model of Genetic Vulnerability to Behavioral Disinhibition and Responsiveness to Reward-Related Cues: Implications for Addiction. Neuropsychopharmacology. 2010;35: 388–400. doi:10.1038/npp.2009.142

76. Fitzpatrick CJ, Gopalakrishnan S, Cogan ES, Yager LM, Meyer PJ, Lovic V, et al. Variation in the Form of Pavlovian Conditioned Approach Behavior among Outbred Male Sprague-Dawley Rats from Different Vendors and Colonies: Sign-Tracking vs. Goal-Tracking. Campolongo P, editor. PLoS ONE. 2013;8: e75042. doi:10.1371/journal.pone.0075042

77. Meyer PJ, Ma ST, Robinson TE. A cocaine cue is more preferred and evokes more frequency-modulated 50-kHz ultrasonic vocalizations in rats prone to attribute incentive salience to a food cue. Psychopharmacology (Berl). 2012;219: 999–1009. doi:10.1007/s00213-011-2429-7

78. Yager LM, Robinson TE. A classically conditioned cocaine cue acquires greater control over motivated behavior in rats prone to attribute incentive salience to a food cue. Psychopharmacology. 2013;226: 217–228. doi:10.1007/s00213-012-2890-y

79. Saunders BT, Robinson TE. Individual variation in the motivational properties of cocaine. Neuropsychopharmacology. 2011;36: 1668–1676. doi:10.1038/npp.2011.48

80. Saunders BT, O'Donnell EG, Aurbach EL, Robinson TE. A Cocaine Context Renews Drug Seeking Preferentially in a Subset of Individuals. Neuropsychopharmacology. 2014;39: 2816–2823. doi:10.1038/npp.2014.131

81. Stephens DN, Crombag HS, Duka T. The challenge of studying parallel behaviors in humans and animal models. Curr Top Behav Neurosci. 2013;13: 611–645. doi:10.1007/7854_2011_133

82. Billings LK, Florez JC. The genetics of type 2 diabetes: what have we learned from GWAS? Ann N Y Acad Sci. 2010;1212: 59–77. doi:10.1111/j.1749-6632.2010.05838.x

83. Cao C, Moult J. GWAS and drug targets. BMC Genomics. 2014;15 Suppl 4: S5. doi:10.1186/1471-2164-15-S4-S5

84. Motsinger-Reif AA, Jorgenson E, Relling MV, Kroetz DL, Weinshilboum R, Cox NJ, et al. Genome-wide association studies in pharmacogenomics: successes and lessons. Pharmacogenet Genomics. 2013;23: 383–394. doi:10.1097/FPC.0b013e32833d7b45

85. Darvasi A, Soller M. Advanced intercross lines, an experimental population for fine genetic mapping. Genetics. 1995;141: 1199–1207.

86. Gonzales NM, Seo J, Hernandez-Cordero AI, St. Pierre CL, Gregory JS, Distler MG, et al. Genome wide association analysis in a mouse advanced intercross line. 2018; doi:10.1101/230920

87. Zhou X, St. Pierre CL, Gonzales NM, Cheng R, Chitre AS, Sokoloff G, et al. Genome-wide association study, replication, and mega-analysis using a dense marker panel in a multi-generational mouse advanced intercross line. 2018; doi:10.1101/387613

88. Pollard DA. Design and Construction of Recombinant Inbred Lines. In: Rifkin SA, editor. Quantitative Trait Loci (QTL). Totowa, NJ: Humana Press; 2012. pp. 31–39. doi:10.1007/978-1-61779-785-9_3

89. Takuno S, Terauchi R, Innan H. The Power of QTL Mapping with RILs. Zhang J, editor. PLoS ONE. 2012;7: e46545. doi:10.1371/journal.pone.0046545

90. Churchill GA, Gatti DM, Munger SC, Svenson KL. The diversity outbred mouse population. Mammalian Genome. 2012;23: 713–718. doi:10.1007/s00335-012-9414-2

91. Svenson KL, Gatti DM, Valdar W, Welsh CE, Cheng R, Chesler EJ, et al. High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population. Genetics. 2012;190: 437–447. doi:10.1534/genetics.111.132597

92. Hansen C, Spuhler K. Development of the National Institutes of Health genetically heterogeneous rat stock. Alcohol Clin Exp Res. 1984;8: 477–479.

93. Johannesson M, Lopez-Aumatell R, Stridh P, Diez M, Tuncel J, Blazquez G, et al. A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: The NIH heterogeneous stock. Genome Research. 2008;19: 150–158. doi:10.1101/gr.081497.108

94. Woods LCS, Mott R. Heterogeneous Stock Populations for Analysis of Complex Traits. In: Schughart K, Williams RW, editors. Systems Genetics. New York, NY: Springer New York; 2017. pp. 31–44. doi:10.1007/978-1-4939-6427-7_2

95. Yalcin B, Nicod J, Bhomra A, Davidson S, Cleak J, Farinelli L, et al. Commercially Available Outbred Mice for Genome-Wide Association Studies. Barsh GS, editor. PLoS Genetics. 2010;6: e1001085. doi:10.1371/journal.pgen.1001085

96. Parker CC, Gopalakrishnan S, Carbonetto P, Gonzales NM, Leung E, Park YJ, et al. Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice. Nature Genetics. 2016;48: 919–926. doi:10.1038/ng.3609

97. Gileta AF, Fitzpatrick CJ, Chitre AS, St. Pierre CL, Joyce EV, Maguire RJ, et al. Genetic characterization of outbred Sprague Dawley rats and utility for genome-wide association studies. 2018; doi:10.1101/412924

98. Flint J, Valdar W, Shifman S, Mott R. Strategies for mapping and cloning quantitative trait genes in rodents. Nature Reviews Genetics. 2005;6: 271–286. doi:10.1038/nrg1576

99. Aldinger KA, Sokoloff G, Rosenberg DM, Palmer AA, Millen KJ. Genetic Variation and Population Substructure in Outbred CD-1 Mice: Implications for Genome-Wide Association Studies. Crusio WE, editor. PLoS ONE. 2009;4: e4729. doi:10.1371/journal.pone.0004729

100. Charles River. CD® (Sprague Dawley) IGS Rat Details. In: CD® (Sprague Dawley) IGS Rat [Internet]. [cited 2 Jul 2018]. Available: https://www.criver.com/products-services/find-model/cd-sd-igs-rat?region=3611

101. Prejean JD, Peckham JC, Casey AE, Griswold DP, Weisburger EK, Weisburger JH. Spontaneous tumors in Sprague-Dawley rats and Swiss mice. Cancer Res. 1973;33: 2768–2773.

102. Clark FM, Yeomans DC, Proudfit HK. The noradrenergic innervation of the spinal cord: differences between two substrains of Sprague-Dawley rats determined using retrograde tracers combined with immunocytochemistry. Neurosci Lett. 1991;125: 155–158.

103. Turnbull AV, Rivier CL. Sprague-Dawley Rats Obtained from Different Vendors Exhibit Distinct Adrenocorticotropin Responses to Inflammatory Stimuli. Neuroendocrinology. 1999;70: 186–195. doi:10.1159/000054475

104. Fuller DD, Baker TL, Behan M, Mitchell GS. Expression of hypoglossal long-term facilitation differs between substrains of Sprague-Dawley rat. Physiological Genomics. 2001;4: 175–181. doi:10.1152/physiolgenomics.2001.4.3.175

105. Bodnar TS, Hill LA, Taves MD, Yu W, Soma KK, Hammond GL, et al. Colony-Specific Differences in Endocrine and Immune Responses to an Inflammatory Challenge in Female Sprague Dawley Rats. Endocrinology. 2015;156: 4604–4617. doi:10.1210/en.2015-1497

106. Brower M, Grace M, Kotz CM, Koya V. Comparative analysis of growth characteristics of Sprague Dawley rats obtained from different sources. Laboratory Animal Research. 2015;31: 166. doi:10.5625/lar.2015.31.4.166

107. Weber K. Differences in Types and Incidence of Neoplasms in Wistar Han and Sprague-Dawley Rats. Toxicologic Pathology. 2017;45: 64–75. doi:10.1177/0192623316672075

108. The STAR Consortium*, Saar K, Beck A, Bihoreau M-T, Birney E, Brocklebank D, et al. SNP and haplotype mapping for genetic analysis in the rat. Nature Genetics. 2008;40: 560–566. doi:10.1038/ng.124

109. Holl K, He H, Wedemeyer M, Clopton L, Wert S, Meckes JK, et al. Heterogeneous stock rats: a model to study the genetics of despair-like behavior in adolescence: Depression and fluoxetine in adolescent outbred rats. Genes, Brain and Behavior. 2018;17: 139–148. doi:10.1111/gbb.12410

110. Chitre AS, Polesskaya O, Holl K, Gao J, Cheng R, Martinez A, et al. Genome wide association study of body weight, body mass index, adiposity, and fasting glucose in 3,173 outbred rats. 2018; doi:10.1101/422428

111. van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, Verstege E, et al. Complexity Reduction of Polymorphic Sequences (CRoPS^{TM}): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. Baxter I, editor. PLoS ONE. 2007;2: e1172. doi:10.1371/journal.pone.0001172

112. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. Orban L, editor. PLoS ONE. 2011;6: e19379. doi:10.1371/journal.pone.0019379

113. Poland JA, Brown PJ, Sorrells ME, Jannink J-L. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. Yin T, editor. PLoS ONE. 2012;7: e32253. doi:10.1371/journal.pone.0032253

114. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. Fay JC, editor. PLoS ONE. 2008;3: e3376. doi:10.1371/journal.pone.0003376

115. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. Orlando L, editor. PLoS ONE. 2012;7: e37135. doi:10.1371/journal.pone.0037135

116. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, et al. Assessing the impact of population stratification on genetic association studies. Nature Genetics. 2004;36: 388–393. doi:10.1038/ng1333

117. Voight BF, Pritchard JK. Confounding from Cryptic Relatedness in Case-Control Association Studies. PLoS Genetics. 2005;1: e32. doi:10.1371/journal.pgen.0010032

118. Paschou P, Drineas P, Lewis J, Nievergelt CM, Nickerson DA, Smith JD, et al. Tracing Sub-Structure in the European American Population with PCA-Informative Markers. Pritchard JK, editor. PLoS Genetics. 2008;4: e1000114. doi:10.1371/journal.pgen.1000114

119. Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scarnicci F, et al. Discerning the Ancestry of European Americans in Genetic Association Studies. PLoS Genetics. 2008;4: e236. doi:10.1371/journal.pgen.0030236

120. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006;38: 904–909. doi:10.1038/ng1847

121. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics. 2010;11: 459–463. doi:10.1038/nrg2813

122. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010;42: 348–354. doi:10.1038/ng.548

123. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics. 2012;44: 821–824. doi:10.1038/ng.2310

124. Cheng R, Lim JE, Samocha KE, Sokoloff G, Abney M, Skol AD, et al. Genome-Wide Association Studies and the Problem of Relatedness Among Advanced Intercross Lines and Other Highly Recombinant Populations. Genetics. 2010;185: 1033–1044. doi:10.1534/genetics.110.116863

125. Gonzales NM, Palmer AA. Fine-mapping QTLs in advanced intercross lines and other outbred populations. Mammalian Genome. 2014;25: 271–292. doi:10.1007/s00335-014-9523-1

126. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. Nature Reviews Genetics. 2010;11: 356.

127. Marigorta UM, Navarro A. High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants. Williams SM, editor. PLoS Genetics. 2013;9: e1003566. doi:10.1371/journal.pgen.1003566

128. Salinas YD, Wang L, DeWan AT. Multiethnic genome-wide association study identifies ethnic-specific associations with body mass index in Hispanics and African Americans. BMC Genetics. 2016;17. doi:10.1186/s12863-016-0387-0

129. Thomas D. Gene–environment-wide association studies: emerging approaches. Nature Reviews Genetics. 2010;11: 259–272. doi:10.1038/nrg2764

130. Wood AR, Hernandez DG, Nalls MA, Yaghootkar H, Gibbs JR, Harries LW, et al. Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. Human Molecular Genetics. 2011;20: 4082–4092. doi:10.1093/hmg/ddr328

131. Seyerle AA, Young AM, Jeff JM, Melton PE, Jorgensen NW, Lin Y, et al. Evidence of Heterogeneity by Race/Ethnicity in Genetic Determinants of QT Interval: Epidemiology. 2014;25: 790–798. doi:10.1097/EDE.0000000000000168

132. Sittig LJ, Carbonetto P, Engel KA, Krauss KS, Barrios-Camacho CM, Palmer AA. Genetic Background Limits Generalizability of Genotype-Phenotype Relationships. Neuron. 2016;91: 1253–1259. doi:10.1016/j.neuron.2016.08.013

133. Kraft P, Zeggini E, Ioannidis JPA. Replication in Genome-Wide Association Studies. Statistical Science. 2009;24: 561–573. doi:10.1214/09-STS290

134. NCI-NHGRI Working Group on Replication in Association Studies, Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, et al. Replicating genotype–phenotype associations. Nature. 2007;447: 655.

135. Palmer C, Pe'er I. Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. Marchini J, editor. PLOS Genetics. 2017;13: e1006916. doi:10.1371/journal.pgen.1006916

136. Evangelou E, Maraganore DM, Ioannidis JPA. Meta-Analysis in Genome-Wide Association Datasets: Strategies and Application in Parkinson Disease. Rutherford S, editor. PLoS ONE. 2007;2: e196. doi:10.1371/journal.pone.0000196

137. Panagiotou OA, Willer CJ, Hirschhorn JN, Ioannidis JPA. The Power of Meta-Analysis in Genome-Wide Association Studies. Annual Review of Genomics and Human Genetics. 2013;14: 441–465. doi:10.1146/annurev-genom-091212-153520

138. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26: 2190–2191. doi:10.1093/bioinformatics/btq340

139. Myers RA, Scott NM, Gauderman WJ, Qiu W, Mathias RA, Romieu I, et al. Genome-wide interaction studies reveal sex-specific asthma risk alleles. Human Molecular Genetics. 2014;23: 5251–5259. doi:10.1093/hmg/ddu222

140. Lam M, Trampush JW, Yu J, Knowles E, Davies G, Liewald DC, et al. Large-Scale Cognitive GWAS Meta-Analysis Reveals Tissue-Specific Neural Expression and Potential Nootropic Drug Targets. Cell Reports. 2017;21: 2597–2613. doi:10.1016/j.celrep.2017.11.028

141. Savage JE, Jansen PR, Stringer S, Watanabe K, Bryois J, de Leeuw CA, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nature Genetics. 2018;50: 912–919. doi:10.1038/s41588-018-0152-6

142. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Research. 2007;17: 240–248. doi:10.1101/gr.5681207

143. Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods. 2008;5: 247–252. doi:10.1038/nmeth.1185

144. Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, et al. High-throughput genotyping by whole-genome resequencing. Genome Research. 2009;19: 1068–1076. doi:10.1101/gr.089516.108

145. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics. 2011;12: 499–510. doi:10.1038/nrg3012

146. Sun X, Liu D, Zhang X, Li W, Liu H, Hong W, et al. SLAF-seq: An Efficient Method of Large-Scale De Novo SNP Discovery and Genotyping Using High-Throughput Sequencing. Aerts J, editor. PLoS ONE. 2013;8: e58700. doi:10.1371/journal.pone.0058700

147. Scheben A, Batley J, Edwards D. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. Plant Biotechnology Journal. 2017;15: 149–161. doi:10.1111/pbi.12645

148. Chen Q, Ma Y, Yang Y, Chen Z, Liao R, Xie X, et al. Genotyping by Genome Reducing and Sequencing for Outbred Animals. Zhao S, editor. PLoS ONE. 2013;8: e67500. doi:10.1371/journal.pone.0067500

149. He J, Zhao X, Laroche A, Lu Z-X, Liu H, Li Z. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Frontiers in Plant Science. 2014;5. doi:10.3389/fpls.2014.00484

150. Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, et al. An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. Liu Z, editor. PLoS ONE. 2013;8: e54603. doi:10.1371/journal.pone.0054603

151. Furuta T, Ashikari M, Jena KK, Doi K, Reuscher S. Adapting Genotyping-by-Sequencing for Rice F2 Populations. G3: Genes|Genomes|Genetics. 2017;7: 881–893. doi:10.1534/g3.116.038190

152. Fu Y-B, Yang M-H. Genotyping-by-Sequencing and Its Application to Oat Genomic Research. In: Gasparis S, editor. Oat. New York, NY: Springer New York; 2017. pp. 169–187. doi:10.1007/978-1-4939-6682-0_13

153. Pértille F, Guerrero-Bosagna C, Silva VH da, Boschiero C, Nunes J de R da S, Ledur MC, et al. High-throughput and Cost-effective Chicken Genotyping Using Next-Generation Sequencing. Scientific Reports. 2016;6: 26929.

154. Wang Y, Cao X, Zhao Y, Fei J, Hu X, Li N. Optimized double-digest genotyping by sequencing (ddGBS) method with high-density SNP markers and high genotyping accuracy for chickens. Xu P, editor. PLOS ONE. 2017;12: e0179073. doi:10.1371/journal.pone.0179073

155. Johnson JL, Wittgenstein H, Mitchell SE, Hyma KE, Temnykh SV, Kharlamova AV, et al. Genotyping-By-Sequencing (GBS) Detects Genetic Structure and Confirms Behavioral QTL in Tame and Aggressive Foxes (Vulpes vulpes). Murphy WJ, editor. PLOS ONE. 2015;10: e0127013. doi:10.1371/journal.pone.0127013

156. De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG. Genotyping-by-Sequencing (GBS): A Novel, Efficient and Cost-Effective Genotyping Method for Cattle Using Next-Generation Sequencing. Nelson JC, editor. PLoS ONE. 2013;8: e62137. doi:10.1371/journal.pone.0062137

157. Gonzales NM, Seo J, Hernandez-Cordero AI, St. Pierre CL, Gregory JS, Distler MG, et al. Genome wide association study of behavioral, physiological and gene expression traits in a multigenerational mouse intercross. 2017; doi:10.1101/230920

158. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Molecular Ecology. 2013;22: 3124–3140. doi:10.1111/mec.12354

159. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. Tinker NA, editor. PLoS ONE. 2014;9: e90346. doi:10.1371/journal.pone.0090346

160. Torkamaneh D, Laroche J, Bastien M, Abed A, Belzile F. Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC Bioinformatics. 2017;18. doi:10.1186/s12859-016-1431-9

161. Wickland DP, Battu G, Hudson KA, Diers BW, Hudson ME. A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. BMC Bioinformatics. 2017;18. doi:10.1186/s12859-017-2000-6

162. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Research. 2002;12: 996–1006. doi:10.1101/gr.229102

163. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000;16: 276–277. doi:10.1016/S0168-9525(00)02024-2

164. Roberts RJ, Macelis D. REBASE--restriction enzymes and methylases. Nucleic Acids Research. 1999;27: 312–313. doi:10.1093/nar/27.1.312

165. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 2014;15. doi:10.1186/s12859-014-0356-4

166. Browning BL, Browning SR. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. The American Journal of Human Genetics. 2009;84: 210–223. doi:10.1016/j.ajhg.2009.01.005

167. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. The American Journal of Human Genetics. 2016;98: 116–126. doi:10.1016/j.ajhg.2015.11.020

168. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. Schork NJ, editor. PLoS Genetics. 2009;5: e1000529. doi:10.1371/journal.pgen.1000529

169. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics. 2012;44: 955.

170. Hannon Lab. FASTX-Toolkit [Internet]. 2010. Available: http://hannonlab.cshl.edu/fastx_toolkit/index.html

171. Herten K, Hestand MS, Vermeesch JR, Van Houdt JK. GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. BMC Bioinformatics. 2015;16. doi:10.1186/s12859-015-0514-3

172. Andrews S. FastQC [Internet]. 2017. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

173. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17: 10. doi:10.14806/ej.17.1.200

174. Purcell S. PLINK/SEQ: A library for the analysis of genetic variation data [Internet]. 2014. Available: http://atgu.mgh.harvard.edu/plinkseq/index.shtml

175. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

176. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20: 1297–1303. doi:10.1101/gr.107524.110

177. Hermsen R, de Ligt J, Spee W, Blokzijl F, Schäfer S, Adami E, et al. Genomic landscape of rat strain and substrain variation. BMC Genomics. 2015;16. doi:10.1186/s12864-015-1594-1

178. Durvasula A, Hoffman PJ, Kent TV, Liu C, Kono TJY, Morrell PL, et al. ANGSD-wrapper: utilities for analyzing next generation sequencing data. doi:10.7287/peerj.preprints.1472v2

179. Ramdas S, Ozel AB, Holl K, Mandel M, Solberg Woods L, Li JZ. Extended regions of suspected mis-assembly in the rat reference genome. 2018; doi:10.1101/332932

180. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330

181. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nature Methods. 2011;9: 179.

182. Steen RG, Kwitek-Black AE, Glenn C, Gullings-Handley J, Van Etten W, Atkinson OS, et al. A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. Genome Res. 1999;9: AP1-8, insert.

183. Jensen-Seaman MI. Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. Genome Research. 2004;14: 528–538. doi:10.1101/gr.1970304

184. Littrell J, Tsaih S-W, Baud A, Rastas P, Solberg-Woods L, Flister MJ. A High-Resolution Genetic Map for the Laboratory Rat. G3: Genes|Genomes|Genetics. 2018; g3.200187.2018. doi:10.1534/g3.118.200187

185. Kanagawa T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). Journal of Bioscience and Bioengineering. 2003;96: 317–323. doi:10.1016/S1389-1723(03)90130-7

186. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology. 2011;12: R18. doi:10.1186/gb-2011-12-2-r18

187. Illumina, Inc. Nextera(R) Library Validation and Cluster Density Optimization: Guidelines for generating high-quality data with Nextera library prepartion kits. [Internet]. 2014. Available: https://www.illumina.com/documents/products/technotes/technote_nextera_library_validation.pdf

188. Flanagan SP, Jones AG. Substantial differences in bias between single-digest and double-digest RAD-seq libraries: A case study. Molecular Ecology Resources. 2018;18: 264–280. doi:10.1111/1755-0998.12734

189. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43: 491–498. doi:10.1038/ng.806

190. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research. 2008;18: 1851–1858. doi:10.1101/gr.078212.108

191. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics. 2011;12: 443–451. doi:10.1038/nrg2986

192. WGS500 Consortium, Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics. 2014;46: 912–918. doi:10.1038/ng.3036

193. Li Z, Wang Y, Wang F. A study on fast calling variants from next-generation sequencing data using decision tree. BMC Bioinformatics. 2018;19. doi:10.1186/s12859-018-2147-9

194. Howie B, Marchini J, Stephens M. Genotype Imputation with Thousands of Genomes. G3: Genes|Genomes|Genetics. 2011;1: 457–470. doi:10.1534/g3.111.001198

195. Huang G-H, Tseng Y-C. Genotype imputation accuracy with different reference panels in admixed populations. BMC Proceedings. 2014;8: S64. doi:10.1186/1753-6561-8-S1-S64

196. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27: 2987–2993. doi:10.1093/bioinformatics/btr509

197. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Scientific Reports. 2015;5: 17875.

198. Nicod J, Davies RW, Cai N, Hassett C, Goodstadt L, Cosgrove C, et al. Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. Nature Genetics. 2016;48: 912–918. doi:10.1038/ng.3595

199. Envigo, Inc. Sprague Dawley® outbred rat | Envigo. In: Envigo [Internet]. [cited 2 Jul 2018]. Available: https://www.envigo.com/products-services/research-models-services/models/research-models/rats/outbred/sprague-dawley-outbred-rat/

200. Charles River Laboratories International, Inc. International Genetic Standardization (IGS) Program [Internet]. 2016. Available: https://www.criver.com/sites/default/files/Technical%20Resources/Charles%20River%20International%20Genetic%20Standardization%20Program.pdf

201. White WJ, Lee CS. The Development and Maintenance of the Crl:CD(SD) IGS BR Rat Breeding System [Internet]. 1998. Available: https://www.crj.co.jp/cms/cmsrs/pdf/company/rm_rm_a_igs_rat_breeding_system.pdf

202. Envigo, Inc. Genetic integrity assurance program | Envigo. In: Envigo [Internet]. 2 Jul 2018 [cited 2 Jul 2018]. Available: https://www.envigo.com/products-services/research-models-services/resources/genetic-integrity-assurance-program/

203. Poiley SM. A systematic method of breeder rotation for non-inbred laboratory colonies. Proc Animal Care Panel. 1960;10: 159–166.

204. Robinson TE, Yager LM, Cogan ES, Saunders BT. On the motivational properties of reward cues: Individual differences. Neuropharmacology. 2014;76: 450–459. doi:10.1016/j.neuropharm.2013.05.040

205. Meyer PJ, Lovic V, Saunders BT, Yager LM, Flagel SB, Morrow JD, et al. Quantifying Individual Variation in the Propensity to Attribute Incentive Salience to Reward Cues. Zhuang X, editor. PLoS ONE. 2012;7: e38987. doi:10.1371/journal.pone.0038987

206. Pitchers KK, Flagel SB, O'Donnell EG, Solberg Woods LC, Sarter M, Robinson TE. Individual variation in the propensity to attribute incentive salience to a food cue: Influence of sex. Behavioural Brain Research. 2015;278: 462–469. doi:10.1016/j.bbr.2014.10.036

207. Pitchers KK, Wood TR, Skrzynski CJ, Robinson TE, Sarter M. The ability for cocaine and cocaine-associated cues to compete for attention. Behavioural Brain Research. 2017;320: 302–315. doi:10.1016/j.bbr.2016.11.024

208. Kawa AB, Bentzley BS, Robinson TE. Less is more: prolonged intermittent access cocaine self-administration produces incentive-sensitization and addiction-like behavior. Psychopharmacology. 2016;233: 3587–3602. doi:10.1007/s00213-016-4393-8

209. Ahrens AM, Meyer PJ, Ferguson LM, Robinson TE, Aldridge JW. Neural Activity in the Ventral Pallidum Encodes Variation in the Incentive Value of a Reward Cue. The Journal of Neuroscience. 2016;36: 7957–7970. doi:10.1523/JNEUROSCI.0736-16.2016

210. Ahrens AM, Singer BF, Fitzpatrick CJ, Morrow JD, Robinson TE. Rats that sign-track are resistant to Pavlovian but not instrumental extinction. Behavioural Brain Research. 2016;296: 418–430. doi:10.1016/j.bbr.2015.07.055

211. Singer BF, Guptaroy B, Austin CJ, Wohl I, Lovic V, Seiler JL, et al. Individual variation in incentive salience attribution and accumbens dopamine transporter expression and function. Dalley J, editor. European Journal of Neuroscience. 2016;43: 662–670. doi:10.1111/ejn.13134

212. Yager LM, Pitchers KK, Flagel SB, Robinson TE. Individual Variation in the Motivational and Neurobiological Effects of an Opioid Cue. Neuropsychopharmacology. 2015;40: 1269–1277. doi:10.1038/npp.2014.314

213. Meyer PJ, Cogan ES, Robinson TE. The Form of a Conditioned Stimulus Can Influence the Degree to Which It Acquires Incentive Motivational Properties. Le Foll B, editor. PLoS ONE. 2014;9: e98163. doi:10.1371/journal.pone.0098163

214. Saunders BT, Yager LM, Robinson TE. Cue-Evoked Cocaine "Craving": Role of Dopamine in the Accumbens Core. Journal of Neuroscience. 2013;33: 13989–14000. doi:10.1523/JNEUROSCI.0450-13.2013

215. Paolone G, Angelakos CC, Meyer PJ, Robinson TE, Sarter M. Cholinergic Control over Attention in Rats Prone to Attribute Incentive Salience to Reward Cues. Journal of Neuroscience. 2013;33: 8321–8335. doi:10.1523/JNEUROSCI.0709-13.2013

216. Morrow JD, Saunders BT, Maren S, Robinson TE. Sign-tracking to an appetitive cue predicts incubation of conditioned fear in rats. Behavioural Brain Research. 2015;276: 59–66. doi:10.1016/j.bbr.2014.04.002

217. Singer BF, Bryan MA, Popov P, Scarff R, Carter C, Wright E, et al. The sensory features of a food cue influence its ability to act as an incentive stimulus and evoke dopamine release in the nucleus accumbens core. Learning & Memory. 2016;23: 595–606. doi:10.1101/lm.043026.116

218. Fitzpatrick CJ, Morrow JD. Pavlovian Conditioned Approach Training in Rats. J Vis Exp. 2016; e53580. doi:10.3791/53580

219. Robinson TE, Flagel SB. Dissociating the Predictive and Incentive Motivational Properties of Reward-Related Cues Through the Study of Individual Differences. Biological Psychiatry. 2009;65: 869–873. doi:10.1016/j.biopsych.2008.09.006

220. Gao J, Gileta A, Palmer A. An efficient variant calling pipeline for double digest genotype-by-sequencing data in heterogeneous stock rats. 2018.

221. Grabowski PP, Morris GP, Casler MD, Borevitz JO. Population genomic variation reveals roles of history, adaptation and ploidy in switchgrass. Molecular Ecology. 2014;23: 4059–4073. doi:10.1111/mec.12845

222. Broad Institute. Picard Tools [Internet]. 2018. Available: http://broadinstitute.github.io/picard/

223. Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence without reference panels. Nature Genetics. 2016;48: 965–969. doi:10.1038/ng.3594

224. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available: http://www.R-project.org/

225. Wigginton JE, Cutler DJ, Abecasis GR. A Note on Exact Tests of Hardy-Weinberg Equilibrium. The American Journal of Human Genetics. 2005;76: 887–893. doi:10.1086/429864

226. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genetics. 2006;2: e190. doi:10.1371/journal.pgen.0020190

227. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: The impact of rare variants. Genome Research. 2013;23: 1514–1521. doi:10.1101/gr.154831.113

228. Thomas Lumley based on Fortran code by Alan Miller. leaps: Regression Subset Selection. R package version 3.0. [Internet]. 2017. Available: https://CRAN.R-project.org/package=leaps

229. Alboukadel Kassambara and Fabian Mundt. factoextra: Extract and Visualize the Results of Multivariate Data Analyses [Internet]. Available: https://CRAN.R-project.org/package=factoextra

230. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88: 76–82. doi:10.1016/j.ajhg.2010.11.011

231. Cheng R, Palmer AA. A Simulation Study of Permutation, Bootstrap, and Gene Dropping for Assessing Statistical Significance in the Case of Unequal Relatedness. Genetics. 2013;193: 1015–1018. doi:10.1534/genetics.112.146332

232. Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D. Improved linear mixed models for genome-wide association studies. Nature Methods. 2012;9: 525–526. doi:10.1038/nmeth.2037

233. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nature Genetics. 2014;46: 100–106. doi:10.1038/ng.2876

234. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature Genetics. 2015;47: 284–290. doi:10.1038/ng.3190

235. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26: 2336–2337. doi:10.1093/bioinformatics/btq419

236. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273: 1516–1517.

237. Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping. Genetics. 1994;138: 963–971.

238. Han B, Kang HM, Eskin E. Rapid and Accurate Multiple Testing Correction and Power Estimation for Millions of Correlated Markers. Storey JD, editor. PLoS Genetics. 2009;5: e1000456. doi:10.1371/journal.pgen.1000456

239. Han B, Eskin E. Interpreting Meta-Analyses of Genome-Wide Association Studies. Kerr K, editor. PLoS Genetics. 2012;8: e1002555. doi:10.1371/journal.pgen.1002555

240. Joo JWJ, Hormozdiari F, Han B, Eskin E. Multiple testing correction in linear mixed models. Genome Biology. 2016;17. doi:10.1186/s13059-016-0903-6

241. Gauderman W, Morrison J. QUANTO documentation. (Technical report no. 157). Los Angeles, CA: Department of Preventive Medicine, University of Southern California; 2001.

242. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol. 2002;155: 478–484.

243. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting FST. Nature Reviews Genetics. 2009;10: 639–650. doi:10.1038/nrg2611

244. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007;81: 559–575. doi:10.1086/519795

245. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4. doi:10.1186/s13742-015-0047-8

246. Clark SL, Aberg KA, Nerella S, Kumar G, McClay JL, Chen W, et al. Combined Whole Methylome and Genomewide Association Study Implicates *CNTN4* in Alcohol Use. Alcoholism: Clinical and Experimental Research. 2015;39: 1396–1405. doi:10.1111/acer.12790

247. Yoshihara Y, Kawasaki M, Tamada A, Nagata S, Kagamiyama H, Mori K. Overlapping and differential expression of BIG-2, BIG-1, TAG-1, and F3: Four members of an axon-associated cell adhesion molecule subgroup of the immunoglobulin superfamily. Journal of Neurobiology. 1995;28: 51–69. doi:10.1002/neu.480280106

248. Mikulska-Ruminska K, Kulik AJ, Benadiba C, Bahar I, Dietler G, Nowak W. Nanomechanics of multidomain neuronal cell adhesion protein contactin revealed by single molecule AFM and SMD. Scientific Reports. 2017;7. doi:10.1038/s41598-017-09482-w

249. Yang Y, Liu W, Fang Z, Shi J, Che F, He C, et al. A Newly Identified Missense Mutation in *FARS2* Causes Autosomal-Recessive Spastic Paraplegia. Human Mutation. 2016;37: 165–169. doi:10.1002/humu.22930

250. Walker MA, Mohler KP, Hopkins KW, Oakley DH, Sweetser DA, Ibba M, et al. Novel Compound Heterozygous Mutations Expand the Recognized Phenotypes of *FARS2* -Linked Disease. Journal of Child Neurology. 2016;31: 1127–1137. doi:10.1177/0883073816643402

251. Vernon HJ, McClellan R, Batista DA, Naidu S. Mutations in *FARS2* and non-fatal mitochondrial dysfunction in two siblings. American Journal of Medical Genetics Part A. 2015;167: 1147–1151. doi:10.1002/ajmg.a.36993

252. Foster JR, Frost D. The History of the Rat. Boorman's Pathology of the Rat. Elsevier; 2018. pp. 7–12. doi:10.1016/B978-0-12-391448-4.00003-4

253. Scott PA, Cierpial MA, Kilts CD, Weiss JM. Susceptibility and resistance of rats to stress-induced decreases in swim-test activity: a selective breeding study. Brain Res. 1996;725: 217–230.

254. Stead JDH, Clinton S, Neal C, Schneider J, Jama A, Miller S, et al. Selective Breeding for Divergence in Novelty-seeking Traits: Heritability and Enrichment in Spontaneous Anxiety-related Behaviors. Behavior Genetics. 2006;36: 697–712. doi:10.1007/s10519-006-9058-7

255. Bertholomey ML, West CHK, Jensen ML, Li T-K, Stewart RB, Weiss JM, et al. Genetic propensities to increase ethanol intake in response to stress: studies with selectively bred swim test susceptible (SUS), alcohol-preferring (P), and non-preferring (NP) lines of rats. Psychopharmacology (Berl). 2011;218: 157–167. doi:10.1007/s00213-011-2381-6

256. Geraldes A, Basset P, Smith KL, Nachman MW. Higher differentiation among subspecies of the house mouse (Mus musculus) in genomic regions with low recombination: RECOMBINATION AND SPECIATION IN MICE. Molecular Ecology. 2011;20: 4722–4736. doi:10.1111/j.1365-294X.2011.05285.x

257. Sanchez-Roige S, Palmer AA, Fontanillas P, Elson S, Adams MJ, Howard DM, et al. Genome-wide association study meta-analysis of the Alcohol Use Disorder Identification Test (AUDIT) in two population-based cohorts (N=141,958). 2018; doi:10.1101/275917

258. Tomie A. Sign-Tracking and Drug Addiction [Internet]. Michigan Publishing, University of Michigan Library; 2018. doi:10.3998/mpub.10215070

259. Pitchers KK, Sarter M, Robinson TE. The hot 'n' cold of cue-induced drug relapse. Learning & Memory. 2018;25: 474–480. doi:10.1101/lm.046995.117

260. Sarter M, Phillips KB. The neuroscience of cognitive-motivational styles: Sign- and goal-trackers as animal models. Behavioral Neuroscience. 2018;132: 1–12. doi:10.1037/bne0000226

261. Langer M, Brandt C, Löscher W. Marked strain and substrain differences in induction of status epilepticus and subsequent development of neurodegeneration, epilepsy, and behavioral alterations in rats. [corrected]. Epilepsy Res. 2011;96: 207–224. doi:10.1016/j.eplepsyres.2011.06.005

262. Segerström L, Roman E. Response: Commentary: Supplier-dependent differences in intermittent voluntary alcohol intake and response to naltrexone in Wistar rats. Frontiers in Neuroscience. 2016;10. doi:10.3389/fnins.2016.00442

263. Berridge KC. Food reward: brain substrates of wanting and liking. Neurosci Biobehav Rev. 1996;20: 1–25.

264. Berridge KC. Reward learning: Reinforcement, incentives, and expectations. Psychology of Learning and Motivation. Elsevier; 2000. pp. 223–278. doi:10.1016/S0079-7421(00)80022-5

265. Berridge KC, Robinson TE. Parsing reward. Trends in Neurosciences. 2003;26: 507–513. doi:10.1016/S0166-2236(03)00233-9

266. Robinson TE, Berridge KC. The neural basis of drug craving: an incentive-sensitization theory of addiction. Brain Res Brain Res Rev. 1993;18: 247–291.

267. Hughson AR, Horvath AP, Holl K, Palmer AA, Solberg Woods LC, Robinson TE, et al. Dissociating addiction-related endophenotypes: Incentive salience attribution, sensation-seeking and novelty-seeking are independent traits in male and female heterogeneous stock rats. 2018; doi:10.1101/421065

268. Robinson TE, Berridge KC. The psychology and neurobiology of addiction: an incentive–sensitization view. Addiction. 2000;95: 91–117. doi:10.1080/09652140050111681

269. Tomie A. Locating reward cue at response manipulandum (CAM) induces symptoms of drug abuse. Neurosci Biobehav Rev. 1996;20: 505–535.

270. Ayduk O, Mendoza-Denton R, Mischel W, Downey G, Peake PK, Rodriguez M. Regulating the interpersonal self: Strategic self-regulation for coping with rejection sensitivity. Journal of Personality and Social Psychology. 2000;79: 776–792. doi:10.1037/0022-3514.79.5.776

271. Saunders BT, Robinson TE. A cocaine cue acts as an incentive stimulus in some but not others: implications for addiction. Biol Psychiatry. 2010;67: 730–736. doi:10.1016/j.biopsych.2009.11.015

272. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. Lewitter F, Kann M, editors. PLoS Computational Biology. 2012;8: e1002822. doi:10.1371/journal.pcbi.1002822

273. Woods LCS, Mott R. Heterogeneous Stock Populations for Analysis of Complex Traits. In: Schughart K, Williams RW, editors. Systems Genetics. New York, NY: Springer New York; 2017. pp. 31–44. doi:10.1007/978-1-4939-6427-7_2

274. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012;28: 2540–2542. doi:10.1093/bioinformatics/bts474

275. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. 2014; doi:10.1101/005165

276. The Wellcome Trust Case Control Consortium, Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. Nature Genetics. 2012;44: 1294.

277. Crowder RJ, Enomoto H, Yang M, Johnson EM, Milbrandt J. Dok-6, a Novel p62 Dok Family Member, Promotes Ret-mediated Neurite Outgrowth. Journal of Biological Chemistry. 2004;279: 42072–42081. doi:10.1074/jbc.M403726200

278. Grandone A, Torella A, Santoro C, Giugliano T, del Vecchio Blanco F, Mutarelli M, et al. Expanding the phenotype of *RTTN* variations: a new family with primary microcephaly, severe growth failure, brain malformations and dermatitis: Expanding the phenotype of *RTTN* variations. Clinical Genetics. 2016;90: 445–450. doi:10.1111/cge.12771

279. Stouffs K, Moortgat S, Vanderhasselt T, Vandervore L, Dica A, Mathot M, et al. Biallelic mutations in RTTN are associated with microcephaly, short stature and a wide range of brain malformations. Eur J Med Genet. 2018; doi:10.1016/j.ejmg.2018.06.001

280. Kheradmand Kia S, Verbeek E, Engelen E, Schot R, Poot RA, de Coo IFM, et al. RTTN Mutations Link Primary Cilia Function to Organization of the Human Cerebral Cortex. The American Journal of Human Genetics. 2012;91: 533–540. doi:10.1016/j.ajhg.2012.07.008

281. Nomura T, Takahashi M, Hara Y, Osumi N. Patterns of Neurogenesis and Amplitude of Reelin Expression Are Essential for Making a Mammalian-Type Cortex. Reh T, editor. PLoS ONE. 2008;3: e1454. doi:10.1371/journal.pone.0001454

282. Weeber EJ, Beffert U, Jones C, Christian JM, Förster E, Sweatt JD, et al. Reelin and ApoE Receptors Cooperate to Enhance Hippocampal Synaptic Plasticity and Learning. Journal of Biological Chemistry. 2002;277: 39944–39952. doi:10.1074/jbc.M205147200

283. Niu S, Renfro A, Quattrocchi CC, Sheldon M, D'Arcangelo G. Reelin Promotes Hippocampal Dendrite Development through the VLDLR/ApoER2-Dab1 Pathway. Neuron. 2004;41: 71–84. doi:10.1016/S0896-6273(03)00819-5

284. Baloyannis SJ. MORPHOLOGICAL AND MORPHOMETRIC ALTERATIONS OF CAJAL-RETZIUS CELLS IN EARLY CASES OF ALZHEIMER'S DISEASE: A GOLGI AND ELECTRON MICROSCOPE STUDY. International Journal of Neuroscience. 2005;115: 965–980. doi:10.1080/00207450590901396

285. Folsom TD, Fatemi SH. The involvement of Reelin in neurodevelopmental disorders. Neuropharmacology. 2013;68: 122–135. doi:10.1016/j.neuropharm.2012.08.015

286. Fatemi SH, Snow AV, Stary JM, Araghi-Niknam M, Reutiman TJ, Lee S, et al. Reelin signaling is impaired in autism. Biol Psychiatry. 2005;57: 777–787. doi:10.1016/j.biopsych.2004.12.018

287. Lakatosova S, Ostatnikova D. Reelin and its complex involvement in brain development and function. Int J Biochem Cell Biol. 2012;44: 1501–1504. doi:10.1016/j.biocel.2012.06.002

288. Maldonado R, Smadja C, Mazucchelli C, Sassone-Corsi P. Altered emotional and locomotor responses in mice deficient in the transcription factor CREM. Proceedings of the National Academy of Sciences. 1999;96: 14094–14099. doi:10.1073/pnas.96.24.14094

289. Mantamadiotis T, Lemberger T, Bleckmann SC, Kern H, Kretz O, Villalba AM, et al. Disruption of CREB function in brain leads to neurodegeneration. Nature Genetics. 2002;31: 47–54. doi:10.1038/ng882

290. Bernstein H-G, Kirches E, Bogerts B, Lendeckel U, Keilhoff G, Zempeltzi M, et al. Wide distribution of CREM immunoreactivity in adult and fetal human brain, with an increased expression in dentate gyrus neurons of Alzheimer's as compared to normal aging brains. Amino Acids. 2013;45: 1373–1383. doi:10.1007/s00726-013-1601-2

291. Wu Q, Sun X, Yue W, Lu T, Ruan Y, Chen T, et al. RAB18, a protein associated with Warburg Micro syndrome, controls neuronal migration in the developing cerebral cortex. Molecular Brain. 2016;9. doi:10.1186/s13041-016-0198-2

292. Miller GM. The emerging role of trace amine-associated receptor 1 in the functional regulation of monoamine transporters and dopaminergic activity: TAAR1 regulation of

monoaminergic activity. Journal of Neurochemistry. 2011;116: 164–176. doi:10.1111/j.1471-4159.2010.07109.x

293. Espinoza S, Ghisi V, Emanuele M, Leo D, Sukhanov I, Sotnikova TD, et al. Postsynaptic D2 dopamine receptor supersensitivity in the striatum of mice lacking TAAR1. Neuropharmacology. 2015;93: 308–313. doi:10.1016/j.neuropharm.2015.02.010

294. Grandy DK, Miller GM, Li J-X. "TAARgeting Addiction"—The Alamo Bears Witness to Another Revolution. Drug and Alcohol Dependence. 2016;159: 9–16. doi:10.1016/j.drugalcdep.2015.11.014

295. Wallace LJ. Trace Amine-Associated Receptor 1. Trace Amines and Neurological Disorders. Elsevier; 2016. pp. 339–347. doi:10.1016/B978-0-12-803603-7.00023-9

296. D'Arcangelo G. Apoer2: A Reelin Receptor to Remember. Neuron. 2005;47: 471–473. doi:10.1016/j.neuron.2005.08.001

297. Lindemann L, Meyer CA, Jeanneau K, Bradaia A, Ozmen L, Bluethmann H, et al. Trace Amine-Associated Receptor 1 Modulates Dopaminergic Activity. Journal of Pharmacology and Experimental Therapeutics. 2007;324: 948–956. doi:10.1124/jpet.107.132647

298. Achat-Mendes C, Lynch LJ, Sullivan KA, Vallender EJ, Miller GM. Augmentation of methamphetamine-induced behaviors in transgenic mice lacking the trace amine-associated receptor 1. Pharmacology Biochemistry and Behavior. 2012;101: 201–207. doi:10.1016/j.pbb.2011.10.025

299. Pei Y, Lee J, Leo D, Gainetdinov RR, Hoener MC, Canales JJ. Activation of the Trace Amine-Associated Receptor 1 Prevents Relapse to Cocaine Seeking. Neuropsychopharmacology. 2014;39: 2299–2308. doi:10.1038/npp.2014.88

300. Jing L, Zhang Y, Li J-X. Effects of the Trace Amine Associated Receptor 1 Agonist RO5263397 on Abuse-Related Behavioral Indices of Methamphetamine in Rats. International Journal of Neuropsychopharmacology. 2015;18: pyu060–pyu060. doi:10.1093/ijnp/pyu060

301. Berridge KC, Robinson TE. Liking, wanting, and the incentive-sensitization theory of addiction. American Psychologist. 2016;71: 670–679. doi:10.1037/amp0000059

302. Liu J-F, Li J-X. TAAR1 in Addiction: Looking Beyond the Tip of the Iceberg. Front Pharmacol. 2018;9: 279. doi:10.3389/fphar.2018.00279

303. Liu J-F, Seaman R, Siemian JN, Bhimani R, Johnson B, Zhang Y, et al. Role of trace amine-associated receptor 1 in nicotine's behavioral and neurochemical effects. Neuropsychopharmacology. 2018;43: 2435–2444. doi:10.1038/s41386-018-0017-9

304. Morris AP. Transethnic meta-analysis of genomewide association studies. Genet Epidemiol. 2011;35: 809–822. doi:10.1002/gepi.20630

305. Lee CH, Eskin E, Han B. Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. Bioinformatics. 2017;33: i379–i388. doi:10.1093/bioinformatics/btx242