UNIVERSITY OF CALIFORNIA SAN DIEGO

Revealing translational and fundamental insights via computational analysis of single-cell sequencing data

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Jessica Lu Zhou

Committee in charge:

Professor Graham McVicker, Chair
Professor Bing Ren, Co-Chair
Professor Kyle Gaulton
Professor Trey Ideker
Professor Abraham Palmer
Professor Francesca Telese

2023

The Dissertation of Jessica Lu Zhou is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

I dedicate my dissertation work to my grandfathers, my 爷爷 and 外公, who could not be here to see me graduate college and embark on my doctoral education journey but played instrumental roles in raising me from birth to adulthood.

TABLE OF CONTENTS

LIST OF SUPPLEMENTAL FILES

Chapter 1

Supplemental File 1.1: zhou_snrna_cell_metrics.csv

Description: Per-nucleus metrics for all nuclei in snRNA-seq dataset after filtering. This table contains selected columns from the metadata table for the Seurat object containing the integrated snRNA-seq data.

Supplemental File 1.2: zhou_snatac_cell_metrics.csv

Description: Per-nucleus metrics for all nuclei in snATAC-seq dataset after filtering. This table contains selected columns from the metadata table for the Signac object containing the integrated snATAC-seq data.

Supplemental File 1.3: zhou_deg_results.txt

Description: All cell type-specific differential gene expression analysis results, obtained using the negative binomial test.

Supplemental File 1.4: zhou_degs_with_eqtls.txt

Description: All DEGs (FDR<10%) that also have eQTLs in the rat brain, with a list of variant IDs for corresponding eQTLs.

Supplemental File 1.5: zhou_kegg_gsea_results.txt

Description: KEGG GSEA results.

Supplemental File 1.6: zhou_diff_chromatin.csv

Description: All cell type-specific differential peak accessibility analysis results, obtained using the negative binomial test.

Supplemental File 1.7: zhou_chromvar_results.tsv

Description: ChromVar analysis results.

Chapter 2

Supplemental File 2.1: zhou_nmu_experiment.xlsx

Description: Data from *NMU* RT-qPCR experiment.

LIST OF FIGURES

Chapter 1

Chapter 2

LIST OF TABLES

Chapter 1

Chapter 2

LIST OF ABBREVIATIONS

RNA-seq	RNA-sequencing
ATAC-seq	Assay for transposase-accessible chromatin with sequencing
scRNA-seq	Single-cell RNA-seq
scATAC-seq	Single-cell ATAC-seq
snRNA-seq	Single-nucleus RNA-seq
snATAC-seq	Single-nucleus ATAC-seq
DEG	Differentially expressed gene
GSEA	Gene set enrichment analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
OCR	Open chromatin region
AI	Addiction index
IVSA	Intravenous self-administration
HS	Heterogeneous stock
ShA	Short access (to drug self-administration)
LgA	Long access (to drug self-administration)
PR	Progressive ratio
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRISPRi	CRISPR interference
GLM	Generalized Linear Model
GLiMMIRS	Generalized Linear Models for Measuring Interactions between Regulatory Sequences

ACKNOWLEDGEMENTS

I would like to express my heartfelt appreciation to my thesis advisor, Dr. Graham McVicker, for their unwavering support, guidance, and encouragement throughout the completion of my dissertation. I am grateful for their patience and dedication in helping me overcome numerous challenges and obstacles that I encountered during the process. Their mentorship has been instrumental in developing my research skills and confidence as a scientist. Likewise, I would like to thank my thesis committee members: Dr. Bing Ren (co-chair), Dr. Francesca Telese, Dr. Abraham Palmer, Dr. Trey Ideker, and Dr. Kyle Gaulton. Their expertise, insightful comments, and constructive feedback have been invaluable in shaping my research and writing.

I would also like to acknowledge the support of the University of California San Diego faculty and staff, especially those affiliated with the Bioinformatics and Systems Biology graduate program, who provided me with valuable resources and opportunities to enhance my academic and professional growth.

Finally, I would like to express my deepest gratitude to my family and friends for their unwavering love, support, and encouragement throughout my academic journey. They have been the driving force behind my success. Thank you all for making this journey possible.

Chapter 1, in part, has been submitted for publication of the material at *Nature Neuroscience*. Zhou, J.L.; de Guglielgmo, G.; Ho, A.J.; Kallupi, M.; Pokhrel, N.; Li, H.; Chitre, A.S.; Munro, D.; Mohammadi, P.; Carette, L.L.; George, O.; Palmer, A.A.; McVicker, G.; Telese, F. The dissertation author was the primary researcher and author of this paper.

Chapter 2, in part, has been submitted for publication of the material. Zhou, J.L.; Guruvayurappan, K.; Chen, H.V.; Chen, A.R.; McVicker, G. The dissertation author was the primary researcher and author of this paper.

VITA

2017 Bachelor of Science in Biomedical Engineering, University of Southern California

2023 Doctor of Philosophy in Bioinformatics and Systems Biology, University of California San Diego

ABSTRACT OF THE DISSERTATION

Revealing translational and fundamental biological insights via computational analysis of single-cell sequencing data

by

Jessica Lu Zhou

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2023

Professor Graham McVicker, Chair
Professor Bing Ren, Co-Chair

Single-cell sequencing has emerged as a powerful tool for dissecting cellular heterogeneity and providing cell type-specific biological insights. Single-cell sequencing technologies have rapidly proliferated over the last decade, leading to an explosion of data generated from such experiments. However, several challenges exist in the computational analysis of single-cell sequencing data due to its large and complex nature, including the need for sophisticated statistical methods to distinguish biologically meaningful signals from noise, the integration of single-cell sequencing data with other types of biological information, and the development of scalable and reproducible computational pipelines that can handle the large and complex nature of the data. In this dissertation, I present two distinct projects analyzing single-cell sequencing

data. The first is of an analytical nature and tackles a translational question. In this project, I built computational pipelines for processing and analyzing single-nucleus RNA- and ATAC-sequencing datasets generated from the amygdalae of genetically diverse heterogenous stock rats, which were subjected to a behavioral protocol for studying addiction-like behaviors following cocaine self-administration. In doing so, I provide a standard reference for analyzing such data as well as reveal cell type-specific insights into the molecular underpinnings of cocaine addiction. The second project is oriented towards methods development and seeks to understand the fundamental biological question of transcriptional regulation. Here, I developed a statistical framework for simulating and modeling data from single-cell CRISPR regulatory screens and used it to perform a genome-wide interrogation of epistatic-like interactions between enhancer pairs. I found that multiple enhancers act together in a multiplicative fashion with little evidence for interactive effects between them. This work revealed novel insights into the collective behavior of multiple regulatory elements and provides a tool that can be applied to future datasets generated from such experiments. This dissertation exemplifies how computational methods can be applied in different contexts to extract meaning from a variety of single-cell sequencing modalities. By tackling both a translational and fundamental biological question, I have showcased the breadth of what can be revealed by studying single-cell sequencing data and the computational methods necessary to extract this information.

INTRODUCTION

Single-cell sequencing is a powerful technology that allows the genetic material of individual cells to be captured and analyzed, providing unprecedented insights into the genomic landscapes of complex biological systems. The origins of single-cell sequencing can be traced back to the early 1990s[1,2], when the first methods for amplifying and analyzing DNA from single cells were developed. However, it was not until the advent of next-generation sequencing technologies in the mid-2000s that single-cell sequencing began to truly revolutionize genomics research. The first report of single-cell transcriptome analysis using a next-generation sequencing platform was published by Tang et al. in 2009[3], which studied a single mouse blastomere and reported improvements over microarray techniques.

Since then, advances in single-cell sequencing technologies have rapidly proliferated and it has become a critical tool for many areas of biology, including developmental biology, immunology, neurobiology, cancer research, and microbiology. By enabling the study of individual cells rather than entire tissues or populations of cells, single-cell sequencing has revealed previously hidden cellular diversity, identified novel cell types and subpopulations, and provided new insights into the mechanisms underlying cell differentiation and disease.

The popularity of single-cell sequencing over the past decade has also led to an explosion of datasets generated from such experiments, which must be carefully analyzed to extract valuable biological insights. However, the computational analysis of single-cell sequencing data presents several notable challenges. First, the data generated by single-cell sequencing technologies are large and complex, consisting of millions of short reads or transcript counts for each individual cell. This requires significant computational resources and specialized algorithms for data processing, normalization, and analysis.

Second, single-cell sequencing data are highly heterogeneous, reflecting the inherent variability in gene expression, chromatin accessibility, or other molecular features across individual cells. This can also include technical noise introduced during sample preparation and sequencing, such as batch effects, dropouts, and amplification bias, which can affect the accuracy and reproducibility of downstream analyses. This requires sophisticated statistical methods for identifying biologically meaningful signals and distinguishing them from noise.

Finally, the interpretation of single-cell sequencing data requires integration with other types of biological information, such as gene annotations, pathway databases, and cell type reference maps. This requires the development of sophisticated computational tools for data integration and visualization, as well as a deep understanding of the biological context and hypotheses under investigation.

For these reasons, conducting proper analyses of single-cell sequencing data with the goal of extracting meaningful and accurate biological insights is no trivial task. In my dissertation, I have performed computational analyses of different single-cell sequencing modalities and demonstrated how the findings can answer both translational and fundamental biological questions. In the first chapter, I conducted an exploratory analysis of single-nucleus RNA- and ATAC-sequencing data generated from the amygdala of rats subjected to a behavioral protocol for intravenous cocaine self-administration. In doing so, I revealed cell type-specific molecular features that are associated with addiction-like behaviors and present standardized computational pipelines for processing and analyzing these data. In my second chapter, I present a statistical framework for simulating and modeling single-cell RNA-sequencing (scRNA-seq) readout from a CRISPR interference (CRISPRi) experiment that targeted putative enhancers in the genome. My models were designed to evaluate whether pairs of enhancers display epistatic-

like interaction effects on target gene expression and is the first genome-wide evaluation of this phenomenon. My findings provide novel insights into the activity of multiple enhancers acting in tandem to answer important questions in the field of gene regulation. Additionally, my statistical framework is one of the first designed specifically for interpretation of CRISPR screens performed in single cells. Such experiments integrate two cutting-edge and powerful technologies that each come with their own set of challenges when it comes to data analysis and interpretation. Thus, in developing my models, I explored the unique statistical considerations that must be accounted for when analyzing data from single-cell sequencing experiments performed in unique contexts.

In conclusion, single-cell sequencing has had a profound impact on genomics research, revealing the complexity and diversity of biological systems in ways that were previously impossible. Its continued development and application are likely to have far-reaching implications for our understanding of biology and for the development of new therapeutic approaches. Therefore, ensuring that we have a thorough understanding of the best practices for how to analyze the data generated by such experiments is an essential area of study. My dissertation details several computational approaches for analyzing data from different types of single-cell sequencing experiments and demonstrates the value of these analyses through the biological insights that they provide, both for translational and fundamental questions in biological research. Altogether, my work has made important contributions towards our understanding of best practices for computational analysis of single-cell sequencing data while also providing impactful biological findings along the way.

REFERENCES

1. Brady G, Barbara M, Iscove NN. Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. Methods Mol Cell Biol. 1990;2(1):17–25.

2. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, Coleman P. Analysis of gene expression in single live neurons. Proc Natl Acad Sci U S A. 1992 Apr 1;89(7):3010–3014. PMCID: PMC48793

3. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May;6(5):377–382. PMID: 19349980

CHAPTER 1: Cocaine addiction-like behaviors are associated with long-term changes in gene regulation, energy metabolism, and GABAergic inhibition within the amygdala

1.1: Abstract

The amygdala processes positive and negative valence and contributes to the development of addiction, but the underlying cell type-specific gene regulatory programs are unknown. We generated an atlas of single nucleus gene expression and chromatin accessibility in the amygdala of outbred rats with low and high cocaine addiction-like behaviors following prolonged abstinence. Between rats with different addiction indexes, we identified thousands of cell type-specific differentially expressed genes enriched for energy metabolism-related pathways that are known to affect synaptic transmission and action potentials. Rats with high addiction-like behaviors showed enhanced GABAergic transmission in the amygdala, which, along with relapse-like behaviors, were reversed by inhibition of Glyoxalase 1, which metabolizes the GABAA receptor agonist methylglyoxal. Finally, we identified thousands of cell type-specific chromatin accessible sites and transcription factor (TF) motifs where accessibility was associated with addiction index, most notably at motifs for pioneer TFs in the Fox, Sox, helix-loop-helix, and AP1 families.

1.2: Introduction

The amygdala is a key brain region involved in regulating a wide range of behaviors, including those related to emotions, motivation and memory[1]. In response to rewarding or aversive environmental stimuli, the amygdala allows organisms to engage in subsequent valence-specific behaviors by determining the value of different stimuli and guiding decision-making based on potential outcomes[1]. The amygdala is implicated in numerous neuropsychiatric disorders including addiction[2–4]. Repeated drug use leads to a heightened sense of pleasure,

which engages the amygdala to form drug-associated memories and reinforces drug-seeking behavior, as the individual is motivated to seek out and use the drug again to experience the rewarding effects[5]. In addition, during withdrawal from addictive drugs, the amygdala mediates negative emotional states, such as anxiety, fear, and irritability[5]. Avoidance of these aversive emotions enhances the incentive value of the drug, leading to sustained drug-seeking behaviors and relapse[6–8]. Because prevention of relapse is the cornerstone of effective treatments for addiction, it is important to understand the amygdala's role in addiction and relapse.

The amygdala is composed of multiple discrete and interconnected subregions, each characterized by highly specialized neuronal populations distinguishable by their morphology and electrophysiological properties[9]. The major subdivisions include the basolateral amygdala (BLA), composed of excitatory glutamatergic neurons and GABAergic inhibitory interneurons, and the central amygdala (CeA), composed of GABAergic neurons[10–12]. While the behavioral function and connectivity of individual subregions of the amygdala have recently been established[1], the mechanisms by which distinct subpopulations of neuronal and non-neuronal cells contribute to its function remains unclear.

Single-cell genomics is a powerful new approach for determining the cellular function and diversity of complex tissues like the amygdala. Single-cell RNA-sequencing (scRNA-seq), which profiles gene expression in individual cells, has identified and cataloged diverse cell types in human, mouse, and non-human primate brains[13–19]. In addition, single-cell assays for transposase-accessible chromatin (scATAC-seq), which profile chromatin accessibility at single cell resolution, have identified regulatory DNA sequences in the rodent and human brain[13,20–26]. Regulatory elements identified by scATAC-seq include promoters and enhancers, which confer

cell type-specificity to gene expression by recruiting sequence-specific transcription factors (TFs)[27–30].

Single cell assays have the potential to reveal, at a molecular level, how specialized amygdalar cell populations are involved in addiction. For example, given that most genetic variants associated with complex human diseases like addiction are located in noncoding regions of the genome[31], snATAC-seq could uncover genetically determined, cell-type specific differences and facilitate functional interpretation of genetic variants[32]. Thus far, however, the application of single-cell assays to the study of addiction-like behaviors in rodents has been limited. Single nucleus RNA-seq (snRNA-seq) has been applied to characterize cellular diversity in brain regions involved in the reward system[33–36] and has been used to analyze transcriptional changes induced by cocaine and morphine[37,38]. However, these prior studies used isogenic rodents, which means that genetically-mediated differences in susceptibility to addiction-like behaviors were not examined. Furthermore, these studies performed experiments following acute, experimenter-administration of drug treatments, which means that they reflect the acute effects of involuntary drug use rather than molecular differences associated with the development of long-lasting addictive-like behaviors. For these reasons, the results from prior single nucleus studies have significant limitations.

To address this knowledge gap, we performed snRNA-seq and snATAC-seq using amygdala tissue from outbred rats obtained from a large genetic study of cocaine addiction-related traits[39]. These rats are subjected to prolonged abstinence from voluntary cocaine intake in a well-validated model of extended access to drug intravenous self-administration (IVSA)[6,39–41]. IVSA is associated with neurochemical changes in key brain regions, which are thought to be similar to those observed in humans with cocaine use disorder[42]. This study used outbred

heterogeneous stock (HS) rats because they have high levels of genetic variation and rich phenotypic diversity[43–46]. By analyzing differences in gene expression and chromatin accessibility in rats with high and low addiction indexes (AI), we identified genes and transcriptional regulators associated with cocaine addiction-like behaviors, including those implicated in energy metabolism and neurotransmitter pathways. Furthermore, using genetic predictions of gene expression, we found that genetic differences contribute to the gene expression differences between high and low AI rats. Finally, we performed pharmacological manipulation of $GABA_A$ receptor signaling in amygdalar tissue slices and in rats to validate insights gained from the transcriptomic data.

1.3: Results

1.3.1: Behavioral characterization of HS rats exhibiting low or high cocaine addiction-like traits

To investigate how chronic cocaine use influences cellular states associated with addiction-like behaviors, we performed snRNA-seq and snATAC-seq on amygdala tissues from HS rats subjected to protracted abstinence (4 weeks) following extended access to cocaine IVSA [39,47–50] (Figure 1.1a). The animals were trained to self-administer cocaine via lever press (0.5 mg/kg/infusion) in operant chambers in short access (ShA, 2h/day, 5 days per week) and long access (LgA, 6h/day, 5 days/week) sessions. We measured the number of cocaine infusions, or lever presses, during each session of the behavioral protocol to quantify escalation of intake, motivation, and compulsive-like behavior. Specifically, we measured escalation as the increase in the mean number of cocaine rewards during each of the LgA sessions compared to the first day of the LgA phase; we measured motivation as the mean number of cocaine rewards over the ShA and LgA sessions under a progressive ratio (PR) schedule of reinforcement, which is when the number of lever presses required to obtain a cocaine infusion was progressively increased;

and we measured compulsive-like behavior as drug taking despite adverse consequences, by pairing 30% of lever presses with an electric foot shock (Figure 1.1b). Based on individual behavioral measures for each rat (Figure 1.1c), we calculated an addiction index (AI)[39] as the average of the normalized values (z-scores) of the three behavioral measures. Prior work has demonstrated that AI measures vulnerability (high AI) or resilience (low AI) to developing cocaine addiction-like behaviors[39,51–53].

We classified rats into high and low AI groups (Figure 1.1d). Both high and low AI rats acquired fewer cocaine rewards in the ShA compared to the LgA sessions of the IVSA protocol (Figure 1.1e, two-way repeated measures ANOVA, addiction index × phase interaction $p<0.0001$, $F_{23,1012}=8.523$). There was no difference between groups in cocaine rewards during ShA sessions. However, we observed a contrasting pattern in escalation during LgA sessions. During LgA sessions, rats with high AI exhibited a progressive escalation of drug intake compared to rats with low AI as evidenced by their increased number of cocaine infusions over the course of this phase of the IVSA protocol (two-way repeated measures ANOVA interaction time x group $F_{13,572}=4.175$, $p < 0.0001$, Figure 1.1e). In contrast, low AI rats did not show escalation during the LgA sessions (Figure 1.1e).

During PR sessions, motivation for cocaine increased in the high AI rats but not in low Al rats when comparing ShA versus LgA (Figure 1.1f, mixed effect model, addiction index × phase interaction, $p=0.0049$, $F_{1,41}=8.83$; Bonferroni corrected $p=0.0001$, post hoc comparisons). Finally, high AI rats showed increased responses despite adverse consequences compared to low AI rats, as demonstrated by the higher number of cocaine infusions when the reward was paired with an electric foot shock (Figure 1.1g, $p<0.001$, unpaired Student's t-test, $t_{44}=3.936$), which may reflect compulsive-like drug use. These results show that we can capture multiple

behavioral aspects that are relevant to cocaine use disorders by using this model of extended access to cocaine IVSA in outbred rats.

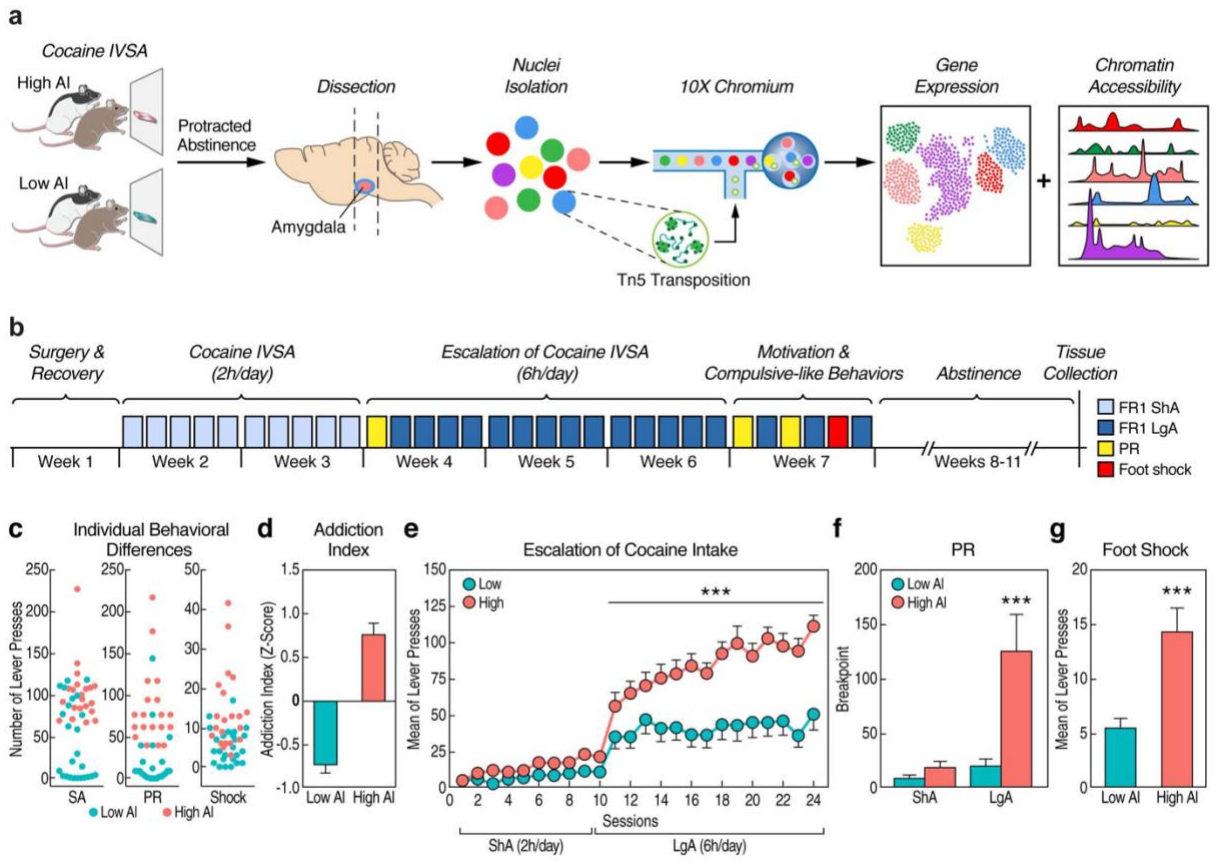


Figure 1.1: Experimental design and rat IVSA cocaine model of addiction.a) Schematic of the study design. **b)** Timeline of the behavioral protocol. **c)** Individual differences in total number of lever presses in self-administration (SA), progressive ratio (PR) and shock-paired (Shock) sessions for each rat. **d)** Mean addiction index scores in high and low AI rats. **e)** Mean number of lever presses across each ShA and LgA IVSA session in high (n=21) and low (n=25) AI rats (*** $p < 0.001$, two-way repeated measures ANOVA interaction time x group $F_{13,572}=4.175$). **f)** Breakpoint analysis of high (n=21) and low (n=25) AI rats under ShA versus LgA (*** $p<0.001$ mixed effect model, addiction index × phase interaction, $p=0.0049$, $F_{1,41}=8.83$). **g)** Mean number of lever presses when paired with electric footshock in high (n=21) and low AI (n=25) rats (***$p<0.001$, unpaired Student's t-test, $t_{44}=3.936$). Error bars in panels d-g represent the standard error of the mean.

1.3.2: snRNA-seq and snATAC-seq defines distinct populations of cell types in the amygdala

The amygdala is thought to contribute to the development of addiction through its regulation of drug-seeking behavior, which, in rats, progressively increases after withdrawal from drug IVSA[6,54]. To identify neuroadaptations that persist in the amygdala after chronic drug exposure during the withdrawal stage, we collected amygdalae after 4 weeks of abstinence from

cocaine IVSA (Figure 1.1a). We purified nuclei and measured the gene expression and chromatin accessibility profiles of individual nuclei by performing snRNA-seq and snATAC-seq with the 10X Genomics Chromium workflow. We performed these experiments on high and low AI rats, as well as naive rats never exposed to cocaine. For snRNA-seq, we used 19 rats including 6 with high AI, 6 with low AI, and 7 naive rats (Table 1.1). For snATAC-seq we used 12 rats including 4 with high AI, 4 with low AI, and 4 naive rats (Table 1.2).

After filtering low quality nuclei and potential doublets based on quality metrics, we obtained a combined total of 163,003 and 81,912 high quality nuclei from the snRNA-seq and snATAC-seq samples, respectively (Supplemental Files 1.1-1.2). Across the snRNA-seq samples, the mean reads per cell varied from 11,967 to 50,343 and the median number of detected genes ranged from 1,293 to 2,855. Across the 12 snATAC-seq samples, the median number of high-quality fragments per nucleus ranged from 7,111 to 22,018. Across samples, we observed means of 8579 and 6826 nuclei per rat in the snRNA-seq and snATAC-seq datasets. The above metrics are consistent with previously published single-nucleus sequencing studies of the amygdala[33,55]. Using these data, we performed normalization, integration across rats, dimensionality reduction and clustering using Seurat[56] (for snRNA-seq) and Signac[57] (for snATAC-seq). In total, we identified 49 cell type clusters in the integrated snRNA-seq dataset and 41 cell type clusters in the integrated snATAC-seq dataset (Supplemental Figure 1.1). Visualization of the integrated data indicated that the clustering is not influenced by batch effects such as sequencing library, percentage of mitochondrial DNA, or individual rats[58] (Supplemental Figure 1.2).

We annotated the snRNA-seq clusters based on the expression of established cell type-specific marker genes[59–63] (Figure 1.2a-b). The major cell types included excitatory neurons

(denoted by expression of *Slc17a7)*, inhibitory GABAergic neurons (*Gad1/Gad2)*, astrocytes (*Gja1)*, microglia (*Ctss)*, mature oligodendrocytes (*Cnp)*, oligodendrocyte precursor cells (OPC) (*Pdgfra)*, and endothelial cells (*Cldn5)* (Figure 1.2c). To annotate the snATAC-seq clusters, we estimated gene activity from pseudo bulk chromatin accessibility at promoter regions of cell marker genes and used these gene activity scores to impute gene expression in the snATAC-seq samples. The imputed gene expression clearly delineates the cell clusters into the same major cell types described above demonstrating strong concordance between our snRNA-seq and snATAC-seq data (Figure 1.2d). In addition to the major cell types, we also identified seven subtypes of inhibitory neurons based on the expression of known cell marker genes (Figure 1.2e). We also sub-clustered the excitatory neurons and identified 18 distinct clusters (Supplemental Figure 1.3), with top markers including known subpopulation markers such as *Cdh13*, *Nr4a2*, *Bdnf*[64].

The total number of nuclei we obtained for each cell type varied substantially (Figure 1.2f). As expected, excitatory and inhibitory neurons are the most common major cell types. The snRNA-seq dataset contains 52,579 (~32.3%) nuclei from inhibitory neurons and 23,943 (~14.7%) nuclei from excitatory neurons (Table 1.3). The snATAC-seq dataset contains 18,208 (~22.2%) nuclei from inhibitory neurons and 20,169 (~24.6%) nuclei from excitatory neurons (Table 1.3). Endothelial cells and some subtypes of inhibitory and excitatory neurons have small numbers of nuclei in the dataset, so for most downstream analyses we focused on the six most common major cell types (Figure 1.2a-b).

To determine how the cell types we identified in the whole amygdala correspond to cell types within spatially defined amygdalar subregions, we generated snRNA-seq data from the CeA and BLA. We found that cell clusters from the CeA and BLA were distinct from one

another, but these regions collectively contained most cell types also identified in the whole amygdala (Supplemental Figure 1.4a). Consistent with the known cell type composition of the CeA and BLA[65], the cell clusters from the CeA co-clustered primarily with inhibitory neurons while the cell clusters from the BLA co-clustered with excitatory neurons (Supplemental Figure 1.4b). Glial cell types from the whole amygdala contained cells from both subregions, except for astrocytes, which mostly co-clustered with cells from the CeA but not the BLA, suggesting that astrocytes might play a specific role in BLA-related function (Supplemental Figure 1.4).

In combination, the snRNA-seq and snATAC-seq datasets that we generated are the first single-cell atlas of molecularly defined cell types in the rat amygdala. The inclusion of multiple high AI, low AI, and naive rats make these datasets an important resource for studying gene expression and chromatin accessibility in the amygdala under both normal conditions as well as in the context of cocaine addiction-like behaviors.

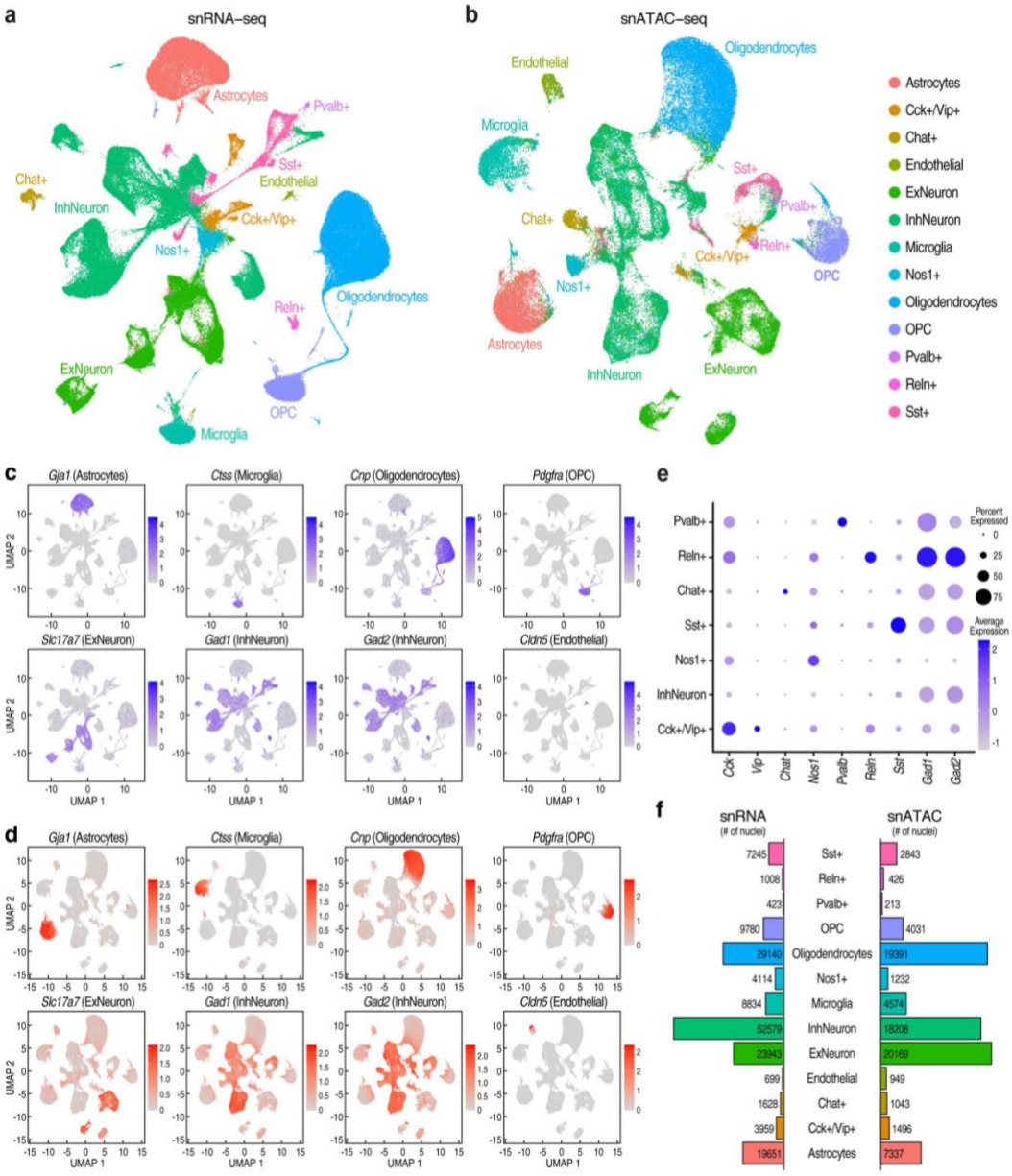


Figure 1.2: Summary of single nucleus RNA-seq and ATAC-seq data from the rat amygdala.

a) Uniform Manifold Approximation and Projection (UMAP) plot of single nucleus RNA-seq (snRNA-seq) data from the rat amygdala. Data are combined across 19 samples, with high, low, and naive addiction index labels. Cells are colored by cluster assignments performed with K-nearest neighbors. We assigned cell type labels to the clusters based on the expression of known marker genes. **b)** UMAP plot of single nucleus ATAC-seq data from 12 rat amygdala samples. snATAC-seq data was integrated with the snRNA-seq data and cluster labels were transferred to the snATAC-seq cells. **c)** Feature plot showing expression of marker genes used to label major subsets of cells: Gja1 (astrocytes), Ctss (microglia), Cnp (oligodendrocytes), Pdgfra (oligodendrocyte precursor cells (OPCs), Slc17a7 (excitatory neurons), Gad1 and Gad2 (inhibitory neurons), and Cldn5 (endothelial cells). **d)** Feature plot showing imputed gene expression of cell type-specific marker genes in snATAC-seq dataset. **e)** Expression of marker genes in cell clusters corresponding to highly specific subsets of inhibitory neurons. The shading and diameter of each circle indicate the estimated mean expression and the percentage of cells within the cluster in which the marker gene was detected. **f)** The number of nuclei assigned to each cell type cluster for the snATAC-seq and snRNA-seq datasets.

Table 1.1: Overview of rats used for snRNA-seq experiments.

Rat sample identifiers are found in RFID column.

RFID	Addiction index	batch	Estimated Number of Cells
933000320047328	High	1	6,573
933000120138592	Low	1	7,512
933000120138586	Naïve	1	8,497
933000320046084	Naïve	1	9,945
933000320046077	Naïve	2	8,021
933000120138609	Low	2	8,781
933000320186802	High	2	12,957
933000320047225	High	2	9,903
933000320046609	High	3	6,512
933000320047001	High	3	10,691
933000320047132	High	3	9,753
933000320186801	Naïve	3	7,668
933000320046621	Naïve	4	6,536
933000320046625	Naïve	4	13,217
933000320047104	Low	4	6,264
933000320045674	High	4	9,834
933000120138730	High	5	9,402
933000120138414	Low	5	8,539
933000320046549	Naïve	6	5,458

Table 1.2: Overview of rats used for snATAC-seq experiments.

Rat sample identifiers are found in RFID column.

RFID	Addiction index	Library date (batch)	Estimated number of cells
933000320186811	High	1-12-21	6491
933000320047166	Low	1-12-21	6354
933000320187130	Naïve	1-12-21	8473
933000320047651	Naïve	1-26-21	6756
933000320047161	Naïve	1-26-21	6951
933000120138730	High	11-14-19	11287
933000120138414	Low	11-14-19	15391
933000320187092	High	1-26-21	6603
933000320047019	High	2-17-21	4658
933000320046611	Low	2-17-21	3191
933000320045785	Low	2-17-21	5969
933000320047174	Naïve	2-23-21	4899

Table 1.3: Overview of cell types in snRNA-seq and snATAC-seq datasets.

Provides number of nuclei of each cell type found in each the snRNA-seq and snATAC-seq datasets, as well as the percentage of all cells in the dataset are represented by each cell type.

cluster	ncells.snRNA	percent.snRNA	ncells.snATAC	percent.snATAC
Astrocytes	19651	12.05560634	7337	8.957173552
Cck+/Vip+	3959	2.428789654	1496	1.82635023
Chat+	1628	0.9987546241	1043	1.273317707
Endothelial	699	0.4288264633	949	1.158560406
ExNeuron	23943	14.68868671	20169	24.6227659
InhNeuron	52579	32.25646154	18208	22.22873327
Microglia	8834	5.419532156	4574	5.58404141
Nos1+	4114	2.523879929	1232	1.50405313
Oligodendrocytes	29140	17.87697159	19391	23.67296611
OPC	9780	5.999889573	4031	4.921134876
Pvalb+	423	0.2595044263	213	0.2600351597
Reln+	1008	0.6183935265	426	0.5200703194
Sst+	7245	4.444703472	2843	3.470797929

1.3.3: Measuring cell type-specific differential gene expression between rats displaying a high versus a low addiction index for cocaine

We used the negative binomial test to identify differentially expressed genes (DEGs) between high and low AI rats in each cell type (Figure 1.3a-b, Supplemental File 1.3). To control for batch effects or violations in the differential expression model assumptions (for example, unaccounted for overdispersion in the data) that can cause deflated (overly significant) p-values, thereby yielding false signals of differential expression[66,67], we performed the same statistical test after permuting the AI labels of the rats. This permutation simulates a null distribution where there is no association between AI and gene expression. This approach is often used to assess p-value calibration in QTL studies[68,69]. While the results from the unpermuted data are highly enriched for low p-values, the p-values from the permuted data resemble the null expectation. This indicates that the highly significant DEGs we identified are not due to poor p-value calibration or batch effects (Supplemental Figure 1.5).

We grouped DEGs into small ($abs(log_2FC)<0.1$) or large effect size groups ($abs(log_2FC)\geq0.1$) and observed that most of the significant DEGs (FDR<10%) have small effect sizes (Supplemental Figure 1.6). In total, we identified 557 unique significant DEGs with large effect sizes in at least one cell type and 8,775 unique significant DEGs with small effect sizes in at least one cell type. The number of significant DEGs between high and low AI rats correlates with the size of the cell type population, which likely reflects greater power to detect differential expression in common cell types. Most of the large effect DEGs (431, or 75%) are also small-effect DEGs in at least one other cell type, indicating that while there are shared patterns of differential expression across cell types, the effect sizes vary across cell types. We also found that significant DEGs were enriched for gene expression quantitative trait loci (eQTLs), which

are genetic variants associated with a gene's expression, in rat brain tissues[70] in almost every cell type tested (Chi-squared test with 1 degree of freedom, $p<0.05$) (Table 1.4). This suggests that heritable differences influence the changes in expression that we observed. Among the most significant DEGs with eQTLs (Supplemental File 1.4), we identified genes with previously reported roles in cocaine or other substance use disorders. For example, *Kcnq3* was differentially regulated across neuronal and glial cell types, and this gene encodes a subunit of a potassium channel implicated in the regulation of reward behavior and susceptibility to drug addiction (Figure 1.3c)[71–73]. Additionally, *Fkbp5* and *Sgk1*, two transcriptional targets of the glucocorticoid receptor, were differentially regulated specifically in glial cell types, and these genes are associated with reward behavior and drug addiction vulnerability (Figure 1.3d-e)[74–76]. These results suggest that genetic differences contribute to the gene expression differences between rats with high and low AI.

The observed DEGs could reflect pre-existing genetic differences or the differential exposure to cocaine in the high versus low AI groups. To further examine the contribution of genetics to observed differences in gene expression, we leveraged genotypes and gene expression data from a previously published reference population of 339 naive HS rats[70]. This allowed us to predict gene expression based on cis-genetic variation in the absence of cocaine exposure. Specifically, we trained models which predict gene expression from SNP genotypes[77] using whole brain bulk RNA-seq from 339 naive HS rats. We then used these models to predict the expression of genes with at least one cis-acting eQTL (8,997 genes) in each of the rats in our snRNA-seq dataset. We compared the differences in mean predicted expression in the high versus low AI rats to the observed differences in expression for each cell type. Before correlating predicted expression to observed expression, we filtered out genes for which the predictive

models had low Pearson r^2, because genes with higher r^2 have more of their variance in expression explained by cis-genetic variation (Table 1.5). Among our major cell types, the observed and predicted expression differences were significantly correlated (Spearman's ρ, $p<0.05$) for microglia, oligodendrocytes and inhibitory neurons (Table 1.5). We found that increasing the stringency of the r^2 cutoff increased the strength of these correlations (Table 1.5). These observations indicate that genetic differences in high versus low AI rats contribute to at least some of the observed differences in expression between high vs. low AI rats. Cocaine is also likely to contribute to the differences in expression; however, the relative contributions of cocaine and genetics are difficult to quantify due to limitations in the genetic predictions of gene expression.

To identify pathways with altered regulation between high and low AI rats, we performed gene set enrichment analysis (GSEA)[78,79] of KEGG pathways using estimates of differential gene expression ($\log_2$ fold change) for each cell type. We identified significant enrichment of several pathways related to addiction (e.g. amphetamine, nicotine, and morphine addiction), neurotransmission (e.g. synaptic vesicle cycle, GABAergic synapse, glutamatergic synapse, and dopaminergic synapse), energy metabolism (e.g. glycolysis, pyruvate metabolism, and oxidative phosphorylation), and others (Figure 1.3f, Supplemental File 1.5). Most cell types showed enrichment of genes belonging to the oxidative phosphorylation pathway, which, together with glucose metabolism, is the main energy source for synaptic activity and action potentials[80,81]. Moreover, different subtypes of inhibitory neurons as well as excitatory neurons were enriched for synaptic vesicle cycle and synapse-related pathways. In combination, these observations suggest that addiction-like behaviors are associated with alterations in the metabolic state of

amygdalar cell populations, which can directly impact neural network activity within the amygdala by affecting neurotransmission and synaptic pathways.

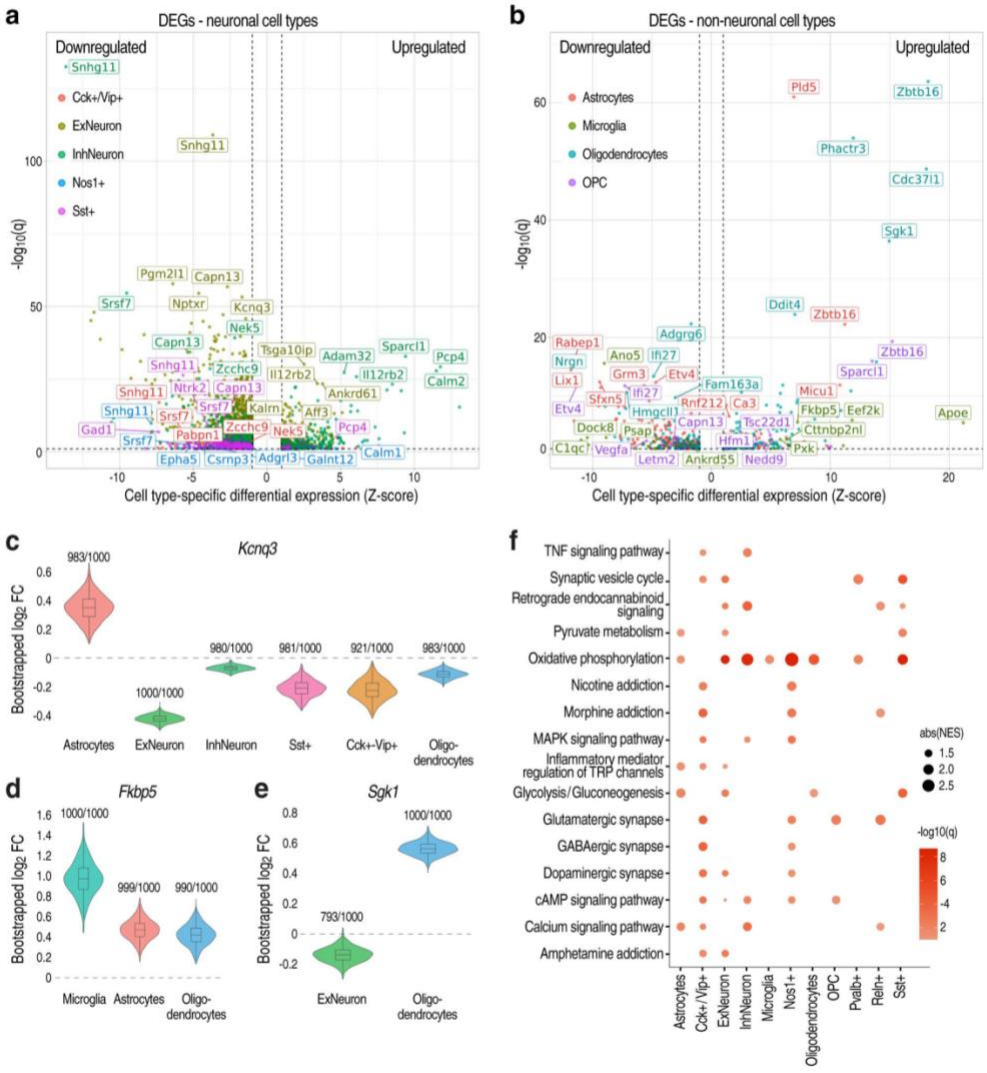


Figure 1.3: Differential gene expression between high and low AI rats.

a) Volcano plot summarizing differential gene expression between high and low AI rats. Points are colored by cell type, and the five most significant (FDR<10%) up- and downregulated genes in each cell type are indicated with labels. Within each cell type, we normalized the log fold changes reported by Seurat with a z-score and plotted the cell type-specific z-scores of the log fold changes on the x-axis (positive fold change = higher expression in high AI rats; negative fold change = higher expression in low AI rats). The -$\log_{10}$ false discovery rate (FDR) corrected p-values (q-values) are plotted on the y-axis. **b)** Volcano plot summarizing differential gene expression between high and low AI rats for non-neuronal (glial) cell type clusters. **c-e)** Violin plots showing distribution of $\log_2$FC from the negative binomial test performed in 1000 bootstrap iterations. Fractions indicate the number of bootstrap iterations in which the log2FC estimate was significantly different than 0. Bootstrap distributions were obtained for cell types in which the following genes had significant differential expression (FDR<10%): **c)** Kcnq3; **d)** Fkbp5; **e)** Sgk1. **f)** KEGG pathways that are enriched for differentially expressed genes by cell type. Size of dot indicates -log10(q) while color indicates normalized enrichment score (NES), which is a metric of GSEA. Only pathways/cell types where q<0.1 are visualized.

Table 1.4: Enrichment of DEGs with eQTLs in the rat brain.

Results of Chi-squared test with Yates' continuity correction for enrichment of significant DEGs (FDR<10%) that also have eQTLs in the rat brain in each cell type.

statistic	p.value	parameter	celltype	q.value
32.00156145	1.54E-08	1	Astrocytes	9.24E-08
5.620791222	0.017748634	1	Cck+-Vip+	0.050195845
263.0825833	3.65E-59	1	ExNeuron	3.29E-58
115.3577062	6.57E-27	1	InhNeuron	5.26E-26
8.589987545	0.003380163	1	Microglia	0.013520653
0.293805179	0.587792337	1	Nos1+	0.587792337
30.98112175	2.61E-08	1	Oligodendrocytes	1.30E-07
5.724274495	0.016731948	1	OPC	0.050195845
43.28438553	4.73E-11	1	Sst+	3.31E-10

Table 1.5: Predicted versus observed differential gene expression.

Spearman correlations (rho, pvalue) between difference in mean predicted expression and observed avg_logFC of expression between high vs. low AI rats for subsets of genes passing each Pearson r^2 cutoff for gene expression prediction models.

celltype	r2_cutoff	correlation	pvalue	ci_low	ci_high	n_genes
Astrocytes	0	0.0187	0.2834	-0.0155	0.0528	3292
Astrocytes	0.025	0.0187	0.3575	-0.0211	0.0584	2426
Astrocytes	0.05	0.0275	0.2503	-0.0194	0.0742	1754
Astrocytes	0.075	0.0369	0.1705	-0.0159	0.0894	1383
Astrocytes	0.1	0.0398	0.1804	-0.0185	0.0978	1133
ExNeuron	0	0.0253	0.1460	-0.0088	0.0595	3292
ExNeuron	0.025	0.0347	0.0876	-0.0051	0.0744	2426
ExNeuron	0.05	0.0427	0.0737	-0.0041	0.0893	1754
ExNeuron	0.075	0.0488	0.0698	-0.0039	0.1012	1383
ExNeuron	0.1	0.0500	0.0926	-0.0083	0.1079	1133
InhNeuron	0	0.0336	0.0538	-0.0006	0.0677	3292
InhNeuron	0.025	0.0465	0.0219	0.0067	0.0862	2426
InhNeuron	0.05	0.0649	0.0065	0.0181	0.1114	1754
InhNeuron	0.075	0.0574	0.0329	0.0047	0.1097	1383
InhNeuron	0.1	0.0646	0.0297	0.0064	0.1224	1133
Microglia	0	0.0356	0.0409	0.0015	0.0697	3292
Microglia	0.025	0.0603	0.0030	0.0206	0.0999	2426
Microglia	0.05	0.0626	0.0088	0.0158	0.1091	1754
Microglia	0.075	0.0620	0.0211	0.0093	0.1143	1383
Microglia	0.1	0.0507	0.0882	-0.0076	0.1086	1133
OPC	0	0.0154	0.3760	-0.0187	0.0496	3292
OPC	0.025	0.0139	0.4944	-0.0259	0.0536	2426
OPC	0.05	0.0128	0.5915	-0.0340	0.0596	1754
OPC	0.075	0.0125	0.6435	-0.0403	0.0651	1383
OPC	0.1	-0.0028	0.9244	-0.0611	0.0554	1133
Oligodendrocytes	0	0.0620	0.0004	0.0279	0.0960	3292
Oligodendrocytes	0.025	0.0741	0.0003	0.0344	0.1136	2426
Oligodendrocytes	0.05	0.0841	0.0004	0.0374	0.1304	1754
Oligodendrocytes	0.075	0.0970	0.0003	0.0445	0.1489	1383
Oligodendrocytes	0.1	0.1026	0.0005	0.0447	0.1599	1133

1.3.4: The development of cocaine addiction-like behaviors is linked to elevated GABAergic transmission in the amygdala

To test the hypothesis that altered metabolic state in amygdalar cells changes neural activity within the amygdala, we focused on GABAergic transmission because alterations of this neurotransmitter system have been previously described in the amygdala in the context of addiction-related phenotypes[2]. Specifically, we measured GABAergic transmission by recording spontaneous inhibitory postsynaptic currents (sIPSCs) in the central amygdala (CeA). CeA slices were collected from a separate cohort of 5 low AI and 5 high AI HS rats that were subjected to prolonged abstinence following the same behavioral protocol described for the snRNA-seq and snATAC seq experiments (Figure 1.4a). As a control, we used CeA slices prepared from 5 age-matched naive HS rats to record baseline GABAergic transmission. There were differences in mean sIPSC frequencies among the groups (one-way ANOVA $F_{2,22}$=6.77, p=0.0051), reflecting a progressive increase in GABAergic transmission from naive to low AI to high AI rats (Figure 1.4b, Supplemental Figure 1.7a), without detectable changes in amplitude (Supplemental Figure 1.7b-c). These results are consistent with the hypothesis that the cocaine addiction-like behaviors exhibited by high AI rats alters GABAergic transmission.

To further investigate the link between GABAergic transmission and energy metabolism in the amygdala with cocaine addiction-like behaviors, we measured siPSCs frequency and amplitude before and after application of S-bromobenzylglutathione cyclopentyl diester (pBBG)[82,83]. pBBG is a small molecule inhibitor of glyoxalase 1 (GLO1), the rate limiting enzyme for the metabolism of methylglyoxal (MG), which is a byproduct of glycolysis that acts as a competitive partial agonist of $GABA_A$ receptors[82]. We found that pBBG reduced the sIPSC frequency compared to vehicle for both high and low AI rats (paired t-tests, t_5=11.83, p=7.6e-5

and t_5=5.07, p=3.9e-3, respectively), but not naive rats (t_5=0.71, p=0.51) (Figure 1.4c-f, Supplemental Figure 1.7a). We observed no effect on iPSCs amplitude (Supplemental Figure 1.7b-c). In most situations, changes in frequency of events indicate presynaptic modulation while changes in amplitude of events reflect postsynaptic modulation; however, previous studies have shown that GABA modulates synaptic transmission presynaptically[84,85]. These findings suggest that Glo1 inhibition may alter presynaptic GABA-A receptor function, leading to reduced GABA release at inhibitory terminals and suppression of inhibitory connections within the CeA.

These results led us to hypothesize that GLO1 inhibition would revert behavioral responses after prolonged abstinence from cocaine IVSA. To test this hypothesis, we measured cue-induced reinstatement of cocaine seeking behavior in a separate cohort of 26 low and high AI rats 30 minutes after systemic injection of pBBG or vehicle[86] following 4 weeks of abstinence from cocaine IVSA (Figure 1.4g). During this test, rats were subjected to the same operant conditions of cocaine IVSA, but without drug availability. Then, reinstatement was triggered by re-exposure to the cocaine infusion-associated light cue. The two-way repeated measures ANOVA showed a significant interaction between AI and pBBG treatment ($F_{1,24}$=6.609, p<0.05), indicating that pBBG versus vehicle reduced cue-induced reinstatement in high AI rats (p-value<0.05, post hoc comparisons with Bonferroni correction), but not in low AI rats (p>0.05). Overall, these results demonstrate that modulating $GABA_A$ transmission via the pharmacological inhibition of GLO1 decreases relapse-like behaviors in animals with high cocaine AI.

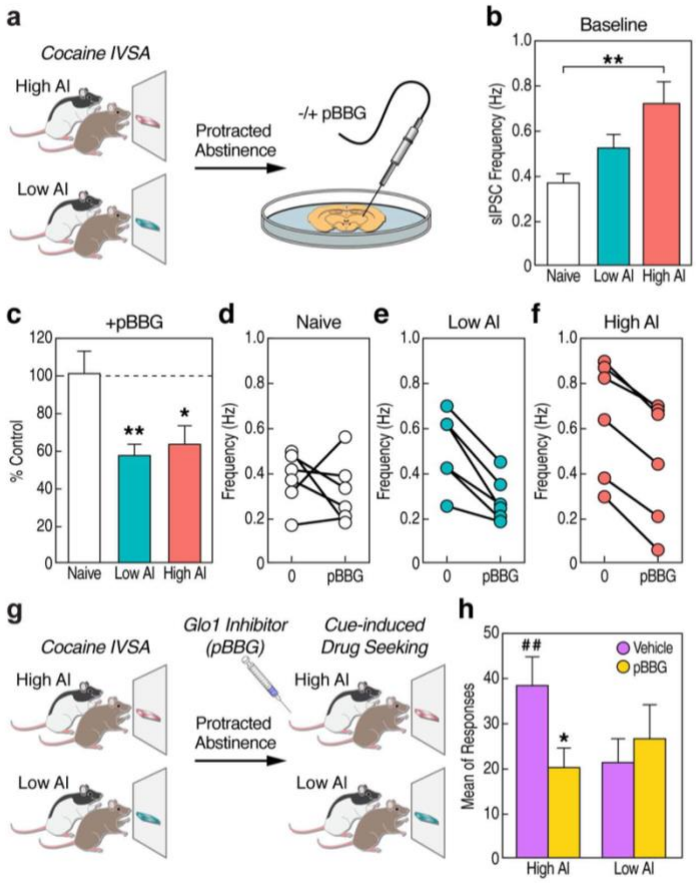


Figure 1.4: Electrophysiology and GLO1 inhibition experiments implicate GABAergic inhibition in cocaine addiction-like behaviors.

a) Schematic showing animal model used for electrophysiology recording in CeA slices from HS rats subjected to 4 weeks of abstinence from cocaine IVSA. Electrophysiological recordings were taken before and after pBBG (S-bromobenzylglutathione cyclopentyl diester) treatment. **b)** Baseline iPSC frequency (measured before pBBG injection). Significant differences in the means between the three groups were observed (**$p < 0.01$; one-way ANOVA $F_{2,22}=6.77$, followed by post-hoc comparison using Tukey's HSD). **c)** sIPSC frequency following pBBG treatment. We observed reduced frequency in the CeA slices from high and low AI rats following pBBG treatment (**$p < 0.01$, *$p<0.05$ following Bonferroni correction; unpaired two-sided Student's t-test). Change in sIPSC frequency following pBBG treatment in **d)** naive, **e)** low, and **f)** high rats. **g)** Schematic of animal model used to test cocaine-seeking behavior. Rats were injected with pBBG following a period of prolonged abstinence and re-exposed to the self-administration chambers in the absence of cocaine. **h)** Following injection of pBBG, high AI rats (n=12) showed significantly higher cocaine-seeking behavior compared to low AI rats (n=14), which was reduced by pBBG treatment (## $p<0.001$, *$p<0.05$ following Bonferroni correction; two-way ANOVA for each measure). Error bars in panels b,c, and h represent the standard error of the mean.

1.3.5: Mapping differences in chromatin accessibility associated with cocaine addiction-like behaviors

To identify regions of accessible chromatin from the snATAC-seq data, we used MACS2[87] to call peaks from the aligned reads for each rat and created a union peak set across

rats. We examined pseudo bulk chromatin accessibility at the transcription start site (TSS) of selected cell type marker genes and observed cell type-specific patterns of accessibility at the expected marker genes of each cell type (Figure 1.5a, Figure 1.2c-d), indicating that the chromatin accessibility corresponds well with the transcriptome measurements.

To better understand the regulatory mechanisms involved in cocaine addiction, we analyzed differences in chromatin accessibility between high and low AI rats. We performed negative binomial[88,89] tests to measure cell type-specific differential chromatin accessibility (Supplemental File 1.6) and compared the observed p-values to those obtained from permuted data (as we did for our DEG analysis). The p-values of the permuted data resemble the null expectation, confirming that the differential peaks between high and low AI are likely true biological differences rather than artifacts (e.g. batch effects) (Supplemental Figure 1.8). In total we identified >20,000 peaks across cell types with significant differential accessibility between the high and low AI groups (FDR<10%) (Supplemental Figure 1.9); however, as with gene expression, most differences were small ($\log_2 FC < 0.1$). This indicates that differences in addiction-like behaviors between rats are associated with modest regulatory changes at a large number of sites.

The differential peaks can be subdivided into those where accessibility is higher (upregulated) or lower (downregulated) in the high AI rats (Supplemental Figure 1.9). In astrocytes, there were roughly equal numbers of up- and downregulated peaks, but the other cell types showed profound directional biases. Excitatory neurons were the most biased with only two detected downregulated peaks, and over 8000 upregulated peaks in the high AI group. Inhibitory neurons showed the opposite bias with over 4000 downregulated peaks but only ~500

upregulated peaks in the high AI group (Supplemental Figure 1.9). These biases likely reflect differences in the activity of specific TFs that control large transcriptional programs.

To determine whether the differential chromatin accessibility is consistent with the differential gene expression, we tested whether the promoters of DEGs are enriched for differential accessibility. We overlapped the significant differentially accessible chromatin peaks in each cell type with the promoters of significant DEGs and computed a log odds ratio (log_2OR) as a measure of enrichment. Across all the major cell types, differentially accessible peaks are enriched (FET, $p<0.05$) at the promoters of DEGs compared to non-DEGs (Figure 1.5b, Table 1.6). We also examined chromatin accessibility at promoter regions for genes belonging to the oxidative phosphorylation pathway because genes within this pathway were enriched for gene expression differences between high vs. low AI rats in most cell types. These genes are also significantly enriched for differentially accessible promoter peaks in inhibitory neurons, excitatory neurons, and oligodendrocytes (Table 1.7). These findings confirm that the observed differences in chromatin accessibility and gene expression are concordant.

The genomic annotations of the significant differential peaks showed that 3.19% of these regions were annotated as promoter or TSS regions (Supplemental Figure 1.10). While this is a small percentage of the peaks, it is consistent with other studies[26]. We then studied the subset of significant differential peaks in each cell type by examining their genomic annotations to determine if they were enriched for promoter/TSS regions compared to the set of all peaks. We observed that differentially accessible peaks were highly enriched in promoter regions, occurring at least four times more frequently than expected given the genomic annotations of all accessible chromatin regions in most of the major cell types (FET, FDR<10%) (Figure 1.5c, Table S1.8). This enrichment may indicate that long-term changes in chromatin associated with addiction-like

behaviors are more concentrated at promoters, or that we have greater statistical power to detect changes at promoters, due to larger effect sizes or greater overall chromatin accessibility.

We hypothesized that differences in chromatin accessibility between high and low AI rats are caused by differential TF activity. To test this hypothesis, we analyzed the snATAC-seq data using ChromVar (Supplemental File 1.7), which identifies TF motifs associated with differential accessibility using sparse single cell data[90]. Many motifs showed significant differences in accessibility between the high and low AI rats, and since many TFs recognize similar motifs, we grouped them into motif clusters and summarized results across cell types (Figure 1.5d).

The motif cluster with the most significant difference in accessibility between high and low AI rats contains motifs for basic helix-loop-helix (bHLH) TFs. This motif cluster has substantially higher accessibility within the excitatory neurons of high AI rats compared to low AI rats (deviance 3.8, p=1e-280), as well as a modest increase in accessibility in inhibitory neurons (deviance 0.38, p=1e-34) (Figure 1.5e-g). The top-ranked motifs in this cluster all harbor the sequence CAGATGG, which is a close match to binding site motifs for multiple neuronal pioneer TFs including those of the bHLH, RFX and FOX families[91,92]. Thus, the widespread increases in chromatin accessibility in excitatory neurons of high AI rats could reflect increased activity of pioneer TFs that recruit chromatin remodelers. However, we did not observe corresponding upregulation in the expression of genes encoding the TFs belonging to these clusters (Supplemental Files 1.3 and 1.7), suggesting that a different mechanism might affect their activity.

We noticed that many motif clusters with increased accessibility in the neurons of high AI rats have decreased accessibility in oligodendrocytes (Figure 1.5d-g). Prominent among these motif clusters are those containing FOX and RFX motifs (Figure 1.5d-g).

Several motif clusters also have opposite effects between excitatory and inhibitory neurons. SOX motifs have decreased accessibility in high AI rats in excitatory neurons but increased accessibility in all other major cell types including inhibitory neurons (Figure 1.5d). MEF2 and Fos (AP1) motifs all have increased accessibility in the excitatory neurons of high AI rats but decreased accessibility in inhibitory neurons (Figure 1.5d). AP1 and MEF2 motifs are of particular interest because they are associated with addiction[93–96] and their expression changes in the brain following chronic exposure to cocaine and other drugs[97–101]. Consistent with these results, we observed that the expression of TFs of the AP1, including Fosl1, Fos, Jun, Junb, and Jund, was decreased in high versus low AI rats (Supplemental Figure 1.11), suggesting that differences in their expression level affect their transcriptional activity. While our analysis cannot pinpoint the precise TFs involved, it implicates many motif clusters that are associated with addiction-like behaviors across thousands of regulatory regions and in a cell type-specific manner.

Accessible chromatin regions harbor cell type-specific regulatory elements[102,103], and enrichment analyses that measure intersections between ATAC-seq peaks and GWAS signals can yield insight into the mechanisms by which genetic variants confer risk[104]. However, cell type-specific measurements of chromatin accessibility are difficult to obtain from human brain tissues. To assess whether our rat snATAC-seq data is meaningful for interpreting human addiction-related traits, we mapped the accessible chromatin peaks to the human reference genome and performed cell type-specific LD score regression[105]. We chose to use summary statistics from well-powered GWAS for alcohol and tobacco use[106,107] because there is significant genetic overlap among GWAS for all known substance use disorders[108] and because available GWAS for cocaine use disorder are much smaller and less powerful. We found

significant enrichments (FDR<10%) of SNP heritability in every trait tested in almost every cell type (Figure 1.5h), with the most significant enrichments in neurons, astrocytes, oligodendrocytes and OPCs. Overall, these results support the hypothesis that, despite the millions of years of evolution separating humans and rats, the regulatory architecture of HS rats is relevant for human addiction-related traits.

Figure 1.5: Analysis of chromatin accessibility and regulatory elements involved in cocaine dependence.

a) Pseudobulk chromatin accessibility at the promoter regions of marker genes for major cell types identified in our analysis. **b)** Significant DEGs (FDR<10%) for each major cell type are enriched for promoters with differentially accessible chromatin accessibility (FDR<10%; Fisher's exact test) in the snATAC-seq data. This indicates that the snRNA-seq and snATAC-seq results are consistent and indicate that long-term transcriptional changes are associated with changes in chromatin accessibility of gene promoters. **c)** Cell type-specific differentially accessible peaks (FDR<10%; Fisher's exact test) are enriched in TSS/promoter regions compared to non-TSS/promoter regions. Error bars in b,c represent 95% confidence intervals for log2 odds ratios (ORs). **d)** Heatmap showing differential activity of various motifs in the significant differential peaks of each cell type. Values indicate average difference of chromVar deviation scores with -log10(p) in parentheses below. There are many cases where motifs display increased activity in the peaks which are more accessible in upregulated peaks in neurons while also displaying decreased activity in downregulated peaks in oligodendrocytes. **e-g)** Volcano plots showing average difference (x-axis) and -log10(q) (y-axis) of chromVAR deviation scores for top 50 motif clusters in **e)** excitatory neurons, **f)** inhibitory neurons, and **g)** oligodendrocytes. **h)** LD score regression results showing significance (-log10p) of enrichment of heritability for several traits related to alcohol and nicotine addiction in cell type-specific accessible chromatin regions (mapped to hg19).

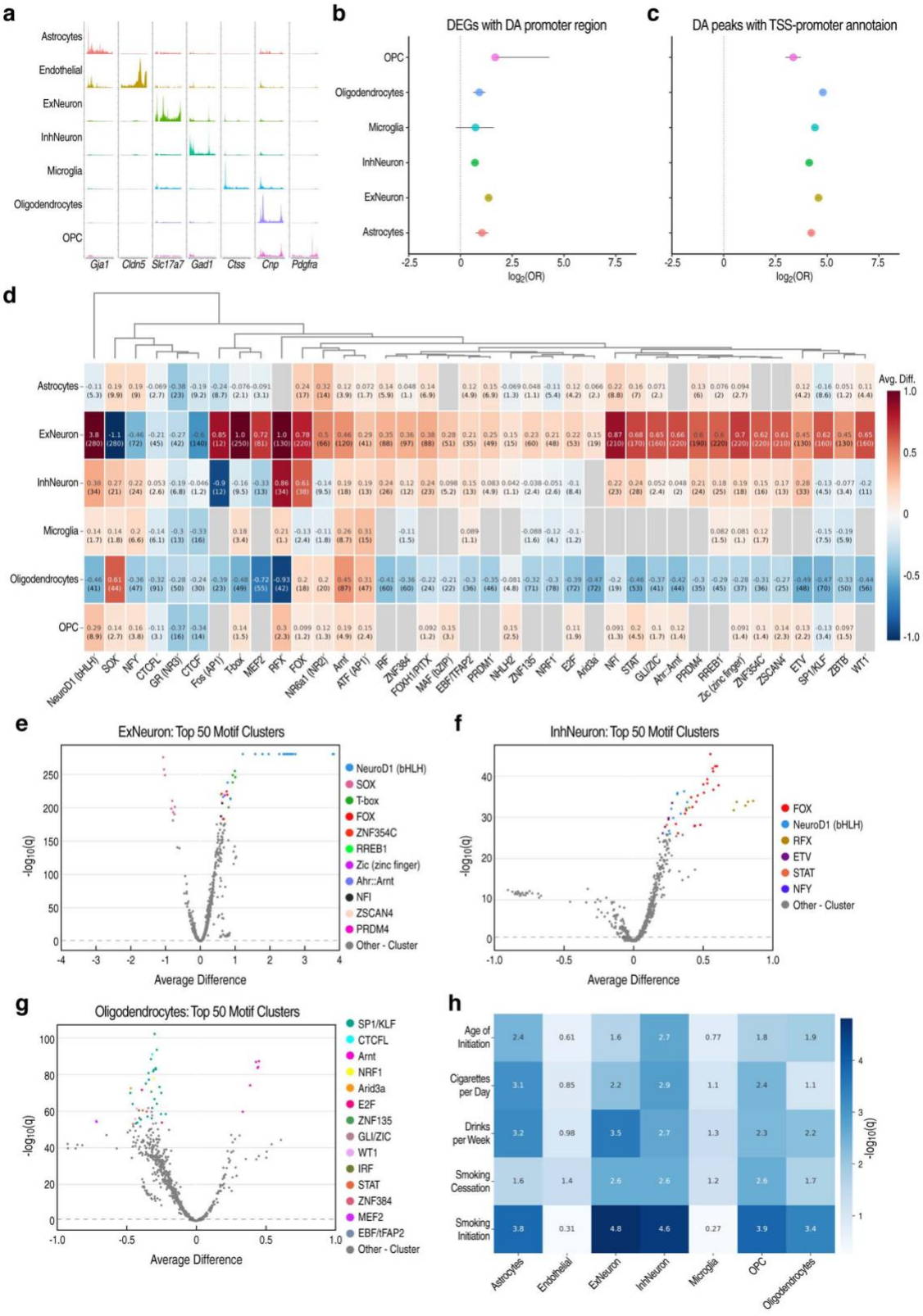

a
Astrocytes
Endothelial
ExNeuron
InhNeuron
Microglia
Oligodendrocytes
OPC
Gja1
Cldn5
Slc17a7
Gad1
Ctss
Cnp
Pdgfra
b
DEGs with DA promoter region
OPC
Oligodendrocytes
Microglia
InhNeuron
ExNeuron
Astrocytes
log2(OR)
c
DA peaks with TSS-promoter annotaion
log2(OR)
d
Avg. Diff.
NeuroD1 (bHLH)
SOX
NFY
CTCFL
GR (NR3)
CTCF
Fos (AP1)
T-box
MEF2
RFX
FOX
NR6a1 (NR2)
Arnt
ATF (AP1)
IRF
ZNF384
FOXH1/PITX
MAF (bZIP)
EBF/TFAP2
PRDM1
NHLH2
ZNF135
NRF1
E2F
Arid3a
NFI
STAT
GLI/ZIC
Ahr::Arnt
PRDM4
RREB1
Zic (zinc finger)
ZNF354C
ZSCAN4
ETV
SP1/KLF
ZBTB
WT1
e
ExNeuron: Top 50 Motif Clusters
-log10(q)
Average Difference
NeuroD1 (bHLH)
SOX
T-box
FOX
ZNF354C
RREB1
Zic (zinc finger)
Ahr::Arnt
NFI
ZSCAN4
PRDM4
Other - Cluster
f
InhNeuron: Top 50 Motif Clusters
-log10(q)
Average Difference
FOX
NeuroD1 (bHLH)
RFX
ETV
STAT
NFY
Other - Cluster
g
Oligodendrocytes: Top 50 Motif Clusters
-log10(q)
Average Difference
SP1/KLF
CTCFL
Arnt
NRF1
Arid3a
E2F
ZNF135
GLI/ZIC
WT1
IRF
STAT
ZNF384
MEF2
EBF/tFAP2
Other - Cluster
h
Age of Initiation
Cigarettes per Day
Drinks per Week
Smoking Cessation
Smoking Initiation
Astrocytes
Endothelial
ExNeuron
InhNeuron
Microglia
OPC
Oligodendrocytes
-log10(q)

Table 1.6: Enrichment of DEGs with differentially accessible promoter regions.

Results of two-sided Fisher's exact test measuring enrichment of DEGs with differentially accessible promoters.

estimate	p.value	conf.low	conf.high	celltype	q.value
18.81516404	0	17.77574265	19.9175995	Astrocytes	0
30.37291625	0.005840132	2.201548611	415.3908706	Endothelial	0.005840132
23.96758633	0	22.74043054	25.23646105	ExNeuron	0
17.57693406	0	16.48571015	18.74700536	InhNeuron	0
21.28725044	0	19.94126124	22.71783229	Microglia	0
27.72361083	0	26.31595904	29.21625885	Oligodendrocytes	0
10.28809974	7.95E-48	7.889775684	13.29876263	OPC	1.06E-47
30.37610297	0.000602346	4.068618798	226.6446137	Pvalb+	0.000688396

Table 1.7: Enrichment of genes belonging to oxidative phosphorylation pathway with differentially accessible promoter regions.

Results of two-sided Fisher's exact test measuring enrichment of differentially accessible promoter regions (FDR<10%) in genes belonging to the oxidative phosphorylation pathway.

estimate	p.value	conf.low	conf.high	celltype
1.590213356	0.027413877	1.036306859	2.4171411	Astrocytes
0	1	0	Inf	Cck+-Vip+
0	1	0	Inf	Chat+
0	1	0	542.2633381	Endothelial
2.603860013	2.75E-06	1.705313547	4.02696631	ExNeuron
2.077830782	0.00082396	1.33980214	3.17681751	InhNeuron
1.50676579	0.06905433	0.945014063	2.350921065	Microglia
0	1	0	181.0640654	Nos1+
3.044756686	6.43E-08	1.985216478	4.741095202	Oligodendrocytes
1.879741784	0.294615891	0.222019752	7.117370798	OPC
0	1	0	106.9419357	Pvalb+
0	1	0	Inf	Sst+

Table 1.8: Enrichment of differentially accessible peaks with TSS/promoter annotations.

Results of two-sided Fisher's exact test measuring enrichment of differential peaks with TSS/promoter annotations.

estimate	p.value	conf.low	conf.high	celltype	q.value
18.81516404	0	17.77574265	19.9175995	Astrocytes	0
30.37291625	0.005840132	2.201548611	415.3908706	Endothelial	0.005840132
23.96758633	0	22.74043054	25.23646105	ExNeuron	0
17.57693406	0	16.48571015	18.74700536	InhNeuron	0
21.28725044	0	19.94126124	22.71783229	Microglia	0
27.72361083	0	26.31595904	29.21625885	Oligodendrocytes	0
10.28809974	7.95E-48	7.889775684	13.29876263	OPC	1.06E-47
30.37610297	0.000602346	4.068618798	226.6446137	Pvalb+	0.000688396

1.4: Discussion

To better understand the molecular basis of addiction and illuminate long-term changes in gene regulation and chromatin accessibility associated with chronic drug use, we have generated an atlas of single-cell gene expression and chromatin accessibility in the amygdala of rats that showed divergent cocaine addiction-like behaviors after a prolonged period of abstinence. Our dataset is the largest resource of cell types in the mammalian amygdala, with over 163,000 nuclei in our snRNA-seq dataset and 81,000 nuclei in our snATAC-seq dataset (Figure 1.2a-b, Supplemental Files 1.1-1.2). The snATAC-seq dataset provides the first map of cell type-specific regulatory elements in the amygdala, which has allowed us to identify TF motifs that may drive addiction-related processes.

Previous single cell transcriptomic studies have focused on the effects of acute passive treatment with psychoactive drugs in rodents[37,38], which cannot fully capture the motivational processes underlying addiction. In contrast, our behavioral protocol involves extended access to cocaine IVSA and reflects several key aspects of cocaine addiction, including escalation of drug

use, enhanced motivation for drug seeking and taking, and persistent drug use despite adverse consequences, which might reflect compulsive-like drug consumption[109]. In addition, using an outbred population of rats with divergent addiction-like traits allowed us to correlate molecular differences not only a high AI phenotype, which reflects vulnerability to severe addiction-like phenotypes, but also to a low AI phenotype, which reflects resiliency to developing behaviors that are hallmarks of addiction. Thus, our study is the first to examine long-term molecular changes in distinct brain cell populations following abstinence from chronic voluntary cocaine use in vulnerable and resilient rats.

One striking finding from our study is that there were thousands of significant cell type-specific differences in gene expression and chromatin accessibility between high and low AI rats, with strong biases in the direction of regulation of open chromatin regions in several major cell types (Supplemental Figures 1.6 and 1.9). Most of these differences were small, which suggests that cocaine addiction-related behaviors may result from the combined action of many small effects on gene expression and chromatin accessibility. Because the HS rats are genetically diverse, the molecular differences between high and low AI rats could arise from genetic differences or from the consumption of different quantities of cocaine. These results are consistent with a polygenic model wherein addiction-like behaviors would result from the collective action of a large number of genetic risk loci with small individual effects. This is a plausible explanation because of the high genetic diversity in the HS rats and because complex traits in humans are highly polygenic[105,110]. Further support for this hypothesis comes from the observation that the majority of DEGs have eQTLs identified in HS rat brains[70] (Supplemental File 1.4), including DEGs such as *Kcnq3, Fkbp5* and *Sgk1* (Figure 1.3a-e).

Alternatively, the effects could be mediated by a relatively small number of TFs that affect many downstream genes and chromatin sites. Because some of the motifs with the strongest chromatin accessibility differences (Figure 1.5e-h) are recognized by pioneer TFs (e.g., BHLH, Sox, Fox), it is tempting to speculate that widespread differences in accessibility are due to pioneer TFs, which have an intrinsic ability to modify chromatin[111]. These explanations are not mutually exclusive, and it is likely that some differences are caused by eQTLs while others are caused by differences in the activity of upstream regulators (which themselves may be affected by genetics or other factors).

In an effort to uncouple pre-existing genetically controlled gene expression differences from cocaine-induced neuroadaptations, we performed an analysis comparing our observed DEGs to differences in expression obtained from genotype-based prediction models (Table 1.5). This allowed us to leverage the genotype data from our cohort of genetically diverse HS rats. We found significant correlations in observed versus predicted differential gene expression between high vs. low AI rats, suggesting that genetics does play a role in long-term transcriptional neuroadaptations that are observed following cocaine use. While the correlation metrics we obtained from our analysis are modest, this is expected due to three limitations of the predictive model. First, the models are trained on whole brain tissue and do not have the same cell type-specific resolution as our snRNA-seq data. Second, the size of the cohort on which the predictive models were trained was quite modest. Third, the models are only capable of capturing a small fraction of variation in expression and do not account for other influences on gene expression. Finally, it is likely that the DEGs we discovered are biased towards highly expressed genes, and eQTLs are less detectable in genes with very low expression. While these limitations make it difficult to quantify how much of the variance in expression is due to genetics, it establishes that

at least some of the differences are due genetic variation (Table 1.5). To properly uncouple pre-existing genetically controlled gene expression differences from cocaine-induced neuroadaptations would require a larger dataset of genotyped rats. One way this could be accomplished is by using polygenic risk scores for addiction-related traits, which will become possible as more rat behavioral GWAS are completed[43,45–47,112].

Human and animal studies have provided genetic and behavioral evidence that alterations in GABAergic transmission are involved in addiction[2,113–117]. Consistent with prior findings showing that GABAergic transmission is enhanced following excessive cocaine use[118], our differential gene expression analysis showed enrichment of genes belonging to the GABAergic synapse pathway (Figure 1.3f) and our electrophysiology results provided evidence for enhanced GABAergic transmission in the high AI rats (Figure 1.4b). Moreover, we found that inhibition of GLO1, the enzyme responsible for MG metabolism, normalizes electrophysiological (Figure 1.4c-f) and behavioral differences (Figure 1.4h) associated with severe addiction-like behaviors. Specifically, while pBBG normalized the increased GABA transmission in electrophysiological recordings for both low and high AI rats (Figure 1.4c), it had a normalizing effect on the drug-seeking behaviors in high AI rats but not low AI rats (Figure 1.4h). This suggests that the inhibitory effects of pBBG on relapse-like behaviors depend on a given threshold of GABAergic transmission. These results corroborate previous findings that MG acts as an endogenous competitive agonist for $GABA_A$ receptors[113,119]. $GABA_A$ receptor agonists used in the context of cocaine-seeking behavior have shown contrasting results leading to both reduction and increase in cocaine-seeking behaviors[120–128]. Since MG is generated in proportion to glycolytic activity of nearly every cell and does not accumulate in synaptic vesicles, it diffuses and may activate $GABA_A$ receptors at synaptic and extra synaptic sites; thus, manipulating the endogenous levels

of MG by GLO1 inhibition probes a mechanism of $GABA_A$ receptor regulation that is fundamentally different from the canonical modulation of synaptic $GABA_A$ receptors. In our study, electrophysiological recordings suggest that there is an increase in GABAergic transmission without changes in postsynaptic currents in the CeA; thus, we speculate that MG-based pharmacological manipulations may alter presynaptic $GABA_A$ receptor function, reducing GABA release at inhibitory terminals and suppressing inhibitory connections within the CeA. Consistently, previous studies demonstrated that the activation of presynaptic $GABA_B$ receptors suppresses inhibitory connection within the CeA[84] and that negative regulation of GABAergic transmission can occur through a presynaptic mechanism[85]. An alternative scenario is that the magnitude of effects is not sufficient to cause detectable changes in amplitude. Overall, these results offer a promising pharmacological target for improving therapeutic approaches for cocaine addiction that was identified by our single cell analysis of the amygdala in high and low AI rats.

While the pharmacological inhibition experiments are not cell type-specific, the pathway enrichment analysis of the transcriptomic data suggest that GABAergic synapse-related genes may be specific to Cck+Vip+ and Nos1+ subtypes of inhibitory neurons. Previous studies manipulating GLO1 activity directly in the mouse amygdala by transgenic expression of *Glo1* or MG microinjection were sufficient to reduce anxiety-like behaviors[129]. Future experiments targeting specific subregions or cell types of the amygdala will be necessary to further characterize the effects of GLO1 inhibition on cocaine addiction-related phenotypes.

The results from the GLO1 inhibition experiments indicate a close connection between localized energy metabolism and neurotransmission[130]. Moreover, genes which are differentially regulated in high versus low AI rats are enriched in pathways related to energy metabolism,

including glycolysis, pyruvate metabolism, and oxidative phosphorylation (Figure 1.3f). Most notably, the expression levels of genes related to oxidative phosphorylation, which determines cellular ATP levels[131], are altered across most amygdalar cell types. Not only is ATP crucial for sustaining electrophysiological activity and cell signaling in the brain[132,133], but it is also required for ATP-dependent chromatin remodeling events initiated by pioneer TFs[134]. This could potentially explain why excitatory and inhibitory neurons show opposite directions of regulation in chromatin accessibility (Supplemental Figure 1.9) and in the enrichment of DEGs in the oxidative phosphorylation pathway (Figure 1.3f). In combination, these observations suggest that an altered metabolic state within the amygdala impacts multiple cellular processes that are involved in vulnerability to and development of addiction. Future experiments that directly manipulate the expression of specific metabolic enzymes or pioneer TFs in a cell type-specific manner will be necessary to fully elucidate their role in addiction.

In conclusion, the cellular atlas created by this study is a valuable resource for understanding cell type-specific gene regulatory programs in the amygdala and their role in the development of cocaine addiction-related behaviors. Our results emphasize the contribution of cellular energetics and the $GABA_A$-mediated signaling to the enduring effects of cocaine use, which led us to perform experiments that manipulate $GABA_A$ transmission via the pharmacological inhibition of GLO1 and identify a novel potential target for treatment of cocaine addiction. We anticipate that future studies will utilize our data to further investigate novel cell type-specific mechanisms involved in addiction.

1.5: Methods

1.5.1: Animals

All protocols were reviewed and approved by the institutional Animal Care and Use Committee at the University of California San Diego. HS rats (Rat Genome Database NMcwiWFsm #13673907, sometimes referred to as N/NIH) which were created to encompass as much genetic diversity as possible at the NIH in the 1980's by outbreeding eight inbred rat strains (ACI/N, BN/SsN, BUF/N, F344/N, M520/N, MR/N, WKY/N and WN/N) were provided by Dr. Leah Solberg Woods (Wake Forest University School of Medicine). To minimize inbreeding and control genetic drift, the HS rat colony consists of 64 or more breeder pairs and is maintained using a randomized breeding strategy, with each breeder pair contributing one male and one female to subsequent generations. To keep track of the rats, their breeding, behavior, organs and genomic info, each rat received a chip with an RFID code. Rats were shipped at 3-4 weeks of age, kept in quarantine for 2 weeks and then housed two per cage on a 12 h/12 h reversed light/dark cycle in a temperature (20-22°C) and humidity (45-55%) controlled vivarium with ad libitum access to tap water and food pellets (PJ Noyes Company, Lancaster, NH, USA). We used 46 HS rats for the behavioral experiments presented in Figure 1.1, of which 20 male rats (high and low AI) were used for the generation of snRNA-seq (Table 1.1) and snATAC-seq data (Table 1.2), along with 11 naive male rats (Tables 1.1 and 1.2). Additionally, 26 of these 46 behaviorally phenotyped rats (13 female, 13 male) were used for the cue-induced reinstatement experiments. For snRNA-seq, we used 19 male rats (6 high AI, 6 low AI, 7 naive) (Table 1.1). For the snATAC-seq, we used 12 male rats (4 high AI, 4 low AI, 4 naive) (Table 1.2). In addition, we used a different cohort of 15 female and male rats (5 high AI, 5 low AI, 5 naive) for the electrophysiology experiments.

1.5.2: Drugs

Cocaine HCl (National Institute on Drug Abuse, Bethesda, MD, USA) was dissolved in 0.9% saline (Hospira, Lake Forest, IL, USA) and administered intravenously at a dose of 0.5 mg/kg/infusion as described below. pBBG was synthesized in the laboratory of Prof. Dionicio Siegel (University of California San Diego, Skaggs School of Pharmacy and Pharmaceutical Sciences). pBBG was dissolved in a vehicle of 8% dimethylsulfoxide, 18% Tween-80, and 74% distilled water and administered intraperitoneally 30 minutes before the test session.

1.5.3: Brain Samples

Brain tissues were obtained from the cocaine brain bank at UCSD[39], which collects tissues from HS rats that are part of an ongoing study of addiction-like behavior[43]. We used 31 HS rats for generating the single-nucleus sequencing data reported in this study, which included 20 rats that had low or high AI for cocaine addiction-related behaviors, using behavioral methods previously described[48] were kept in their home cages and never subjected to the catheter implantation or the behavioral protocol for cocaine IVSA. Brain tissues were collected after four weeks of abstinence from cocaine self-administration, which has been used in prior studies to examine long-lasting effects of self-administration[47,135–140]. Brain tissues were extracted and snap-frozen (at −30°C). Cryosections of approximately 500 microns (Bregma -1.80 - 3.30mm) were used to dissect the amygdala on a −20°C frozen stage, including the CeA, BLA, and medial amygdala from both hemispheres. Punches from three sections were combined for each rat. In addition, 6 ACI/EurMcw rats were used for the dissection of the CeA and BLA.

1.5.4: Single-cell library preparation, sequencing, and alignment

snRNA-seq libraries from the whole amygdala tissues were performed by the Center for Epigenomics, UC San Diego using the Droplet-based Chromium Single-Cell 3’ solution (10x

Genomics, v3 chemistry). Briefly, frozen tissue was homogenized via glass dounce. Nuclei were then resuspended in 500 µL of nuclei permeabilization buffer (0.1% Triton-X-100 (Sigma-Aldrich, T8787), 1X protease inhibitor, 1 mM DTT, and 1U/µL RNase inhibitor (Promega, N211B), 2% BSA (Sigma-Aldrich, SRE0036) in PBS). Sample was incubated on a rotator for 5 min at 4°C and then centrifuged at 500 rcf for 5 min (4°C, run speed 3/3). Supernatant was removed and pellet was resuspended in 400 µL of sort buffer (1 mM EDTA 0.2 U/µL RNase inhibitor (Promega, N211B), 2% BSA (Sigma-Aldrich, SRE0036) in PBS) and stained with DRAQ7 (1:100; Cell Signaling, 7406). Up to 75,000 nuclei were sorted using a SH800 sorter (Sony) into 50 µL of collection buffer consisting of 1 U/µL RNase inhibitor in 5% BSA; the FACS gating strategy sorted based on particle size and DRAQ7 fluorescence. Sorted nuclei were then centrifuged at 1000 rcf for 15 min (4°C, run speed 3/3) and supernatant was removed. Nuclei were resuspended in 35 µL of reaction buffer (0.2 U/µL RNase inhibitor (Promega, N211B), 2% BSA (Sigma-Aldrich, SRE0036) in PBS) and counted on a hemocytometer. 12,000 nuclei were loaded onto a Chromium Controller (10x Genomics). Libraries were generated using the Chromium Single-Cell 3′ Library Construction Kit v3 (10x Genomics, 1000075) with the Chromium Single-Cell B Chip Kit (10x Genomics, 1000153) and the Chromium i7 Multiplex Kit for sample indexing (10x Genomics, 120262) according to manufacturer specifications. cDNA was amplified for 12 PCR cycles.

For snATAC-seq libraries from the whole amygdala tissues, nuclei were purified from frozen tissues using an established method[141]. Briefly, frozen amygdala tissue was homogenized using a 2 ml glass dounce with 1 ml cold 1x Homogenization Buffer (HB). The cell suspension was filtered using a 70 µm Flowmi strainer (BAH136800070, Millipore Sigma) and centrifuged at 350g for 5 min at 4°C. Nuclei were isolated by iodixanol (D1556, Millipore Sigma) density

gradient. The nuclei iodixanol solution (25%) was layered on top of 40% and 30% iodixanol solutions. Samples were centrifuged in a swinging bucket centrifuge at 3,000g for 20 min at 4°C. Nuclei were isolated from the 30-40% interface. Libraries were generated using the Chromium Next GEM Single Cell ATAC v1.1 (10x Genomics, PN-1000175) with the Chromium Next GEM Chip H Single Cell Kit (10x Genomics, 1000162) and the Chromium i7 Multiplex Kit for sample indexing (10x Genomics, 1000212) according to manufacturer specifications. DNA was amplified for 8 cycles.

For snRNA libraries from BLA and CeA, frozen brain tissues were obtained from the ACI/EurMcw rat strain, one of the HS rat founder strains. For nuclei isolation, brain punches from 3 rats for each region were pooled and homogenized in homogenization buffer (0.26 M sucrose, 0.03 M KCl, 0.01 M MgCl2, 0.02 M Tricine-KOH pH 7.8, 0.001 M DTT, 0.5 mM spermidine, 0.15 mM Spermine, 0.3% NP40) using with 1ml glass Dounce homogenizers. The homogenate was filtered with a 70-um strainer filter (SP Bel-Art, cat no 136800070) and centrifuged for 5 minutes at 350 RCF. The nuclei were purified with an iodixanol gradient (Sigma-Aldrich # 92339-11-2) by layering a 25% Iodixanol-nuclei mixture on top of 30% and 40% Iodixanol solutions. After centrifugation at 4^{o}C 3,000 RCF for 20 minutes, nuclei were collected from the 30-40% interface. Nuclei were washed in ATAC-RSB-Tween buffer (0.01 M Tris-HCl pH 7.5, 0.01 M NaCl, 0.003 M MgCl2, 0.1% Tween-20) and then resuspended in 1X nuclei buffer (10x Genomics, PN 2000207). 12,000 nuclei were loaded on the 10x Genomics Chromium Controller for GEM generation. RNAse inhibitors (Roche Diagnostics, # 03335402001) were added to all buffers (1U/ul). snRNA-seq library was performed using the Chromium Next GEM Single Cell Multiome Reagent Kit A (# 1000282) following Chromium Next GEM Single Cell Multiome ATAC + Gene Expression Reagent Kits User Guide (10X

Genomics). After the transposition reaction, nuclei were encapsulated and barcoded. Next-generation sequencing libraries were constructed following the User Guide.

For each library type, SPRISelect reagent (Beckman Coulter, B23319) was used for size selection and clean-up steps. Final library concentration was assessed by Qubit dsDNA HS Assay Kit (Thermo-Fisher Scientific) and post library QC was performed using Tapestation High Sensitivity D1000 (Agilent) to ensure that fragment sizes were distributed as expected. Final libraries were sequenced using the NovaSeq6000 (Illumina).

1.5.5: Behavioral experiments

Intravenous catheterization and behavioral testing of rats used for the generation of snRNA-seq and snATAC-seq were carried out using an established protocol of extended access to cocaine IVSA, progressive ratio (PR) testing, and foot shock, as reported previously[39,48,49]. Briefly, after surgical implantation of intravenous catheters and a week of recovery, HS rats were trained to self-administer cocaine (0.5 mg/kg/infusion) in 10 short access (ShA) sessions (2h/day, 5 days per week). Next, the animals were allowed to self-administer cocaine in 14 long access (LgA) sessions (6h/day, 5 days/week) to measure the escalation of drug intake (Figure 1.1e). Following the escalation phase, rats were screened for motivation using the PR test and for persistent drug-seeking despite adverse consequences using contingent foot-shock. The breakpoint (Figure 1.1f) was defined as the maximal number of presses completed before a 60-minute period during which a rat does not complete the next schedule. Rats were classified as vulnerable (high AI), defined by having high addiction-like behavior, versus resilient (low AI), defined as having low addiction-like behavior, via a median split[51,52]. AI was computed by averaging normalized measurements (z-scores) for the three behavioral tests: escalation of drug intake, motivation, and compulsive-like behavior, or drug taking despite adverse consequences[142]

(Figure 1.1c-d). The z-scores were calculated as $Z = (x-\mu)/\sigma$, where x is the raw value, μ is the mean of the cohort, and σ is the standard deviation of the cohort. Brain tissues were collected after four weeks of abstinence. For the pBBG studies, we used rats with low and high AI distinct from those used for the snRNA-seq and snATAC-seq experiments. Four weeks after the last drug self-administration session, the rats were placed back in the self-administration chambers without the availability of cocaine. The number of responses to the previous drug-paired lever (cocaine seeking behavior) was measured 30 minutes after intraperitoneal injection of pBBG (15 mg/kg/ml) or its vehicle, in a Latin square design. The 30 minutes time point was selected based on a previous study[86]. Data were analyzed using Prism 9.0 software (GraphPad, San Diego, CA, USA). Self-administration data were analyzed using repeated-measures analysis of variance (ANOVA) or mixed effect model followed by Bonferroni post-hoc tests when appropriate. For pairwise comparisons, data were analyzed using the unpaired *t*-test. The data are expressed as mean ± SEM unless otherwise specified. Values of $p < 0.05$ were considered statistically significant.

1.5.6: Electrophysiology

Slices of the CeA were prepared from rats after 4 weeks of protracted abstinence from cocaine IVSA following the same behavioral protocol described above or age-matched naive rats. High AI (n=5), low AI (n=5) and naive (n=5) rats were used for patch clamp baseline recordings. These rats were distinct from those used for snRNA-seq and snATAC-seq. Slices from each group were also used to record iPSCs after pBBG treatment. The naive rats received sham IV surgery. The rats were deeply anesthetized with isoflurane and brains were rapidly removed and placed in oxygenated (95% O_2, 5% CO_2) ice-cold cutting solution that contained 206 mM sucrose, 2.5 mM KCl, 1.2 mM NaH_2PO_4, 7 mM $MgCl_2$, 0.5 mM $CaCl_2$, 26 mM

$NaHCO_3$, 5 mM glucose, and 5 mM Hepes. Transverse slices (300 μm thick) were cut on a Vibratome (Leica VT1200S; Leica Microsystems) and transferred to oxygenated artificial cerebrospinal fluid (aCSF) that contained 130 mM NaCl, 2.5 mM KCl, 1.2 mM NaH_2PO_4, 2.0 mM $MgSO_4 \cdot 7H_2O$, 2.0 mM $CaCl_2$, 26 mM $NaHCO_3$, and 10 mM glucose. The slices were first incubated for 30 min at 35°C and then kept at room temperature for the remainder of the experiment. Individual slices containing CeA were transferred to a recording chamber that was mounted on the stage of an upright microscope (Olympus BX50WI). The slices were continuously perfused with oxygenated aCSF at a rate of 3 mL/min. Neurons were visualized with a 60X water-immersion objective (Olympus), infrared differential interference contrast optics, and a charge coupled device camera (EXi Blue; QImaging). Whole-cell recordings were performed using a Multiclamp 700B amplifier (10 kHz sampling rate, 10 kHz low-pass filter) and Digidata 1440A and pClamp 10 software (Molecular Devices). Patch pipettes (4–6 MΩ) were pulled from borosilicate glass (Warner Instruments) and filled with 70 mM $KMeSO_4$, 55 mM KCl, 10 mM NaCl, 2 mM $MgCl_2$, 10 mM Hepes, 2 mM Na-ATP, and 0.2 mM Na-GTP. Pharmacologically isolated sIPSCs were recorded in the presence of the glutamate receptor blockers, CNQX (Tocris #0190) and APV (Tocris #189), and the GABA-B receptor antagonist CGP55845 (Tocris #1246). Experiments with a series resistance of >15 MΩ or >20% change in series resistance were excluded from the final dataset. pBBG (2.5uM) was acutely applied in the bath. The frequency, amplitude, and kinetics of sIPSCs were analyzed using semi-automated threshold-based mini detection software (Easy Electrophysiology) and visually confirmed. Data were analyzed using Prism 9.0 software (GraphPad, San Diego, CA, USA) with one-way ANOVA followed by post hoc Tukey HSD test or with paired t-tests. The data are expressed as

mean ± SEM unless otherwise specified. Values of $p < 0.05$ were considered statistically significant.

1.5.7: Alignment of snRNA-seq and snATAC-seq reads

Raw base call (BCL) files were used to generate FASTQ files using Cell Ranger 3.1.0 for snRNA-seq data, Cellranger ATAC v.2.0.0 for snATAC-seq data, and Cell Ranger ARC v.2.0.0 for processing Chromium Single Cell Multiome ATAC + Gene Expression sequencing data. For demultiplexing raw base call (BCL) files generated by the sequencers into FASTQ files, we used the `cellranger mkfastq` command for RNA-seq reads, `cellranger-atac mkfastq` for ATAC-seq reads, and `cellranger-arc mkfastq` for the CeA and BLA samples which were generated using the multiome kit[143,144]. Next, we used `cellranger count` and `cellranger-atac count` to align the snRNA-seq and snATAC-seq reads, respectively, to a custom rat reference genome based on the rn6 reference genome downloaded from the UCSC genome browser[145–147]. The reference genome files for cell ranger were created from FASTA and genome annotation files for Rattus norvegicus Rnor_6.0 (Ensembl release 98)[148] and JASPAR2022 motifs[149]. BLA and CeA samples were aligned to the same reference genome using `cellranger-arc count`. We then filtered cells based on quality control metrics and performed barcode and UMI counting for the RNA-seq and ATAC-seq reads.

1.5.8: Quality control and preprocessing of snRNA-seq data

All snRNA-seq preprocessing was performed with Seurat v3.2.3[56]. For each sample, we loaded the filtered feature barcode matrices containing only detected cellular barcodes returned by `cellranger count` into Seurat. We computed the number of unique genes detected in each cell (nFeature_RNA); the total number of molecules detected within a cell (nCount_RNA); and the percentage of reads mapping to the mitochondrial genome (percent.mt) (Supplemental File 1.1).

nFeature_RNA is informative because low-quality cells or empty droplets will typically have very low numbers of detected genes while doublets or multiplets will exhibit very high gene counts. nCount_RNA is a metric that correlates with nFeature_RNA. We examined percent.mt because low-quality or dying cells typically exhibit a high degree of mitochondrial contamination. We removed all cells for which the value of any of these metrics was more than three standard deviations from the mean within the sample ($x > \mu \pm 3\sigma$). Next, we normalized the count data for each sample using sctransform[150] with percent.mt as a covariate.

1.5.9: Integrating snRNA-seq data across samples and clustering

To integrate snRNA-seq data across all our samples, we used reciprocal principal component analysis (RPCA), as implemented in Seurat[56,151]. First, we identified 2000 highly variable features (genes) across all of the samples to use as integration features using the `SelectIntegrationFeatures()` function, which we passed as anchor features (`anchor.features`) to the `PrepSCTIntegration()` function. Next, we performed dimensionality reduction with PCA on each sample using `RunPCA()`. After this, we ran the `FindIntegrationAnchors()` function to find a set of anchors between the list of Seurat objects from all of our samples using the same anchor features passed to `PrepSCTIntegration()`, specifying RPCA as the dimensional reduction method to be performed for finding anchors (`reduction = rpca`) and the first 30 RPCA dimensions to be used for specifying the k-nearest neighbor search space. Two rats (1 high AI, 1 low AI) were used as reference samples for the integration. We used the resulting anchor set to perform dataset integration across all of our samples using `IntegrateData()`. We clustered the integrated dataset by constructing a K-nearest neighbor (KNN) graph using the first 30 PCs followed by the Louvain algorithm, implemented in Seurat using the `FindNeighbors()` function followed by `FindClusters()`. Finally, we ran PCA once again on the integrated dataset and

visualized the data using uniform manifold approximation and projection (UMAP). Visualization of the integrated data in two-dimensional space indicated that batch correction was successful (Supplemental Figure 1.2a-c). To compare CeA and BLA subregion samples with the whole amygdala, we subsampled whole amygdala samples from the naïve rats in our study and performed the same integration technique. The integrated subregion data was visualized using UMAP.

1.5.10: Cell type assignment for snRNA-seq data

We identified marker genes of each cluster in our integrated snRNA-seq dataset using MAST[152], implemented with the `FindMarkers()` function in Seurat. Cell type identities were assigned based on expression of known marker genes of cell types known to be found in the amygdala.

1.5.11: Cell type-specific gene expression analysis for snRNA-seq data

Within each cell type, we tested for DEGs between the high AI rats and the low AI rats. We used the negative binomial test[88,89] implemented with the `FindMarkers()` function in Seurat to identify differential expression between groups, using percent.mt and the library prep date as covariates. We used the avg_log2FC value returned by the `FindMarkers()` function as an estimate of effect size. We did not pre-filter genes for testing based on log-fold change or minimum fraction of cells in which a gene was detected. This approach allowed us to detect weaker signals because we tested all observed genes in the dataset. Multiple testing correction was performed using the Benjamini-Hochberg method and we used a false discovery rate of 10% as a significance threshold (FDR<10%). Permutation tests were performed using the same methods, covariates, and filtering options but with shuffled addiction index labels. Results from

permuted and unpermuted data were compared by visualizing the distributions of uncorrected p-values with QQ-plots (Supplemental Figure 1.5).

We used ClusterProfiler[153] to perform gene set enrichment analysis (GSEA) of KEGG pathways for DEGs from each cell type. A ranked list of the avg_logFC values for all genes evaluated with our negative binomial test was given as input to GSEA. Multiple testing correction for GSEA results was performed using Benjamini-Hochberg adjustment, with statistical significance assessed at a FDR<10%.

All rat eQTLs described in the paper come from the RatGTEx portal (https://ratgtex.org/download/). Specifically, we downloaded their table of conditionally independent cis-eQTLs, which only includes eQTLs passing a significance threshold of FDR<0.05. We examined cis-eQTLs in the following brain tissues: BLA, Brain, IL, LHb, NAcc, NAcc2, OFC, PL, PL2. For each cis-eQTL in the database, only the top associated SNP is given, but some genes have more than one cis-eQTL in a tissue, meaning there are multiple loci with statistically independent associations with the gene's expression. We measured enrichment of significant DEGs (FDR<10%) that also had eQTLS in the rat brain using the Chi-squared test, implemented using the `chisq.test()` function in R.

To obtain bootstrap distributions of DEG effect sizes, we resampled nuclei with replacement 1000 times. Resampling was performed separately for nuclei from high and low AI rats so that the sample size of each set remained consistent. For each bootstrap iteration we recorded the p-value and the coefficients for the hi/low AI condition from the negative binomial regression performed by Seurat's `FindMarkers()' function. We then rescaled the coefficient to be in units of log2 fold change. We note that the log2FC estimates obtained by this method correspond to the p-values but differ slightly from Seurat's avg_log2FC estimates because

Seurat's avg_log2FC calculations introduce a pseudocount and do not consider the effects of covariates. The distribution of resulting bootstrap fold-change estimates and q-values were visualized as violin plots using Plotly in Python (Figure 1.3c-e).

1.5.12: Comparing observed gene expression differences to predicted gene expression differences based on cis-genetic variation

To estimate the genetic component of gene expression variation in the brain, conditionally independent cis-eQTLs and their allelic fold change (aFC) estimates for whole brain hemisphere tissue were downloaded from the RatGTEx Portal (https://ratgtex.org/download/). Using aFC as effect size, gene expression was predicted from genotypes using eQTL linear models[77] (https://github.com/PejLab/gene_expr_pred). Predicted relative expression was thus obtained for 26 rats whose genotypes were available, and only for genes with at least one significant cis-eQTL. Genes with zero-variance predictions were removed, resulting in predictions for 8,997 genes. To further prioritize genes by estimated prediction accuracy, gene expression was predicted for the same 339 rats that were used to map the whole brain hemisphere eQTLs. Pearson correlation r^2 was calculated between those predictions and observed log-TPM expression from the same rats. We then measured the difference between mean predicted expression in high vs low AI rats and compared it against the avg_logFC estimates obtained by Seurat's FindMarkers() function. Spearman's correlation coefficient (ρ) was calculated between the difference in mean predicted expression and the observed avg_logFC. We performed these tests multiple times using different r^2 cutoffs for the gene expression prediction models to filter genes (Table. S3). Spearman's correlation coefficients (ρ) and the associated p-values were calculated using `scipy.stats.spearmanr()`. Confidence

intervals were calculated using the formula $tanh(tanh^{-1}(\rho) \pm \frac{1.96}{\sqrt{N-3}})$. Spearman ρ confidence intervals were visualized using Plotly in Python.

1.5.13: Quality control and preprocessing of snATAC-seq data

All snATAC-seq data preprocessing was performed with MACS2[87] (for peak calling) and Signac[57]. Although peak calling is performed during alignment by `cellranger-atac count`, we chose to call peaks separately using MACS2 because Cell Ranger's peak calling function has been observed to merge multiple distinct peaks into a single region[154]. To minimize loss of informative features for clustering and downstream analysis, we first called peaks on the snATAC-seq BAM files for each rat with MACS2 (`macs2 callpeak -t {input} -f BAM -n {sample} --outdir {output} {params} --nomodel --shift -100 --ext 200 --qval 5e-2 -B --SPMR`). We confirmed that MACS2 calls more peaks and peaks with smaller widths compared to Cell Ranger (Supplemental Figure 1.12). Next, we merged overlapping peaks across all our samples to generate a combined peak set using BEDtools[155] (`bedtools merge`). We generated a new peak by barcode matrix for each sample using this combined peak set and all detected cellular barcodes using the `FeatureMatrix()` function in Signac. We used these new matrices to create ChromatinAssay objects in Signac with the BSgenome.Rnorvegicus.UCSC.rn6[146] reference genome using the `CreateChromatinAssay()` function. From these ChromatinAssay objects we created Seurat objects with `CreateSeuratObject()`, which are compatible with functions from the Seurat package. We computed several quality control metrics for each sample: nucleosome banding pattern (nucleosome_signal); transcriptional start site (TSS) enrichment score (TSS.enrichment); total number of fragments in peaks (peak_region_fragments); and fraction of fragments in peaks (pct_reads_in_peaks) (Supplemental File 1.2). We removed all cells for which the value of any of these metrics was more than two standard deviations from the mean

within the sample ($x > \mu \pm 2\sigma$). We removed one rat (FTL_463_M757_933000320046135) from our dataset, due to the very low number of detected fragments per cell in this sample (Supplemental Figure 1.13).

1.5.14: Integrating snATAC-seq data across samples and clustering

Each sample was normalized using term frequency-inverse document frequency (TF-IDF) followed by singular value decomposition (SVD) on the TF-IDF matrix using all features (peaks)[57,154]. The combined steps of TF-IDF followed by SVD are known as latent semantic indexing (LSI)[156,157]. Non-linear dimensionality reduction and clustering were performed using UMAP and KNN following the same procedure used, respectively, just as we did for the snRNA-seq data. We merged the data across all samples within Signac and repeated the process of LSI in the same manner as it was applied to individual samples. We then integrated the merged dataset using Harmony[158] implemented by Signac, integrating over the sample library variable to account for the effects of different sequencing libraries used for different samples. We observed successful reduction of batch effects following integration (Supplemental Figure 1.2d-f). We once again performed non-linear dimensionality reduction and clustering with UMAP and KNN, respectively. Notably, LSI, UMAP and KNN are used for visualization purposes; raw counts were used for downstream differential accessibility analyses.

1.5.15: Label transfer and cell type assignment for snATAC-seq data

We created a gene activity matrix for the integrated snATAC-seq data using the `GeneActivity()` function in Signac. This counts the number of fragments per cell overlapping the promoter region of a given gene and uses that value as a gene activity score. Gene activity serves as a proxy for gene expression as gene expression is often correlated with promoter accessibility. The gene activity scores were log normalized using the `NormalizeData()` function

in Seurat with the normalization method set to `LogNormalize` and the scaling factor set to the median value of nCount_RNA across all cells, based on the gene activity scores. Cell type identities were assigned by integrating the snATAC-seq data with the integrated snRNA-seq data and performing label transfer[56] as described in Signac. Briefly, this approach identifies shared correlation patterns in the gene activity matrix and the scRNA-seq dataset to identify matched biological states across the two modalities. The process returns a classification score for each cell for each cell type defined in the scRNA-seq data. Each cell was assigned the cell type identity with the highest prediction score. Additionally, by identifying matched cells in the snRNA-seq dataset, we were able to impute RNA expression values for each of the cells in our snATAC-seq dataset. This enabled us to perform correlative analyses of chromatin accessibility and gene expression in later downstream analyses, as it produced a pseudo-multimodal dataset.

1.5.16: Differential chromatin accessibility analysis of snATAC-seq data

Similar to our differential analyses of the snRNA-seq data, we tested for differentially accessible genomic regions between nuclei from the high versus low AI rats within each cell type. We used the negative binomial test[150,159] implemented with the `FindMarkers()` function from Seurat to model the raw snATAC-seq count data using peak_region_fragments, library batch date, and rat sample ID as covariates. Multiple testing correction was performed using Benjamini-Hochberg adjustment and a false discovery rate below 10% (FDR<10%) was used to determine statistical significance. Permutation tests were performed in the same manner as for the differential gene expression analyses (using the same statistical test and covariates with shuffled addiction index labels).

1.5.17: Partitioned heritability analysis

We downloaded summary statistics for the Liu et al. 2019 GWAS of tobacco and alcohol use[106] and used the munge_sumstats.py script from LD Score (LDSC)[105] to parse the summary statistics file into the proper format for downstream analyses. We used the sets of significant differential peaks (FDR<10%) for each cell type as foreground peaks and DNaseI hypersensitivity profiles for 53 epigenomes from ENCODE Honeybadger2. We used the UCSC liftOver tool to convert the foreground peaks from rn6 to hg19. There was no need to lift over the background peaks as Honeybadger2 is already in hg19. Next, we generated partitioned LD scores for the background and foreground regions. We used the make_annot.py script to make annotation files and the ldsc.py script to compute annotation-specific LD scores. We used the European 1000 Genomes Phase 3 PLINK[160] files to compute the LD scores. Finally, using the baseline model and standard regression weights from the LDSC Partitioning Heritability tutorial, we ran a cell type-specific partitioned heritability analysis with the LD scores we computed.

1.5.18: Annotation of accessible chromatin regions

Before performing any differential analyses, we first used the annotatePeaks.pl script from the HOMER suite to annotate accessible chromatin regions and significant differential peaks (FDR<10%) for each cell type in our integrated dataset[161]. For each cell type, we performed a Fisher's Exact Test to measure the enrichment of genomic regions annotated as a promoter region within the differential peaks compared to the set of all peaks in the dataset and observed significant results for all cell types tested. Specifically, we compared the ratio of peaks annotated as promoter regions to non-promoter regions in the significant differential peaks (FDR<10%) versus all other peaks.

1.5.19: Fisher's Exact Tests

We first performed a Fisher's Exact Test to measure enrichment of DEGs (FDR<10%) with differentially accessible promoters. We defined the latter as the case where the promoter region of a gene overlaps a significant differentially accessible peak (FDR<10%). We obtained gene coordinates from the TxDb.Rnorvegicus.UCSC.rn6.refGene annotation package and defined promoter regions as being 1500 bases upstream and 500 bases downstream of the TSS (`promoters(genes(TxDb.Rnorvegicus.UCSC.rn6.refGene), upstream = 1500, downstream = 500)`). We then generated a confusion matrix from the following four values: the number of DEGs with differentially accessible promoters; the number of DEGs with non-differentially accessible promoters; the number of non-DEGs with differentially accessible promoters; and the number of non-DEGs with non-differentially accessible promoters. We then performed a Fisher's Exact Test to measure enrichment of differentially accessible peaks (FDR<10%) which were annotated as TSS/promoter regions by HOMER (annotatePeaks.pl). We generated a confusion matrix from the following four values: the number of differential peaks with a TSS/promoter annotation; the number of differential peaks without a TSS/promoter annotation; the number of non-differential peaks (FDR>10%) with a TSS/promoter annotation; and the number of non-differential peaks (FDR>10%) without a TSS/promoter annotation.

1.5.20: Measuring differential activity of transcription factors with chromVAR

We measured cell type specific motif activities using chromVAR to test for per motif deviations in accessibility between nuclei from high versus low AI rats. Motif data was pulled from the JASPAR2020 database, and integrated with snATAC-seq data using the `AddMotifs()` function in Signac. After adding motifs to our snATAC-seq dataset, we ran chromVAR with the `RunChromVAR()` wrapper in Signac. Differential analysis of chromVAR deviation scores was

performed using the Wilcoxon rank-sum test between high AI rats versus lowly addicted rats within each cell type. Differential testing was performed using Seurat's `FindMarkers()` function with the mean function set as `rowMeans()` to calculate average difference in deviation scores between groups. Multiple testing correction was performed using Benjamini-Hochberg adjustment and a false discovery rate below 10% (FDR<10%) was used to determine statistical significance. Motif clusters were defined using the provided cluster numbers from JASPAR's matrix clustering-results and the names of the clusters were annotated by hand based on which TFs were present in each cluster. When selecting clusters to visualize, we retrieved the top 50 motifs (FDR<10%) per cell-type and highlighted their respective clusters. Volcano plots and heatmap data were generated using Plotly in Python. Hierarchical ordering of heatmap clusters was generated with Plotly's `figure_factory.create_dendrogram()` function, which wraps the `cluster.hierarchy.dendrogram()` function in SciPy.

1.6: Acknowledgements

Chapter 1, in part, has been submitted for publication of the material at *Nature Neuroscience*. The dissertation author was the primary researcher and author of this paper.

1.7: References

1. Janak PH, Tye KM. From circuits to behaviour in the amygdala. Nature. 2015 Jan 15;517(7534):284–292. PMCID: PMC4565157

2. Roberto M, Gilpin NW, Siggins GR. The Central Amygdala and Alcohol: Role of γ-Aminobutyric Acid, Glutamate, and Neuropeptides. Cold Spring Harb Perspect Med. 2012 Dec;2(12):a012195. PMCID: PMC3543070

3. Buffalari DM, See RE. Amygdala Mechanisms of Pavlovian Psychostimulant Conditioning and Relapse. In: Self DW, Staley Gottschalk JK, editors. Behavioral Neuroscience of Drug Addiction [Internet]. Berlin, Heidelberg: Springer; 2010 [cited 2022 Aug 4]. p. 73–99. Available from: https://doi.org/10.1007/7854_2009_18

4. Schumann CM, Bauman MD, Amaral DG. Abnormal structure or function of the amygdala is a common component of neurodevelopmental disorders. Neuropsychologia. 2011 Mar;49(4):745–759. PMCID: PMC3060967

5. Koob GF. Anhedonia, Hyperkatifeia, and Negative Reinforcement in Substance Use Disorders. In: Pizzagalli DA, editor. Anhedonia: Preclinical, Translational, and Clinical Integration [Internet]. Cham: Springer International Publishing; 2022 [cited 2023 Apr 21]. p. 147–165. Available from: https://doi.org/10.1007/7854_2021_288

6. Pickens CL, Airavaara M, Theberge F, Fanous S, Hope BT, Shaham Y. Neurobiology of the incubation of drug craving. Trends Neurosci. 2011 Aug;34(8):411–420. PMCID: PMC3152666

7. Kalivas PW, Volkow ND. The Neural Basis of Addiction: A Pathology of Motivation and Choice. AJP. American Psychiatric Publishing; 2005 Aug;162(8):1403–1413.

8. Kilts CD, Schweitzer JB, Quinn CK, Gross RE, Faber TL, Muhammad F, Ely TD, Hoffman JM, Drexler KPG. Neural Activity Related to Drug Craving in Cocaine Addiction. Archives of General Psychiatry. 2001 Apr 1;58(4):334–341.

9. Aerts T, Seuntjens E. Novel Perspectives on the Development of the Amygdala in Rodents. Front Neuroanat. 2021 Dec 9;15:786679. PMCID: PMC8696165

10. Arrigucci R, Bushkin Y, Radford F, Lakehal K, Vir P, Pine R, Martin D, Sugarman J, Zhao Y, Yap GS, Lardizabal AA, Tyagi S, Gennaro ML. FISH-Flow, a protocol for the concurrent detection of mRNA and protein in single cells using fluorescence in situ hybridization and flow cytometry. Nat Protoc. 2017 Jun;12(6):1245–1260. PMCID: PMC5548662

11. Ehrlich I, Humeau Y, Grenier F, Ciocchi S, Herry C, Lüthi A. Amygdala Inhibitory Circuits and the Control of Fear Memory. Neuron. Elsevier; 2009 Jun 25;62(6):757–771. PMID: 19555645

12. Ciocchi S, Herry C, Grenier F, Wolff SBE, Letzkus JJ, Vlachos I, Ehrlich I, Sprengel R, Deisseroth K, Stadler MB, Müller C, Lüthi A. Encoding of conditioned fear in central amygdala inhibitory circuits. Nature. Nature Publishing Group; 2010 Nov;468(7321):277–282.

13. Yao Z, Liu H, Xie F, Fischer S, Adkins RS, Aldridge AI, Ament SA, Bartlett A, Behrens MM, Van den Berge K, Bertagnolli D, de Bézieux HR, Biancalani T, Booeshaghi AS, Bravo HC, Casper T, Colantuoni C, Crabtree J, Creasy H, Crichton K, Crow M, Dee N, Dougherty EL, Doyle WI, Dudoit S, Fang R, Felix V, Fong O, Giglio M, Goldy J, Hawrylycz M, Herb BR, Hertzano R, Hou X, Hu Q, Kancherla J, Kroll M, Lathia K, Li YE, Lucero JD, Luo C, Mahurkar A, McMillen D, Nadaf NM, Nery JR, Nguyen TN, Niu SY, Ntranos V, Orvis J, Osteen JK, Pham T, Pinto-Duarte A, Poirion O, Preissl S, Purdom E, Rimorin C, Risso D, Rivkin AC, Smith K, Street K, Sulc J, Svensson V, Tieu M, Torkelson A, Tung H, Vaishnav ED, Vanderburg CR, van Velthoven C, Wang X, White OR, Huang ZJ, Kharchenko PV, Pachter L, Ngai J, Regev A, Tasic B, Welch JD, Gillis J, Macosko EZ, Ren B, Ecker JR, Zeng H, Mukamel EA. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. Nature. 2021;598(7879):103–110. PMCID: PMC8494649

14. Bakken TE, Jorstad NL, Hu Q, Lake BB, Tian W, Kalmbach BE, Crow M, Hodge RD, Krienen FM, Sorensen SA, Eggermont J, Yao Z, Aevermann BD, Aldridge AI, Bartlett A,

Bertagnolli D, Casper T, Castanon RG, Crichton K, Daigle TL, Dalley R, Dee N, Dembrow N, Diep D, Ding SL, Dong W, Fang R, Fischer S, Goldman M, Goldy J, Graybuck LT, Herb BR, Hou X, Kancherla J, Kroll M, Lathia K, van Lew B, Li YE, Liu CS, Liu H, Lucero JD, Mahurkar A, McMillen D, Miller JA, Moussa M, Nery JR, Nicovich PR, Niu SY, Orvis J, Osteen JK, Owen S, Palmer CR, Pham T, Plongthongkum N, Poirion O, Reed NM, Rimorin C, Rivkin A, Romanow WJ, Sedeño-Cortés AE, Siletti K, Somasundaram S, Sulc J, Tieu M, Torkelson A, Tung H, Wang X, Xie F, Yanny AM, Zhang R, Ament SA, Behrens MM, Bravo HC, Chun J, Dobin A, Gillis J, Hertzano R, Hof PR, Höllt T, Horwitz GD, Keene CD, Kharchenko PV, Ko AL, Lelieveldt BP, Luo C, Mukamel EA, Pinto-Duarte A, Preissl S, Regev A, Ren B, Scheuermann RH, Smith K, Spain WJ, White OR, Koch C, Hawrylycz M, Tasic B, Macosko EZ, McCarroll SA, Ting JT, Zeng H, Zhang K, Feng G, Ecker JR, Linnarsson S, Lein ES. Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature. 2021;598(7879):111–119. PMCID: PMC8494640

15. Scala F, Kobak D, Bernabucci M, Bernaerts Y, Cadwell CR, Castro JR, Hartmanis L, Jiang X, Laturnus S, Miranda E, Mulherkar S, Tan ZH, Yao Z, Zeng H, Sandberg R, Berens P, Tolias AS. Phenotypic variation of transcriptomic cell types in mouse motor cortex. Nature. 2021;598(7879):144–150. PMCID: PMC8113357

16. Booeshaghi AS, Yao Z, van Velthoven C, Smith K, Tasic B, Zeng H, Pachter L. Isoform cell-type specificity in the mouse primary motor cortex. Nature. 2021;598(7879):195–199. PMCID: PMC8494650

17. Kozareva V, Martin C, Osorno T, Rudolph S, Guo C, Vanderburg C, Nadaf N, Regev A, Regehr WG, Macosko E. Author Correction: A transcriptomic atlas of mouse cerebellar cortex comprehensively defines cell types. Nature. 2022;602(7896):E21. PMCID: PMC8828463

18. Bhaduri A, Sandoval-Espinosa C, Otero-Garcia M, Oh I, Yin R, Eze UC, Nowakowski TJ, Kriegstein AR. An atlas of cortical arealization identifies dynamic molecular signatures. Nature. 2021;598(7879):200–204. PMCID: PMC8494648

19. Berg J, Sorensen SA, Ting JT, Miller JA, Chartrand T, Buchin A, Bakken TE, Budzillo A, Dee N, Ding SL, Gouwens NW, Hodge RD, Kalmbach B, Lee C, Lee BR, Alfiler L, Baker K, Barkan E, Beller A, Berry K, Bertagnolli D, Bickley K, Bomben J, Braun T, Brouner K, Casper T, Chong P, Crichton K, Dalley R, de Frates R, Desta T, Lee SD, D'Orazi F, Dotson N, Egdorf T, Enstrom R, Farrell C, Feng D, Fong O, Furdan S, Galakhova AA, Gamlin C, Gary A, Glandon A, Goldy J, Gorham M, Goriounova NA, Gratiy S, Graybuck L, Gu H, Hadley K, Hansen N, Heistek TS, Henry AM, Heyer DB, Hill D, Hill C, Hupp M, Jarsky T, Kebede S, Keene L, Kim L, Kim MH, Kroll M, Latimer C, Levi BP, Link KE, Mallory M, Mann R, Marshall D, Maxwell M, McGraw M, McMillen D, Melief E, Mertens EJ, Mezei L, Mihut N, Mok S, Molnar G, Mukora A, Ng L, Ngo K, Nicovich PR, Nyhus J, Olah G, Oldre A, Omstead V, Ozsvar A, Park D, Peng H, Pham T, Pom CA, Potekhina L, Rajanbabu R, Ransford S, Reid D, Rimorin C, Ruiz A, Sandman D, Sulc J, Sunkin SM, Szafer A, Szemenyei V, Thomsen ER, Tieu M, Torkelson A, Trinh J, Tung H, Wakeman W, Waleboer F, Ward K, Wilbers R, Williams G, Yao Z, Yoon JG, Anastassiou C, Arkhipov A, Barzo P, Bernard A, Cobbs C, de Witt Hamer PC, Ellenbogen RG, Esposito L, Ferreira M, Gwinn RP,

Hawrylycz MJ, Hof PR, Idema S, Jones AR, Keene CD, Ko AL, Murphy GJ, Ng L, Ojemann JG, Patel AP, Phillips JW, Silbergeld DL, Smith K, Tasic B, Yuste R, Segev I, de Kock CPJ, Mansvelder HD, Tamas G, Zeng H, Koch C, Lein ES. Author Correction: Human neocortical expansion involves glutamatergic neuron diversification. Nature. 2022;601(7893):E12. PMCID: PMC8770134

20. Di Bella DJ, Habibi E, Stickels RR, Scalia G, Brown J, Yadollahpour P, Yang SM, Abbate C, Biancalani T, Macosko EZ, Chen F, Regev A, Arlotta P. Molecular Logic of Cellular Diversification in the Mouse Cerebral Cortex. Nature. 2021 Jul;595(7868):554–559. PMCID: PMC9006333

21. Ziffra RS, Kim CN, Ross JM, Wilfert A, Turner TN, Haeussler M, Casella AM, Przytycki PF, Keough KC, Shin D, Bogdanoff D, Kreimer A, Pollard KS, Ament SA, Eichler EE, Ahituv N, Nowakowski TJ. Single-cell epigenomics reveals mechanisms of human cortical development. Nature. 2021;598(7879):205–213. PMCID: PMC8494642

22. Zhang Z, Zhou J, Tan P, Pang Y, Rivkin AC, Kirchgessner MA, Williams E, Lee CT, Liu H, Franklin AD, Miyazaki PA, Bartlett A, Aldridge AI, Vu M, Boggeman L, Fitzpatrick C, Nery JR, Castanon RG, Rashid M, Jacobs MW, Ito-Cole T, O'Connor C, Pinto-Duartec A, Dominguez B, Smith JB, Niu SY, Lee KF, Jin X, Mukamel EA, Behrens MM, Ecker JR, Callaway EM. Epigenomic diversity of cortical projection neurons in the mouse brain. Nature. 2021;598(7879):167–173. PMCID: PMC8494636

23. Li YE, Preissl S, Hou X, Zhang Z, Zhang K, Qiu Y, Poirion OB, Li B, Chiou J, Liu H, Pinto-Duarte A, Kubo N, Yang X, Fang R, Wang X, Han JY, Lucero J, Yan Y, Miller M, Kuan S, Gorkin D, Gaulton KJ, Shen Y, Nunn M, Mukamel EA, Behrens MM, Ecker JR, Ren B. An atlas of gene regulatory elements in adult mouse cerebrum. Nature. 2021;598(7879):129–136. PMCID: PMC8494637

24. Liu H, Zhou J, Tian W, Luo C, Bartlett A, Aldridge A, Lucero J, Osteen JK, Nery JR, Chen H, Rivkin A, Castanon RG, Clock B, Li YE, Hou X, Poirion OB, Preissl S, Pinto-Duarte A, O'Connor C, Boggeman L, Fitzpatrick C, Nunn M, Mukamel EA, Zhang Z, Callaway EM, Ren B, Dixon JR, Behrens MM, Ecker JR. DNA methylation atlas of the mouse brain at single-cell resolution. Nature. 2021;598(7879):120–128. PMCID: PMC8494641

25. Domcke S, Hill AJ, Daza RM, Cao J, O'Day DR, Pliner HA, Aldinger KA, Pokholok D, Zhang F, Milbank JH, Zager MA, Glass IA, Steemers FJ, Doherty D, Trapnell C, Cusanovich DA, Shendure J. A human cell atlas of fetal chromatin accessibility. Science. 2020 Nov 13;370(6518):eaba7612. PMCID: PMC7785298

26. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, Qiu Y, Li YE, Gaulton KJ, Wang A, Preissl S, Ren B. A single-cell atlas of chromatin accessibility in the human genome. Cell. 2021 Nov 24;184(24):5985-6001.e19. PMCID: PMC8664161

27. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, Kaul R, Halow J, Van Nostrand EL, Freese P, Gorkin DU, Shen Y, He Y, Mackiewicz M, Pauli-Behn F, Williams BA, Mortazavi A, Keller CA, Zhang XO, Elhajjajy SI, Huey J, Dickel DE, Snetkova V, Wei X, Wang X, Rivera-Mulia JC, Rozowsky J, Zhang

J, Chhetri SB, Zhang J, Victorsen A, White KP, Visel A, Yeo GW, Burge CB, Lécuyer E, Gilbert DM, Dekker J, Rinn J, Mendenhall EM, Ecker JR, Kellis M, Klein RJ, Noble WS, Kundaje A, Guigó R, Farnham PJ, Cherry JM, Myers RM, Ren B, Graveley BR, Gerstein MB, Pennacchio LA, Snyder MP, Bernstein BE, Wold B, Hardison RC, Gingeras TR, Stamatoyannopoulos JA, Weng Z. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583(7818):699–710. PMCID: PMC7410828

28. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nature Reviews Genetics. 2014 Apr;15(4):272–286.

29. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015 Jul 23;523(7561):486–490. PMCID: PMC4685948

30. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, Lee C, Regalado SG, Read DF, Steemers FJ, Disteche CM, Trapnell C, Shendure J. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell. Elsevier; 2018 Aug 23;174(5):1309-1324.e18. PMID: 30078704

31. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015 Oct 15;24(R1):R102–R110. PMCID: PMC4572001

32. 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. PMCID: PMC8549541

33. Tran MN, Maynard KR, Spangler A, Huuki LA, Montgomery KD, Sadashivaiah V, Tippani M, Barry BK, Hancock DB, Hicks SC, Kleinman JE, Hyde TM, Collado-Torres L, Jaffe AE, Martinowich K. Single-nucleus transcriptome analysis reveals cell-type-specific molecular signatures across reward circuitry in the human brain. Neuron. Elsevier; 2021 Oct 6;109(19):3088-3103.e5. PMID: 34582785

34. Chen R, Blosser TR, Djekidel MN, Hao J, Bhattacherjee A, Chen W, Tuesta LM, Zhuang X, Zhang Y. Decoding molecular and cellular heterogeneity of mouse nucleus accumbens. Nat Neurosci. 2021 Dec;24(12):1757–1771. PMCID: PMC8639815

35. He J, Kleyman M, Chen J, Alikaya A, Rothenhoefer KM, Ozturk BE, Wirthlin M, Bostan AC, Fish K, Byrne LC, Pfenning AR, Stauffer WR. Transcriptional and anatomical diversity of medium spiny neurons in the primate striatum. Current Biology. Elsevier; 2021 Dec 20;31(24):5473-5486.e6. PMID: 34727523

36. Phillips RA, Tuscher JJ, Black SL, Andraka E, Fitzgerald ND, Ianov L, Day JJ. An atlas of transcriptionally defined cell populations in the rat ventral tegmental area. Cell Reports [Internet]. Elsevier; 2022 Apr 5 [cited 2022 Aug 4];39(1). Available from: https://www.cell.com/cell-reports/abstract/S2211-1247(22)00364-3 PMID: 35385745

37. Avey D, Sankararaman S, Yim AKY, Barve R, Milbrandt J, Mitra RD. Single-Cell RNA-

Seq Uncovers a Robust Transcriptional Response to Morphine by Glia. Cell Rep. 2018 Sep 25;24(13):3619-3629.e4. PMCID: PMC6357782

38. Savell KE, Tuscher JJ, Zipperly ME, Duke CG, Phillips RA, Bauman AJ, Thukral S, Sultan FA, Goska NA, Ianov L, Day JJ. A dopamine-induced gene expression signature regulates neuronal function and cocaine response. Science Advances. American Association for the Advancement of Science; 2020 Jun 1;6(26):eaba4221.

39. Carrette LLG, Guglielmo G de, Kallupi M, Maturin L, Brennan M, Boomhower B, Conlisk D, Sedighim S, Tieu L, Fannon MJ, Velarde N, Martinez AR, Kononoff J, Kimbrough A, Simpson S, Smith LC, Shankar K, Ramirez FJ, Chitre AS, Lin B, Polesskaya O, Woods LCS, Palmer AA, George O. The Cocaine and Oxycodone Biobanks, Two Repositories from Genetically Diverse and Behaviorally Characterized Rats for the Study of Addiction. eNeuro [Internet]. Society for Neuroscience; 2021 May 1 [cited 2021 Aug 2];8(3). Available from: https://www.eneuro.org/content/8/3/ENEURO.0033-21.2021 PMID: 33875455

40. Chen BT, Yau HJ, Hatch C, Kusumoto-Yoshida I, Cho SL, Hopf FW, Bonci A. Rescuing cocaine-induced prefrontal cortex hypoactivity prevents compulsive cocaine seeking. Nature. Nature Publishing Group; 2013 Apr;496(7445):359–362.

41. Cohen A, Koob GF, George O. Robust Escalation of Nicotine Intake with Extended Access to Nicotine Self-Administration and Intermittent Periods of Abstinence. Neuropsychopharmacology. Nature Publishing Group; 2012 Aug;37(9):2153–2160.

42. Koob GF, Buck CL, Cohen A, Edwards S, Park PE, Schlosburg JE, Schmeichel B, Vendruscolo LF, Wade CL, Whitfield TW, George O. Addiction as a Stress Surfeit Disorder. Neuropharmacology. 2014 Jan;76(0 0):10.1016/j.neuropharm.2013.05.024. PMCID: PMC3830720

43. Solberg Woods LC, Palmer AA. Using Heterogeneous Stocks for Fine-Mapping Genetically Complex Traits. Methods Mol Biol. 2019;2018:233–247. PMCID: PMC9121584

44. Hansen C, Spuhler K. Development of the National Institutes of Health genetically heterogeneous rat stock. Alcohol Clin Exp Res. 1984 Oct;8(5):477–479. PMID: 6391259

45. Saar K, Beck A, Bihoreau MT, Birney E, Brocklebank D, Chen Y, Cuppen E, Demonchy S, Flicek P, Foglio M, Fujiyama A, Gut IG, Gauguier D, Guigo R, Guryev V, Heinig M, Hummel O, Jahn N, Klages S, Kren V, Kuhl H, Kuramoto T, Kuroki Y, Lechner D, Lee YA, Lopez-Bigas N, Lathrop GM, Mashimo T, Kube M, Mott R, Patone G, Perrier-Cornet JA, Platzer M, Pravenec M, Reinhardt R, Sakaki Y, Schilhabel M, Schulz H, Serikawa T, Shikhagaie M, Tatsumoto S, Taudien S, Toyoda A, Voigt B, Zelenika D, Zimdahl H, Hubner N. SNP and haplotype mapping for genetic analysis in the Rat. Nat Genet. 2008 May;40(5):560–566. PMCID: PMC5915293

46. Baud A, Hermsen R, Guryev V, Stridh P, Graham D, McBride MW, Foroud T, Calderari S, Diez M, Ockinger J, Beyeen AD, Gillett A, Abdelmagid N, Guerreiro-Cacais AO, Jagodic M, Tuncel J, Norin U, Beattie E, Huynh N, Miller WH, Koller DL, Alam I, Falak S, Osborne-Pellegrin M, Martinez-Membrives E, Canete T, Blazquez G, Vicens-Costa E,

Mont-Cardona C, Diaz-Moran S, Tobena A, Hummel O, Zelenika D, Saar K, Patone G, Bauerfeind A, Bihoreau MT, Heinig M, Lee YA, Rintisch C, Schulz H, Wheeler DA, Worley KC, Muzny DM, Gibbs RA, Lathrop M, Lansu N, Toonen P, Ruzius FP, de Bruijn E, Hauser H, Adams DJ, Keane T, Atanur SS, Aitman TJ, Flicek P, Malinauskas T, Jones EY, Ekman D, Lopez-Aumatell R, Dominiczak AF, Johannesson M, Holmdahl R, Olsson T, Gauguier D, Hubner N, Fernandez-Teruel A, Cuppen E, Mott R, Flint J. Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nat Genet. 2013 Jul;45(7):10.1038/ng.2644. PMCID: PMC3821058

47. Carrette LLG, Corral C, Boomhower B, Brennan M, Crook C, Ortez C, Shankar K, Simpson S, Maturin L, Solberg Woods LC, Palmer AA, de Guglielmo G, George O. Leptin Protects Against the Development and Expression of Cocaine Addiction-Like Behavior in Heterogeneous Stock Rats. Front Behav Neurosci. 2022 Mar 3;16:832899. PMCID: PMC8934439

48. Characterization of cocaine addiction-like behavior in heterogeneous stock rats | bioRxiv [Internet]. [cited 2022 Jun 9]. Available from: https://www.biorxiv.org/content/10.1101/2021.07.22.453410v2

49. Sedighim S, Carrette LL, Venniro M, Shaham Y, de Guglielmo G, George O. Individual differences in addiction-like behaviors and choice between cocaine versus food in Heterogeneous Stock rats. Psychopharmacology (Berl). 2021 Dec;238(12):3423–3433. PMCID: PMC8889911

50. George O, Mandyam CD, Wee S, Koob GF. Extended Access to Cocaine Self-Administration Produces Long-Lasting Prefrontal Cortex-Dependent Working Memory Impairments. Neuropsychopharmacology. 2008 Sep;33(10):2474–2482. PMCID: PMC2760333

51. Belin D, Balado E, Piazza PV, Deroche-Gamonet V. Pattern of intake and drug craving predict the development of cocaine addiction-like behavior in rats. Biol Psychiatry. 2009 May 15;65(10):863–868. PMID: 18639867

52. Deroche-Gamonet V, Belin D, Piazza PV. Evidence for addiction-like behavior in the rat. Science. 2004 Aug 13;305(5686):1014–1017. PMID: 15310906

53. Belin D, Deroche-Gamonet V. Responses to Novelty and Vulnerability to Cocaine Addiction: Contribution of a Multi-Symptomatic Animal Model. Cold Spring Harb Perspect Med. Cold Spring Harbor Laboratory Press; 2012 Nov 1;2(11):a011940. PMID: 23125204

54. Koob GF, Volkow ND. Neurobiology of addiction: a neurocircuitry analysis. Lancet Psychiatry. 2016 Aug;3(8):760–773. PMCID: PMC6135092

55. Yu B, Zhang Q, Lin L, Zhou X, Ma W, Wen S, Li C, Wang W, Wu Q, Wang X, Li XM. Molecular and cellular evolution of the amygdala across species analyzed by single-nucleus transcriptome profiling. Cell Discov. Nature Publishing Group; 2023 Feb 14;9(1):1–22.

56. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius

M, Smibert P, Satija R. Comprehensive integration of single-cell data. Cell. 2019 Jun 13;177(7):1888-1902.e21. PMCID: PMC6687398

57. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. Nat Methods. Nature Publishing Group; 2021 Nov;18(11):1333–1341.

58. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. Nature Publishing Group; 2018 May;36(5):411–420.

59. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, Häring M, Braun E, Borm LE, La Manno G, Codeluppi S, Furlan A, Lee K, Skene N, Harris KD, Hjerling-Leffler J, Arenas E, Ernfors P, Marklund U, Linnarsson S. Molecular Architecture of the Mouse Nervous System. Cell. 2018 Aug 9;174(4):999-1014.e22. PMCID: PMC6086934

60. Saunders A, Macosko E, Wysoker A, Goldman M, Krienen F, de Rivera H, Bien E, Baum M, Wang S, Goeva A, Nemesh J, Kamitaki N, Brumbaugh S, Kulp D, McCarroll SA. Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell. 2018 Aug 9;174(4):1015-1030.e16. PMCID: PMC6447408

61. Tasic B, Yao Z, Graybuck LT, Smith KA, Nguyen TN, Bertagnolli D, Goldy J, Garren E, Economo MN, Viswanathan S, Penn O, Bakken T, Menon V, Miller J, Fong O, Hirokawa KE, Lathia K, Rimorin christine, Tieu M, Larsen R, casper T, Barkan E, Kroll M, Parry S, Shapovalova NV, Hirschstein D, Pendergraft J, Sullivan HA, Kim TK, Szafer A, Dee N, Groblewski P, Wickersham ian, cetin A, Harris JA, Levi BP, Sunkin SM, Madisen L, Daigle TL, Looger L, Bernard A, Phillips J, Lein E, Hawrylycz M, Svoboda K, Jones AR, Koch christof, Zeng H. Shared and distinct transcriptomic cell types across neocortical areas. Nature. 2018 Nov;563(7729):72–78. PMCID: PMC6456269

62. Yao Z, van Velthoven CTJ, Nguyen TN, Goldy J, Sedeno-Cortes AE, Baftizadeh F, Bertagnolli D, Casper T, Chiang M, Crichton K, Ding SL, Fong O, Garren E, Glandon A, Gouwens NW, Gray J, Graybuck LT, Hawrylycz MJ, Hirschstein D, Kroll M, Lathia K, Lee C, Levi B, McMillen D, Mok S, Pham T, Ren Q, Rimorin C, Shapovalova N, Sulc J, Sunkin SM, Tieu M, Torkelson A, Tung H, Ward K, Dee N, Smith KA, Tasic B, Zeng H. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. Cell. 2021 Jun 10;184(12):3222-3241.e26. PMCID: PMC8195859

63. BRAIN Initiative Cell Census Network (BICCN). A multimodal cell census and atlas of the mammalian primary motor cortex. Nature. 2021 Oct;598(7879):86–102. PMCID: PMC8494634

64. O'Leary TP, Sullivan KE, Wang L, Clements J, Lemire AL, Cembrowski MS. Extensive and spatially variable within-cell-type heterogeneity across the basolateral amygdala. eLife. 9:e59003. PMCID: PMC7486123

65. Beyeler A, Dabrowska J. Neuronal diversity of the amygdala and the bed nucleus of the stria terminalis. Handb Behav Neurosci. 2020;26:63–100. PMCID: PMC7423190

66. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010 Oct;11(10):10.1038/nrg2825. PMCID: PMC3880143

67. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, Chen J. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biology. 2020 Jan 16;21(1):12.

68. Banovich NE, Lan X, McVicker G, Geijn B van de, Degner JF, Blischak JD, Roux J, Pritchard JK, Gilad Y. Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. PLOS Genetics. Public Library of Science; 2014 Sep 18;10(9):e1004663.

69. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, Stephens M, Gilad Y, Pritchard JK. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. Nature Publishing Group; 2012 Feb;482(7385):390–394.

70. Munro D, Wang T, Chitre AS, Polesskaya O, Ehsan N, Gao J, Gusev A, Woods LCS, Saba LM, Chen H, Palmer AA, Mohammadi P. The regulatory landscape of multiple brain regions in outbred heterogeneous stock rats [Internet]. bioRxiv; 2022 [cited 2022 Aug 17]. p. 2022.04.07.487560. Available from: https://www.biorxiv.org/content/10.1101/2022.04.07.487560v1

71. Liu M, Tan X, Liu E, Hang Z, Song R, Mu S, Han W, Yue Q, Sun J. Inactivation of the Lateral Hypothalamus Attenuates Methamphetamine-Induced Conditioned Place Preference through Regulation of Kcnq3 Expression. Int J Mol Sci. 2022 Jun 30;23(13):7305. PMCID: PMC9266452

72. Tsuboi D, Otsuka T, Shimomura T, Faruk MO, Yamahashi Y, Amano M, Funahashi Y, Kuroda K, Nishioka T, Kobayashi K, Sano H, Nagai T, Yamada K, Tzingounis AV, Nambu A, Kubo Y, Kawaguchi Y, Kaibuchi K. Dopamine drives neuronal excitability via KCNQ channel phosphorylation for reward behavior. Cell Reports [Internet]. Elsevier; 2022 Sep 6 [cited 2023 Apr 21];40(10). Available from: https://www.cell.com/cell-reports/abstract/S2211-1247(22)01133-0 PMID: 36070693

73. Hansen HH, Andreasen JT, Weikop P, Mirza N, Scheel-Krüger J, Mikkelsen JD. The neuronal KCNQ channel opener retigabine inhibits locomotor activity and reduces forebrain excitatory responses to the psychostimulants cocaine, methylphenidate and phencyclidine. Eur J Pharmacol. 2007 Sep 10;570(1–3):77–88. PMID: 17628530

74. Cruz B, Vozella V, Carper BA, Xu JC, Kirson D, Hirsch S, Nolen T, Bradley L, Fain K, Crawford M, Kosten TR, Zorrilla EP, Roberto M. FKBP5 inhibitors modulate alcohol drinking and trauma-related behaviors in a model of comorbid post-traumatic stress and alcohol use disorder. Neuropsychopharmacol. Nature Publishing Group; 2022 Nov 18;1–11.

75. Levran O, Peles E, Randesi M, Li Y, Rotrosen J, Ott J, Adelson M, Kreek MJ. Stress-related genes and heroin addiction: a role for a functional FKBP5 haplotype.

Psychoneuroendocrinology. 2014 Jul;45:67–76. PMCID: PMC4316666

76. Heller EA, Kaska S, Fallon B, Ferguson D, Kennedy PJ, Neve RL, Nestler EJ, Mazei-Robison MS. Morphine and cocaine increase serum- and glucocorticoid-inducible kinase 1 activity in the ventral tegmental area. J Neurochem. 2015 Jan;132(2):243–253. PMCID: PMC4302038

77. Mohammadi P, Castel SE, Brown AA, Lappalainen T. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. Genome Res. 2017 Nov 1;27(11):1872–1884.

78. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545–15550. PMCID: PMC1239896

79. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. Nature Publishing Group; 2003 Jul;34(3):267–273.

80. Kasischke KA, Vishwasrao HD, Fisher PJ, Zipfel WR, Webb WW. Neural activity triggers neuronal oxidative metabolism followed by astrocytic glycolysis. Science. 2004 Jul 2;305(5680):99–103. PMID: 15232110

81. Attwell D, Laughlin SB. An energy budget for signaling in the grey matter of the brain. J Cereb Blood Flow Metab. 2001 Oct;21(10):1133–1145. PMID: 11598490

82. Distler MG, Plant LD, Sokoloff G, Hawk AJ, Aneas I, Wuenschell GE, Termini J, Meredith SC, Nobrega MA, Palmer AA. Glyoxalase 1 increases anxiety by reducing GABAA receptor agonist methylglyoxal. J Clin Invest. 2012 Jun 1;122(6):2306–2315. PMCID: PMC3366407

83. Perez CL, Barkley-Levenson AM, Dick BL, Glatt PF, Martinez Y, Siegel D, Momper JD, Palmer AA, Cohen SM. A Metal-Binding Pharmacophore Library Yields the Discovery of a Glyoxalase 1 Inhibitor. J Med Chem. 2019 Feb 14;62(3):1609–1625. PMCID: PMC6467756

84. Delaney AJ, Crane JW, Holmes NM, Fam J, Westbrook RF. Baclofen acts in the central amygdala to reduce synaptic transmission and impair context fear conditioning. Sci Rep. 2018 Jul 2;8:9908. PMCID: PMC6028433

85. Li C, Pleil KE, Stamatakis AM, Busan S, Vong L, Lowell BB, Stuber GD, Kash TL. Presynaptic inhibition of GABA release in the BNST by kappa opioid receptor signaling. Biol Psychiatry. 2012 Apr 15;71(8):725–732. PMCID: PMC3314138

86. de Guglielmo G, Conlisk DE, Barkley-Levenson AM, Palmer AA, George O. Inhibition of Glyoxalase 1 reduces alcohol self-administration in dependent and nondependent rats. Pharmacol Biochem Behav. 2018 Apr;167:36–41. PMCID: PMC5866249

87. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based Analysis of ChIP-Seq (MACS). Genome Biology. 2008 Sep 17;9(9):R137.

88. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–140. PMCID: PMC2796818

89. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014 Dec 5;15(12):550.

90. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nature Methods. Nature Publishing Group; 2017 Oct;14(10):975–978.

91. Matsuda T, Irie T, Katsurabayashi S, Hayashi Y, Nagai T, Hamazaki N, Adefuin AMD, Miura F, Ito T, Kimura H, Shirahige K, Takeda T, Iwasaki K, Imamura T, Nakashima K. Pioneer Factor NeuroD1 Rearranges Transcriptional and Epigenetic Profiles to Execute Microglia-Neuron Conversion. Neuron. Elsevier; 2019 Feb 6;101(3):472-485.e7. PMID: 30638745

92. Glahs A, Zinzen RP. Putting chromatin in its place: the pioneer factor NeuroD1 modulates chromatin state to drive cell fate decisions. EMBO J. 2016 Jan 4;35(1):1–3. PMCID: PMC4718001

93. Cruz FC, Rubio FJ, Hope BT. Using c-fos to study neuronal ensembles in corticostriatal circuitry of addiction. Brain Res. 2015 Dec 2;1628(0 0):157–173. PMCID: PMC4427550

94. Zhang Y, Crofton EJ, Li D, Lobo MK, Fan X, Nestler EJ, Green TA. Overexpression of DeltaFosB in nucleus accumbens mimics the protective addiction phenotype, but not the protective depression phenotype of environmental enrichment. Frontiers in Behavioral Neuroscience [Internet]. 2014 [cited 2022 Aug 4];8. Available from: https://www.frontiersin.org/articles/10.3389/fnbeh.2014.00297

95. Bali P, Kenny PJ. Transcriptional mechanisms of drug addiction. Dialogues Clin Neurosci. 2019 Dec;21(4):379–387. PMCID: PMC6952748

96. Walker DM, Cates HM, Loh YHE, Purushothaman I, Ramakrishnan A, Cahill KM, Lardner CK, Godino A, Kronman HG, Rabkin J, Lorsch ZS, Mews P, Doyle MA, Feng J, Labonté B, Koo JW, Bagot RC, Logan RW, Seney ML, Calipari ES, Shen L, Nestler EJ. Cocaine self-administration alters transcriptome-wide responses in the brain's reward circuitry. Biol Psychiatry. 2018 Dec 15;84(12):867–880. PMCID: PMC6202276

97. Nestler EJ, Barrot M, Self DW. ΔFosB: A sustained molecular switch for addiction. Proceedings of the National Academy of Sciences. 2001 Sep 25;98(20):11042–11046.

98. Hope BT, Nye HE, Kelz MB, Self DW, Iadarola MJ, Nakabeppu Y, Duman RS, Nestler EJ. Induction of a long-lasting AP-1 complex composed of altered Fos-like proteins in brain by

chronic cocaine and other chronic treatments. Neuron. Elsevier; 1994 Nov 1;13(5):1235–1244. PMID: 7946359

99. Nye HE, Nestler EJ. Induction of chronic Fos-related antigens in rat brain by chronic morphine administration. Mol Pharmacol. American Society for Pharmacology and Experimental Therapeutics; 1996 Apr 1;49(4):636–645. PMID: 8609891

100. Nye HE, Hope BT, Kelz MB, Iadarola M, Nestler EJ. Pharmacological studies of the regulation of chronic FOS-related antigen induction by cocaine in the striatum and nucleus accumbens. J Pharmacol Exp Ther. American Society for Pharmacology and Experimental Therapeutics; 1995 Dec 1;275(3):1671–1680. PMID: 8531143

101. Moratalla R, Elibol B, Vallejo M, Graybiel AM. Network-Level Changes in Expression of Inducible Fos–Jun Proteins in the Striatum during Chronic Cocaine Treatment and Withdrawal. Neuron. Elsevier; 1996 Jul 1;17(1):147–156. PMID: 8755486

102. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Gräf S, Huss M, Keefe D, Liu Z, London D, McDaniell RM, Shibata Y, Showers KA, Simon JM, Vales T, Wang T, Winter D, Zhang Z, Clarke ND, Birney E, Iyer VR, Crawford GE, Lieb JD, Furey TS. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 2011 Oct;21(10):1757–1767. PMCID: PMC3202292

103. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet. Nature Publishing Group; 2019 Apr;20(4):207–220.

104. Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front Genet. 2020 May 13;11:424. PMCID: PMC7237642

105. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics. Nature Publishing Group; 2015 Mar;47(3):291–295.

106. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, Datta G, Davila-Velderrain J, McGuire D, Tian C, Zhan X, Choquet H, Docherty AR, Faul JD, Foerster JR, Fritsche LG, Gabrielsen ME, Gordon SD, Haessler J, Hottenga JJ, Huang H, Jang SK, Jansen PR, Ling Y, Mägi R, Matoba N, McMahon G, Mulas A, Orrù V, Palviainen T, Pandit A, Reginsson GW, Skogholt AH, Smith JA, Taylor AE, Turman C, Willemsen G, Young H, Young KA, Zajac GJM, Zhao W, Zhou W, Bjornsdottir G, Boardman JD, Boehnke M, Boomsma DI, Chen C, Cucca F, Davies GE, Eaton CB, Ehringer MA, Esko T, Fiorillo E, Gillespie NA, Gudbjartsson DF, Haller T, Harris KM, Heath AC, Hewitt JK, Hickie IB, Hokanson JE, Hopfer CJ, Hunter DJ, Iacono WG, Johnson EO, Kamatani Y, Kardia SLR, Keller MC, Kellis M, Kooperberg C, Kraft P, Krauter KS, Laakso M, Lind PA, Loukola A, Lutz SM, Madden PAF, Martin NG, McGue M, McQueen MB, Medland SE, Metspalu A, Mohlke KL, Nielsen JB, Okada Y, Peters U, Polderman TJC, Posthuma D, Reiner AP, Rice JP, Rimm E, Rose RJ, Runarsdottir V, Stallings MC, Stančáková A, Stefansson H, Thai KK, Tindle HA,

Tyrfingsson T, Wall TL, Weir DR, Weisner C, Whitfield JB, Winsvold BS, Yin J, Zuccolo L, Bierut LJ, Hveem K, Lee JJ, Munafò MR, Saccone NL, Willer CJ, Cornelis MC, David SP, Hinds DA, Jorgenson E, Kaprio J, Stitzel JA, Stefansson K, Thorgeirsson TE, Abecasis G, Liu DJ, Vrieze S. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nature Genetics. Nature Publishing Group; 2019 Feb;51(2):237–244.

107. Polimanti R, Walters RK, Johnson EC, McClintick JN, Adkins AE, Adkins DE, Bacanu SA, Bierut LJ, Bigdeli TB, Brown S, Bucholz KK, Copeland WE, Costello EJ, Degenhardt L, Farrer LA, Foroud TM, Fox L, Goate AM, Grucza R, Hack LM, Hancock DB, Hartz SM, Heath AC, Hewitt JK, Hopfer CJ, Johnson EO, Kendler KS, Kranzler HR, Krauter K, Lai D, Madden PAF, Martin NG, Maes HH, Nelson EC, Peterson RE, Porjesz B, Riley BP, Saccone N, Stallings M, Wall TL, Webb BT, Wetherill L, Edenberg HJ, Agrawal A, Gelernter J. Leveraging genome-wide data to investigate differences between opioid use vs. opioid dependence in 41,176 individuals from the Psychiatric Genomics Consortium. Molecular Psychiatry. Nature Publishing Group; 2020 Aug;25(8):1673–1687.

108. Hatoum AS, Johnson EC, Colbert SMC, Polimanti R, Zhou H, Walters RK, Gelernter J, Edenberg HJ, Bogdan R, Agrawal A. The addiction risk factor: A unitary genetic vulnerability characterizes substance use disorders and their associations with common correlates. Neuropsychopharmacology. 2022 Sep;47(10):1739–1745. PMCID: PMC9372072

109. Wade CL, Vendruscolo LF, Schlosburg JE, Hernandez DO, Koob GF. Compulsive-Like Responding for Opioid Analgesics in Rats with Extended Access. Neuropsychopharmacology. 2015 Jan;40(2):421–428. PMCID: PMC4443956

110. Visscher PM, Yengo L, Cox NJ, Wray NR. Discovery and implications of polygenicity of common diseases. Science. American Association for the Advancement of Science; 2021 Sep 24;373(6562):1468–1473.

111. Zaret KS. Pioneer Transcription Factors Initiating Gene Network Changes. Annu Rev Genet. 2020 Nov 23;54:367–385. PMCID: PMC7900943

112. Chitre AS, Polesskaya O, Holl K, Gao J, Cheng R, Bimschleger H, Martinez AG, George T, Gileta AF, Han W, Horvath A, Hughson A, Ishiwari K, King CP, Lamparelli A, Versaggi CL, Martin C, St. Pierre CL, Tripi JA, Wang T, Chen H, Flagel SB, Meyer P, Richards J, Robinson TE, Palmer AA, Woods LCS. Genome wide association study in 3,173 outbred rats identifies multiple loci for body weight, adiposity, and fasting glucose. Obesity (Silver Spring). 2020 Oct;28(10):1964–1973. PMCID: PMC7511439

113. Stephens DN, King SL, Lambert JJ, Belelli D, Duka T. GABAA receptor subtype involvement in addictive behaviour. Genes Brain Behav. 2017 Jan;16(1):149–184. PMID: 27539865

114. Koob GF, Nestler EJ. The neurobiology of drug addiction. J Neuropsychiatry Clin Neurosci. 1997;9(3):482–497. PMID: 9276849

115. Koob GF. A role for GABA mechanisms in the motivational effects of alcohol.

Biochemical Pharmacology. 2004 Oct 15;68(8):1515–1525.

116. Dixon CI, Morris HV, Breen G, Desrivieres S, Jugurnauth S, Steiner RC, Vallada H, Guindalini C, Laranjeira R, Messas G, Rosahl TW, Atack JR, Peden DR, Belelli D, Lambert JJ, King SL, Schumann G, Stephens DN. Cocaine effects on mouse incentive-learning and human addiction are linked to α2 subunit-containing GABAA receptors. Proc Natl Acad Sci U S A. 2010 Feb 2;107(5):2289–2294. PMCID: PMC2836671

117. Augier E, Barbier E, Dulman RS, Licheri V, Augier G, Domi E, Barchiesi R, Farris S, Nätt D, Mayfield RD, Adermark L, Heilig M. A molecular mechanism for choosing alcohol over an alternative reward. Science. American Association for the Advancement of Science; 2018 Jun 22;360(6395):1321–1326.

118. Kallupi M, Wee S, Edwards S, Whitfield TW, Oleata CS, Luu G, Schmeichel BE, Koob GF, Roberto M. Kappa Opioid Receptor-Mediated Dysregulation of GABAergic Transmission in the Central Amygdala in Cocaine Addiction. Biol Psychiatry. 2013 Oct 1;74(7):520–528. PMCID: PMC3773286

119. McMurray KMJ, Ramaker MJ, Barkley-Levenson AM, Sidhu PS, Elkin P, Reddy MK, Guthrie ML, Cook JM, Rawal VH, Arnold LA, Dulawa SC, Palmer AA. Identification of a novel, fast acting GABAergic anti-depressant. Mol Psychiatry. 2018 Feb;23(2):384–391. PMCID: PMC5608625

120. Williams CL, Buchta WC, Riegel AC. CRF-R2 and the Heterosynaptic Regulation of VTA Glutamate during Reinstatement of Cocaine Seeking. J Neurosci. 2014 Jul 30;34(31):10402–10414. PMCID: PMC4115144

121. Bentzley BS, Aston-Jones G. Inhibiting subthalamic nucleus decreases cocaine demand and relapse: Therapeutic potential. Addict Biol. 2017 Jul;22(4):946–957. PMCID: PMC5010790

122. Shinohara F, Kamii H, Minami M, Kaneda K. The Role of Dopaminergic Signaling in the Medial Prefrontal Cortex for the Expression of Cocaine-Induced Conditioned Place Preference in Rats. Biol Pharm Bull. 2017;40(11):1983–1989. PMID: 29093348

123. Mitchell SJ, Maguire EP, Cunningham L, Gunn BG, Linke M, Zechner U, Dixon CI, King SL, Stephens DN, Swinny JD, Belelli D, Lambert JJ. Early-life adversity selectively impairs α2-GABAA receptor expression in the mouse nucleus accumbens and influences the behavioral effects of cocaine. Neuropharmacology. 2018 Oct;141:98–112. PMCID: PMC6178871

124. Sun W, Yuill MB. Role of the GABAa and GABAb Receptors of the Central Nucleus of the Amygdala in Compulsive Cocaine-seeking Behavior in Male Rats. Psychopharmacology (Berl). 2020 Dec;237(12):3759–3771. PMCID: PMC7686280

125. Pelloux Y, Minier-Toribio A, Hoots JK, Bossert JM, Shaham Y. Opposite Effects of Basolateral Amygdala Inactivation on Context-Induced Relapse to Cocaine Seeking after Extinction versus Punishment. J Neurosci. 2018 Jan 3;38(1):51–59. PMCID: PMC5761436

126. Madangopal R, Ramsey LA, Weber SJ, Brenner MB, Lennon VA, Drake OR, Komer LE, Tunstall BJ, Bossert JM, Shaham Y, Hope BT. Inactivation of the infralimbic cortex decreases discriminative stimulus-controlled relapse to cocaine seeking in rats. Neuropsychopharmacology. 2021 Oct;46(11):1969–1980. PMCID: PMC8429767

127. Pantazis CB, Aston-Jones G. Lateral septum inhibition reduces motivation for cocaine: reversal by diazepam. Addict Biol. 2019 Mar 21;10.1111/adb.12742. PMCID: PMC6754816

128. Cruz AM, Spencer HF, Kim TH, Jhou TC, Smith RJ. Prelimbic cortical projections to rostromedial tegmental nucleus play a suppressive role in cue-induced reinstatement of cocaine seeking. Neuropsychopharmacology. 2021 Jul;46(8):1399–1406. PMCID: PMC8209220

129. McMurray KMJ, Du X, Brownlee M, Palmer AA. Neuronal overexpression of Glo1 or amygdalar microinjection of methylglyoxal is sufficient to regulate anxiety-like behavior in mice. Behav Brain Res. 2016 Mar 15;301:119–123. PMCID: PMC4728018

130. Harris JJ, Jolivet R, Attwell D. Synaptic Energy Use and Supply. Neuron. 2012 Sep 6;75(5):762–777.

131. Boyer PD. What makes ATP synthase spin? Nature. Nature Publishing Group; 1999 Nov;402(6759):247–249.

132. Du F, Zhu XH, Zhang Y, Friedman M, Zhang N, Uğurbil K, Chen W. Tightly coupled brain activity and cerebral ATP metabolic rate. Proc Natl Acad Sci U S A. 2008 Apr 29;105(17):6409–6414. PMCID: PMC2359810

133. Erecińska M, Silver IA. ATP and Brain Function. J Cereb Blood Flow Metab. SAGE Publications Ltd STM; 1989 Feb 1;9(1):2–19.

134. Swinstead EE, Paakinaho V, Presman DM, Hager GL. Pioneer factors and ATP-dependent chromatin remodeling factors interact dynamically: A new perspective. Bioessays. 2016 Nov;38(11):1150–1157. PMCID: PMC6319265

135. Lepack AE, Werner CT, Stewart AF, Fulton SL, Zhong P, Farrelly LA, Smith ACW, Ramakrishnan A, Lyu Y, Bastle RM, Martin JA, Mitra S, O'Connor RM, Wang ZJ, Molina H, Turecki G, Shen L, Yan Z, Calipari ES, Dietz DM, Kenny PJ, Maze I. Dopaminylation of histone H3 in ventral tegmental area regulates cocaine seeking. Science. 2020 Apr 10;368(6487):197–201. PMCID: PMC7228137

136. Fulton SL, Mitra S, Lepack AE, Martin JA, Stewart AF, Converse J, Hochstetler M, Dietz DM, Maze I. Histone H3 dopaminylation in ventral tegmental area underlies heroin-induced transcriptional and behavioral plasticity in male rats. Neuropsychopharmacology. 2022 Sep;47(10):1776–1783. PMCID: PMC9372029

137. Werner CT, Mitra S, Martin JA, Stewart AF, Lepack AE, Ramakrishnan A, Gobira PH, Wang ZJ, Neve RL, Gancarz AM, Shen L, Maze I, Dietz DM. Ubiquitin-proteasomal regulation of chromatin remodeler INO80 in the nucleus accumbens mediates persistent

cocaine craving. Sci Adv. 2019 Oct 9;5(10):eaay0351. PMCID: PMC6785264

138. Werner CT, Mitra S, Auerbach BD, Wang ZJ, Martin JA, Stewart AF, Gobira PH, Iida M, An C, Cobb MM, Caccamise A, Salvi RJ, Neve RL, Gancarz AM, Dietz DM. Neuroadaptations in the dorsal hippocampus underlie cocaine seeking during prolonged abstinence. Proc Natl Acad Sci U S A. 2020 Oct 20;117(42):26460–26469. PMCID: PMC7585028

139. Calipari ES, Godino A, Salery M, Damez-Werno DM, Cahill ME, Werner CT, Gancarz AM, Peck EG, Jlayer Z, Rabkin J, Landry JA, Smith ACW, Defilippi P, Kenny PJ, Hurd YL, Neve RL, Dietz DM, Nestler EJ. Synaptic Microtubule-Associated Protein EB3 and SRC Phosphorylation Mediate Structural and Behavioral Adaptations During Withdrawal From Cocaine Self-Administration. J Neurosci. 2019 Jul 17;39(29):5634–5646. PMCID: PMC6636087

140. Carpenter MD, Hu Q, Bond AM, Lombroso SI, Czarnecki KS, Lim CJ, Song H, Wimmer ME, Pierce RC, Heller EA. Nr4a1 suppresses cocaine-induced behavior via epigenetic regulation of homeostatic target genes. Nat Commun. 2020 Jan 24;11:504. PMCID: PMC6981219

141. Duttke SH, Montilla-Perez P, Chang MW, Li H, Chen H, Carrette LLG, Guglielmo G de, George O, Palmer AA, Benner C, Telese F. Glucocorticoid Receptor-Regulated Enhancers Play a Central Role in the Gene Regulatory Networks Underlying Drug Addiction. Frontiers in Neuroscience [Internet]. 2022 [cited 2022 Jul 22];16. Available from: https://www.frontiersin.org/articles/10.3389/fnins.2022.858427

142. Guglielmo G de, Carrette LL, Kallupi M, Brennan M, Boomhower B, Maturin L, Conlisk D, Sedighim S, Tieu L, Fannon MJ, Martinez A, Velarde N, Kononoff J, Kimbrough A, Simpson S, Smith LC, Shankar K, Crook C, Avelar A, Schweitzer P, Woods LCS, Palmer AA, George O. Characterization of cocaine addiction-like behavior in heterogeneous stock rats [Internet]. bioRxiv; 2021 [cited 2022 Aug 19]. p. 2021.07.22.453410. Available from: https://www.biorxiv.org/content/10.1101/2021.07.22.453410v2

143. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR, Shah P, Bell JC, Jhutty D, Nemec CM, Wang J, Wang L, Yin Y, Giresi PG, Chang ALS, Zheng GXY, Greenleaf WJ, Chang HY. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat Biotechnol. Nature Publishing Group; 2019 Aug;37(8):925–936.

144. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH. Massively parallel digital transcriptional profiling of single cells. Nat Commun. Nature Publishing Group; 2017 Jan 16;8(1):14049.

145. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC, Nassar LR, Maulding ND, Lee CM, Lee BT, Hinrichs AS, Fyfe AC, Fernandes JD, Diekhans M, Clawson H, Casper J, Benet-Pagès A, Barber GP, Haussler D, Kuhn RM, Haeussler M, Kent WJ. The UCSC Genome Browser database: 2021 update. Nucleic Acids Research. 2021 Jan 8;49(D1):D1046–D1057.

146. Team TBD. BSgenome.Rnorvegicus.UCSC.rn6: Full genome sequences for Rattus norvegicus (UCSC version rn6). 2014.

147. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler and D. The Human Genome Browser at UCSC. Genome Res. 2002 Jun 1;12(6):996–1006.

148. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, Berry A, Bhai J, Bignell A, Billis K, Boddu S, Brooks L, Charkhchi M, Cummins C, Da Rin Fioretto L, Davidson C, Dodiya K, Donaldson S, El Houdaigui B, El Naboulsi T, Fatima R, Giron CG, Genez T, Martinez JG, Guijarro-Clarke C, Gymer A, Hardy M, Hollis Z, Hourlier T, Hunt T, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Marugán JC, Mohanan S, Mushtaq A, Naven M, Ogeh DN, Parker A, Parton A, Perry M, Piližota I, Prosovetskaia I, Sakthivel MP, Salam AIA, Schmitt BM, Schuilenburg H, Sheppard D, Pérez-Silva JG, Stark W, Steed E, Sutinen K, Sukumaran R, Sumathipala D, Suner MM, Szpak M, Thormann A, Tricomi FF, Urbina-Gómez D, Veidenberg A, Walsh TA, Walts B, Willhoft N, Winterbottom A, Wass E, Chakiachvili M, Flint B, Frankish A, Giorgetti S, Haggerty L, Hunt SE, IIsley GR, Loveland JE, Martin FJ, Moore B, Mudge JM, Muffato M, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Dyer S, Harrison PW, Howe KL, Yates AD, Zerbino DR, Flicek P. Ensembl 2022. Nucleic Acids Research. 2022 Jan 7;50(D1):D988–D995.

149. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, Fornes O, Leung TY, Aguirre A, Hammal F, Schmelter D, Baranasic D, Ballester B, Sandelin A, Lenhard B, Vandepoele K, Wasserman WW, Parcy F, Mathelier A. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Research. 2022 Jan 7;50(D1):D165–D173.

150. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. bioRxiv. Cold Spring Harbor Laboratory; 2019 Mar 18;576827.

151. Richards LM, Riverin M, Mohanraj S, Ayyadhury S, Croucher DC, Díaz-Mejía JJ, Coutinho FJ, Dirks PB, Pugh TJ. A comparison of data integration methods for single-cell RNA sequencing of cancer samples [Internet]. bioRxiv; 2021 [cited 2022 May 19]. p. 2021.08.04.453579. Available from: https://www.biorxiv.org/content/10.1101/2021.08.04.453579v1

152. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, Linsley PS, Gottardo R. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA

sequencing data. Genome Biol [Internet]. 2015 [cited 2020 Oct 26];16. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4676162/ PMCID: PMC4676162

153. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu X, Liu S, Bo X, Yu G. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021 Aug 28;2(3):100141.

154. Stuart T, Srivastava A, Lareau C, Satija R. Multimodal single-cell chromatin analysis with Signac. bioRxiv. Cold Spring Harbor Laboratory; 2020 Nov 10;2020.11.09.373613.

155. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–842.

156. Cusanovich DA, Daza R, Adey A, Pliner H, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J. Multiplex Single Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing. Science. 2015 May 22;348(6237):910–914. PMCID: PMC4836442

157. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990;41(6):391–407.

158. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P ru, Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods. Nature Publishing Group; 2019 Dec;16(12):1289–1296.

159. Yirga AA, Melesse SF, Mwambi HG, Ayele DG. Negative binomial mixed models for analyzing longitudinal CD4 count data. Sci Rep. Nature Publishing Group; 2020 Oct 7;10(1):16742.

160. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep;81(3):559–575. PMCID: PMC1950838

161. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010 May 28;38(4):576–589. PMCID: PMC2898526

CHAPTER 2: Genome-wide analysis of CRISPR perturbations indicates that enhancers act multiplicatively and without epistatic-like interactions

CHAPTER 2: Genome-wide analysis of CRISPR perturbations indicates that enhancers act multiplicatively and without epistatic-like interactions

2.1: Abstract

A single gene may be regulated by multiple enhancers, but how they work in concert to regulate transcription is poorly understood. Prior studies have mostly examined enhancers at single loci and have reached inconsistent conclusions about whether epistatic-like interactions exist between them. To analyze enhancer interactions throughout the genome, we developed a statistical framework for CRISPR regulatory screens that utilizes negative binomial generalized linear models that account for variable guide RNA (gRNA) efficiency. We reanalyzed a single-cell CRISPR interference experiment that delivered random combinations of enhancer-targeting gRNAs to each cell and interrogated interactions between 3,808 enhancer pairs. We found that enhancers act multiplicatively with one another to control gene expression, but our analysis provides no evidence for interaction effects between pairs of enhancers regulating the same gene. Our findings illuminate the regulatory behavior of multiple enhancers, and our statistical framework provides utility for future analyses studying interactions between enhancers.

2.2: Introduction

Cis-regulatory elements (CREs), which include enhancers, direct transcription and shape cellular identity, growth, and biological function. Most genes are regulated by multiple enhancers[1,2], yet we lack a detailed understanding of how enhancers act together to influence gene expression. When multiple enhancers for a gene are active in the same cell type, it is often assumed that they act additively—that is, their combined effect is equal to the sum of their

individual effects[3]. However, enhancers may also act non-additively, and interactions between regulatory elements may modulate their effects on gene expression[3–10].

To date, most studies of regulatory elements have examined their effects independently, and studies of regulatory element interactions have focused on a small number of loci and have reached contradictory conclusions[4–8]. For example, a study of the *a-globin4* gene found that its expression is best explained by simple additivity between constituent elements of its super enhancer[7]. In addition, a study that deleted three constituent enhancers of a super enhancer for *Wap3* found no evidence of synergy between the studied enhancers and differences in the magnitudes of effect that each constituent enhancer had on the target gene, with all three enhancers necessary to induce full induction of the gene during pregnancy[8]. Reexamination of both of these super enhancer datasets found that the effects of the constituent enhancers on the target genes were best described by a logistic generalized linear model (GLM), but that beyond this there was no significant evidence for interactions between enhancers[5]. Contrary to these findings, a recent study of the *MYC* locus described both synergistic and additive enhancer-enhancer interactions, where enhancers separated from one another by larger genomic distances are more likely to have synergistic interactions and enhancers located closer to one another are more likely to have additive interactions[9]. Altogether, these studies have been limited to the examination of a small number of genes and enhancers and their results are difficult to interpret due to their conflicting findings and the lack of explicit definitions and consistent terminology for different models of enhancer activity.

Recent technological advances have made it possible to couple CRISPR-induced genome perturbations with single-cell RNA sequencing[10–16]. Because single-cell CRISPR perturbation experiments can induce multiple genomic perturbations in each cell, they can be used to identify

interactions, or epistatic-like effects, between targeted sequences. Specifically, when such experiments are designed to target regulatory elements, they yield cells wherein multiple regulatory elements are simultaneously perturbed. This feature of these datasets can be harnessed to measure the combined effects of multiple regulatory elements, such as enhancers, on gene expression.

Here, we present GLiMMIRS (Generalized Linear Models for Measuring Interactions between Regulatory Sequences), a statistical analysis framework that can be applied to single cell CRISPR perturbation experiments to quantify the effects of multiple regulatory elements on gene expression and identify interactions between them. GLiMMIRS has both data simulation and modeling components and importantly, accounts for variations in gRNA efficiency, a key variable in the interpretation of CRISPR experiments that has typically been ignored when analyzing data from them[10,11,17–19]. We applied GLiMMIRS to a multiplexed, single-cell CRISPR interference (CRISPRi) experiment that targeted putative enhancers in K562 cells[11]. We conducted a power analysis, which found that this dataset provides sufficient power to detect strong interactions between enhancers, but low power to detect weak interactions. Our analysis strongly supports a model in which most enhancers act multiplicatively to affect the expression of their target genes, but we find no evidence for the presence of additional interactions between them.

2.3: Results

2.3.1: Variation in guide efficiency should be considered when estimating enhancer effects from CRISPR perturbations

To analyze the combined effect of multiple enhancers on gene expression, we leveraged data from a multiplexed, single-cell CRISPRi screen performed in K562 cells[11]. In this screen,

gRNAs were designed to target putative enhancers and enhancer-gene pairs were identified by associating perturbed enhancers with differences in the expression of nearby genes (Figure 2.1a). Due to the high multiplicity of infection (MOI) used in this experiment, many gRNAs targeting different enhancers are present within each cell (Figure 2.1a). While the high MOI was intended to increase power to detect enhancer-gene pairs, we leveraged this feature of the dataset to quantify how pairs of enhancers regulate the expression of common target genes and to detect potential interaction effects between them (Figure 2.1b). In particular, we focused on cells which received gRNAs targeting pairs of enhancers within 1Mb of the same gene, which we designate as the putative target gene[20–22].

Most enhancers in this dataset were targeted by two different gRNAs. The original study did not distinguish between gRNAs that targeted the same enhancer; however, it is important to consider differences in guide efficiency when examining the combined effects of multiple enhancers in a CRISPR screen. This is because the joint effect of both enhancer perturbations can appear smaller or larger than expected if one of the targeting guides has low efficiency. To illustrate this concept, we examined two enhancers of *NMU*, which were among the most significant enhancer-gene pairs discovered by the original study. We performed CRISPRi experiments to perturb the enhancers of *NMU* using guide designs from the paper (Figure 2.1c, Supplemental File 2.1). We quantified gene expression following each perturbation using reverse transcription-quantitative polymerase chain reaction (RT-qPCR) and found that one of the two gRNAs targeting the first enhancer (enhancer A, gRNAs A1 and A2) caused much larger reductions in *NMU* expression (Figure 2.1d). Differences in guide efficiency like the ones we observed for gRNAs A1 and A2 can give false signals of epistatic-like interactions if different guides targeting the same enhancers are treated as equivalent. For example, if by chance most of

the cells which contained guides targeting both enhancer A and B contained gRNA A1 (rather than the inefficient A2), then the joint effect of targeting both enhancers could be greatly overestimated.

To examine variation in guide efficiency, we estimated the efficiency of the gRNAs included in the experiment using GuideScan 2.0[23]. Predicted guide efficiency varies substantially across the guide library (Figure 2.1e), indicating that it is important to consider this variable when analyzing enhancer interactions using this dataset.

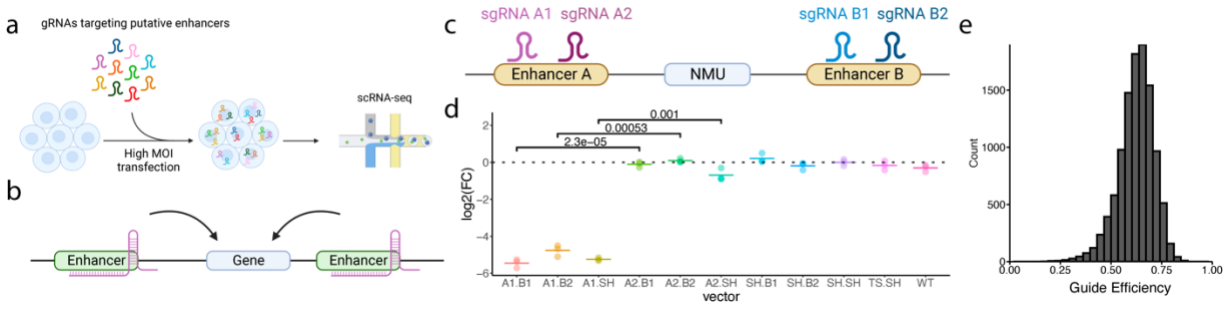


Figure 2.1: Variation in guide efficiency should be considered when estimating enhancer effects from CRISPR perturbations.a) Schematic of the Gasperini et al. experiment. A library of gRNAs targeting putative enhancers was transduced into cells with a high multiplicity of infection (MOI), resulting in multiple perturbations per cell. The identities of the gRNAs and their effects on gene expression were read out with single-cell RNA-seq (scRNA-seq). **b)** Schematic of two enhancers acting on the same gene. We seek to quantify the effect on multiple enhancers acting on a single gene. **c)** Schematic of CRISPR perturbation experiment targeting enhancers of NMU with two gRNAs per enhancer. **d)** Results of CRISPRi RT-qPCR experiment perturbing NMU enhancers for three technical replicates. For each NMU enhancer (enhancers A and B), two gRNAs were used (A1, A2 and B1, B2, respectively) and delivered on the same vector. Vectors containing gRNA A1 resulted in larger fold changes in NMU expression than their counterparts containing gRNA A2 instead (denoted p-values come from unpaired Welch's two-sided t-tests against the null hypothesis that there is no difference in mean fold change (FC) between vectors using gRNA A1 vs. gRNA A2. SH = safe harbor). TS = NMU transcription start site, WT = wild type K562 cells expressing dCas9-KRAB without any gRNAs, horizontal bar = mean $\log_2$(FC). See also Supplemental File 2.1. **e)** Distribution of guide efficiency values predicted by GuideScan 2.0 for the gRNAs used in the Gasperini et al. experiment.

2.3.2: GLiMMIRS provides a modeling and simulation framework for quantifying enhancer effects from CRISPR screens

We developed GLiMMIRS, a dual modeling and simulation framework for analyzing data from CRISPR screens to evaluate the effects of regulatory elements on target genes. We first sought to evaluate the utility of a model that incorporates guide efficiency by testing a

simple model that considers just one enhancer acting on one gene, which we refer to as the GLiMMIRS baseline model (GLiMMIRS-base) (Figure 2.2a). For each enhancer and gene of interest, we fit a generalized linear model (GLM) with a negative binomial distribution to the observed scRNA-seq counts. The predictor of interest in this model is the probability that the enhancer is perturbed, $X_{perturb}$. We calculated the value of $X_{perturb}$ using the efficiencies of the targeting sgRNAs which are present in each cell (see Methods). In addition to considering guide efficiency, we also included covariates to account for cell cycle[24] and other relevant variables (see Methods) [24].

To evaluate the performance of GLiMMIRS-base, we developed a simulation framework for single-cell CRISPRi screens (Figure 2.2b, see Methods) and used it to generate a dataset resembling the Gasperini et al.[11] experimental dataset, with gRNAs targeting the enhancers of predetermined target genes. This is the simulation component of GLiMMIRS (GLiMMIRS-sim), designed for evaluation of our baseline scenario. This provided us with a set of ground truth coefficient values which we could use to benchmark our model. We generated scRNA-seq counts for each gene by sampling from a negative binomial distribution defined by gene-specific parameters (Figure 2.2b, Methods). We then fit our baseline model to the simulated count data and compared the estimated model coefficients to the "ground truth" values used in the simulation. The coefficient of determination (R^2, see Methods) between the estimated enhancer effect coefficients and the ground truth values was higher ($R^2 = 0.657, MSE = 0.52$, Pearson's $r = 0.862$) when we implemented our model with a perturbation probability, $X_{perturb}$ (see Methods), compared to a model that used a simple indicator value representing the presence or absence of targeting gRNAs for the enhancer being modeled ($R^2 = -0.449, MSE = 2.195$, Pearson's $r = 0.811$) (Figure 2.2c, Table 2.1). This is because the model that uses the indicator

value systematically underestimates the enhancer effect, by assuming that the presence of a gRNA completely inhibits the target site even when the gRNA has low efficiency. We also generated "noisy" guide efficiency values with GLiMMIRS-sim (Supplemental Figure 2.1a-b) to account for uncertainty in predicted guide efficiencies[25–28]. These noisy guide efficiency values were calculated as a function of true guide efficiency and a noise-controlling constant D (see Methods), where D is inversely related to the amount of noise in the efficiency value. We found that fitting to the simulated data using the values of $X_{perturb}$ computed from the noisy guide efficiencies still performed better than an indicator variable under low noise ($D = 100$; $R^2 = 0.642, MSE = 0.542$, Pearson's $r = 0.854$) and medium noise ($D = 10$; $R^2 = 0.499, MSE = 0.752$, Pearson's $r = 0.789$). Under a simulation with very noisy guide efficiencies, the coefficient estimates correlated very poorly with the ground truth due to the presence of some extreme outliers ($D = 1$; $R^2 = -6107.575, MSE = 8937.909$, Pearson's $r = 0.03$) (Supplemental Figure 2.1c, Table 2.2). In summary, accounting for guide efficiency improves the accuracy in coefficient estimates and is robust to moderate noise in the guide efficiency estimates.

We then applied GLiMMIRS-base to the Gasperini et al.[11] dataset and compared the p-values obtained from our GLM to those from the published analysis. We detected a similar number of significant enhancer-gene pairs (588 validated by GLiMMIRS-base out of the 664 reported by Gasperini et al.[11]), but with lower p-values for most of the highly significant pairs. Our p-values are well-calibrated, and when applied to permuted data (where gRNA identities are assigned to different cells) the p-value distribution matches the null expectation (Figure 2.2d). These results establish that accounting for guide efficiency offers advantages over an indicator variable for gRNA presence and suggest that including cell cycle scores as additional covariates

in GLiMMIRS may further boost power to detect enhancer-gene pairs. Having established the validity of our approach for the simpler scenario of single enhancers acting on single genes, we proceeded to study the effects of pairs of enhancers on single genes.

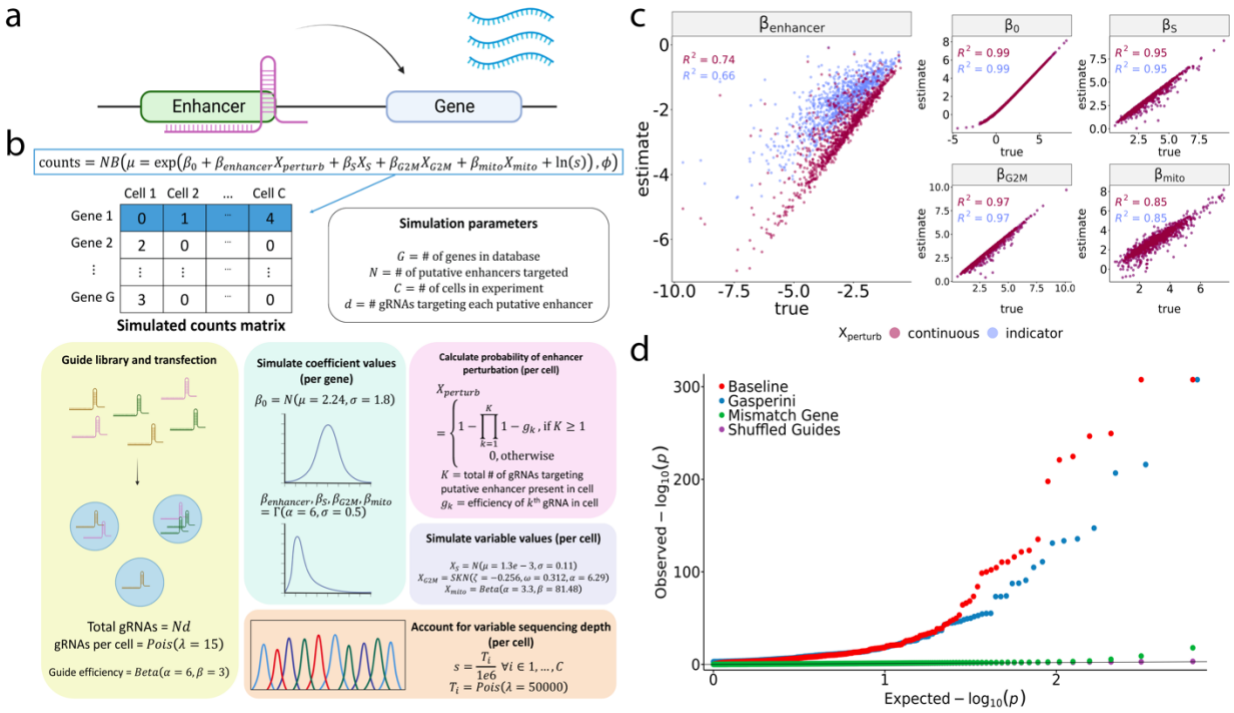


Figure 2.2: GLiMMIRS provides a modeling and simulation framework for quantifying enhancer effects from CRISPR screens.

a) A schematic of GLiMMIRS-base, wherein we evaluate the effect of a single putative enhancer on a single target gene. We model count data with a negative binomial generalized linear model (GLM). **b)** Schematic of data simulation for a single-cell CRISPRi experiment perturbing enhancers. Coefficient values (β) were simulated for each gene and corresponding variable values (X) were simulated for each cell. $X_{perturb}$ was calculated as a function of simulated guide efficiency. Values were sampled from distributions that resembled the empirical data whenever possible. We also simulated a per cell scaling factor to account for sequencing depth. **c)** Scatterplot comparing true versus estimated coefficient values for 1000 genes modeled with GLiMMIRS-base. These genes were designated as "true" target genes in the simulation, meaning their enhancers were targeted in the simulated experiment. Shows results of fitting to simulated data using a value of $X_{perturb}$ calculated from guide efficiency (continuous), representing perturbation probability, versus an indicator variable (indicator). A pseudocount of 0.01 was added to the counts (see also Supplemental Figure 2.4). Coefficients of determination (R^2) are shown. See also Table 2.1. **d)** Quantile-quantile plot of observed versus expected -log$_{10}$p indicates similarity between GLiMMIRS-base and the results published by Gasperini et al. The baseline values (orange) indicate the results of GLiMMIRS-base. The Gasperini values (green) indicate the previously published results. Mismatch gene and scrambled perturbation are negative control models. Mismatch gene (purple) compares an enhancer with a randomly assigned gene expression vector, while scrambled perturbation (yellow) shuffles the vector of guide perturbation probabilities.

Table 2.1: Fitting GLiMMIRS-base to simulated data comparing perturbation probability to indicator variable for $X_{perturb}$.

Pearson correlation (r), mean squared error (MSE), and coefficient of determination (R^2) between true and estimated coefficient values for each coefficient in the baseline model when fitting with guide-efficiency derived value of $X_{perturb}$ versus with an indicator (0/1) value for $X_{perturb}$.

X.perturb	term	r	p_val	MSE	R2
continuous	(Intercept)	0.99681599	0	0.03726403	0.98812193
continuous	guide.eff	0.86197003	1.07E-296	0.52009733	0.65668657
continuous	s.score	0.97391272	0	0.10758061	0.93555405
continuous	g2m.score	0.98275344	0	0.06522213	0.95666934
continuous	percent.mito	0.92464345	0	0.25841741	0.8216003
indicator	(Intercept)	0.99681564	0	0.03727181	0.98811945
indicator	guide.eff	0.81093353	1.65E-234	2.19548345	-0.4492268
indicator	s.score	0.97391118	0	0.10758667	0.93555042
indicator	g2m.score	0.98275542	0	0.06521277	0.95667556
indicator	percent.mito	0.92465712	0	0.25838458	0.82162297

Table 2.2: Fitting GLiMMIRS-base to simulated data comparing different levels of noise in guide efficiency estimates.

Pearson correlation (r), mean squared error (MSE) and coefficient of determination (R^2) between true and estimated coefficient values for each coefficient in the baseline model when fitting with perturbation probabilities ($X_{perturb}$) calculated from different sets of noisy guide efficiency estimates.

D	term	r	p_val	MSE	R2
1	(Intercept)	0.99677719	0	0.03735124	0.98801518
1	guide.eff	0.02978161	0.34874436	8937.90936	-6107.5746
1	s.score	0.97358575	0	0.10833667	0.93470673
1	g2m.score	0.98214717	0	0.06559903	0.95510971
1	percent.mito	0.92459247	0	0.25991558	0.82114178
10	(Intercept)	0.99680445	0	0.03727269	0.98807399
10	guide.eff	0.78897122	4.10E-213	0.75236876	0.49898029
10	s.score	0.97391847	0	0.10767907	0.93555831
10	g2m.score	0.98275896	0	0.06528683	0.95666923
10	percent.mito	0.92468192	0	0.25864625	0.82156289
100	(Intercept)	0.99681607	0	0.03726286	0.98812231
100	guide.eff	0.8542484	8.50E-286	0.54202322	0.6422134
100	s.score	0.97391254	0	0.10758247	0.93555293
100	g2m.score	0.98275351	0	0.06522047	0.95667044
100	percent.mito	0.92464094	0	0.25841917	0.82159908

2.3.3: GLiMMIRS-int detects interactions between pairs of enhancers

To model the effects of pairs of enhancers on a target gene, we modified GLiMMIRS-base by replacing the enhancer term $\beta_{enhancer}X_{perturb}$ with three new terms to represent: 1) the

first enhancer in the pair ($\beta_A X_A$); 2) the second enhancer in the pair ($\beta_B X_B$); and 3) an epistatic-like interaction between the enhancers ($\beta_{AB} X_{AB}$). As with the baseline model above, the values of the X_A and X_B predictors are the probability that the respective enhancers are perturbed. Likewise, the value of X_{AB} is the probability that both enhancers are simultaneously perturbed and is also estimated from the predicted guide efficiencies. This new model, which evaluates interaction effects between pairs of enhancers, is the GLiMMIRS interactions model (GLiMMIRS-int).

To identify pairs of enhancers to test in the experimental data, we identified target sites from the Gasperini et al. experiment, or putative enhancers, which were both located within 1MB of a common target gene as testable enhancer pairs. We found a total of 795,616 testable enhancer pairs from the set of enhancers targeted in the Gasperini et al.[11] study. Since cells must contain perturbations of multiple enhancers to determine whether there is an interaction effect between the enhancers, we evaluated the number of cells containing gRNAs targeting both enhancers within testable pairs. While the majority of testable enhancer pairs are simultaneously perturbed in fewer than 10 cells, several hundred enhancer pairs are simultaneously targeted in at least 10 cells (Figure 2.3a).

We performed a power analysis to evaluate our power for detecting interactions at different MOI, represented by different values of λ (see Methods) (Supplemental Figure 2.2a), and different magnitudes of (fixed) interaction effect sizes (Figure 2.3b-c). To do this, we used GLiMMIRS-sim to generate ground truth data for evaluating interactions between enhancer pairs (see Methods). In our power analysis, we defined positive cases as enhancer pairs with a true interaction effect on their target gene and negative cases as pairs of enhancers with individual effects on the target gene but no interaction effect. As expected, we observed that power to detect

interaction effects scales with the magnitude of the interaction effect size as well as the MOI, which controls the number of testable cells (Supplemental Figure 2.2b-c). Our power analysis indicated that we have low power (<25%) to detect interactions of small effect sizes (<2), particularly at low MOIs ($\lambda = 15,25$). This is likely because the number of testable cells, or cells containing gRNAs targeting both enhancers in a testable pair, are very low (Figure 2.3a). The scenario $\lambda = 15$ from our power analysis most closely resembles the empirical data (Supplemental Figure 2.2, Figure 2.3a), indicating that we have moderate power (>50%) to detect large interaction effects (≥ 5) and low power to detect smaller effects. Thus, with the experimental dataset analyzed in our study, we expect that we will have sufficient power to detect strong interaction effects between enhancers but be unable to draw conclusions about the presence or absence of weak interactions.

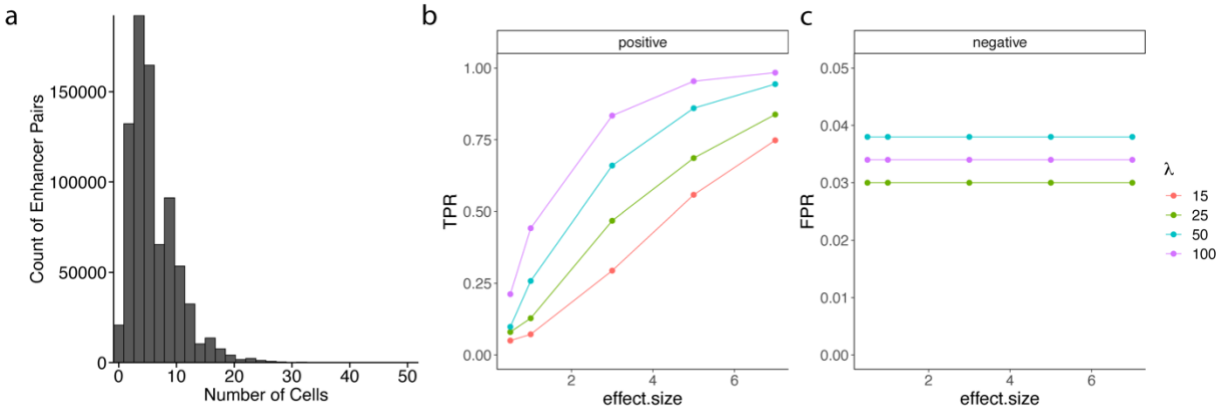


Figure 2.3: GLiMMIRS-int detects interactions between pairs of enhancers.

a) Distribution of the frequency of all testable target site pairs in the Gasperini et al. dataset. Criteria for testable pairs are defined as pairs of target sites, or putative enhancers, located within 1MB of a common target gene that are simultaneously perturbed in the same cells. **b-c)** Results of power analysis for ability to detect interaction effects in simulated datasets with varying multiplicities of infection λ (MOI) and effect sizes (x-axis). We calculated **b)** true positive rate (TPR), or power, from the "positive" ground truth enhancer pairs with interaction effects that we simulated, and **c)** false positive rate (FPR) from the "negative" control enhancer pairs without interaction effects that we simulated.

2.3.4: Enhancers act multiplicatively to control gene expression, but analysis of CRISPR perturbations provide no evidence for interactions

We next applied GLiMMIRS to the Gasperini et al.[11] CRISPRi dataset to study enhancer-enhancer interactions. To survey for interactions between enhancers, we defined two sets of testable enhancer pairs throughout the genome: a smaller, high-confidence set and a larger, unbiased set of testable pairs (see Methods). The high-confidence set consisted of 330 testable pairs and corresponding target genes where each of the individual enhancers had a previously reported regulatory effect on the target gene. The unbiased set consisted of all testable pairs that were perturbed in a minimum of 20 cells, regardless of any previously established relationship between each individual enhancer and the target gene. The unbiased set contained 3,808 enhancer pairs and target genes.

We first examined whether the combined effects of multiple enhancers on gene expression were better described by a multiplicative or additive model. To this end, we fit two versions of GLiMMIRS-int to the 330 enhancer pairs and their target genes in the high-confidence set: an additive model, in which we used an identity link function and a multiplicative model, in which we used a log link function. We then compared the model fits with Akaike Information Criterion (AIC). This approach is similar to that used by Dukler et al.[5] to compare additive, exponential and logistic models for two genes. In all cases, the multiplicative model provided a better fit, indicating that the combined effect of enhancers is better described by a multiplicative model (Figure 2.4a). Thus, we used the multiplicative form of GLiMMIRS-int in all subsequent analyses.

We applied GLiMMIRS-int to the 330 enhancer pairs in the unbiased set and observed no significant interaction terms (Likelihood Ratio Test, FDR<0.1) (Figure 2.4b). When applying

GLiMMIRS-int to the 3,808 enhancer pairs where each constituent enhancer did not necessarily have a significant effect on gene expression, we identified 4 significant interaction term effects with this model (Likelihood Ratio Test, FDR<0.1) (Figure 2.4b). These interactions were observed at the *EXOC8*, *BABAM2*, *H2BC12*, and the *ZBED9* gene loci, and all significant interaction terms were positive (Figure 2.4c)

We examined the distribution of single-cell RNA-seq read counts for the four genes with significant interaction terms, focusing on the cells that received guides targeting both corresponding enhancers. For all four genes, we noted that there was a single outlier cell with high read counts that received both guides (Figure 2.4d). Since GLM coefficients and p-values can be influenced by outliers, we performed a bootstrap analysis of the interaction coefficients (β_{AB}), which is less sensitive to outliers. For each of the enhancer pairs and their corresponding target genes, we resampled cells with replacement 100 times, fit GLiMMIRS-int to the resampled data, and recorded the β_{AB} estimates. The 99% bootstrap confidence intervals for β_{AB} for all four genes spanned zero (Figure 2.4e). We additionally performed a permutation test of β_{AB} to obtain p-values that are more robust to outliers. We shuffled the assignments of gRNAs in cells for the gRNAs targeting both enhancers in each pair jointly 10,000 times, and fit GLiMMIRS-int to the permuted data to obtain a null distribution of interaction coefficients. Two of the p-values obtained by this approach were nominally significant (p=0.0077 and p=0.0003 by two-sided permutation test) but would not withstand multiple testing correction given the total number of tests performed (Figure 2.4b). In combination, these results indicate that the four significant interaction terms are largely driven by cells with outlier expression of the target gene, and that there is insufficient evidence to reject the null hypothesis of no interactions between enhancers.

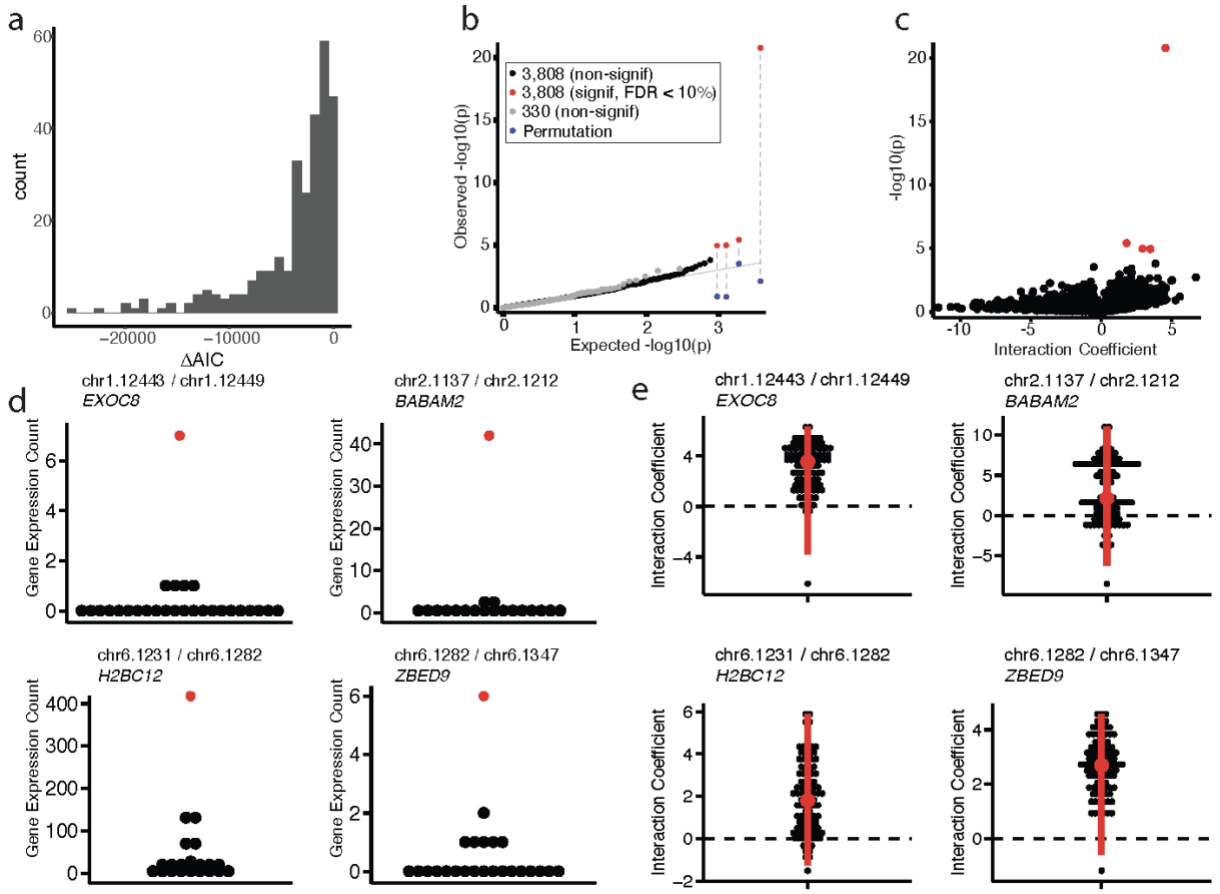


Figure 2.4: Enhancers act multiplicatively to control gene expression, but analysis of CRISPR perturbations provide no evidence for interactions.

a) Distribution of *Δ*AIC, the difference in Akaike Information Criterion between the best fitting model and the lesser model for 330 high confidence enhancer pairs and corresponding target genes from Gasperini et al. In every case evaluated, the multiplicative model fit better than the additive model. **b)** Quantile-quantile plot of interaction coefficient p-values for 330 high confidence enhancer pairs, where each individual enhancer had significant effects on the target gene expression, and 3,808 unbiased enhancer pairs, where each constituent enhancer did not necessarily have a significant effect on gene expression. No enhancer pairs among the 330 high confidence pairs had significant interaction coefficients after multiple testing correction (gray). Four significant interactions were observed for the 3,808 unbiased pairs at the *EXOC8*, *BABAM2*, *H2BC12*, and *ZBED9* gene loci (red) (FDR<0.1). Permutation test p-values for these four loci are shown in blue. Non-significant cases from the 3,808 unbiased pairs are shown in black. **c)** Volcano plot of interaction coefficients for the 3,808 unbiased pairs; significant interaction coefficients (FDR<0.1) are indicated in red. **d)** Gene expression counts from cells containing guides targeting both enhancers in a testable pair for the four genes with significant interaction terms. For all four genes, among the cells containing gRNAs targeting both enhancers in a pair, there contained a single outlier cell with extreme gene expression counts (red). **e)** Bootstrapping analysis of the four significant enhancer interactions. Red dots indicate the median coefficient estimate; red lines indicate 99% confidence intervals.

2.4: Discussion

CRISPR perturbations provide a new way to measure how combinations of enhancers regulate gene expression. We reanalyzed data from a single-cell CRISPRi experiment designed to map enhancers to the genes that they regulate. Since this dataset transduced guide RNAs with a high MOI, multiple enhancers near to (within 1MB of) the same gene were sometimes perturbed within the same cells, making it possible to analyze the joint effects of multiple enhancers on a common target gene. Our analysis supports a model in which enhancers act multiplicatively to control gene expression. Such a model was previously proposed by Dukler et al.[5], whose analysis of two loci in the genome supported either a logistic or multiplicative model of regulatory activity over an additive model[5]. Our genome-wide analysis confirms that a multiplicative model of enhancer activity fits the data in our analysis very well. The multiplicative model consistently provides a better fit than an additive model (Figure 2.4a) and statistics obtained from applying our multiplicative model to 3,808 unbiased testable pairs in the experimental data closely resemble those expected under the null hypothesis of no enhancer interactions (Figure 2.4b). The logistic model would be considered a refinement of a multiplicative model in which the expression of a gene has a maximum threshold that can be achieved by the activity of its enhancers. However, we cannot formally distinguish between logistic and multiplicative models with our dataset because this would require examining interactions between more than two enhancers for a single gene.

A limitation of the dataset that we analyzed is that even with a high MOI and a large number of sequenced cells, only a small subset of enhancer pairs could be interrogated. Specifically, we only tested 3,808 out of a possible 795,616 testable enhancer pairs because most enhancer pairs satisfying our testing criteria were not simultaneously perturbed in a sufficient

number of cells. Furthermore, we only had sufficient power to detect interactions that exerted at least a moderately strong effect on expression (e.g. 29.4% power to detect interactions with an absolute effect size of 3 or greater at a simulated MOI of $\lambda = 15$). Many of these power limitations could be overcome through CRISPRi experiments designed specifically to probe enhancer interactions. For example, a high MOI CRISPRi experiment could be performed in which a much smaller number of candidate enhancers are targeted so that testable pairs are frequently perturbed simultaneously in the same cells. Multiple guides could also be transduced on the same vectors so that nearby enhancers are guaranteed to be targeted in many cells. This latter approach was recently used to estimate enhancer interactions at the MYC locus[9].

Further limitations of our analysis are that we only analyzed data from a single cell line under a single condition, and it is possible that enhancer interactions are more prevalent under dynamic conditions or in different cell types.

Despite the above limitations, our results argue against the presence of strong epistatic interactions between enhancers. If such interactions do exist, they must be infrequent, of small effect, or restricted to specific cell types or conditions. How can these observations be reconciled with prior reports of enhancer redundancy or synergy? A possible explanation is that an interaction term is required by additive models because the combined effects of multiple enhancers is greater (synergistic) or less than (redundant) than expected under an additive model. However, these deviations from additivity may be naturally accounted for by a multiplicative model without the need for an interaction term. For example, under a multiplicative model, perturbation of a weak enhancer may have a small or negligible effect on expression but would have a much more substantial effect when combined with a perturbation to a strong enhancer. An

additive model would require an interaction term to describe these results and the enhancers would appear to be 'redundant'.

A recent study by Lin et al. analyzed enhancer interactions at the *MYC* locus using pairs of CRISPR guides and reported additive interactions between nearby enhancers, and synergistic interactions between distant enhancers[9]. In our dataset, we did not observe any differences in interactions between enhancers that were close together or far apart (Supplemental Figure 2.3); however, it is difficult to compare our results with those from Lin et al. for two reasons. First, the high-throughput screen in Lin et al. was performed using cell proliferation as readout, rather than gene expression, thereby assuming that proliferation was proportional to *MYC* expression. Second, while Lin et al. examined how selected pairs of enhancers affect the expression of *MYC* and other genes, their analysis relied on log relative expression obtained by RT-qPCR, which is not directly comparable to scRNA-seq expression estimates.

Future studies which examine enhancer interactions will benefit from GLiMMIRS, which uses a generalized linear model that accounts for guide efficiency, differences in per-cell sequencing depth and several covariates. We note that it is important to consider a multiplicative model as the baseline expectation when looking for enhancer interactions, and when interactions are identified it is important to consider the possibility that the results are driven by a small number of outlier cells. To increase power to detect weak interactions, CRISPR experiments that are specifically designed to examine enhancer interactions are desirable. Our study motivates the further study of enhancer interactions in more cell types and conditions, to which GLiMMIRS can be applied to yield novel insights into regulatory element interactions and their effects on transcription.

2.5: Methods

2.5.1: CRISPRi perturbation of NMU enhancers

We identified two target sites of interest, A and B, for the gene *NMU*, each of which was targeted by two gRNAs in the Gasperini et al.[11] experiment (A1 and A2 targeting enhancer A; B1 and B2 targeting enhancer B). Pairs of gRNAs were designed by FlashFry[29] to target enhancers A and B at the same time, using 2 gRNAs per site. The gRNA pairs included the following: NMU_tss+NMU_tss (positive control), Safe_harbor (SH)+SH (negative control), A_sgRNA1+SH, A_sgRNA2+SH, SH+B_sgRNA1, SH+B_sgRNA2, A_sgRNA1+B_sgRNA1, A_sgRNA1+B_sgRNA2, A_sgRNA2+B_sgRNA1, A_sgRNA2+B_sgRNA2. Pairs of gRNAs were cloned into pLV-dCas9-KRAB-puro (Addgene #71236) following published methods[30,31]. Briefly, DNA oligos carrying pairs of guides were synthesized by IDT and cloned into pLV-dCas9-KRAB-puro plasmids by Gibson assembly reactions. Lentivirus was generated by co-transfecting the plasmid with PsPAX2 (Addgene #12260) and pMD2.G (Addgene #12259) in 293FT cells obtained from the Salk Institute Stem Cell Core. Lentivirus was harvested 48h post transfection. K562 cells (ATCC #CCL-243) were transduced by the lentiviruses using spinoculation. 72h after transduction, K562 cells with viral genome integration were selected by puromycin for 48 h. Total RNA from live K562 cells was extracted and reverse transcribed using SuperScript IV First-Strand Synthesis System (Thermo Fisher Scientific #18091050) with random hexamers. *NMU* expression was quantified by reverse transcription quantitative PCR (RT-qPCR). CRISPR gRNA designs and PCR primers used in experiment can be found in Supplemental File 2.1.

2.5.2: Data from Gasperini et al.

Data from the at-scale screen in the Gasperini et al. study are available at GEO accession number GSE120861. Guide spacer sequences were obtained from Supplementary Table 2 in the Gasperini et al. study[11]. The single-cell RNA-seq expression matrix from the at-scale screen was downloaded from the GEO file 'GSE120861_at_scale_screen.exprs.mtx'. The cell barcodes were determined from the GEO file 'GSE120861_at_scale_screen.cells.txt'. Gene names were determined from the GEO file 'GSE120861_at_scale_screen.genes.txt'. The expression matrix had 207,324 cell barcodes and 13,135 gene names. Covariate information as well as cell-guide mapping information was determined from the GEO file: 'GSE120861_at_scale_screen.phenoData.txt.gz'.

2.5.3: Computing guide efficiencies

We first collected the 13,189 guide RNA sequences used in the at-scale screen previously published by Gasperini et al.[11], which were published in Supplementary Table 2 of their study. We then appended 'NGG' to each 20 bp spacer sequence for compatibility with GuideScan 2.0[23]. We then used the GuideScan 2.0 gRNA sequence search tool (https://guidescan.com/grna) with the organism 'hg38' and the enzyme 'cas9' parameters to predict efficiencies for the 20bp guide RNA spacer sequences. We used the "Cutting.Efficiency" values outputted from GuideScan as our guide efficiency values.

Out of the 13,189 guide RNA sequences, 762 guide RNAs were designed to target transcription start sites, 101 guide RNAs were designed as non-targeting controls, 14 guide RNAs were designed as positive controls targeting the globin locus, and the remaining 12,312 guide RNAs were designed to target candidate enhancer sequences.

From the 12,312 enhancer-targeting guide RNAs, 1,415 guide RNAs did not find a match, had multiple off-targets, or had multiple perfect matches in the GuideScan 2.0 database. We excluded these 1,415 guide RNA sequences from downstream analysis.

2.5.4: Computing cell cycle scores

Cell cycle scores were computed from the single-cell RNA-sequencing gene expression matrix from the at-scale screen previously published by Gasperini et al.[11] using the Seurat R package.

Since the Seurat R package uses gene names from the Hugo Gene Nomenclature Committee, gene names were converted from their Ensembl Gene ID to HGNC symbol (https://www.genenames.org/) using the BioMart[32] tool from Ensembl[33] with the "hsapiens_gene_ensembl" dataset. Of the 13,135 genes in the at-scale expression matrix, 349 genes were not recognized by BioMart and 591 genes did not successfully map from Ensembl Gene ID to HGNC symbol. For the total 940 genes that could not be mapped from Ensembl Gene ID to HGNC symbol, the Ensembl Gene ID was imputed as the gene name for downstream analysis with Seurat.

To determine cell cycle scores, we used pre-defined sets of genes associated with S and G2M phases from the Seurat library. We log-normalized the data, identified variable features, and scaled the expression matrix using functions defined in Seurat. We then used the cell cycle scoring function with the predefined S and G2M gene sets in Seurat to compute cell cycle scores for each cell in the at-scale screen. To visualize the separation of cells based on their cell cycle scores, we performed a principal component analysis in Seurat using the S and G2M gene sets as features.

2.5.5: Model fitting and implementation

All models were fitted by maximum likelihood using the `glm.nb()` function from the MASS package in R[34]. Every model described in this work is a negative binomial generalized linear model with a log link function.

2.5.6: Defining a baseline model for a single enhancer acting on a single target gene

Our baseline model tests for the simple case where a single enhancer acts on a single gene. The model is a generalized linear model which assumes a log link function and that the single-cell RNA-seq tag counts of each gene are negative binomially-distributed. In other words, $y = NB(\mu, \phi)$ where y represents the scRNA-seq counts of the genes, ϕ represents the dispersion parameter of the negative binomial distribution, and μ is the mean parameter of the negative binomial distribution. The mean parameter is specified by a linear predictor passed through an exponential (inverse log-link) function: $\mu = exp(\beta_0 + \beta_{enhancer}X_{perturb} + \beta_S X_S + \beta_{G2M}X_{G2M} + \beta_{mito}X_{mito} + \beta_{gRNAs}X_{gRNAs} + \beta_{batch}X_{batch} + ln(s))$. In this expression, we have gene-specific coefficients and cell-specific predictor values. β_0 is the intercept and represents the baseline gene expression before the influence of any other relevant factors on gene expression. $\beta_{enhancer}$ represents the effect of a perturbed target site (putative enhancer) on its target gene. β_S and β_{G2M} are coefficients that represent the effect of the S and G2M cell cycle states, respectively. β_{mito} is a coefficient representing the effect of percentage of mitochondrial DNA. Finally, β_{gRNAs} is a coefficient representing the effect of total counts of gRNAs observed within a given cell. β_{batch} is a coefficient representing the effect of the prep batch, from the Gasperini et al. 2019 experiment. We incorporate measures of guide efficiency in the variable $X_{perturb}$. This variable is calculated for each cell based on the efficiencies of every gRNA targeting the target site being modeled which are present in the cell. Specifically, $X_{perturb}$ is calculated for

any given cell and target site as $1 - \prod_{k=1}^{K}(1 - g_k)$, where K is the total number of gRNAs targeting the target site found in the cell and g_k is the efficiency of the k^{th} gRNA. Because we interpret guide efficiency as the probability that a gRNA successfully perturbs its designated target site, the expression for $X_{perturb}$ can be interpreted as the joint probability of a perturbation in each cell based on all the gRNAs targeting the sites that are present in that cell. X_S and X_{G2M} are S and G2M cell cycle scores, respectively, for each cell. X_{mito} is the percentage of mitochondrial DNA in a cell. X_{gRNAs} is the total number of gRNAs observed in a cell. X_{batch} is the prep batch (from Gasperini et al. 2019). Finally, s is an offset term for the model that serves as a scaling factor controlling for variable sequencing depth across cells. It is calculated as $s = \frac{T}{1e6}$, where T is the total scRNA-seq counts in a cell summed across all genes in the expression count matrix. Prior to fitting the models, we added a pseudocount of 0.01 to the scRNA-seq counts of the gene being modeled for all cells to prevent inflation of coefficients (see 2.5.12: Defining a model for an enhancer pair acting on a single target gene).

2.5.7: Simulating data for single enhancers acting on single genes

To begin, we define some simulation parameters, including the total number of cells, C; the total number of genes, G; the total number of target sites, N; and the number of gRNAs targeting each site, d. Note that the total number of target sites, N, is also the total number of target genes, as this simulation assumes that each target site is a unique enhancer for a unique gene. To generate a simulated dataset, we need to simulate sets of coefficient values for each gene ($\beta_0, \beta_{enhancer}, \beta_S, \beta_{G2M}, \beta_{mito}$) as well as corresponding variable values for each cell ($X_{perturb}, X_S, X_{G2M}, X_{mito}$, and scaling factor s). We also need to simulate the gRNA library and assign gRNAs to cells, as well as assign guide efficiencies to gRNAs (which will be used to calculate $X_{perturb}$). These values are used to calculate a value of μ for defining a negative

binomial distribution from which simulated counts for a given gene will be drawn. Specifically, $\mu = exp(\beta_0 + \beta_{enhancer}X_{perturb} + \beta_S X_S + \beta_{G2M}X_{G2M} + \beta_{mito}X_{mito} + ln(s))$. The terms for total gRNA counts per cell and batch are omitted from the simulation for simplicity, and are also omitted when fitting the baseline model to the simulated data. The dispersion parameter for the negative binomial distribution will be constant across all genes and estimated from the empirical data. For the simulated dataset described in our paper, we used values of $G = 13000, N = 1000, d = 2$.

We first simulated values of $\widehat{\beta_0}$, or estimated baseline coefficients, for each gene. To do this, we randomly selected a subset of 1,000 genes and 10,000 cells from the Gasperini et al. 2019 at scale experiment and fit the counts for these genes to negative binomial distributions using maximum likelihood estimation (MLE). Specifically, we define the mean parameter of the negative binomial here as $\mu = exp(\widehat{\beta_0} + ln(s))$. Note that here s is calculated from the total counts for the gene observed across the subset of 10,000 cells using the formula defined in the previous section. This simplified model has no covariates, but does account for the scaling factor, as the goal is to simply get a sense of what coefficient values reflect the empirical data. After modeling the counts from the random subset of data, we visualized the distribution of estimated $\widehat{\beta_0}$(from which μ is calculated) and dispersion parameters for each gene tested. From what we observed, we picked a fixed dispersion value of $\phi = 1.5$ for defining the negative binomial distribution for generating simulated count data. We also observed that the distribution of $\widehat{\beta_0}$ estimated from the subset of the at scale experiment were roughly normally distributed. Therefore, we fit these estimated $\widehat{\beta_0}$ values to a normal distribution with MLE to obtain parameters for defining a normal distribution from which to sample β_0 values for the simulated dataset. We obtained parameters for the normal distribution of $\mu \approx 2.24$ and $\sigma \approx 1.8$, so we

sampled G times from $N(\mu = 2.24, \sigma = 1.8)$ to yield baseline coefficients for all the genes in the simulated dataset.

To assign guides to cells, we first determined the number of gRNAs in each cell in our simulated dataset by sampling from a Poisson distribution defined as $Pois(\lambda = 15)$. This value of λ comes from the fact that in the Gasperini et al. 2019 experiment, they observed a median of approximately 15 unique gRNAs per cell. Thus, we sampled C times from the distribution defined by $Pois(\lambda = 15)$ to obtain the number of unique gRNAs in each cell. To assign gRNAs to each cell, we sampled g times without replacement from the set of all gRNAs in our library, where g is the total number of gRNAs in each cell (determined in the previous step) and the gRNA library is denoted as a sequence of integers $1,2,\ldots,dN$. Information about which gRNAs are found in which cells are stored in a one hot encoded matrix.

We defined guide efficiency for each gRNA by sampling from a left-skewed Beta distribution, to represent the fact that an experimental design would select for gRNAs with higher efficiencies). For our simulation we used a Beta distribution defined as $Beta(a = 6, b = 3)$.

Next, we created a mapping of gRNAs to target genes. For each target site, or putative enhancer, we randomly select an integer from $1,2,\ldots,G$ to represent the target gene of the candidate enhancer (indexers are used as gene identifiers). This is done without replacement to simulate a case where we are attempting to study enhancers of distinct genes, and yields a vector of length N, which we will replicate d times to yield a complete mapping of gRNAs to target genes. In this vector of length Nd, the index of a given value in the vector represents the gRNA identifier.

Enhancer effect sizes are represented by the coefficient $\beta_{enhancer}$ and are assigned on a per-gene basis. These values represent the effect that an enhancer has on the expression of its target gene. To do this, we sampled from a gamma distribution and multiplied the values by -1 to yield a negative value, representative of the expectation that successful repression of an enhancer will most likely decrease target gene expression. We wanted the values to be on a comparable scale with the expected baseline expression, β_0, while also not being so small that they would be difficult for the model to detect changes in expression. We chose to sample values of $\beta_{enhancer}$ from a gamma distribution defined by $\Gamma(\alpha = 6, \sigma = 0.5)$, and all values drawn from the distribution were multiplied by -1 to represent a negative effect on target gene expression, which is the expectation when an enhancer is repressed.

$X_{perturb}$ is calculated for each cell as a function of guide efficiencies for the gRNAs targeting the putative enhancer of interest found in that cell. Specifically, it is calculated for each cell as $X_{perturb} = 1 - \prod_{k=1}^{K}(1 - g_k)$ where K is the total number of gRNAs targeting the putative enhancer of the gene being simulated/modeled that are present in the cell and g_k is the guide efficiency of the kth gRNA in this set of targeting gRNAs. $X_{perturb} = 0$ when $K = 0$ (Figure 2.2b). We compared the performance of using this variable in our model against the performance of using a binary indicator variable that simply represents the presence of any gRNA targeting the gene being simulated/modeled in each cell.

We generated cell cycle scores for each cell in our simulated dataset using a similar approach to the one we used for sampling β_0 values. That is, we first fit models to the empirical data to identify a distribution to draw simulated values from such that they would reflect the distribution of the real data. We first calculated S and G2M cell cycle scores for the empirical data using Seurat's CellCycleScoring() function[35–38]. We observed that while the S cycle scores

calculated from the empirical data appeared to be normally distributed, the G2M scores appeared to show a right skewed distribution. Thus, we fit the empirical S cycle scores to a normal distribution and the empirical G2M scores to a skew normal distribution with MLE. We used the estimated parameters to define distributions for sampling S and G2M scores for the simulated dataset. Specifically, we sampled C times from a normal distribution defined by $N(\mu = -1.296e-3, \sigma = 0.11)$ and a skew normal distribution defined by $N(\zeta = -0.256, \omega = 0.312, \alpha = 6.29, \tau = 0)$ to obtain simulated S and G2M scores, respectively.

We generated corresponding values of β_S and β_{G2M} by sampling from the same distribution used to generate the enhancer effect sizes, or the gamma distribution defined by $\Gamma(\alpha = 6, \sigma = 0.5)$.

Percentage of mitochondrial DNA per cell is simulated using the same approach used to simulate the cell cycle scores and baseline expression values (β_0). We fit to the empirical percentages of mitochondrial DNA per cell. We fit to a beta distribution using MLE, and used the resulting parameter estimates to define a new beta distribution from which we sampled simulated values of percentage of mitochondrial DNA. This beta distribution was defined as $Beta(a = 3.3, b = 81.48)$.

Coefficients for the effect size of percentage of mitochondrial DNA, β_{mito}, were simulated per gene by sampling from the same gamma distribution used to sample the other coefficients ($\beta_{enhancer}, \beta_S, \beta_{G2M}$). This is the gamma distribution defined as $\Gamma(\alpha = 6, \sigma = 0.5)$.

Finally, we simulated scaling factor values, s, for each cell in our simulated experiment, which were used to calculate values of μ for simulating counts for each gene. To do this, we simulated values of T, or total counts per cell, for each cell by sampling from a Poisson

distribution defined by $Pois(\lambda = 50000)$, where 50000 is the expected number of reads observed in each cell in a scRNA-seq experiment.

2.5.8: Simulating noisy guide efficiencies

The noisy guide efficiency estimate, w, for a given gRNA in our simulated dataset was sampled from a new Beta distribution parameterized by a' and b', which are calculated from the "true" simulated guide efficiency for the gRNA, w, and a dispersion-controlling constant D. We wanted the noisy guide efficiency to be sampled from a Beta distribution whose mean is equivalent to the "true" guide efficiency value; thus, $w = \frac{a'}{a'+b'}$. We defined the dispersion-controlling constant D as $D = a' + b'$. From this, it follows that $a' = Dw$ and $b' = D - a'$. Like so, we calculated values of a' and b' from which to draw the noisy guide efficiency estimate for a given gRNA in our simulated guide library. The magnitude of D is inversely proportional to the amount of noise (Supplemental Figure 2.1a-b).

2.5.9: Fitting baseline model to simulated data

To fit the baseline model to simulated data, we used a negative binomial GLM with a mean defined by the same log-link function described for generating simulated counts: $\mu = exp(\beta_0 + \beta_{enhancer}X_{perturb} + \beta_S X_S + \beta_{G2M}X_{G2M} + \beta_{mito}X_{mito} + ln(s))$. Models were fitted by MLE. Each model can be described as $y = NB(\mu, \phi)$, where y is the simulated counts for the gene being modeled, and all variable values ($X_{perturb}, X_S, X_{G2M}, X_{mito}$) come from the per-cell values from the simulated dataset. We omit β_{gRNA} when fitting to the simulated data for simplicity.

2.5.10: Evaluating performance of baseline model on simulated data

Our simulated dataset had N target sites, or genes that were regulated by an enhancer perturbed in the experiment. For each of these genes, we computed the Pearson correlation

(Pearson's r and p-value) between the estimated coefficients, derived from fitting the baseline model to the simulated data, and the "true" coefficients, which were the "ground truth" coefficient values that we generated for the simulation and used to parameterize the distribution from which the simulated counts were drawn. We also calculated MSE for these values. Finally, we calculated the correlation of determination (R^2) as a measure of the model performance, as $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$, where SS_{res} is the sum of squared residuals and SS_{tot} is the total sum of squares. Specifically, we calculated SS_{res} as the sum of squared differences between the true and estimated coefficient values, and SS_{tot} as the sum of squared differences between each estimated coefficient value and the average of all estimate values for the coefficient. These metrics are summarized in Table 2.1 for the continuous vs. indicator forms of $X_{perturb}$ and in Table 2.2 for the three different sets of noisy simulated guide efficiencies.

2.5.11: Fitting baseline model to experimental data

For running a single enhancer-gene pair analysis on the experimental data, we obtained the 664 previously published enhancer-gene pairs from the Gasperini et al.[11] paper using information provided in Supplemental Table 1 from the paper. Using these 664 previously published enhancer-gene pairs, we retrieved all experimental gRNAs targeting these enhancers, and filtered gRNAs where there was no valid guide efficiency from GuideScan 2.0. We then obtained the preparation batch, cell gRNA count, and percent mitochondrial reads covariates from their experimental data published on GEO and excluded cells without covariate values for our downstream modeling. To account for sequencing depth, we used the at-scale gene expression matrix and counted the number of transcripts per cell. We then divided these values by 1e-6 to obtain values for each cell which we included in our linear model through the offset() function. Prior to running the models, a pseudocount of 0.01 was added to the scRNA-seq counts

for each cell. Models were then fitted using the nb.glm() function in the MASS R package using a log-link function and optimizing via maximum likelihood estimation. In the at-scale model, there were 207,324 cells total. After filtering for cells without covariate values, there were 205,797 cells that were included in the modeling process. The scrambled perturbation negative control was obtained by scrambling the vector of guide efficiencies prior to modeling. The mismatch gene negative control set was obtained by randomly sampling a gene for a given enhancer from the set of 664 previously published enhancer-gene pairs.

2.5.12: Defining a model for an enhancer pair acting on a single target gene

Our model for evaluating interactions between enhancers is quite similar to our baseline model, except we replace $\beta_{enhancer}$ with three new coefficients: $\beta_A, \beta_B, \beta_{AB}$. Referring to the two enhancers in the pair being modeled as enhancers A and B: β_A represents the effect of enhancer A on the target gene; β_B represents the effect of enhancer B on the target gene; β_{AB} represents the interaction effect between enhancers A and B on the target gene. X_A, X_B, X_{AB} represent the perturbation probabilities of enhancer A, enhancer B, and both enhancers, respectively. The new negative binomial GLM has a mean defined as: $\mu = exp(\beta_0 + \beta_A X_A + \beta_B X_B + \beta_{AB} X_{AB} + \beta_S X_S + \beta_{G2M} X_{G2M} + \beta_{mito} X_{mito} + \beta_{gRNAs} X_{gRNAs} + \beta_{batch} X_{batch} + ln(s))$. They are calculated in the same manner as $X_{perturb}$ from the baseline model.

When fitting linear models, we observed inflated β_{AB} coefficients associated with cases where all cells containing gRNAs for both enhancers A and B showed no expression of the target gene. To prevent this inflation of the coefficients, we added a pseudocount of 0.01 to all the gene expression counts. When including a pseudocount in our modeling process, we observed a reduction in outliers in our enhancer effect sizes (Supplemental Figure 2.4).

2.5.13: Defining testable pairs of enhancers for interactions

We defined testable enhancer pairs as any pairs of target sites, or putative enhancers, from the Gasperini et al. 2019 experiment which were located within 1MB of a common target gene. We also defined two subsets of testable pairs based on certain filtering criteria: a smaller, high confidence set of 330 enhancer pairs and their corresponding target genes, and a larger unbiased set of 3,808 enhancer pairs and corresponding target genes. To define our high confidence set, we restricted the set of all testable pairs to those where both individual enhancers in the pair had previously established evidence of a regulatory effect on the target gene based on the analysis performed by Gasperini et al.[11] in their original study. To define our unbiased set, we simply looked for testable pairs that were simultaneously perturbed in a minimum of 20 cells; that is, there must be 20 cells receiving at least one of the gRNAs targeting each of the enhancers in the pair. We did not require either enhancer to have prior evidence of a regulatory effect on the target gene, thereby allowing for the possibility of regulatory effects that only arise in the presence of an interaction with another enhancer. In all cases, we also discarded enhancer pairs if all the gRNAs for either enhancer in the pair had undefined guide efficiency estimates.

2.5.14: Simulating data for enhancer pairs acting on a single target gene

We adapt the simulation framework used for simulating data for a single enhancer acting on a single gene. However, we have additional parameters to determine the number of "ground truth" enhancer pairs with and without an interaction effect between them. We refer to these as "positive" (N_{pos}) and "negative" (N_{neg}) pairs, respectively. These are selected from the set of all possible pairwise combinations of N target sites defined for our simulation. Note that for the case of an enhancer pair acting on a single gene, N represents the total number of putative enhancers rather than the total number of target genes. After randomly selecting N_{pos} and N_{neg} pairs

without replacement from the set of possible pairs, we then randomly select the same number of genes without selection from the set of possible genes $(1, \ldots, G)$ to be the target genes of those pairs. For the simulation described in this paper, we selected values of $N_{pos} = N_{neg} = 500$ and a total of $N = 1000$ target sites.

2.5.15: Simulating data for power analysis

Most aspects of the data simulation are identical to the data simulation for a single enhancer acting on a single gene. The coefficients β_A and β_B are drawn from the same distribution as $\beta_{enhancer}$. However, for the power analysis, we assign several different fixed values of β_{AB} for genes that are acted upon by an interaction effect between enhancers (e.g., the target genes of "positive" enhancer pairs). For genes that are not acted upon by any interaction effect, $\beta_{AB} = 0$. The other parameter that we modulate in the simulations is the value of λ for the Poisson distribution used to sample the number of unique gRNAs found in each cell. This is representative of multiplicity of infection, or MOI, so for each value of λ that we want to test with our power analysis, we generate different numbers of gRNAs per cell (Supplemental Figure 2.2) and use these sets of values to generate different mappings of gRNAs in cells. This yields a different one-hot encoded matrix for each value of lambda, which will also lead to different sets of values of X_A, X_B, and X_{AB}, as greater MOI may result in more gRNAs for a target site found in each cell and greater perturbation probabilities. Simulated counts are generated from a negative binomial distribution parameterized by $NB(\mu, \phi)$, where $\mu = exp(\beta_0 + \beta_A X_A + \beta_B X_B + \beta_{AB} X_{AB} + \beta_S X_S + \beta_{G2M} X_{G2M} + \beta_{mito} X_{mito} + ln(s))$ and $\phi = 1.5$ (determined from modeling empirical data, see Methods for simulating data for single enhancers acting on a single gene). We generated a set of simulated counts for each value of λ and interaction effect size. For our power analysis, we used values of $\lambda = 15,25,50,75,100$ and $\beta_{AB} = 0.5,1,3,5,7$.

2.5.16: Power analysis

For our power analysis, we fit our model to the simulated data for the "positive" and "negative" pairs to obtain true positive rates (TPR) and true negative rates (TNR), respectively. We calculated the proportion of models that correctly called significant interaction terms, β_{AB}, for the "positive" cases to obtain TPR. We calculated the proportion of models that correctly called no significant interaction terms, β_{AB}, for the "negative" cases to obtain TNR.

2.5.17: Comparing multiplicative to additive model

To compare the fits of multiplicative vs. additive models of enhancer pair activity, we defined each model under the null hypothesis (H_0), where there is no interaction term (for simplicity). For the multiplicative model under H_0, we use the canonical log-link function and define the mean of the negative binomial, μ, as:

$\mu = exp(\beta_0 + \beta_A X_A + \beta_B X_B + \beta_S X_S + \beta_{G2M} X_{G2M} + \beta_{mito} X_{mito} + \beta_{gRNAs} X_{gRNAs} + \beta_{batch} X_{batch} + ln(s))$. For the additive model under H_0, we use the identity link function where the mean is simply equivalent to the linear predictor without transformation, defined as:

$\mu = s(\beta_0 + \beta_A X_A + \beta_B X_B + \beta_{AB} X_{AB} + \beta_S X_S + \beta_{G2M} X_{G2M} + \beta_{mito} X_{mito} + \beta_{gRNAs} X_{gRNAs} + \beta_{batch} X_{batch})$. We applied each model to the 330 testable pairs from the experimental data where each enhancer in the pair had evidence of being an enhancer for the target gene based on the analysis by Gasperini et al. We compare model fits by examining the Akaike Information Criterion (AIC), with a lower AIC indicating a better fit. We calculated ΔAIC by subtracting the AIC of the lesser model from the AIC of the best fitting model. Since we found that the multiplicative model fit better in every case we tested, every ΔAIC reported in our study reflects the AIC of the additive model subtracted from the AIC of the multiplicative model.

2.5.18: Fitting interaction model to empirical data

For analyzing both sets of enhancer pairs tested in our analysis, we followed an identical procedure to the baseline model scenario, with the exception of adding a second enhancer effect vector, and allowing for interactions between the two enhancer vectors using built-in functionality within the glm.nb() function in the MASS R package.

2.5.19: Bootstrapping of significant interaction coefficients

We first performed bootstrapping to generate empirical distributions for the four significant interaction terms identified in our genome-wide analysis of enhancer pairs. We resampled all the cells in our dataset with replacement and refit our enhancer pair linear models with their associated covariates to obtain the bootstrapped empirical interaction coefficients. We then used the bootstrapped interaction coefficient estimates to derive 99% confidence intervals for the interaction coefficient using quantiles.

2.5.20: Permutation test for significant interaction coefficients

To determine permutation-based p-values associated with the observed significant interaction coefficients, we generated a null distribution of interaction coefficients by shuffling the perturbation probability vectors for enhancer 1 and enhancer 2 jointly, such that the same numbers of cells would have both enhancers perturbed. After performing 1000 permutations, we computed two-tailed p-values by counting the number of interaction coefficients with a magnitude greater than our observed significant interaction coefficient and dividing by the total number of permutations performed.

2.5.21: Schematic figures

All schematic figures created with BioRender.com.

2.6: Acknowledgements

Chapter 2, in part, has been submitted for publication of the material. The dissertation author was the primary researcher and author of this paper.

2.7: References

1. Hong JW, Hendrix DA, Levine MS. Shadow Enhancers as a Source of Evolutionary Novelty. Science. 2008 Sep 5;321(5894):1314. PMCID: PMC4257485

2. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jørgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Müller F, Forrest ARR, Carninci P, Rehli M, Sandelin A. An atlas of active enhancers across human cell types and tissues. Nature. 2014 Mar 27;507(7493):455–461. PMCID: PMC5215096

3. Visel A, Akiyama JA, Shoukry M, Afzal V, Rubin EM, Pennacchio LA. Functional Autonomy of Distant-Acting Human Enhancers. Genomics. 2009 Jun;93(6):509–513. PMCID: PMC2683195

4. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. Super-Enhancers in the Control of Cell Identity and Disease. Cell. 2013 Nov 7;155(4):934–947.

5. Dukler N, Gulko B, Huang YF, Siepel A. Is a super-enhancer greater than the sum of its parts? Nature Genetics. 2017 Jan;49(1):2–3.

6. Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y, Plajzer-Frick I, Pickle CS, Kato M, Garvin TH, Pham QT, Harrington AN, Akiyama JA, Afzal V, Lopez-Rios J, Dickel DE, Visel A, Pennacchio LA. Enhancer redundancy provides phenotypic robustness in mammalian development. Nature. 2018 Feb;554(7691):239–243.

7. Hay D, Hughes JR, Babbs C, Davies JOJ, Graham BJ, Hanssen L, Kassouf MT, Marieke Oudelaar AM, Sharpe JA, Suciu MC, Telenius J, Williams R, Rode C, Li PS, Pennacchio LA, Sloane-Stanley JA, Ayyub H, Butler S, Sauka-Spengler T, Gibbons RJ, Smith AJH, Wood WG, Higgs DR. Genetic dissection of the α-globin super-enhancer in vivo. Nat Genet. 2016 Aug;48(8):895–903. PMCID: PMC5058437

8. Shin HY, Willi M, HyunYoo K, Zeng X, Wang C, Metser G, Hennighausen L. Hierarchy within the mammary STAT5-driven Wap super-enhancer. Nat Genet. 2016 Aug;48(8):904–911. PMCID: PMC4963296

9. Lin X, Liu Y, Liu S, Zhu X, Wu L, Zhu Y, Zhao D, Xu X, Chemparathy A, Wang H, Cao Y, Nakamura M, Noordermeer JN, La Russa M, Wong WH, Zhao K, Qi LS. Nested epistasis enhancer networks for robust genome regulation. Science. 2022 Sep 2;377(6610):1077–1085. PMID: 35951677

10. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, Adamson B, Norman TM, Lander ES, Weissman JS, Friedman N, Regev A. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 2016 Dec 15;167(7):1853-1866.e17. PMID: 27984732

11. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A, Schreiber J, Noble WS, Trapnell C, Ahituv N, Shendure J. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell [Internet]. 2019 Jan 3 [cited 2019 Jan 8];0(0). Available from: https://www.cell.com/cell/abstract/S0092-8674(18)31554-X PMID: 30612741

12. Allen F, Behan F, Khodak A, Iorio F, Yusa K, Garnett M, Parts L. JACKS: joint analysis of CRISPR/Cas9 knockout screens. Genome Res. 2019 Mar 1;29(3):464–471. PMID: 30674557

13. Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, Schuster LC, Kuchler A, Alpar D, Bock C. Pooled CRISPR screening with single-cell transcriptome readout. Nature Methods. 2017 Mar;14(3):297–301.

14. Replogle JM, Norman TM, Xu A, Hussmann JA, Chen J, Cogan JZ, Meer EJ, Terry JM, Riordan DP, Srinivas N, Fiddes IT, Arthur JG, Alvarado LJ, Pfeiffer KA, Mikkelsen TS, Weissman JS, Adamson B. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nature Biotechnology. Nature Publishing Group; 2020 Mar 30;1–8.

15. Hill AJ, McFaline-Figueroa JL, Starita LM, Gasperini MJ, Matreyek KA, Packer J, Jackson D, Shendure J, Trapnell C. On the design of CRISPR-based single-cell molecular screens. Nature Methods. Nature Publishing Group; 2018 Apr;15(4):271–274.

16. Xie S, Duan J, Li B, Zhou P, Hon GC. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. Molecular Cell. 2017 Apr 20;66(2):285-299.e5.

17. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. Nature Publishing Group; 2014 Apr;32(4):381–386.

18. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017 Oct;14(10):979–982. PMCID: PMC5764547

19. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods. 2017 Mar;14(3):309–315. PMCID: PMC5330805

20. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012 Feb 3;148(3):458–472. PMID: 22265598

21. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. Nature. 2012 Apr 11;485(7398):376–380. PMCID: PMC3356448

22. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, Gribnau J, Barillot E, Blüthgen N, Dekker J, Heard E. Spatial partitioning of the regulatory landscape of the X-inactivation center. Nature. 2012 Apr 11;485(7398):381–385. PMCID: PMC3555144

23. Perez AR, Pritykin Y, Vidigal JA, Chhangawala S, Zamparo L, Leslie CS, Ventura A. GuideScan software for improved single and paired CRISPR guide RNA design. Nat Biotechnol. Nature Publishing Group; 2017 Apr;35(4):347–349.

24. Kowalczyk MS, Tirosh I, Heckl D, Rao TN, Dixit A, Haas BJ, Schneider RK, Wagers AJ, Ebert BL, Regev A. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res. 2015 Dec;25(12):1860–1872. PMCID: PMC4665007

25. Kim HK, Kim Y, Lee S, Min S, Bae JY, Choi JW, Park J, Jung D, Yoon S, Kim HH. SpCas9 activity prediction by DeepSpCas9, a deep learning–based model with high generalization performance. Science Advances. American Association for the Advancement of Science; 2019 Nov 6;5(11):eaax9249.

26. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J, Root DE. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. Nature Publishing Group; 2016 Feb;34(2):184–191.

27. Xiang X, Corsi GI, Anthon C, Qu K, Pan X, Liang X, Han P, Dong Z, Liu L, Zhong J, Ma T, Wang J, Zhang X, Jiang H, Xu F, Liu X, Xu X, Wang J, Yang H, Bolund L, Church GM, Lin L, Gorodkin J, Luo Y. Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. Nat Commun. Nature Publishing Group; 2021 May 28;12(1):3238.

28. Konstantakos V, Nentidis A, Krithara A, Paliouras G. CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. Nucleic Acids Research. 2022 Apr 22;50(7):3616–3637.

29. McKenna A, Shendure J. FlashFry: a fast and flexible tool for large-scale CRISPR target design. BMC Biology. 2018 Jul 5;16(1):74.

30. Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, Lin KC, Huang H, Liu T, Marina RJ, Jung I, Shen Y, Guan KL, Ren B. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nature Methods. 2017 Jun;14(6):629–635.

31. Chen HV, Lorenzini MH, Lavalle SN, Sajeev K, Fonseca A, Fiaux PC, Sen A, Luthra I, Ho AJ, Chen AR, Guruvayurappan K, O'Connor C, McVicker G. Deletion mapping of regulatory elements for GATA3 in T cells reveals a distal enhancer involved in allergic diseases. Am J Hum Genet. 2023 Apr 6;110(4):703–714. PMCID: PMC10119147

32. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. BioMart – biological queries made easy. BMC Genomics. 2009 Jan 14;10(1):22.

33. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, Berry A, Bhai J, Bignell A, Billis K, Boddu S, Brooks L, Charkhchi M, Cummins C, Da Rin Fioretto L, Davidson C, Dodiya K, Donaldson S, El Houdaigui B, El Naboulsi T, Fatima R, Giron CG, Genez T, Martinez JG, Guijarro-Clarke C, Gymer A, Hardy M, Hollis Z, Hourlier T, Hunt T, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Marugán JC, Mohanan S, Mushtaq A, Naven M, Ogeh DN, Parker A, Parton A, Perry M, Piližota I, Prosovetskaia I, Sakthivel MP, Salam AIA, Schmitt BM, Schuilenburg H, Sheppard D, Pérez-Silva JG, Stark W, Steed E, Sutinen K, Sukumaran R, Sumathipala D, Suner MM, Szpak M, Thormann A, Tricomi FF, Urbina-Gómez D, Veidenberg A, Walsh TA, Walts B, Willhoft N, Winterbottom A, Wass E, Chakiachvili M, Flint B, Frankish A, Giorgetti S, Haggerty L, Hunt SE, IIsley GR, Loveland JE, Martin FJ, Moore B, Mudge JM, Muffato M, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Dyer S, Harrison PW, Howe KL, Yates AD, Zerbino DR, Flicek P. Ensembl 2022. Nucleic Acids Research. 2022 Jan 7;50(D1):D988–D995.

34. Venables WN, Ripley BD. Modern Applied Statistics with S [Internet]. Fourth. New York: Springer; 2002. Available from: https://www.stats.ox.ac.uk/pub/MASS4/

35. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. Cell. Elsevier; 2019 Jun 13;177(7):1888-1902.e21. PMID: 31178118

36. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. Nature Publishing Group; 2018 May;36(5):411–420.

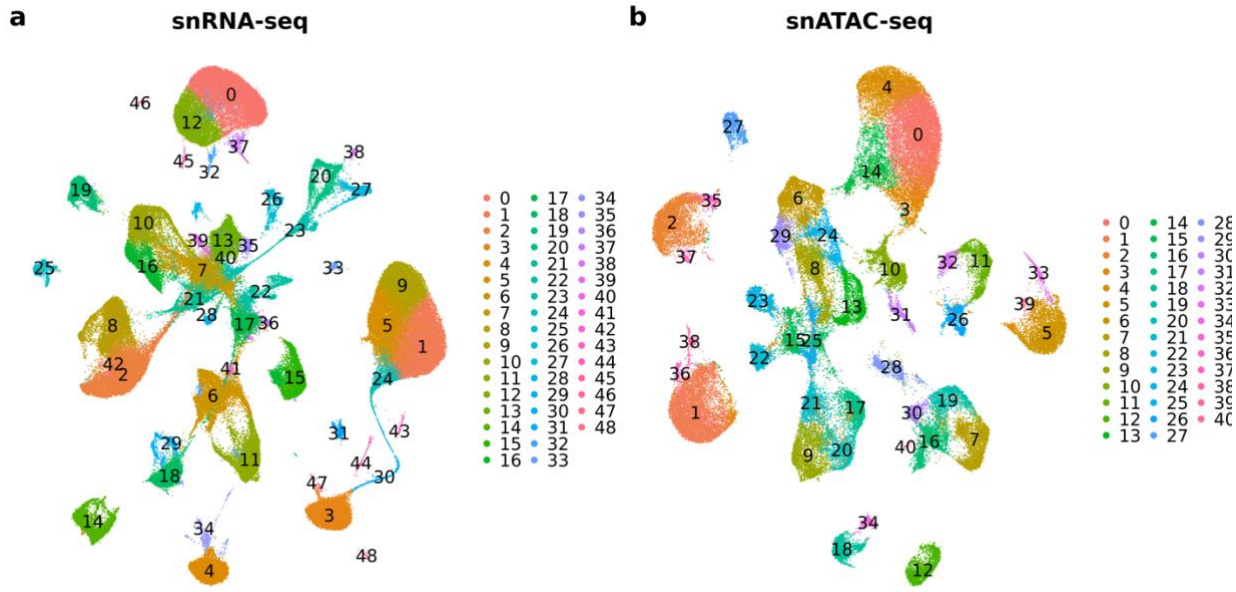
37. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. Nature Publishing Group; 2015 May;33(5):495–502.

38. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. Integrated analysis of multimodal single-cell data. Cell. Elsevier; 2021 Jun 24;184(13):3573-3587.e29. PMID: 34062119

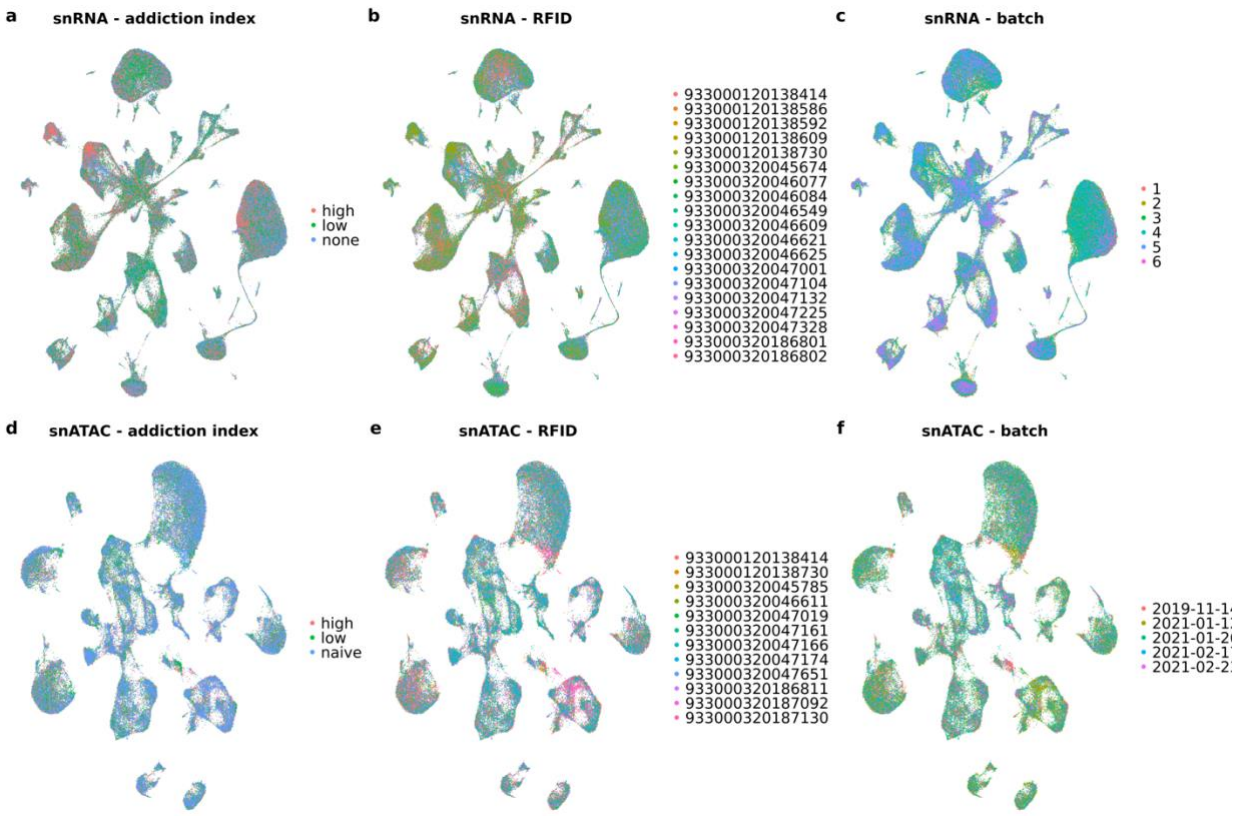
APPENDIX

Supplemental Figures

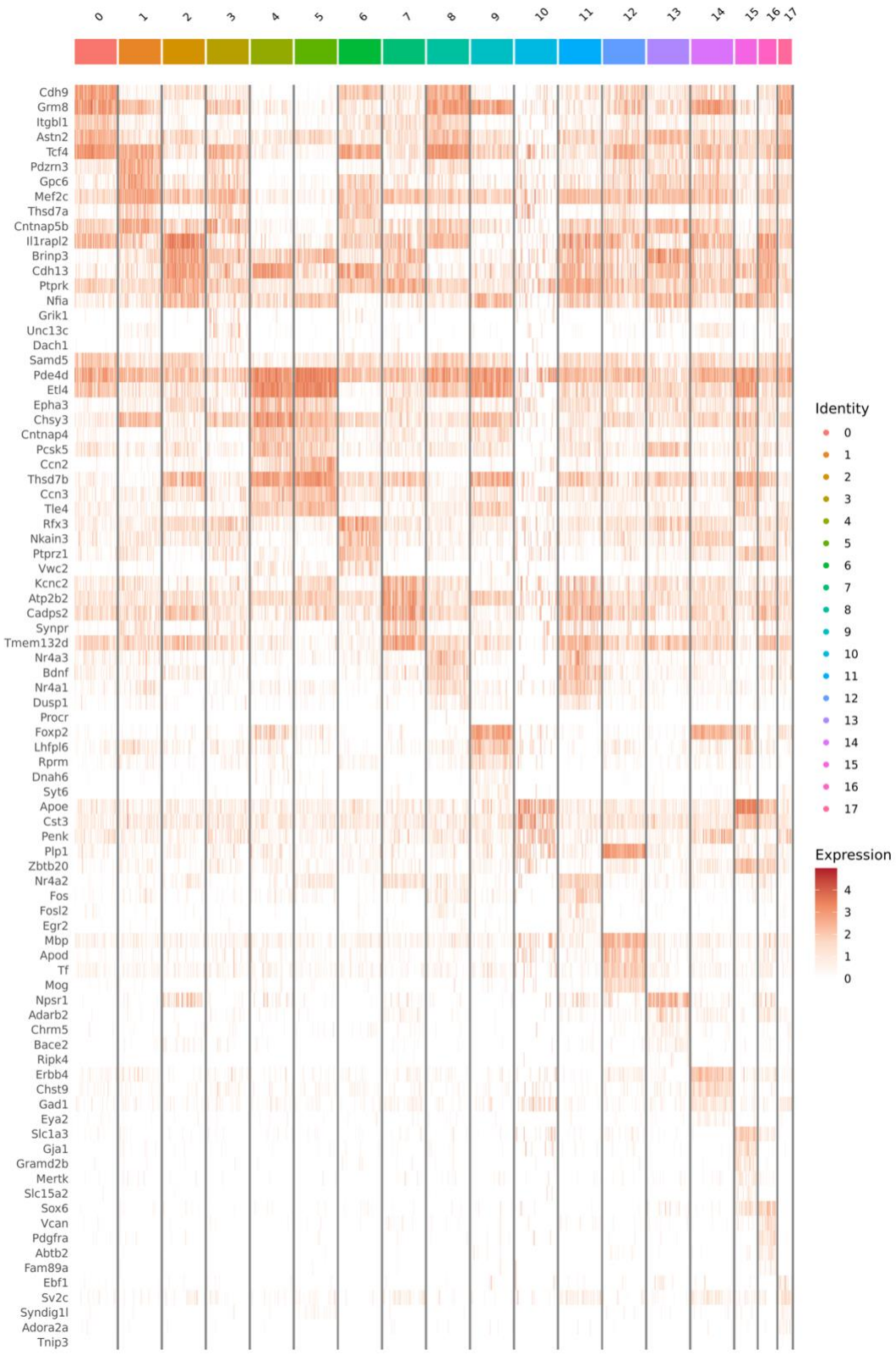
Chapter 1



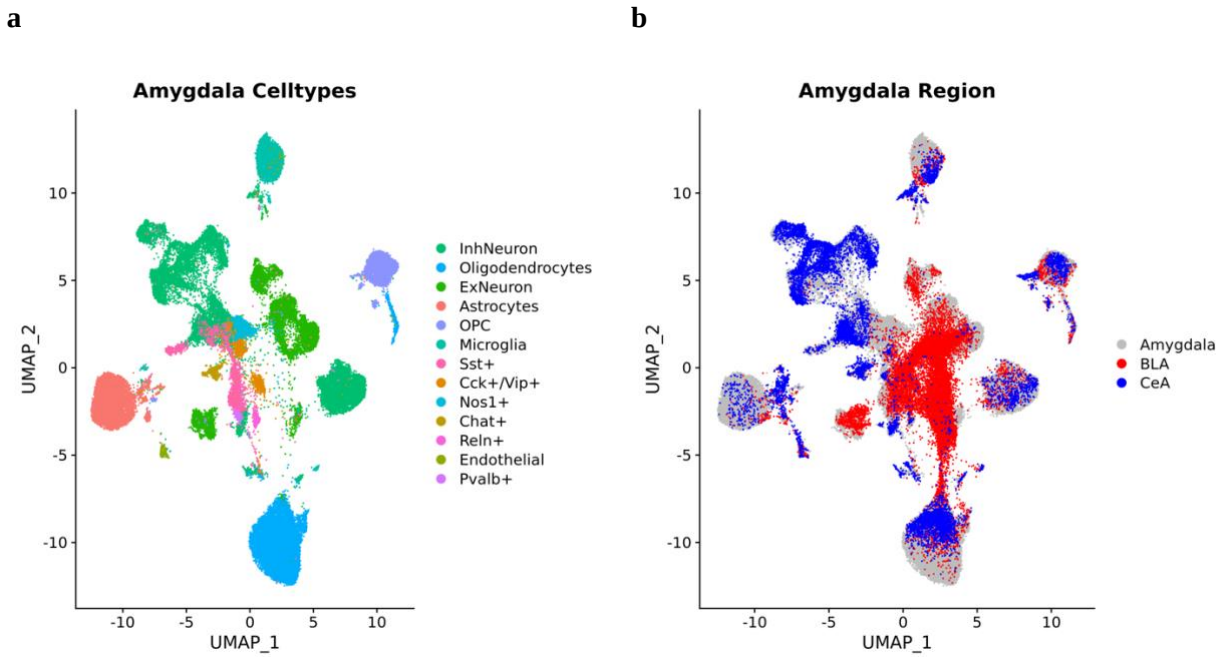
Supplemental Figure 1.1: UMAP visualization of the clusters identified in integrated single-cell data sets. (**a**) Clustering of integrated snRNA-seq dataset revealed 49 clusters. We first performed a k-nearest neighbors analysis (KNN) using the first 30 dimensions calculated by reciprocal principal component analysis (PCA). This was implemented with the FindNeighbors() function in Seurat. Next we used a modularity optimization technique using the Louvain algorithm to cluster the data, implemented with the FindClusters() function in Seurat with a resolution parameter of 0.8. (**b**) Clustering of integrated snATAC-seq data revealed 41 clusters. Latent semantic indexing (LSI) was used for dimensionality reduction rather than PCA. The first 30 dimensions minus the first dimension were used for KNN and clustering and the algorithm used for clustering was the smart local moving (SLM) algorithm. These steps were implemented with the same Seurat functions.



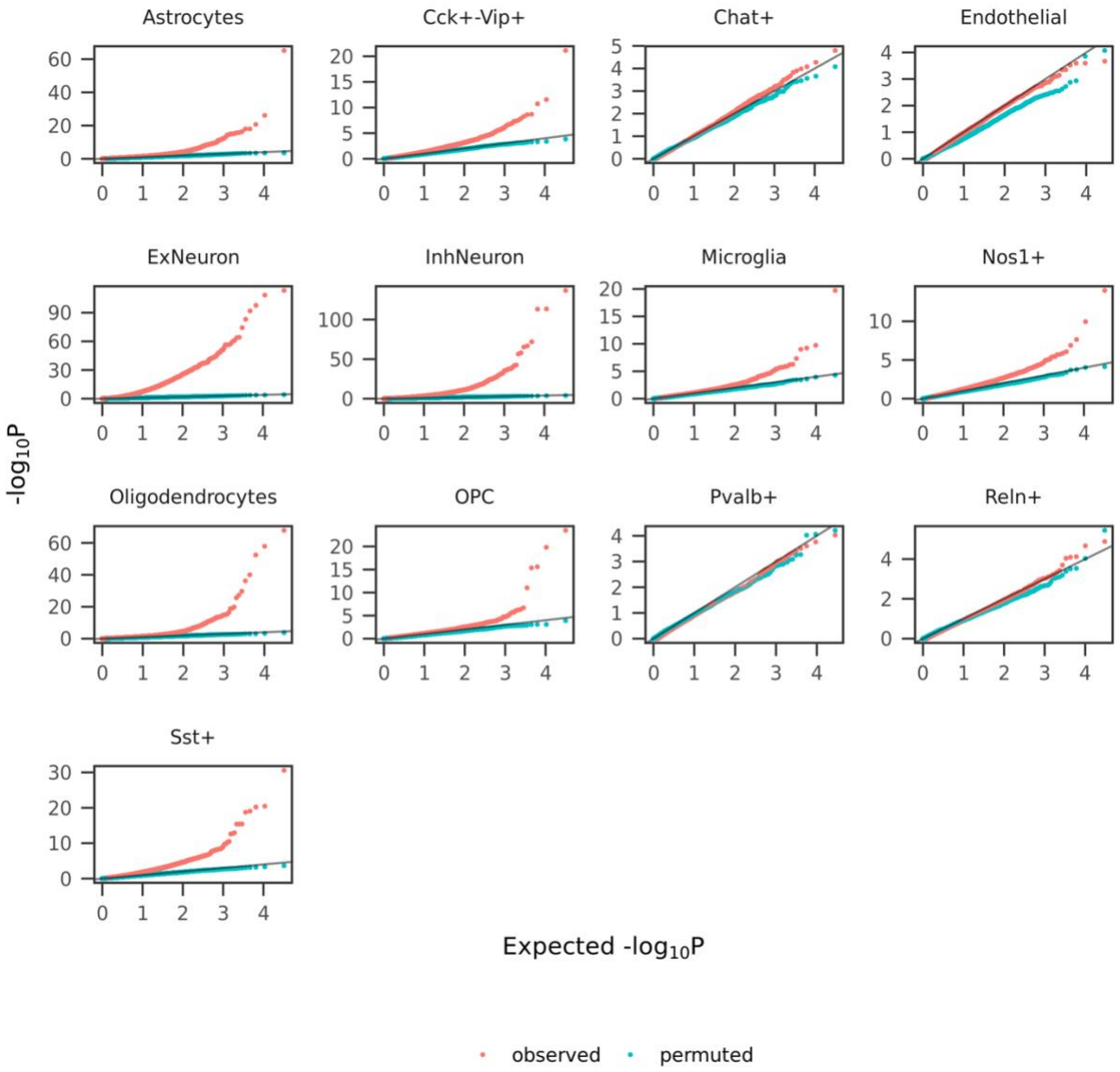
Supplemental Figure 1.2: UMAPS of snRNA-seq and snATAC-seq profiles, respectively, following batch correction of integrated datasets, grouped on: addiction index (**a, d**), rat sample (**b, e**), and batch information (**c, f**). These plots demonstrate that cells do not cluster by any of these covariates following batch correction. Integration and batch correction of the snRNA-seq dataset was performed using SCTransform while Harmony was used for the snATAC-seq dataset.



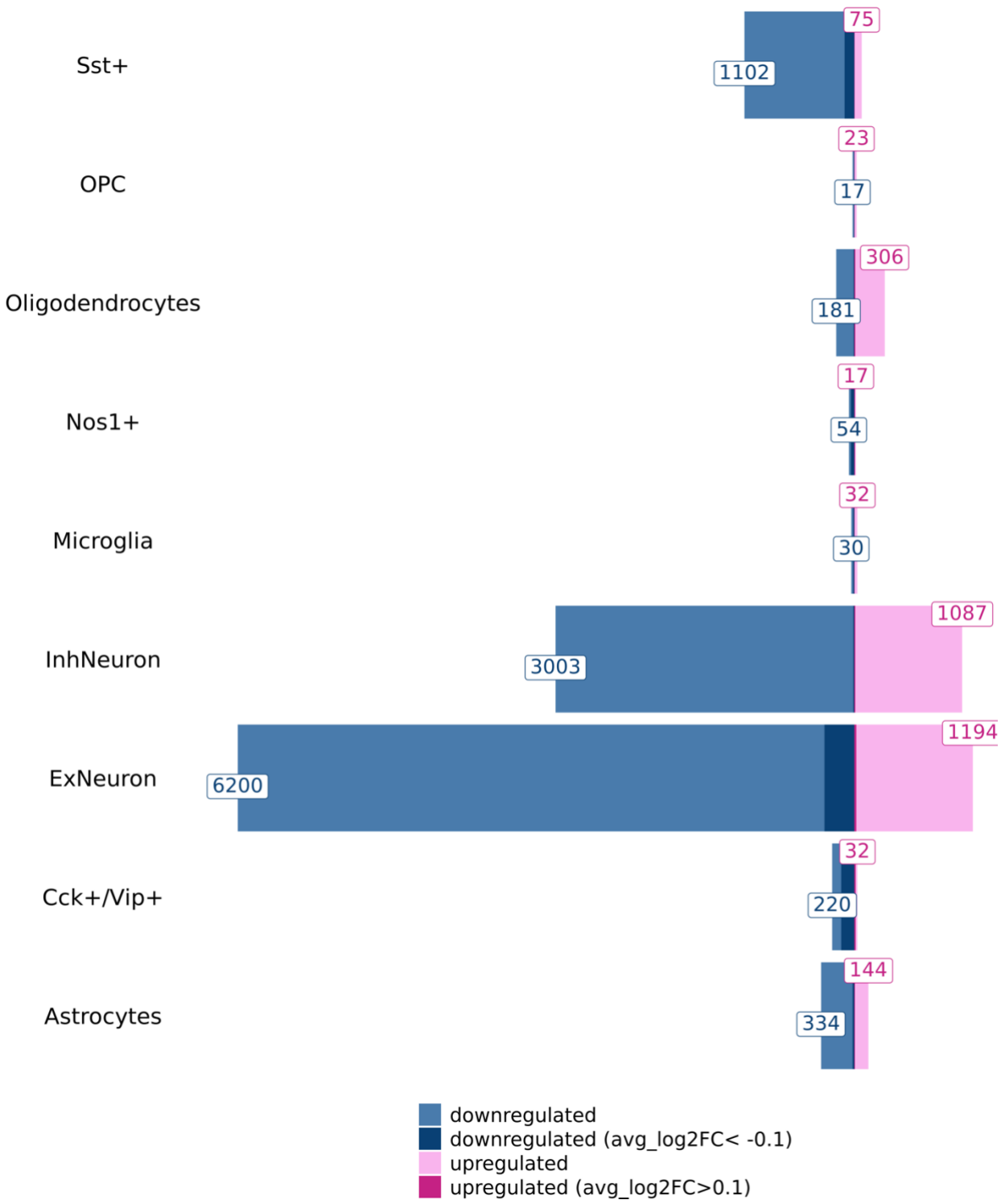
Supplemental Figure 1.3: Heatmap of top five marker gene expression within subclustered excitatory neurons.



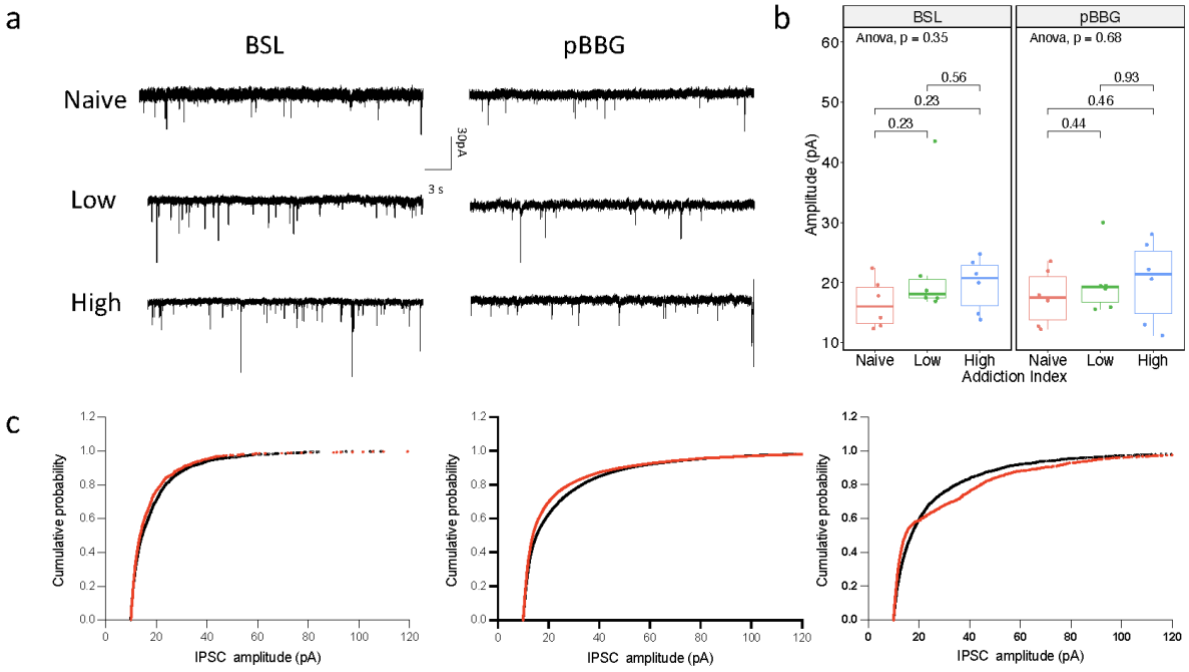
Supplemental Figure 1.4: Co-clustering of snRNA-seq from a CEA sample, a BLA sample, and the whole amygdala samples from all the naive rats in our study. **a)** UMAP with cells colored by cell type cluster assignments. **b)** UMAP with cells colored by source tissue, where "Amygdala" refers to the set of all amygdala samples from the naive rats in our study.



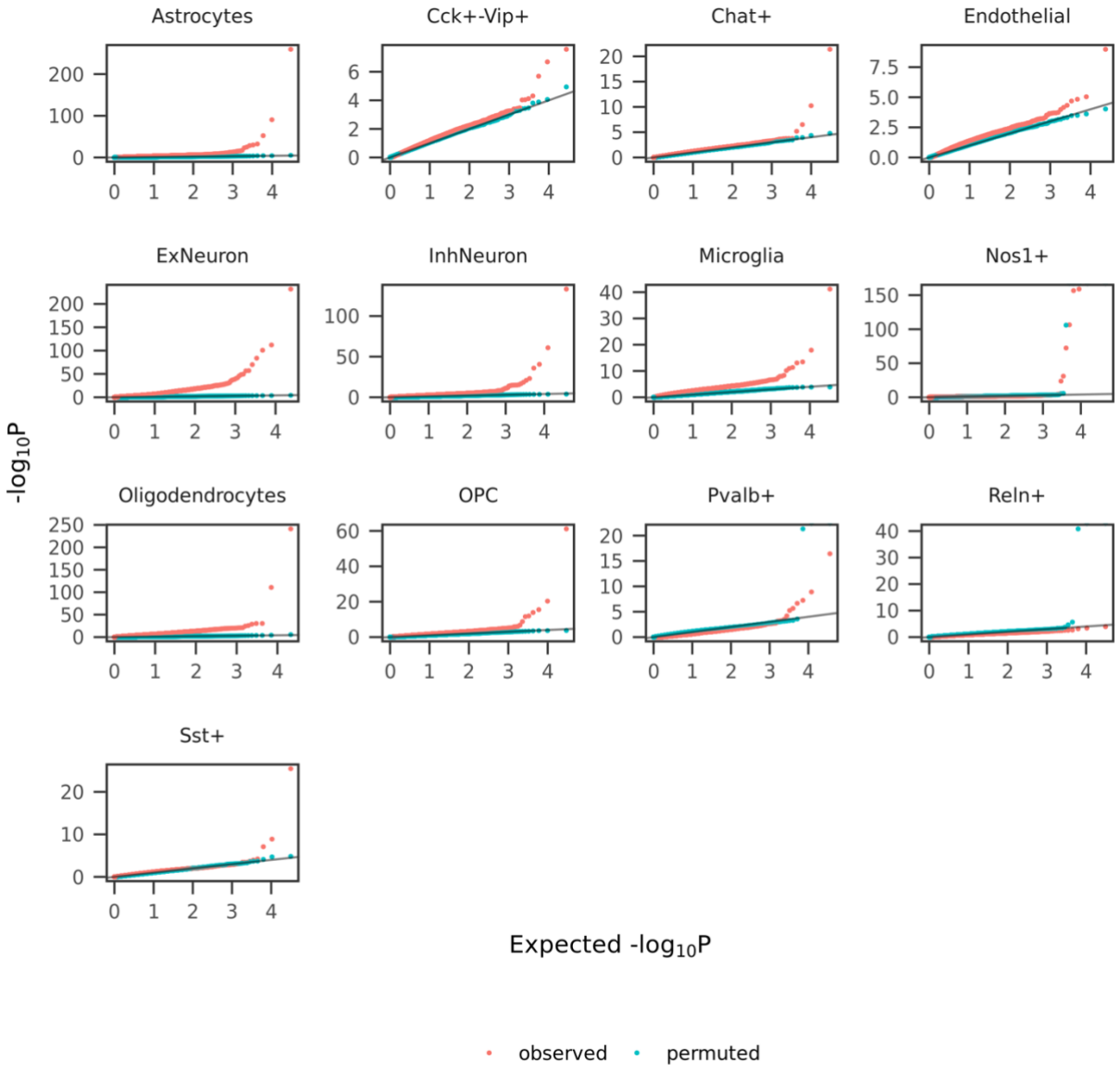
Supplemental Figure 1.5: QQ plots showing distribution of p-values for our differential gene expression analysis performed on our observed versus permuted data (AI labels associated with each cell were shuffled). The negative binomial test was the statistical test used for the analysis of both the observed and permuted datasets.



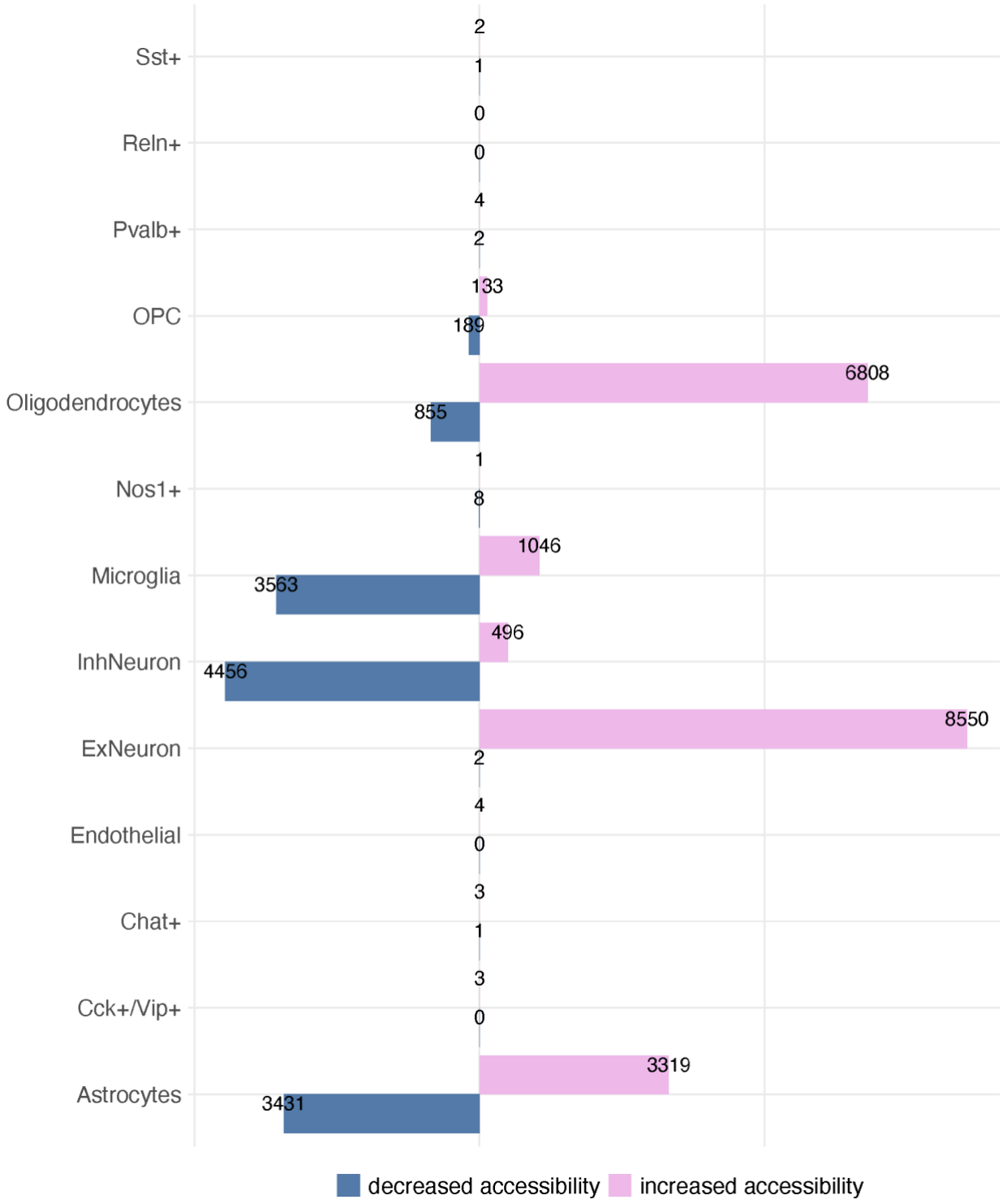
Supplemental Figure 1.6: Barplot showing numbers (labeled) of significant (FDR<10%) up- and downregulated DEGs by cell type. Darker shades indicate DEGs with a large fold change (abs(avg_log_2FC)0.1).



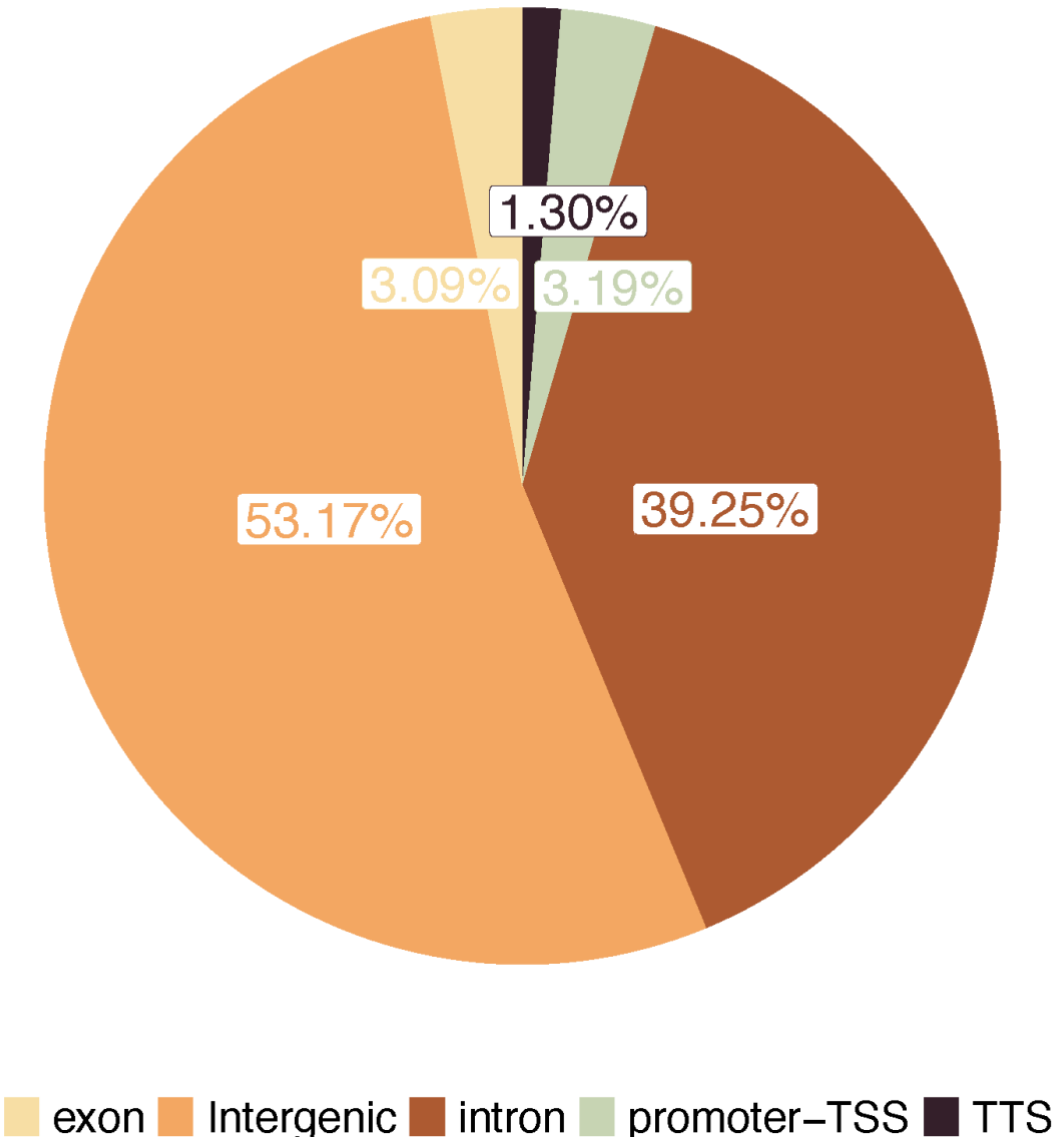
Supplemental Figure 1.7: Summary of electrophysiology experiments studying GABA transmission in the central amygdala. **a)** Representative traces of sIPSC frequencies for baseline (BSL) and following treatment with pBBG (pBBG) in naive, low AI and high AI rats. **b)** ANOVA test comparing mean amplitude in BSL vs. pBBG across naive, low AI and high AI rats (degrees of freedom = 2). **c)** Cumulative probability plots of the peak amplitude for naive, low AI and high AI rats.



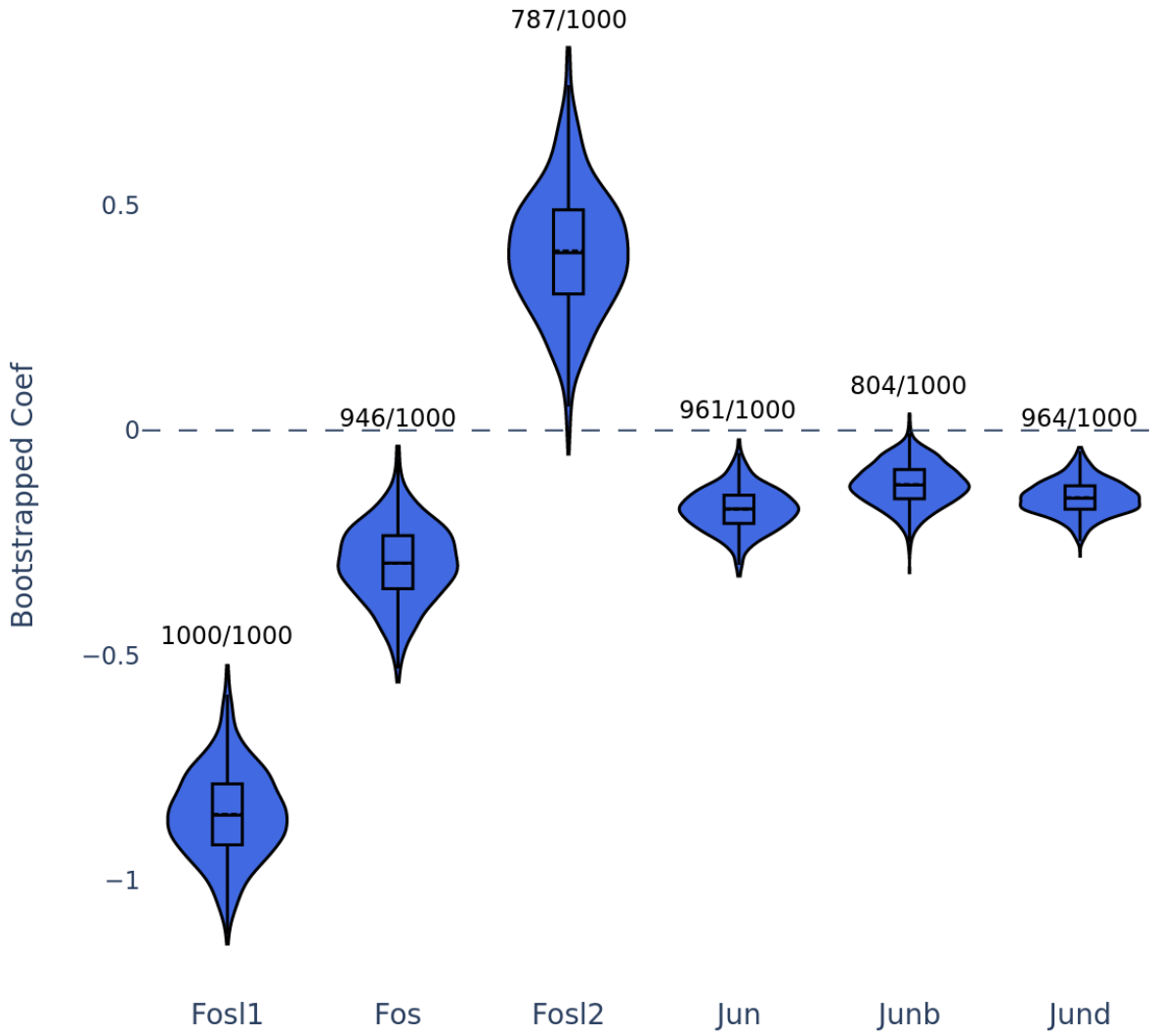
Supplemental Figure 1.8: QQ plots showing distribution of p-values for our differential peak accessibility analysis performed on our observed versus permuted data (AI labels associated with each cell were shuffled). The negative binomial test was used for the analysis of both the observed and permuted datasets.



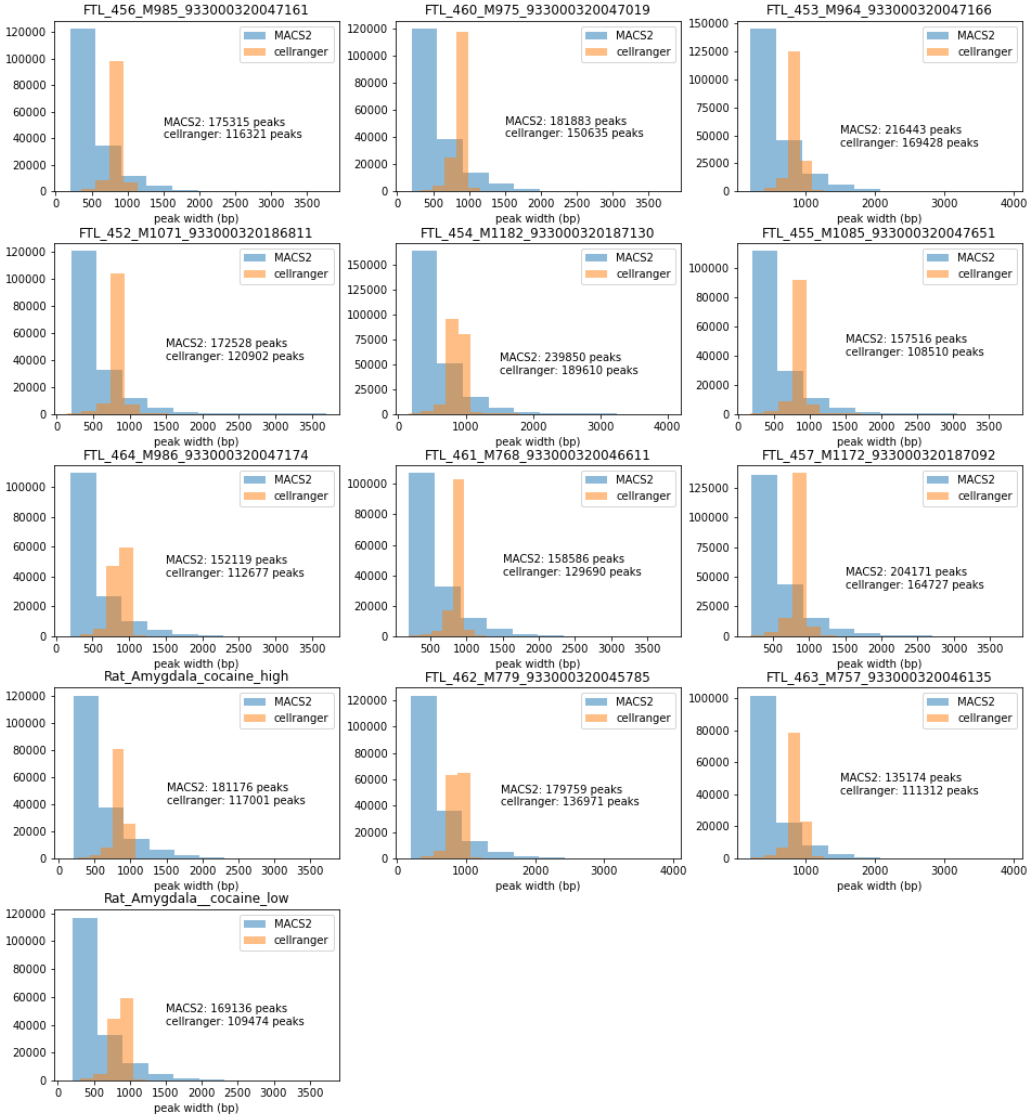
Supplemental Figure 1.9: Bar plot showing number of significant (FDR<10%) differentially accessible peaks between high vs. low rats in each cell type.



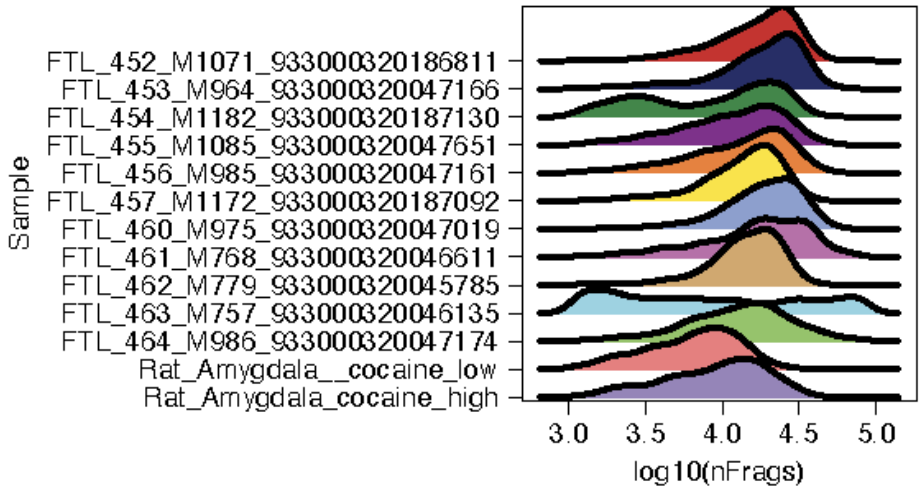
Supplemental Figure 1.10: Pie chart showing genomic annotations of all OCRs in our snATAC-seq dataset across all rats.



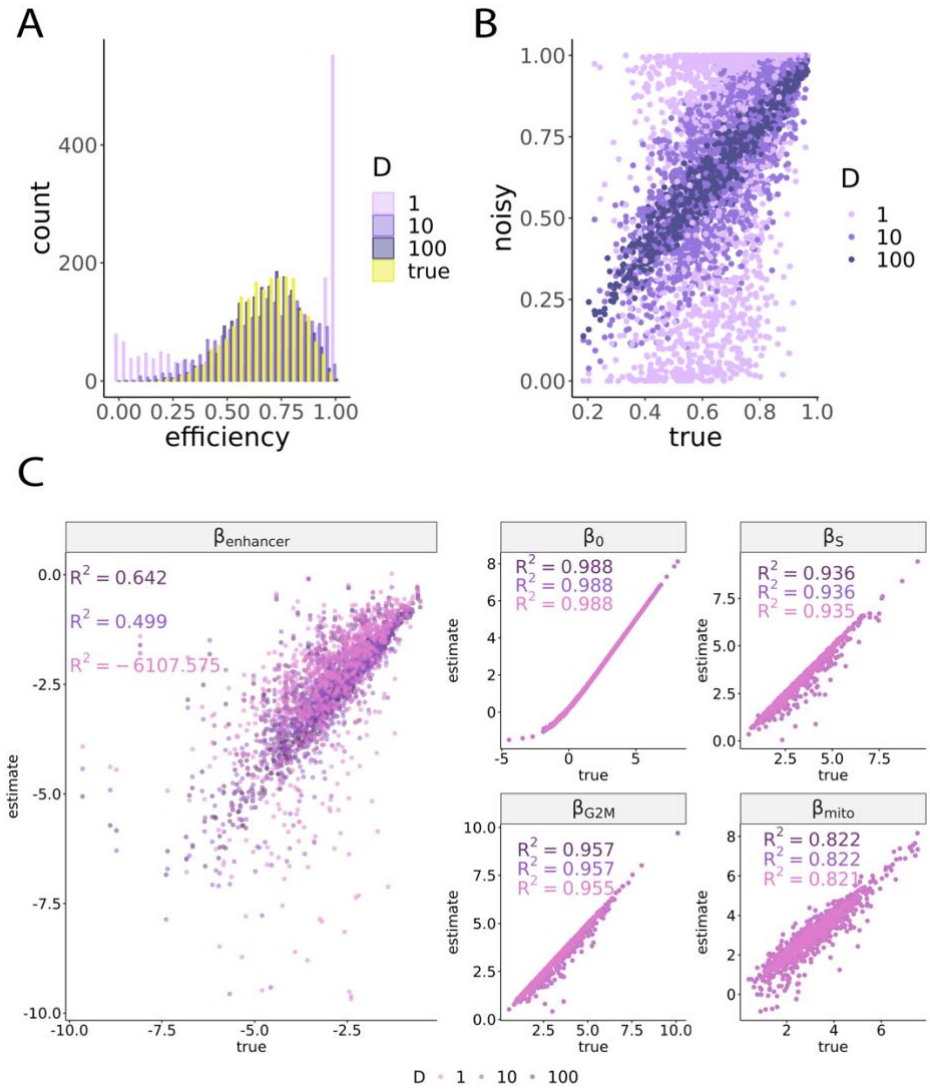
Supplemental Figure 1.11: Violin plots showing DEG analysis in InhNeurons over 1000 bootstrap iterations. Each violin shows the distribution of log2FC results per iteration. The fraction represents the number of significant iterations (FDR<10%).



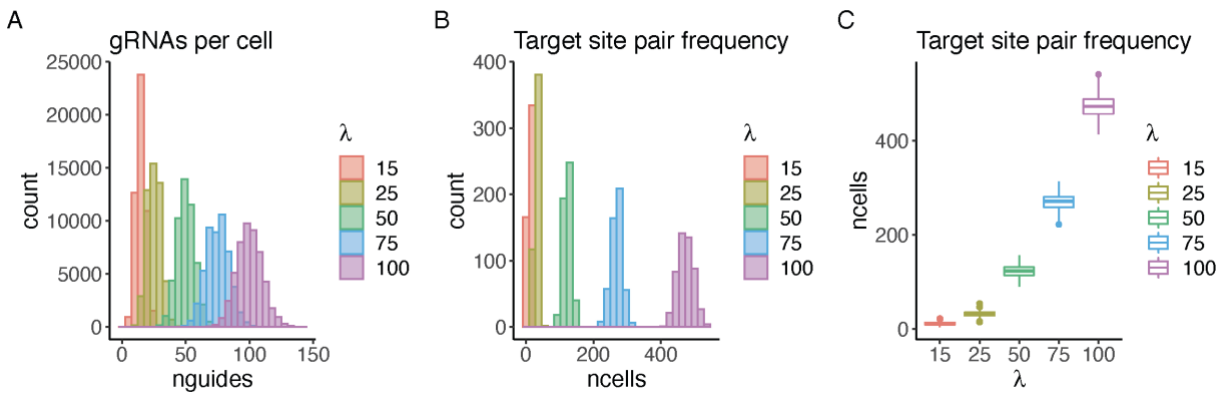
Supplemental Figure 1.12: Histograms showing distribution of peak sizes for peaks called by MACS2 (on the BAM files for the snATAC-seq data) versus CellRanger's internal peak calling algorithm. MACS2 calls smaller, more precise peaks.



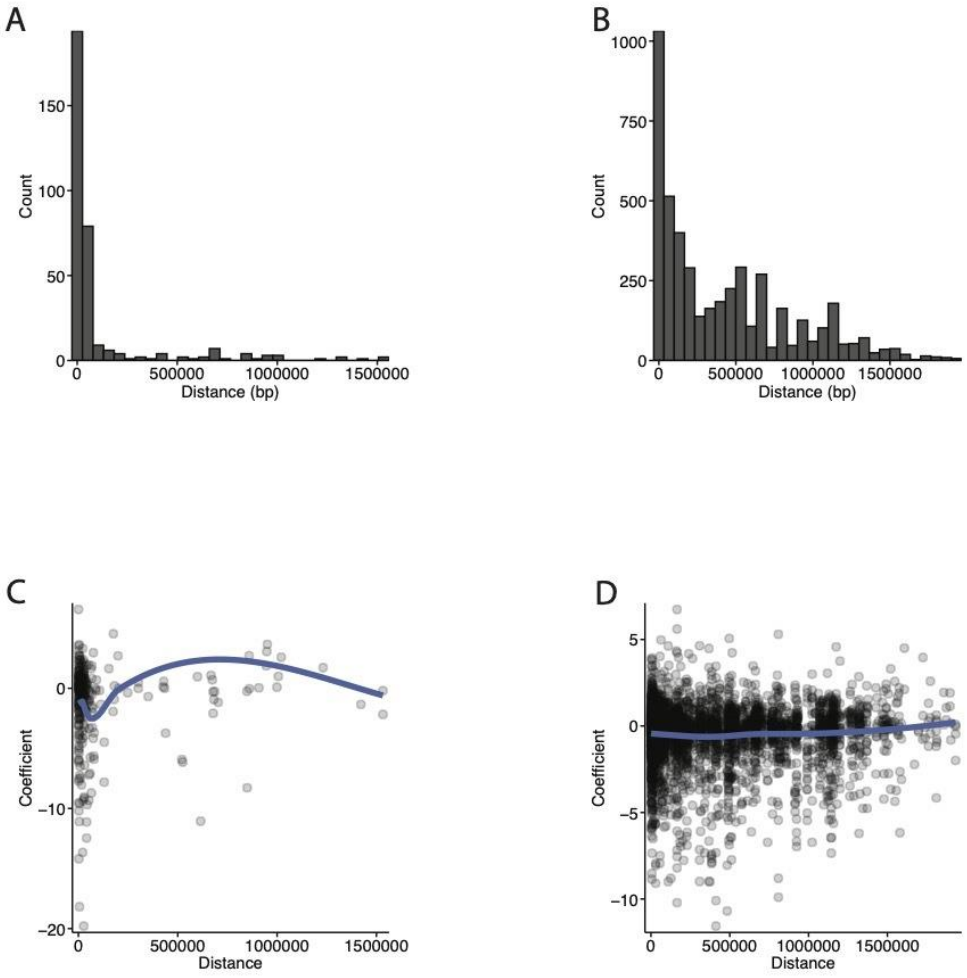
Supplemental Figure 1.13: Ridge plot quantifying the number of unique fragments (log10(nFrags)) per sample in the ATAC. Sample FTL_463_M757_933000320046135 was removed at this step and not included in any of our downstream snATAC-seq analyses due to its low number of fragments.



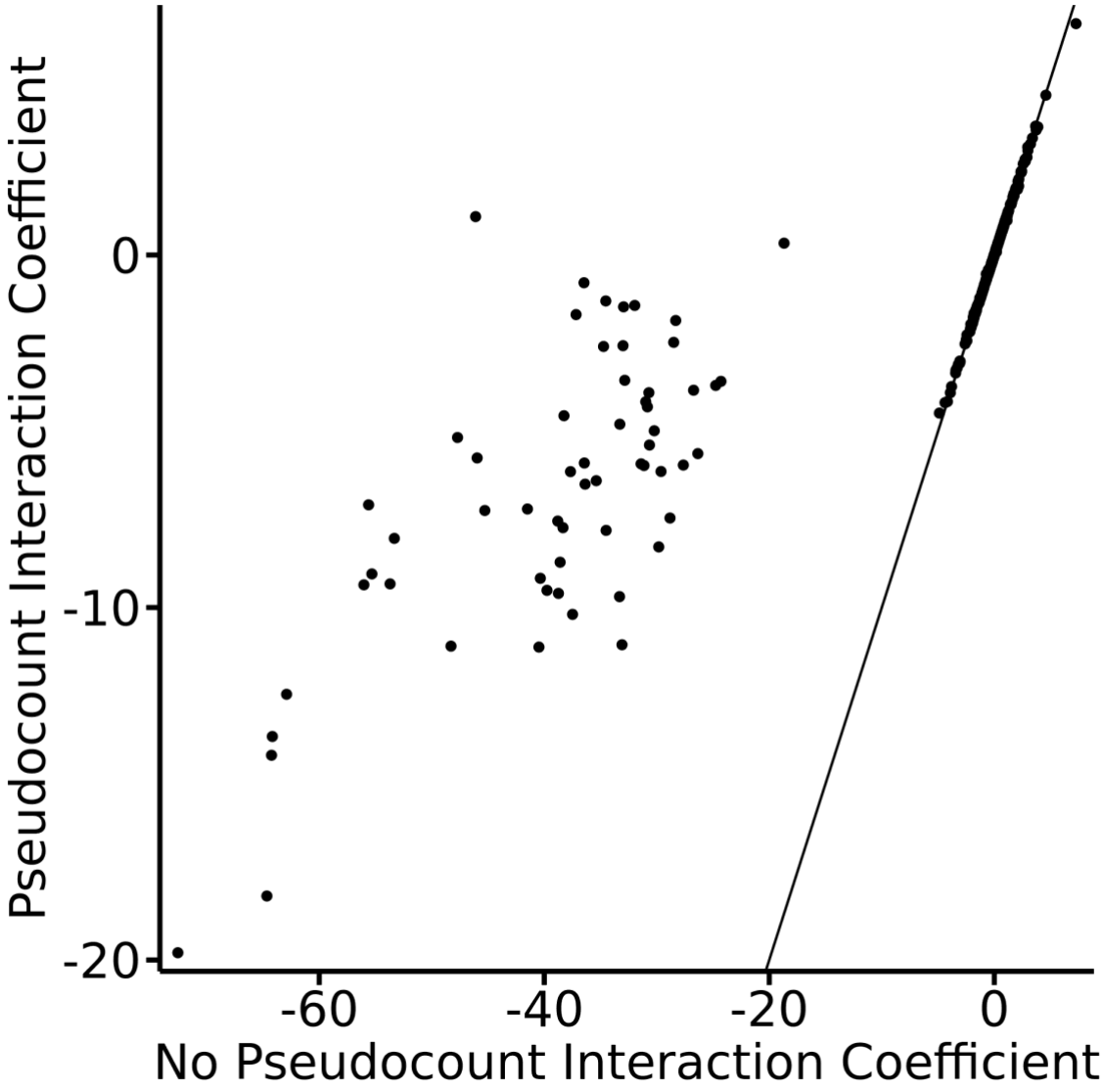
Supplemental Figure 2.1: Simulated noisy gRNA efficiency values and their effects on coefficient estimates. **a)** Histogram of noisy and true guide efficiencies from simulations with different values of D. D is the dispersion-controlling coefficient used to control "noise." **b)** Scatterplot comparing noisy guide efficiencies to true guide efficiencies with different values of D. **c)** Scatterplot comparing true versus estimated coefficient values for each gene evaluated with GLiMMIRS-base. These plots summarize the results of fitting the model to 1000 genes in the simulated dataset which were designated as "true" target genes (genes whose enhancers were perturbed by gRNAs in the simulated experiment). Plot shows results of fitting to simulated data using the three different sets of noisy guide efficiencies. A pseudocount of 0.01 was applied to the counts for all cells. Coefficients of determination (R2) are shown. 36 outliers fall outside the axis range and are not visible in the enhancer panel for the set where D=1.



Supplemental Figure 2.2: Simulated MOI corresponds to the number of testable pairs. a) Distribution of the number of unique gRNAs per cell at different simulated MOIs (represented by the λ parameter in the Poisson distribution). **b-c)** Distribution of target site pair frequencies, or the number of cells receiving gRNAs targeting both sites in a "ground truth" (positive) pair of interacting enhancers, at different simulated MOIs (λ).



Supplemental Figure 2.3: Interaction coefficient estimates do not vary with distance between enhancers. a) Distribution of distances between enhancer pairs for the 330 enhancer pair set. **b)** Distribution of distances between enhancer pairs for the 3,808 enhancer pair set. **c)** Distance between enhancer pairs and magnitude of interaction coefficients for 330 enhancer pair set tested. **d)** Distance between enhancer pairs and magnitude of interaction coefficients for 3,808 enhancer pair set tested. Blue lines in **c)** and **d)** are loess curves fitted to the data.



Supplemental Figure 2.4: Outlier interaction coefficient estimates are moderated by introduction of a pseudocount. Magnitude of interaction term coefficients for 330 enhancer-enhancer pairs when adding vs. not adding a pseudocount of 0.01 to adjust the gene expression. The inclusion of a pseudocount greatly reduces the magnitude of outlier interaction coefficient estimates (note difference in x and y axis scales).

REFERENCES

1. Brady G, Barbara M, Iscove NN. Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. Methods Mol Cell Biol. 1990;2(1):17–25.

2. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, Coleman P. Analysis of gene expression in single live neurons. Proc Natl Acad Sci U S A. 1992 Apr 1;89(7):3010–3014. PMCID: PMC48793

3. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May;6(5):377–382. PMID: 19349980