Conserved mammalian cell types and epigenomics of the striatum and cortex to identify targets and consequences of addiction genetic risk variants

by

BaDoi Nguyen Phan

Bachelor of Science in Biomedical Engineering, Johns Hopkins University, 2016

Submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Carnegie Mellon University

2023

CARNEGIE MELLON UNIVERSITY

School of Computer Science

Thesis Number

This dissertation was presented

by

BaDoi Nguyen Phan

It was defended on

July 28, 2023

and approved by

Ryan W. Logan, PhD, Department of Psychiatry UMass Chan Medical School

Kathryn Roeder, PhD, Department of Statistics Carnegie Mellon University

Dennis Kostka, PhD, Department of Developmental Biology University of Pittsburgh

Dissertation Director:
Andreas R. Pfenning, PhD, Computational Biology Department Carnegie Mellon University

This research is funded in part by the National Institutes of Health grants F30DA053020, T32GM008208, and 1DP1DA046585. The Computational Biology Department at Carnegie Mellon University specifically disclaims responsibility for any analyses, interpretations, or conclusions. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, donors, or the U.S. Government.

Copyright © by BaDoi Nguyen Phan

2023

Conserved mammalian cell types and epigenomics of the striatum and cortex to identify targets and consequences of addiction genetic risk variants

BaDoi Nguyen Phan, PhD

Carnegie Mellon University, 2023

Substance use disorders are serious mental health conditions that contribute to human suffering and early death. They are influenced by a combination of genetic and environmental factors that lead to the dysregulation of reward areas in the brain, such as the striatum and cortex. To gain a better understanding of how genetics contribute to substance use disorders in these specific brain regions, we delve into the molecular identities of cell types conserved across rodents and primates. By utilizing integrative bioinformatics and machine learning techniques, we can dissect the genetic risk associated with substance use disorders using comparative single-cell genomics across mammalian species and investigate the consequences of opioid use disorder using a cohort of postmortem human brains. Our research yielded several significant findings. First, we examined the epigenomic landscapes of striatal and cortical cell types in mice and postmortem human brains. This examination helped identify the specific tissue and cellular targets affected by different genetic risk variants for substance use disorders. Leveraging comparative epigenomics principles and machine learning, we extended our analysis to include multi-mammalian species' striatal cell type epigenomics. This approach allowed us to finely map human genetic risk variants to conserved cell type-specific putative enhancers, revealing potential causal variants and their shared cellular targets across species. Additionally, we investigated the transcriptional dysregulation of these conserved striatal cell types in postmortem human brains affected by opioid use disorder and overdose. Our findings highlighted cell type-specific signatures of DNA damage, neurodegeneration, and neuroinflammation, which were further confirmed in chronic opioid exposure studies conducted on nonhuman primate striatal cell types. Lastly, to pave the way for future gene therapy targeting these specific cell types, we again employed machine learning to select cell type-specific candidate enhancers. These enhancers were tailored to activate transgene transcription only in molecularly defined neuronal subtypes and were validated in both nonhuman primate and rodent striatum and cortex. In summary, our research demonstrates that genetic risk variants for substance use disorders impact conserved mammalian genomic loci, which can be effectively modeled using comparative "omics" and computational approaches.

Table of Contents

Acknowledgements xv

1.0 Substance use disorder genetics and the conserved reward system 1

1.1 Substance use disorders 1

1.2 Genetics of Substance Use Disorders 2

1.3 Single cell molecular biology to connect the genetics to addiction behavior 4

1.4 Conservation of the genome and implications for biomedical research 6

2.0 Addiction-associated genetic variants implicate brain cell type- and region-specific *cis*-regulatory elements in addiction neurobiology 15

2.1 Abstract 16

2.1.1 Significance Statement: 16

2.2 Introduction 17

2.3 Materials and Methods 19

2.3.1 ATAC-seq data processing pipeline 19

2.3.1.1 Publicly available datasets 20

2.3.2 Mapping mouse open chromatin region (OCR) orthologs 23

2.3.3 Experimental Design 24

2.3.4 Statistical Analyses 27

2.3.5 Linkage-Disequilibrium Score Regression (LDSC) 28

2.3.6 LDSC Regression GWAS Enrichment Backgrounds 35

2.3.7 Interpretation of Convolutional Neural Network models 36

2.4 Results 37

2.4.1 Genetic risk for substance use traits is associated with the neuronal epigenomes of reward areas 37

2.4.2 Mouse-human conserved cell type-specific open chromatin enrich for addiction risk loci **46**

2.4.3 Convolutional Neural Network (CNN) models of mouse cell type-specific CRE activity refine human NeuN+ OCRs for GWAS enrichment **54**

2.4.4 CNN models predict allele-specific activity of addiction-associated GWAS SNPs in neuronal subtypes **60**

2.5 Discussion **66**

2.5.1 Data availability **73**

2.6 Acknowledgement: **73**

2.6.1 Author Contributions: **73**

2.6.2 Funding: **74**

2.6.3 Conflicts of Interest: **74**

3.0 Fine-mapping candidate neuropsychiatric regulatory variants using cell type-aware comparative genomics **75**

3.1 Abstract **76**

3.1.1 One-Sentence Summary **76**

3.2 Main Text **77**

3.3 Supplementary Text **86**

3.3.1 Cross-species caudate cell type ortholog identification **86**

3.3.2 Cell-TACIT performance evaluation **87**

3.3.3 Pairwise cross-species LD score regression **91**

3.3.4 Polyfun Fine-mapping **93**

3.3.5 Additional fine-mapped loci **94**

3.3.6 Limitations **96**

3.4 Materials & Methods **100**

3.4.1 Animal Care & Use **100**

3.4.2 Non-Human Primate Single Nuclei ATAC-seq **101**

3.4.3 Rhesus Macaque snATAC-seq read processing **102**

3.4.4 Reporter Assay Molecular Cloning **103**

3.4.5 AAV production and delivery **104**

3.4.6 Immunofluorescence and Microscopy **105**

3.4.7 ATAC Data Processing **106**

3.4.7.1 Human caudate snATAC-seq preprocessing **106**

3.4.7.2 Mouse caudoputamen cell types reprocessing **106**

3.4.7.3 Bulk M1 cortex and liver ATAC-seq data processing and peak ortholog mapping **107**

3.4.8 Cross-species peak ortholog mapping **107**

3.4.9 LD score regression in hg38 coordinates **109**

3.4.9.1 GWAS for neurological and psychiatric disorders **110**

3.4.10 Cell-TACIT Model training and evaluation **111**

3.4.10.1 Cell-TACIT Model evaluation **113**

3.4.10.2 Cross-species peak cell type annotations and predictions **114**

3.4.10.3 Cell-TACIT Age to prioritized fine-mapped GWAS loci **116**

3.4.11 eQTL Dataset Processing & Analyses **117**

3.4.12 Functionally informed fine-mapping **118**

3.4.13 Selection of conserved, active cell type enhancers and validation **120**

3.4.13.1 Reporter Assay Sequence Predictions **121**

3.4.13.2 Image Analyses and Statistics **121**

3.4.14 Statistical corrections for multiple comparisons **122**

3.5 Acknowledgments **123**

3.5.1 Funding **123**

3.5.2 Author contributions **123**

3.5.3 Data and materials availability **124**

3.5.4 Inclusion & Diversity **124**

4.0 Single nuclei transcriptomics of the human and monkey striatum implicates DNA damage, neuroinflammation, and neurodegeneration signaling in opioid use disorder **125**

4.1 Abstract **127**

4.2 Introduction **128**

4.3 Results **130**

4.3.1 High quality single nuclei transcriptomics identifies canonical and low abundant cell types in human dorsal striatum **130**

4.3.2 Neuronal and glial cell type-specific expression of opioid receptors in human striatum **134**

4.3.3 Identification of cell type-specific transcription factor and gene regulatory networks in human dorsal striatum **136**

4.3.4 Differentially expressed genes in specific striatal cell types associated with OUD **137**

4.3.5 Pathways in neuronal subtypes converge on neurodegenerative processes in OUD **141**

4.3.6 Enrichment of DNA damage markers in specific neuronal subtypes in OUD **143**

4.3.7 Enriched pathways among glial cells support molecular signatures related to neuroinflammation and neurodegeneration in OUD **147**

4.3.8 Sex-specific transcriptional alterations in striatal cell types associated with OUD 152

4.4 Discussion 155

4.5 Methods 160

4.5.1 Human Subjects 160

4.5.2 Isolation of nuclei from human postmortem brain tissue and library preparation: 160

4.5.3 Nonhuman Primate Subjects 161

4.5.4 Chronic Morphine Dosing in Nonhuman Primate Subjects 162

4.5.5 Rhesus brain tissue preparation 162

4.5.6 Single nucleus RNA sequencing in rhesus brain tissue 163

4.5.7 Single nuclei RNAseq data processing 163

4.5.8 Labeling striatum cells with a macaque snRNA reference dataset 165

4.5.9 Annotating interneurons with mouse marker genes 166

4.5.10 Differential gene and cellular state expression analysis in humans 167

4.5.11 Gene set enrichment analyses 169

4.5.12 Transcription factor-gene regulatory network analyses 170

4.5.13 Analyses of cellular activation of gene sets 171

4.5.14 Statistical Analyses 172

4.5.15 Data Availability 174

4.6 Acknowledgements 174

4.6.1 Funding 174

4.6.2 Author Contributions 175

5.0 Conclusion 176

Appendix A Chapter 2 Extended Data Legends 179

Appendix B Chapter 3 Supplemental Figures 180

Appendix C Chapter 3 Supplemental Data Legends..... 237

Appendix D Chapter 4 Extended Data Legend..... 240

Bibliography 243

List of Tables

Table 1. cSNAIL ATAC-seq mouse sample summary 27

List of Figures

Figure 1. A hierarchical model of biological units proposed to be conserved across species. 7

Figure 2. Shared and unique genetic architecture of genetic risk variants of addiction-associated traits. 38

Figure 3. Substance use and risky behavior GWAS risk variants enrich within reward region- and cell type-specific epigenomic profiles.....40

Figure 4. Sensitivity of stratified LDSC regression for cell type- and region-specific in the GWAS trait enrichment requires well-powered GWAS in relevant cell types....45

Figure 5. Cell type specificity of cSNAIL ATAC-seq in mouse cortex and striatum 48

Figure 6. Cell type-specific enrichment of substance use traits are conserved in mouse-human orthologous open chromatin regions. 50

Figure 7. GWAS enrichment in addiction- and non-addiction-related traits using mapped mouse orthologs of tissue- and cell type-specific open chromatin regions. 53

Figure 8. Convolutional Neural Network (CNN) model performance and selection of candidate functional SNPs. 55

Figure 9. Cell type-specific convolutional neural network (CNN) models refine human NeuN+ enrichments for substance use genetic risk GWAS. 59

Figure 10. Convolutional neural network (CNN) models for predicting cell type-specific open chromatin predict activity of addiction GWAS SNPs 61

Figure 11. Locus plots of addiction-associated SNPs predicted to act in striatal and cortical cell types. 65

Figure 12. Summary of LDSC GWAS enrichments in human and mouse-human orthologous bulk tissue and cell type open chromatin 68

Figure 13. Single nuclei RNA-sequencing of postmortem brain to identify specific cell types in human striatum 131

Figure 14. Marker genes of annotated cell types in the human dorsal striatum 133

Figure 15. Differential proportion of striatal cell types across biospecimens and OUD diagnosis 134

Figure 16. Expression patterns of striatal opioid receptor and endogenous ligands 136

Figure 17. Upset plot of overlapping differentially expressed genes in OUD 138

Figure 18. Transcriptional alterations in neuronal cell types in human striatum associated with OUD 139

Figure 20. Chronic morphine exposure and single nucleus RNA-seq of the rhesus macaque striatum 145

Figure 19. Elevated markers of DNA damage in neurons associated with OUD and chronic morphine 146

Figure 21. Transcriptional alterations in specific glial cell types in human striatum associated with OUD 151

Figure 22. Sex-biased transcriptional alterations in striatal cell types associated with OUD 154

Figure 23. Parallel cellular and molecular pathways in neurodegeneration and opioid use disorder 156

Acknowledgements

I would like to thank my advisor, Andreas R. Pfenning, for his support and guidance. Throughout my time in the joint MD/PhD and Computational Biology PhD training programs at the University of Pittsburgh (Pitt) and Carnegie Mellon University (CMU), he has worked tirelessly to transform his research program into the premier hub connecting neuroscience, genomics, and machine learning. He models being an innovative thinker, thorough researcher, nurturing mentor, and fearless leader. In providing the lab with insight and encouragement, he unites disparate fields pushing the limits to study the neurogenomics of complex behaviors. By example, he has taught me to be kind and generous to myself and others.

I would also like to thank the rest of the Pfenning lab and the funding sources including the National Institute on Drug Abuse at the NIH and the Computational Biology Department at CMU. All the work in this dissertation are collaborative efforts within and outside the Pfenning lab. Without the experimental skill and visions of Ashley R. Brown, Alyssa J. Lawler, Jing He (William R. Stauffer lab), and Madelyn H. Ray (Ryan W. Logan lab), none of this work would have been possible. I am especially grateful to Chaitanya Srinivasan, Jing He, and Madelyn H. Ray for allowing me to contribute to their research projects as parts of my dissertation. The insights of other lab members, Irene M. Kaplow, Michael Kleyman, Michael Leone, and the many mentees that I had the fortune to work with have also been an immense pleasure. I am particularly thankful for the under-mentioned but forever appreciated star of the Pfenning lab, Ashley R. Brown, whose technical knowledge is overflowing, keeps the lab running, and most critically keeps Andreas to task in his ambitious research program.

The joint MD/PhD training program and the Computational Biology PhD program at Pitt and CMU have been nurturing communities for my professional and person growth. My co-advisor, Ryan W. Logan, has always been available with thoughtful insights on the

neurobiology of addiction and insightful thoughts on career moves and perspectives. My other committee members, Kathryn Roeder and Dennis Kostka, have been encouraging advisors and provided valuable guidance along my scientific training. The two training programs themselves has been unique learning environments, thanks to the leadership of the program directors. Notably, I would like to thank my MD/PhD program career advisor who was also my department chair during the majority of my PhD, Russell S. Schwartz, and I would like to thank the director of the MD/PhD program, Richard A. Steinman, along with the organization and executive of Kathy Hansell-Prigg, Nicole E. Stenger, and Kelly M. Gentille, who are central hubs of the programs that enable my training and transitions to be smooth.

Beyond the formal training programs, I would also like to thank my first mentors at the Lieber Institute for Brain Development who are the reason I entered the MD/PhD training program: Andrew E. Jaffe who taught me that data science is a fun skill to practice that happens to be highly marketable, Brady J. Maher who fostered my interests in the neurobiology of development, and Thomas M. Hyde who took a big chance on a naïve queer kid to impart the importance of basic molecular psychiatry research for the advancement of patient care.

Outside of work, I would like to thank Ruby A. Holland, the rest of my friends, and Calvin for making my years in Pittsburgh feel like home. I would like to thank my parents for their support in every way possible of my protracted education away from home; my mother for teaching me perseverance, and my father for teaching endurance.

1.0 Substance use disorder genetics and the conserved reward system

BaDoi N. Phan[1,2]

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University, 15213
[2]Medical Scientist Training Program, School of Medicine, University of Pittsburgh, 15213

This chapter was written in part by ChaGPT version May 2023 prompted by ideas from the author and edited to the extent of author's knowledge for accuracy to convey the final message.

1.1 Substance use disorders

Substance use disorders are complex behaviors that is influenced by both genetic and environmental factors. The Diagnostic and Statistical Manual 5th edition (DSM-5), describes 11 features of substance use disorders ranging from physical symptoms (withdrawal, tolerance, cravings, relapse) to social symptoms (neglecting personal and professional responsibilities, consuming time and resources to maintain use, etc)[1]. These criteria are not specific to any substance of abuse. More importantly, these features distinguish the subset of the population with substance use disorders from the larger population who use potentially addictive substances without acquiring a use disorder.

The impacts of addiction on human lives and society are markedly clear. Over the years, the prevalence of substance use disorders in America continues to rise. Within these numbers are people who used both illicit drugs (heroin, methamphetamine) and those who misuse of prescription drugs (pain relievers, fast acting anxiolytics). The National Survey on Drug Use and Health estimates that more than 46 million Americans above 12 years old have an alcohol or drug use disorder in 2021, which is 16% of Americans in this age range[2]. The Center for

Disease Control and Prevention estimated more than 100,000 Americans died due to an overdose in 2021[3], 2.9% of all deaths, rivaling deaths due to diabetes or Alzheimer's Disease[4]. The etiologies of human addiction behaviors remain a warranted and growing area of research and many hope that that new insights would lead to strategies to effectively mitigate negative consequences of substance use disorders.

Although the problem of substance use disorders is primarily consequence of human social engineering, many groups have demonstrated that other model species exhibit addictive behaviors to substances of abuse. Certain animal models, mostly rats, have been used to explore aspects of substance use behaviors in operant and non-operant paradigms. Some behavioral models of addiction mirror the DSM-5 criteria for substance use disorders, such as measures of drug seeking to model cravings or substance deprivation and re-exposure to model relapse[5–7]. Detractors of animal models for the study of addiction have noted that certain drugs and certain species cannot acquire the range of drug use behaviors limiting translatability of the human conditions[8]. While I do not wish to speculate on the validity of behavior-based animal models of substance use disorders, I propose to explore other ways to model addiction in non-human organisms: through the genetics factors of addiction and their impacts on conserved genomic mechanisms shared between species.

1.2 Genetics of Substance Use Disorders

To date, many studies through multiple means have identified that genetics contribute to risk for substance use disorder. Twin and familial-based heritability studies of substance use disorders have noted moderate to high degrees of genetic influences (30-70%) in substance use disorder replicated across multiple twin cohorts and substances of abuse[9,10]. A subset of genetic heritability for substance use disorders can be measured by common single nucleotide

polymorphisms across the human genome in genome-wide association studies (GWAS). GWAS profiling in thousands to millions of individuals with substance use behaviors and substance use disorders, largely of white European ancestry, have identified numerous genomic loci that contribute to genetic risk for substance use disorders[11,12]. Notably, some studies have shown that the genetic risk for substance use disorders may be distinct from genetic risk for non-disorder substance use behaviors[13–16], furthering the observation that not all individuals who use substances acquire a use disorder[17]. One recent study even showed that the heritability of substance use disorders lie on a spectrum that interacts with increasing environmental risk factors connecting the two etiologies of addiction comes[18]. Further meta analyses have combined GWAS studies across multiple substance use disorders and related traits to show that there exists a common genetic factor underlying addiction genetic risk[19,20]. A summary of these genetic studies alongside large bodies of neurobiological and neuroimaging findings have contributed to a working Genetically Informed Neurobiology of Addiction model where addiction is "a developmental process that comprise of escalating problem use"[21]. At the core of this model are polygenic risk factors for substances of abuse which are acted upon by environmental factors (stress, substance, socioeconomics) that contribute to brain (dys)function of nigrostriatal, prefrontal, and limbic circuits mapping onto binge-intoxication, preoccupation or craving, and negative affect/withdrawal responses, respectively[21]. The authors of this model acknowledge that while the links exist between genetics, neurobiology, and neuroimaging, deep chasms remain that divide these fields warranting further research to bridge the gaps between addiction genetics and addiction biology.

1.3 Single cell molecular biology to connect the genetics to addiction behavior

Of the ~40,000 genes within a mammalian genome, a smaller combination of genes is used in each cell type's function and determine its identity[22]. Similarly, only a minority of the ~3 billion nucleotides of the genome are in use by any cell type[23]—some to encode the transcribed genes and most in open chromatin. Open chromatin refers to the accessible regions of DNA within the genome that are not tightly packaged or condensed into a compact structure. These open chromatin regions allow regulatory elements, such as transcription factors and other proteins, to bind and interact with the DNA, thereby influencing gene expression[24,25]. The tucked away DNA are thought to be used by other cells, tissues, or contexts[23].

Single-cell transcriptomics sequencing (scRNA-seq) and single-cell epigenomics sequencing, including the open chromatin assays (scATAC-seq), have emerged as powerful tools to understand the regulatory mechanisms underlying the diversity of cell types within complex tissues such as the brain[26,27]. Single cell transcriptomics enables us to examine gene expression profiles within individual cells, providing insights into the diversity and heterogeneity of cell populations within the brain[26]. Single cell epigenomics profiles enable identification of which genomic loci contribute to cellular heterogeneity of an organism from one genome[24]. By applying single cell approaches to dissect the component parts of complex tissues, we can identify specific cell types and molecular pathways that are differentially affected by genetic variants of substance use disorders. In the context of substance use disorder genetic variants, the differences in DNA sequence between individuals may influence how those regions of the DNA are able to regulate genome usage to affect the cellular responses within the reward circuitry and other brain regions involved in addiction. Ongoing research efforts already being to profile how genetic variation in the general human populations relate to tissue-specific[28–30] and cell type-specific gene transcription variation[31,32]. Through scRNA-

seq, researchers can compare which genes may be differentially primed to respond between individuals with high and low gene risk profiles, as well as between different stages of addiction. This approach helps to identify differentially expressed genes and regulatory elements associated with addiction, including those involved in neuronal signaling, synaptic plasticity, reward processing, and stress response. Through scATAC-seq, we may be able to identify which genetic variants may lead the functional effects contributing to transcriptional variation in their association with complex behaviors. By integrating epigenomics with transcriptomic profiles we maybe be we may be able to link common genetic variation for substance use disorders to regulation of gene expression associated with the stages of addiction.

Outside of the applications in human genetics, others have applied cell type-specific multi-omics methods to study the molecular consequences in various stages of addiction in postmortem human tissues human and model organisms. These studies more directly measure the transcriptional and epigenetic changes associated with various stages of addiction to identify the specific genes and regulatory elements involved in addiction across different cell types and tissues related to the neural circuits of reward[33–35]. While both these multi-omics and cell type-resolution studies can be applied to understand both human genetics of addiction and consequences of addiction at fine detail, the translatability of these studies lead to the bigger question—how do genetic variation alter cellular function to produce different behaviors?

Overall, the application of single cell approaches in the study of substance use disorders enables a comprehensive understanding of the molecular and cellular mechanisms underlying addiction, a behavior that seem to be shared across mammalian species and beyond. They help to identify key regulatory elements, genes, and cell types involved in addiction, and many people hope these methods will lead to the development of therapies and interventions for patients before, during, and after substance abuse and acquisition of substance use disorders.

1.4 Conservation of the genome and implications for biomedical research

Translating findings from large genome-wide association studies to knowledge of addiction neurobiology and actionable targets for substance use disorders face large barriers. First, narrowing down which neurobiological units in model species are targeted by genetic risk variants remains elusive and complex considering the diversity of cell types and contexts that comprise the neural circuits of reward. The genetic variants associated with complex human traits fall within non-coding regions of the genome and a large number due to linkage disequilibrium[36]—resulting in significantly associated variants that are not expected to carry functional contributions to genetic risk for the trait. Complex traits continue to be demonstrated to be polygenic, or even omnigenic[37], at common genetic variants, so attempts to aggregate genome-wide significant variants to protein coding genes then manipulating the gene orthologs in other species are inherently flawed. Finally, while genetic association studies may identify genetic risk factors for substance use disorders, curbing human suffering due to addiction will require development of effective preventative and therapeutic measures.

We approach the challenge of translating the human genetic findings of substance use disorders through the lens of evolution[38]. Along the thought attributed to George E. P. Box that "all models are wrong, but some models are useful", discovering which parts of the model organisms' biological systems are conserved despite millennia of divergence from human lessens the concerns that studying human conditions in model organisms is wrong toward useful. Conveniently, the neural circuits of reward are well conserved in mammals. From a neuroanatomical stance, many have observed that the brain regions and connectivity of the reward areas have orthologs in primates[39], rodents[40], and invertibrates[41,42]. We extend these observations claiming that the components of the reward system are also conserved across mammals from cell types down to the genome (**Figure 1**).

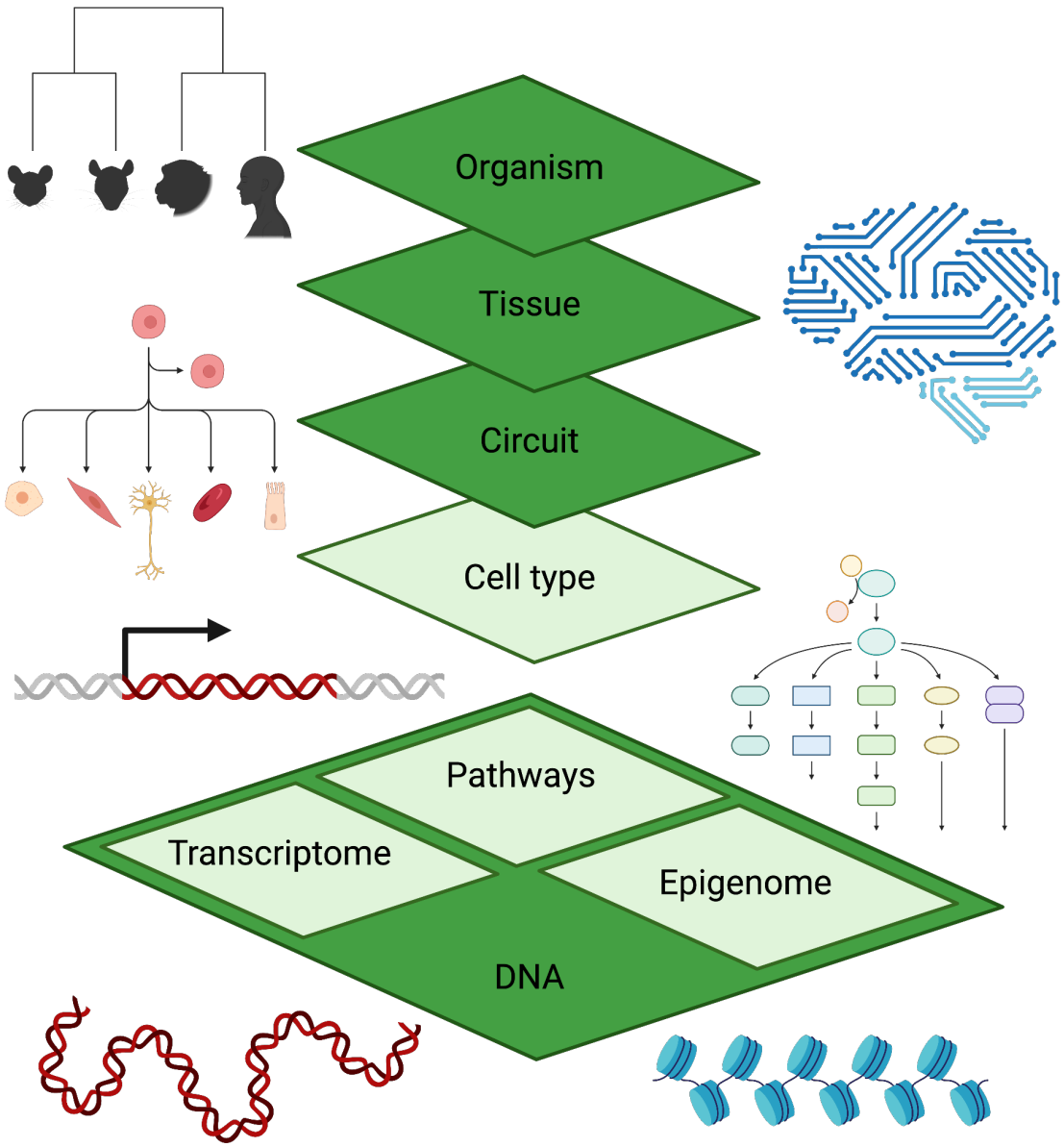


Figure 1. A hierarchical model of biological units that may be conserved across species. *Each layer is represented by icons along the margins. From the top left in clockwise order: hierarchy of mouse, rat, monkey, and human for "Organism"; a graphic brain circuits for "Tissue" and "Circuit"; a signaling cascade for "Pathways"; DNA in nucleosomes for "Epigenome"; DNA for "DNA"; a graphic of a transcribed gene for "Transcriptome"; a hierarchy of cell types for "Cell types". Elements in green have evidence conservation across mammalians. The elements in light green represent the components proposed to be conserved across mammals that remain to be investigated in this dissertation.*

At the level of DNA, comparative genomics efforts have shown that the human genome is undeniably conserved from nucleotide sequence[43] to whole chromosomes[44] compared to closely- and distantly-related mammals. At least 10.7% of the human genome is conserved across 240 mammalian species[43] and even more of the human genome have mappable orthologs to the genomes of rodents and primates, 49% and 85% respectively[45,46]. Genome sequence conservation indicate that the functions of a genome would also be conserved—the genes and gene regulation. A comparison of human and mouse functional DNA elements show a majority are shared[47] and under genomic constraint[48]. Considering human genetic risk variants lie in the noncoding genome, a comparative approach to how the genome maintains similar functions across these species, perhaps at the level of cell type-resolution with epigenomics or transcriptomics, may bridge the gap between directly modeling the human genetics of complex traits within the conserved genomic loci of other model organisms.

The mammalian genome may have different levels of function, as illustrated in **Figure 1**, that can apply evolutionary pressures to remain conserved across hundreds of millions of years. Because an organism's genome is the scaffold upon which transcriptomes, epigenomics, molecular pathways, and cellular function are derived, we propose to measure where in the genome these functions are conserved as well as their contexts to inform genomic modeling of human genetic risk. The measurable conservation of the mammalian genome can be broadly described by what advances in molecular biology can measure. Here is where cell type-specific data, such scRNA-seq or scATAC-seq, from orthologous tissues of multiple species can provide valuable insights on what is conserved from these genomes and what is not.

At the level of genes, many ubiquitously transcribed genes are conserved due to their critical functions the survival of the cells and ultimately the species. They typically represent essential cellular functions and housekeeping genes that are required for basic cellular processes. Other genes have variable expression patterns, cell-type-specific genes and genes

associated with cellular states or phenotypes. Recent multi-species comparative single cell transcriptomics studies have also identified that these identity-defining are also conserved in their expression patterns across species[49–52] and can be exploited algorithmically to use whole cell gene expression patterns to match orthologous cell types between species together[53–55]. These second group of conserved genes, where cellular contexts are unique between tissues and cell types and these contexts are conserved between species, are key to understanding how perturbation of these genes between individuals of a species may have different disease risks. Interestingly, the genes whose nucleotide sequences are most conserved across mammals seem to contribute to neural or embryonic tissues[43] while those that are the least conserved between mammals relate to innate immunity or are the barriers for a species to the environment. The conservation of many context-specific genes between humans and rodents are largely known and appreciated by the biomedical research community springing forth decades of effort to manipulate conserved genes in whole animals or restricted to certain contexts to determine function and dysfunction. With some exceptions[56], conserved genes have contributed to many successes in biomedical research—an approach that continues to be widely popular to model and study the genetic bases of human disease[57].

3. Epigenetic Variability: Epigenetic modifications, including DNA methylation and histone modifications, play a crucial role in regulating gene expression and maintaining cellular identity. The epigenetic landscape can exhibit variability between cells, reflecting differential gene regulation and cellular plasticity. scRNA-seq can be combined with epigenomic profiling techniques to investigate the epigenetic variability across different cell types or states within a population. This integration helps understand the impact of epigenetic modifications on gene expression patterns and cellular heterogeneity.

Within an organism, different cells may use the genome in complex ways to affect its function. Between organisms, many tissues and cell types have described by anatomists and physiologists to have orthologs across with humans. Usually, anatomic ortholog of tissues and cell types is taken for granted to mean conserved function; however, a prudent approach defining orthologs would measure how function at the levels of transcriptomes, epigenomes, and molecular pathways that derive from the genomes would be conserved noting along the way the notable divergence.

By analyzing single-cell data, researchers can gain a comprehensive understanding of the stages of conservation within the genome. This information helps identify the core cellular functions, cell-type-specific genes, dynamic gene expression patterns, and epigenetic variability that contribute to cellular diversity, plasticity, and overall tissue function. Furthermore, studying the conservation of the genome at the single-cell level can shed light on disease mechanisms, development, and tissue regeneration processes.

The chromatin structure consists of DNA wrapped around proteins called histones, forming a complex known as nucleosomes. The tight packaging of DNA around nucleosomes can restrict the access of transcriptional machinery to genes, thereby repressing their expression. However, certain regions of the genome remain relatively open, allowing regulatory elements to access and modulate gene expression.

Open chromatin regions are typically associated with active gene transcription and regulatory elements, such as promoters, enhancers, and other regulatory sequences. These regions are characterized by specific molecular features, including DNA accessibility, reduced nucleosome occupancy, and specific histone modifications. Accessibility refers to the ease with which proteins can interact with the DNA, influencing gene expression. Reduced nucleosome occupancy means that these regions have fewer nucleosomes, making the DNA more

accessible. Additionally, specific histone modifications, such as histone acetylation and methylation, can mark open chromatin regions and promote gene activation.

The identification and mapping of open chromatin regions have been facilitated by experimental techniques like DNase-seq, ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing), and MNase-seq (Micrococcal Nuclease sequencing). These methods can assess DNA accessibility and provide a genome-wide view of open chromatin regions.

The study of open chromatin is crucial for understanding gene regulation and cellular processes. By mapping these regions, researchers can identify potential regulatory elements, enhancers, and other non-coding regions that play important roles in gene expression and cellular function. It also provides insights into the dynamic nature of gene regulation and can aid in understanding disease mechanisms, developmental processes, and cell type-specific gene expression programs.

Overall, the concept of open chromatin highlights the importance of DNA accessibility and the role it plays in modulating gene expression and cellular function. The mapping of open chromatin regions using techniques like ATAC-seq and DNase-seq has significantly contributed to our understanding of gene regulation and provides a foundation for further investigation into various biological processes.

Studying open chromatin of the genome between multiple species presents several challenges and limitations due to the inherent differences in genomic organization, regulatory elements, and evolutionary divergence. Some of the key limitations include:

1. Genomic Differences: Different species have variations in their genome sizes, gene content, and genomic architecture. These variations can impact the identification and comparison of open chromatin regions between species. For example, the presence of repetitive

DNA elements or structural variations in one species may complicate the identification and alignment of homologous open chromatin regions in another species.

2. Regulatory Element Divergence: Regulatory elements, such as promoters and enhancers, can exhibit significant divergence between species. This divergence can affect the accessibility and functionality of open chromatin regions. Regulatory elements that are conserved across species may be more easily comparable, but species-specific regulatory elements may have distinct functions and may not be readily identified using standard comparative approaches.

3. Limited Reference Genomes: In some species, particularly non-model organisms or those with complex genomes, reference genomes may be incomplete or unavailable. This can hinder the identification and mapping of open chromatin regions in these species, as the absence of a comprehensive reference genome makes it challenging to accurately align and annotate the regulatory elements.

4. Technical Variability: Experimental techniques used to study open chromatin, such as ATAC-seq or DNase-seq, may have inherent technical variability, which can affect the reproducibility and comparability of data between species. Differences in sample preparation, sequencing technologies, and data analysis pipelines can introduce biases and make direct comparisons between species challenging.

5. Evolutionary Dynamics: The evolutionary dynamics of regulatory elements can lead to changes in open chromatin regions even between closely related species. Evolutionary events such as gene duplications, gene losses, and genomic rearrangements can alter the regulatory landscape and make it difficult to directly compare open chromatin profiles.

6. Functional Conservation: Although open chromatin regions are associated with gene regulation, the functional consequences of these regions can vary between species. Different species may have distinct regulatory mechanisms and gene regulatory networks, making it

challenging to directly infer functional similarities or differences based solely on open chromatin profiles.

Despite these limitations, comparative studies of open chromatin between species can still provide valuable insights into the evolution and conservation of regulatory elements. By leveraging the available genomic data and employing sophisticated comparative genomics approaches, researchers can gain a deeper understanding of the similarities and differences in open chromatin landscapes across multiple species. Such comparative studies can uncover conserved regulatory elements and shed light on the evolutionary processes shaping gene regulation.

Comparing open chromatin in the human genome to open chromatin in other mammals involves several considerations and approaches. Here are some key methods and considerations for such comparisons:

1. Comparative Genomics: Comparative genomics involves comparing genomic sequences and annotations across different species. By aligning and comparing the genomes of humans and other mammals, researchers can identify regions of sequence conservation. These conserved regions may include open chromatin regions associated with regulatory elements. Comparative genomics can provide insights into shared regulatory elements and functional conservation across species.

2. Orthologous Regions: Identifying orthologous regions, which are corresponding genomic regions between different species that have evolved from a common ancestral sequence, is crucial for comparing open chromatin. Orthology prediction methods can help determine which regions of the genomes in different mammals are likely to correspond to each other. By comparing open chromatin profiles in these orthologous regions, researchers can assess the conservation of chromatin accessibility and regulatory potential.

3. Cross-Species Chromatin Accessibility Assays: Experimental techniques such as ATAC-seq or DNase-seq can be applied to multiple mammalian species to generate chromatin accessibility profiles. By using a consistent assay across different species, researchers can directly compare open chromatin regions and identify similarities or differences. This approach allows for the identification of conserved regulatory elements and species-specific variations in chromatin accessibility.

4. Functional Validation: Comparisons of open chromatin regions can be complemented by functional validation experiments. For instance, regulatory elements identified in the human genome can be tested for their functionality in other mammals through reporter assays or transgenic animal models. These functional experiments provide evidence for the conservation or divergence of regulatory elements across species.

5. Phylogenetic Analysis: Phylogenetic analysis can help infer the evolutionary relationships between species and provide insights into the timing and mechanisms of chromatin accessibility changes. By considering the phylogenetic distances between species, researchers can gain a better understanding of the conservation or divergence of open chromatin regions over evolutionary time.

It is important to note that while these approaches can provide valuable information on open chromatin conservation across mammals, there can also be species-specific differences due to evolutionary divergence, differences in regulatory mechanisms, or unique biological features. Therefore, careful interpretation and integration of comparative genomics, experimental data, and functional studies are necessary to gain a comprehensive understanding of open chromatin dynamics and regulatory evolution between human and other mammalian genomes.

2.0 Addiction-associated genetic variants implicate brain cell type- and region-specific *cis*-regulatory elements in addiction neurobiology

Chaitanya Srinivasan[1*], BaDoi N. Phan[1,2*], Alyssa J. Lawler[3], Easwaran Ramamurthy[1], Michael Kleyman[1], Ashley R. Brown[1], Irene M. Kaplow[1,4], Morgan E. Wirthlin[1,4], Andreas R. Pfenning[1,3,4,‡]

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University, 15213
[2]Medical Scientist Training Program, School of Medicine, University of Pittsburgh, 15213
[3]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, 15213
[4]Neuroscience Institute, Carnegie Mellon University, 15213
[*]CS and BNP contributed equally to the work
[‡]Direct correspondence to apfenning@cmu.edu

This chapter is reproduced, in full, with permission as an author from research published in the Journal of Neuroscience under copyright license (CC-BY) Creative Commons Attribution 4.0 International License.

SRINIVASAN C, PHAN BN, LAWLER AJ, RAMAMURTHY E, KLEYMAN M, BROWN AR, KAPLOW IM, WIRTHLIN ME, PFENNING AR. **ADDICTION-ASSOCIATED GENETIC VARIANTS IMPLICATE BRAIN CELL TYPE- AND REGION-SPECIFIC CIS-REGULATORY ELEMENTS IN ADDICTION NEUROBIOLOGY.** J NEUROSCI. 2021 OCT 27;41(43):9008-9030. DOI: 10.1523/JNEUROSCI.2534-20.2021. EPUB 2021 AUG 30. PMID: 34462306; PMCID: PMC8549541.

2.1 Abstract

Recent large genome-wide association studies (GWAS) have identified multiple confident risk loci linked to addiction-associated behavioral traits. Most genetic variants linked to addiction-associated traits lie in non-coding regions of the genome, likely disrupting cis-regulatory element (CRE) function. CREs tend to be highly cell type-specific and may contribute to the functional development of the neural circuits underlying addiction. Yet, a systematic approach for predicting the impact of risk variants on the CREs of specific cell populations is lacking. To dissect the cell types and brain regions underlying addiction-associated traits, we applied stratified LD score regression to compare GWAS to genomic regions collected from human and mouse assays for open chromatin, which is associated with CRE activity. We found enrichment of addiction-associated variants in putative CREs marked by open chromatin in neuronal (NeuN+) nuclei collected from multiple prefrontal cortical areas and striatal regions known to play major roles in reward and addiction. To further dissect the cell type-specific basis of addiction-associated traits, we also identified enrichments in human orthologs of open chromatin regions of female and male mouse neuronal subtypes: cortical excitatory, D1, D2, and PV. Lastly, we developed machine learning models to predict mouse cell type-specific open chromatin, enabling us to further categorize human NeuN+ open chromatin regions into cortical excitatory or striatal D1 and D2 neurons and predict the functional impact of addiction-associated genetic variants. Our results suggest that different neuronal subtypes within the reward system play distinct roles in the variety of traits that contribute to addiction.

2.1.1 Significance Statement:

We combine statistical genetic and machine learning techniques to find that the predisposition to for nicotine, alcohol, and cannabis use behaviors can be partially explained

by genetic variants in conserved regulatory elements within specific brain regions and neuronal subtypes of the reward system. Our computational framework can flexibly integrate open chromatin data across species to screen for putative causal variants in a cell type- and tissue-specific manner for numerous complex traits.

2.2 Introduction

Substance use disorders (SUD) have increased in prevalence over the last three decades, with an estimated 100 million cases worldwide[58,59]. Pharmacological interventions are limited in their ability to cure addiction due to physiological and logistical barriers[60,61]. As the societal epidemic of substance use grows, there is a greater need to understand the neurobiology of substance use behaviors and addiction.

The reward circuits co-opted in addiction as well as the associated neural cell types are highly conserved across primates and rodents[41,62]. Many studies have shown that addictive substances promote impulsive and compulsive behavior by activating the mesolimbic dopamine system, in which dopaminergic inputs from the ventral tegmental area project to medium spiny neurons (MSN) of the nucleus accumbens (NAc) in the ventral striatum (STR)[63]. Glutamatergic inputs to the NAc from the amygdala, frontal cortex, and hippocampus contribute to behavioral motivation through the extrapyramidal motor system[63]. Subsequently, the NAc sends outputs to nuclei of the ventral pallidum, which are critical for processing and modulating substance reward signal[63]. The development of compulsive substance-seeking is hypothesized to be linked to recruitment of the dorsal STR, which together with the prefrontal cortical regions regulates a variety of reward- and addiction-related phenotypes[63,64]. These findings emphasize that substance abuse behavior involves the interplay of the brain regions and cell types that make up the reward system.

Increasing evidence reveals strong genetic links to substance use risk[65–68] and SUD[69–72]. Genome-wide association studies (GWAS) report that genetic risk for substance use shares underlying architecture with other neuropsychiatric disorders[65,67]. These risk variants tend to lie in non-coding regions of the human genome rather than the protein-coding regions that have clearer functional consequence to link genes to traits[73]. These genetic variants, including single nucleotide polymorphisms (SNPs), can disrupt transcription factor binding in cis-regulatory elements (CREs) with varying impact on gene regulation and downstream neural circuitry. Many CREs have tissue- and cell type-specific activity[23], suggesting that cell types and tissues underlying addiction may be uniquely targeted by genetic variants at these CREs. GWAS for nicotine-, alcohol-[67], and cannabis-use traits[65] have identified multiple confident risk loci and SNPs linked to addiction-associated phenotypes with brain-specificity, yet their effects on the CREs of specific brain regions and cell types involved in addiction pathophysiology are an open area of inquiry.

A comparison of GWAS to functional annotations of the human genome have yielded estimates that over 90% of SNPs associated with complex phenotypes lie within potentially functional non-coding regions, which are marked by epigenetic features including open chromatin[74,75]. Linkage disequilibrium (LD) of significant SNPs complicates the identification of causal variants contributing to genetic risk[36], as multiple SNPs that are tightly linked will inherently all have similar association with a disorder, even if not all of them are necessarily causal. Regression of SNP LD scores against GWAS summary statistics (LDSC regression) is the dominant method for relating human genetics to functional annotations. LDSC regression partitions risk SNPs identified by GWAS into the tissues or cell types in which genetic variation in CREs may contribute to heritability of complex traits[74,76]. Yet, the functional consequences of risk SNPs in CRE sequences cannot be reliably inferred from DNA sequences alone[77]. By synthesizing recent developments in epigenomic assays[78–80] and machine learning[81–85], it is

possible to predict cell types affected by addiction-associated genetic variation and propose cell type-specific hypotheses on the pathogenesis of addiction.

Here, we implement a framework that links the genetic predisposition to addiction-associated traits to specific brain regions and cell types within them by identifying those that have open chromatin regions that are enriched for SNPs identified by GWAS. We first intersect SNPs measured by GWAS with human and mouse bulk tissue- and cell type-specific open chromatin regions to identify putative region- and cell type-specific CREs that may be impacted by genetic variation associated with addiction-related traits. To overcome limits of cellular resolution in the human brain, we apply convolutional neural network models trained on transgenically labelled neuron populations in the reward system of mice to predict the cell type-specificity of GWAS-associated SNPs in the human genome. We further apply these models to the problem of screening for putative causal SNPs within dense loci reported in GWAS for addiction-associated traits. This pipeline, to our knowledge, describes the first integrative analyses across species, brain regions, and cell types to screen for candidate causal addiction-associated genetic risk variants in dense loci with numerous significant SNPs in LD.

2.3 Materials and Methods

2.3.1 ATAC-seq data processing pipeline

We processed raw FASTQ files of ATAC-seq experiments with the official ENCODE ATAC-seq pipeline[86] accessed by https://github.com/ENCODE-DCC/atac-seq-pipeline. We ran this pipeline using the `mm10` genome assembly for mouse and the `hg38` genome for human with the following settings: `smooth_win = 150`, `multimapping = 0`, `idr_thresh = 0.1`, `cap_num_peak = 300,000`,

`keep_irregular_chr_in_bfilt_peak = true`. We grouped biological replicates (e.g. samples from the same tissue region or condition) when processing data to obtain individual de-duplicated, filtered bam files. We removed samples that had low periodicity indicated by ENCODE quality control metrics (https://www.encodeproject.org/atac-seq/) and reprocessed the remaining replicates with the pipeline. Using the high-quality replicates, we obtained reproducible (IDR) peaks for each condition. Unless otherwise stated, we used the "optimal" reproducible set of peaks for downstream analyses.

2.3.1.1 Publicly available datasets

Fullard *et al.* NeuN-sorted ATAC-seq of human postmortem brain[87]:

We identified OCRs overlapping addiction-related variants through analysis of human postmortem brain ATAC-seq in which cells were sorted into NeuN-positive and NeuN-negative groups via fluorescence activated nuclei sorting (FANS); the brain regions we used were dorsolateral prefrontal cortex (DLPFC), orbitofrontal cortex (OFC), ventrolateral prefrontal cortex (VLPFC), anterior cingulate cortex (ACC), superior temporal gyrus (STC), inferior temporal gyrus (ITC), primary motor cortex (PMC), insula (INS), primary visual cortex (PVC), amygdala (AMY), hippocampus (HIP), mediodorsal thalamus (MDT), nucleus accumbens (NAc), and putamen (PUT). We downloaded data from the Sequence Read Archive (SRA) through Gene Expression Omnibus (GEO) accession #GSE96949. We separated samples by cell type and reprocessed them with the ENCODE pipeline as detailed above, aligning reads to `hg38`. All datasets were high-quality according to the ENCODE metrics for epigenomic datasets (https://www.encodeproject.org/atac-seq/). We used the "optimal reproducible peaks" for each cell type and brain region as foregrounds in GWAS LDSC enrichment with the Honeybadger2 OCR set as the background set (see LDSC Regression GWAS Enrichment Backgrounds).

Corces *et al*. single cell chromatin accessibility profiling[88]:

We downloaded 24 clusters of IDR peaks of human isocortex, striatum, hippocampus, and substantia nigra in BED format through GEO accession #GSE147672. These clusters represent cell populations defined by Corces *et al.* from the measured brain regions. We assigned clusters to cell populations as described in Corces *et al.*: astrocyte (clusters 13, 17), hippocampal excitatory (clusters 3-4), isocortical astrocyte (cluster 15), isocortical excitatory (cluster 1), isocortical inhibitory (cluster 11), microglia (cluster 24), neuron (cluster 7), nigral astrocyte (cluster 14), nigral neurons (clusters 5-6), nigral oligodendrocyte precursor (cluster 10), oligodendrocyte (clusters 19-23), oligodendrocyte precursor (clusters 8-9), striatal astrocyte (cluster 16), and striatal inhibitory cells (clusters 2, 12). We did not include cluster 18, which corresponds to a doublet. We merged coordinates from clusters assigned to the same cell types to define foreground sets for LDSC regression GWAS enrichment. We merged the foreground sets with the Honeybadger2 OCR set to define the background set (LDSC regression GWAS Enrichment Backgrounds).

Lake *et al.* human occipital cortex scTHS-seq[89]:

We downloaded BED-formatted cell type-specific differential OCRs from occipital cortex scTHS-seq of excitatory neurons (EXC), inhibitory neurons (IN), astrocytes (AST), endothelial cells (END), oligodendrocyte precursor cells (OPC), oligodendrocytes (OLI), and microglia (MIC) from the GEO subseries #GSE97887. We used the `hg38` OCR coordinates as foregrounds in LDSC regression GWAS enrichment with the Honeybadger2 OCR set as the background set (LDSC regression GWAS Enrichment Backgrounds).

Mo *et al.* mouse INTACT-sorted nuclei ATAC-seq[80]:

We downloaded FASTQ files of *R26-CAG-LSL-Sun1-sfGFP-Myc* transgenic mouse lines for cell type-specific ATAC-seq performed using the INTACT method from the accession #GSE63137. *Mo et al.* isolated INTACT-enriched nuclei from three cell types: excitatory neurons (EXC, *Camk2a-cre*), vasoactive intestinal peptide neurons (VIP, *Vip-cre*), and

parvalbumin neurons (PV, *Pvalb-cre*). We reprocessed the data with the Kundaje Lab open chromatin pipeline using the mm10 genome (`https://github.com/kundajelab/atac_dnase_pipelines`). We mapped reproducible mouse ATAC-seq peaks for each cell type to `hg38` using halLiftover with the 12-mammals Cactus alignment[90,91] followed by HALPER[92] (Mapping mouse OCR orthologs) to produce a foreground set of orthologous human sequences for LDSC regression GWAS enrichment[93]. We mapped the ENCODE `mm10` DNaseI-hypersensitive peak set[94] to `hg38` (Mapping mouse OCR orthologs) and used successfully mapped hg38 orthologs of mm10 OCRs a background set for mouse foreground enrichments. Furthermore, we used this dataset to evaluate differential accessibility in cSNAIL-INTACT PV and PV-negative ATAC-seq samples and develop convolutional neural network models of cell type-specific open chromatin (see Methods below).

Human negative control foregrounds[95–98]:

We downloaded raw ATAC-seq profiles of human adult female and male stomach ATAC-seq generated by Snyder *et al*. (ENCSR337UIU, ENCSR851SBY, respectively), female human embryonic liver DNase-seq generated by Stamatoyannopoulos *et al*. (ENCSR562FNN), and human embryonic lung DNase-seq generated by Stamatoyannopoulos *et al*. (ENCSR582IPV) from `https://www.encodeproject.org/`. We processed these files using the ENCODE pipeline as detailed above to obtain optimal reproducible `hg38` peaks. We also downloaded BED files of human adipocyte and preadipocyte ATAC-seq profiles generated by Cannon *et al*. from GEO accession number #GSE110734. We mapped these BED coordinates from `hg19` to `hg38` using liftOver to define negative control foregrounds for human LDSC regression GWAS enrichment. We merged the human negative control foregrounds and Fullard *et al.* foregrounds with the Honeybadger2 OCR set to define the background for human negative control foreground enrichments.

Human-orthologous negative control foregrounds[99]:

We also downloaded raw ATAC-seq data profiled in female mouse kidney, female mouse liver, and male mouse lung generated by Liu *et al.* from SRA accession #SRP167062 to define human-orthologous negative control foregrounds. We processed these files using the ENCODE pipeline as detailed above to get optimal reproducible peaks. We mapped optimal reproducible peaks from `mm10` to `hg38` using halLiftover with the 12-mammals Cactus alignment followed by HALPER (Mapping mouse OCR orthologs) to define negative control foregrounds for human-orthologous LDSC GWAS enrichments. We merged all human orthologous foregrounds with the human orthologs of the ENCODE `mm10` DNaseI-hypersensitive peak set to define a background for human-orthologous LDSC GWAS enrichments.

2.3.2 Mapping mouse open chromatin region (OCR) orthologs

We employed halLiftover[91] with the 12-mammals Cactus alignment[90] followed by HALPER (https://github.com/pfenninglab/halLiftover-postprocessing)[92] to map `mm10` mouse reproducible OCRs to `hg38` human orthologs in order to perform LDSC regression GWAS enrichment. The Cactus multiple sequence alignment file[90] has 12 genomes, including `mm10` and `hg38`, aligned in a reference-free manner, allowing us to leverage multi-species alignments to confidently identify orthologous regions across species. halLiftover uses a Cactus-format multiple species alignment to map BED coordinates from a query species to orthologous coordinates of a target species, and HALPER constructs contiguous orthologs from the outputs of halLiftover[92]. We ran the `orthologFind.py` function from HALPER on the outputs of halLiftover using the following parameters: `-max_frac 5.0 -min_frac 0.05 -protect_dist 5 -narrowPeak -mult_keepone`. In general, 70% of mouse brain ATAC-seq reproducible

peaks were able to be mapped to confident human orthologs. To map the ENCODE `mm10` mouse DHS background, which does not contain summit information, to `hg38` we used the mouse coordinates of position with the most species aligned in a region to define the summit. Only for the `mm10` mouse DHS background set, for which a significant proportion of regions could not be confidently mapped to `hg38`, we flanked the original assembly coordinates by 300 bp to increase OCR mapping from 54% to 64%.

2.3.3 Experimental Design

To augment and compare to mouse cell type-specific ATAC-seq datasets generated in this study, we performed bulk tissue ATAC-seq from cortex (CTX) and dorsal striatum/nucleus accumbens (CPU) of 7- and 12-week-old C57Bl/6J mice (N = 2 each age) from both sexes **(**

Table 1**)** as described in Buenrostro *et al.*, 2015[100] with the following minor differences in buffers and reagents. We euthanized mice with isoflurane, rapidly decapitated to extract the brain, and sectioned it in ice-cold oxygenated aCSF (119mM NaCl, 2.5 mM KCl, 1mM NaH_2PO_4(monobasic), 26.2mM $NaHCO_3$, 11mM glucose) at 200-micron sections on a vibratome (Leica VT1200). We further micro-dissected sections for cortex and dorsal striatum on a stereo microscope and transferred dissected regions into chilled lysis buffer[100]. We dounce homogenized the dissected brains in 5mL of lysis buffer with the loose pestle (pestle A) in a 15mL glass dounce homogenizer (Pyrex #7722-15). We washed nuclei lysate off the pestle with 5mL of lysis buffer and filtered the nuclei through a 70-micron cell strainer into a 50mL conical tube. We washed the dounce homogenizer again with 10mL of BL buffer and transferred the lysate through the 70-micron filter (Foxx 1170C02). We pelleted the 20 mL of nuclei lysate at 2,000 x *g* for 10 minutes in a refrigerated centrifuge at 4°C. We discarded the supernatant and resuspended the nuclei in 100-300 microliters of water to approximate a concentration of 1-2 million nuclei/ mL. We filtered the nuclei suspension through a 40-micron

cell strainer. We stained a sample of nuclei with DAPI (Invitrogen #D1206) and counted the sample to measure 50k nuclei per ATAC-seq transposition reaction. The remaining steps follow the Buenrostro *et al.*, 2015[100] protocol for tagmentation and library amplification. We shallowly sequenced barcoded ATAC-seq libraries at 1-5 million reads per sample on an Illumina MiSeq and processed individual samples through the ENCODE pipeline for initial quality control. We used these QC measures (clear periodicity, library complexity, and minimal bottlenecking) to filter out low-quality samples and re-pooled a balanced library for paired-end deep sequencing on an Illumina NextSeq to target 30 million uniquely mapped fragments per sample after mitochondrial DNA and PCR duplicate removal. These raw sequencing files entered processing through the ENCODE ATAC-seq pipeline as above by merging technical replicates and grouping biological replicates by brain region for each pipeline run.

The cSNAIL genome (pAAV-Ef1a-DIO-Sun1-Gfp-WPRE-pA) contains *loxP* sites to invert the *Sun1-Gfp* fusion gene and integrate into the nuclear membrane of cells expressing the *Cre* gene, allowing these cell populations to be profiled for various genomic assays (Lawler et al., 2020 in press J. Neuro). We packaged the cSNAIL genome with AAV variant PHP.eB (pUCmini-iCAP-PHP.eB) in AAVpro(R) 293T cells (Takara, cat #632273). Viviana Gradinaru provided us with the pUCmini-iCAP-PHP.eB (http://n2t.net/addgene:103005; RRID: Addgene 103005)[101]. We precipitated viral particles with polyethylene glycol, isolated with ultracentrifugation on an iodixanol density gradient, and purified in PBS with centrifugation washes and 0.2µM syringe filtration. We injected each mouse with 4.0 x 10^{11}vg into the retro-orbital cavity under isoflurane anesthesia. We allowed the virus to incubate in the animal for 3-4 weeks to reach peak expression. We closely monitored the health of the animals throughout the length of the virus incubation and did not note any concerns.

On the day of the cSNAIL mouse or bulk tissue ATAC-seq experiments, we dissected brain regions from fresh tissue and extracted nuclei in the same manner as described for bulk

tissue experiments. Then, we sorted the nuclei suspension into Sun1GFP+ (Cre+) and Sun1GFP- (Cre-) fractions using affinity purification with Protein G Dynabeads (Thermo Fisher, cat. 10004D). A pre-clearing incubation with beads and nuclei for 10-15 minutes removes effects from non-specific binding events. Next, we incubated the remaining free nuclei with anti-GFP antibody (Invitrogen, #G10362) for 30 minutes to bind Sun1GFP. Finally, we added new beads to the solution to conjugate with the antibody and incubated the reaction for an additional 20 minutes. The pre-clear step and all incubations took place in wash buffer (0.25M Sucrose, 25mM KCl, 5mM $MgCl_2$, 20mM Tricine with KOH to pH 7.8, and 0.4% IGEPAL) at 4°C with end-to-end rotation. After the binding process, we separated bead-bound nuclei on a magnet, washed three times with wash buffer, and filtered through a 20µM filter to ensure purity. We resuspended nuclei in nuclease-free water for input into the ATAC-seq tagmentation reaction. We performed nuclei quantification and tagmentation in the same manner described for bulk tissue ATAC-seq above.

We list in

Table 1 the number of animals, the genotypes, and which regions collected for ATAC-seq experiments in this study. All transgenic mouse strains in this study were originally generated on C57BL/6J backgrounds and lines were maintained on a C57BL/6J background throughout breeding. The general breeding strategy was homozygous transgenic mice with C57BL/6J mice to produce heterozygous transgenic offspring for experiments, except for Sst-Cre mice which were homozygous for the transgene. To minimize genetic drift within an isolated population, breeding C57BL/6J mice and transgenic mice were routinely refreshed after 1-5 generations with stock animals from The Jackson Laboratory or from other colonies at Carnegie Mellon University. The mice did not contain additional transgenes outside of cell type-specific Cre. N=2 PValb-2a-Cre samples from CPU/NAc region had received a sham surgery with saline injection into the external globus pallidus 5 days before they were

sacrificed. N=2 D1-Cre samples from both CPU and NAc regions had received headcap surgeries 3 weeks before they were sacrificed. Both PValb-2a-Cre and D1-Cre were overall healthy at time of sacrifice. We collected cSNAIL samples from both sexes where possible (

Table 1).

Table 1. cSNAIL ATAC-seq mouse sample summary

Genotype, number of replicates, sex, region and number of replicates per region, and cell type information for cSNAIL samples.

Genotype	Replicates	Sex (Female /Male)	Region and Replicate per region	Cell type
C57BL/6 WT	N=4	2 F, 2 M	CTX= 4, CPU/NAc = 4	bulk
Pvalb-cre	N=5	3 F (CTX) 1 F, 1 M (CPU/NAc)	CTX= 3, CPU/NAc = 2	PV
Sst-cre	N=2	1 F, 1 M	CTX= 2, CPU/NAc = 2	SST
Drd1-cre	N=2	2 F	CPU=2, NAc=2	D1
Adora2a-cre	N=2	2 F	CPU=2, NAc=2	D2

2.3.4 Statistical Analyses

We created a consensus set of non-overlapping IDR peaks from the ATAC-seq pipeline for cSNAIL ATAC-seq and Mo *et al.* INTACT samples (`Tissue`: Ctx, CPU, and NAc ; `Celltype`: EXC, PV, SST, VIP, D1, D2). We extended the peak set 200bp up- and down-stream, count overlapping fragments with `Rsubread v2.0.1` using the de-duplicated BAM files from the pipeline[51], and created with `DESeq2 v1.26.0` a variance-stabilized count matrix aware of experimental `Group` (combination of `Tissue` and `Celltype`) with `~Group`[102]. We plotted the principle component analysis in **Figure** 5A for the first two components with this variance-stabilized count matrix. We used `Deeptools v3.5.0` to

convert the same BAM files to normalized bigWig files and average over replicates of the same `Group`[103]. We plotted the tracks using `pyGenomeTracks v3.5` around marker genes for each cell type (*Slc17a7*, *Drd1*, *Adora2a*, *Pvalb*, *Sst*, *Vip*)[104] **Figure** 5B. We computed the mean accessibility for each `Group` 2kb up- and down-stream the transcription start sites (TSS) and correlated $\log_{10}$ (TSS accessibility + 1) with gene expression $\log_{10}$(meta gene counts + 1) of Drop-Seq annotated cell types from prefrontal cortex and striatum[105]. We used the Saunders *et al.* tissue subcluster metagene profiles (sum of gene expression in all cells) and summed subclusters to cluster-level metagene profiles. Most tissue cluster metagene profiles corresponded to cSNAIL ATAC-seq celltype and tissue profiles, except for cSNAIL cortical PV+ samples were matched to Saunders *et al.* cortical MGE+ interneuron clusters.

2.3.5 Linkage-Disequilibrium Score Regression (LDSC)

We computed the conditioned heritability of CREs for GWAS variants using the stratified LDSC regression pipeline for cell type-specific enrichment as outlined in https://github.com/bulik/ldsc/wiki/Cell-type-specific-analyses[106]. We downloaded the GWAS summary statistics files and processed them with the LDSC `munge_sumstats` function to filter rare or poorly imputed SNPs with default parameters. We munged the summary statistics files for HapMap3 SNPs excluding the MHC regions downloaded at `http://ldsc.broadinstitute.org/static/media/w_hm3.noMHC.snplist.zip`. We inspected GWAS file to ensure the effect allele, non-effect allele, sample size, p-value, and signed summary statistic for each SNP in each GWAS were included and appropriate for LDSC. The addiction-associated GWAS measure genetic predisposition for age of smoking initiation (`AgeofInitiation`)[67], heaviness of smoking (`CigarettesPerDay`)[67], having ever regularly smoked (`SmokingInitiation`)[67],

current versus former smokers (SmokingCessation)[67], alcoholic drinks per week (DrinksPerWeek)[67], cannabis consumption (Cannabis)[65], and risk tolerance (RiskyBehavior)[66]. GWAS traits related to addiction include multisite chronic pain (ChronicPain)[107] and number of coffee cups drank per data (CoffeePerDay)[108]. Other addiction-related traits come from underpowered GWAS including opioid dependence (OpioidDep)[109], cocaine dependence (CocaineDep)[110], and diagnosis of obsessive-compulsive disorder (OCD)[111]. GWAS from strong brain-related traits used are schizophrenia risk (Schizophrenia)[112], highest level of educational attainment (EduAttain)[113], and sleep duration (SleepDuration)[114]. The non-brain related traits measure genetic liability for lean body mass (LBM)[115], bone mineral density (BMD)[116], and coronary artery disease (CAD)[117].

We estimated LD scores for each foreground set and corresponding background set **(see LDSC Regression GWAS Enrichment Backgrounds)** with the LDSC regression pipeline make_annot and ldsc functions using hg38 1000 Genomes European Phase 3 European super-population (1000G EUR) cohort plink files downloaded from https://data.broadinstitute.org/alkesgroup/LDSCORE/GRCh38/. An example of an ATAC-seq optimal set of reproducible peaks mapped to hg38 in narrowPeak format is annotated with 1000G EUR LD scores using the following call:

```
python make_annot.py \
--bed-file optimal_peak.narrowPeak.gz \
--bimfile 1000G.EUR.hg38.${chr}.bim \
--annot-file foreground.${chr}.annot
```

We downloaded the baseline v1.2 files for cell type-specific enrichment in hg38 coordinates from the same link above as well as the corresponding weights `weights.hm3_noMHC' file excluding the MHC region from https://data.broadinstitute.org/alkesgroup/LDSCORE/. HapMap SNPs

and corresponding weights file used in the LDSC analyses only refer to the SNP rsIDs, rather than genomic coordinates, so only the baseline and LD statistics used to annotate the foreground and background files must be in `hg38` coordinates. In accordance with the LDSC regression script input format, we created an `enrichment.ldcts' file listing the annotated foreground/background pair for each foreground set. We estimated the conditional heritability enrichment using the `ldsc` function, which integrates the foreground and background LD score estimates, munged GWAS SNP data, baseline variant data, and variants weights. The final function call to GWAS enrichment was as follows:

```
python ldsc.py --h2-cts $Munged_GWAS \
--ref-ld-chr baseline_v1.2/baseline. \
--w-ld-chr weights.hm3_noMHC. \
--ref-ld-chr-cts enrichment.ldcts \
--out $Output_Label
```

The pipeline was run using the `--h2-cts` parameter produces the conditional coefficient estimate of tau_C (the additive difference in heritability per SNP in SNPs inside vs. outside the foreground conditional on the background and baseline annotations)[74,93], coefficient error, and coefficient p-value estimates. We adjusted for multiple testing using the false discovery rate on coefficient p-values of the LD score regression coefficients (`alpha = 0.05`) on all 18 GWAS traits intersected on within the same foreground/background set. A significant FDR-value indicates enrichment of the foreground genomic regions for GWAS SNPs relative to the background. Lastly, we computed genetic correlations in **Figure** 2 between GWAS of addiction-associated traits using the pre-munged summary statistics as described by Bulik-Sullivan *et al.*[118].

We trained a set of convolutional neural network (CNN) models to learn the regulatory code of a given cell type from the DNA sequences underlying the cell type's OCRs. The models

take in one-hot encoded 501bp genomic sequences to predict 1 for an OCR or 0 for a non-OCR sequence. Positive sequences were centered on IDR peak summits that are annotated to be in introns and non-promoter non-coding regions, and negative sequences are approximately ten times the number of positives sequences that are G/C-matched and not overlapping IDR peaks. We excluded promoters, which we defined as within 5,000bp of the TSS (based on GRCH38.p12), and exons because distal sequences have been shown to confer more enhancer cell type-specificity and be more predictive of expression levels of regulated genes[23]. We constructed the negative set by first building a sequence repository `$BGDIR` according to https://bitbucket.org/CBGR/biasaway_background_construction/src/master/ from the mouse `mm10` genome using 501bp sequences. Then we used the `biasaway`[119,120] command-line interface to generate negative sequences with the matching nucleotide distribution along a sliding window of the 501bp IDR peak sequence:

biasaway c --foreground $FORGEGROUND_FASTA --nfold 10 --deviation 2.6 --step 50 --seed 1 –winlen 100 --bgdirectory $BGDIR

We employed a 5-fold cross validation chromosome hold-out scheme to train 5 models per set of IDR peaks, enabling us to evaluate the stability and consistency of learned regulatory patterns. (A model that was training a fold did not see sequences during training from the validation set for that fold, and no model saw the test set until final model performance evaluation.) Sequences from these chromosomes were used as the validation set for each fold:

fold1: {`chr6, chr13, chr21`}

fold2: {`chr7, chr14, chr18`}

fold3: {`chr11, chr17, chrX`}

fold4: {`chr9, chr12`}

fold5: {`chr10, chr8`}.

We used sequences from chromosomes {`chr1, ch2, chr19`} for the test set.

We trained the models with Keras v2.3.0-tf (https://keras.io/) implemented through Tensorflow v2.2.0 and used stochastic gradient descent (SGD) with Nesterov momentum to learn model parameters that minimized the binary cross entropy loss. All models used the same CNN architecture after a grid-search of hyperparameters found stable and high validation performance by area under the precision-recall curve (auPRC) in an architecture with five Conv1D layers [kernel_size = 11, filters = 200, activation= 'relu', kernel_regularizer=l2(1e-10)] with a Dropout layer (rate = 0.25) between each pair of consecutive layers, then one MaxPooling1D layer (pool_size = 26, strides = 26), one Flatten layer, one Dense layer [units = 300, activation=`relu', kernel_regularizer=l2(1e-10)], one Dropout layer (rate = 0.25), a final output Dense layer [units=1, activation = 'sigmoid', kernel_regularizer=l2(1e-10)], and a final Dropout layer (rate = 0.25) before the sigmoid output layer. [71] to train each fold with batch_size= 1000, epochs = 23, num_cycles = 2.35, base_learning_rate = 1e-2, max_learning_rate = 1e-1,base_momentum = .85,max_momentum = 0.99. With these hyperparameters, we trained models across folds to predict positive OCRs of all measured cell types against an approximately 1:10 positive:negative class ratio. We computed classifier performance metrics including weighted accuracy (using threshold = 0.5), weighted f1_score (using threshold = 0.5), area under receiver operating characteristic (auROC), and area under precision-recall curve (auPRC). We selected the reported hyperparameters because they maximized the validation auPRC, which we prioritized due to the class imbalance. We report the test set auPRC, F1 score, and false positive rate in **Figure** 8A. We provide both the scripts and trained Keras models at . https://github.com/pfenninglab/addiction_gwas_enrichment.

We used CNN model scores to classify whether a peak from *Fullard et al.* NeuN+ open chromatin data is active in a neuronal subtype {EXC, D1, D2}. We took NeuN+ IDR "optimal peaks" from regions significantly enriched for addiction-associated traits {OFC, VLPFC, DLPFC, ACC, STC, PUT, NAc, **Figure** 2A}, extracted 501bp DNA sequences of each centered on the summit, and scored each peak with cell type-specific machine learning models trained with the appropriate tissue context (e.g., score cortical NeuN+ peaks with a model trained with cortical EXC cell type). We averaged scores across models from different cross-validation folds from the same cell types and classified NeuN+ peaks with scores greater than 0.5 as active in that cell type, as this threshold was the most discriminative in classifying positive validation set sequences (**Figure** 8B). We defined these CNN-prioritized peaks as foregrounds for the LDSC regression GWAS enrichment analyses as described above. We created a consensus set of peaks merging all model-prioritized peaks and the Honeybadger2 set of OCRs to be the matched background, and we performed GWAS enrichment and computed FDR for all 18 GWAS traits (only enrichments for addiction-associated GWAS shown, **Figure** 9).

We collected the addiction-associated SNPs by submitting the summary statistics files for the seven addiction-associated traits {AgeofInitiation, CigarettesPerDay, SmokingInitiation, SmokingCessation, DrinksPerWeek, Cannabis, RiskyBehavior} to the FUMA webserver[121]. FUMA computed LD R^2 based on the 1000 Genomes European (1000G EUR) super-population reference[122] via PLINK[123], linked GWAS-significant lead SNPs to off-lead SNPs in LD with the lead, and annotated functional consequences of genetic variants via ANNOVAR based on ENSEMBL build 85 human gene annotations[124]. ANNOVAR functional gene annotations for a SNP are as defined in the primary publication and online: https://annovar.openbioinformatics.org/en/latest/user-

guide/gene/. We scored all effect and non-effect alleles with each set of CNN models, averaged predictions across folds, and calibrated CNN scores that predict activity using the set of validation positive OCRs. We computed the ΔSNP probability effect by taking the difference between the effect allele and the non-effect allele. Most SNPs reported by GWAS are not expected to be the causal variant for a trait, so the distribution of ΔSNP probability can be used to define a null distribution. We computed the p-value that an allele has a non-zero ΔSNP probability by fitting a normal distribution of null ΔSNP probabilities. We corrected for multiple testing using the method swfdr `v1.12.0` to compute q-values to control for a false-discoveries conditioned on potentially informative covariates[125]. Weighted FDR-correction methods, including swfdr, have been shown to be robust to uninformative covariates and increase power to detect real differences for informative covariates while controlling false discoveries[126]. We conditioned the proportion of expected null p-values on the following covariates (**Figure** 7E, step 4): the difference in GC content of the 501 surrounding the SNP compared to the average GC content of positive sequences used to train each model (GC content), the minor allele frequency (MAF) based on the European ancestry subjects in the 1000G reference panel, whether the SNP overlapped a Fullard et al. NeuN+ OCR (inNeuN peak), and whether a SNP was fine-mapped and predicted to be causal by CAUSALdb using the European LD structure and an ensemble of statistical fine-mapping tools (isCausal) (FINEMAP, CAVIARBF, PAINTOR)[127–130]. We applied an alpha of 0.05 on the false-discovery q-values for all 14,790 SNPs scored across 5 sets of CNNs to determine significantly large enough ΔSNP effects.

To accompany cell type-specific activity predictions, we downloaded SNPs that are reported as *cis* expression quantitative trait loci (eQTL) in human cortex, frontal cortex (DLPFC), anterior cingulate cortex (ACC), caudate, putamen, and the nucleus accumbens (NAc) from the GTEX Consortium from

https://www.gtexportal.org/home/datasets[28,29]. We identified genes for which at least one of the 170 SNPs is an eQTL and plotted them as arcs in **Figure** 10B and **Figure** 7. Locus plots are generated with pyGenomeTracks v3.5 tools.

For **Figure** 10A, we compared calibrated SNP probabilities of the either effect or non-effect allele across each model and grouped them by whether they overlapped a cortical or striatal NeuN+ OCR, NeuN- OCR, both, or neither, depending on whether the model was for EXC or D1/D2 neuronal subtypes, respectively. We computed 2-tailed t-tests between groups and corrected for multiple comparisons with the family-wise Bonferroni method for N=18 tests from three models and (4 choose 2) six possible comparisons per model. * $P < 0.05/N$, ** $P < 0.01/N$, *** $P < 0.001/N$.

2.3.6 LDSC Regression GWAS Enrichment Backgrounds

We found that LDSC regression GWAS enrichment analysis is sensitive to the selected background set of matched regions. To construct appropriate background sets for each GWAS enrichment, we used the ENCODE and RoadMap Honeybadger2[23] and Mouse DHS peak sets for the respective human and mouse-based open chromatin GWAS enrichment. The Honeybadger2 set, downloaded from https://personal.broadinstitute.org/meuleman/reg2map/, consists of DNaseI-hypersensitive OCRs across 53 epigenomes consisting of promoter, enhancer, and dyadic regions. Honeybadger2 is an appropriate epigenetic reference for enrichment of cell type-specific open chromatin from various foregrounds such as the Fullard *et al.* and Lake *et al*. Honeybadger2 regions allow the LDSC algorithm to properly account for the heritability from OCRs of a particular cell type or regions rather than because they tend to be more conserved, are enriched for ubiquitously active transcription factor motifs, or other factors distinguishing open chromatin from heterochromatin. The human orthologs of the ENCODE

Mouse DHS peak set, downloaded through the ENCODE ATAC-seq pipeline at http://mitra.stanford.edu/kundaje/genome_data/mm10/ataqc/mm10_univ_dhs_ucsc.bed.gz, is a set of peaks combined from mouse DNaseI-hypersensitivity OCRs from ENCODE and provides a background for human orthologs of mouse OCRs. The mm10 mouse DHS regions were mapped to hg38 as described in Mapping mouse OCR orthologs. For each respective foreground-background pairing, the foreground regions were merged with the background reference to ensure the background always contained the foreground set. The mouse background was used to calculate the significance of the relationship between mouse OCRs and GWAS for addiction-associated traits to control for a possible association between the degree to which a region is conserved and its likelihood in influencing the predisposition to an addiction-associated trait.

2.3.7 Interpretation of Convolutional Neural Network models

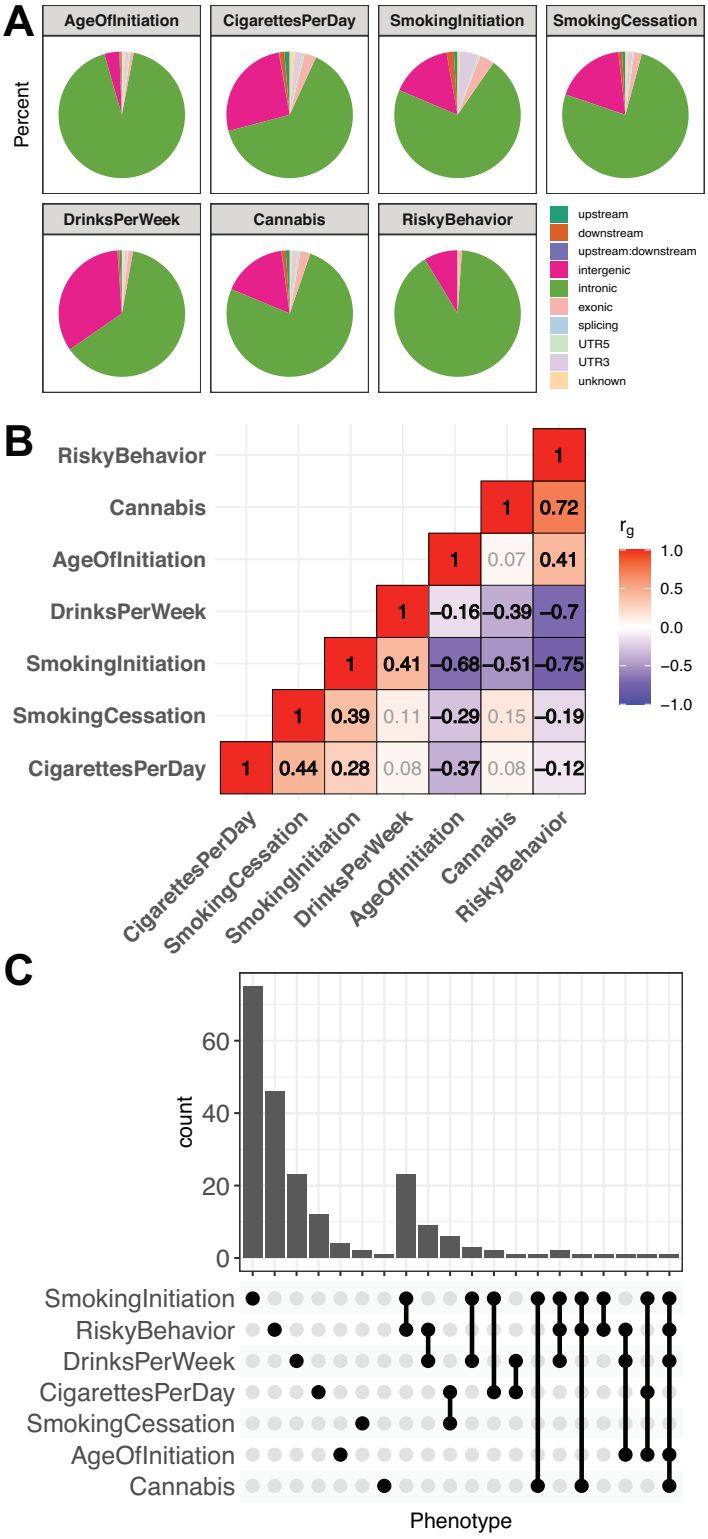
To ensure that the classification task decisions relied on biologically relevant sequence signatures and not artifacts, we performed model interpretation using Deep SHAP v0.37.0 [131,132] and TF-MoDISco[133]. For a random subsample of 2,000 true positive sequences from the validation set, we generated per-base importance scores and hypothetical importance scores relative to a reference set of 500 true negative sequences from the validation set. These scores describe the contribution of each base toward a positive model classification, which is a predicted OCR in the given cell type. TF-MoDISco is an importance score-aware method that clusters commonly important subsequences, called "seqlets," to construct the motifs that the model learned. We ran TF-MoDISco v0.4.2.3 with the options sliding_window_size=11, flank_size=3, min_seqlets_per_task=3000, trim_to_window_size=11, initial_flank_to_add=3, final_flank_to_add=3, kmer_len=7,

`num_gaps=1`, and `num_mismatches=1`. The resulting motifs were filtered to remove rare patterns with fewer than 100 supporting seqlets. Then, the motifs were visualized and associated with known motifs using Tomtom[134] version 5.3.3 with the HOCOMOCO v11 FULL database and default parameters (**Appendix A Extended Data Figure** 7-1).

2.4 Results

2.4.1 Genetic risk for substance use traits is associated with the neuronal epigenomes of reward areas

Recent well-powered GWAS have identified dozens of candidate genetic risk loci associated with seven addiction-associated traits: age of smoking initiation (`AgeOfInitiation`), average number of cigarettes smoked per day (`CigarettesPerDay`), having ever regularly smoked (`SmokingInitiation`), being a former versus current smoker (`SmokingCessation`), the number of alcoholic drinks per week (`DrinksPerWeek`), lifetime cannabis use (`Cannabis`), and risk tolerance (`RiskyBehavior`)[65–67]. These GWAS measure reward, risk tolerance, and various substance use behaviors, thereby providing a means of studying genetic variation associated with addiction. We found that 72-98% of addiction-associated genetic variants lie in non-coding regions of the genome (**Figure** 2A). Of those risk variants, 47-85% lie in introns, which is a substantial over-representation in each GWAS (odds ratio, $OR_{AgeOfInitiation}$ =2.3, $OR_{Cannabis}$ = 2.3, $OR_{CigarettesPerDay}$ = 1.4, $OR_{DrinksPerWeek}$ = 1.6, $OR_{RiskyBehavior}$ = 1.4, $OR_{SmokingCessation}$ = 1.8, $OR_{SmokingInitiation}$ =1.3, Fisher's Exact $P_{Bonferroni} < 8 \times 10^{-79}$). Furthermore, pairwise genetic correlations of risk alleles in these seven GWAS indicated shared and distinct genetic architecture across addiction-associated traits (r_g, **Figure** 2B). Although common genetic



variants are shared between addiction-associated traits on a genome-wide scale, the reported significant loci are often unique to a particular trait and are densely packed with SNPs in high LD (**Figure** 2C). SNPs that are associated with the seven traits span 205 non-overlapping loci across the human genome and include on average 71 SNPs (minimum 1, median 22.5, maximum 1780) within each locus that are either genome-wide significant ($P_{GWAS} < 5 \times 10^{-8}$) or in high LD with the lead SNPs ($R^2 > 0.8$, **Figure** 8-1).

Figure 2. Shared and unique genetic architecture of genetic risk variants of addiction-associated traits.

(A) Pie chart of ANNOVAR-annotated SNP function of addiction-associated trait lead and off-lead SNPs in LD R2 > 0.8. Dark colors indicate un-transcribed/non-coding annotations, light for transcribed/exonic annotations. SNP annotation labels are according to ANNOVAR using ENSEMBL build 85

gene annotations (Methods). (B) Pair-wise LDSC genetic correlation (rg) matrix of seven addiction-associated traits. FDR-significant correlations at shown in bold, non-significant in gray (FDR < 0.05). (C) Upset plot of non-overlapping genomic loci shared or unique to each addiction-associated trait. Genomic loci are clustered and identified by shared GWAS-significant SNPs and genomic region overlap.

We investigated whether genetic variants implicated by addiction-associated GWAS show a tendency to cluster at putative cis-regulatory elements (CREs) of the brain using a stratified LDSC regression approach (Statistical Analyses, Methods), which looks for an enrichments of significant SNPs from GWAS in human annotations[93,106]. We applied LDSC to compare the seven addiction-associated GWAS to open chromatin region (OCR) annotations of sorted neuronal (NeuN+) and glial (NeuN-) nuclei across 14 brain regions[87] (**Figure** 3A). We found that genetic variants associated with `SmokingInitiation`, `SmokingCessation`, `DrinksPerWeek`, and `Cannabis` significantly enriched in NeuN+ OCRs of brain regions known and speculated to contribute to reward and addiction [88] (FDR < 0.05). We found that genetic variants associated with `SmokingInitiation` and `Cannabis` shared enrichment in NeuN+ prefrontal cortical OCRs (from orbitofrontal cortex and dorsolateral prefrontal cortex; SmokingInitiation-OFC FDR = 9.1E-03, Cannabis-OFC FDR = 1.4E-02, SmokingInitiation-DLPFC FDR = 9.1E-03, Cannabis_DLFC FDR = 2.6E-02), while those associated with `SmokingCessation` and `DrinksPerWeek` shared enrichment in NeuN+ striatal OCRs (both putamen and NAc; SmokingCessation-PUT FDR = 2.5E-02, SmokingCessation-NAC FDR = 4.3E-03, DrinksPerWeek-PUT FDR = 3.3E-02, DrinksPerWeek-NAC FDR = 1.0E-02). The enrichments of NeuN+ OCRs indicate that genetic variation in epigenomes of neuronal populations from frontal cortex and striatum contribute to

addiction liability. The difference in NeuN+ enrichments between regions across addiction-associated traits can likely be explained by the difference in proportions and identities of neuronal subtypes of each area, so we sought to dissect the different neuronal subtypes contributing to these enrichments.

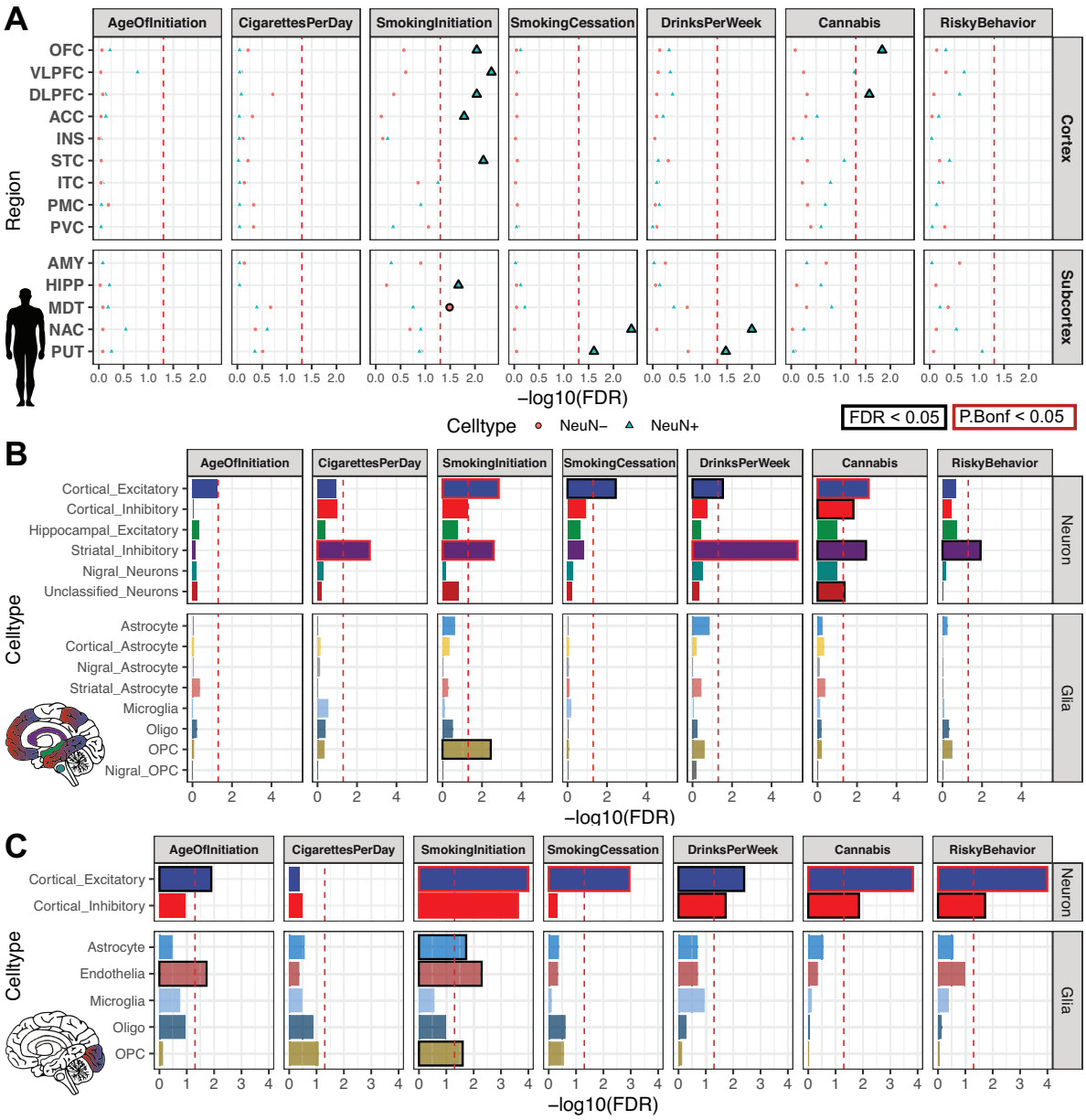


Figure 3. Substance use and risky behavior GWAS risk variants enrich within reward region- and cell type-specific epigenomic profiles.

Stratified LDSC regression (GWAS enrichment) finds enrichment of substance use and risky behavior traits in region-specific and cell type-specific open chromatin profiles of human

postmortem brain.

(A) GWAS enrichment false-discovery rates in ATAC-seq of 14 postmortem human brain regions coupled with NeuN-labeled fluorescence activated nuclei sorting. Brain regions are stratified by cortical and subcortical regions, with cortical regions ordered frontal to caudal. Sorted cell types within each brain region are denoted by shape (blue triangle for NeuN+/neuronal, red circle for NeuN-/glial). FDR-adjustment was performed across all enrichments on the Fullard et al. dataset for ***Figure*** *3. Brain regions reported to be significantly enriched (FDR <= 0.05, black; Bonferroni P-value <= 0.05, red) are plotted with bolded points, and the significance threshold is represented by a dashed red line. (B) Barplot of GWAS enrichment false-discovery rates in single cell open chromatin profiles of cell clusters in isocortex, hippocampus, and striatum. Cell types in brain regions that are significantly enriched (FDR <= 0.05) are plotted with bolded bars, and the significance threshold is represented by a dashed red line. (C) Barplot of GWAS enrichment false-discovery rates in single cell THS-seq OCRs of major cell clusters in occipital cortex. Cell types in brain regions that are significantly enriched (FDR <= 0.05) are plotted with bolded bars, and the significance threshold is represented by a dashed red line. Traits assessed are age of smoking initiation (`AgeofInitiation`), average number of cigarettes per day for ever smokers (`CigarettesPerDay`), having ever regularly smoked (`SmokingInitiation`), current versus former smokers (SmokingCessation), number of alcoholic drinks per week (`DrinksPerWeek`), lifetime cannabis use (`Cannabis`), and risky behavior (`RiskyBehavior`). OFC: orbitofrontal cortex, VLPFC: ventrolateral prefrontal cortex, DLPFC: dorsolateral prefrontal cortex, ACC: anterior cingulate cortex, INS: insula, STC: superior temporal gyrus, ITC: inferior temporal gyrus, PMC: primary motor cortex, PVC: primary visual cortex, AMY: amygdala, HIPP: hippocampus, MDT: mediodorsal thalamus, NAc: nucleus accumbens, PUT: putamen, Ast:*

astrocyte, End: endothelial, Ex: excitatory neuron, In: inhibitory neuron, Mic: microglia, Oli: oligodendrocyte, Opc: oligodendrocyte precursor.

Broad marker-gene based labeling approaches, such as using NeuN to label neurons, do not capture the rich diversity of neuronal subtypes; bulk NeuN+ open chromatin signal represents an average signal from heterogeneous neuronal subtypes, each with distinct epigenomic landscapes, gene regulation, and network connectivity. Hence, NeuN-labeled open chromatin profiles likely do not capture OCRs unique to less populous neuronal subtypes. The difference in proportions of neuronal subtypes between brain regions may also contribute to brain region-specific NeuN+ OCR enrichment for GWAS variants of addiction-associated traits. We therefore applied LDSC regression GWAS enrichment on single cell open chromatin profiles from human postmortem isocortical, striatal, hippocampal, nigral (**Figure** 3B) and occipital cortical cell types (**Figure** 3C)[88,89]. We found that addiction-associated genetic variants largely enriched in both excitatory and inhibitory neuronal OCRs. Genetic variants associated with `SmokingInitiation` (FDR = 1.4E-03, $P_{Bonferroni}$ = 1.8E-02), `SmokingCessation` (FDR = 3.6E-03), `DrinksPerWeek` (FDR = 3.6E-03), and `Cannabis` enriched in isocortical excitatory neuron OCRs (**Figure** 3B). We found enrichment of genetic variants associated with `CigarettesPerDay` (FDR = 2.2E-03, $P_{Bonferroni}$ e = 3.3E-02), `SmokingInitiation` (FDR = 2.5E-03, $P_{Bonferroni}$ = 4.4E-02) `DrinksPerWeek` (FDR = 4.8E-06, $P_{Bonferroni}$ = 3.4E-05), `Cannabis` (FDR = 3.4E-03), and `RiskyBehavior` (FDR = 1.1E-02) in striatal inhibitory neurons. Genetic variants associated with `Cannabis` also enriched in isocortical inhibitory neuron (FDR = 1.5E-02) and unclassified neuron OCRs (FDR = 4.4E-02). Among the glial cell types, only oligodendrocyte precursor cell OCRs were enriched for an addiction-associated trait (`SmokingInitiation`; FDR = 3.6E-03). We found enrichment of genetic variants

associated with `AgeOfInitiation` (FDR = 1.2E-02) and `SmokingCessation` (FDR = 1.0E-04, $P_{Bonferroni}$ = 6.1E-04) in OCRs of occipital cortical excitatory neurons. We found no enrichment of genetic variants associated with `CigarettesPerDay` for OCRs of occipital cortex cell types. Genetic variants associated with `SmokingInitiation`, which enriched in astrocyte (FDR = 1.8E-02), endothelial (FDR = 5.1E-03), inhibitory (FDR = 2.4E-04, $P_{Bonferroni}$ = 2.1E-03), and oligodendrocyte precursor cell OCRs (FDR = 2.5E-02) from occipital cortex, shared enrichment in NeuN- OCRs of mediodorsal thalamus (**Figure** 3A; FDR = 3.2E-02). Interestingly, genetic variants associated with `SmokingCessation`, which showed enrichment for striatal NeuN+ OCRs, enriched only for OCRs of occipital cortical excitatory neurons and not cortical inhibitory neurons (FDR = 1.1E-03, $P_{Bonferroni}$ = 1.3E-02). Sorted bulk ATAC-seq only showed enrichment of `SmokingCessation` associated genetic variants in OCRs of NeuN+ striatal regions, which are largely composed of inhibitory MSNs. We overall found that the enrichments of addiction-associated genetic variants in Corces *et al.* isocortex OCRs agreed with those in Lake *et al.* occipital cortex OCRs. Single-cell epigenomics of human postmortem brain can further dissect the genetic risk for substance-use traits into neuronal subtypes that otherwise would not be parsed with bulk tissue assays.

We confirmed that our pipeline for LDSC regression on NeuN-sorted OCRs from 14 brain regions can reproduce the GWAS enrichments published by Fullard et al. While our approach uses OCRs from reproducible ATAC-seq peaks rather than differentially accessible peaks, we found consistent enrichments of genetic variants associated with schizophrenia risk (Schizophrenia; VLPFC_NeuN+ FDR = 9.3E-06, OFC_NeuN+ FDR = 3.1E-05, STC_NeuN+ FDR = 6.6E-05, NAC_NeuN+ FDR = 9.1E-05, DLPFC_NeuN+ FDR = 9.1E-05, ACC_NeuN+ FDR = 1.0E-04, ITC_NeuN+ FDR = 1.0E-04, PUT_NeuN+ FDR = 2.6E-04, HIPP_NeuN+ FDR = 8.4E-04, PMC_NeuN+ FDR = 2.9E-03, INS_NeuN+ FDR = 1.4E-02, MDT_NeuN+ FDR = 2.4E-02), highest level of educational attainment (EduAttain;

NAC_NeuN+ FDR = 8.4E-06, VLPFC_NeuN+ FDR = 3.7E-04, PUT_NeuN+ FDR = 1.1E-03, OFC_NeuN+ FDR = 1.3E-03, STC_NeuN+ FDR = 1.6E-03, DLPFC_NeuN+ FDR = 2.4E-03, STC_NeuN- FDR = 6.7E-03 , HIPP_NeuN+ FDR = 6.7E-03, ITC_NeuN+ FDR = 1.4E-02, MDT_NeuN+ FDR = 1.4E-02, VLPFC_NeuN- FDR = 2.4E-02, ACC_NeuN+ FDR = 2.4E-02, MDT_NeuN- FDR = 2.8E-02, PVC_NeuN- FDR = 3.2E-02, PMC_NeuN+ FDR = 3.5E-02), and habitual sleep duration (SleepDuration; STC_NeuN+ FDR = 2.7E-04, VLPFC_NeuN+ FDR = 3.7E-04, PUT_NeuN+ FDR = 4.0E-04, NAC_NeuN+ FDR = 5.0E-04, DLPFC_NeuN+ FDR = 1.6E-03, ITC_NeuN+ FDR = 1.4E-02, OFC_NeuN+ FDR = 2.4E-03, MDT_NeuN+ FDR = 4.1E-03) (Figure 4B). We did not find enrichment in brain OCRs of genetic variants identified in several low-powered GWAS (cocaine dependence (CocaineDep)[110], opioid dependence (OpioidDep)[109], and obsessive-compulsive disorder (OCD)[111], each of which had included fewer than 5000 individuals with the trait (Figure 4A). In addition, we found no enrichments in brain OCR for several well-powered studies of traits related to addiction behaviors, including multi-site chronic pain (ChronicPain)[107] and quantity in cups of coffee drank per day (CoffeePerDay)[108]. We also found no enrichment in brain OCRs for anthropometric traits, including coronary artery disease (CAD)[117], bone mineral density (BMD)[116], and lean body mass (LBM)[115] (Figure 4B, C). Lastly, we validated that human OCRs from non-brain tissues would not enrich for risk variants associated with brain traits. We gathered publicly available OCRs from stomach ATAC-seq, adipocyte ATAC-seq, preadipocyte ATAC-seq, liver DNase-seq, and lung DNase-seq profiles[95–98] (**Figure** 7D) and performed LDSC regression on the total 18 GWAS from above. To our expectation, we did not find enrichments of stomach, liver, or lung OCRs for genetic variants associated with brain-related traits. We did find enrichment of BMD in lung OCRs (FDR = 9.1E-04, PBonferroni = 9.1E-04), a connection previously recognized[135–137]. The secondary GWAS enrichments in other traits and foregrounds demonstrate two trends: a GWAS trait would enrich if the GWAS

was properly powered to detect genetic risk variants, and the foreground regions are from cell types or tissue of that trait's potential etiological origin.

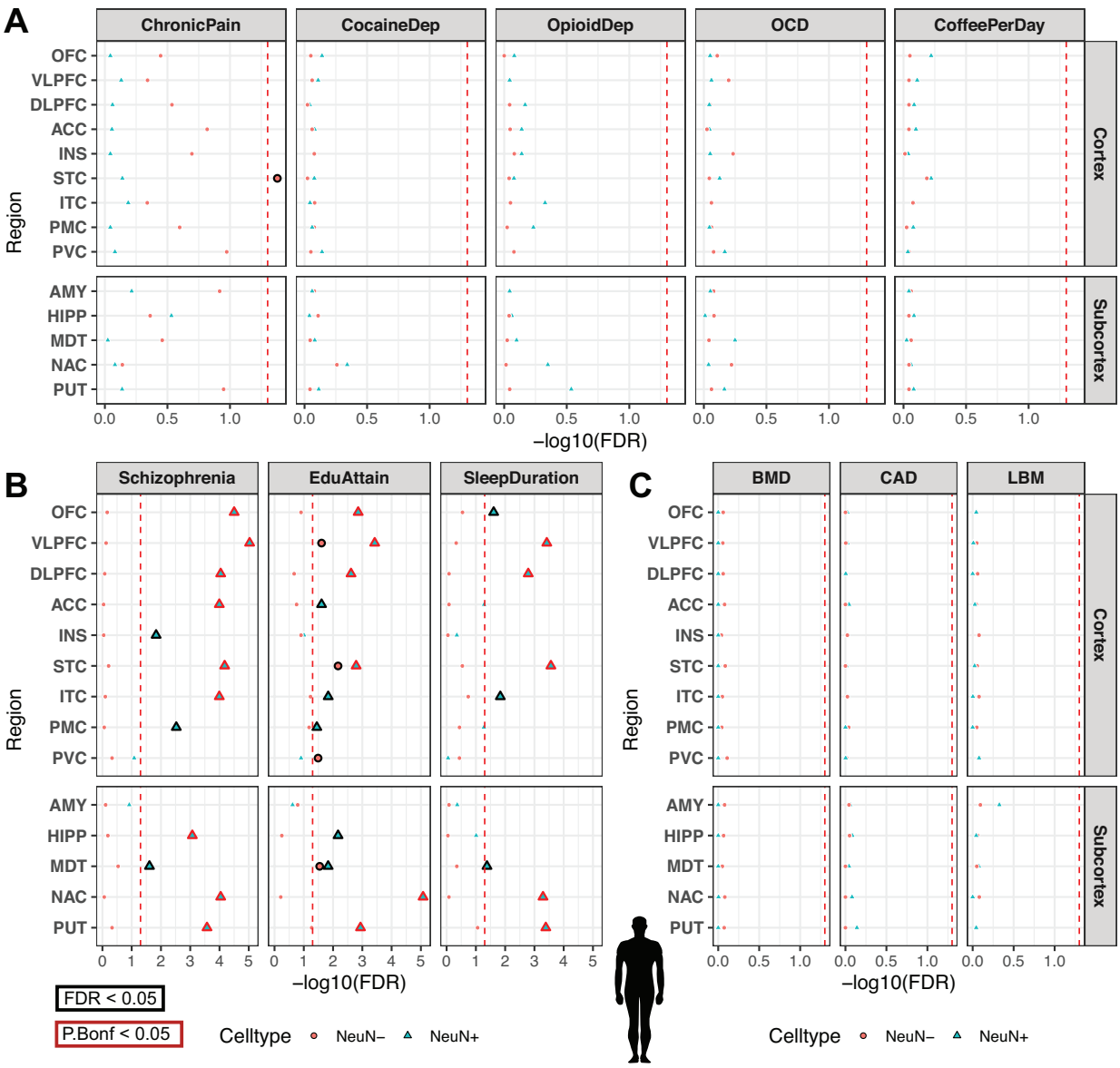


Figure 4. Sensitivity of stratified LDSC regression for cell type- and region-specific in the GWAS trait enrichment requires well-powered GWAS in relevant cell types.

GWAS enrichment plots with false-discovery rates in ATAC-seq of 14 postmortem human brain regions coupled with NeuN-labeled fluorescence activated nuclei sorting. Regions are stratified by cortical and subcortical regions, with cortical regions ordered frontal to caudal. Sorted cell types within each brain region are denoted by shape (blue triangle for NeuN+/neuronal, red circle for NeuN-/glial). Cell types in brain regions that are enriched at

FDR <= 0.05 are plotted with bigger shapes and with black outlines and enriched at Bonferroni P-value <= 0.05 with red outlines.

(A) GWAS enrichment of addiction- or substance use-associated traits: multi-site chronic pain (`ChronicPain`), cocaine dependence (`CocaineDep`), opioid dependence (`OpioidDep`), diagnosis of obsessive-compulsive disorder (`OCD`), and cups of coffee drank per day (`CoffeePerDay`). The GWAS for OCD, opioid dependence, and cocaine dependence are reportedly underpowered to detect genetic liability for these traits (N_{case}< 5,000).

(B) GWAS enrichment in well-powered brain-related traits show cell type- and region-specific enrichment: educational attainment (`EduAttain`), schizophrenia risk (Schizophrenia), habitual sleep duration (`SleepDuration`).

(C) GWAS enrichment in non-brain associated traits do not show cell type- or region-specific enrichment: heel bone-mineral density (`BMD`), coronary artery disease (`CAD`), and lean body mass (`LBM`).

2.4.2 Mouse-human conserved cell type-specific open chromatin enrich for addiction risk loci

In order to further interrogate the different neuronal subtypes that comprise the enrichment of addiction-associated genetic variants in OCR sets measured by Fullard *et al.*, Corces *et al.*, and Lake *et al.* (**Figure** 3, **Figure** 4), we performed targeted epigenomic experiments in mouse on isolated neuronal subtypes from key brain regions of the reward circuit: frontal cortex (CTX), caudoputamen (CPU), and the nucleus accumbens (NAc). We isolated nuclei from specific cell types for ATAC-seq using a modified version of the INTACT approach[80] called *cre*-specific nuclei anchored independent labeling (cSNAIL, Experimental Design, Methods). cSNAIL-INTACT was applied to isolate nuclei marked by *Pvalb*, *Sst*, *Drd1*, and *Adora2a* in *cre-driver* lines using a shortened form of the *Sun1-Gfp* fusion protein

packaged with AAV-PHP.eb and delivered through retro-orbital injection (**Figure** 6A). We show that cell type-targeting provided markedly distinct genome-wide ATAC-seq profiles compared to bulk tissue ATAC-seq alone (**Figure** 5A). cSNAIL ATAC-seq specifically captured nuclei with increased accessibility around the marker gene that was driving *Cre* recombinase expression (**Figure** 5B). Accessibility around cSNAIL ATAC-seq transcription start sites (TSS) strongly correlated with matched pseudobulk gene expression in the same cell type and tissue (Methods, both Pearson and Spearman correlation $P_{bonf} < 2 \times 10^{-16}$, **Figure** 5C,D). We applied HALPER, an approach that leverages reference-free multi-species genome alignments to produce 1-1 contiguous CRE orthologs[92], to reliably map ~70% of mouse neuronal subtype OCRs to their human orthologs in the `hg38` human reference genome (Statistical Methods) for LDSC regression GWAS analysis.

Our GWAS enrichment analysis of human orthologs from mouse OCRs (mouse-human orthologs) measured in various neuronal subtypes and bulk tissue (**Figure** 6B) show that genetic variants associated with `SmokingInitiation` and `Cannabis` shared enrichment in cortical EXC and PV neuron OCRs from both Mo *et al.* (SmokingInitiation- CTX PV+ FDR = 2.6E-02, Cannabis- CTX PV+ FDR = 8.9E-03, SmokingInitiation- CTX EXC+ FDR = 3.6E-02, Cannabis- CTX EXC+ FDR = 1.4E-02) and this study (Pfenning data; SmokingInitiation- CTX PV+ FDR = 2.6E-02, Cannabis- CTX PV+ FDR = 1.3E-02). Genetic variants associated with `Cannabis` further enriched in CTX bulk tissue OCRs (FDR = 1.8E-02), which could be attributed to signal from cortical EXC and PV neuron populations. Cortical PV neuron OCRs further enriched with genetic variants associated with `DrinksPerWeek` (FDR = 1.7E-02). `SmokingCessation` associated genetic variants distinctly enriched in cortical VIP neuron OCRs (FDR = 1.6E-02).

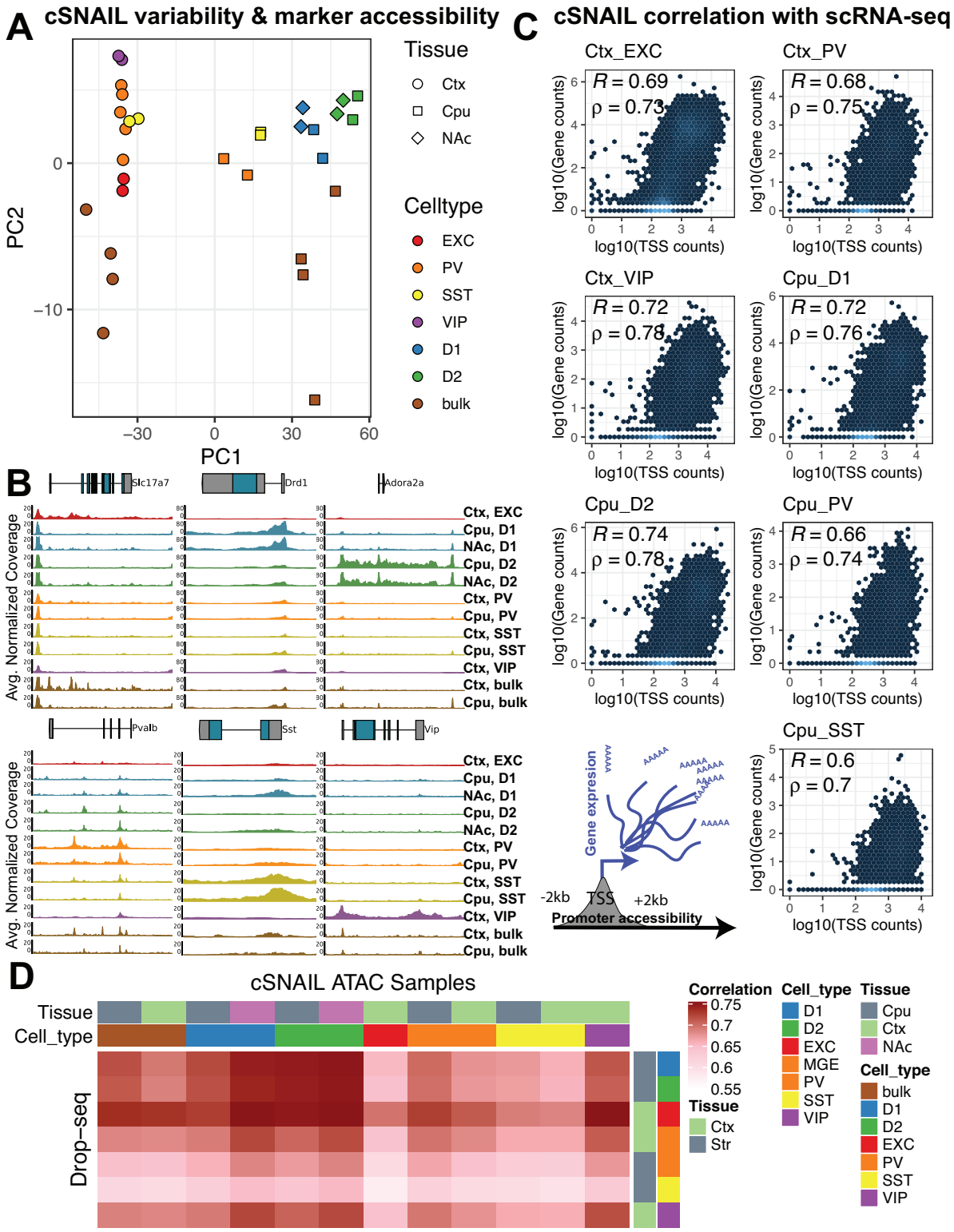


Figure 5. Cell type specificity of cSNAIL ATAC-seq in mouse cortex and striatum

(A) Principle component plots of chromatin accessibility counts from cre-dependent Sun1-GFP Nuclear Anchored Independent Labeled (cSNAIL) ATAC-seq from cre-driver lines (Methods, sample sizes in ***Appendix A Extended data***

Table 1). Major axes of variation separate cell types by tissue source (PC1) and cell type versus bulk ATAC-seq (PC2).

(B) Normalized coverage track plots around marker genes demarcating cell type-specificity of cSNAIL ATAC-seq samples.

(C) Density correlation plot of normalized chromatin accessibility log counts around the transcription start site (TSS) correlated with matched pseudo-bulk cell type log gene counts from Drop-seq of mouse cortex and striatum. Drop-seq cell types meta-gene profiles report sum gene counts for cell clusters from frontal cortex and striatum. Pearson's and Spearman's correlation are denoted with R and ρ, respectively.

(D) Pairwise correlation matrix of TSS chromatin accessibility log counts with Drop-seq pseudo-bulk log gene counts from cortical and striatal cell clusters.

Within neuronal subtypes from CPU and NAc, we found enrichment of genetic variants associated with all measured addiction-associated traits in CPU D2 MSN (AgeOfInitiation FDR = 1.6E-02, CigarettesPerDay FDR = 7.9E-03, SmokingInitiation FDR = 1.6E-03, SmokingCessation FDR = 3.5E-02, DrinksPerWeek FDR = 1.4E-03, Cannabis FDR = 7.1E-04, Cannabis $P_{Bonferroni}$ = 1.9E-02, RiskyBehavior FDR = 2.6E-02) and NAc D2 MSN OCRs (AgeOfInitiation FDR = 2.5E-02, CigarettesPerDay FDR = 9.8E-03, SmokingInitiation FDR = 1.6E-02, SmokingCessation FDR = 7.9E-03, DrinksPerWeek FDR = 3.7E-03, Cannabis FDR = 6.6E-04, Cannabis $P_{Bonferroni}$ = 1.6E-02, RiskyBehavior FDR = 1.6E-03). Genetic variants associated with all measured traits excluding `SmokingInitiation` and `RiskyBehavior` all enriched in NAc D1 MSN OCRs (AgeOfInitiation FDR = 1.4E-02, CigarettesPerDay FDR = 1.6E-02, SmokingCessation FDR = 3.6E-02, DrinksPerWeek FDR = 1.7E-02, Cannabis FDR = 5.1E-03). CPU D1 MSN OCRs were enriched with genetic variants associated with all measured traits excluding `RiskyBehavior` (AgeOfInitiation FDR = 3.1E-02,

CigarettesPerDay FDR = 3.8E-03, SmokingInitiation FDR = 1.5E-02, SmokingCessation FDR = 1.6E-02, DrinksPerWeek FDR = 3.7E-02, Cannabis FDR = 7.8E-03). We found that CPU bulk tissue OCRs were enriched with genetic variants associated with all measured addiction-associated traits excluding `AgeOfInitiation` and `RiskyBehavior` (CigarettesPerDay FDR = 3.4E-03, SmokingInitiation FDR = 1.6E-02, SmokingCessation FDR = 9.2E-03, DrinksPerWeek FDR = 1.6E-02, Cannabis FDR = 1.6E-02). Distinctly, CPU PV+ and SST+ neuron OCRs enriched with genetic variants associated with `Cannabis` (CPU PV+ FDR = 9.5E-03, CPU SST+ FDR = 3.3E-02).

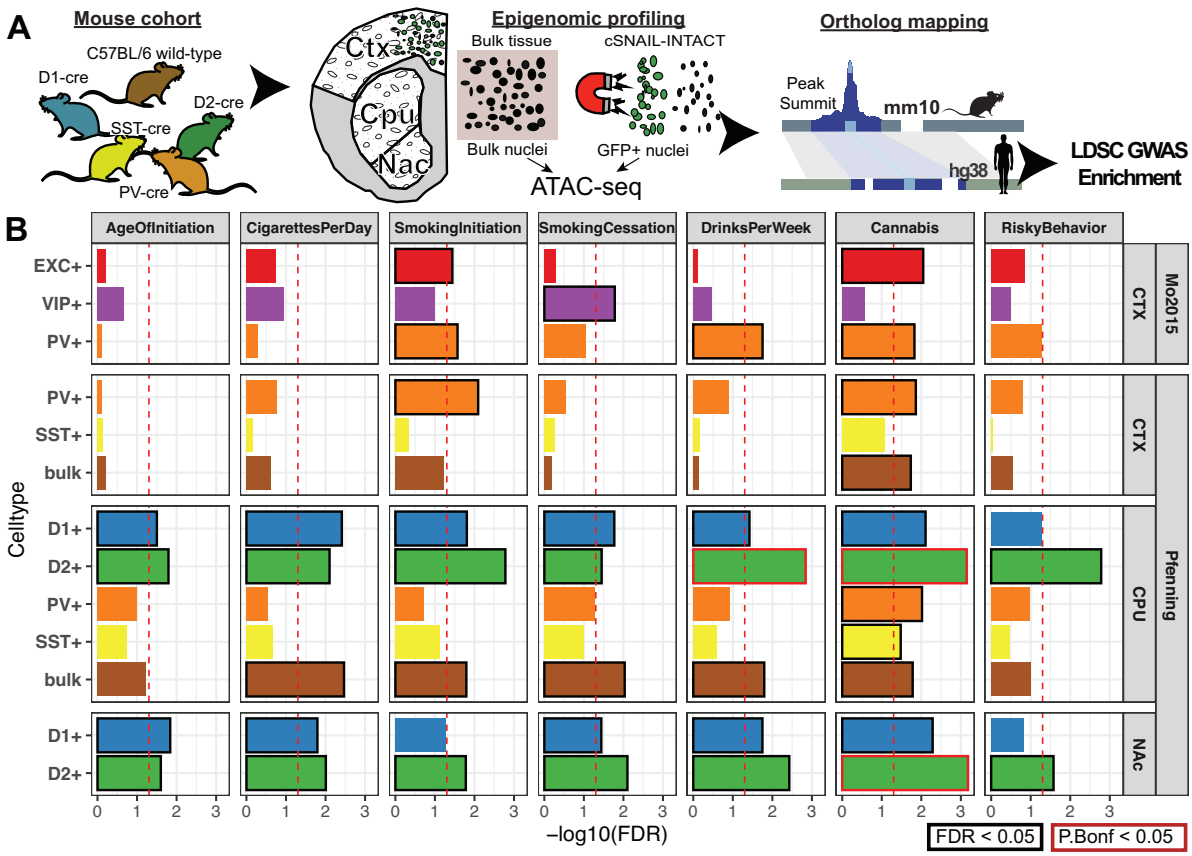


Figure 6. Cell type-specific enrichment of substance use traits are conserved in mouse-human orthologous open chromatin regions.

(A) Experimental design to map human orthologous regions from mouse ATAC-seq of bulk cortex (CTX), dorsal striatum (CPU), and nucleus accumbens (NAc) of cre-dependent Sun1-

GFP Nuclear Anchored Independent Labeled (cSNAIL) nuclei of D1-cre, A2a-cre, PValb-2a-cre, and SST-cre mice. cSNAIL ATAC-seq experiments report enriched (+) nuclei populations. (B) Stratified LD score regression finds enrichment of substance use and risky behavior traits for brain region and cell type specific ATAC-seq open chromatin profiles of mouse brain. Replication of enrichment is shown using INTACT-enriched OCRs from Mo et al. of cortical excitatory (EXC+), vasoactive intestinal peptide interneuron (VIP+), and parvalbumin interneuron (PV+). Enrichments that are enriched at FDR <= 0.05 are plotted with black outlines and Bonferroni P-value <= 0.05 with red outlines. FDR-adjusted p-value was performed across all mouse-human ortholog GWAS enrichment across this ***Figure****.*

Corresponding to our analysis of human brain OCRs, we also confirmed the specificity of mouse-human orthologous CRE enrichments for genetic variants associated with addiction-related, brain-related, and non-brain related traits (**Figure** 7). We found enrichments of genetic variants associated with `ChronicPain` in cortical PV neuron OCRs from both Mo *et al.* (FDR = 3.9E-02) and this study (**Figure** 7A; FDR = 7.9E-03). Within striatal cell types, we found that CPU D2 and NAc D1 MSN OCRs were enriched for genetic variants associated with `ChronicPain` (CPU D2+ FDR = 4.9E-02, NAc D1+ FDR = 2.2E-02). In contrast, CPU D1 and NAc D2 MSN OCRs were enriched for genetic variants associated with `OpioidDep` (CPU D1+ FDR = 4.5E-02, NAc D2+ FDR = 2.0E-02). Genetic variants associated with `OpioidDep` also enriched in CPU PV OCRs (FDR = 4.7E-02). `Schizophrenia`, `EduAttain`, and `SleepDuration` associated genetic variants all enriched in OCRs of all measured cell types (**Figure** 7B). None of these mouse-human orthologs enriched for genetic variants associated with non-brain-related traits: `BMD`, `CAD`, and `LBM` (**Figure** 7C). We validated that our approach to map OCRs from mouse to human did not bias enrichment to brain traits by performing GWAS enrichment on OCRs from mouse non-brain tissues (kidney,

liver, and lung) (**Figure** 7D). As expected, we did not find an enrichment for genetic variants associated with a brain-related trait. We did find that mouse-human orthologs of lung OCRs enrich for BMD (FDR = 8.3E-03), which concords with the enrichment of human lung OCRs.

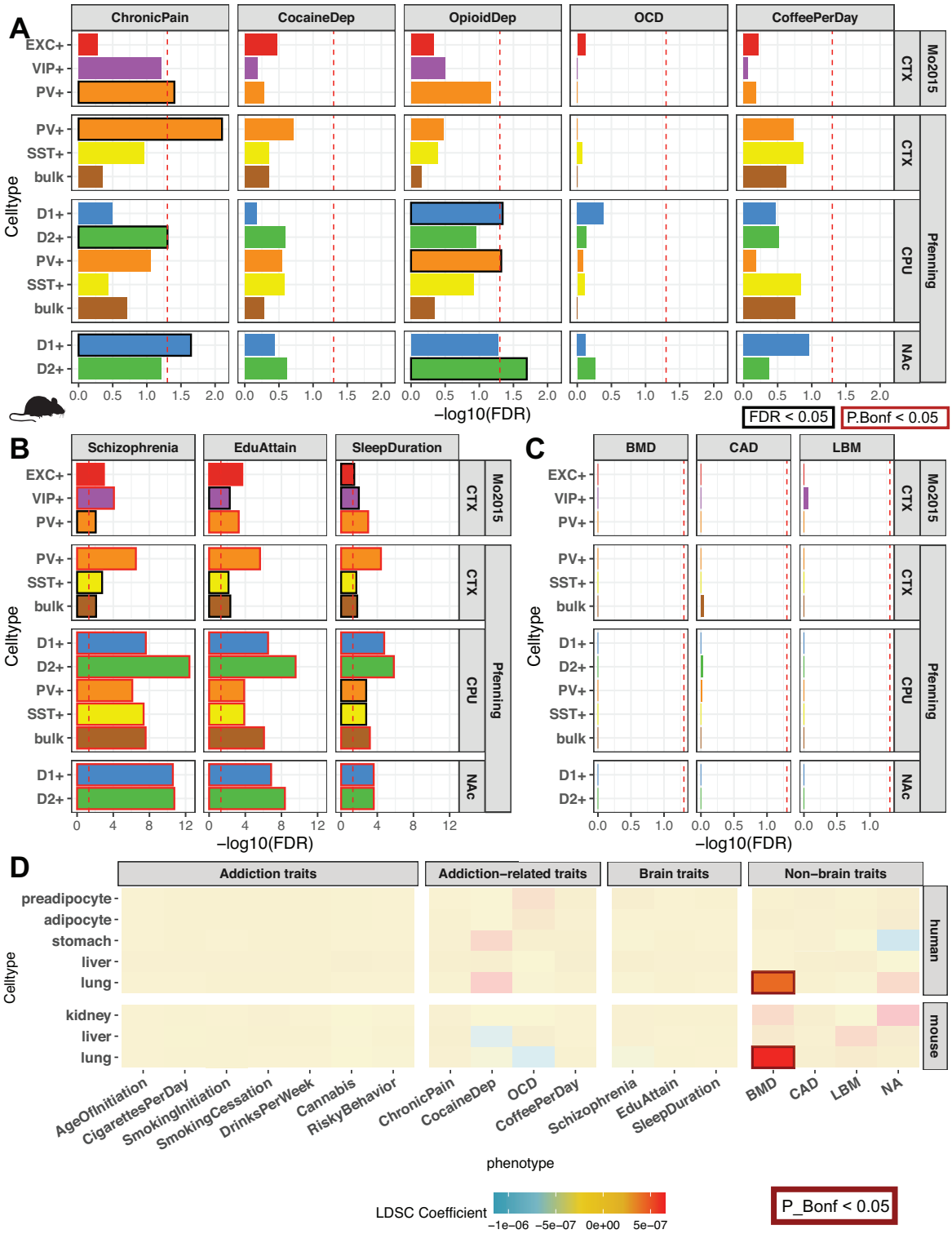


Figure 7. GWAS enrichment in addiction- and non-addiction-related traits using mapped mouse orthologs of tissue- and cell type-specific open chromatin regions.

GWAS enrichment plots with false-discovery rates in human orthologous regions mapped from mouse ATAC-seq of bulk cortex (CTX), dorsal striatum (CPU), and nucleus accumbens (NAc) or cre-dependent Sun1-GFP Nuclear Anchored Independent Labeled (cSNAIL) nuclei of D1-cre, D2-cre, and PV-cre mice. cSNAIL ATAC-seq experiments report both enriched (+) and de-enriched (-) nuclei populations. Enrichments that are enriched at FDR < 0.05 are plot with black outlines. Replication of enrichment is shown using INTACT-enriched OCRs from Mo et al. of cortical excitatory (EXC+), vasoactive intestinal peptide interneuron (VIP+), and parvalbumin interneuron (PV+).

(A) GWAS enrichment of addiction- or substance use-associated traits: multi-site chronic pain (`ChronicPain`), cocaine dependence (`CocaineDep`), opioid dependence (`OpioidDep`), diagnosis of obsessive-compulsive disorder (`OCD`), and cups of coffee drank per day (`CoffeePerDay`). The GWAS for OCD, opioid dependence, and cocaine dependence are reportedly underpowered to detect genetic liability for these traits (N_{case}< 5,000).

(B) GWAS enrichment in well-powered brain-related traits show cell type- and region-specific enrichment: educational attainment (`EduAttain`), schizophrenia risk (`Schizophrenia`), habitual sleep duration (`SleepDuration`).

(C) GWAS enrichment in non-brain associated traits do not show cell type- or region-specific enrichment: heel bone-mineral density (`BMD`), coronary artery disease (`CAD`), and lean body mass (`LBM`).

(D) Heatmap of LDSC regression coefficients of GWAS enrichment for all measured GWAS in non-brain OCRs from human or mouse-human mapped orthologs. Tissues for which OCRs are significantly enriched (FDR < 0.05, black; Bonferroni P-value <= 0.05, red) with GWAS variants are outlined with a bolded box.

2.4.3 Convolutional Neural Network (CNN) models of mouse cell type-specific CRE activity refine human NeuN+ OCRs for GWAS enrichment

The genetic tools available for mouse research allowed us to isolate the nuclei of specific neuronal subtypes and generate deep open chromatin profiles at greater cellular resolution. However, a lack of directly measured mouse-human conservation could lead to false negatives and false positives in the cell type-specificity of CREs at specific loci that add noise to GWAS comparisons. To leverage the strengths of the mouse and human approaches, we developed a procedure to predict the neuronal subtype-specificity of human OCRs using machine learning models trained in mouse. The OCR profile of each neuronal subtype is largely a result of a developmental cascade of transcription factors that cooperatively recognize and bind to specific sequence elements in the genome, resulting in a neuronal subtype-specific open chromatin profile[25]. These complex combinations of sequence features comprise a regulatory code that links genome sequence to neuronal subtype-specific open chromatin. This regulatory code can be effectively learned using convolutional neural networks (CNNs) and has been demonstrated to be highly conserved between mouse and human[83,138].

The concordant pattern of enrichment for addiction associated genetic variants in human and mouse-human orthologous OCRs suggested that risk variants may affect the regulatory activity of neuronal subtypes that are conserved between human and mouse. We therefore devised and trained a collection of CNN binary classification models to learn the genome sequence features that distinguish OCRs for cortical excitatory (EXC) neurons, striatal D1 MSNs, and striatal D2 MSNs[83–85,138]. For each set of reproducible OCRs from the mouse INTACT and cSNAIL data, we trained models to predict the reproducible peaks from approximately ten times the number of nucleotide content-matched negative sequences (Methods). Our models made confident predictions on held-out test sequences as reported by high F1-scores, high area under the precision-recall curves (**Figure** 8A), and low false positive

rates at a threshold of 0.5 (**Figure** 8B). These models reproducibly learned transcription factor motif families that are enriched in human neuronal subtypes of cortex (MEF2, JUN) and striatum (POU, NRF1, ZFHX3), as previously reported by Fullard *et al.* (**Figure** 8F, **Appendix A Extended Data Figure** 8-1).

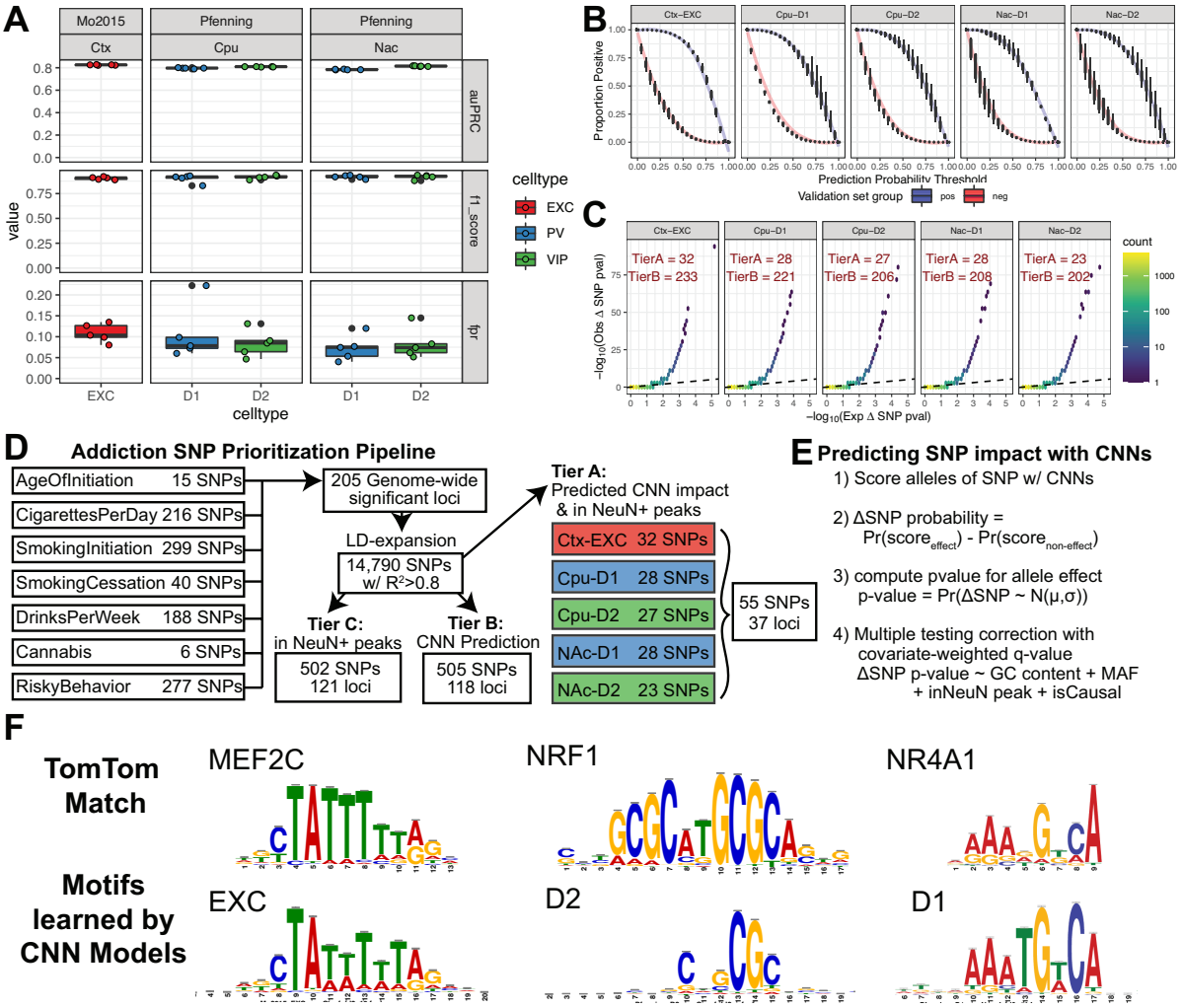


Figure 8. Convolutional Neural Network (CNN) model performance and selection of candidate functional SNPs.

(A) Performance metrics for convolutional neural network (CNN) models show high specificity on the test sets of positive peaks or 10x nucleotide content-matched negatives. Test set performance metrics are reported for area under the precision-recall curve (auPRC), F1-score (using threshold = 0.5), and false positive rates across all possible thresholds (Methods).

Models were trained on IDR peaks of mouse cortical excitatory neurons (Ctx-EXC) and D1 and D2 medium spiny neurons from caudoputamen (CPU) and nucleus accumbens (NAc).

(B) The models best discriminate the proportion of positives and negative sequences at a threshold of 0.5. Plots show the proportion of positives (blue) or negatives (red) that are called "positive" across CNN output thresholds from 0 to 1 averaged across folds for each set of CNN models.

(C) Quantile-quantile plots of p-values of calibrated ΔSNP probabilities (Methods) from a normal distribution after centering by the mean and scaling by the standard deviation of delta SNP probabilities across all SNPs (n=14,790 SNPs) for each set of CNN models. A hexbin plot was used to visualize over-plotting, where every hexagon is colored by the number of SNPs in that bin. The black dotted line denotes the equality line y = x. The number of significant SNPs at false discovery q-value < 0.05 at Tier A or B are reported for each cell type and tissue (Methods).

*(D) Schematic to select for predicted causal impact addiction-associated GWAS SNPs. The pipeline begins with SNPs across addiction-associated GWAS aggregated to 205 non-overlapping GWAS loci across 14,790 SNPs after LD-expansion to include those in LD R^2 > 0.8 (**Appendix A Extended Data Figure** 8-**2**). SNPs are further prioritized into three tiers. Tier C includes SNPs that only overlap Fullard et al. NeuN+ ATAC-seq peaks, Tier B includes SNPs with only predicted significant differential allelic impact by on CNN-predicted CRE activity at q-value < 0.05, and Tier A include SNPs satisfying both criteria (Method).*

(E) Outline of predicting differential CRE activity between alleles using calibrated CNN probabilities of CRE activity while controlling for false discovery with informative covariates (Methods).

(F) Example motif matches from ***Appendix A Extended Data Figure*** *8-**1** of TomTom known transcription factor consensus motifs and the learned important features in CNN models for cortical excitatory and striatal D1 and D2 MSNs.*

We reasoned that NeuN+ OCR signal, which is comprised of OCR signals from several neuronal subtypes, can be parsed into its component cell types by CNNs that are trained to predict OCR activity in those component cell types. This enables the study of human addiction genetics at a cell type-level resolution from high-quality tissue-level open chromatin profiles. To discern whether NeuN+ OCR enrichments for addiction-associated genetic variants come from the same cell types observed in **Figure** 3, we applied our trained CNN models to predict whether bulk cortical or striatal NeuN+ OCRs have activity in either cortical EXC or striatal D1 and D2 neurons, respectively (**Figure** 9A). We did not conduct these analyses for PV, SST, or VIP interneurons because they comprise a much lower percentage of cortical and striatal neurons than the other neuron types[139,140]. We ran LDSC regression GWAS enrichments on the sets of NeuN+ OCRs predicted to be specific to cortical EXC, striatal D1, and striatal D2 neurons. Genetic variants associated with `SmokingInitiation`, which initially were enriched for occurring in OCRs of various NeuN+ frontal cortical areas (**Figure** 2A), were enriched for occurring in NeuN+ OCRs predicted to be active in EXC neurons (**Figure** 9B; VLPFC FDR = 5.0E-03, DLPFC FDR = 9.4E-03, STC FDR = 1.0E-02, ACC FDR = 1.5E-02, OFC FDR = 1.5E-02). Genetic variants associated with `Cannabis`, which were enriched for occurring in NeuN+ cortical OCRs (**Figure** 2A), were also enriched for occurring in NeuN+ OCRs predicted to be active in EXC neurons (OFC FDR = 8.7E-03, DLPFC FDR = 1.7E-02, VLPFC FDR = 3.6E-02). The enrichments of excitatory cortical cell type-specific OCRs for `SmokingInitiation` and `Cannabis` associated genetic variants agree with the results from the analysis of the Fullard *et al*., Corces et al., and Lake *et al*. OCR datasets (**Figure** 2).

Genetic variants associated with `SmokingCessation` and `DrinksPerWeek`, which were enriched for occurring in PUT and NAc NeuN+ OCRs (**Figure** 2A), were enriched for occurring in OCRs predicted to be active in D1 and D2 MSNs of PUT and NAc (SmokingCessation-D2 NAc FDR = 2.3E-03, SmokingCessation-D1 NAc FDR = 5.4E-03, SmokingCessation-D1 PUT FDR = 1.1E-02, SmokingCessation-D2 PUT FDR = 1.9E-02, DrinksPerWeek-D2 PUT FDR = 5.8E-03, DrinksPerWeek-D1 PUT FDR = 1.2E-02, DrinksPerWeek-D1 NAc FDR = 2.6E-02, DrinksPerWeek -D2 NAc FDR = 2.9E-02). Thus, our new framework that we applied to these addiction-related traits (outlined in **Figure** 9A) refines addiction genetic risk signal to neuronal subtypes within specific brain regions. This framework can be applied to CREs from any tissue-cell type combination for which bulk tissue open chromatin measurements are available from human and cell type open chromatin measurements are available from another vertebrate[138,141].

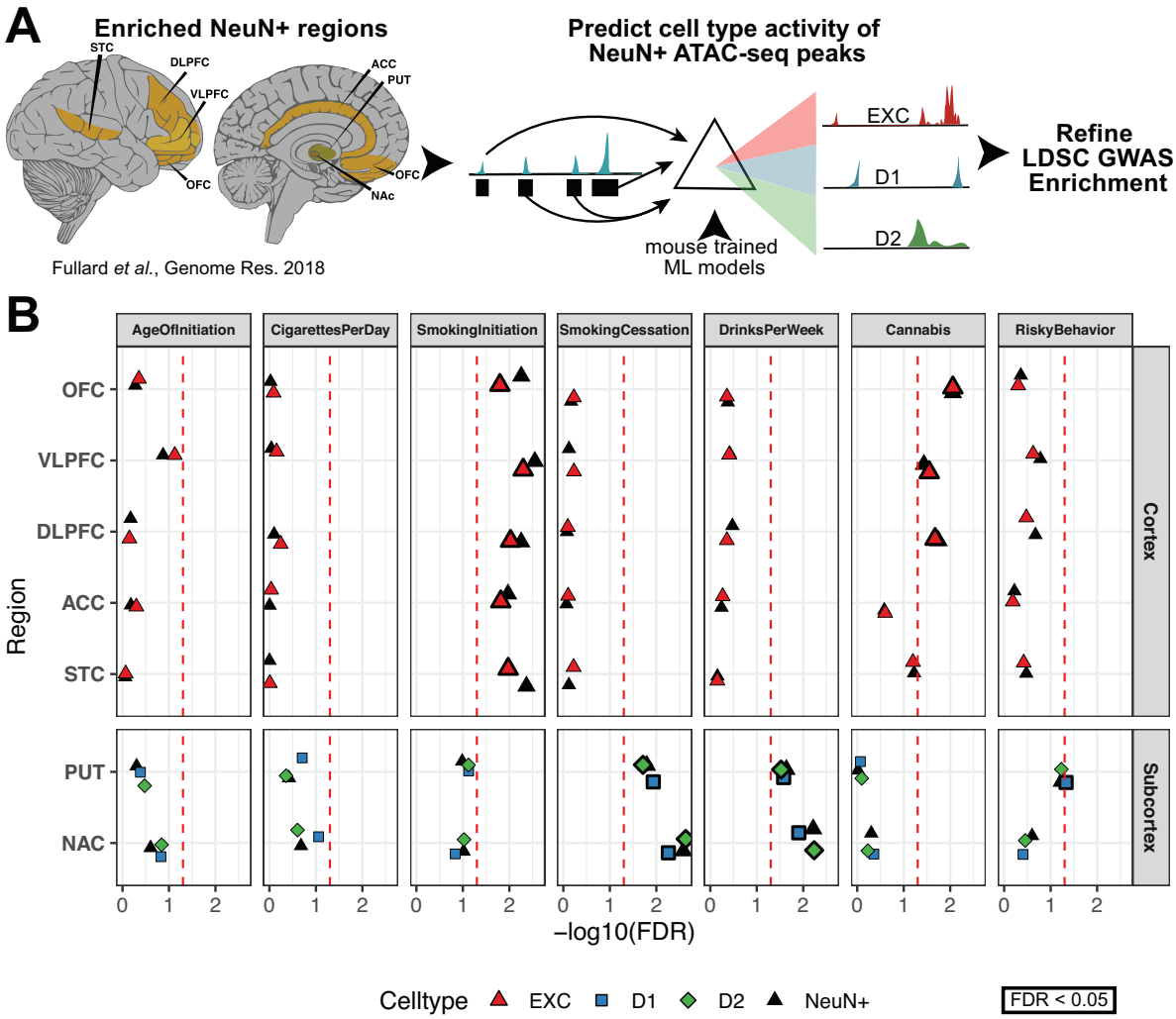
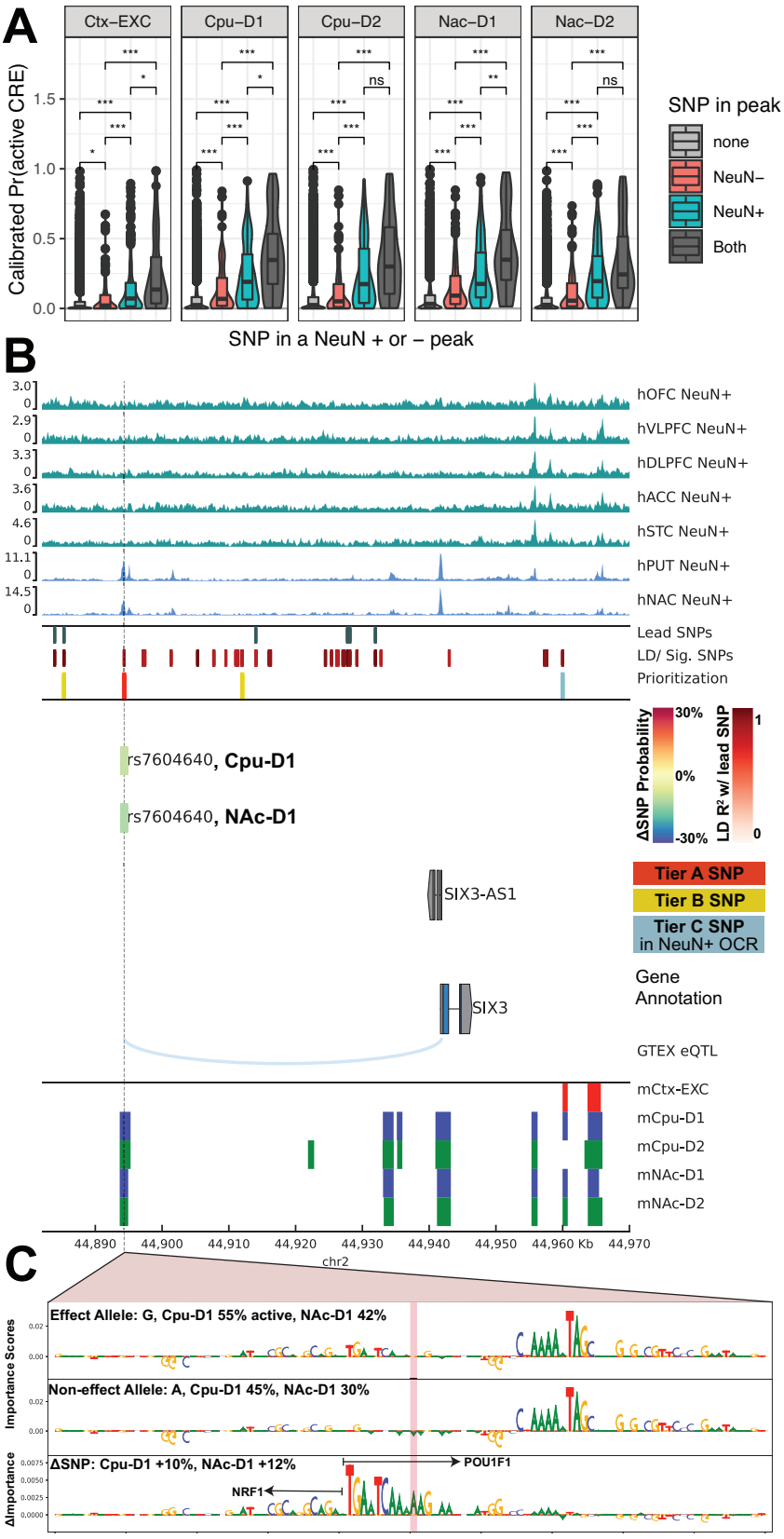


Figure 9. Cell type-specific convolutional neural network (CNN) models refine human NeuN+ enrichments for substance use genetic risk GWAS.

(A) Schematic to predict cell type-specific activity of NeuN+ ATAC-seq peaks enriched from brain regions assayed in Fullard et al. using CNN models trained on mouse cell type-specific ATAC-seq peaks. CNN-predicted OCRs are used as input for computing GWAS enrichment. (B) Stratified LD score regression of addiction associated traits in Fullard et al. NeuN+ OCRs predicted to be cell type-specific by machine learning models of open chromatin. Cell types are colored by the source mouse cell type-specific OCRs from ***Figure*** *9A. Original enrichments from* ***Figure*** *5A are reproduced in black. Larger, bolded points are significant for FDR < 0.05 (red dotted line).*

2.4.4 CNN models predict allele-specific activity of addiction-associated GWAS SNPs in neuronal subtypes

Lastly, we applied our convolutional neural network (CNN) models to screen addiction-associated genetic variants for predicted functional activity in EXC, D1, and D2 neuronal subtypes. CNN-based approaches have been demonstrated to fine-map dense risk loci and select candidate causal genetic variants[83–85,88,142], yet none have been applied in the context of addiction-associated genetic risk or in the cell types that we have assayed. We identified 14,790 unique SNPs that were collected across the seven addiction-associated GWAS to score for differential neuronal subtype OCR activity (Methods). We expect that many SNPs reported from GWAS are significantly associated with traits due to LD rather than being the true causal variant. When scored with our CNN models, the 96.2% of addiction-associated SNPs that either do not lie in any OCR or in only NeuN- OCRs have low probabilities of being active in excitatory, D1, or D2 neuronal subtypes. We also found that these SNPs have significantly lower predicted probabilities of activity than the remaining 3.8% of addiction-associated SNPs in any NeuN+ OCR ($P_{Bonferroni} < 0.05$, **Figure** 10A). We then predicted the probability of activity for both the effect and non-effect alleles and estimated the differential impact of the alleles in order to fine-map candidate causal effect SNP and target neuronal subtypes and tissues. Most SNPs do not have predicted differential allelic activity (delta SNP) in a neuronal subtype, while a handful of SNPs have larger differential activity that deviate from a normal distribution when visualized on quantile-quantile plots (**Figure** 8C, Methods). We outline in **Figure** 8D an approach to prioritize the candidate causal SNPs by two lines of evidence: 1) a predicted differential neuronal subtype OCR activity with large effect size that is controlled for false discovery (q-value < 0.05, Methods) and 2) having physical overlap with measured human NeuN+ OCR in Fullard *et al.* (**Figure** 8D). We are able to prioritize 55 SNPs spanning 37 loci to Tier A that both have significant predicted ΔSNP probability effect and overlap a Fullard *et*



al. NeuN+ OCR, 505 SNPs to Tier B that only have predicted ΔSNP probability effect, and 502 SNPs to Tier C that overlap NeuN+ open chromatin but do not have a predicted significant ΔSNP probability effect (**Appendix A Extended Data Figure** 8-2).

Figure 10. Convolutional neural network (CNN) models for predicting cell type-specific open chromatin predict activity of addiction GWAS SNPs

(A) Cell type-activity predicted probability active by each set of CNN models of cell type activity for genome-wide significant SNPs and off-lead SNPs in LD $R^2 > 0.8$ with the lead SNPs. Activity scores for SNPs

*are stratified by overlap with Fullard et al. cortical or striatal NeuN+ (teal), NeuN- peaks (salmon), both (dark gray), or neither (light gray). Significance symbols denote Bonferroni-adjusted p-values from 2-tailed t-tests for N=18 possible pairwise comparisons, N.S. not significant, * P < 0.05/N, ** P < 0.01/N, *** P < 0.001/N.*

(B) Locus plot for candidate SNPs with predicted function SNP impact in cortical excitatory and striatal D1, and D2 MSN cell types. Genome tracks from top to bottom: human (h) NeuN+ MACS2 ATAC-seq fold change signal of cortical and striatal brain regions enriched in ***Figure*** *5A. SNP tracks show lead SNPs from seven addiction-associated GWAS and the SNPs either in LD with the lead SNPs (Lead SNPs) or independently significant SNPs (LD/ Sig. SNPs). Each SNP is colored by increasing red intensity that indicates the degree of LD with a lead SNP. Prioritized candidate causal SNPs by predicted differential cell type activity and overlap with Fullard et al. NeuN+ OCRs are plot as (red for Tier A, yellow for Tier B, and teal for Tier C, Methods). Tier A SNP rs7604640 is predicted to have a strong ΔSNP effect by CPU-D1 and NAc-D1 CNN models, and the bars are colored by the % change in probability active. Gene annotation tracks plot GENCODE genes from the GRCh38 build. eQTL link tracks of FDR-significant GTEX cis-eQTL from cortical or striatal brain regions, and orthologs of mouse (m) putative CREs mapped from excitatory or striatal neuronal subtypes measured by cSNAIL ATAC-seq. Cell type colors label cortical excitatory neurons (EXC; red), D1 medium spiny neurons (D1; blue), or D2 medium spiny neurons (D2; green).*

(C) Representative importance scores of 50bp flanking either side of the SNP rs7604640 that measure relative contribution of that sequence being active in D1 MSNs. CNN importance score interpretations are shown for effect and non-effect alleles, and the difference in importance scores reveal the relatively more important DNA motif in the effect allele that matches consensus POU1F1 motif overlapping the rs7604640 SNP. The model interprets this

POU1F1 motif and a nearby NRF1 motif as contributing to the effect allele having more activity in D1 MSNs.

One such SNP from Tier A, rs7604640, lies in human NeuN+ open chromatin specific to striatum 46kb upstream of the *SIX3* locus on chromosome 2. rs7604640 overlaps human orthologs of mouse OCRs in only D1 and D2 neurons and we predict the effect allele of rs7604640 has an increased probability of open chromatin activity in D1 OCRs of the striatum compared to the non-effect allele (**Figure** 10B). rs7604640 is one of many off-lead SNPs identified in the `SmokingInitiation` GWAS (P_{GWAS} =3.04 x 10^{-12}) and is in high LD with the SNP rs163522 (R^2 = 0.856, P_{GWAS} =1.11 x 10^{-11}), which is independently significant from the lead SNP, rs1004787 (R^2 = 0.630, P_{GWAS} =5.27 x 10^{-17}). rs7604640 was reported by HaploRegv4 to overlap a POU1F1 motif [98], which our D1 models predict to contribute towards increased probability of being active in D1 MSNs (**Figure** 10C). Furthermore, this SNP is a known *cis*-eQTL for the antisense *SIX3-AS1* gene in striatal regions from the Genotype-Tissue Expression (GTEX) project[28–30,143]. Anti-sense gene expression is one mechanism of regulating their sense gene[144,145], and deletion of the gene *SIX3* has been shown to inhibit development of D2 medium spiny neurons[146]. Altogether, this evidence formulates the hypothesis that common genetic variant rs7604640 has D1 MSN-specific, allelic impact on open chromatin activity in a mouse-human conserved putative CRE regulating the MSN regulator *SIX3*.

In addition to rs7604640, we report four loci with 1-4 candidate SNPs each in Tier A that may be putative causal SNPs with cell type-specific activity in addiction-associated traits (**Figure** 11). The SNPs in these loci all have reported eQTL in frontal cortex or striatum tissues from GTEx, and they overlap corresponding NeuN+ OCRs and mouse-human orthologous OCRs. In some cases, our prioritized Tier A SNPs were predicted to have ΔSNP effects (Methods) in only striatal MSNs, showcasing our framework's ability to predict cell type-

specific impact. These SNPs include rs11191352 ($P_{SmokingInitiation}$=2.12 x 10^{-7}, **Figure** 11A), rs9826458 ($P_{RiskyBehavior}$= 4.36 x 10^{-21}, $P_{SmokingInitiation}$=1.21 x 10^{-14}, **Figure** 11B), and rs9844736 ($P_{RiskyBehavior}$= 3.04 x 10^{-7}, $P_{SmokingInitiation}$=3.58 x 10^{-7}, **Figure** 11C). In a few cases, our models predicted SNPs to have strong ΔSNP effects across both cortical excitatory and striatal cell types. These include two SNPs in the highly pleiotropic MAPT-CRHR1 locus that are 152bp apart and in perfect LD with each other, rs11575895 and rs62056779 (**Figure** 11D). The prioritized SNPs in the MAPT-CRHR1 locus are genome-wide significant for 5 of the 7 addiction-associated traits (**Figure** 8-1) and the locus has been implicated in other neuropsychiatric traits such as Alzheimer's Disease[147,148]. We provide the summary of CNN predictions in these reported loci across all 14,790 analyzed SNPs along with the accompanying annotations that we incorporated into our prioritization of candidate causal SNPs and their predicted cell types (**Appendix A Extended Data Figure** 8-1).

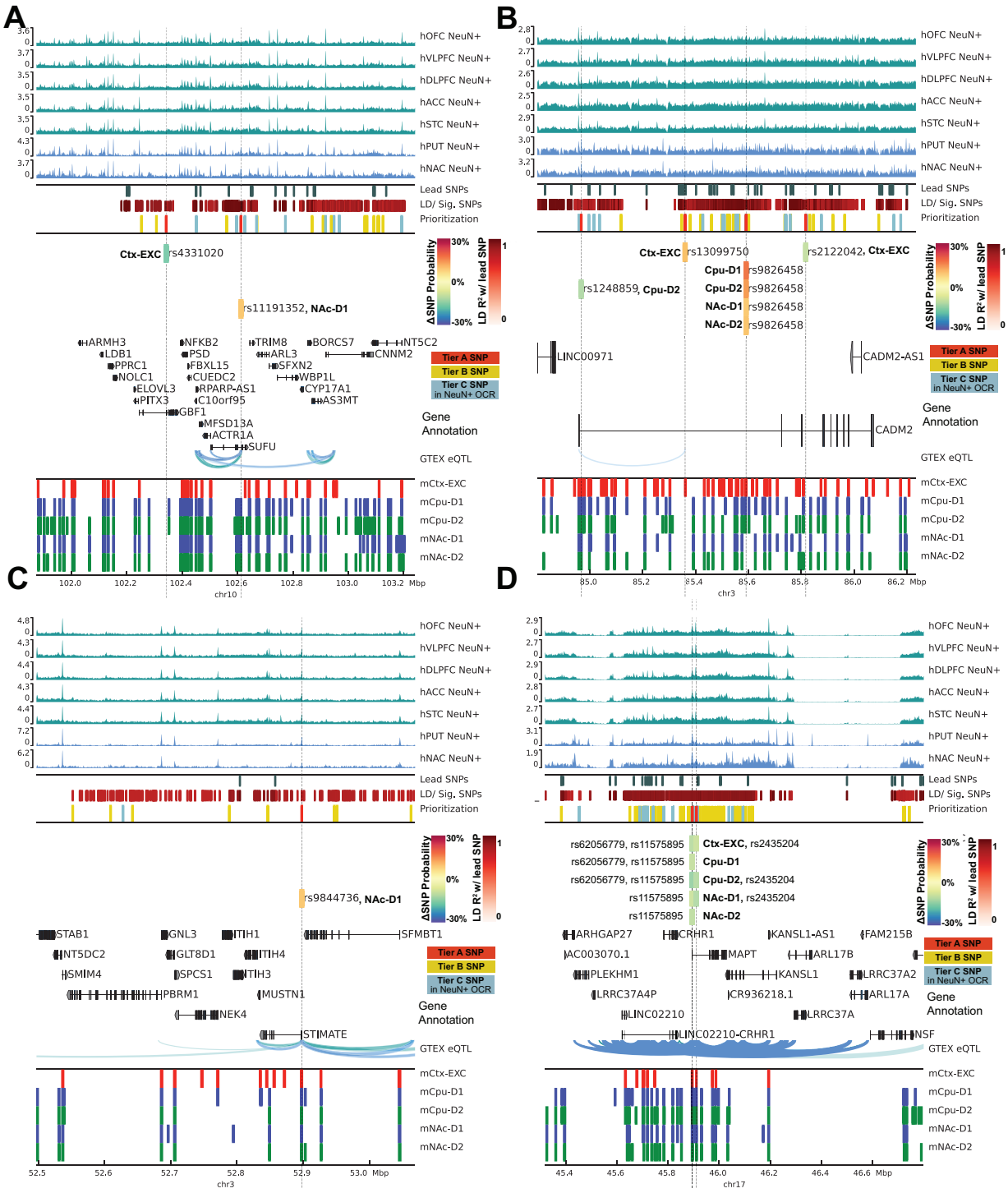


Figure 11. Locus plots of addiction-associated SNPs predicted to act in striatal and cortical cell types.

Locus plot across four additional loci with Tier A SNPs with predicted function SNP impact in cortical excitatory and striatal D1 and D2 MSN cell types. Genome tracks from top to bottom: human (h)NeuN+ MACS2 ATAC-seq fold change signal of cortical and striatal brain regions

*enriched in **Figure** 5A. SNP tracks plot lead SNPs aggregated across seven addiction-associated GWAS, the SNPs in LD with the lead SNPs (Lead SNPs) or independently significant SNPs (LD/ Sig. SNPs). Each SNP is colored by red increasing in intensity by the degree of LD with a lead SNP. Prioritized candidate causal SNPs by predicted differential cell type activity and overlap with Fullard et al. NeuN+ OCRs are plot as (red for Tier A, yellow for Tier B, and teal for Tier C, Methods). Tier A SNP rs7604640 is predicted to have strong ΔSNP effect by CPU-D1 and NAc-D1 CNN models and the bars are colored by the % change in probability active. Gene annotation tracks plot GENCODE genes from the GRCh38 build. eQTL link tracks of FDR-significant GTEX cis-eQTL from cortical and striatal brain regions, and orthologs of mouse (m) putative CREs mapped from excitatory or striatal neuronal subtypes measured by cSNAIL ATAC-seq. NeuN+ ATAC-seq tracks and eQTL links are colored by source brain region as cortical (teal) or striatal (blue). Cell type colors label cortical excitatory neurons (EXC; red), D1 medium spiny neurons (D1; blue), or D2 medium spiny neurons (D2; green).*

2.5 Discussion

In this study, we demonstrate the first analyses integrating neuronal subtype OCRs across human and mouse brain epigenomics using CNN models to select candidate addiction-associated SNPs acting at putative neuronal subtype-specific CREs. We trained CNN models to predict neuronal subtype-specific activity of OCRs and used the models to predict whether addiction-associated genetic variants in risk loci impact putative CRE function. Our findings link the genetic heritability of addiction-associated behaviors to the OCR profiles of neuronal subtypes and brain regions and present specific hypotheses describing how genetic variants may impact gene regulation in addiction-associated behaviors. These analyses in conjunction

suggest that genetic variation-associated nicotine, alcohol, and cannabis use behaviors may impact putative CREs in different combinations of excitatory (EXC), D1, and D2 neuronal subtypes. These findings provide a foundation for future investigations into the cell type-specific genetic mechanisms underlying addiction-related traits. More broadly, our cross-species integrative computational framework leverages high-resolution cell-type targeted epigenomics in model organisms to interpret the genetic risk variants of complex traits in humans.

We initially found that addiction-associated genetic variants were enriched in human NeuN+ OCRs of the prefrontal cortex and striatum, known areas involved in addiction and reward circuitry[149,150] (**Figure** 12A). Genetic variants associated with `SmokingInitiation` and `Cannabis`, initiating behaviors of substance use, were enriched in NeuN+ OCRs of prefrontal areas including DLPFC, VLPFC, and OFC (**Figure** 3A). These OCRs were predicted to be active in cortical excitatory neurons of these brain regions (**Figure** 9B). Addiction-associated genetic variants that enrich in OCRs of cortical excitatory neurons in these areas may reduce corticostriatal activation from prefrontal cortex to inhibit behaviors predisposing the initiation of substance use[63,151]. These genetic variants may contribute to reduced prefrontal self-control reward, leading to behaviors observed in addiction such as impulsivity, reduced satiety, and enhanced motivation to procure drugs. In addition, we found enrichment of striatal NeuN+ OCRs for genetic variants associated with `SmokingCessation` and `DrinksPerWeek` (**Figure** 3A). In **Figure** 9B, we showed that these genetic variants are predicted to affect open chromatin in both D1 and D2 MSNs, which are coordinators of mesocorticostriatal dopamine systems. Genetic variants affecting open chromatin in these MSN subtypes may predispose individuals to increased alcohol use (`DrinksPerWeek`) or decreased nicotine use (`SmokingCessation`), perhaps driving the neuroplastic changes in D1 and D2 MSNs observed in human and rodent drug dependence

studies[152–158]. While drug addiction has been attributed to various areas of the reward circuit, our investigations into heritable genetic risk for addiction-associated traits unravel how regulatory DNA sequence variation in OCRs of projection neurons in implicated areas link genetic risk to neural circuits to behavior.

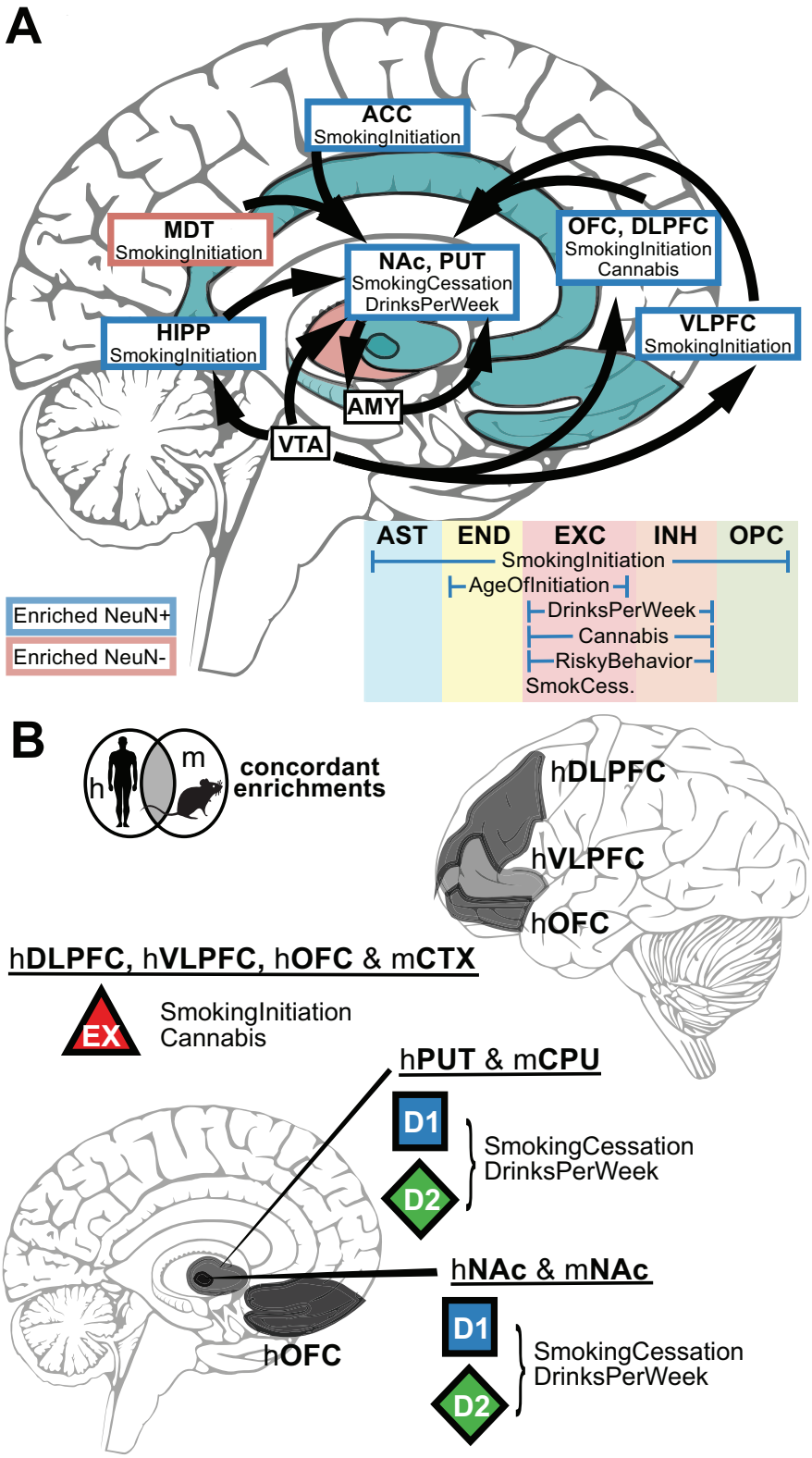


Figure 12. Summary of LDSC GWAS enrichments in human and mouse-human orthologous bulk tissue and cell type open chromatin

(A) Schematic of human NeuN-labeled bulk tissue and occipital cortex cell types from ***Figure*** *2C for which addiction-associated genetic variants were significantly enriched (FDR < 0.05) in OCRs. Brain regions are labelled by the cell type*

that enriched (NeuN+: blue box/shading; NeuN-: red box/shading) spatially along with the trait(s) for which OCRs were found significantly enriched with risk variants. Occipital cortex cell types from ***Figure*** *5C are listed along with the trait(s) for which OCRs were found significantly enriched with risk variants.*

(B) Schematic of addiction-associated genetic variants that share enrichments from human brain regions and neuronal subtypes in both human and mouse-human orthologous open chromatin. Brain graphic adapted from Fullard et al.[35]

Since key component cell types of the reward circuit such as D1 and D2 MSNs have not been profiled for high-quality open chromatin measurements in a human reference genome to our best knowledge, we leveraged high-quality mouse cell type open chromatin measurements using a cross-species OCR mapping framework. We first conducted ATAC-seq of MSN and interneuron subtypes in mouse brain to identify neuronal subtype-specific OCRs. Then, we used a multiple genome sequence alignment framework to identify the orthologous regions of the human genome. By leveraging reference-genome free CRE ortholog mapping tools, we retained high-quality cell type-specific measurements within relevant brain regions of the reward circuit, enabling analysis of cell populations from brain regions where we lack primary human open chromatin profiles. Across these brain regions, we found remarkably concordant enrichments of cell type OCRs between mouse and human profiles as well as shared enrichments between traits (**Figure** 12B). Genetic variants associated with both `SmokingInitiation` and `Cannabis` enriched in mouse-human orthologous OCRs of cortical EXC (**Figure** 6B), concordant with enrichments in human cortical NeuN+ OCRs (**Figure** 3A), which were predicted to include EXC neurons (**Figure** 10B). Genetic variants from these two traits showed replicable enrichment in human EXC neuron OCRs of isocortex and occipital cortex (**Figure** 3B, C), providing strong evidence that genetic variation in cortical

excitatory neuron OCRs confers susceptibility to nicotine and cannabis use behaviors. The enrichments of genetic variants associated with `Cannabis` in isocortical inhibitory neuron OCRs (**Figure** 3B) and mouse-human orthologous OCRs of cortical PV neurons (**Figure** 6B) suggest that genetic variation in cortical PV neuron OCRs also confer susceptibility of cannabis use behavior. Within striatal regions, D1 and D2 MSN mouse-human orthologous OCRs enriched for genetic variants of all measured addiction-associated traits (**Figure** 6A), with strongest concordance in human OCRs for genetic variants associated with `SmokingCessation` and `DrinksPerWeek` (**Figure** 3A, **Figure** 12B). The enrichments in conserved OCRs of MSN subtypes in the dorsal striatum and nucleus accumbens unsurprisingly emphasize known roles of MSNs of both areas to drive and maintain addiction behaviors[159,160]. Our validations of enrichments both at the tissue and cell type level across human and human-orthologous OCRs agree with LDSC regression GWAS enrichments of non-coding regions around differentially expressed genes in DLPFC and NAc measured from postmortem human subjects who were diagnosed with opioid use disorder vs. neuropsychiatric controls[161]. Due to the conservation of reward circuit between mouse and human, our approach is able to unravel the cell types in which genetic variation at the epigenome level predisposes addiction-related traits even from measurements in organisms that have not been exposed to addictive substances. Further, this level of OCR conservation is present at the level of excitatory cell types in cortical brain regions[162]. This may explain why we found enriched cell types in occipital cortex (**Figure** 3C), which is not well-defined for its role in addiction-related traits.

In an orthogonal approach to mapping mouse-human orthologous OCRs, we devised and trained convolutional neural network (CNN) models to classify the neuronal subtype membership of mouse and human NeuN+ OCRs in order to refine GWAS enrichments of bulk tissue to the major neuronal subtypes of cortex and striatum. This approach can provide further

validation for enrichments of human and mouse-human orthologous OCRs in cell types and tissues. Refinement of NeuN+ OCRs revealed that addiction-associated traits enriched for two clusters of cell types and brain regions. The first group, which displays concordant cortical excitatory enrichments between human and mouse, consists of `SmokingInitiation` and `Cannabis` (**Figure** 9B), and the second group, which displays concordant D1 and D2 MSN enrichments, consists of `SmokingCessation` and `DrinksPerWeek`. A draw-back of assigning human NeuN+ OCR membership to individual cell types lies in the considerably low representation of interneurons in both cortical and striatal neuron populations - as low as 12% in neocortex[139,140] and ~5% in striatum[163,164]. NeuN+ open chromatin profiles alone do not always capture OCRs unique to rare interneurons, some of which had OCRs identified by human single-cell assays and mouse-human orthologs enriched for addiction GWAS variants (**Figure** 6B,C). As a result, we did not train CNN models for PV, SST, or VIP interneurons. However, the striking enrichments of OCRs from certain interneuron populations for addiction GWAS variants begin to demonstrate these populations' roles in the addiction neural circuits[165–171]. The overall concordance of enrichments across human and mouse-human orthologous OCRs suggests a conserved regulatory code between mouse and human CREs. Correspondence in the neural circuitry has been well-appreciated between human studies and rodent models of addiction[172,173], and our study further demonstrates that mouse-human conserved OCRs may explain considerable heritability of addiction-associated traits. This makes animal models suitable not only for studying the neural circuits of addiction but also cell-type-specific gene-regulatory mechanisms of addiction.

We used several selection criteria along with CNN models to predict the functional impact of genetic variants associated with addiction-related traits (**Figure** 12, **Figure** 8, **Figure** 8-1). The fine-mapping pipeline described effectively narrows down a set of 14,790 SNPs to a putatively functional set of 55 Tier A candidate causal SNPs that can be experimentally tested

to determine which brain regions and neuronal subtypes they would have function in. The candidate functional SNPs that our models prioritize demonstrate how a candidate SNP within a locus, such as rs7604640 (**Figure** 10B), might act in distinct neuronal subtypes and brain regions. Cell type-and brain region-specificity adds complexity to identifying how genetic variation may alter gene regulation to predispose individual to addiction-associated traits. Our approach often reported one to four candidates per loci, even in stretches of SNPs in perfect LD such as the MAPT-CRHR1 locus (**Figure** 8D). This reflects the idea that many SNPs in the same loci are significantly associated with addiction due to LD with only one or a few causal SNPs and are unlikely to influence addiction-associated genetic predisposition. We report many candidate SNPs that also overlap mouse-human orthologs from the same predicted cell type raise the idea that altering the conserved regulatory DNA sequence may be a mechanism of cell type-specific gene regulatory tuning in a population or even across species[174].

Our study depends solely on assays of open chromatin as a proxy for putative CREs. While our study utilizes well characterized cell types from *cre*-driver lines against a C57BL/6J genetic background, we recognize the limitation of relying on one mouse strain in light of evidence that mouse genetic backgrounds have unique tissue-specific open chromatin[175]. Epigenetic assays for chromatin conformation, histone modifications, and methylation would further inform how putative CREs regulate nearby gene expression. While eQTL studies do not control for inflated associations due to LD and report gene expression differences from bulk tissue, we do note that our approach prioritizes several SNPs known to be *cis*-eQTLs in relevant brain regions, which indirectly affirms our framework's ability to select SNPs with functional impacts on gene regulation. Although *cis*-EQTLs are often not cell type- or tissue-specific, our findings of risk loci in brain regions implicated in addiction-related traits reflect a strength of our approach in discerning brain-specific signal. In order to validate our predictions,

it will be necessary to further investigate candidate genetic variants such as rs7604640 (**Figure** 10B) in future studies using a fluorescence reporter assay or *in situ* hybridization studies. These methods can measure regulatory activity differences between risk and non-risk alleles to verify our predictions of SNP impact on putative CREs and indicate whether the reported differences in regulatory activity are cell type-specific. The candidate SNPs we identified provide possible mechanisms linking differences in genetic make-up with the genes, cell types, and brain regions that could influence addiction and substance use behaviors (**Figure** 10).

2.5.1 Data availability

Code used to run intermediate and final analyses reported in this paper are available on the GitHub page: https://github.com/pfenninglab/addiction_gwas_enrichment. Sequencing output files for data generated in this study are deposited on the GEO at accession GSE161374. Questions and comments about data and analyses may be directed to the corresponding author.

2.6 Acknowledgement:

We would like to thank members of the Eric Yttri lab and Aryn Gittis lab at Carnegie Mellon University for providing *Drd1-cre, Adora2a-cre,* and *PValb-2a-Cre* mice for cell type-specific ATAC-seq experiments.

2.6.1 Author Contributions:

Conceptualization: ARP, BNP, CS; ATAC-seq data processing: AJL, ER, IMK, BNP; GWAS enrichment investigation: BNP, CS, ER, MK; Machine learning models: AJL, BNP, ER, IMK; bulk tissue ATAC-seq: AJL, ARB, MEW; cSNAIL ATAC-seq: AJL, ARB; writing

(original draft): BNP, CS; review and editing: AJL, ARP, BNP, CS, ER, IMK; funding acquisition, resources, & supervision: ARP;

2.6.2 Funding:

National Institute of General Medical Sciences training grant T32GM008208 (BNP), National Institute On Drug Abuse of the National Institutes of Health Award Number F30DA053020 (BNP). Sloan Foundation Fellowship (ARP), National Institute on Drug Abuse Avenir Award 1DP1DA046585 (ARP), National Science Foundation Graduate Student Research Fellowship DGE1745016 (AJL), Carnegie Mellon Brainhub Presidential Fellowship (ER), Carnegie Mellon Computational Biology Department Lane Postdoctoral Fellowship (IMK)

2.6.3 Conflicts of Interest:

AJL, ER, and ARP are inventors of the cSNAIL patent. Other authors do not declare any conflict of interest.

3.0 Fine-mapping candidate neuropsychiatric regulatory variants using cell type-aware comparative genomics

BaDoi N. Phan[1,2], Alyssa J. Lawler[3,4†], Jing He[5†], Ashley R. Brown[1,†], Irene M. Kaplow[1,6], Amanda Kowalczyk[1,6], Chaitanya Srinivasan[1], Grant A. Fox[1], Daniel E. Shäffer[1], Ziheng Chen[3], William R. Stauffer[5], Andreas R. Pfenning[1,6]*

Affiliations:
[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University; Pittsburgh, PA, USA.
[2]Medical Scientist Training Program, School of Medicine, University of Pittsburgh; Pittsburgh, PA, USA.
[3]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University; Pittsburgh, PA, USA.
[4]Current Affiliation: Broad Institute of Harvard and MIT; Cambridge, MA, USA.
[5]Neurobiology Department, School of Medicine, University of Pittsburgh; Pittsburgh, PA, USA.
[6]Neuroscience Institute, Carnegie Mellon University; Pittsburgh, PA, USA.
[7]Current Affiliation: Allen Institute; Seattle, WA, USA.
†These authors contributed equally
*Corresponding author. Email: apfenning@cmu.edu

3.1 Abstract

Measures of nucleotide sequence conservation across species are useful for identifying functional genomic loci, but can fail when regulatory function is maintained, often in a cell-type specific manner, even when sequence is not. We introduce Cell-TACIT, the Cell Type-Aware Conservation Inference Toolkit, to identify trait-associated regulatory variants. Cell-TACIT integrates sequence conservation scores with cell type-specific open chromatin data collected from a few mammalian species to impute function for hundreds more. Applying Cell-TACIT to neuropsychiatric trait loci identifies higher heritability enrichment and more fine-mapped variants than nucleotide conservation and human chromatin data alone. Our *in vivo* reporter assays validate predictions for enhancers with risk variants near the *DRD2* schizophrenia risk locus. By integrating genome conservation and multi-species open chromatin data, Cell-TACIT prioritizes variants within regions of conserved regulatory function for *in vivo* characterization and addresses a major challenge in translating disease associations to mechanistic understanding.

3.1.1 One-Sentence Summary

Regulatory functions conserved across mammals underlying cell type evolution reveal functions of human risk variants.

3.2 Main Text

As the number of subjects involved in genome-wide association studies (GWAS) of neuropsychiatric traits increases, a greater number of associated loci are detected[176]. However, only 4.2% of associated single nucleotide polymorphisms (SNPs) are in coding regions[177]. The rest fall outside known protein-coding genes, often in regulatory elements that significantly contribute to the heritability of common traits and disorders[23,74,178]. To interpret the impact of regulatory SNPs, DNA sequence conservation metrics across primates, mammals, and vertebrates have provided powerful lenses for inferring the risk impact of genetic variation within the human genome, as elaborated in a companion paper[74,179–184]. However, genome constraint approaches cannot resolve which tissues or cell types are implicated in the trait or disease process [185]. This presents particular challenges for identifying the cell types affected by risk SNPs in the brain, which has extreme cell type diversity. A complementary approach measures the cell type-specific activity of putative *cis*-regulatory elements using assays like transposase-accessible chromatin with sequencing (ATAC-seq) and has proved increasingly useful for interpreting the cell type targets of human traits [88,186–192]. Integrating comparative genomics and cell type-specific experimental evidence holds great promise for inferring the impacts of genetic variation, but empirical studies of accessible chromatin are resource prohibitive, severely limiting the cell types and species for which data are available.

Despite several hundred million years of divergence, some genomic elements retain cell type-specific gene regulation across Metazoa by shuffling and reforming critical transcription factor-binding sites [193], suggesting that borrowing information across species could help make up for the absence of extensive empirical data on gene regulatory function. Since conservation of regulatory function does not always require nucleotide-level sequence conservation [138,185,193], machine learning models are useful for inferring functional

conservation from nuanced sequence patterns. We previously demonstrated that convolutional neural networks learn transcription factor binding sites related to cell type-specificity [187] and accurately identify orthologous DNA sequences across species that retain tissue-specific open chromatin activity [185].

We present the Cell Type-Aware Conservation Inference Toolkit (Cell-TACIT), a machine learning framework that learns the conserved functional activity of non-coding genomic sequence from cross-species matched cell type open chromatin data and imputes regulatory activity across hundreds of placental mammals included in a reference-free genome alignment (**Figure 1A-C**). In a companion paper[194], we demonstrate our ability to predict differences in tissue- and cell type-specific open chromatin at orthologous genomic regions of other species and use those predictions to identify regulatory loci undergoing convergent evolution. Here, we leverage both comparative genomics and cell type-specific experimental evidence to infer the impacts of non-coding genetic variation in neuropsychiatric human traits.

We first estimated SNP-heritability using nucleotide conservation metrics, specifically phyloP[183] computed from 240 placental mammal genomes. These phyloP scores showed unprecedented power to interpret human trait-associated loci[184] but had inconsistent heritability enrichment for neuropsychiatric traits (**Figure 1D**, left, phyloP.cons from 240 mammals, Alzheimer's Disease heritability enrichment = 69.7; schizophrenia HE = 3.82)[43]. In the case of schizophrenia, others have speculated that evolutionary annotations are not capable of substantially explaining the heritability of schizophrenia due to the human-specificity of the phenotype of implicated loci [195–198] despite success from analyses from mouse brain single-cell transcriptomic profiles to identify putative cell types underlying schizophrenia [199–201]. The framework of Cell-TACIT provides that opportunity to compare two alternative models for the evolution of regulatory elements associated with schizophrenia and other neuropsychiatric

traits: do the regulatory functions of these regions emerge anew in the human lineage, or do they have deeply conserved function that is often missed by traditional methods?

To provide a foundation for comparative genomic analysis of neural cell types, we built a cross species atlas of chromatin accessibility of the caudate nucleus, a brain region with well-established cell type homologies[187,202] that has been heavily implicated in psychiatric disorders, addiction behavior, and sleep behavior [87,187,199,200,203]. We performed single nucleus assays for transposase-accessible chromatin (snATAC-seq) experiments in rhesus macaque (**Appendix B Figure S1A-B**) and combined this with published snATAC-seq from human and mouse [88,186,187]. We found 8 caudate cell types at a resolution with matched orthologs across humans, rhesus macaques, and mice (**Figure 1A**, **Appendix B S1C-F**, Supplemental Text, Cross-species caudate cell type ortholog identification). The neuronal subtypes consist of interneurons, the majority of which are *PVALB*$^{+}$, canonical medium spiny neurons (MSNs) marked by dopamine receptors D1 and D2 [62,202], and a distinct population of neurons that shares features with D1/D2-hybrid medium spiny neurons previously reported in macaque and mouse single cell transcriptomic studies [88,204–206].

We mapped human open chromatin regions (OCRs) from caudate cell types across 240 Zoonomia genomes to find conserved cell type regulatory grammar, i.e. combinations of transcription factor binding sites that drive expression (**Figure 1B**) [45,207]. The OCRs of each cell type encompass many cell type-specific transcription factor binding motifs across species (**Appendix B Figure S2, Appendix C Data S1)**, and the subset of human cell type OCRs mappable to rhesus macaque or mouse overlap OCRs measured active in those species (**Appendix B Figure S3)**. These data show caudate neural cell types retain distinct regulatory codes within the Euarchontoglires genomes, indicating that Cell-TACIT models can infer cross-species open chromatin status from orthologous DNA sequences. We trained Cell-TACIT models for each neural cell type and demonstrated their accuracy at predicting cell

type-specific open chromatin status of orthologous DNA (**Figure 1C**, **Appendix B Figure S4-S8**, **Appendix C Data S2-3**, Supplemental Text, Cell-TACIT performance evaluation).

We conducted stratified LD score regression with caudate cell type OCRs in 52 neuropsychiatric traits spanning neurodegeneration, psychiatric disorders, substance abuse disorders, sleep and neurological traits, and 12 non-brain traits (**Figure 1D, Appendix C Data S4**). We find that conditional independent regression effect sizes are larger in shared human-macaque or human-mouse cell type OCRs than OCRs from any one species alone (**Appendix C Data S5,** Supplemental Text, Pairwise cross-species LD score regression). This suggests that the conservation of cell type-specific open chromatin can help to finemap loci identified through GWAS for neuropsychiatric disorders.

Using Cell-TACIT predictions of caudate OCRs, we created conserved cell type open chromatin annotation tracks for each clade of mammals increasingly diverged from humans (**Appendix C Data S6**). We show in both brain- and non-brain related traits that cell type OCRs with orthologs predicted to be active in distant species had higher heritability enrichment than other human cell type OCRs (**Figure 2**, colored bars; **Appendix B Figure S11-26** "Cell-TACIT Score"). We found an increase in heritability enrichment in the cell type of likely etiology, such as Alzheimer's Disease (AD) with microglia ($HE_{AD\text{-}Micro}$ = 177 - 278) [208,209], schizophrenia with D2 MSNs ($HE_{SCZ\text{-}D2}$ = 64.1 - 93.8) [201], and SmokingInitiation with D1 MSNs ($HE_{SmokInit\text{-}D1}$ = 19.5 - 43.1) [67]. Since it may be more difficult to map OCRs to species more distantly related to humans, this trend may have been driven by biases in evolutionary distance instead of biologically meaningful differences across species. We therefore confirmed that heritability enrichments are not higher due to mappability to distant species (**Figure 2**, black bars, $HE_{AD\text{-}Microglia}$ = 126 - 166; $HE_{SCZ\text{-}D2}$ = 58.8 - 62.2; $HE_{SmokInit\text{-}D1}$ = 20.8 - 24.2) [43]. Notably, these heritability enrichments are substantially higher than enrichments calculated when using single-nucleotide PhyloP conservation (**Figure 2**, left). Overall, the heritability

enrichments from Cell-TACIT annotations suggest that the conserved regulatory code of distinct cell types, rather than nucleotide-level genome sequence conservation, holds considerably more genetic heritability for neuropsychiatric traits.

We summarized the predicted conserved activity across placental mammals of each non-coding, distal open chromatin region into the Cell-TACIT Age, a metric capturing the million years divergence from humans weighted by predicted activity (**Figure 1E**). In essence, the Cell-TACIT Age measures the millions of years that a human OCR has persisted to retain activity in a cell type of another species. We conducted an orthogonal set of heritability enrichment analyses excluding coding regions and promoters, which have different levels of evolutionary constraints, and binning open chromatin regions into equal Cell-TACIT Age quartiles to control for annotation size. Despite lower power for genome-wide enrichment analyses by using fewer OCRs per bin, we find that neuropsychiatric traits have higher heritability enrichments in "older" Age quartiles. Such traits include self-reported depression in D2 MSNs, age of smoking initiation in D1 MSNs, and cross-psychiatric disorders in D1 Hybrid MSNs (**Appendix B Figures S11-26**, "Cell-TACIT Age" columns). The full set of heritability enrichment analyses comparing various cross-species snATAC-seq, Zoonomia PhyloP, and Cell-TACIT annotations are reported in the Supplement (**Appendix B Figure S11-26, Appendix C Data S7**).

We examined the value of Cell-TACIT Age to identify functional genetic variation by comparing to two expression quantitative trait loci (eQTL) datasets [210,211]. We first looked at Age quartiles that overlap eQTLs from the caudate nucleus from the GTEx Consortium (**Methods**). We found that caudate GTEx eQTLs were uniformly enriched in the "youngest" Age quartiles of all cell types but mostly depleted from "older" Age quartiles (**Figure 3A**, logistic regression Bonferroni p-value < 0.05, **Appendix C Data S8**). Since Cell-TACIT models were at the resolution of neural cell types, the lack of enrichment in "older" Age

quartiles may be due to differences between bulk tissue and single cell gene expression. We then compared eQTLs from single–cell RNA-seq of the human prefrontal cortex where four major glial cell types have one-to-one correspondence with our models. The matched Cell-TACIT models had the top enrichment for cell type eQTLs (**Figure 3B**, Bonferroni p-value < 0.05, **Appendix C Data S8**). This effect was notable for microglia and oligodendrocyte eQTLs. Astrocyte eQTLs had lower enrichment for astrocyte Age quartiles, perhaps because astrocytes have known brain region-specific gene expression, especially between the prefrontal cortex and the caudate nucleus [212,213]. Interestingly, "older" Age quartiles in matched cell types have decreasing enrichment for eQTLs–suggesting that predicted "younger" regulatory regions have less time for negative selection and enjoy variable effects on gene regulation. For cell type eQTLs that do not have a matching Cell-TACIT model, "older" Age quartiles are not enriched for eQTLs (**Appendix B Figure S28A**, Bonferroni p-value < 0.05). While imperfect, comparing to brain and cell type eQTLs demonstrates that Cell-TACIT Age is specific at ranking genetic variants with regulatory function.

We used the Cell-TACIT Age to prioritize SNPs from the subset of neuropsychiatric traits with moderately- and well-powered GWAS. We performed functionally informed fine-mapping using our cross-species caudate cell type OCRs and updated annotations of primate and mammalian genome sequence constraint (**Appendix B Figure S27, Appendix C Data S9-10,** Supplemental Text, Polyfun Fine-mapping) [214,215]. Fine-mapped neuropsychiatric trait SNPs relative to other annotations or non-brain traits are more likely to be enriched in "older" Cell-TACIT Age quartiles (**Figure 3C,** FDR < 0.05, **Appendix C Data S8**). "Older" medium spiny neuron Age quartiles had the highest enrichments for SNPs associated with psychiatric, sleep, substance abuse, and general neurological traits. In addition, two microglia Age quartiles were enriched for neurodegenerative disorder-associated SNPs. Fine-mapped neuropsychiatric trait SNPs were not enriched in other neural cell types, likely indicating they are not the

etiological targets of risk variants (**Appendix B Figure S28B,** FDR > 0.05). Furthermore, the opposing trends of enrichment in Cell TACIT Age regions between fine-mapped GWAS SNPs and eQTLs may be due to differences in selection pressure on genetic variation contributing to human disorders and unassociated variants [216,217].

A cluster of pleiotropic SNPs near the *DRD2* locus were fine-mapped as candidate causal SNPs for multiple traits [67,113,201,218]. Dopamine signaling and changes in dopamine receptor D2 availability in the caudate nucleus have long been associated with schizophrenia [219,220], with recent findings showing that mouse MSN marker genes were enriched for schizophrenia risk SNPs [199,200]. These SNPs overlap many D2 MSN OCRs, some predicted to remain active across placental mammals and some predicted to lose activity in species diverged from humans (**Figure 3D**). The top two fine-mapped SNPs, rs7933981 and rs61687445, overlap the same open chromatin region accessible across MSNs (**Figure 4A,** Candidate A). A nearby SNP, rs7928017, is in perfect linkage disequilibrium with the top variants and overlaps another MSN OCR (**Figure 4A,** Candidate B). All three SNPs are eQTLs for *DRD2* in the postmortem human brain [210]. The phyloP conservation scores at these SNPs indicate that these nucleotides are conserved across placental mammals at the single-nucleotide level (candidate A $\text{phyloP}_{rs7933981} = 7.0$ and $\text{phyloP}_{rs61687445} = 3.01$; candidate B $\text{phyloP}_{rs7928017} = 2.34$) [183], and only the first two SNPs are conserved in primates using the PhastCons method, which accounts for conservation of neighboring nucleotides (**Figure 4A**, $\text{PhastCons}_{rs7933981} = \text{PhastCons}_{rs61687445} = 1.0$, $\text{PhastCon}_{rs7928017} = 0$) [182]. Multiple public functional annotations available through the UCSC genome browser indicate that candidate B has more regulatory activity than Candidate A (**Appendix B Figure S29**). When statistical fine-mapping, genome conservation, and epigenomic annotations provide equivocal or conflicting evidence of SNP impact, we propose that Cell-TACIT can shed light on cell type-specific regulatory function.

In contrast to existing functional evidence, we calculated a higher neuronal Cell-TACIT Age for candidate A relative to B (**Figure 4C**). Both the human and mouse orthologs of candidate A have higher predicted neuronal activity than B (**Figure 4D**, **Appendix C Data S11**, HSA and MMA vs. HSB and MMB). We compared our predictions with bulk ATAC-seq of the caudate nucleus in Egyptian fruit bats conducted by a companion study. We measured multiple reproducible open chromatin regions at the fruit bat ortholog of candidate A yet none at the ortholog of candidate B. Even though Cell-TACIT models were trained on data from two primate species and mouse, they correctly predicted conserved open chromatin activity at this locus in a bat species, 94 million years diverged from humans.

To test cell type-specific gene regulatory function of the two candidate regulatory elements upstream of the *DRD2* locus, we performed adeno-associated virus-based *in vivo* reporter assays of the two candidate enhancers (**Figure 1F, 4B, Appendix B Figure S30**). We co-stained transduced mouse brains for the reporter mCherry and the neuronal marker NeuN and found that both candidate sequences were able to express the mCherry protein; namely, they act as enhancers (**Figure 4E, Appendix B Figure S31**). We detected more mCherry+ nuclei in injections of candidate A relative to candidate B (linear mixed effect t-test $p_{A\ vs.\ B}$= 0.012, N=3 per reporter, **Appendix C Data S12**). Whether we used the human or mouse ortholog of the enhancers did not significantly affect the detected level of neuronal mCherry+ nuclei (linear mixed effect t-test $p_{human\ vs\ mouse}$ = 0.914, **Appendix C Data S12**). We noticed that the predicted trends in neurons and glia correspond to the measured mCherry+ nuclei counts: Cell-TACIT had predicted weak glial gene regulatory activation and stronger neuronal activation, and the correlation between Cell-TACIT predictions and the mCherry+ nuclei measurements is strong (Pearson's R = 0.794, p = 0.019, **Figure 4F**).

Our *in vivo* reporter experiments demonstrate that Cell-TACIT quantitatively predicts the conserved gene regulatory activity of both human and mouse orthologs of a candidate

sequence in the correct cell type-specific expression patterns. Both tested human sequences have varying degrees of cell type-specificity, so the fine-mapped SNPs within these enhancers may affect *DRD2* transcription in caudate neurons at either or both enhancers. In addition to the *DRD2* locus, we highlight other loci with fine-mapped SNPs with predicted cell type-specific gene regulatory impacts (**Appendix B Figure S32-34**, Supplemental Text, Additional fine-mapped loci).

The case of schizophrenia, where deficits affect higher cognition and is largely human-specific [196,198,221], highlights the power of Cell-TACIT to interpret the genetic basis of complex neuropsychiatric traits. Although some schizophrenia variants have nucleotide-level conservation, more tend to be found at loci where open chromatin is predicted to be highly conserved across placental mammals in specific caudate neuron types. Our findings support the biological model where regulatory elements containing trait-associated variants can tolerate turnover at individual nucleotide positions while maintaining regulatory activity in a particular cell type across distant species. Cell-TACIT enables future work using chromatin accessibility data from matched cell types in multiple species to prioritize human trait-associated variants for *in vivo* characterization and untangle their roles in complex tissues and traits.

3.3 Supplementary Text

3.3.1 Cross-species caudate cell type ortholog identification

We observed that using more highly variable features used to compute the PCA dimensionality reduction during `ArchR`'s iterative LSI clustering (`varFeatures`) enabled us to find discriminating features for separating D1 and D2 medium spiny neuron (MSN) clusters in caudate samples. (Using fewer variable features during iterative LSI clustering separated glia from neurons, producing only 1 MSN cluster.) From the clustering results that used more variable features, we were clearly able to discern the same 8 caudate subtypes in human and rhesus snATAC-seq. We aligned the He and Kleyman *et al.* D1/2 hybrid subtype, which corresponds to Corces *et al.* striatonigral MSN subtype, using marker gene activity scores (e.g. *RXFP1*, *FOXP2*, and *DRD1*, **Appendix B Figure S1C-D**) and the high-dimensional co-clustering the `ArchR` cross-cell type "`PeakMatrix`" subsetted to peaks with unique, usable OCR orthologs between human and rhesus macaque, described below [88,206,222]. All other macaque and human caudate cell types also corresponded to a unique rhesus macaque cell type cluster. These orthogonal computational approaches suggest that caudate cell types are well-conserved between humans and rhesus macaques.

When we performed the sub-clustering for mouse snATAC-seq as for human and rhesus macaques, we were unable to resolve D1 vs. D2 subtypes. Neither the aggregate cell type track plots for D1 and D2 MSN subtypes around marker genes such as *Drd1* and *Drd2* nor the per-cell gene activity scores in these genes corroborated the published CATlas snATAC-seq labels. We suspect the lower per-nuclei fragment depth (<10,000 unique fragments per nuclei) resulted in low sampling of accessible chromatin in D1- and D2-specific OCRs. However, we were able to identify the CATlas-labeled neuronal "MXD" L2-cluster subtype (with high *Tac1*, *Foxp2*, and *Chrm4* gene activity scores found restricted to the mouse

striatum) as the orthologous cell types to the primate D1H MSN cell type by marker gene accessibility (**Appendix B Figure S1F**). From this CATlas snATAC-seq dataset of the mouse caudoputamen, we were able to identify the striatonigral MSN, astrocytes, oligodendrocytes, oligodendrocyte precursors, and microglia cell types. We called reproducible peaks as above for these 5 cell types. We paired these snATAC cell type OCR profiles with high quality cSNAIL ATAC-seq from validated D1-*cre*, A2a-*cre* (D2), and Pvalb-*cre* transgenic mouse lines (**Appendix B Figure S1E**) [187].

We mapped the human, rhesus macaque, and mouse cell type open chromatin regions (OCRs) across species to demonstrate that the cell types share a common regulatory code. Roughly 90% of the OCRs are mappable within primates and 60% between human and more distantly related mammalian species, including mice (**Figure 1B**). We computed the motifs underlying the accessible peaks, including the MSN D1H subtype, we found a large number of cell type-specific motifs (found in fewer than 4 cell types) discovered within each species that were also identified in another species **(Appendix B Figure S2**). The OCRs mappable to rhesus macaque and mouse genomes have similar proportions of promoters and distal elements, including those peaks that are shared between human-rhesus macaque and human-mouse (**Appendix B Figure S3**). These data together at single cell and cell type-specific ATAC-seq strongly suggest that we are able to identify orthologous cell types across species, the cell type-specific marker peaks are shared across species, and the transcription factor motifs that make up the cell type-specific regulatory code are conserved.

3.3.2 Cell-TACIT performance evaluation

We trained Cell-TACIT models for each caudate cell type (**Figure 1C**, **Methods**)[194] using only human sequences to determine if a model trained using data from only one species could make accurate predictions in species not used in training [185,223]. Our positive set was our

human OCRs, and our negative set was the human sequences from the final negative set nonCellEnhLargeGC. These models, which refer to as HgOnly models, had decent accuracy for predicting open chromatin (auROC = 0.854 ± 0.0580, tpr = 0.751 ± 0.0916, mean ± st. dev.). Since these models were evaluated using an 9.46-to-1 negative-to-positive ratio, we also evaluated the area under the precision-recall curve (auPRC), F1 score, and false positive rate (fpr) and found that those were generally high considering how imbalanced the dataset was (auPRC.adj = 0.420 ± 0.157, f1_score = 0.829 ±, 0.0578) and had low false positive rates (fpr = 0.204 ± 0.0811). These models also achieved performance that was substantially better than expected by chance for predicting cell type-specific open chromatin.

Since previous studies have suggested that models trained on data for multiple species provide more accurate open chromatin predictions than models trained using data from only 1 species [185], we then used our data from human, macaque, and mouse to train Cell-TACIT models for each caudate cell type (**Figure 1C**, Methods) [194], using OCRs from human, macaque, and mouse against the final negative set nonCellEnhLargeGC. Across the 8 cell types, the HgRmMm models had even higher performance for predicting open chromatin than the HgOnly models did (auROC = 0.880 ± 0.0528, tpr = 0.750 ± 0.0889). Here, HgRmMm models were evaluated using an 11.1-to-1 negative-to-positive ratio–150k validation negatives per cell type, 15k validation positives per cell type) and found that performance was also generally high for such an imbalanced dataset (auPRC.adj = 0.465 ± 0.177, f1_score = 0.864 ± 0.0338, fpr = 0.159 ± 0.0501). **Appendix B Figure S4** and **Appendix C Data S2** further breaks down the model validation set performance during negative set selection and hyper parameter tuning across all negative sets for HgOnly and HgRmMm models to show that the final negative set maintains a balance of sensitivity and specificity genome-wide to predict

open chromatin activity. **Appendix C Data S3** reports the general test set performance metrics of the `HgOnly` and `HgRmMm` models using the `nonCellEnhLargeGC` negative set.

To assess cell type-specificity of our Cell-TACIT models, we predict OCRs accessible in only the target cell type as open and OCRs only accessible in all other cell types as closed. Based on this definition of cell type exclusive positive OCRs and union of any other OCR negatives, we identified on average 26.5k cell type validation negative OCRs and 748 cell type validation positive OCRs. `HgOnly` models had cell type-specific true positive rates of 0.714 ± 0.116, true negative rates of 0.757 ± 0.119, and negative predictive value of 0.988 ± 0.0103. `HgRmMm` models had cell type-specific true positive rates of 0.731 ± 0.0860, true negative rates of 0.799 ± 0.0707, and negative predictive value of .988 ± 0.016. The gain in cell type-specific `tnr` and `tpr` suggests that `HgRmMm` models are able to be more specific and sensitive. Since we also had more negatives than positives for this evaluation, we evaluated the `auPRC` for this task and found that it was generally high ($\texttt{auPRC.adj}_{\texttt{HgOnly}}$ = 0.186± 0.163, $\texttt{auPRC.adj}_{\texttt{HgRmMm}}$ = 0.245 ± 0.197). **Appendix B Figure S5** and **Appendix C Data S2** further breaks down the cell type-specific validation set performance during negative set selection and hyper parameter tuning across all models to conclude that the final set of models do maintain cell type specificity in OCR predictions. **Appendix C Data S3** reports the cell type-specific test set performance metrics of the `HgOnly` and `HgRmMm` models using the `nonCellEnhLargeGC` negative set.

To assess the ability of our Cell-TACIT models to detect species-specific open chromatin changes, we predict the activity state in cell type OCRs where the ortholog in another species is accessible or where the ortholog is inaccessible. Based on this definition of OCR cross-species change, we identified on average for a cell type 3,990 species-specific validation OCRs where in one species had lost OCR activity and 3801 validation OCRs where there is

sustained activity across species. The HgOnly models had an average true positive rate of 0.757 ± 0.152, true negative rate of 0.5559 ± 0.160, and negative predictive value of 0.698 ± 0.206, **Appendix B Figure S6B**. The HgRmMm models had an average true positive rate of 0.735 ± 0.0185, true negative rate of 0.624 ± 0.149, and negative predictive value of 0.713 ± 0.204, **Appendix B Figure S6A**. These results suggest that, with more species-specific examples, the models are better able to detect accurate loss in cell type OCR activity across species. Curiously, the HgRmMm cross species model was better at predicting true maintenance or loss of OCR activity in rhesus macaque and mouse (tnr_{hg38} = 0.526, $tnr_{rheMac10}$ = 0.648, tnr_{mm10} = 0.697, tpr_{hg38} = 0.612, $tpr_{rheMac10}$ = 0.849, tpr_{mm10} = 0.743). This may reflect the underlying data quality where the human snATAC-seq samples came from frozen postmortem human brains while the rhesus macaque snATAC-seq and cSNAIL mouse ATAC-seq came from freshly dissected tissue.

When we stratified species-specific prediction accuracy across cell types, the HgOnly often successfully predicted changes in chromatin status between species, and the HgRmMm models were even more specific (**Appendix B Figure S6C**). D1 MSN, D2 MSN, and Pvalb interneuron models had high auROC and auPRC across all three species OCR sets. Species-specificity auROC and auPRC in human D1H MSNs and glia were lower compared to the rhesus macaque and mouse OCRs sets. The higher auROC and auPRC in rhesus macaque and mouse suggests the models may be able to predict cross-species changes well within the primate lineage in spite of the poor human performance, which could be a result of the data coming from frozen brain tissue. **Appendix B Figures S7A** and **S7B** and **Appendix C Data S2** further breaks down the species-specific validation set performance during negative set selection and hyper parameter tuning across all models to conclude that the final set of models do maintain cell type specificity in OCR predictions. **Appendix C Data S3** reports the species-

specific test set performance metrics of the `HgOnly` and `HgRmMm` models using the `nonCellEnhLargeGC` negative set.

Lastly, we assessed the phylogeny-matched correlations by determining the correlation between the model predictions and the years diverged from source OCRs. We do not expect all species to use the same set of OCRs across species [207]. Furthermore, we expect the OCRs measured in one species will be less active on a genome-wide level when mapped across species in more distant species due to evolutionary drift and usage of cell type-specific gene regulation [177]. We mapped human OCRs across species and observed the expected decline in the number of OCRs that can be mapped in increasingly distant species (**Figure 1B**, **Appendix B Figure S7**). This mappability may also be lower in certain species due to lower quality genomes or representation within Zoonomia [214]. From these orthologous DNA regions, we found a negative correlation with the average Cell-TACIT raw scores of each species and the number of years diverged from humans, **Appendix B Figure S8A**. When we performed the calibration on the Cell-TACIT scores, we found that the expected negative correlation improved and were comparable across cell type models (**Methods**, **Appendix B Figure S10B**). This phylogeny-matched correlation analysis suggests our models were able to predict the regulatory decay of cell type OCRs across species.

3.3.3 Pairwise cross-species LD score regression

We applied genome-wide SNP-heritability enrichment analyses using stratified LD score regression (LDSC) on human neurological and psychiatric traits comparing the changes in pairwise OCR overlap in three species to human OCRs to complement enrichment analyses made in OCRs that to be active across hundreds of mammalian genomes using Cell-TACIT. We considered annotations using human cell type OCRs, mappability of cell type OCRs to distant mammalian genomes, and conservation of cell type OCRs in model species (Methods).

Neurological and psychiatric traits with a wide range of SNP heritability and power to detect genome-wide significant loci were selected across GWAS (**Appendix B Figure S9A**), and their genetic architectures were shared between traits (**Appendix B Figure S9B**). We estimated contributions to GWAS traits with stratified LD score regression (LDSC) using the per-SNP independent effect of having a risk SNP in cell type OCRs, τ^*, conditioned on the baseline annotations of the genome (Methods). We compared the changes to τ^* from partitioning GWAS traits by cell type OCRs from humans, mappable to or from rhesus macaque or mouse, and cell type OCR orthologs shared between humans and the model organism (numerical results in **Appendix C Data S5**). First, we found that neuronal OCRs tend to be significantly enriched across the panel of largely brain-related traits (**Appendix B Figure S10A**) compared to glial OCRs (**Appendix B Figure S10B**). Neuronal cell type OCRs measured in macaque and mappable to human or measured in humans and mappable to macaque had larger effect sizes than the larger set of human OCRs (**Figure S10A**, columns 1-2, **Appendix C Data S5**, linear regression slope= 0.36-0.89, FDR= 0.0044 - 4.68 x 10^{-20}). OCR mappability to mouse had even larger effect sizes than human cell type OCRs (**Appendix B Figure S10A**, columns 4-5, **Appendix C Data S5**, linear regression slope= 0.517-1.76, FDR = 0.004-6.77 x 10^{-63}). Neuronal OCRs shared between species had larger effect sizes on top of being able to map to those species alone (**Appendix B Figure S10A**, columns 3,6, **Appendix C Data S5**, linear regression slope= 0.65-1.24, FDR= 0.0249-2.34 x 10^{-10}). These results suggest that complex brain traits are highly enriched in human neuronal OCRs that are conserved in both the DNA sequence and regulatory genomic function within orthologous cell types in rhesus macaque and mouse.

3.3.4 Polyfun Fine-mapping

Based on the genome-wide enrichment for trait heritability using cell type OCR profiles from humans and model species and companion work on mammalian and primate genome constraint scores [176] above many other functional annotations, we expect that these additional conservation annotations will improve base-resolution fine-mapping of genetic variants associated with common traits. We applied functionally-informed fine-mapping (polyfun) with these annotations and the polyfun baseline_LF annotations for 43 brain-related traits from GWAS with significant overlap with the UKBB or European ancestry populations for which there is pre-computed LD matrices across >20 million genetic variants (MAF > 0.005) [206,210]. Across these traits, 41 GWAS returned 7,423 SNPs with non-marginal posterior inclusion probability (PIP) of being the candidate causal SNPs (PIP > 0.1, **Appendix C Data S9**), with 748 SNPs highly confidently fine-mapped as candidate causal (PIP > 0.95).

We annotated 1,233 SNPs to regions constrained in mammals across the whole genome (phyloP ≥ 2.270, Posterior inclusion probability of being the causal SNP, PIP > 0.1) and 693 of these are constrained in mammals at distal, non-exonic regions of the human genome. In contrast, more fine-mapped SNPs, 946 unique SNPs, were distal, non-exonic and overlap a caudate cell type OCR with non-zero Cell-TACIT Age (**Appendix C Data S10**, PIP > 0.1). Both of these measures of mammalian conservation correlate with higher fine-mapping PIP scores. Although Cell-TACIT Age and phyloP are correlated with each other, both are still correlated with higher PIP scores when adjusting for each other ($\text{p-value}_{\text{phyloPcons}} = 9.99 \times 10^{-6}$, $\text{p-value}_{\text{Cell-TACIT-Age}} = 0.0166$, nested mixed effect linear regression ANOVA).

The number of non-marginal fine-mapped traits (PIP > 0.10) significantly correlates with the power of the GWAS, estimated by the z-score from h^2 estimation, (linear regression $P_{h2_zscore} = 0.00240$) but not the total heritability of the traits ($P_{h2g} = 0.0645$). This suggests that

larger GWAS in the less well-powered traits may identify more fine-mapped SNPs. Most fine-mapped SNPs were from well-powered neurological traits that tend to be easier to phenotype across large populations ($Neuro_{avg}$ = 547 SNPs per trait, BMI=2,823 SNPs, PIP > 0.10). Psychiatric traits that require more expertise to diagnose and report had mere fractions in number of fine-mapped SNPs as neurological traits ($Psych_{avg}$ = 168 SNPs per trait, 586 Schizophrenia SNPs, PIP > 0.10, **Appendix B Figure S27**). Substance abuse and substance abuse disorder traits had the fewest fine-mapped SNPs per trait (SU_{avg} = 40.3 SNPs per trait, SUD_{avg} = 5.6 SNPs per trait). Many of these GWAS have yet to reach the point of detecting all trait-associated loci [187], so functionally informed fine-mapping and interpretation of fine-mapped SNPs will need to be performed in the future when more of these GWAS become sufficiently powered. We provide the polyfun annotations from this study online (**Data Availability**) to facilitate re-fine-mapping of subsequent GWAS.

3.3.5 Additional fine-mapped loci

We used the Cell-TACIT Age to improve fine-mapping of SNPs from several well-powered GWAS. In the neuroticism locus around the *GRM3* gene (**Appendix B Figure S32**), many SNPs were fine-mapped to be candidate causal SNPs with varying degrees of confidence. The top two SNPs, rs274622 and rs802425, had nearly equal PIP scores, yet only the latter SNP overlapped a D2 MSN OCR with high Cell-TACIT D2 Age (chr7:86,665,609, D2 Age=38). The first SNP may still be a candidate causal variant, but it may be involved through a different cell type or perhaps different mechanism. Both SNPs were found by GTEx to be eQTLs for *GRM3* within the cerebrum GTEx tissue and *GRM3* expression was enriched in D2 MSNs of the macaque [174]. We pose that both variants may be functional.

We used Cell-TACIT to refine the fine-mapped loci from moderately powered GWAS. We previously reported that smoking-related traits had genome-wide enrichment for D1 and

D2 mouse neuron subtypes from the striatum [210] (**Appendix B Figure S33**). We improved our tissue and cell type predictions with Cell-TACIT MSN_D1 models and predicted that the SNP rs148428140, from the SmokingCessation and CigarettesPerDay GWAS, had the highest Cell-TACIT D1 Age within D1 OCRs in the region (chr9:133,609,189, D1 Age=42) (**Figure 3A**). This D1 OCR had high predicted activity in almost every species compared to other OCRs in the locus.

We investigated a locus associated with Alzheimer's Disease (AD) in AD and AD-by-proxy GWAS, which had strong, conserved neuroimmune signatures [225] (**Appendix B Figure S34**). We found two SNPs, rs2632516 and rs2526377, within microglia OCRs with high Cell-TACIT Microglia Ages, (chr17:58,331,292, Microglia Age: 44; chr17:58,332,310, Microglia Age: 15, respectively). These SNPs and other fine-mapped SNPs in the locus fell in regions constrained across mammals (phyloP ≥ 2.270) and overlapped with human microglia OCRs. High Cell-TACIT Microglia Age points to these SNPs as likely candidate causal variants contributing to AD. Both SNPs were eQTLs for the nearby gene *TSPOAP1* and a gene outside the plotted region, *TRIM3*, in blood and skin GTEx tissues [226]; rs2632516 is an eQTL in peripheral blood cells of systemic lupus erythematosus patients for the microRNA miR-142 [227], which regulates other AD-associated gene regulation when manipulated in human induced pluripotent neural progenitor cells and mouse hippocampus [223]. These examples and many other loci exist where Cell-TACIT cross-species predictions and Cell-TACIT Age could suggest which SNPs in cell type OCR may be active to prioritize downstream testing (**Appendix B Data S6**). Regulatory conservation inferences may guide how human cell type-specific genetic variant could be studied in orthologous OCRs of model and non-model species (e.g. rodents, non-human primates, prairie voles)

These examples and many other loci exist where Cell-TACIT cross-species predictions and Cell-TACIT Age could suggest which SNPs in cell type OCR may be active to prioritize

downstream testing (**Data S10**). Regulatory conservation inferences may guide how human cell type-specific genetic variants could be studied in orthologous OCRs of model and non-model species (e.g. rodents, non-human primates, prairie voles).

3.3.6 Limitations

Cell-TACIT, which was introduced here jointly with the companion paper[194], is the first method to associate differences in regulatory activity between species in specific cell types. Cell-TACIT and the experimental validations we demonstrate in this paper relate the regulatory conservation of cell types across placental mammals to the heritability of human traits in ways that are useful but have limitations.

A major limitation of Cell-TACIT is that species, sample quality, sample preparation, research facilities, and data annotation are often confounded with each other. To the best of our abilities, we replaced or augmented the data to represent the highest-quality cell type OCR profiles from human, rhesus macaque, and mouse. The lower performance in our cross-species analyses of glia and D1H MSN cell types in humans but not rhesus macaque suggests that our model training approach may need to be further optimized, that the regulatory code in some of the cell types for which we have models is only partially conserved between euarchonta and glires, or that there are confounding factors due to differences between the labs in which the data were generated that prevented some of our models from making consistently accurate clade-specific open and closed chromatin predictions. These sub-optimally accurate clade-specific open and closed chromatin predictions decrease the reliability of results. Due to the challenges in obtaining open chromatin from most cell types and inherent differences in species' lifestyles, standardizing conditions across species for data generation is often impossible.

Our expansion of Cell-TACIT has additional limitations beyond sub-optimal performance of some models for some evaluation metrics. We made predictions across the placental mammalian phylogeny, but our models were trained using only species from primates and mouse, so the regulatory code that they learned may not be conserved in all of the species where we made predictions. Determining if the regulatory code in specific cell types is conserved beyond euarchontoglires requires regulatory genomics data from specific cell types in at least one mammal outside of euarchontoglires, which may be infeasible to collect due to the difficulty of obtaining tissue samples from most non-euarchontoglires mammals. In addition, we trained our models using only sequences at least 20,000 bp from the nearest TSS so that they would learn the enhancer and not promoter regulatory code, but, for some of our analyses, we made predictions at OCRs at least 3,000 bp from the nearest TSS to increase the number of OCRs we considered; some of these predictions may be inaccurate because some of these OCRs may be promoters, and further research is needed to improve the ability of convolutional neural network models for regulatory DNA sequence prediction to distinguish between promoters and enhancers. Furthermore, many neurological conditions may be caused by regulatory elements that are accessible only when a neuron is activated, so such regulatory elements may not be present in our data [24].

Our models make binary predictions, but open chromatin in a cell type representing multiple cells is a continuous signal; an exciting extension to this work would be to predict open chromatin signal across species [228] and compute an improved Cell-TACIT Age metric that incorporates the level of predicted open chromatin. We anticipate that, as experimental technology for obtaining cell type-specific [229,230] and single-cell [231] regulatory genomics data as well as our ability to distinguish between promoters and enhancers and compare open chromatin signals between species improves, Cell-TACIT's reliability and applicability will increase.

We presented *in vivo* cell type-specific reporter assays that demonstrate the strength of Cell-TACIT predictions, yet these experiments are limited by the low throughput and method to ascertain cell type-specificity. We took a targeted approach to test the conserved regulatory activity by testing two candidate sequences using both human and mouse orthologs. The experimental burden to synthesize, clone, package into high titer AAV virus that crosses the blood brain barrier, and transduce entire brain regions precludes testing more human loci or open chromatin regions or more orthologs from distantly related mammalian species. We were size-limited by DNA synthesis technologies and only tested the central 227bp of the candidate enhancers rather than the full 501bp sequence. Our computational models predicted the shorter 227bp to be weaker and less cell type-specific than the full 501bp sequences (**Figure 4C**, mouse and human squares vs. **Figure 4D**), a result that experimentally aligns with previous reports on how the length of synthesized candidate enhancer affects the measured strength and variation of reporter gene transcription in reporter assays [232]. The peripheral DNA sequences–in combination with the central region–may contain cell type-specific regulatory motifs or stabilize enhancers during transcription. Our reporter assay experiments leverages AAV technology to efficiently transduce high copy-numbers of our candidate enhancers into the mouse brain, but this measures only the candidate enhancer in an episomal context rather than in the native genome. Orthogonal approaches using CRISPR-knockout or CRISPR-inactivation combined with cell type targeting can be useful to confirm our results [233,234].

Within the field of rodent neurobiology that studies the caudoputamen, transgenic reporter strains are the cornerstone to selectively measure or manipulate D1 and D2 MSNs. D1 and D2 receptors have strong homologies with other G-protein coupled receptors, including D3, D4, and D5 receptors, so few antibodies have been successfully validated to specifically label D1 or D2 MSNs [235,236]. The membrane-trafficking property of dopamine receptors proves an additional challenge to using D1 or D2 antibodies as cell type-specific counterstains in our

reporter assays. Since we predicted our candidate enhancer sequences to be equally active in both D1 and D2 MSNs but differentially active in neurons vs. glia, we circumvented these problems by using antibodies against the nuclear NeuN stain to counterstain caudoputamen neurons, which are overwhelmingly D1 or D2 MSNs [57]. Ascertaining finer cell type-specific impacts of other human candidate enhancers would have to rely on single- or double-labeling D1 or D2 reporter rodent strains [237], which excludes testing candidate enhancers in other species, or using single-cell transcriptomics methods, which add different sets of challenges [238].

Our *in vivo* experiments tested the gene regulation of mouse brains, irrespective of behavioral stratification, environmental exposures, or experimental conditions. Future work could expand these *in vivo* experiments to identify interactions between human risk variants, gene regulatory mechanisms, and the specific neural circuits underlying behavior in rodents, primates, or non-traditional model organisms of traits and behaviors. Some experiments can be paired with established paradigms such as in intravenous drug self-administration for substance use and withdrawal behaviors or multi-arm bandit gambling tasks for reward and risk-taking behaviors. The translational models on the behavior side complement the regulatory predictions provided by Cell-TACIT to dissect the gene-to-behavior relationships in specific neural circuits of the brain that ultimately are key to understanding the human genetics of complex traits and behavior.

A number of human traits that we analyze–especially schizophrenia–do not have known related behavioral paradigms or phenotypes measurable in other species. While we find increased heritability enrichment in measured or predicted active regulatory elements mappable from distant species, we do not presume that those species have these traits. Rather, our findings suggest that, across mammalian evolution, the conserved regulatory elements contribute to increased fitness, so genetic variation at these elements in the human genome may

predispose humans to diseases or disorders. Experimental manipulation of these conserved regulatory elements in behaving model organisms may be able to reveal the biological mechanisms and pathophysiology of human genetic variants, as animal knockout models have done for rare mono-genetic human disorders [239]. For example, investigation of the locus from **Figure 3D,** and **Figure 4A**, a region containing eQTLs of DRD2 and accessible in human and mouse D2 neurons, may identify pathological regulation of the D2 receptors, altering critical basal ganglia dopamine signaling.

3.4 Materials & Methods

3.4.1 Animal Care & Use

All non-human primate animal procedures were in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the University of Pittsburgh's Institutional Animal Care and Use Committee (IACUC) (Protocol ID, 19024431). Rhesus macaque monkeys were single- or pair-housed with a 12h-12h light-dark cycle. Monkey P was a 5-year-old female (5.4kg).

All mouse animal procedures were in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by Carnegie Mellon University's IACUC (Protocol ID, 201600003). The mice used for reporter assay experiments were C57BL/6J wild type mice or D1-*cre* and A2a-*cre* mice maintained on a C57BL/6J background from the GENSAT Consortium (Jackson Laboratory, Bar Harbor, ME). All mice were 2-5 months old at the time of the sacrifice. Mice experiments included representation from both sexes. All animals were housed with a 12 hour light cycle, and experiments were

performed at the same time of day relative to lights on. Animals for the data primary to this study received no treatments other than the retro-orbital AAV injections.

3.4.2 Non-Human Primate Single Nuclei ATAC-seq

To maximize nuclei viability for single nuclei ATAC-seq (below), we followed a harvesting protocol described in [240,241]. Briefly, we sedated Monkey P with ketamine and maintained general anesthesia with isoflurane. We transported the monkey to a surgery suite and placed it in a stereotaxic frame (Kopf Instruments). We removed the calvarium and then perfused the monkey with 3-4 liters of ice cold artificial cerebrospinal fluid (ACSF; 124 mM NaCl, 5 mM KCl, 2 mM MgSO4, 2 mM CaCl2, 23 mM NaHCO3, 3 mM NaH2PO4, 10 mM glucose; pH 7.4, osmolarity 290–300 mOsm) oxygenated with 95 % O2:5 % CO2. We then opened the dura and removed the brain. We sliced the brain into slabs on a custom 3-D printed matrix along the rostral-caudal axis. We dissected the caudate nucleus under a dissection microscope.

We isolated nuclei according to the 10X Genomics "Demonstrated Protocol: Nuclei Isolation for Single Cell ATAC Sequencing". Briefly, we cut brain tissue into small pieces and triturated the tissue in NbActiv1 buffer (BrainBits) using wide- and regular-bore pipette tips sequentially to dissociate cells. We filtered the cells through a 30 um MACS SmartStrainer (130-098-458) into an Eppendorf tube. After centrifuging, we used a lysis buffer (1mM Tris-HCl, 1mM NaCl, 0.3 mM $MgCl_2$, 0.01% Tween-20, 0.01% Nonidet P40 Substitute, 0.001%. Digitonin, 0.1% BSA) to lyse the cells for 5 min. We washed the lysed cells once with 1 ml chilled wash buffer (1 mM Tris-HCl, 1mM NaCl, 0.3 mM $MgCl_2$, 0.1% BSA, 0.01% Tween-20) and resuspended the nuclei pellet with Diluted Nuclei Buffer (10X Genomics, PN-2000153). We filtered the nuclei through a strainer and counted the nuclei.

We used 10x Chromium Single Cell ATAC Library & Gel Bead Kit (10X Genomics, Cat# PN-1000110) for monkey P. We transposed the nuclei and generated libraries according to 10x Genomics protocol and briefly incubated the nuclei suspensions in a transposition mix that includes transposase, which preferentially fragments the DNA in open chromatin. We generated Gel beads-in-emulsion (GEMs) after running through a 10x Genomics Chromium controller. We barcoded open chromatin within GEMs in a Bio-Rad PCR machine (Cat# C1000). We purified DNA with Dynabeads (10X Genomics, PN-2000048) and SPRIselect reagent (Beckman Coulter, Cat# B23318) after breaking the emulsion with a recovery agent (10X Genomics, 220016). Then, we amplified and indexed the samples by sample index PCR and purified them with SPRIselect reagent. We analyzed quality of the libraries using Agilent Bioanalyzer 2100 and quantified the libraries by qPCR using a KAPA Library Quantification Kit (KAPA Biosystems, Cat# KK4824). We pooled together libraries and loaded them onto NovaSeq 6000 S4 Flow Cell. We sequenced N=2 technical replicates of the caudate nucleus from monkey P to the depth of 150,000 x 150 bp paired-end reads per nuclei (Figure S1A, B).

3.4.3 Rhesus Macaque snATAC-seq read processing

We de-multiplexed sequencing reads with `dex-fastq` from the SnapTools package (https://github.com/r3fang/SnapTools) [242], trimmed them for the transposase adapter sequence "`CTGTCTCTTATACACATCT`" on the 3' end of reads with `cutadapt` [222], and aligned them to the `rheMac10` genome with `bowtie2` with the "`--sensitive --no-discordant --no-mixed --minins 25 --maxins 1000`" parameters to map fragments between 25bp and 1000bp long [206,243]. After aligning the reads, we removed PCR duplicates with `samtools` [244] and created arrow files for downstream analyses with the `ArchR` package [222,245]. We followed the approach by Jing, Kleyman *et al.* to map the higher quality `GRCh38.p13` human RefSeq gene annotations onto the rheMac10 genome with the

liftOff tool [206] and used this to create a custom rheMac10 ArchR gene and genome annotation as outlined within the ArchR manual. Using this liftOff annotation, we were able to compute ArchR gene activity scores around the orthologous rhesus macaque regions of human genes, enabling us to obtain scores for substantially more genes and complete gene bodies than we would have been able to obtain using base rheMac10 gene annotations. Within ArchR, we removed inferred doublet nuclei from each sample with cutEnrich = .5 and filterRatio = 1. Then we performed iterative latent-semantic index (LSI) clustering with 4 iterations and 30,000-150,000 variable features followed by Harmony batch correction [246], UMAP visualization, and Louvain clustering, where we used default settings for all methods. We identified snATAC-seq clusters by co-clustering with Seurat v3's canonical correlation analysis (CCA) on ArchR gene activity scores [222] with previously published single nucleus RNA-seq gene expression of the rhesus macaque striatum [247]. We called peaks across technical replicates with macs2 [248], identified reproducible peaks between replicates for each cell type, created a consensus peakset across cell types, and created a peak by cell "PeakMatrix" through ArchR [249].

3.4.4 Reporter Assay Molecular Cloning

We created the template *in vivo* reporter plasmid construct, pAAV-Hsp68-pNLS/mCh-3rNLS/ReporterAssayEmpty (pAAV-ReporterAssayEmpty), with VectorBuilder Cloning Services (VectorBuilder Inc, Chicago, IL). The template plasmid contains inverted terminal repeats (ITRs) standard for AAV production, Hsp68 minimal promoter, synthetic intron [187,250], mCherry with two SV40 Large T antigen nuclear localization signals (PKKKRKVED) on the N- and C- terminals [101], and the Chloramphenicol resistance gene as a place-holder for the candidate enhancer between two pairs of multiple cloning sites. After receiving the plasmid,

we also added three RNA nuclear localization motifs (AGCCC) outside the coding sequence and before the SV40 poly-adenylation signal [232] (**Appendix B Figure S33**).

To clone in the candidate enhancers, we synthesized the human enhancer sequence by IDT as gBlocks™ with common 5' (GAGTCTGAACCTGTGTGCTA) and 3' (TCAGAAGATCGGAAGAGCATCGTAG) sequences for downstream PCR amplification (Integrated DNA Technologies). We used PCR to amplify synthesized inserts using the KAPA HiFi HotStart ReadyMix protocol (Roche, KK2601) with primers to add BsmBI restriction enzyme sites compatible with Golden Gate cloning (F: GGCTACCGTCTCACATGGAGTCTGAACCTGTGTGCTA, R: GGCTACCGTCTCGACCACTACGATGCTCTTCCGATCT). We linearized the pAAV-ReporterAssayEmpty plasmid using the Q5 HotStart HiFi ReadyMix protocol (NEB, M0494) with primers to both add BsmBI restriction enzyme sites and excise the chloramphenicol place-holder (F: GGCTACCGTCTCGTGGTACCTAGCATCTCAC, R: GGCTACCGTCTCCCATGTTTCTCCATTTGGCGC). We cloned the enhancers into the pAAV-RAe plasmids with the NEB Golden Gate Assembly Kit (NEB, E1602) following manufacturer instructions. We validated cloning accuracy with Sanger sequencing.

3.4.5 AAV production and delivery

We packaged reporter assay constructs into individual AAV preps packaged with the AAV-PhP.eB capsid that crosses the blood-brain barrier and has high transduction of neural cell types as previously described [88]. Briefly, AAV was produced in AAVpro(R) 293T cells (Takara, Kyoto, Japan; #632273) by co-transfection of the genome pAAV Reporter Assay plasmid, an AAV helper plasmid, and pUCmini-iCAP-PHP.eB. pUCmini-iCAP-PHP.eB was a gift from Viviana Gradinaru (http://n2t.net/addgene:103005; RRID: Addgene 103005) [88]. We precipitated the AAV particles with Polyethylene Glycol (PEG 8000, Sigma-Aldrich, St. Louis,

MO; cat. P2139-500G) from the cell media, lysed the cells with Salt Active Nuclease enzyme (SAN, ArcticZymes, Sykehusveien, Norway, cat. 70900-201), and purified AAV on an iodixanol gradient (OptiPrep, Sigma-Aldrich, cat. D1556-250ML) with ultracentrifugation for 2.5 hours at 350,000 x g at 18°C. We filtered and concentrated the virus in PBS+F-68 (PBS, Pluronic(™) F-68 Non-ionic Surfactant (100x) 0.001%) using Amicon Ultra-15 centrifugation filters (Millipore, Burlington, MA; #UFC905024) and resuspended in PBS. We measured the viral titer for each with the AAVpro(R) Titration Kit (Takara, #6233), diluted to the same concentration of 2.32 x 10^9 vector genomes (vg) / µL, and stored in single-use aliquots at −80 °C until injection.

We used three cages of 4 adult age- and sex-matched littermates for each retro-orbital virus injection. We anesthetized each animal with 2-3% isoflurane until we did not observe any pedal withdrawal reflex and maintained anesthesia with 1-2% isoflurane during the procedure. Then, we injected each reporter assay AAV (1.16 x 10^{11} vg in 50 µL) with a viral transduction control AAV expressing nuclear-localizing GFP (4 x 10^{11} vg in 50 µL) of into the retro-orbital cavity and treated the eye with 0.5% Proparacaine Hydrochloride Ophthalmic Solution. We monitored the animals while the virus incubated for 2 weeks before brains were harvested.

3.4.6 Immunofluorescence and Microscopy

We fixed animals with 4% paraformaldehyde (PFA) transcardial perfusion and incubated the brains in 4% PFA overnight after dissection. We cut 50-60 µm thick coronal slices with a vibrating microtome (Leica VT1000 S). We performed immuno-fluorescence staining of transduced mouse brains to assay for reporter gene presence in caudoputamen neurons, which overwhelmingly are canonical D1 or D2 MSNs in equal proportions [186] and are where the rodent caudoputamen is the orthologous region to primate caudate nucleus. We co-stained free-floating sections with anti-mCherry primary antibody (Cell Signaling; E5D8F) to

amplify signal with AlexaFluor 594 secondary antibody (Cell Signaling Technology, Danvers, MA; #8889) as well as anti-NeuN primary antibody (ABcam; ab279297) to stain for neurons with AlexaFluor 405 (Thermo Fisher; A48268). We took images of the caudoputamen blinded to experimental conditions with a brightfield epi-fluorescence microscope under 20x objective (Nikon, Eclipse Ti2) with two to three 3x3 tiled images per animal capturing different coronal views of the dorsal striatum in both the left or right hemispheres. Each image had 0.365 micron per pixel in the x- and y- directions, and we reduced background noise using the NIS-Elements denoising software.

3.4.7 ATAC Data Processing

3.4.7.1 Human caudate snATAC-seq preprocessing

We downloaded human snATAC-seq reads aligned to the `hg38` genome from GSE147672 [187]. We created a human caudate snATAC-seq ArchR project (N=3 biological replicates) and assigned glial and neuronal subtype labels from the publication's table `SupplementaryDataSet2_scATAC-QC-And-Cluster-Residence_v1.xlsx`. We subclustered the main MSN cluster to resolve D1 and D2 MSN subtypes by increasing the number of variable features during iterative LSI and otherwise used the Corces *et al.* [187] caudate cell type labels: D1 MSN, D2 MSN, striatonigral (SN) MSN, *PVALB*+ interneurons, astrocytes, oligodendrocytes, oligodendrocyte precursors (OPC), and microglia.

3.4.7.2 Mouse caudoputamen cell types reprocessing

We downloaded the de-multiplexed sequencing reads for the caudoputamen (regions 4D and 5E) samples of mouse snATAC-seq from the `NeMO` archive (N=4 biological replicates, https://nemoarchive.org) [185]. We aligned de-multiplexed reads with the same method that we used for rhesus macaque, aligned reads with `bowtie2` to the `mm10` mouse genome, and created arrow files and ArchR projects for the mouse. We assigned the neuronal and glial

subtypes from the cell barcodes available through the CATlas CellBrowser webportal (http://catlas.org). For D1 MSNs, D2 MSNs, and interneurons, most of which had high *PVALB* gene activity scores in primate snATAC-seq, we instead used the corresponding, previously published high-quality cre-dependent Sun1GFP nuclear-anchored independent labeled (cSNAIL) ATAC-seq datasets from mouse caudoputamen, which we downloaded from Srinivasan, Phan *et al.* GSE161374 [251].

3.4.7.3 Bulk M1 cortex and liver ATAC-seq data processing and peak ortholog mapping

We compared the above cell type ATAC-seq profiles with cross-species matched bulk ATAC-seq profiles. For these data, we used the processed house mouse motor cortex ATAC-seq data from [185], the processed rhesus macaque and Norway rat motor cortex ATAC-seq data from [185] that were originally presented in [251], and the processed Egyptian fruit bat motor cortex ATAC-seq data from GSE187366 (access token: `exsheqcazjsnxsx`) [194,252]. For rhesus macaque and Egyptian fruit bat, we used the subset of orofacial motor cortex peaks that overlapped hand/wing motor cortex peaks, as described in the companion[194]. We used the processed liver ATAC-seq data from [45], where the house mouse data was originally presented in [92], the rhesus macaque and Norway rat data that were originally presented in [253], and the cow and pig data that were originally presented in [91]. For these datasets, we used the open chromatin regions that were filtered to be likely enhancers in [92] and [194]. We used the orthologs in the Zoonomia genomes obtained using the Zoonomia v2 CACTUS [74,93] that were obtained in [194].

3.4.8 Cross-species peak ortholog mapping

For all snATAC-seq and cSNAIL-ATAC-seq cell type peaks from the primate caudate nucleus and mouse caudoputamen, we kept peaks in the narrowPeak format to retain the summit information for each peak, as critical transcription factor motifs have been shown to be enriched near open chromatin peak summits [118]. For ArchR snATAC-peaks, we achieved this

by taking the summits of reproducible peaks in each cell type and extending them 250bp in both directions. We named each peak uniquely with "refGenome:chr:start:stop:summit" to track it and its unique, usable orthologs when mapping across species. We mapped peaks across Zoonomia genomes using the Zoonomia version 2 CACTUS alignment [254] (https://cglgenomics.ucsc.edu/data/cactus/) with `halLiftover` [74,191] and identified unique, usable orthologs with HALPER's `orthologFind.py` script with parameters `-max_frac 1.5 -min_len 50 -protect_dist 10 -narrowPeak` (https://github.com/pfenninglab/halLiftover-postprocessing) [208,209,254–256]. We mapped human snATAC-seq peaks for 8 cell types across the 240 genomes in the Zoonomia alignment. We mapped the house mouse and rhesus macaque snATAC-seq and cSNAIL ATAC-seq peaks to the human genome. Prior to applying halLiftover and HALPER for the rhesus macaque peaks, we used the `liftOver` tool [111,201,257–263] to map `rheMac10` coordinates to coordinates in `rheMac8`, which is the genome in the 241 Zoonomia alignment. We mapped both the peaks (chr:start-stop) and the summits (chr:summit), keeping only peaks where a summit was mappable within its corresponding peak.

To evaluate whether the regulatory codes in the identified caudate cell types were sufficiently conserved across between primates and mouse to train Cell-TACIT models that could achieve high lineage-specific OCR accuracy, we compared the enrichments of TF motifs occurring in mouse OCRs to those enriched for occurring in human OCRs, as in [194]. We then ran MEME-ChIP (*131*) on the OCRs identified in each cell type from each respective species, and we used the respective `-db` to the CIS-BP version 2.00 databases (*86*) for *Homo sapiens*, *Macaca mulatta*, and *Mus musculus*. We used default settings along with the parameters `-spamo-skip`, and `-fimo-skip`. After that, we compared the motif enrichments of each cell type across species.

3.4.9 LD score regression in hg38 coordinates

We performed linkage disequilibrium score regression (LDSC) analyses within the `hg38` coordinates by applying liftOver across the published bed files baseline annotation (v1.1) and baselineLD annotations (v2.2) for cell type-specific LDSC and partitioned heritability enrichment LDSC analyses, respectively, as recommended by original authors (https://alkesgroup.broadinstitute.org/LDSCORE/, https://github.com/bulik/ldsc/wiki/LD-Score-Estimation-Tutorial) [13,14,264–266]. We used the 1000G SNPs mapped to `hg38` coordinates across super-populations (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions) [208,209,255–257] and computed LD scores as previously outlined using the `--ld-wind-kb 1000` parameter. Using these LD scores, we computed genetic correlations between traits [111,201,258–264], LD scores for additional annotations in `hg38` coordinates, and partitioned heritability analyses as previously described [14,65,67,108,265].

For LDSC cell type-specific analyses (https://github.com/bulik/ldsc/wiki/Cell-type-specific-analyses), we sought to estimate the independent contribution of an annotation after regressing out a background annotation and the baseline (v1.1) annotations. This coefficient is further standardized to the number of SNPs in the annotation and the per-SNP estimated narrow-sense heritability, the τ^* as described by [13,14,265–267]. This allowed us to compare both the significant contribution of a cell type annotation to a trait as well as the relative effect sizes between traits with varying power and heritability. We compared τ^* in human cell type OCRs to cell type OCRs from other species across significant LDSC traits with a standard linear regression model for each cell type, `coeff_diff ~ model_species/coeff_mean + peaktype`, and the numerical results are reported in **Appendix C Data S5**. The regression variables are defined as follows:

`coeff_diff`: τ^* from using other species information minus human OCR τ^*

`coeff_mean`: mean τ^* from using other species information and human OCR τ^*

`model_species`: categorical levels `rheMac10` or `mm10`

`peaktype`: categorical levels `Model_Org -> Hg`, `Hg -> Model_Org`, or `OCR_Ortholog`.

For LDSC partitioned heritability LDSC analyses (https://github.com/bulik/ldsc/wiki/Partitioned-Heritability), we sought to estimate the genome-wide enrichment of heritability of SNPs within an annotation compared to outside the annotation normalized to how many SNPs fall within the annotation and the total heritability of the trait. We computed heritability enrichment of human cell type OCRs mappable or mappable and predicted active in increasingly distant mammalian orders.

3.4.9.1 GWAS for neurological and psychiatric disorders

We collected summary statistics of 64 genome-wide association studies (GWAS) spanning neurological and psychiatric traits related to the brain and cognitive functions along with control traits known to be involved with other tissues and organs (**Appendix C Data S4**). These traits have been grouped grossly into six clinical domains of effect: neurodegenerative (Degen) [66,107,113,218,268–272], psychiatric disorders from diagnosis or with related measures (Psych) [114,203,273], substance abuse behavior (SU) [274–276], substance abuse disorder by diagnosis or with related measures of disordered use (SUD) [115–117,269,277,278], neurological trait not otherwise specified (Neuro) [278–280], sleep-related traits (Sleep) [185], metabolism traits from blood biochemistry (Metabolism) [119], or other traits (Other) [185]. Three GWAS (Cannabis use disorder, opioid dependence vs. opioid exposure, and post-traumatic stress disorder) analyzed in this study had been conducted in African-ancestry or admixed populations.

Complex traits have varying degrees of SNP heritability (h^2), and genome-wide association studies to current date have varying degrees of power to detect trait-associated loci. Furthermore, the complexity and interrelatedness of complex traits prove a challenge for attributing overlap or non-overlap of cell type-specific measurements of related GWAS. We provide two auxiliary analyses to ascertain the relationship between the selected traits: estimates of the GWAS power based on confidence of the h^2 estimation by LDSC and by genetic correlation, r_g, (https://github.com/bulik/ldsc/wiki/Heritability-and-Genetic-Correlation).

3.4.10 Cell-TACIT Model training and evaluation

We trained Cell Type-Aware Conservation Inference Toolkit (Cell-TACIT) convolutional neural network models (CNN) in the caudate nucleus using cell type OCRs from snATAC-seq or cell type-sorted ATAC-seq from multiple species described below. We applied theCNN architecture as previously described in Srinivasan, Phan *et al.* (5 convolutional layers, 1 dense layer, and a sigmoid activation) and trained CNNs with a one-cycle policy (OCP) cyclic learning rate scheme [279–281]. Using that architecture, we specifically did a grid-search to find that the following parameters produced robust models according to cross-validation performance: `batch size=1000`, `L2 regularization=1e-10`, and `dropout=0.25`. Similarly, we found that the following OCP learning parameters produced the most robust models: `epoch=23`, `cycles=2.35`, learning rates varying between `1e-2` to `1e-1`, and momentum varying between `0.85` to `0.99`. We trained 5-fold cross-validated models using the same chromosome hold-out scheme as in our previous work [92], where OCRs on the test set (`chr1` and `chr2`) are never seen by any models or evaluations for hyper-parameter selection, and models for each fold do not see OCRs on its own validation chromosome hold-out sets: `fold1: chr[6, 13, 21]`, `fold2: chr[7, 14, 18]`,

fold3: chr[11, 17, X], fold4: chr[9, 12], and fold5: chr[8, 10]. In the cross-species setting, the peak assignment to test, validation, and training sets are determined by the hg38 chromosome if the peak is mappable to human, or the hg38 chromosome of the mappable peaks in synteny with the unmappable peak. To identify peaks in synteny with an unmappable peak, we sorted the peaks by the source mouse or macaque chromosome and coordinates, converted the results to a data table, and filled in the missing hg38 chromosome to peaks not mappable to hg38 by the previous or next non-missing hg38 entry by running dplyr::fill(direction = "updown"). We trained models using 1 species, human (HgOnly), to evaluate whether a model trained on one species can make accurate predictions in species not used in training [281,282] or using all 3 species, human-rhesus-mouse (HgRmMm) to evaluate the gain in model performance using cell type-matched, cross-species open chromatin data. The 3-species HgRmMm training paradigm established the final caudate Cell-TACIT models because these models had as good or better performance than the models trained using only 1 species, a result further supported by multiple previous studies [187].

We trained models using the following negative sets to identify the negative set with the best balance in overall performance and specific performance at detecting cell type OCR specificity and changes in OCR conservation. All models used positive sets from the reproducible OCR set detected in their cell types across species, where we removed OCRs overlapping exons and within 20,000bp from a TSS to thoroughly exclude promoter elements, as described in the companion paper[194] because promoters and enhancers within the same cell type have been shown to be bound by only partially overlapping groups of transcription factors [284,285]. The following negative sets were used to train 5-fold cross-validated models, and the nonCellEnhLargeGC negative set is the final negative set for caudate Cell-TACIT models.

1. nonEnhNeg: We obtained orthologs of the positive set OCRs mapped to other species that did not overlap any reproducible or non-reproducible OCR of the cell

type used in model training [282,283]. In the human-rhesus-mouse 3-species setting, the `nonEnhNeg` set is 40-60% the size of the positive set. This negative set aims to push models to learn cases when cell type OCR activity is lost although genome sequences can still be aligned.

2. `largeGC`: We obtained the large set of GC-matched negatives to the positive set within each species using the biasaway tool with the following parameters: `biasaway c --nfold 10 --deviation 2.6 --step 50 --seed 1 --winlen 100` [211]. While 10 times GC-matched negatives are generated, we further filter them to exclude exonic and promoter regions and regions overlapping any reproducible or non-reproducible OCR of the target cell type, resulting in 4-5 times as many GC-matched negatives as positives. This negative set aims to account for non-specific biases due to GC-content of or other functional activity of the genome unrelated to cell type OCR activity.

3. `nonEnhLargeGC`: We obtained the union of negatives from `nonEnhNeg` and `largeGC`. This negative set was 4.5-5.5x larger than the positive set.

4. `nonCellEnhLargeGC`: We obtained the union of negatives from `nonEnhLargeGC` and OCRs within each species from other cell types (`nonCellNeg`). These additional negatives come from the positive sets of the other 7 cell types that do not overlap any reproducible or non-reproducible OCR of the target cell type, which is about 4-5 times the positive set. This negative set combines the first two and also aims to train models to learn cell type-specific OCRs rather than tissue- or cell type-invariant OCRs. When combined, the `nonCellEnhLargeGC` negative set is 9-11x larger than the positive set.

3.4.10.1 Cell-TACIT Model evaluation

We evaluated Cell-TACIT models across folds for overall performance on the validation set using the area under the receiver-operator curve (`auROC`), area under the Precision-Recall curve (`auPRC`), area under the Precision-Recall curve subtracting by the ratio of positives to negatives (P:N) to account for various P:N ratios across models (`auPRC.adj`), f1 score (`f1_score`), false positive rate (`fpr`), true positive rate (`tpr`), and negative predictive value (`npv`). We also evaluated these models for cell type-specific prediction of OCRs using tpr and true negative rate (`tnr`), where we evaluated results for each model-cell type combination separately. In this case, `tpr` and `tnr` are calculated using the `nonCellNeg` validation set. Similarly, we evaluated our models' abilities to predict whether open chromatin

is conserved, in which we calculated `tnrs` using the `nonEnhNeg` validation set [284]. For metrics that required categories (e.g `fpr`), rather than scores (e.g. `auROC`), we used the threshold > 0.5 to classify a DNA sequence as open and <=0.5 as closed. Lastly, we accessed the average Cell-TACIT predictions of cell type OCRs on average when mapped to increasing distant species, the phylogeny-matched correlations as described in [285,286]. The negative sets vary in size, which will affect the evaluation metrics affected by class imbalance. For the reported evaluation metrics, the following values would be the baseline (random), `auROC` = 0.5, `auPRC.adj` = 0, `f1_score` = 0.5, `tpr` = 0.5, and `tnr` = 0.5.

3.4.10.2 Cross-species peak cell type annotations and predictions

We predicted open chromatin at orthologs of human caudate cell type OCRs that were mappable [214] to each Zoonomia genome. This creates a peak-by-species matrix for each cell type with Cell-TACIT predictions, where there is a row for each human OCR for a cell type and a column for each Zoonomia genome. We entered NAs for entries in species that did not have a unique, usable DNA ortholog and recorded the Cell-TACIT prediction if there was a unique, usable DNA ortholog. For each species, we obtained the median TimeTree estimate of million years ago (MYA) diverged from humans (e.g. rhesus macaque is 28.8 MYA, house mouse is 89 MYA) [214] (**Appendix C Data S6**, tab: `Zoonomia_Genome_Information`, column: "`Time Since Split from Human (TimeTree median)`"). We averaged across predicted OCR ortholog open chromatin in species at the same MYA divergence from humans and within the same taxonomic order for Boreroeutherian species or taxonomic clade for Afrotherians and Xenarthrans (e.g. all predictions mappable to Chiroptera). This created 20 groupings from human to the most distant clades, which represent the predicted evolutionary history of each OCR as far as it has a mappable ortholog (**Appendix C Data S6**, tab: `Zoo_meta_clades`, column: "`group_meta`"). For each cell type, we classified whether each human OCR is predicted to be accessible in each of the 20 groupings by taking the average

for each group and defining predicted accessible as having a prediction of at greater than 0.5. We used these binary annotations for Cell-TACIT LDSC analyses (**Figure 1D**, **Figure 2, Appendix B Figure S11-26**).

As introduced in Srinivasan, Phan *et al.*[287], we calibrated the raw Cell-TACIT model score by the percentage of the validation set that had that score or lower to scale the strength of CNN nonlinear sigmoid activation across model folds and cell types. This was accomplished by taking the Cell-TACIT score of the validation positive set to create an empirical cumulative distribution (`ecdf` R function) to scale all new predictions by the fraction of validation positives with a lower Cell-TACIT model score. This produced calibrated scores near 0 for the validation negative set and expandrf the dynamic range for prediction values in the positive set. We applied this calibration to the continuous score for the average across the species groupings for Cell-TACIT predictions (**Figure 1E**, **Figure 3D, Figure S30-32**) and individual species predictions (**Figure 1F**, **Figure 4C,D**). For heatmaps of Cell-TACIT predictions, we used the following Newick tree created by collapsing the Zoonomia species tree from Foley *et al.* [288] (`Zoonomia_ChrX_lessGC40_241species_30Consensus.tree`) to select species per clade groups, and we used the binomial names to obtain the each species's silhoeutte from http://phylopic.org to plot alongside mammalian phylogeny trees created with `ggtree` and `ggtreeExtra` R packages [253]:

```
((Loxodonta_africana:0.1204058581,Tolypeutes_matacus:0.1229192
742):0.01145639324,(((Oryctolagus_cuniculus:0.1608919201,Rattu
s_norvegicus:0.2965409366):0.006031885449,(Tupaia_chinensis:0.
1385286196,(Galeopterus_variegatus:0.08750990202,(Eulemur_flav
ifrons:0.07870638791,(Callithrix_jacchus:0.05610238745,(Macaca
_mulatta:0.02760672357,(Nomascus_leucogenys:0.01503824207,(Pon
go_abelii:0.01184875793,(Gorilla_gorilla:0.006121144092,(Pan_t
```

roglodytes:0.004607178674,Homo_sapiens:0.004410717579):0.00190
302907):0.005963570605):0.002256803519):0.00775016763):0.01406
957514):0.05984338905):0.01269656686):0.002845408602):0.002347
302752):0.01705378614,(Uropsilus_gracilis:0.207400545,(Rousett
us_aegyptiacus:0.1237216561,(Sus_scrofa:0.1368474811,(Equus_ca
ballus:0.07947626741,(Canis_lupus_familiaris:0.1226195774,Mani
s_pentadactyla:0.1145549768):0.006550057057):0.002145230337):0
.002284646667):0.005799996479):0.02130276027):0.01145639324);.

3.4.10.3 Cell-TACIT Age to prioritized fine-mapped GWAS loci

We created a metric called Cell-TACIT Age that quantifies the evolutionary persistence of open chromatin conservation for a human cell type OCR. Specifically, Cell-TACIT Age for a cell type OCR is the sum of the MYA of each 20 species grouping weighted by the average calibrated predicted Cell-TACIT score of each group (**Figure 1E**). OCRs that do not have any mappable ortholog to a group do not have Cell-TACIT scores. We chose to call this calculation composed of Cell-TACIT scores across clades the Cell-TACIT Age because it is derived from species MY divergence from humans and to differentiate it from scores of a species or scores averaged over species of a clade. We assigned fine-mapped SNPs a Cell-TACIT Age for each overlapping cell type OCR that is distal with a less strict cutoff >3kb from a TSS and non-exonic based on gene annotations of the human genome. Cell-TACIT Age calculations are most appropriate for distal open chromatin elements. To convert Cell-TACIT Age into quantiles for eQTL and LDSC analyses, we ordered non-exonic, distal cell type open chromatin regions by the Age value and divided the regions into equal quartiles. We had also tried quintiles, but this created too few genome-wide regions per quantile bin to estimate heritability by LDSC or other enrichment analyses.

3.4.11 eQTL Dataset Processing & Analyses

To identify enrichment of genetic variants associated with gene regulation in Cell-TACIT annotations, we used the full summary statistics of two eQTL datasets. We downloaded the first dataset, eQTLs from bulk caudate RNA-seq from the GTEx consortium v8 [214], from the GTEx google bucket at the link `gs://gtex-resources/GTEx_Analysis_v8_QTLs/GTEx_Analysis_v8_eQTL_all_associations/Brain_Caudate_basal_ganglia.v8.allpairs.parquet`. We downloaded the second dataset, eQTLs from single-cell single nucleus-RNA-seq of the prefrontal cortex [288], from Zenodo link https://zenodo.org/record/6104982#.YtLARy-B1hE. We calculated the FDR by applying the Benjamini-Hochberg method [288] to the summary statistics nominal p-values of each dataset to identify variants that were significantly correlated with allele-specific gene expression either in bulk caudate nucleus or prefrontal cortical cell types after multiple hypothesis correction. Some variants were tested multiple times with different neighboring genes, so we selected the most significant eQTL p-value for each eQTL. We annotated the variant `hg38` location to be in one of the Cell-TACIT Age quartiles or outside these annotations. We computed enrichment odds ratio that a variant is a significant eQTL (FDR < 0.05) and overlaps a Cell-TACIT Age quartile (vs. not being in a Cell-TACIT Age annotation) using a generalized linear model with a binomial link `` `signif_FDR ~ quantile` `` with the `speedglm` R-package for >10 million variants [254]. The enrichment analyses are well-powered by the >10 million of eQTL variants tested, so we used the Bonferroni correction at alpha = 0.05 to determine if eQTLs tend to be in Cell-TACIT Age quartiles. We did the Bonferroni correction over 4 Cell-TACIT Age quartiles by 8 Cell-TACIT cell type annotations by 8 single-cell eQTL sets or 1 bulk caudate eQTL set (**Appendix C Data S8**). For plotting,

we transformed the estimated log odds ± standard error to the regular odds ratio (**Figure 3A-B, Appendix B Figure S28A**).

3.4.12 Functionally informed fine-mapping

We applied functionally informed fine-mapping (`polyfun`), an approach for prioritizing candidate-causal variants from GWAS by accounting for LD and partition heritability (h^2) at the SNP level across any set of genomic annotations [289], to the 43 traits from the above list of 64 traits with significant population overlap with the UK Biobank (**Appendix C Data S4**, column "runPolyfun"). To do this, we pre-computed polyfun LD matrices on >18 million imputed SNPs [289]**.** We augmented the polyfun baselineLF v2.2 annotations with improved annotations of genome constraint using accelerated and conserved mammalian phyloP (top 0-4%, 8 annotations), primate PhastCons (top 0-5%, 5 annotations), conserved and accelerated phyloP at the FDR 5% threshold (PhyloP ≥ 2.270 and ≤ -2.270, 2 annotations), phastCons covering the top 3.5% of SNPs (PhastCons ≥ 0.961, 1 annotation) [289], and 6 ENCODE3 cCRE annotations from promoters, distal, and proximal enhancers [291], and we excluded 12 older primate and mammalian PhastCons annotations. Lastly, we included annotations for caudate cell type OCRs measured in humans, rhesus macaque (used human orthologs), and house mouse (used human orthologs) for a total of 24 regulatory genomic and functional conservation caudate cell type annotations. We created these annotations by lifting over bed files of each annotation from `hg38` to `hg19` [290] and computing LD scores using the precomputed UKBB LD matrices [291]. We excluded the few SNPs not found in the UKBB LD matrices from analyses (`--allow-missing`). We performed all polyfun analyses using the SuSIE as the underlying fine-mapper for a maximum of 10 candidate causal SNPs within a 1Mb locus (`--method susie --max-num-causal 10`) [294]. We fine-mapped any locus containing a least a SNP with GWAS p-value below a cutoff selected based on (`--pvalue-`

`cutoff`) the power of each GWAS; the cutoffs were $p < 5x10^{-7}$ for traits with few genome-wide significant loci (Alzheimer's Disease, OCD, PTSD, OpioidExposure, AlcoholDependence, OpioidDependence, and Dozing) and $p < 5x10^{-8}$ otherwise [292]. All polyfun runs, except Obsessive-Compulsive Disorder (OCD) and Opioid Exposure (OE_vs_UX), returned fine-mapped SNPs with non-negligible posterior probabilities of being predicted causal SNPs (PIP > 0.10), where the number of fine-mapped causal SNPs correlated with the power of each GWAS. We reported fine-mapped SNPs in **Appendix C Data S9** alongside the SNP's phyloP score. We annotated fine-mapped SNPs that overlap a non-coding and non-promoter cell type OCR with the Cell-TACIT Age of that OCR along with the fraction of bases in the OCR that has a phyloP score ≥ 2.270 ("`consFrac`") (**Data S10, Appendix B Figure S27**).

We compared the fine-mapped variants associated with human traits to the Cell-TACIT annotations. We combined our tested fine-mapped neuropsychiatric traits with similarly fine-mapped traits conducted in a companion study[184]. We excluded the fine-mapped SNPs from `'BMI,'` `'Intelligence score,'` `'Neuroticism,'` and `'Morning Person,'` traits already in our set of 43-fine mapped traits, to create a "non-brain" background set of SNPs. Because some traits were from under-powered GWAS to find few fine-mapped SNPs (e.g. Anorexia), while other traits have well-powered studies to find more fine-mapped SNPs (e.g. BMI), we grouped SNPs together by their trait groups to power enrichment analyses across 6 trait domains ("Degen," "Psych," "SU," "Neuro," "Sleep," "non-brain," as defined in **Appendix C Data S4**). Variants may be pleiotropic–fine-mapped from multiple traits within a trait group, so we selected the highest PIP of each variant and filtered out variants with marginal probability of being fine-mapped variants as previously done (PIP > 0.10) [294]. We annotated the fine-mapped variants by the `hg38` location to be in one of the Cell-TACIT Age quartiles or outside these annotations and similarly computed enrichment odds ratio that a variant is a

non-marginal fine-mapped brain variant (vs. "non-brain") and overlaps a Cell-TACIT Age quartile (vs. not being in a Cell-TACIT Age annotation) using a generalized linear model with a binomial link `trait_group ~ quantile.` We had fewer fine-mapped SNPs across trait groups and Cell-TACIT quartiles compared to >10 million eQTLs, so we used the FDR correction at alpha = 0.05 to determine if fine-mapped SNPs tend to be in Cell-TACIT Age quartiles. We did the FDR correction over 4 Cell-TACIT Age quartiles by 8 Cell-TACIT cell type annotations by 5 trait groups (**Appendix C Data S8**). For plotting, we transformed the estimated log odds ± standard error to the regular odds ratio. (**Figure 3C, Appendix B Figure S28B**).

3.4.13 Selection of conserved, active cell type enhancers and validation

To test the prediction of conserved striatal cell type enhancer activity across species, we selected two human MSN candidate enhancers and the corresponding mouse orthologs to test *in vivo* with a reporter assay plasmid, similar to previous strategies for validating candidate cell type-specific enhancers [295]. We selected two human candidate enhancers upstream of the *DRD2* locus to validate Cell-TACIT model predictions of conserved cell type enhancers according to the following criteria: 1) human sequences predicted active by Cell-TACIT D1 and D2 models, 2) the mouse orthologous region are predicted active or inactive by Cell-TACIT models, 3) have open or closed chromatin in D1 and D2 canonical medium spiny neurons in human and mouse that are consistent with Cell-TACIT model predictions, and 4) contain fine-mapped SNPs from neuropsychiatric trait GWAS with PIP > 0.10. The first human sequence, called HSA (hg38 chr11:113567227-113567453), is an open chromatin peak summit-centered 227 bp region that contains fine-mapped SNPs rs7933981 allele G and rs61687445 allele C reported in **Figure 3B**. The mouse ortholog, called MMA (mm10 chr9:49210951-49211177), overlaps mouse D1 and D2 open chromatin region and is predicted

by Cell-TACIT models to be strongly active in mouse D1 and D2 MSNs. The second sequence, called HSB (hg38 chr11:113577834-113578060), contains rs7928017 allele C. The mouse ortholog, called MMB (mm10 chr9:49196535-49196761), does not overlap any mouse D1 or D2 open chromatin region and is not predicted by D1 and D2 Cell-TACIT models to have conserved activity.

3.4.13.1 Reporter Assay Sequence Predictions

We predicted the open chromatin of the orthologs of human candidate enhancer sequence in clades of placental mammals diverged from humans (**Figure 4C**) and of the shorter, central 227 bp of each open chromatin region that was used to clone into the reporter plasmids (**Figure 4D**). We padded the shorter sequences with N's to be the 501bp required to predict open chromatin with Cell-TACIT models. To get predictions at the level of neurons and glia for comparing to NeuN-stained mouse brains, we computed the weighted average across each Cell-TACIT model prediction by the proportion of that cell type in the human caudate nucleus, estimated from the proportion of cell types in the *Corces et al.* snATAC-seq across 3 subjects [296]. The estimated proportions within neurons are interneurons = 7.67%, D1 MSN = 42.2%, D2 MSN = 44.9%, and D1H MSN = 5.3%. The estimated proportions within glia are astrocytes = 6.97%, microglia = 10.4%, oligodendrocytes = 76.2%, and oligodendrocyte precursors 6.46%. To calculate the weighted average standard error, we used the `wt.var()` function in the `Hmisc` R package, which calculates the weighted variance that we converted to the weighted standard error with N=4 Cell-TACIT models for either the average neuron or glia predictions.

3.4.13.2 Image Analyses and Statistics

We counted nuclei in each fluorescent channel using the python package DeepCell pre-trained Mesmer nuclear detection with default parameters aside from the micro-per-pixel correction factor [161,297–300] and post-processed the segmented nuclei with scikit-image [301]. We

excluded nuclei towards the perimeter of the tiled image not from the mouse caudoputamen. We filtered the segmentation by shape factors that are likely artifacts rather than biological nuclei with the `regionprops_table` function from scikit-image v0.19.3 [161,297], diameter between 5 μm and 25 μm, and segmentation solidity greater than 92.5%. We assigned mCherry- and GFP-detected nuclei as NeuN+ if they had more than 5% overlap with NeuN nuclei. We summarized the segmentation across biological replicates to determine the mean mCherry nuclei per mm^2 and standard error of the mean as plotted in Figure 4E.

We analyzed the number of mCherry detected nuclei per mm^2 as linear mixed effect model accounting for cohort batch effects, viral transduction control GFP nuclei, and repeated imaging from each animal using the R package lme4 [302]. We applied statistical model to account enhancer of origin and species of origin: `mCherry ~ Enhancer + Species + Cage + GFP + (1 | Animal_ID)`.

3.4.14 Statistical corrections for multiple comparisons

In the τ^* LD score regression analyses, we report the p-values that cell type annotations have a non-zero conditionally independent effect. The coefficients estimate the annotation's contribution to heritability conditionally independent from the baseline annotations and a background set of cell type OCRs and open chromatin regions. The τ^* p-value is adjusted for FDR correction across all combinations of 64 traits and 8 cell types measured in 7 cell type OCR sets, relevant to **Appendix B Figure S4**, **Appendix C Data S5**.

In the heritability enrichment LD score regression analyses, we report the p-values that cell type annotations have non-zero heritability enrichment. The enrichment estimates the proportion of heritability in the SNPs in the annotation over the number of SNPs in the annotation and the total heritability of the trait. This enrichment p-value is adjusted for FDR correction across all combinations of 64 traits and 8 cell types measured in 26 annotations sets

(Cell-TACIT predictions in species, PhyloP annotations, and genome mappability), relevant to **Figure 2**, **Appendix B Figure S11-26**, **Appendix C Data S7**. Statistical significance is determined by FDR lower than an alpha of 0.05.

In the eQTL enrichment logistic regression analyses, we report the p-values that the log odds of significant eQTLs in Cell-TACIT Age quartile bins has a coefficient that is non-zero. The log odds p-value is adjusted for Bonferroni correction across all combinations of 1 caudate eQTL region or 8 single cell eQTL cell types by 8 Cell-TACIT cell type models, relevant to **Figure 3A-C**, **Appendix B Figure S28**, **Appendix C Data S8**.

3.5 Acknowledgments

We thank members of the Pfenning lab, Diane P. Genereux, Ruby A. Holland, Michael Leone, Patrick F. Sullivan, Samuel J. Dienel, and members of the Zoonomia Consortium for discussions and feedback. We thank Jacquelyn Breter for taking excellent care of animals.

3.5.1 Funding

This study was funded by the following sources: National Institutes of Health NIDA grant DP1DA046585 (ARP), National Institutes of Health NIDA grant F30DA053020 (BNP), National Institutes of Health NIMH grant UG3MH120094 (WRS), Carnegie Mellon University Computational Biology Department Lane Fellowship (IMK), Carnegie Mellon University Neuroscience Institute (AK).

3.5.2 Author contributions

Conceptualization, Writing – original draft: BNP, ARP. Methodology: BNP. Visualization: BNP, AK, AJL, Software: BNP, DES. Investigation: BNP, JH, ARB, GAF, CS,

IMK. Formal Analysis: BNP, AJL, AK. Resources: ARP, AJL, ARB, JH, WRS. Data Curation: BNP, AJL, ARB, IMK. Writing – review & editing: all authors. Supervision: ARP. Funding acquisition: ARP, WRS

3.5.3 Data and materials availability

Code for analyses and cross-species convolutional neural network models of caudate cell types used in this study are available at https://github.com/pfenninglab/Zoonomia_Cell-TACIT_fine-mapping. Open chromatin regions from human, macaque, and mouse are uploaded within the github repository. Rhesus macaque snATAC-seq sequencing reads are deposited to GEO at under accession `#GSE188983` accessible with the token `glwhuamqdlgphgx`. Polyfun annotations for fine-mapping and cell type ATAC-seq peaks mapped across 241 mammals, are deposited upon publication at https://kilthub.cmu.edu and the UCSC genome browser as custom loadable tracks. During the review process, these data are available at: https://www.icloud.com/iclouddrive/0mNrBC3-i_0ieyF72miR9s4FQ#polyfun_annotations and https://www.icloud.com/iclouddrive/0bDnv_SXJJHl6Ruc2GVk-ZG_Q#zoonomia_peaks. Other intermediate files will be available by request to the corresponding author.

3.5.4 Inclusion & Diversity

One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community.

4.0 Single nuclei transcriptomics of the human and monkey striatum implicates DNA damage, neuroinflammation, and neurodegeneration signaling in opioid use disorder

BaDoi N. Phan[1,2,3], Madelyn H. Ray[4,5], Xiangning Xue[6], Robert J. Fenster[7,8], Stephen J. Kohut[7,9], Jack Bergman[7,9], Suzanne N. Haber[7,10], Kenneth M. McCullough[11], Madeline K. Fish[13], Jill R. Glausier[14], Qiao Su[1], Allison E. Tipton[13], David A. Lewis[14], Zachary Freyberg[14,15], George C. Tseng[6], Shelley J. Russek[4,12,13], Yuriy Alekseyev[16], Kerry J. Ressler[7,8], Marianne L. Seney[14], Andreas R. Pfenning[1,2†], Ryan W. Logan[4,17,18†]

[1]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, United States, 15213.
[2]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, United States, 15213.
[3]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States, 15213.
[4]Department of Pharmacology, Physiology & Biophysics, Boston University Chobanian and Avedisian School of Medicine, Boston, MA, United States, 02118.
[5]Whitaker Cardiovascular Institute, Boston University School of Medicine, Boston, MA, United States, 02118.
[6]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, United States, 15213.
[7]Department of Psychiatry, Harvard Medical School, Boston, MA, United States, 02115.
[8]Division of Depression and Anxiety, McLean Hospital, Department of Psychiatry, Harvard Medical School, Belmont, MA, 02478.
[9]Behavioral Biology Program, McLean Hospital, Belmont, MA, 02478.
[10]Department of Pharmacology and Physiology, University of Rochester, School of Medicine, Rochester, NY, 14642.
[11]Basic Neuroscience Division, Department of Psychiatry, Harvard Medical School, McLean Hospital, Belmont, MA, 02478.
[12]Center for Systems Neuroscience, Boston University, Boston, MA, United States, 02118.
[13]Graduate Program for Neuroscience, Boston University, Boston, MA, United States, 02118.
[14]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States, 15219.
[15]Department of Cell Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States, 15219.
[16]Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, MA, United States, 02118.
[17]Department of Psychiatry, University of Massachusetts Chan Medical School, Worcester, MA, United States, 01605.
[18]Department of Neurobiology, University of Massachusetts Chan Medical School, Worcester, MA, United States, 01605.

This chapter is reproduced, in full, with permission as an author from research provided in pre-print format on BioRxiv. The copyright holder for this preprint is the authors.

SINGLE NUCLEI TRANSCRIPTOMICS OF HUMAN AND MONKEY STRIATUM IMPLICATES DNA DAMAGE, NEUROINFLAMMATION, AND NEURODEGENERATION SIGNALING IN OPIOID USE DISORDER. BADOI N. PHAN, MADELYN H. RAY, XIANGNING XUE, ROBERT J. FENSTER, STEPHEN J. KOHUT, JACK BERGMAN, SUZANNE N. HABER, KENNETH M. MCCULLOUGH, MADELINE K. FISH, JILL R. GLAUSIER, QIAO SU, ALLISON E. TIPTON, DAVID A. LEWIS, ZACHARY FREYBERG, GEORGE C. TSENG, SHELLEY J. RUSSEK, YURIY ALEKSEYEV, KERRY J. RESSLER, MARIANNE L. SENEY, ANDREAS R. PFENNING, RYAN W. LOGAN. BIORXIV 2023.05.17.541145; DOI: HTTPS://DOI.ORG/10.1101/2023.05.17.541145

†Co-corresponding Authors:

Ryan W. Logan, PhD

364 Plantation Street, LRB 611

University of Massachusetts Chan Medical School

Worcester, MA 01605

ryan.logan@umassmed.edu

Andreas R. Pfenning, PhD

Gates Hillman Center Room 7711

Carnegie Mellon University

Pittsburgh, PA 15213

apfenning@cmu.edu

4.1 Abstract

The striatum in the brain is involved in various behavioral functions, including reward, and disease processes, such as opioid use disorder (OUD). Further understanding of the role of striatal subregions in reward behaviors and their potential associations with OUD requires molecular identification of specific striatal cell types in human brain. The human striatum contains subregions based on different anatomical, functional, and physiological properties, with the dorsal striatum further divided into caudate and putamen. Both caudate and putamen are associated with alterations in reward processing, formation of habits, and development of negative affect states in OUD. Using single nuclei RNA-sequencing of human postmortem caudate and putamen, we identified canonical neuronal cell types in striatum (*e.g.,* dopamine receptor 1 or 2 expressing neurons, D1 or D2) and less abundant subpopulations, including D1/D2 hybrid neurons and multiple classes of interneurons. By comparing unaffected subjects to subjects with OUD, we found neuronal-specific differences in pathways related to neurodegeneration, interferon response, and DNA damage. DNA damage markers were also elevated in striatal neurons of rhesus macaques following chronic opioid administration. We identified sex-dependent differences in the expression of stress-induced transcripts (*e.g., FKBP5*) among astrocytes and oligodendrocytes from female subjects with OUD. Thus, we describe striatal cell types and leverage these data to gain insights into molecular alterations in human striatum associated with opioid addiction.

4.2 Introduction

Rates of deaths from opioid overdose and people diagnosed with opioid use disorder (OUD) are continuing to rise in the United States[303–306]. Efforts to generate new treatment strategies and to bolster avenues of therapeutic development to treat opioid addiction require a deeper understanding of the changes that occur in the human brain with chronic opioid use. To date, few studies have investigated the cellular and molecular changes in human brain associated with OUD[307–309]. Recently, we reported alterations in molecular pathways related to neuroinflammation, extracellular matrix[310,311], and synaptic signaling in postmortem brains from unaffected subjects and subjects with OUD[250,312,313]. Markers for microglia were also significantly enriched in OUD subjects, along with significant changes in pathways related to neurodegenerative disorders.

Prior studies highlight putative roles for several pathways related to neurodegeneration in OUD, including dysfunctional aggregation of the tau protein[308,311,314] and activation of proinflammatory cascades. Activation of inflammatory signaling is initiated by a myriad of processes in the brain. Across several neurodegenerative disorders, DNA damage is associated with neuroinflammation and strongly implicated in early pathogenesis[150,315–317] of Alzheimer's disease, Parkinson's Disease, and Amyotrophic Lateral Sclerosis[317]. Elevated DNA damage can activate microglia to release proinflammatory cytokines[318], leading to the activation of oxidative stress[206], metabolic, and immune signaling cascades in other glial cell types and neurons[319,320]. Thus, emerging evidence in human brain reveals changes in pathways involved in the pathogenesis of neurodegenerative disorders to be associated with opioid addiction, with strong implications for the involvement of pathways that regulate DNA damage and neuroinflammation in OUD. Further investigating these molecular alterations associated with

OUD using single cell technologies may yield important insights into the potential roles for DNA damage and neuroinflammation in specific cell types in human brain.

Single cell RNA-based approaches provide high-throughput analyses to characterize the transcriptomes of specific cell types in the human brain. Studies using single nuclei RNA-sequencing (snRNA-seq) can resolve neural cell type specificity from disease-associated molecular alterations previously found using other approaches (*e.g.,* bulk tissue RNA-seq). By stratifying pathways across different cell types, the aim is to uncover new disease-associated molecular mechanisms related to specific neuronal and glial cell types. In human brain, the striatum is anatomically divided into ventral and dorsal regions, whereby the dorsal region comprises the caudate and the putamen. Altered functioning of the dorsal striatum is implicated in several psychiatric and neurological disorders. In OUD, the dorsal striatum is of particular interest, given the brain region's critical roles in reward processing, habitual drug-seeking, craving, and relapse, as well as the development of negative affective states during opioid withdrawal[318]. However, only a few studies have characterized the various cell types in the human dorsal striatum and essentially none have explored transcriptional alterations in striatal cell types in subjects with OUD.

To identify transcriptional alterations associated with OUD in specific striatal cell types, we conducted snRNA-seq on both the caudate and putamen of dorsal striatum from unaffected comparison subjects and subjects with OUD (98,848 total nuclei from both caudate and putamen of 6 unaffected subjects and 6 subjects with OUD, evenly divided between females and males). Using these data, we identified transcriptionally and anatomically heterogeneous cell types in human striatum, followed by differential expression analysis comparing unaffected and OUD subjects. We found significant transcriptional changes in neuroinflammatory and neurodegenerative pathways, implicating neuronal DNA damage responses and activated glial responses of microglia in OUD. Elevated DNA damage markers

were also found in specific neuronal cell types in the striatum of rhesus macaques following chronic opioid administration. Additionally, we explored the interaction between sex and disease, finding sex- and cell type-specific transcriptional programs associated with OUD. As a resource, we deposited the annotated and filtered single nuclei transcriptomes for human caudate and putamen in unaffected and OUD subjects on the CZ CELLxGENE Discover Portal.

4.3 Results

4.3.1 High quality single nuclei transcriptomics identifies canonical and low abundant cell types in human dorsal striatum

To investigate the molecular features of human dorsal striatum, we collected both caudate and putamen from unaffected subjects (n=6, 3 females and 3 males, for each subregion). Unaffected subjects were then compared to subjects with OUD to identify cell type-specific molecular alterations associated with opioid addiction (n=6, 3 females and 3 males, for each subregion). Both striatal regions were dissected from frozen postmortem tissue samples and processed to generate nuclei suspensions for snRNA-seq (**Figure** 1b). Based on quality control assessments, we used a total of 98,848 sequenced nuclei for subsequent analyses (~70% of captured nuclei with an average of 5,757 nuclei per caudate and 3,440 nuclei per putamen; **Figure** 13d). The number of nuclei or percent nuclei captured were similar across all subjects (p=0.40, linear regression, **Appendix D Table** S1). Each nucleus was deeply sequenced at 70% saturation rates, detecting an average of 3,513 ± 77 genes (mean ± standard error) and 13,635 ± 560 unique transcripts per nucleus per subject (**Figure** 13e). The average number of genes detected and the number of unique mapped identifiers for transcripts were

similar between unaffected and OUD subjects (p=0.55, linear regression, **Appendix D Table S1**).

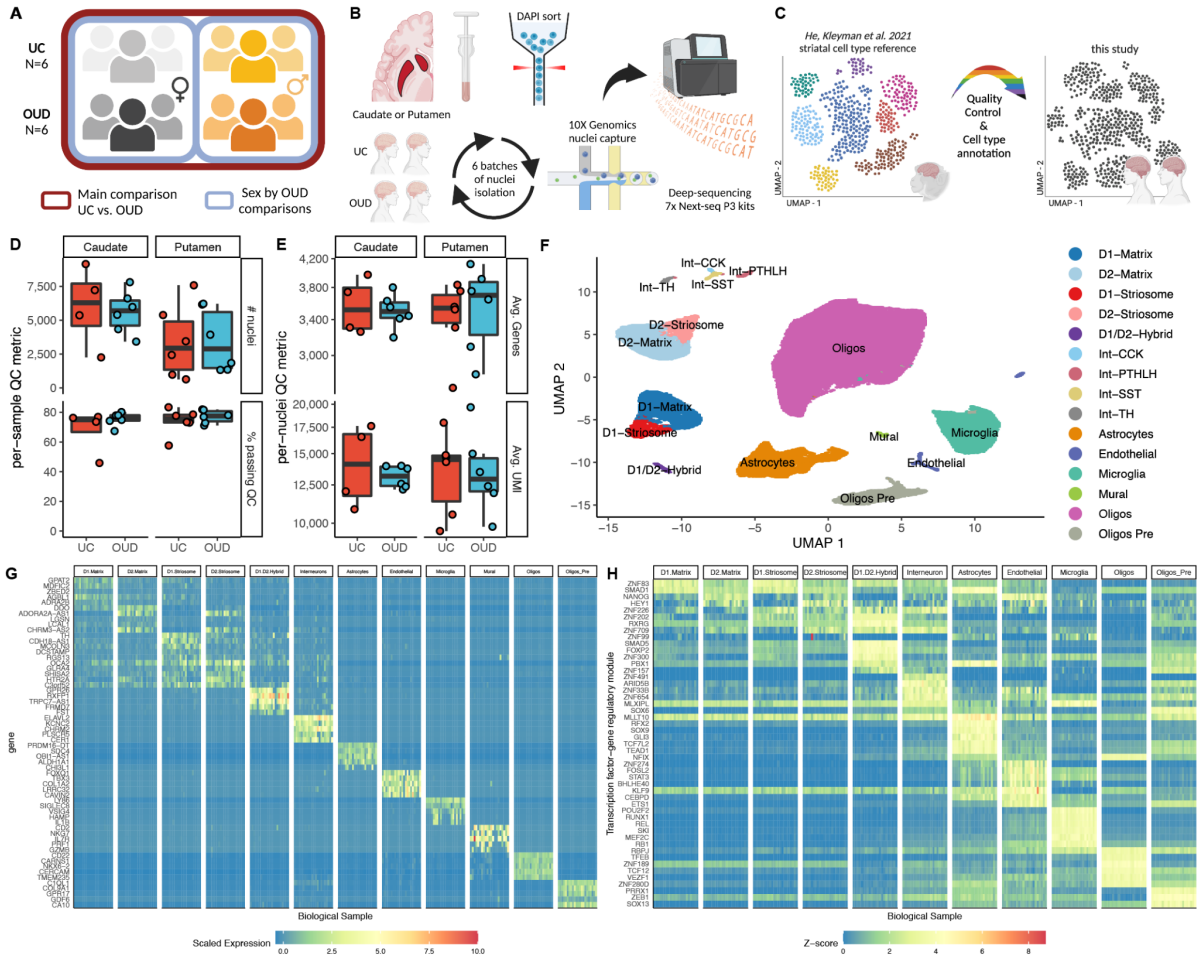


Figure 13. Single nuclei RNA-sequencing of postmortem brain to identify specific cell types in human striatum

(A) Post-mortem human brain cohort design and analysis to compare unaffected comparison (UC) subjects with subjects with opioid use disorder (OUD).

(B) Schematic of the sample collection process to isolate single nuclei and generate single nuclei RNA-seq libraries in balanced batches from the caudate nucleus and putamen.

(C) Single nuclei clustering and label annotation of dorsal striatal cell types using a high quality reference[319].

(D) Per-sample quality control (QC) metrics and (E) average per-nuclei QC metrics across tissue sources and diagnoses.

(F) A low dimensionality projection of striatal cell types after QC filtering and annotation.

(G) Cell type-specific marker genes and (H) marker transcription factor-gene regulatory networks discovered from single nuclei RNA-seq striatal tissues.

We used multiple rounds of graph-based clustering, cell label transfers and refinement from a high-quality, high-resolution snRNA-seq dataset from non-human primate striatum[322,323] (**Figure** 13c). The first round identified major cell classes of glia and neurons: astrocytes (*AQP4*+), endothelial cells (*CLDN*+), microglia (*CX3CR1*+), mural cells (*CD44*+), oligodendrocytes (*ASPA*+), oligodendrocyte precursors (OPC, *PDGFRA*+), and neurons (*RBFOX3*+, *GAD2*+) (**Figure** 14). The second round identified major striatal neuronal subtypes, including medium spiny neurons (MSN, *PPP1R1B*+) and interneurons (*LHX6*+) (**Figure** 14). Recapitulating the heterogeneous striatal neuronal subtypes previously described in our non-human primate datasets[320,321], our snRNA-seq analysis identified canonical medium spiny neuron (MSNs), including *DRD1*+ and *DRD2*+ MSNs within distinct neurochemical compartments, the striosome (*KCNIP1*+) and matrix (*EPHA4*+).

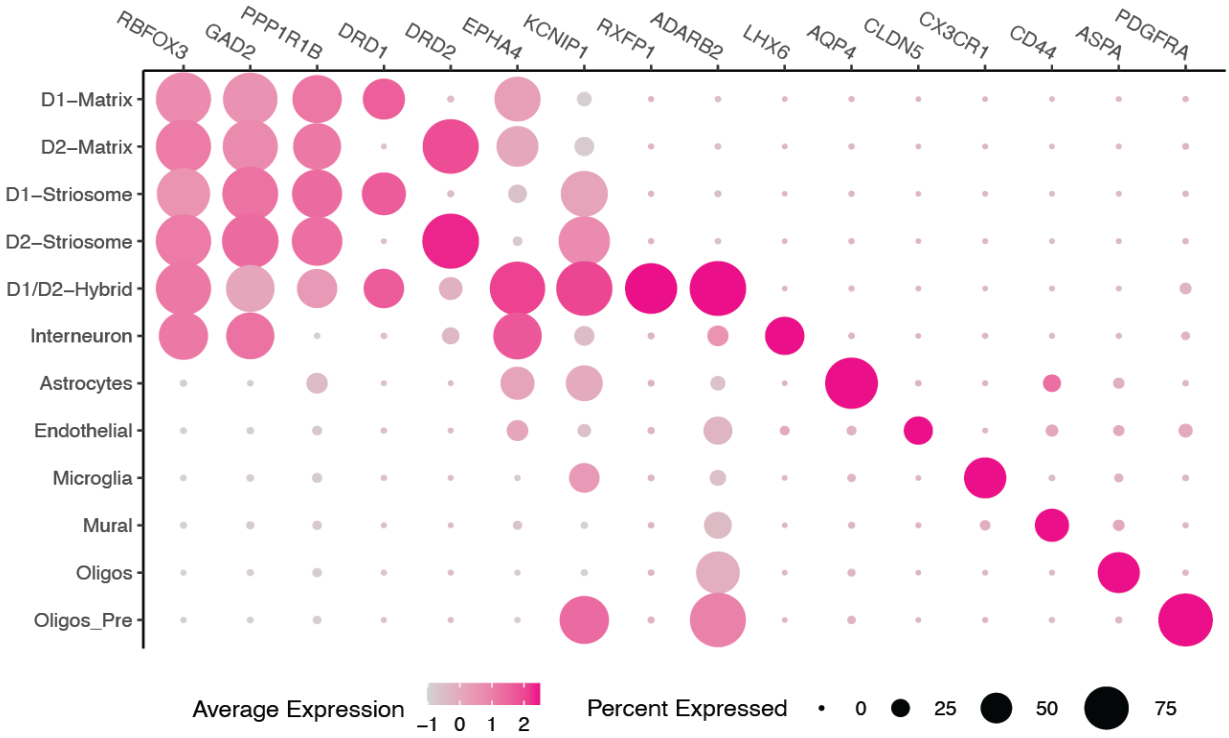


Figure 14. Marker genes of annotated cell types in the human dorsal striatum

Dot plots of the marker genes used to annotate the various cell types of the dorsal striatum based on a non-human primate reference. The normalized expression patterns are averaged across all cells and individuals.

In addition, we identified less abundant neuronal subtypes including MSNs expressing both *DRD1* and *DRD2* (D1/D2 hybrid cells, known as eccentric MSNs or D1H in mice)[161,298], mural cells, and several types of interneurons (*e.g.*, *CCK*+, *PVALB*+, *SST*+, and *TH*+; **Figure** 13f). Previously, interneuronal subtypes have only been described in mouse brain single cell RNA-seq datasets[322]. For subsequent analyses, we pooled interneuronal subtypes because of their comparatively lower abundance to other neurons, and largely shared transcriptional signatures. Overall, relative proportions of low or high abundance striatal cell types were consistent across brain regions, between sexes, and between unaffected and OUD subjects (FDR=0.25, mixed effects linear regression, **Figure** 15, **Appendix D Table S**2).

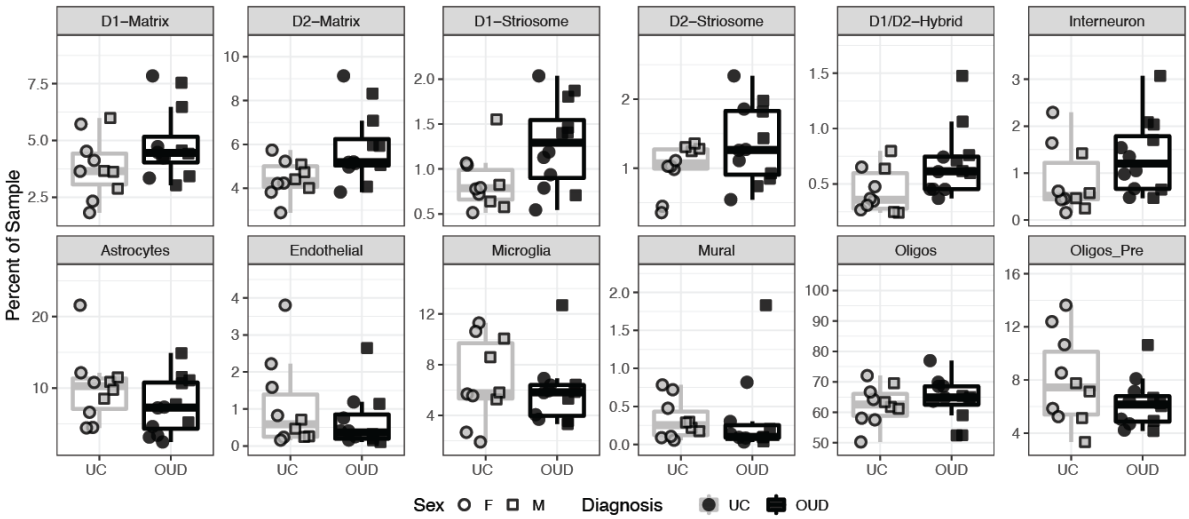


Figure 15. Differential proportion of striatal cell types across biospecimens and OUD diagnosis

Boxplot showing the relative percent of each cell type detected in each biospecimen that passes quality control. No significant differences were observed between unaffected comparison and OUD individuals.

4.3.2 Neuronal and glial cell type-specific expression of opioid receptors in human striatum

Opioids, including endogenous opioids, enkephalin, dynorphin, and β-endorphin, differentially activate several classes of opioid receptors, including mu, delta, and kappa, encoded by *OPRM1, OPRD1,* and *OPRK1*, respectively. Previous efforts have mapped striatal expression of opioid receptors in non-human primates and rodents, characterizing species-specific and cell type-specific patterns of expression[323,324]. We investigated the expression of opioid receptors and cognate ligands within human dorsal striatum, revealing several notable results related to specific cell types and potential species differences in expression. First, *OPRM1* was detected in each MSN subtype with the highest level of expression in D1/D2 hybrid and D2-striosome MSNs (**Figure** 16), indicating that, in contrast to rodents, *OPRM1* expression is also in the striosome in human brain. *OPRM1* in other human striatal regions is

consistent with prior findings in non-human primate striatum[325]. Additionally, we reproducibly detected *OPRM1* in microglia across subjects and biological samples, consistent with findings in rodents[161,299], establishing a potential pathway for opioids to directly modulate neuroimmune processes[326].

D1-striosome MSNs expressed the gene encoding the preproprotein, prodynorphin (*PDYN*), while D2-matrix and D2-striosome MSNs coexpressed *OPRD1* and proenkephalin (*PENK*) at markedly higher levels than other striatal cell types. Preferential expression of *OPRD1* and *PENK* in D2 MSNs in human striatum is consistent with rodent striatum[327,328]. Across MSNs, we detected *OPRK1* expression at many folds lower than *OPRM1*, *OPRD1*, *PENK*, or *PDYN*, suggesting that targets of the dynorphin-kappa opioid receptor signaling by D1-striosome MSNs may primarily be outside of dorsal striatum[329]. Between unaffected and OUD subjects, we found no significant differences in opioid receptor and/or endogenous ligand expression within specific cell types. Other mechanisms involved in the post-transcriptional and/or post-translational regulation of opioid receptor and ligand expression, along with receptor binding activity, may be altered in OUD.

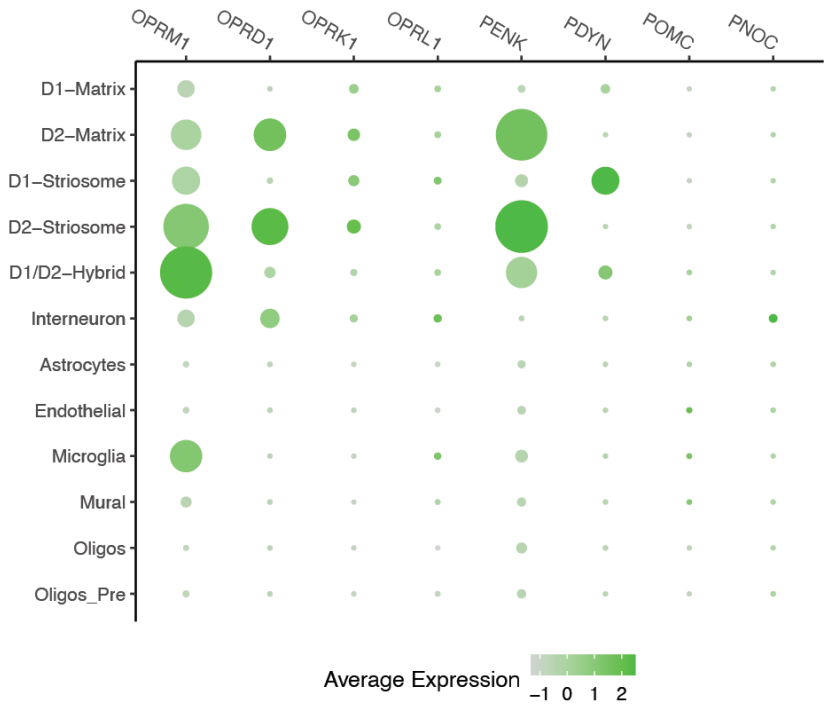


Figure 16. Expression patterns of striatal opioid receptor and endogenous ligands

Dot plots of the opioid receptor and endogenous ligands across the annotated striatal cell types. The normalized expression patterns are averaged across all cells and individuals.

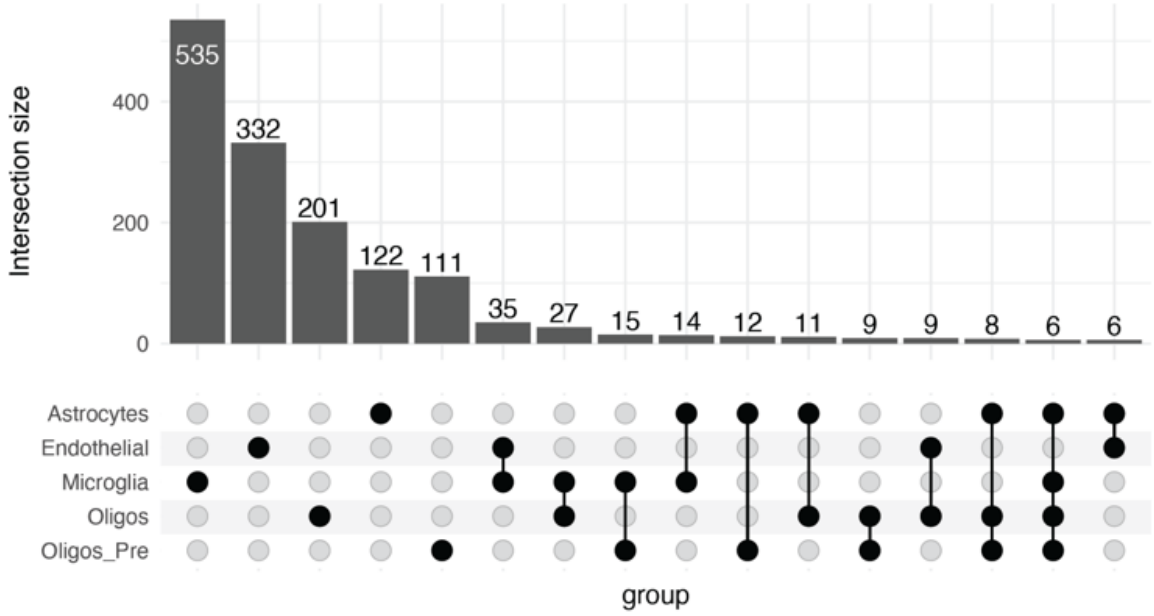
4.3.3 Identification of cell type-specific transcription factor and gene regulatory networks in human dorsal striatum

Following cell annotations, we investigated novel transcriptional networks that represent putative regulatory networks in specific striatal cell types. First, we identified marker genes for each of the annotated cell types that were reproducible across biological replicates (mean ± standard error, 722 ± 172 specific marker genes per cell type, **Figure** 13g, **Appendix D Table S3**). Glial cell types yielded more marker genes (1,134 ± 204) compared to neuronal cell types (310 ± 131). Next, we inferred gene regulatory networks directly from each of the cell type-specific marker datasets using machine-learning (SCENIC)[330,331]. Briefly, SCENIC is based on the premise that transcriptional regulation of gene expression is, in part, based on transcription factor binding to proximal gene promoters. Based on single cell gene expression and the enrichment of promoter binding sites, we used SCENIC to build transcription factor gene regulatory modules that were specific to each of the neural cell types identified in human striatum. Collectively, we found many transcription factors and gene regulatory modules per cell type (47 ± 15 specific modules; glial subtypes, 90±19; neuronal subtypes, 12 ± 6; **Figure** 13h, **Appendix D Table S4**) Together, the annotated cell type clusters in human dorsal striatum, their associated marker genes, and related transcription factor gene regulatory modules offer new insights into the molecular signaling pathways that contribute to the heterogeneity of cellular identity in human striatum.

4.3.4 Differentially expressed genes in specific striatal cell types associated with OUD

We assembled a cohort of subjects with OUD based on time since clinical diagnosis of at least four years and deceased from opioid overdose, as described previously[332]. Subjects with OUD were matched with unaffected subjects on sex, age, postmortem interval (PMI), and RNA integrity number (RIN). (**Appendix D Table S5**, n=6 unaffected subjects and n=6 OUD subjects with 3 females and 3 males for both caudate and putamen; **Figure** 13a). We first aggregated gene expression profiles across cell types to generate pseudobulk profiles for each major cell type. We then directly compared gene expression profiles across neuronal and glial subtypes between unaffected and OUD subjects to identify differentially expressed genes (DEGs), weighted for cell type proportions between groups, while also considering covariates (age, PMI, RIN, cell type abundance, gene detection rate, and surrogate variables). We identified 1,765 DEGs across any cell types (197 ± 43 genes per cell type; FDR < 0.05, **Appendix D Table S6**), with more DEGs found in glial cells relative to neurons (**Figure** 18a, 3a). Further, DEGs in OUD subjects were largely unique to each glial cell type, while DEGs

overlapped more between neuronal cell types (



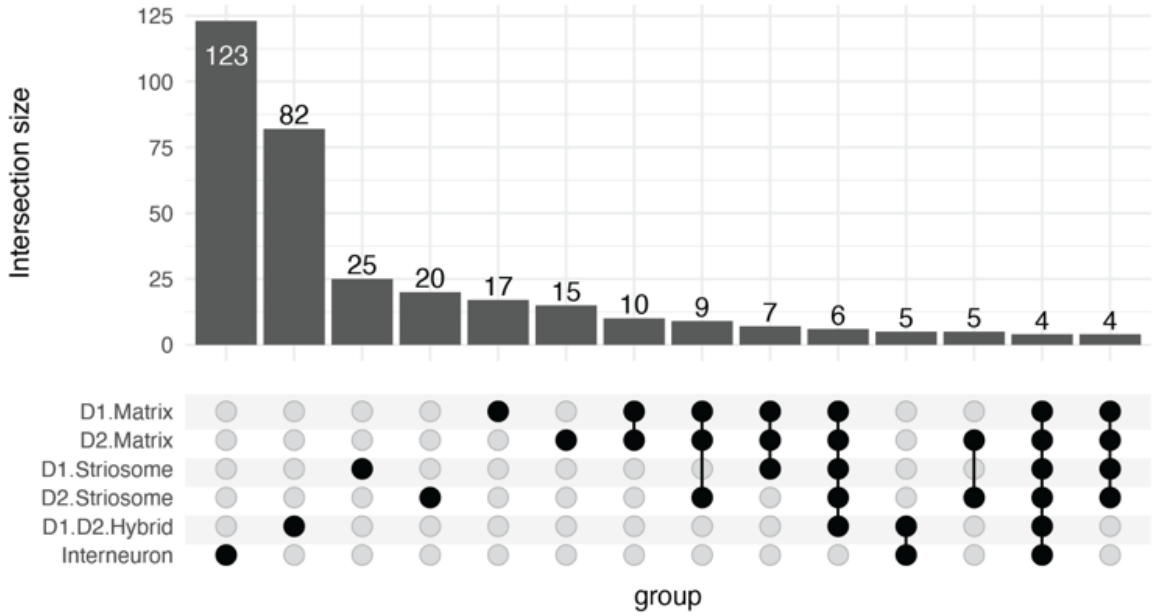


Figure 17).

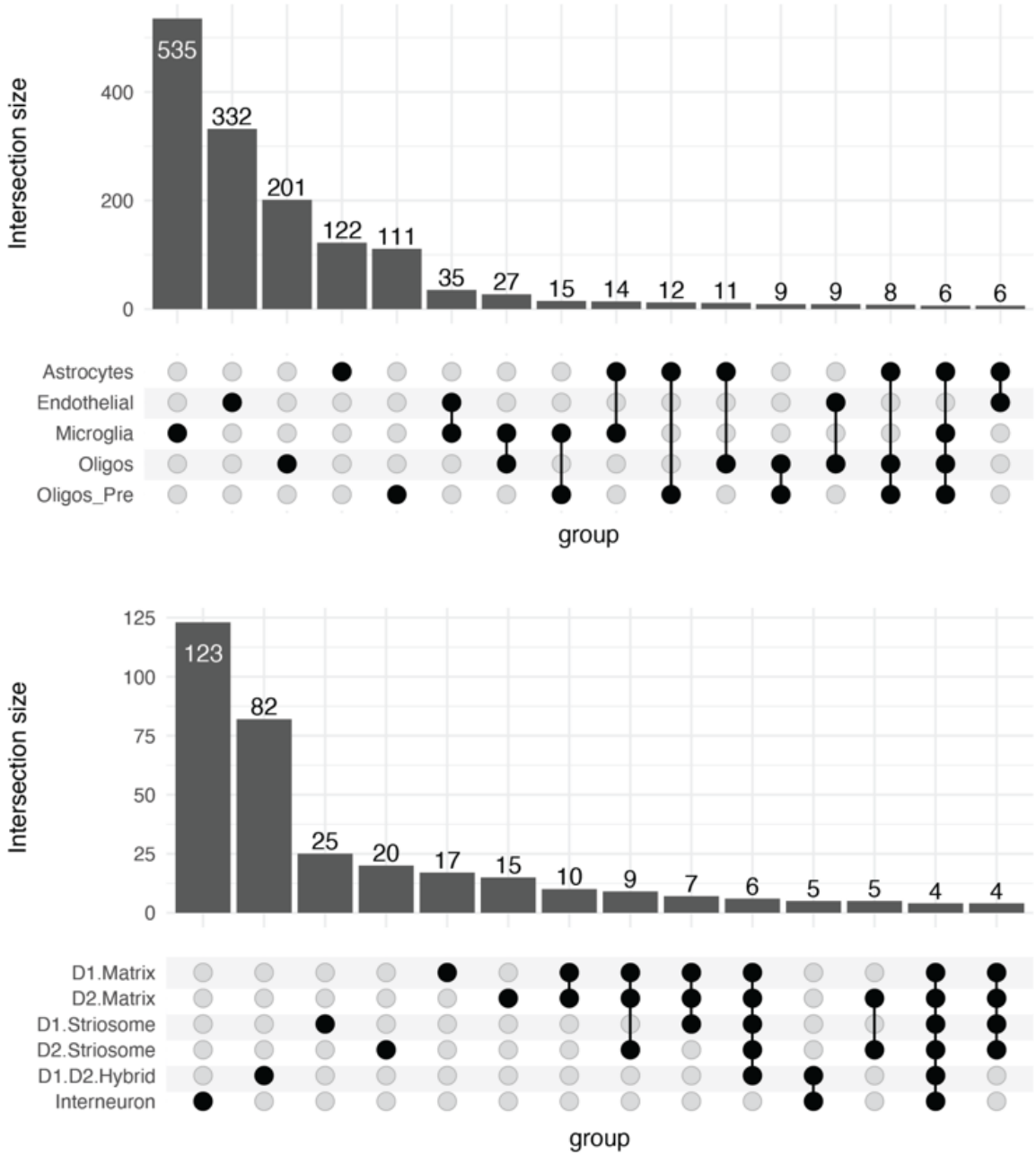


Figure 17. Upset plot of overlapping differentially expressed genes in OUD

Upset plot showing the overlaps in significantly differentially expressed genes (FDR < 0.05) across glia (top) or neurons (bottom). The histogram shows how many genes are in each intersection of cell types that share DEGs.

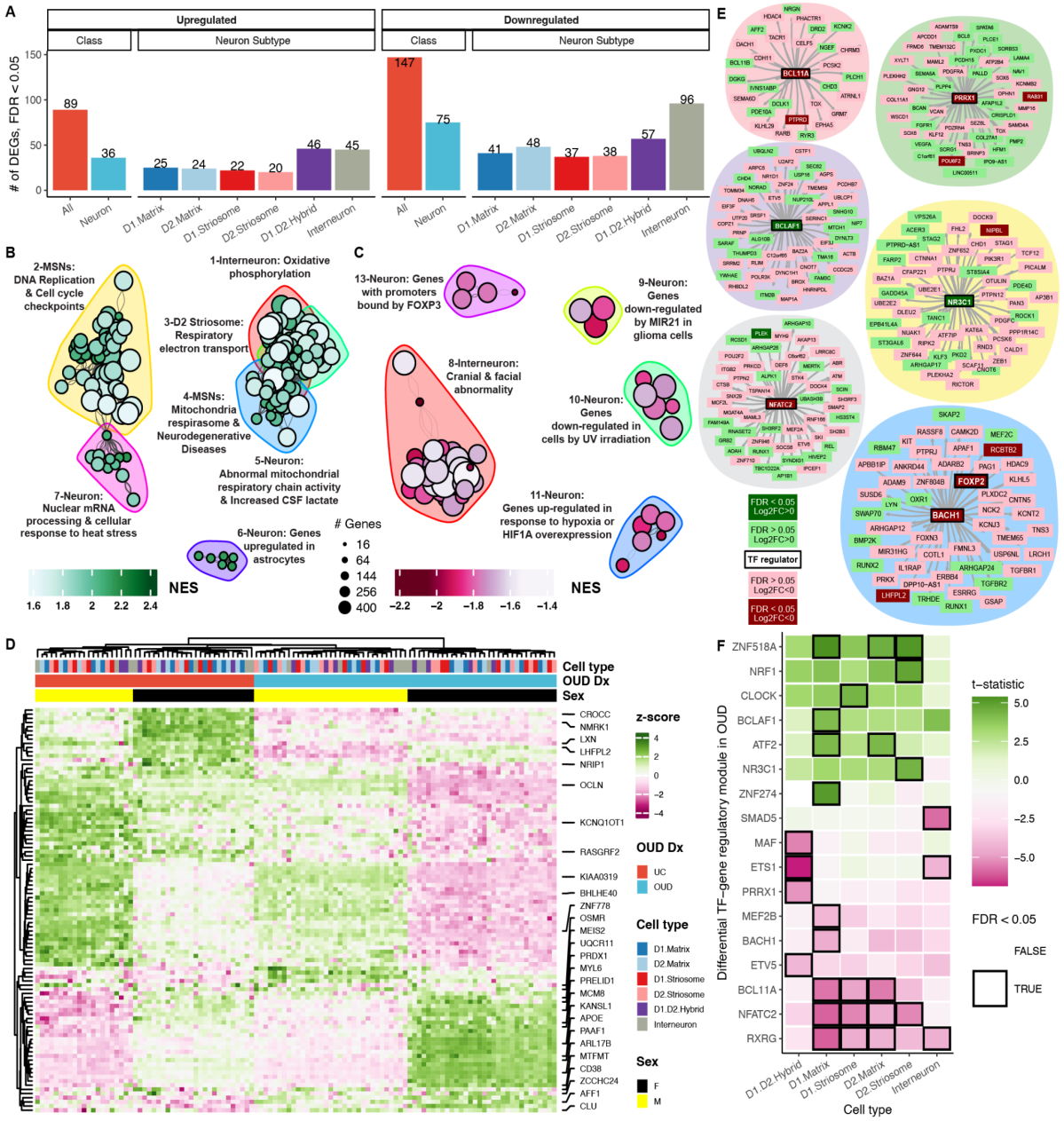


Figure 18. Transcriptional alterations in neuronal cell types in human striatum associated with OUD

(A) Barplot of differentially expressed genes (DEGs) either up- or down-regulated in striatal neurons at a false discovery rate less than 0.05 (FDR < 0.05).

(B-C) Network plot of significant clusters and enriched pathways, where each point is a significantly enriched pathway in a neuronal cell type and lines represent the proportion of shared genes between two pathways. Each point is colored by the normalized enrichment score (NES), indicating whether a pathway is enriched in up- or down-regulated genes. The color

outlines represent the unique clusters of interconnected pathways by shared genes and the nearby text labels the unique cluster number and briefly summarizes which cell types and pathways are represented.

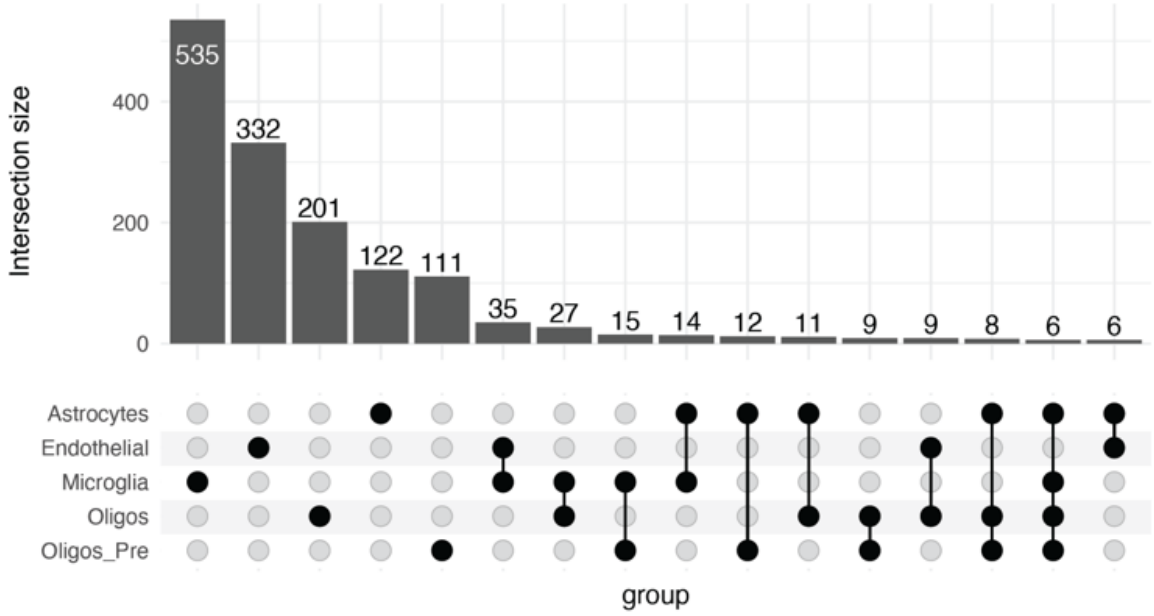
(D) Expression heatmap of z-normalized pseudobulk gene expression, where the rows are DEGs and the columns are the biological replicates labeled by cell type, diagnosis, and sex of each subject. Genes that are represented by the enriched pathways in (B-C) are labeled along the right margins.

(E) Cluster maps of select transcription factor gene regulatory networks. Each transcription factor or gene is colored to denote the direction and significance of DE in neuronal subtypes in OUD.

(F) A heatmap plotting the t-statistic of differentially activated transcription factor gene regulatory networks across neuronal subtypes in OUD.

Using gene set enrichment analyses (GSEA), we identified pathways significantly enriched in aggregate cell types and across specific cell types. Across cell types, we found 286 pathways (34 ± 7 pathways per cell type; FDR < 0.05, **Appendix D Table S7**) and 51 differentially regulated transcription factor regulatory modules (5.1 ± 0.7 per cell type, FDR < 0.05, **Appendix D Table S8**). The cell type-specific transcriptional differences in OUD subjects were represented largely by distinct biological processes between neurons and glia. Among MSNs, DEGs had a higher tendency to be shared between spiny neuron subtypes, while DEGs tended to be more distinct between glial subtypes and between interneuron subtypes. We identified a higher number of DEGs in interneuron subtypes and D1/D2 hybrid MSNs compared to canonical D1 or D2 MSNs. Further, DEGs in D1/D2 hybrid MSNs did not overlap with canonical MSNs, suggesting these sparser populations of striatal neurons displayed a

unique transcriptional response to OUD (



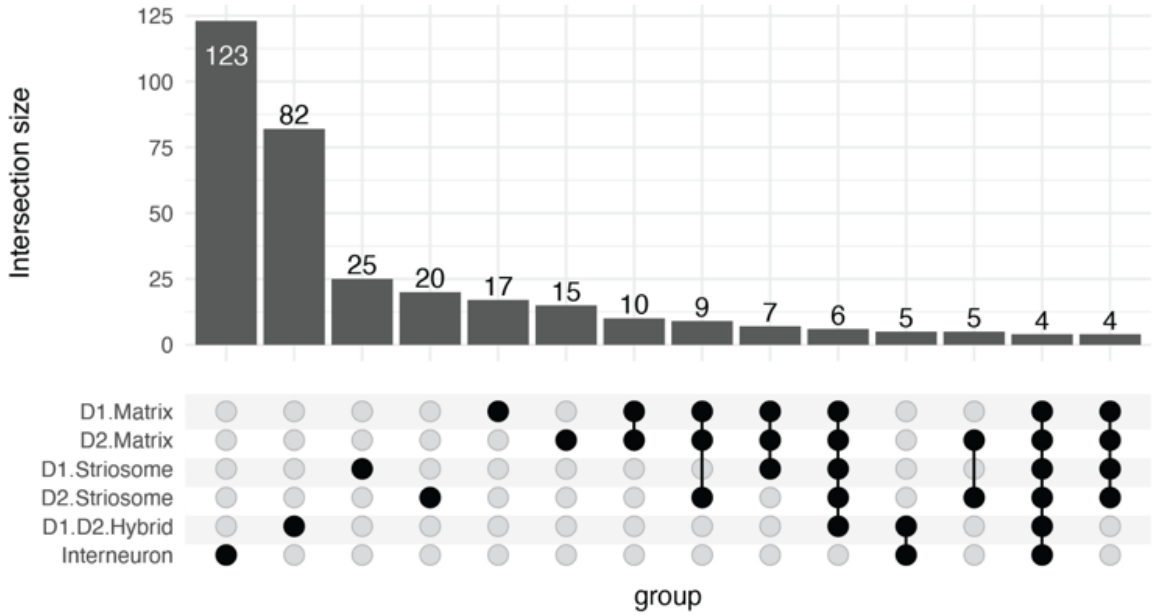


Figure 17).

4.3.5 Pathways in neuronal subtypes converge on neurodegenerative processes in OUD

We conducted separate pathway analyses on upregulated and downregulated transcripts between unaffected and OUD subjects. Among neuronal subtypes, several upregulated

pathways in OUD were associated with neurodegenerative processes. In MSNs, enriched pathways from upregulated DEGs in OUD included DNA replication and cell cycle re-entry (**Figure** 18b, **Appendix D Table S9**), a common feature in neurodegeneration[333–335]. For example, both *MCM8* ($\log_2$ fold-change, log2FC > 1.06, FDR < 0.041) and myosin light-chain 6, *MYL6*, (log2FC >1.13, FDR < 0.045) were significantly upregulated in striatal MSNs. MCM8 is involved in DNA replication, forming a complex with MCM9 to facilitate repair of double-stranded DNA breaks[333–335]. MYL6 is a key factor in rho-GTPase signaling, known to facilitate glutamate receptor endocytosis[336] and neuroplasticity[337,338], and may be involved in DNA damage via the activation of RAC1[339–341]. Other genes related to neurodegeneration included *APOE*[342], which was upregulated in striosome MSNs and D2-matrix MSNs (log2FC >2.50, FDR < 0.025, **Figure** 18d, **Figure** 20d), and *GPX4*, which was upregulated in D1-striosome MSNs (log2FC>1.17, FDR < 0.044) in OUD subjects. Both *APOE* and *GPX4* are involved in oxidative stress within neurons, where GPX4 buffers the accumulation of reactive oxygen species to prevent apoptosis[343]. Accumulation of reactive oxygen species and oxidative stress in neurons involves alterations in redox signaling and mitochondrial respiration[342]. Indeed, across multiple MSN subtypes, genes involved in mitochondrial respiration were significantly upregulated in OUD, including a complex III subunit of the mitochondrial respiratory chain, *UQCR11* (log2FC >1.36, FDR < 0.042), and a subunit of complex I, *NDUFA4* (log2FC>1.29, FDR < 0.030).

Consistent with cell type-specific alterations in pathways related to neurodegeneration, we identified several transcription factor gene regulatory modules that were significantly upregulated in different neuronal subtypes (FDR < 0.05) linked to cell senescence, DNA damage, and inflammation, and stress (**Figure** 18). For example, gene regulatory module linked to BCLAF1 was upregulated in OUD, a transcription factor involved in the transduction of NFkB-dependent signaling and the activation of DNA-damage-induced senescence[42344]

(**Figure** 18e). Another module was associated with the upregulation of the neuron-specific glucocorticoid receptor transcription factor, NR3C1, previously related to the impact of psychosocial stress on human brain[345,346] (**Figure** 18e). Other stress-related transcription factor gene regulatory modules were upregulated in OUD, including ATF2[307,347,348] and ZNF518A[349–352]. Both transcription factor modules were upregulated primarily in D1-matrix and D2-matrix MSNs (FDR < 0.04). ATF2 is induced by stress in mouse dorsal striatum[353–356] and phosphorylated downstream of delta opioid receptor activation[357–361], suggesting delta opioid receptor activation of ATF2-dependent transcription in striatal matrix MSNs[311] may be due to an interplay between stress and opioids in OUD.

Compared to upregulated modules, we identified almost twice as many significantly downregulated transcription factor gene regulatory modules in OUD subjects (Fig 2c). We identified several downregulated pathways that are related to neuroprotection, further implicating processes involved in neurodegeneration in OUD. For example, the humanin-like 8 gene, *MTRNR2L8*, is significantly downregulated in OUD across every neuronal subtype (log2FC <-3.79, FDR < 0.0054, **Figure** 18d). *MTRHR2L8* and the mitochondrial paralog, *MTRNR2*, may act as neuroprotective factors in response to neurodegeneration and the induction of DNA damage[362–364]. The top DEGs within clusters of pathways involved in UV irradiation, another potential link to DNA damage (FDR < 0.029, **Figure** 18c), included *FOXO1*, *TLE4*, *SHOC2*, *MEIS2*, and *THBS1* (log2FC <-0.64, FDR < 0.0422) in interneurons and CLASP1, PIK3C3, and MTUS1 (log2FC <-0.64, FDR < 0.042) in D1/D2 hybrid MSNs. Additionally, downregulation of MEF2B and BACH1 transcription factor gene regulatory modules were found in D1-matrix MSNs in OUD, accompanied by downregulation of MAF, EST1, ETV1, and PRRX1 modules in D1/D2 hybrid MSNs (FDR < 0.043), each of which are implicated in neurodegeneration[365]. Notably, BACH1, BCL11A, and PRRX1 have downstream targets with concordant differential expression in OUD (*e.g., FOXP2*, *POUF3*,

PTPRD, *RAB31*, *RCBTB2*, Fig 2e), indicating coordinated changes in striatal gene networks implicated in neurodegeneration associated with opioid addiction. Additional links to neurodegeneration included modules for NFATC2 and RXRG (FDR < 0.042). Both nuclear factor of activated T-cell C2 (NFATC2) and the retinoid X receptor gamma (RXRG) of the RXR family of receptors link changes in neuronal activity associated with addictive drugs to processes involved in neuroprotection, neurodegeneration,[311,366] and reward-related behaviors[161].

4.3.6 Enrichment of DNA damage markers in specific neuronal subtypes in OUD

To investigate more broadly whether specific processes related to neurodegeneration were altered in OUD, we tested the possible enrichment of DNA damage markers in specific cell types[313,367] between unaffected and OUD subjects (**Figure** 20a). Overall, we confirmed an upregulation of DNA damage gene markers within striatal neurons of OUD subjects (p=0.046; linear regression t=2.18; **Figure** 20b). Within neuronal subtypes, we found a significant increase in DNA damage gene signature specifically within interneurons of OUD subjects (p=0.035, linear regression t=2.14, **Figure** 20c, **Appendix D Table S10**). Together with the significant changes in the expression of *APOE* and *MTRNR2L8* in OUD (**Figure** 20d), our findings suggest striatal neurons may incur DNA damage in response to opioids and other factors, such as stress, neuroinflammation, and hypoxia, associated with opioid addiction. Interestingly, we found a significant downregulation of pathways involved in hypoxia (**Figure** 18c, **Appendix D Table S9**), including decreased expression of genes that are activated in response to hypoxic events (*e.g., NMRK1, BHLHE40*)[368], or overexpression of hypoxia-inducible transcription factor, HIF1A. Therefore, while respiratory depression is a hallmark of opioid overdose, downregulation of hypoxia-responsive genes points toward possible

compensation in neurons secondary to periodic hypoxia and altered redox states, consistent with rodent models of substance use[369–371].

Given the potential confounding factors in our postmortem human brain cohort between chronic opioid use disorder and overdose, we explored how chronic opioids could alter transcriptional markers of DNA damage in striatal neurons in nonhuman primates (**Figure** 20e). Rhesus macaques were dosed with morphine twice daily for 5-6 months for a total of 1,500 mg morphine before tissue collection. Morphine treatment subjects were matched for age, sex, and weight to N=4 non-morphine exposed subjects (**Figure** 19a). The nucleus accumbens of both treatment groups were processed together for snRNA-seq (Methods, **Figure** 19b). Similar to our human dataset, cell types were clustered across rhesus macaque subjects, consistent with our previously described cell populations in rhesus striatum (**Figure** 19c,d). Our gene signature analysis identified a marginal increase of DNA damage marker enrichment among neurons of rhesus macaques chronically treated with morphine ($P = 0.054$, linear regression, **Figure** 20f), with significant increases in specific striatal neuronal cell types ($P < 0.00066$, linear regression, **Figure** 20f). Thus, elevated neuronal DNA damage in non-human primates corroborates those findings in postmortem brain tissue from subjects with OUD, providing additional, cross-species validation of the impact of chronic opioids on striatum.

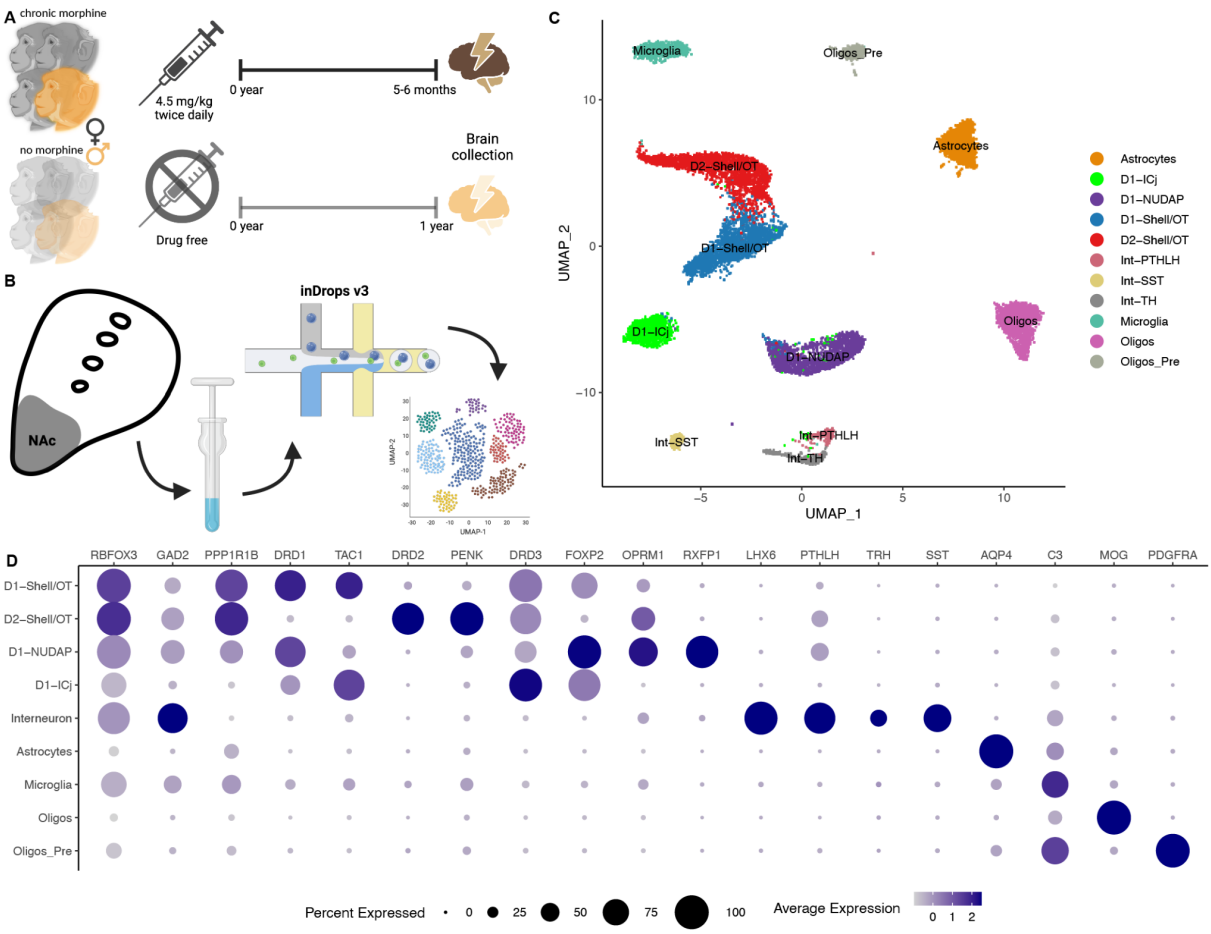


Figure 19. Chronic morphine exposure and single nucleus RNA-seq of the rhesus macaque striatum

(A) Schematic of chronic morphine dosing strategy for 5-6 months prior to brain tissue collection in N=4 rhesus macaque subjects to a total of 1,500 mg of total morphine exposure. N=4 control treatment subjects are matched by age, sex, and weight to the chronic morphine cohort and are unexposed to morphine or other drugs for 1 year prior to brain tissue collection.

(B) Schematic of nucleus extraction from a tissue punch of the nucleus accumbens for inDrops single nuclei RNA-seq library synthesis and clustering.

(C) UMAP projection of post-quality control and annotation of the rhesus macaque striatal cell types using a high-quality reference.

(D) Dot plot showing the marker genes used to validate the annotations of the chronic morphine rhesus macaque striatal single nucleus RNA-seq dataset.

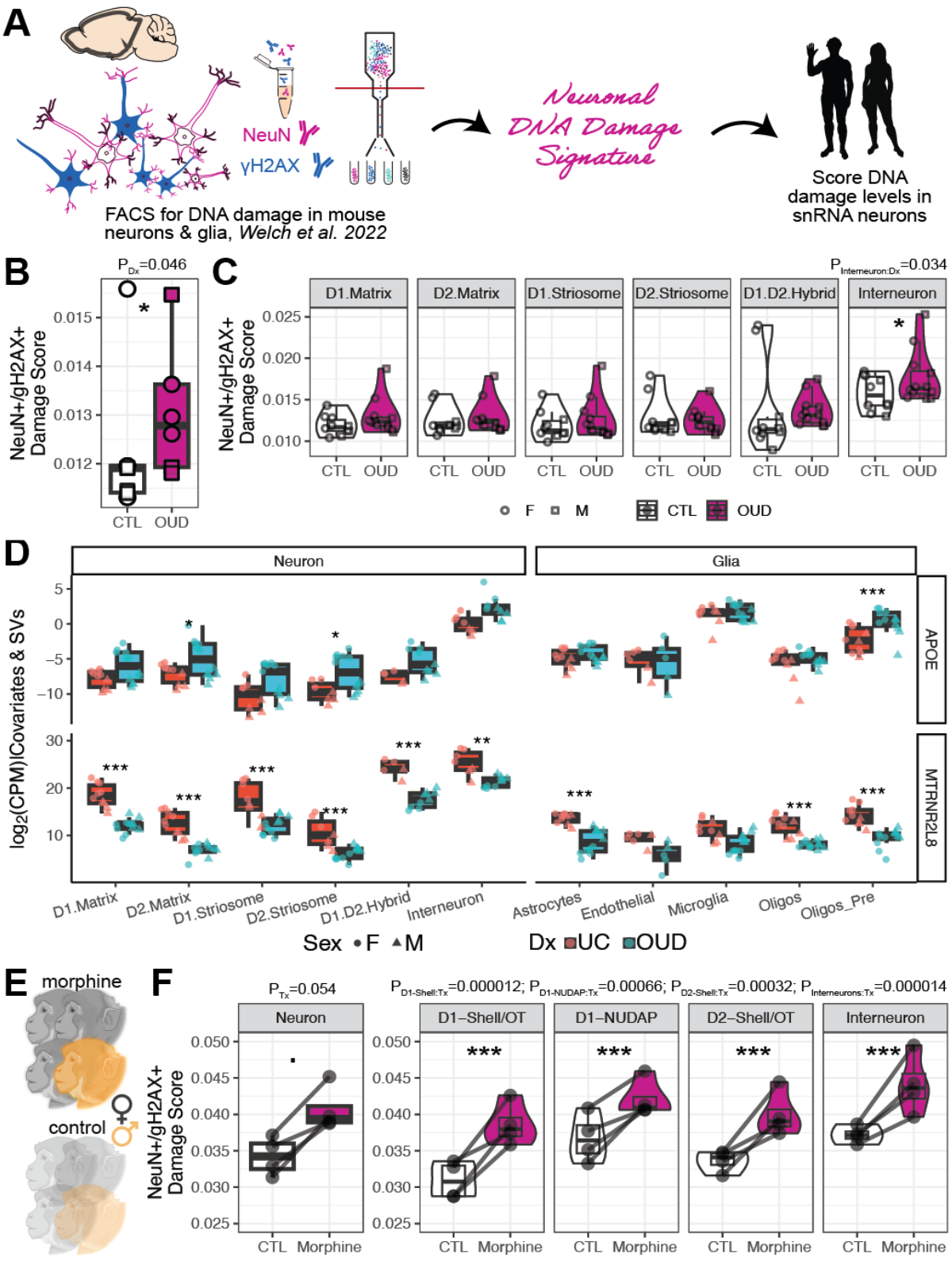


Figure 20. Elevated markers of DNA damage in neurons associated with OUD and chronic morphine

(A) Schematic of the application of DNA damage gene signatures from a mouse model of Alzheimer's Disease17 to score DNA damage in human striatal neurons.

(B) Boxplot of subject-level pseudobulk average DNA damage scores across striatal neurons between unaffected and OUD subjects (PDx = 0.046, linear regression).

(C) Violin-boxplot of cell type-level pseudobulk average DNA damage scores between unaffected and OUD subjects. The significant effect of diagnosis P-values from linear regressions are reported above each plot.

(D) Boxplots showing the log2(counts per million) normalized gene expression values regressing covariates and surrogate variables for the APOE and MTRNR2L8 genes stratified by neuronal and glial cell types. Each point is a biological replicate colored by diagnosis and shaped by biological sex.

(E) Schematic of chronic morphine exposure or unexposed rhesus macaques (N=4 subjects per treatment).

(F) Boxplots of subject level pseudobulk average DNA damage scores across striatal neurons or neuronal subtypes between matched control subjects or those exposed to chronic morphine. Each point represents an individual monkey and lines connected matched pairs. The significant effect of chronic morphine treatment P-values from linear regressions are reported above each plot. (P < 0.05; ** P < 0.01; *** P < 0.001).*

4.3.7 Enriched pathways among glial cells support molecular signatures related to neuroinflammation and neurodegeneration in OUD

Compared to neurons, we identified more than twice as many DEGs in glial cells between unaffected and OUD subjects (**Figure 17a**). The number of DEGs were independent of cell type proportions, as microglia and endothelial cells had the most DEGs among glial cells, and expression patterns were largely distinct between glial cell types (**Figure 15**). An

exception was the robust enrichment of genes involved in interferon response across multiple glial cell types, including astrocytes and oligodendrocytes. Interferon response genes were significantly upregulated in glial cell types in OUD subjects (FDR < 0.047, **Appendix D Table S11**). Upregulated hub genes within the interferon response pathway included: *CMP2*, *HLA-F*, *IFI44*, *PPM1B*, and *RSAD2*. Furthermore, we observed an upregulation of the NFKB1 transcription factor gene regulatory module in astrocytes and the upregulation of the STAT3 module in oligodendrocyte precursor cells (OPC, **Figure** 21f, FDR< 0.036). Several of the predicted NFKB1 gene targets are also differentially expressed including the inhibitor of this pathway, *NFKBIA* (log2FC = 1.86, FDR = 0.039). Other glial cell types showed non-significant activation of this module, with the weakest being microglia. In response to cell stress, including neurodegeneration, microglia release pro-inflammatory cytokines that may trigger an interferon response in other cell types, potentially responsible for the more broad patterns of interferon activation we observe across multiple cell types in human striatum associated with OUD[299].

Altered neuroinflammatory signaling in microglia has previously been reported in postmortem brains from subjects with OUD[372–374]. Several DEGs in microglia of OUD subjects lend further support for associations between augmented inflammatory signaling, synaptic plasticity, and neurodegenerative processes and OUD. For example, the PAX-FOXO1, CDH1, and neuron projection signaling pathways were enriched among microglia DEGs (FDR < 0.049). PAX-FOXO1 signaling regulates cellular senescence in the brain and is involved in the pathogenesis of several neurodegenerative disorders[373,375], while CDH1 signaling is involved in glial cell migration and axonal projections[300]. In the neuronal projection pathway, *ADGRB3* was among the top downregulated genes in microglia of OUD subjects (FDR = 3.5e-7). *ADGRB3* (known as BAI3, brain-specific angiogenesis inhibitor 3) encodes for a protein that has high-affinity for complement, C1q, an innate immune component released by neurons to

eliminate damaged or inactive synapses[300]. Downregulation of *ADGRB3* in microglia may lead to aberrant synaptic formation. In support of synaptic changes associated with opioid addiction, several genes involved in glutamatergic and GABAergic synaptic functions and plasticity were altered in subjects with OUD (astrocytes: metabotropic glutamate receptor 5, *GRM5*, log2FC = -3.13, FDR = 1.06 E -14 and glutamate ionotropic receptor AMPA type subunit 1, *GRIA1*, log2FC = -1.98, FDR = 2.8 E -8; oligodendrocytes: GABA type A receptor subunit beta1, *GABRB1*, log2FC = -0.717, FDR = 0.023; microglia: GABA type A receptor subunit gamma 2, *GABRG2*, log2FC = -1.45, FDR = 0.038). Together, our cell type-specific findings suggest an interplay between microglial-dependent signaling and synaptic plasticity in OUD, consistent with our previous bulk transcriptomics findings in other striatal subregions from subjects with OUD[376]. With single nuclei resolution, we report that the elevation of neuroinflammatory pathways related to microglial activation in OUD is likely due to transcriptional alterations within microglia, rather than pronounced changes in the number of microglia within striatum in subjects with OUD (FDR = 0.48, linear mixed effect regression for differential abundance, **Figure 15**, **Appendix D Table S2**).

Endothelial cells displayed the second most DEGs behind microglia between unaffected and OUD subjects. Upregulated pathways in endothelial cells included several growth factors, including VEGFR, EGFR, and PDGFR signaling[377,378] (**Figure** 21b). These growth factors regulate nociception and opioid tolerance via actions in peripheral circulation and spinal cord[379], suggesting similar pathways in endothelial cells may be involved in alterations associated with OUD in human brain[380]. Other pathways included leukocyte chemotaxis and dendrite morphogenesis. Among the densely connected endothelial pathways, the top significantly upregulated hub gene was the brain-specific angiogenesis inhibitor 1-associated protein 2, *BAIAP2* (log2FC=2.31, FDR = 0.0017), and, interestingly, the

Alzheimer's related microtubule-associated protein tau, *MAPT* (log2FC = 1.58, FDR=0.036). The upregulation of *MAPT* may contribute to neurovascular brain insults in OUD[381–384,385,386].

The identification of differentially expressed transcription factor gene regulatory modules across glial cell types further implicated neurodegenerative processes in OUD. In oligodendrocyte precursor cells, the CEBPB module was upregulated in OUD (**Figure** 21e-f, FDR = 0.031), with prior studies showing CEBPB directly regulates the expression of *APOE* in Alzheimer's disease-related pathologies[387]. Notably, *APOE* is concordantly upregulated in oligodendrocyte precursor cells (**Figure** 20d, log2FC = 2.56, FDR = 5.9e-4), along with other predicted CEBPB targets (**Figure** 20e). The RXRG TF-gene module, which was downregulated in neuronal subtypes, was also downregulated in microglia (**Figure** 20e, FDR = 0.027). An inferred target of RXRG is *FOXP2* (**Figure** 20e), which was recently significantly associated with OUD and other substance use disorders at the genome-wide level[388,389]. FOXP2 expression in microglia may be unique to humans compared to other primates[303,304]. Thus, downregulation of the FOXP2 transcription factor gene regulatory module in human microglia may link the genetic risk of OUD to neurodegenerative processes.

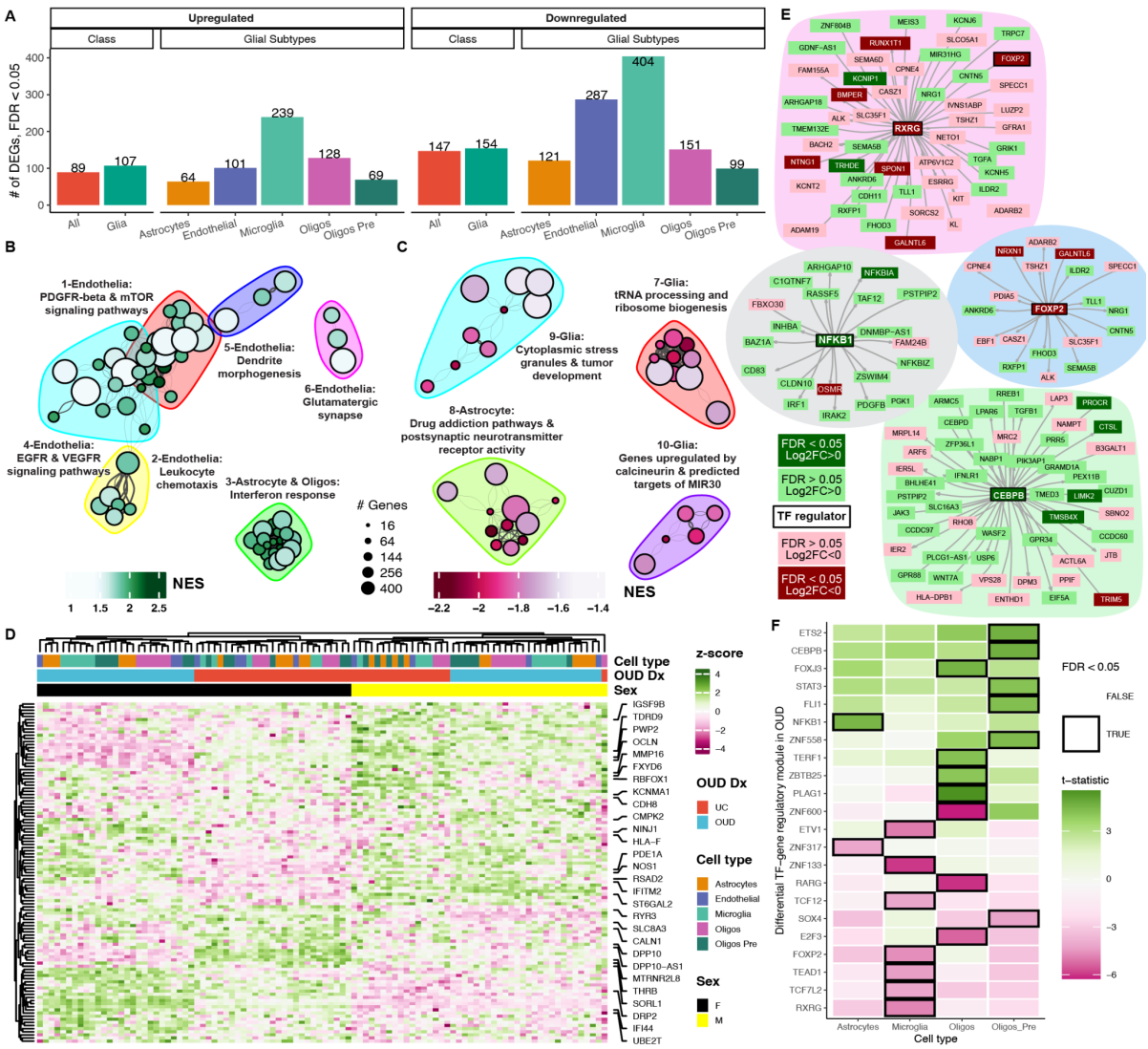


Figure 21. Transcriptional alterations in specific glial cell types in human striatum associated with OUD

(A) Barplot of differentially expressed genes (DEGS) either up- or down-regulated in striatal glia at a false discovery rate less than 0.05 (FDR < 0.05).

(B-C) Network plot of significant clusters and enriched pathways, where each point is a significantly enriched pathway in a glial cell type and lines represent the proportion of shared genes between two pathways. Each point is colored by the normalized enrichment score (NES), indicating whether enriched in up- or down-regulated genes. The color outlines represent the unique clusters of interconnected pathways by shared genes and the nearby text labels the unique cluster number and briefly summarizes which cell types and pathways are represented.

(D) Expression heatmap of z-normalized pseudobulk gene expression where the rows are DEGs and the columns are the biological replicates labeled by cell type, diagnosis, and sex of each individual. Genes that are represented by the enriched pathways in (B-C) are labeled along the right margins.

(E) Cluster maps of select transcription factor gene regulatory networks each transcription factor or gene is colored by the direction and significance of being differentially expressed in glial cell types in OUD.

(F) A heatmap plotting the t-statistic of differentially activated transcription factor gene networks across glial cell types in OUD.

4.3.8 Sex-specific transcriptional alterations in striatal cell types associated with OUD

Rates of OUD and responses to opioids are dependent on sex[390]. Between unaffected and OUD subjects, we found a significant effect of sex on DEGs in both neuronal and glial cell types. Overall, the impact of sex on DEGs in OUD was magnified in neurons (**Figure** 18d) relative to glia (**Figure** 21d). We therefore conducted secondary analyses to identify female-biased and male-biased transcriptional alterations in neurons and glia associated with OUD (**Appendix D Table S12**). Pathways enriched primarily in females with OUD were associated with the upregulation of DNA repair processes in MSNs, upregulation of interferon response in glia, and upregulation of synaptic-related function in astrocytes (Figure 22a, **Appendix D Table S13**, **Appendix D Table S15**). We also identified the FK506 binding protein 5, *FKBP5*, as a potential factor in the role of sex in opioid use. FKBP5[391] in the brain has been proposed as a key factor mediating the relationship between sex and OUD[392–395]. *FKBP5* was significantly upregulated in astrocytes, oligodendrocytes, and oligodendrocyte precursor cells of female OUD subjects (Figure 22b; $\log2FC_{F\ glia}>1.30$, $FDR_{F\ glia} < 0.0165$, $\log2FC_{M}= -1.62$, $FDR_{M\ interneurons} < 0.040$). Prior studies in rodents identified sex differences in *Fkbp5* expression

in dorsal striatum in response to opioids[396]. We investigated whether transcriptional alterations by opioids in rodents were enriched in human dorsal striatum by sex, cell type, and brain region. Indeed, rodent gene sets were significantly enriched in DEGs from glia and oligodendrocytes of female OUD subjects ($FDR_{F\ glia} < 0.027$, **Appendix D Table S16**). Cross-species enriched DEGs were primarily associated with the downregulation of synaptic regulatory proteins, particularly in astrocytes (Figure 22a, cluster 6). The genes encoding synaptic regulatory proteins spanned both pre- and postsynaptic compartments, consistent with the role of astrocytes in synaptic modulation and the tripartite synapse (Figure 22c, **Appendix D Table S14**, **Appendix D Table S15**). In contrast, male-biased transcriptional alterations in OUD were mostly found in microglia, MSNs, and interneurons (Figure 22d, clusters 12, 15, 16). Male-biased pathways consisted of respiratory electron transport and related mitochondrial complexes (Figure 22e). Collectively, our evidence of sex differences in cell type-specific molecular signaling suggests an augmented, magnified glial cell response in females compared to males in OUD.

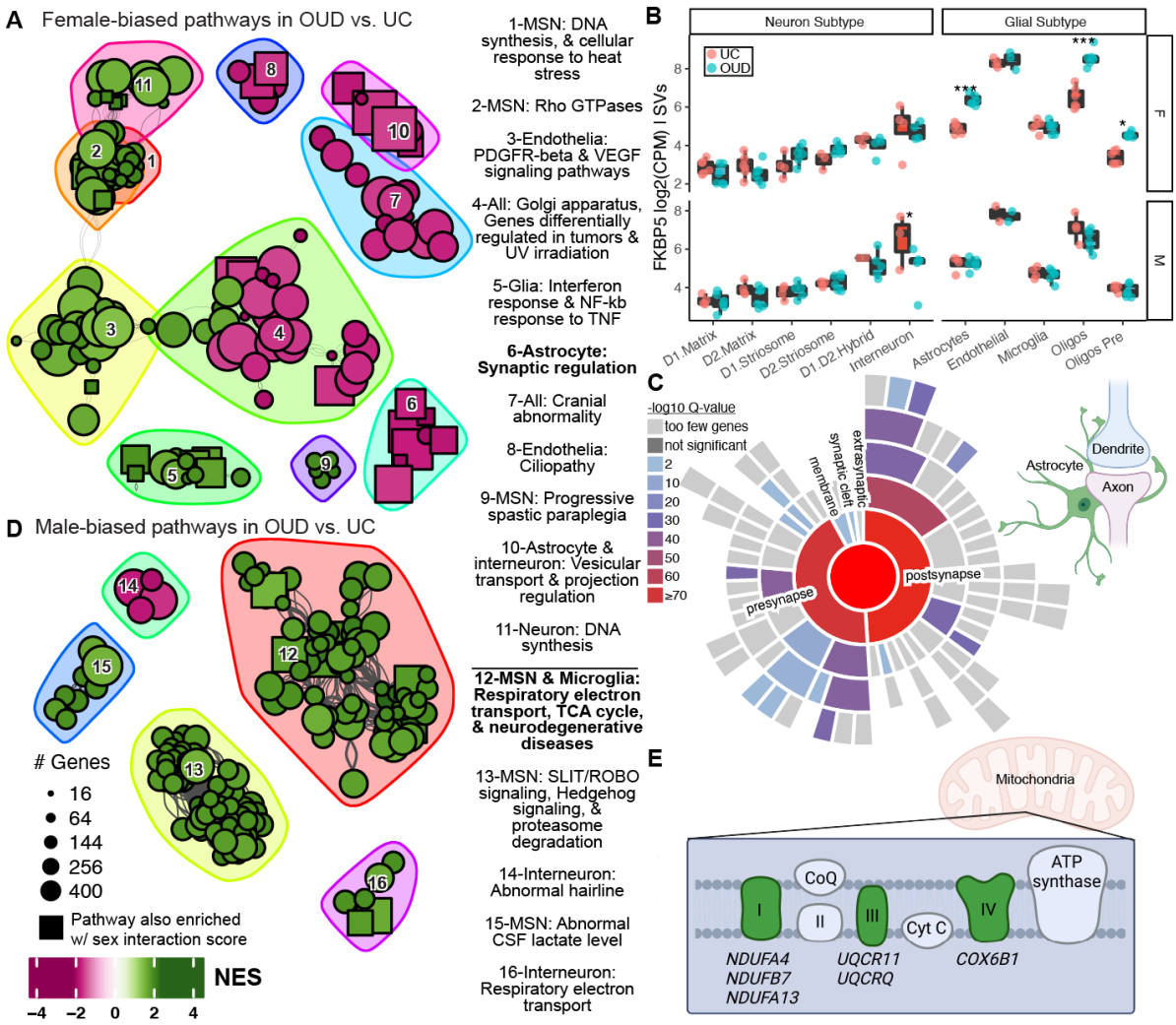


Figure 22. Sex-biased transcriptional alterations in striatal cell types associated with OUD

(A) Network plot of significant clusters and enriched pathways. Here, enriched pathways include only DEGs calculated within female subjects if less significant than those calculated within male subjects. Pathways with a square point are also enriched using an alternate calculation of a sex-interaction score. Clusters are labeled and the corresponding summaries listed along the right side of the cluster plot.

(B) Boxplots showing sex- and cell type-specificity of differential expression of FKBP5. The boxplot labels the log2(counts per million) normalized gene expression values regressing out covariates and surrogate variables. Each point is a biological replicate colored by diagnosis (PDx within Sex < 0.05; ** PDx within Sex < 0.01; *** PDx within Sex < 0.001).*

(C) A sunburst plot representing curated synaptic gene sets from SynGO from genes enriched in cluster 6, a female-biased set of pathways enriched in astrocytes.

(D) A network plot as in (A); however, displaying the male-biased pathways.

(E) Diagram showing the male-biased DEGs that are components of the electron transport gene.

4.4 Discussion

Using single nuclei transcriptomics technologies, we identified both canonical neuronal and glial cell types in human striatum (*e.g.,* D1 and D2 MSNs), along with several of the less abundant cell types, including subclasses of striatal interneurons and D1/D2 hybrid neurons. We also characterized the expression of specific opioid receptor and endogenous ligand subclasses across cell types, highlighting both expression patterns unique to human striatum and those patterns consistent with non-human primate and rodent findings. In addition, we investigated the impact of opioid addiction on transcriptional patterns among specific striatal cell types between unaffected and OUD subjects. Our single nuclei findings provide further insights into the specific cell types that may be impacted in OUD, while highlighting several putative mechanisms in human striatal subtypes of neurons and glia involved in opioid addiction. A common theme that emerged from our analyses was the involvement of processes related to neuroinflammation and neurodegeneration in OUD, including a broad interferon response among glial cells and elevated DNA damage in neurons (**Figure 23**). DNA damage markers were identified in both the striatum of humans and monkeys, suggesting chronic opioid use associated with OUD leads to augmented DNA modifications (*e.g.,* double-stranded breaks, chromatin accessibility). Elevated markers for neuroinflammation, the involvement of microglia-dependent signaling in striatum, and alterations in synaptic signaling in OUD are

consistent with previous bulk transcriptomics findings from postmortem brains of subjects with OUD.

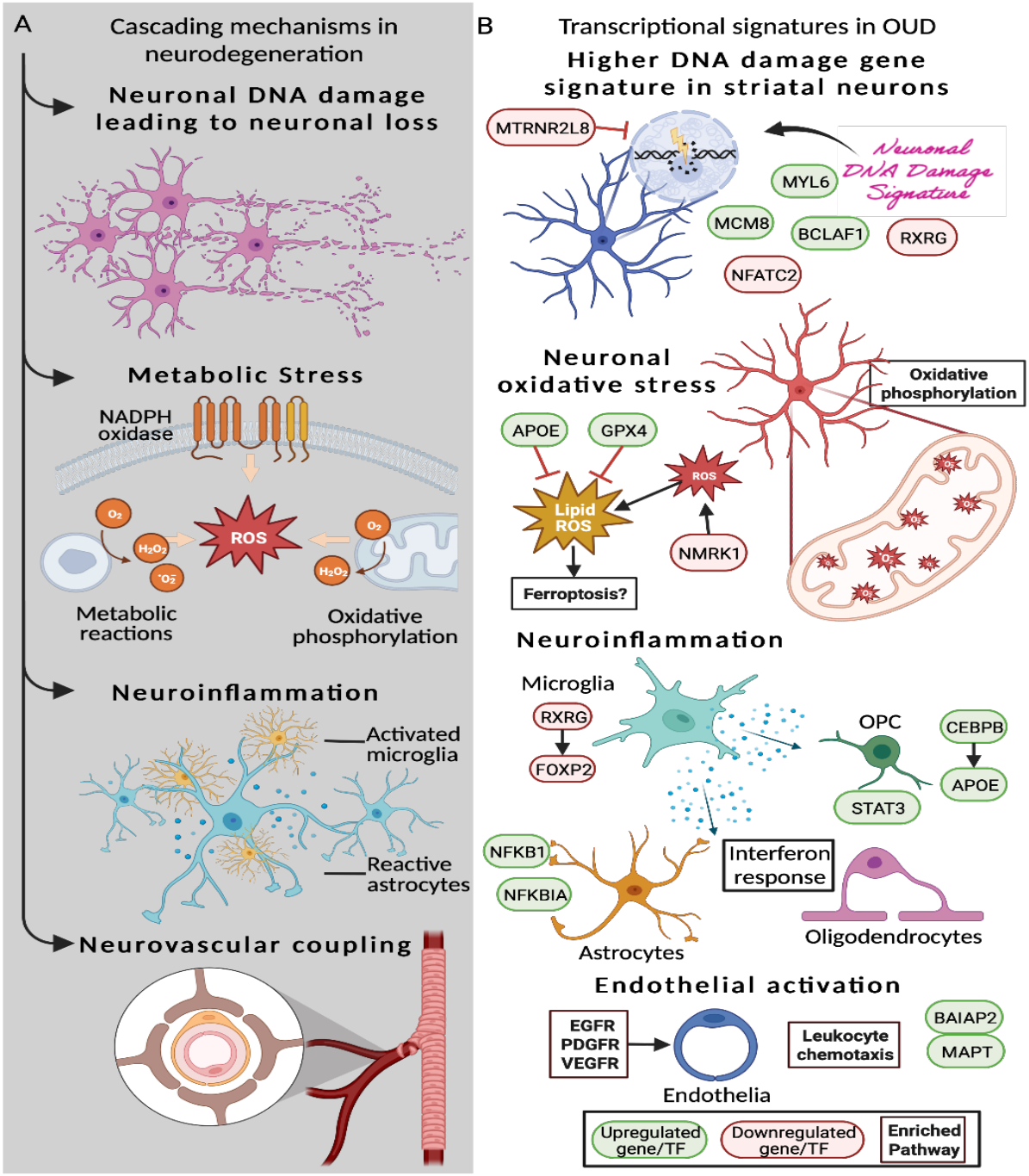


Figure 23. Parallel cellular and molecular pathways in neurodegeneration and opioid use disorder

(A) Molecular cascade of mechanisms that are shared across neurodegenerative processes. Figure created in BioRender, an adaptation from ref126.

(B) A summary of cell type-specific transcriptional alterations in the human postmortem dorsal striatum that parallel the cascading mechanisms in neurodegeneration.

Many of the genes and pathways enriched in specific striatal cell types of OUD subjects have been implicated in neurodegenerative disorders (Figure 23a). DNA damage, in particular, has been broadly studied in Alzheimer's disease[318], Parkinson's disease[397], and ALS[398]. Integrity of DNA is critical for neural cells to prevent insertions, deletions, or mutations that can lead to aberrant synaptic plasticity and/or neurodegeneration. Double-stranded breaks and subsequent DNA repair may allow neurons to rapidly respond at the transcriptional level to changes in neuronal activity[310,311]. Opioids likely lead to persistent changes in neuronal activity and synaptic plasticity within striatum. Prolonged or repeated alterations in neuronal activity combined with dysfunction within DNA repair pathways hampered by persistent cell stress (*e.g.,* oxidative stress) may lead to the accumulation of DNA damage. Neurons may be particularly vulnerable to DNA damage[311,366]. Our results suggest elevated DNA damage in neurons of OUD subjects, particularly within striatal interneurons (Figure 23b). Several subclasses of striatal interneurons are found in striatum[399], possessing an array of transcriptional and physiological properties. We separately clustered TH+, CCK+, PTHLH+, and SST+ interneurons in human striatum. The functionality of interneurons in the striatum is diverse, with several subclasses modulating the excitability and inhibition of MSNs in response to various substances[400], including opioids[306,312,401]. Although we were unable to reference the overall elevation of DNA damage signature to a specific subset of interneurons, our findings suggest striatal interneurons may be particularly vulnerable to a loss of DNA integrity and repair mechanisms in response to changes in neuronal activity by opioids and related perturbations.

The DNA damage response in neurons can lead to a proinflammatory response in microglia and other glial cells[402,403]. In OUD, there were a relatively large number of genes that were upregulated specifically in microglia, enriched for genes involved in neuroinflammation. Microglial activation may propagate to other neural cell types, evident by an increase in interferon responses[404]. Indeed, we found significantly upregulated interferon response in both neuronal and glial cells of OUD subjects. Interferon responses may modulate opioid actions in the brain and could be involved in the emergence of symptoms of opioid withdrawal[405].

Other pathways related to neurodegenerative disorders include oxidative stress, mitochondrial respiration, and neuroprotection. The coordinated alterations in the expression of genes involved in these pathways may reflect compensation to changes in neural activity impacted by opioid use. For example, glutamatergic hyperactivity at striatal MSNs[362–364] may cascade towards excitotoxicity and elevated DNA damage[406–408], both of which have been recently associated with OUD[408–410]. GRM5 was specifically downregulated in astrocytes in OUD. GRM5 is typically transiently expressed in astrocytes to detect and respond to extracellular glutamate, suggesting downregulation in OUD may be due to changes in glutamatergic activity in striatal neurons[411–413]. Alterations in other glutamatergic and GABAergic transcripts were also found in specific cell types in OUD (*e.g.,* downregulation of *GRIA1* in astrocytes and upregulation of *GABRG2* in microglia). In parallel, enrichment of neurodegenerative and neuronal activity markers in certain striatal cells in OUD may be a result of high energetic demand on MSNs to maintain a hyperpolarized state, a possible mechanism of vulnerability of striatal MSNs proposed in Huntington's disease[414]. We also found OUD-associated changes in neuronal energetics with the downregulation of nicotinamide riboside kinase 1, *NMRK1*, and the basic helix-loop-helix family member e40, *BHLHE40* (DEC1), both of which are involved in regulating cellular metabolism, oxidative stress, and

inflammation[415,416]. Together, the broad set of changes in genes and pathways associated with OUD were associated with various processes involved in neurodegeneration.

Sex differences in the vulnerability to substance use, development of substance use disorders, and treatment outcomes, along with the biological response to drugs are evident from both preclinical and clinical studies[417]. Our findings suggest sex-specific alterations in the transcriptional response to opioids in specific striatal cell types. Based on the transcriptional patterns, we found a more pronounced inflammatory response associated with striatal microglia in females compared to males with OUD, suggesting microglia in the human brain also exhibit sex-specific responses to stress and substances[418]. In addition, we identified the upregulation of *FKBP5* as a potential female-specific factor in several glial subtypes in OUD. *FKBP5* acts as a co-chaperone of the glucocorticoid receptor activated in response to stress, with major implications in the pathology of several psychiatric disorders and the impact of stress on substance withdrawal, craving, and relapse[419]. Further, the transcriptional alterations within microglia, oligodendrocytes, and astrocytes in female OUD subjects were significantly enriched for *FKBP5* target genes identified in rodents administered opioids. Genes related to synaptic functions were also enriched in astrocytes of females, while enrichment was mostly in MSNs and interneurons in male OUD subjects, suggesting sex-specific alterations of striatal neuronal and glial cell signaling in opioid addiction.

Several limitations of our resource stem from challenges of profiling single cell transcriptomes in postmortem human brains. While the nuclear transcriptome correlates on whole with the transcriptomes of other subcellular compartments, there are notable differences in mRNA trafficking within certain neural cell types and synapses. Furthermore, the nuclear transcriptome does not capture post-transcriptional regulation. Further, the sex-biased differential expression in OUD is likely influenced by cultural and behavioral factors that are confounded with biological sex. Lastly, while this resource will be useful to identify the

consequences of OUD in the postmortem brain transcriptome, integration of our findings with other types of approaches from genetics to functional genomics in clinical cohorts to animal models of opioid addiction could contribute to a greater understanding of opioid addiction and aid in the realization of new treatment strategies.

4.5 Methods

4.5.1 Human Subjects

Postmortem human brain samples were obtained during autopsies conducted by the Allegheny County Office of the Medical Examiner (Pittsburgh, PA). Consent was obtained from next-of-kin and procedures were approved by the University of Pittsburgh's committee for Oversight of Research and Clinical Training Involving Decedents and Institutional Review Board for Biomedical Research. Each subject meeting diagnostic criteria for OUD at the time of death (n = 6) was matched with an unaffected comparison subject (n = 6) for sex and as closely as possible for age and PMI (see **Appendix D Table S4**). The duration of illness for each OUD subject was at least four years prior to death. For all subjects, the caudate and putamen were identified on fresh-frozen coronal tissue blocks using anatomical landmarks and tissue was collected via cryostat, using an approach that minimizes contamination from white matter and other striatal subregions and ensures RNA preservation.

4.5.2 Isolation of nuclei from human postmortem brain tissue and library preparation:

Nuclei were isolated from 24 biospecimens of frozen human postmortem brain tissue (12 subjects (Unaffected/OUD) x 2 brain regions = 24 samples). Samples weighing 10-15mg were homogenized using ~10 strokes per glass pestle in 7mL glass douncers with 5mL nuclei isolation medium containing DAPI . Homogenate was filtered using a 40um mesh strainer

(FisherScientific #48680). Nuclei were sorted for DAPI fluorescence using a BD FACS Aria at the Boston University Flow Cytometry Core. Approximately 100,000 nuclei were sorted into 7ul of 0.04% bovine serum albumin (MilliporeSigma #126615) in phosphate buffered saline (ThermoFisher #10010031). Nuclei were counted using a hemocytometer and assessed for concentration and debris. 7,000 nuclei were targeted per sample except for one sample with lower concentration where 5,000 nuclei were targeted. The 10x Chromium process was performed and next generation sequencing libraries were prepared using the 10x genomics single cell 3' gene expression dual index kit.

Libraries were sequenced at the Boston University Single Cell Sequencing Core. The pool of snRNA-seq libraries were sequenced on 7 Next-seq P3 flow cells with an intermediate re-pooling scheme to optimize for 50-80% sequencing saturation, > 8,000 average UMI per cell. Between sequencing runs, we preliminarily aligned the sequencing reads as outlined below to assess quality per sample and estimate the number of viable nuclei and sample complexity. We identified two samples, C-13291 and C-612, to have low QC metrics due to wetting failures, mean UMI per cell <1,000 and estimated # cells >50,000. These samples were excluded from subsequent re-pooling and further analyses (**Appendix D Supplemental Table 1.3** tab STARsolo QC).

4.5.3 Nonhuman Primate Subjects

Eight adult rhesus monkeys (*macaca mulatta*) weighing between 6.0 and 13.0 kg served as subjects in the present study. Subjects lived individually in stainless-steel enclosures under a 12-hr light-dark cycle with side and front visual access to other conspecifics. All subjects had continuous access to water and were fed a diet of High Protein Monkey chow (Purina Mills International, Brentwood, MO), fresh fruit, and vegetables. Environmental enrichment (mirrors, toys, foraging boards, music, etc.) was also provided daily. Subject health and well-

being were monitored daily by trained technical and veterinary staff. Animal husbandry and research was conducted in accordance with the guidelines provided by the Institute of Laboratory Animal Resources[422] as adopted and promulgated by the U.S. National Institutes of Health. The facility is licensed by the U.S. Department of Agriculture and all experimental protocols were approved by the Institutional Animal Care and Use Committee at McLean Hospital.

4.5.4 Chronic Morphine Dosing in Nonhuman Primate Subjects

Four subjects (3 male, 1 female; mean age: 12 years, range: 9-18 years) received daily morphine treatment for 5 months (see below); four additional subjects, matched for age, weight, and sex, served as experimental controls (3 male, 1 female; mean age: 14 years, range: 11-21 years). Control subjects had a history of nicotine or cocaine exposure but were drug free for ~1 year prior to tissue collection.

Each morphine-treated subject was trained to come to the front of the enclosure for twice daily intramuscular (IM) injections (<0.5 cc) of morphine sulfate (NIDA Drug Supply Program) dissolved in 0.9% saline. Injections were administered at 09:30 and 17:00 hours. A gradual dosing escalation method (0.5 log unit increase every 3 days) was used to achieve the terminal dosage of 9 mg/kg/day (i.e., 4.5 mg/kg, BID). This dosage was selected to produce moderate opioid dependence[420]. All subjects received approximately 1,500 mg of morphine over the 5-6-month period of chronic dosing. On the last day and approximately 3 hrs following the last IM morphine injection, each subject received an IM injection of ketamine (10 mg/kg) followed by 5.0 ml IV of a pentobarbital-based euthanasia solution (Beuthanasia-D).

4.5.5 Rhesus brain tissue preparation

After sacrifice, brains were rapidly dissected into slabs and frozen on metal plates in liquid nitrogen vapor. Time between animal sacrifice and tissue freezing was between 1-2

hours for all animals. Brains were kept stored at -80 C until punching, when they were punched on a microtome with guidance from a macaque anatomist (Dr. S. Haber). Punches were stored at -80 C until the day of nuclear isolation and encapsulation.

4.5.6 Single nucleus RNA sequencing in rhesus brain tissue

Nuclei were isolated as in previous studies[421], with minor modifications. Punches were placed into buffer HB (0.25 M sucrose, 25 mM KCl, 5 mM $MgCl_2$, 20 mM Tricine-KOH pH 7.8, 1 mM DTT, 0.15 mM spermine, 0.5mM spermidine, protease inhibitors) and placed in a Dounce homogenizer for 10 strokes each with loose and tight pestles. A 5% IGEPAL solution was added to a final concentration of 0.3% followed by five additional dounce strokes, then the lysate was filtered through a 40-µm strainer. Nuclei were mixed with an equal volume of 50% iodixanol and then layered on top of an iodixanol gradient of 40% and 30% layers in a 2 mL dolphin microcentrifuge tube. Nuclei were spun by centrifuging at 10,000 x g for 4 min at 4°C and then collected by aspiration at the interface of the 30% and 40% iodixanol layers. Nuclear concentration and prep quality were ascertained by loading on a hemocytometer and were diluted to a concentration of 80-100K and 15% iodixanol with Buffer HB prior to loading on InDrops v3 platform. Single-nuclei suspensions were encapsulated into droplets, lysed, and the RNA within each droplet was reverse-transcribed using a unique nucleotide barcode as described previously[422]. Approximately 6,000 nuclei, in two batches of 3,000 nuclei each were processed per library and sequenced on Illumina NovaSeq S2 chips (at a density of approximately 20,000 reads/nucleus).

4.5.7 Single nuclei RNAseq data processing

We aligned single nuclei RNA-seq (snRNA-seq) reads to the human genome (GRCh38.p13) or rhesus macaque genome (rheMac10) for each output with the turn-key single-cell transcriptomics method STARsolo, which is folds faster than the CellRanger pipeline and

equally accurate (v2.7.9a) [423]. For the macaque samples, we used a set of gene annotations by mapping the human gene annotations to the rheMac10 genome using the `liftoff` tool[424]. These alternate rheMac10 gene annotations are deposited to Carnegie Mellon University's Kilthub repository resource at ref[425].We chose parameters for the STARsolo UMI quantification to closely replicate the 10X Cell-Ranger pipeline v6 and use the filtered genome and gene annotation available from 10X Genomics (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest, Human reference 2020-A). We ran STARsolo to allow for pre-mRNA gene counts as well as exonic counts for nuclear RNA and to separately count introns and exons for RNA velocity analyses, (`--soloFeatures GeneFull Velocyto`). We used the following parameters to correct cell barcodes, de-duplicate transcripts by their unique molecular identifier (UMI), assign UMI counts to genes, and pre-filter cells that are likely empty droplets (`--soloType Droplet --soloCBmatchWLtype 1MM --soloCellFilter EmptyDrops_CR --soloMultiMappers EM --soloUMIdedup 1MM_CR`). We pre-process the UMI gene x cell count matrix to reduce inherent biases in the technology. We identified likely ambient RNA contamination with SoupX [426,427], empty droplets with DropletQC [425], doublets with scds [431], and damaged nuclei with miQC [318]. For each of these analyses, each sample (GEM well) was analyzed separate from each other. We ran SoupX as described to estimate the fraction of ambient RNA from both raw and unfiltered UMI count matrices from STARsolo and perform ambient RNA removal aware of the cell clusters in the filtered matrix. For just the SoupX analyses, we clustered the cells with Seurat v4 [428–430] with `FindClusters(algorithm = 2, resolution = 0.5)`. For DropletQC, we used the intronic and exonic UMI counts per cell per gene from STARsolo to get the fraction of intronic UMI per cell (referred to as the nuclear fraction). We identified empty droplets with default DropletQC parameters `(nf_rescue = 0.50, umi_rescue = 1000)`. We identified droplets with scds's

hybrid algorithm using the function `cxds_bcds_hybrid` to estimate doublet scores and called doublets on cells with `scds.hybrid_score` > 1.0. We identified damaged cells with miQC which uses a Bayesian EM algorithm to learn the relationship between mitochondria UMI counts and number of captured genes. We used the posterior probability cutoff of 0.75 to call damaged cells by miQC.

To combine cells together across samples, we normalize the UMI counts with the variance-stabilizing `SCTransform` and `glmGamPoi` on each sample [431] and jointly embed cells across samples with reciprocal PCA integration [432] as outlined in https://satijalab.org/seurat/articles/integration_rpca.html. In this joint embedding, we over-clustered the dataset with `FindClusters(algorithm = 2, resolution = 1)` and removed any cluster with more than 10% of cells flagged by `miQC`, `scds`, or `DropletQC` as low-quality biased-clusters in the data.

4.5.8 Labeling striatum cells with a macaque snRNA reference dataset

We annotated our cells from the human or monkey striatum to a recently published high-resolution snRNA-seq reference dataset of the non-human primate striatum [433] using Seurat v4. We downloaded the monkey snRNA-seq processed, annotated gene UMI counts for all cells and MSNs from GSE167920. *He, Kleyman et al.* had aligned the snRNA-seq reads to the rheMac10 genome using the GRCh38 gene annotation liftover to rheMac10, so the gene-wise labels represent the UMI counts on the rheMac10 genome most orthologous to human. For both full nuclei and MSN subset datasets, we re-processed the macaque cells with `SCTransform` , `glmGamPoi`, and reciprocal PCA with default parameters as above to enable label transfer using the most recent integration algorithms in Seurat.

To transfer cell annotations from the reference macaque striatum dataset to the human or macaque striatum cells, we perform two label transfers at increasing resolutions: one with

all cells and another with just MSNs. As *He, Kleyman et al.* described, the differences between transcriptionally and anatomically distinct MSN subtypes are subtle, so we split the annotations into two steps to optimally annotate the cells. The first label transfers the cell classes (Oligodendrocytes, MSNs, Interneurons, etc.) from the macaque to the human dataset with the Seurat functions `FindTransferAnchors(reduction = 'rpca')` and `TransferData`. Next, we identified cells or cell clusters that were labeled as MSNs and transferred MSN subtype labels (D1.Striosome, D2.Striomsome, etc.) from the macaque to human datasets. We filtered out cells where the cell class or cell subtype labels have max prediction scores less than 0.5 as these tend to represent noisy predictions due to low quality cells from either datasets. We confirmed accurate label transfer at the cell class and cell subtype levels with published marker genes and similar proportions across subjects and samples.

Even with the robust cutoffs that we applied to this dataset to remove likely low quality or doublet cells, we find a residual subset of the data that contain these cells. Upon clustering, doublet cells tend to project into the UMAP space as long streaks between two well-defined cell types. Low-quality cell types would project into the UMAP space as amorphous cell types without clear boundaries. Using these embedding features, we selected these clusters with Seurat's `FindClusters(resolution = 1)` function, confirmed that they have the indicative QC metrics, and removed them from analyses.

4.5.9 Annotating interneurons with mouse marker genes

To demonstrate the high resolution of our datasets, we annotated the striatal interneuron using previously characterized mouse markers of these subtypes [434]. We sub-clustered the interneurons labeled by the macaque dataset and annotated them manually as interneuron subtypes best labeled by the marker genes *TH*, *PTHLH*, *SST*, or *CCK*. The previously published macaque snRNA-seq dataset had too few of these interneurons sampled from N=2 monkeys,

so these subtypes were under-represented to be sub-clustered. In this study, we sampled more broadly from N=12 individuals (human) or N=8 individuals (rhesus macaque). While these interneuron subtypes are clearly distinct in this dataset and in the neural circuits, they still are under-represented to power certain downstream. For these analyses, we analyze these cells together as "`Interneurons`".

4.5.10 Differential gene and cellular state expression analysis in humans

To investigate the gene expression differences in OUD and unaffected individuals, we used the pseudo-bulk aggregation of gene expression profiles. Many have shown that pseudo-bulk based case-control differential expression analyses robustly detect gene-level differences with lower false discovery due to repeated measures from single cells of the same individual[435–437]. The raw UMI counts were added together from the same individual, brain region, and cell type to create the pseudo bulk profiles. We aggregate the interneuron subtypes together as "`Interneurons`". We filtered out pseudo-bulk profiles aggregated from more than 20 cells. We filtered out mitochondrial, ribosomal, and low-expressing genes with less than 5 average UMI counts. We retained 20,203 genes and 210 pseudobulk profiles to apply the `voom-limma` method[388] for differential gene expression analyses and the `sva` method to construct surrogate variables to identify un-modeled sources of transcriptomic variation. These statistical methods together address several challenges in analyzing cell type differential expression across multiple axes of meaningful biological variation: 1) accounting for correlated pseudo-bulk samples shared across individuals with the `duplicateCorrelation(block = Subject.ID)`, 2) estimating quality weights for adjusting for cell type proportions with the function `voomWithQualityWeights()`, and 3) calculating un-modeled variation in high-dimensional single cell data with the `sva()` function. To plot the gene expression profiles, we use the normalized counts per million (CPM) of each pseudobulk profile corrected for the batch

effects and surrogate variables unrelated to the OUD diagnosis using the `cleaningY()` function (https://github.com/LieberInstitute/jaffelab). For heatmap visualizations of the gene expression profiles, we also z-normalized the corrected expression profile of each gene grouped by cell type, since gene expression is highly cell-type specific.

We calculated the differential expression in OUD vs. unaffected comparison individuals for two sets of hypotheses:

1) `What is the differential expression in each cell type averaged across brain regions and sex?`

2) `What is the differential expression within the subset of female or male individuals?`

To achieve these comparisons, we used one linear model with a nested variable capturing the interaction between cell type, OUD diagnosis, Sex, and Region, `Celltype_Dx_Sex_Region`. The nested variable allows for contrasting subsets of the data to calculate differential expression for each family of hypotheses at different cell type resolutions: `~ 0 + Celltype_Dx_Sex_Region + Age + PMI + RIN + GDR + 23 SVs`, (GDR: gene detection rate, SV: surrogate variables). We estimate the effect of OUD vs. unaffected individuals across each annotated cell type, across neuronal subtypes (`Neuron`), across glial types (`Glia`), and across all cell types (`All`). For the first set of hypotheses at the annotated cell type level, we used the simple contrast `CelltypeA_OUD_Sex?_Region? - CelltypeA_Ctl_Sex?_Region?` with samples from both Sex and Regions to obtain the average effect across those variables (`?`, the wildcard placeholder for F, M, Caudate, or Putamen). For the second set hypotheses, we only included individuals within each Sex to obtain the effect of OUD vs. unaffected within each subset.

To augment cell type-level differential analyses, we also performed differential gene expression analysis grouping together neurons, glia, or all cell types. Cellular identity is

hierarchically related, so grouping cell types together aids identification of shared transcriptomic processes in OUD. For the grouped cell types, we used the same linear model adding together the effects of each cell type in the group and weight the differential expression by the relative proportion of each cell type: `ProportionA*(CelltypeA_OUD_Sex?_Region? - CelltypeA_Ctl_Sex?_Region?)`. In this dataset, the average proportions of each cell type in each sample are: `D1-Matrix = 4.4%`, `D2-Matrix = 5.2%`, `D1-Striosome = 1.1%`, `D2-Striosome = 1.2%`, `D1/D2-Hybrid = 0.57%`, `Interneuron = 1.1%`, `Astrocyte = 8.6%`, `Endothelia = 0.83%`, `Microglia = 6.2%`, `Oligodendrocyte = 63%`, and `Oligodendrocyte precursor = 1.1%`. We report the full list of differentially expressed genes for each set of hypotheses in Tables S6 and S12.

We calculated the differential expression in chronic morphine exposure vs. non-morphine treat rhesus macaque subjects. We used one linear model with a nested variable capturing the interaction between cell type, morphine treatment and matching stats. `~ 0 + Celltype_Tx + Pair + SVs`. We estimate the effect of morphine treatment vs. no-morphine individuals across each annotated cell type, across neuronal subtypes (`Neuron`), across glial types (`Glia`), and across all cell types (`All`). For the first set of hypotheses at the annotated cell type level, we used the simple contrast `CelltypeA_morphine - CelltypeA_control`.

4.5.11 Gene set enrichment analyses

We identified pathways that are differentially altered in OUD using gene set enrichment analyses with the molecular signatures database [325,438] and the `fgsea` and `msigdbr` R-packages [438,439]. We included the hallmark gene pathways, curated gene sets from BioCarta, KEGG, Canonical Pathways, Reactome, and WikiPathways, and ontology gene sets. We also

downloaded and included curated pathways from SynGO in enrichment analyses[325]. We report the full list of enriched pathways corrected for multiple hypotheses within each set of hypotheses in **Appendix D Table S7**. We clustered redundant/related pathways using custom R-scripts and the `igraph` R-package creating networks of pathways connected by overlapping genes as previously described [325]. We visualize the clustered pathways using `igraph` functions and report both clustered and singleton pathways alongside each **Figure** in tables S7, S9, and S11.

To intersect the genes found to be differentially expressed in OUD within females and males, we also compared to differentially expressed genes identified by microarray studies of striatum of male mice exposed to drugs[437]. Piechota et al. identified from hierarchical clustering 2 main gene signatures that became pronounced time lapsed since drug exposure, A and B, with subsets of the second B1, B2, and B3. We identified the human orthologs of these gene sets and performed GSEA as above on the differentially expressed genes identified across the full cohort, on the subset of female and male subjects, or the caudate and putamen samples. We report the main findings in the text and the full set of enrichments in **Appendix D Table S16**.

4.5.12 Transcription factor-gene regulatory network analyses

The single nuclei RNA-seq dataset in this study provides a new resource to generate gene pathways. To discover new pathways from our single cell RNA-seq, we applied the pySCENIC Protocol to build transcription factor (TF)- gene regulatory networks [311]. To potentially capture individual-specific TF-gene regulatory relationships, we ran the `GRNBoost2` with multiprocessing to infer TF-gene relationships from the `pySCENIC` package on each individual separately [438,439]. To speed up the `GRNBoost2` run time, and reduce bias by cell type proportion, we downsampled oligodendrocytes (~60% of all cells) to

the next most prevalent cell type, Astrocytes (~8%), reducing the `GRNBoost2` run time by 50%. Furthermore, we subset to the ~20k genes that are expressed and analyzed in the pseudobulk differential expression analysis further reducing GRNBoost2 runtime by 30%. Since `GRNBoost2` is a stochastic method, we ran this step 3-4 times for each individual and selected the TF-gene relationships that were reproducible >80% across runs. Similarly, to aggregate the TF-gene relationships across individuals, we aggregated TF-gene relationships that were detected in 80% of 12 individuals. Across these aggregation steps, we averaged the importance scores for selected TF-gene relationships. We selected likely direct TF-gene relationships by filtering relationships where the gene's promoter contain the TF's binding motifs using the `cisTarget` method implement in `pySCENIC` package using the pre-built TF-motif promoter binding databases for human accessed from https://resources.aertslab.org/cistarget/databases/ in January 2023[325]. We report the list of discovered TF-gene relationships in the **Appendix D Table S17**.

4.5.13 Analyses of cellular activation of gene sets

We calculated the level of activity of each collection of gene sets with the `AUCell` method as described previously[325] on the list of TF-gene regulatory networks or custom gene sets (described later). For the TF-gene relationships identified above, we calculated the differentially active TF-gene modules between unaffected and OUD individuals by creating the pseudobulk average of each TF-module by individual, brain region, and cell type. We performed linear regressions to assess the effect of OUD diagnosis accounting for covariates RIN, Age, Sex, Region, and number of aggregated cells. We corrected multiple testing of the OUD effect on TF-gene module activity across all cell types and TF modules and report the full cell type by sample linear regression are reported in **Appendix D Table S8**. We similarly performed a similar pseudobulk average at the biological sample level aggregating over

individual and brain regions. We visualized the significant TF-gene module activity differences set at FDR < 0.05 in glia and neurons and of select modules with the `igraph` R-package[437].

In addition to the TF-gene relationships, we assessed the state of DNA damage in OUD using two published gene signatures of DNA damage[311]. The neuronal DNA damage signature were collected by Welch et al. from fluorescence activated sorting of mouse brain nuclei gated for `NeuN+`:`gH2AX+` nuclei. Using the human orthologs of these mouse genes, we scored our single nuclei RNA-seq datasets with the `AUCell` method to study the differences in DNA damage gene signatures between unaffected and OUD individuals or no-morphine or chronic morphine treated rhesus macaque individuals. We applied the same pseudobulk average of DNA damage scores and reported the linear regression interaction between OUD and cell type in DNA damage signatures in **Appendix D Table S10**. We similarly apply the same pseudobulk average of DNA damage scores and report the linear regression interaction between chronic morphine exposure and cell type regressing out the match pair in DNA damage signatures.

4.5.14 Statistical Analyses

We correct for multiple hypotheses tested in this study using the false discovery rate (FDR) less than alpha = 0.05. In different expression analyses, we perform FDR correction across all cell types tested. Similarly for gene set enrichment analysis and AUCell pseudobulk analyses, we perform FDR correction across all pathways and differentially expressed genes across cell types. We compute the FDR correction with the `swfdr` R package which increases power by leveraging Storey's q-value and modeling the relationship of null P-values and independent variables such as the average logCPM or the number of genes in a pathway[125,126]. In differential gene expression analyses, we calculate the FDR within each cell type comparison (`adj.P.Val.Within`) and between all cell type comparisons

(adj.P.Val.Between). We report the more conservative correction between cell types in the main text and Figures, adj.P.Val.Between. Measured metrics (gene expression counts, quality control metrics) or derived metrics (AUCell gene activity scores) are repeated measurements of the same cohort of human or rhesus macaque individuals. Metrics that are differentially expressed across multiple cell type conditions are reported and summarized at the largest significant adjusted P-value and lowest effect size magnitude. Genes or metrics that are only significantly differentially expressed across one cell type are reported with exact adjusted P-values and effect sizes rounded to 2 significant digits. Exact P-values and effect sizes are reported in the supplementary data. All statistical analyses performed use two-sided test statistics.

For human samples, linear regressions of differential AUCell modules, or other quality control metrics at the pseudobulk analyses were performed across biospecimens (N=22 biospecimens) controlling for relevant covariates including brain region extracting the interaction between OUD diagnosis with cell type. The general linear model for this linear regression for sample level is ~ Dx + Age + PMI + RIN + Sex and for cell type level is ~ Dx:Celltype + Celltype + Age + PMI + RIN where we extract the interaction term to determine OUD diagnosis-specific changes in a cell type. For rhesus macaque samples, linear regressions of differential expression and differential AUCell modules were performed across biospecimens (N=8 biospecimens) controlling for matched pair grouping variable, which accounts for sex, age, and weight. The general model for this linear regression at the individual level is ~ Tx + Pair and for cell type level is ~ Celltype:Tx + Celltype + Pair.

4.5.15 Data Availability

Single nuclei RNA-seq data processing and downstream analyses of the human dorsal striatum in this paper are collected in the github repository https://github.com/pfenninglab/Logan_Striatum_snRNA-seq. Single nuclei RNA-seq processing and downstream analyses of the rhesus macaque nucleus accumbens in this paper are collected in the github repository https://github.com/pfenninglab/McLean_chronic_opioid_monkey_snRNA-seq. The raw sequencing reads and annotated Seurat objects for both human and rhesus macaque studies are uploaded to GEO under SuperSeries accession number GSE233279. A browsable webportal of the human dorsal striatum single nuclei transcriptomes are on the CZ CellxGene Discovery webportal at https://tinyurl.com/humanStriatum.

4.6 Acknowledgements

Thank you to the staff and technicians who work diligently as part of the Brain Tissue Donation Program at the University of Pittsburgh. Postmortem human brain tissue was provided by the University of Pittsburgh Brain Tissue Donation Program and the National Institutes of Health NeuroBioBank at the University of Pittsburgh. One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community.

4.6.1 Funding

The research reported in this article was supported by: Research was supported by NIH HEAL Initiative under National Heart, Lung, and Blood Institute (R01HL150432, RWL); and National Institute on Drug Abuse (R01DA051390, MLS and RWL; DP1DA046585, ARP; F30DA053020, BNP; R01DA047130, SJK).

4.6.2 Author Contributions

RWL conceptualized and managed the project. RWL and ARP provided supervision for data analyses and interpretation. BNP, ARP, RWL, SJK, and MLS obtained funding for the project. BNP, MHR, JFG, DAL, ZF, MLW, ARP, and RWL designed and implemented the project. MHR, MKF, AET, SJR, and YA conducted human tissue homogenization, single nuclei extraction, and single nuclei library preparations. RJF, SJK, JB, SNH, KMM, and KJR conceptualized and managed the rhesus macaque project, including tissue homogenization, single nuclei extraction, and single nuclei library preparations. BNP completed the computational analyses and BNP, MHR, QS, and RWL provided software, visualization, and validation. BNP, XX, QS, GCT, ARP, and RWL designed and interpreted analyses and outcomes. RWL, JG, DAL, and ARP provided the tissue resources and computational infrastructure for analyses. BNP, MHR, MKF, ARP, and RWL wrote the original manuscript with Figures. All authors participated in reviewing and editing the manuscript for publication.

5.0 Conclusion

BaDoi N. Phan[1,2]

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University, 15213
[2]Medical Scientist Training Program, School of Medicine, University of Pittsburgh, 15213

This chapter was written in part by ChaGPT version May 2023 prompted by ideas from the author and edited to the extent of author's knowledge for accuracy to convey the final message.

In a study aimed at understanding the genetic basis of addiction-related behaviors, researchers integrated neuronal subtype open chromatin regions (OCRs) from human and mouse brain epigenomics. They used convolutional neural network (CNN) models to identify addiction-associated single nucleotide polymorphisms (SNPs) that potentially affect neuronal subtype-specific regulatory elements. The findings suggest that genetic variants associated with nicotine, alcohol, and cannabis use behaviors may impact putative CREs (cis-regulatory elements) in different combinations of excitatory, D1, and D2 neuronal subtypes. This research provides insights into the cell type-specific genetic mechanisms underlying addiction-related traits and demonstrates the utility of cross-species computational frameworks for interpreting genetic risk variants in humans using model organisms.

The study initially identified addiction-associated genetic variants enriched in neuronal subtypes of the prefrontal cortex and striatum, brain regions involved in addiction and reward circuitry. These variants were found in OCRs active in cortical excitatory neurons and striatal D1 and D2 medium spiny neurons (MSNs). The results suggest that genetic variation in these neuronal subtypes may contribute to behaviors observed in addiction, such as impulsivity and enhanced motivation to procure drugs. The study also revealed enrichments of addiction-

associated genetic variants in other brain regions and neuronal subtypes, providing insights into the genetic regulation of nicotine, alcohol, and cannabis use behaviors.

To overcome the lack of high-quality open chromatin measurements in human reference genomes for key cell types involved in addiction, the researchers leveraged mouse cell type open chromatin data and identified orthologous regions in the human genome. This approach allowed them to analyze cell populations in brain regions where primary human open chromatin profiles were lacking. The study demonstrated remarkable concordance in enrichments of cell type OCRs between mouse and human profiles, indicating a conserved regulatory code between species. The findings highlight the suitability of animal models for studying neural circuits and cell-type-specific gene regulation in addiction.

In addition to integrating mouse-human orthologous OCRs, the researchers developed and trained CNN models to classify neuronal subtype membership based on open chromatin regions. This refinement of the analysis provided further validation for enrichments of human and mouse-human orthologous OCRs in specific cell types and brain regions. The approach identified two clusters of cell types and brain regions associated with addiction-related traits: one cluster enriched in cortical excitatory neurons linked to smoking initiation and cannabis use, and another cluster enriched in D1 and D2 MSNs associated with smoking cessation and alcohol consumption. The study also highlighted the role of interneuron populations in addiction neural circuits and demonstrated the complexity of genetic variation and gene regulation in cell type-specific contexts.

To predict the functional impact of genetic variants associated with addiction-related traits, the researchers employed a fine-mapping pipeline and CNN models. They identified a set of candidate causal SNPs that could potentially impact gene regulation in specific brain regions and neuronal subtypes. These candidate SNPs provide possible mechanisms for linking genetic variation with the genes, cell types, and brain regions that influence addiction and

substance use behaviors. The study acknowledged the limitations of relying solely on open chromatin assays and suggested the incorporation of other epigenetic assays for a more comprehensive understanding of gene regulation.

Overall, this study provides insights into the genetic mechanisms underlying addiction-related traits by integrating neuronal subtype OCRs across human and mouse brain epigenomics. The findings highlight the role of specific cell types and brain regions in addiction and suggest a conserved regulatory code between species. The research offers a foundation for future investigations into cell type-specific genetic mechanisms and demonstrates the value of computational frameworks and animal models for understanding complex traits and diseases.

Appendix A Chapter 2 Extended Data Legends

Extended Data Figure 7-1.

TomTom matches with motifs learned by CNN models in each cell type and fold to contribute to a strong positive prediction. Learned important features were interpreted by DeepSHAP and clustered into unique seqlets by TF-Modisco (Methods).

Extended Data Figure 7-2. Addiction-associated genetic variants annotated with cell type and brain region functional markers

*Addiction-associated genetic variants from the main seven GWAS (**Figure** 7) that were scored by CNN models along with computed raw CNN scores, predicted probability active, and ΔSNP probabilities, and tier of predicted candidate causal SNP. Each entry is recorded for a distinct SNP, predicted CNN model, and GWAS trait. Additional columns reporting are annotated by FUMA [72] and CAUSALdb [78]. SNPs are annotated in this study to overlap with human NeuN+ OCRs [35]. A complete legend describing column headers is in the first sheet of the table.*

Appendix B Chapter 3 Supplemental Figures

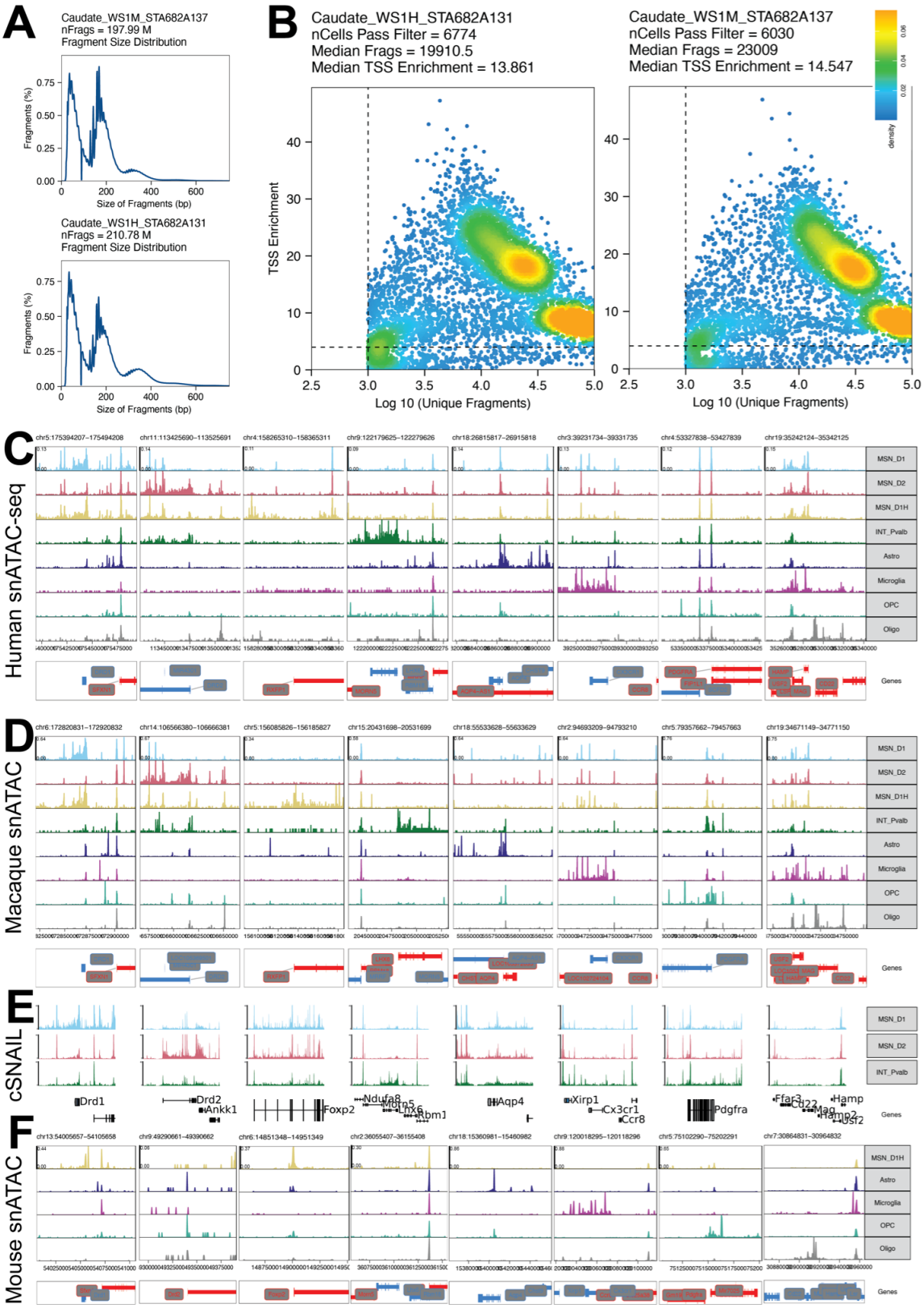


Figure S1. Identification of caudate cell types in human, macaque, and mouse snATAC-seq

Periodicity plots of 2 technical replicates of rhesus macaque caudate snATAC-seq libraries show high quality fragment distribution matching expedited pitch and periodicity of intact chromatin.

Quality control plots of per cell number of fragments and transcription start site enrichment score.

Pseudo-bulk ATAC-seq signal plots generated with ArchR around marker gene loci in 8 caudate cell types from human snATAC-seq support cell type-specific of identified clusters from Figure 1A [88,222]*. Canonical marker genes for each track are in order DRD1, DRD2, RXFP1, LHX6, AQP4, PDGFRA, and MAG to collectively discriminate the 8 caudate cell types, as previously reported* [206]*.*

Pseudo-bulk ATAC-seq signal plots around marker gene loci in rhesus macaque caudate snATAC-seq. Canonical marker genes for each track are the same as (C).

ATAC-seq signal plots around marker gene loci mouse caudoputamen using cSNAIL in D1-, D2-, and Pvalb-cre mouse strains to label MSN_D1, MSN_D2, and INT_Pvalb neuronal nuclei [187]*. Canonical marker genes for each track are in order DRD1, DRD2, FOXP2, LHX6, AQP4, PDGFRA, and MAG to collectively discriminate the 8 caudate cell types.*

Pseudo-bulk ATAC-seq signal plots around marker gene loci mouse caudoputamen snATAC-seq for MSN_SN, and glial cell types. Marker gene tracks are the same as (E). MSN_D1: D1 medium spiny neurons, MSN_D2: D2 medium spiny neurons, PVALB/Pvalb+ interneurons, Astro: astrocytes, Oligo: oligodendrocytes, OPC: oligodendrocyte precursors.

Cell type	Shared motifs identified across species
MSN_D1	AR, CENPB, GATA3, GFI1, HLF, HOXB1, HSF2, HSF4, LBX2, MEF2B, MGA, MYBL2, NFATC1, NKX2-6, NPAS4, PAX6, RHOXF1, SIM1, SIX5, SOX2, STAT1, STAT3, STAT4, STAT6, TBX1, TBX10, TBX15, UBP1, XBP1, ZNF274, ZNF282, ZNF75A, ZNF784
MSN_D2	AR, CEBPA, CEBPB, CEBPD, CENPB, GATA3, GFI1, HLF, HOXB1, HSF2, HSF4, IRX2, IRX5, MGA, MYBL2, NFATC1, NKX2-6, NPAS4, NR1I3, OVOL1, PAX6, SIM1, SIX5, STAT1, STAT3, STAT4, STAT6, T, TBX1, TBX10, TBX15, UBP1, XBP1, ZNF274, ZNF75A, ZNF784
MSN_SN	ALX3, CENPB, DLX3, DLX5, EMX1, EMX2, EVX1, GBX2, GSX1, GSX2, HESX1, HMX1, HOXD1, HOXD3, LBX1, LBX2, MEOX2, MGA, NOTO, OTP, PHOX2B, POU2F2, POU3F4, POU6F2, RAX, RHOXF1, TBX10, TFEB, VAX1, VAX2, VSX1, VSX2, ZNF784
INT_Pvalb	BHLHE22, BHLHE23, HOXB1, NEUROG1, NEUROG2, NEUROG3, POU3F3, TFEB, UBP1, XBP1
Astro	EN1, HNF4A, HSF1, NFKB1, NR1I3, NR2C1, NR5A2, PGR, PRDM4, RARA, RARG, RELA, RXRA, THRA, ZKSCAN3, ZNF75A
Microglia	ELF5, ETV2, HSF1, IRF1, IRF3, IRF6, NR1I3, NR2C1, NR2F6, PRDM1, RARA, RELA, SPIB, SPIC, STAT2, ZNF282
OPC	RXRA, SOX10, ZKSCAN3
Oligo	HNF4A, NEUROG1, NEUROG2, NEUROG3, NR2C1, NR5A2, PGR, RARA, RUNX3, RXRA, SOX10, SOX11, SOX12, SOX13, SOX14, SOX15, SOX17, SOX18, SOX21, SOX3, SOX5, SOX6, SOX7, SOX8, SOX9, SRY, TCF7, THRA, ZKSCAN3

Figure S2. Shared regulatory code in cross-species motif analyses of distal open chromatin regions

MEME-chip significantly enriched transcription factor binding sites identified to be shared in each cell type across human, rhesus macaque, and mouse caudate cell type.

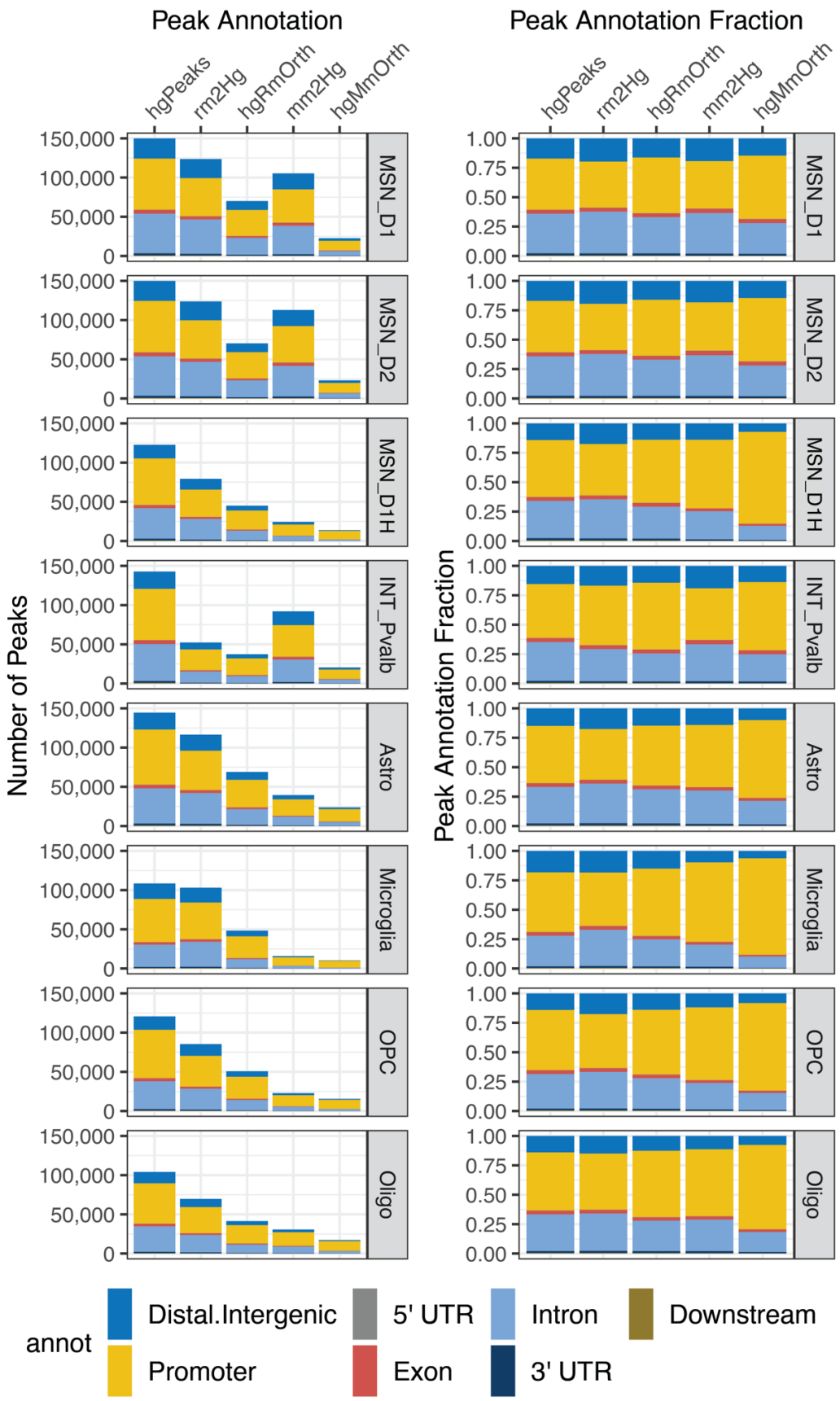


Figure S3. Human cell type OCRs are less mappable across more distantly diverged species

(Left) The number of human cell type OCRs that can be mapped to another species. Each point is a species. The x-axis represents the number of million years (MY) that species diverged from humans. The point color is the membership of the species with a mammalian taxonomic order. (Right) the proportion of all human cell type OCRs that are mappable to distantly related species.

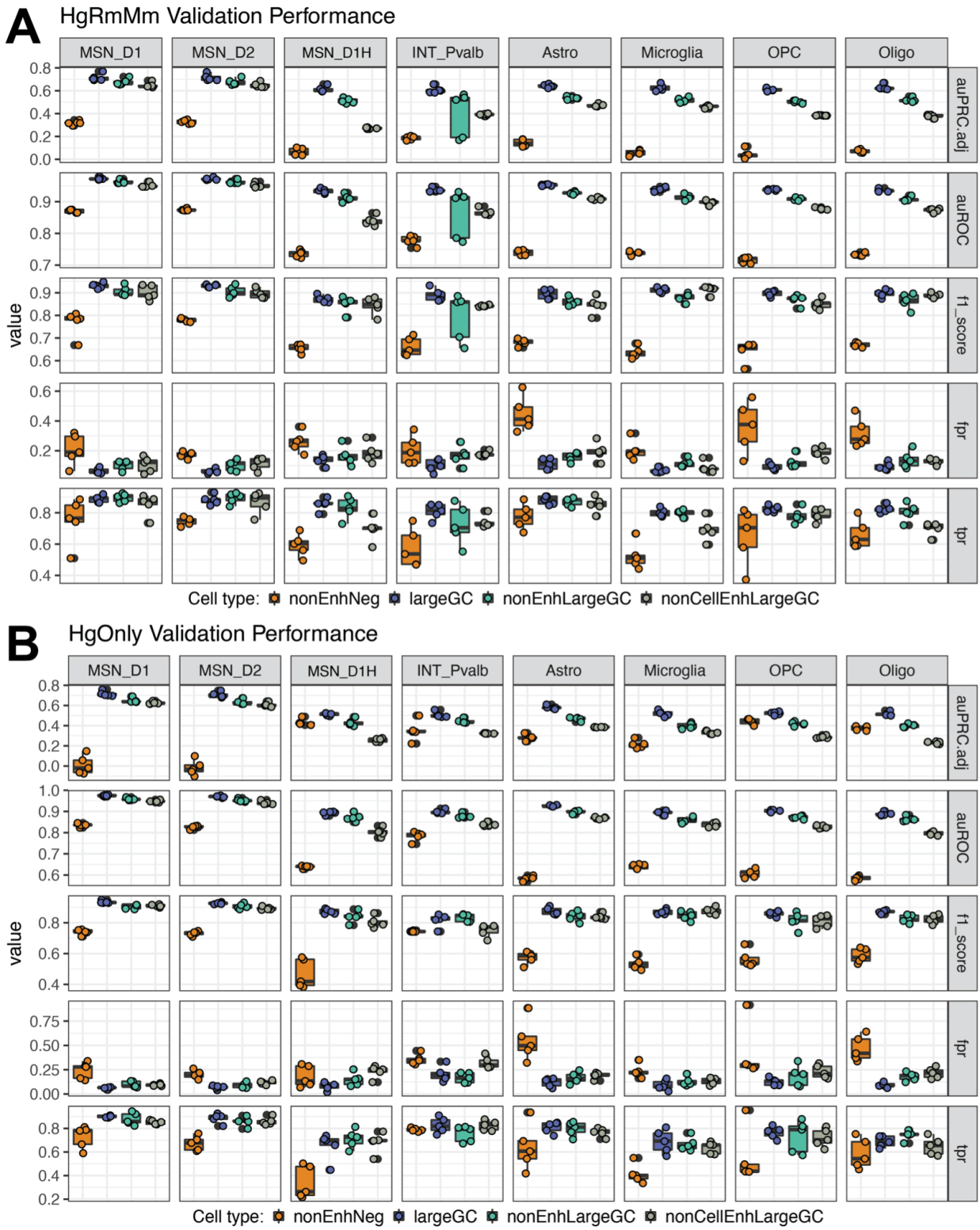


Figure S4. Cell-TACIT caudate cell type models cross-validation performance with single- and multi-species data across various negative sets.

Negative sets evaluation of orthologs of cell type OCRs in other species that are not accessible (nonEnhNeg), large set of GC-matched genomic regions (largeGC), union of the nonEnhNeg

and largeGC (nonEnhLargeGC), union of nonEnhLargeGC, and accessible regions in other cell types (nonCellEnhLargeGC). (A) Overall cross-validation performance metrics of Cell-TACIT models trained with OCRs detected in 3 species across a number of negative sets (HgRmMm). (B) Overall cross-validation performance of models trained with only human sequences (HgOnly). Each point is a different cell type, cross-validation fold, negative set combination. Metrics (details in Methods): area under the Precision-Recall curve adjusted for the proportion of the minority class (auPRC.adj); area under the receiver-operator curve (auROC); false positive rate (fpr); true positive rate (tpr).

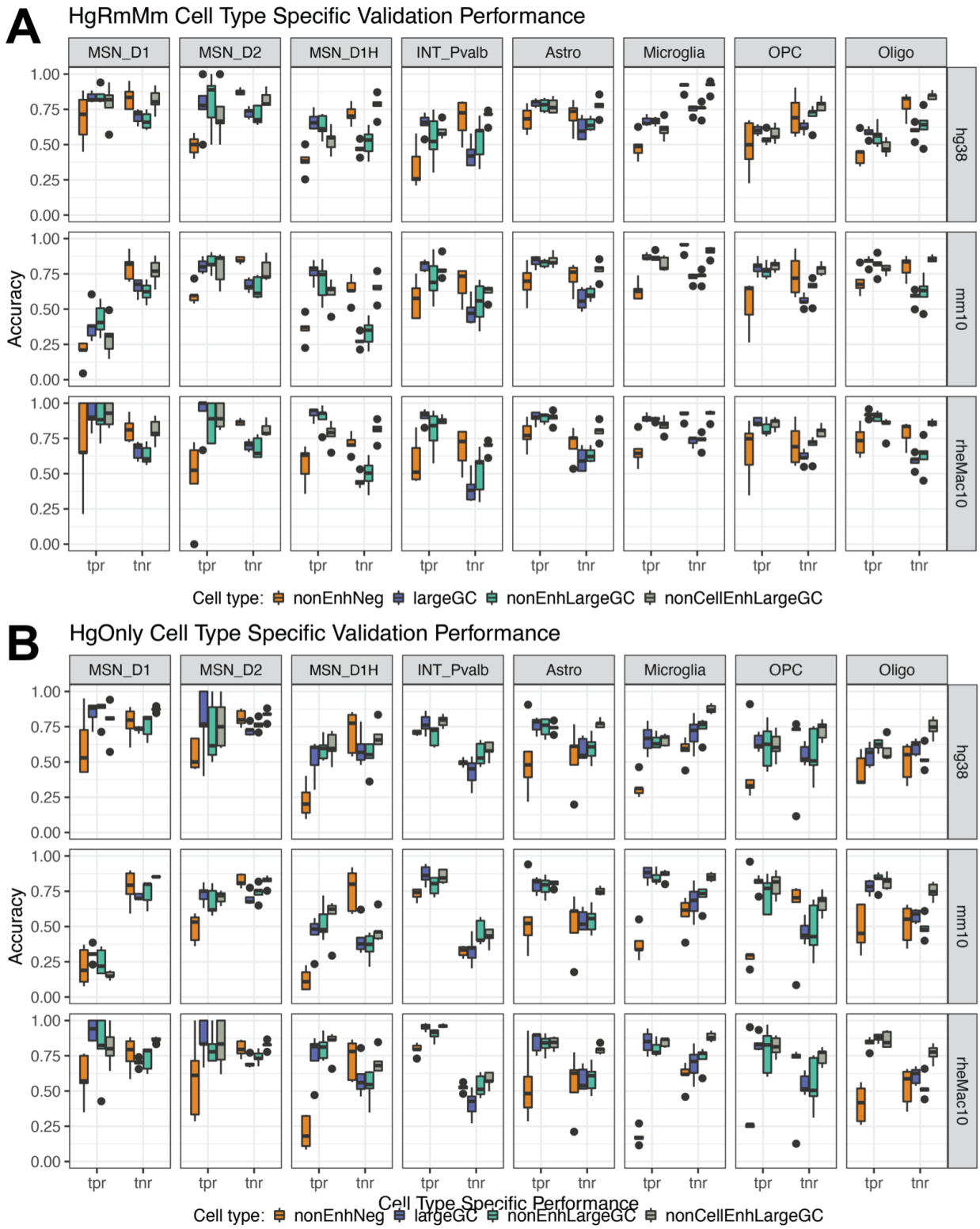


Figure S5. Negative set comparisons across subsets of data designed to evaluate models' ability to predict cell type-specific open chromatin.

Negative sets evaluation of cell type-specific positive and negative cross-validation for Cell-TACIT models trained using negative sets from orthologs of cell type OCRs in other species

that are not accessible (nonEnhNeg), large set of GC-matched genomic regions (largeGC), union of the nonEnhNeg and largeGC (nonEnhLargeGC), union of nonEnhLargeGC and accessible regions in other cell types (nonCellEnhLargeGC).

(A) *Cell type-specific positive and negative cross-validation performance metrics of Cell-TACIT models trained with OCRs detected in 3 species across a number of negative sets (HgRmMm).*

(B) *Cell type-specific positive and negative cross-validation performance metrics of Cell-TACIT models trained with OCRs detected in models trained with only human sequences (HgOnly).*

Each point is a different cell type, cross-validation fold, negative set combination.

Metrics (details in Methods): true positive rate (tpr); true negative rate (tnr).

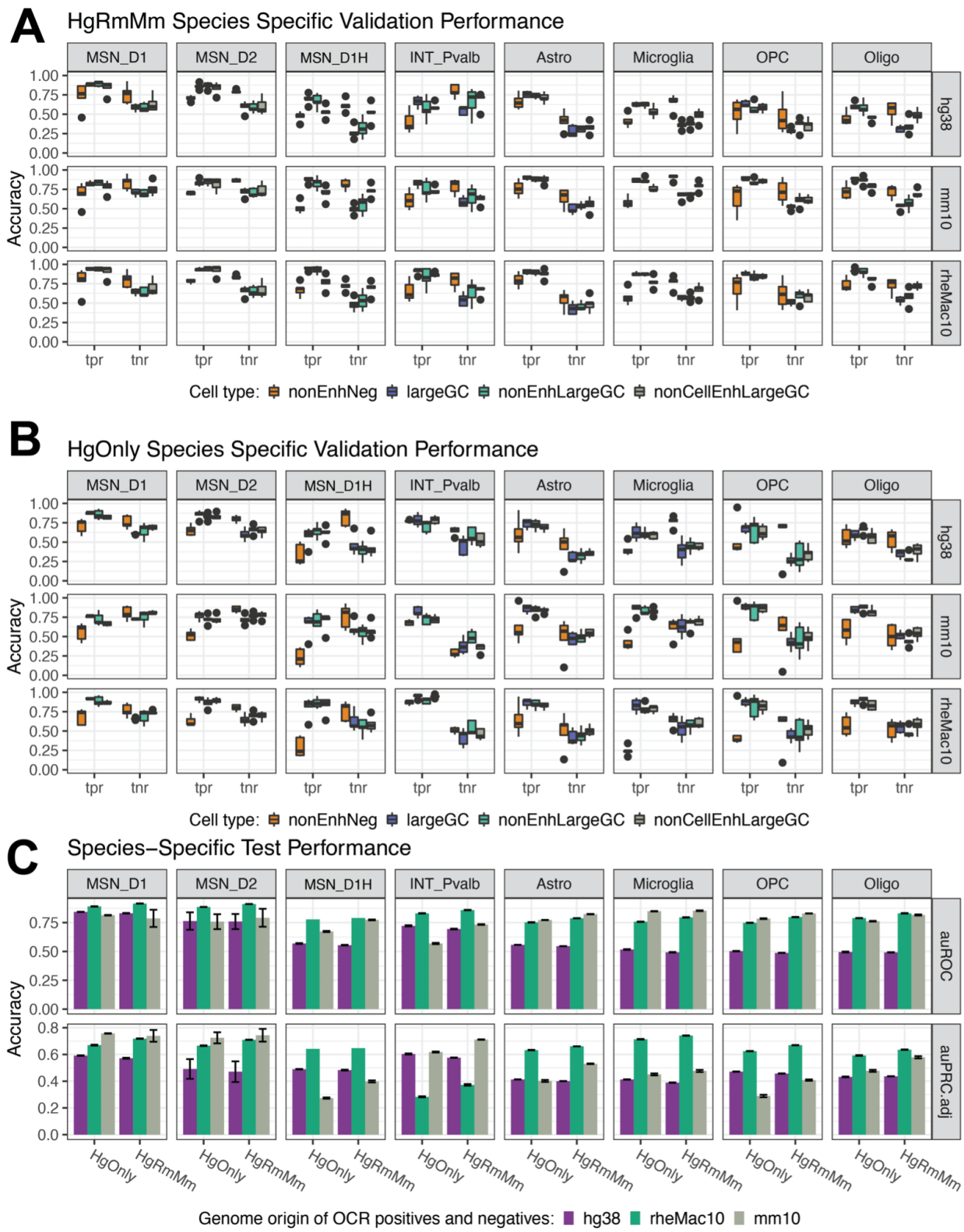


Figure S6. Negative set comparisons across subsets of data designed to evaluate models' ability to predict species-specific open chromatin changes.

(A) *Species-specific positive and negative cross-validation performance metrics of Cell-TACIT models trained with OCRs detected in 3 species across a number of negative sets (HgRmMm).*

(B) *Cell type-specific positive and negative validation set performance metrics of Cell-TACIT models trained with OCRs detected in models trained with only human sequences (HgOnly).* ***(C)*** *Test set performance using the species-specificity positive and negative sets using auPRC.adj and auROC across both HgRmMm and HgOnly models show both models have similar species-specific prediction accuracies. In parts* ***(A)*** *and* ***(B)****, each point is a different cell type, cross-validation fold, negative set combination. Positives come from OCRs in one species whose ortholog is still accessible in another species. Negatives are inaccessible regions whose orthologs are accessible in other species. Metrics (details in Methods): true positive rate (tpr); true negative rate (tnr); area under the precision-recall curve adjusted for the proportion of the minority class (auPRC.adj); area under the receiver-operator curve (auROC).*

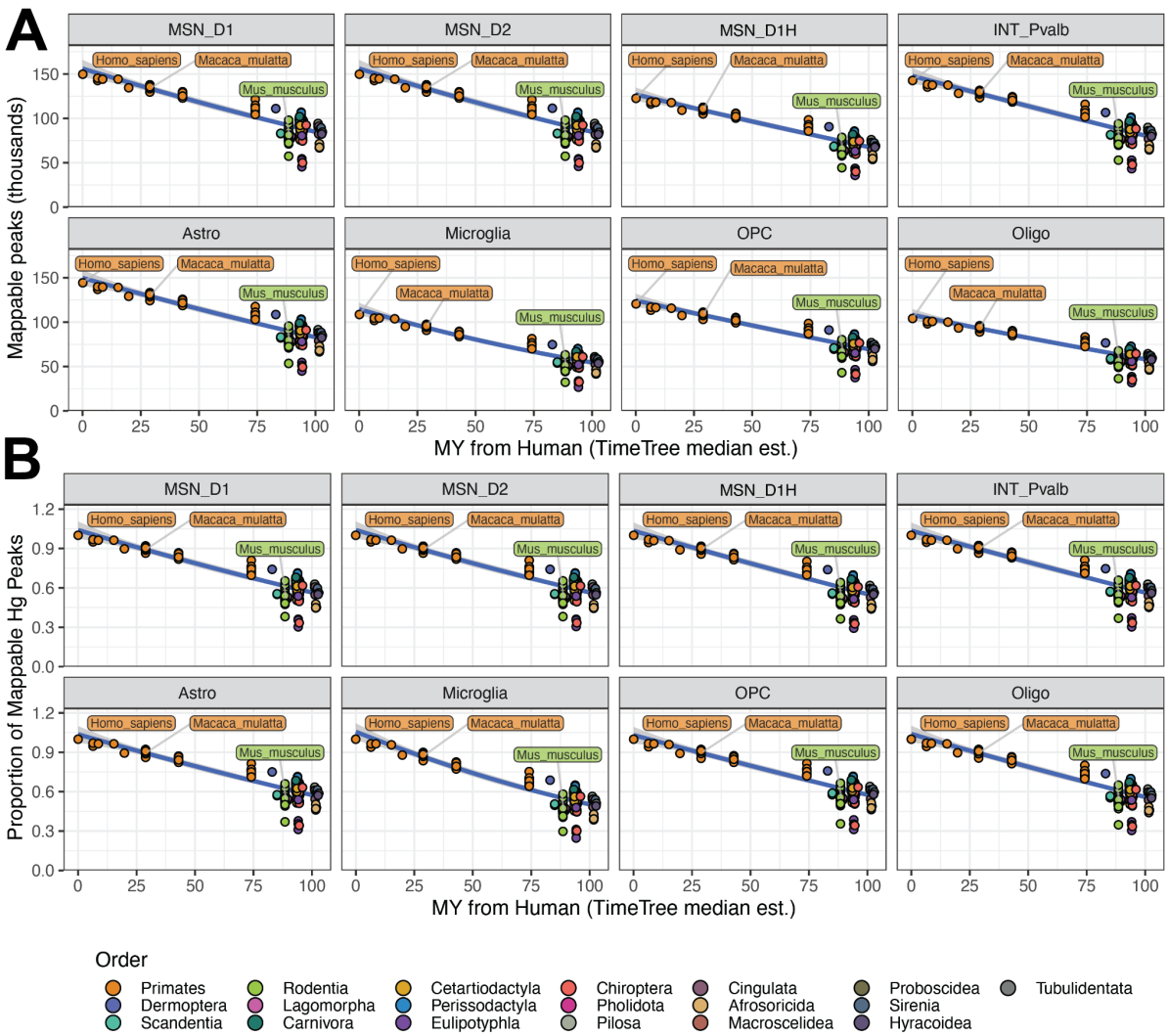


Figure S7. Mappability of macaque and mouse cell type OCRs to the human genome and overlap with human cell type OCRs.

Cell type orthologs (OCRs in a cell type shared across humans and another species) are a subset of peaks mappable from the other species. The relative proportions of annotations of cell type OCRs are generally similar across species and cell types. hgPeaks: human cell type OCRs; rm2hg: macaque peaks mappable to the human genome; hgRmOrth: cell type OCRs accessible in both human and macaque; mm2hg: mouse peaks mappable to the human genome; hgMmOrth: cell type OCRs accessible in both human and mouse. UTR: untranslated region, annot: annotation.

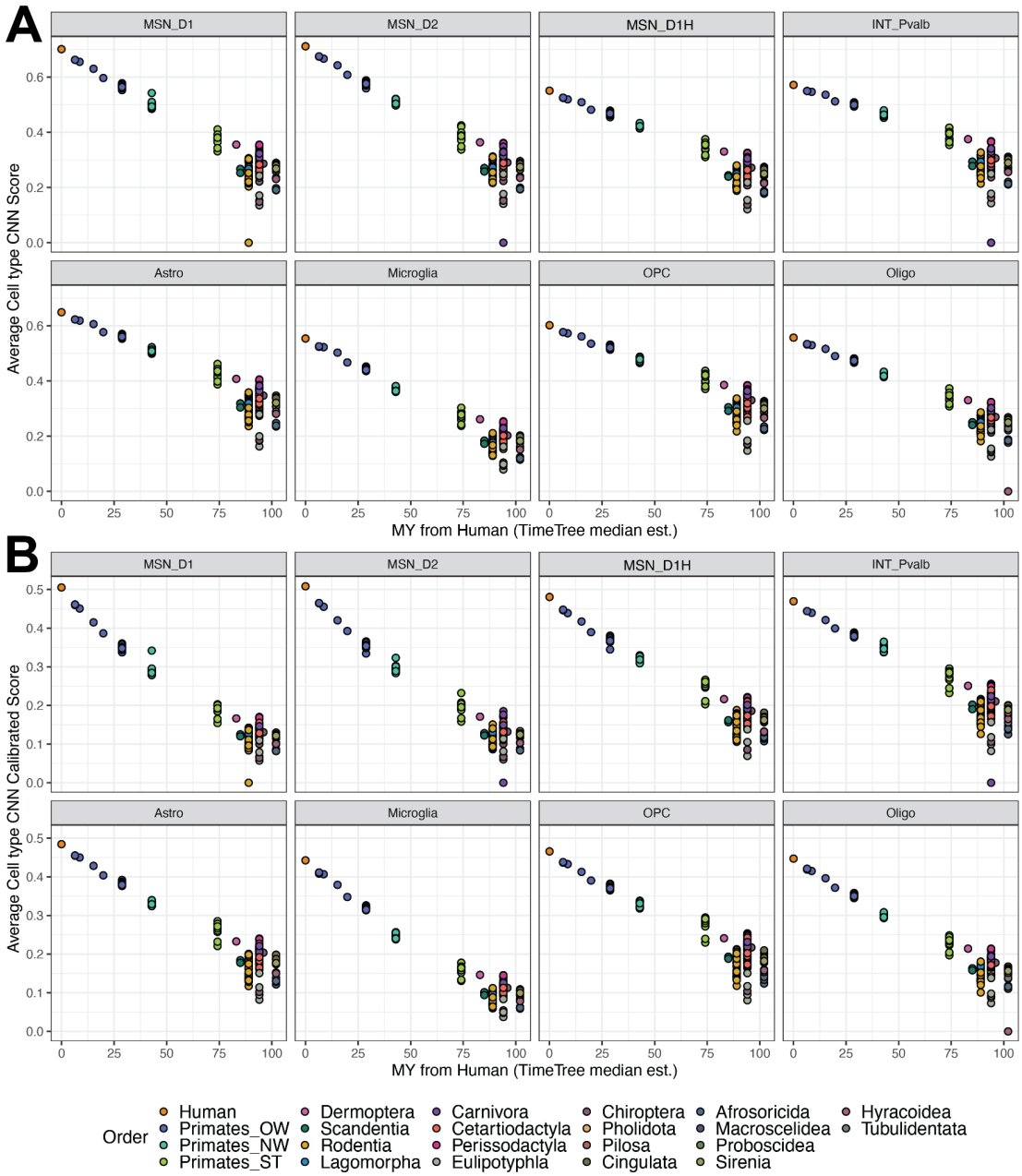


Figure S8. Cell-TACIT predictions of human cell type OCR orthologs in distant species reflect sequence-level drift and loss of regulatory function with species divergence.

(A) *The average Cell-TACIT scores of cell type orthologs of human OCRs in each species versus the million years (MY) that species diverged from humans. Each point is a species and the point color is the membership of the species with a mammalian taxonomic order.* ***(B)*** *the Cell-TACIT scores that are calibrated to the percentage of the model validation positives with a lower score. More negative correlations are seen with these calibrated scores and the MY diverged from humans.*

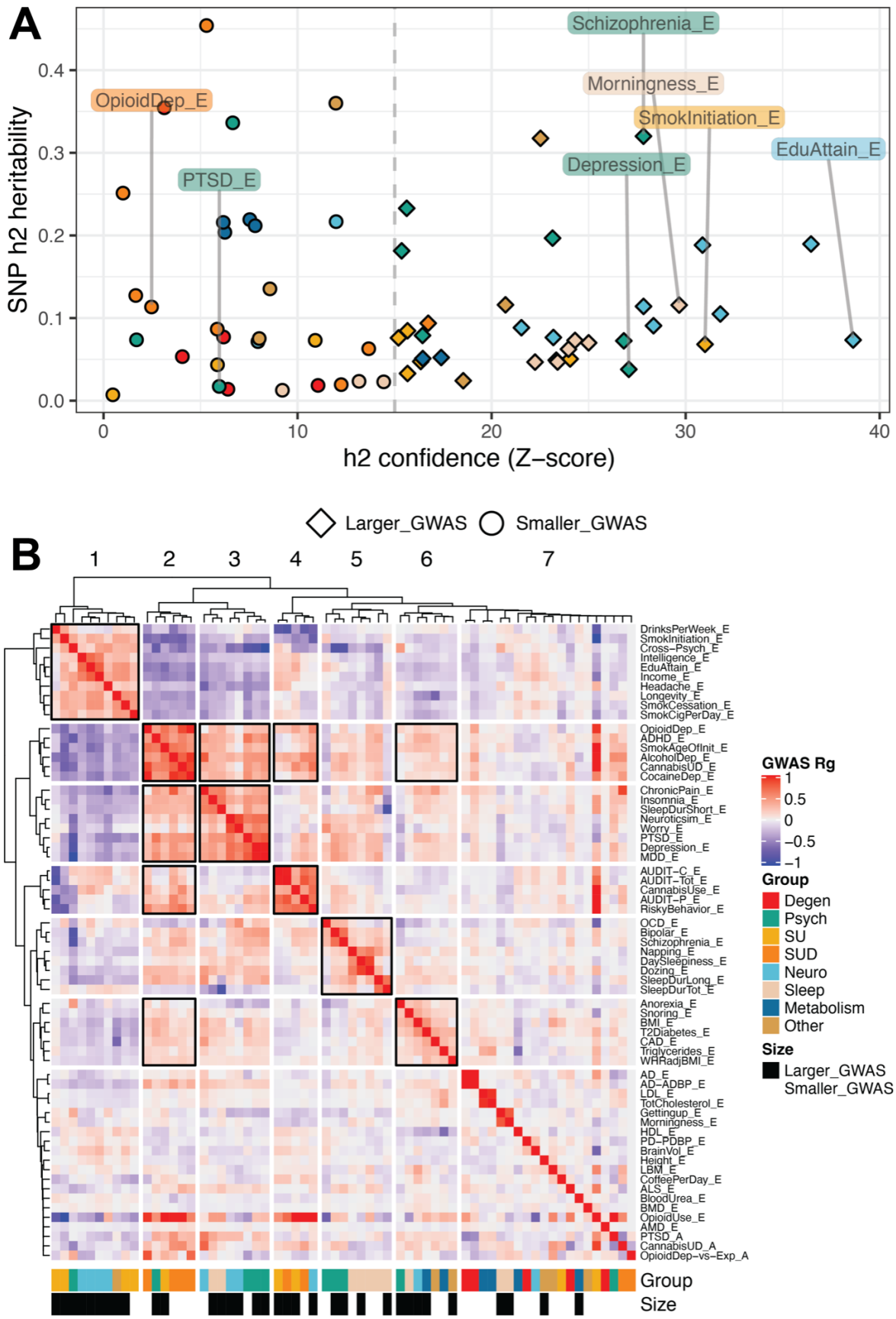

A
SNP h2 heritability
h2 confidence (Z−score)
OpioidDep_E
PTSD_E
Schizophrenia_E
Morningness_E
SmokInitiation_E
Depression_E
EduAttain_E
Larger_GWAS
Smaller_GWAS
B
1
2
3
4
5
6
7
DrinksPerWeek_E
SmokInitiation_E
Cross−Psych_E
Intelligence_E
EduAttain_E
Income_E
Headache_E
Longevity_E
SmokCessation_E
SmokCigPerDay_E
OpioidDep_E
ADHD_E
SmokAgeOfInit_E
AlcoholDep_E
CannabisUD_E
CocaineDep_E
ChronicPain_E
Insomnia_E
SleepDurShort_E
Neuroticsim_E
Worry_E
PTSD_E
Depression_E
MDD_E
AUDIT−C_E
AUDIT−Tot_E
CannabisUse_E
AUDIT−P_E
RiskyBehavior_E
OCD_E
Bipolar_E
Schizophrenia_E
Napping_E
DaySleepiness_E
Dozing_E
SleepDurLong_E
SleepDurTot_E
Anorexia_E
Snoring_E
BMI_E
T2Diabetes_E
CAD_E
Triglycerides_E
WHRadjBMI_E
AD_E
AD−ADBP_E
LDL_E
TotCholesterol_E
Gettingup_E
Morningness_E
HDL_E
PD−PDBP_E
BrainVol_E
Height_E
LBM_E
CoffeePerDay_E
ALS_E
BloodUrea_E
BMD_E
OpioidUse_E
AMD_E
PTSD_A
CannabisUD_A
OpioidDep−vs−Exp_A
Group
Size
GWAS Rg
1
0.5
0
−0.5
−1
Group
Degen
Psych
SU
SUD
Neuro
Sleep
Metabolism
Other
Size
Larger_GWAS
Smaller_GWAS

Figure S9. Shared genetic architecture of neurological and psychiatric traits from GWAS of varying heritability and power.

(A) Scatter plot of LD score regression-estimated SNP heritability (h_{2g}) by the confidence of estimation (h_{2g} z-score) across 64 GWAS. Each study is a point colored by the trait group (legend in panel B). Round points with h_{2g} z-score < 15 indicate smaller, less well-powered GWAS. Diamond points with h_{2g} z-score > 15 indicate larger, more well-powered GWAS.

(B) Genetic correlation heat map clustered with hierarchical clustering into 7 trait groups with shared genetic architectures. GWAS are labeled by the trait and the source population (_E for European ancestry subjects and _A for African ancestry subjects). Black-outlined blocks indicate blocks of correlated traits for which more than 70% of the trait-trait correlations have a positive genetic correlation (r_g) greater than 0.10. Traits are also annotated with group traits and whether the trait had a larger or smaller GWAS from (A).

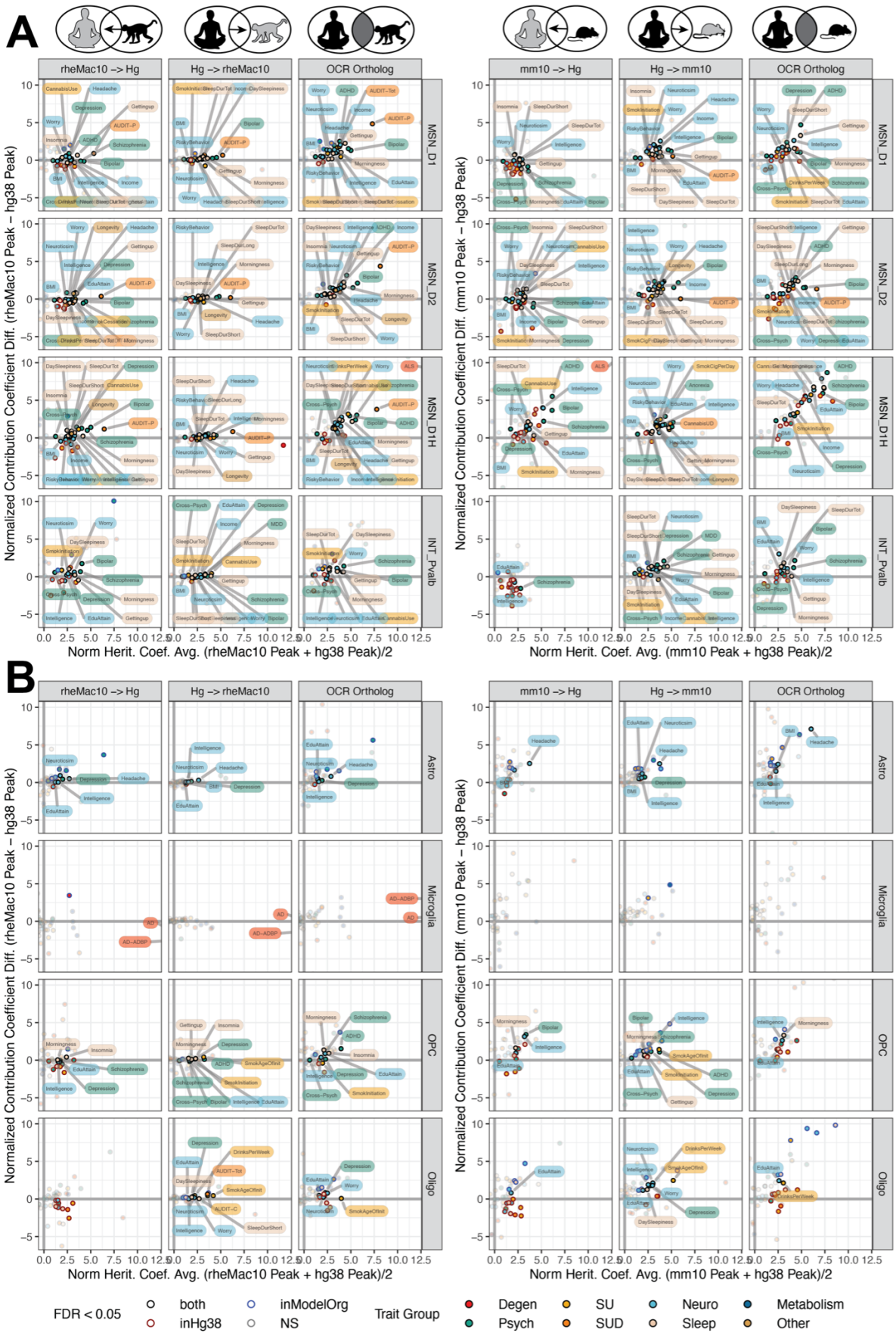

A
rheMac10 –> Hg
Hg –> rheMac10
OCR Ortholog
mm10 –> Hg
Hg –> mm10
OCR Ortholog
MSN_D1
MSN_D2
MSN_D1H
INT_Pvalb
Normalized Contribution Coefficient Diff. (rheMac10 Peak – hg38 Peak)
Normalized Contribution Coefficient Diff. (mm10 Peak – hg38 Peak)
Norm Herit. Coef. Avg. (rheMac10 Peak + hg38 Peak)/2
Norm Herit. Coef. Avg. (mm10 Peak + hg38 Peak)/2
B
Astro
Microglia
OPC
Oligo
FDR < 0.05
both
inHg38
inModelOrg
NS
Trait Group
Degen
Psych
SU
SUD
Neuro
Sleep
Metabolism
Other

Figure S10. Pairwise human-model species caudate cell type OCR comparisons demonstrate complex traits enriched within conserved open chromatin in the human genome.

Conditional independence (τ, normalized heritability coefficient) LD score regression (LDSC) analyses of caudate cell type open chromatin regions (OCRs) using human, rhesus macaque, and mouse open chromatin. Each point is a GWAS colored by the trait group (Data S1). The τ* was estimated for each GWAS in 7 OCR sets for each cell type: 0) human OCRs; 1) rhesus macaque OCRs mapped to human genome, rheMac10 -> Hg; 2) human OCRs that can be mapped to rhesus macaque, Hg -> rheMac10; 3) overlap of human and rhesus OCRs and OCR orthologs 4) mouse OCRs that can be mapped to human, mm10 -> Hg; 5) human OCRs that can be mapped to mouse, Hg -> mm10; and 6) overlap of human and mouse OCRs and OCR ortholog. The τ* for OCR sets mappable or open chromatin conserved in other species (1-6) are compared to human OCRs (0) using a mean-difference plot where the average τ* are plotted on the x-axis (normalized heritability coefficient average) and the change in τ* with mappability or orthology in model species are plotted on the y-axis (normalized heritability coefficient difference). Points above the y=0 line demonstrate higher τ* with genome or open chromatin conservation information above human cell type OCRs alone. Significant τ* LDSC estimates are outlined with red for significance in human OCRs (inHg38), significant in Model Species (inModelOrg, 1-6), or both (FDR < 0.05, computed over all 8 cell types, 64 traits, 7 OCR set comparisons). Traits that does not have significant τ* estimates are plotted with transparency.* ***(A)*** *Neuronal cell type comparisons: MSN_D1, MSN_D2, MSN_SN, and INT_Pvalb.* ***(B)*** *Glial cell type comparisons: Astro, Microglia, OPC, and Oligo.*

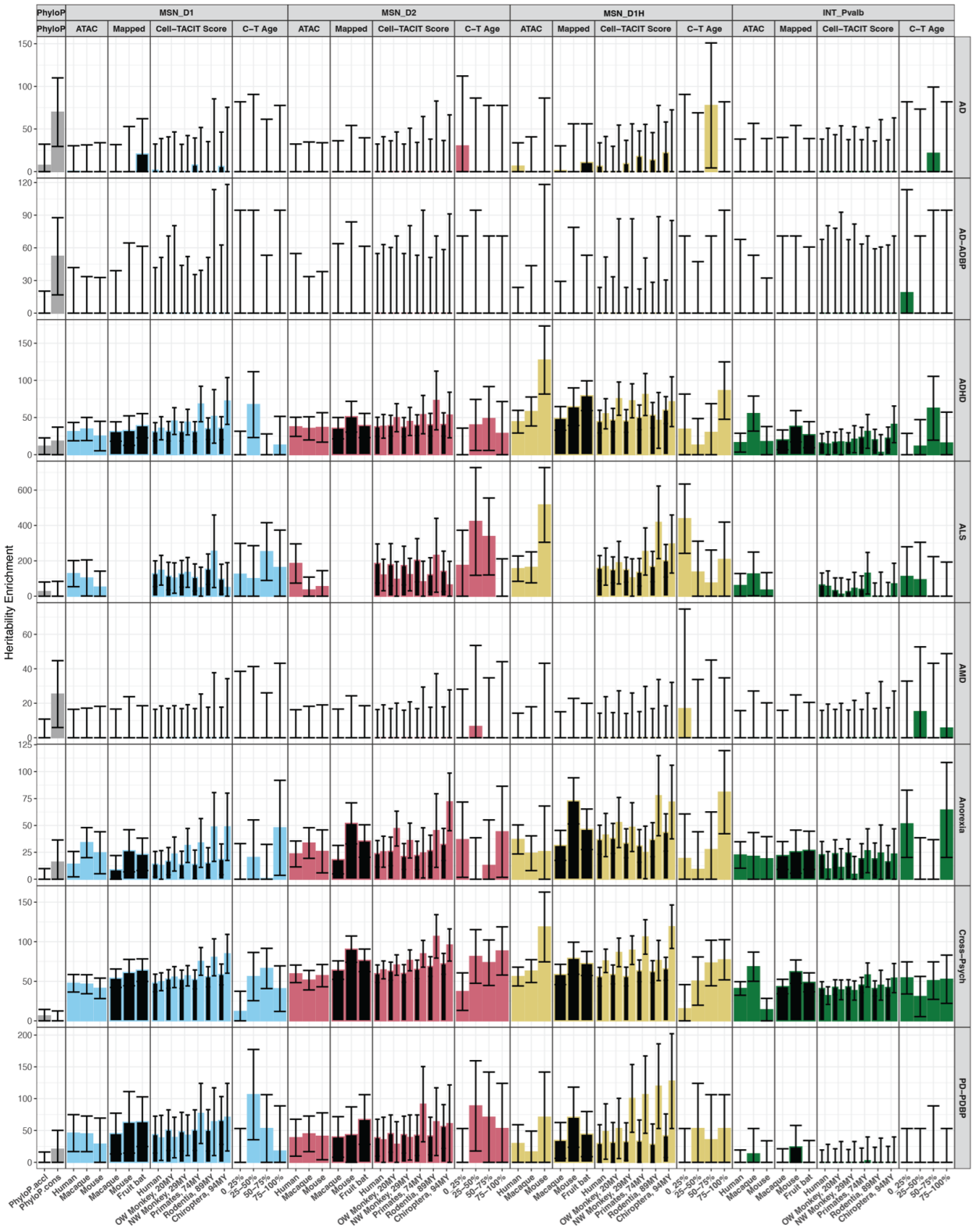


Figure S11. Heritability enrichment of neurodegenerative and psychiatric traits and the conserved neuronal epigenome

Heritability enrichments in D1, D2, and D1H MSNs, Pvalb interneurons cell type OCRs from a number of sources showed cell type OCRs measured or predicted activity in other species

relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars).

C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: Alzheimer's Disease, AD; Alzheimer's disease or AD-by-proxy (AD-ADBP); ADHD Amyotrophic Lateral Sclerosis (ALS); Age-related macular degeneration (AMD); Parkinson's Disease or Parkinson's Disease-by-proxy (PD-PDBP).

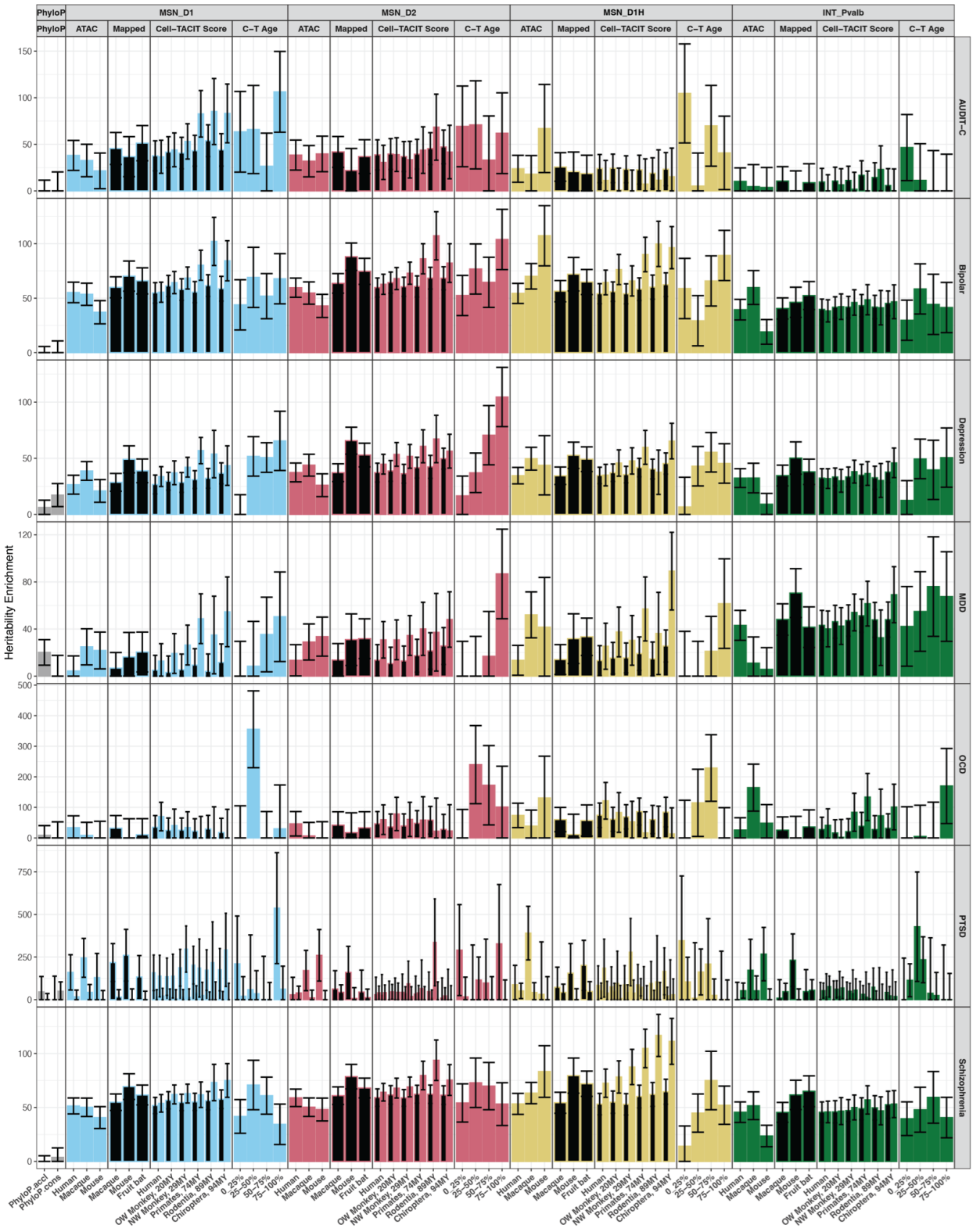


Figure S12. Heritability enrichment of substance abuse and psychiatric traits and the conserved neuronal epigenome

Heritability enrichments in D1, D2, and D1H MSNs, Pvalb interneurons cell type OCRs from a number of sources showed cell type OCRs measured or predicted activity in other species

relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: Alcohol use-disorder identification test, consumption domain (AUDIT-C); MDD, major depressive disorder; OCD, obsessive-compulsive disorder; PTSD, post-traumatic disorder.

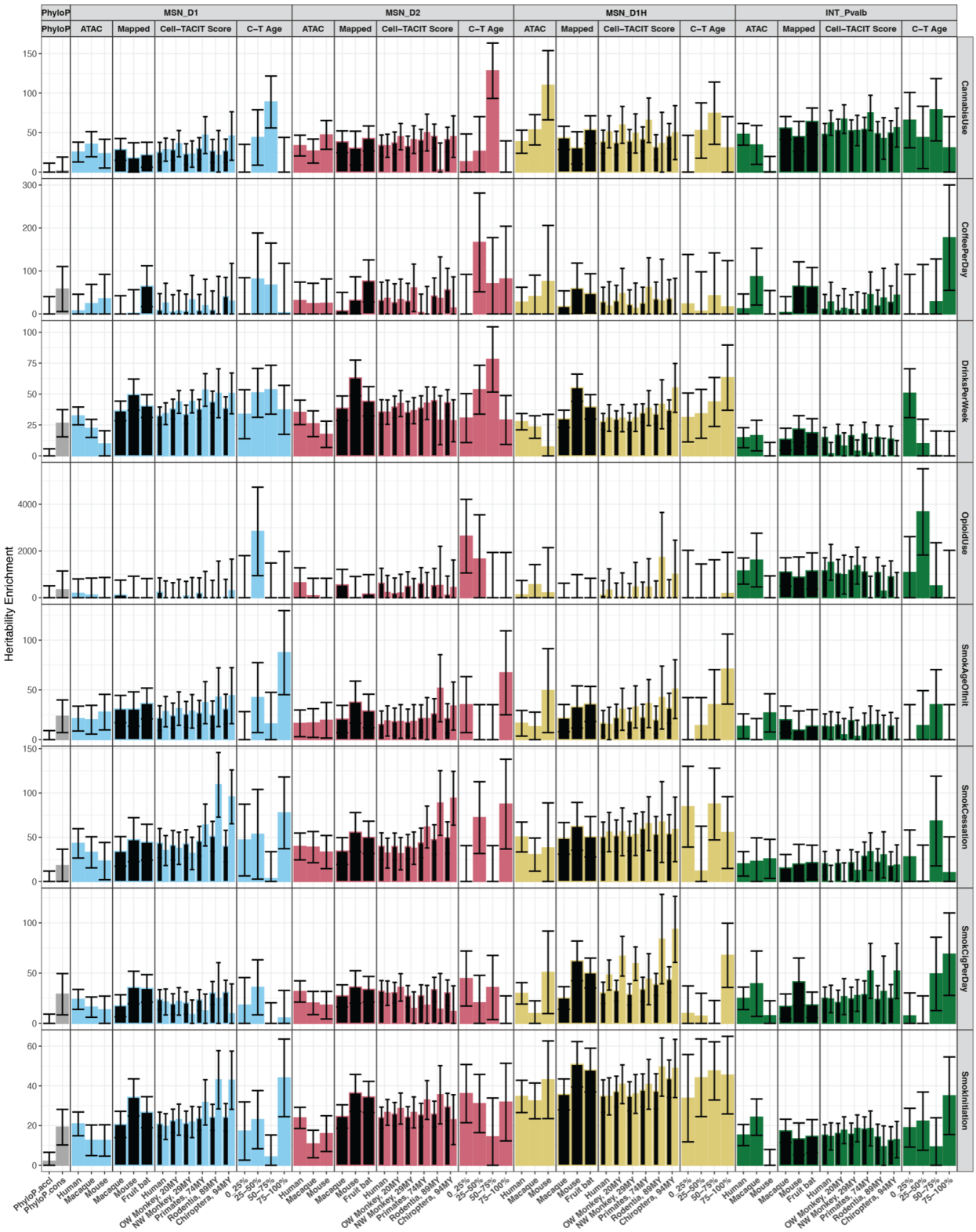


Figure S13. Heritability enrichment of substance abuse traits and the conserved neuronal epigenome

Heritability enrichments in D1, D2, and D1H MSNs, Pvalb interneurons cell type OCRs from a number of sources showed cell type OCRs measured or predicted activity in other species

relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: age of smoking initiation (SmokAgeInit); previous smoker vs. current smoker (SmokCessation); current smoker vs. non-smoker (SmokInitiation).

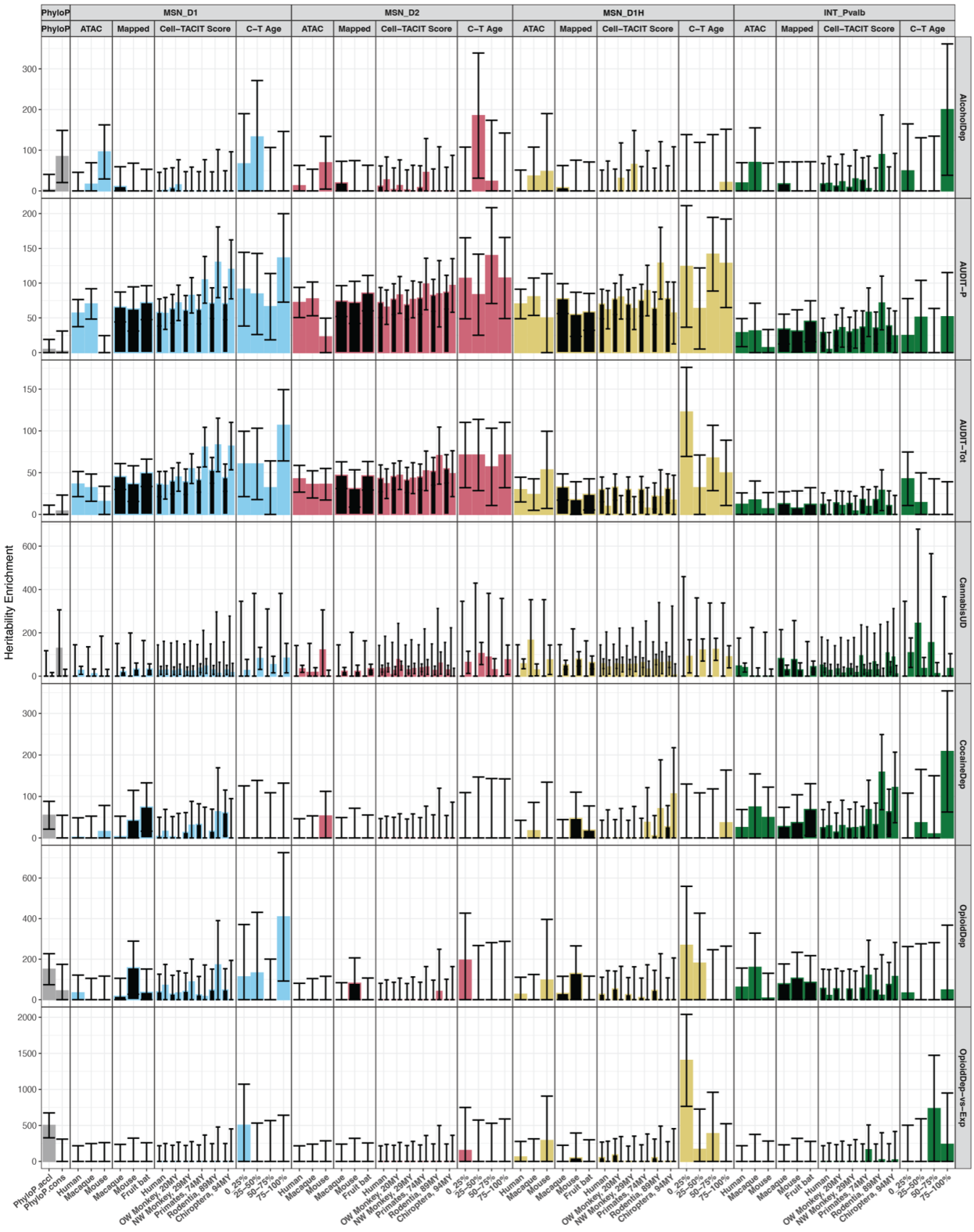


Figure S14. Heritability enrichment of substance use disorder traits and the conserved neuronal epigenome

Heritability enrichments in D1, D2, and D1H MSNs, Pvalb interneurons cell type OCRs from a number of sources showed cell type OCRs measured or predicted activity in other species

relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: Alcohol dependence (AlcDependence); alcohol use disorder identification test, problematic use domain (AUDIT-P); alcohol use disorder identification test, total score (AUDIT-Tot); cannabis use disorder (CannabisUD); Cocaine dependence (CocaineDep); Opioid dependence (OpioidDep); Opioid dependence vs. opioid exposure (OpioidDep-vs-Exp).

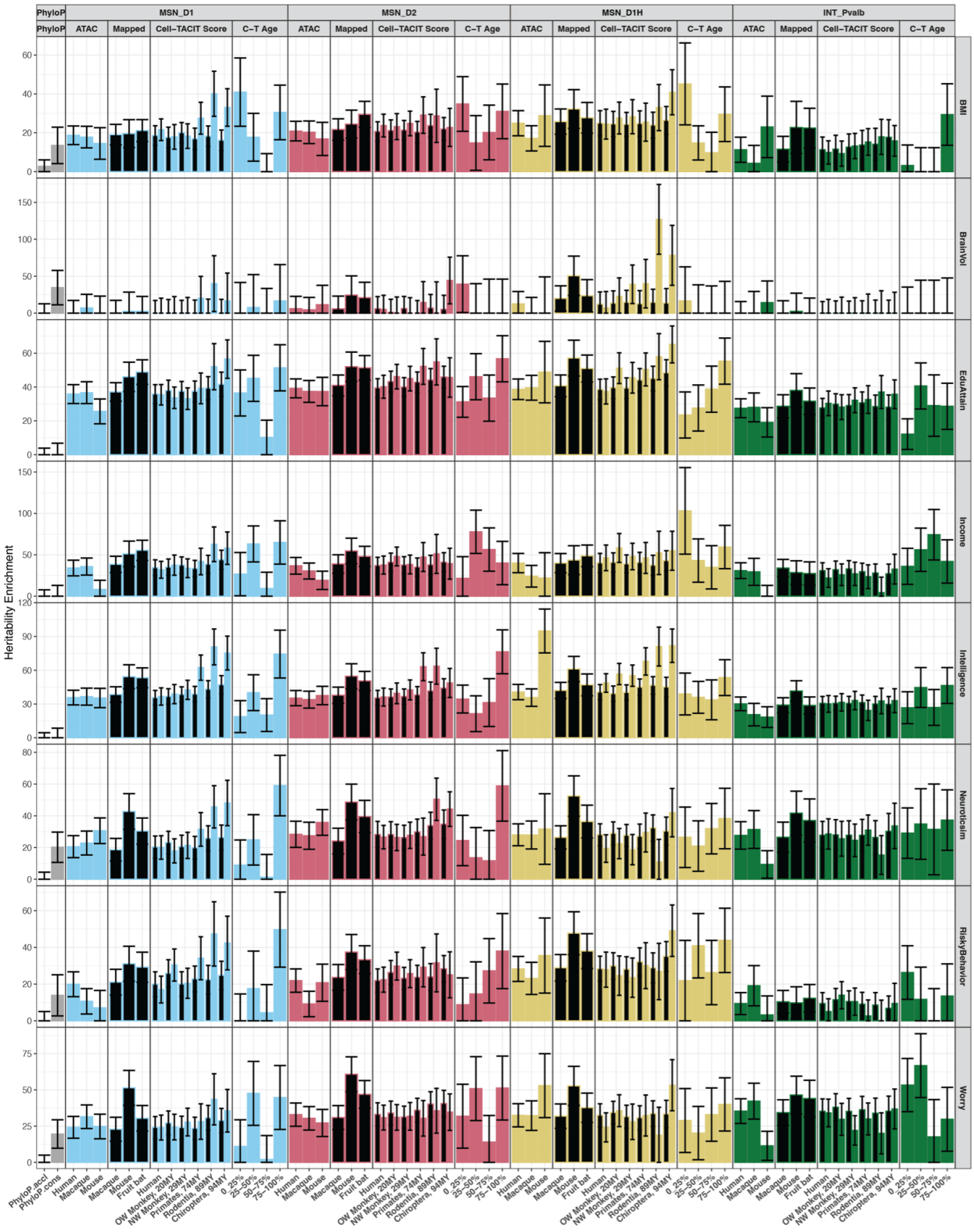


Figure S15. Heritability enrichment of neurological traits and the conserved neuronal epigenome

Heritability enrichments in D1, D2, and D1H MSNs, Pvalb interneurons cell type OCRs from a number of sources showed cell type OCRs measured or predicted activity in other species relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: Body-mass index (BMI), brain volume (BrainVol), educational attainment (EduAttain).

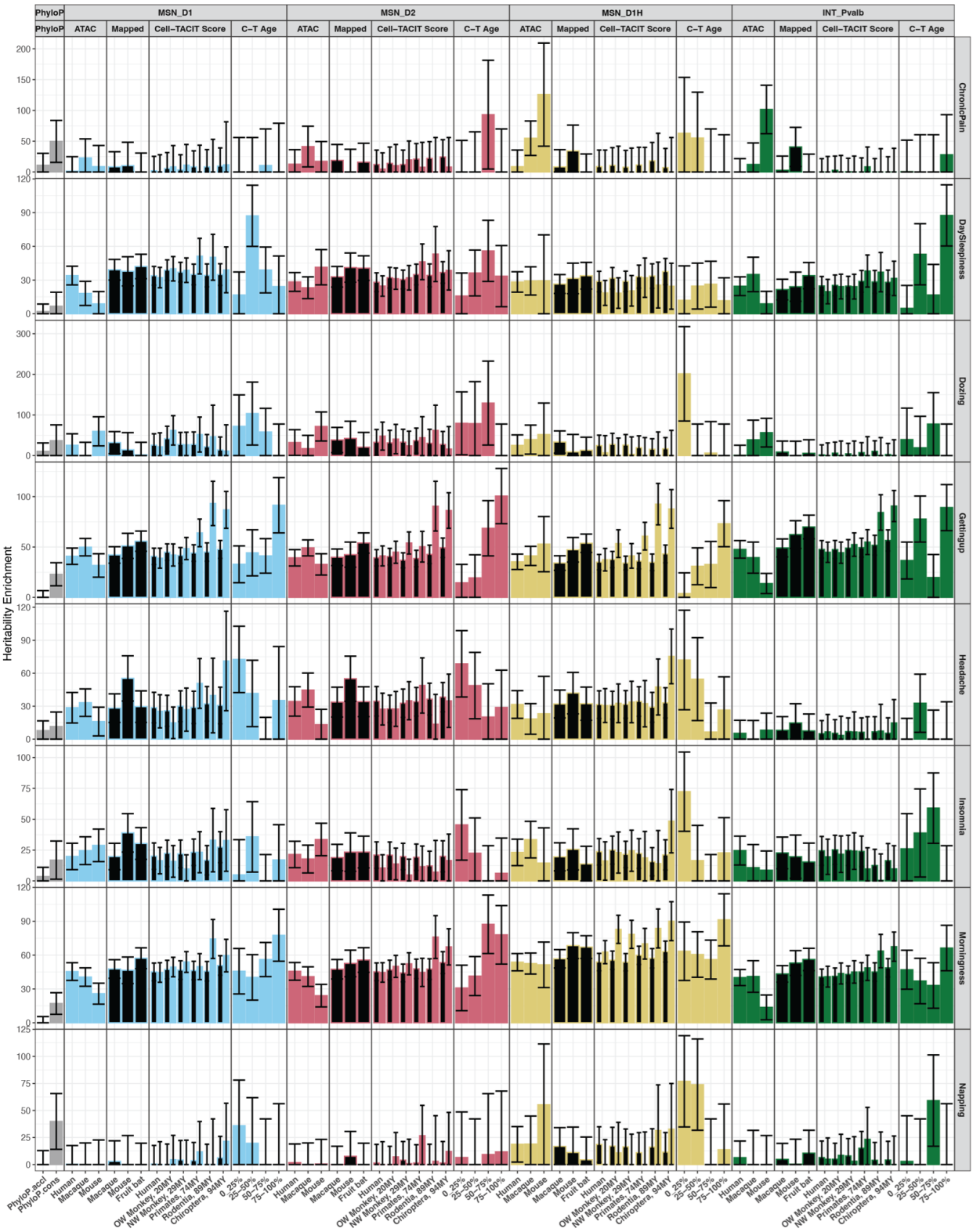


Figure S16. Heritability enrichment of chronic pain and sleep-related traits and the conserved neuronal epigenome

Heritability enrichments in D1, D2, and D1H MSNs, Pvalb interneurons cell type OCRs from a number of sources showed cell type OCRs measured or predicted activity in other species

relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: multi-site chronic pain (ChronicPain), daytime sleepiness (DaySleepiness), ease of getting up in the morning (GettingUp), preference for morning vs. evening (morningness).

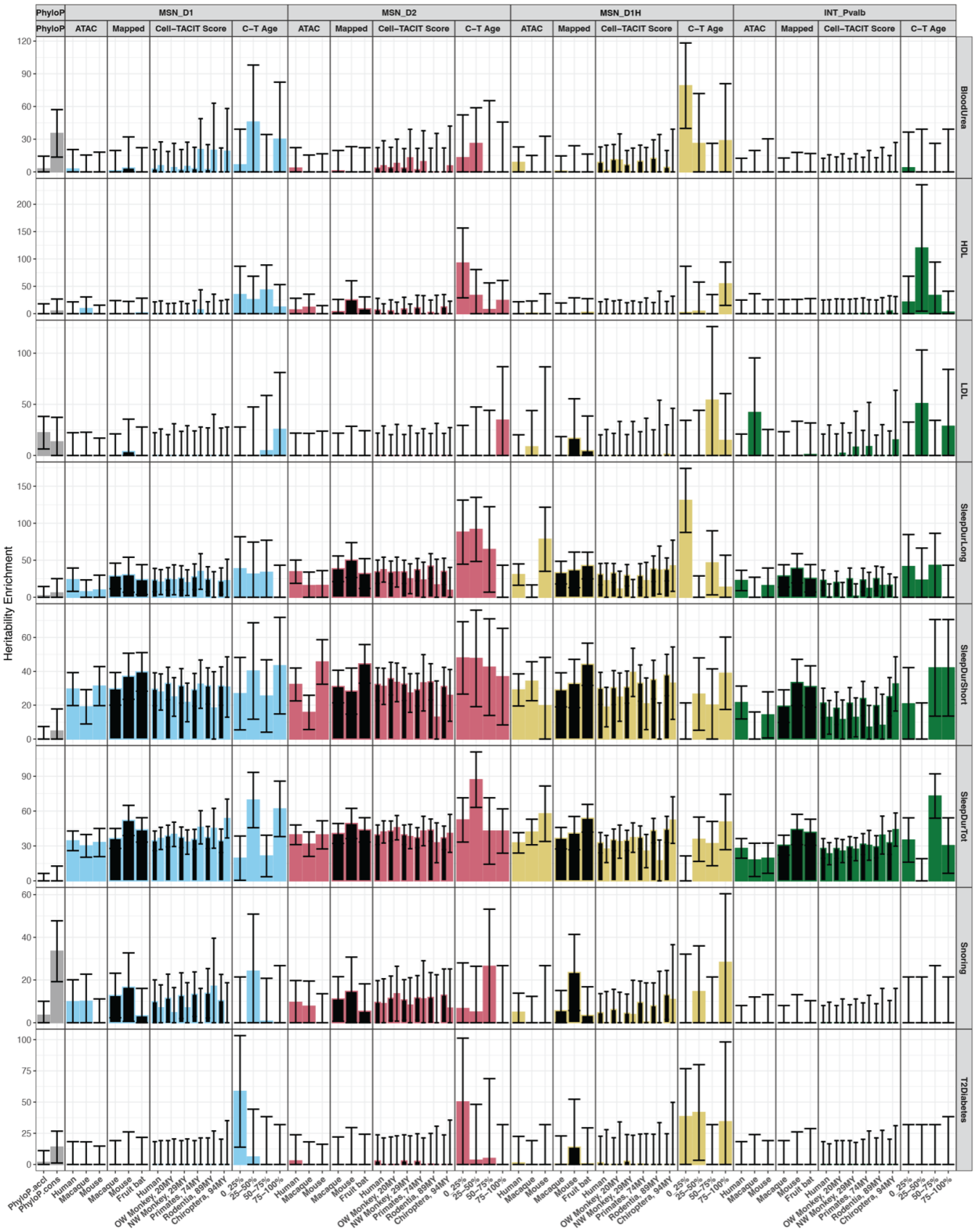


Figure S17. Heritability enrichment of metabolic and sleep-related traits and the conserved neuronal epigenome

Heritability enrichments in D1, D2, and D1H MSNs, Pvalb interneurons cell type OCRs from a number of sources showed cell type OCRs measured or predicted activity in other species

relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: high-density lipoprotein (HDL), low-density lipoprotein (LDL), long sleep duration (SleepDurLong), short sleep duration (SleepDurShort), total hours of sleep (SleepDurTot), type 2 diabetes mellitus (T2Diabetes).

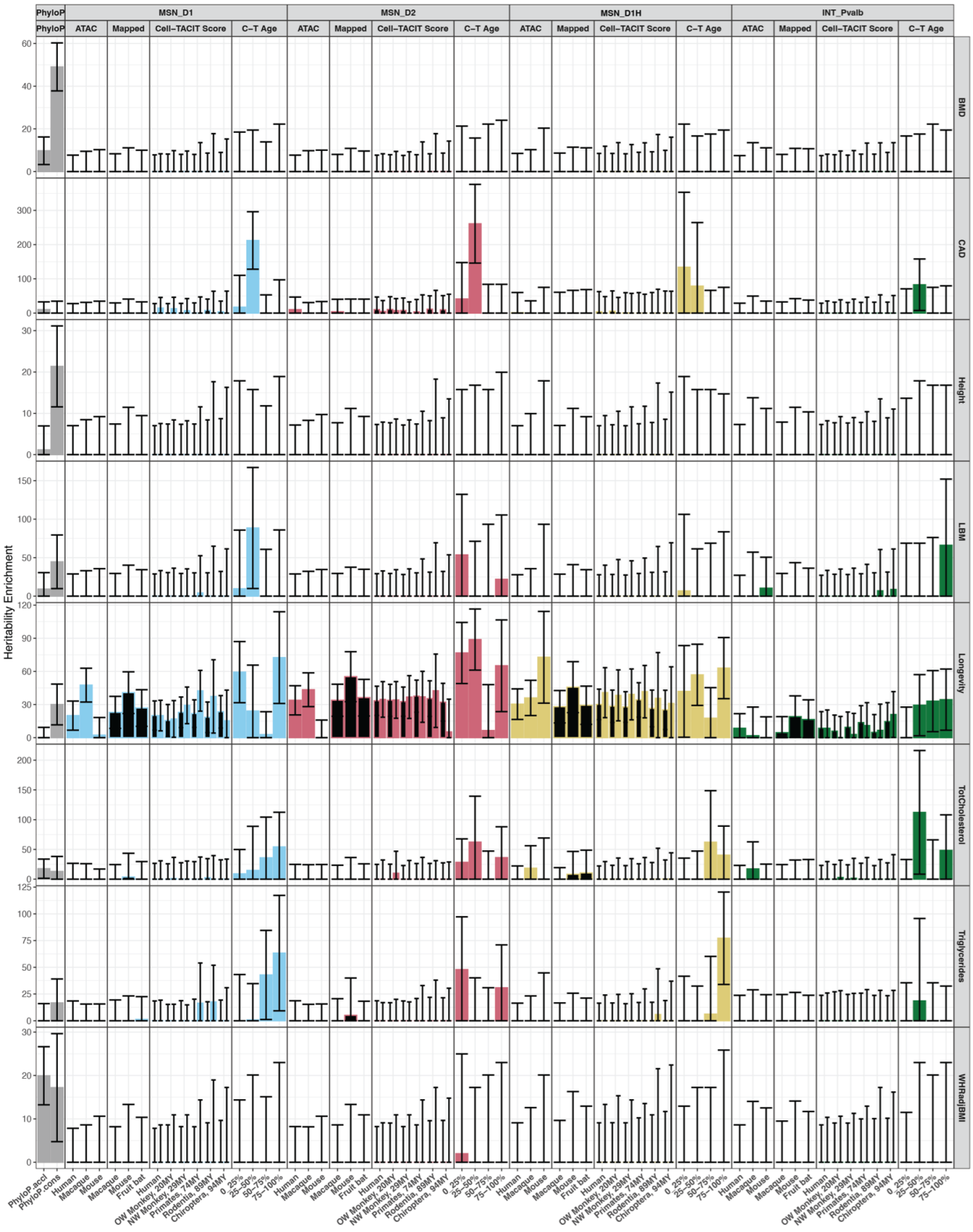


Figure S18. Heritability enrichment of anthropomorphic and sleep-related traits and the conserved neuronal epigenome

Heritability enrichments in D1, D2, and D1H MSNs, Pvalb interneurons cell type OCRs from a number of sources showed cell type OCRs measured or predicted activity in other species

relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: heel T-score bone mineral density (BMD), coronary artery disease (CAD), lean body mass (LBM), parental longeivity (Longevity), total cholesterol (TotCholesterol), waist-hip ratio adjusted for body mass index (WHRadjBMI).

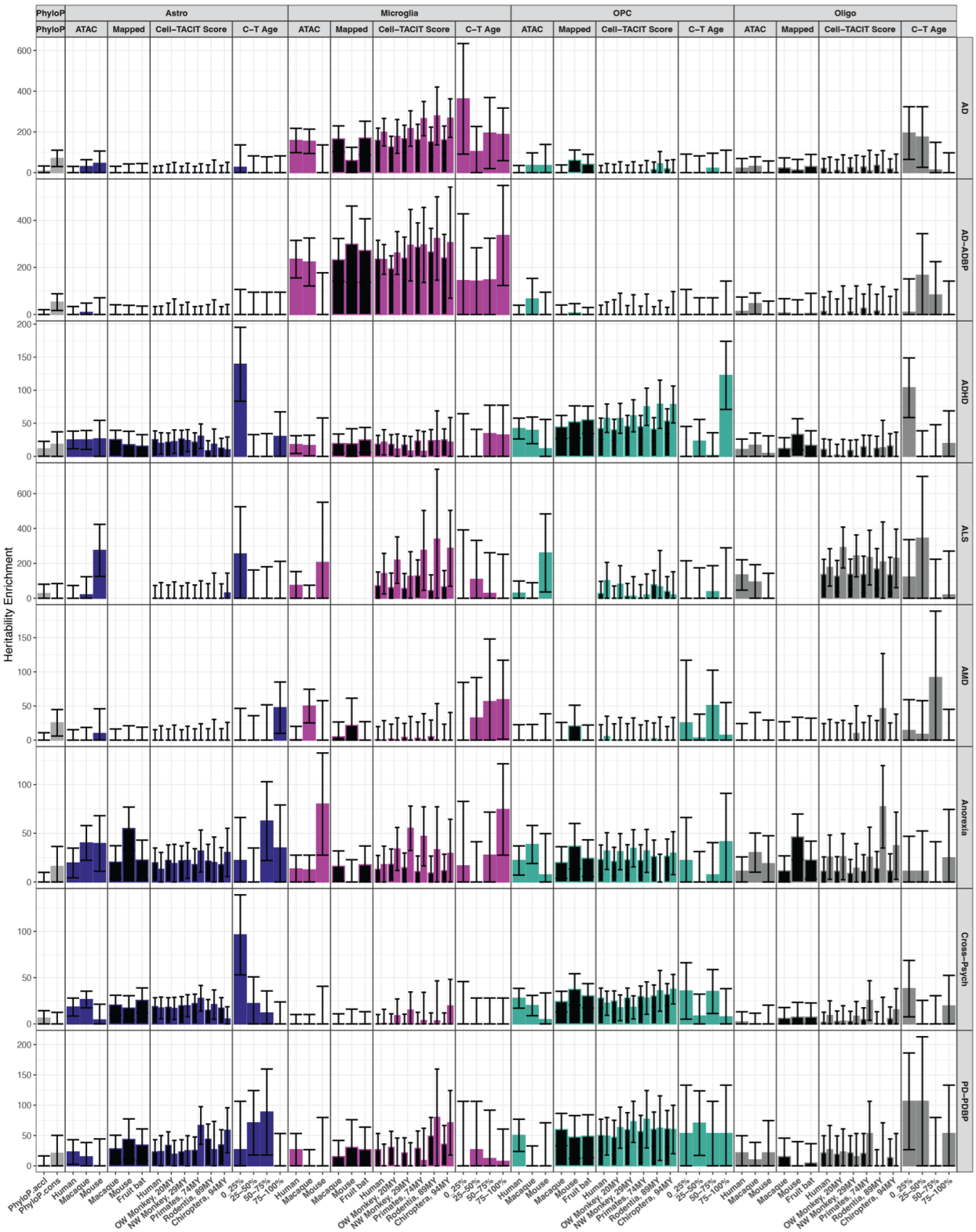


Figure S19. Heritability enrichment of neurodegenerative and psychiatric traits and the conserved glia and liver epigenome

Heritability enrichments in astrocyte, microglia, oligodendrocyte precursor cells, oligodendrocyte, and liver cell type OCRs from a number of sources showed cell type OCRs

measured or predicted activity in other species relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: Alzheimer's Disease, AD; Alzheimer's disease or AD-by-proxy (AD-ADBP); ADHD Amyotrophic Lateral Sclerosis (ALS); Age-related macular degeneration (AMD); Parkinson's Disease or Parkinson's Disease-by-proxy (PD-PDBP).

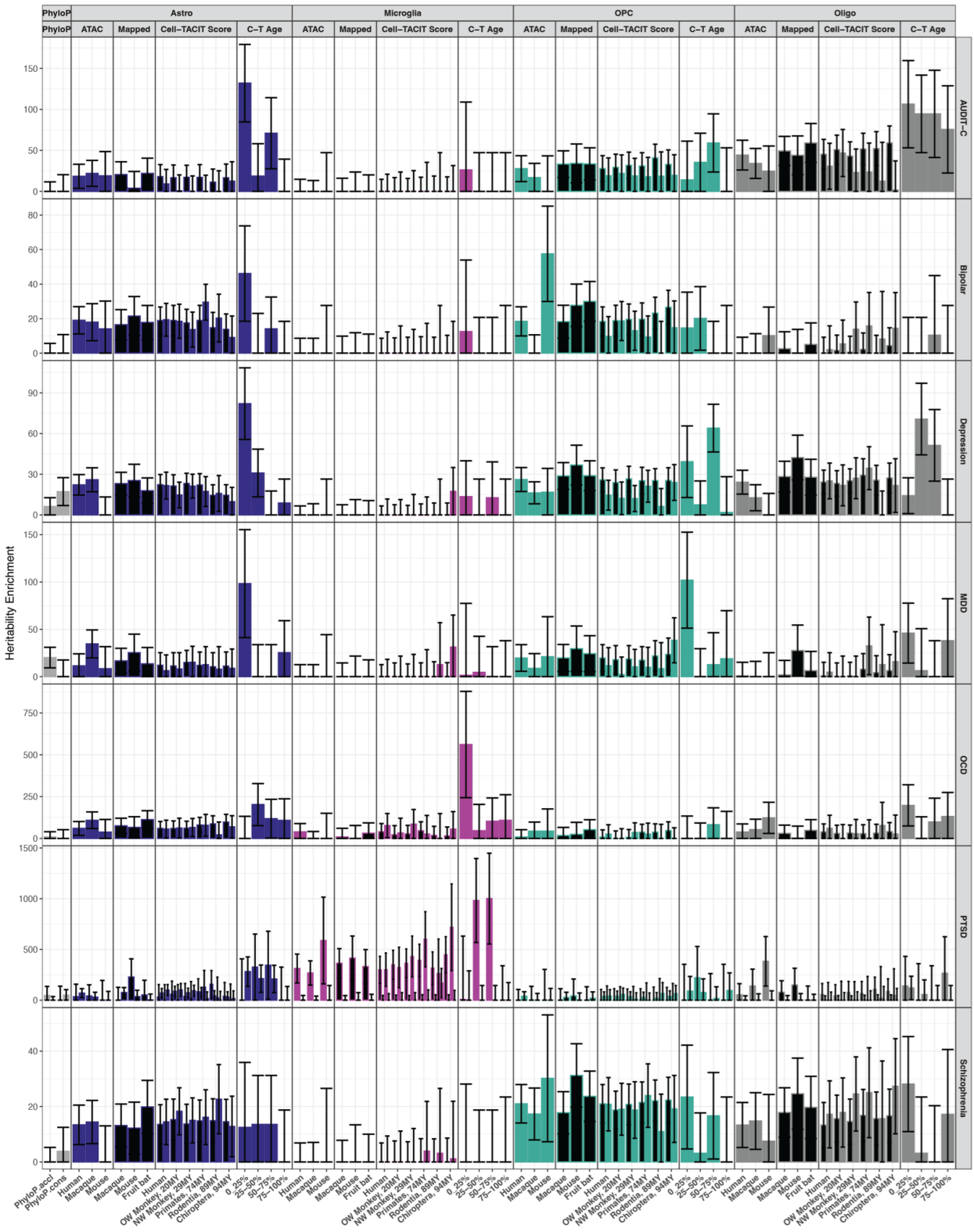


Figure S20. Heritability enrichment of substance use and psychiatric traits and the conserved glia and liver epigenome

Heritability enrichments in astrocyte, microglia, oligodendrocyte precursor cells, oligodendrocyte, and liver cell type OCRs from a number of sources showed cell type OCRs measured or predicted activity in other species relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: Alcohol use-disorder identification test, consumption domain (AUDIT-C); major depressive disorder (MDD); obsessive-compulsive disorder (OCD); post-traumatic disorder (PTSD).

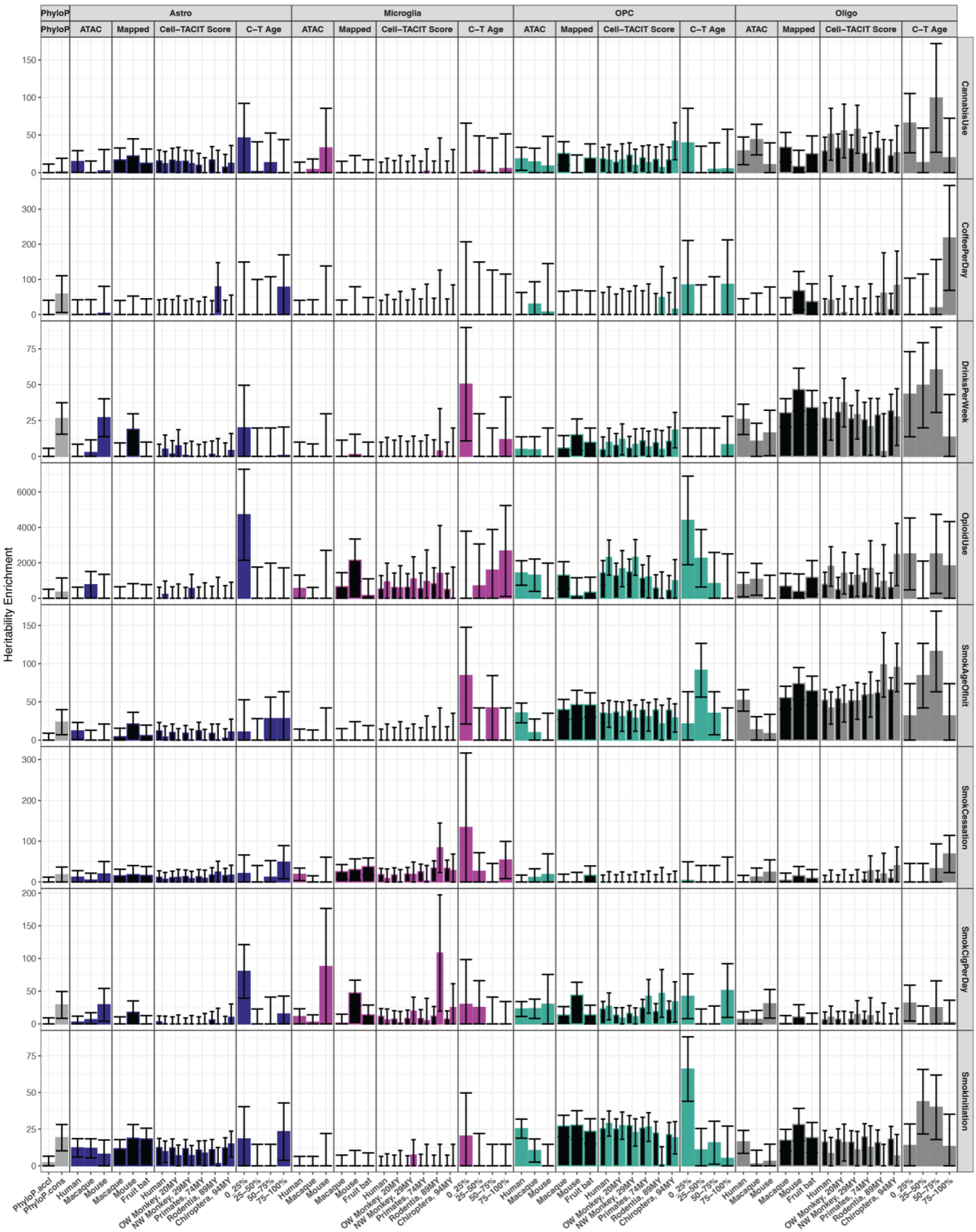


Figure S21. Heritability enrichment of substance use traits and the conserved glia and liver epigenome

Heritability enrichments in astrocyte, microglia, oligodendrocyte precursor cells, oligodendrocyte, and liver cell type OCRs from a number of sources showed cell type OCRs

measured or predicted activity in other species relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: age of smoking initiation (SmokAgeInit); previous smoker vs. current smoker (SmokCessation); current smoker vs. non-smoker (SmokInitiation).

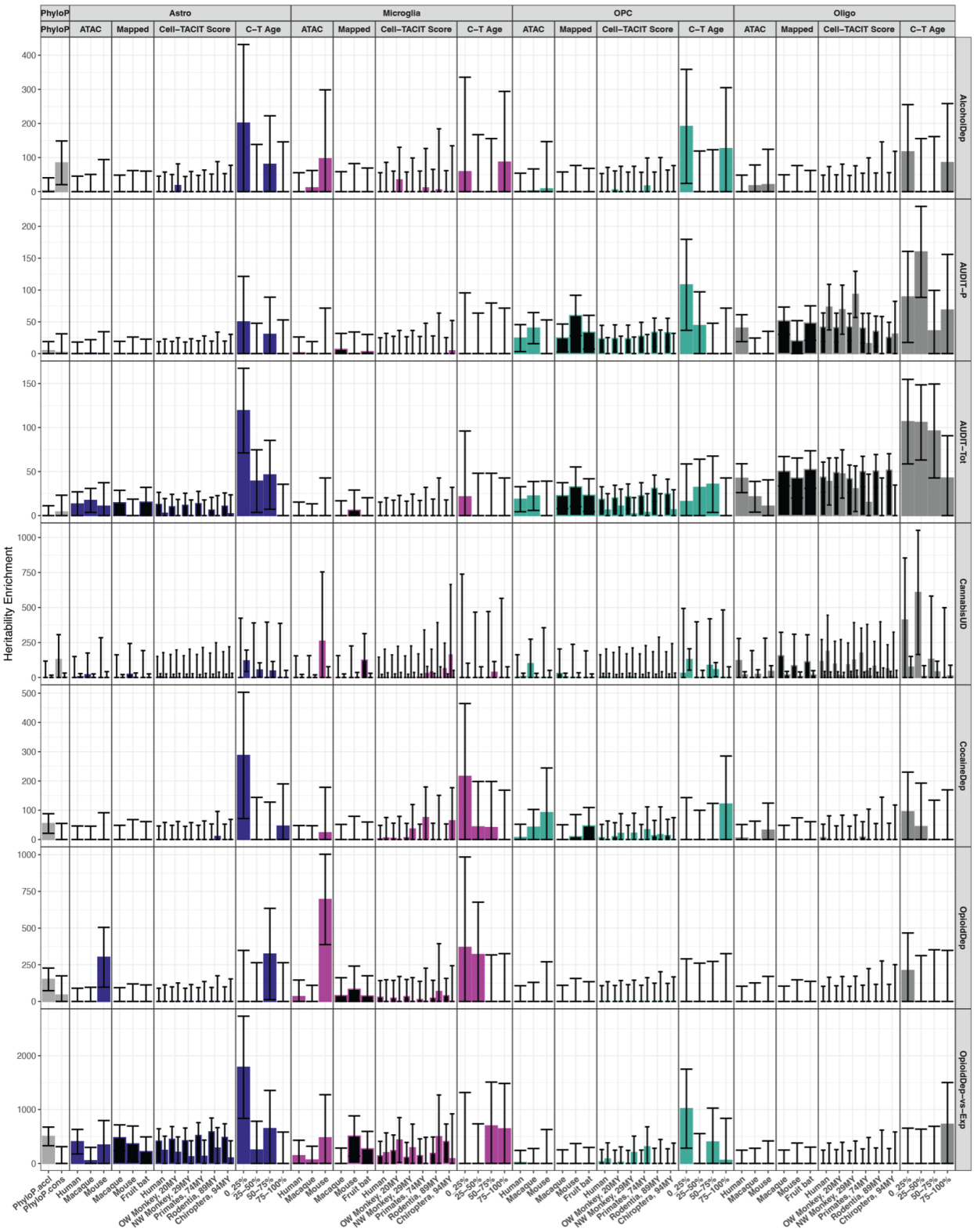


Figure S22. Heritability enrichment of substance use disorder traits and the conserved glia and liver epigenome

Heritability enrichments in astrocyte, microglia, oligodendrocyte precursor cells, oligodendrocyte, and liver cell type OCRs from a number of sources showed cell type OCRs measured or predicted activity in other species relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: Alcohol dependence (AlcDependence); alcohol use disorder identification test, problematic use domain (AUDIT-P); alcohol use disorder identification test, total score (AUDIT-Tot); cannabis use disorder (CannabisUD); Cocaine dependence (CocaineDep); Opioid dependence (OpioidDep); Opioid dependence vs. opioid exposure (OpioidDep-vs-Exp).

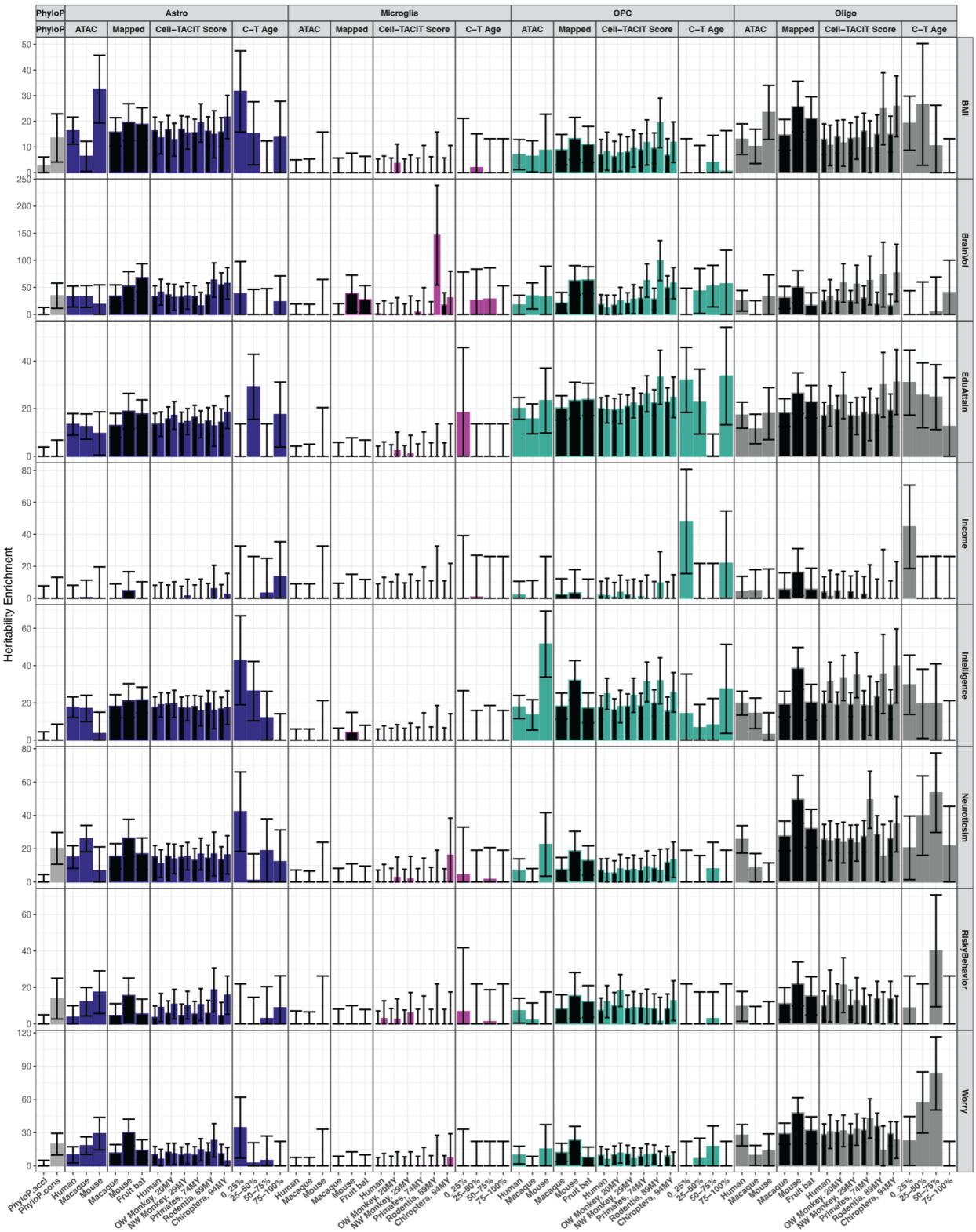


Figure S23. Heritability enrichment of neurological traits and the conserved glia and liver epigenome

Heritability enrichments in astrocyte, microglia, oligodendrocyte precursor cells, oligodendrocyte, and liver cell type OCRs from a number of sources showed cell type OCRs

measured or predicted activity in other species relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: Body-mass index (BMI), brain volume (BrainVol), educational attainment (EduAttain).

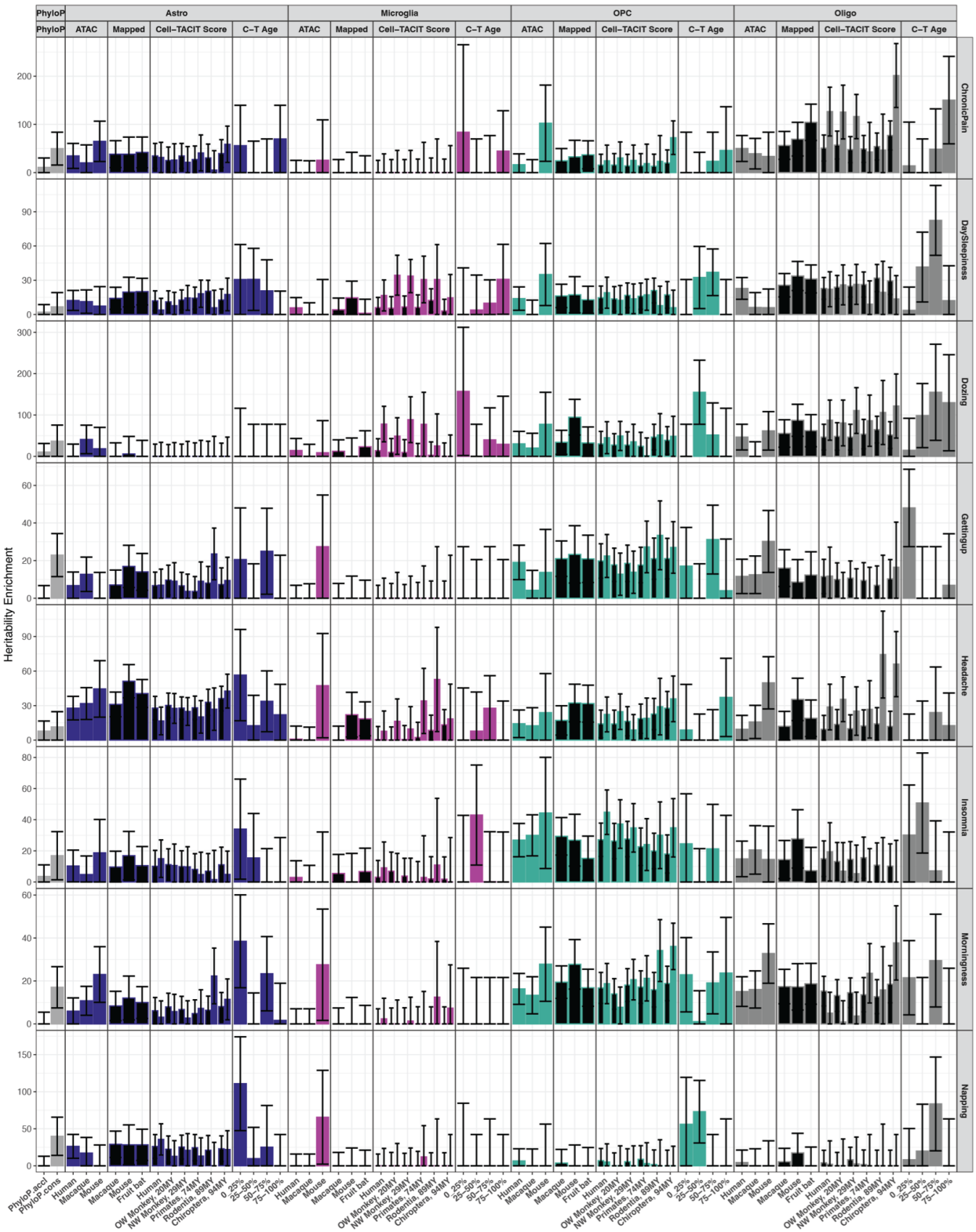


Figure S24. Heritability enrichment of chronic pain and sleep-related traits and the conserved glia and liver epigenome

Heritability enrichments in astrocyte, microglia, oligodendrocyte precursor cells, oligodendrocyte, and liver cell type OCRs from a number of sources showed cell type OCRs

measured or predicted activity in other species relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: multi-site chronic pain (ChronicPain), daytime sleepiness (DaySleepiness), ease of getting up in the morning (GettingUp), preference for morning vs. evening (morningness).

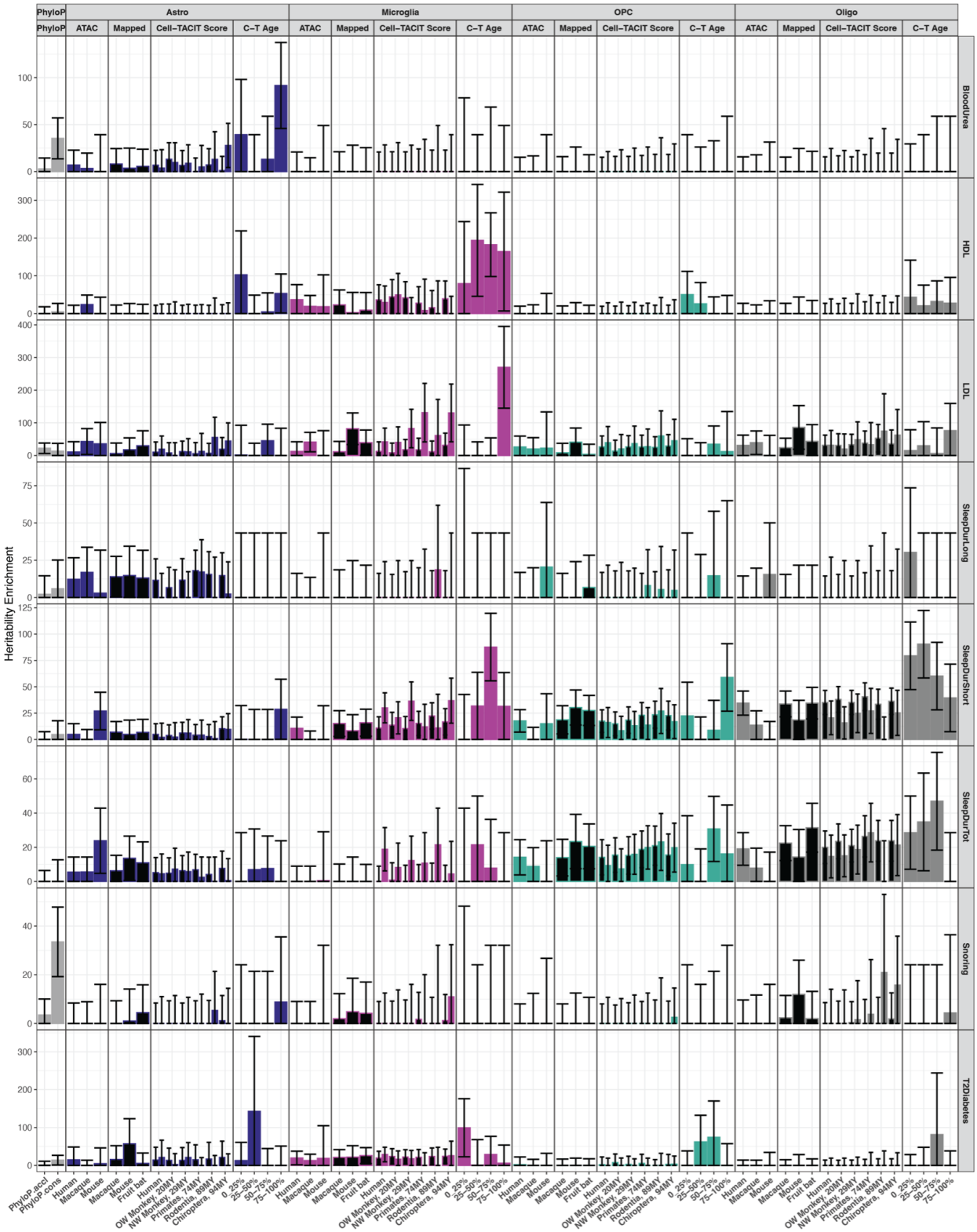


Figure S25. Heritability enrichment of metabolic and sleep-related traits and the conserved glia and liver epigenome

Heritability enrichments in astrocyte, microglia, oligodendrocyte precursor cells, oligodendrocyte, and liver cell type OCRs from a number of sources showed cell type OCRs

measured or predicted activity in other species relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: high-density lipoprotein (HDL), low-density lipoprotein (LDL), long sleep duration (SleepDurLong), short sleep duration (SleepDurShort), total hours of sleep (SleepDurTot), type 2 diabetes mellitus (T2Diabetes).

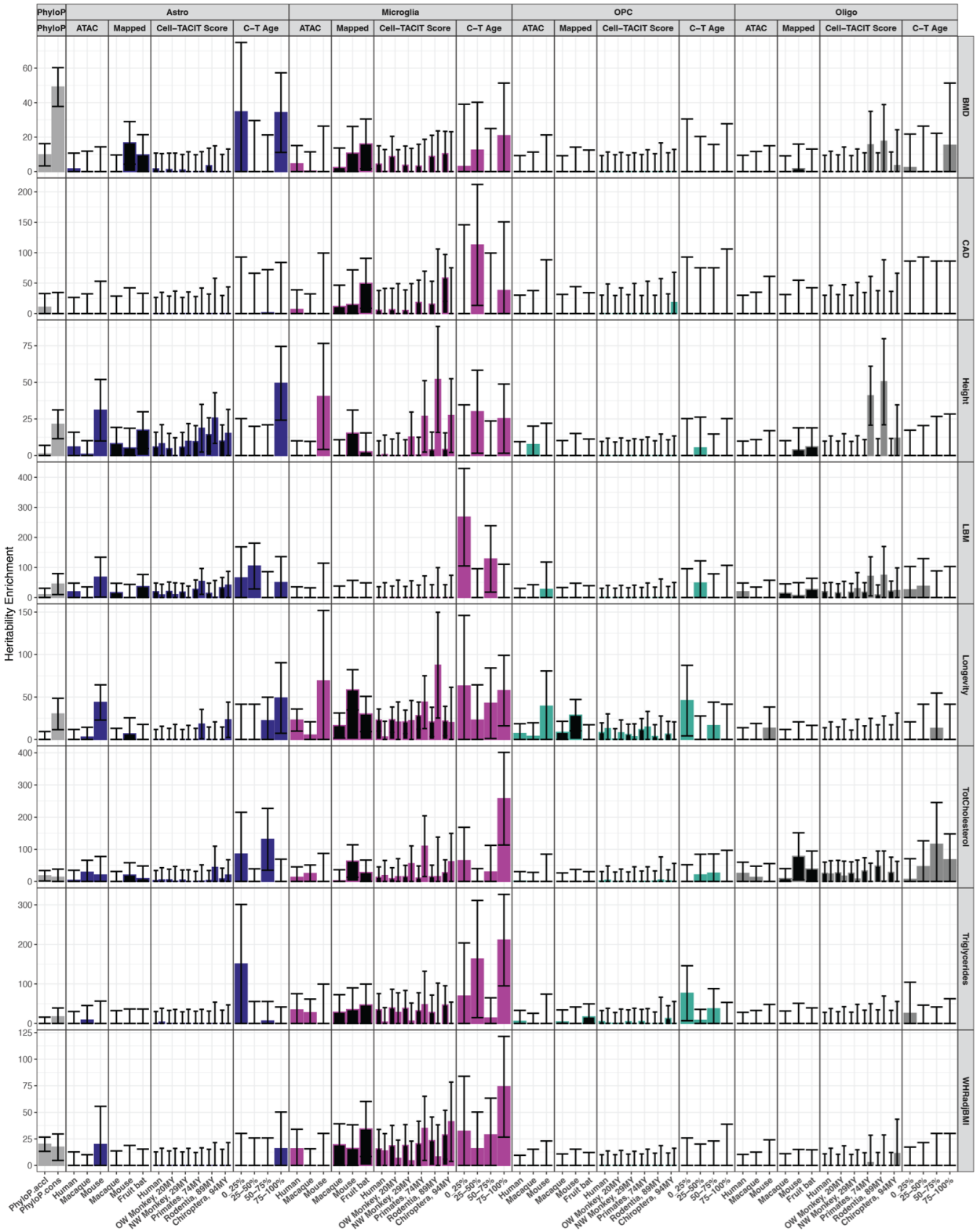


Figure S26. Heritability enrichment of anthropomorphic and sleep-related traits and the conserved glia and liver epigenome

Heritability enrichments in astrocyte, microglia, oligodendrocyte precursor cells, oligodendrocyte, and liver cell type OCRs from a number of sources showed cell type OCRs

measured or predicted activity in other species relevant to human complex traits. Barplots are the heritiability enrichment from LD score regression analyses with error bars plotting standard error of the enrichment.

phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species (black bars) or mappable and predicted active by Cell-TACIT machine learning models (colored bars). C-T Age: Cell-TACIT Age estimating whether human cell type open chromatin is predicted active in closely or across distantly diverged placental mammals. The top quartile corresponds to open chromatin regions that are "older" in placental mammalian genomes.

Trait: heel T-score bone mineral density (BMD), coronary artery disease (CAD), lean body mass (LBM), parental longevity (Longevity), total cholesterol (TotCholesterol), waist-hip ratio

adjusted for body mass index (WHRadjBMI).

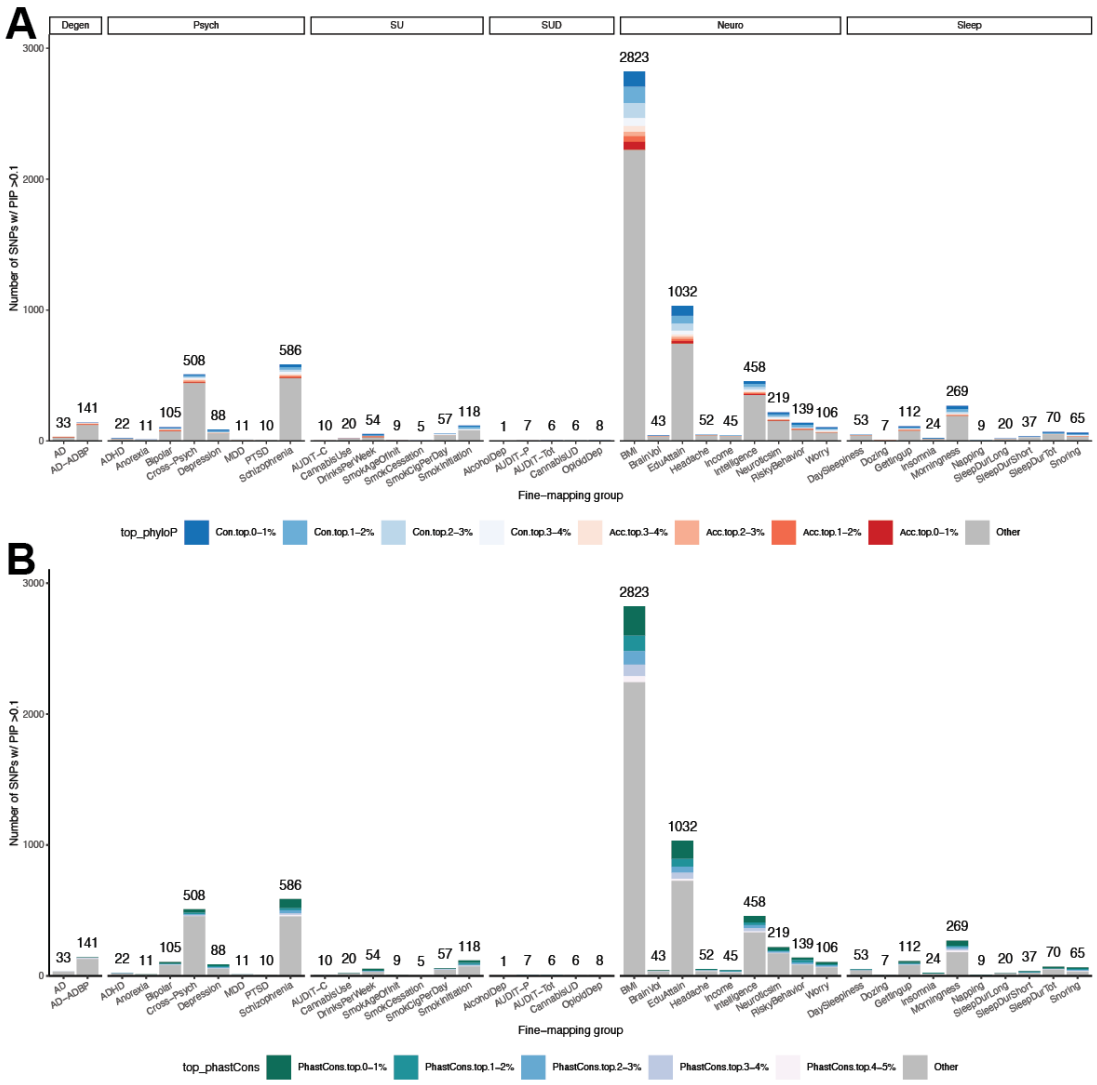


Figure S27. Fine-mapping of neuropsychiatric traits show varying number of genetic variants related to power of GWAS to detect genome-wide loci

(A) *Candidate causal SNPs in neuropsychiatric traits with non-negligible posterior inclusion probabilities (PIP> 0.1). SNPs are annotated by overlap with the top mammalian phyloP accelerated or constrained groups.*

(B) As in (A) and SNPs are annotated by overlap with top constrained primate Phastcons groups. Neurodegenerative (Degen), psychiatric disorder (Psych), substance use (SU), substance use disorder (SUD), neurological (Neuro), traits as in Data S1.

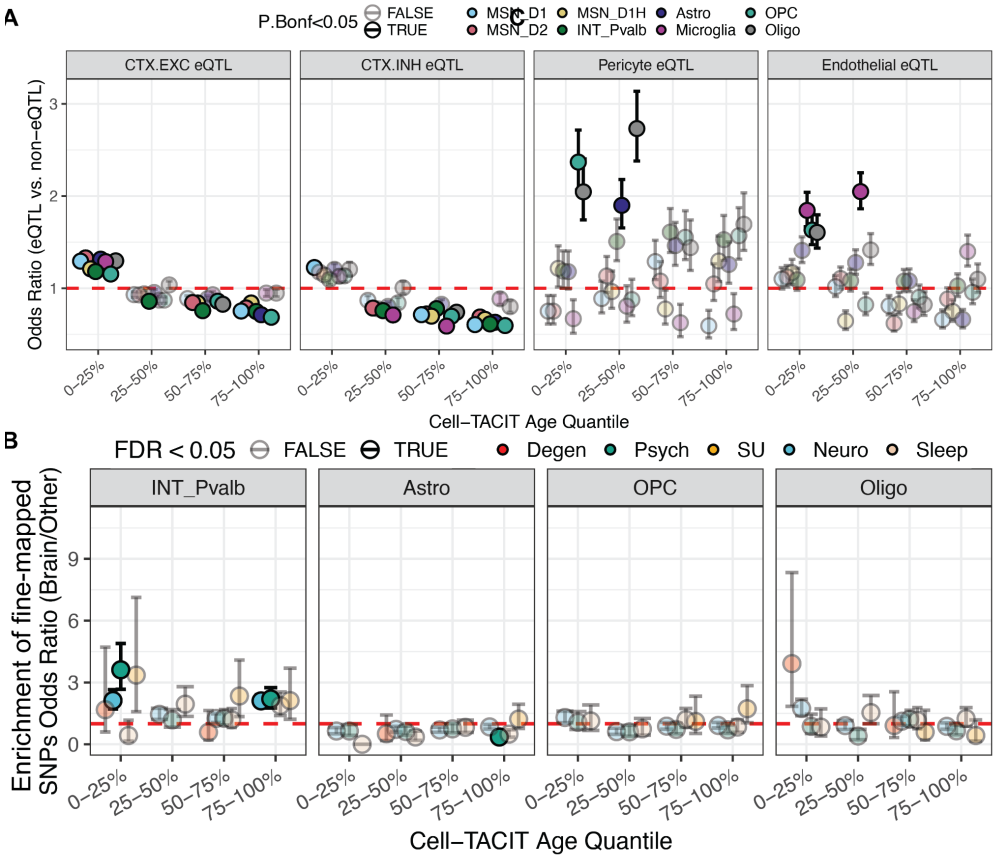


Figure S28. Comparison of Cell-TACIT Ages with off-target eQTLs and fine-mapped neuropsychiatric trait genetic variants.

(A) *Stripchart plotting the odds ratio (±standard error of the mean) that significant eQTLs from single-cell prefrontal cortex cell types with Cell-TACIT Age quartiles. These prefrontal cortex cell types do not have one-to-one correspondence with Cell-TACIT caudate cell types. Cell types with one-to-one correspondence are shown in* ***Figure 3B.*** *Each point is colored by the Cell-TACIT model. Solid points are significant enrichment at Bonferroni corrected P-value < 0.05.* ***(B)*** *Enrichment stripchart for fine-mapped neuropsychiatric trait SNPs with PIP > 0.10 in Cell-TACIT Age quartiles for the other caudate neural cell types. Each point is colored by enrichment for SNPs fine-mapped in neurodegenerative traits (Degen), psychiatric traits (Psych), substance use traits (SU), neurological traits (Neuro), or sleep traits (Sleep). The top Cell-TACIT cell types are shown in* ***Figure 3C****. Solid points are significant enrichment at FDR corrected P-value < 0.05.*

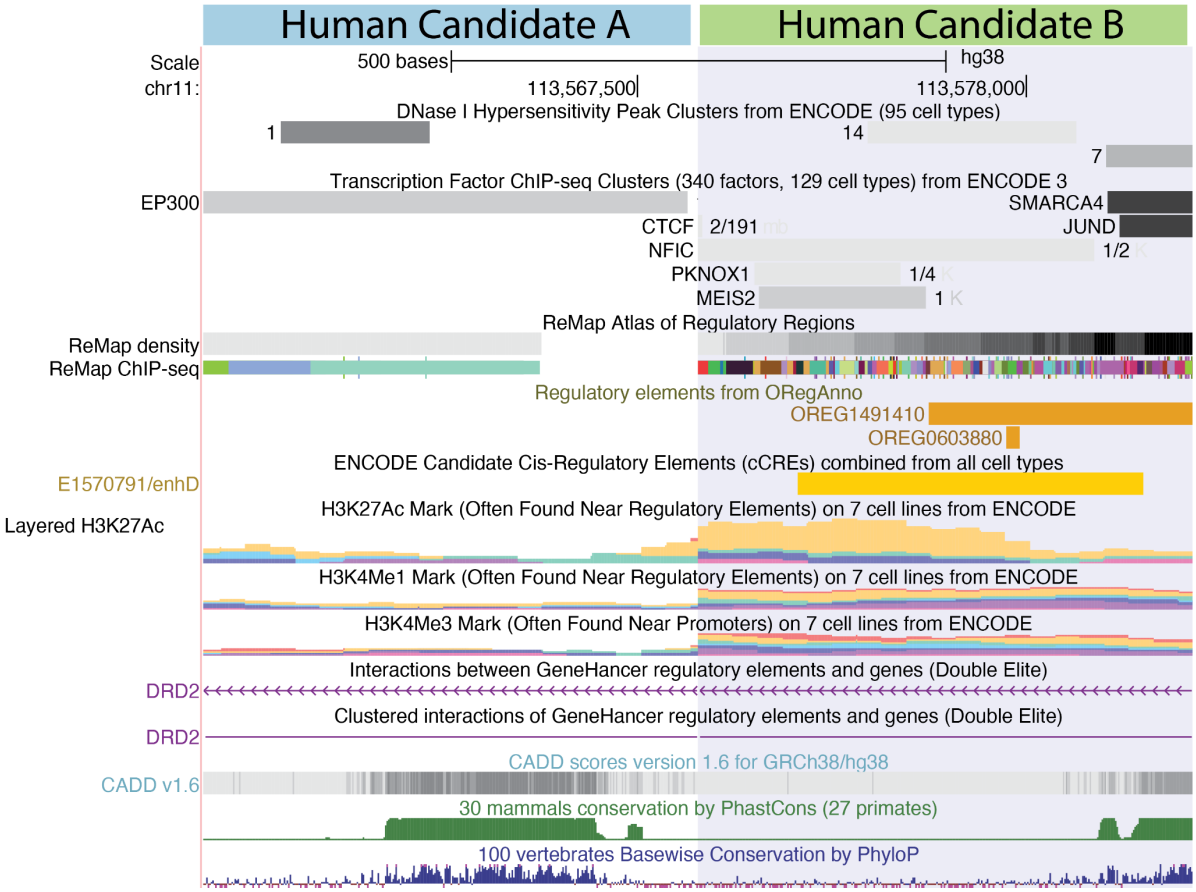


Figure S29. UCSC genome tracks for two candidate enhancer sequences upstream of the *DRD2* locus.

Genome tracks of human candidate enhancer A (hg38, chr11:113,567,061-113,567,561) and B (hg38, chr11:113,577,668-113,578,168) with epigenomic tracks available through the UCSC genome browser. The tracks from top to bottom are the ENCODE DNAse I hypersensitivity peaks, ENCODE3 ChIP-seq peaks, ORegAnno database annotations, ENCODE3 cCRE regions, ENCODE histone ChIP-seq multi-layer plots, GeneHancer links, CADD SNP impact predicted scores, and older nucleotide conservation tracks.

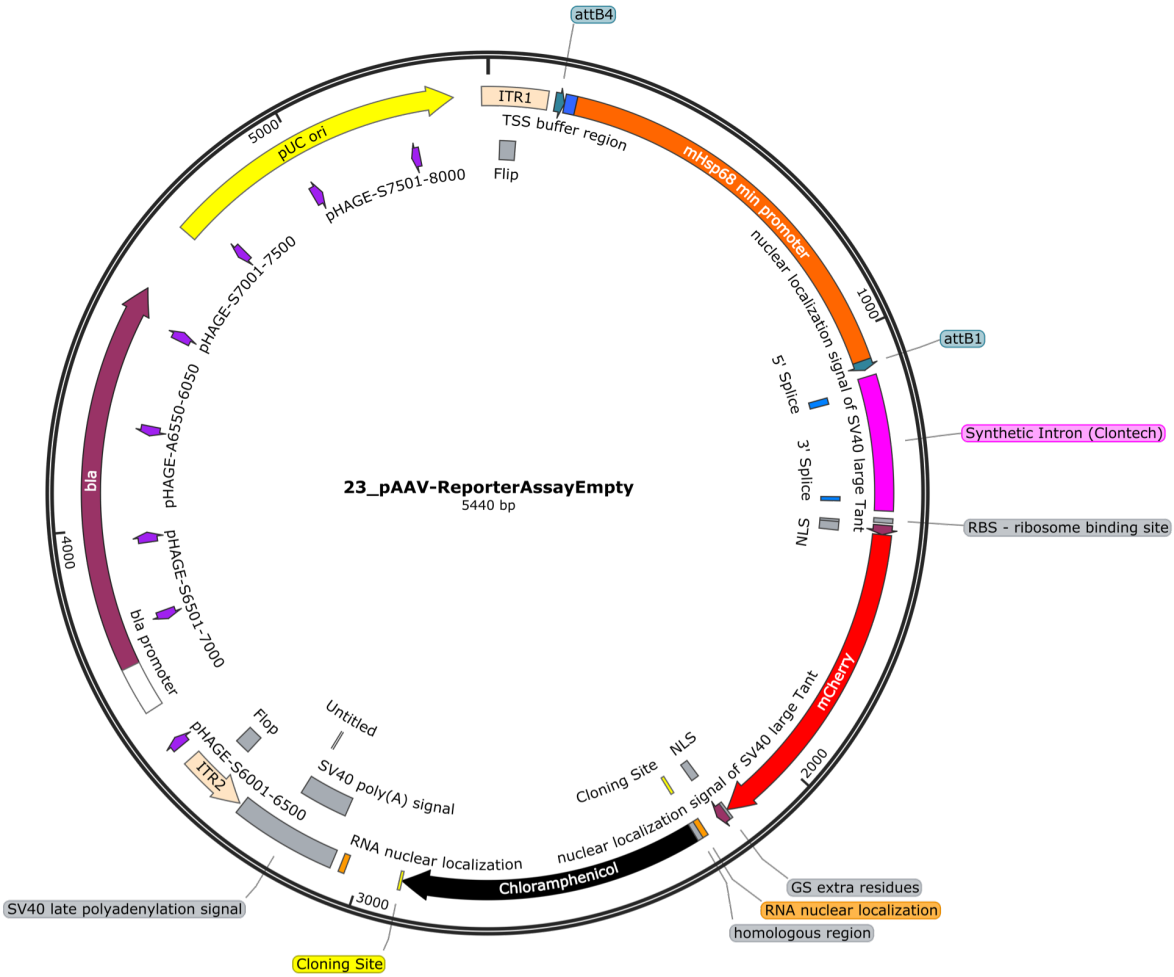


Figure S30. Plasmid template for adeno-associated virus-based *in vivo* reporter assays.

Plasmid map for mCherry reporter assay to measure transcriptional activation by candidate enhancer sequences. This plasmid contains inverted terminal repeat (ITR) compatible with adeno-associated virus (AAV) serotype 2-based packaging systems and capsids such as PhP.eB. This empty plasmid has the chloramphenicol resistance gene as the placeholder where candidate enhancers to be tested can be cloned in between multiple cloning sites. The plasmid contains the ampicillin resistance gene bla, the Hsp68 minimal promoter, an mRNA stabilizing synthetic intron, a mCherry protein open reading frame, two protein nuclear localization motifs, three mRNA nuclear localization motifs, and a polyA signal.

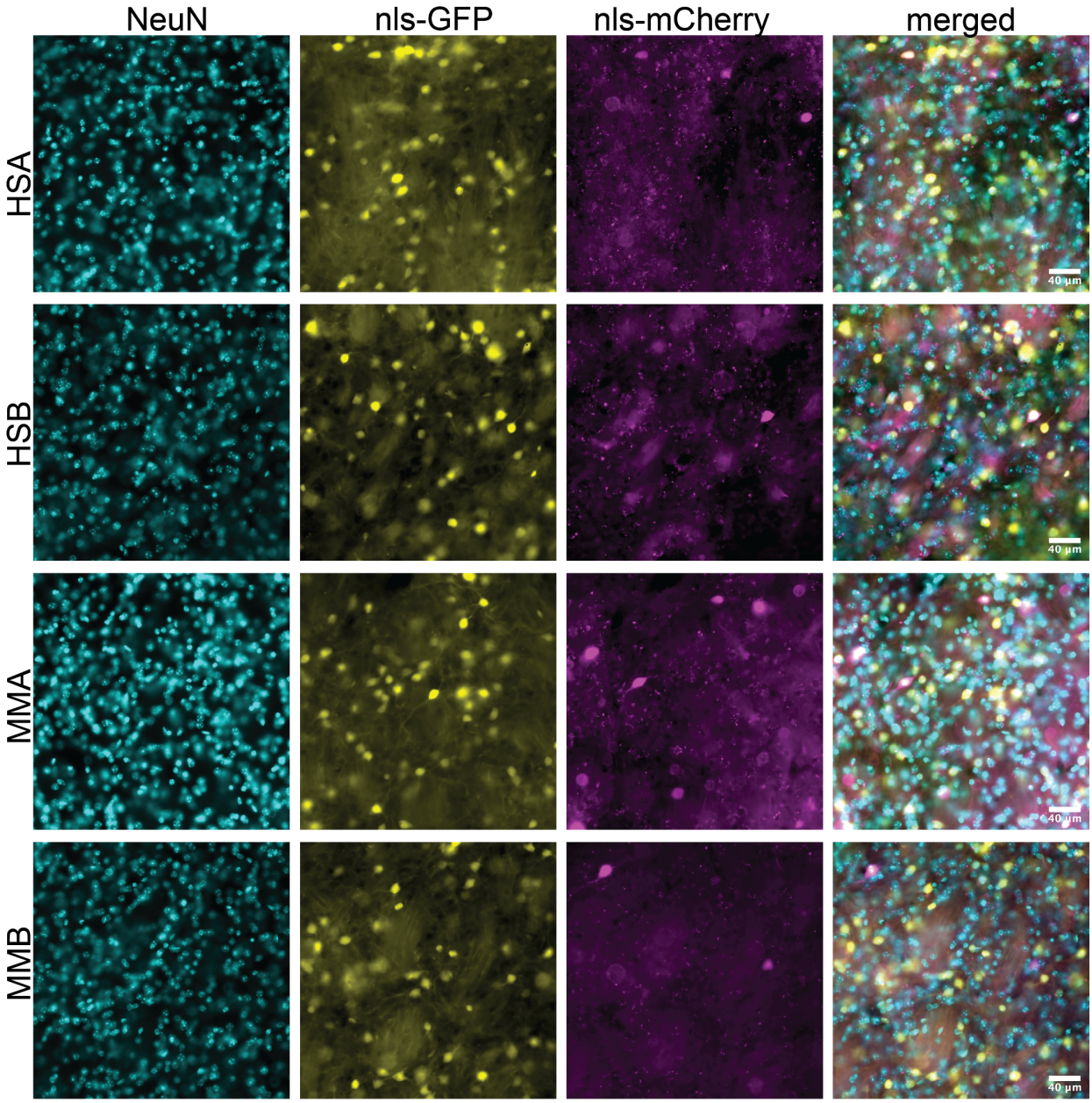


Figure S31. An *in vivo* reporter assay to test conserved cell type regulatory activity of candidate enhancers.

Multiplexed fluorescence micrographs acquired at 20X magnification of the mouse caudoputamen with injections of a transduction control virus delivering a nuclear localising (nls) GFP with the reporter assay virus delivering nls-mCherry under the control of a minimal promoter and the candidate enhancer sequence (A or B) from the human (HS) or mouse genomes (MM). The immunofluorescence from left to right label NeuN, GFP, mCherry, and the composite image.

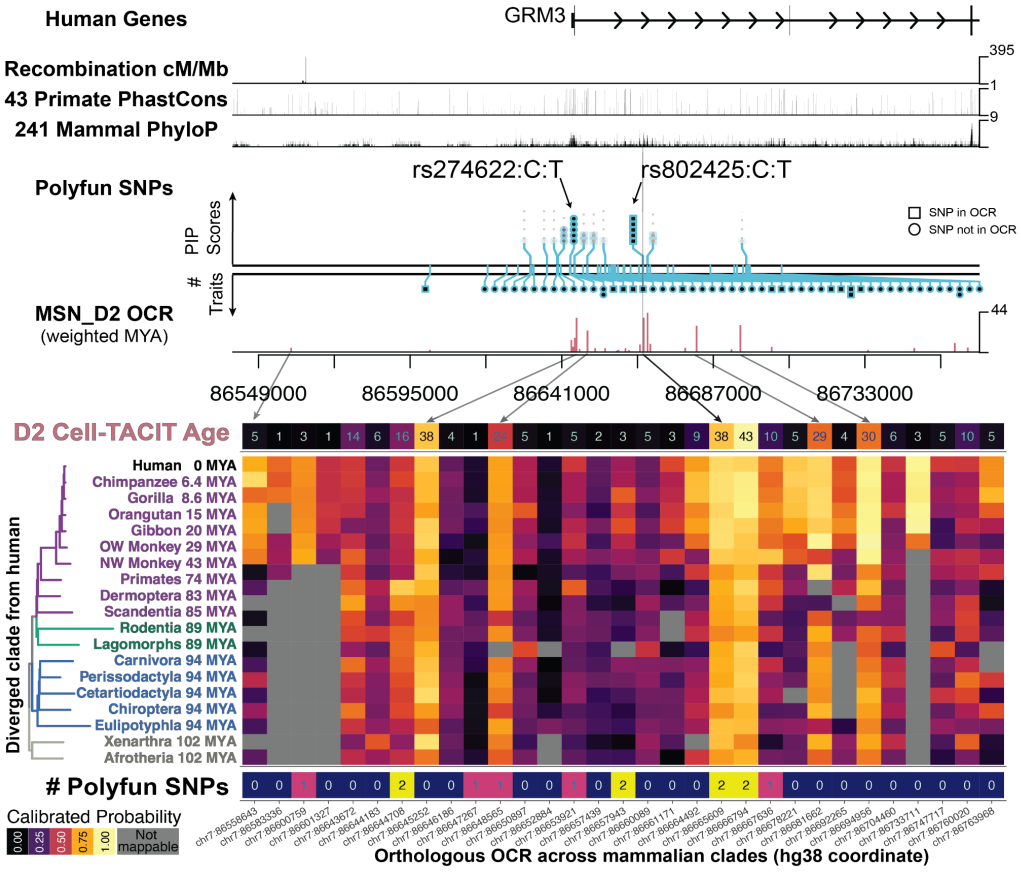


Figure S32. Cell-TACIT annotations prioritize fine-mapped SNPs of neuroticism to conserved, active D2 open chromatin regions.

Locus plot at the GRM3 gene locus with SNPs and Cell-TACIT annotations for D2 MSN models identifies candidate regulatory elements with fine-mapped SNPs from the neuroticism GWAS. Cell-TACIT Age (top row) heatmap for D2 MSNs summarizes the predicted activity of each open chromatin region in the track plot across increasingly distant clades of placental mammals (middle rows). The phylogenetic tree annotates the millions of years ago (MYA) that the common ancestor of each group of placental mammals and humans has diverged. Multiple D2 MSN open chromatin regions contain fine-mapped SNPs in this locus (bottom row). Lollipop plot feature height reflects the maximum PIP of the fine-mapped SNPs, and features are jittered to better visualize clusters of fine-mapped SNPs across a large genomic region. Mb: million bases. Arrows align cell type OCR tracks to corresponding columns in the heatmap. OCRs not mappable to a clade are plotted in gray.

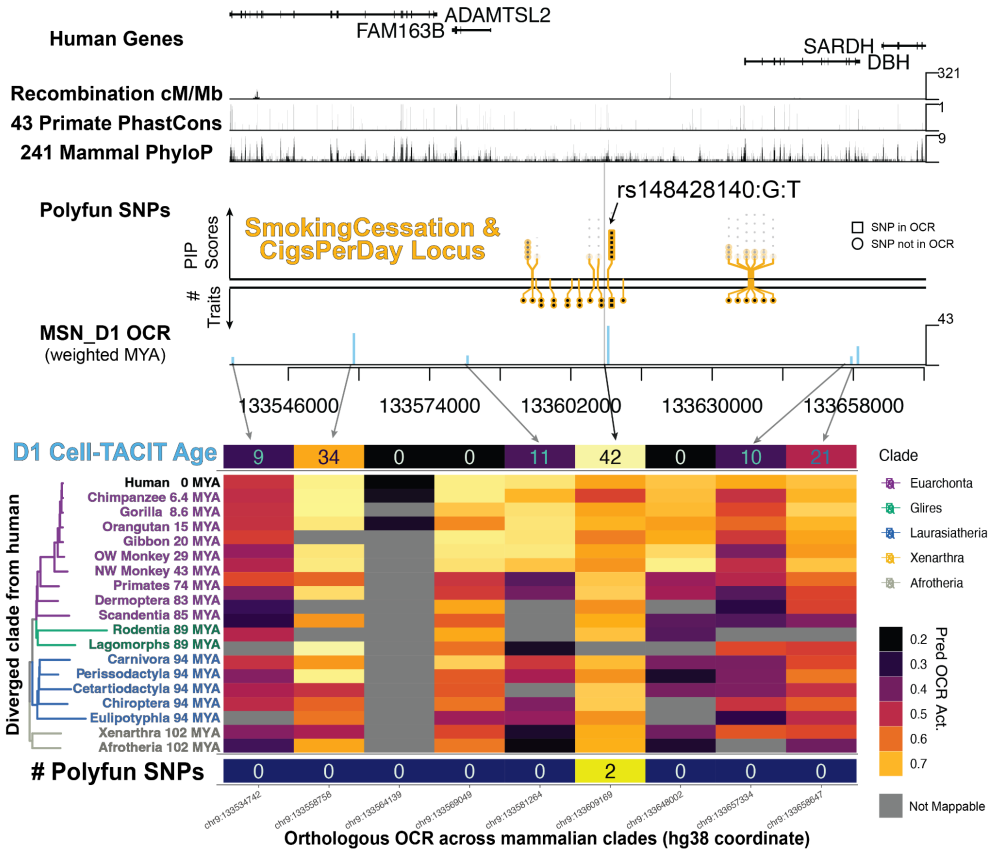


Figure S33. Cell-TACIT annotations prioritize fine-mapped SNPs of smoking use traits to conserved, active D1 MSN open chromatin regions.

Locus plot at a dense gene locus including ADAMTSL2 and other genes with SNPs and Cell-TACIT annotations for D1 MSN models identifies candidate regulatory elements with fine-mapped SNPs from smoking cessation and cigarettes per day GWAS. Cell-TACIT Age (top row) heatmap for D1 MSNs summarizes the predicted activity of each open chromatin region in the track plot across increasingly distant clades of placental mammals (middle rows). The phylogenetic tree annotates the millions of years ago (MYA) that the common ancestor of each group of placental mammals and humans has diverged. One D1 MSN open chromatin region contains fine-mapped SNPs in this locus (bottom row). Lollipop plot feature height reflects the maximum PIP of the fine-mapped SNPs, and features are jittered to better visualize clusters of fine-mapped SNPs across a large genomic region. Mb: million bases. Arrows align cell type OCR tracks to corresponding columns in the heatmap. OCRs not mappable to a clade are plotted in gray.

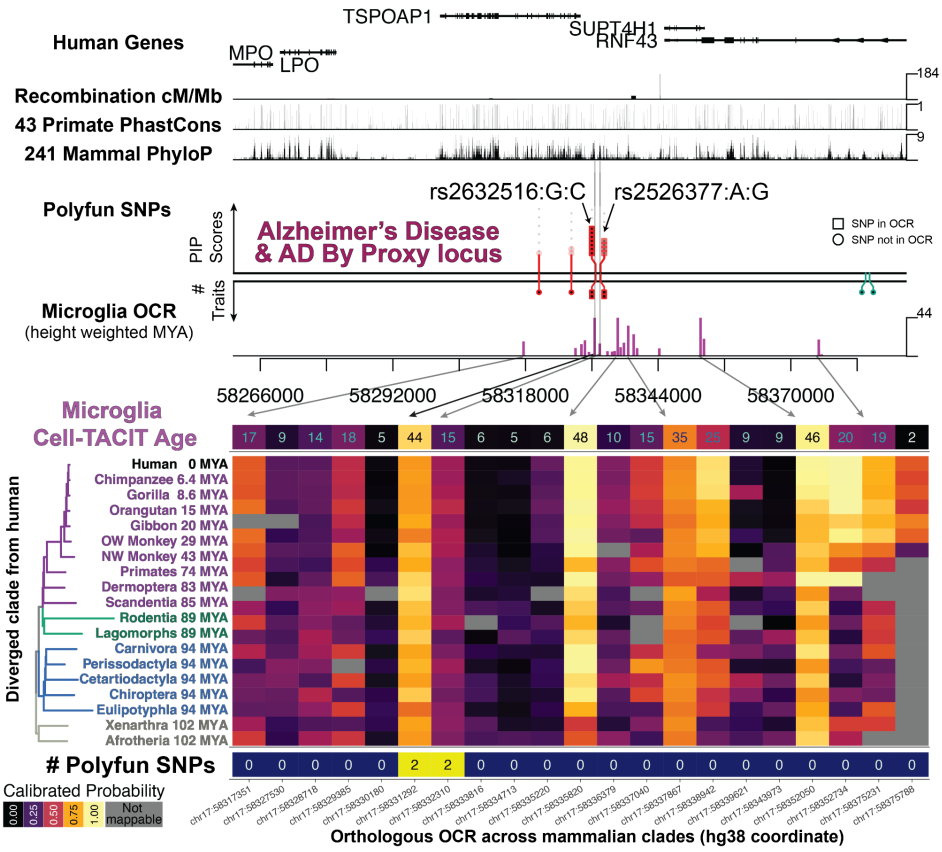


Figure S34. Cell-TACIT annotations prioritize fine-mapped SNPs of Alzheimer's Disease to conserved, active microglia open chromatin regions.

Locus plot at a dense gene locus including TSPOAP1 and other genes with SNPs and Cell-TACIT annotations for microglia models identifies candidate regulatory elements with fine-mapped SNPs from Alzheimer's Disease (AD) and AD-by-proxy GWAS. Cell-TACIT Age (top row) heatmap for microglia summarizes the predicted activity of each open chromatin region in the track plot across increasingly distant clades of placental mammals (middle rows). The phylogenetic tree annotates the millions of years ago (MYA) that the common ancestor of each group of placental mammals and humans has diverged. Two microglia open chromatin regions contain fine-mapped SNPs in this locus (bottom row). Lollipop plot feature height reflects the maximum PIP of the fine-mapped SNPs, and features are jittered to better visualize clusters of fine-mapped SNPs across a large genomic region. Mb: million bases. Arrows align cell type OCR tracks to corresponding columns in the heatmap. OCRs not mappable to a clade are plotted in gray.

Appendix C Chapter 3 Supplemental Data Legends

Data_S1_cross-species_meme-chip_motifs.xlsx

Table of MEME-chip motif enrichments computed in each cell type and source species from cell type OCR sequences before splitting into test, validation, and training sets for Cell-TACIT model training and evaluations.

Data_S2_Cell-TACIT_model_validation_performance_values.xlsx

Numerical values for model validation performance. ***Tabs are split by overall model validation set performance for overall positives and negatives, cell type-specific positives and negatives, and species-specific positives and negatives.***

Data_S3_Cell-TACIT_model_test_performance_values.xlsx

Numerical values for model test performance. ***Tabs are split by overall model test set performance for overall positives and negatives, cell type-specific positives and negatives, and species-specific positives and negatives.***

Data_S4_list_of_GWAS_traits.xlsx

List of genome-wide association studies (GWAS) analyzed in this work. The references for each GWAS are listed in shorthand in the data column reference and in Methods. The order that these traits appear in this list also dictate the traits that occur in ***Figure S11-26****. N_UKB indicates the GWAS contains significant overlap with the UK Biobank and the GWAS has a sample size as indicated for inclusion in polyfun fine-mapping. No entries indicated that the GWAS was not run in polyfun fine-mapping.*

Data_S5_ldsc_coefficient_caudate_celltype_cross-species.xlsx

Numerical data underlying the linear regression of the points plot in ***Figure S10*** *on the conditional independence effect sizes between OCR sets for each cell type. Term is the linear regression term used Methods. "model_" terms specifies the offset effect based on which*

genome an OCR set came from, mm10 or rheMac10 relative to hg38. Additional columns are the numerical data underlying ***Figure S10*** *from conditional independent LD score regression analyses across 64 GWAS traits, 8 cell types, and 7 cross-species open chromatin sets. The first tab describes the column names, and each tab is divided by GWAS trait in order as listed in* ***Data S4****. This orientation was selected for fast filtering and lookup by GWAS.*

Data_S6_Zoonomia_species.xlsx

The tabs in this data table list the species in the Zoonomia CACTUS, the taxonomic orders used for the 20 clade comparisons.

Data_S7_heritability_enrichments_Corces2020_phyloP_CellTACIT-Age.xlsx

Numerical data underlying ***Figure S11-26*** *from heritability enrichment LD score regression analyses across 64 GWAS traits, 8 cell types, and 26 cross-species open chromatin sets. The first tab describes the column names, and each tab is divided up by GWAS trait in order as listed in* ***Data S4****. This orientation was selected for fast filtering and lookup by GWAS. phyloP: mammalian phyloP constrained (phyloP.cons) or accelerated regions (phyloP.accel), gray bars.*

ATAC-seq: cell type OCRs measured in human, human and macaque, or human and mouse.

Cell-TACIT Score: human cell type OCRs mapabble to distant species or mappable and predicted active by Cell-TACIT machine learning models.

Cell-TACIT Age: distal human cell type OCRs binned into the youngest (0-25%) to oldest (75-100%) quartiles of being active across distantly related species.

Data_S8_Cell-TACIT-Age_eQTL_and_fine-mapped_SNPs.xlsx

Numerical outputs from logistic regression enrichment analyses of eQTLs or fine-mapped SNPs in Cell-TACIT Age quartiles. These columns are the numerical data for ***Figure*** *3A****-C, S28****.*

Data_S9_polyfun_caudate_finemapped_snps_20210518.xlsx

Output of functionally-informed fine-mapping by polyfun and SuSIE for 43 traits (41 returned SNPs). The first tab describes the column names, and each tab is divided up by GWAS trait in order as listed in ***Data S1****. This orientation was selected for fast filtering and lookup by GWAS.*

Data_S10_CellTACIT_Age_caudate_finemapped_snps_20220718.xlsx

Cell-TACIT Age annotations for polyfun fine-mapped SNPs from Data S5 if those SNPs overlapped non-exonic human cell type OCRs. Each tab is divided up by the Cell-TACIT age annotated across all traits and fine-mapped SNPs. This orientation was selected for fast filtering and lookup by GWAS. Additional columns here are "cell type" for the source Cell-TACIT Age model, "consFrac" is the proportion of the cell type OCR that has a Zoonomia phyloP score > 2.270, and the relevant "Cell-TACIT Age".

Data_S11_reporter_assay_CellTACIT-predictions.xlsx

Cell-TACIT Age and Score predictions for two candidate enhancers near the DRD2 locus. These columns are the numerical data for ***Figure 4C, D, F****.*

Data_S12_reporter_assay_stats_and_data.xlsx

Reporter assay segmentation outputs per image and statistical outputs underlying Figure 4E, F.

Appendix D Chapter 4 Extended Data Legend

Table S1. Comparison of quality control metrics by OUD diagnosis

False discovery rate-corrected t-tests comparing aggregated single nuclei RNA-se quality control metrics between UC and OUD individuals across the caudate or putamen.

Table S2. Comparison of proportion of cell types by OUD diagnosis

False discovery rate-corrected multiple mixed effect linear regression for any difference in cell type proportion between UC vs. OUD controlling for covariates as fixed effects and individual as a random effect.

Table S3. Significant marker genes of striatal cell types

Significant marker genes for each striatal cell type using the Seurat `FindMarkers()` *function with default parameters to compute one-vs-all comparison using Wilcoxon-rank sum test. We computed the FDR across all cell types and significant marker genes are reported at FDR < 0.05.*

Table S4. Significant marker TF-gene modules of striatal cell types

Significant marker TF-gene module activation for each striatal cell type using t-tests comparing the TF-gene module AUCell scores of each cell type with all others. The false discovery rate was computed across all cell types and significant marker genes are reported at FDR < 0.05.

Table S5. Postmortem individual-level and tissue-level metadata

Description of the biospecimen-level and individual-level metadata describing relevant covariates and opioid use disorder diagnoses.

Table S6. Differentially expressed genes in UC vs. OUD within striatal cells

Outputs of differential expression across striatal cells or the aggregate "meta cells", `All`, `Neuron`, and `Glia`. The differential expression FDR was calculated using the `swfdr` method to increase power by modeling power with an independent factor, average expression. We calculated FDR within each cell type comparison (`adj.P.Val.Within`) and between all cell type comparisons (`adj.P.Val.Between`) and report the more conservative correction between cell types in the main text and Figures, `adj.P.Val.Between`.

Table S7. Enriched pathways in UC vs. OUD, main comparison

Outputs of the gene-set pathway enrichment analyses using the main differential expression from UC vs. OUD. We calculated FDR across all cell type comparisons using `swfdr` modeling power by the number of genes in a gene set.

Table S8. Enriched TF-gene modules in UC vs. OUD

Outputs of differential pseudo-bulk average transcription factor-gene regulatory network activation in striatal cell types using linear regression to control for covariates.

Table S9. Clustered neuronal pathways in UC vs. OUD, main comparison

Neuronal pathways from Table S7 with additional clustering to identify related, enrich pathways. The cluster number corresponds to those in Figure 18B-C.

Table S10. Comparison of neuronal DNA damage score by OUD diagnosis

Outputs of differential pseudo-bulk average DNA damage score activation in striatal neurons using linear regressions to control for covariates.

Table S11. Clustered glial pathways in UC vs. OUD, main comparison

Glial pathways from Table S7 with additional clustering to identify related, enrich pathways. The cluster number corresponds to those in Figure 21B-C.

Table S12. Differentially expressed genes in UC vs. OUD subset to female or male

Similar outputs to Table S6 from differential expression across striatal cells or the aggregate "meta cells", `All`, `Neuron`, and `Glia`. These differential expression tables assess UC vs. OUD in female or male individuals (SexF and SexM, respectively).

Table S13. Enriched female-biased pathways in UC vs. OUD

Similar outputs to Table S7 showing gene-set pathway enrichment analyses using the differential expression from UC vs. OUD that were biased to female individuals with OUD.

Table S14. Enriched male-biased pathways in UC vs. OUD

Similar outputs to Table S7 showing gene-set pathway enrichment analyses using the differential expression from UC vs. OUD that were biased to male individuals with OUD.

Table S15. Enriched pathways with a sex interaction in UC vs. OUD

Similar outputs to Table S7 showing gene-set pathway enrichment analyses using the differential expression from UC vs. OUD that had a differential interaction score between DEGs in females vs. in male individuals.

Table S16. Enriched pathways from Piechota et al. 2010

Gene-set enrichment analysis using gene sets from Piechota et al. 2010, gene pathways after drug exposure in male mouse striatum, with differentially expressed genes.

Table S17. Detected transcription factor (TF)-gene relationships

The transcription factor-gene regulatory relationships identified using the pySCENIC protocol to build gene regulatory networks. These TF-gene relationships were identified across multiple runs and across multiple individuals with and without OUD.

Bibliography

1. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (DSM-5 (R))*. (American Psychiatric Association Publishing, 2013).

2. U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. 2021 National Survey of Drug Use and Health. *National Survey on Drug Use and Health (NSDUH)* https://www.samhsa.gov/data/release/2021-national-survey-drug-use-and-health-nsduh-releases (2022).

3. CDC/National Center for Health Statistics. Overdose Deaths In 2021 Increased Half as Much as in 2020 - But Are Still Up 15%. *www.cdc.gov* https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2022/202205.htm (2022).

4. Xu, J., Murphy, S., Kochanek, K. & Arias, E. *Mortality in the United States, 2021*. https://www.cdc.gov/nchs/products/databriefs/db456.htm (2022) doi:10.15620/cdc:122516.

5. Spanagel, R. Animal models of addiction. *Dialogues Clin. Neurosci.* **19**, 247–258 (2017).

6. Smith, M. A. Nonhuman animal models of substance use disorders: Translational value and utility to basic science. *Drug Alcohol Depend.* **206**, 107733 (2020).

7. Kuhn, B. N., Kalivas, P. W. & Bobadilla, A.-C. Understanding addiction using animal models. *Front. Behav. Neurosci.* **13**, 262 (2019).

8. Field, M. & Kersbergen, I. Are animal models of addiction useful? *Addiction* **115**, 6–12 (2020).

9. Prescott, C. A., Khoddam, R. & Arpawong, T. E. Genetic risk for substance abuse and addiction: Family and twin studies. *eLS* 1–11 Preprint at https://doi.org/10.1002/9780470015902.a0005230.pub2 (2016).

10. Agrawal, A. & Lynskey, M. T. Are there genetic influences on addiction: evidence from family, adoption and twin studies. *Addiction* **103**, 1069–1081 (2008).

11. Gelernter, J. & Polimanti, R. Genetics of substance use disorders in the era of big data. *Nat. Rev. Genet.* **22**, 712–729 (2021).

12. Deak, J. D. & Johnson, E. C. Genetics of substance use disorders: a review. *Psychol. Med.* **51**, 2189–2200 (2021).

13. Johnson, E. C. *et al.* A large-scale genome-wide association study meta-analysis of cannabis use disorder. *Lancet Psychiatry* **7**, 1032–1045 (2020).

14. Sanchez-Roige, S. *et al.* Genome-wide association study meta-analysis of the Alcohol Use Disorders Identification Test (AUDIT) in two population-based cohorts. *Am. J. Psychiatry* **176**, 107–118 (2019).

15. Toikumo, S. *et al.* Multi-ancestry meta-analysis of tobacco use disorders based on electronic health record data prioritizes novel candidate risk genes and reveals associations with numerous health outcomes. *medRxiv* (2023) doi:10.1101/2023.03.27.23287713.

16. Zhou, H. *et al.* Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nat. Neurosci.* **23**, 809–818 (2020).

17. Schlag, A. K. Percentages of problem drug use and their implications for policy making: A review of the literature. *Drug Sci. Policy Law* **6**, 205032452090454 (2020).

18. Barr, P. B. *et al.* Clinical, environmental, and genetic risk factors for substance use disorders: characterizing combined effects across multiple cohorts. *Mol. Psychiatry* **27**, 4633–4641 (2022).

19. Hatoum, A. S. *et al.* The addiction risk factor: A unitary genetic vulnerability characterizes substance use disorders and their associations with common correlates. *Neuropsychopharmacology* **47**, 1739–1745 (2022).

20. Hatoum, A. S. *et al.* Multivariate genome-wide association meta-analysis of over 1 million subjects identifies loci underlying multiple substance use disorders. *Nat. Ment. Health* **1**, 210–223 (2023).

21. Bogdan, R., Hatoum, A. S., Johnson, E. C. & Agrawal, A. The Genetically Informed Neurobiology of Addiction (GINA) model. *Nat. Rev. Neurosci.* **24**, 40–57 (2023).

22. García-Ortega, L. F. & Martínez, O. How many genes are expressed in a transcriptome? Estimation and results for RNA-seq. *PLoS One* **10**, e0130262 (2015).

23. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

24. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

25. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).

26. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nat. Commun.* **11**, 4307 (2020).

27. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: Recording the past and predicting the future. *Science* **358**, 69–75 (2017).

28. Consortium, G. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

29. Consortium, G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

30. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

31. Bryois, J. *et al.* Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat. Neurosci.* **25**, 1104–1112 (2022).

32. Luo, J. *et al.* Expression quantitative trait locus studies in the era of single-cell omics. *Front. Genet.* **14**, (2023).

33. Winter, J. J., Rodríguez-Acevedo, K. L., Dittrich, M. & Heller, E. A. Early life adversity: Epigenetic regulation underlying drug addiction susceptibility. *Mol. Cell. Neurosci.* **125**, 103825 (2023).

34. Tsankova, N., Renthal, W., Kumar, A. & Nestler, E. J. Epigenetic regulation in psychiatric disorders. *Nat. Rev. Neurosci.* **8**, 355–367 (2007).

35. Nestler, E. J. & Lüscher, C. The molecular basis of drug addiction: linking epigenetic to synaptic and circuit mechanisms. *Neuron* **102**, 48–59 (2019).

36. Bush, W. S. & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.* **8**, e1002822 (2012).

37. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

38. Berto, S., Liu, Y. & Konopka, G. Genomics at cellular resolution: insights into cognitive disorders and their evolution. *Hum. Mol. Genet.* **29**, R1–R9 (2020).

39. Haber, S. N. & Knutson, B. The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology* **35**, 4–26 (2010).

40. Balsters, J. H., Zerbi, V., Sallet, J., Wenderoth, N. & Mars, R. B. Primate homologs of mouse cortico-striatal circuits. *Elife* **9**, (2020).

41. Scaplen, K. M. & Kaun, K. R. Reward from bugs to bipeds: a comparative approach to understanding how reward circuits function. *J. Neurogenet.* **30**, 133–148 (2016).

42. Lowenstein, E. G. & Velazquez-Ulloa, N. A. A fly's eye view of natural and drug reward. *Front. Physiol.* **9**, 407 (2018).

43. Christmas, M. J. *et al.* Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023).

44. Damas, J. *et al.* Evolution of the ancestral mammalian karyotype and syntenic regions. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2209139119 (2022).

45. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).

46. Consortium, Z. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).

47. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

48. Andrews, G. *et al.* Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. *Science* **380**, eabn7930 (2023).

49. BRAIN Initiative Cell Census Network (BICCN). A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86–102 (2021).

50. Chen, D. *et al.* Single cell atlas for 11 non-model mammals, reptiles and birds. *Nat. Commun.* **12**, 7083 (2021).

51. Colquitt, B. M., Merullo, D. P., Konopka, G., Roberts, T. F. & Brainard, M. S. Cellular transcriptomics reveals evolutionary identities of songbird vocal circuits. *Science* **371**, eabd9704 (2021).

52. Park, S. H. E., Ortiz, A. K. & Konopka, G. Corticogenesis across species at single-cell resolution. *Dev. Neurobiol.* **82**, 517–532 (2022).

53. Miller, J. A. *et al.* Common cell type nomenclature for the mammalian brain. *Elife* **9**, (2020).

54. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).

55. Kriebel, A. R. & Welch, J. D. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nat. Commun.* **13**, 780 (2022).

56. Ataman, B. *et al.* Evolution of Osteocrin as an activity-regulated factor in the primate brain. *Nature* **539**, 242–247 (2016).

57. Phan, B. N. *et al.* A myelin-related transcriptomic profile is shared by Pitt-Hopkins syndrome models and human autism spectrum disorder. *Nat. Neurosci.* **23**, 375–385 (2020).

58. 2016 Alcohol, G. B. D. & Collaborators, D. U. The global burden of disease attributable to alcohol and drug use in 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet. Psychiatry* **5**, 987–1012 (2018).

59. Eddie, D., Greene, M. C., White, W. L. & Kelly, J. F. Medical burden of disease among individuals in recovery from alcohol and other drug problems in the united states: findings from the national recovery survey. *J. Addict. Med.* **13**, 385–395 (2019).

60. Pear, V. A. *et al.* Urban-rural variation in the socioeconomic determinants of opioid overdose. *Drug Alcohol Depend.* **195**, 66–73 (2019).

61. Pullen, E. & Oser, C. Barriers to substance abuse treatment in rural and urban communities: counselor perspectives. *Subst. Use Misuse* **49**, 891–901 (2014).

62. Grillner, S. & Robertson, B. The basal ganglia over 500 million years. *Curr. Biol.* **26**, R1088–R1100 (2016).

63. Koob, G. F. & Volkow, N. D. Neurocircuitry of addiction. *Neuropsychopharmacology* **35**, 217–238 (2010).

64. Goldstein, R. Z. & Volkow, N. D. Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. *Nat. Rev. Neurosci.* **12**, 652–669 (2011).

65. Pasman, J. A. *et al.* GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal influence of schizophrenia. *Nat. Neurosci.* **21**, 1161–1170 (2018).

66. Karlsson Linnér, R. *et al.* Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* **51**, 245–257 (2019).

67. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).

68. Erzurumluoglu, A. M. *et al.* Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. *Mol. Psychiatry* **25**, 2392–2409 (2020).

69. Dick, D. M. The genetics of addiction: where do we go from here? *J. Stud. Alcohol Drugs* **77**, 673–675 (2016).

70. Kendler, K. S. & Prescott, C. A. Cannabis use, abuse, and dependence in a population-based sample of female twins. *Am. J. Psychiatry* **155**, 1016–1022 (1998).

71. Kendler, K. S. & Prescott, C. A. Cocaine use, abuse and dependence in a population-based sample of female twins. *Br. J. Psychiatry* **173**, 345–350 (1998).

72. Waaktaar, T., Kan, K.-J. & Torgersen, S. The genetic and environmental architecture of substance use development from early adolescence into young adulthood: a longitudinal twin study of comorbidity of alcohol, tobacco and illicit drug use. *Addiction* **113**, 740–748 (2018).

73. Jensen, K. P. A review of genome-wide association studies of stimulant and opioid use disorders. *Mol. Neuropsychiatry* **2**, 37–45 (2016).

74. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).

75. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

76. Visscher, P. M. *et al.* 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

77. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).

78. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin* **8**, 57 (2015).

79. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

80. Mo, A. *et al.* Epigenomic signatures of neuronal diversity in the mammalian brain. *Neuron* **86**, 1369–1384 (2015).

81. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10**, e1003711 (2014).

82. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).

83. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).

84. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).

85. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).

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

87. Fullard, J. F. *et al.* An atlas of chromatin accessibility in the adult human brain. *Genome Res.* **28**, 1243–1252 (2018).

88. Corces, M. R. *et al.* Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020).

89. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).

90. Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011).

91. Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).

92. Zhang, X., Kaplow, I. M., Wirthlin, M., Park, T. Y. & Pfenning, A. R. HALPER facilitates the identification of regulatory element orthologs across species. *Bioinformatics* **36**, 4339–4340 (2020).

93. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

94. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).

95. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

96. Cannon, M. E. *et al.* Open chromatin profiling in adipose tissue marks genomic regions with functional roles in cardiometabolic traits. *G3 (Bethesda)* **9**, 2521–2533 (2019).

97. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).

98. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

99. Liu, C. *et al.* An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Scientific data* **6**, 65 (2019).

100. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1-21.29.9 (2015).

101. Chan, K. Y. *et al.* Engineered AAVs for efficient noninvasive gene delivery to the central and peripheral nervous systems. *Nat. Neurosci.* **20**, 1172–1179 (2017).

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

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

104. Ramírez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).

105. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, (2018).

106. Bulik-Sullivan, B. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

107. Johnston, K. J. A. *et al.* Genome-wide association study of multisite chronic pain in UK Biobank. *PLoS Genet.* **15**, e1008164 (2019).

108. Coffee *et al.* Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Mol. Psychiatry* **20**, 647–656 (2015).

109. Cheng, Z. *et al.* Genome-wide Association Study Identifies a Regulatory Variant of RGMA Associated With Opioid Dependence in European Americans. *Biol. Psychiatry* **84**, 762–770 (2018).

110. Cabana-Domínguez, J., Shivalikanjli, A., Fernàndez-Castillo, N. & Cormand, B. Genome-wide association meta-analysis of cocaine dependence: Shared genetics with comorbid conditions. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **94**, 109667 (2019).

111. (iocdf-Gc), I. O. C. D. F. G. C. & (ocgas), O. C. G. A. S. Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Mol. Psychiatry* **23**, 1181–1188 (2018).

112. the Psychiatric Genomics Consortium, S. W. G. of. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

113. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

114. Dashti, H. S. *et al.* Genome-wide association study identifies genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates. *Nat. Commun.* **10**, 1100 (2019).

115. Zillikens, M. C. *et al.* Large meta-analysis of genome-wide association studies identifies five loci for lean body mass. *Nat. Commun.* **8**, 80 (2017).

116. Kemp, J. P. *et al.* Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat. Genet.* **49**, 1468–1475 (2017).

117. Howson, J. M. M. *et al.* Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nat. Genet.* **49**, 1113–1119 (2017).

118. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

119. Khan, A., Riudavets Puig, R., Boddie, P. & Mathelier, A. BiasAway: command-line and web server to generate nucleotide composition-matched DNA background sequences. Preprint at https://biasaway.uio.no (2020).

120. Worsley Hunt, R., Mathelier, A., Del Peso, L. & Wasserman, W. W. Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics* **15**, 472 (2014).

121. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

122. Consortium, 1000 Genomes Project *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

123. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

124. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

125. Boca, S. M. & Leek, J. T. A direct approach to estimating false discovery rates conditional on covariates. *PeerJ* **6**, e6035 (2018).

126. Korthauer, K. *et al.* A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.* **20**, 1–21 (2019).

127. Wang, J. *et al.* CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies. *Nucleic Acids Res.* **48**, D807–D816 (2020).

128. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

129. Chen, W. *et al.* Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**, 719–736 (2015).

130. Kichaev, G. *et al.* Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33**, 248–255 (2017).

131. Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **41**, 647–665 (2014).

132. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. *arXiv [cs.CV]* (2017).

133. Shrikumar, A. *et al.* Technical note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. *arXiv [cs.LG]* (2018).

134. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).

135. Zeng, X. *et al.* Association of bone mineral density with lung function in a Chinese general population: the Xinxiang rural cohort study. *BMC Pulm. Med.* **19**, 239 (2019).

136. Lee, I. S. *et al.* Relationship between pulmonary function and bone mineral density in the Korean National Health and Nutrition Examination Survey. *Korean J. Intern. Med.* **31**, 899–909 (2016).

137. Kim, C. *et al.* Differences in bone density on chest CT according to smoking status in males without chronic obstructive lung disease. *Sci. Rep.* **9**, 10467 (2019).

138. Chen, L., Fish, A. E. & Capra, J. A. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLoS Comput. Biol.* **14**, e1006484 (2018).

139. Lefort, S., Tomm, C., Floyd Sarria, J.-C. & Petersen, C. C. H. The excitatory neuronal network of the C2 barrel column in mouse primary somatosensory cortex. *Neuron* **61**, 301–316 (2009).

140. Beaulieu, C. Numerical data on neocortical neurons in adult rat, with special reference to the GABA population. *Brain Res.* **609**, 284–292 (1993).

141. Minnoye, L. *et al.* Cross-species analysis of enhancer logic using deep learning. *Genome Res.* **30**, 1815–1834 (2020).

142. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).

143. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).

144. Barman, P., Reddy, D. & Bhaumik, S. R. Mechanisms of Antisense Transcription Initiation with Implications in Gene Expression, Genomic Integrity and Disease Pathogenesis. *Non-coding RNA* **5**, (2019).

145. Pelechano, V. & Steinmetz, L. M. Gene regulation by antisense transcription. *Nat. Rev. Genet.* **14**, 880–893 (2013).

146. Xu, Z. *et al.* SP8 and SP9 coordinately promote D2-type medium spiny neuron production by activating Six3 expression. *Development* **145**, (2018).

147. Hoffman, J. L. *et al.* Alcohol drinking exacerbates neural and behavioral pathology in the 3xTg-AD mouse model of Alzheimer's disease. *Int. Rev. Neurobiol.* **148**, 169–230 (2019).

148. Ramamurthy, E. *et al.* Cell type-specific histone acetylation profiling of Alzheimer's disease subjects and integration with genetics. *Front. Mol. Neurosci.* **15**, 948456 (2022).

149. Volkow, N. D., Wang, G.-J., Tomasi, D. & Baler, R. D. Unbalanced neuronal circuits in addiction. *Curr. Opin. Neurobiol.* **23**, 639–648 (2013).

150. Koob, G. F. & Volkow, N. D. Neurobiology of addiction: a neurocircuitry analysis. *The lancet. Psychiatry* **3**, 760 (2016).

151. Volkow, N. D. & Morales, M. The brain on drugs: from reward to addiction. *Cell* **162**, 712–725 (2015).

152. Fehr, C. *et al.* Association of low striatal dopamine d2 receptor availability with nicotine dependence similar to that seen with other drugs of abuse. *Am. J. Psychiatry* **165**, 507–514 (2008).

153. Volkow, N. D. *et al.* Low level of brain dopamine d2 receptors in methamphetamine abusers: association with metabolism in the orbitofrontal cortex. *Focus* **1**, 150–157 (2003).

154. Volkow, N. D. *et al.* Decreased striatal dopaminergic responsiveness in detoxified cocaine-dependent subjects. *Nature* **386**, 830–833 (1997).

155. Wang, G. J. *et al.* Dopamine D2 receptor availability in opiate-dependent subjects before and after naloxone-precipitated withdrawal. *Neuropsychopharmacology* **16**, 174–182 (1997).

156. Volkow, N. D. *et al.* Decreases in dopamine receptors but not in dopamine transporters in alcoholics. *Alcohol. Clin. Exp. Res.* **20**, 1594–1598 (1996).

157. Cheng, Y. *et al.* Distinct synaptic strengthening of the striatal direct and indirect pathways drives alcohol consumption. *Biol. Psychiatry* **81**, 918–929 (2017).

158. Wilar, G., Shinoda, Y., Sasaoka, T. & Fukunaga, K. Crucial Role of Dopamine D2 Receptor Signaling in Nicotine-Induced Conditioned Place Preference. *Mol. Neurobiol.* **56**, 7911–7928 (2019).

159. Ferguson, S. M. *et al.* Transient neuronal inhibition reveals opposing roles of indirect and direct pathways in sensitization. *Nat. Neurosci.* **14**, 22–24 (2011).

160. Ji, X. *et al.* Dopamine Receptors Differentially Control Binge Alcohol Drinking-Mediated Synaptic Plasticity of the Core Nucleus Accumbens Direct and Indirect Pathways. *J. Neurosci.* **37**, 5463–5474 (2017).

161. Seney, M. L. *et al.* Transcriptional Alterations in Dorsolateral Prefrontal Cortex and Nucleus Accumbens Implicate Neuroinflammation and Synaptic Remodeling in Opioid Use Disorder. *Biol. Psychiatry* **90**, (2021).

162. Yao, Z. *et al.* A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell* **184**, 3222-3241.e26 (2021).

163. Tepper, J. M. & Koós, T. Gabaergic interneurons of the striatum. in *Handbook of basal ganglia structure and function, second edition* vol. 24 157–178 (Elsevier, 2017).

164. Krienen, F. M. *et al.* Innovations present in the primate interneuron repertoire. *Nature* **586**, 262–269 (2020).

165. Jiang, C. *et al.* Morphine coordinates \SST and PV interneurons in the prelimbic cortex to disinhibit pyramidal neurons and enhance reward. *Mol. Psychiatry* (2019) doi:10.1038/s41380-019-0480-7.

166. Schall, T. A., Wright, W. J. & Dong, Y. Nucleus accumbens fast-spiking interneurons in motivational and addictive behaviors. *Mol. Psychiatry* (2020) doi:10.1038/s41380-020-0683-y.

167. Lee, J. H. *et al.* Dopaminergic regulation of nucleus accumbens cholinergic interneurons demarcates susceptibility to cocaine addiction. *Biol. Psychiatry* **88**, 746–757 (2020).

168. Bracci, E., Centonze, D., Bernardi, G. & Calabresi, P. Dopamine excites fast-spiking interneurons in the striatum. *J. Neurophysiol.* **87**, 2190–2194 (2002).

169. Lansink, C. S., Goltstein, P. M., Lankelma, J. V. & Pennartz, C. M. A. Fast-spiking interneurons of the rat ventral striatum: temporal coordination of activity with principal cells and responsiveness to reward. *Eur. J. Neurosci.* **32**, 494–508 (2010).

170. Wiltschko, A. B., Pettibone, J. R. & Berke, J. D. Opposite effects of stimulant and antipsychotic drugs on striatal fast-spiking interneurons. *Neuropsychopharmacology* **35**, 1261–1270 (2010).

171. Ribeiro, E. A. *et al.* Transcriptional and physiological adaptations in nucleus accumbens somatostatin interneurons that regulate behavioral responses to cocaine. *Nat. Commun.* **9**, 3149 (2018).

172. Farrell, M. R., Schoch, H. & Mahler, S. V. Modeling cocaine relapse in rodents: Behavioral considerations and circuit mechanisms. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **87**, 33–47 (2018).

173. Berke, J. D. & Hyman, S. E. Addiction, dopamine, and the molecular mechanisms of memory. *Neuron* **25**, 515–532 (2000).

174. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518**, 365–369 (2015).

175. Halow, J. M. *et al.* Tissue context determines the penetrance of regulatory DNA variation. *Nat. Commun.* **12**, 2850 (2021).

176. Sullivan, P. F. *et al.* Psychiatric Genomics: An Update and an Agenda. *Am. J. Psychiatry* **175**, 15 (2018).

177. Sullivan, P. & Zoonomia Consortium. Evolutionary constraint uniquely informs human disease mechanism.

178. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).

179. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).

180. Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).

181. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).

182. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

183. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, (2010).

184. Sullivan, P. F. *et al.* Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023).

185. Kaplow, I. M. *et al.* Inferring mammalian tissue-specific regulatory conservation by predicting tissue-specific differences in open chromatin. *BMC Genomics* **23**, (2022).

186. Li, Y. E. *et al.* An atlas of gene regulatory elements in adult mouse cerebrum. *Nature* **598**, 129–136 (2021).

187. Srinivasan, C. *et al.* Addiction-associated genetic variants implicate brain cell type- and region-specific cis-regulatory elements in addiction neurobiology. *J. Neurosci.* (2021) doi:10.1523/JNEUROSCI.2534-20.2021.

188. Pan, Z. *et al.* Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nat. Commun.* **12**, 5848 (2021).

189. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324.e18 (2018).

190. Hujoel, M. L. A., Gazal, S., Hormozdiari, F., van de Geijn, B. & Price, A. L. Disease heritability enrichment of regulatory elements is concentrated in elements with ancient sequence age and conserved function across species. Preprint at https://doi.org/10.1101/420166.

191. Dey, K. K. *et al.* Evaluating the informativeness of deep learning annotations for human complex diseases. *Nat. Commun.* **11**, 4703 (2020).

192. van de Geijn, B. *et al.* Annotations capturing cell type-specific TF binding explain a large fraction of disease heritability. *Hum. Mol. Genet.* **29**, 1057–1067 (2020).

193. Wong, E. S. *et al.* Deep conservation of the enhancer regulatory code in animals. *Science* **370**, (2020).

194. Kaplow IM *et al.* Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. *Science* **Apr 28;380**, eabm7993 (2023).

195. Huo, Y., Li, S., Liu, J., Li, X. & Luo, X.-J. Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. *Nat. Commun.* **10**, 670 (2019).

196. Bhattacharyya, U., Bhatia, T., Deshpande, S. N. & Thelma, B. K. Genetic variations in evolutionary accelerated regions disrupt cognition in schizophrenia. *Psychiatry Res.* **314**, 114586 (2022).

197. Srinivasan, S. *et al.* Probing the Association between Early Evolutionary Markers and Schizophrenia. *PLoS One* **12**, e0169227 (2017).

198. Banerjee, N. *et al.* Recently evolved human-specific methylated regions are enriched in schizophrenia signals. *BMC Evol. Biol.* **18**, 63 (2018).

199. Skene, N. G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).

200. Bryois, J. *et al.* Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nat. Genet.* **52**, 482–493 (2020).

201. Consortium, T. S. W. G. of T. P. G., The Schizophrenia Working Group of the Psychiatric Genomics Consortium, Ripke, S., Walters, J. T. R. & O'Donovan, M. C. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *Nature* (2022) doi:10.1038/s41586-022-04434-5.

202. Grillner, S., Robertson, B. & Stephenson-Jones, M. The evolutionary origin of the vertebrate basal ganglia and its role in action selection. *J. Physiol.* **591**, 5425–5431 (2013).

203. Jansen, P. R. *et al.* Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat. Genet.* **51**, 394–403 (2019).

204. Gokce, O. *et al.* Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq. *Cell Rep.* **16**, 1126–1137 (2016).

205. Stanley, G., Gokce, O., Malenka, R. C., Südhof, T. C. & Quake, S. R. Continuous and Discrete Neuron Types of the Adult Murine Striatum. *Neuron* **105**, 688-699.e8 (2020).

206. He, J. *et al.* Transcriptional and anatomical diversity of medium spiny neurons in the primate striatum. *Curr. Biol.* (2021) doi:10.1016/j.cub.2021.10.015.

207. Genereux, D. P. *et al.* A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).

208. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).

209. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).

210. Consortium, T. G. & The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* vol. 369 1318–1330 Preprint at https://doi.org/10.1126/science.aaz1776 (2020).

211. Bryois, J. *et al.* Cell-type specific cis-eQTLs in eight brain cell-types identifies novel risk genes for human brain disorders. Preprint at https://doi.org/10.1101/2021.10.09.21264604.

212. Herrero-Navarro, Á. *et al.* Astrocytes and neurons share region-specific transcriptional signatures that confer regional identity to neuronal reprogramming. *Science Advances* vol. 7 Preprint at https://doi.org/10.1126/sciadv.abe8978 (2021).

213. Chai, H. *et al.* Neural Circuit-Specialized Astrocytes: Transcriptomic, Proteomic, Morphological, and Functional Evidence. *Neuron* **95**, 531-549.e9 (2017).

214. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).

215. Sullivan, P. F. & Zoonomia Consortium. Leveraging Base Pair Mammalian Constraint to Understand Genetic Variation and Human Disease.

216. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. Preprint at https://doi.org/10.1101/2022.05.07.491045.

217. Glassberg, E. C., Gao, Z., Harpak, A., Lan, X. & Pritchard, J. K. Evidence for Weak Selective Constraint on Human Gene Expression. *Genetics* **211**, 757–772 (2019).

218. Nagel, M. *et al.* Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet.* **50**, 920–927 (2018).

219. Hirvonen, J. *et al.* Increased caudate dopamine D2 receptor availability as a genetic marker for schizophrenia. *Arch. Gen. Psychiatry* **62**, 371–378 (2005).

220. Abi-Dargham, A. *et al.* Increased baseline occupancy of D2 receptors by dopamine in schizophrenia. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 8104–8109 (2000).

221. Mathalon, D. H. & Ford, J. M. Neurobiology of schizophrenia: search for the elusive correlation with symptoms. *Front. Hum. Neurosci.* **6**, 136 (2012).

222. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).

223. Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.* **16**, e1008050 (2020).

224. Roller, M. *et al.* LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. *Genome Biol.* **22**, 62 (2021).

225. Latini, A. *et al.* Altered expression of miR-142, miR-155, miR-499a and of their putative common target in systemic lupus erythematosus. *Epigenomics* **13**, 5–13 (2021).

226. Ghanbari, M. *et al.* A functional variant in the miR-142 promoter modulating its expression and conferring risk of Alzheimer disease. *Hum. Mutat.* **40**, 2131–2145 (2019).

227. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).

228. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* **17**, (2020).

229. Lawler, A. J. *et al.* Machine learning sequence prioritization for cell type-specific enhancer design. *Elife* **11**, (2022).

230. Gemberling, M. P. *et al.* Transgenic mice for in vivo epigenome editing with CRISPR-based systems. *Nat. Methods* **18**, 965–974 (2021).

231. Stojanovic, T. *et al.* Validation of dopamine receptor DRD1 and DRD2 antibodies using receptor deficient mice. *Amino Acids* **49**, 1101–1109 (2017).

232. Graveland, G. A. & DiFiglia, M. The frequency and distribution of medium-sized neurons with indented nuclei in the primate and rodent neostriatum. *Brain Res.* **327**, (1985).

233. Ren, K. *et al.* Striatal Distribution and Cytoarchitecture of Dopamine Receptor Subtype 1 and 2: Evidence from Double-Labeling Transgenic Mice. *Front. Neural Circuits* **11**, 57 (2017).

234. Gerfen, C. R., Paletzki, R. & Heintz, N. GENSAT BAC Cre-Recombinase Driver Lines to Study the Functional Organization of Cerebral Cortical and Basal Ganglia Circuits. *Neuron* vol. 80 1368–1383 Preprint at https://doi.org/10.1016/j.neuron.2013.10.016 (2013).

235. Öztürk, B. E. *et al.* scAAVengr, a transcriptome-based pipeline for quantitative ranking of engineered AAVs with single-cell resolution. *Elife* **10**, (2021).

236. Hrvatin, S. *et al.* A scalable platform for the development of cell-type-specific viral drivers. *Elife* **8**, (2019).

237. Davenport, A. T., Grant, K. A., Szeliga, K. T., Friedman, D. P. & Daunais, J. B. Standardized method for the harvest of nonhuman primate tissue optimized for multiple modes of analyses. *Cell Tissue Bank.* **15**, 99–110 (2014).

238. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).

239. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* vol. 17 10 Preprint at https://doi.org/10.14806/ej.17.1.200 (2011).

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

241. Warren, W. C. *et al.* Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science* **370**, (2020).

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

243. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa1016.

244. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

245. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* vol. 177 1888-1902.e21 Preprint at https://doi.org/10.1016/j.cell.2019.05.031 (2019).

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

247. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, (2006).

248. Zanta, M. A., Belguise-Valladier, P. & Behr, J.-P. Gene delivery: A single nuclear localization signal peptide is sufficient to carry DNA to the cell nucleus. *Proceedings of the National Academy of Sciences* vol. 96 91–96 Preprint at https://doi.org/10.1073/pnas.96.1.91 (1999).

249. Zhang, B. *et al.* A Novel RNA Motif Mediates the Strict Nuclear Localization of a Long Noncoding RNA. *Mol. Cell. Biol.* **34**, 2318 (2014).

250. Lawler, A. J. *et al.* Cell Type-Specific Oxidative Stress Genomic Signatures in the Globus Pallidus of Dopamine-Depleted Mice. *J. Neurosci.* **40**, 9772–9783 (2020).

251. Wirthlin, Morgan, Kaplow, Irene M., Lawler, Alyssa J., He, J., Phan., BaDoi, N., Brown, Ashley R., Stauffer, William R., Pfenning, Andreas R. The Regulatory Evolution of the Primate Fine-Motor System. (2020).

252. Wirthlin, M. E. *et al.* Vocal learning-associated convergent evolution in mammalian regulatory elements and proteins. (2021).

253. Halstead, M. M. *et al.* A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. *BMC Genomics* **21**, 698 (2020).

254. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 11484–11489 (2003).

255. van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).

256. Fritsche, L. G. *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.* **48**, 134–143 (2016).

257. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).

258. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).

259. Watson, H. J. *et al.* Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nat. Genet.* **51**, 1207–1214 (2019).

260. Mullins, N. *et al.* Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat. Genet.* **53**, 817–829 (2021).

261. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).

262. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).

263. Nievergelt, C. M. *et al.* International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci. *Nat. Commun.* **10**, 4558 (2019).

264. Cross-Disorder Group of the Psychiatric Genomics Consortium. Electronic address: plee0@mgh.harvard.edu & Cross-Disorder Group of the Psychiatric Genomics Consortium. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* **179**, 1469-1482.e11 (2019).

265. Polimanti, R. *et al.* Leveraging genome-wide data to investigate differences between opioid use vs. opioid dependence in 41,176 individuals from the Psychiatric Genomics Consortium. Preprint at https://doi.org/10.1101/765065.

266. Cabana-Domínguez, J., Shivalikanjli, A., Fernàndez-Castillo, N. & Cormand, B. Genome-wide association meta-analysis of cocaine dependence: Shared genetics with comorbid conditions. Preprint at https://doi.org/10.1101/374553.

267. Walters, R. K. *et al.* Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat. Neurosci.* **21**, 1656–1669 (2018).

268. Hill, W. D. *et al.* Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income. *Nat. Commun.* **10**, 5741 (2019).

269. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).

270. Jansen, P. R. *et al.* Genome-wide meta-analysis of brain volume identifies genomic loci and genes shared with intelligence. *Nat. Commun.* **11**, 5606 (2020).

271. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).

272. Meng, W. *et al.* A Genome-Wide Association Study Finds Genetic Associations with Broadly-Defined Headache in UK Biobank (N=223,773). *EBioMedicine* **28**, 180–186 (2018).

273. Wang, H. *et al.* Genome-wide association analysis of self-reported daytime sleepiness identifies 42 loci that suggest biological subtypes. *Nat. Commun.* **10**, 3503 (2019).

274. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).

275. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).

276. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9**, 2941 (2018).

277. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).

278. Timmers, P. R. *et al.* Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *Elife* **8**, (2019).

279. Grossman, S. R. *et al.* Positional specificity of different transcription factor classes within enhancers. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E7222–E7230 (2018).

280. Vandel, J., Cassan, O., Lèbre, S., Lecellier, C.-H. & Bréhélin, L. Probing transcription factor combinatorics in different promoter classes and in enhancers. *BMC Genomics* **20**, 103 (2019).

281. *Architecture of the human regulatory network derived from ENCODE data*. *Nature* vol. 489 91–100 (2012).

282. Blair Hedges, S. & Kumar, S. *The Timetree of Life*. (OUP Oxford, 2009).

283. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).

284. Foley, N. M. *et al.* A genomic timescale for placental mammal evolution. (2021).

285. Yu, G., Lam, T. T.-Y., Zhu, H. & Guan, Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol. Biol. Evol.* **35**, 3041–3043 (2018).

286. Wang, L.-G. *et al.* Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol. Biol. Evol.* **37**, 599–603 (2020).

287. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* (1995).

288. Enea, M. Fitting Linear and Generalized Linear Models to Large Data Sets [R package speedglm version 0.3-4]. (2022).

289. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* vol. 82 1273–1300 Preprint at https://doi.org/10.1111/rssb.12388 (2020).

290. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).

291. Lambert, J. T. *et al.* Parallel functional testing identifies enhancers active in early postnatal mouse brain. *Elife* **10**, (2021).

292. Greenwald, N. F. *et al.* Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**, 555–565 (2021).

293. van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).

294. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

295. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using**lme4**. *Journal of Statistical Software* vol. 67 Preprint at https://doi.org/10.18637/jss.v067.i01 (2015).

296. National Institute on Drug Abuse. Drug Overdose Death Rates. *National Institute on Drug Abuse* https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates (2023).

297. Puig, S. *et al.* Uncovering circadian rhythm disruptions of synaptic proteome signaling in prefrontal cortex and nucleus accumbens associated with opioid use disorder. *bioRxiv* (2023) doi:10.1101/2023.04.07.536056.

298. Nagamatsu, S. T. *et al.* CpH methylome analysis in human cortical neurons identifies novel gene pathways and drug targets for opioid use disorder. *Front. Psychiatry* **13**, 1078894 (2022).

299. Xue, X. *et al.* Molecular rhythm alterations in prefrontal cortex and nucleus accumbens associated with opioid use disorder. *Transl. Psychiatry* **12**, 123 (2022).

300. Mendez, E. F. *et al.* Angiogenic gene networks are dysregulated in opioid use disorder: evidence from multi-omics and imaging of postmortem human brain. *Mol. Psychiatry* **26**, 7803–7812 (2021).

301. Ray, M. H., Williams, B. R., Kuppe, M. K., Bryant, C. D. & Logan, R. W. A Glitch in the Matrix: The Role of Extracellular Matrix Remodeling in Opioid Use Disorder. *Front. Integr. Neurosci.* **16**, 899637 (2022).

302. Kovacs, G. G. *et al.* Heroin abuse exaggerates age-related deposition of hyperphosphorylated tau and p62-positive inclusions. *Neurobiol. Aging* **36**, 3100–3107 (2015).

303. Madabhushi, R., Pan, L. & Tsai, L.-H. DNA damage and its links to neurodegeneration. *Neuron* **83**, 266–282 (2014).

304. Shanbhag, N. M. *et al.* Early neuronal accumulation of DNA double strand breaks in Alzheimer's disease. *Acta Neuropathologica Communications* vol. 7 Preprint at https://doi.org/10.1186/s40478-019-0723-5 (2019).

305. Wang, Z.-X., Li, Y.-L., Pu, J.-L. & Zhang, B.-R. DNA damage-mediated neurotoxicity in Parkinson's disease. *Int. J. Mol. Sci.* **24**, 6313 (2023).

306. Yang, J.-L., Sykora, P., Wilson, D. M., 3rd, Mattson, M. P. & Bohr, V. A. The excitatory neurotransmitter glutamate stimulates DNA repair to increase neuronal resiliency. *Mech. Ageing Dev.* **132**, 405–411 (2011).

307. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

308. Araújo, B. *et al.* Neuroinflammation and Parkinson's disease-from neurodegeneration to therapeutic opportunities. *Cells* **11**, 2908 (2022).

309. Geloso, M. C. *et al.* The Dual Role of Microglia in ALS: Mechanisms and Therapeutic Approaches. *Front. Aging Neurosci.* **9**, 242 (2017).

310. Song, X. *et al.* DNA Repair Inhibition Leads to Active Export of Repetitive Sequences to the Cytoplasm Triggering an Inflammatory Response. *J. Neurosci.* **41**, 9286–9307 (2021).

311. Welch, G. M. *et al.* Neurons burdened by DNA double-strand breaks incite microglia activation through antiviral-like signaling in neurodegeneration. *Sci Adv* **8**, eabo4662 (2022).

312. Martin, L. J. & Chang, Q. DNA Damage Response and Repair, DNA Methylation, and Cell Death in Human Neurons and Experimental Animal Neurons Are Different. *J. Neuropathol. Exp. Neurol.* **77**, 636–655 (2018).

313. Liu, L. *et al.* Cross-Talking Pathways of Forkhead Box O1 (FOXO1) Are Involved in the Pathogenesis of Alzheimer's Disease and Huntington's Disease. *Oxid. Med. Cell. Longev.* **2022**, (2022).

314. Becher, B., Spath, S. & Goverman, J. Cytokine networks in neuroinflammation. *Nat. Rev. Immunol.* **17**, 49–59 (2017).

315. Shokri-Kojori, E. *et al.* Brain opioid segments and striatal patterns of dopamine release induced by naloxone and morphine. *Hum. Brain Mapp.* **43**, (2022).

316. Shokri-Kojori, E., Wang, G.-J. & Volkow, N. D. Naloxone precipitated withdrawal increases dopamine release in the dorsal striatum of opioid dependent men. *Transl. Psychiatry* **11**, 445 (2021).

317. Zilverstand, A., Huang, A. S., Alia-Klein, N. & Goldstein, R. Z. Neuroimaging Impaired Response Inhibition and Salience Attribution in Human Drug Addiction: A Systematic Review. *Neuron* **98**, (2018).

318. Muñoz-Manchado, A. B. *et al.* Diversity of Interneurons in the Dorsal Striatum Revealed by Single-Cell RNA Sequencing and PatchSeq. *Cell Rep.* **24**, 2179-2190.e7 (2018).

319. Sharif, N. A. & Hughes, J. Discrete mapping of brain Mu and delta opioid receptors using selective peptides: quantitative autoradiography, species differences and comparison with kappa receptors. *Peptides* **10**, 499–522 (1989).

320. Maduna, T. *et al.* Microglia Express Mu Opioid Receptor: Insights From Transcriptomics and Fluorescent Reporter Mice. *Front. Psychiatry* **9**, 726 (2018).

321. Coffey, K. R. *et al.* A cAMP-Related Gene Network in Microglia Is Inversely Regulated by Morphine Tolerance and Withdrawal. *Biol Psychiatry Glob Open Sci* **2**, 180–189 (2022).

322. Dai, K. Z. *et al.* Dopamine D2 receptors bidirectionally regulate striatal enkephalin expression: Implications for cocaine reward. *Cell Rep.* **40**, 111440 (2022).

323. Crittenden, J. R. *et al.* Striosome-dendron bouquets highlight a unique striatonigral circuit targeting dopamine-containing neurons. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11318–11323 (2016).

324. Hong, S. *et al.* Predominant Striatal Input to the Lateral Habenula in Macaques Comes from Striosomes. *Curr. Biol.* **29**, (2019).

325. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

326. Kim, D. & Tsai, L.-H. Linking cell cycle reentry and DNA damage in neurodegeneration. *Ann. N. Y. Acad. Sci.* **1170**, 674–679 (2009).

327. Nishimura, K. *et al.* Mcm8 and Mcm9 form a complex that functions in homologous recombination repair induced by DNA interstrand crosslinks. *Mol. Cell* **47**, (2012).

328. Lee, K. Y. *et al.* MCM8-9 complex promotes resection of double-strand break ends by MRE11-RAD50-NBS1 complex. *Nat. Commun.* **6**, 7744 (2015).

329. Osterweil, E., Wells, D. G. & Mooseker, M. S. A role for myosin VI in postsynaptic structure and glutamate receptor endocytosis. *J. Cell Biol.* **168**, 329–338 (2005).

330. Auer, M., Hausott, B. & Klimaschewski, L. Rho GTPases as regulators of morphological neuroplasticity. *Ann. Anat.* **193**, 259–266 (2011).

331. Tolias, K. F., Duman, J. G. & Um, K. Control of synapse development and plasticity by Rho GTPase regulatory proteins. *Prog. Neurobiol.* **94**, 133–148 (2011).

332. Fritz, G. & Henninger, C. Rho GTPases: Novel Players in the Regulation of the DNA Damage Response? *Biomolecules* **5**, 2417 (2015).

333. Ayton, S., Faux, N. G. & Bush, A. I. Ferritin levels in the cerebrospinal fluid predict Alzheimer's disease outcomes and are regulated by APOE. *Nat. Commun.* **6**, 1–9 (2015).

334. Reichert, C. O. *et al.* Ferroptosis Mechanisms Involved in Neurodegenerative Diseases. *Int. J. Mol. Sci.* **21**, (2020).

335. Belaidi, A. A. *et al.* Apolipoprotein E potently inhibits ferroptosis by blocking ferritinophagy. *Mol. Psychiatry* 1–10 (2022).

336. Murphy, M. P. How mitochondria produce reactive oxygen species. *Biochem. J* **417**, 1–13 (2009).

337. Shao, A.-W. *et al.* Bclaf1 is an important NF-κB signaling transducer and C/EBPβ regulator in DNA damage-induced senescence. *Cell Death Differ.* **23**, 865–875 (2016).

338. Yu, Z., Zhu, J., Wang, H., Li, H. & Jin, X. Function of BCLAF1 in human disease. *Oncol. Lett.* **23**, (2022).

339. McGowan, P. O. *et al.* Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. *Nat. Neurosci.* **12**, 342–348 (2009).

340. Holmes, L., Jr *et al.* Aberrant Epigenomic Modulation of Glucocorticoid Receptor Gene (NR3C1) in Early Life Stress and Major Depressive Disorder Correlation: Systematic Review and Quantitative Evidence Synthesis. *Int. J. Environ. Res. Public Health* **16**, (2019).

341. Glucocorticoid receptor gene (NR3C1) methylation processes as mediators of early adversity in stress-related disorders causality: A critical review. *Neurosci. Biobehav. Rev.* **55**, 520–535 (2015).

342. Green, T. A. *et al.* Induction of Activating Transcription Factors (ATFs) ATF2, ATF3, and ATF4 in the Nucleus Accumbens and Their Regulation of Emotional Behavior. *J. Neurosci.* **28**, 2025–2032 (2008).

343. Pattinson, C. L. *et al.* Excessive daytime sleepiness is associated with altered gene expression in military personnel and veterans with posttraumatic stress disorder: an RNA sequencing study. *Sleep* **43**, (2020).

344. Shahabi, N. A., McAllen, K. & Sharp, B. M. Phosphorylation of activating transcription factor in murine splenocytes through delta opioid receptors. *Cell. Immunol.* **221**, (2003).

345. Banghart, M. R., Neufeld, S. Q., Wong, N. C. & Sabatini, B. L. Enkephalin Disinhibits Mu Opioid Receptor-Rich Striatal Patches via Delta Opioid Receptors. *Neuron* **88**, 1227–1239 (2015).

346. Jiang, Z. G. & North, R. A. Pre- and postsynaptic inhibition by opioids in rat striatum. *J. Neurosci.* **12**, 356–361 (1992).

347. Tajima, H. *et al.* Evidence for in vivo production of Humanin peptide, a neuroprotective factor against Alzheimer's disease-related insults. *Neurosci. Lett.* **324**, (2002).

348. Gurunathan, S., Jeyaraj, M., Kang, M.-H. & Kim, J.-H. Mitochondrial Peptide Humanin Protects Silver Nanoparticles-Induced Neurotoxicity in Human Neuroblastoma Cancer Cells (SH-SY5Y). *Int. J. Mol. Sci.* **20**, (2019).

349. Ahuja, M. *et al.* Bach1 derepression is neuroprotective in a mouse model of Parkinson's disease. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

350. Tolve, M. *et al.* The transcription factor BCL11A defines distinct subsets of midbrain dopaminergic neurons. *Cell Rep.* **36**, (2021).

351. Simon, R., Wiegreffe, C. & Britsch, S. Bcl11 Transcription Factors Regulate Cortical Development and Function. *Front. Mol. Neurosci.* **13**, (2020).

352. Cirnaru, M.-D. *et al.* Unbiased identification of novel transcription factors in striatal compartmentation and striosome maturation. *Elife* **10**, (2021).

353. Benedito, A. B. *et al.* The transcription factor NFAT3 mediates neuronal survival. *J. Biol. Chem.* **280**, 2818–2825 (2005).

354. Hudry, E. *et al.* Inhibition of the NFAT pathway alleviates amyloid β neurotoxicity in a mouse model of Alzheimer's disease. *J. Neurosci.* **32**, (2012).

355. Manocha, G. D. *et al.* NFATc2 Modulates Microglial Activation in the AβPP/PS1 Mouse Model of Alzheimer's Disease. *J. Alzheimers. Dis.* **58**, (2017).

356. Vihma, H., Luhakooder, M., Pruunsild, P. & Timmusk, T. Regulation of different human NFAT isoforms by neuronal activity. *J. Neurochem.* **137**, (2016).

357. Godino, A. *et al.* Transcriptional control of nucleus accumbens neuronal excitability by retinoid X receptor alpha tunes sensitivity to drug rewards. *Neuron* (2023) doi:10.1016/j.neuron.2023.02.013.

358. Krezel, W. *et al.* Impaired locomotion and dopamine signaling in retinoid receptor mutant mice. *Science* **279**, 863–867 (1998).

359. Krzyzosiak, A. *et al.* Retinoid x receptor gamma control of affective behaviors involves dopaminergic signaling in mice. *Neuron* **66**, 908–920 (2010).

360. Wietrzych-Schindler, M. *et al.* Retinoid x receptor gamma is implicated in docosahexaenoic acid modulation of despair behaviors and working memory in mice. *Biol. Psychiatry* **69**, 788–794 (2011).

361. Krzyżosiak, A. *et al.* Vitamin A5/X controls stress-adaptation and prevents depressive-like behaviors in a mouse model of chronic stress. *Neurobiology of stress* **15**, (2021).

362. Cook, M. E., Jarjour, N. N., Lin, C.-C. & Edelson, B. T. Transcription Factor Bhlhe40 in Immunity and Autoimmunity. *Trends Immunol.* **41**, 1023–1036 (2020).

363. Lin, C.-C. *et al.* Bhlhe40 controls cytokine production by T cells and is essential for pathogenicity in autoimmune neuroinflammation. *Nat. Commun.* **5**, 1–13 (2014).

364. Zafar, A., Ng, H. P., Kim, G.-D., Chan, E. R. & Mahabeleshwar, G. H. BHLHE40 promotes macrophage pro-inflammatory gene expression and functions. *FASEB J.* **35**, e21940 (2021).

365. Logan, R. W. *et al.* NAD+ cellular redox and SIRT1 regulate the diurnal rhythms of tyrosine hydroxylase and conditioned cocaine reward. *Mol. Psychiatry* **24**, 1668–1684 (2019).

366. Roy, E. R. *et al.* Type I interferon response drives neuroinflammation and synapse loss in Alzheimer disease. *J. Clin. Invest.* **130**, 1912–1930 (2020).

367. Thompson, J. A. & Ziman, M. Pax genes during neural development and their potential role in neuroregeneration. *Prog. Neurobiol.* **95**, (2011).

368. Silies, M. & Klämbt, C. APC/CFzr/Cdh1-dependent regulation of cell adhesion controls glial migration in the Drosophila PNS. *Nat. Neurosci.* **13**, 1357–1364 (2010).

369. Bolliger, M. F., Martinelli, D. C. & Südhof, T. C. The cell-adhesion G protein-coupled receptor BAI3 is a high-affinity receptor for C1q-like proteins. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 2534–2539 (2011).

370. Kakegawa, W. *et al.* Anterograde C1ql1 signaling is required in order to determine and maintain a single-winner climbing fiber in the mouse cerebellum. *Neuron* **85**, 316–329 (2015).

371. Neniskyte, U. & Gross, C. T. Errant gardeners: glial-cell-dependent synaptic pruning and neurodevelopmental disorders. *Nat. Rev. Neurosci.* **18**, 658–670 (2017).

372. Gamble, M. C. *et al.* Mu-opioid receptor and receptor tyrosine kinase crosstalk: Implications in mechanisms of opioid tolerance, reduced analgesia to neuropathic pain, dependence, and reward. *Front. Syst. Neurosci.* **16**, (2022).

373. Puig, S., Donica, C. L. & Gutstein, H. B. EGFR Signaling Causes Morphine Tolerance and Mechanical Sensitization in Rats. *eNeuro* **7**, (2020).

374. Barkai, O. *et al.* Platelet-derived growth factor activates nociceptive neurons by inhibiting M-current and contributes to inflammatory pain. *Pain* **160**, (2019).

375. Lopez-Bellido, R. *et al.* Growth Factor Signaling Regulates Mechanical Nociception in Flies and Vertebrates. *J. Neurosci.* **39**, 6012–6030 (2019).

376. Canepa, E. & Fossati, S. Impact of Tau on Neurovascular Pathology in Alzheimer's Disease. *Front. Neurol.* **11**, 573324 (2020).

377. Xia, Y. *et al.* C/EBPβ is a key transcription factor for APOE and preferentially mediates ApoE4 expression in Alzheimer's disease. *Mol. Psychiatry* **26**, 6002–6022 (2020).

378. Yao, Y. *et al.* A delta-secretase-truncated APP fragment activates CEBPB, mediating Alzheimer's disease pathologies. *Brain* **144**, (2021).

379. Deak, J. *et al.* Genome-wide association study in individuals of European and African ancestry and multi-trait analysis of opioid use disorder identifies 19 independent genome-wide significant risk loci. *Mol. Psychiatry* **27**, 3970–3979 (2022).

380. Ma, S. *et al.* Molecular and cellular evolution of the primate dorsolateral prefrontal cortex. *Science* **377**, eabo7257 (2022).

381. Gerdle, B. *et al.* Prevalence of widespread pain and associations with work status: a population study. *BMC Musculoskelet. Disord.* **9**, (2008).

382. Ailes, E. C. *et al.* Opioid prescription claims among women of reproductive age--United States, 2008-2012. *MMWR Morb. Mortal. Wkly. Rep.* **64**, (2015).

383. McHugh, R. K. *et al.* Gender differences in a clinical trial for prescription opioid dependence. *J. Subst. Abuse Treat.* **45**, (2013).

384. Gjersing, L. & Bretteville-Jensen, A. L. Gender differences in mortality and risk factors in a 13-year cohort study of street-recruited injecting drug users. *BMC Public Health* **14**, (2014).

385. Lee, R. S. *et al.* DNA methylation and sex-specific expression of FKBP5 as correlates of one-month bedtime cortisol levels in healthy individuals. *Psychoneuroendocrinology* **97**, (2018).

386. Nold, V. *et al.* Impact of Fkbp5 × early life adversity × sex in humanised mice on multidimensional stress responses and circadian rhythmicity. *Mol. Psychiatry* **27**, 3544–3555 (2022).

387. Serdarevic, M., Striley, C. W. & Cottler, L. B. Gender differences in prescription opioid use. *Curr. Opin. Psychiatry* **30**, 238 (2017).

388. Piechota, M. *et al.* The dissection of transcriptional modules regulated by various drugs of abuse in the mouse striatum. *Genome Biol.* **11**, R48 (2010).

389. McClung, C. A., Nestler, E. J. & Zachariou, V. Regulation of Gene Expression by Chronic Morphine and Morphine Withdrawal in the Locus Ceruleus and Ventral Tegmental Area. *J. Neurosci.* **25**, 6005 (2005).

390. Gonzalez-Hunt, C. P. & Sanders, L. H. DNA damage and repair in Parkinson's disease: Recent advances and new opportunities. *J. Neurosci. Res.* **99**, 180–189 (2021).

391. Hill, S. J. *et al.* Two familial ALS proteins function in prevention/repair of transcription-associated DNA damage. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E7701–E7709 (2016).

392. Caldecott, K. W., Ward, M. E. & Nussenzweig, A. The threat of programmed DNA damage to neuronal genome integrity and plasticity. *Nat. Genet.* **54**, 115–120 (2022).

393. Madabhushi, R. *et al.* Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. *Cell* **161**, 1592–1605 (2015).

394. Stott, R. T., Kritsky, O. & Tsai, L.-H. Profiling DNA break sites and transcriptional changes in response to contextual fear learning. *PLoS One* **16**, e0249691 (2021).

395. Pollina, E. A. *et al.* A NPAS4-NuA4 complex couples synaptic activity to DNA repair. *Nature* **614**, 732–741 (2023).

396. Wu, W. *et al.* Neuronal enhancers are hotspots for DNA single-strand break repair. *Nature* **593**, 440–444 (2021).

397. Tepper, J. M. *et al.* Heterogeneity and Diversity of Striatal GABAergic Interneurons: Update 2018. *Front. Neuroanat.* **12**, 91 (2018).

398. Arttamangkul, S., Platt, E. J., Carroll, J. & Farrens, D. Functional independence of endogenous μ- and δ-opioid receptors co-expressed in cholinergic interneurons. *Elife* **10**, (2021).

399. Zhu, Y. *et al.* Opioid-induced fragile-like regulatory T cells contribute to withdrawal. *Cell* **186**, 591-606.e23 (2023).

400. Koob, G. F. Drug Addiction: Hyperkatifeia/Negative Reinforcement as a Framework for Medications Development. *Pharmacol. Rev.* **73**, 163–201 (2021).

401. Didier, M. *et al.* DNA strand breaks induced by sustained glutamate excitotoxicity in primary neuronal cultures. *J. Neurosci.* **16**, 2238–2250 (1996).

402. Hearing, M. C. *et al.* Reversal of morphine-induced cell-type-specific synaptic plasticity in the nucleus accumbens shell blocks reinstatement. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 757–762 (2016).

403. Hearing, M., Graziane, N., Dong, Y. & Thomas, M. J. Opioid and Psychostimulant Plasticity: Targeting Overlap in Nucleus Accumbens Glutamate Signaling. *Trends Pharmacol. Sci.* **39**, 276–294 (2018).

404. Panatier, A. & Robitaille, R. Astrocytic mGluR5 and the tripartite synapse. *Neuroscience* **323**, 29–34 (2016).

405. Morigaki, R. & Goto, S. Striatal Vulnerability in Huntington's Disease: Neuroprotection Versus Neurotoxicity. *Brain Sci* **7**, (2017).

406. Lynch, W. J., Roth, M. E. & Carroll, M. E. Biological basis of sex differences in drug abuse: preclinical and clinical studies. *Psychopharmacology* **164**, 121–137 (2002).

407. Huhn, A. S., Berry, M. S. & Dunn, K. E. Review: Sex-Based Differences in Treatment Outcomes for Persons With Opioid Use Disorder. *Am. J. Addict.* **28**, 246–261 (2019).

408. McKee, S. A. & McRae-Clark, A. L. Consideration of sex and gender differences in addiction medication response. *Biol. Sex Differ.* **13**, 34 (2022).

409. Bekhbat, M. & Neigh, G. N. Sex differences in the neuro-immune consequences of stress: Focus on depression and anxiety. *Brain Behav. Immun.* **67**, 1–12 (2018).

410. Barko, K. *et al.* Brain region- and sex-specific transcriptional profiles of microglia. *Front. Psychiatry* **13**, 945548 (2022).

411. Zannas, A. S., Wiechmann, T., Gassen, N. C. & Binder, E. B. Gene-Stress-Epigenetic Regulation of FKBP5: Clinical and Translational Implications. *Neuropsychopharmacology* **41**, 261–274 (2016).

412. Levran, O. *et al.* Stress-related genes and heroin addiction: a role for a functional FKBP5 haplotype. *Psychoneuroendocrinology* **45**, 67–76 (2014).

413. Nylander, I. *et al.* Evidence for a Link Between Fkbp5/FKBP5, Early Life Social Relations and Alcohol Drinking in Young Adult Rats and Humans. *Mol. Neurobiol.* **54**, 6225–6234 (2017).

414. National Research Council, Division on Earth and Life Studies, Institute for Laboratory Animal Research & Committee for the Update of the Guide for the Care and Use of Laboratory Animals. *Guide for the Care and Use of Laboratory Animals: Eighth Edition*. (National Academies Press, 2011).

415. Holtzman, S. G. & Villarreal, J. E. Operant behavior in the morphine-dependent rhesus monkey. *J. Pharmacol. Exp. Ther.* **184**, 528–541 (1973).

416. Borgatti, D., Bergman, J. & Kohut, S. Effects of lorcaserin on opioid-induced observable behavior in rhesus monkeys. *FASEB J.* **34**, 1–1 (2020).

417. Renthal, W. *et al.* Characterization of human mosaic Rett syndrome brain tissue by single-nucleus RNA sequencing. *Nat. Neurosci.* **21**, 1670–1679 (2018).

418. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

419. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv* 2021.05.05.442755 (2021) doi:10.1101/2021.05.05.442755.

420. Phan, B. & Pfenning, A. Alternate gene annotations for rat, macaque, and marmoset for single cell RNA and ATAC analyses. *https://kilthub.cmu.edu/articles/dataset/alternate_gene_annotations_for_rat_macaque_and_marmoset_for_single_cell_rna_and_atac_analyses/21176401* https://doi.org/10.1184/R1/21176401.v1 (2022) doi:10.1184/R1/21176401.v1.

421. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, (2020).

422. Muskovic, W. & Powell, J. E. DropletQC: improved identification of empty droplets and damaged cells in single-cell RNA-seq data. *Genome Biol.* **22**, 1–9 (2021).

423. Bais, A. S. & Kostka, D. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics* **36**, 1150–1158 (2019).

424. Hippen, A. A. *et al.* miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data. *PLoS Comput. Biol.* **17**, e1009290 (2021).

425. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

426. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).

427. Ahlmann-Eltze, C. & Huber, W. glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics* **36**, 5701–5702 (2020).

428. Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 1–12 (2020).

429. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* (2018).

430. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).

431. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).

432. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

433. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021) doi:10.1101/060012.

434. Koopmans, F. *et al.* SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron* **103**, 217-234.e4 (2019).

435. Kowalczyk, A., Chikina, M. & Clark, N. Complementary evolution of coding and noncoding sequence underlies mammalian hairlessness. *Elife* **11**, (2022).

436. Kowalczyk, A., Partha, R., Clark, N. L. & Chikina, M. Pan-mammalian analysis of molecular constraints underlying extended lifespan. *Elife* **9**, (2020).

437. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. (2006).

438. Van de Sande, B. *et al.* A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**, 2247–2276 (2020).

439. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2018).