5-3-2024

Neuroepigenetic Mechanisms of Brain Development: From Technology to Biological Insights

Allen Yen
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Developmental, Regenerative, and Stem Cell Biology

Dissertation Examination Committee:
Joseph D. Dougherty, Chair
Kristen L. Kroll
Robi D. Mitra
David M. Ornitz
Allegra A. Petti

Neuroepigenetic Mechanisms of Brain Development: From Technology to Biological Insights

by
Allen Yen

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2024
St. Louis, Missouri

Table of Contents

List of Figures

Supplemental Figures

List of Tables

Acknowledgments

Embarking on a PhD journey marks a significant milestone in the continuous adventure of personal and professional growth—a journey that is never traveled solo. Along this path, numerous individuals have played pivotal roles, turning challenges into opportunities and dreams into realities. To all these remarkable supporters, I express my deepest appreciation and dedicate this acknowledgement.

First and foremost, I am eternally grateful to my family, who have been my foundational support at every stage of my life. Your unwavering love and consistent encouragement have been instrumental in shaping me into the person I am today. Mom and Dad, thank you for raising me and supporting me. Mom, your consistent support, and belief in me has been foundational to my journey. You have always encouraged me to pursue my passions and interests. Dad, I admire your foresight in anticipating needs and orchestrating details, your exceptional ability to plan ahead, and your vast knowledge across numerous domains. It's truly inspiring to see how you blend practical wisdom with a broad understanding of the world to navigate and prepare for the future so effectively. Erica, I would like to extend a heartfelt thanks to you for your guidance. Following your journey is like having a life hack or a sneak peek into my future—it's like you're blazing the trail and I'm just cruising along with a "been there, done that" map thanks to you. It's almost like having a life cheat code, making everything seem a bit less daunting. This support network has also grown in the last few years. Phuong, you have also invariably offered sound life advice and provide perspectives that are both enlightening and profound. You also have a unique talent for savoring the joyful moments of life while also demonstrating utmost dedication and focus when it comes to serious work and commitments. This admirable blend of lightheartedness and diligence is both rare and inspiring. Zoey and Miles, it’s been a remarkable experience watching the two of you grow up and learn at such an impressive pace. Witnessing your rapid development and acquisition of new skills and knowledge is truly a joy. Your curiosity and eagerness to explore the world around you are infectious, and it’s a privilege to observe your journey into young learners.

I am forever grateful for Lisa Hsieh, who fostered my enthusiasm for mathematics in high school. She taught me not only mathematics, but also invaluable life lessons. Amid the pressures of schoolwork and extracurriculars, she emphasized the critical balance between hard work and rest. She encouraged me to pause, inhale deeply, and appreciate the natural ebb and flow of life—recognizing when to immerse in tasks and when to rest. Her advice underpins much of my productivity and wellness today. My gratitude extends to all she has offered: her insight, her support, and her enduring example. Lisa will always remain in my thoughts and in my memories.

I would like to thank Robert Chow for giving me my first glimpse and hands-on lab experience as an intern at StemCyte. My most vivid memory was the first time I walked into the cord blood storage facility, where there were rows and rows of liquid nitrogen cryogenic storage tanks, each large enough to engulf a person. The sheer scale and methodical organization were almost cinematic, akin to a scene out of science fiction movie. I’m forever thankful that you trusted me as a fresh high school graduate with centrifuges, controlled-rate freezers, liquid nitrogen tanks, and HLA typing experiments. This was my first venture into the industrial lab setting, an experience that shaped my perception of the scientific world, embedding itself as a cornerstone of my early career aspirations.

Word cannot express my gratitude to my mentor and friend Chih-Lin Hsieh (unrelated to Lisa). You were the first to teach me essential lab skills, but also nurtured my curiosity. Your teachings instilled in me a profound appreciation for the investigative spirit that drives scientific inquiry. You're not confined to the lab as you could effortlessly run a farmer's market from your backyard, given your abundant harvest of vegetables, fruits, and herbs. Your ability to cultivate growth, be it in plants or budding scientists, is truly remarkable. I am thankful that our relationship extends beyond professional bounds and evolved into a lasting friendship. I hope to mirror your impact, inspiring future generations of scientists with the same generosity and spirit that you have bestowed upon me.

I am immensely grateful to Victoria Bolotina for all your exceptional mentorship and support. You saw the potential in me for a career in science well before I recognized it myself. The opportunity to work in your lab was an invaluable experience, offering me not just skills in microscopy and stem cell culture, but a true behind-the-scenes view of the research world. Your guidance in navigating the complexities of grant writing, interactions with journal editors and reviewers, the essentials of data organization, figure creation, and networking at scientific conferences has been fundamental to my growth. I am profoundly thankful for your mentorship, your belief in my abilities, and your instrumental role in shaping me into the scientist I have become today.

I want to thank all my dear friends associated with Boston—those I initially met there, those who have moved away, and those who have recently made it their home. This city, which continues to hold a piece of my heart even after relocation to Saint Louis, has been the backdrop to so many valued friendships. So to Chris, Steve, James, Tom, Jena, Kyle, Pavania, Beatriz, Stuart, Jamie, Jason, Vladimir, and Qingde, thank you for being such a significant part of my life, for keeping our connections strong, and always giving me a reason to look forward to going back.

Relocating to Saint Louis opened the door to an extraordinary community, where I've had the privilege of forming a new support network. I am fortunate to have Joe Dougherty as my mentor. Your visionary approach and commitment to exploring bold, innovative ideas are incredibly inspiring. Moreover, your ability motivate is exceptional. I recall feeling anxious about our first biannual evaluation thinking that I hadn't made enough progress on my projects, yet I left it invigorated and eager to pursue our established goals. I am thankful for your confidence in me to engage in what could only be described as "pie in the sky" or "cowboy" experiments using cutting-edge technologies under your guidance. Rob, you've essentially been a co-mentor to me and thank you for your steadfast support. The number of beers wagered on various results and theories added a fun and spirited layer to our work, turning even the most rigorous analyses into memorable and enjoyable moments. I must also express my appreciation to Kristen Kroll, Allegra Petti, and David Ornitz. Your invaluable contributions played a crucial role in enhancing the quality of my work. I am deeply grateful for each of your expertise and support throughout this process.

I would like to thank both past and present members of our lab—Jiayang, Simona, Din, Sarah, Sneha, Colin, Titi, Mari, Mike, Tony, Gareth, Katie, Susan, Rebecca, Stu, Tomás, Francesca, Bernie, Rachel, Stephen, and Darshan, for creating an environment that is not only productive but also truly enjoyable. A special shoutout to Kelli our amazing lab manager, whose efforts ensured that our lab functions seamlessly. I also want to thank past and present members of the Mitra lab—Xuhua, Juanru, Arnav, Matt, and Josh. I would like to thank Scale Biosciences, especially Fatjon Leti, Dominic Skinner, and Maggie Nakamoto, for coming to the rescue with their assay and making Chapter 5 and 6 possible. Outside the lab, I'd like to thank Alex, Lorenzo, Melvin, Jamison, Sam, Pat, and Amy for all the fun memories.

Finally, I cannot express enough gratitude to Maggie, your unwavering support has been my anchor through this journey. Your encouragement through every challenge, your genuine happiness at my achievements, and your constant, comforting presence have been invaluable. I deeply admire that you strike a perfect balance between your professional dedication and playful spirit. While you take your responsibilities and care for your patients seriously, you also know the importance of not taking life too seriously, like with spontaneous Nerf dart wars. You remind me of balance, ensuring I rest and recharge when needed. I cherish the time we spend together creating meals with real food, as well as in Overcooked, where our culinary adventures take a more frantic turn. As we close this chapter, I am filled with excitement and anticipation for the adventures and growth that await us in the next. I love you.

In closing, I am grateful for everyone who has supported me on this journey. Each of you have contributed in your own unique way to my achievements. Your guidance, inspiration, and encouragement have been instrumental in my growth, and I carry forward not only the knowledge, but also the invaluable relationships nurtured along the way. Thank you all for being part of my story.

Allen Yen

Washington University in St. Louis

May 2024

Dedicated to my parents, Teresa and Hubert Yen.

ABSTRACT OF THE DISSERTATION

Neuroepigenetic Mechanisms of Brain Development:

From Technology to Biological Insights

by

Allen Yen

Doctor of Philosophy in Biology and Biomedical Sciences

Developmental, Regenerative, and Stem Cell Biology

Washington University in St. Louis, 2024

Professor Joseph D. Dougherty, Chair

Each cell in the brain has the same genomic sequence, yet they can have vastly different phenotypes and function. This diversity is a result of complex genetic and signaling pathways, and knowing how these are regulated is key to understanding physiological development and how pathogenic dysfunctions arise. Genomic methods such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) and assay for transposase-accessible chromatin with sequencing (ATAC-seq) have played pivotal roles in dissecting these processes. However, a key limitation is that the cells are destroyed upon observation of their molecular states, which preclude our ability to correlate historical epigenetic information with future readouts of cell function or animal behavior. This dissertation explores technological advancements and their application in studying neurodevelopmental disorders. First, I introduce the design and iterative development of Calling Cards, a method that records transient molecular events, enabling retrospective analysis of

gene regulatory elements and gene expression. This is particularly pertinent for analyzing cellular outcomes that are undetermined at the time of a molecular event. To demonstrate the unique insight that this can provide, I apply Calling Cards in the mouse brain to understand how the observed consequences of neurodevelopmental disorders can be associated with historical molecular events.

This dissertation is structured as follows: Chapter 1 introduces foundational concepts and context, particularly focusing on epigenetics of the developing brain. Chapter 2 offers a detailed guide to bulk Calling Cards, equipping researchers at all levels to conduct and analyze these experiments independently, featuring enhanced reagents and protocols for improved assay sensitivity and flexibility. This chapter also includes a guided tutorial of custom software for data processing, facilitating broader access and application of this technology. Chapter 3 details the generation of transgenic Calling Cards mouse lines, sharing the insights gained from this approach. Chapter 4 applies bulk Calling Cards to examine epigenetic differences in brain masculinization and their role in sex-dependent gene expression, revealing candidate genomic regions associated with neurodevelopmental disorders may be influenced by perinatal hormonal fluctuations.

The dissertation continues into the realm of single-cell genomics, presenting the combinatorial indexing Calling Cards method, which allows for the concurrent analysis of the transcriptome and gene regulatory elements. Chapter 5 is a study that focuses on defining the core phenotype of a syndromic neurodevelopmental disorder. It uses the transcriptomic data to delineate how cortical development goes awry in a bona-fide model of autism and intellectual disability. Chapter 6 delves into the methodological advancements of adapting Calling Cards to the combinatorial indexing platform and shows preliminary analysis of the gene regulatory elements, setting the stage for future research. Finally, Chapter 7 concludes this dissertation by summarizing the key findings and significance of each chapter and proposing potential future directions.

In summary, the presented body of work expands our understanding of epigenetic gene regulation in brain development and neurodevelopmental disorders, offering new perspectives on the epigenetic underpinnings of these processes. The advancements in Calling Cards technology presented herein aims to equip the scientific community with innovative tools for exploring biological phenomena across various fields and disciplines.

Chapter 1: Epigenetics of the developing brain

1.1 Preface

This chapter contains contents from a published manuscript:

MYT1L in the making: emerging insights on functions of a neurodevelopmental disorder gene

Jiayang Chen*, Allen Yen*, Colin P. Florian*, Joseph D. Dougherty

* Authors contributed equally. The order of co-first authors was determined by rounds of Super Smash Brothers.

Translational Psychiatry 12, 292. July 22, 2022

1.2 Introduction

The mammalian brain is the most complex organ in the body and the complex cellular morphologies, connections, and functions continue to challenge neuroscientists today. To effectively grasp the brain's complexities, it is helpful to deconstruct it into more manageable units. By systematically analyzing the brain from broader regions to specific circuits, individual cells, and their molecular characteristics, we can achieve a more nuanced understanding of its various components. Proper brain development requires that cells are generated in the proper order, number, and location. The fundamental processes of neurodevelopment exhibit similarities across invertebrates and vertebrates, indicating that certain underlying mechanisms of neuronal specification and temporal organization are evolutionarily conserved as reviewed in (Holguera and Desplan, 2018). The expansion of the cerebral cortex is what makes the human brain distinct from all other animals. It is comprised of an estimated 16 billion neurons and 61 billion non-neuronal cells that can be organized into over 50 distinct anatomical areas (Azevedo et al., 2009; Glasser et al., 2016). The dorsolateral prefrontal cortex (dlPFC), a nexus for higher cognitive functions and complex social behaviors, becomes especially pertinent as molecular and cellular disruptions in the dlPFC circuitry have been implicated in many neuropsychiatric diseases (Allard, 2012; Goldstein and Volkow, 2011; Grimm et al., 2008; Koenigs and Grafman, 2009; Smucny et al., 2022). Therefore, to unravel the complexities of neurodevelopmental disorders, we must first understand the molecular and cellular mechanisms underlying cell diversity and function.

Since the nineteenth century, analysis of cell morphology has been a cornerstone of biological research, emphasizing that the shape of a cell is closely linked to its function. This principle led to many early significant discoveries, including Ramón y Cajal's neuron doctrine

(Ramón y Cajal, 1954), which fundamentally changed our understanding of neural connectivity. Using the Golgi stain technique, Cajal described what he called “espinas”, or thorns, on the surface of Purkinje cells, which was the first documented description of dendrites (Ramón y Cajal, 1888). After seeing these protrusions in various species, he speculated that these spines must receive axonal inputs from other neurons and serve as a point of contact between other cells, meaning that neurons are independent units that can connect to each other. As methodologies evolved, researchers began categorizing cells not only by their shape but also by their function and electrophysiological properties. Today, in the genomics era, we can classify cell types based on molecular markers identified by high throughput analysis and cataloging of mature neurons. Despite these advances, these atlases may not fully represent the breadth of neuronal diversity, as they capture only a static picture of the cellular landscape at a particular moment. To truly grasp brain complexity, we must explore the dynamics of neuronal specification, migration, maturation, and integration into functional networks throughout various stages of neurodevelopment and maturation.

1.3 Epigenetic modifications in neurodevelopment

Mammalian brain development requires a complex cascade of gene expression patterns in a temporal and spatial manner to generate the diversity of cell types for proper neural function. During neurogenesis, progenitor cells can either undergo symmetrical division to maintain its identity through self-renewal or proliferation, or it can undergo asymmetrical division, leading to the emergence of a daughter cell with a new identity or state (Betschinger and Knoblich, 2004; Clevers, 2005; Yamashita et al., 2005). Despite having identical genomic sequences, cells differentiate into varied types through epigenetic processes, where heritable changes cannot be explained by the genomic sequence alone (Deans and Maggert, 2015). One of these epigenetic mechanisms is DNA methylation. The most widely studied pattern is when the fifth carbon position of cytosine is methylated (5mC) in CpG dinucleotides at gene promoter regions (Law and Jacobsen, 2010). While methylation is thought to be associated with transcriptional repression, some transcription factors were found to have enhanced promoter binding in methylated promoter regions (Yin et al., 2017). This suggests that this epigenetic mark has multiple functional roles and functions that are dependent on their context (Jones, 2012).

Expanding upon this framework, another layer of transcriptional regulation involves histone modifications, which modulate the architecture of chromatin and thus the accessibility of DNA to transcriptional machinery. Post-translational modification of histone proteins can alter their interaction with DNA; for instance, interactions that strengthen histone-DNA interactions lead to tightly packed nucleosomes and heterochromatin formation, while those that decrease these interactions result in more open chromatin structure conducive to gene activation. Acetylation of lysine 9 and lysine 27 on histone H3 (H3K9ac and H3K27ac) is commonly found at the enhancers and promoters of actively transcribed genes. These acetylation marks are integral to regulating processes such as cell cycle regulation, proliferation, and differentiation (Lee and Lee, 2010; Murao et al., 2016; Zhang et al., 2020). Additionally, histone methylation can either activate or repress transcription, depending on the site of methylation. Unlike acetylation, which can alter the charge of histones and impact their interaction with DNA directly, methylation does not change the charge and hence a more nuanced effect. For instance, H3K4me1 is associated with active enhancer regions, H3K4me3 with active promoters, and H3K9me3 and H3K27me3 with transcriptionally silenced genomic regions (Di Nisio et al., 2021). These modifications are dynamically regulated by specific enzymes, often referred to as “writers” and “erasers,” and the proper balance is essential in regulating the activity of genetic programs and the downstream cellular processes (Husmann and Gozani, 2019).

The functional impact of histone modifications is interpreted by “reader” proteins, which guide the recruitment of various complexes to modify the chromatin landscape, thus promoting or inhibiting transcription. An example is Brd4, a member of the bromodomain and extra-terminal domain (BET) protein family, which recognizes acetylated lysines and, through interaction with the positive transcription elongation factor b (pTEFb) complex, enhances chromatin accessibility and transcriptional activity (Dey et al., 2003; Jang et al., 2005; LeRoy et al., 2012; Wu and Chiang, 2007; Yang et al., 2005). BET proteins are crucial in cell fate specification and maintaining neuronal function by regulating neurotransmitter receptors, ion channels, neuroplasticity, and cognition (Korb et al., 2015; Sartor et al., 2015; Sullivan et al., 2015). Small molecule BET inhibitors such as JQ1 prevents BRD4 from binding to chromatin, displacing the Mediator complex and RNA Polymerase II from enhancers, resulting in reduced target gene expression (Bhagwat et al., 2016; Crump et al., 2021; Filippakopoulos et al., 2010; Kanno et al., 2014; Lovén et al., 2013). Given their integral role in these pathways, disruptions in histone modifications or their effector proteins can lead to pathologies, underscoring the importance of understanding these epigenetic patterns and their functional outcomes.

1.4 Methods to profile transcription factors, histone modifications, and enhancers

To understand how transcription is regulated, it becomes necessary to map protein-DNA interactions and histone modifications across the entire genome. Having the binding profile of a particular transcription factor (TF) and the associated transcriptional machinery is fundamental to deciphering gene regulatory networks that underlie biological processes. The interactions between chromatin states and transcriptional regulation are complex. Therefore, a comprehensive profiling of the epigenome across various biological contexts and cell types is necessary to understand physiological processes and to identify pathogenic deviations that lead to disease.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has been the gold standard technique for mapping the genome-wide distribution of DNA-binding proteins, nucleosomes, and histone modifications (Johnson et al., 2007). The process involves crosslinking the DNA-binding protein to the DNA, followed by fragmenting the chromatin into smaller fragments that usually range from 200-600bp. These fragments are then enriched using a specific antibody to isolate the DNA-protein complex, and the resulting material is prepared into sequencing libraries. By sequencing the fragments, investigators can identify the precise locations of the protein-DNA interactions across the genome. Despite its utility, ChIP-seq has some drawbacks. It is time-consuming, requires crosslinking which does not preserve the native protein-DNA interactions, and requires a substantial number of cells and sequencing depth. To overcome these challenges, newer techniques such as CUT&RUN (Skene et al., 2018; Skene and Henikoff, 2017) and CUT&TAG (Kaya-Okur et al., 2020, 2019) have been developed, which are more efficient mapping approaches with lower cell number and sequencing requirements. The quality of the data still relies heavily on the specificity and quality of the antibodies; any lack of specificity can result in high background noise or poor enrichment of the target complex. Additionally, like many other genomic techniques, they are inherently destructive: to analyze the protein-DNA interactions, the cells or tissues must be harvested, which provides only a snapshot at the time of sample collection and eliminates the possibility of subsequent observations in the same sample. These advances and limitations highlight the dynamic nature of genomics research and the continuous need for methodological improvements.

Calling Cards technology offers a novel approach to complement existing methods like ChIP-seq, CUT&RUN, and CUT&TAG by overcoming some of their limitations. Calling Cards is a molecular recording strategy that allows for the capture of cumulative protein-DNA interactions over time (Cammack et al., 2020; Lalli et al., 2022; Moudgil et al., 2020b). The technology hinges on two key components: a fusion of a DNA-binding protein with the *piggyBac* transposase and a self-reporting transposon (SRT). Within the cell, the transposase inserts the SRT into the genome near where it binds. This integration of the SRT functions as a permanent record of where the DNA-binding protein bound to the genome, which can be recovered at a later time using next-generation sequencing. This retrospective analysis can offer unique insights into biological processes that are not possible with chromatin immunoprecipitation methods. That can be especially pertinent to but not limited to developmental biology studies involving dynamic cellular processes. It allows scientists to trace back the interactions that might influence critical outcomes, such as cell differentiation or lineage specification. Importantly, Calling Cards does not rely on antibodies and provides a flexible and versatile framework to study virtually any DNA-binding protein of interest, expanding the scope of genomic investigation. Calling Cards provides a temporal record of protein-DNA interactions, which establishes a foundation for exploring the genetic underpinnings of complex traits and diseases, especially where traditional methods might not capture the entire scope of molecular interactions over dynamic processes like neurodevelopment.

1.5 Emerging insights on functions of a neurodevelopmental disorder gene

Human genetic studies recently associated the gene Myelin Transcription Factor 1 Like (*MYT1L)* with neurodevelopmental disorders (NDDs) (Blanchet et al., 2017; Coursimault et al., 2021; de Ligt et al., 2012; Loid et al., 2018; Sanders, 2015; Satterstrom et al., 2020; The DDD Study et al., 2014; Wang et al., 2016; Windheuser et al., 2020). Specifically, MYT1L loss of function (LoF) is associated with intellectual disability (ID) and autism spectrum disorder (ASD), while MYT1L duplication has been observed in patients with schizophrenia (SCZ) (Mansfield et al., 2020). Yet, the mechanism by which MYT1L variants contribute to disease pathology is still unknown.

MYT1L, along with Myelin Transcription Factor 1 (MYT1) and Suppression of Tumorigenicity 18 (ST18/MYT3), is part of the three-gene MYT/neural zinc finger (NZF) transcription factor (TF) family. These TFs are characterized by DNA binding C2HC-type zinc fingers, and a MYT1 domain, which is hypothesized to function as a transcriptional repressor (Mall et al., 2017; Romm et al., 2005). While all three TFs are found to be expressed in the developing brain, MYT1L has specifically been shown to enhance neuronal differentiation (Mall et al., 2017; Matsushita et al., 2014). Seminal studies have shown that overexpression of ASCL1 and BRN2 reprograms fibroblasts into functional neurons in vitro and the addition of MYT1L significantly increases conversion efficiency (Wapinski et al., 2013). However, the exact role of MYT1L during this transdifferentiation process remains poorly understood. As a member of MYT/NZF protein family, it is thought that MYT1L represses its target genes' expression, reminiscent of the known repressive functions of MYT1. Indeed, in vitro neuronal transdifferentiation studies demonstrated that MYT1L represses non-neuronal gene expression, while promoting neuronal differentiation

(Mall et al., 2017). On the other hand, both in vitro and in vivo studies indicate MYT1L can activate gene expression with a comparable magnitude to reported repression, suggesting that it can also function as an activator (Chen et al., 2021; Manukyan et al., 2018). Further studies are needed to resolve its true molecular function in biologically relevant contexts.

1.5.1 The association of MYT1L mutation and human disease

Human genetic studies have identified genetic mutations in transcription factors and chromatin remodelers (*MECP2*, *CHD8*, *SETD5*, etc.) as causes for various forms of neuropsychiatric disorders, ID, ASD, and SCZ (Amir et al., 1999; Deliu et al., 2018; Katayama et al., 2016; Sanders, 2015; Satterstrom et al., 2020; The DDD Study et al., 2014). One of these newly associated factors is MYT1L.

With the increased integration of genome sequencing into the clinic over the last 10 years, *MYT1L* mutations, mostly de novo, have consistently been found in patients with early onset neurological disorders. Currently, there are over 100 described patients with MYT1L mutations, with 80% of them harboring potential *MYT1L* LoF mutations and others harboring *MYT1L* partial duplications (Blanchet et al., 2017; Coursimault et al., 2021; Mansfield et al., 2020; Windheuser et al., 2020). *MYT1L* LoF mutations include deletions, frameshift, and single nucleotide variations (SNVs), which are predicted to cause decreases in mRNA production or aberrant protein functions. Notably, missense mutations from clinical but not general-population studies cluster in the central zinc finger domains and the MYT1 domain (Adzhubei et al., 2010; Karczewski et al., 2020) (**Figure 1A**), the most confident structures predicted by AlphaFold (**Figure 1B**), indicating these domains might be crucial for the protein’s functions (Jumper et al., 2021; Varadi et al., 2022).

Among patients with *MYT1L* LoF mutations, ID, ASD, and developmental delay are the most common symptoms. Other phenotypes include seizures, syndromic obesity, microcephaly, macrocephaly, and muscular hypotonia. This constellation of symptoms has now been recognized as MYT1L Syndrome or 2p25.3 Deletion Syndrome (Blanchet et al., 2017; Coursimault et al., 2021; Mansfield et al., 2020; Windheuser et al., 2020). In addition, most patients with *MYT1L* partial duplications were reported to either have ID, ASD, or both. It seems these developmental impacts of MYT1L haploinsufficiency indicate a well-conserved role for the protein: across two labs with independently generated lines, MYT1L haploinsufficient mice were also shown to have obesity, hyperactivity, and social deficits (Chen et al., 2021; Wöhr et al., 2022).

Finally, regarding *MYT1L* duplications in humans, although 33% of *MYT1L* duplication patients presented with SCZ exclusively, all but one of those duplications contain neighboring gene *PXDN*, indicating *MYT1L* may not be the only contributing factor in the region for SCZ risk (Mansfield et al., 2020). The association of both LoF and putative duplications with disease indicates that neurobiology is very sensitive to the levels of MYT1L activity and identifying the loci that are influenced by altered MYT1L levels might aid in understanding the downstream pathophysiology. Therefore, in the following sections, we summarize previous studies on MYT1L to provide mechanistic insights into its cellular and molecular functions under different contexts.

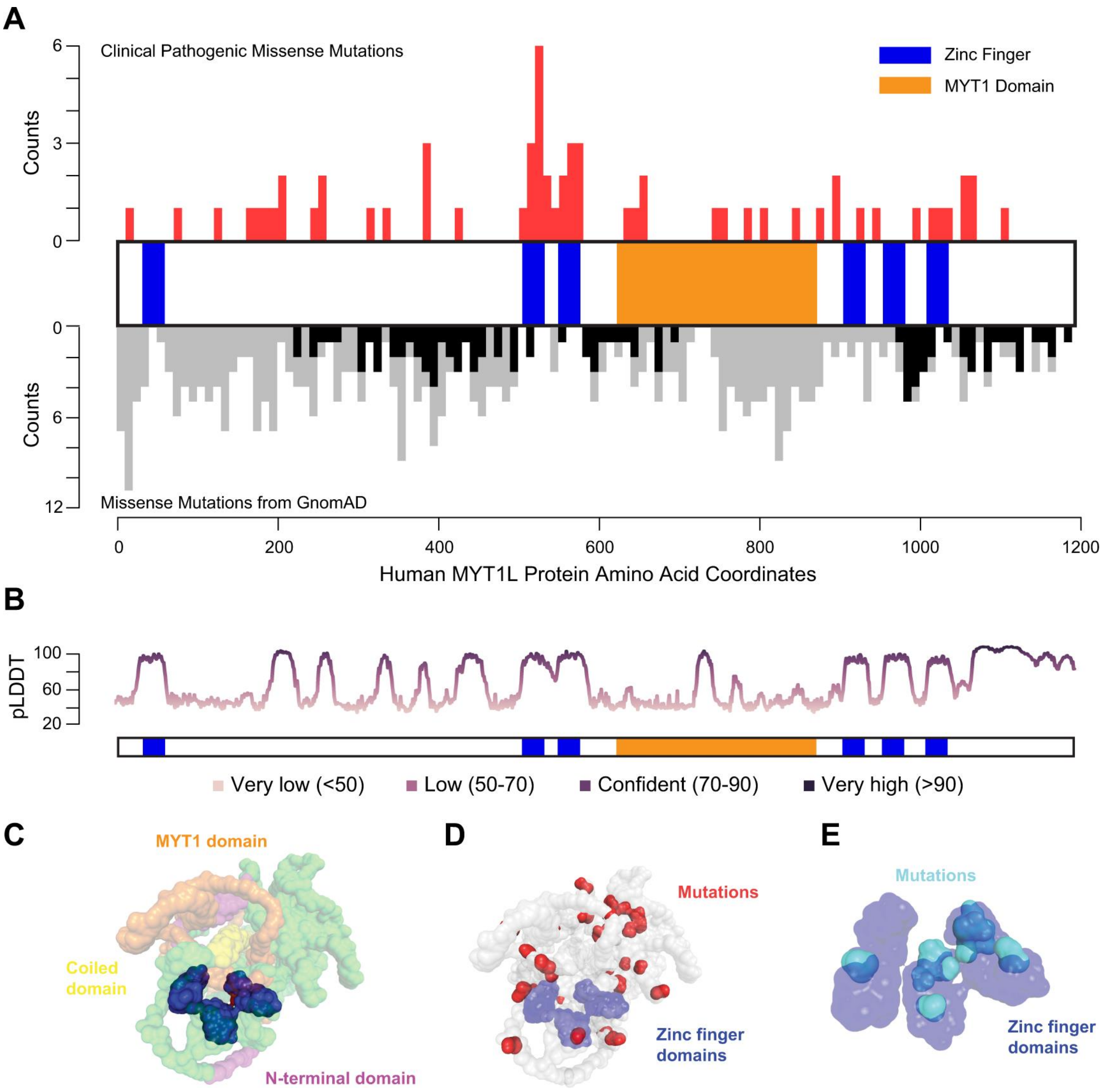


Figure 1: Schematic of human MYT1L domains and predicted protein structure by AlphaFold

(A) Distribution of missense mutations described in clinical studies (top, red) compared to a general population sample (gnomAD, bottom, with gray bars displaying all missense mutations and black bars displaying 'possible damaging mutations' as predicted by PolyPhen2). 'Possible damaging mutations' in the general population are largely excluded from the regions mutated in clinical samples. (**B)** AlphaFold's calculated confidence measure (pLDDT score) per-residue of the model's prediction based on the IDDT-Cα metric. **(C)** 3D AlphaFold structure (AF-Q9UL68-F1) prediction of MYT1L protein showing the N-terminal domain (magenta), MYT1 domain (orange), coiled domain (yellow), and six zinc finger domains (blue) coming in proximity with each other to form a putative DNA-binding pocket. Unannotated regions are shown in green. (https://alphafold.ebi.ac.uk/entry/Q9UL68). (**D)** Loss of function mutations from patient reports are found throughout the protein. Those not within the annotated zinc finger domains (blue) are shown in red. **(E)** Isolated and magnified view of the zinc finger domains (blue) shows patient mutations (cyan) cluster in the zinc fingers.

1.5.2 MYT1L functions to promote neuronal maturation

Neuronal identity is determined by the effects of a combination of basic helix-loop-helix (bHLH) TFs (i.e., ASCL1, NEUROD1, and NEUROG1) as well as other developmentally expressed TFs such as BRN2 and MYT1L. In vitro overexpression studies have shown that the pioneer factor ASCL1 is sufficient for induction of neuronal traits, but overexpression in combination with other factors such as BRN2, and especially, MYT1L is necessary for efficient fibroblasts conversion to neurons as well as the maturation of the induced neurons (iNs) (Mall et al., 2017; Tomaz, 2016; Vierbuchen et al., 2010). Ultimately, many of these studies suggest that MYT1L and other members of the MYT family primarily function to preserve neuronal phenotypes as it has been shown that MYT1L is mostly expressed during the post-specification phase when cell populations have become post-mitotic. Furthermore, none of the MYT family members were observed to be expressed by in situ hybridization in germinal zones containing mostly undifferentiated cells (Kameyama et al., 2011; Matsushita et al., 2014), and very little overlap (5%) was seen with SOX2 positive progenitors (Chen et al., 2021). Interrogation of specific domains of MYT1L has further defined its role in neuronal conversion. For example, Mall et al. (2017) showed that, when fused to an activating element (VP64), the DNA binding domains of MYT1L displayed a dominant-negative effect on ASCL1-mediated neuronal conversion. In addition, just a 423-amino-acid fragment (i.e., amino acids 200–623), which contains the N-terminal domain and the middle two zinc fingers, was functionally indistinguishable from full-length MYT1L. Surprisingly, this fragment does not contain the MYT1 domain.

In contrast to the overexpression studies discussed above, knockdown of MYT1L via short hairpin (sh) RNAs resulted in a reduction of neuronal maturation gene programs such as neurite outgrowth, axonal development, synaptic transmission, and extracellular matrix composition, which hints that MYT1L also acts as an activator (Kepa et al., 2017). It has also been reported that MYT1L was found to be deleted (~5%) and downregulated (>80%) in glioblastomas, suggesting that gliomagenesis requires neutralization of terminal neural differentiation (Hu et al., 2013). Furthermore, others have shown that MYT1L and MYT1 expression can slow tumor growth in glioblastoma cell line models via repression of pro-proliferative genes (Melhuish et al., 2018). However, impacts on glia in vivo are likely not direct since MYT1L expression has not been consistently observed in glia (Chen et al., 2021; Kim et al., 1997).

Spatiotemporal expression of MYT family TFs is finely tuned across development, specifically during neuronal maturation. Of the MYT family, Myt1 and Myt3 are expressed the earliest at embryonic (E) day 9.5 as suggested by in situ hybridization (Matsushita et al., 2014). Quantitative RT-PCR results showed that Myt1 and Myt1l were upregulated from E10.5 to E15.5 and then downregulated postnatally (**Figure 2A**) (Matsushita et al., 2014). In addition, Myt1l mRNA levels increase across neurogenesis in mice, and low levels are sustained in adulthood, which mirrors human expression patterns (Matsushita et al., 2014). In mice, MYT1L protein levels were sustained from E14 (beginning) to postnatal (P) day 1 and declined thereafter (Chen et al., 2021), but remained detectable indefinitely. The earliest time point of detectable Myt1l expression occurs at E9.5 in the ventrolateral portion of the spinal cord, again where newborn neurons are found. In addition, at E12.5, BrdU staining to identify proliferating cells hardly overlapped with Myt1l expression, further supporting that Myt1l-positive cells were mostly post-mitotic

(Matsushita et al., 2014). Indeed, across the multiple CNS regions examined (spinal cord, hindbrain, midbrain, cortex, and retina), Myt1l mRNA was upregulated when neurons began to differentiate (**Figure 2B**) and overlapped with markers of neurons. Overall, analysis of Myt1l expression pattern and time course further supports the assumption that it is responsible for neuronal maturation and preservation of cell fate.

Several in vivo studies have also shed light on MYT1L's necessity for neuronal maturation. In zebrafish, knocking down human MYT1L orthologs, myt1la and myt1lb, by antisense morpholinos (MO) results in almost complete loss of oxytocin (OXT) and arginine vasopressin (AVP) in the neuroendocrine pre-optic area of the hypothalamus, suggesting MYT1L LoF might affect neuroendocrine system development (Blanchet et al., 2017). This could either represent loss of these neurons, or loss of their maturation since neuropeptide expression occurs relatively late in neuronal maturation (Almazan et al., 1989). In a MYT1L Syndrome mouse model that displays MYT1L haploinsufficiency, precocious neuronal differentiation from progenitors to immature neurons was observed upon MYT1L loss during early brain development (Chen et al., 2021) (**Figure 2C**). This suggests MYT1L LoF leads to loss of proliferating cells during development and correspondingly a smaller brain in the adult, providing a mechanistic explanation for the human patients' microcephaly. In addition, assessment in adults revealed MYT1L heterozygous mice show impaired neuronal maturation in terms of transcriptional profiles, neuronal morphology, and potentially neuronal electrical properties (Chen et al., 2021). In summary, MYT1L may have multiple roles in neurodevelopment, with strong evidence that at least one may be promoting neuronal maturation.

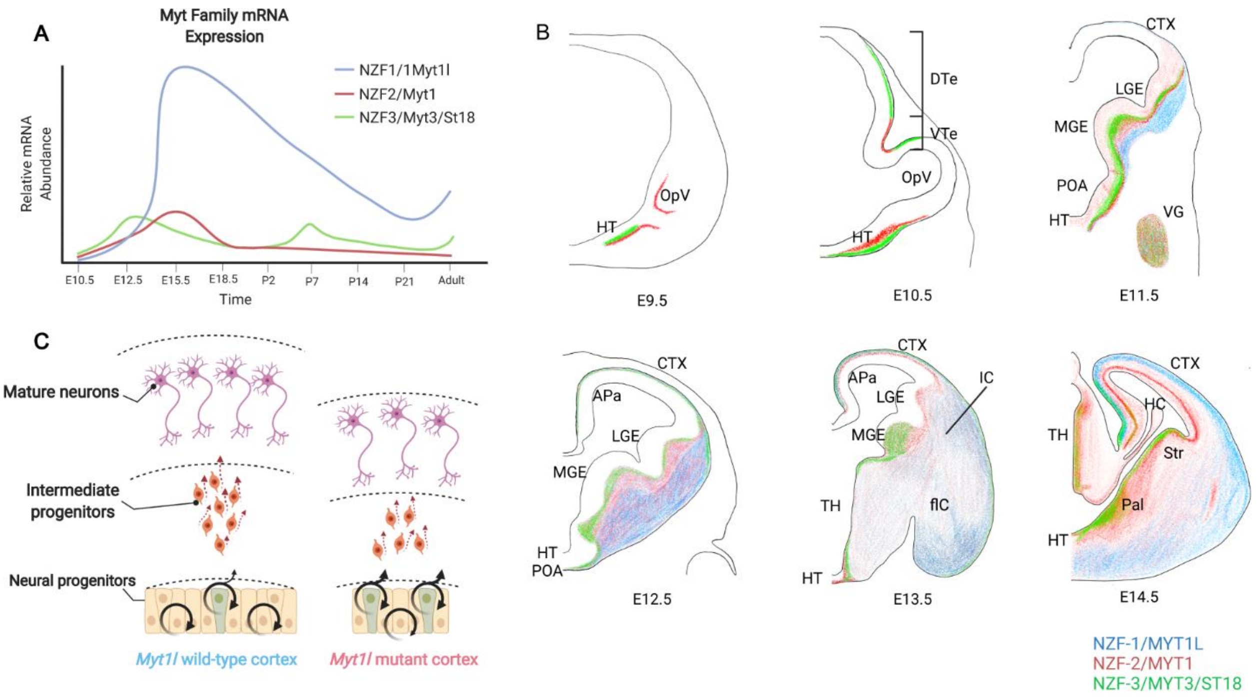


Figure 2: Mouse embryonic brain expression patterns of MYT family transcription factors

(A) Quantitative RT-PCR summarized as relative mRNA expression of Myt1 (red), Myt1l (blue), and Myt3 (green) in the developing mouse from E10.5 to adult, adapted from Matsushita et al. 2014. (**B)** Color coded summary of published in situ hybridization data from Matsushita et al. showing the spatial expression pattern of MYT1, MYT1L, and MYT3 in the developing cortex. (**C)** The diagram shows a hypothesized mechanism of microcephaly in Myt1l mutant mice at E14. APa, archipallium; BG, basal ganglia; CTX, cortex; DTe, dorsal telencephalon; fIC, fibers of the internal capsule; HC, hippocampus; HT, hypothalamus; IC, internal capsule; LGE, lateral ganglionic eminence; MGE, medial ganglionic eminence; OpV, optic vesicle; Pal, pallidum; POA, preoptic area; Str, striatum; TH, thalamus; Vg, trigeminal ganglion; VTe, ventral telencephalon.

1.5.3 Is MYT1L an activator or a repressor?

Once MYT1L binds to DNA, whether it functions as a transcriptional activator, repressor, or both, is still not clearly understood. In vitro transdifferentiation studies have represented MYT1L as a repressor of non-neuronal gene programs (Mall et al., 2017; Wapinski et al., 2013), while other in vivo studies have found evidence that MYT1L activates neuronal genes (Chen et al., 2021; Kepa et al., 2017). Early in vitro studies show that MYT1L was able to activate a hRARβ promoter-luciferase reporter as well as a Pit-1 enhancer/promoter luciferase reporter in CV-1 and

HeLa cells (Jiang et al., 1996). Furthermore, MYT1 and MYT1L were directly compared using an in vitro reporter with a synthetic promoter carrying seven copies of the AAAGTTT motif separated by nine nucleotides (Manukyan et al., 2018). In this assay, overexpression of full-length MYT1 repressed transcription while overexpression of full-length MYT1L activated transcription of the reporter in HeLa, A549, and U87 cells, which all have relatively low or no endogenous MYT1 and/or MYT1L expression. In cultured neuronal cells, shRNA-mediated knockdown of MYT1L resulted in reduced expression of neuronal transcripts associated with neurite outgrowth, axonal development, and synaptic transmission (Kepa et al., 2017). This is consistent with recent data from a germline MYT1L heterozygous mouse model showing increased expression of "early fetal" genes in prefrontal cortex of adult mice, resulting in an immature transcriptional signature compared to wild-type (WT) mice (Chen et al., 2021).

MYT1L also contains a repressive MYT1 domain. Compared to the N-terminal activation domain, the MYT1 domain appears to be highest conserved region second to the middle and C-terminal zinc fingers, containing the Ser/Thr-rich region in MYT1 and MYT1L (Jiang et al., 1996; Mall et al., 2017) (**Figure 1A**), and appears repressive in most studies so far. Mechanistically, through a yeast-two-hybrid screen, the central domain of MYT1 was shown to interact with the corepressor SIN3B. Since this region is conserved across the MYT family, it was also shown that MYT1L interacted with SIN3B using a Gal4 assay (Romm et al., 2005), and other studies have supported the conclusion that the central, MYT1 domain can interact with the corepressor SIN3B (Mall et al., 2017). Specifically, the interaction between MYT1 and MYT1L with SIN3B can result in transcriptional repression via histone deacetylase (HDAC) interaction with SIN3B (Romm et al., 2005). When directed to promoter regions by MYT1 and MYT1L, the SIN3B-HDAC complex can remove activating chromatin modifications, resulting in less accessibility and ultimately, repression (Romm et al., 2005).

The seemingly divergent functions of the activating N-terminal domain and repressive MYT1 domain make it challenging to classify MYT1L as a transcriptional activator or repressor. Altogether, these focused studies on the molecular domains of MYT1L suggest that the role of MYT1L is context dependent and may largely function as an activator in vivo. Follow-up studies are needed to determine if the role of MYT1L remains the same in adulthood after neurodevelopment has been completed.

To analyze the molecular and cellular role of MYT1L during neurodevelopment, a detailed time-course analysis of chromatin accessibility and TF binding is required. Single-cell/nuclei technologies can be leveraged to identify the cis regulatory landscapes and trajectories of the different cell types that make up the brain (Preissl et al., 2018). This general approach can be used with the MYT1L Syndrome mouse model to map altered gene regulatory programs and resulting impact on cellular proportions upon loss of MYT1L. Traditional methods to assay the TF activation or repression utilize fluorescence or luciferase-based reporter constructs for a quantitative readout of downstream activities. While these are highly sensitive and reproducible, they are not suitable for high-throughput screening of hundreds of putative regulatory elements. Massively parallel reporter assays (MPRAs) are an approach that can be used to test the cis regulatory function of thousands of DNA sequences in one experiment and can be deployed in vivo in a cell-type-specific manner (Lagunas et al., 2023). The main limitation is that these ~150 bp synthetic libraries are taken out of their original context, so additional validation experiments are necessary. Looking at chromatin accessibility and MYT1L TF binding together with functional assays could provide insight into the context-dependent role of MYT1L as an activator and/or repressor.

1.6 Mouse model of MYT1L syndrome

In vitro studies and analyses have significantly advanced our understanding of MYT1L's role in neuronal identity and maturation, but they are limited in their ability to replicate the intricate and dynamic biological processes that occur within living tissues. Therefore, developing a mouse model of MYT1L is critical to deepen our knowledge of its cellular and molecular functions in a physiological context. This model facilitates the study of MYT1L within living, developing, and interacting neural networks. This enables investigations to understand its contribution to neural differentiation and maturation by studying the effects of MYT1L haploinsufficiency directly. Therefore, we engineered a mouse model that harbors a mutation in exon7 (chr12:29849338, c.3035dupG, S710fsX), analogous to a human MYT1L patient mutation found in exon10 (Chen et al., 2021).

Analyses indicate that while homozygous knockout (KO) embryos are viable, they do not survive postnatally. In these KO mice, MYT1L transcripts and proteins are undetectable, whereas heterozygous (Het) mice exhibit a 25% reduction in MYT1L levels. This haploinsufficiency manifests in phenotypes such as obesity, reduced white matter volume, and microcephaly (Chen et al., 2021). Transcriptomic studies reveal that neural progenitors in these mice differentiate prematurely, leading to an immature transcriptional and chromatin landscape. These deficiencies resulted in behavioral abnormalities including hyperactivity, decreased sociability, and muscle weakness. Notably, these phenotypes were more pronounced in males, recapitulating the sex ratio bias observed in autism and intellectual disability.

Through this mouse model, we can investigate MYT1L's mechanisms and roles within a physiological framework, addressing critical questions about the vulnerability of specific cell types to MYT1L deficiency and identifying the cellular basis for the observed clinical phenotypes.

1.7 Conclusions

The body of concepts reviewed here illuminates the essential role of MYT1L in neurodevelopment, underscoring its multifaceted functions in guiding differentiation and maturation. The epigenetic landscape is critical in regulating these processes, reflecting the complexity of brain development. The MYT1L syndrome mouse model is invaluable for elucidating MYT1L's cellular and molecular function in vivo, offering a more dynamic and contextualized assessment than from in vitro studies alone. This model not only deepens our understanding of MYT1L's biological roles, but also paves the way for future explorations into interventions that target MYT1L-mediated pathways.

Over the last decade, there has been a surge of research productivity investigating the role of the epigenetic landscape in various fields. This reflects the wealth of tools available for profiling histone modifications, DNA-binding protein occupancy, and chromatin states in the current genomic era. While traditional genomic and epigenomic assays offer detailed profiles, they are snapshot methods that capture only the cell state at a moment in time when the sample was harvested. Calling Cards technology seeks to address this limitation by recording protein-DNA interactions over time, linking past interactions to future cellular outcomes. This technology is particularly valuable for studying developmental processes or other dynamic processes.

This body of work attempts to serve a dual purpose: firstly, to generalize the use of Calling Cards technology, making it accessible and applicable to researchers across disciplines, not just those specialized in genomics. Secondly, to harness this technology to advance our understanding of how developmental processes are programmed into the genome, and how their disruption can result in neurodevelopmental disorders such as MYT1L syndrome. These advancements and their applications not only address current key questions in the field today, but also enable us to ask new questions tomorrow.

Chapter 2: The complete experimentation guide for bulk Calling Cards

2.1 Preface

This chapter contains contents from the following published manuscripts:

Calling Cards: A customizable platform to longitudinally record protein-DNA interactions over time in cells and tissues.

Allen Yen, Chase Mateusiak, Simona Sarafinovska, Mariam A. Gachechiladze, Juanru Guo, Xuhua Chen, Arnav Moudgil, Alex J. Cammack, Jessica Hoisington-Lopez, MariaLynn Crosby, Michael R. Brent, Robi D. Mitra, Joseph D. Dougherty

Current Protocols, 3: e928. September 27, 2023

Measuring transcription factor binding and gene expression using self-reporting transposon calling cards and transcriptomes.

Matthew Lalli, Allen Yen, Urvashi Thopte, Fengping Dong, Arnav Moudgil, Xuhua Chen, Jeffrey Milbrandt, Joseph D. Dougherty, Robi D. Mitra

NAR Genomics and Bioinformatics, Volume 4, Issue 3, August 31, 2022.

2.2 Abstract

Calling Cards is a platform technology to record a cumulative history of transient protein-DNA interactions in the genome of genetically targeted cell types. The record of these interactions is recovered by next generation sequencing. Compared to other genomic assays, whose readout provides a snapshot at the time of harvest, Calling Cards enables correlation of historical molecular states to eventual outcomes or phenotypes. To achieve this, Calling Cards uses the *piggyBac* transposase to insert self-reporting transposon (SRT) “Calling Cards” into the genome, leaving permanent marks at interaction sites. Calling Cards can be deployed in a variety of *in vitro* and *in vivo* biological systems to study gene regulatory networks involved in development, aging, and disease. Out of the box, it assesses enhancer usage but can be adapted to profile specific transcription factor binding with custom transcription factor (TF)-*piggyBac* fusion proteins. The Calling Cards workflow has five main stages: delivery of Calling Cards reagents, sample preparation, library preparation, sequencing, and data analysis. Here, we first present a comprehensive guide for experimental design, reagent selection, and optional customization of the platform to study additional TFs. Then, we provide an updated protocol for the five steps, using reagents that improve throughput and decrease costs, including an overview of a newly deployed computational pipeline. This protocol is designed for users with basic molecular biology experience to process samples into sequencing libraries in 1-2 days. Familiarity with bioinformatic analysis and command line tools is required to set up the pipeline in a high-performance computing environment and to conduct downstream analyses.

2.2 Introduction

Transcription factors (TFs) and DNA regulatory elements interact to drive proper spatial and temporal patterns of gene expression. Transcriptional dysregulation caused by mutations in TFs or regulatory elements can result in disease (reviewed in (Chatterjee and Ahituv, 2017; Lee and Young, 2013)). In addition, transcriptional and epigenetic changes are often studied to understand disease processes, even when the disease is driven by other causes. Next generation sequencing has enabled genome-wide analysis of protein-DNA interactions by chromatin immunoprecipitation followed by sequencing (ChIP-seq), however, ChIP-seq requires high quality antibodies, relatively large amounts of starting material, and only provides a snapshot of states at the time of harvest. Thus, without performing a time course experiment, one cannot attribute historical molecular events with eventual cell states. Furthermore, the need for large amounts of starting material precludes the widespread use of these approaches in specific cell types in complex tissues. Newer immunotethering approaches such as CUT&RUN and CUT&TAG enable experiments with less input material and lower sequencing depths, but the dependency on antibodies remains and is not easily adaptable to query targeted cell populations. To address these limitations, we developed Calling Cards, a customizable platform that records protein-DNA interactions over time using a genetically encoded system. Calling Cards can be adapted and deployed in cell lines and tissues across different biological contexts, without the use of antibodies, and in specific, genetically targeted cell types.

Calling Cards relies upon two key components: a TF-*piggyBac* transposase fusion and a self-reporting transposon (SRT), which is a *piggyBac* transposon that contains a tdTomato reporter. When tethered to a TF, the *piggyBac* transposase inserts SRTs into the genome near TF binding sites, leaving permanent “Calling Cards”, which can then be recovered by sequencing and mapped with base pair resolution (**Figure 3A**). Use of the hyperactive *piggyBac* (hyPB) increased the overall number of transposition events while maintaining a similar insertion pattern (Moudgil et al., 2020b; Yusa et al., 2011). This accumulation of Calling Cards insertions provides a cumulative recording of TF binding over the assayed period. Examples of Calling Cards data are shown in **Figure 3B-F**. A TF of interest can be assayed through cloning TF-transposase fusion proteins. As an alternative to using TF-fusions, the naive *piggyBac* transposase can be leveraged for its natural affinity for BRD4, a TF that recognizes acetylated lysine residues and found to be highly enriched in super enhancers (Yoshida et al., 2017), and shown to be important in driving transcription of genes that define cell identity (Wang et al., 2012, 2008). Thus, unfused *piggyBac* can be used to record BRD4-bound enhancer usage.

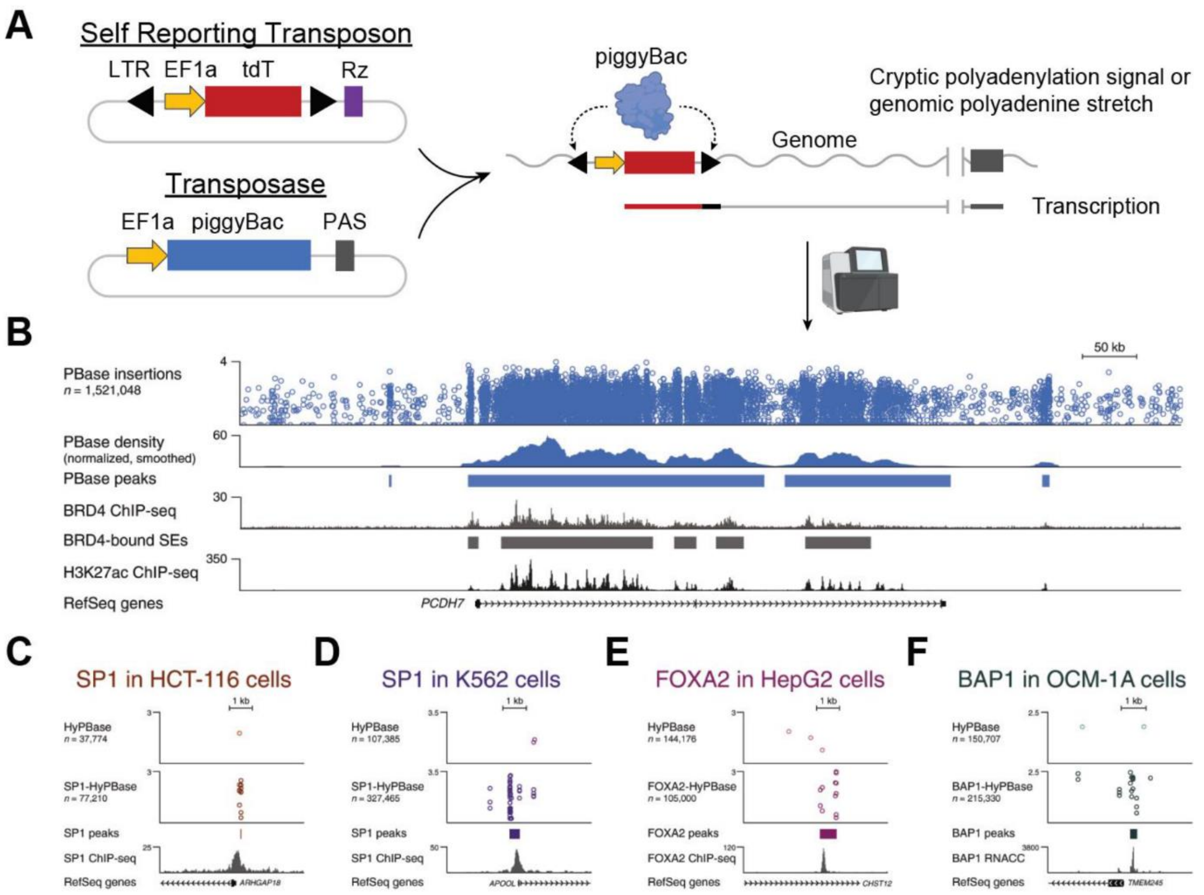


Figure 3: Example tracks showing recording of BRD4-bound super enhancers and TF binding sites using Calling Cards

(A) Diagram of the self reporting transposon (SRT) and *piggyBac* transposase constructs. When expressed in cells, the *piggyBac* transposase inserts the SRT into the genome at sites of protein-DNA interaction leaving a permanent mark, or Calling Card. The location of Calling Cards insertions can be recovered through RNA sequencing. **(B)** The top track shows the genomic locations of SRT insertions in cells transfected with Calling Cards at the PCDH7 locus. The normalized density of Calling Cards correlates with BRD4 and H3K27ac ChIP-seq peaks. **(C-F)** Fusion of hyPB with a variety of TFs works to redirect Calling Cards across different cell lines. This figure is adapted and reprinted from (Moudgil et al., 2020b) with permission from Elsevier.

While Calling Cards can be thought of as an alternative to CUT&RUN or ChIP-seq, its recording feature also enables additional kinds of experimental questions. First, it can cumulatively record protein-DNA interactions over time, providing an integrated snapshot that could replace time series data. Second, it can be used to correlate early enhancer usage to later cell fate decisions.

Some examples include: how can a seemingly homogenous cluster of pluripotent stem cells give rise to many different cell types? How can genetically identical organisms have distinct biological responses to the same stimulus? Current standard genomic technologies and assays are inherently destructive since to analyze the molecular state, the harvested cells are destroyed. By recording protein-DNA interactions over time, historical molecular events such as transcription factor binding, or historical epigenetic states can be linked to current cell states. Linking recorded molecular states to eventual outcomes could be broadly applicable to many areas of research, including but not limited to developmental biology, aging, and gene-environment interactions.

In this protocol, we provide a resource to guide researchers, especially those new to genomic assays, to design and execute a Calling Cards experiment. We have created a streamlined workflow (**Figure 4**) to simplify the selection of reagents needed to perform the desired experiment (**Figure 5**), discuss methods to validate the constructs, and provide recommendations for reagent delivery methods with suggested controls. Specifically, Basic Protocol 1 describes how to create a plasmid pool of barcoded SRTs that can be used for *in vitro* transfections or AAV packaging for *in vivo* transductions. It also describes intracerebroventricular injections, a relatively simple and robust method to reliably deliver AAVs directly into the cerebral lateral ventricles and CNS in early postnatal mice. Support Protocol 1 describes an important step to validate the barcode distribution of the plasmid or AAV pool using next-generation sequencing. Next, Basic Protocol 2 outlines the steps to harvest RNA and perform first-strand synthesis from Calling Cards-containing samples. Support Protocol 2 describes an optional, but recommended, step to perform the first quality control assessment of samples by analyzing the abundance of tdTomato-containing SRT transcripts by qPCR. This determines if samples should or should not be carried through the remainder of the protocol. This step can minimize unnecessary labor and usage of reagents. Basic

Protocol 3 details PCR amplification of Calling Cards transcripts, bead cleanup, tagmentation, indexing, and final bead cleanup. The tagmentation step is required to cleave long fragments into smaller sizes so they are compatible with short-read sequencing platforms. The indexing PCR adds on a unique sequence to each library and enables multiple samples to be pooled and sequenced together on the same run. Additionally, there are quality control (QC) steps built into the protocol after each Basic Protocol to monitor progress through the procedure. Basic Protocol 4 describes library pooling strategies and recommended parameters for short-read next generation sequencing platforms. Lastly, Basic Protocol 5 provides a high-level guide on using the Nextflow Calling Cards bioinformatic pipeline to prepare the raw sequencing data into a format that can be used for downstream analysis.

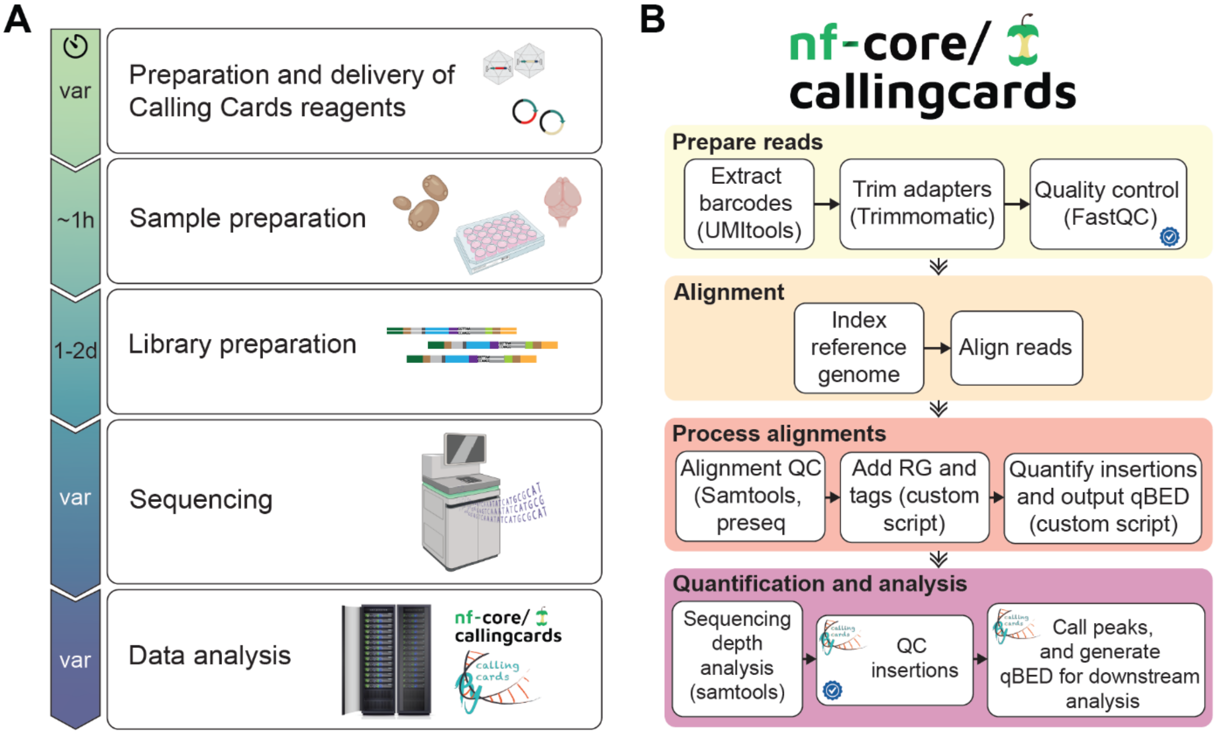

A
var
~1h
1-2d
var
var
Preparation and delivery of Calling Cards reagents
Sample preparation
Library preparation
Sequencing
Data analysis
nf-core/ callingcards
calling cards
B
nf-core/ callingcards
Prepare reads
Extract barcodes (UMItools)
Trim adapters (Trimmomatic)
Quality control (FastQC)
Alignment
Index reference genome
Align reads
Process alignments
Alignment QC (Samtools, preseq
Add RG and tags (custom script)
Quantify insertions and output qBED (custom script)
Quantification and analysis
Sequencing depth analysis (samtools)
QC insertions
Call peaks, and generate qBED for downstream analysis

Figure 4: General wet lab and computational workflows of a Calling Cards experiment

(A) The wet lab protocol is split into five main stages: 1) the viral or plasmid Calling Cards reagents are prepared and delivered into the target cells/tissue; 2) The sample is harvested; 3) the sequencing libraries are prepared; 4) the libraries are sequenced on Illumina NGS platforms; and 5) the generated FASTQ files are processed through the Calling Cards Nextflow pipeline and other downstream computational softwares. **(B)** The computational pipeline is distributed as a self-contained package that will process FASTQ files to Calling Card qBED files. The pipeline is divided into four main chunks: 1) the reads are prepared by extracting sample barcodes, trimming Illumina adapters, and standard quality control; 2) the reads are aligned to a reference genome; 3) the alignments undergo standard quality control and sample barcodes are added to headers of each read then collated into a qBED file; 4) the output files can be used for downstream analysis such as differential peak analysis and motif enrichment analysis. The blue check mark represents steps where QC metrics will be written to a file in the output and analysis directory.

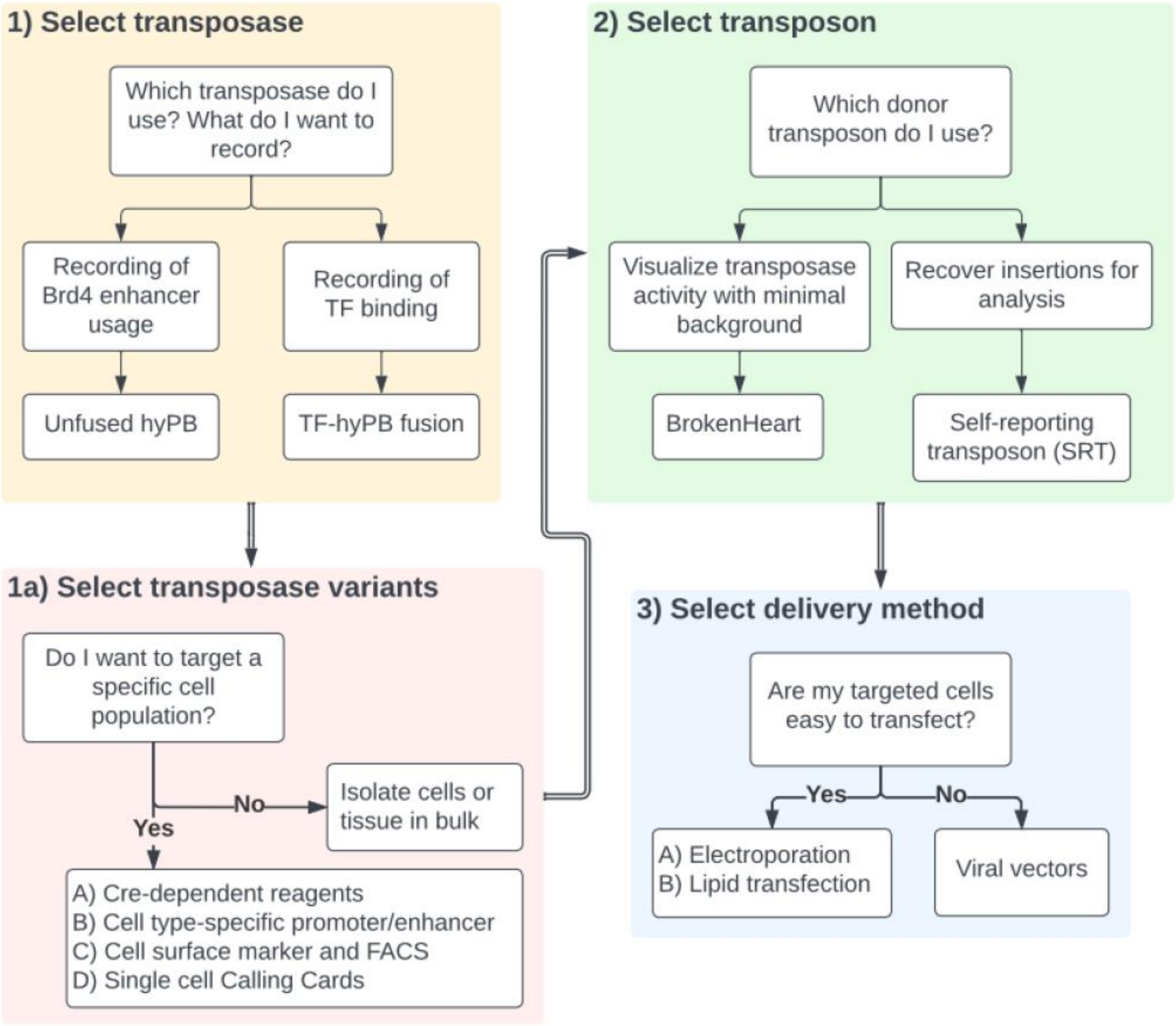

1) Select transposase
Which transposase do I use? What do I want to record?
Recording of Brd4 enhancer usage
Recording of TF binding
Unfused hyPB
TF-hyPB fusion
2) Select transposon
Which donor transposon do I use?
Visualize transposase activity with minimal background
Recover insertions for analysis
BrokenHeart
Self-reporting transposon (SRT)
1a) Select transposase variants
Do I want to target a specific cell population?
No
Yes
Isolate cells or tissue in bulk
A) Cre-dependent reagents
B) Cell type-specific promoter/enhancer
C) Cell surface marker and FACS
D) Single cell Calling Cards
3) Select delivery method
Are my targeted cells easy to transfect?
Yes
No
A) Electroporation
B) Lipid transfection
Viral vectors

Figure 5: Decision tree for selecting Calling Cards reagents for desired readouts

There are various transposase and donor transposon variants depending on the biological question and goal. The first decision is to decide between using an unfused or TF-fused transposase (step 1). If Calling Cards recording is desired in a genetically defined cell population, "FrontFlip-hyPB" Cre-dependent transposase options are available (step 1a.A). Alternatively, a cell type-specific promoter can be used to drive expression of hyPB (e.g. Nestin-hyPB to target neural progenitors) (step 1a.B). A constitutive hyPB can be used for ubiquitous expression followed by enrichment of target cell population by FACS (step 1a.C). The final option is to conduct a single cell Calling Cards experiment (step 1a.D; see (Moudgil et al., 2020b) for details). The decision of donor transposon is made in step 2, followed by delivery method in step 3.

2.3 Strategic planning

Expertise needed to implement the protocol

Basic molecular biology skills are required to successfully perform this protocol. If performing *in vivo* experiments, basic animal handling and husbandry skills are also necessary. The use of viral vectors requires proper safety training and laboratory approval according to institutional guidelines. For those new to preparing sequencing libraries, it may be beneficial to consult a sequencing center or service provider for optimal design and ordering of indexing primers as described in the section "Consideration for primer selection and ordering for sequencing libraries." The sequencing of libraries requires the use of Illumina high throughput sequencing instruments that are typically found within genomics core facilities or commercial fee-for-service sequencing companies. For data analysis, a Linux based high-performance computing environment or small dedicated server is necessary for computationally intensive tasks. If one is not available, pay-as-you-go cloud computing platforms such as Amazon Web Services (AWS), Google Cloud, or Microsoft Azure can be used. Proper setup of the computational environment benefits from familiarity with Nextflow and container runtimes such as Charliecloud, Singularity, or Docker. Basic to moderate skills with computational and bioinformatic analysis using command line tools, packages, and job schedulers is required. If not available within the lab, this computational expertise, similar to what would be required for RNA-seq, ChIP-seq or ATAC-seq workflows, may be available through local genomics cores that provide sequencing services.

Design of custom TF-hyPB fusion proteins (optional)

Calling Cards requires the presence of two components within a cell: the hyperactive *piggyBac* (hyPB) transposase and the donor transposon. Using Calling Cards "out of the box" with unfused wild-type hyPB can identify BRD4-bound super enhancers. By creating a TF-hyPB fusion, one can redirect the insertion of Calling Cards to specific TF binding sites. We have previously used this for several TFs in yeast (e.g., Gal4, Gal80, Ste12, Bas1, Pho2, Gcn4, and Pho4) (Wang et al., 2007) and vertebrate systems *in vitro* (e.g., SP1, FOXA2, BAP1, ASCL1, MYOD1, NEUROD2, and NGN1) (Cammack et al., 2020; Lalli et al., 2022; Moudgil et al., 2020b; Yen et al., 2018) and *in vivo* (e.g., SP1) (Cammack et al., 2020). The general workflow to design, create, and validate a TF-hyPB fusion is described in **Figure 6**.

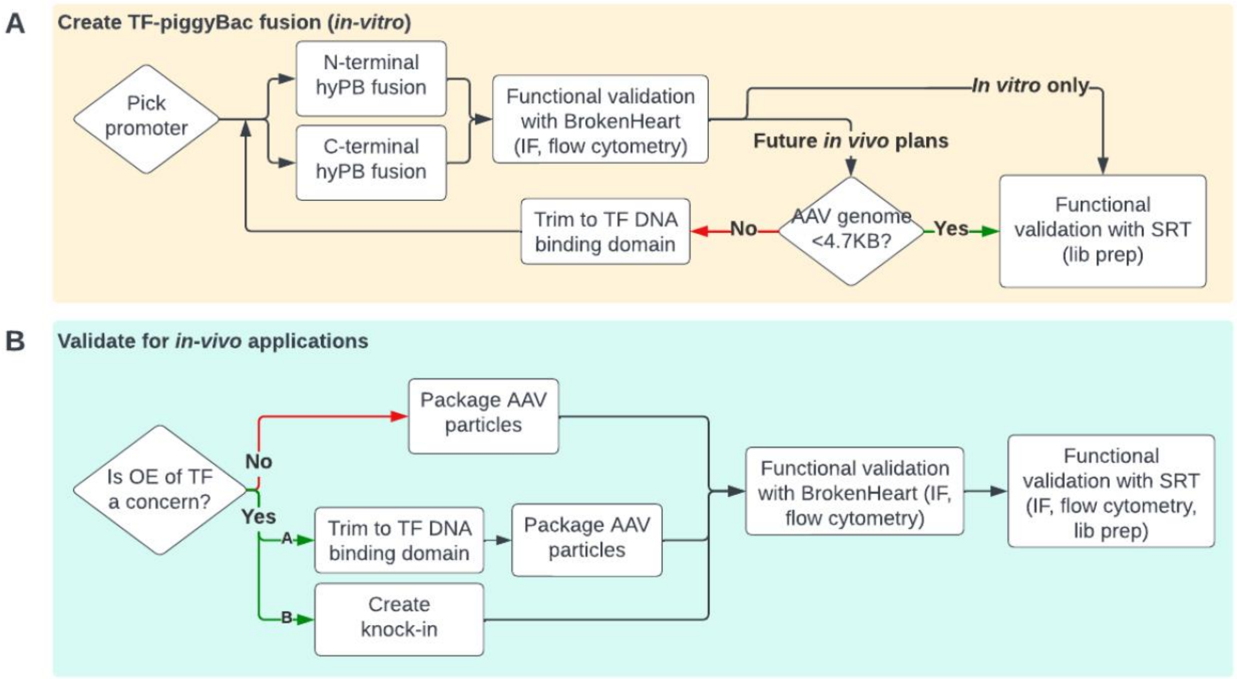

A
Create TF-piggyBac fusion (in-vitro)
Pick promoter
N-terminal hyPB fusion
C-terminal hyPB fusion
Functional validation with BrokenHeart (IF, flow cytometry)
In vitro only
Future in vivo plans
AAV genome <4.7KB?
No
Yes
Trim to TF DNA binding domain
Functional validation with SRT (lib prep)
B
Validate for in-vivo applications
Is OE of TF a concern?
No
Yes
A
B
Package AAV particles
Trim to TF DNA binding domain
Package AAV particles
Create knock-in
Functional validation with BrokenHeart (IF, flow cytometry)
Functional validation with SRT (IF, flow cytometry, lib prep)

Figure 6: General workflows for creating TF-hyPB fusions

(A) Steps to create a TF-hyPB fusion construct. Functional validation with immunofluorescence or flow cytometry is recommended to be performed using the BrokenHeart donor transposon due to its complete absence of fluorescence background (**Supplemental Figure 1C-E**), compared to the minimal background of the SRT. Final functional validation is performed using the SRT to generate libraries for downstream analysis. **(B)** Additional considerations for *in vivo* applications.

The first step is to select a suitable promoter for your targeted application. Ubiquitous promoters such as EF1a (X. Wang et al., 2017), CMV early enhancer-chicken β-actin (CAG)

(Alexopoulou et al., 2008), and phosphoglycerate kinase 1 (PGK) (McBurney et al., 1991) have been successfully used to drive expression of the transposase. Next, clone both an N- and C-terminal TF-hyPB fusion construct. As TFs have diverse binding modes, having both versions will allow empirical determination of which fusion has the most efficient transposase activity, yet maintains specificity for the TF-targeted motif. Once the completed expression construct is cloned, it is important to sequence the entire plasmid to ensure that all elements are intact, especially the repetitive AAV ITRs which are prone to deletions. Low cost, commercial long read sequencing (e.g., Plasmidsaurus or Genewiz) can be a useful resource to sequence the entire plasmid without primer walking, to determine if multiple plasmids species are within the submitted sample, and to resolve repetitive regions such as AAV ITRs that are often difficult for traditional Sanger sequencing.

Once the sequence is confirmed, the TF-hyPB fusion can be functionally validated by transfection with the BrokenHeart plasmid (BrokenHeart) *in vitro* into an easily transfectable cell line such as HEK293 or mouse neuroblastoma N2A cells. BrokenHeart is a tdTomato reporter that is interrupted, or "broken", with a donor transposon (**Supplemental Figure 1**). Without transposase activity, the expressed tdTomato protein would be not functional and no fluorescence would be observed. In the presence of transposase activity, the transposase will excise the transposon within BrokenHeart and the tdTomato coding sequence is restored and the functional protein will be expressed and fluoresce. (**Supplemental Figure 1C-E**). This can be used as a validation transposon to screen and confirm the function and activity of custom TF-hyPB fusions because of its low background fluorescence compared to SRTs (**Supplemental Figure 1F-H**), however, insertions cannot be recovered from RNA. To recover Calling Cards insertions from RNA, the SRT will have to be used (**Figure 5**, **Supplemental Figure 1I,J**). N- and C- terminal fusions can be quantitatively compared by microscopy or FACS to determine the relative activity compared to an unfused hyPB. In general, we see that fusions to any TF reduce transposase activity, yet this reduced activity is still sufficient to mediate transposition. This suggests that transposon levels, rather than transposase activity, is typically the rate limiting process (Nakazawa et al., 2009; Wu et al., 2006), and this is consistent with our observations of various donor/transposon ratios *in vitro* (**Supplemental Figure 2**)

If future *in vivo* experiments with AAVs are desired, it will be important to ensure that the length of the AAV transfer genome (ITR to ITR) is less than ~4.7kB, the maximum cargo size of AAV particles for efficient viral packaging (Wu et al., 2010). If the sequence is larger, steps to trim away bases that are not critical for TF DNA binding will be necessary and functional validation should be redone to confirm TF-hyPB fusion activity. Lentiviral vectors with larger packaging capacities have also been used to deliver TF-hyPB fusions *in vivo* with limited success, likely due to the more restricted spread of lentivirus in the brain.

Finally, after BrokenHeart validation, functional validation recovering insertion sites with SRTs will reveal if TF-hyPB fusion directs Calling Cards insertions at expected TF binding sites to confirm that fusion of hyPB does not alter the specificity or TF binding properties. If available, TF ChIP-seq data from the same cell lines can provide a benchmark to validate that the TF-hyPB fusion is functioning as expected. If not, detection of the TF's canonical motif using tools such as HOMER (Heinz et al., 2010) or the MEME suite (Bailey et al., 2015) can also indicate on-target activity. Likewise, comparison to profiles from unfused hyPB can confirm redirection of binding. After the construct passes all QC and functional validation steps described above, then it can be packaged into AAV particles and injected *in vivo* into the target tissue or experimental system of choice.

A potential concern when expressing the Calling Cards reagents into cells through transfection or transduction is that the TF-hyPB fusion protein, and inadvertently the TF itself, is expressed above endogenous levels, and could thus alter gene expression. If overexpression is a concern, there are multiple approaches that can be used. The first (**Figure 6B**), is to trim to the minimal TF DNA binding domain, as mentioned for reducing size. Removal of other effector domains may render the TF-hyPB fusion sufficient to bind DNA but not interact with co-factors, thus minimizing effects of overexpression. Another approach is to create a knock-in hyPB as a fusion to the endogenous TF locus of a cell or mouse line. For all approaches, functional validation with BrokenHeart and SRT should be carried out at an appropriate time point, allowing enough time for the AAV to mature and Calling Cards reagents to express at high levels. In brain tissue, AAVs are often allowed to mature 10+ days, although we have been able to recover sufficient insertions as early as 2 days after injection *in vivo* (**Figure 7A-C**). AAV maturation can be tracked over time with tdTomato RT-qPCR, which correlates with insertion density (**Figure 7C-E).**

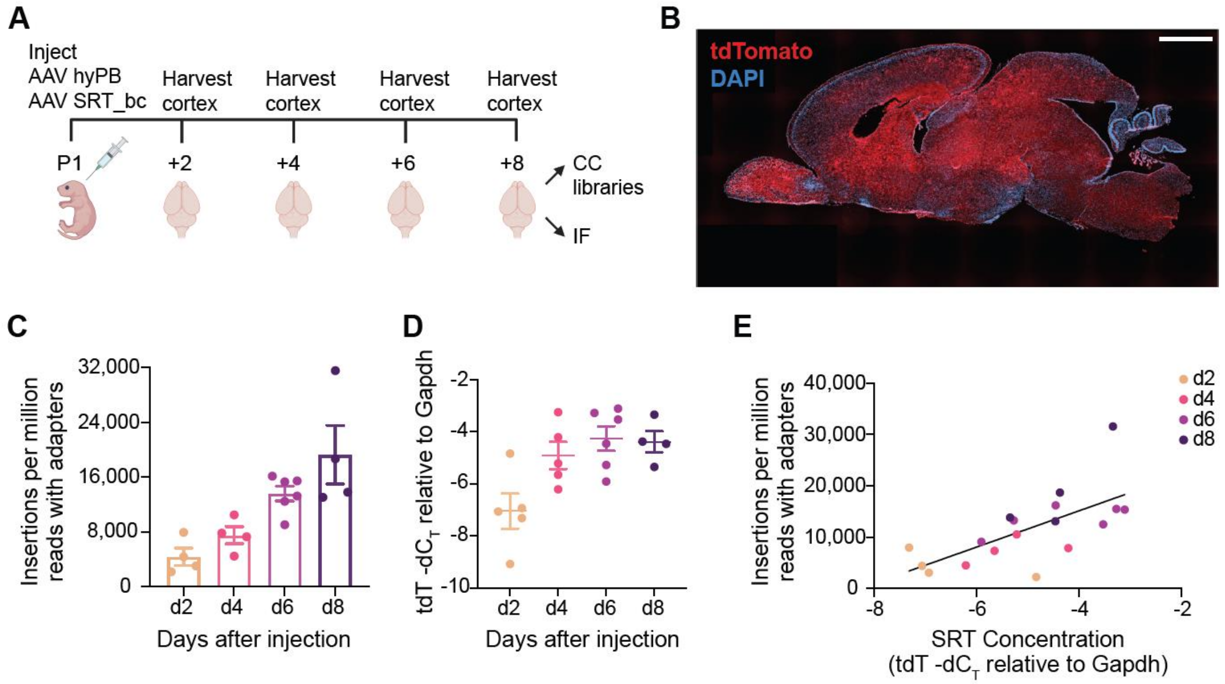


Figure 7: Timeline of Calling Cards activity in the mouse brain after AAV delivery

(A) Schematic of AAV Calling Cards time course experimental design. **(B)** Sagittal section of a brain harvested 2 days after neonatal intracerebroventricular injection with Calling Cards reagents, showing widespread expression of SRT-derived tdTomato. Scale bar 1 mm. **(C)** Insertion counts recovered at each time point, normalized to read depth. n = 4-6 hemi-cortices. **(D)** SRT concentration, measured by RT-qPCR as tdTomato -dC$_T$ relative to Gapdh, over time. **(E)** Insertion counts recovered as a function of SRT concentration. Simple linear regression, $R^2 = 0.4447$, $p = 0.0025$.

Considerations for primer selection and ordering for sequencing libraries

In the final steps of library preparation, a dual indexing strategy is used to multiplex and increase the number of samples that can be sequenced per run to decrease costs. Here, we refer to "indexes" as sequences within Illumina adapters that are sequenced independently of the standard Read1/Read2 (R1/R2) sequences, while "barcodes" are within the R1/R2 reads. The OM-PB primer contains the Illumina P5 adapter, Index2/i5, TruSeq Read1, 3 bp primer barcode, and partial *piggyBac* LTR (see **Figure 8B**, **Supplemental Figure 3**). It is ~100 bp long and can be cost prohibitive to synthesize an OM-PB primer with unique Index2/i5s for each sample. Instead, we have found it most effective to assign unique Index1/i7 indexes to each sample and have a small collection of up to 8 OM-PB primers with the same Index2/i5 yet varying the 3bp primer barcode with a hamming distance of 2 (GCA, ATC, CTA, ACG, CGT, TGC, GAT, and TAG). If more OM-PB primers are needed, additional combinations can be created by switching the Index2/i5. Note that correctly assigning these indexes and primer barcodes is critical to sample demultiplexing. If needed, consult your sequencing core for suggestions on designing optimal primers for multiplexing and compatibility with their workflow. These are separate entities from SRT barcodes, which are used to identify unique insertions in the same locus within each sample.

While there are many strategies to allocate indexes and barcodes, one potential approach is to designate an Index2/i5 to an experiment, primer barcodes to different animals within that experiment (biological replicates), and Index1/i7 as different tissue samples or technical replicates for each animal, if multiple samples are prepared.

We have found Calling Cards to be reproducible across biological replicates as shown by the high correlation of normalized insertions per million (**Supplemental Figure 4**). Thus, for a given experiment, we typically create at least three replicate libraries per animal, with at least 3 animals per experimental group. When sequenced to saturation, this is typically sufficient to collectively recover at least 500k unique insertions per experimental group, which is a threshold we have found to be reliable for peak calling (**Supplemental Figure 11**). The additional benefit of uniquely indexed samples is that deeper sequencing of specific samples can be re-pooled and sequenced without worry of index/barcode clashing.

Estimated costs to perform a Calling Cards experiment

The cost of the experiment is driven by the complexity of the biological question, and the expected effect size. The complexity of the biological question will determine how many conditions, and thus the amount of reagents that are required for a successful project. Further, the expected effect size can be used to determine the number of replicates needed per condition, which will further influence cost. Estimating replicate numbers is best done via a power analysis, which involves determining the sample size needed to detect an estimated or observed effect size. Typically, more replicates are needed to be statistically powered to detect small differences (such as subtle changes in the same cell type) which would increase the project cost, while you may only need a few samples to reliably detect large differences, such as the differences between cell types or distinct TFs. Here, we provide the estimated cost breakdown per replicate and for a hypothetical *in vivo* experiment to look for large effects that uses 6 total samples (3 biological replicates across 2 conditions) in **Table 3.** This is simply meant to ensure that investigators are aware that the costs of a Calling Cards experiment are similar to other genomics experiments, and thus to carefully design cost-effective experiments.

Limitations

While Calling Cards enables the recording of protein-DNA interactions over time, the readout of the method provides a cumulative history of enhancer usage or TF binding and is unable to resolve the temporal order of insertions (e.g., which insertions occurred first vs. those that occurred at the end of the recording period). To obtain some time information, one could harvest samples in a time course and resolve unique Calling Cards insertion peaks by computationally subtracting common regions. This approach was used to map SP1 binding and expression of early and late genes in the developing mouse cortex (Cammack et al., 2020).

Another limitation is that current Calling Cards reagents record continuously from delivery of the viral reagent (hyPB and/or SRT) until sacrifice of the animal. Recording only during specified time points would require development of drug inducible transposases, which would open up additional experimental opportunities.

The protocol described here is based on starting from a total RNA sample derived from potentially many cells (a “bulk” sample), thus the Calling Cards data represent the average signal from all the different cell types expressing the transposase. If the sample is heterogeneous like the brain, the resulting data should be interpreted taking this into account: the bulk Calling Cards data represents the average insertions across all the cell types expressing transposase. However, Calling Cards is compatible with droplet-based microfluidic platforms such as 10x Genomics to identify enhancer usage or TF binding with single cell/nucleus resolution. This innovation circumvents the need for SRT barcodes, as the cell-barcodes inherent to the single cell platforms can serve this purpose. The protocol adaptations for these single-cell Calling Cards are covered in (Moudgil et al., 2020b). Of note, the Nextera Mate Pair Sample Prep Kit (Illumina FC-132-1001) used for single cell Calling Cards library preps has been discontinued by the manufacturer. An in-house protocol is being developed and will be published when completed, though interested readers can reach out to us sooner. An alternate non-single cell approach to measure Calling Cards from specific cell types is to use Cre-dependent reagents (Cammack et al., 2020).

Another limitation, based on observations from single cell Calling Cards data, is that the number of insertions per cell is relatively low (<100). While this decreases the potential deleterious effects of transposon insertions in key regulatory elements, since any given cell has few insertions, the recovery of Calling Cards insertions from rare cell types in sufficient numbers poses a challenge. The number of biological replicates needed to achieve a 500k unique insertion threshold may be a limitation and should be considered during experimental design.

The number of Calling Cards insertions is also dependent upon the delivery and expression of the SRT and *piggyBac* transposase. The copy number of SRTs is important in determining the success of a Calling Cards experiment. When transfecting cells with plasmids, one can begin with a 1:1 plasmid cocktail of SRT:transposase and can further optimize by increasing the ratio and amount of SRT (**Supplemental Figure 2**). The same applies to the AAV Calling Cards reagents as sufficient time is needed for the AAV to mature, express the transgenes, and functionally hop into the genome. Our preliminary data demonstrates functional hopping as early as 2 days after injection (**Figure 7**), though insertions increase with time as expected.

Basic Protocol 3 represents an optimized library preparation protocol (**Figure 8**) to enable robust recovery of Calling Cards insertions even when in lower abundance (e.g., from a relatively sparse cell type labeled by a Cre line), however, if extremely rare cell types are targeted using either a transgenic or molecular approach, FACS enrichment of TdTomato positive cells/nuclei prior to RNA extraction may be necessary to enrich for cells with Calling Cards insertions.

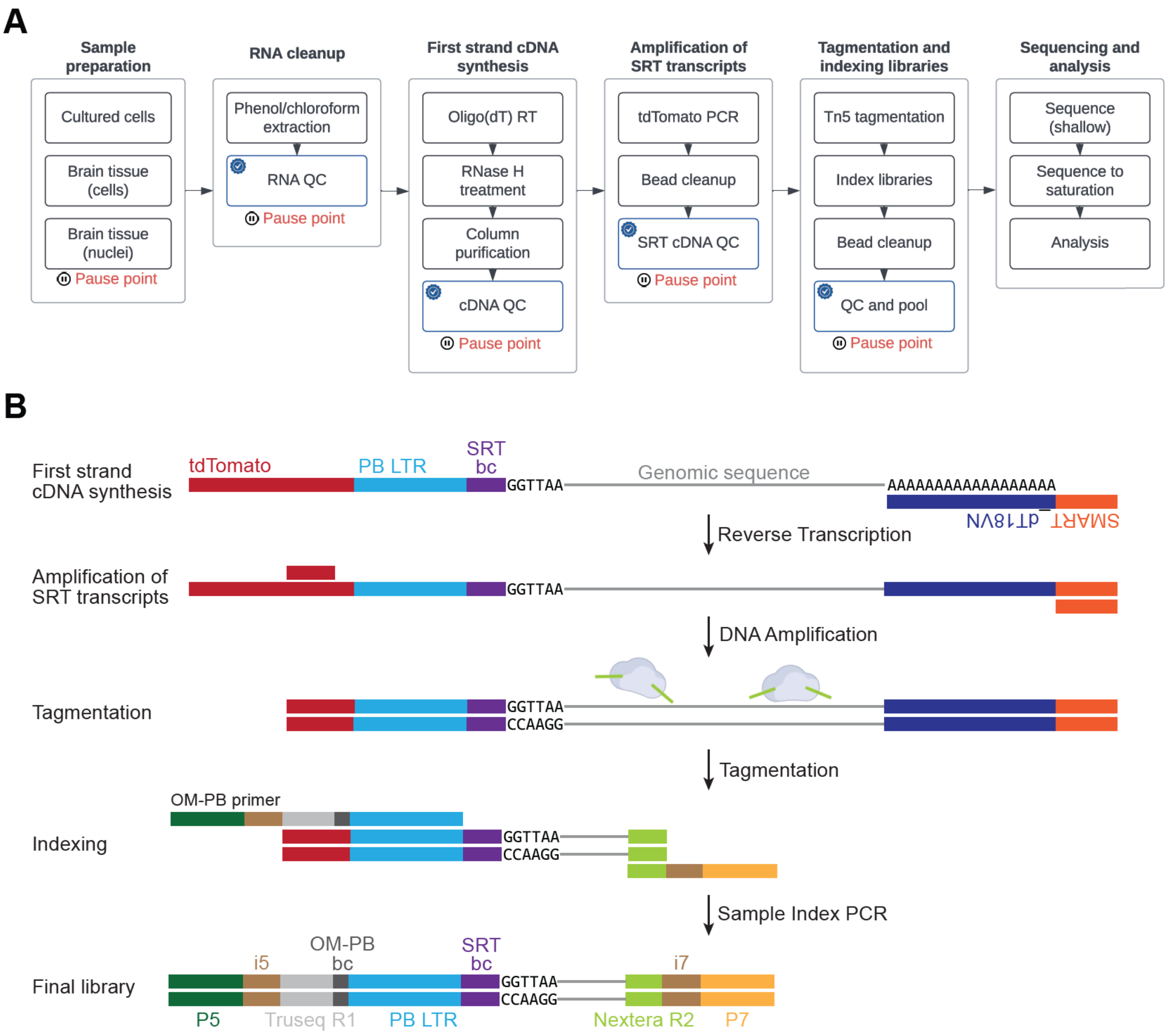


Figure 8: Experimental workflow for bulk Calling Cards library preparation

(A) The sequencing library preparation protocol is broken down into several main sections. Recommended quality control (QC) checkpoints are noted by the blue checkmark. Appropriate pause points are shown in red. **(B)** A cartoon depicting how the libraries are prepared and the final library structure that is loaded onto the sequencer.

2.4 Advancements in methodology

Optimized library preparation for in vivo tissues

The Calling Cards platform is a versatile tool designed to capture longitudinal TF-binding or BRD4 enhancer usage, exhibiting proven efficacy in vitro with cell lines such as HCT118 and K562, which can be transfected via electroporation. High expression levels of the transposase and SRT are crucial since the total number of insertions per cell is dependent on the expression of the transgenes. I theorize that enhancing SRT expression could further improve system efficiency. Analysis of Calling Cards data involves peak calling, which identifies genomic regions with significant enrichments of SRT insertions, then tests these candidate peaks for statistical significance. The resolution in detecting subtle peaks over background is directly associated with the number of insertions, emphasizing the need for a robust protocol to maximize the number of insertions and the recovery during the library preparation steps across a wide range of biological contexts.

The versatility of Calling Cards extends to in vivo applications using AAV vectors, although in vivo transgene expression is typically not as efficient as in vitro experiments. In general, transfection efficiencies in vitro can reach upwards of 90%, contrasting with 30-70% for AAV-mediated cortical labeling in vivo. AAVs are known to drive high transgene expression, however depending on the injection route and target brain region, the transgene expression levels are usually lower than those from in vitro experiments. Consequently, brain tissue samples may contain substantial amounts of non-SRT RNA, complicating library preparation due to excessive non-specific PCR templates. By using the in vitro library preparation directly with in vivo samples without modifications, 20-60% of the sequencing reads meeting filtering criteria and are thus removed from downstream analyses.

To mitigate background RNA interference, using additional starting material proved beneficial. By maximizing the input RNA of Maxima H Minus Reverse Transcriptase, which is capable of processing up to 5 µg of total RNA, enhanced library quality and usable read percentage. For experiments with limited starting material, an oligo-dT polyA RNA capture strategy yielded similar improvements.

Subsequent PCR amplification still encountered mispriming issues, potentially attributed to residual RNA and SMART_dT18VN primer interactions. Introducing a column purification step post-reverse transcription effectively removed potential PCR contaminants, ensuring that only purified single-stranded DNA was used as input for the PCR amplification of SRTs. Maximizing the PCR input to utilize up to 100 ng of the first strand synthesis product also enhanced the library quality.

These methodological refinements have significantly boosted in vivo library quality, with usable fragment rates now consistently exceeding 90%. This improvement not only reduces sequencing costs but also amplifies the recoverable insertion data per sample, marking a substantial advance in the methodology over the original protocol (**Figure 9**).

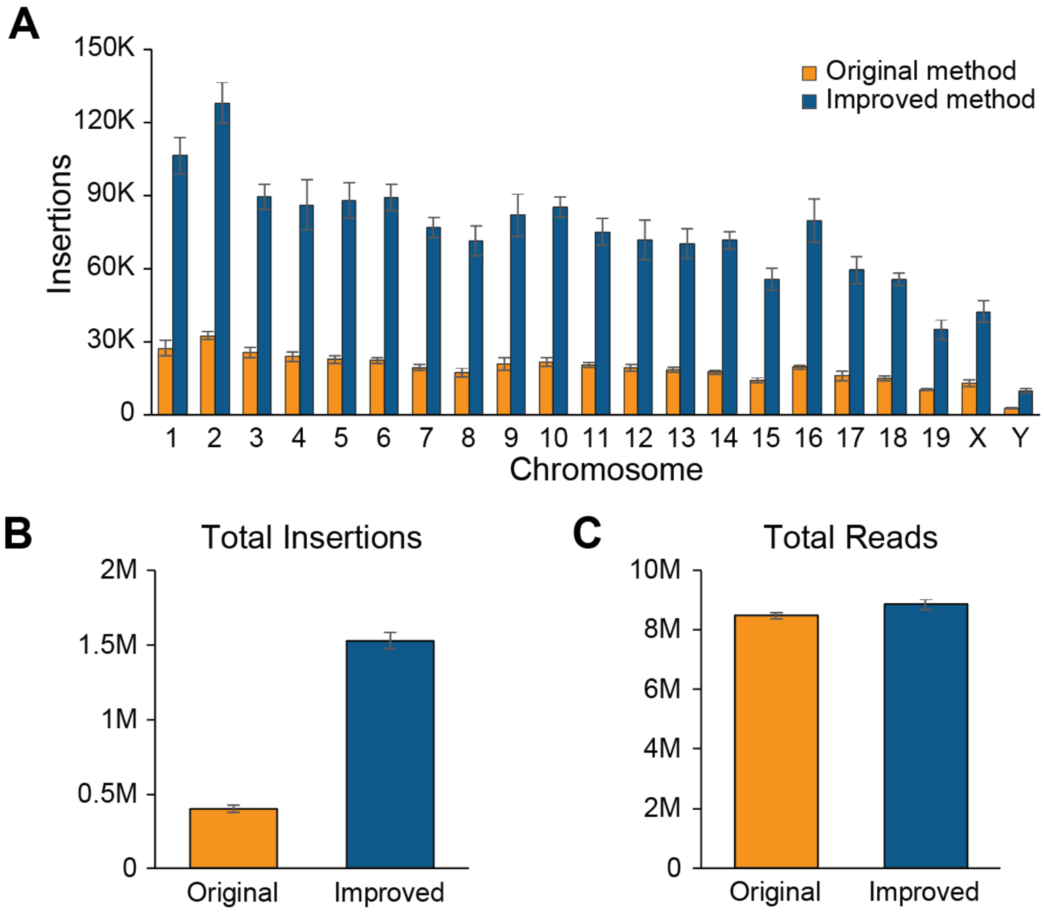


Figure 9: Improved library preparation protocol increases the number of recovered insertions

(A) Summary bar plot showing the distribution of recovered Calling Cards insertions per chromosome for the original method used in (Moudgil et al., 2020b) and (Cammack et al., 2020) in orange. The improved method (blue) shows increased recovery of insertions. **(B)** Bar graph shows an approximately 3-fold increase in recovery of insertions with similar level of sequencing depth. **(C)** Bar graph shows that the libraries prepared from the original method and improved method were sequenced to a similar sequencing depth.

Design and validation of barcoded Calling Cards reagents

Early protocols recovered inserted transposons from genomic DNA (Wang et al., 2012), but the advent of SRTs allows for the facile recovery of calling cards through RNA sequencing. RNA-mediated mapping of transposon insertions is more efficient than previous DNA-based protocols, and this protocol enables the simultaneous identification of TFBS and changes in gene expression in single cells (Moudgil et al., 2020b). However, in bulk experiments on populations of cells, the RNA-mediated protocol is technically cumbersome, requiring a large number of replicates to identify independent insertions in the same genomic locus.

Current implementations of the mammalian calling card protocol employ a hyper-active *piggyBac* transposase (Yusa et al., 2011). An inherent constraint of this transposase is its requirement for a ‘TTAA’ tetranucleotide sequence for transposon insertion. As a result, multiple independent calling card insertions often occur at the same genomic location in different cells. Since the identification of TF binding sites is based on transposition count rather than read density, if these independent insertions are not distinguished, it limits the dynamic range of bulk calling card experiments. Current best practices for in vivo Calling Cards experiments require a large number of biological replicates (typically 10) for each condition to increase the number of insertions that can be detected at a given TTAA location (Cammack et al., 2020). While this improves the quantitative readout of these experiments, experimental cost and labor scale linearly with the number of replicates. Therefore, as an alternative approach, we sought to embed a unique barcode within the terminal repeat (TR) of the self-reporting transposon, the best location to enable reliable recovery without barcode swapping. Doing so is challenging, however, because all published sequences of the *piggyBac* transposon TRs are completely invariant, indicating strong sequence constraints on TR function which might preclude barcode insertion (Li et al., 2005; Morellet et al., 2018; Solodushko et al., 2014; Y. Wang et al., 2017).

Here, we performed targeted mutagenesis of the *piggyBac* terminal repeat sequence to identify sites that could accommodate barcodes in calling card experiments. We discovered at least four consecutive nucleotides within the TR that were tolerant of a range of mutations without major reductions in transposition efficiency. As a resource to the scientific community, we have developed a set of barcoded *piggyBac* SRT plasmids and modified the calling card analysis software to utilize these barcodes.

To test whether barcoded SRTs can also function in vivo and reduce this need for technical replicates, we performed calling card experiments with barcoded and non-barcoded SRTs in the mouse cortex. We packaged tdTomato SRT plasmids with or without barcodes into AAVs and delivered them to cortex of mice as described (Cammack et al., 2020; Moudgil et al., 2020b). Unfused *piggyBac* has an insertion preference at super-enhancers which are a class of enhancers regulating genes linked to cell identity (Whyte et al., 2013; Yoshida et al., 2017). Leveraging this property, calling cards have been used to read out these important regulatory elements (Cammack et al., 2020; Kfoury et al., 2021; Moudgil et al., 2020b). To record these sites in vivo, we co-transduced mouse cortexes with unfused *piggyBac* and barcoded or non-barcoded SRTs.

After 21 days, we collected similar amounts of brain tissue from mice injected with barcoded or non-barcoded SRTs and prepared calling card libraries (**Figure 10A**). As with our *in vitro* experiments, all 25 unique barcodes were integrated into the genome and efficiently recovered (**Figure 10B**). Lower recovery of 2/25 barcodes may reflect imbalances in vector DNA pooling prior to AAV packaging. After normalizing by the total depth of sequencing, we found that use of barcodes improved the recovery of SRTs and yielded around 2-fold more genomic insertions than non-barcoded counterparts (**Figure 10C**). Genome-wide, integrations of barcoded and non-barcoded SRTs were highly concordant. Visualizing insertions and called peaks across the genome demonstrates this concordance (**Figure 10D**). Analysis of genomic features of SRT insertion sites revealed similar insertional preferences (**Figure 10E**). We recovered more insertions in promoter regions using barcoded SRTs, suggesting the unbarcoded SRTs might have had especially limited dynamic range in these regions. This would be expected as some of these loci are expected to contain strong binding sites or few 'TTAA' sequences, limiting the quantification of recurrent non-barcoded insertions.

Next, we performed functional enrichment analysis of genes located near insertions. Based on the tropism of AAV9, we expected the vast majority of insertions to be in neuronal cells, with some insertions in astrocytes. Accordingly, barcoded and non-barcoded insertion sites were located near genes strongly enriched for neurological functions including synapse organization, forebrain development, and axonogenesis (**Figure 10F**). Functional enrichment was similar for insertions of barcoded and non-barcoded SRTs and consistent with our previous findings (Cammack et al., 2020; Moudgil et al., 2020b). Altogether, these results demonstrate that barcoded SRTs can recover biologically relevant binding events *in vivo* and outperform non-barcoded SRTs in the number of unique insertions at a fixed sequencing depth, while significantly reducing labor and reagent costs.

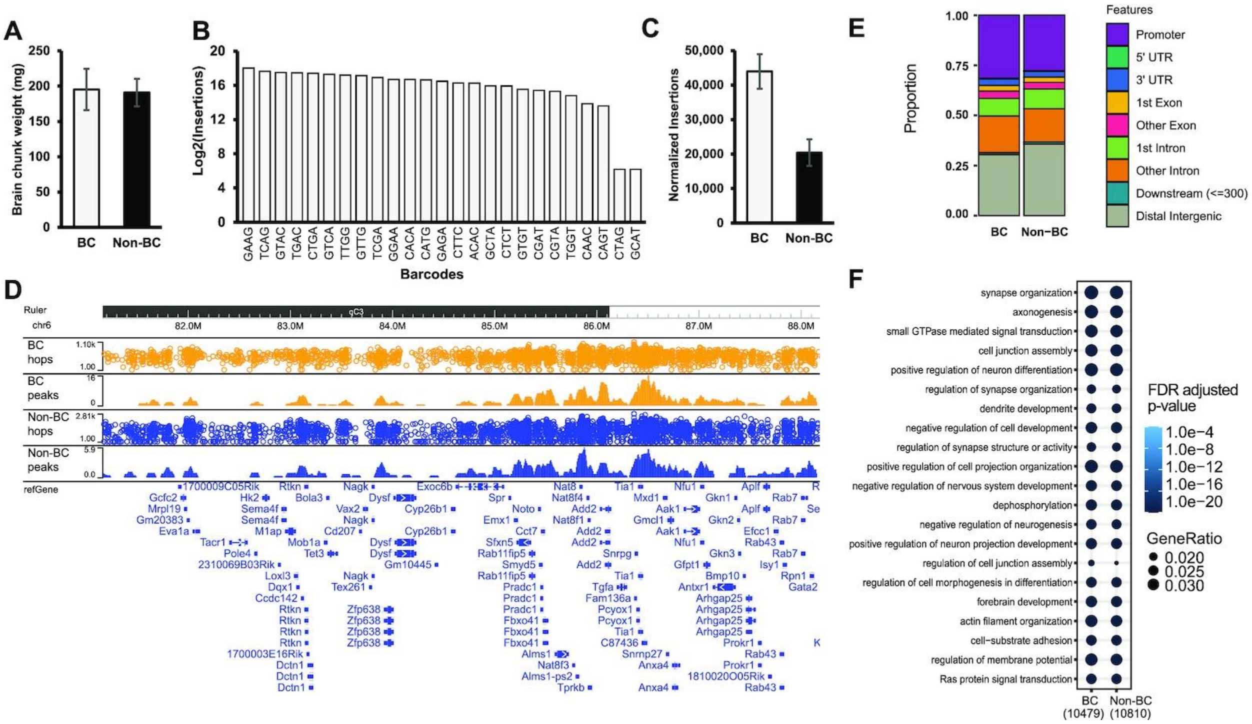


Figure 10: Comparison of barcoded and non-barcoded SRT Calling Cards in the mouse brain

(A) Equivalent amounts of brain tissue were collected after in vivo calling card experiments using a pool of 25 barcoded (BC) or non-barcoded (non-BC) SRT donors delivered by AAV. n = 4 for BC and 3 for Non-BC. **(B)** Number of genomic insertions recovered for each barcoded SRT. **(C)** Number of genomic insertions recovered at the same depth of sequencing for barcoded and non-barcoded SRTs. **(D)** Browser view of genomic insertions and called peaks for barcoded and non-barcoded SRTs. **(E)** Genomic features of peaks called by barcoded and non-barcoded experiments. **(F)** KEGG pathway enrichment comparison of genes near peaks called by barcoded and non-barcoded experiments.

Design and validation of nuclear Calling Cards reagents

One of the central goals of medicine is to associate genotype with phenotype. Single-cell genomics has emerged as a powerful and essential methodology for unraveling cellular heterogeneity in various disease states and complex tissues. However, isolating high-quality, intact cells for transcriptome or epigenomic studies can be challenging due to factors like cell size, fragility, or connectivity. Moreover, cell isolation from fresh frozen tissues is generally not possible. As an alternative, nuclei have been demonstrated to be viable for analyses of frozen or tough-to-dissociate tissues, with some studies demonstrating that single-nucleus RNA sequencing (snRNAseq) might outperform single-cell RNA sequencing (snRNAseq) for hard-to-dissociate samples (Wu et al., 2019; Wen et al., 2022). One limitation of using nuclei is their lower RNA content and lack of cytoplasmic RNA which can be sufficient for identifying cell types but could hinder detailed cell state analysis (Bakken et al., 2018; Thrupp et al., 2020). Both single-cell and single-nucleus RNAseq pose advantages and experimental challenges. To make Calling Cards technology more universally adoptable across a wide array of studies and sample types for the broader research community, I have developed a novel nuclear Calling Cards SRT construct. This innovation enables the use of nuclei for Calling Cards, thereby broadening its utility and enhancing our capability to dissect cellular heterogeneity and protein-DNA interactions in challenging tissue types.

Efficient transfection or transduction requires plasmids or episomes to not only get into the cell, but also into the nucleus where transcription occurs. Standard Calling Cards constructs allow for SRT transcription and subsequent cytoplasmic localization, observable via live-cell or tissue imaging (Cammack et al., 2020; Lalli et al., 2022; Moudgil et al., 2020b). To retain the SRTs in the nucleus, I engineered a construct incorporating a histone H2B to the tdTomato SRT's N-terminal side and 3 SV40 nuclear localization sequences (NLS) at the C-terminal side (**Figure 11A**). To validate that the addition of these NLS elements does not affect the function of the SRT, the constructs were packaged into AAV9 viral particles and injected into the mouse cortex. Nuclear-localized tdTomato expression was confirmed, indicating successful and functional SRT expression and nuclear retention (**Figure 11B**).

Further analysis via flow cytometry analysis on isolated cells and nuclei from the cortices revealed substantial transduction rates, with the 70-80% of the cells/nuclei highly expressing tdTomato (**Figure 11C**). Sequencing analysis revealed a modest 15% reduction in insertion recovery from the nuclear-targeted H2B-SRT compared to its cytosolic counterpart, despite similar sequencing depths (**Figure 11D, E**). These findings confirm that the H2B-SRT construct Calling Cards in nuclear samples, potentially accommodating a broader spectrum of samples and research objectives.

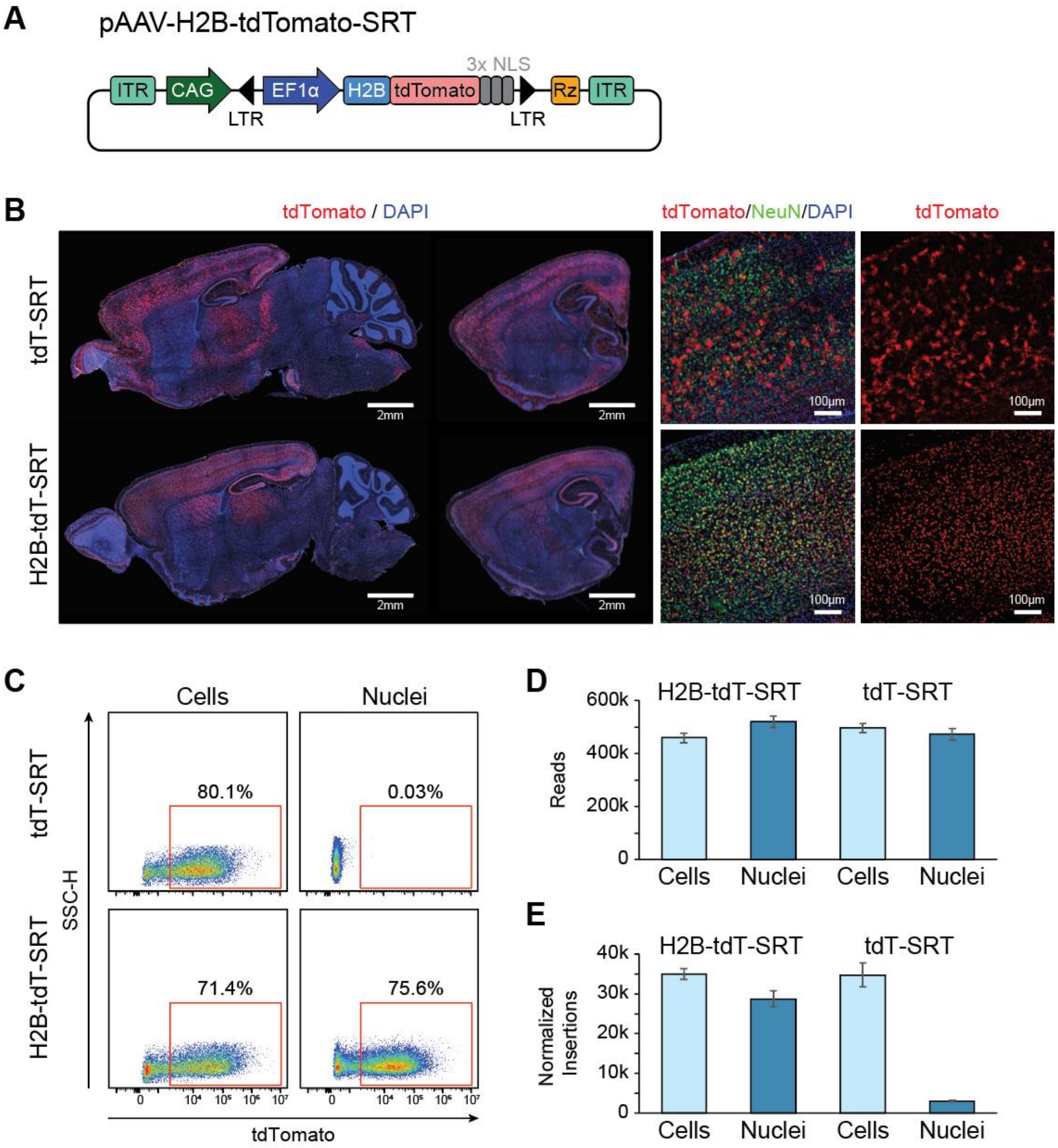


Figure 11: Design and validation of nuclear Calling Cards using H2B-SRT

(A) Schematic of the pAAV-H2B-tdTomato-SRT-3xSV40. The H2B fused tdTomato SRT is driven by a CAG promoter and flanked by 3x SV40 sequenecs at the 3' end of the SRT. The ribozyme (Rz) is present to suppress expression of non-integrated SRTs. **(B)** Images of sagittal sections of mouse brains that have been injected. The images on the right show images taken at higher magnification, showing that SRT expression in many cells and nuclei in the cortex. **(C)** Scatterplots of flow cytometry figures where the cells/nuclei have isolated from injected brains and gated for objects, singlets, and viability for cells. **(D)** Summary bar graph showing that the libraries were sequenced similarly. **(E)** Summary bar graph showing that a slight reduction in the number of insertions recovered in nuclei expressing H2B-tdT-SRT compared to standard tdT-SRT.

2.5 Basic Protocol 1: Preparation and delivery of Calling Cards reagents

This two-track protocol describes the procedure to transfect plasmids into *in vitro* cultures or inject AAV reagents into the developing mouse brain. Prior to beginning this protocol, high quality endotoxin-free Calling Cards plasmids or purified AAVs are required. Plasmid DNA can be prepared using many available commercial maxiprep kits. If an academic core or commercial virus packaging service is not available to generate AAVs, the procedure is outlined and described here (Challis et al., 2019). This protocol describes neonatal intracerebroventricular injections as an example, but AAVs can be injected into the target tissue of choice.

> ***NOTE:*** *All animal studies were approved by and were performed in accordance with the guidelines of the Animal Care and Use Committee of Washington University in Saint Louis, School of Medicine and conform to NIH guidelines of the care and use of laboratory animals.*

> ***NOTE:*** *Viral vectors are biohazardous materials and investigators must be trained according to governmental and institutional regulations and standard operating procedures.*

The complete list of reagents and equipment needed for this protocol can be found in **Table 4**.

Calling Cards plasmid preparation (timing: variable)

Once you receive the bacterial stabs from Addgene, it is recommended to create glycerol stocks or purified plasmid for each of the plasmids for long-term storage. A complete guide can be found on their website (https://www.addgene.org/protocols/create-glycerol-stock/).

NOTE: *it is highly recommended to validate each plasmid by long-read sequencing. Fully sequencing the entire plasmid is advantageous because it can reveal unexpected products, deletions, recombinations, and concatemers. Additionally, it can sequence through repetitive elements such as ITRs, which can form secondary structures and are typically troublesome for Sanger sequencing reactions. Confirmation of intact ITRs is crucial as mutations in this region can affect packaging efficiency.*

The following protocol describes the general steps to amplify plasmid constructs. To create a balanced plasmid pool of all the barcoded self-reporting transposons (plasmid #7 in **Table 5)**, start at Step 1. For individual constructs, begin at Step 4.

1. Quantify the concentration of each of the purified plasmids by nanodrop or Qubit.

   NOTE: *the Qubit assay is recommended for sample concentrations that are <100 ng/µl.*
2. Pool equal masses (e.g., 100 ng) of each plasmid into a clean 1.5 ml microcentrifuge tube.
3. Pulse vortex to mix and briefly centrifuge.
4. Quantify the concentration of the pool by nanodrop or Qubit.
5. Transform plasmid pool into NEB Stable Competent *E. coli* cells or equivalent bacterial strain according to manufacturer’s instructions. For individual plasmids, proceed to the next step. For the barcoded plasmid pool, skip step 6 and proceed to step 7.

   NOTE: *use of strains that are optimized for plasmids containing repeat elements with reduced recombination and endonuclease activity (e.g. NEB Stable or One Shot Stbl3) are recommended.*
6. For individual plasmids: spread 50-100 µl onto a selection plate and incubate overnight at 37°C.

7. For individual plasmids: use a sterile pipette tip or inoculation loop to pick a colony from the plate and drop into a 150ml liquid LB+antibiotic culture.

   For barcoded plasmid pool: after 1 hr of outgrowth, directly inoculate a 150 ml liquid LB+antibiotic culture.

8. Incubate the liquid culture in a shaking incubator for 12-18 hr at 37°C or 24 hr at 30°C.

   NOTE: *the standard step of plating the outgrowth on selection plates is skipped to streamline amplification of the plasmid pool and recovery of all barcoded elements. The outgrowth can be plated; however, the whole plate would need to be scraped then inoculated into a liquid overnight culture.*

9. After the incubation, harvest bacteria by centrifuging at 5000 xg for 10 mins at 4°C. Discard supernatant and purify plasmid DNA using the ZymoPURE II Plasmid Maxiprep kit according to manufacturer's instructions.

   NOTE: *to create a glycerol stock, take 500 µl of the liquid culture and mix with 500 µl 40% glycerol prior to centrifuging. This can be stored in cryovials at -80°C for years. Freeze and thaw cycles should be avoided.*

   NOTE: *it is highly recommended to perform the endotoxin-removal step during plasmid purification. The QIAGEN Plasmid Maxi kit or other equivalent kit can also be used to obtain high-quality endotoxin-free plasmid DNA.*

10. Quantify concentration by nanodrop or Qubit.

11. For individual plasmids and barcoded plasmid pool: sequence to verify important plasmid features (e.g, AAV ITRs, tdTomato transgene, and PB LTR) on the entire plasmid by long-read sequencing prior to AAV packaging.

For barcoded plasmid pool: proceed to Support Protocol 1 to use next-generation sequencing to assess barcode distribution of plasmid pool.

NOTE: *Sanger sequencing can also be done to sequence verify plasmid features, however multiple sequencing primers are typically needed to cover an entire plasmid. Additionally, Sanger cannot detect multiple plasmid products within the tube and typically requires special reaction conditions to sequence through ITRs.*

12. (Recommended, but optional). Assess endotoxin levels to ensure virtually undetectable levels of <2.5 endotoxin units [EU]/ml using the Endosafe nextgen-PTS (Charles River) Assay. Consult manufacturer's documentation and user guide for specific instructions (https://criver.com.sg/products/lal-rapid-catridge-technology/item/download/61_6f1703d2a3b8d985c3483c7d8c0ddbe5).

    NOTE: *depending on the injection route and tissue, trace amounts of endotoxin can trigger strong adaptive immune responses and lead to inflammatory response, especially when injected intravenously.*

This is a safe stopping point and the plasmid pool can be stored at -20°C, or proceed to the next step.

Delivery of Calling Cards reagents *in vitro* (timing: variable)

There are numerous methods and commercial reagents available to transfect plasmids into cells. Polyethyleneimine (PEI) is a robust and cost-effective reagent that has been optimized for high transgene expression in a variety of cell lines. A generalized transfection protocol can be found here (Yang et al., 2017).

Production of adeno-associated viruses (timing: variable)

Adeno-associated viruses (AAVs) are commonly used to deliver stable long-term expression of the gene of interest into cells or tissues. All endotoxin-free plasmid preparations of Calling Cards constructs can be packaged into AAV particles with the desired serotype. The packaging can be done in-house following this protocol (Challis et al., 2019), or performed by academic core facilities or commercial vendors providing AAV packaging as a service.

Intracerebroventricular injection (timing: 1 hr)

> ***NOTE:*** *All animal studies were approved by and were performed in accordance with the guidelines of the Animal Care and Use Committee of Washington University in Saint Louis, School of Medicine and conform to NIH guidelines of the care and use of laboratory animals.*

Delivery of AAVs into the ventricle of P0-2 mouse pups is an easy and efficient route to label cells in the developing cortex (Cammack et al., 2020). A key benefit of delivery at this age is that it allows for efficient CNS labeling yet is considered a non-surgical injection since the skull bone has not yet hardened. Additionally, the procedure for a whole litter is more rapid than adult stereotactic injection and the survival rate is very high. Investigators and end users should be trained according to institutional guidelines prior to using viral vectors in the laboratory. See **Supplemental Figure 6A, B** for an example of the setup.

> ***NOTE:*** *if available in the animal facility, work with viral vectors and animals in a biosafety cabinet.*

13. Remove dam and stud from home cage and place in clean cage away from pups to avoid stressing the dam.
14. Set a heating pad to “low” (warm to the touch) and place the home cage on top.
15. Prepare an appropriate amount of AAV cocktail by thawing viral aliquots on ice. A total of 6 µl virus (3 µl transposase and 3 µl donor transposon) is injected per animal. Depending on the concentration, dilute each AAV with sterile PBS to 1.0x10^{13} vg/ml and mix the transposase and donor transposon AAVs 1:1. Keep the AAV cocktail on ice. (*Optional):* Add 0.05% Fast Green FCF dye to the AAV cocktail to facilitate visualization of the injection into the ventricles.
16. Fill a beaker with freshly made 0.5% sodium hypochlorite (10% Clorox) solution to decontaminate any material or liquids that come in contact with AAVs.
17. Prepare and clean a Hamilton syringe with 5 full volume washes with sterile deionized water, 5 washes with 80% ethanol, followed by 5 washes with sterile deionized water.
18. Load the AAV cocktail into a Hamilton syringe.

    NOTE: *Avoid excessive pipetting and handling to prevent bubbles as injection of air into the ventricle is often fatal.*
19. Place a wet paper towel on a bed of ice. Anesthetize pups by placing them on the ice for 5-8 mins (maximum of 15 mins) to induce hypothermia. The paper towel ensures that their skin is not making direct contact with the ice. When unresponsive to physical stimuli, the pups are sufficiently anesthetized.

    NOTE: *it is recommended to work in appropriately sized batches to ensure that the pups are not left on ice for more than 15 mins.*

20. Working quickly but carefully, inject 1 µl of virus per site at 3 sites (noted as x’s in **Supplemental Figure 6C**) per hemisphere at a rate of 1 µl per second. After each injection, wait 5 seconds before withdrawing the needle to minimize backflow. Repeat for the opposite hemisphere.
21. Place pup in home cage on heating pad for recovery.
22. Repeat injection steps 20 and 21 for the remaining pups.
23. Leave pups on the heating pad for 5-10 mins until they have returned to normal body temperature, indicated by movement, and return to pink color.
24. Carefully return pups to home cage, followed by the dam and stud.

    NOTE: *rub gloves through dirty home cage bedding prior to handling the pups to mask the foreign scent of nitrile gloves. This should decrease the chance of cannibalization.*
25. Dispose of any leftover virus or unwanted contaminated material that came into contact with AAV into the beaker with bleach. After 5 mins, decontaminated unwanted materials can be disposed of properly.
26. Depending on experimental design, allow time for animals to age and AAVs to express. Collection can occur as early as 2 days, or as late as adulthood.

2.6 Support Protocol 1: Next-generation sequencing quantification of barcode distribution within self-reporting transposon plasmid pool and adeno-associated virus genome

Since the barcoded SRT plasmid pool contains multiple barcodes, attention is needed to ensure that all barcodes are represented evenly, amplified, and that none are lost throughout this process. Having an evenly distributed plasmid pool is optimal for transformation, outgrowth, and for ultimately generating an evenly distributed pool of packaged AAV particles. While Sanger sequencing provides the consensus sequence of the pool, next-generation sequencing can reveal the actual representation of the library. To do so, a fragment of the plasmid containing the SRT barcodes should be amplified by PCR and standard Illumina sequencing adapters will be ligated. These products will then be purified, quantified, and submitted for sequencing. Similarly, the AAV genomes from the packaged viral particles can be harvested and analyzed for the actual distribution of barcodes that were packaged.

The complete list of reagents and equipment needed for this protocol can be found in **Table 6**.

<u>Isolation of viral genome from AAV particles (timing: 3.5-4 hr)</u>

The protocol is adapted from the ‘Digest virus particles to release DNA’ section of Support Protocol 1: Determination of rAAV Titers by the Dot-Blot Assay (Gray et al., 2011).

1. Prepare DNA digestion reaction according to the table below in a 1.5 ml microcentrifuge tube.

Component	**1 reaction (µl)**
10x DNase I reaction buffer	10
2 U/µl DNase I	1
Concentrated AAV stock	2
Nuclease-free water	87
Total	*100*

2. Pulse vortex and centrifuge to collect liquid. Incubate for 1 hr at 37°C.
3. Add 1 µl of 0.5 M EDTA to each tube to inactivate the DNase. Pulse vortex and centrifuge to collect liquid. Incubate for 10 min at 75°C.
4. Add 2 µl Proteinase K to the reaction. Pulse vortex and centrifuge to collect liquid. Incubate for 2 hr at 50°C.
5. Heat inactivate the Proteinase K by incubating for 10 mins at 95°C.
6. Allow the tubes to cool and clean up reactions using QIAquick PCR Purification kit according to manufacturer's instructions.

<u>Library preparation (timing: 1h)</u>

7. Prepare a 1-step sequencing library to assess barcode distribution of the pool.

Component	1 reaction (µl)
Q5 HF 2x Master Mix	5
10 µM SRT_bc_QC_F primer	0.5
10 µM SRT_bc_QC_F primer	0.5
5 ng AAV genome	Variable
Nuclease-free water	Variable
Total	*10*

NOTE: *The primer sequences can be found in* ***Table 3****. The volumes can be scaled up if the concentration of the AAV genome is too low.*

8. Transfer tubes to a preheated PCR machine and begin thermocycling with the following parameters.

Temperature	Time	Cycles
98°C	30 sec	1
98°C	5 sec	10 cycles
55°C	10 sec	
72°C	30 sec	
4°C	Hold	1

This is a safe stopping point and the PCR amplicons can be stored at -20°C, or proceed to the next step.

Bead cleanup and quantification (timing: 30min)

NOTE: *For all wash steps, use freshly prepared 80% ethanol.*

9. Bring AMPure XP beads to room temperature for at least 30 mins. Vortex completely to resuspend beads immediately prior to use.
10. Add 40 µl of nuclease-free water to the PCR reaction to bring the volume up to 50 µl.
11. Add 50 µl AMPure XP beads (1X) to the sample and mix thoroughly by pipetting or vortexing.
12. Incubate on bench for 5 min.
13. Place on the magnetic rack for 2 min or until the solution clears. Without disturbing the beads, aspirate and discard the supernatant.
14. Add 200 µl 80% ethanol, making sure not to disturb the bead pellet. Incubate for 30 sec.
15. Aspirate supernatant and discard.
16. Repeat wash by adding 200 µl 80% ethanol to each sample. Incubate for 30 sec.
17. Aspirate supernatant and discard.
18. Air dry pellet for 1 min or until the beads become matte and lose their shine.

    NOTE: *do not over dry the beads (they will appear cracked) as this will decrease elution efficiency!*
19. Pulse-spin the strip tubes to collect any remaining ethanol. Place on the magnetic rack and aspirate and discard residual ethanol.

20. Remove the strip tubes from the magnetic rack. Add 20 µl Buffer EB to elute PCR products. Mix thoroughly by pipetting.
21. Incubate on bench for 2 min.
22. Place on a magnetic rack for 1 min, or until supernatant is clear.
23. Transfer 20 µl supernatant to a new tube.

    NOTE: *it is important to ensure that there is no bead carryover as this can affect downstream steps. The supernatant should be completely clear*.
24. Quantify the concentration and visualize PCR products by running a 1:10 diluted sample on an Agilent D1000 ScreenTape device.
25. Submit for shallow low depth sequencing.

    NOTE: *>10k reads is sufficient to quantify the barcode distribution within the pool of AAV genomes.*

Data analysis (timing: variable)

The distribution of 4 bp SRT barcodes found within the pool of packaged AAV genomes can be assessed by counting the number of reads associated with each SRT barcode. A simple way to analyze this is to search within the *_R1.fastq.gz file for the string CTTTNNNNGGTTAA, where NNNN represents the SRT barcode within the entire SRT construct and count the number of unique occurrences of each string. A barcode whitelist (barcode_whitelist.txt) is provided to only include counts from the known 25 possible barcodes. The counts are then saved to a tab-delimited file named output.txt, which can be used to visualize the distribution as shown in **Figure 12**. Small differences in barcode distribution (e.g. a 2-3-fold difference between the most and least expressed barcode) are normal and expected.

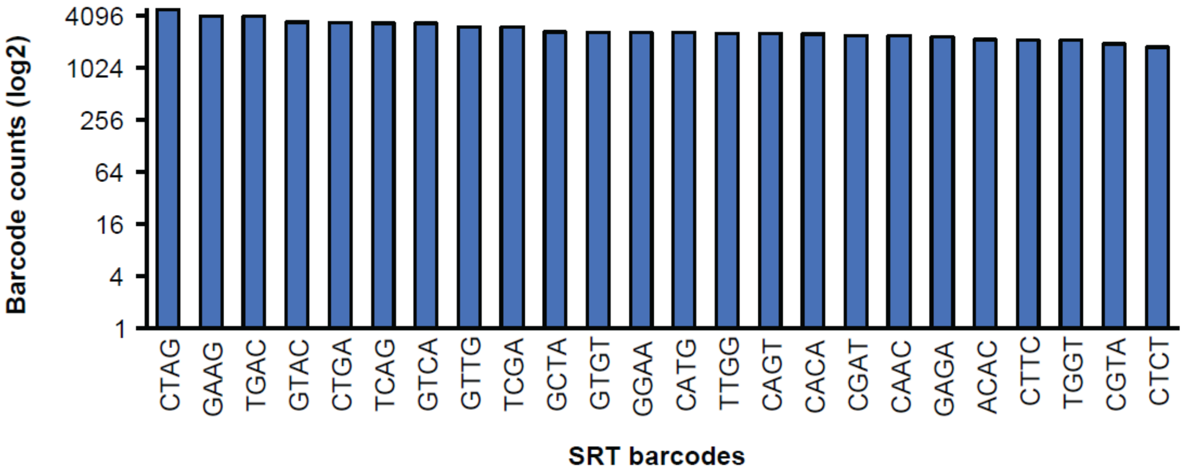


Figure 12: Representative distribution of SRT barcodes in AAV genomes

(A) Schematic showing the expected base calls for Read1 beginning with the 3 bp OM-PB primer barcode, followed by 25 bp of the hyPB long terminal repeat, a 4 bp SRT barcode, then GGTTAA preceding genomic sequence. **(B)** Bar plot showing the relatively equal barcode representation from a library prepared from pooled AAV genomes.

26. Use a text editor to generate the barcode_whitelist.txt file. Each line should contain the query string. The full list is provided below.

```
CTTTCTAGGGTTAA
CTTTGAAGGGTTAA
CTTTTGACGGTTAA
CTTTGTACGGTTAA
CTTTCTGAGGTTAA
CTTTTCAGGGTTAA
CTTTGTCAGGTTAA
CTTTGTTGGGTTAA
CTTTTCGAGGTTAA
CTTTGCTAGGTTAA
CTTTGTGTGGTTAA
CTTTGGAAGGTTAA
CTTTCATGGGTTAA
CTTTTTGGGGTTAA
CTTTCAGTGGTTAA
CTTTCACAGGTTAA
CTTTCGATGGTTAA
CTTTCAACGGTTAA
```

```
CTTTGAGAGGTTAA
CTTTACACGGTTAA
CTTTCTTCGGTTAA
CTTTTGGTGGTTAA
CTTTCGTAGGTTAA
CTTTCTCTGGTTAA
```

27. In the terminal, the following command will search for lines in the input file that match any query strings in the barcode_whitelist.txt. The unique occurrence of each barcode is then counted and sorted in descending order. The results are saved in a tab-delimited file (output.txt).

```
$ zgrep -f barcode_safelist.txt AAVgenomes_R1.fastq.gz \
| grep -F -o -f barcode_safelist.txt \
| sort \
| uniq -c \
| sort -nr > output.txt
```

2.7 Basic Protocol 2: Sample preparation and RNA purification

There are numerous methods and commercial kits to purify total RNA. The steps outlined below describe a generic procedure to lyse cells and purify RNA using a phenol-chloroform extraction method. Specific cell lines or cultures may require some optimization or additional steps.

The complete list of reagents and equipment needed for this protocol can be found in **Table 7**.

Option A: Harvesting *in vitro* cultures

1. Carefully aspirate the growth medium.
2. Add 1 ml cold TRIzol per 2.5x10^7 cells directly onto cells and pipette up and down several times until cells have been lysed and homogenized. This can be scaled down or up as needed.

   NOTE: *washing cells with DPBS prior to addition of TRIzol can lead to mRNA degradation and is thus not recommended.*
3. Transfer to a clean microcentrifuge tube and incubate for 5 mins at RT to allow for complete dissociation of nucleoprotein complexes. Samples can be frozen at -80C until further processing.
4. Add 0.2 ml chloroform for every 1 ml TRIzol used and shake vigorously for 30 sec.
5. Incubate for 7 mins at RT.
6. Centrifuge samples for 15 mins at 12,000 xg at 4°C.

   NOTE: *the mixture separates into a pink organic phase on the bottom, a thin white interphase in the middle, and a colorless upper aqueous phase.*
7. Harvest the colorless upper aqueous phase to a new tube and record the volume.

NOTE: Do not touch the interphase or organic phases with the pipette tip when removing the aqueous phase! The interphase and organic phases can be saved and frozen at -80°C if desired for DNA and protein recovery.

8. Proceed to RNA purification.

Option B: Harvesting brain tissue (timing: 10 min/mouse)

1. Deeply anesthetize a mouse with isoflurane.
2. (Optional). Conduct a trans-cardial perfusion using ice-cold DPBS. A protocol can be found here (Wu et al., 2021).

   ***NOTE:** the perfusion with DPBS is only to clear tissues of blood and not required for RNA purification and library preparations. To perform immunostaining to visualize the localization of AAV-transduced cells, perfusion and fixation with paraformaldehyde is recommended. Tissue for library preparations should not be fixed with paraformaldehyde.*
3. Harvest the brain and dissect specific regions if needed.
4. Transfer to a clean 1.5 ml RNase-free microcentrifuge tube.

   ***NOTE:** the tissue can be snap-frozen with liquid nitrogen and stored at -80°C until further processing with minimal impact on the quality of final Calling Cards libraries. Harvested tissue can also be stored in RNAlater Stabilization Solution (ThermoFisher AM7020).*

The steps outlined below describe a general procedure to homogenize mouse brain tissue and purify total RNA using a phenol chloroform extraction method. Any RNA cleanup approach that produces high quality DNA-free total RNA should be compatible with Calling Cards. Different tissue types may require some optimization.

5. Weigh the tissue and return to microcentrifuge tube.

NOTE: *if working with frozen tissue, do not allow tissue to thaw prior to homogenization.*

6. Add 500 µl cold TRIzol to the frozen tissue and immediately homogenize using 10 strokes with a cordless handheld homogenizer or until no visible tissue chunks remain.

   NOTE: *to minimize bubbles and loss of homogenate from the tube, keep the pestle tip below the liquid level.*

7. Add additional volumes of cold TRIzol to bring the total volume to 1 ml for each 100 mg tissue.
8. Incubate for 5 mins at RT. If processing multiple samples, repeat steps 13-15 to homogenize the remaining samples and start the timer after the last sample.
9. Add 0.2 ml chloroform for every 1 ml TRIzol used and shake vigorously for 30 sec.
10. Incubate for 7 mins at RT.
11. Centrifuge samples for 10 mins at 12,000 xg at 4°C.

    NOTE: *the mixture separates into a pink organic phase on the bottom, a thin white interphase in the middle, and a colorless upper aqueous phase.*

12. Harvest the colorless upper aqueous phase to a new tube and record the volume.

    NOTE: *Do not touch the interphase or organic phases with the pipette tip when removing the aqueous phase! The interphase and organic phases can be saved and frozen at -80°C if desired.*

13. Proceed to RNA purification.

<u>RNA purification (timing: 60 mins)</u>

The Zymo RNA Clean & Concentrator-5 Kit and QIAGEN RNEasy Plus Mini Kit have been tested and validated for brain tissue and *in vitro* cultured cells. Generally, any RNA cleanup and purification kit that yields high quality RNA can be used. Optimizations may be needed for other tissue types or kits. The following section describes the protocol using the Zymo RNA Clean & Concentrator-5 kit.

> ***NOTE:*** *Clean all working surfaces and pipettes with RNaseZap before setting up the workspace. Use sterile, RNase- and DNase-free, and filter barrier tips throughout the procedure.*
>
> ***NOTE:*** *An on-column DNase treatment included in the Zymo kit is described below. If this does not sufficiently remove genomic DNA, a more effective DNase treatment in solution can be performed using TURBO DNA-free DNase Treatment and Removal Reagents (ThermoFisher AM1907), or equivalent.*

14. Add 2 volumes of Zymo RNA Binding Buffer to each tube.

    NOTE: *For example, if you collected 200 µl aqueous phase, add 400 µl Binding Buffer.*

15. Add 1 volume of 100% ethanol to each tube. Mix thoroughly by pipetting.

    NOTE: *For example, if you have 600 µl aqueous phase+Binding Buffer from the previous step, add 600 µl ethanol.*

16. Transfer 700 µl to a column at a time. Centrifuge for 30 sec at 12,000 xg at RT to bind RNA. Discard flow-through and repeat until all remaining solution has been run through the column.

    NOTE: *A RNase- and DNase-free vacuum manifold can also be used for steps 16, 17, and 20-22 to process many samples in parallel.*

17. Add 400 µl Zymo RNA Wash Buffer to the column and spin at 12,000 xg for 1 min at RT. Discard flow through and move to a new collection tube.

18. Prepare the DNase mixture.

Component **Add in order**	**1 reaction (µl)**	**8 reactions + 10% (µl)**
DNase Digestion Buffer	35	283.5
DNase I	5	40.5
Total	*40*	*324*

19. Add 40 µl DNase mix directly to the membrane of each column. Incubate for 15 min at RT.
20. Add 400 µl Zymo RNA Prep Buffer and spin at 12,000 xg for 30 sec at RT. Discard flow-through.
21. Add 700 µl Zymo RNA Wash Buffer and spin at 12,000 xg for 30 sec at RT. Discard flow-through.
22. Add 400 µl Zymo RNA Wash Buffer and spin at 12,000 xg for 1 min at RT. Discard flow-through.
23. Spin empty column at 12,000 xg for 1 min at RT to remove residual Wash Buffer.
24. Transfer the column to a pre-labeled RNase-free tube. Add 30-100 µl nuclease-free water directly to the silica bed. Incubate at RT for 1 min, then spin at 12,000 xg for 2 min.

    NOTE: *the amount of water will depend on tissue/cell type, expected RNA yield, and desired concentration.*
25. Use 1 µl to measure A_{260}/A_{280} and concentration using a Nanodrop.

    NOTE: *Values should be near 1.95-2.00.*
26. Assess quality, integrity, and concentration of RNA using a High Sensitivity RNA ScreenTape or Bioanalyzer RNA 6000 Pico Assay.

    NOTE: *Alternatively, the Qubit RNA High Sensitivity Assay and Qubit RNA IQ Assay can be used.*

This is a safe stopping point and the RNA can be stored at -80°C, or proceed to first strand cDNA synthesis.

First strand cDNA synthesis (timing: 2 hr)

The goal of this section of the protocol is to reverse transcribe polyA mRNAs using an oligo(dT) primer that also adds on a common primer sequence (Picelli et al., 2014). After first strand synthesis, RNase H is used to degrade the RNA strand within RNA/DNA duplexes. The cDNA products are then column purified as carryover SMART_dT18VN primer was found to increase nonspecific PCR products in the subsequent steps.

> ***NOTE:*** *Clean all working surfaces and pipettes with RNaseZap before setting up the workspace. Use sterile, RNase- and DNase-free, and filter barrier tips throughout the procedure.*

Use a thermocycler with a heated lid set to 105°C for all incubations throughout this protocol.

27. Set up the following PCR program on a thermocycler and preheat to 65°C.

Temperature	Time	Cycles
65°C	5 min	1
4°C	Hold	1
50°C	60 min	1
85°C	10 min	1
4°C	Hold	1

28. Add the following components into individual PCR strip tubes on ice.

Component	1 reaction (µl)
Nuclease-free water	Variable
Template RNA*	Variable
100 µM SMART_dT18VN	0.5
10 mM dNTPs	1.0
Total	*14.0*

*Application note: Up to 5 µg total RNA can be used for cDNA synthesis. It is recommended to maximize the starting RNA amounts to maximize recovery of Calling Cards insertions.

29. Mix by pipetting and centrifuge briefly.
30. Transfer tubes to preheated thermocycler and incubate at 65°C for 5 min.
31. After incubation, place tubes on ice and preheat the thermocycler to 50°C and the next phase of the program.
32. Add the following components into each tube on ice.

Component	**1 reaction (µl)**
5x RT Buffer	4
100 µM RNaseOUT RNase Inhibitor	1
Maxima H Minus Reverse Transcriptase (200 U/µl)	1
Total	*6*

33. Mix by pipetting and centrifuge briefly.
34. Transfer tubes to preheated thermocycler and resume the program to incubate at 50°C for 60 mins, 85°C for 10 mins, and 4°C hold.
35. Once first strand cDNA synthesis is complete, add 2 U RNase H to each reaction. Mix and centrifuge briefly.
36. Incubate in a thermocycler at 37°C for 20 mins.
37. Clean up PCR reactions with QIAquick PCR Purification Kit (Qiagen) according to manufacturer’s instructions.
    a. Add 105 µl (5 volumes) Buffer PB to each reaction and mix by pipetting.
    b. Transfer to QIAquick column and spin at 10,000 xg for 1 min.
    c. Add 700 µl Buffer PE to wash the column and spin at 10,000 xg for 1 min.

d. Remove residual wash buffer by spinning at 10,000 xg for 1 min.

e. Transfer column to a clean RNase- and DNase-free tube.

f. Elute cDNA by adding 30 µl Buffer EB directly to the membrane and incubate for 1 min at RT.

g. Centrifuge at 10,000 xg for 1 min.

38. Quantify single-stranded cDNA concentration and yield using the Qubit ssDNA assay**.**

This is a safe stopping point and the cDNA can be stored at 4°C for 48 hours or -20°C for weeks. There is an optional library density qPCR assay to assess tdTomato expression as a proxy for "library density," which is a measure of the relative abundance of Calling Cards transcripts in a sample. This is designed to determine if samples should or should not be carried through the remainder of the protocol, which can minimize unnecessary labor and usage of reagents. Proceed to Support Protocol 2 for the qPCR assay or Basic Protocol 3 to continue with the sequencing library preparation.

2.8 Support Protocol 2: Library density quantitative PCR

A library density qPCR assay can be performed to assess relative expression of tdTomato containing SRTs as a preliminary quality control checkpoint. Results can inform a go/no-go decision for individual samples to prevent unnecessary labor and use of reagents. Example qPCR results are shown in **Figure 13.**

The complete list of reagents and equipment needed for this protocol can be found in **Table 8**.

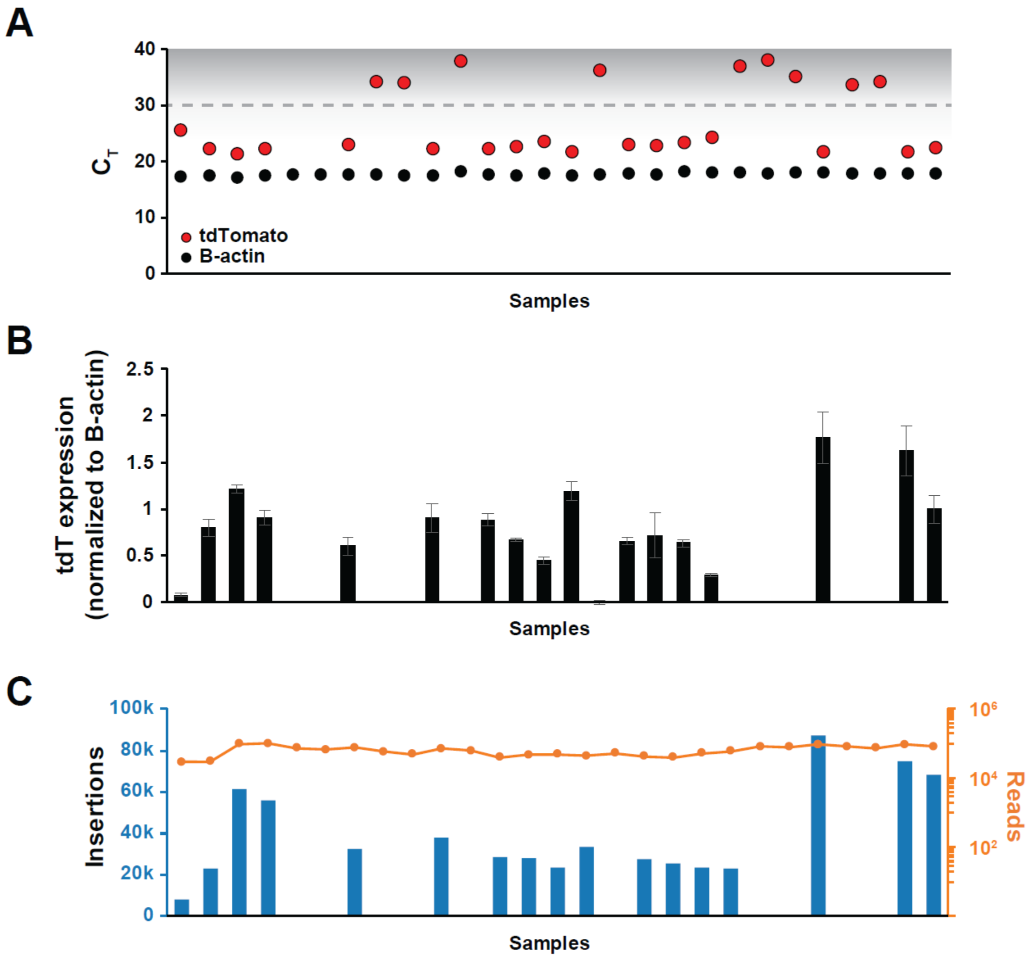


Figure 13: Representative quality control of cDNA and library complexity qPCR

(A) Example data from a qPCR library density assay using samples with a range of Calling Cards expression. Samples with a C_T>30 have minimal number of tdTomato transcripts and can be omitted from downstream processing. **(B)** Plot showing tdTomato expression normalized to B-actin expression. **(C)** Plot showing the number of recovered insertions (blue) with number of sequencing reads (orange) per sample. In this example, insertions were recovered from samples with a tdTomato C_T<30.

1. Determine an appropriate mass of cDNA input based on the single-stranded cDNA concentrations all samples obtained at the end of Basic Protocol 2. A range of 1-10 ng cDNA is compatible with this assay. Use the same mass input across all samples.
2. Prepare the appropriate amount of reagents based on the number of reactions according to the table below. At least 3 replicates of each reaction are recommended. Include no template (nuclease-free water only) control reactions to identify PCR contamination.

Component	**1 reaction (µl)**
2x PowerUp SYBR Green Master Mix	5
5 µM Forward primer*	1
5 µM Reverse primer*	1
cDNA template	Variable
Nuclease-free water	Variable
Total	*10*

*tdTomato, B-actin, and Gapdh primer sequences can be found in **Table 9**.

3. Transfer the reaction mixes to each well of a 384-well plate according to the plate design and layout.
4. Seal the plate with an optical adhesive cover and centrifuge at 100 xg for 1 min to remove any air bubbles and ensure all liquid is at the bottom of each well.
5. Place the plate in the qPCR machine and set up the instrument to run using the following cycling conditions followed by a melt curve.

Temperature	**Ramp Rate**	**Time**	**Cycles**
95°C	1.9°C/s	20 sec	1
95°C	1.9°C/s	1 sec	40 cycles
60°C	1.9°C/s	20 sec	
95°C	1.9°C/s	15 sec	1
60°C	1.9°C/s	60 sec	1
95°C	0.05°C/s	15 sec	1

6. Once the run is finished, calculate the relative expression of tdTomato to β-actin (or other housekeeping gene).
    a. For each replicate, normalize C_T values for tdTomato to the C_T values for β-actin. This is expressed as $\Delta C_T = C_{T,tdTomato} - C_{T,\beta-actin}$.
    b. Exponentially transform ΔC_T for each replicate to calculate expression. This is expressed by $2^{-\Delta C_T}$.
    c. Calculate the mean and standard deviation of expression replicates.
    d. Raw C_T values as well as relative tdTomato expression can be plotted as shown in **Figure 13**.

NOTE: *Samples with no or extremely low tdTomato expression should be omitted from subsequent library preparation steps. Thresholds for determining failed samples will need to be determined empirically, but generally C_T values greater than 30 cycles will likely not yield libraries to recover Calling Cards insertions. Examples are shown in* ***Figure 13****.*

2.9 Basic Protocol 3: Sequencing library preparation

This protocol contains two main steps to create sequencing libraries from cDNA prepared from samples containing Calling Cards: 1) amplification of self-reporting transcripts; and 2) tagmentation and indexing PCR. There are quality control measures following each step to monitor progress through the protocol. At the end of this basic protocol, you will have sequencing libraries for each of your samples.

The complete list of reagents and equipment needed for this protocol can be found in **Table 10**.

Amplification of self-reporting transcripts (timing: 3 hr, 20 mins hands-on)

This PCR step will amplify cDNA containing SRTs for downstream tagmentation, indexing, and sequencing.

1. Set up and preheat a thermocycler with the following PCR program.

Temperature	Time	Cycles
95°C	3 min	1
98°C	20 sec	20 cycles*
65°C	30 sec	
72°C	5 min	
72°C	10 min	1
4°C	Hold	1

***NOTE:** the number of cycles will need to be optimized per tissue/cell type to generate enough PCR product for downstream tagmentation. It is important to avoid overamplification, which can introduce amplification bias and PCR duplicates.*

2. Add the following components into individual PCR tubes on ice.

Component	1 reaction (µl)
2x Kapa HiFi HotStart Readymix	12.5
25 µM TdTomato_F1 Forward primer	0.5
25 µM SMART Reverse primer	0.5

Template cDNA	Up to 100 ng
Nuclease-free water	Variable
Total	*25*

**__NOTE:__ It is recommended to maximize the amount of template cDNA into this PCR step to maximize recovery of Calling Cards insertions (__Figure 14__).*

3. Transfer tubes to the preheated thermocycler and start the program.

This is a safe stopping point and the DNA can be stored at 4°C for 48 hours or -20°C for weeks, or proceed to the next step.

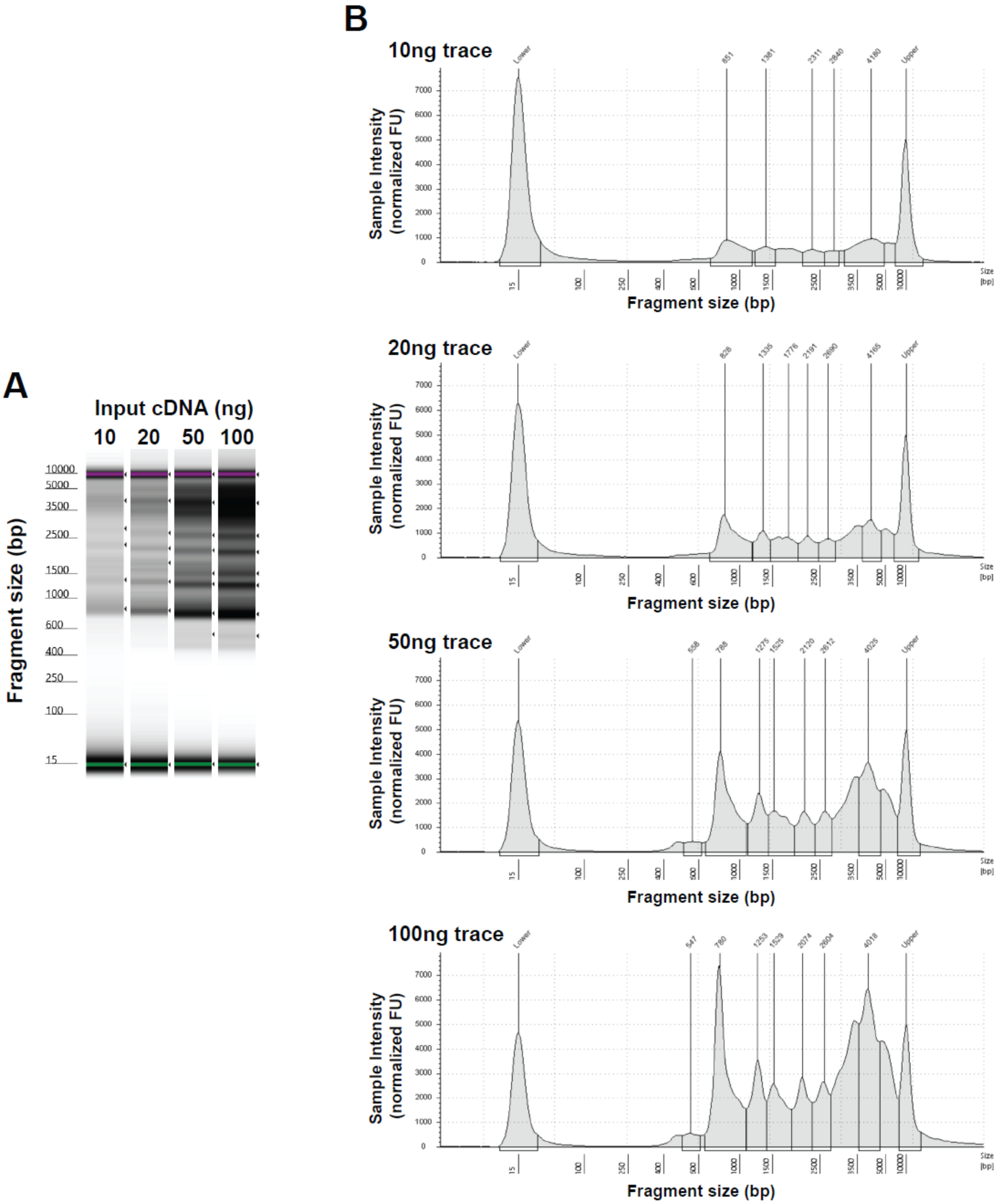


Figure 14: Titration of input cDNA into PCR to amplify SRTs

(A) Example Tapestation gel images and **(B)** electropherogram traces demonstrating that larger amounts of starting material (up to 100 ng) can be used to increase yield of SRTs.

Bead cleanup of PCR products and QC (timing: 30-45 mins)

Magnetic beads are used to clean up the SRT PCR products. If this is your first time working with magnetic beads, it is recommended to consult the manufacturer's documentation for best practices and tips. For all wash steps, use freshly prepared 80% ethanol.

NOTE: *AMPureXP beads (Beckman Coulter) and Mag-Bind TotalPure NGS (Omega BioTek) magnetic beads have both been tested and validated. The protocol specifies AMPure XP beads, however they are interchangeable without any volume changes.*

4. Bring AMPure XP beads to room temperature for at least 30 mins.

   NOTE: *vortex to completely resuspend beads immediately prior to use.*
5. Add 25 µl of nuclease-free water to each PCR reaction to bring the volume up to 50 µl.
6. Add 30 µl AMPure XP beads (0.6X ratio) to each sample and mix thoroughly by pipetting or vortexing.
7. Incubate on bench for 5 min.
8. Place on the magnetic rack for 1 min or until the solution clears. Without disturbing the beads, aspirate and discard the supernatant.
9. Add 200 µl 80% ethanol, making sure not to disturb the bead pellet. Incubate for 30 sec.
10. Aspirate supernatant and discard.
11. Repeat wash by adding 200 µl 80% ethanol to each sample. Incubate for 30 sec.
12. Aspirate supernatant and discard.
13. Pulse-spin the strip tubes to collect any remaining ethanol. Place on the magnetic rack and aspirate and discard residual ethanol.
14. Air dry the pellet for 1 min or until the beads become matte and lose their shine.

NOTE: *do not over dry the beads (they will appear cracked) as this will decrease elution efficiency.*

15. Remove the strip tubes from the magnetic rack. Add 11 µl Buffer EB to elute PCR products. Mix thoroughly by pipetting.
16. Incubate on bench for 2 min.
17. Place on a magnetic rack for 1 min, or until supernatant is clear.
18. Transfer 11 µl supernatant of each sample to a new strip tube.

    NOTE: *it is important to ensure that there is no bead carryover as this can affect downstream steps. The supernatant should be completely clear.*
19. Quantify the concentration and visualize PCR products by running a 1:10 diluted sample on an Agilent High Sensitivity D5000 ScreenTape device or Bioanalyzer High Sensitivity DNA kit. Example gel images and traces are shown in **Figure 15A,B**.

    NOTE: *Alternatively, the Qubit dsDNA High Sensitivity kit can be used to quantify concentration of SRTs.*

This is a safe stopping point and the DNA can be stored at 4°C for 48 hours or -20°C for weeks, or proceed to the next step.

Tagmentation and indexing PCR (timing: 1-1.5 hr)

In this tagmentation reaction, the Nextera XT transposome will enzymatically cleave the cDNA into smaller fragments and tag them with a Nextera overhang. The subsequent indexing will add a unique identifier to each library and enables multiple libraries to be pooled and sequenced together. Proper tagmentation and indexing of the library to a final size of 200-1000 bp is optimal for efficient clustering on the sequencer flow cell.

20. Set up and preheat a thermocycler with the following PCR program.

Temperature	Time	Cycles
55°C	5 min	1
4°C	Hold	1
72°C	3 min	1
95°C	30 sec	1
95°C	10 sec	18 cycles*
52°C	30 sec	
72°C	30 sec	
72°C	5 min	1
4°C	Hold	1

****NOTE:** The number of cycles will need to be optimized per tissue/cell type to generate enough PCR product for downstream sequencing.*

21. Dilute amplified SRT PCR products to 300 pg/µl with nuclease-free water.

22. Add the following components into individual nuclease-free tubes on ice.

Component	**1 reaction (µl)**
Nextera Tagment DNA (TD) Buffer	10
Amplicon Tagment Mix (ATM)	5
Nuclease-free water	3
300 pg/µl SRT DNA	2
Total	*20*

23. Pipette to mix and briefly spin down. Transfer to the thermocycler preheated to 55°C and incubate for 5 min.

24. After the incubation, add 5 µl Neutralization Tagment (NT) Buffer to stop the tagmentation reaction. Pipette to mix and briefly spin down. Incubate at room temperature for 5 mins.

25. Add the following to each PCR tube.

Component	**1 reaction (µl)**
Nextera PCR Master (NPM) Mix	15
Nuclease-free water	8
10 µM barcoded OM-PB primer	1
10 µM Nextera N7 indexing primer	1
Total	*25*

NOTE: *each replicate should receive a unique barcode-index combination (see "Considerations for primer selection and ordering for sequencing libraries" under the Strategic Planning section). See **Table 9** and Error! Reference source not found. for primer sequences.*

26. Pipette to mix and briefly spin down. Transfer to the thermocycler preheated to 95°C and start the program.

This is a safe stopping point and the tagmented DNA can be stored at 4°C for 48 hours or -20°C for weeks, or proceed to the next step.

Bead cleanup of PCR products and QC (timing: 30-45 mins)

Magnetic beads are used to size-select and clean up the tagmented and indexed PCR products.

NOTE: *AMPureXP beads (Beckman Coulter) and Mag-Bind TotalPure NGS (Omega BioTek) magnetic beads have both been tested and validated. The protocol specifies AMPure XP beads, however they are interchangeable without any volume changes.*

27. Bring AMPure XP beads to room temperature for at least 30 mins.

    NOTE: *vortex to completely resuspend beads immediately prior to use.*
28. Add 30 µl AMPure XP beads (0.6X ratio) to each sample and mix thoroughly by pipetting or vortexing.
29. Incubate on bench for 5 min.
30. Place on the magnetic rack for 1 min or until the solution clears. Without disturbing the beads, aspirate and discard the supernatant.
31. Add 200 µl 80% ethanol, making sure not to disturb the bead pellet. Incubate for 30 sec.

32. Aspirate supernatant and discard.

33. Repeat wash by adding 200 µl 80% ethanol to each sample. Incubate for 30 sec.

34. Aspirate supernatant and discard.

35. Pulse-spin the strip tubes to collect any remaining ethanol. Place on the magnetic rack and aspirate and discard residual ethanol.

36. Air dry the pellet for 1 min or until the beads become matte and lose their shine.

    NOTE: *do not over dry the beads (they will appear cracked) as this will decrease elution efficiency.*

37. Remove the strip tubes from the magnetic rack. Add 11 µl Buffer EB to elute PCR products. Mix thoroughly by pipetting.

38. Incubate on bench for 2 min.

39. Place on a magnetic rack for 1 min, or until supernatant is clear.

40. Transfer 11 µl supernatant of each sample to a new strip tube.

    NOTE: *it is important to ensure that there is no bead carryover as this can affect downstream steps. The supernatant should be completely clear.*

41. Quantify the concentration and visualize PCR products by running a 1:10 diluted sample on an Agilent High Sensitivity D5000 ScreenTape device or Bioanalyzer High Sensitivity DNA kit. Libraries should be smoothly distributed between 200-1000bp. Example gel images and traces are shown in **Figure 15C,D**.

    NOTE: *the Qubit High Sensivity dsDNA Assay Kit can be used here as an alternative to the Tapestation or Bioanalyzer. However, the Qubit only provides concentration and does not determine the sizing of DNA fragments which is needed to accurately calculate molarity.*

This is a safe stopping point and the libraries can be stored at 4°C for 48 hours or -20°C for weeks, or proceed to the next step.

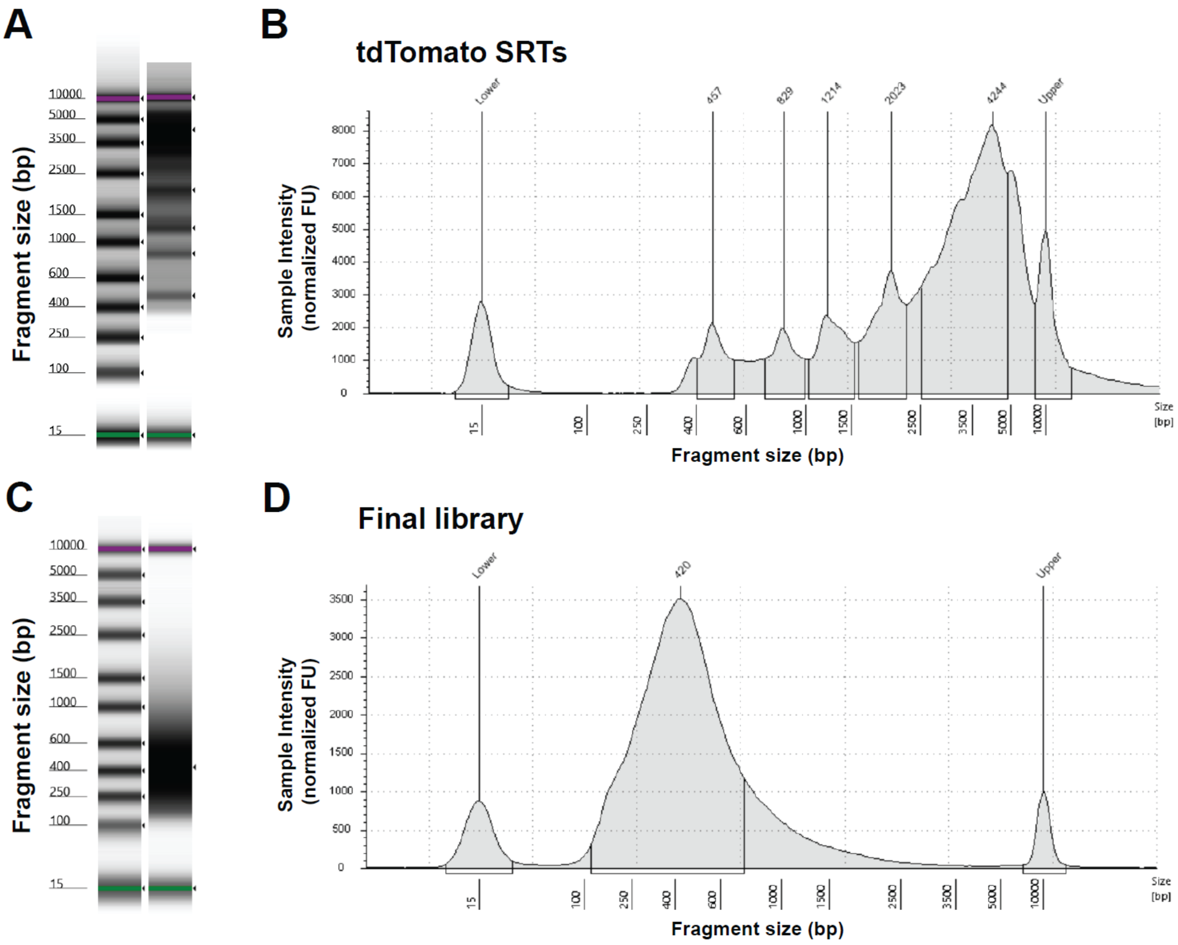


Figure 15: Representative quality control of SRTs and final library

Representative gel images and electropherogram traces for tdTomato SRTs after bead cleanup **(A, B)** and final sequencing library **(C, D)**.

2.10 Basic Protocol 4: Library pooling and sequencing

Libraries that were generated using the same protocol and uniquely indexed can be pooled into the same tube and sequenced together. Balancing and pooling libraries is relatively simple, but accurate quantification of concentration is essential for cluster generation on the sequencer flow cell. Calling Cards libraries are typically low yield libraries and thus bead-based normalization methods are not recommended. Normalizing by qPCR is a sensitive method to quantify adaptor-ligated templates and is the recommended method, however it is relatively time consuming and requires more reagents than other methods. An alternative is to pool samples equimolar based on Tapestation or Bioanalyzer values. The final combined library concentration should ideally be at least 5 nM, however low molarity pools ~1 nM have been successfully sequenced without any loss of quality.

The complete list of reagents and equipment needed for this protocol can be found in **Table 11**.

<u>Library quantification by qPCR (timing: 1.5 hr)</u>

This protocol is recommended, but not required. If libraries are being pooled based on Tapestation or Bioanalyzer molarity values, skip to the next section "Library pooling by Tapestation or Analyzer".

1. Prepare samples and plate according to kit manufacturer's instructions.

   NOTE: *For the Kapa kit, be sure to obtain the kit with the compatible passive reference dye formulation for your qPCR instrument. Similarly, for the NEBNext kit, only add an* *appropriate volume of ROX according to qPCR instrument requirements. Consult the qPCR instrument technical manual to verify ROX parameters.*

2. Run the qPCR assay according to the manufacturer's instructions.
3. *Optional:* run a melt curve analysis to detect adaptor dimers.
4. Determine the concentration of the library samples using the standard curve generated by the DNA standards.
5. For each library, use the sample's average fragment size to determine molarity.
6. Normalize each sample to the sample with the lowest molarity and pool samples equimolar for sequencing.

   NOTE: *samples with a molarity <1 nM should not be included in the pool. The final pool should ideally be at least 5 nM.*
7. Submit for Illumina next-generation sequencing.

Read	**Read1**	**i7 Index**	**i5 Index**	**Read2**
Purpose	Calling Cards insertion locus	Sample Index	Sample Index	Insert
Length**	75	10	10	75

NOTE: *a read length of 75 bp is the minimum recommended read length. Shorter reads may result in reduced alignment rates.*

Library pooling by Tapestation or Bioanalyzer (timing: 15 mins)

This protocol is only needed if libraries are not quantified by qPCR.

1. Normalize each sample to the sample with the lowest concentration and pool samples equimolar for sequencing.

NOTE: *samples with a molarity <1 nM should not be included in the pool. The final pool should ideally be at least 5 nM.*

2. Submit for sequencing.

<u>Sequencing (timing: variable)</u>

Sequencing uses standard Illumina NGS sequencers, typically provided by a local genomics core facility, MGI@GTAC, or commercial sequencing vendor. If it is available and time efficient with your local genomics core, first-pass shallow sequencing at a depth of at least 100k single or paired end 150 bp reads per sample within the pooled library is recommended to perform preliminary sequencing QC. This step ensures high quality library preparations prior to committing and investing in resources for deep sequencing. Metrics such as sequencing read counts, base call qualities, adapter content, alignment scores, and overrepresented sequences are representative of the whole pool even at low sequencing depth.

The full sequencing depth needed to recover all Calling Cards insertions in a sample is variable depending on the complexity of the library and number of insertions. In our experience, 10-25M reads per sample is sufficient to reach ~90% saturation for most samples (**Supplemental Figure 5**). To estimate sequencing saturation, unique reads and alignments within .bam files can be downsampled to simulate less sequencing. These points can be plotted, and a logarithmic growth curve can be fitted to the points to estimate the number of reads needed to reach >90% sequencing saturation.

The MiSeq, MiniSeq, NextSeq550, and NovaSeq6000 platforms have been successfully used to sequence Calling Cards libraries. The Genome Technology Access Center at the

McDonnell Genome Institute (GTAC@MGI; https://gtac.wustl.edu/) is familiar with and offers a service to sequence Calling Cards libraries on the NovaSeq6000 platform, available to any laboratory. General guidelines for in-house sequencing, or to share with your sequencing provider, are provided below.

Due to the low complexity of bases in the first 38 bases of R1, we recommend that Calling Cards libraries are not sequenced by themselves on their own lane, but rather multiplexed with other library types. Otherwise, 50% co-loading with a complex sequencing library (e.g. PhiX DNA) is required for MiniSeq, NextSeq550, and NovaSeq6000 platforms. For the MiSeq, a cluster density of 750 k/mm^2 is targeted and libraries are loaded at a concentration of ~30 nM with 10-15% PhiX. For the MiniSeq and NextSeq550 platforms, pooled libraries are loaded at a concentration of ~3 nM to aim for a cluster density of 150 k/mm^2. For the NovaSeq6000 platform, libraries were loaded at 1.5 nM and sequenced on a S4 flow cell using the Xp workflow and a 151x10x10x151 sequencing recipe according to the manufacturer protocol. Reads were demultiplexed by sample index (i5/i7) using bcl2fastq allowing for 1 mismatch without adapter trimming (see official documentation for details; https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html).

1. Transfer raw fastq files from source location to your cluster or compute environment using research data management software (e.g. Globus) or command-line tools (e.g. rsync).

2.11 Basic Protocol 5: Data analysis

The bioinformatics analysis pipeline has been built as a Nextflow workflow, which containerizes each process (Ewels et al., 2020). This not only greatly simplifies the maintenance of software dependencies, but also enables the deployment across a variety of different compute environments. A high-level overview is provided here and complete documentation can be found on the nf-core/callingcards main page (https://nf-co.re/callingcards).

> ***NOTE:*** *this pipeline is being actively developed and there is a chance that some code written here may be different than the latest version. This guide is written based on the first stable release v0.0.1. See https://nf-co.re/callingcards for the most up to date documentation.*

> ***NOTE:*** *for assistance setting up your compute environment or the Nextflow pipeline on your system, consult your system administrator or service help desk. For Calling Cards specific questions, please contact authors at* https://github.com/nf-core/callingcards/issues *for nf-callingcards or https://github.com/The-Mitra-Lab/pycallingcards/issues for py-callingcards.*

The complete list of equipment needed for this protocol can be found in **Table 12**.

<u>Install and configure nextflow (timing: variable)</u>

1. Install Nextflow 22.10.4 or later on your system. Details can be found on the official documentation (https://www.nextflow.io/docs/latest/index.html).
2. Install a container engine of choice (e.g., Docker, Singularity, Shifter, Podman, or Charliecloud). For this guide, we will be using Singularity.

3. Test the installation and configuration with a minimal data set. This only tests the proper configuration of the pipeline and local compute environment. Note that some config profiles may be needed to instruct Nextflow how to fetch the required software. In this example, test and singularity profiles are chained. The path to a custom local_config file is specified in the -c parameter. It is recommended to check if a config file is already available for your cluster from nf-core/configs (https://nf-co.re/usage/configuration). If one is not available, a tutorial for writing one can be found here (https://nf-co.re/docs/usage/tutorials/step_by_step_institutional_profile).

```
$ nextflow run nf-core/callingcards \
     -profile test,singularity \
     -c /ref/jdlab/software/nextflow/wustl_htcf.config \
     --outdir /scratch/jdlab/allen/test_run
```

If Nextflow is correctly configured, the pipeline should complete in a few minutes. The resulting test_run directory should have a file structure like that found in **Figure 11**. Let's check the qBED file to make sure the output is correct by counting the number of recovered insertions, or hops. There should be 101 insertions.

```
$ wc -l test_run/human_AY53-1_50_T1/hops/human_AY53-1_50_T1.qbed
101 test_run/human_AY53-1_50_T1/hops/human_AY53-1_50_T1.qbed
```

Let's see where insertions were found in the genome by looking at the first 10 lines of the qBED file. Each line of the qBED file corresponds to a unique insertion. The columns represent chromosome, start coordinate, end coordinate, sequencing depth of that insertion, strand, and SRT barcode.

```
$ head test_run/human_AY53-1_50_T1/hops/human_AY53-1_50_T1.qbed
chr1    35235   35239   3       -       CTAG
chr1    76228   76232   2       +       CTAG
chr1    200400  200404  1       +       CTAG
chr1    807528  807532  2       +       CTAG
```

```
chr1    807619  807623  1       +       CTAG
chr1    810640  810644  1       +       CTAG
chr1    812850  812854  1       +       CTAG
chr1    815191  815195  1       -       CTAG
chr1    838289  838293  3       -       CTAG
chr1    1144424 1144428 1       -       CTAG
```

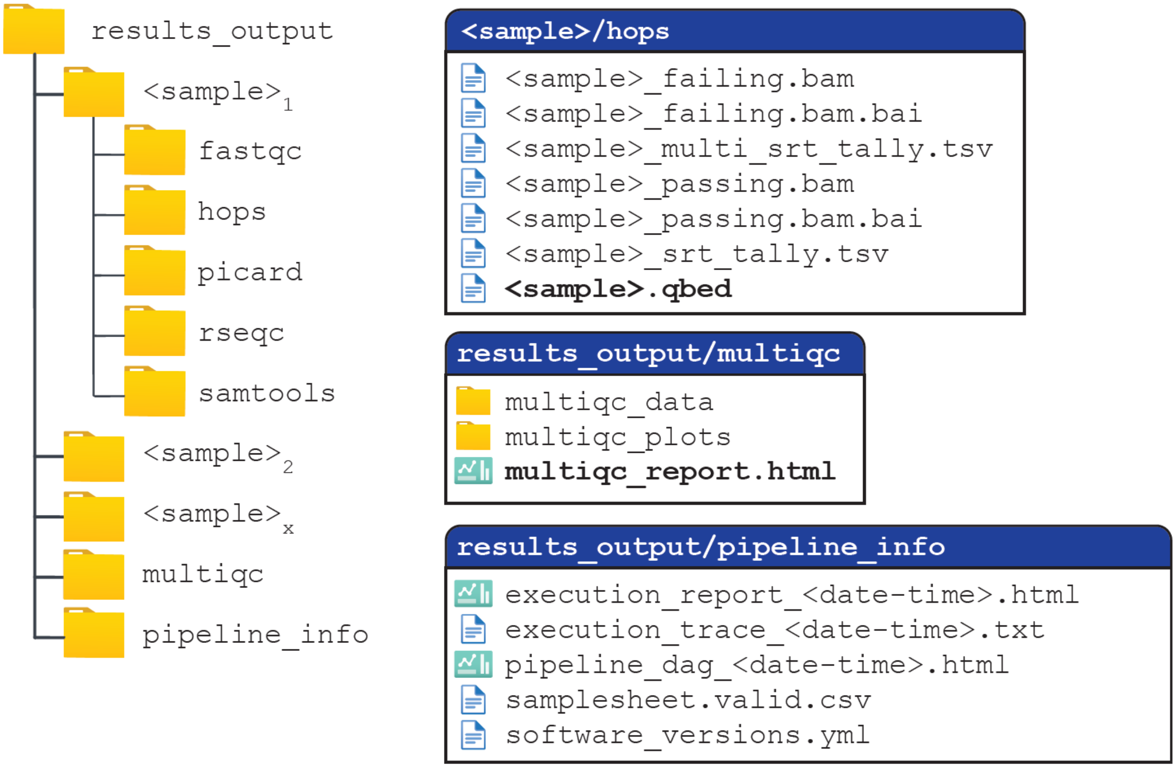


Figure 16: Generic file structure of nf-core/callingcards pipeline output

The tree file structure upon completion of the pipeline is shown on the left. The name of the top-level directory is specified as the --ouput parameter within the params.json file. For each line in the samplesheet.csv, there will be a separate directory with the sample name provided in the first field. Within each sample directory, /hops contain the .qbed file listing Calling Cards insertions. Intermediate files that may be helpful in troubleshooting can also be found here. A html report can be found at multiqc/multiqc_report.html, where basic sequencing and alignment statistics are aggregated and displayed. Diagnostic reports and information can be found in the pipeline_info directory.

Run nf-core/callingcards on your own data to generate qBED files (timing: variable)

Now with the pipeline set up, you are ready to run a full dataset. This can be your own data, or we have also provided an example set that is publicly available at Gene Expression Omnibus (GEO) with the accession number GSE223926. The code below is shown using this example.

1. Using a text editor, create a samplesheet.csv with each row containing comma separated fields described in the table below. Each row represents an independent sample to be analyzed.

Fields	Description
sample	Custom sample name. This entry will be identical for multiple sequencing libraries/runs from the same sample. Spaces in sample names are automatically converted to underscores (_).
fastq_R1	Full path to FastQ file for Illumina short reads 1. File has to be gzipped with the extension .fastq.gz or .fq.gz.
fastq_R2	*(Optional).* Full path to FastQ file for Illumina short reads 2. File has to be gzipped with the extension .fastq.gz or .fq.gz.
barcode_details	Full path to the barcode details json file for a given sample.

Below is the samplesheet.csv for this example.

```
sample,fastq_1,fastq_2,barcode_details
cortex_rep1,raw/cortex_rep1_R1.fastq.gz,,barcode_details-TAG.json
hindbrain_rep1,raw/hindbrain_rep1_R1.fastq.gz,,barcode_details-TAG.json
midbrain_rep1,raw/midbrain_rep1_R1.fastq.gz,,barcode_details-TAG.json
cortex_rep2,raw/cortex_rep2_R1.fastq.gz,,barcode_details-TAG.json
hindbrain_rep2,raw/hindbrain_rep2_R1.fastq.gz,,barcode_details-TAG.json
midbrain_rep2,raw/midbrain_rep2_R1.fastq.gz,,barcode_details-TAG.json
cortex_rep3,raw/cortex_rep3_R1.fastq.gz,,barcode_details-TAG.json
hindbrain_rep3,raw/hindbrain_rep3_R1.fastq.gz,,barcode_details-TAG.json
midbrain_rep3,raw/midbrain_rep3_R1.fastq.gz,,barcode_details-TAG.json
```

Note that here, we are not using the fastq_R2 field, so it is intentionally left blank.

2. Create a barcode_details.json that provides SRT barcode information. An example json for a standard experiment using all available SRT barcodes is provided below (see table below for detailed description of each field). More complex experimental designs such as multiplexing transcription factors with subsets of SRT barcodes will require editing the SRT barcode array within the components object. Note that the 3 bp OM-PB primer barcode will need to be specified in the r1_pb object. In this example, we use the 'TAG' barcode.

Fields	Description
batch	*(Optional).* Name of the experiment, e.g. experiment_1234.
tf	*(Optional).* Either ID or symbol of the TF.
r1	Keys correspond to barcode component names. Each component is a map with keys trim which is boolean. Set to true to trim off this portion of the read (NOTE: not used in mammals pipeline. This functionality is instead controlled by UMITools). Index specifies where this component occurs in the read.
r2	same as r1.
components	Each key corresponds to a component in r1 or r2 and must be preceded by r1 or r2 as appropriate. Each value is a map where the key map is required and lists the expected sequence(s) for that component. The key r1_pb holds the value of the 3bp OM-PB primer barcode. Optional additional keys are match_allowance to allow mismatches >0, bam_tag which is used to add the sequence extracted from the read to the aligned read, match_type which controls how the sequences are matched (default is edit_distance and appropriate in most circumstances), and require is a boolean which, when set to false, allows any number of mismatches in the component without penalty.
insert_seq	A list of sequences expected to be present on the 5' end of the aligned read.
max_mismatch	The maximum number of mismatches allowed in a barcode. By default, this is equal to the sum of the mismatches allowed on each component. Set this to less than that sum to fail barcodes which pass component checks but have more than x number of mismatches overall.

```
{
       "batch": "",
       "tf": "",
       "r1": {
              "pb": {"trim": true,
                     "index": [0,3]},

              "ltr1": {"trim": true,

                     "index": [3,28]},
              "srt": {"trim": true,
                     "index":[28,32]},
              "ltr2": {"trim": true,
                     "index": [32,38]}
              },
       "r2":{},
              "components": {
                     "r1_pb": {"map":["TAG"],
                            "match_allowance": 0,
                            "bam_tag": "PB"},

                     "r1_ltr1": {"map": ["CGTCAATTTTACGCAGACTATCTTT"],
                            "match_type": "edit_distance",
                            "match_allowance": 0,
                            "require": true,
                            "bam_tag": "L1"},
                     "r1_srt": {"map": ["CTAG", "CAAC", "CTGA", "GCAT", "GTAC",
                                        "CACA", "TGAC", "GTCA", "CGAT", "CTCT",
                                        "GAAG", "TCGA", "CATG", "GTTG", "CTTC",
                                        "GCTA", "GAGA", "GTGT", "CGTA", "TGGT",
                                        "GGAA", "ACAC", "TCAG", "TTGG", "CAGT"],

                                   "match_type": "edit_distance",

                                   "match_allowance": 0,
                                   "require": true,
                                   "bam_tag": "ST",
                                   "annotation": true},
                     "r1_ltr2": {"map": ["GGTTAA"],
                            "match_type": "edit_distance",
                            "match_allowance": 0,
                            "require": true,
                            "bam_tag": "L2"}
                     },
       "insert_seq": ["TTAA"],
       "max_mismatch": 0
}
```

3. Create a params.json file which will save the input parameters in json format. See table below for additional details of the json fields. An example is provided below using the default bwamem2 aligner and thus we need a bwamem2-indexed reference genome. A

list of all available modules can be found at https://nf-co.re/modules. If another aligner is used, provide the appropriate indexed reference genome.

Fields	Description
input	Full path to the samplesheet.csv.
outdir	Full path to the output directory where the analysis file will be written to.
aligner	Specification for which alignment software to use. Available options are: bwa, bwamem2, bowtie, or bowtie2.
fasta	Full path to the reference genome.
fasta_index	Full path to the indexed reference genome.
gtf	Full path to the reference genome annotation file. File has to be gzipped with the extension .gtf.gz.
bwamem2_index	Full path to the bwamem2 index reference genome.
save_intermediate	Boolean (true/false) to determine whether intermediate analysis are saved. This is typically not needed, however this can be helpful for troubleshooting issues.

```
{
"input" : "samplesheet.csv",
"outdir": "brain_region_analysis",
"aligner": "bwamem2",
"fasta"                                                                  :
"/ref/jdlab/data/genomes/Mus_musculus/UCSC/mm10/Sequence/WholeGenomeFasta/genom
e.fa",
"gtf"                                                                    :
"/ref/jdlab/data/genomes/Mus_musculus/UCSC/mm10/Annotation/gencode.vM23.annotat
ion.gtf.gz",
"fasta_index" : "/ref/jdlab/data/bwamem-2_index/bwamem2_mm10/genome.fa.fai",
"bwamem2_index" : "/ref/jdlab/data/bwamem-2_index/bwamem2_mm10",
"save_intermediate" : false
}
```

4. To run the pipeline with default parameters, use the following command (see https://github.com/nf-core/callingcards/blob/master/conf/default_mammals.config for more details).

```
$ nextflow run nf-core/callingcards \
      -profile default_mammals,singularity \
      -c /ref/jdlab/software/nextflow/wustl_htcf.config \
      -params-file /scratch/jdlab/allen/test_run
```

The Nextflow pipeline checks the input samplesheet.csv, params.json, indexes the reference genome if not already done and in the working directory, prepares the reads, extracts SRT barcodes, trims adapter sequences, aligns the reads to a reference genome, processes the aligned reads to map the insertion locus, and outputs a qBED file (**Figure 4B**). First let's organize the qBED files by copying them from their individual sample directories into a common directory.

```
$ find . -iname "*.qbed" -type f -exec cp {} ./qbeds \;
```

Now let's see the number of recovered insertions by counting the number of lines of each qBED file.

```
$ wc -l *.qbed
   810198 cortex_rep1.qbed
   784854 cortex_rep2.qbed
   590342 cortex_rep3.qbed
   232303 hindbrain_rep1.qbed
   140234 hindbrain_rep2.qbed
   155584 hindbrain_rep3.qbed
   655623 midbrain_rep1.qbed
   373405 midbrain_rep2.qbed
   202208 midbrain_rep3.qbed
  3944751 total
```

This file is then used to call peaks, calculate significance, motif enrichment, and other downstream analyses.

5. If we examine the first 10 lines of cortex_rep1.qbed, the outputs should match. Each line of the qBED file corresponds to a unique insertion. The columns represent chromosome, start coordinate, end coordinate, sequencing depth of that insertion, strand, and SRT barcode.

```
$ head cortex_rep1.qbed
chr1    3119238 3119242 29      -       TTGG
chr1    3121337 3121341 16      +       CATG
```

```
chr1    3168440 3168444 7       +       GTTG
chr1    3198446 3198450 24      +       GTCA
chr1    3198534 3198538 1       +       CTAG
chr1    3199307 3199311 45      -       CAAC
chr1    3200147 3200151 3       -       GAAG
chr1    3200511 3200515 17      -       TGAC
chr1    3200561 3200565 14      -       CACA
chr1    3207813 3207817 3       -       TGGT
```

Analysis of qBED files (timing: variable)

With a qBED file of unique insertions, the next steps are to call peaks to identify genomic regions enriched with Calling Cards insertions, perform differential peak analysis, and identify nearby genes. The concept of Calling Cards peak calling is similar to analyzing chromatin immunoprecipitation-sequencing (ChIP-seq) experiments with some distinct differences. Rather than using aligned reads, the insertions themselves will be used for peak calling. Additionally, the distributions of read and insertion densities are distinct, so we have developed Py-callingcards (https://github.com/The-Mitra-Lab/pycallingcards) to analyze Calling Cards data. The repository contains tutorials using example data and complete API documentation.

Visualization of Calling Cards data (timing: variable)

The WashU Epigenome Browser (http://epigenomegateway.wustl.edu) (Li et al., 2019) provides a streamlined platform to visualize Calling Cards data using the qBED format, which reports transposon insertions as discrete points along the genome (x-axis) and the number of reads associated with that insertion (y-axis) (Moudgil et al., 2020a). This creates a scatterplot-like depiction that can be used to visualize densities of Calling Cards insertions across the genome (**Figure 3A**). To load qBED tracks into the WashU Epigenome Browser, the file must first be sorted, compressed, and indexed.

6. Install a stable release of HTSlib (https://github.com/samtools/htslib).

7. The qBED file needs to be properly sorted by chromosome, start coordinate, and end coordinate prior to viewing on the genome browser. If not already sorted, the following command will sort the file and its output will be saved as a new file.

```
$ sort -k1V -k2n -k3n cortex_rep1.qbed > cortex_rep1_sorted.qbed
```

8. Block compress the qBED file. The output file will be a new file with the .gz extension and the original will be removed.

```
$ bgzip cortex_rep1_sorted.qbed
```

9. Index the compressed file in bed format. This will create cortex_rep1_sorted.qbed.gz.tbi.

```
$ tabix -p bed cortex_rep1_sorted.qbed.gz
```

10. Store the compressed qBED (*.qbed.gz) and index (*.qbed.gz.tbi) together locally or in a web-accessible directory.
11. On the browser, select the desired reference genome. In this example, we will use mm10.
12. Upload the compressed qBED and index pair together to the browser. Select qBED as the track file type. If you are using local files, select Tracks > Local Tracks. If using remote files, select Tracks > Remote Tracks. Additional tracks and track types (ie. BED, bedGraph, bigWig, etc.) can also be compressed, indexed, and simultaneously loaded in a similar way by repeating steps 7-9).
13. Adjust track parameters such as color, height, scale, opacity, and marker size from within the browser interactively.

2.12 Commentary

Historical development of the protocol

Calling Cards was first developed and used to study TF binding in *Saccaromyces cerevisiae*. This yeast method uses a Ty5 integrase and engineered TF-Sir4 protein fusions to leave permanent “Calling Cards” where the engineered Sir4 has bound to the genome (Wang et al., 2008). To expand upon this technology, the transposase system was switched to the *piggyBac* transposon/transposase system due to its high transposition efficiency and activity across species (yeast, zebrafish, insects, mouse, and human), broadening the technology’s applicability (Wang et al., 2012). In these early protocols, to recover Calling Cards insertions, genomic DNA was cleaved with restriction endonucleases, fragments were circularized, and amplified using inverse PCR from the transposon, and sequenced to capture transposon-genome junctions. The genomic sequences were then mapped to identify the location of the insertions. However, when Calling Cards began to be ported to complex systems like the mammalian brain, the small number of insertions per cell (<30) and the scale of the mammalian genome (~3 Gb), made recovery of sufficient numbers of transposons challenging.

The development of self-reporting transposons (SRTs) solved this problem by having each insertion produce multiple copies of its own transposon-genome junction by transcribing RNA from within the transposon. Thus, insertions could be mapped from RNA with base-pair resolution using specialized RNA sequencing (RNA-seq). In parallel, and with the same starting material, standard RNA-seq can provide a simultaneous readout of gene expression from the same cells, in a manner compatible with single-cell sequencing technologies.

The SRT is a *piggyBac* transposon that contains an EF1α promoter driving a tdTomato reporter without a polyadenylation signal. When inserted in the genome, RNA polymerase II will transcribe the SRT (tdTomato) into the flanking genomic region until it reaches a cryptic polyadenylation signal or genomic polyadenine stretch. It was found that RNA-based libraries were much more sensitive than the DNA-based method and recovered nearly 50% more insertions (Moudgil et al., 2020b). Since each insertion produces multiple RNA copies of itself, one limitation of using this technology (outside of single-cell applications) was that one cannot differentiate 1) multiple independent insertions at the same locus across cells from 2) one insertion at the same locus whose RNA was sequenced multiple times. Initially this was resolved by preparing multiple independent aliquots of adjacent tissue for RNA extraction, followed by barcoding each sample by PCR (Cammack et al., 2020), thus exponentially amplifying the amount of sample handling. However, we more recently resolved this issue by introducing transcribed barcodes into the SRT itself. This feature dramatically reduces the required number of replicates, decreases cost and labor requirements, and increases the sensitivity of detecting multiple insertions at the same locus (Lalli et al., 2022). Finally, we have found that higher expression and copy number of the donor transposon is beneficial to maximize the number of insertions per cell, as transposons are more limiting than transposase activity (Yusa et al., 2011) (**Supplemental Figure 2A**), and now can recommend optimized ratios in this protocol.

The enhanced sensitivity of the SRTs relative to inverse PCR also enabled the application of Calling Cards to address questions in complex tissues, like the mammalian brain. Thus, to direct Calling Cards to genetically defined cell populations, we developed Cre-dependent hyPB transpose systems. We initially adopted a FLEx (or "fl̲ip-ex̲cision") switch vector design so that a cassette encoding hyPB is in the antisense orientation and should not produce functional protein.

In the presence of Cre (e.g., when delivered by AAV into a Cre-expressing mouse brain), recombination occurs, the sequence is flipped into the sense orientation, driving hyPB expression. However, with FLEx Calling Cards, we had noticed a low level of hyPB activity in Cre negative animals, suggesting some transcription of hyPB from the antisense strand of the AAV. To eliminate this background transposition, we redesigned the FLEx cassette to insert an intron into the middle of hyPB and position key LoxP sites within it. The vector thus contains the front half of the hyPB protein antisense to the second half of the protein. Thus, without Cre, both sense or antisense strands would not produce functional protein. This "FrontFlip-hyPB" construct was shown to effectively eliminate background transposition (Cammack et al., 2020), and the general design may be helpful in suppressing anti-sense leakiness in other Cre-dependent enzyme systems.

Thus, Calling Cards using barcoded SRTs, optionally combined with cell-type specific designs, are the latest reagents that can be used to record BRD4-bound enhancer usage or TF-DNA binding events in specific cell types in complex systems.

Applications of the method

Calling Cards has been successfully used to record BRD4-bound enhancer usage with unfused hyPB or transcription factor binding with custom TF-hyPB fusions *in vivo* and *in vitro* (Cammack et al., 2020; Kfoury et al., 2021; Moudgil et al., 2020b; Wang et al., 2012, 2008). Using this method, it was found that sex-biased BRD4-bound enhancer activity in glioblastoma (GBM) underlies sex differences in stem cell function and tumorigenicity in GBM. Pharmacological or genetic inhibition of BRD4 in male and female patient-derived GBM cell lines revealed that male GBM cells are more sensitive to BET inhibition and has important clinical implications (Kfoury et al., 2021). In addition, we have included a vignette here showing how Calling Cards can be used to record regional differences in enhancer usage across brain development. Additionally, the development of Cre recombinase dependent constructs with reduced background (ie. FrontFlip-hyPB) or use of cell type-specific promoters enables targeted recording in genetically defined populations (Cammack et al., 2020). While this protocol derives from our experience mainly on mouse neural tissue and cell types, tailoring Calling Cards for a particular cell type, tissue, or model organism may require optimization of reagent delivery, RNA isolation, and some PCR steps during library preparation. We also note that the tdTomato reporter within the SRT is useful to not only visualize cells but can also be used to enrich cell populations with FACS. We have developed a collection of modular Calling Cards reagents that can be adapted to be used across a wide range of applications (see **Table 5** for list of available reagents). Calling Cards is positioned to generate unique datasets that enable the analysis of observed current cell states with historical molecular and epigenetic states.

Comparison with other methods

There is a myriad of genomic methods to investigate the epigenome, transcriptome, and chromatin state, and the assay of choice should be determined by the biological question (**Table 1)**. While many methods produce snapshots of states at tissue harvest, Calling Cards is one of two methods we know of that produces data that represents a cumulative catalog of protein-DNA interactions over time, using a genetically encoded system. The other method, DamID (Vogel et al., 2007), identifies binding sites of a protein by using a DNA methyltransferase (DAM) fusion protein, which methylates an adenine within a GATC sequence in close proximity to the binding site. Since adenine methylation does not occur naturally in eukaryotes, mapping the location of the methylated adenine nucleotides is then interpreted as a proxy for the binding site. The cumulative view of activity can be an appealing method to obtain the ground truth of TF binding or enhancer usage as this is recorded over time and is not sensitive to collection time. Further, since it is genetically encoded, it can be delivered to specific cell types. Since Calling Cards starts from an RNA sample, the mRNA transcriptome can be analyzed in parallel from the same cells and input material, allowing a measure of the final transcriptional state in parallel to the recorded TF binding profile. This approach is thus unique from DamID datasets, which cannot provide quantitative measures of transcription in addition to the record of DAM fusion protein-DNA interactions. RNA Pol II can be fused with DAM to profile transcription start sites and gene expression, but the ability to simultaneously profile a TF is lost.

In comparison to non-recording methods of assessing DNA-protein interactions, in test cases examined so far, Calling Cards data is concordant with ChIP-seq data (Moudgil et al., 2020b). Thus, Calling Cards offers a powerful complementary approach to ChIP-seq/CUT&RUN and together, the resulting datasets can provide a more complete picture of TF biology and gene regulatory elements. Furthermore, experiments that require screening multiple time points, multiple TFs, or association of historical molecular events with eventual cell states are applications where one might prefer Calling Cards over ChIP-seq. However, for systems where genetic delivery of reagents is not an option (e.g., human brain), non-Calling Cards methods are required. Finally, Calling Cards has been demonstrated to work both *in vitro* and *in vivo* and is a versatile tool that can be broadly used in models to study development and diseases that accumulate changes over time.

Table 1: Comparison with other genomic methodologies

	ATAC-seq	**ChIP-seq**	**CUT&RUN**	**DamID**	**Calling Cards**
Enzyme type	Tn5	n/a	Protein A-MNase fusion	DNA adenine methyltransferase (fused or unfused)	*piggyBac* transposase (fused or unfused)
Sequence bias	Tn5 bias (G/C rich sequences)	n/a	n/a	GATCs	TTAAs
Antibody needed	No	Yes	Yes	No	No
Target specific cell population	Possible, FACS enrich based on cell type-specific marker	Possible, FACS enrich based on cell type-specific marker	Possible, FACS enrich based on cell type-specific marker	Cell type-specific expression of Dam fusion proteins with Cre	Cell type-specific expression of *piggyBac* transposase with Cre; FACS for tdTomato
Input material needed	50,000 nuclei	1M for abundant proteins (e.g. RNA Pol II)	100,000-500,000 cells	10,000 for abundant proteins (e.g. RNA Pol II)	20,000 cells or nuclei
Readout	Snapshot of chromatin state	Snapshot of TF binding or histone modifications	Snapshot of TF binding or histone modifications	Longitudinal recording of protein-DNA interactions	Longitudinal recording of protein-

					DNA interactions and transcriptome
Multi-ome readout	No	No	No	Yes, parallel transcriptome when using Pol2-Dam fusion	Yes, parallel transcriptome
Technical considerations	Data quality is dependent on high quality cell/nuclei preps and handling	Quality is dependent on abundance of target protein and quality of antibody	Quality is dependent on abundance of target protein and quality of antibody	Resolution limited to GATC motif in the genome and Dam can be toxic	Multiple replicates significantly increase power of calling differentially hopped regions
Library prep duration	1d	1-2d	1-2d	2-3d	2d
Sequencing depth	10-20M reads per sample	10-40M reads per sample	~5M reads per sample	25-50M reads per sample	10-25M reads per sample
Compatibility with single cell platforms	Yes	Yes, but only for abundant targets	Yes, but only for abundant targets	Yes, but only for abundant targets	Yes

Understanding Results

Interpretation of quality control checkpoints

Considering the protocol length and cost of reagents, careful quality control throughout is recommended to ensure high quality datasets and efficient use of reagents and time. During the development and optimization of the protocol, we have identified key parameters that are important for the successful generation of high-quality sequencing libraries and data. These include the amount of starting material, amount of DNA input into each PCR step, amount of final library, and the number of recovered insertions. Interestingly, we found relatively permissive limits of variation for QC steps, meaning that the protocol is robust at recovering insertions given proper library preparation steps. We generally found that maximization of input material at each step that allows a range of input leads to the most recovered insertions. For exceptionally complex libraries, multiple library preparations from the same starting material may be necessary to capture their full diversity.

The first QC checkpoint quantifies the integrity of the purified total RNA. A minimum RNA integrity number (RIN) of 8 with minimal signs of degradation is recommended for most cases, however this can vary depending on the type of sample. *In vitro* cell samples can be as high as 10, whereas FFPE samples will typically be low. RINs lower than 7 generally do not sequence well, however purification/enrichment kits can be used prior to the library preparation. In addition to RIN, genomic DNA contamination can be seen as an additional peak above the 28S band or as any unexpected signal between the rRNA peaks. The genomic DNA can not only cause inaccurate RNA quantification, but can also interfere with downstream steps, so an additional DNase I treatment (in addition to the DNase I treatment during RNA cleanup) is recommended prior to proceeding further.

The next QC step is the concentration quantification of single stranded cDNA following the first strand synthesis. The presence of the SMART_dT18VN RT primer during the subsequent PCR amplification of SRTs has been found to increase nonspecific products likely due to priming to endogenous polyA stretches. Thus, complete removal of primer using a simple spin column-based PCR purification kit is an important step prior to PCR amplification of SRTs. Library density assessment of tdTomato containing cDNA fragments using quantitative RT-PCR is an optional but valuable QC step. Here, the samples can be initially screened for the abundance of tdTomato and determined if certain samples not passing a threshold should be omitted from subsequent library preparation steps (**Figure 13**). In our experience, $C_{T,tdTomato}$>30 and tdTomato expression values normalized to β-actin that are <0.1 will still make libraries, however, the number of recovered insertions is limited. The specific values will likely vary based on sample type and transfection/transduction efficiency and will require some initial empirical testing, but the potential benefit of saving labor and reagent costs is worthwhile. Accurate quantification of ssDNA using a fluorometric method is highly recommended to maximize the amount of input material (up to 100 ng) that goes into the next PCR reaction to amplify tdTomato containing SRTs (**Figure 14**).

After bead cleanup of SRTs, the electropherogram of the amplified SRTs should resemble **Figure 15A,B**. A mostly smooth distribution of products with a bias towards larger products should be expected from 300-5000 bp. Some banding has been observed to occur and has not been found to affect library quality, however, strong and distinct bands without larger PCR products likely indicates a low number or diversity of SRTs, or can be artifacts resulting from PCR overamplification (**Supplemental Figure 7A, B**).

The final QC step is to quantify the library concentration and fragment size distribution. The tagmented library should be a smooth distribution between 200-1000 bp (**Figure 15C, D**). Presence of a ~100 bp band indicates insufficient removal of the OM-PB primer and another round of bead cleanup is recommended to remove the primer as this can have a detrimental effect on sequencing (**Supplemental Figure 7C, D**)

Troubleshooting

Troubleshooting advice and suggestions for commonly observed issues can be found in the table below.

Problem	**Possible reason**	**Comments and suggestions**
RNA is degraded or low quality (low RIN, low A260/A280 ratio)	Starting material was improperly handled/stored	Snap freeze tissue in liquid nitrogen. If starting with a frozen sample, do not let the sample thaw prior to RNA cleanup. If starting with fresh tissue, dissect quickly and minimize time to homogenization.
	RNase contamination	Clean all surfaces, equipment, and gloves with RNase removing detergents.
Genomic DNA contamination	DNase treatment was insufficient	Repeat DNase treatment or switch from an on-column to an in-tube DNase I treatment method.
Low RNA yield	Overloaded homogenization	Homogenize samples in 1ml TRIzol per 100mg tissue. Scale volume up if needed.
	Residual organic solvent contamination	Precipitate RNA and increase the number of ethanol washes.

	Insufficient homogenization or lysis	Ensure no visible cell aggregates or tissue chunks remain during homogenization.
Low cDNA yield	Degraded RNA	Check RNA quality by BioAnalyzer or Tapestation.
	Residual contamination from RNA purification steps	Precipitate RNA and wash with ethanol to remove EDTA, guanidinium, and proteases.
	Low amount of RNA input	Increase the amount of starting RNA by combining biological replicates.
Low SRT yield	Low expression of Calling Cards constructs	Increase the amount of plasmid or virus used to deliver Calling Cards reagents.
Distinct banding in SRT QC	Low complexity library	Increase the amount of plasmid or virus used to deliver Calling Cards reagents.
No product after bead cleanup	Failed bead cleanup	Use fresh 80% ethanol and do not overdry the beads. Do not disturb bead pellet during washes.
Low concentration (<1nM) of final library	Insufficient PCR amplification	Increase the number of cycles for the indexing PCR step.
	Failed bead cleanup	Use fresh 80% ethanol and do not overdry the beads. Do not disturb bead pellet during washes.
Final library fragment size distribution not 200-1000bp	Failed tagmentation	Undertagmentation yielding larger products is a result of too much input material. Overtagmentation leads to smaller fragments and is a result of too little input material. Verify 600pg DNA input using a fluorometric quantification method.
Sequencing reads are the OM-PB primer sequence followed by G’s	OM-PB indexing primer carryover	Beware of a ~100bp band in the electropherogram QC of the final library. Repeat bead cleanup and QC.

Low proportion of reads with adapter	PCR mispriming due to low library density	Combine biological replicates to increase signal (SRT fragments) to background.
Presence of overrepresented sequences (>0.1% of library) that is not biologically relevant	PCR mispriming (may correspond to prominent bands in SRT QC), genomic DNA contamination, primer carryover	Ensure proper cleanup of DNA with silica column or magnetic beads.
Low number of recovered insertions	Low expression of Calling Cards constructs	Increase the amount of plasmid or virus used to deliver Calling Cards reagents.
	Low library complexity	If starting from a rare cell population, purify the target population by FACS or enrich their RNA by a method like TRAP (Heiman et al., 2014). Otherwise add more biological replicates
Poor sequencing coverage of each insertion (<5 reads/unique insertion)	Insufficient sequencing depth	Sequence deeper.

Computational analysis of Calling Cards data

Analysis can be broken down into three main stages: 1) generation of the qBED files, 2) peak calling and insertion counting, and 3) downstream analysis. The workflow and steps involved for the first two stages are diagrammed in **Figure 4B**. Further downstream analysis (e.g., differential peaks between samples, motif enrichment, gene ontology analysis) is not covered in this protocol. First, the raw reads are prepared by filtering for reads containing SRT barcodes with UMItools, trimming adapter sequences with trimmomatic, and standard quality control such as per base sequence quality scores and per base sequence content with FastQC. The reads are then aligned and the SRT insertions are mapped to the reference genome. The modularity of the nextflow pipeline enables the end user to easily select the desired alignment software. The default aligner is bwamem2. A full list of preconfigured nf-core modules can be found at https://nf-co.re/modules. In the same file, the header of aligned reads is tagged with the OM-PB barcode, index1/i7, index2/i5, and SRT barcode, allowing us to identify the insertions' sample of origin. The resulting bam file is indexed and parsed into a qBED file which lists the unique insertions. This file can be used as input into peak callers to call peaks for genomic regions that are enriched with Calling Cards insertions.

Sequencing saturation

Following an initial sequencing of a sample, one can determine if deeper sequencing would be useful to recover more binding events. For Calling Cards data, sequencing saturation is a measure of how many times an insertion has been read. This provides a way to estimate how much new information (ie. unique insertions) is likely to be gained by sequencing deeper. Within each qBED file, the fourth column represents the number of reads associated with each insertion. To evaluate saturation, bam files were downsampled across set ratios (0.001, 0.003, 0.01, 0.03, 0.1, and 0.3 of the full library) using samtools to simulate shallower sequencing.

```
$ samtools view \
-b \
--subsample-seed 123 \
--subsample <ratio> \
cortex_rep1_passing.bam > cortex_rep1_passing_sampled_<ratio>.bam
```

An example plot of downsampled samples with varying amounts of insertions is shown in **Supplemental Figure 5A**. A final recommended sequencing depth can be estimated where the plots start to asymptote. Peaks were called on each of the downsampled files (**Supplemental Figure 5B, C**). While the size of peaks was not as sensitive to the number of insertions, the number of called peaks is directly proportional to the number of insertions and reads, as expected. It is also evident that the number of reads per insertion also increases with the number of reads (**Supplemental Figure 5D**). A left-skewed cumulative density indicates under-sequencing as there are many insertions that were only sequenced once. A shift to the right with a relatively low number of insertions with low coverage demonstrates that the sample is approaching sequencing saturation. With deeper sequencing, the resolution of the called peaks increases. This is evident based on the observation that the fraction of overlapping peaks does not change upon downsampling, despite the dramatic increase in the number of peaks (**Supplemental Figure 5E**).

Assessing reproducibility of across replicates

To assess the reproducibility of Calling Cards data across biological and/or technical replicates, the qBED files for each experimental condition or genotype can be concatenated to create a file containing all insertions. Each insertion is tagged with unique index sequences and can be used to identify the sample of origin. Peaks are called on the aggregated insertions. The similarity between a pair of replicates can be analyzed by plotting the number of insertions (normalized to sequencing depth) found within each peak region from each of the replicates as a scatterplot, with values from one replicate on the x axis, and the other on the y axis. To assess multiple replicates, a scatterplot matrix can be plotted. Replicates with high similarity should have a symmetrical linear correlation (**Supplemental Figure 4**).

Peak calling and differential peak analysis

After the insertions have been recovered and listed in the qBED file format, the next steps are to call peaks to identify genomic regions enriched with Calling Cards insertions, perform differential peak analysis to identify differences between experimental conditions, and identify nearby genes. The Py-callingcards package (https://github.com/The-Mitra-Lab/pycallingcards) has been deployed to perform these steps. Complete documentation and tutorials with example datasets are also provided.

Multi-omics data integration

Using a TF-hyPB fusion transposase, Calling Cards records cumulative TF binding events. If homologous high-quality ChIP-seq datasets are available for the TF of interest, Calling Cards data can be directly compared to assess the degree of overlap between the orthogonal methods. Depending on how similar the ChIP-seq experimental conditions are, one should expect a reasonable amount of intersection, keeping in mind that Calling Cards data may yield more peaks given that it records over time. When using Calling Cards with the unfused *piggyBac* transposase, BRD4-bound enhancers are recorded and can be similarly validated with BRD4 ChIP-seq. Additionally, it can be informative to compare these putative enhancers with ATAC-seq, H3K27ac, and H3K4me1 ChIP-seq peaks which reveal open chromatin, active enhancers, and promoters (Cammack et al., 2020), as BRD4 ChIPseq can correspond somewhat to these. Generally, these Calling Cards peaks should be de-enriched in regions with repressive H3K27me3 marks. The combination of these datasets can be used to establish enhancer-gene regulatory interactions, which can be further confirmed with chromosome conformation capture technologies such as Hi-C or HiChIP.

Comparison to RNA-seq data is also sometimes informative especially when looking at promotor-localized peaks (and keeping in mind the usual caveats and challenges of linking enhancer regions to specific genes). Differentially expressed genes can be identified from bulk RNA-seq data using DEseq2 (Love et al., 2014) and correlated with TF binding or enhancer usage Calling Cards peaks. Associating these TF binding-gene pairs can be useful to those studying gene regulation. Functions to perform these analyses are also contained within Py-callingcards.

Files to submit for publication

Finally, with analysis completed, as with all high throughput sequencing data, Calling Cards data should be uploaded to a publicly accessible repository such as Gene Expression Omnibus (GEO) concurrent with publication of corresponding manuscripts. To be maximally useful to the community, each submission should include a metadata spreadsheet, all raw FASTQ files prior to any processing (QC filters, adapter trimming, etc.), and processed data files. These can include insertions (qBED), genomic coordinates of called peaks (BED), density tracks of insertions (bedGraph), insertions per peak counts matrix for all samples in the study, and a list of differentially hopped regions. The qBED, BED, and bedGraph files are useful as they can then be used by any user to visualize the data on the WashU Epigenome Browser, or for comparison to their own datasets. We also recommend including a summary table in publications, reporting sequencing metrics, QC metrics, and the number of recovered insertions (example shown below in **Table 2**).

Table 2: Summary of bulk Calling Cards experiments

Sample	**Constructs**	**Biological replicates**	**Insertions**	**Reads**	**Mean coverage**
Mouse cortex	AAV9-hyPB AAV9-tdT-SRT_bc	3	2,185,394	154,165,254	70.5
Mouse midbrain	AAV9-hyPB AAV9-tdT-SRT_bc	3	1,231,236	130,084,501	105.7
Mouse hindbrain	AAV9-hyPB AAV9-tdT-SRT_bc	3	528,121	91,084,025	172.5

Time Considerations

Basic Protocol 1: Preparation and delivery of Calling Cards reagents

Calling Cards plasmid preparation (timing: variable)

Delivery of Calling Cards reagents (timing: variable)

Production of adeno-associated viruses (timing: variable)

Intracerebroventricular injection (timing: 1 hr)

Support Protocol 1: NGS quantification of barcode distribution within SRT plasmid pool and AAV genome

Isolation of viral genome from AAV particles (timing: 3.5-4 hr)

Library preparation (timing: 1h)

Bead cleanup and quantification (timing: 30 min)

Data analysis (timing: variable)

Basic Protocol 2: Sample preparation and RNA purification

Harvesting in vitro cultures (timing: variable)

Harvesting brain tissue (timing: 10 min/mouse)

First strand cDNA synthesis (timing: 2 hr)

Support Protocol 2: Library density qPCR (timing: 1.5 hr)

Basic Protocol 3: Sequencing library preparation

RNA purification (timing: 60 min)

Amplification of self-reporting transcripts (timing: 3 hr, 20 mins hands-on)

Bead cleanup of PCR products and QC (timing: 30-45 min)

Tagmentation and indexing PCR (timing: 1-1.5 hr)

Bead cleanup of PCR products and QC (timing: 30-45 min)

Basic Protocol 4: Library pooling and sequencing

Library quantification by qPCR (timing: 1.5 hr)

Library pooling (timing: 15 min)

Sequencing (timing: variable)

Basic Protocol 5: Data analysis

Install and configure nextflow (timing: variable)

Run nf-core/callingcards on your own data to generate qBED files (timing: variable)

Analysis of qBED files (timing: variable)

Visualization of Calling Cards data (timing: variable)

2.13 Data and code availability

The data that support the protocol are available in Gene Expression Omnibus (GEO) at https://www.ncbi.nlm.nih.gov/, reference number GSE223926. The Nextflow bioinformatic pipeline and latest documentation can be found at https://nf-co.re/callingcards.

2.14 Internet resources

https://github.com/nf-core/callingcards/blob/master/conf/default_mammals.config: This is a link to an example Nextflow config file with default parameters to analyze mammalian Calling Cards data.

https://nf-co.re/modules: This is a link to a searchable list of all nf-core modules that are available to be used.

https://github.com/nf-core/callingcards/issues: Found a bug? Have a feature request? We welcome any submissions big or small through github.

https://nfcore.slack.com/channels/callingcards: This is the official slack channel that is monitored by the developers and authors. Feel free to drop in to ask questions or just say 'hi'!

https://www.addgene.org/kits/mitra-barcoded-transposon/: This is a link to an Addgene plasmid kit that contains individual barcoded self-reporting transposons. These can be grown up and pooled into one large pool or multiple subpools.

https://www.addgene.org/protocols/create-glycerol-stock/: This is a link to a guide from Addgene on how to create bacterial glycerol stocks.

https://www.nextflow.io/docs/latest/index.html: This is a link to the Nextflow documentation.

https://nf-co.re/docs: This is a link to Nextflow documentation on how to use nf-core pipelines, which is a community curated set of developed pipelines that adhere to a common set of guidelines.

https://github.com/The-Mitra-Lab/pycallingcards: This is a link to the pycallingcards package for Calling Cards analysis and visualization.

http://epigenomegateway.wustl.edu: This is a link to the Washington University Epigenome Browser.

https://epigenomegateway.readthedocs.io/en/latest/: This is a link to the WashU Epigenome Browser documentation.

https://support-docs.illumina.com/SHARE/AdapterSeq/Content/SHARE/AdapterSeq/AdapterSequencesIntro.htm: This is a link to Illumina documentation that lists adapter sequences and validated index sequences.

2.15 Acknowledgements

We would like to thank the Genome Technology Access Center at the McDonnell Genome Institute (GTAC@MGI) for their sequencing expertise and services. This work was supported by the Hope Center Viral Vectors Core at Washington University School of Medicine. Imaging was performed in part through the use of the Washington University Center for Cellular Imaging (WUCCI). This work was supported by National Institute of Mental Health (RF1MH117070, RF1MH126723 to R.D.M. and J.D.D.), National Institute of General Medical Sciences (35GM141012 to M.R.B.), National Human Genome Research Institute (T32HG000045 to A.Y.), Autism Science Foundation (22-007 to S.S.), and National Institute of Neurological Disorders and Stroke (T32NS121881 to M.A.G.).

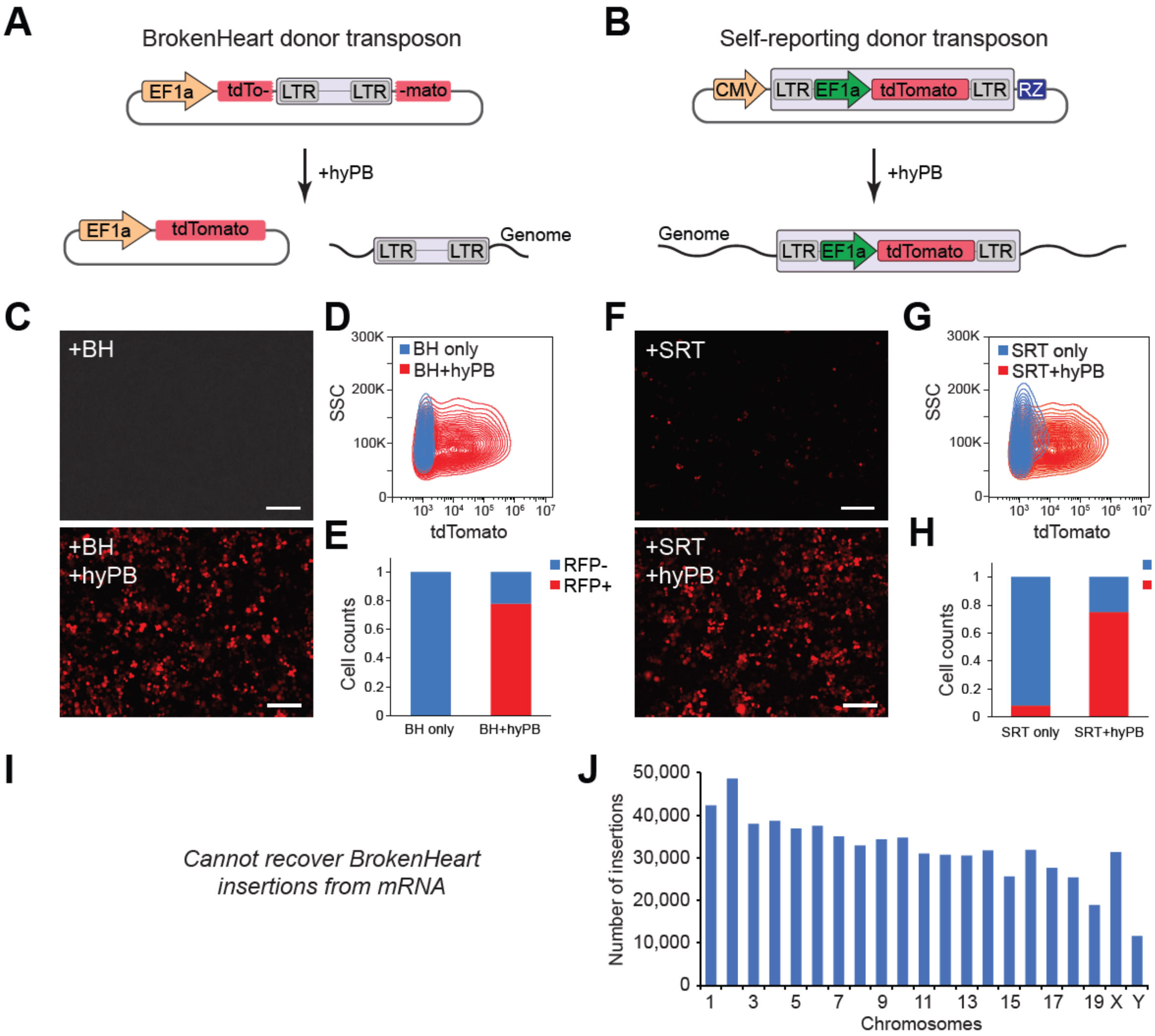
2.16 Author contributions

Conceptualization: A.Y., C.M., A.M., A.J.C., R.D.M., and J.D.D.; Methodology: A.Y., C.M., J.G., X.C., A.M., and A.J.C.; Software: A.Y., C.M., J.G., A.M., and R.D.M.; Validation: A.Y., A.M., and A.J.C.; Formal analysis: A.Y., C.M., S.S., and M.A.G.; Investigation: A.Y., S.S., M.A.G., J.H-L., and M.C.; Resources: M.R.B., R.D.M., and J.D.D.; Data curation: A.Y., C.M., and J.G.; Writing-original draft: A.Y. and S.S.; Writing-review and editing: A.Y., C.M., S.S., M.A.G., J.G., X.C., J.H-L., M.C., M.R.B., R.D.M., and J.D.D.; Visualization: A.Y., C.M., S.S., and M.A.G.; Supervision: A.Y., R.D.M., and J.D.D.; Project administration: A.Y., R.D.M., and J.D.D.; Funding acquisition: A.Y., S.S., M.A.G., M.R.B., R.D.M., and J.D.D.

2.17 Disclosures

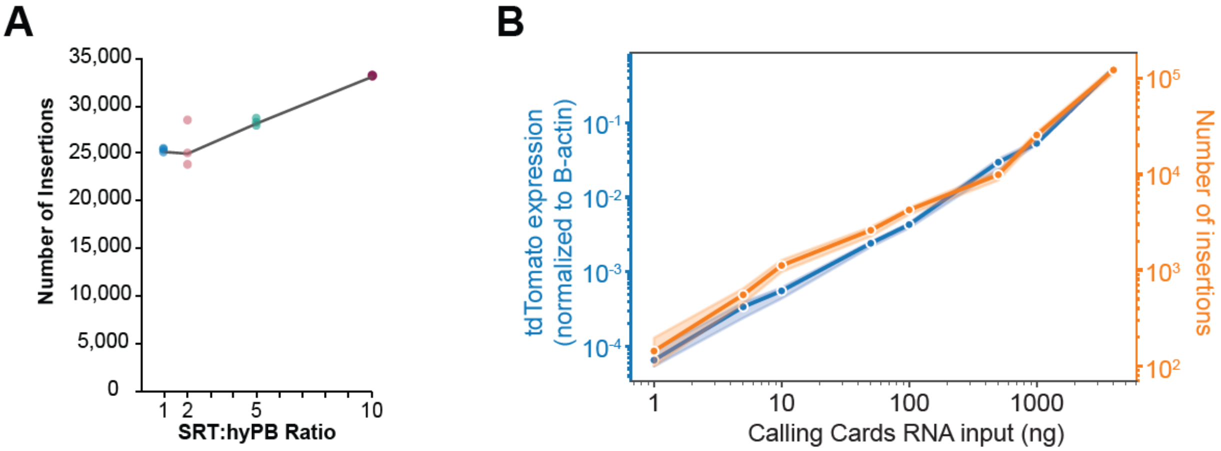
A.M. and R.D.M have filed a patent application for self-reporting transposon (SRT) technology.

2.18 Supplemental figures and tables



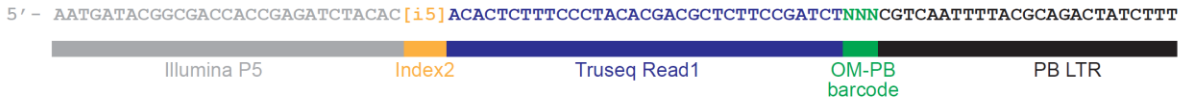
Supplemental Figure 1: Comparison of BrokenHeart (BH) and self-reporting transposons (SRT)

(A,B) Schematics showing differences in BrokenHeart and SRT transposition mechanisms. TdTomato fluorescence is detectable only with transposase activity in BrokenHeart conditions, while some background is observed with SRTs. **(C)** Representative images of HEK293 cells transfected with BrokenHeart only or BrokenHeart+hyPB. Scale bar: 50um. **(D)** Contour plot showing virtually no tdTomato fluorescence in donor only control. **(E)** Quantification of cell proportions of RFP negative and RFP positive cells. **(F-H)** Similar analysis as **C-E**, but with SRT donor instead of BrokenHeart. Scale bar: 50um. **(I)** Sequencing libraries cannot be made from BrokenHeart Calling Card libraries, whereas libraries from SRT Calling Cards **(J)** can be prepared from mRNA.

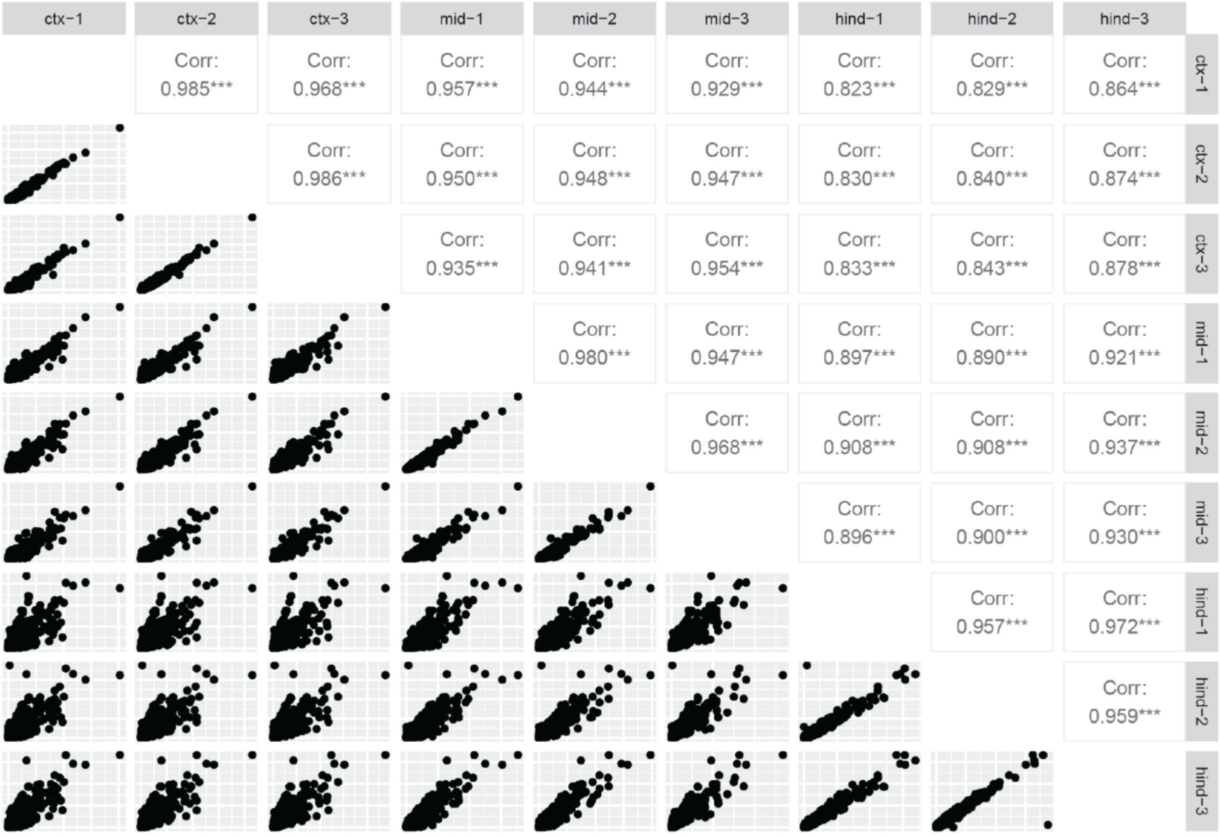


Supplemental Figure 2: Optimization of experimental conditions to maximize Calling Cards insertions

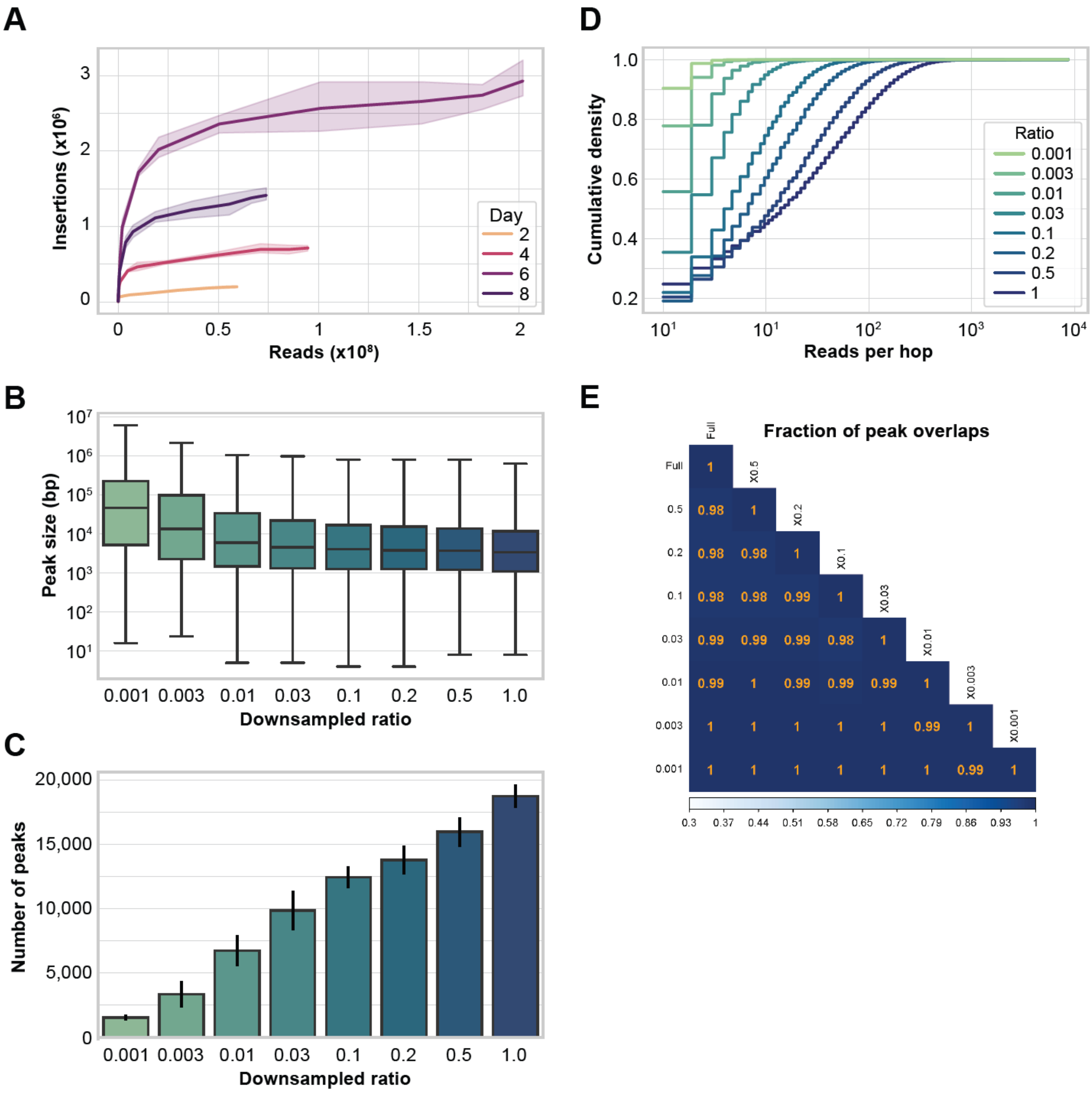
(A) Normalized insertions per 50k reads at four plasmid ratios of self-reporting transposon (SRT) to hyper *piggyBac* transposase (hyPB) (1:1, 2:1, 5:1, and 10:1) show that increasing transposon availability increases recovery of insertions in HEK293 cells. **(B)** TdTomato expression determined by quantitative RT-PCR (blue) and recovered insertions (orange) as a function of total RNA (4ug) spiked with a range of RNA containing Calling Cards insertions.


5′- AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNCGTCAATTTTACGCAGACTATCTTT
Illumina P5
Index2
Truseq Read1
OM-PB barcode
PB LTR

Supplemental Figure 3: Sequence and structure of OM-PB primer

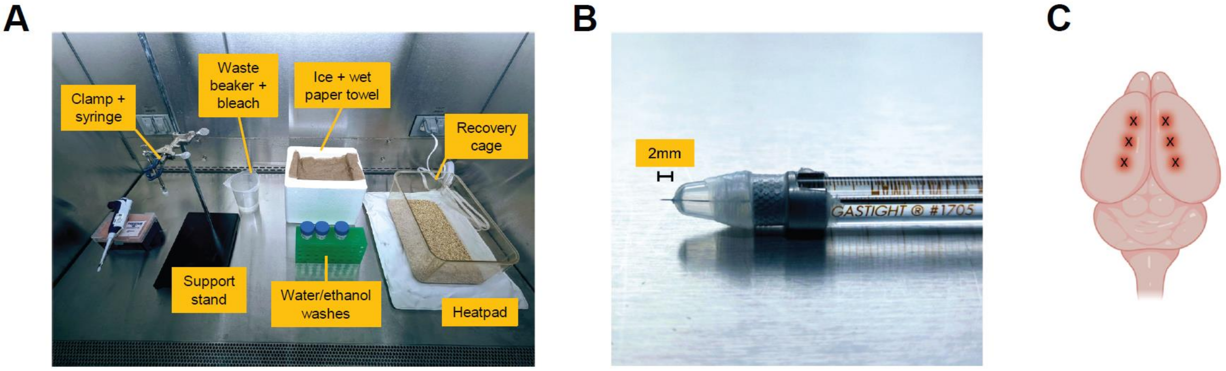


Supplemental Figure 4: Scatterplot correlation matrix of biological replicates of cortex, midbrain, and hindbrain samples



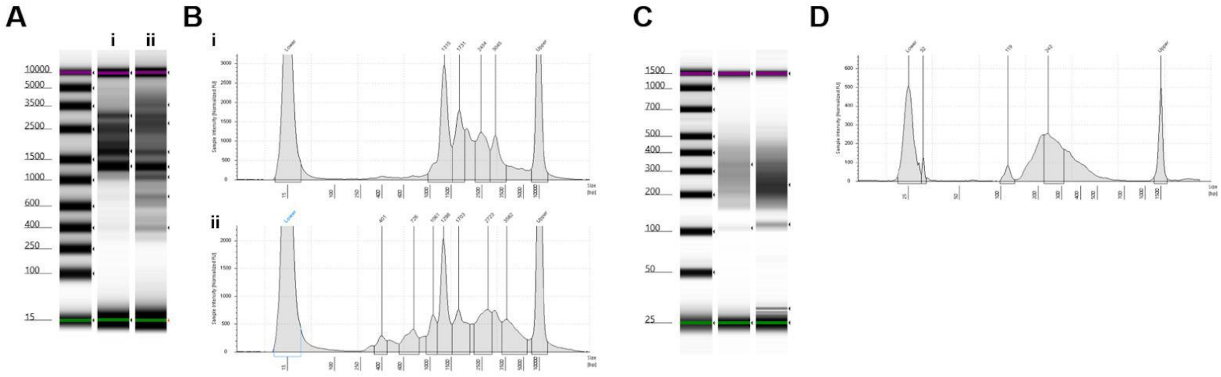
Supplemental Figure 5: Sequencing saturation and called peaks

(A) Plot of insertion densities at various downsampled read depths for samples collected 2-8 days after injection with Calling Cards reagents (see **Figure 7** for experimental design). Panels B-E are based on the deeply sequenced Day 8 sample in A. The BAM file was downsampled at set ratios to simulate a range of sequencing depths. **(B)** Box plots showing the distribution of the sizes of called peaks. **(C)** Bar plots demonstrating that the number of called peaks increases with deeper sequencing. **(D)** Cumulative density plots of the number of reads per Calling Card insertion. **(E)** Heatmap showing that the called peak regions are virtually identical with different sequencing depths. Taken all together, shallower sequencing will lead to few broad peaks while deeper sequencing will increase the resolution and result in more peaks that are narrower.



Supplemental Figure 6: Setup for intracerebroventricular (ICV) injections

(A) An example layout of materials and equipment that are needed for neonatal ICV injections of AAVs within a biosafety cabinet. **(B)** Close-up photograph of the hamilton syringe and custom needle guard to ensure a consistent injection depth of ~2mm. **(C)** Cartoon schematic depicting the approximate anatomical locations of the injections for a single animal.



Supplemental Figure 7: Tapestation traces of samples that do not pass QC

Panel **(A)** shows the gel image of samples with abnormal SRTs. Panel **(B)** shows the electropherogram of the two samples. There is a lack of the distribution as seen in Figure 9A,B. Panel **(C)** shows the gel image of tagmented libraries where lane 2 does not pass QC due to the strong presence (>5% area of total region) of a ~120bp peak. Panel **(D)** shows the electropherogram of lane 2.

Table 3: Estimated costs for a bulk Calling Cards experiment

Item / Service / Step	Approximate cost per replicate	6 samples (3 replicates, 2 conditions)
AAV packaging (hyPB)	$15	$90
AAV packaging (SRT)	$15	$90
RNA isolation and purification	$8	$48
First strand synthesis	$8	$48
SRT amplification	$3	$18
Tagmentation and indexing	$37	$222
Bead cleanups	$1	$6
Quality control and analysis	$13	$78
Primers	$2	$12
Sequencing costs	$38	$228
Animal costs	Variable	Variable
Total	*$140+*	*$840+*

AAV: adeno-assoiated virus; hyPB: hyperactive piggyBac; SRT: self-reporting transposon

Table 4: Reagents and materials for Basic Protocol 1

Reagents or Equipment	Vendor and Catalog Number
Plasmids, supplied as bacterial stabs	Various; see **Table 5** for complete list
NEB Stable Competent E. coli (High Efficiency)	New England Biolabs C3040H, or equivalent
Lysogeny Broth, Miller	Beckman, Dickinson, and Company 244610
Carbenicillin disodium salt	Sigma-Aldrich C1389
Kanamycin sulfate	Sigma-Aldrich K1377
ZymoPURE II Plasmid Maxiprep Kit or EndoFree Plasmid Maxi Kit	Zymo D4203 or Qiagen 12362
DMEM	ThermoFisher 11965092
Fetal Bovine Serum	Atlanta Biologicals S11150, or equivalent
Trypsin-EDTA (0.25%)	ThermoFisher 25200056
Polyethylenimine hydrochloride MAX (MW 40,000)	Polysciences 24765
OptiMEM	ThermoFisher 31985070
Purified AAV particles with high titer	Various
0.5% sodium hypochlorite	Various
NanoDrop Spectrophotometer	ThermoFisher ND-2000
Qubit 3.0 Fluorometer	ThermoFisher Q33216
1.5ml microcentrifuge tubes	Midwest Scientific MID15C, or equivalent
Vortex-Genie 2 Mixer, Variable speed	Scientific Industries SI0236, or equivalent
Thermomixer R	Eppendorf 05-400-205, or equivalent
Baffled Erlenmeyer flasks	Sigma-Aldrich CLS44441L-6EA, or equivalent
MaxQ 8000 Incubated Shaker	ThermoFisher SHKE8000, or equivalent
Nalgene PPCO Centrifuge Bottles	ThermoFisher 3141-0250, or equivalent
Sorvall LYNX 6000 Centrifuge	ThermoFisher 75006590, or equivalent
Fiberlite F14-6 x 250y Fixed-Angle Rotor	ThermoFisher 096-062075, or equivalent
Precision Econotherm Incubator	ThermoFisher 51221126, or equivalent
Mini Microcentrifuge	Midsci MF12, or equivalent
Endosafe nexgen-PTS	Charles River PTS150K
Endosafe Compendial LAL Cartridges	Charles River PTS2001
Gas Tight syringe with small removable needle, 50ul	Hamilton 80930
Custom syringe needles (33G, 0.5in, point style 4, 12 deg bevel angle)	Hamilton 7803-15
Custom syringe needle guard	See **Supplemental Figure 6B**

Table 4: Reagents and materials for Basic Protocol 1, *continued*

Laboratory support stand	Grainger 23YW89, or equivalent
Multipurpose clamp	Grainger 404R44, or equivalent
Heating pad	Sunbeam 731-500, or equivalent
Ice bucket with lid	Thomas Scientific 20A00F928, or equivalent
Beaker	Cole-Parmer EW-34502-42, or equivalent

Table 5: Calling Cards plasmids and AAV constructs

#	Construct[a]	Component	Description	Addgene
1	pAAV-rBrokenHeart (constitutive)	Donor transposon	A hyPB donor transposon interrupting a tdTomato reporter. Transposase activity will rescue tdTomato coding sequence and induce tdTomato expression.	203394
2	pAAV-BrokenHeart (cre-on)	Donor transposon	A hyPB donor transposon interrupting a tdTomato reporter. In the presence of Cre, transposase activity will rescue tdTomato coding sequence and induce tdTomato expression.	86950
3	pAAV-PB-SRT-tdTomato	Donor transposon	Non-barcoded *piggyBac* self-reporting transposon with tdTomato marker.	154889
4	pAAV-H2B-tdTomato-SRT	Donor transposon	Nuclear localized *piggyBac* self-reporting transposon with tdTomato marker. Contains a N-terminal H2B and C-terminal SV40 NLS.	203393
5	pAAV-PB-SRT-tdTomato_BC[x]	Donor transposon	Individual barcoded *piggyBac* self-reporting transposon with tdTomato markers. [x] is the barcode number.	193166 - 193187
6	pAAV-PB-SRT-Puro_BC[x]	Donor transposon	Individual barcoded *piggyBac* self-reporting transposons with a puromycin resistance cassette, [x] is the barcode number. These can be combined into one large pool or multiple subpools for transfection or packaging.	193143 - 193165
7	Barcoded SRT Calling Cards Collection	Donor transposon	This collection contains all tdTomato and Puromycin SRTs (from #5 and #6 above) in bacterial glycerol stocks in a convenient 96-well plate format.	1000000213

Table 5: Calling Cards plasmids and AAV constructs, *continued*

8	myc-hyPB (pRM1225)	Transposase	Wild-type hyPB transposase that drives the insertion of SRTs at BRD4 bound super enhancers. Contains a N-terminal Myc epitope tag.	Contact authors[b]
9	Sp1_621C-hyPB FLEx (pRM1718)	Transposase	Truncated SP1 containing C-terminal 621 amino acids which includes the DNA binding domain fused with hyPB. In the presence of Cre, this fusion protein redirects insertion of SRTs to SP1 TF binding sites, which are promoters found in unmethylated open chromatin.	Contact authors[b]
10	Myc-hyPB-FrontFlip (Cre-on) (pRM1888)	Transposase	Split hyPB transposase with chimeric intron containing LoxP sites. In the presence of Cre, the intron is spliced and functional N-terminal myc tagged hyPB is produced. This reduces Cre-independent background transposition relative to classic FLEx cassettes.	Contact authors[b]
11	pAdDeltaF6	AAV helper plasmid	Plasmid that expresses E4, E2A, and VA (for standard AAV packaging, if needed).	112867
12	pAAV2/9n	AAV RepCap plasmid	Plasmid that expresses Rep2 and Cap9 (for standard AAV packaging, if needed).	112865

AAV, adeno-assoiated virus; hyPB, hyperactive piggyBac; NLS, nuclear localization signal; SRT, self-reporting transposon; TF, transcription factor.

[a] [x] indicates the barcode number.

[b] piggyBac donor transposon plasmids and AAV packaging plasmids are distributed through Addgene's website (https://www.addgene.org/). Contact Addgene or authors for information about access to plasmids.

Table 6: Reagents and materials for Support Protocol 1

Reagents or Equipment	**Vendor and Catalog Number**
DNase I, RNase-free	New England Biolabs M0303
0.5 M EDTA, pH 8.0, RNase-free	Invitrogen AM9260G
Proteinase K, Molecular Biology Grade	New England Biolabs P8107S
QIAquick PCR Purification Kit	Qiagen 28104
Q5 Hot Start High-Fidelity Master Mix	New England Biolabs M0494
AMPure XP Reagent or Mag-Bind TotalPure NGS beads	Beckman Coulter A63882 or Omega Biotek M1378-02
T100 Thermal Cycler	Biorad 1861096
Magnetic Separation Rack for PCR strip tubes	Permagen MSR812, or equivalent

Table 7: Reagents and materials for Basic Protocol 2

Reagents or Equipment	Vendor and Catalog Number
DPBS, no calcium, no magnesium	ThermoFisher 14190144
Isoflurane	Various
TRIzol Reagent	ThermoFisher 15596026
Chloroform	Sigma-Aldrich C2432
RNaseZap	ThermoFisher AM9782, or equivalent
RNA Clean & Concentrator-5 or RNEasy Plus Mini Kit	Zymo R1014 or Qiagen 74134
Molecular biology grade water	Corning 46-000-CM, or equivalent
High Sensitivity RNA ScreenTape and Sample Buffer	Agilent 5067-5579 and Agilent 5067-5580
Qubit RNA Assay Kits	ThermoFisher Q32852, Q10211, or Q33224
SMART_dT18VN primer	Custom synthesized by Integrated DNA Technologies (IDT)
Maxima H Minus Reverse Transcriptase	ThermoFisher EP0752
dNTP mix (10mM ea, PCR grade)	ThermoFisher 18427088
RNaseOUT Recombinant Ribonuclease Inhibitor	ThermoFisher 10777019
RNase H	New England Biolabs M0297L or ThermoFisher 18021071
QIAquick PCR Purification Kit	Qiagen 28106
Buffer EB	Qiagen 19086
Qubit ssDNA Assay Kit	ThermoFisher Q10212
Handheld homogenizer	SP Bel-Art F65100-0000, or equivalent
RNase-free pestles for 1.5 ml tubes	Fisher Scientific 12-141-364
1.5ml microcentrifuge tubes	Midwest Scientific MID15C, or equivalent
Refrigerated centrifuge	Eppendorf 5430R, or equivalent
Rotor for 1.5ml and 2ml tubes	Eppendorf FA-45-30-11, or equivalent
Nanodrop Spectrophotometer	ThermoFisher ND-2000
Tapestation 4200 System	Agilent G2991BA
Tempassure PCR 8-tube strips	USA Scientific 1402-4700, or equivalent
T100 Thermal Cycler	Biorad 1861096, or equivalent
Qubit 3.0 Fluorometer	ThermoFisher Q33216
Qubit Assay Tubes	Q32856

Table 8: Reagents and materials for Support Protocol 2

Reagents or Equipment	Vendor and Catalog Number
TdTomato qPCR primers	Custom synthesized by Integrated DNA Technologies (IDT)
Molecular biology grade water	Corning 46-000-CM, or equivalent
PowerUp SYBR Green Master Mix	ThermoFisher A25743, or equivalent
MicroAmp Optical 384-Well Reaction Plate	ThermoFisher 4309849, or equivalent
MicroAmp Optical Adhesive Film	ThermoFisher 4311971, or equivalent
QuantStudio 6 Flex Real-Time PCR System	ThermoFisher 4485691, or equivalent system

Table 9: Primer sequences for Chapter 2

Primer Name[a]	**Sequence (5' → 3')**[b]	**Use**	**Length (bp)**
SRT_bc_QC_F	AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCGTCAATTTTACGCAGACTATCTT	Amplify SRT barcode fragment for NGS	88 without Index2
SRT_bc_QC_R	CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGAGCGTGGATAGCAGTGGAATCC	Amplify SRT barcode fragment for NGS	88 without Index2
SMART_dT18VN	AAGCAGTGGTATCAACGCAGAGTACGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN	First-strand synthesis	59
SRT_tdTomato_F1	TCCTGTACGGCATGGACGAG	Amplify tdTomato SRTs	20
SMART_Rev	AAGCAGTGGTATCAACGCAGAGT	Amplify tdTomato SRTs	23
OMPB_BC-TAG_Index2_x	AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGCGTCAATTTTACGCAGACTATCTTT	Indexing PCR	90 without Index2
OMPB_BC-GCA_Index2_X	AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCACGTCAATTTTACGCAGACTATCTTT	Indexing PCR	90 without Index2
OMPB_BC-CTA_Index2_X	AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTACGTCAATTTTACGCAGACTATCTTT	Indexing PCR	90 without Index2
OMPB_BC-ACG_Index2_X	AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTACGCGTCAATTTTACGCAGACTATCTTT	Indexing PCR	90 without Index2

(Continued)

Table 9: Primer sequences for Chapter 2, *continued*

Primer Name[a]	Sequence (5’ → 3’)[b]	Use	Length (bp)
OMPB_BC-GAT_Index2_x	AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATCGTCAATTTTACGCAGACTATCTTT	Indexing PCR	90 without Index2
OMPB_BC-ATC_Index2_X	AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTATCCGTCAATTTTACGCAGACTATCTTT	Indexing PCR	90 without Index2
N7_Index_X	CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG	Indexing PCR	39 without Index1
β-actin_qPCR_For	AGAGGGAAATCGTGCGTGAC	qPCR	20
β-actin_qPCR_Rev	CAATAGTGATGACCTGGCCGT	qPCR	21
TdTomato_qPCR_For	CAAGCTGAAGGTGACCAAGG	qPCR	20
TdTomato_qPCR_Rev	CCGTCCTCGAAGTTCATCAC	qPCR	20
Gapdh_qPCR_For	AGGTCGGTGTGAACGGATTTG	qPCR	21
Gapdh_qPCR_Rev	GGGGTCGTTGATGGCAACA	qPCR	19

[a] The three bases highlighted in blue indicate the location of the OM-PB primer barcode. For more information, refer to “ Considerations for primer selection and ordering for sequencing libraries” in Chapter 2.3.

[b] The [i5] and [i7] represent placeholders where unique indexes can be added for multiplexing multiple samples to be sequenced in a single run. For more information about standard Illumina indexes, see https:// supportdocs.illumina.com/SHARE/AdapterSeq/Content/SHARE/AdapterSeq/AdapterSequencesIntro.htm, or consult your sequencing core or service provider.

Table 10: Reagents and materials for Basic Protocol 3

Reagents or Equipment	Vendor and Catalog Number
Standard desalted primers	Custom synthesized by Integrated DNA Technologies (IDT)
Molecular Biology Grade Water	Corning 46-000-CM, or equivalent
KAPA HiFi HotStart ReadyMix	Roche KK2601
AMPure XP Reagent or Mag-Bind TotalPure NGS beads	Beckman Coulter A63882) (Omega Biotek M1378-02
Ethyl alcohol 200 Proof (ACS/USP Grade)	Pharmco-Aaper 11100020S, or equivalent
Qubit dsDNA High Sensitivity Assay Kit	ThermoFisher Q32851
High Sensitivity D5000 ScreenTape	Agilent 5067-5592
High Sensitivity D5000 Ladder and Sample Buffer	Agilent 5067-5593
Nextera XT DNA Library Preparation Kit	Illumina FC-131-1096
96 well metal cooling block	Argos Technologies 63615-04, or equivalent
T100 Thermal Cycler	Biorad 1861096, or equivalent
Magnetic Separation Rack for PCR strip tubes	Permagen MSR812, or equivalent
Tapestation 4200 System	Agilent G2991BA
Qubit 3.0 Fluorometer	ThermoFisher Q33216
Qubit Assay Tubes	Q32856
Nanodrop Spectrophotometer	ThermoFisher ND-2000

Table 11: Reagents and materials for Basic Protocol 4

Reagents or Equipment	Vendor and Catalog Number
Kapa Library Quantification Kit or NEBNext Library Quant Kit for Illumina	Roche KK4824 or New England Biolabs E7630
MicroAmp Optical 384-Well Reaction Plate	ThermoFisher 4309849, or equivalent
MicroAmp Optical Adhesive Film	ThermoFisher 4311971, or equivalent
QuantStudio 6 Flex Real-Time PCR System	ThermoFisher 4485691, or equivalent
96 well metal cooling block	Argos Technologies 63615-04

Table 12: Materials for Basic Protocol 5

Equipment or software	Vendor and Catalog Number
POSIX compatible system (e.g., Linux, macOS, etc.)	Various
8-core Intel or AMD processor (16 cores recommended)	Various
64GB RAM	Various
500GB free disk space	Various
Bash 3.2 or later	Various
Shared filesystem*	Various
Batch scheduling system* (e.g., SLURM, SGE, LSF, etc.)	Various

*This is only required if running in cluster mode.

Chapter 3: Design, development, and validation of transgenic Calling Cards mice

3.1 Introduction

In the previous chapter and publications, I showed that exogenous expression of Calling Cards reagents using plasmid or AAV vectors induced high expression of Calling Cards in cultured cells and in the mouse brain (Cammack et al., 2020; Moudgil et al., 2020b; Yen et al., 2023). This technique offers significant potential as a tool for recording transient molecular states in live tissues. For instance, we know that progenitor cells respond to temporal and spatial morphogenic signals resulting in the expression of precise combinations of transcription factors (TFs) that determine cell fate. However, many of these TFs are expressed for only a brief moment during the cells' differentiation and maturation trajectories and therefore are not detectable after these transient events occur. Having the ability to record these historical molecular interactions and associate them with subsequent phenotypes offers valuable insights into developmental biology and other fields necessitating such analyses.

Current Calling Cards reagents utilize AAV vectors to express *piggyBac* transposase and self-reporting transposons (SRTs) *in vivo,* which limit the application to cell populations that are physically accessible and amenable to injections. Additionally, stable transgene expression requires a cascade of cellular transduction mechanisms which typically occurs over a period of several days (Aschauer et al., 2013; Berry and Asokan, 2016). While *in utero* AAV injections into embryos are feasible, a surgical procedure is necessary to access the fetal environment and the yield and location of transduced cells is dependent on viral tropism and age of injection. In contrast, an even spread of transduced cells is observed following intracerebroventricular injections at post-natal day 1 (P1). However, a limitation of P1 injections of Calling Cards is that many key cell fate determinations and developmental processes have already occurred. To address these limitations, I sought to develop and evaluate triple transgenic mice in which a Cre-dependent *piggyBac* transposase, SRTs, and Cre recombinase are present in the genome to enable recording of molecular states in genetically defined cell populations, negating the need for viral injections.

In this study, I generated a hemizygous PGK-GFP-SRT(Tg/+) donor transposon mouse line and validated the functionality of a the conditional ROSA26$^{\text{LSL-PB}}$(fl/fl) *piggyBac* transposase line (Rad et al., 2015), and crossed to a variety of Cre mouse lines, generating triple transgenics that should display Calling Card activity. As controls, I generated mice that had one or two of these components, and delivered the remaining component by virus. While the single and double transgenic animals performed largely as anticipated, the triple transgenic animals either failed to yield viable litters or did not generate sufficient numbers of insertions for meaningful experimental analysis. Despite the challenges in creating transgenic Calling Cards mouse lines, the insights gained have informed our strategies for future attempts using alternate approaches.

3.2 Results

3.2.1 Generation and validation of donor transposon PGK-GFP-SRT mice

One of the core components of Calling Cards is the self-reporting transposon (SRT). While the AAV reagents described in the previous chapter use a tandem dimer Tomato (tdTomato) reporter, I developed SRTs containing enhanced GFP (eGFP) because it is thought to be less cytotoxic when expressed at high levels and it enables direct comparisons between transgenic eGFP and AAV tdTomato SRTs in the same tissues with both immunofluorescence and sequencing libraries. The transgenic construct consisted of a moderate strength PGK promoter for constitutive expression driving expression of a GFP SRT within *piggyBac* long terminal repeats (LTRs), as prior efforts with stronger promotors had also failed (data not shown). A hammerhead ribozyme (Yen et al., 2004), positioned immediately downstream of the 3' *piggyBac* LTR, minimized background artifacts by cleaving non-transposed SRTs (Moudgil et al., 2020b). This PGK-GFP-SRT construct was microinjected into the male pronuclei of fertilized oocytes (**Figure 17A**). Approximately 78 copies of the transgene were inserted in the genome at a singly inherited locus of unknown position via transgenesis (**Figure 17B, C**). This high copy count is beneficial as the SRT copy number can directly influence the functional insertion rate per cell (**Supplemental Figure 2**).

Next, I evaluated the PGK-GFP-SRT(Tg/+) mice to verify and validate that these transgenic SRTs are functional and record transient molecular interactions in similar patterns when benchmarked against the established AAV reagents. This involved transcranial injections of AAV9-hyPB into P0-1 neonatal pups, followed by a 21-day recording period. Post-harvest, each brain was bisected: one hemisphere for immunofluorescence to quantify the number of cells that contain Calling cards insertions and to visualize the localization of transduced cells, and the other hemisphere for sequencing to map the insertion sites and quantify their abundance (**Figure 17A**). The immunofluorescence results confirmed widespread GFP expression across brain regions (**Figure 17B)** and transduction in both neurons and astrocytes (**Figure 17C, D**). Sequencing analysis demonstrate recovery of Calling Cards insertions across all chromosomes, validating the functionality of the SRT array (**Figure 17E**). Notably, the insertion density appeared consistent across chromosomes when adjusting for chromosome length, with the exception of chromosome 11, suggesting potential integration of the transgene array therein. Furthermore, the insertion pattern demonstrated both high reproducibility across animals (**Figure 17F**) and strong correlation with AAV-Calling Cards, substantiating the utility and effectiveness of the transgenic SRT system.

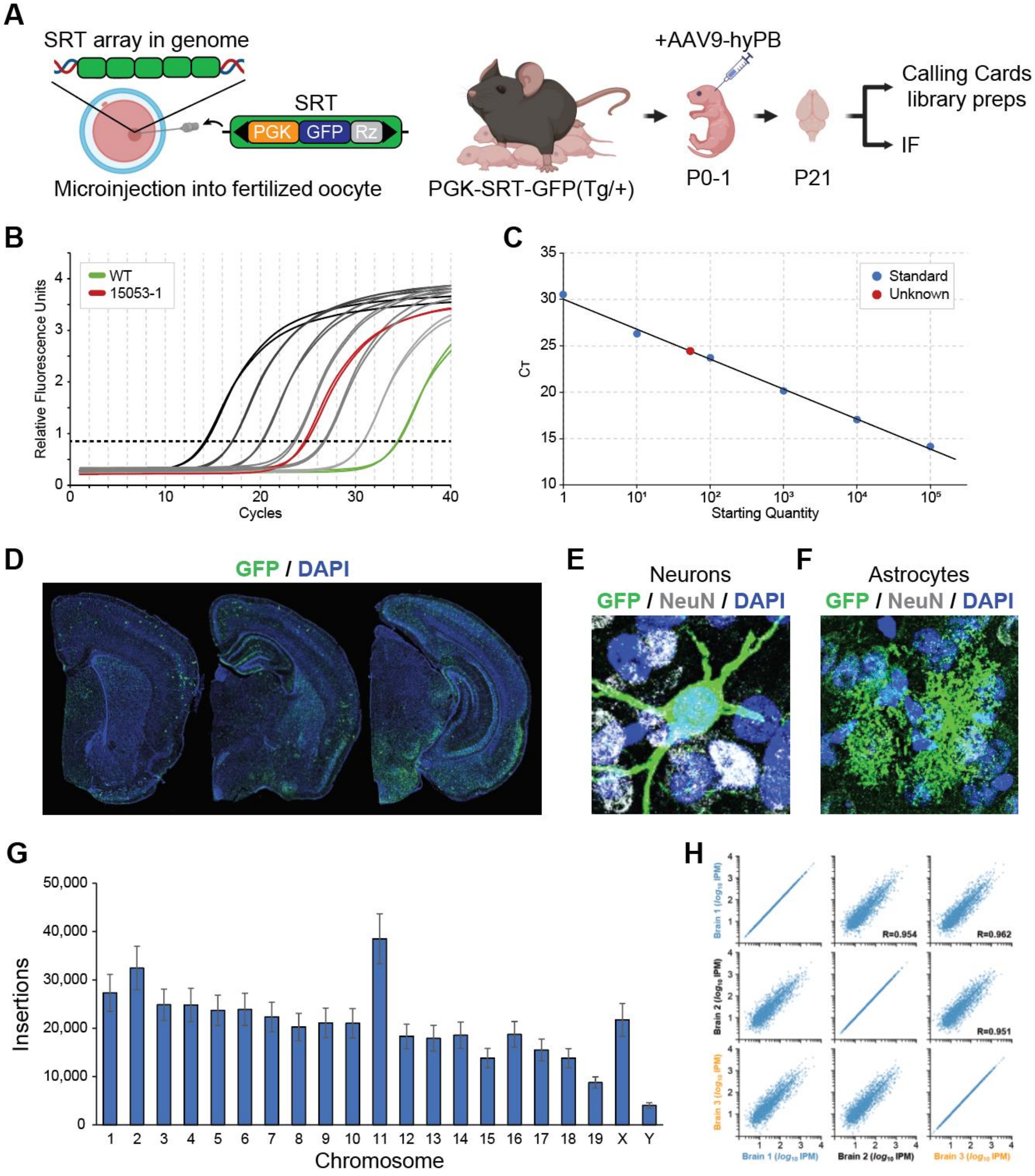


Figure 17: Generation and validation of PGK-GFP-SRT(Tg/+) mice

(A) Schematic showing the SRT array integrated into the genome after pronuclear microinjection of the linearized PGK-GFP-SRT construct into a fertilized C57Bl/6 egg. The PGK-GFP-SRT(Tg/+) litters were injected with AAV9-hyPB at P0-1 and harvested at P21 for sequencing library preparations and immunofluorescence analysis. **(B)** Amplification curves from a quantitative real-time PCR (qPCR) run for copy number analysis. Known quantities of transgenes were run in duplicates and are shown in the grey lines. Amplification curves for a WT sample with no GFP is shown in green and a transgenic founder is shown in red. **(C)** The resulting standard curve constructed from the known copy numbers can be used to empirically determine the transgene copy number of the founder. **(D)** Representative images of coronal sections along the anterior-posterior axis showing cells with Calling Cards insertions (green) and nuclei (blue). **(E, F)** A representative image of a neuron in **E** and astrocyte in **F** with Calling Cards insertions (green) amongst other neurons (grey) do not contain insertions. Nuclei are stained blue. **(G)** Summary graph (average±s.d.) showing the number of insertions recovered per chromosome across 3 biological replicates **(H)** Correlation plots showing high degree of correlation between biological replicates.

3.2.2 Validation of Cre-dependent *piggyBac* mice

The other core component of Calling Cards is the *piggyBac* transposase. To enhance the system's versatility and accessibility for the broader scientific community, I utilized the homozygous *Rosa26*$^{\textit{LSL-PB}}$(fl/fl) mice (Rad et al., 2015), which incorporates a *loxP*-flanked stop (LSL) cassette upstream of *piggyBac* in the ROSA26 locus (**Figure 18A**). This particular locus is favored for its capacity to drive widespread gene expression in mice. In this setup, the *piggyBac* transposase can be specifically activated in targeted cell types by removing the LSL cassette using Cre recombinase. To validate the mouse, I crossbred the LSL-PB(fl/fl) mice with Actin-Cre(Cre/+) mice, resulting in the removal of the LSL cassette, leading to activation of *piggyBac* transposase in all cells. Next, I administered transcranial injections of AAV9-tdTomato-SRT into neonatal P1 pups to test the transposase's efficiency within this transgenic system. Immunofluorescence analysis revealed widespread tdTomato expression throughout the cortex of LSL-PB(fl/fl)/Actin-Cre(Cre/+) double transgenic mice, consistent with Actin-Cre's universal activity (**Figure 18B**). In contrast, Cre-negative control animals also injected with AAV9-tdTomato-SRT displayed only marginal tdTomato expression, indicating a low background transposition rate. Sequencing data confirmed recovery of insertions across all chromosomes, with similar insertion densities after adjusting for chromosome length (**Figure 18C, D**). Notably, there was no enrichment of insertions on chromosome 11 with AAV-delivered SRTs as I saw with transgenic SRTs, suggesting this characteristic is unique to transgenic SRT applications. This data confirms the transposition activity of transgenic *piggyBac*.

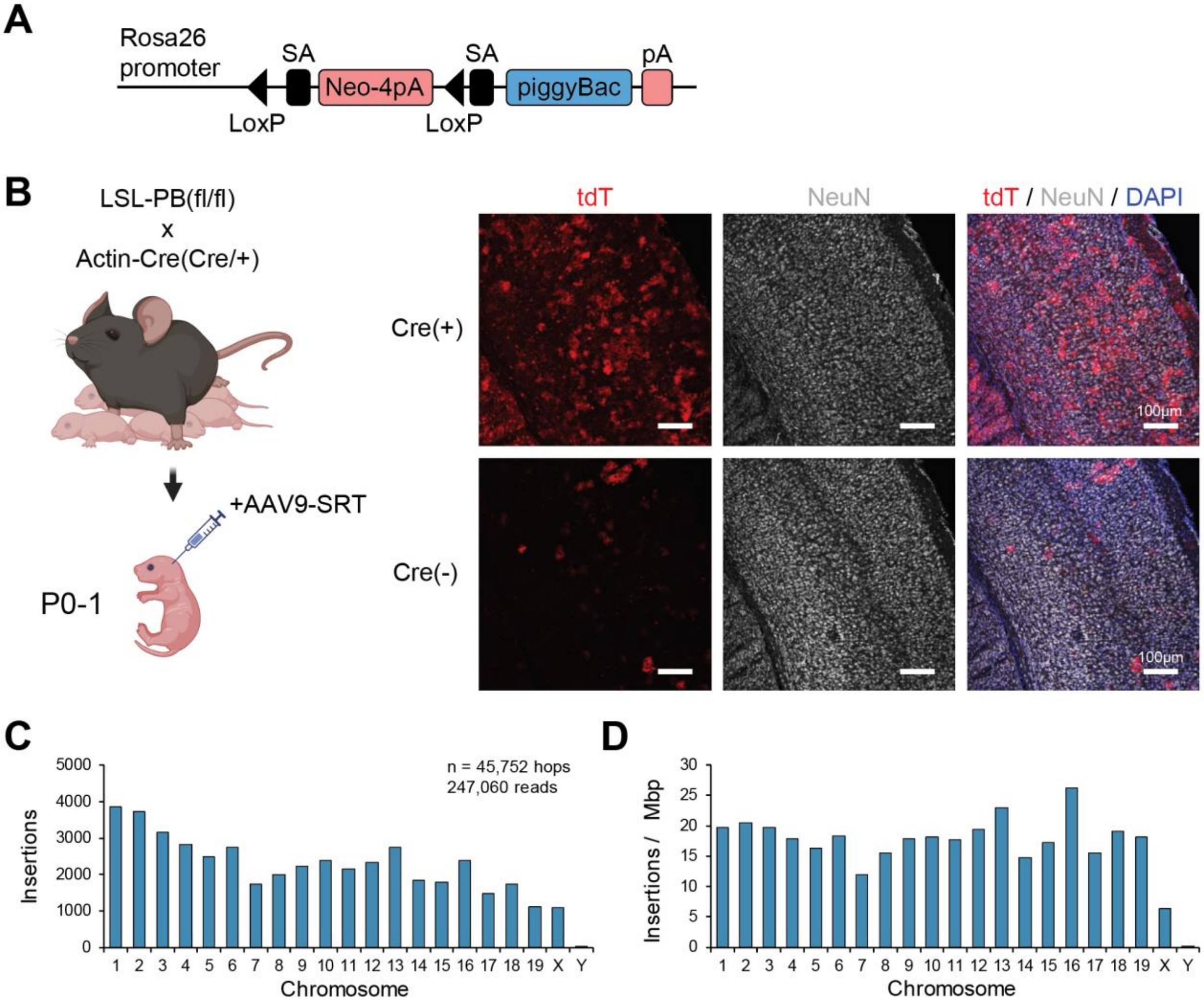


Figure 18: Validation of LSL-PB(fl/fl) transposase mouse

(A) Schematic of the conditional *piggyBac* transposase mouse line *ROSA26*$^{LSL\text{-}PB}$(fl/fl). Without Cre recombinase, there is no expression of *piggyBac* due to the Neo-4pA stop cassette. In the presence of Cre, the stop cassette is excised. **(B)** The LSL-PB(fl/fl) line was crossed with heterozygous Actin-Cre(Cre/+) mice to activate *piggyBac* transposase in all cells. AAV9-tdTomato-SRT was trancranially injected into P0-1 neonatal pups. Immunofluorescence staining of P21 brain sections show robust activation of Calling Cards (red) in many NeuN-positive neurons (grey) of the cortex in Cre-positive P21 animals. Nuclei were stained blue with DAPI. Some background is apparent in Cre-negative conditions. **(C)** Summary data showing recovered insertions were distributed across all chromosomes. **(D)** The summary data in C was normalized for chromosome length.

3.2.3 Complexities of triple transgenic Calling Cards mice

Calling Cards technology offers a unique way to capture permanent records of transient protein-DNA interactions, allowing for the quantitative study of historical cell states. Unlike conventional that only provide a "snapshot" of gene expression at a specific moment, Calling Cards can facilitate retrospective analysis, proving especially valuable in scenarios where there is a significant amount of time between the molecular event of interest and the time the data is collected. This can be especially promising in areas such as embryonic development, cell fate determination, and neurodevelopment, where it can offer valuable insights into the gene expression programs that govern these complex processes.

The results demonstrating the effectiveness of both the transgenic PGK-GFP-SRT(Tg/+) donor transposon line and the cre-dependent LSL-PB(fl/fl) transposase line were a significant milestone. I proceeded to crossbreed these two lines, generating PGK-GFP-SRT(Tg/+)/LSL-PB(fl/fl) double transgenic animals, which may potentially be a valuable generalized tool for researchers, as it requires only an additional cross with a specific Cre driver line to activate Calling Cards in a cell type of interest. To validate the Cre-dependence, I transcranially injected AAV8-Ef1a-mCherry-IRES-Cre into the ventricles of P0-1 PGK-GFP-SRT(Tg/+)/LSL-PB(fl/fl) neonatal pups. Immunofluorescence and genomic analysis confirmed that Calling Cards were activated in mCherry-positive cells, suggesting that this Cre-dependent Calling Cards mouse line was suitable for further testing with Cre driver lines.

Formation of the three primary germ layers—ectoderm, mesoderm, and endoderm—during gastrulation is a critical phase in the developing organism and is known to be marked with significant changes in gene expression (Chan et al., 2019; Pijuan-Sala et al., 2019). However, whether the epigenetic landscape governing cell fate determination and patterning between molecular layers is still unclear. By using an Actin-Cre line, Calling Cards can be turned on in all cells to record enhancer usage during this critical developmental process. The eventual lineage information can then be used to identify gene regulatory elements that contributed to the specification of the germ layers. This can be a powerful tool to assess enhancer activity during early development, which is a particularly challenging environment to access using AAVs. However, out of over 80 pups from multiple breeding pairs, only 1 animal was the desired triple transgenic PGK-GFP-SRT(Tg/+)/LSL-PB(fl/fl)/Actin-Cre(Cre/+) genotype, compared to the 20 that would be expected from Mendelian inheritance. This suggests a potentially detrimental impact of Calling Cards insertions at this crucial embryonic stage. Analysis of the one surviving triple transgenic animal showed GFP+ cells distributed across the brain, with slight enrichments in the retrosplenia granular cortex region in the visual cortex and the granular layer of the olfactory bulbs (**Figure 19**). Genomic analysis recovered an average of 1633 insertions from brain tissue, 845 insertions from liver, and 182 insertions from muscle tissue. Thus, the surviving mouse had very little transposon activity. One possible explanation for the drop in developmental viability is that a Calling Card insertion, which is approximately the size of a EF1α promoter and tdTomato reporter, landed into a key regulatory region during early embryogenesis and induced cell death. At this stage, each cell may be indispensable, and loss of a single cell can impact the viability of the embryo. This single animal could have survived if the Calling Cards insertions did not land in a critical genomic region and were epigenetically silenced by PIWI-interacting RNAs or other defense mechanisms against transposons (Czech and Hannon, 2016). These insertions will then be immobile but passed down to all its cellular progeny, explaining why I see low numbers of insertions across various tissues.

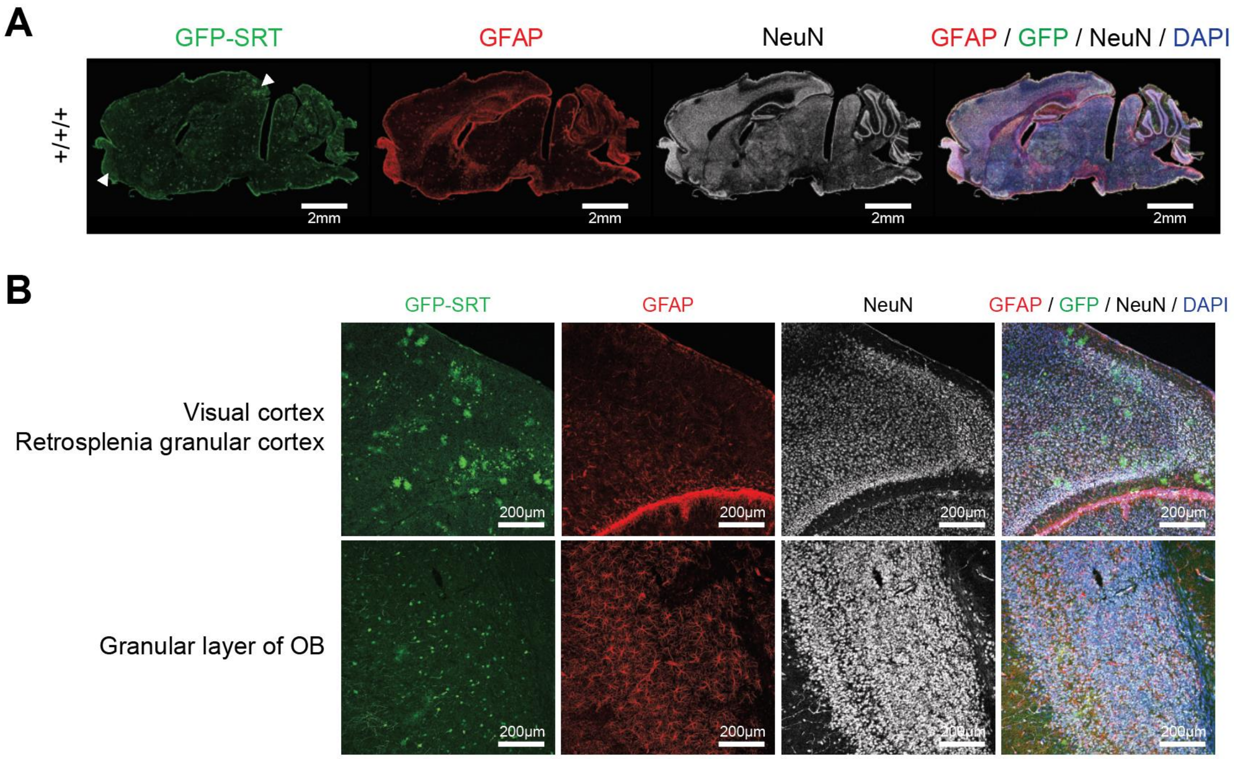


Figure 19: Analysis of the single viable Actin-CC mouse

(A) Images of immunofluorescence sections from triple transgenic (PB$^+$/GFP-SRT$^+$/Actin-Cre$^+$) show SRTs (GFP; green), astrocytes (GFAP; red), and neurons (NeuN; gray). Nuclei are stained blue with DAPI. **(B)** Zoomed in images of the visual and retrosplenia granular cortex and the granular layer of the olfactory bulb regions indicated by the arrowheads in **A**.

To avoid early embryogenesis, the subsequent experiment targeted neural progenitors at E12 via crossing with the Nestin-Cre(Cre/+) line. This line successfully produced the expected numbers of triple transgenic mice (**Figure 20A**), there was a notable 28±2% reduction in brain size and volume (**Figure 20B**), which correlated with a 31±3% decrease in brain weight (**Figure 20C**), suggesting potential adverse effects of Calling Cards on neural progenitor cells. Therefore, I next tried Baf563b-cre, which targets post-mitotic neurons, thus avoiding expression in progenitors. This yielded viable offspring with the expected genotype ratios and expression of

GFP-SRTs in neurons (**Figure 21A,B**), although the genomic analysis uncovered a disproportionate number of insertions on Chromosome 4. This enrichment was observed across multiple brain regions, suggesting that this could originate from a spontaneous insertion in a progenitor cell that was silenced, and then passed down to all progeny (**Figure 21C**).

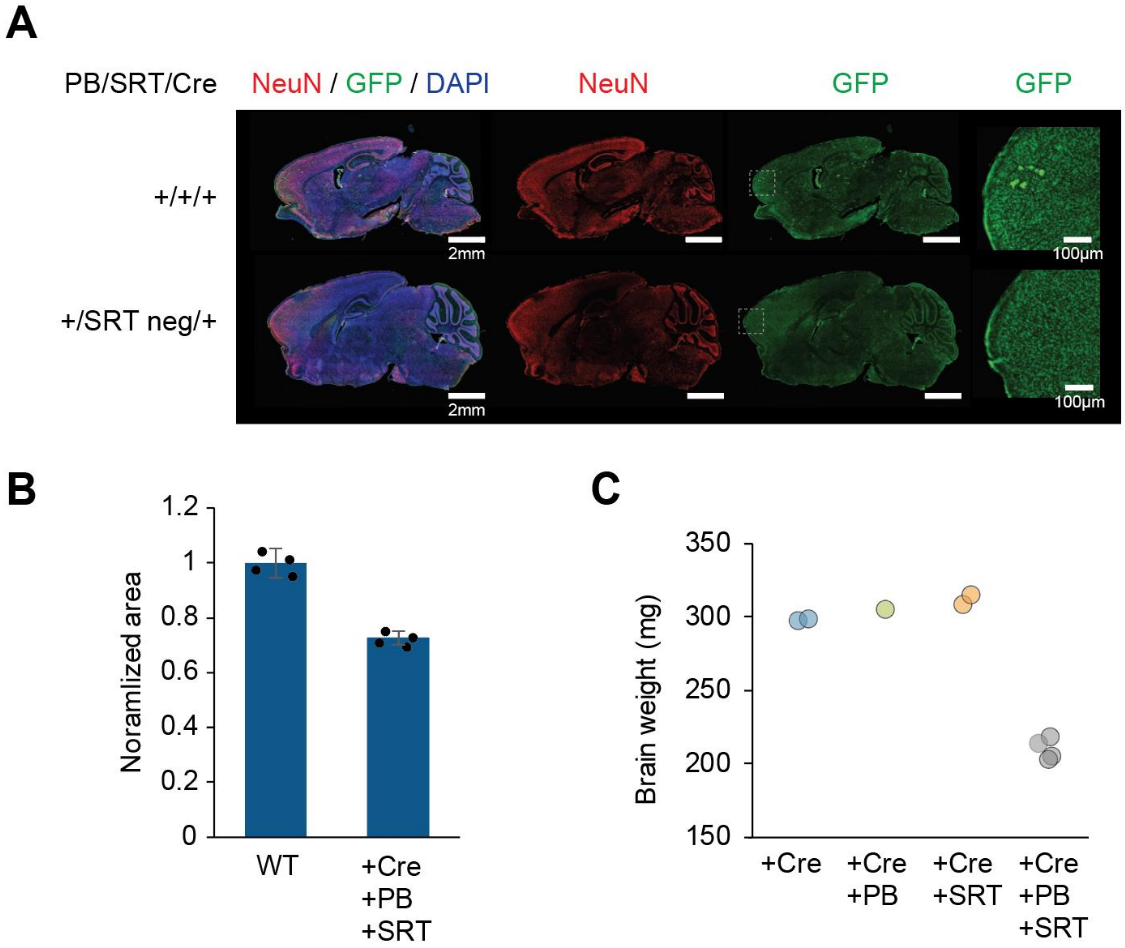


Figure 20: Nestin-CC mice have decreased brain volume and weight

(A) Images of immunofluorescence sections from triple transgenic (PB$^+$/GFP-SRT$^+$/Nestin-Cre$^+$) and double transgenic (PB$^+$/GFP-SRT$^-$/Nestin-Cre$^+$) animals showing localization in neurons (NeuN; red) and SRT insertions (GFP; green). Nuclei are stained blue with DAPI. The expression of GFP in the SRT negative sample may indicate either high background signal, but very few insertions were recovered by sequencing suggesting that they are not true insertions. **(B)** The area of the brain sections from Nestin-CC mice was quantified and normalized against WT. Comparative analysis shows a decrease in brain size. **(C)** The brain weights were measured during harvesting. Triple positive Nestin-CC brains show a decrease in weight only when all 3 transgenes were expressed.

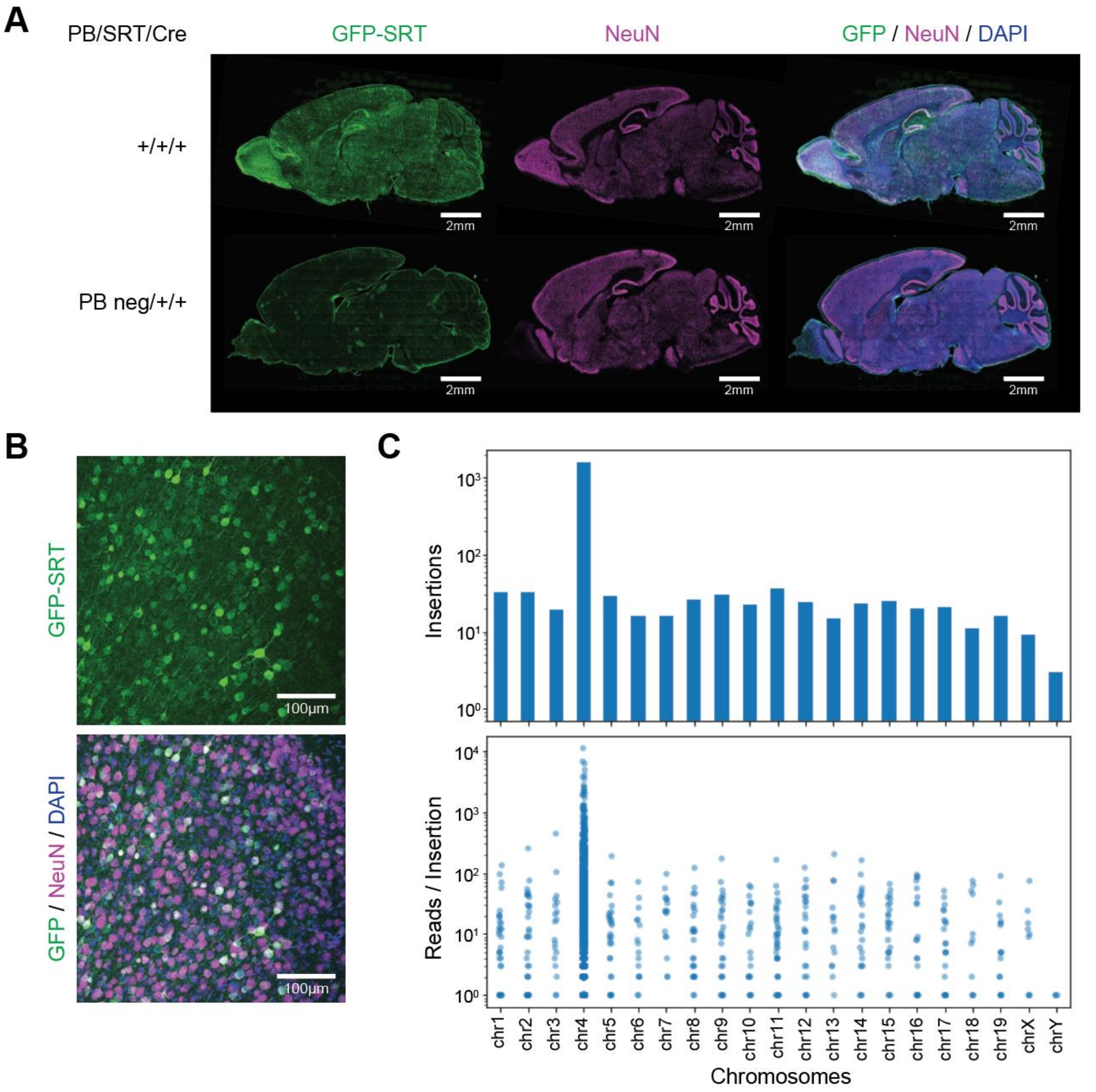


Figure 21: Baf53b-CC mice have abnormal distribution of Calling Cards insertions

(A) Images of immunofluorescence sections from triple transgenic (PB^+/$GFP\text{-}SRT^+$/$Baf53b\text{-}Cre^+$) and double transgenic (PB^-/$GFP\text{-}SRT^+$/$Nestin\text{-}Cre^+$) animals showing broad and high expression of SRT insertions (GFP; green) in neurons (NeuN; magenta) across the entire brain. Nuclei are stained blue with DAPI. **(B)** Higher magnification immunofluorescence images showing specific localization of SRT insertions (GFP; green) in neurons (NeuN; magenta). **(C)** Analysis of sequencing Calling Cards libraries shows a significant enrichment of insertions in Chromosome 4.

Further exploration using cell type-specific Cre lines, such as Vgat-Cre(Cre/+) for GABAergic neurons and Rbp4-Cre(Cre/+) for layer V (L5) pyramidal neurons, resulted in distinct outcomes. Notably, the Vgat-Cre driven Calling Cards demonstrated that the correct populations of cells were being targeted (**Figure 22**). In contrast, comparisons between transgenic and AAV-Calling Cards, show discrepancies in GFP and tdTomato expression. AAV9-tdTomato-SRT was transcranially injected into triple positive PGK-GFP-SRT(Tg/+)/LSL-PB(fl/fl)/Rbp4-Cre(Cre/+) as well as double positive LSL-PB(fl/fl)/Rbp4-Cre(Cre/+) P1 neonatal pups, where *piggyBac* expression was driven by Rbp4-Cre. Interestingly, analysis of immunofluorescence staining show that the AAV9-tdTomato-SRTs were found in the expected expression pattern, labeling cells in L5 of the cortex, however the transgenic GFP SRTs were enriched in the thalamus and hindbrain in cells that morphologically looked like astrocytes (**Figure 23**). While the GFP-only expressing cells can represent cells where insertions occurred during embryonic development, I would expect that the GFP and tdTomato-SRTs colocalize in the cortex. This suggests that the *piggyBac* transposase is working as expected, but the SRTs are not. In both cases, the number of recovered insertions were too low for robust statistical analyses. Given that all Cre lines tested so far target cell populations ranging from early to late embryonic development, the final line I tested was CaMKIIα-Cre, which is expressed postnatally around P21. Immunofluorescence results suggest that Calling Cards insertions were present, however the sequencing results also showed low recovery of insertions.

The transgenic lines containing one or two of the necessary Calling Cards components functioned as expected when the final component was expressed via AAV. However, attempts to create triple transgenic mice carrying piggyBac transposase, SRT, and Cre recombinase in the genome encountered various failures. I observed that expressing Calling Cards during early embryonic development resulted in non-viable animals and that neural progenitors might be particularly susceptible to Calling Cards insertions, leading to microcephaly. Furthermore, while using lines with Cre expression controlled by Baf53b, Vgat, and CaMKIIα promoters did activate Calling Cards in the targeted cell populations, the overall number of insertions recovered was too low for well-powered experiments (**Figure 24**). Although this attempt did not produce the anticipated transgenic Calling Cards mice, the insights gained have informed our strategies for the next series of experiments to develop these mice.

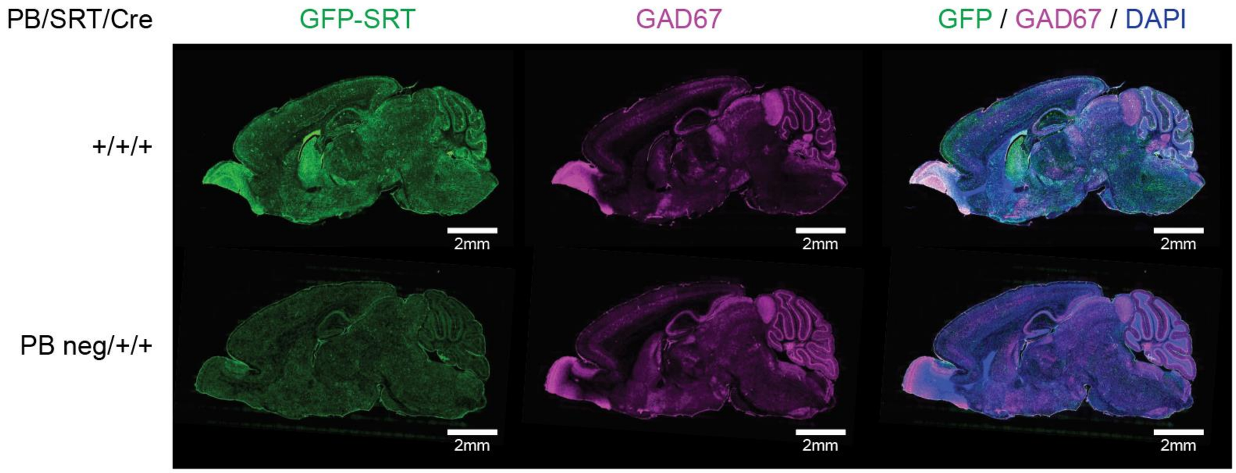


Figure 22: Vgat-CC mice successfully targets GABAergic neurons

Images of immunofluorescence sections from triple transgenic (PB^+/GFP-SRT^+/Vgat-Cre^+) and double transgenic (PB^-/GFP-SRT^+/Vgat-Cre^+) animals showing broad and high expression of SRT insertions (GFP; green) in the inhibitory neurons (GAD67; magenta) in the striatum, olfactory bulb, and cerebellum. Nuclei are stained blue with DAPI.

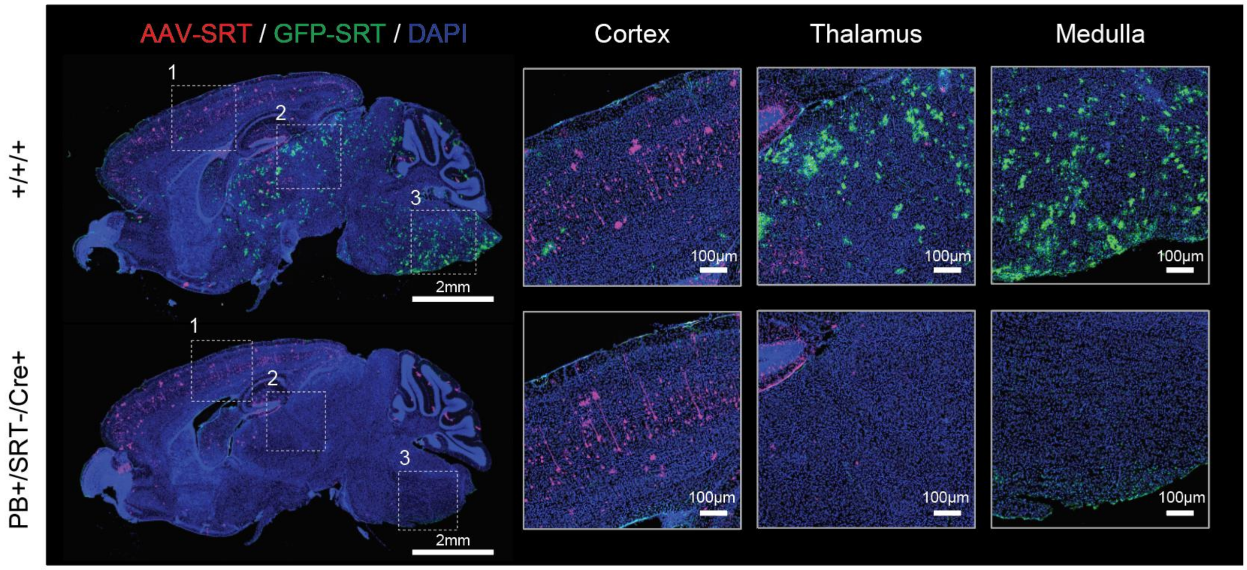


Figure 23: AAV-SRTs do not correlate with transgenic Rbp4-SRTs

Triple transgenic (PB$^+$/GFP-SRT$^+$/Rbp4-Cre$^+$) or double transgenic (PB$^+$/GFP-SRT$^-$/Rbp4-Cre$^+$) animals were injected with AAV9-tdTomato-SRT at P1 and harvested at P21. On the left, immunostained images of sagittal sections show cells expressing AAV-SRTs in magenta and cells expressing transgenic GFP-SRTs driven by Rbp4-PB in green. Nuclei are stained blue with DAPI. On the right, images zoomed into the cortex, thalamus, and medulla brain regions are shown.

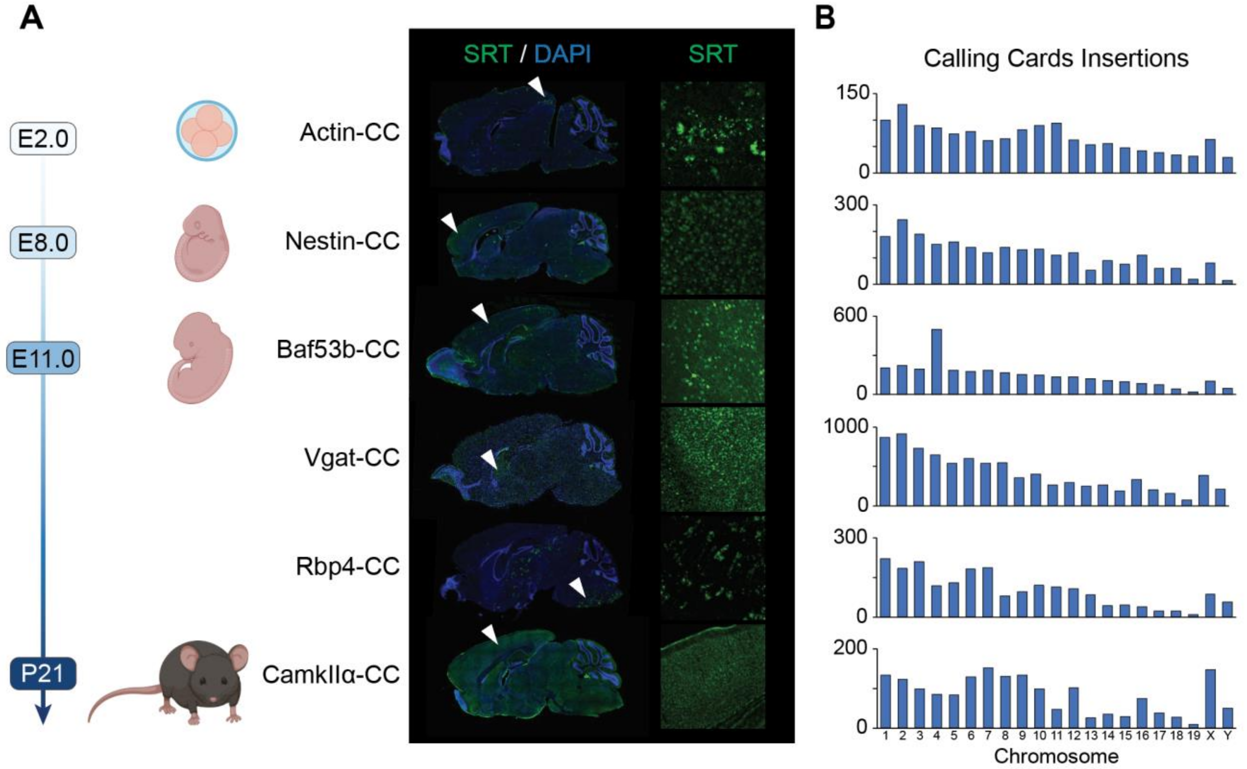


Figure 24: Testing a range of Cre driver lines to activate transgenic Calling Cards

(A) Schematic showing the developmental timeline of when Calling Cards should approximately be turned on depending on the Cre driver line. On the right, immunofluorescence images of show entire sagittal brain sections and the localization of GFP-SRTs. The white arrows point to the location of the zoomed image on the right. **(B)** Summary bar graph showing the number of recovered insertions per chromosome per transgenic Calling Cards line.

3.3 Discussion

Here, I tested a panel of Cre-driver lines that span embryonic to postnatal development from Actin-Cre that turns on during early embryogenesis, Nestin-Cre that targets neural progenitor cells, Baf53b-Cre that labels all post-mitotic neurons, Vgat-Cre for inhibitory neurons, Rbp4-Cre for layer V pyramidal neurons, and CaMKIIα-Cre for postnatal cortical and hippocampal neurons. My findings suggest that transgenic Calling Cards may be detrimental to early progenitor populations and may cause cell death as seen with Actin and Nestin-Cre crosses. This adverse effect seems to be mitigated when targeting more differentiated, post-mitotic neurons using Baf53b, Vgat, Rbp4, and CaMKIIα Cre lines. An interesting observation was made with the Baf53b-Cre-driven Calling Cards, where the vast majority of insertions mapped to chromosome 4. While immunofluorescence staining generally confirmed appropriate GFP expression localization and, in some cases, showed strong GFP expression, the yield of recovered insertions was disappointingly low. This was not attributed to issues with sequencing library preparation, as evidenced by robust tdTomato expression and insertion recovery in triple transgenic animals injected using AAV9 Calling Cards vectors. The correct localization of the transgenic and AAV SRTs suggests effective functionality of the transgenic *piggyBac* transposase, hinting at possible underlying biological complexities affecting the transgenic SRTs. These findings underscore the nuanced interaction between transgenic Calling Cards and cellular context, emphasizing the need for alternative approaches to engineer Calling Cards into the genome.

Given the varying outcomes observed between transgenic and AAV-mediated expression of the SRT, particularly in the case with Rbp4-Cre where both modalities were present in the same animal, we can speculate two possible factors contributing to these differences. The first is related to the timing of SRT expression: transgenic SRTs are integrated from embryonic day 0, whereas AAV-SRTs are introduced postnatally at P0. The second involves the genetic mechanisms at play: transgenic SRTs require excision from the genome, causing a double-strand break (DSB), whereas AAV constructs remain episomal, potentially reducing genomic stress. It is not far-fetched that many DSBs due to high Calling Cards activity can trigger genome instability and apoptosis.

A plausible theory for the low number of insertions can go back to ancient battles between transposons and their host organisms. To safeguard their integrity, mammalian genomes have co-opted sophisticated endogenous defense mechanisms, such as PIWI-interacting RNAs (piRNAs) and Argonaute proteins, aimed specifically at balancing the beneficial and detrimental consequences of transposon activity (Ernst et al., 2017; Wang and Lin, 2021; Wilhelm and Bernard, 2016). These mechanisms could effectively silence SRTs, particularly given that each contains a PGK promoter. Such silencing could occur both at the transcriptional and post-transcriptional levels, particularly if the SRT array containing tens to hundreds of copies is perceived as an anomaly given its high transcriptional activity and is silenced.

Although the attempt to establish transgenic Calling Cards disappointingly fell short of expectations, the insights garnered from these experiments shed valuable light on the intricate dynamics between transposons and host defense mechanisms. These findings will inform and refine our future alternative approaches for transgenic Calling Cards.

3.4 Materials and methods

Generation of PGK-GFP-SRT mouse line

The PGK-GFP-SRT transgene construct was linearized by performing a double digest of 10 µg circular plasmid pRM1671 with PstI-HF (New England Biolabs R3140) and KpnI-HF (New England Biolabs R3142) in rCutSmart Buffer (New England Biolabs B6004S) at 37°C for 1 hour. The products were run on a 1% agarose gel with GelGreen (Biotium 41005) and visualized using the Visi-Blue Transilluminator (UVP 95-0431-01). The 1.9 kb desired band was cut out of the gel using a clean scalpel and purified using the QIAEX II Gel Extraction Kit (Qiagen 20021) according to manufacturer's instructions. The purified linear DNA fragment was eluted using 0.2 µm filtered Transgene Injection Buffer (10mM Tris, 0.1mM EDTA, pH 7.4) and the concentration was quantified using the Qubit High sensitivity dsDNA Quantification Kit (ThermoFisher Q32854). The linear PGK-GFP-SRT was injected into 15 newly fertilized eggs from C57BL/6 mice each day for a total of 3 days. The animals that are born were then screened for GFP using the standard genotyping protocol described below. All founders were bred with WT C57BL/6 mice and the lines were maintained as heterozygotes. Each line was validated for transposon activity with transcranial injections of AAV9-hyPB and the candidate line was chosen based on the one where most Calling Cards insertions were recovered. The sperm from two proven male breeders was cryopreserved and stored in two separate locations for security.

Animals

All animal studies were approved by and performed in accordance with the guidelines of the Animal Care and Use Committee of Washington University in Saint Louis, School of Medicine and conform to NIH guidelines of the care and use of laboratory animals. The animals were housed in controlled environments with a 12-hour light-dark cycle, constant temperature and relative humidity, and *ad libitum* access to food and water. The following mouse lines were used: B6.FVB-

Tmem163$^{Tg(ACTB\text{-}cre)2Mrt}$/EmsJ (Actin-Cre; Jackson Laboratories strain #033984), B6.Cg-Tg(Nes-cre)1Kln/J (Nestin-Cre; Jackson Laboratories strain #003771), STOCK Tg(Actl6b-Cre)4092Jiwu/J (Baf53b-Cre; Jackson Laboratories strain #027826), B6J.129S6(FVB)-$^{Slc32a1tm2(cre)Lowl}$/MwarJ (Vgat-Cre; Jackson Laboratories strain #028862), B6.FVB(Cg)-Tg(Rbp4-cre)KL100Gsat/Mmucd (Rbp4-Cre; MMRRC 037128-UCD), B6.Cg-Tg(Camk2a-cre)T29-1Stl/J (CamKIIα-Cre; Jackson Laboratories strain #005359), and C57BL/6J (Jackson Laboratories strain #000664). The *Rosa26*$^{LSL\text{-}PB}$(fl/fl) mice were generously provided as a gift from Dr. Roland Rad. The transgenic lines were refreshed every 8-10 generations by backcrossing to freshly obtained C57BL/6J males and females from Jackson Laboratories. Upon weaning at P21, the animals were group-housed by sex and genotype.

Genotyping

Tissue (tail biopsy, ear punch, or toe clipping) was obtained from each animal and placed in a PCR tube. 100 µl lysis buffer (25mM NaOH, 0.2mM EDTA, pH 12) was added to each tube and incubated at 99°C for 60 min in a thermocycler. Once the samples cooled to room temperature, 100 µl 40 mM Tris-HCl pH 5 was added to neutralize the alkaline lysis buffer. The crude lysate containing genomic DNA (gDNA) was stored at 4°C. Cre driver lines were genotyped using Cre-F and Cre-R primers, which amplified a 450 bp product. ROSA26$^{LSL\text{-}PB}$ line was genotyped using BpA5F and Rosa3R primers, which amplified a 250 bp product. To differentiate between (fl/fl) and (fl/+), the Rosa5F and Rosa3R primers were used, which amplified a 450 bp product. All genotyping PCRs were multiplexed with β-actin_For and β-actin_Rev primers as this not only confirms the presence of gDNA, but also minimizes non-specific amplification. For each reaction, 1ul crude gDNA was mixed with 5 µl OneTaq Quick-Load 2X Master Mix (New England Biolabs

M0271), 1 µl 10µM SRY For/Rev primer mix, 1 µl 10µM β-actin For/Rev primer mix, and 2 µl ddH2O. See **Table 13** for primer sequences. PCR products were run on a 1% agarose gel and visualized with GelRed (Biotium 41003).

Intracerebroventricular injections

Injections were performed as described in the Intracerebroventricular Injection section within Basic Protocol 1 found in (Yen et al., 2023). Briefly, the pups were anesthetized on ice and a total of 6 µl (3 µl per hemisphere, 1 µl per site) was injected into the ventricles of P0-1 pups using a 50 µl Hamilton syringe. After the injections, the pups were kept warm on a heating pad until they were returned to their home cage.

Bulk Calling Cards library preparations

Tissue homogenization, RNA isolation, and library preparation steps are described in Basic Protocol 2 and 3 found in (Yen et al., 2023). Briefly, the dissected brain tissue was cut into 10 chunks to identify up to 10 independent insertion events at any given insertion locus, snap-frozen in the vapor phase of liquid nitrogen, then stored at -80°C until further processing. For homogenization, the tissue chunk homogenized in Trizol Reagent (ThermoFisher 15596018) and total RNA was harvested using the RNA Clean & Concentrator Kit-25 (Zymo Research R1018) with slight modifications as described in (Yen et al., 2023). GFP-SRTs were PCR amplified using the EGFP-C_For primer in place of the SRT_tdTomato_F1 primer for tdTomato-SRTs (sequence in Error! Reference source not found.). Bulk sequencing libraries were generated and sequenced o n the Illumina platform. Calling Cards found at the same insertion site were considered distinct if they had distinct barcodes (ie. came from different tissue chunks). Insertions that pass filtering were treated equally during analysis, regardless of read depth.

Immunofluorescence and imaging

Animals were deeply anesthetized with Isoflurane in an induction chamber until unresponsive to toe pinch. The mouse was initially perfused with ice-cold PBS to clear the circulatory system of blood, then with ice-cold 4% (w/v) paraformaldehyde for 10 mins. Following PFA perfusion, the tissues of interest were harvested, dissected for processing, and drop fixed in a tube containing 4% (w/v) PFA overnight at 4°C. Then the tissue was cryoprotected in 15% (w/v) sucrose, then 30% (w/v) sucrose at 4°C, then frozen in plastic molds (Polysciences 18646A-1) containing OCT compound (Fisher Scientific 23-730-571). The tissue blocks were kept at -80°C until further processing. Tissue was cut into 35 µm-thick sagittal or coronal free-floating sections for immunostaining. The sections were permeabilized with 0.1% (v/v) Triton X-100 for 15 mins and blocked with 5% (v/v) normal donkey serum (Jackson ImmunoResearch 014-000-121) for 60 mins. The primary antibodies used were chicken anti-GFP (Aves Lab GFP-1020) at 1:1000 dilution, rabbit anti-RFP (Rockland 600-401-379) at 1:500 dilution, and mouse anti-NeuN (Millipore Sigma MAB377) at 1:100 dilution. Secondary antibodies used were donkey anti-chicken Alexa Fluor488 (ThermoFisher A78948), donkey anti-rabbit Alexa Fluor568 (ThermoFisher A10042), and donkey anti-mouse Alexa Fluor647 (ThermoFisher A31571). 1 µg/ml DAPI (ThermoFisher D1306) was used to stain the nuclei blue. Sections were mounted onto slides with Prolong Gold anti-fade mounting medium (ThermoFisher P36934) and sealed with nail polish. Low magnification widefield slidescans of entire stained sections were captured using a 10x objective on the Axioscan 7 (Zeiss). High magnification confocal images were captured using 20x or 63x objectives on the LSM700 AxioImager Z2 (Zeiss).

3.5 Data and code availability

The raw data, processed data, and code used to analyze the data are available upon request.

3.6 Acknowledgements

We thank members of the Dougherty and Mitra laboratories for helpful discussions and feedback; M. Wallace and the Mouse Genetics Core for their services and expertise in generating the PGK-GFP-SRT transgenic mouse; M. Li and the Hope Center Viral Vectors Core at Washington University School of Medicine for generating the AAV vectors; the DNA Sequencing and Innovation Lab (DSIL) and the Genome Technology Access Center at the McDonnell Genome Institute (GTAC@MGI) for their sequencing expertise and services; and the Washington University Center for Cellular Imaging (WUCCI). This work was funded by grants from the National Institute of Mental Health (RF1MH117070, RF1MH126723, and R01MH124808 to J.D.D. and R.D.M.). A.Y. was supported in part by T32HG000045 from the National Human Genome Research Institute.

3.7 Author contributions

Project conceptualization: A.Y., R.D.M., and J.D.D. Method development, experiments, and data collection: A.Y. Formal analysis: A.Y. Figures and data visualization: A.Y., R.D.M., and J.D.D. Project coordination: A.Y., R.D.M., and J.D.D. Funding acquisition: R.D.M. and J.D.D.

3.8 Tables

Table 13: Primer sequences for Chapter 3

Primer Name	Sequence (5’ →3’)	Product
SRY_For	TTGTCTAGAGAGCATGGAGGGCCATGTCAA	273 bp
SRY_Rev	CCACTCCTCTGTGACACTTTAGCCCTCCGA	
GFP_For	CCTACGGCGTGCAGTGCTTCAGC	350 bp
GFP_Rev	CGGCGAGCTGCACGCTGCGTCCTC	
β-actin_For	AGAGGGAAATCGTGCGTGAC	150 bp
β-actin_Rev	CAATAGTGATGACCTGGCCGT	
BpA5_For	GCTGGGGATGCGGTGGGCTC	250 bp
Rosa3_Rev	GGCGGATCACAAGCAATAATAACCTGTAGTTT	
Rosa60_For	CTCTCCCAAAGTCGCTCTG	450 bp
Rosa62_Rev	TACTCCGAGGCGGATCACAAGC	
Cre_For	CCGGTCGATGCAACGAGTGATGAGGTTC	443 bp
Cre_Rev	GCCAGATTACGTATATCCTGGCAGCG	
MYT1L_Comm_For	CCAAGTCCTGTCCTACCCAAGT	380 bp
MYT1L-WT_Rev	TCTTGCTACACGTGCTACT	
MYT1L-Mut_Rev	TCTTGCTACACGTACTGGA	

(Continued)

Table 13: Primer sequences for Chapter 3, *continued*

Primer Name[a]	**Sequence (5’ →3’)**[b]	**Product**
SMART_dT18VN	AAGCAGTGGTATCAACGCAGAGTACGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN	n/a
SMART_Rev	AAGCAGTGGTATCAACGCAGAGT	n/a
EGFP-C_For	CATGGTCCTGCTGGAGTTCGTG	n/a
OMPB_BC-GAT_Index2_X	AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATCGTCAATTTTACGCAGACTATCTTT	n/a
N7_Index_X	CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG	n/a

[a] The three bases highlighted in blue indicate the location of the OM-PB primer barcode. For more information, refer to “ Considerations for primer selection and ordering for sequencing libraries” in Chapter 2.3.

[b] The [i5] and [i7] represent placeholders where unique indexes can be added for multiplexing multiple samples to be sequenced in a single run. For more information about standard Illumina indexes, see https:// supportdocs.illumina.com/SHARE/AdapterSeq/Content/SHARE/AdapterSeq/AdapterSequencesIntro.htm, or consult your sequencing core or service provider.

Chapter 4: Decoding sex-specific gene expression: The role of hormones, transposable elements, and epigenetics in brian development

4.1 Preface

This chapter contains contents from a manuscript in preparation:

Decoding sex-specific gene expression: The role of hormones, transposable elements, and epigenetics in brain development

Allen Yen, Robi D. Mitra, Joseph D. Dougherty

4.2 Abstract

The exploration of sex differences in gene expression and their impact on brain masculinization is crucial to understanding fundamental developmental processes. While it is acknowledged that the brain's transcriptional landscape varies between sexes, particularly during pivotal perinatal hormonal surges, the role of enhancer elements in mediating these differences remains less understood. In this context, Brd4-bound enhancers are of particular interest due to their proven significance in defining cell identity and influencing transcriptional states. Utilizing the Calling Cards technology, this study provides a detailed examination of Brd4-bound enhancer usage across three developmental stages, providing insights into the transcriptional dynamics underpinning brain development and sexual differentiation.

I focus on cataloging Brd4-bound enhancer activity before the onset of steroid hormone influence, aiming to uncover inherent sex-specific enhancer usage. Subsequent phases investigate the impact of the perinatal testosterone surge and its lasting effects beyond this hormonal peak, employing an approach to record enhancer activity and associated gene regulatory elements to eventual epigenetic states across male and female developmental trajectories.

By integrating Calling Cards data with genomic and epigenetic analyses, this study delineates a nuanced landscape of sex-differential enhancer activity, uncovering potential mechanisms through which early hormonal environments shape neural circuitry and behavior. Moreover, by probing the intersections between transposable elements and enhancer regions, this study elucidates novel facets of neurodevelopmental biology, offering a comprehensive resource that deepens our comprehension of the molecular underpinnings of sex-specific brain development and its implications for neuropsychiatric disorders.

4.3 Introduction

Biological phenotypes, whether in humans or other organisms like mice, show sex-specific characteristics. These differences can partially be attributed to hormones and sex chromosomes, however the full extent of how sex differences contribute to normal physiology and how diseases and disorders arise when these processes go wrong is still not fully understood. There have been large consortiums efforts like the Genotype-Tissue Expression (GTEx) (Lonsdale et al., 2013) that have identified tissue and cell type specific effects and has been critical in associating gene expression, transcription factor binding, chromatin state, and genome-wide association studies to link gene with function, and re-analysis of this data shows subtle yet reproducible changes in gene expression between the sexes across all regions of the brain (Fass et al., 2023).

The developing brain undergoes intricate processes influenced by internal genetic differences and external factors during neurodevelopment. While males and females share some common genetic regulators that orchestrate brain development, their developmental trajectories differ due to varying responses to external influences such as the uterine environment and postnatal exposures and experiences. A key factor that differentiates these developmental pathways is the exposure to gonadal hormones during critical periods (Matsumoto et al., 2003; Ogawa et al., 2000). This hormonal surge activates epigenetic regulators (Bramble et al., 2016), thereby establishing cellular memories in alignment with the organization-activational hypothesis (Phoenix et al., 1959). This hypothesis serves as a cornerstone for understanding sexual dimorphism in the brain, suggesting that hormonal signals during key developmental stages not only prepare the neural circuits for sex-specific behaviors, but also prime these circuits for activation later in life.

Experiences throughout life interact with genetic predispositions to shape the organism's or individual's behavior, termed context-dependent epigenetics (Crews, 2011). This can be driven by internal factors like testosterone by binding to cytosolic androgen receptors, which then function as ligand-activated transcription factors to activate expression of target genes (Davey and Grossmann, 2016). Moreover, recent findings also highlight the role of estrogen receptor-α (ERα) in modulating sex differences in gene expression within known sexual dimorphic brain regions and its implications for behavior (Gegenhuber et al., 2022). The interaction between intrinsic genetic differences and external factors prompts a deeper investigation of the epigenetic landscape and gene regulatory elements. Unraveling the molecular mechanisms by which early hormonal environments shape brain development and function is critical to understanding neuropsychiatric disorders, which consistently exhibit marked sex biases in prevalence, symptomatology, and treatment responses.

Sexual dimorphism emerges within the first few weeks of postnatal development, following a perinatal testosterone surge that is rapidly cleared from circulation within a few hours (McCarthy, 2008). While conventional genomic methods provide snapshots of the epigenetic states at discrete time points, requiring sequential sampling to capture the evolving developmental landscape, Calling Cards records an integrated and continuous epigenetic profile. This approach enables the correlation of early molecular interactions and gene regulatory elements with eventual cellular states. Calling Cards consists of two main components: a self-reporting transposon (SRT) and a *piggyBac* transposase, which specifically targets acetylated genomic regions marked by Brd4, indicative of active enhancers (Cammack et al., 2020; Moudgil et al., 2020b). Upon binding, the *piggyBac* transposase inserts the SRT into the genome, effectively tagging these interaction locations with a permanent mark. The accumulations of these insertions throughout the experimental timeline yields a record of enhancer usage, which can be analyzed using next-generation sequencing to present a comprehensive view of epigenetic activity during crucial developmental periods. In this study, I used Calling Cards to map enhancer activity across 3 epochs: (1) before the testosterone surge to determine preexisting sex-biased enhancers; (2) through the testosterone surge to pinpoint hormone-responsive elements; and (3) after the surge to evaluate whether testosterone induces lasting epigenetic alterations. This integrated analysis illuminates sex-specific gene regulatory mechanisms, offering valuable insights into the epigenetic underpinnings of sexual dimorphism. One class of regulatory sequences that are able to modify host gene expression that have been underexplored especially in its role to drive sex-biased gene expression are transposable elements.

Transposable elements (TEs) are mobile genetic elements that make up at least 50% of the human genome, having expanded throughout evolution (International Human Genome Sequencing Consortium et al., 2001). A significant class of TEs are retrotransposons, which replicate in the genome by transcribing a RNA intermediate, akin to retroviral replication. These retrotransposons can be categorized into two types: those with long terminal repeats (LTR) and non-LTR elements. The non-LTR elements are the most common and have also lost their ability to retrotranspose. Along the non-LTR elements, the most prevalent are the long interspersed nuclear elements (LINEs), with only a minority that are still active (Philippe et al., 2016). Additionally, the short interspersed nuclear elements (SINEs) represent another group of active TEs that have garnered attention due to their potential to influence genomic integrity and function. These elements, particularly when they insert into gene promoters or bodies, can have mutagenic effects, altering gene expression or disrupting gene function, thereby playing a critical role in genomic variation and evolution. There have been studies that show that increased LINE-1 activity has been associated with schizophrenia and depression (Doyle et al., 2017; Liu et al., 2016). Recent research indicates that Alu elements, a subset of SINEs, are prevalent at enhancer-promoter RNA interaction sites, suggesting a role in gene regulation (Liang et al., 2023). Additionally, dormant regulatory sequences within TEs can be epigenetically activated, potentially leading to oncogene activation and tumorigenesis (Babaian and Mager, 2016; Jang et al., 2019). These findings underscore TEs' complex roles in gene expression and disease. However, their contribution to sex-specific gene expression remains to be fully understood, representing a critical area for further investigation to understand the genetic and epigenetic mechanisms underlying sex differences in biology.

4.4 Results

4.4.1 Sex differential Brd4 enhancer usage across embryonic development

The link between brain masculinization and sex-specific variations in gene expression is well-documented. It is not clear if these transcriptional differences were attributable to sex-specific enhancer activity and if transcriptional states were stable or transient. Here, I focused on Brd4-bound enhancers because there is established evidence showing that these enhancers play key roles in establishing cell identity (Dowen et al., 2014; Hnisz et al., 2013; Whyte et al., 2013). Calling Cards is a platform technology that enables the molecular recording of transient protein-DNA interactions over time, enabling the association of observed phenotypes with historical cell states (Cammack et al., 2020; Moudgil et al., 2020b).

I first sought to catalog Brd4-bound enhancer usage in three developmental phases. The initial phase examines early development, specifically before the influence of steroid hormones, to determine if sex-specific enhancer usage exists independently of hormonal effects that typically modulate gene transcription. The second phase aligns with the perinatal testosterone surge in males to observe how this hormonal event influences enhancer activity. The third phase extends beyond this surge, through a hormonally quiescent period, up to the onset of the pubertal hormonal surge (**Figure 25**). By cataloging enhancer usage across these stages, I can identify enhancers that display distinct patterns of activity between males and females, and associate potential genes that may be downstream of these gene regulatory elements.

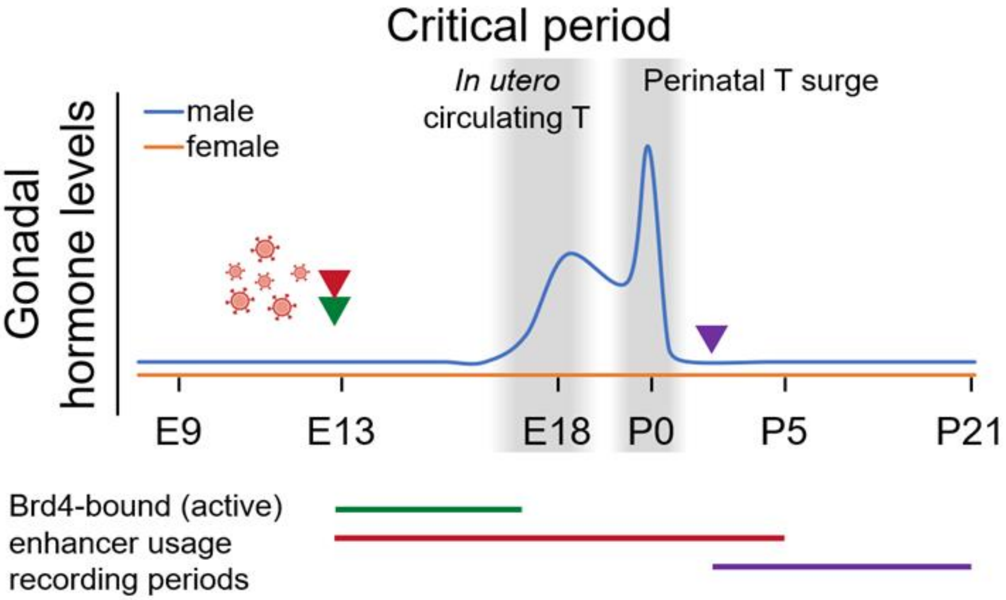


Figure 25: Calling Cards records perinatal Brd4-bound enhancer activity in male and female mice

AAV9 Calling Cards were used to record enhancer usage across three developmental windows: pre-testosterone surge (E13-E17; green), before and throughout testosterone surge (E13-P5; red), and after the testosterone surge (P2-P21; purple). The arrowheads indiviate delivery of Calling Cards reagents to initial epigenetic recording, and animals were harvested at the end of the bar to end recording. The blue and orange lines illustrate how testosterone levels fluctuate in males and females respectively throughout embryonic and early postnatal development (adapted from (McCarthy, 2008)). The arrows indicate the developmental age at which the AAVs were injected for their respective groups.

To record enhancer usage during the pre-testosterone surge phase, I injected Calling Cards into the ventricles of E13 embryos and harvested the brains at E17, prior to the perinatal testosterone surge in males. After this brief 4-day viral transduction, a substantial number of Calling Cards insertions were recovered—880,441 in males (n=5) and 1,376,183 in females (n=9)—suggesting that adeno-associated viruses (AAVs) are capable of inducing Calling Cards transgene expression and recording activity in a relatively short period of time (**Figure 26A**, see **Table 14** for sample details). To analyze hormone-responsive Brd4-enhancer activity, I injected AAV-Calling Cards into the ventricles of another cohort of E13 embryos which recorded enhancer usage through the P0-P1 perinatal testosterone surge, before harvesting the brain tissue at P5, a time point long after the circulating testosterone has been cleared. There were 3,091,069 recovered insertions in males (n=6) and 2,910,552 in females (n=5) (**Figure 26A**, see **Table 14** for sample details), with this greater number than the E17 harvest consistent with a longer recording window. Finally, to determine whether enhancer usage is only altered transiently in the presence of testosterone or if it is fundamentally changed, a third cohort of animals were injected post-testosterone surge at P2 and were analyzed at P21, a period that is hormonally-quiescent and before the pubertal hormone surge. This final group yielded 5,695,851 insertions in males (n=6) and 4,607,802 in females (n=5) (**Figure 26A**, see **Table 14** for sample details). Pairwise comparisons of the distribution of Calling Cards insertions within peaks show high correlation between samples within the same age group across sexes, with age accounting for most of the variance (**Supplemental Figure 8**). The Calling Cards insertion profile from the E13-E17 group was more similar to the females than males from the E13-P5 group, indicating that Calling Cards effectively captured the differential epigenetic activity induced by the hormonal surge. The lower correlation of these two groups with P2-P21 males and females may be due to the postnatal timing of the injections of Calling Cards at P2 which likely targeted a different population of cells. Additionally, the comparison of males and females from P2-P21 shows some differences, suggesting lasting organizational effects of the testosterone surge, even in the absence of circulating hormones. Overall, the analysis of total insertions across these periods revealed a consistent recording rate of slightly more than 500,000 insertions per day, demonstrating the capability of Calling Cards to continuously capture ongoing epigenetic activity throughout these developmental stages (**Supplemental Figure 9**).

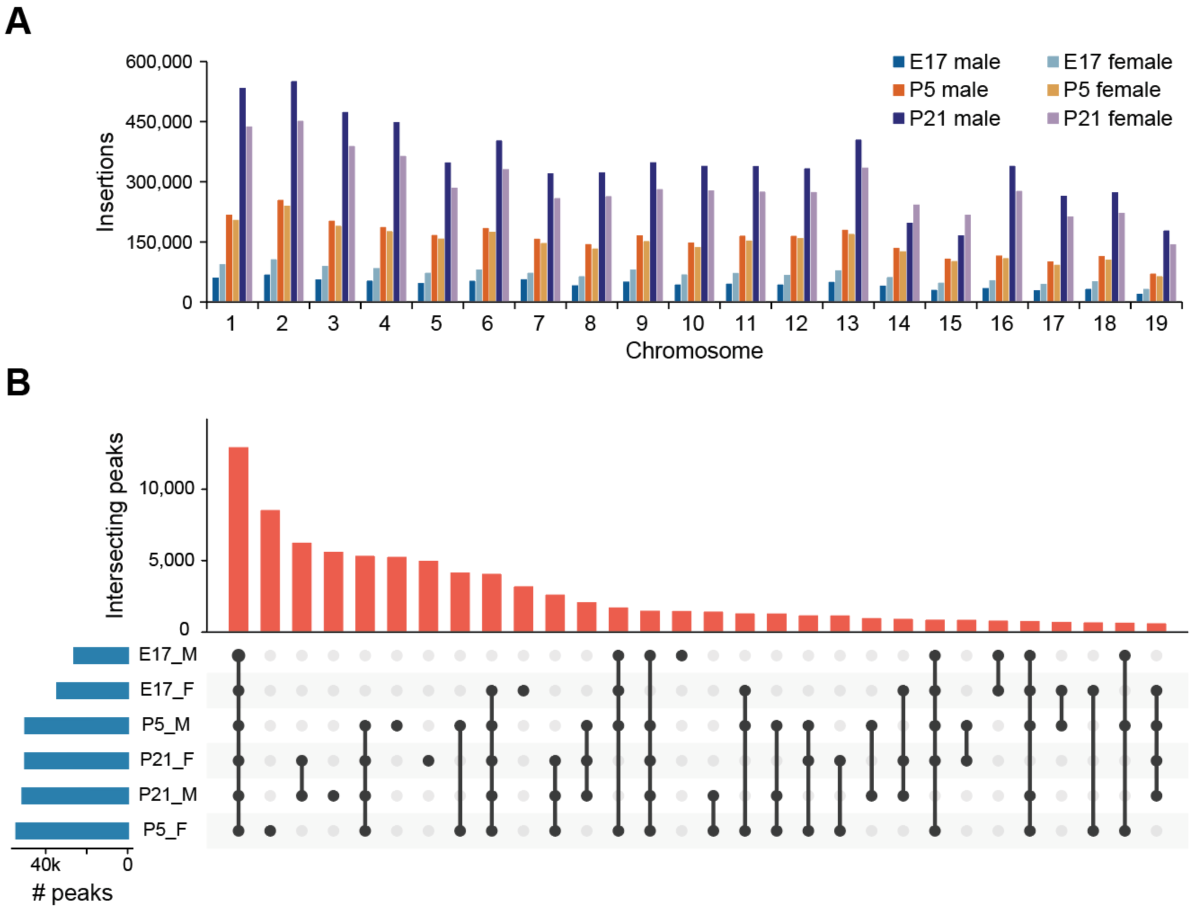


Figure 26: Calling Cards records enhancer usage across development

(A) Bar plot showing the distribution of recovered insertions per chromosome across male and female samples from the pre-testosterone surge, throughout surge, and post-surge cohorts. The insertions from chrX and chrY were also recovered but omitted from analysis. See **Table 14** for sample details. **(B)** Upset plot showing the intersection of Brd4 peak regions.

Next, I identified male and female-specific genomic regions that showed a high number of Calling Cards insertions during different developmental stages using CCcaller from Pycallingcards (Guo et al., 2024). A total of 12,928 peaks were found to overlap across all samples from E13-P21, representing 24-49% of the peaks in each individual sample (**Figure 26B**). GO analysis of the nearest gene revealed an enrichment of genes related to key neurodevelopmental processes like axon guidance, axonogenesis, and ephrin receptor signaling, underscoring their importance across developmental stages. To discern sex-specific variations, I first combined male and female insertions within the E13-E17 cohort to identify a joint set of peaks. Then I performed differential analysis using Fisher's exact test to identify peaks that had significantly different numbers of male or female insertions within each peak region (**Figure 27A-C**). This illustrated that even prior to the hormonal surge, some epigenetic differences exist between the male and female developing brain.

To understand which DNA-binding TFs might be driving the recruitment of Brd4 to these regions, I conducted a binned motif enrichment analysis by grouping the sex-biased peaks into bins based on their log-fold change in male or female enrichment. The motif enrichment is then calculated for each bin, which normalizes the differences in sequence composition and returns high confidence motif calls. This uncovered 8 TFs associated with male Calling Cards peaks and 30 TFs with female peaks (**Figure 27D**). Key female-biased TFs included ESRRB, RARA, RARB, and RXR. One of the highly enriched motifs was IRF9, which has been shown to be a key upstream regulator of sex-biased functional pathways before the appearance of sex hormones (Deegan et al., 2019). Further GO analysis linked male-biased peaks to immune function and transcription regulation (**Figure 27E**), while female-biased peaks correlated with neural development and differentiation (**Figure 27F**), suggesting early sex differences in enhancer usage that hormones may later amplify.

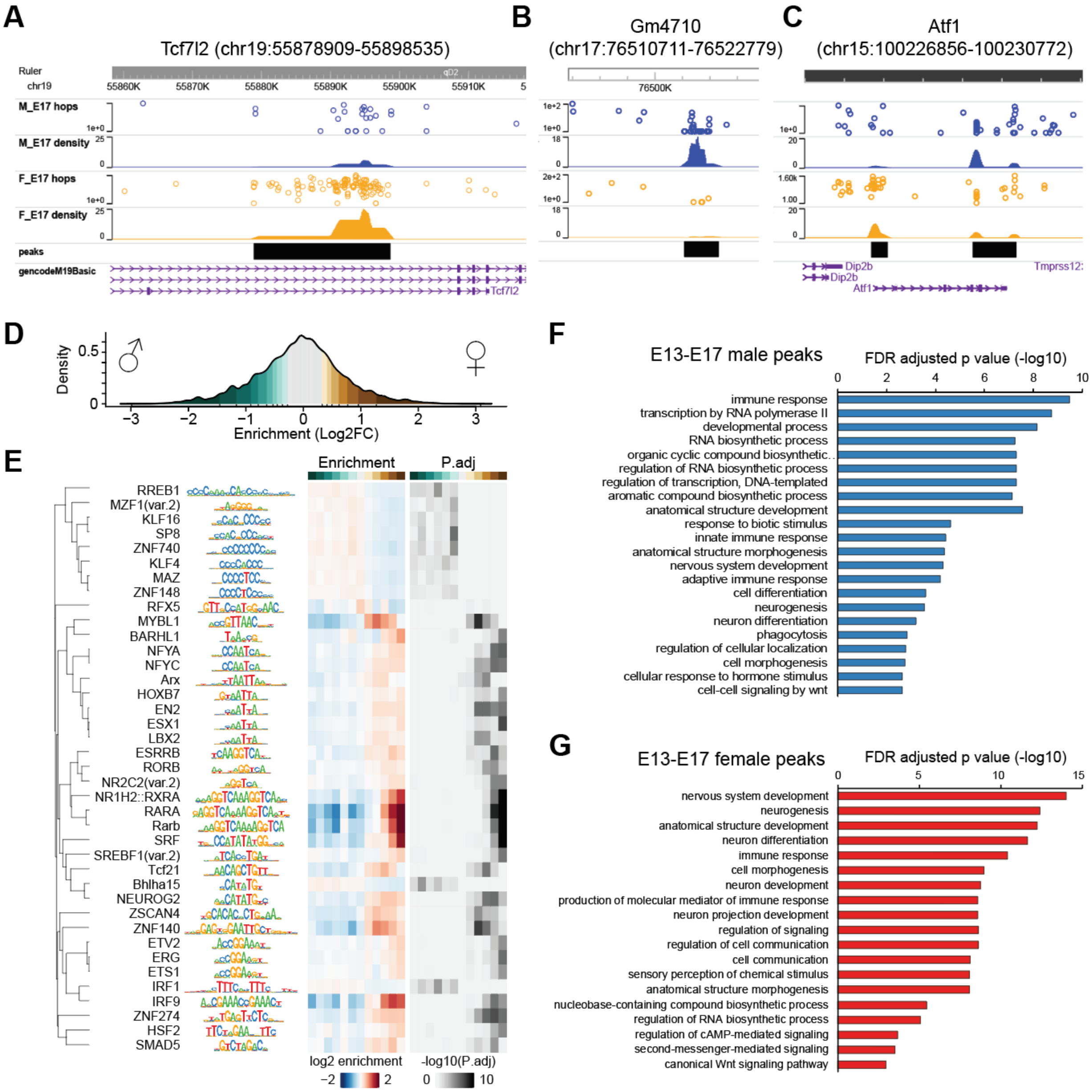


Figure 27: Identification of sex differential gene regulatory elements prior to perinatal testosterone surge

Genome browser tracks show a representative view of the Tcf7l2 region that shows female-biased Calling Cards insertions in **(A)**, non-coding region Gm4710 that shows male-biased insertions in **(B)**, and both male- and female-biased regions within the same Atf1 region in **(C)**. **(D)** Distribution of log2-fold changes in Calling Cards insertions.

The distribution is then binned; dark green indicates regions with high enrichment of male insertions, while brown denotes those significantly enriched in females. **(E)** Motif enrichment analysis of the binned genomic regions identify TFs that were found in male-biased regions (green) and female-biased regions (brown). The sequence logo representation of the point weight matrix of each motif is shown. The left heatmap shows the log2 enrichment score and the right heatmap shows the -log10 of the adjusted P value. GO analysis of genes associated with male-biased peaks are in **(F)** and females in **(G)**.

4.4.2 Sex differential Brd4 enhancer usage across brain masculinization

The organizational-activational hypothesis posits that perinatal testosterone exposure masculinizes the brain by organizing the tissue in such a way that they are activated and respond differently to gonadal hormones during puberty, thus underpinning behavioral sex differences (Phoenix et al., 1959). The brief testosterone surge in males around birth, lasting only hours, is proposed to lastingly shape neural circuits during this pivotal developmental window, though the precise genomic and molecular mechanisms remain partly elusive. Given that Brd4 has been implicated in driving sex-specific gene expression and its associated enhancers are crucial for cellular differentiation (Kfoury et al., 2021; Lee et al., 2017). I aimed to explore whether the perinatal testosterone peak might activate Brd4 enhancers, influencing sex-biased gene expression.

Using Calling Cards as described above, Brd4-bound enhancers were profiled in male and female cortices from E13 to P5, revealing 34,953 shared peaks, with an additional 8,528 unique to females and 5,253 to males (**Figure 26B**). GO analysis indicated that male-biased peaks were enriched for pathways like axon guidance and response to transforming growth factor beta (TGFβ), along with complement activation and phagocytosis (**Figure 28A**). In contrast, female-biased peaks were linked to mRNA metabolic processes, mRNA stabilization, and Notch signaling (**Figure 28B**). The observed male-specific enrichment of pathways such as the complement system and phagocytosis might reflect a heightened responsiveness or a particular sensitivity to testosterone.

To identify if there are any sex biased gene regulator elements associated with these genomic regions, I conduced a binned motif enrichment analysis of the male and female peaks. This revealed a significant increase in detected TF motifs during the perinatal testosterone surge period, showing an increase from 39 motifs to 239 motifs post-testosterone surge, with 120 in female and 111 in male peaks (**Figure 28C**). Notably, male-enriched motifs included TFs like Zic2 and Foxa2, which are known to have sexual dimorphic activities (Dhakal et al., 2020; Kfoury et al., 2021). This analysis yields a comprehensive catalog of sex-differential enhancer activity and gene regulatory elements, offering insights for further exploration of sex-based gene expression differences during brain development.

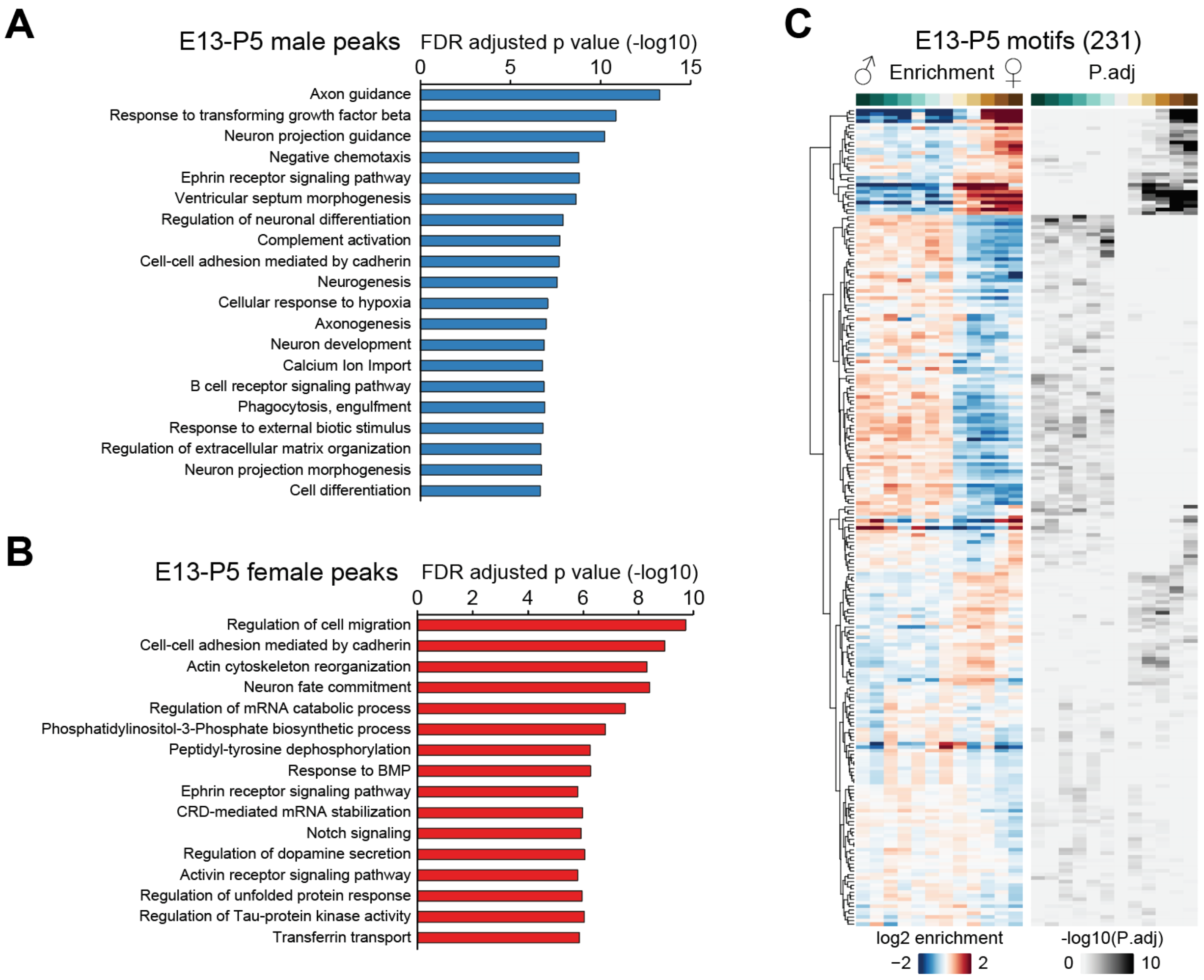


Figure 28: Characterization of Brd4 enhancer usage across brain masculinization

GO analysis of enhancers used during the perinatal testosterone surge from male enriched peaks in **(A)** and female enriched peaks in **(B)**. **(C)** Motif enrichment analysis of the binned genomic regions identify TFs that were found in male-biased regions (green) and female-biased regions (brown). The left heatmap shows the log2 enrichment score and the right heatmap shows the -log10 of the adjusted P value.

4.4.3 Sex differential Brd4 enhancer usage across postnatal brain maturation

To determine whether the perinatal testosterone surge exerts sustained effects on enhancer utilization, Calling Cards were injected into pups at P2, a time when the perinatal testosterone should be cleared from circulation (McCarthy, 2008), and enhancer usage was recorded until P21.

This P2-P21 dataset was analyzed similar to the E13-P5 dataset. First, joint peaks were called which identified 5,597 male-specific and 4,969 female-enriched peaks. GO analysis linked male peaks to processes such as axon guidance, Semaphorin-Plexin signaling pathway, and heparan sulfate proteoglycan biosynthetic processes, which are crucial for neural circuit development (**Figure 29A**). Female peaks were enriched in functions related to ion transport, GTPase signal transduction, and TOR signaling, which are essential for cell communication and survival (**Figure 29B**). Considering the complexity of the forebrain and cortex, alterations in guidance or morphogenetic pathways could critically affect neural wiring and, hence, cognitive outcomes. Enhanced activity at enhancers near genes involved in these pathways might alter neuronal receptor complex compositions, potentially reflecting organization effects induced by early testosterone exposure.

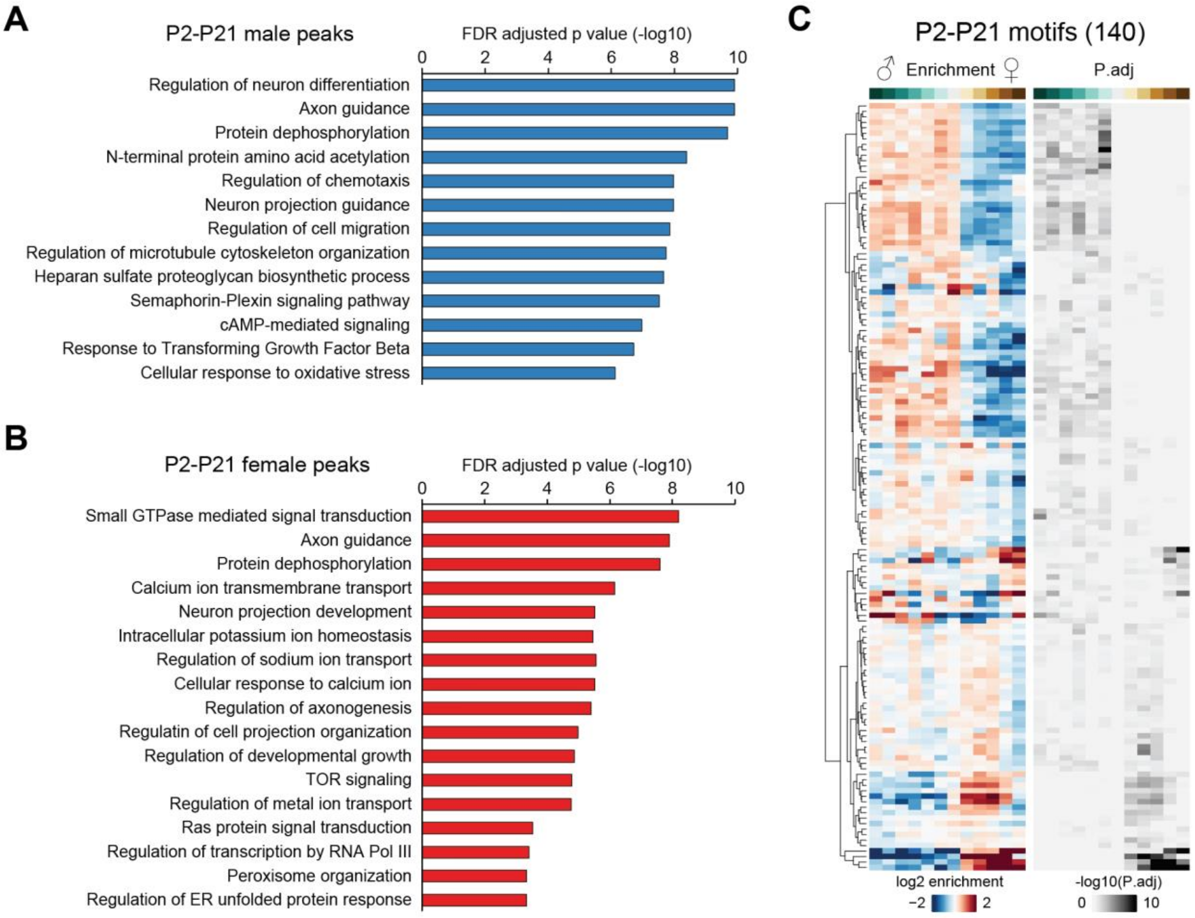


Figure 29: Characterization of Brd4 enhancer usage across postnatal brain maturation

GO analysis of enhancers used during postnatal brain maturation from male enriched peaks in **(A)** and female enriched peaks in **(B)**. **(C)** Motif enrichment analysis of the binned genomic regions identify TFs that were found in male-biased regions (green) and female-biased regions (brown). The left heatmap shows the log2 enrichment score and the right heatmap shows the -log10 of the adjusted P value.

To evaluate if the testosterone pulse modified enhancer profiles via epigenetic changes, I compared enhancer activity across different developmental stages in males and females. Developmental stage accounted for most of the variation in insertion patterns rather than sex, with the data showing limited overlap between the E13-E17 and E13-P5 datasets (**Figure 30A**). Since these were injected as two separate cohorts, slight variation in the gestational age of the embryos can affect which cells were at the ventricular zone at the time of in utero injections. Another potential source of variation can be if the chromatin accessibility changes induced by testosterone have placed the Calling Cards insertions in a region of closed chromatin. Thus, the SRT is not expressed and those insertions are not recovered. Notably, post-testosterone phases show a marked increase in motif enrichment within enhancer regions, indicating significant alterations in the epigenetic landscape and enhancers used, suggesting the potential for lasting impact from the perinatal hormone surge (**Figure 30B**). A small number of motifs were commonly enriched in male samples across all the time points: MYBL1, RFX5, NEUROG2, SRF, and ZNF274 (**Figure 30C, D**). The zinc-finger protein ZNF274 has been shown to recruit the histone H3 lysine 9 (H3K9) methyltransferase SETDB1 to repress maternal gene expression in neurons (Langouët et al., 2018).

The use of Calling Cards here has not only traced enhancer activity through a pivotal neurodevelopmental period, but also established a comprehensive catalog of active genomic regions. This dataset serves as a resource for neurodevelopmental biologists, providing insights into the epigenetic mechanisms that may underpin developmental and sex-specific differences in brain function.

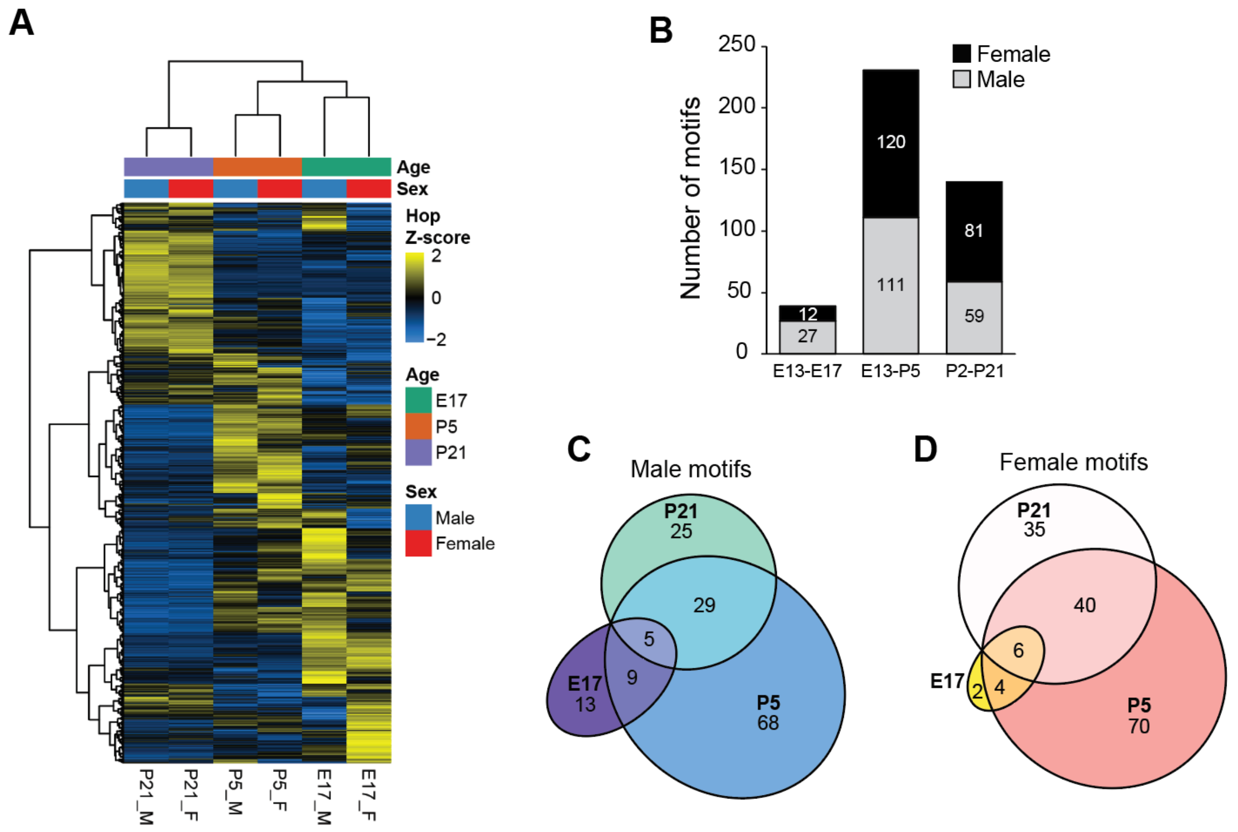


Figure 30: Enhancer activity and motifs across brain development, masculinization, and maturation

(A) Heatmap showing the Z-scaled number of insertions per peak region (rows) per sample (columns). **(B)** Summary plot showing the number of male and female-specific enriched motifs per developmental phase based on Calling Cards peaks. **(C, D)** Venn diagrams showing the number of intersecting motifs across developmental phases in males and females. The E13-E17 cohort is labeled as E17, E13-P5 cohort is labeled as P5, and the P2-P21 cohort is labeled as P21.

4.4.4 Sex differential enrichment of insertions in transposable elements

Historically, the role of sex differences in biological research has been undervalued, yet recent initiatives highlight how sex chromosomes and hormones critically affect gene expression and epigenetic patterns. A study identified 4,164 genes with sex-differentiated expression at puberty, aligned with a hormonal surge impacting brain development and behavior (Shi et al., 2016). Despite these findings, the specific mechanisms underpinning sex-biased gene expression during early brain masculinization are not fully understood.

Recent studies emphasize the significance of transposable elements (TEs), which are mobile genetic sequences, in modulating gene expression throughout embryonic and postnatal brain development. Here, I aim to investigate whether TEs could act as key regulatory elements, modulating sex-specific gene expression during the perinatal testosterone surge, which may offer new insights into the intricate interplay between regulatory elements and sex hormones during critical developmental windows.

To identify TE-associated Calling Cards peaks, I filtered the peaks to only those that are within 5kb of an annotated TSS and intersected with the RepeatMasker database to identify interspersed repeats and low complexity DNA sequences (Tarailo-Graovac and Chen, 2009). This approach resulted in 1,857 peaks containing 189,495 insertions for the E13-E17 group, 4,295 peaks containing 374,418 insertions for the E13-P5 group, and 13,294 peaks containing 458,731 insertions for the P2-P21 group (**Figure 31A**). Classification of insertions revealed a predominance of LINEs, SINEs, and ERVs (**Figure 31B**). Interestingly, the probability of a Calling Cards insertion to land within TEs like LINE-1, LTR-ERVs, and Alu SINEs was higher in males across developmental stages (**Figure 31C**), suggesting an intrinsic sex bias in these genomic features.

Notably, this bias was present across all three epochs studied, suggesting it does not depend on the testosterone surge.

Next, I performed motif enrichment analysis to investigate the regulatory mechanisms of these sex-biased TEs. I found 72 motifs in males and 42 motifs in females (complete list with consensus sequences can be found in **Table 17** and **Table 18**). The most significant motifs in females were linked to interferon (IFN) signaling, including IRF1, IRF2, and IFN-stimulated response elements (ISREs). In males, the prominent motifs were associated with nuclear receptors such as SF1, the estrogen responsive element (ERE), retinoic acid receptor gamma (RARg), and androgen response element (ARE). This pattern aligns with the known testosterone surge in males and the activation of nuclear receptors. These findings suggest that TEs in gene promoter regions may serve as regulatory elements that can be one of the factors that drive sex-specific gene expression. Moreover, the ongoing mobility of these TEs suggests they could act as dynamic modules influencing gene regulation by biological sex. Further studies are needed to test this speculative model.

Exploring the relevance to neuropsychiatric disorders, I utilized the SFARI-gene (Abrahams et al., 2013; Banerjee-Basu and Packer, 2010) and ARCHS4 databases (Lachmann et al., 2018) to examine associations with autism spectrum disorders. Notably, TE-regulated genes from both pre-and post-testosterone surge samples showed significant overlap with SFARI high confidence autism-related genes (score 1 or 2), suggesting these TE-influenced genes might contribute to sex-specific vulnerabilities in neuropsychiatric conditions (**Figure 32**). This study introduces a novel paradigm for understanding sex-biased gene expression, demonstrating that TEs in gene promoter regions not only serve as dynamic regulatory elements, but also suggest mechanisms through which sex-specific genetic regulation may contribute to differential susceptibility in neuropsychiatric disorders.

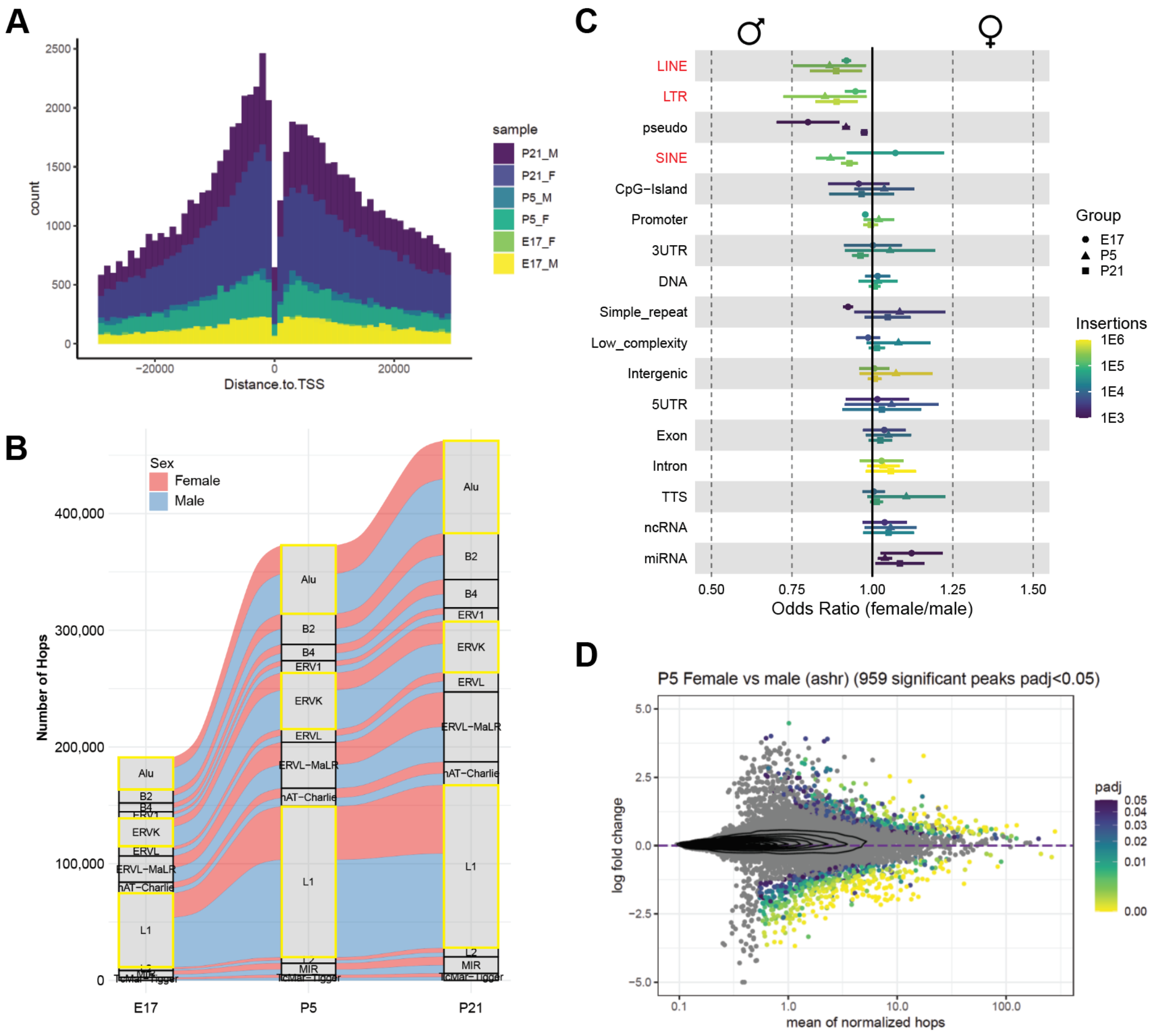


Figure 31: Calling Cards insertions in transposable elements

(A) Histogram showing the distribution of Calling Cards peaks relative to annotated transcription start sites (TSS). **(B)** Alluvial plot showing the number of male and female Calling Cards insertions within promoter regions that overlap with different classes of transposable elements. L1, ERVK, and Alu elements (highlighted in yellow) were found to be sex-biased. **(C)** Forest plot showing the odds ratio of a Calling Cards insertion landing with different classes of TEs or genetic elements. Element names highlighted in red were found to be sex-biased. **(D)** MA plot showing the differential Calling Cards insertions in peaks within TEs. Peaks with male enrichment are shown with negative log fold changes, while female enriched peaks are shown with positive log fold changes.

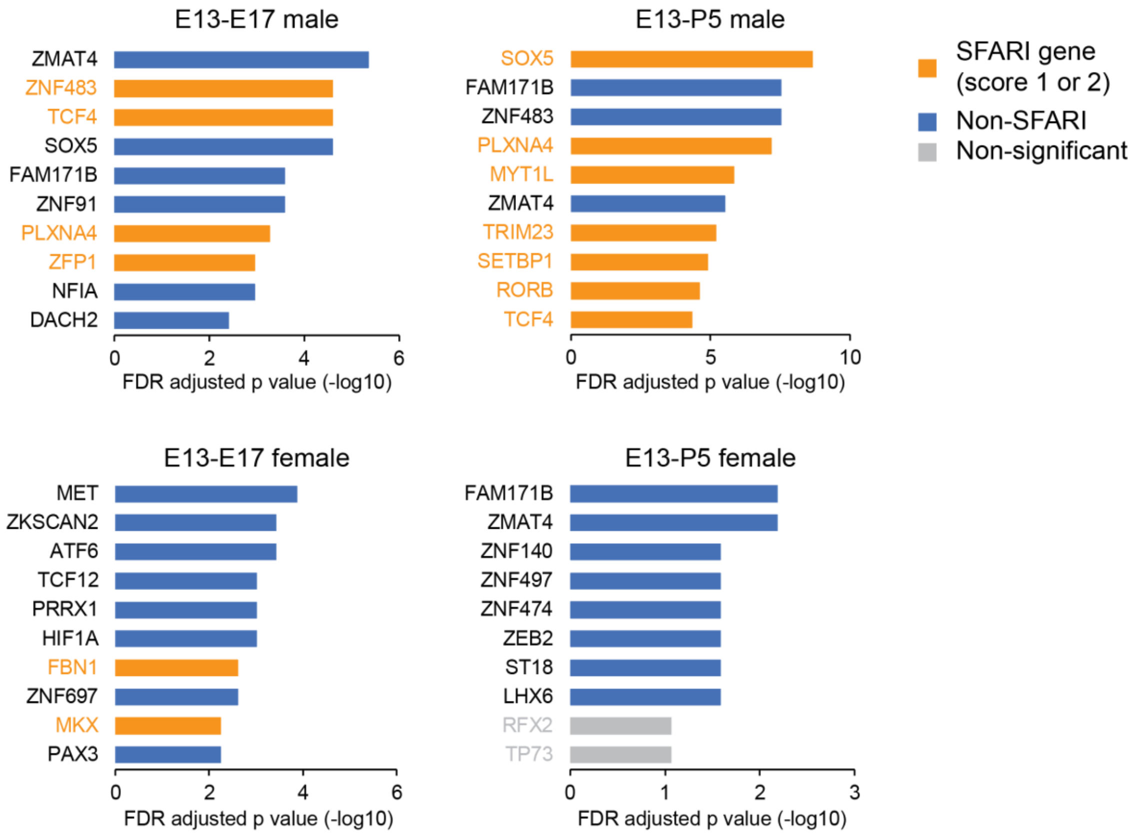


Figure 32: Male-biased TEs are associated with high-confidence autism genes

Putative genes that are regulated by the sex-biased TEs identified by Calling Cards. Genes that are highlighted in orange are those found in the SFARI gene list with a score of 1 or 2. Grey bars represent non-significant genes after FDR adjustment.

4.5 Discussion

In this study, I investigated the impact of sex differences on brain development, focusing on Brd4-bound enhancers and their role in orchestrating sex-specific gene expression. Through the application of Calling Cards technology, I profiled enhancer usage across critical neurodevelopmental phases, uncovering distinct patterns that might underlie sex-differential brain maturation and function. Notably, the analysis across three developmental windows—before, during, and after the perinatal testosterone surge—revealed significant variations in enhancer activity, hinting at both the organizational influence of early hormone exposure and the enduring nature of these epigenetic marks.

This study extended beyond traditional enhancer analysis by examining the role of transposable elements (TEs) as dynamic regulatory elements that may influence sex biases of gene expression. This approach revealed differences in enhancer activities and associated transcription factor motifs, providing new insights into the genomic mechanisms underlying sex differences during brain development. The significant overlap of TE-regulated genes with those implicated in autism spectrum disorders suggests a possible link between sex-biased gene regulation and the higher incidence of autism in males. Based on these findings, we can hypothesize a model whereby TEs containing sex-responsive transcription factor motifs could modulate gene expression in a sex-specific manner. Another level of regulation can be that if these normally silenced TEs might become epigenetically activated through hormone signaling or other sex-dependent mechanism, leading to the adoption of these TE promoters for gene expression, a process known as promoter exaptation (Jang et al., 2019; Shah et al., 2023). A gene can come under the influence of sex if a TE containing a sex-responsive transcription factor motif is in the promoter of that gene. Another level of regulation can be if the normally silenced TE becomes epigenetically activated through a hormonal signaling or sex-dependent mechanism, leading to promoter exaptation and gene expression. Further studies into the interaction between TEs and sex-specific gene regulation could provide a deeper understanding of the genetic basis of sex differences and their role in predisposing individuals to sex-biased diseases.

Overall, this research not only corroborates the pivotal role of hormonal and chromosomal factors in brain sexual differentiation but also expands our understanding of the genetic and epigenetic mechanisms that contribute to sex-specific neural trajectories. By establishing a comprehensive catalog of enhancer usage and identifying key regulatory elements implicated in sex-biased gene expression, this work provides valuable resources for future studies aiming to unravel the complexities of brain development and the etiology of sex-differentiated behaviors and neuropsychiatric disorders.

4.6 Materials and methods

Animals and tissue collection

All animal studies were approved by and performed in accordance with the guidelines of the Animal Care and Use Committee of Washington University in Saint Louis, School of Medicine and conform to NIH guidelines of the care and use of laboratory animals. The animals were housed in controlled environments with a 12-hour light-dark cycle, constant temperature and relative humidity, and *ad libitum* access to food and water. Timed pregnant CD-1 IGS mice (strain code 022) were ordered from Charles River.

Genotyping

Tissue (tail biopsy or toe clipping) was obtained from each animal and placed in a PCR tube. 100 µl lysis buffer (25mM NaOH, 0.2mM EDTA, pH 12) was added to each tube and incubated at 99°C for 60 min in a thermocycler. Once the samples cooled to room temperature, 100ul 40mM Tris-HCl pH 5 was added to neutralize the alkaline lysis buffer. The crude lysate containing genomic DNA (gDNA) was stored at 4°C. For each sample, a multiplexed reaction was performed to genotype the SRY allele to determine sex (see **Table 16** for sequences). For SRY, 1 µl crude gDNA was added to a master mix containing 5 µl OneTaq Quick-Load 2X Master Mix (New England Biolabs M0271), 1 µl 10µM SRY For/Rev primer mix, 1 µl 10µM β-actin For/Rev primer mix, and 2 µl ddH2O. Thermocycling conditions were as follows: 94°C for 3 min; 35 cycles of: 94°C for 10 sec, 60°C for 20 sec, 68°C for 20 sec; 68°C for 5 min; and 4°C hold. Multiplexing β-actin not only confirms the presence of gDNA, but also minimizes non-specific amplification of the MYT1L mutant band in WT samples. PCR products were run on a 1% agarose gel and visualized with GelRed (Biotium 41003).

Generation of AAV9 viral particles

Endotoxin-free plasmid preparations of pAAV-hyPB and pool of barcoded pAAV-TdTomato-SRT_bc (Addgene Kit #11000000213) were done using the ZymoPURE II Plasmid Maxiprep kit (Zymo D4202). These vectors were packaged into AAV9 viral particles by triple transfection into HEK293T cells by the Hope Center Viral Vectors Core at Washington University School of Medicine. The AAV9 particles were purified by iodixanol gradient ultracentrifugation, and the titer was determined by qPCR. Endotoxin levels were assessed using the Endosafe nextgen-PTS (Charles River) Assay.

In-utero AAV injections

Timed pregnant CD-1 IGS mice were acquired from Charles River and were designated for in-utero AAV injections at E13. Prior to the surgical procedure, equal volumes of AAV9-hyPB and AAV9-TdTomato-SRT_bc were mixed and kept on ice until needed for the injections. The surgical area was prepared with sterile surgical drapes and the pre-sterilized tools were laid out, taking care to maintain sterile conditions. The pregnant dam was anesthetized using isoflurane (Covetrus 11695-6777-2) in an induction chamber, followed by a subcutaneous injection of 0.1 mg/kg buprenorphine SR into the interscapular area for post-operative analgesia. Ophthalmic gel (Pivetal 46066-753-55) was applied to protect the eyes before the head was positioned in a nose cone connected to a vaporizer (Midmark Matrx VIP 3000) delivering 2% isoflurane with oxygen for anesthesia maintenance during the procedure. The surgical site on the abdomen was cleared of hair and sanitized with three applications of 80% ethanol and betadine surgical scrub (Avrio Health 304970-0A). A sterile drape (Dynarex 4410) with an opening over the abdomen was positioned over the mouse. During the laparotomy, a midline incision through the skin and muscle layers exposed the embryos. The uterine horns were carefully removed from the abdominal cavity and placed on top of the surgical drape. 1 µl AAV cocktail was injected into the ventricles of each embryo, except for the two medial embryos in each uterine horn. After all injections were completed, the uterine horns were put back into the abdomen. The muscle layer was then sutured with a 5-0 silk suture (Surgical Specialties Corp 774B) in a running stitch with a lock knot every 3 passes. Isoflurane was reduced to 1% to speed recovery while the skin incision was closed with 5-0 nylon sutures (Ethicon 668G) with interrupted stitches. Post-operation, the mouse was moved to a clean cage partially on a heating pad to recover, under observation for immediate postoperative responses and discomfort. A cardboard tube was added to the cage to provide environmental enrichment and monitored twice a day for the next 48 hours. These checks included wound and suture inspection and for any signs of discomfort. If necessary, the outer nylon sutures were removed after 10 days.

Intracerebroventricular injections

Injections were performed as described in the Intracerebroventricular Injection section within Basic Protocol 1 found in (Yen et al., 2023). Briefly, the pups were anesthetized on ice and a total of 6 µl (3 µl per hemisphere, 1 µl per site) was injected into the ventricles of P2 pups using a 50 µl Hamilton syringe. After the injections, the pups were kept warm on a heating pad until they were returned to their home cage.

Tissue collection

To harvest the E17 embryos, the pregnant dam was euthanized using carbon dioxide. The embryos were rapidly dissected from the uterine horns of the mouse in ice-cold HBSS. The embryos were removed from the amniotic pouches, decapitated, and the brains were dissected from the skulls. The cortices were harvested and flash frozen in liquid nitrogen, and stored at -80°C. Tail tissue was collected from each embryo for gDNA isolation and SRY genotyping. The cortex tissues from the P5 and P21 mice were similarly dissected, flash frozen in liquid nitrogen, and stored at -80°C. Toe tissue was collected from each animal for gDNA isolation and SRY genotyping to confirm sexes.

Bulk Calling Cards library preparation

The frozen tissue was homogenized in Trizol (ThermoFisher 15596026) using a handheld homogenizer (SP Bel-Art F65100-0000) with plastic pestles (Fisher Scientific 12-141-364) in 1.5ml centrifuge tubes. Total RNA was purified using the Zymo RNA Clean & Concentrator-25 kit (Zymo R1017). A detailed step-by-step protocol is provided in Basic Protocol 2 in (Yen et al., 2023). The RNA integrity and concentration of the purified RNA were then quantified using RNA Screentape (Agilent 5067-5579). A library density quantitative PCR assay for TdTomato was performed to identify samples that had low expression of SRTs and were unlikely to make high quality libraries. Details on this protocol can be found in Support Protocol 2 in (Yen et al., 2023). The sequencing libraries were prepared according to Basic Protocol 3, pooled and sequenced according to Basic Protocol 4 in (Yen et al., 2023) to a target depth of approximately 10 million read pairs per sample.

Sequencing

Pooled dual indexed libraries were submitted to the Genome Technology Access Center at the McDonnell Genome Institute (GTAC@MGI) for sequencing. For their workflow, the concentration of each library was determined using the KAPA Library Quantification Kit according to the manufacturer's protocol. Target sequencing depth was determined prior to pooling and samples were pooled in ratios based on the targeted depth and concentrations to produce cluster counts appropriate for the Illumina NovaSeq 6000 instrument. Normalized libraries were sequenced on a NovaSeq 6000 S4 Flow Cell using the XP workflow and a 151x10x10x151 sequencing recipe according to the manufacturer's protocol. Base calls were converted to fastq format and demultiplexed using the onboard DRAGEN software to run BCL Convert.

Bulk Calling Cards analysis

The raw FASTQ files were processed using the nf-core/callingcards pipeline (Yen et al., 2023). The resulting qbed files were filtered to keep only insertions with more than 2 reads. The CCcaller peak caller from Pycallingcards (Guo et al., 2024) was used to call background-free peaks from the bulk Calling Cards data using the following parameters: *maxbetween*: 2000, *maxbetween.pvaluecutoff*: 0.01, and *pseudocounts*: 0.1. The insertions across different conditions were combined to generate a joint set of peaks. The data was then split to explore sample and condition-specific enrichments within these joint peaks, which was then used for differential peak analysis using Fisher's exact test. For further analysis of the peak regions, *annotatePeaks.pl* for peak annotation and *findMotifsGenome.pl* for motif enrichment analysis were used from the HOMER suite of tools (Heinz et al., 2010). RepeatMasker was used to annotate interspersed repeats and low complexity DNA sequences in the mm10 genome (Tarailo-Graovac and Chen, 2009). Common genome arithmetic operations such as merging, intersecting, and counting genome regions were performed using the Bedtools utilities (Quinlan and Hall, 2010).

4.7 Data and code availability

All data and code used in this study are available upon request.

4.8 Acknowledgements

I thank members of the Dougherty and Mitra laboratories for helpful discussions and feedback; A. Oldenborg for teaching me in utero injections; M. Crosby and J. Hoisington-Lopez at the DNA Sequencing and Innovation Lab (DSIL) and the Genome Technology Access Center at the McDonnell Genome Institute (GTAC@MGI) for their sequencing expertise and services. This work was funded by grants from the National Institute of Mental Health (RF1MH117070, RF1MH126723, and R01MH124808 to J.D.D. and R.D.M.). A.Y. was supported in part by T32HG000045 from the National Human Genome Research Institute.

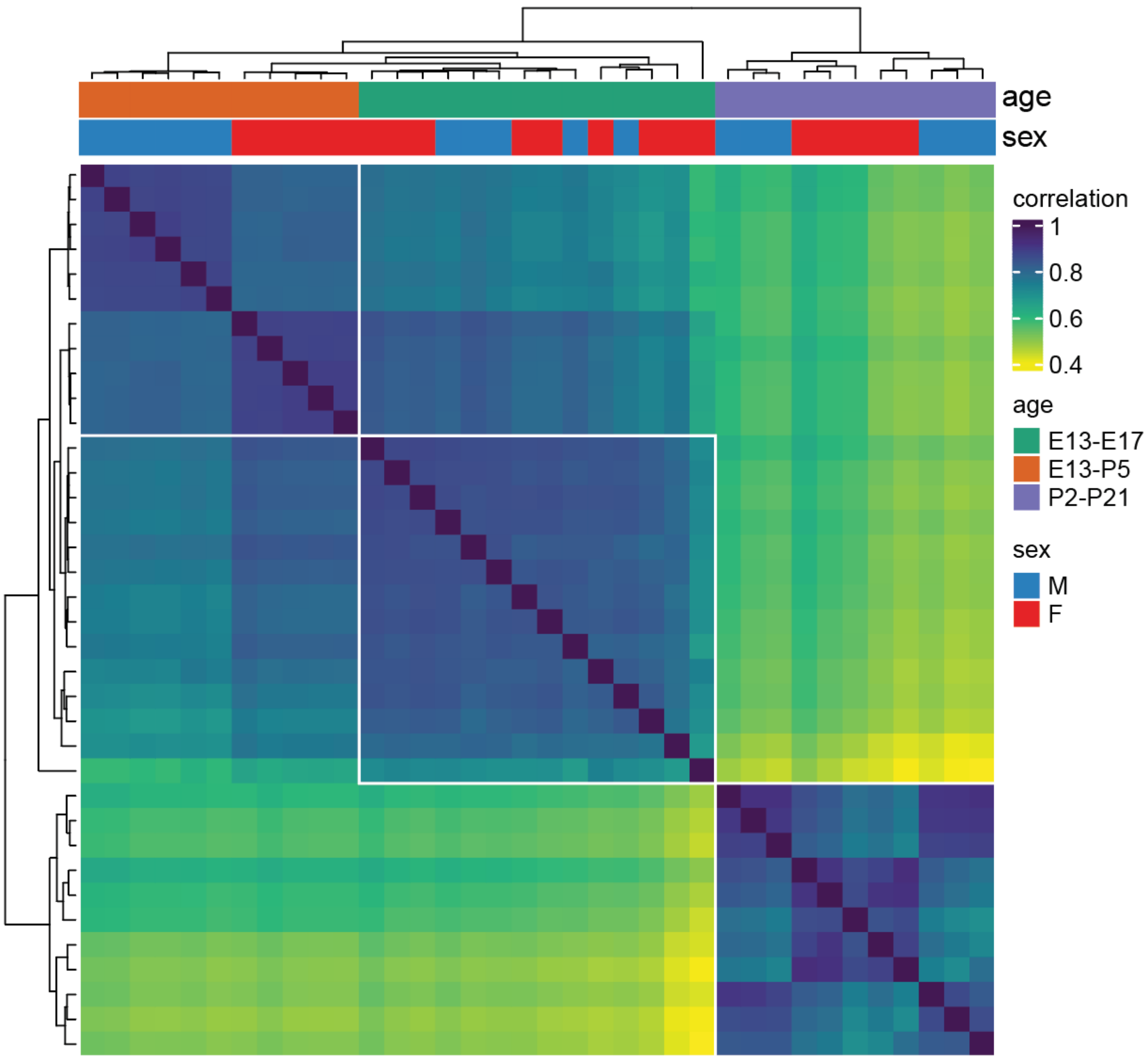
4.9 Author contributions

Project conceptualization: A.Y., R.D.M., and J.D.D. Method development, experiments, and data collection: A.Y. Formal analysis: A.Y., R.D.M., and J.D.D. Figures and data visualization: A.Y., R.D.M., and J.D.D. Project coordination: A.Y., R.D.M., and J.D.D. Funding acquisition: R.D.M. and J.D.D.

4.10 Disclosures

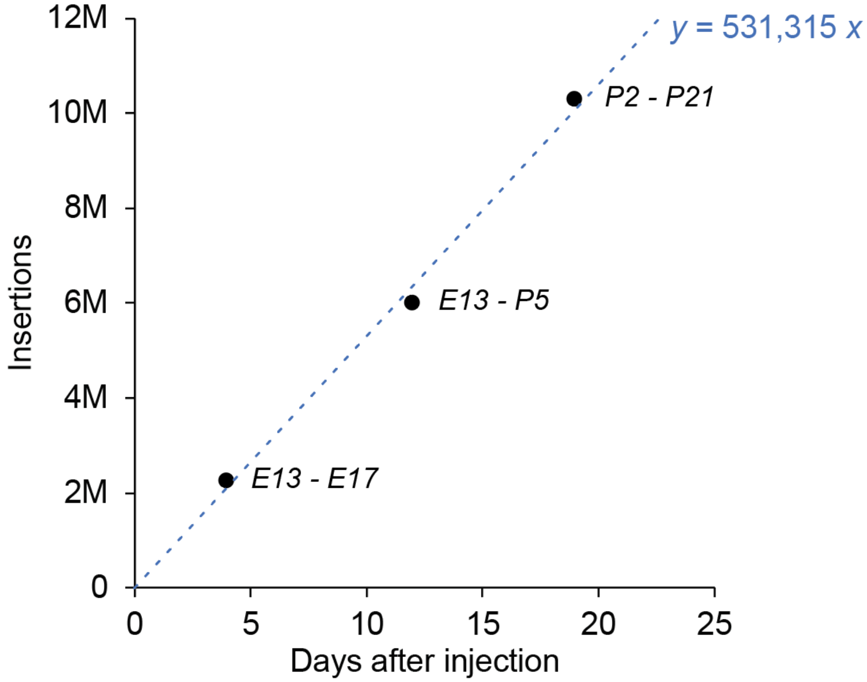
R.D.M has filed a patent application for self-reporting transposon (SRT) technology.

4.11 Supplementary figures and tables



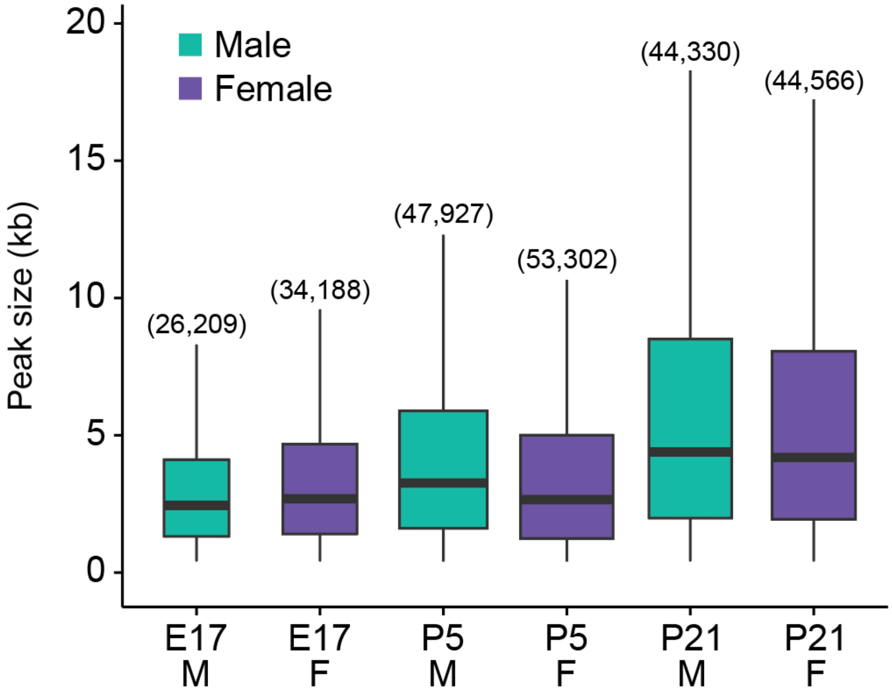
Supplemental Figure 8: Sample correlations based on insertions per called peak

Heatmap showing the computed Pearson correlation coefficients based on the number of insertions per peak. This shows that samples within each age group have high correlation. The E13-E17 and E13-P5 groups correlate more closely than with the postnatal P2-P21 group. To aid visualization, the age groups are clustered and outlined with white squares.



Supplemental Figure 9: Calling Cards insertion rate is constant over extended recording times

The number of days after injection of AAV-Calling Cards into the mouse brain is plotted against recovered insertions. The nearly linear relationship demonstrates that the rate of insertions is constant even after 3 weeks. A linear regression estimates an accumulation of 531,315 insertions per day.



Supplemental Figure 10: Sizes of Brd4 enhancer regions

Boxplots showing the distribution of sizes of Calling Cards peaks representing putative Brd4 enhancers sizes. The total number of peaks per group is noted in parentheses above.

Table 14: Summary of individual samples for the sex differences Calling Cards experiments

Sample	Insertions	Reads	Mean Coverage
F_E17_1-1	25,003	3,189,465	127.6
F_E17_1-2	203,799	10,937,388	53.7
F_E17_1-3	314,779	10,876,214	34.6
F_E17_1-4	242,264	13,383,793	55.2
F_E17_2-2	110,548	8,725,407	78.9
F_E17_3-2	183,221	10,010,624	54.6
F_E17_3-4	96,864	9,630,549	99.4
F_E17_3-8	126,099	8,649,483	68.6
F_E17_3-10	73,606	12,407,581	168.6
M_E17_2-5	147,446	7,850,563	53.2
M_E17_3-3	142,304	9,312,161	65.4
M_E17_3-5	347,726	35,067,128	100.8
M_E17_3-7	147,932	11,309,794	76.5
M_E17_3-9	95,033	10,143,344	106.7
F_P5_2-3	372,659	2,690,961	7.2
F_P5_3-1	254,539	8,877,578	34.9
F_P5_3-2	754,024	52,688,736	69.9
F_P5_3-3	759,203	36,670,573	48.3
F_P5_3-4	770,127	66,395,391	86.2
M_P5_1-4	310,289	9,249,870	29.8
M_P5_1-7	288,817	5,765,502	20.0
M_P5_2-7	354,213	7,630,424	21.5
M_P5_3-1	749,684	49,685,946	66.3
M_P5_3-4	719,868	55,263,404	76.8
M_P5_3-5	668,198	71,325,063	106.7
F_P21_1-1	1,097,133	28,672,025	26.1
F_P21_1-3	756,326	32,598,204	43.1
F_P21_1-4	1,003,657	38,751,295	38.6
F_P21_1-6	718,122	24,655,095	34.3
F_P21_1-10	1,032,564	34,506,985	33.4
M_P21_1-11	763,515	19,772,311	25.9
M_P21_1-12	922,776	17,492,384	19.0
M_P21_1-2	1,126,401	35,402,373	31.4
M_P21_1-5	1,031,665	39,096,400	37.9
M_P21_1-7	877,388	34,138,688	38.9
M_P21_1-9	974,106	35,780,469	36.7

Table 15: Summary of groups for the sex differences Calling Cards experiments

Recording period	Sex	Replicates	Total Insertions	Total Reads	Mean Coverage
E13-E17	Female	9	1,376,183	87,810,504	63.8
E13-E17	Male	5	880,441	73,682,990	83.7
E13-P5	Female	5	2,910,552	167,323,239	57.5
E13-P5	Male	6	3,091,069	198,920,209	64.4
P2-P21	Female	5	4,607,802	159,183,604	34.5
P2-P21	Male	6	5,695,851	181,682,625	31.9

Table 16: Primer sequences for Chapter 4

Primer Name	Sequence (5’ → 3’)	Product
SRY_For	TTGTCTAGAGAGCATGGAGGGCCATGTCAA	273 bp
SRY_Rev	CCACTCCTCTGTGACACTTTAGCCCTCCGA	
β-actin_For	AGAGGGAAATCGTGCGTGAC	150 bp
β-actin_Rev	CAATAGTGATGACCTGGCCGT	
β-actin_qPCR_F	AGAGGGAAATCGTGCGTGAC	n/a
β-actin_qPCR_R	CAATAGTGATGACCTGGCCGT	n/a
TdTomato_qPCR_F	CAAGCTGAAGGTGACCAAGG	n/a
TdTomato_qPCR_R	CCGTCCTCGAAGTTCATCAC	n/a
SMART_dT18VN	AAGCAGTGGTATCAACGCAGAGTACGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN	n/a
SMART_Rev	AAGCAGTGGTATCAACGCAGAGT	n/a
SRT_tdTomato_F1	TCCTGTACGGCATGGACGAG	n/a
OMPB_BC-GAT_Index2_X	AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATCGTCAATTTTACGCAGACTATCTTT	n/a
N7_Index_X	CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG	n/a

[a] The three bases highlighted in blue indicate the location of the OM-PB primer barcode. For more information, refer to “ Considerations for primer selection and ordering for sequencing libraries” in Chapter 2.3.

[b] The [i5] and [i7] represent placeholders where unique indexes can be added for multiplexing multiple samples to be sequenced in a single run. For more information about standard Illumina indexes, see https:// supportdocs.illumina.com/SHARE/AdapterSeq/Content/SHARE/AdapterSeq/AdapterSequencesIntro.htm, or consult your sequencing core or service provider.

Table 17: Motifs enriched in male-biased transposable elements with Calling Cards peaks after testosterone surge (E13-P5)

Motif	Class	Consensus	P-value	Log P-value	q-value (Benjamini)
Otx2	Homeobox	NYTAATCCYB	1.00E-174	-401.90	0.0000
Max	bHLH	RCCACGTGGYYN	1.00E-159	-368.10	0.0000
DUX4	Homeobox	NWTAAYCYAATCAWN	1.00E-140	-324.50	0.0000
Duxbl	Homeobox	TAAYCYAATCAA	1.00E-136	-314.50	0.0000
ZNF341	Zf	GGAACAGCCG	1.00E-83	-192.10	0.0000
GSC	Homeobox	RGGATTAR	1.00E-65	-151.40	0.0000
MNT	bHLH	DGCACACGTG	1.00E-60	-140.00	0.0000
SF1	NR	CAAGGHCANV	1.00E-60	-139.10	0.0000
ERE	NR	VAGGTCACNSTGACC	1.00E-57	-133.40	0.0000
Rfx5	HTH	SCCTAGCAACAG	1.00E-57	-132.40	0.0000
GATA3	Zf	AGATSTNDNNDSAGATAASN	1.00E-55	-128.70	0.0000
PRDM10	Zf	TGGTACATTCCA	1.00E-55	-127.30	0.0000
ZNF189	Zf	TGGAACAGMA	1.00E-50	-116.60	0.0000
GRHL2	CP2	AAACYKGTTWDACMRGTTTB	1.00E-50	-116.20	0.0000
BORIS	Zf	CNNBRGCGCCCCCTGSTGGC	1.00E-49	-114.40	0.0000
c-Myc	bHLH	VVCCACGTGG	1.00E-48	-111.20	0.0000
TEAD2	TEA	CCWGGAATGY	1.00E-45	-104.90	0.0000
NFY	CCAAT	RGCCAATSRG	1.00E-38	-88.20	0.0000
Sox7	HMG	VVRRAACAATGG	1.00E-36	-85.05	0.0000
Atf7	bZIP	NGRTGACGTCAY	1.00E-34	-78.40	0.0000
Klf9	Zf	GCCACRCCCACY	1.00E-32	-75.53	0.0000
Egr2	Zf	NGCGTGGGCGGR	1.00E-31	-72.17	0.0000
PU.1	ETS	AGAGGAAGTG	1.00E-30	-71.30	0.0000
n-Myc	bHLH	VRCCACGTGG	1.00E-30	-70.86	0.0000
NF1	CTF	CYTGGCABNSTGCCAR	1.00E-29	-68.24	0.0000
RARg	NR	AGGTCAAGGTCA	1.00E-28	-66.50	0.0000
RUNX	Runt	SAAACCACAG	1.00E-27	-63.10	0.0000
ARE	NR	RGRACASNSTGTYCYB	1.00E-27	-62.90	0.0000
CLOCK	bHLH	GHCACGTG	1.00E-27	-62.63	0.0000
Gli2	Zf	YSTGGGTGGTCT	1.00E-25	-58.82	0.0000
Tbx20	T-box	GGTGYTGACAGS	1.00E-25	-57.82	0.0000
RORgt	NR	AAYTAGGTCA	1.00E-23	-54.04	0.0000
KLF5	Zf	DGGGYGKGGC	1.00E-21	-49.47	0.0000
NRF1	NRF	CTGCGCATGCGC	1.00E-21	-48.43	0.0000
Hand2	bHLH	TGACANARRCCAGRC	1.00E-20	-46.68	0.0000
HRE	HSF	BSTTCTRGAABVTTCYAGAA	1.00E-13	-31.99	0.0000
VDR	NR	ARAGGTCANWGAGTTCANNN	1.00E-13	-31.65	0.0000
YY1	Zf	CAAGATGGCGGC	1.00E-13	-31.26	0.0000

Table 17: Motifs enriched in male-biased transposable elements with Calling Cards peaks after testosterone surge (E3-P5), *continued*

TEAD4	TEA	CCWGGAATGY	1.00E-12	-29.58	0.0000
RUNX2	Runt	NWAACCACADNN	1.00E-11	-25.37	0.0000
NPAS	bHLH	NVCACGTG	1.00E-10	-23.74	0.0000
ZNF264	Zf	RGGGCACTAACY	1.00E-10	-23.04	0.0000
Sox17	HMG	CCATTGTTYB	1.00E-09	-23.02	0.0000
Six1	Homeobox	GKVTCADRTTWC	1.00E-09	-22.08	0.0000
CTCF	Zf	AYAGTGCCMYCTRGTGGCCA	1.00E-09	-21.62	0.0000
ZNF382	Zf	GNCTGTASTRNTGBCTCHTT	1.00E-08	-19.62	0.0000
HOXA2	Homeobox	GYCATCMATCAT	1.00E-07	-18.08	0.0000
GATA	Zf	NNNNNBAGATAWYATCTVHN	1.00E-07	-17.00	0.0000
EKLF	Zf	NWGGGTGTGGCY	1.00E-07	-16.50	0.0000
ISRE	IRF	AGTTTCASTTTC	1.00E-06	-16.11	0.0000
IRF2	IRF	GAAASYGAAASY	1.00E-06	-15.86	0.0000
NRF	NRF	STGCGCATGCGC	1.00E-06	-14.62	0.0000
Rfx6	HTH	TGTTKCCTAGCAACM	1.00E-05	-13.51	0.0000
PAX6	Homeobox	NGTGTTCAVTSAAGCGKAAA	1.00E-05	-13.41	0.0000
BMAL1	bHLH	GNCACGTG	1.00E-05	-13.32	0.0000
ZNF669	Zf	GARTGGTCATCGCCC	1.00E-05	-12.77	0.0000
T1ISRE	IRF	ACTTTCGTTTCT	1.00E-05	-12.63	0.0000
Usf2	bHLH	GTCACGTGGT	1.00E-05	-12.36	0.0000
NFkB-p65	RHD	WGGGGATTTCCC	1.00E-05	-12.36	0.0000
NFE2L2	bZIP	AWWWTGCTGAGTCAT	1.00E-05	-12.11	0.0000
ZBTB33	Zf	GGVTCTCGCGAGAAC	1.00E-04	-10.34	0.0002
USF1	bHLH	SGTCACGTGR	1.00E-04	-10.04	0.0003
IRF1	IRF	GAAAGTGAAAGT	1.00E-04	-9.73	0.0004
MafB	bZIP	WNTGCTGASTCAGCANWTTY	1.00E-04	-9.42	0.0005
MITF	bHLH	RTCATGTGAC	1.00E-03	-8.79	0.0009
Pax8	Homeobox	GTCATGCHTGRCTGS	1.00E-03	-8.54	0.0011
RUNX1	Runt	AAACCACARM	1.00E-03	-8.25	0.0015
STAT5	Stat	RTTTCTNAGAAA	1.00E-02	-6.69	0.0071
IRF8	IRF	GRAASTGAAAST	1.00E-02	-6.66	0.0072
GATA3	Zf	AGATGKDGAGATAAG	1.00E-02	-6.23	0.0110
Atf2	bZIP	NRRTGACGTCAT	1.00E-02	-5.02	0.0358
SpiB	ETS	AAAGRGGAAGTG	1.00E-02	-4.84	0.0423

Table 18: Motifs enriched in female-biased transposable elements with Calling Cards peaks (E13-P5)

Motif	Class	Consensus	P-value	Log P-value	q-value (Benjamini)
T1ISRE	IRF	ACTTTCGTTTCT	1.00E-946	-2179	0.0000
ZNF41	Zf	CCTCATGGTGYCYTWYTCCC TTGTG	1.00E-906	-2086	0.0000
ZNF382	Zf	GNCTGTASTRNTGBCTCHTT	1.00E-879	-2024	0.0000
DUX4	Homeobox	NWTAAYCYAATCAWN	1.00E-804	-1853	0.0000
ISRE	IRF	AGTTTCASTTTC	1.00E-589	-1358	0.0000
PAX6	Homeobox	NGTGTTCAVTSAAGCGKAAA	1.00E-508	-1171	0.0000
IRF2	IRF	GAAASYGAAASY	1.00E-341	-785.7	0.0000
Duxbl	Homeobox	TAAYCYAATCAA	1.00E-300	-692.4	0.0000
IRF1	IRF	GAAAGTGAAAGT	1.00E-214	-493.4	0.0000
Gli2	Zf	YSTGGGTGGTCT	1.00E-195	-450.5	0.0000
GATA3	Zf	AGATGKDGAGATAAG	1.00E-189	-435.9	0.0000
HRE	HSF	BSTTCTRGAABVTTCYAGAA	1.00E-172	-397.2	0.0000
GRHL2	CP2	AAACYKGTTWDACMRGTTTB	1.00E-158	-364	0.0000
Egr2	Zf	NGCGTGGGCGGR	1.00E-93	-215.2	0.0000
Oct2	Homeobox	ATATGCAAAT	1.00E-86	-199.1	0.0000
Six1	Homeobox	GKVTCADRTTWC	1.00E-85	-195.9	0.0000
NF1	CTF	CYTGGCABNSTGCCAR	1.00E-79	-182.7	0.0000
Sox7	HMG	VVRRAACAATGG	1.00E-79	-182	0.0000
SpiB	ETS	AAAGRGGAAGTG	1.00E-66	-153.6	0.0000
VDR	NR	ARAGGTCANWGAGTTCANNN	1.00E-65	-150.3	0.0000
ERE	NR	VAGGTCACNSTGACC	1.00E-63	-145.4	0.0000
ZNF136	Zf	YTKGATAHAGTATTCTWGGTN GGCA	1.00E-51	-119.1	0.0000
CLOCK	bHLH	GHCACGTG	1.00E-44	-102.5	0.0000
Hand2	bHLH	TGACANARRCCAGRC	1.00E-38	-89.77	0.0000
Max	bHLH	RCCACGTGGYYN	1.00E-38	-89.23	0.0000
TEAD2	TEA	CCWGGAATGY	1.00E-26	-60.21	0.0000
Oct11	Homeobox	GATTTGCATA	1.00E-23	-53.13	0.0000
STAT1	Stat	NATTTCCNGGAAAT	1.00E-18	-41.66	0.0000
Tcf3	HMG	ASWTCAAAGG	1.00E-15	-35.33	0.0000
Srebp2	bHLH	CGGTCACSCCAC	1.00E-12	-28.5	0.0000
ZNF669	Zf	GARTGGTCATCGCCC	1.00E-10	-25.31	0.0000
CArG	MADS	CCATATATGGNM	1.00E-09	-20.81	0.0000
GATA3	Zf	AGATSTNDNNDSAGATAASN	1.00E-08	-19.83	0.0000
E2F	E2F	TTSGCGCGAAAA	1.00E-08	-19.47	0.0000
USF1	bHLH	SGTCACGTGR	1.00E-07	-17.47	0.0000
Brn1	Homeobox	TATGCWAATBAV	1.00E-07	-16.83	0.0000

Table 18: Motifs enriched in female-biased transposable elements with Calling Cards peaks (E13-P5), ***continued***

Usf2	bHLH	GTCACGTGGT	1.00E-06	-14.84	0.0000
HINFP	Zf	TWVGGTCCGC	1.00E-06	-14.04	0.0000
bHLHE40	bHLH	KCACGTGMCN	1.00E-05	-13.71	0.0000
YY1	Zf	CAAGATGGCGGC	1.00E-05	-12.92	0.0000
IRF8	IRF	GRAASTGAAAST	1.00E-04	-11.26	0.0001
FXR	NR	AGGTCANTGACCTB	1.00E-04	-10.46	0.0003

Chapter 5: MYT1L deficiency impairs excitatory neuron trajectory during cortical development

5.1 Preface

This chapter contains contents from a manuscript under review:

MYT1L deficiency impairs excitatory neuron trajectory during cortical development

Allen Yen, Xuhua Chen, Dominic D. Skinner, Fatjon Leti, MariaLynn Crosby, Jessica Hoisington-Lopez, Yizhe Wu, Jiayang Chen, Robi D. Mitra, Joseph D. Dougherty

bioRxiv, March 7, 2024. doi: https://doi.org/10.1101/2024.03.06.583632.

5.2 Abstract

Mutations that reduce the function of MYT1L, a neuron-specific transcription factor, are associated with a syndromic neurodevelopmental disorder. Furthermore, MYT1L is routinely used as a proneural factor in fibroblast-to-neuron transdifferentiation. MYT1L has been hypothesized to play a role in the trajectory of neuronal specification and subtype specific maturation, but this hypothesis has not been directly tested, nor is it clear which neuron types are most impacted by MYT1L loss. In this study, we profiled 313,335 nuclei from the forebrains of wild-type and MYT1L-deficient mice at two developmental stages: E14 at the peak of neurogenesis and P21, when neurogenesis is complete, to examine the role of MYT1L levels in the trajectory of neuronal development. We found that MYT1L deficiency significantly disrupted the relative proportion of cortical excitatory neurons at E14 and P21. Significant changes in gene expression were largely concentrated in excitatory neurons, suggesting that transcriptional effects of MYT1L deficiency are largely due to disruption of neuronal maturation programs. Most effects on gene expression were cell autonomous and persistent through development. In addition, while MYT1L can both activate and repress gene expression, the repressive effects were most sensitive to haploinsufficiency, and thus more likely mediate MYT1L syndrome. These findings illuminate the intricate role of MYT1L in orchestrating gene expression dynamics during neuronal development, providing insights into the molecular underpinnings of MYT1L syndrome.

5.3 Introduction

Every brain cell shares the same genetic code, yet they exhibit a wide range of functions. This diversity arises because different cell lineages enact different gene expression programs that direct each cell in the embryonic brain to develop in a highly orchestrated manner. Disruption of these processes can lead to abnormal neurodevelopment and result in impaired cognition, communication, and adaptive behavior, as seen in profound autism and intellectual disability (ID) (Lord et al., 2018; Willsey et al., 2022). Notably, many genes associated with such neurodevelopmental disorders (NDDs) are expressed early during brain development and are involved in gene regulation and synaptic function (Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium et al., 2019; Fu et al., 2022). Studies using post-mortem human brain tissue provide evidence that cortical excitatory neurons are commonly dysregulated in autism (Gandal et al., 2022; Velmeshev et al., 2023). However, since these are end of life studies, whether this a cause or consequence of autism is unclear.

One such NDD associated gene is Myelin Transcription Factor 1 Like (MYT1L), which is highly expressed exclusively in postmitotic neurons in the embryonic brain and sustained at lower levels throughout life (Kepa et al., 2017; Matsushita et al., 2014). Early fibroblast-to-neuron transdifferentiation studies demonstrate that MYT1L promotes neuronal cell fate by repressing non-neuronal lineage programs (Mall et al., 2017; Vierbuchen et al., 2010). Similarly, *in vivo* epigenetic studies of normal development show that MYT1L promotes neuronal differentiation by recruiting the SIN3B repressive complex to promoters and enhancers of postmitotic neurons to suppress early developmental programs (Chen et al., 2023). Indeed, loss of MYT1L in multiple mouse models resulted in upregulation of a fetal gene expression signature (Chen et al., 2021; Kim et al., 2022; Weigel et al., 2023). To date, three pivotal studies have delved into the in vivo functions of MYT1L by creating transgenic mouse models. Each study uniquely disrupted a different exon of MYT1L (6 in (Wöhr et al., 2022), 7 in (Chen et al., 2021), and 9 in (Kim et al., 2022)). The animal models are valuable tools to study the molecular and cellular consequences of MYT1L haploinsufficiency and the mice recapitulate many of the clinical presentations such as hyperactivity, structural malformations, obesity, and behavioral deficits (Chen et al., 2021; Kim et al., 2022; Weigel et al., 2023). However, it remains largely unknown how MYT1L haploinsufficiency influences the trajectory of neuronal differentiation *in vivo*, and whether the development of specific neuronal subtypes is particularly susceptible to the loss of MYT1L. Moreover, it is unclear if there is a critical moment in each cell's developmental window during which MYT1L function is indispensable, as understanding this timeline could delineate when the transcriptional dynamics and developmental processes are amenable to interventions.

Detailed atlases mapping the gene expression profiles of thousands of cell types across the entire mouse brain have significantly advanced our understanding of brain organization under typical conditions (Di Bella et al., 2021; La Manno et al., 2021; Liu et al., 2023; Yao et al., 2023; Zhang et al., 2023; Zu et al., 2023). Building upon this foundational knowledge, we can now explore how genetic perturbations affect neurodevelopment, specifically investigating the impact of disrupting a gene regulatory network through the loss of a single TF on this atlas. Given the widespread expression pattern of MYT1L in neurons, it is unclear if specific neuronal subtypes are more sensitive to MYT1L deficiency. Likewise, previous studies using bulk RNA sequencing have shown that MYT1L deficiency affects genes associated with the cell cycle (Chen et al., 2023, 2021; Weigel et al., 2023), differentiation (Mall et al., 2017; Vierbuchen et al., 2010), and proliferation(Melhuish et al., 2018). However, a limitation of bulk sequencing is that it only provides average gene expression data from a mixed population of cells, making it challenging to discern the precise origin of observed differences. For example, MYT1L haploinsufficiency results in an increased expression of developmental gene expression programs in vitro and in the post-natal brain (Chen et al., 2021; Mall et al., 2017; Weigel et al., 2023), but it remains unclear whether the observed differences are due to an increased proportion of immature progenitors or whether post-mitotic neurons are generated in proper numbers, but fail to mature completely and become trapped in an intermediate state. Furthermore, MYT1L functions as both a transcriptional repressor(Mall et al., 2017; Romm et al., 2005) and activator (Chen et al., 2021; Manukyan et al., 2018), but the variations in its role by cell type or developmental stage, as well as the sensitivity of the activated or repressed gene targets to disruption, are still unclear. Although loss of MYT1L leads to precocious differentiation during development (Chen et al., 2021) and sustained activation of developmental programs in the adult brain (Chen et al., 2023; Mall et al., 2017), the implications for neuronal development trajectory and cell-type specific fate specification remain unknown. Utilizing single cell transcriptomics, we can obtain a high-resolution mapping of dynamic developmental processes and elucidate how the loss of MYT1L contributes to the observed differential gene expression patterns.

In this study, we profiled a total of 313,335 nuclei to investigate the molecular and cellular consequences of MYT1L haploinsufficiency at the peak of neurogenesis (E14) and when neurogenesis is complete (P21). Our findings indicate that MYT1L deficiency primarily impacts excitatory neurons. We further identified that genes regulated by MYT1L, whether activated or repressed, exhibit cell type-specific responses to MYT1L haploinsufficiency. A significant number of dysregulated genes were TFs or epigenetic regulators temporally expressed during specific time windows, highlighting lineage specific gene regulatory networks. In summary, our findings provide insights on how MYT1L haploinsufficiency disrupts embryonic and postnatal neurodevelopment. We have identified key transcriptional networks and defined the vulnerable cell types and developmental stages that potentially contribute to the pathogenesis of MYT1L syndrome.

5.4 Results

5.4.1 Loss of MYT1L disrupts proportions of excitatory and inhibitory neurons

To characterize the role of MYT1L during peak neurogenesis and to understand the acute consequences of MYT1L haploinsufficiency and loss on cell fate specification and maturation, we applied a combinatorial indexing approach (Cao et al., 2017; Martin et al., 2023) to profile and analyze transcription from 216,830 nuclei from the developing forebrain of embryonic day 14 (E14) MYT1L knockout (KO), heterozygous (Het), and wild type (WT) animals (**Figure 33A, B**). We find that all cell types are well represented across all genotypes (median genotype LISI score(Luecken et al., 2022)=2.7) (**Figure 33C**). We identified 26 clusters representing 7 broad neural cell types, which were further classified into three subtypes of radial glial cells (*Hes1* and *Nestin* positive), 3 subtypes of intermediate progenitor cells (*Neurog2* and *Eomes* positive) fated to be excitatory neurons, 3 subtypes of inhibitory intermediate progenitor cells (*Dlx1* and *Nkx2.1* positive), 8 subtypes of excitatory neurons (*Neurod6* and *Tbr1* positive), 9 subtypes of inhibitory neurons (*Gad1* and *Gad2* positive), Cajal-Retzius cells, oligodendrocyte progenitor cells, and microglia (**Figure 33C, D**). We assigned cell cycle scores based on cell cycle phase marker gene expression and confirmed that the progenitors were mostly in G2M or S, while the post-mitotic neurons were in G1/G0 (**Figure 33E**) and expressed MYT1L (**Figure 33F**). The progenitor cells segregated into two distinct populations—the root clusters which gave rise to divergent excitatory and inhibitory neuron developmental trajectories. This profile of cellular diversity indicated that we captured a developmental window encompassing differentiation and maturation processes, enabling us to investigate the molecular and cellular consequences of loss of MYT1L in the developing E14 cortex.

Because MYT1L is highly expressed in virtually all neurons during neurogenesis (**Figure 33F**), we sought to assess the short-term consequences of its deficiency on overall cell type proportions. We found subtle but statistically significant disruptions to the abundance of post-mitotic immature excitatory neurons (Im ExN_3), deep layer excitatory neurons (Im L5-6 ExN_1, Im L5-6 ExN_2, L5-6 ExN_1, L5-6 ExN_2, and Im L6 ExN), immature inhibitory neurons (Im InhN_3), and specific subtypes of inhibitory neurons (somatosensory cortex (SI), Darpp32+ D1-D2, and CEA-BST) (**Figure 33H**). Radial glia and inhibitory intermediate progenitors were mostly unaffected by the loss of MYT1L. Non-cycling immature excitatory neurons (Im ExN_3) in the subventricular zone (SVZ) were the most developmentally immature cells from the excitatory trajectory that were affected, showing an increase in abundance in KOs compared to WT which could be a result of precocious neuronal differentiation.

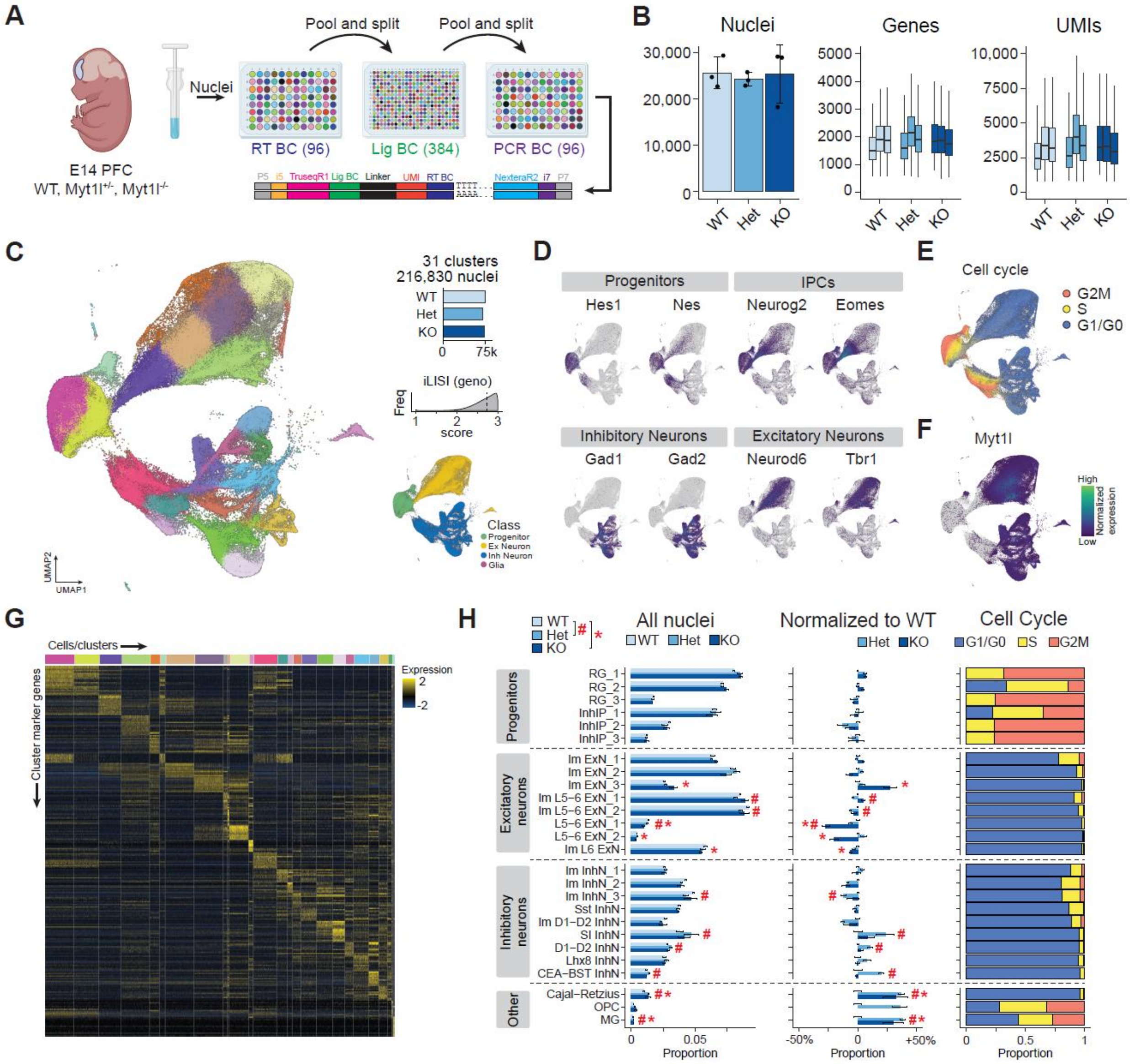


Figure 33: Single nucleus transcriptional profiling of E14 forebrain in MYT1L animals

(A) Schematic showing dissection of forebrain tissue, isolation of nuclei, combinatorial barcoding, and generation of snRNAseq libraries. **(B)** General library statistics showing average nuclei per genotype, median genes per nuclei, and median UMIs per nuclei. **(C)** Uniform manifold approximation and projection (UMAP) of 216,830 nuclei from MYT1L WT, Het, and KO animals colored by cell type. The bar plot shows the total number of nuclei per genotype across biological replicates. The histogram shows the local inverse Simpson's index (LISI) score has a median of 2.7, indicating that the genotypes are well mixed and integrated. The bottom right UMAP inset shows all nuclei color coded by cell class. **(D)** Top markers for progenitors, intermediate progenitor cells (IPCs), inhibitory neurons, and excitatory neurons. **(E)** UMAP of all nuclei color coded by cell cycle score based on cell cycle genes. **(F)** UMAP plot showing expression of MYT1L in postmitotic excitatory and inhibitory neurons. **(G)** Heatmap showing the top marker gene expression (rows) for cells in each cluster (columns). **(H)** Bar plots show the average relative proportions of nuclei in each annotated cell cluster for MYT1L WT, Het, and KO genotypes (left). These proportions are normalized to WT (center). The composition of each cluster by cell cycle phase is shown on the right.

5.4.2 Loss of MYT1L disrupts excitatory neuron development

We then conducted a differential expression analysis to analyze the molecular signatures of each cell type and determine which subtype exhibited the most significant transcriptional changes due to MYT1L deficiency. Individual clusters were aggregated into pseudobulk groups and then we used DESeq2 (Love et al., 2014) to conduct pairwise analyses between WT and KO genotypes to uncover the most pronounced expression differences within each cell type. We identified 1,174 unique differentially expressed genes (DEGs; BH adjusted P-value < 0.1; expression level change ≥ 15%) between WT and KO, of which 781 were upregulated in KO cells and 415 were downregulated compared to WT (**Figure 34A**). There were only 11 genes that were not exclusively up or downregulated across all cell types. DEGs that were found to be unique to a single cluster accounted for 54% (633/1174) of the DEGs, demonstrating disruption of both ubiquitously expressed and cell type-specific genes. Notably, deep layer excitatory neurons, especially immature L6 neurons, harbored the majority of DEGs (**Figure 34A**), indicating that these neurons may be particularly sensitive to loss of MYT1L, which is consistent with their disrupted cell proportions (**Figure 33H**). Progenitor cells, which do not yet express MYT1L, showed no DEGs, demonstrating that the molecular and cellular consequences of MYT1L deficiency are cell intrinsic to the cells that express MYT1L. This suggests that there are no signals from the differentiating neurons that robustly influence the transcriptional identity of the proliferating progenitor pool at E14.

Given that MYT1L homozygotes do not survive postnatally, and the human disorder is caused by haploinsufficiency, it is of interest to analyze Hets. Consequently, we investigated whether the DEGs were dose-responsive to the number of MYT1L copies, or if there were non-linear effects of MYT1L loss. Furthermore, whether this pattern of regulation was the same for activated and repressed genes may suggest which function is most critical to the disorder. Therefore, we modeled the number of functional alleles as an ordinal factor and classified 522 genes that were upregulated in KOs as MYT1L-repressed genes, while the 451 genes that were upregulated in WTs were considered MYT1L-activated genes. We found that MYT1L-repressed genes were more sensitive to the gene dose of MYT1L than genes activated by MYT1L (**Figure 34B**). In MYT1L-activated genes, expression levels in Hets were similar to WTs suggesting that these target genes are sufficiently activated even with decreased levels of MYT1L. However, the MYT1L-repressed genes exhibited a nearly linear gene-dosage response and in some cases Hets were more similar to KOs (**Figure 34B**). This suggests that MYT1L repressed targets become highly upregulated with the loss of a single MYT1L allele.

Prior bulk RNAseq of the E14 MYT1L Het mouse cortex revealed an immature transcriptional signature when compared to WT(Chen et al., 2021). To evaluate if a particular cell type was driving this effect, we performed gene ontology (GO) enrichment analysis for the DEGs in each cluster. We found that MYT1L-repressed genes (KO>WT) are represented in development, neuron migration, and cell fate commitment pathways and were the top enriched pathways in MYT1L KO neurons (**Figure 34C**), while MYT1L-activated genes (WT>KO) are involved in synapse organization, axonogenesis, and neurotransmitter secretion and transport (**Figure 34D**). Together, this revealed that loss of MYT1L results in an immature developmental transcriptional state. By analyzing the DEGs across cell types, we found that MYT1L-repressed genes had a more functionally diverse response to loss of MYT1L compared to those that are activated by MYT1L. This suggests that the suppression of developmental genes is critical to ensure proper neuronal maturation.

We next asked if our lists of MYT1L-activated and -repressed genes are direct or indirect targets of MYT1L. To test this, we integrated our E14 snRNA-seq dataset with an age and region-matched E14 forebrain MYT1L CUT&RUN dataset that cataloged 560 high-confidence MYT1L binding sites within promoter sequences(Chen et al., 2023). 58 direct MYT1L targets were found to be differentially expressed and most of these showed dose-dependent responses to MYT1L. Comparing KOs to WTs, 47 (36 were dose-dependent) were upregulated and 16 (9 were dose-dependent) were downregulated consistently across cell types, reinforcing that MYT1L functions as a transcriptional repressor at ~80% of consistent targets. Additionally, this demonstrates that the DEGs were largely driven by indirect effects of MYT1L deficiency. Indeed, differentially expressed MYT1L targets were significantly enriched for TFs (81/560; P=2.2x10^{-16}, Fisher's exact test) (**Figure 34E**).

We then performed network analysis on all the DEGs identified through cluster pseudobulk analysis to gain insight into functional interactions and putative upstream regulators(Szklarczyk et al., 2023). We identified modules by clustering the protein-protein interaction network based on functional annotations and found that the modules were significantly enriched with transcription regulators and epigenetic factors. Many of these had higher expression in the L5-6 ExN_1 cluster from MYT1L KO animals (**Figure 34F**). This provides evidence that MYT1L can be a transcriptional regulator that not only influences its direct target genes, but also downstream indirect targets within a gene network.

As studies have shown that some autism risk genes disrupt excitatory and inhibitory neurogenesis, we then extended our analysis to test if our observed transcriptional disruptions converge with genes associated with neurodevelopmental disorders. We intersected the excitatory and inhibitory neuron DEGs (**Figure 34A**) with a list of 932 high-confidence autism-related genes from the SFARI database with a score of 1 or 2 (Abrahams et al., 2013; Sanders et al., 2015; Satterstrom et al., 2020). We observed a significant overlap between DEGs and the SFARI genes, with 178 out of 1174 DEGs (15.2%) being shared ($P=2.1x10^{-12}$, chi-square test with Yates' continuity correction). This overlap comprised 146 genes from the excitatory neuron clusters and 32 genes were from the inhibitory neuron clusters. As autism genes have a bias towards genes highly expressed in neurons (Ouwenga and Dougherty, 2015), we wanted to test if this overlap was merely driven by a neuron bias. Therefore, we randomly sampled 932 highly expressed genes in these clusters a thousand times and examined the overlap with SFARI genes. The median overlap was 14 compared to the 178 seen here, indicating a greater than 10-fold enrichment for SFARI genes among MTY1L DEGs. Deeply examining the 178 genes, many of the autism associated DEGs such as *Ext1* and *Phf21a* exhibited pronounced effects in deep layer excitatory neurons (**Supplemental Figure 11**). This finding indicates that the pathways perturbed by MYT1L deficiency have a signature similar to pathways disrupted by a subset of autism genes involved in axon guidance, neuronal migration, and chemical synaptic transmission, suggesting convergence with key autism-related genes and pathways (Velmeshev et al., 2023). Overall, these observed transcriptomic changes reveal molecular changes that preferentially affect the maturation and function of deep layer excitatory neurons.

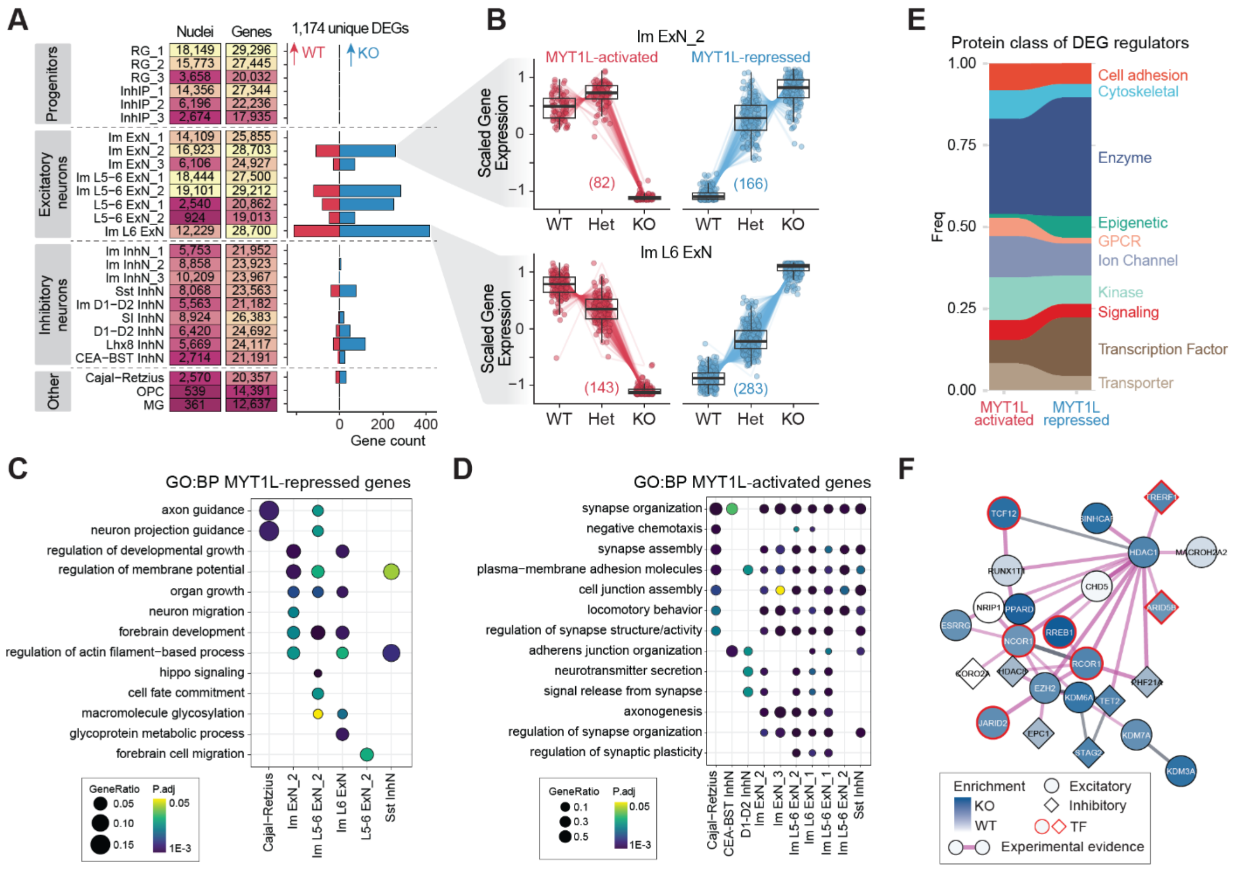


Figure 34: Gene expression changes across cell types

(A) Summary plot showing the numbers of nuclei and genes detected in each cluster. The bar plot shows the number of differentially expressed genes that are upregulated in WT (red) or upregulated in KO (blue). **(B)** MYT1L gene dose-dependent gene expression patterns of DEGs in Im ExN_2 and Im L6 ExN clusters separated by those that are MYT1L-activated (loss of expression in KO) and MYT1-represssed (gain of expression in KO). **(C, D)** Dotplot showing enriched GO biological processes in MYT1L-repressed **(C)** and MYT1L-activated **(D)** DEGs. **(E)** Plot showing the frequency of annotated protein classes of the DEGs. **(F)** STRING physical protein-protein interaction networks for a coregulatory module. Darker blues indicate the gene was higher expressed in KO, a red outline signifies that the gene is a TF, and the shape indicates the cell type.

5.4.3 Loss of MYT1L disrupts transcriptional maturation

Our differential analysis shows that MYT1L deficiency is associated with an immature transcriptional signature in excitatory neurons. While an explanation is that these genes are simply dysregulated, we hypothesize that MYT1L deficient neurons progress along their developmental trajectories at a slower pace, resulting in an immature gene signature. Additionally, we hypothesize that there is a critical moment during differentiation when MYT1L function is most needed to guide the developmental trajectory. To test these hypotheses, we assessed the differences in maturation trajectories leading up to and through this developmental window between genotypes. We used Monocle3 (Trapnell et al., 2014) to reconstruct a pseudotemporal trajectory (**Figure 35A**) independent of our prior cluster definitions. This models the cell state as a continuum of dynamic changes, allowing us to quantify gene expression changes as the cell progresses through differentiation. We observed subtle yet widespread disruption to the distribution of Het and KO nuclei compared to WT in pseudotime states (**Figure 35B-D**). This suggests a developmentally immature signature that can be missed when looking only at cell proportion based on cluster markers.

To identify drivers of excitatory neuron development, we analyzed the expression of TFs along pseudotime in WT cells, providing us with a putative timeline of gene activation and expression from progenitors to differentiated excitatory neurons (**Figure 35E**). Using these temporal profiles, we can then test if loss of MYT1L causes variations in the timing of expression of specific TFs which could profoundly impact the developmental trajectory of excitatory neurons. We found 27 TFs that showed disrupted timing of expression as a result of MYT1L deficiency using the Kullback-Leibler divergence test (**Figure 35F,G**). These TFs were generally de-repressed in Hets and KOs and are involved in developmental regulation (*Dlx5*, *Dlx6*, and *Hoxd10*), control of cell cycle progression (*Hbp1*), neurogenesis (*Nhlh2*, *Lmx1a*, and *Insm2*), and epigenetic regulation (*Tet2* and *Prdm*). To identify where MYT1L may have the greatest effect, we intersected all the excitatory pseudotemporal TFs with the E14 MYT1L CUT&RUN peaks (Chen et al., 2023) and found an enrichment of direct MYT1L targets during a transient period shortly after the transition from progenitor to postmitotic neuron, suggesting its important role during this critical moment. We also found that six genes within this pseudotime bin (*Efna4*, *Ccng2*, *Nbr1*, *Frmd4b*, *Sorsb2*, and *Midn*) were targets of ZBTB12, a molecular gatekeeper known to safeguard the unidirectional transition of progenitors to differentiated states(Han et al., 2023). Together, this provides a pseudotime-resolved sequence of MYT1L target gene expression, and identification of a critical developmental window, where alterations in these patterns during this sensitive moment may lead to disruptions in neuronal differentiation and maturation.

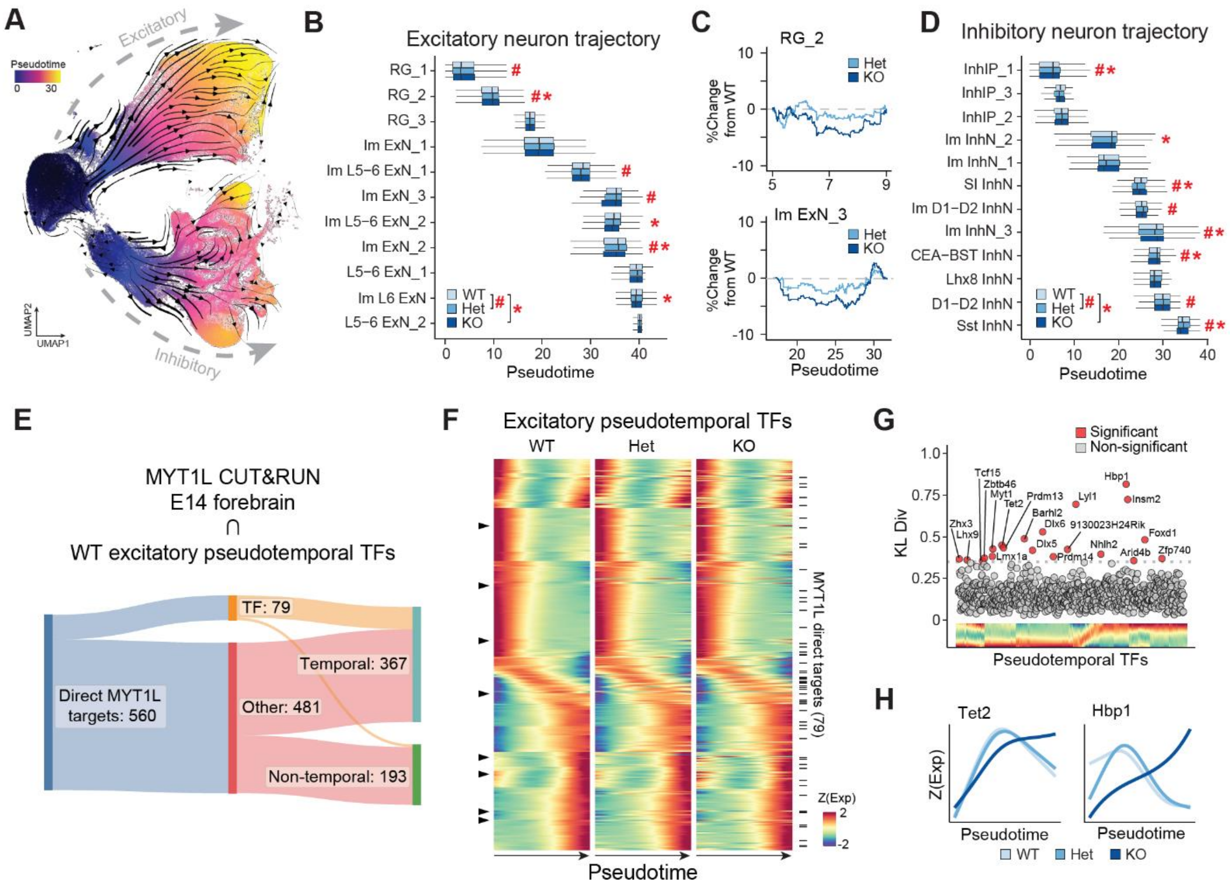


Figure 35: Loss of MYT1L disrupts excitatory neuron maturation

(A) UMAP plot of all nuclei from all genotypes colored by pseudotime and overlaid with RNA velocity trajectories. **(B)** Boxplots showing distributions nuclei from excitatory neurons along pseudotime. Kolmogorov-smirnov tests were performed to test for differences in distributions. P<0.05 for WT and Het comparisons are noted with a #, while statistically significant WT and KO comparisons are noted with a *. **(C)** Representative plots showing the relative differences of distribution in the RG_2 and Im ExN_3 clusters of MYT1L Het and KO nuclei compared to the WT distribution. **(D)** Boxplots showing the distributions of inhibitory neuron nuclei along pseudotime. P<0.05 for WT and Het comparisons are noted with a #, while statistically significant WT and KO comparisons are marked with a *. **(E)** Diagram showing the number of direct MYT1L targets identified by CUT&RUN that had a dynamic gene expression profile across pseudotime. **(F)** Heatmaps showing scaled expression of WT (left), Het (middle), and KO (right) excitatory neuron pseudotemporal genes. Each row represents a gene and sorted according to their expression peak in pseudotime. The black tick marks on the right note the rows in which genes are MYT1L direct targets determined by CUT&RUN in **E**. Black triangles on the left denote a subset of genes as examples with disrupted timing of expression. **(G)** Scatterplot showing the Kullback-Liebler divergence metric to identify differential pseudotemporal expression profiles in KOs compared to WT. **(H)** Representative traces of the differential pseudotemporal gene expression profiles for Tet2 and Hbp1 across genotypes.

5.4.4 Sensitivity of excitatory neurons persist throughout neurodevelopment

To investigate the long-term effects of MYT1L deficiency on both cell proportion and transcriptional changes, we conducted a snRNAseq analysis of the cortex in juvenile male and female WT and MYT1L Het animals at P21, when neurogenesis is largely completed. Analysis of KOs are not possible as they are not viable postnatally (Chen et al., 2021; Kim et al., 2022; Weigel et al., 2023). We analyzed snRNAseq data from 96,505 nuclei to identify 19 types of excitatory neurons spanning cortical layers, 11 subtypes of inhibitory neurons, and 8 non-neuronal types using hierarchical correlation mapping, referencing the taxonomies and subclass annotations from the Allen Brain Cell (ABC) Atlas (Yao et al., 2023) (**Figure 36A, Supplemental Figure 12**). When analyzing overall proportions of excitatory neurons, we observed significantly fewer L2/3 IT ENT and L4/5 IT neurons in MYT1L Het cortices compared to WT, while there were increased numbers in L6 CT, L6 IT, and L6b/CT neurons in the Hets (**Figure 36B**). By P21, 985 cluster pseudobulk DEGs were detected, of which 576 were unique to a single cell type and nearly exclusive to excitatory neurons (**Figure 36C, Supplemental Figure 13**). Similar to the E14 DEGs, we observed an increased number of upregulated DEGs upon loss of MYT1L in Hets, indicating de-repression. L6 neurons were the most affected, with modest effects on upper L2/3 intrathalamic (IT) cortex and mid-layer L4/5 IT neurons. 48% (280/576) of these DEGs overlapped with MYT1L CUT&RUN targets from adult prefrontal cortex (Chen et al., 2023) (**Figure 36C**). This significant overlap implies that a substantial portion of the DEGs observed at P21 may be directly influenced by MYT1L binding to their promoter regions. GO analysis revealed that DEGs upregulated in WT were associated with axon guidance, synaptic cell adhesion, and neurotransmission (**Figure 36D**), while genes upregulated in Hets were enriched in pathways related to nervous system development and axonogenesis (**Figure 36E**). This reflects the E14 enrichment analysis and demonstrates that MYT1L Het excitatory neurons are immature compared to WT and that while the magnitude of effect on neurodevelopment and maturation is greatest during embryonic development, some deficiency is sustained throughout early postnatal development.

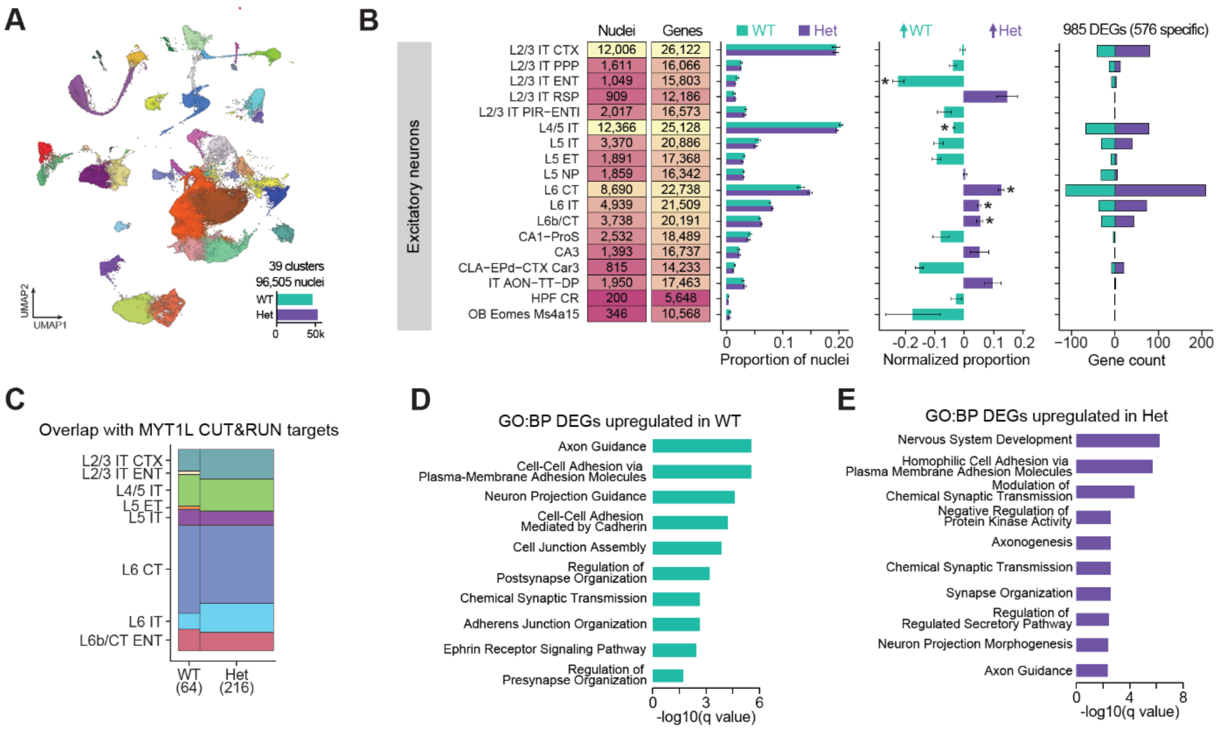


Figure 36: Single nucleus transcriptional profiling of P21 cortex in MYT1L animals

(A) UMAP projection showing 96,505 nuclei in 39 clusters from MYT1L WT and Het animals. **(B)** Summary plot showing the numbers of nuclei and genes detected in each cluster (left). Bar plots show the average relative proportions of nuclei in each annotated excitatory neuron cluster for MYT1L WT and Het genotypes. These proportions are normalized to WT (center bar plot). The right bar plot shows the number of pseudobulk differentially expressed genes that are upregulated in WT (cyan) or upregulated in Het (purple). **(C)** Mosaic plot showing the proportions of DEGs that overlap with MYT1L direct targets identified by CUT&RUN. The total numbers of overlapping genes are indicated in parentheses below the genotype labels. **(D, E)** GO analysis of biological processes of DEGs that are upregulated in WT and upregulated in Hets.

To deepen our understanding of the developmental progression from E14 progenitors to terminally differentiated cell types at P21, we integrated the two datasets together, analyzing a total of 313,335 nuclei. This integration revealed distinct developmental pathways for excitatory and inhibitory neurons, branching out from the clusters of progenitors (**Figure 37A**). For the subsequent analyses, we focused on the excitatory neuron trajectory encompassing 191,217 nuclei. We found that the E14 L5-6 ExN_1 and Im L6 ExN clusters were transcriptionally similar to the P21 L6 and L6b corticothalamic (CT) clusters and were observed at a transition zone between the ages. This indicates that the deep layer neurons are the first to exhibit markers indicative of a terminally differentiated cell type. The E14 Im ExN clusters showed a developmental trajectory towards the upper layer cortical neurons.

We next sought to test the hypothesis that MYT1L heterozygosity disrupted cell-type specific transcriptional maturation by P21. First, to understand the biological processes underlying the WT maturation of immature E14 L6 ExNs compared to their mature counterparts at P21, we performed gene set enrichment analyses (GSEA) on their gene expression profiles. As expected, we observed a de-enrichment of cell fate specification genes and an enrichment of neurotransmitter receptor activity genes in P21 L6 neurons (**Figure 37B,C**). Next, through differential gene expression analysis, we identified 4986 genes with significant differences between WT E14 and WT P21 L6 ExNs, highlighting a signature for neuronal maturation. Then, we further investigated the impact of MYT1L deficiency on the expression of these maturation-associated genes by comparing these 4986 genes with DEGs in P21 Het and WT L6 ExNs. Our analysis revealed that 35-42% of the P21 DEGs in the MYT1L deficient neurons overlapped with the maturation gene set (**Figure 37D**) ($P=1.3x10^{-4}$, chi-square test with Yates' continuity correction), indicating that about half of the transcriptional effects of MYT1L deficiency can be summarized as a disruption in neuronal maturation programs. To get more insight into the disrupted pathways, we analyzed these shared maturation genes (**Figure 37E**) as well as the genes that were only dysregulated at P21 (**Figure 37F**). We found that P21 DEGs upregulated in WTs showed an overall increase of expression of maturation-associated genes at P21 (upper right quadrant) and are related to synaptic transmission, GABA signaling, and ion transport (**Figure 37G**), while only a few E14 L6 maturation genes were upregulated (**Figure 37E**, lower right quadrant). In contrast, there was a significant number of P21 DEGs upregulated in Hets that showed higher expression of maturation genes at E14 (**Figure 37E**, lower left quadrant) which were related to synapse organization, axon guidance, and regulation of cell migration (**Figure 37H**). Finally, we observed a significant number of P21 Het DEGs that showed higher expression at P21 (**Figure 37E**, upper left quadrant), suggesting that MYT1L deficiency results in atypical expression of some developmental gene programs. GO analysis revealed an enrichment in pathways related to protein dephosphorylation and proteoglycan processes. Moreover, genes associated with semaphorin receptor binding genes, specifically *Sema4a*, *Sema7a*, and *Sema4d* were found to be upregulated. Notably, *Sema4d* is recognized for its role as an intrinsic inhibitor of axonal pathfinding (Moreau-Fauvarque et al., 2003; Yamaguchi et al., 2012; Zhang et al., 2014). This suggests that elevated levels of *Sema4d* in MYT1L Hets may impair axonal development and formation of synaptic connections. Collectively, this integrated analysis demonstrates that the P21 Het L6 ExNs exhibit an immature transcriptional signature, and the dysregulated genes suggests disrupted axon development and neurotransmitter signaling. The convergence of these findings underscores the critical role of MYT1L in guiding neuronal maturation and establishes a link between MYT1L heterozygosity and the perturbation of essential developmental pathways in L6 excitatory neurons.

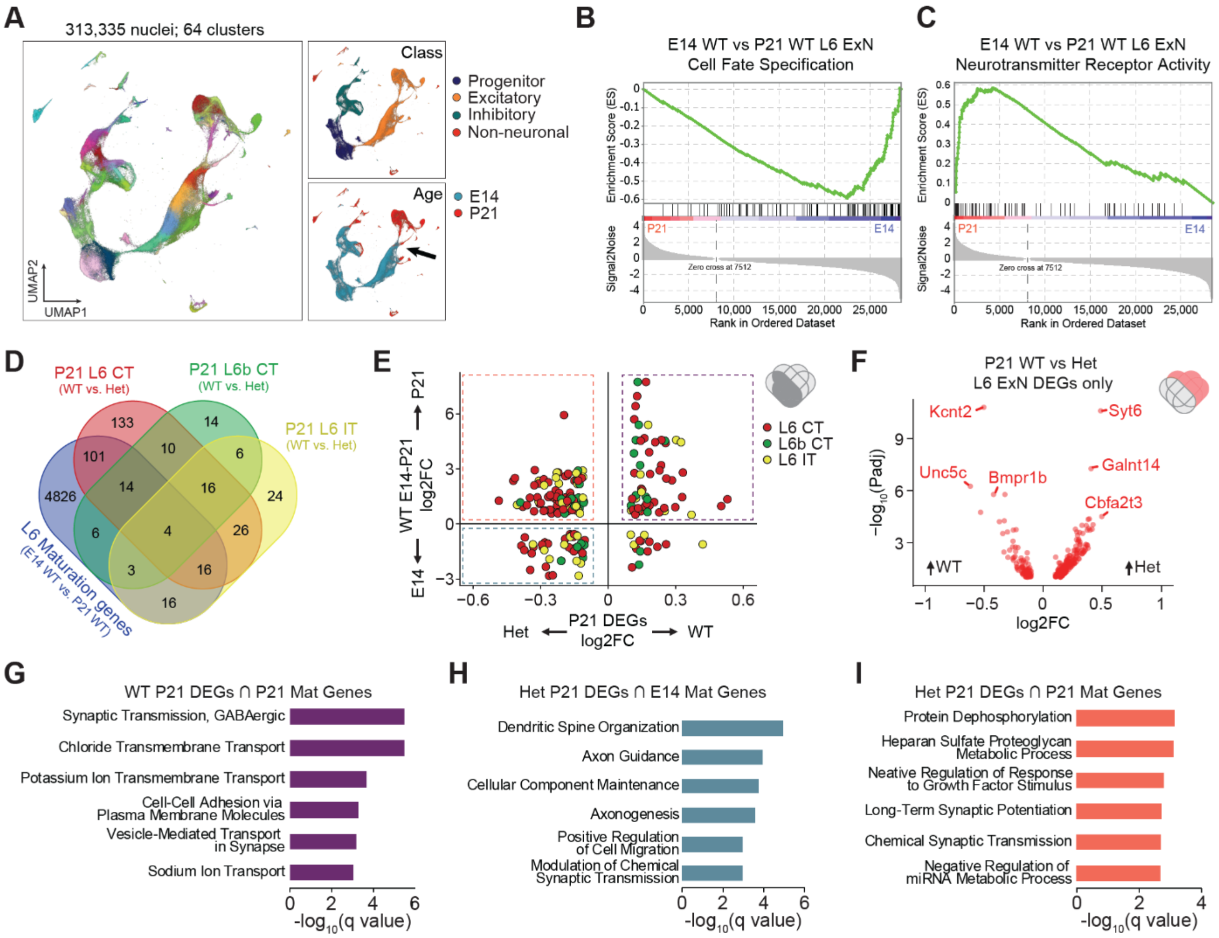


Figure 37: Integrated analysis of E14 and P21 nuclei

(A) UMAP projection showing integrated data from E14 and P21 datasets. The top right inset shows the cells colored by cell class. The bottom right inset shows the cells colored by age, with an arrow indicating the L6 transition zone. **(B,C)** GSEA analysis of the expression dataset comparing E14 L6 ExN to P21 L6 ExNs. In **B**, the green line shows the de-enrichment of genes in the Cell Fate Specification GSEA list in P21 L6 ExNs compared to E14 L6 ExNs. In **C**, the plot shows enrichment of genes associated with Neurotransmitter Receptor Activity in P21 compared to E14. **(D)** Venn diagram showing the number of overlapping genes in the maturation-associated gene set (E14 WT L6 ExN vs. P21 WT L6 ExNs) with DEGs from WT and Het P21 L6 clusters. **(E)** Scatterplot comparing the magnitude and direction of effect of the common genes between the WT L6 maturation genes and P21 L6 ExNs (shaded region indicated in venn diagram cartoon in the upper right corner). The x axis represents the log2FC from WT and Het P21 DEGs from the 3 L6 ExN clusters. The y axis is the log2FC of WT E14 L6 ExN and WT P21 L6 ExN DEGs. The upper right quadrant (purple) are genes upregulated in P21 WT and also generally at P21. The upper left (orange) quadrant are genes upregulated in P21 Hets that are also generally increased at P21. The bottom right (teal) are genes that are upregulated in P21 Hets, that are upregulated at E14. **(F)** Volcano plot showing the log2 fold change and direction of effect of DEGs from P21 L6 ExN DEGs that are not shared with the E14-P21 WT maturation gene list. **(G-I)** GO analyses of DEGs that are in the selected quadrants of **E**.

5.5 Discussion

In this study, we analyzed the transcriptomes of 313,335 nuclei across neurodevelopment in a model of MYT1L mutation, allowing us to delineate the molecular and cellular consequences of loss of MYT1L. Leveraging single-cell atlases of the mouse brain as references, we have advanced our understanding of how the disruption of a single TF can perturb neurodevelopment and maturation processes. We implemented pseudotime and RNA velocity to quantitatively assess neuronal transcriptional maturation. Our results reveal that although MYT1L is expressed in all neurons, deep layer excitatory neurons are particularly susceptible to MYT1L haploinsufficiency, resulting in an immature transcriptomic signature. This signature can be a result of precocious differentiation of earlier progenitors, a slower transition from progenitors into excitatory neurons, or cells stalled in a partially differentiated state. This deficiency causes a delay in neuronal maturation at E14, and the dysregulation of the regulatory programs that control neuronal maturation persist through P21. We also demonstrate that MYT1L primarily functions as a transcriptional repressor, affecting gene expression programs linked to key developmental processes like axon guidance, neuron migration, and cell fate commitment. We found these MYT1L-repressed pathways were gene dose-responsive, with even slight reductions in MYT1L levels leading to substantial upregulation of these genes. In contrast, genes activated by MYT1L, mainly those involved in synaptic function and neurotransmission, were more tolerant of haploinsufficiency. Additionally, our findings show that the dysregulated genes were enriched with TF and epigenetic regulators, which can initiate a cascade of downstream effects stemming from MYT1L perturbation. Using a brain region and age matched MYT1L CUT&RUN dataset, most effects at E14 were indirect and the percentage of direct effects increased at P21. Because MYT1L recruits the SIN3B deacetylation complex, it is possible that many of the "indirect" regulatory targets as measured by CUT&RUN are in fact direct as the deacetylated histones can have repressive epigenetic “memories” (Ramaswami et al., 2020), and a deeper analysis of histone state may disentangle this. Nonetheless, the overall results highlight the cell type-specific and developmental stage-specific nature of MYT1L’s function.

To date, there are three transgenic mouse models that disrupt different exons of MYT1L that converge on a hyperactivity phenotype, while other behaviors are varied likely because different assays were used (Chen et al., 2021; Kim et al., 2022; Wöhr et al., 2022). A study by Weigel et al. (Weigel et al., 2023) performed scRNAseq on the neonatal (P0) forebrain from the mice described in Wohr et al. (Wöhr et al., 2022) and found a decreased number of newly formed neurons in the subventricular zone. Additionally, the authors observed an increased expression of non-neuronal gene expression programs which can perturb neuronal cell identity. Here, we observed a slight upregulation of mouse embryonic fibroblast (MEF) signature genes at P21, albeit with a minor effect size. They also show L5/6 neurons as having the greatest number of DEGs, consistent with the work here. Interestingly, they also observed a moderate increase in the proportion of striatal inhibitory neurons in MYT1L Hets, however, additional experiments are needed to interpret these findings.

In the current era of genomics bulk RNAseq and snRNAseq are primary techniques for examining transcriptional landscapes across various biological perturbations. However, the concordance between DEGs identified through bulk and single-cell approaches is generally low. This discrepancy can be attributed in part to factors such as RNA capture efficiency, gene dropouts, and data sparsity. By leveraging a previously reported bulk RNAseq dataset from Chen et al. (Chen et al., 2021), we perform a pairwise comparison with the pseudobulk aggregated snRNAseq dataset to discern whether the noted differences stem from variations in cell proportions or from intrinsic transcriptional changes within each cell type. In analyzing our E14 data, we observe that the discrepancies in cell proportions for MYT1L Het samples compared to WT are generally within a 10% margin, albeit with some exceptions, including the Im ExN_3, L5-6 ExN_1, and L5-6 ExN_2 clusters. Upon examining clusters that exhibit a substantial number of pseudobulk DEGs, such as Im L6 ExN, Im ExN_2, and Im L5-6 ExN_2, we identify only minor shifts in cell proportions. For these specific clusters, it appears that differential expression is predominantly driven by transcriptional alterations within the cell types, rather than changes in their proportions.

While MYT1L is a neuron-specific TF, it remains uncertain whether its loss may exert non-cell autonomous effects on surrounding glia during postnatal development. Our analysis revealed a relatively higher proportion of oligodendrocytes and microglia in P21 MYT1L Hets compared to WT, but we did not detect any DEGs in these cell types. However, with relatively low numbers of cells and gene counts in these clusters, there may be differences below our threshold to detect. These findings suggest the possibility of MYT1L-mediated effects on oligodendrocytes and microglia, yet further investigations with increased cell numbers are needed to elucidate the nature and extent of these effects. What was abundantly clear in our data at both time points was the profound, cell type-specific transcriptional responses to MYT1L deficiency, especially in deep layer excitatory neurons.

A striking observation from our analysis of pseudobulk DEGs reveals that around 15% of these DEGs overlap with SFARI gene candidates and display significant dysregulation in deep layer excitatory neurons, particularly the L5-6 ExN_1, L5-6 ExN_2, and Im L6 ExN clusters (Error! Reference source not found.). Interestingly, expression levels of these genes were elevated in KOs compared to WTs, hinting at the possible loss of a repressive mechanism. Despite the majority of these genes not being identified as direct targets in the E14 MYT1L CUT&RUN analysis, it is important to note that the CUT&RUN dataset only includes gene targets based on MYT1L occupancy in promoter regions, omitting potential targets influenced by distal regulatory elements, as it remains challenging to systematically link long-distance enhancers to specific gene targets. Nevertheless, the observed differential expression allows us to hypothesize that MYT1L may function as a transcriptional regulator, influencing SFARI gene expression directly or indirectly, or through mechanisms like epigenetic memory. Ultimately, the disrupted pathways we've identified represent a core set of pathways that are critical for proper neurodevelopment.

Overall, our comprehensive analyses across developmental stages and MYT1L deficiency's impact underscore its pivotal role in neuronal maturation and development. These findings reveal that the developmental trajectory and transcriptional landscape of excitatory neurons are markedly altered by MYT1L deficiency, with effects persisting from early neurogenesis through adolescence. This study not only advances our understanding of the genetic and molecular foundations of neuronal development, but also demonstrates how we can deeply characterize genetic perturbations at scale to investigate the enduring impact of MYT1L on the maturation and function of neurons.

5.6 Materials and methods

Animals and tissue collection

All animal studies were approved by and performed in accordance with the guidelines of the Animal Care and Use Committee of Washington University in Saint Louis, School of Medicine and conform to NIH guidelines of the care and use of laboratory animals. The animals were housed in controlled environments with a 12-hour light-dark cycle, constant temperature and relative humidity, and *ad libitum* access to food and water. The C57BL/6-*Myt1l*^{em1Jdd}/J (Myt1l S710fsX (Chen et al., 2021); Jackson Laboratories 036428) line was maintained with breeding pairs consisting of a Myt1l Het and an in-house C57BL/6J mouse. The transgenic line was refreshed every 8-10 generations by backcrossing to freshly obtained C57BL/6J males and females from Jackson Laboratories. Upon weaning at P21, the animals were group-housed by sex and genotype. To obtain homozygous animals for embryonic studies, timed pregnant Myt1l Het x Het breeding pairs were set up and vaginal plugs were checked the following morning. The first morning after the plug was found was considered to be E0.5. E14-14.5 embryos were rapidly dissected from the uteri of the mice in HBSS on ice. The pups were decapitated, and the brains were quickly extracted from the skulls. The meninges was removed, the forebrain was dissected, flash frozen in liquid nitrogen, and stored at -80°C. Tail tissue was collected from each embryo for gDNA isolation and genotyping. Cortical tissue from P21 pups were harvested and stored using the same method.

Genotyping

Tissue (tail biopsy, ear punch, or toe clipping) was obtained from each animal and placed in a PCR tube. 100 µl lysis buffer (25mM NaOH, 0.2mM EDTA, pH 12) was added to each tube and incubated at 99°C for 60 min in a thermocycler. Once the samples cooled to room temperature, 100 µl 40mM Tris-HCl pH 5 was added to neutralize the alkaline lysis buffer. The crude lysate containing genomic DNA (gDNA) was stored at 4°C. Three reactions were performed for each animal to genotype the WT allele, MYT1L mutant allele, and SRY to determine sex (see **Table 19** for sequences). The PCR conditions for genotyping with allele specific PCR primer pairs involved mixing 1 µl of the crude gDNA with 5 µl Phusion High-Fidelity PCR Master Mix (New England Biolabs M0531), 1 µl 10µM MYT1L_Comm_For/Rev primer mix, 1 µl 10µM β-actin For/Rev primer mix, and 2 µl ddH2O. Thermocycling conditions were as follows: 98°C for 3 min; 35 cycles of: 98°C for 10 sec, 61°C for 20 sec, 72°C for 20 sec; 72°C for 5 min; and 4°C hold. For SRY, 1 µl crude gDNA was added to a master mix containing 5 µl OneTaq Quick-Load 2X Master Mix (New England Biolabs M0271), 1 µl 10µM SRY For/Rev primer mix, 1 µl 10µM β-actin For/Rev primer mix, and 2 µl ddH2O. Thermocycling conditions were as follows: 94°C for 3 min; 35 cycles of: 94°C for 10 sec, 60°C for 20 sec, 68°C for 20 sec; 68°C for 5 min; and 4°C hold. Multiplexing β-actin not only confirms the presence of gDNA, but also minimizes non-specific amplification of the MYT1L mutant band in WT samples. PCR products were run on a 1% agarose gel and visualized with GelRed (Biotium 41003).

<u>Nuclei isolation and fixation</u>

In this study, nuclei from E14 and P21 prefrontal cortices were isolated from flash frozen tissue. The brain tissues were Dounce homogenized in ice-cold homogenization buffer (10mM Tris-HCl pH 7.4, 10mM NaCl, 3mM $MgCl_2$, 1mM DTT, 1X cOmplete EDTA-free Protease Inhibitor (Roche 4693132001), and 0.2 U/µl RNasin Inhibitor (Promega N2515) using a 2 ml KIMBLE KONTES Dounce Tissue Grinder (DWK 885300-002) with 15 strokes with the “A” large clearance pestle, followed by 15 strokes of the “B” small clearance pestle. The homogenate was transferred to a 15 ml centrifuge tube. Walls of the homogenizer tubes were washed with 1ml of homogenization buffer and combined with the homogenate in the 15 ml tube. The nuclei were pelleted by centrifugation in a swinging bucket rotor at 500x g for 5 mins at 4°C. The supernatant was aspirated and discarded.

For E14 samples, the pellets were washed twice with 1ml nuclei wash buffer (DPBS, 1% BSA, and 0.2 U/µl RNase inhibitor), filtered through a 40µm Flowmi cell strainer (Millipore Sigma BAH136800040), and counted using a hemocytometer with Trypan Blue.

For P21 samples, the pellets after the first centrifugation were resuspended in 1 ml homogenization buffer. A gradient centrifugation step using 25:35 Iodixanol was performed to purify the nuclei from cellular debris and myelin generated during tissue dissociation. To make the 25% Iodixanol layer, 1 ml 50% iodixanol was added to 1 ml of the homogenate containing the nuclei and debris. This was carefully layered on top of 2 ml 35% iodixanol in a clear polycarbonate tube (Beckman 355672) and centrifuged at 10,000x g for 30 min at 4°C with deceleration turned off. After the gradient centrifugation, myelin and cellular debris remaining at the top of the 25% iodixanol layer and was aspirated and discarded. The purified nuclei at the interface of the two Iodixanol layers was collected and transferred to a clean 15 ml centrifuge tube. The volume was brought up to 6 ml with nuclei wash and resuspension buffer and pelleted by centrifuging at 500x g for 5 min at 4°C. The supernatant was carefully removed, washed once with nuclei wash buffer to ensure removal of carryover Ioxidanol, filtered through a 40µm Flowmi cell strainer, and counted using a hemocytometer with Trypan Blue.

For both E14 and P21 samples, 500k-2.5M nuclei were resuspended in 500 µl calcium and magnesium-free DPBS and used as input into the ScaleBio Sample Fixation Kit (Scale Biosciences 2020001) protocol according to manufacturer’s instructions. After fixation, the nuclei were counted once more and checked for quality using a microscope with a 60x objective. The nuclei were then stored at -80°C until all samples have been collected and fixed.

Single-nucleus RNAseq library preparation

Libraries were prepared from fixed E14 and P21 nuclei separately. For the E14 timepoint, a total of 9 samples (3 biological replicates of a mix of males and females per MYT1L WT, Het, and KO genotypes) were used. For the P21 timepoint, a total of 12 samples consisting of 3 biological replicates per sex per MYT1L WT and Het genotypes were used. The day of the library preparation, the frozen fixed nuclei were thawed on ice and each sample was counted twice using a hemocytometer.

For the E14 samples, the ScaleBio Single Cell RNA Sequencing Kit v1.0 (Scale Biosciences 2020008) was used according to manufacturer's instructions. Nuclei from each sample were loaded at 10,000 nuclei per well to the 96-well Indexed RT Oligo Plate to add the RT barcode and UMI onto each transcript during reverse transcription. By loading each sample into a distinct set of wells, the RT barcodes can serve as sample identifiers, enabling all genotypes to be processed on the same plate in a single batch per age. The nuclei from each well were then collected and pooled using the Scale Biosciences' supplied collection funnel, mixed, and distributed across the 384-well Indexed Ligation Oligo Plate where the Ligation Barcode was added to each UMI-RT barcoded transcript. Then, the nuclei were once again collected and pooled using another collection funnel and counted with a hemocytometer with Trypan Blue. A total of 1,600 nuclei were distributed per well of the 96-well Final Distribution Plate. In each well, second strand synthesis was performed followed by a cleanup step. The PCR products were then tagmented followed by an indexing PCR step to add a third barcode to each well. 5ul from each of the 96 libraries were pooled and cleaned using 0.8X SPRIselect beads (Beckman Coulter B23317). The average fragment size of the final library was quantified using a High Sensitivity D5000 Screentape (Agilent). The library concentration was quantified using the NEBNext Library Quant

Kit for Illumina (New England Biolands E7630S). The libraries were sequenced on a shared S4 flowcell on a NovaSeq6000 (Illumina) instrument to a target depth of 10,000 reads per nucleus.

For the P21 samples, the protocol described above was followed through the cleanup step. Prior to tagmentation, 3 µl (half of the total volume) was transferred to a clean 96-well PCR plate to create a Calling Cards Final Distribution Plate. This plate was set aside to pilot single-nucleus Calling Cards (snCC) library preparation, the results of which will be reported in a future methods paper. The remaining 3 µl was used for the remainder of the ScaleBio protocol with slight modifications. To account for the reduced volume of template input, the volumes for all subsequent steps have been halved to keep all reaction proportions the same. Additionally, the Indexing PCR program was increased to 16 cycles instead of 14. The libraries were pooled, cleaned, and quantified as described above according to manufacturer's instructions. This library pool was sequenced on a shared 25B flowcell on a Novaseq X (Illumina) instrument to a target depth of 10,000 reads per nucleus.

Single-nucleus RNAseq analysis

Base calls were converted to FASTQ format and demultiplexed by Index1 barcode by the Genome Technology Access Center at the McDonnell Genome Institute (GTAC@MGI). Combinatorial barcode demultiplexing, barcode processing, adapter trimming, read mapping to the mm10 reference genome, single-nuclei counting, and generation of the feature-barcode matrices were done using ScaleBio Single-cell RNA Nextflow Workflow v1.4 (https://github.com/ScaleBio/ScaleRna). The count matrices were brought into Seurat for downstream analyses. For quality control, a UMI-gene cutoff of 800-6000 UMIs and 300-3000 genes was used, followed by filtering out multiplets by DoubletFinder (McGinnis et al., 2019) for each sample. After filtering, a total of 216,830 nuclei across all samples remained, with a median of 3,204 UMIs and 1,819 genes per nucleus. The count matrices were log2 normalized, centered, and scaled using a scaling factor of 10,000. The top 3000 most variable genes were identified using dispersion and mean expression thresholds. PCA was then performed followed by dimensionality reduction by UMAP and unsupervised clustering using the Louvain algorithm using the FindClusters function of Seurat. A range of values was tested for the resolution parameter and a clustering tree was plotted using clustree (Zappia and Oshlack, 2018) to determine 0.8 as the optimal resolution for the E14 dataset. Cluster marker genes were defined by grouping the clusters by genotype, setting logfc.threshold=0.5 and min.pct=0.25, and comparing the fold changes between pct.1 and pct.2 using FindConservedMarkers. Cell type annotations were done referencing literature and the ABC atlas. Pseudobulk differential expression analysis between genotypes within cell types was performed using DESeq2 (Love et al., 2014). Pseudotime and trajectory analysis was done using Monocle3 (Trapnell et al., 2014). To prepare the data for RNA velocity analysis, .loom files containing the spliced and unspliced counts matrices were constructed from the .bam files and feature-barcode matrices using velocyto (La Manno et al., 2018) and the mm10 genome. RNA velocity was then computed with the dynamical model of scVelo (Bergen et al., 2020), which was then used in CellRank's (Lange et al., 2022) VelocityKernel to compute a macrostate transition matrix to classify initial, terminal or intermediate cell states.

For analysis of the P21 brains, a similar workflow as described above was used. After quality control, 96,505 nuclei remained with a median of 3,447 UMIs and 1,597 genes per nucleus. Cell type annotations was done exclusively using the ABC Atlas (Yao et al., 2023). The total library counts were normalized, centered, and scaled using a scaling factor of 10,000. The dimensionality of the data was reduced first with principal component analysis on 100 components based on the top 3,000 most variable genes. The graph was then embedded and visualized in two dimensions using UMAP. The nuclei were clustered using the Louvain algorithm using the FindClusters function of Seurat. A range of values were tested for the resolution parameter and a clustering tree was plotted using clustree to determine 0.8 as the optimal resolution for the P21 dataset.

Statistics

No statistical methods were used to predetermine sample sizes. Samples were generally littermates and genotypes were assigned randomly by the sperm at conception, with no input from investigators. The investigators were not blinded to the samples, but all samples were processed in parallel in the same batch.

5.7 Data and code availability

The data generated in this study can be downloaded in raw and processed forms from the Gene Expression Omnibus (GSE262368) and Neuroscience Multi-omic Data Archive (NeMO). The E14 MYT1L CUT&RUN dataset was downloaded from Gene Expression Omnibus (GEO: GSE222072). The ScaleBio Single-cell RNA Nextflow Workflow to process raw sequencing data into feature-barcode matrices can be found on Github (https://github.com/ScaleBio/ScaleRna). The code used to process, analyze, and visualize the data can be found at bitbucket (https://bitbucket.org/jdlabteam/yen-et-al-myt1l-snrnaseq/src/main/).

5.8 Acknowledgements

I thank members of the Dougherty and Mitra laboratories for helpful discussions and feedback, particularly M. Vasek for critical reading and copyediting of the manuscript; Y. Li for helpful analysis suggestions; the DNA Sequencing and Innovation Lab (DSIL) and the Genome Technology Access Center at the McDonnell Genome Institute (GTAC@MGI) for their sequencing expertise and services; and F. Schlesinger and B. Biddy for bioinformatic support for the ScaleBio pipeline. This work was funded by grants from the National Institute of Mental Health (RF1MH117070, RF1MH126723, and R01MH124808 to J.D.D. and R.D.M.). A.Y. was supported in part by T32HG000045 from the National Human Genome Research Institute.

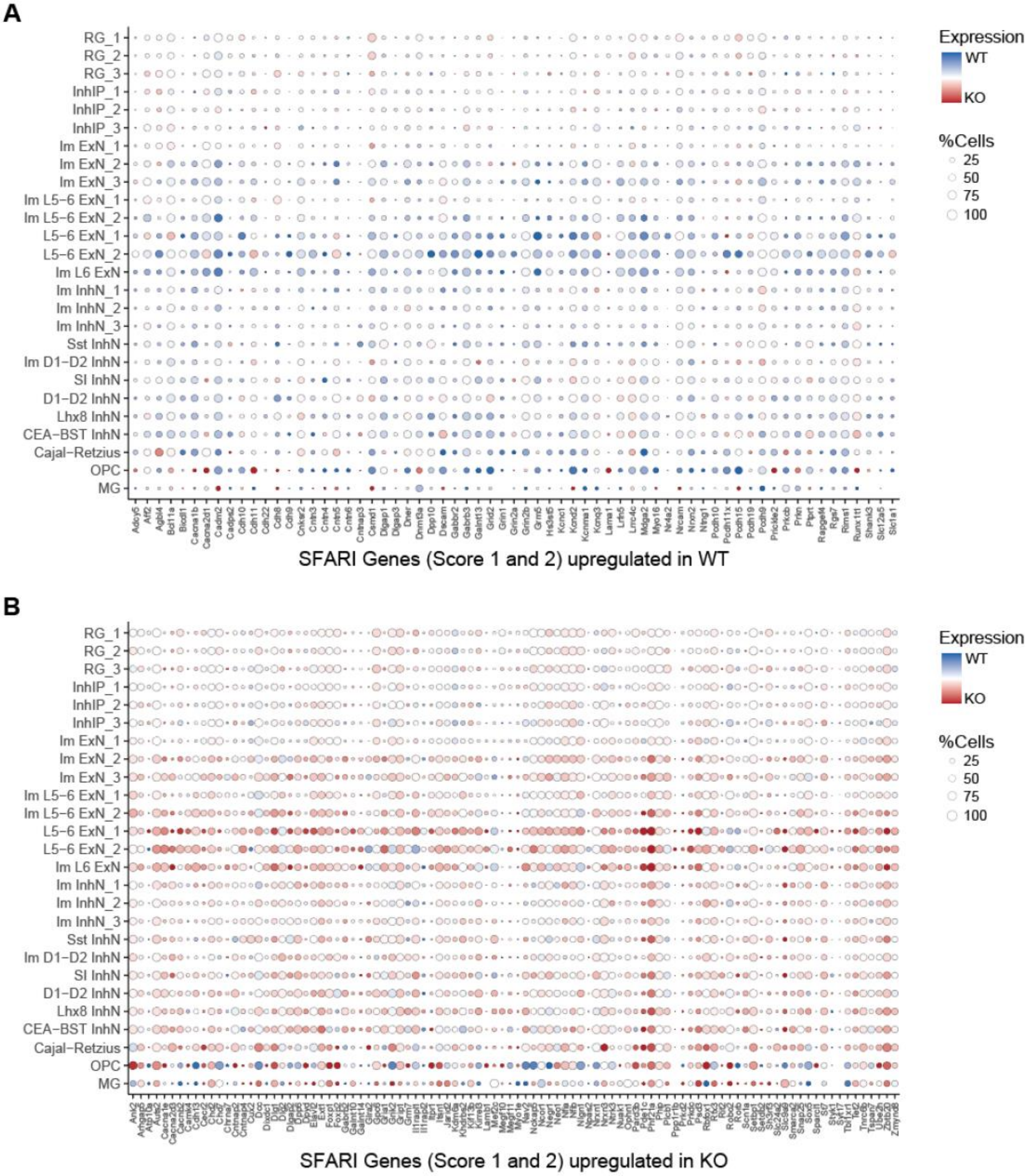
5.9 Author contributions

Project conceptualization: A.Y., R.D.M., and J.D.D. Method development, experiments, and data collection: A.Y., X.C., D.S., F.L., M.C., J.H-L, and J.C. Formal analysis: A.Y., Y.W., R.D.M., and J.D.D. Figures and data visualization: A.Y., R.D.M., and J.D.D. Writing-original draft: A.Y. and J.D.D. Writing-review and editing: A.Y., J.C., R.D.M., and J.D.D. Project coordination: A.Y., R.D.M., and J.D.D. Funding acquisition: R.D.M. and J.D.D.

5.10 Disclosures

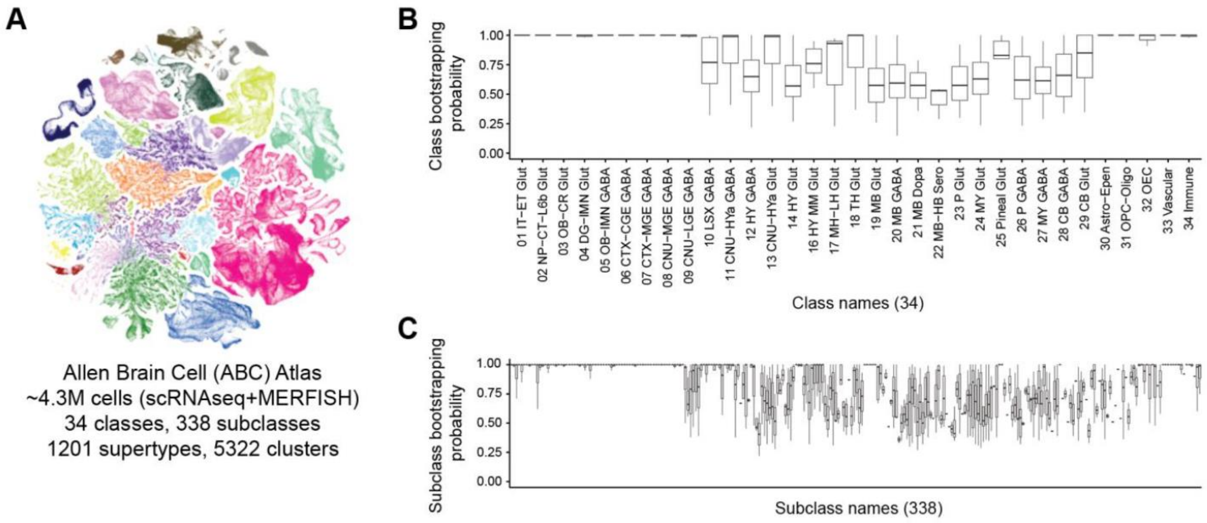
D.D.S. and F.L. are employees of Scale Biosciences. R.D.M has filed a patent application for self-reporting transposon (SRT) technology.

5.11 Supplemental figures and tables



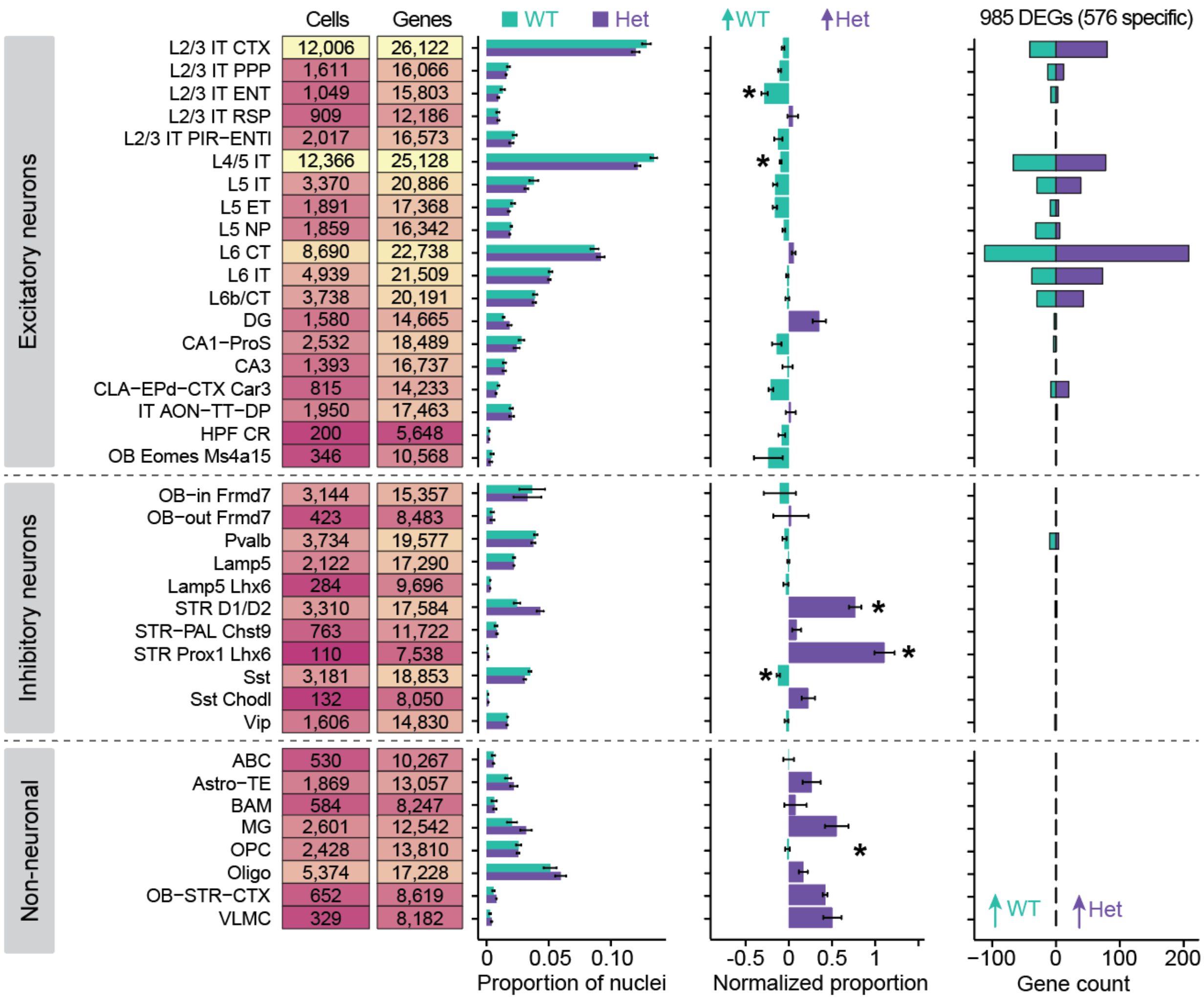
Supplemental Figure 11: Cell type-specific dysregulated genes associated with autism

164 differentially expressed genes from WT and KO E14 were found to be high confidence SFARI genes with a score of 1 or 2. The dotplot shows the genes as columns and cell types as rows for **(A)** genes that were upregulated in WT (blue) and **(B)** genes that were upregulated in KO (red). The relative size of each circle represents the percentage of cells expressing the gene.



Supplemental Figure 12: Standardized cell annotation of P21 dataset using the Allen Brain Cell Atlas

(A) Representative UMAP from the Allen Institute's Allen Brain Cell (ABC) Atlas (https://portal.brain-map.org/atlases-and-data/bkp/abc-atlas) that consists of transcriptomes and anatomical location of millions of cells. Using MapMyCells (https://portal.brain-map.org/atlases-and-data/bkp/mapmycells) and hierarchical correlation mapping, cells from this study were mapped onto the atlas. For each cell, a random set of 90% of the marker genes was selected, then mapped to the atlas by traversing the taxonomy by starting with classes, then proceeding to subclasses, supertypes, and clusters. This was repeated 100 times to obtain the bootstrapping probability. The bootstrapping probability for class is shown in **(B)** and subclass is shown in **(C)**. Labels with a high bootstrapping probability (close to 1) are considered high confidence labels and subclass names were used to annotate nuclei from this study.



Supplemental Figure 13: Single nucleus transcriptional profiling of P21 cortex in MYT1L animals

Summary plot showing the numbers of nuclei and genes detected in all clusters from P21 animals. The excitatory neuron subset was shown in (**Figure 36B**). Bar plots show the average relative proportions of nuclei in each annotated excitatory neuron cluster for MYT1L WT and Het genotypes. These proportions are normalized to WT (center bar plot). The right bar plot shows the number of pseudobulk differentially expressed genes that are upregulated in WT (cyan) or upregulated in Het (purple).

Table 19: Primer sequences for Chapter 5

Primer Name	Sequence (5’ →3’)	Product
SRY_For	TTGTCTAGAGAGCATGGAGGGCCATGTCAA	273 bp
SRY_Rev	CCACTCCTCTGTGACACTTTAGCCCTCCGA	
β-actin_For	AGAGGGAAATCGTGCGTGAC	150 bp
β-actin_Rev	CAATAGTGATGACCTGGCCGT	
MYT1L_Comm_For	CCAAGTCCTGTCCTACCCAAGT	380 bp
MYT1L-WT_Rev	TCTTGCTACACGTGCTACT	
MYT1L-Mut_Rev	TCTTGCTACACGTACTGGA	

Chapter 6: Single-nucleus Calling Cards with combinatorial indexing

6.1 Preface

This chapter contains contents from a collaboration with Scale Biosciences.

6.2 Introduction

Single-cell transcriptomics have emerged as one of the standard methods for analyzing cellular diversity within complex tissues. Despite the insights provided by the high-resolution transcriptomic data, it only captures a fraction of the complexity of the molecular networks. This recognition has spurred the development of techniques capable of multimodal "omic" measurements, which can concurrently measure whole genome attributes (Macaulay et al., 2016, 2015), chromatin accessibility (Cao et al., 2018; Chen et al., 2019), DNA modifications (Guo et al., 2013; Hu et al., 2016; Smallwood et al., 2014), histone modifications (Bartosovic et al., 2021; Rang et al., 2022), and proteomic profiles (Kochan et al., 2015; Soh et al., 2016) in parallel with the transcriptome. Such methodologies aim to provide an enriched and integrative view of these molecular networks, facilitating the correlation of their dysfunctions with pathologies.

Despite these advancements, there remains a notable gap in effectively characterizing transcription factor (TF) binding and enhancer dynamics at the single-cell level. Addressing this, Moudgil et al. adapted the Calling Cards technology for application with the 10x Genomics platform, thereby enabling the longitudinal recording of protein-DNA interactions at the single cell level. The authors demonstrated the adaptability of the single-cell Calling Cards approach using various TFs fused to the hyperactive *piggyBac* (hyPB) transposase. Additionally, they demonstrated its application in vivo by profiling BRD4 binding in the postnatal mouse cortex, identifying distinct regulatory elements in astrocytes and neurons. However, a key limitation of single-cell Calling Cards is its capacity to process up to 10,000 cells per sample per reaction. This constraint necessitates conducting multiple reactions to analyze larger cell populations or to study complex tissues comprehensively, significantly inflating costs. This issue is particularly challenging when undertaking comparative analyses across different conditions or integrating biological replicates, emphasizing the need for more scalable and cost-efficient solutions.

To overcome these limitations, I have adapted Calling Cards to the ScaleBio combinatorial indexing platform. This integration enables the analysis of hundreds of thousands of cells within a single experiment, significantly scaling throughput in a cost-effective manner. In Chapter 5, I demonstrated how the ScaleBio scRNAseq platform could reveal nuanced biological insights that are only achievable through the analysis of many transcriptomes. In this particular experiment, adeno-associated virus (AAV) expressing Calling Cards reagents were injected into the ventricles of neonatal MYT1L WT and heterozygous pups, facilitating widespread cortical labeling (**Figure 11B**, **Figure 38**). Over the initial three weeks of postnatal development, Calling Cards recorded BRD4 binding, which is a readout of enhancer activity. Using this method, the primary objectives of this study were to ascertain whether MYT1L mutations lead to differential enhancer activities in a cell type specific manner, to identify cell types most impacted by this epigenetic perturbation, and to determine if these regulatory elements can be associated with the transcriptomic signatures described in Chapter 5. Here, I introduce the development of single-nucleus Calling Cards using combinatorial indexing and share findings from the initial pilot experiments. While I focus exclusively on demonstrating its application in nuclei through a specific testcase, this method is also compatible with cells.

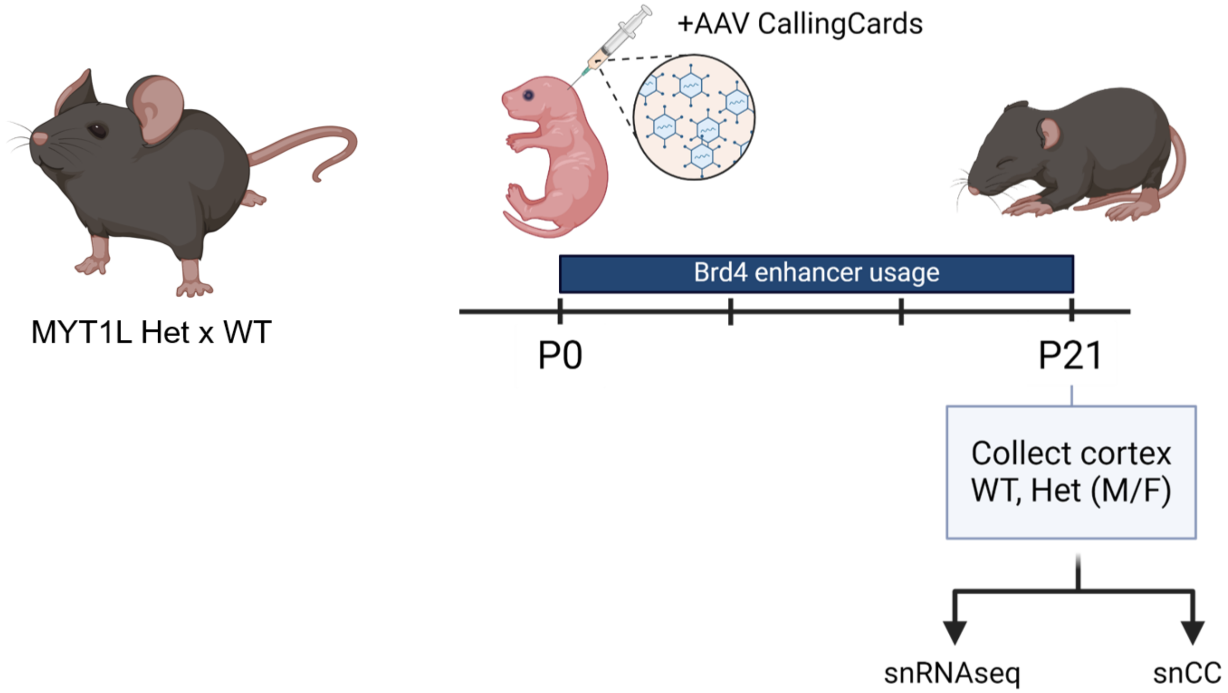


Figure 38: Schematic of experiment to pilot single-nucleus Calling Cards

The experimental workflow for this pilot experiment. P0-1 pups from MYT1L WT x Het breeding pairs were injected transcranially with a 1:1 mixture of AAV9-hyPB and AAV9-H2B-tdTomato-SRT.The brains were harvested at P21 and the cortices were dissected. Nuclei were isolated from the tissue for the parallel analysis of transcriptome with ScaleBio snRNAseq and enhancer usage by single-nucleus Calling Cards. The final cohort contained 3 biological replicates each of male and female MYT1L WT and Het animals.

6.3 Results

The goal of the workflow is to recover SRTs from the snRNAseq libraries so that the Calling Cards insertions can be mapped to the genome and associated with a cell type. To achieve this, the library preparation began with the standard ScaleBio protocol for the first two rounds of barcoding to add the RT and ligation barcodes, followed by second strand synthesis and enzymatic cleanup. The PCR products were then divided: half was transferred to a new 96-well plate to create a duplicate plate (**Figure 39A**). One plate proceeds with the remaining steps of the ScaleBio protocol to produce snRNAseq libraries, while the other plate undergoes amplification of SRTs for the Calling Cards libraries. This amplification uses a biotinylated primer set that also introduces well-specific PCR barcodes, matching those in the snRNAseq library to link SRTs to their respective cell barcodes. Next, the amplified SRTs were then circularized to bring the transposon-genome junction on the 5' end of the mRNA close to the Ligation, UMI, and RT barcodes which are at the 3' end. Post-circularization, the products are sheared and ligated with adapters to prepare the fragments for high-throughput short read sequencing using custom primers. To analyze the Calling Cards sequencing data, the reads were first filtered to remove read pairs that do not contain SRTs, then the passing reads were used as input into the ScaleBio analysis pipeline to perform the barcode parsing, cell demultiplexing, and alignment (**Figure 39B**). The aligned transcripts were then annotated to identify the location of the SRT with base-pair resolution. The cell barcode from the Calling Cards were then cross-referenced with the snRNAseq data to identify the cell type. This method, in comparison to the previously published 10x Genomics approach for Calling Cards, demonstrated a greater than fivefold increase in sequencing efficiency and output, making it a significant improvement in data acquisition (**Supplemental Figure 14**). This improvement is likely because the PCR barcode was switched to the i5 side and sequenced as the custom Index2 read, while the previous method relied on custom Index1 reads to demultiplex which was unreliable because of the custom priming strategy.

The Calling Cards insertions are cataloged in the qBED format (Moudgil et al., 2020a), which is a modified version of the standard BED format. This format is designed for compatibility with established tools like bedtools (Quinlan and Hall, 2010) for downstream analysis. Analysis of the Calling Cards library revealed that a total of 205,937 insertions were recovered across all cells, with a relatively balanced distribution by genotype: 97,944 insertions in WT samples and 107,993 in Het samples (**Figure 39C**). Notably, 87% of the nuclei (84,285 out of 96,505) were represented in this dataset, indicating that the majority contained Calling Cards insertions (**Figure 39D, E**). Further analysis focused on the distribution of insertions among different cell types. Excitatory neurons exhibited the highest number of insertions, with inhibitory neurons following (**Figure 40A**). In contrast, non-neuronal cells such as astrocytes, microglia, and oligodendrocytes displayed fewer insertions, which aligned with the patterns of tdTomato expression (**Figure 40B**). This discrepancy might be attributed to the tropism of the AAV9 capsids for neurons promoting robust transgene expression in these cells, or potentially to the inherently lower RNA content in glial cells, which could reduce the detection sensitivity of Calling Cards SRTs. While the recovery of cell barcodes was substantial, the average number of insertions per nucleus was low, typically ranging from 1 to 2 (**Figure 40C**). This sparsity of insertions per nucleus poses challenges for reliable peak calling and impedes the feasibility of nuanced differential analyses, particularly at the cell type level. Although the depth of the current dataset constrains robust comparative analyses between genotypes per cluster, aggregating the data enables preliminary comparative insights, particularly among excitatory neurons.

To identify genomic regions that had high insertion densities of Calling Cards, a joint set of 4,057 regions from WT and Het samples were called using CCcaller, an optimized peak calling algorithm within Pycallingcards (**Figure 41A**) (Guo et al., 2024). Following this, insertions from each genotype were intersected with these joint peaks to count the number of insertions within each peak region. I then sought to analyze transcription factors, given that the Calling Cards were targeted to BRD4 binding sites representative of enhancer regions. The analysis identified 16 transcription factors (TFs) with a higher density of insertions in WT samples and 28 TFs with increased insertions in Het samples when compared to WT (**Figure 41B-D**). Interestingly, seven of the TFs—*PBX1*, *ZFPM2*, *MEIS2*, *THRB*, *KCTD1*, *HDAC9*, and *MEF2C*—are recognized as direct targets of MYT1L, determined by CUT&RUN, and exhibited differential Calling Cards insertion patterns. This data, while indicative, remains preliminary. A larger dataset containing more insertions will be necessary to perform a more comprehensive differential analysis and to discern more subtle distinctions.

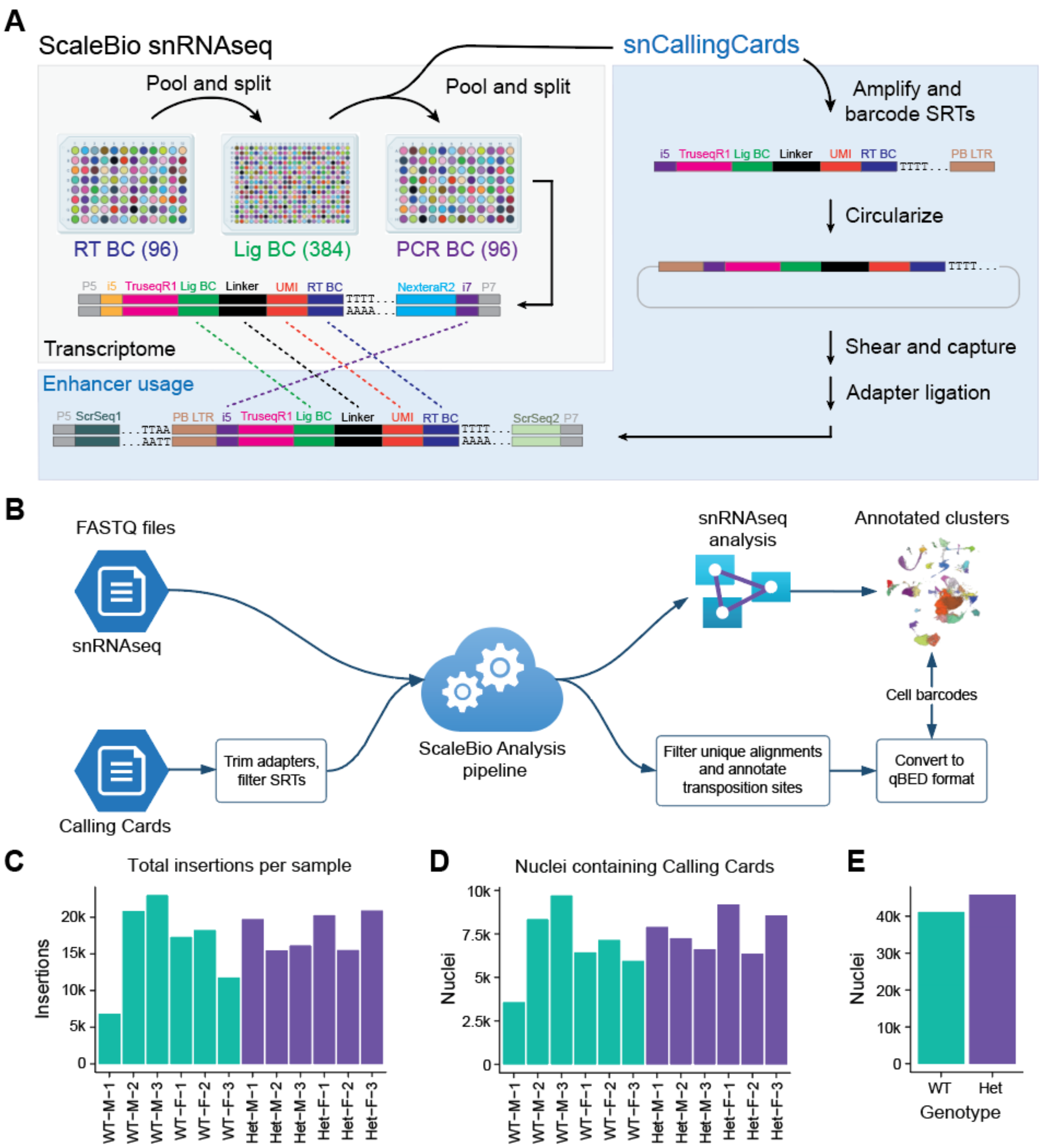


Figure 39: Integration of Calling Cards with the ScaleBio platform

(A) Schematic illustrating the general workflow beginning with the ScaleBio combinatorial barcoding. After the second round of barcoding, half of the products are split towards the snCallingCards protocol. Here, the SRTs are amplified, barcoded, circularized, then ligated with Illumina adapters. After sequencing, the barcode combinations are matched from the two libraries to associate Calling Cards insertions to specific cell types. **(B)** A diagram of the computational workflow to analyze the multimodal data. The snRNAseq library is analyzed using the ScaleBio RNA analysis pipeline and processed for downstream analysis. The Calling Cards libraries are first trimmed and any reads that are not Calling Cards SRTs are filtered out. Then the ScaleBio analysis pipeline is used to align the reads to the reference genome and demultiplex the barcodes. Finally, the insertion sites are annotated and output in the qBED format. The cell barcodes that are common between the qBED and snRNAseq datasets can be selected for further downstream analysis. **(C)** Bar graph displays the count of Calling Cards insertions identified in each sample. **(D)** Bar graph indicates the number of nuclei per sample found to contain Calling Cards insertions. **(E)** Summary graph presents the aggregated number of nuclei containing Calling Cards insertions.

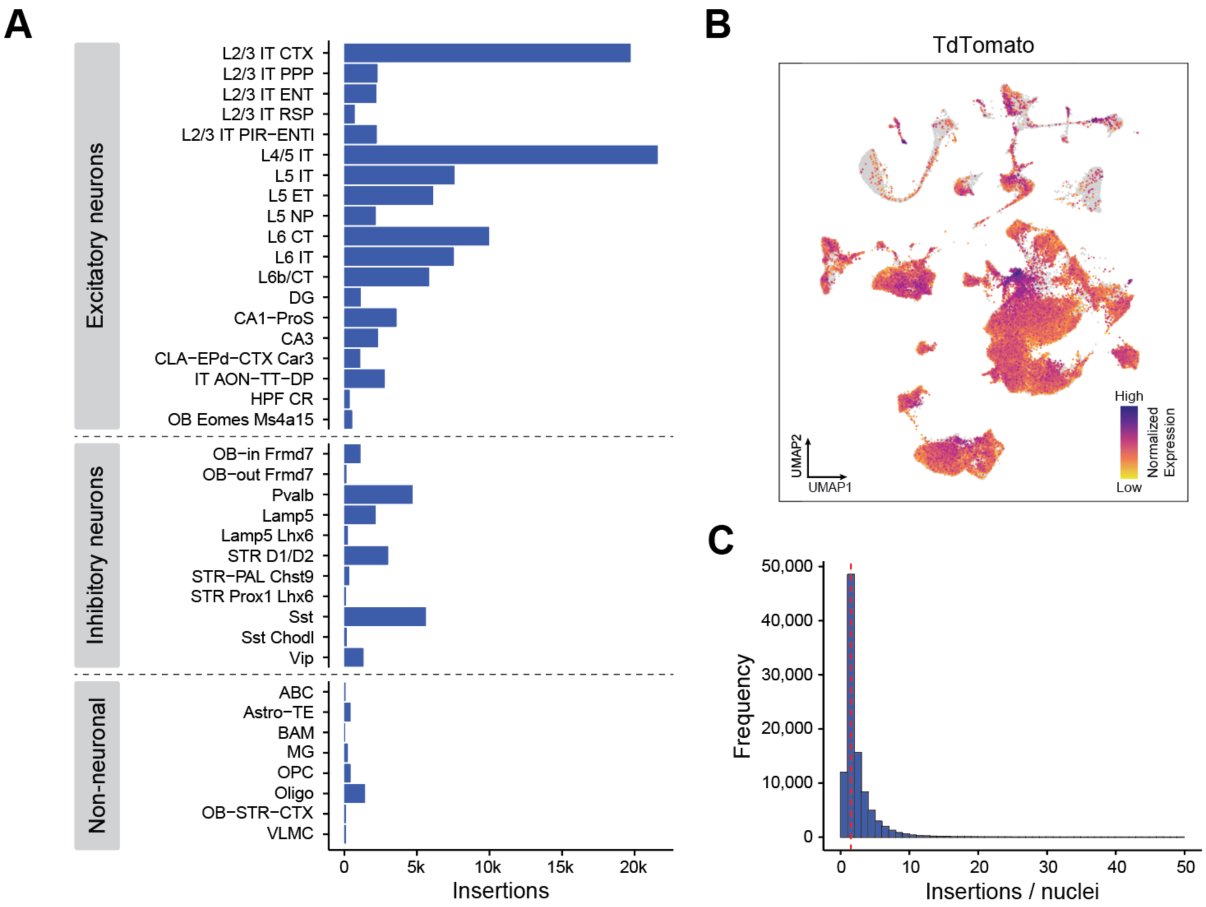


Figure 40: Calling Cards insertions and tdTomato expression across cell types

(A) Bar plots shows the total number of Calling Cards insertions recovered for each annotated cell cluster for MYT1L WT and Het genotypes. **(B)** UMAP plot of all nuclei overlaid with a color scale showing the normalized expression of TdTomato from the SRT. **(C)** Histogram of the number of recovered insertions per nuclei. The red line indicates that the median number of insertions per nuclei was 1.

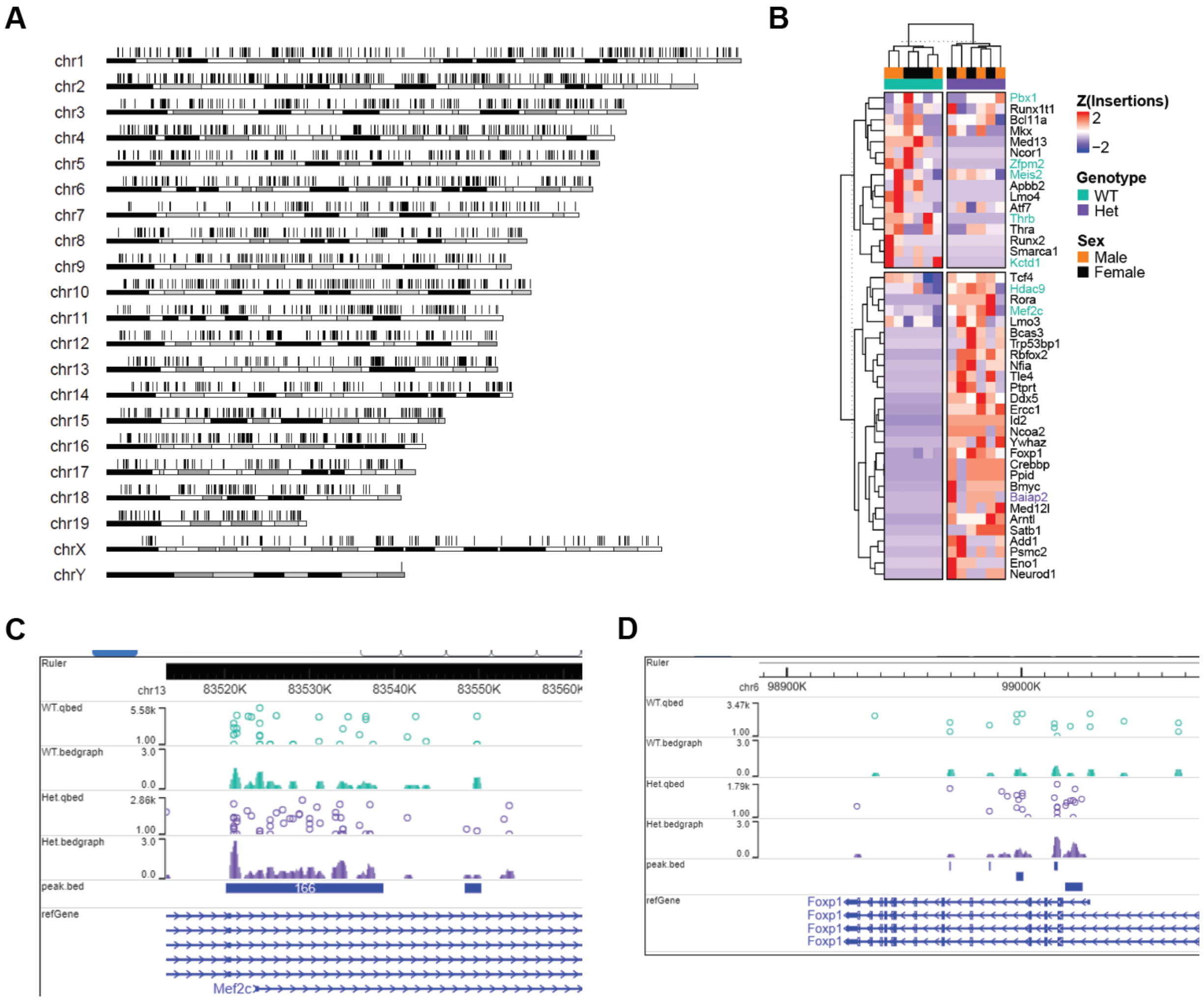


Figure 41: Analysis of Calling Cards insertions in excitatory neurons

(A) Genomic regions enriched with Calling Cards insertions from excitatory neuron clusters across all chromosomes. **(B)** Heatmap showing transcription factors that showed differential enrichment of Calling Cards insertions within peaks. **(C, D)** Screenshots from the WashU Epigenome Browser showing Calling Cards tracks and called peak regions for the gene Mef2c in **C** and Foxp1 in **D**. Each open circle point represents a unique insertion that was recovered and the computed density plot in the track below. The WT tracks are shown in turquoise, and the Het tracks are shown in purple.

6.4 Discussion

By integrating Calling Cards technology with the ScaleBio platform, I have significantly enhanced sequencing efficiency, which has facilitated the initial identification of differential BRD4 enhancer usage between MYT1L WT and Het excitatory neurons. This has the potential to uncover novel insights into the regulatory dynamics of gene expression. The ScaleBio platform allows for the analysis of hundreds of thousands of cells, making it possible to include multiple biological replicates and apply more rigorous statistical analyses to detect subtle changes. This framework can enable future research aimed at exploring the complex interplay between genomic architecture and cellular function.

One of the limitations of single-cell/nucleus Calling Cards is the sparsity of the collected data. For instance, when viral Calling Cards reagents were applied to the postnatal mouse cortex, 111,382 insertions were recovered from 35,950 cells, averaging 3.1 insertions per cell, with 73.7% of the cells with at least one insertion (Moudgil et al., 2020b). By contrast, in the same study using K562 cells—a human immortalized lymphoblast cell line—transfected by electroporation with Calling Cards plasmid reagents, 327,465 insertions were detected among 21,554 cells, with an average of 15.3 insertions per cell and 95.8% of cells containing at least one insertion. Given the comparable sequencing depth per insertion between the mouse cortex (109.6) and K562 cells (137.0), it appears that the delivery method of the transgene can be a factor that can significantly impact the number of insertions per cell. Specifically, in vitro electroporation tends to be highly efficient, introducing multiple vector copies into each cell, whereas intracerebroventricular injection using AAV vectors achieves widespread cortical labeling but only a few viral particles infect each cell. This results in a broad but sparse expression per cell due to the limited copies of the SRT, thereby yielding a low overall number of insertions per cell.

If SRTs are the limiting reagent, several strategies can be employed to increase the number of insertions per cell. One approach involves increasing the SRT:*piggyBac* transposase ratio, thereby amplifying the available SRT for transposition. Preliminary in vitro experiments suggest that this adjustment enhances the number of insertions per cell. It appears that only a minimal amount of transposase is required to catalyze transposition, implying that an abundance of SRTs could promote increased numbers of insertions. Another strategy might focus on delivering Calling Cards to specific tissue regions rather than aiming for broad coverage. For instance, stereotaxic injections to deliver Calling Cards reagents directly into a targeted brain region could result in a greater concentration of reagents per cell, as opposed to the more diluted effects observed with widespread but less focused injection routes.

The integration of Calling Cards with combinatorial indexing is a start to advancing our capability to associate enhancer usage and gene expression at a single cell level, but also at a scale that enables the exploration of cellular heterogeneities and molecular pathways that define neurodevelopmental processes. Future iterations of this technology, couple with refined analytical strategies, not only stand to reveal the underpinnings of cellular identity and function, potentially offer insights into the genetic and epigenetic mechanisms of biological complexity and disease pathology.

6.5 Materials and methods

As this is an ongoing collaboration with Scale Biosciences, the specific details of reagents, primer sequences, thermocycler conditions, and sequencing parameters will be described in an upcoming manuscript or whitepaper.

Animals

All animal studies were approved by and performed in accordance with the guidelines of the Animal Care and Use Committee of Washington University in Saint Louis, School of Medicine and conform to NIH guidelines of the care and use of laboratory animals. Details on animal husbandry and maintenance of the MYT1L line is described in detail in Chapter 5.6.

Generation of AAV9 viral particles

Endotoxin-free plasmid preparations of pAAV-hyPB and pAAV-H2B-TdT-SRT (Addgene 203393) were done using the ZymoPURE II Plasmid Maxiprep kit (Zymo D4202). These constructs were packaged into AAV9 viral particles by Virovek using their recombinant baculovirus production protocol with Sf9 insect cells. The AAV9 particles were purified, the titer was determined by qPCR, and standardized to $1x10^{13}$ vg/ml. Endotoxin levels were assessed using the Endosafe nextgen-PTS (Charles River) Assay.

Intracerebroventricular injection of AAV-CallingCards

The AAV9-hyPB and AAV9-H2B-TdT-SRT vectors were mixed 1:1 and injected into the ventricles of P0-1 pups from MYT1L WT x Het breeding pairs as described in (Yen et al., 2023). At P5, toe tissue was collected for identification and genotyping. At P21, the animals were deeply anesthetized with isoflurane and perfused with ice-cold DPBS. The brain was harvested, TdTomato fluorescence was verified using a handheld fluorescence flashlight (Nightsea Xite-GR), the cortex was dissected, and the tissue was flash frozen in liquid nitrogen and stored at -80°C. The tissue was processed following the “Nuclei isolation and fixation” section as described in Chapter 5.6.

Single-nucleus Calling Cards library preparation

The preparation of snCC libraries began with the standard ScaleBio snRNAseq library protocol, which is detailed in **Figure 39A** and described in Chapter 5.6. Following the ScaleBio protocol’s cleanup step, half the volume (3 µl) of each well of the Final Distribution Plate was transferred into a 96-well plate, resulting in a duplicate Final Distribution Plate for Calling Cards. This plate was used as the template for SRT amplification and barcoding. The PCR products were then circularized, sheared, and indexed using a strategy based on the protocol that was published in (Moudgil et al., 2020b). The libraries were sequenced on an Illumina Nextseq platform with 50% PhiX or balanced Nextera libraries.

6.6 Acknowledgements

I thank members of the Dougherty and Mitra laboratories for helpful discussions and feedback; M. Crosby and J. Hoisington-Lopez at the DNA Sequencing and Innovation Lab (DSIL) and the Genome Technology Access Center at the McDonnell Genome Institute (GTAC@MGI) for their sequencing expertise and services; and D. Skinner, F. Leti, and M. Nakamoto for their support. This work was funded by grants from the National Institute of Mental Health (RF1MH117070, RF1MH126723, and R01MH124808 to J.D.D. and R.D.M.). A.Y. was supported in part by T32HG000045 from the National Human Genome Research Institute.

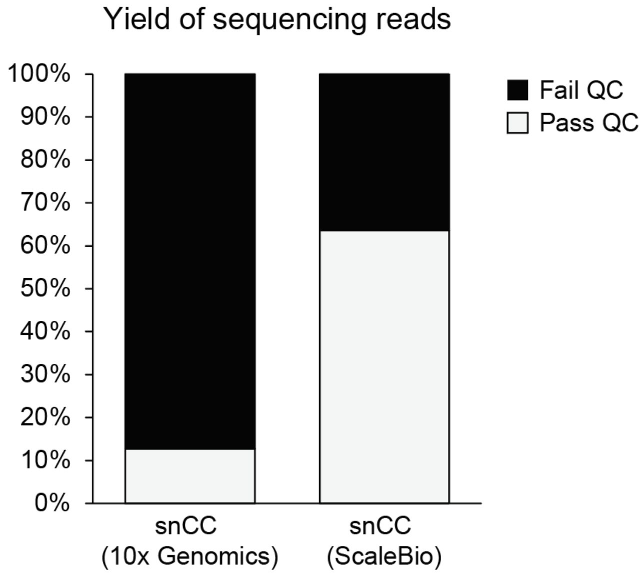
6.7 Author contributions

Project conceptualization: A.Y., R.D.M., and J.D.D. Method development, experiments, and data collection: A.Y., X.C., M.C., and J.H-L. Formal analysis: A.Y., R.D.M., and J.D.D. Figures and data visualization: A.Y., R.D.M., and J.D.D. Project coordination: A.Y., R.D.M., and J.D.D. Funding acquisition: R.D.M. and J.D.D.

6.8 Disclosures

F.L. and M.N. are employees of Scale Biosciences. R.D.M has filed a patent application for self-reporting transposon (SRT) technology.

6.9 Supplemental figures and tables



Supplemental Figure 14: Improved sequencing efficiency and output

Stacked bar plot showing that the overall yield of passing reads from the ScaleBio snCC library preparation and sequencing was increased over five-fold compared to the 10x Genomics approach.

Chapter 7: Conclusions and future directions

7.1 Summary of the dissertation

This dissertation comprises projects focused on two main themes: developing technology to enable novel analyses beyond the capabilities of existing tools and utilizing these innovations to explore epigenetic processes in neurodevelopment.

Chapter 2 outlines the enhancements made to the bulk Calling Cards protocol, including library preparation and sequencing optimizations, the introduction of barcoded and nuclear SRTs, and a computational workflow to streamline Calling Cards data analysis. These improvements aim to broaden the technology's application across various research domains, extending beyond genomics specialists.

In Chapter 3, I detail our efforts to create transgenic Calling Cards mouse lines, aiming to expand the technology's reach by facilitating studies on early development and specific cell populations with spatial and temporal control using Cre recombinase. Despite successfully creating these mice, the Calling Cards did not function as anticipated. Our findings indicate potential sensitivity of certain cell populations to Calling Cards insertions and suggest that endogenous defense against transposons may hinder insertions during early development. Nevertheless, these results provided valuable insights for refining the technology.

Chapter 4 demonstrates the optimized bulk Calling Cards protocol's effectiveness through a study examining enhancer activity during brain masculinization triggered by perinatal testosterone. I discovered that certain sex-differential enhancers were associated with transposable elements like LINEs and Alu SINEs, and correlate with key autism-related genes. This suggests a potential role for these elements in sex-specific gene expression and sex differences of autism prevalence.

Chapter 5 and 6 discuss the development and application of a single-nucleus Calling Cards, particularly in the context of neurodevelopmental studies using the MYT1L mouse model. Chapter 5 presents findings on how MYT1L loss affects excitatory neuron development, while Chapter 6 explores methodological advances to integrate combinatorial barcoding with Calling Cards to analyze larger cell populations, demonstrating the technique's feasibility.

Overall, this dissertation documents the advancement of Calling Cards technology and its application to neurodevelopmental research, highlighting the potential of these tools in uncovering new insights into complex biological processes. The enhancements and applications described enable future explorations into the genetic and epigenetic frameworks that influence neurodevelopment, and demonstrate the technology's broader applicability across various scientific disciplines, enriching a wide array of research fields and fostering innovation.

7.2 Significance, innovation, and future directions

The complete experimentation guide for bulk Calling Cards

This guide aims to make Calling Cards technology accessible to researchers from various fields, regardless of their experience level with genomic tools. Calling Cards is a powerful technique offering longitudinal insights into protein-DNA interactions, which is not possible with conventional techniques. This capability is particularly valuable for tracing events where the initiating molecular interaction and its biological outcome are temporally separate. For instance, the activity of lineage-specifying transcription factors (TFs) and the manifestation of related cell fates can significantly vary across cell types, an area where existing snapshot approaches may not suffice.

Despite its strengths, Calling Cards has limitations, such as the need for exogenous expression in target cells or tissues, which can lead to overexpression of the TF and potentially disrupt natural TF binding or modify gene pathways, introducing artificial changes. To mitigate these effects, the guide emphasizes the need for proper controls and offers comprehensive advice on experimental design, implementation, and troubleshooting, along with general protocols for sample handling and analysis that should align with many, albeit not all, sample types. We suggest optimization strategies for exceptional cases.

The RNA-based detection of SRTs enhances the sensitivity of recovering Calling Cards insertions but also introduces certain limitations. The transient nature of RNA means that sustained transcription is necessary to recover the insertions, particularly if insertions in initially open chromatin regions become inaccessible over time due to changes in chromatin state or methylation-induced silencing. A proposed future direction is to employ the bacteriophage T7 promoter to drive SRT expression. Using this approach, the SRT insertions will not be actively transcribed in the living cells and would thus be “silent” insertions. To collect the sample for Calling Cards, genomic DNA can be harvested and used as the template for in vitro transcription with T7 RNA Polymerase. This method can enable the generation of complete RNA transcripts from all insertions, regardless of RNA stability or chromatin state, thus offering a more comprehensive recovery of insertions.

Additionally, this guide highlights cost-reducing technological advancements to Calling Cards experiments, such as barcode integration in SRTs, which streamlines sample handling and reduces the use of reagents. The development of nuclear SRT permits tissue preservation for later processing, and we introduce a portable and reproducible software pipeline to process Calling Cards data into a format that is compatible with many standard packages for downstream analyses.

The creation and optimization of new methodologies like Calling Cards is exciting, particularly when witnessing the inventive applications and discoveries they enable in the scientific community. It is my hope that this guide will empower researchers to harness this technology creatively, unlocking novel insights across diverse scientific realms.

Design, development, and validation of transgenic Calling Cards mice

Currently, the delivery of Calling Cards to target cells or tissues is achieved through electroporation with plasmid constructs or by using adeno-associated viruses (AAVs) for tissue transduction. The latter meth's efficacy is particularly influenced by the injection's location and timing, as minor variations can affect the cell populations being targeted. This delivery method is less suitable for studying early developmental stages, which are crucial for understanding cell fate determination.

To overcome this limitation, we sought to develop transgenic Calling Cards mice. In these mice, the *piggyBac* transposase and SRT components are integrated into the genome, allowing us to target specific cell populations at desired times and locations using Cre recombinase. While the single and double transgenic combinations of transposase, SRT, and Cre recombinase were successful, the triple transgenic mice did not function as expected. I have several hypotheses for why the transgenic Calling Cards approach faced challenges. One possibility is that the SRT elements are being silenced by the cells' defense mechanisms against perceived viral threats, leading to DNA methylation or even cell apoptosis. Another concern is the potential for the sizable SRT insertions (~2kb) to interfere with critical enhancer regions, affecting gene regulation or genome stability. Finally, our observations suggest that stem cells and progenitor cells might be more susceptible to disruptions from Calling Cards insertions than mature neurons are.

Despite these setbacks, such failed experiments are invaluable learning opportunities. They provide critical insights and form the foundation for refining our methods. The knowledge gained from these trials is guiding our strategies for future iterations, illustrating the iterative nature of scientific discovery. The T7 promoter driving SRT expression is a promising strategy to circumvent issues associated with insertional activation leading to potential silencing. The T7-SRTs are considered “silent” within the genome, which might allow them to avoid triggering endogenous silencing mechanisms. Moreover, the T7-SRT construct is more compact at 670bp, which, though may still be disruptive if directly inserted into an enhancer region, is smaller and can be less perturbing, reducing the risk of genomic instability and cell death. The upcoming analysis of preliminary data from using the T7 promoter in various cells and tissues will be crucial to evaluate the viability of this approach.

Gene regulatory loci in transposable elements

In this study, I delved into the impact of sex differences on brain development, emphasizing the role of Brd4-bound enhancers in directing sex-specific gene expression. Employing Calling Cards technology, I profiled enhancer usage during pivotal neurodevelopmental stages, identifying unique patterns that potentially drive sex-based differences in brain maturation and functionality. My analysis spanned three crucial developmental periods—before, during, and after the perinatal testosterone surge—identifying shifts in enhancer activity that illuminate the formative effects of early hormonal influences and the lasting imprints of these epigenetic modifications.

By examining the associations of enhancers and transposable elements (TEs), I found that these TEs can be versatile regulatory entities that shape sex-specific gene expression. The discovery of distinct enhancer activities and associated transcription factor motifs provides fresh perspectives on the genomic architecture of sex differences in brain development. The observed correlation between TE-mediated gene regulation and autism spectrum disorder-related genes offers an avenue to investigate the genetic basis behind the male prevalence in autism.

The findings of this study affirm the critical influence of hormonal and genetic determinants in neural sexual differentiation and broaden our grasp of the intricate genetic and epigenetic networks underpinning sex-distinct neural pathways. By crafting an extensive inventory of enhancer usage and pinpointing crucial regulatory elements linked to sex-biased gene expression, this data can be a resource for future investigations into the intricacies of brain development and the origins of sex-specific behavioral patterns and neuropsychiatric conditions.

MYT1L deficiency impairs excitatory neurons trajectory during cortical development

In this study, we undertook an expansive transcriptomic analysis of over 300,000 nuclei to elucidate the impact of MYT1L mutation on neurodevelopment, representing a significant leap in our comprehension of transcription factor roles in neuronal maturation. This work also showcases the application of single-cell technologies beyond cell type cataloging to make atlases, using them to define the fundamental phenotype of a disorder. By integrating sophisticated analytical tools like pseudotime and RNA velocity, our research establishes a new benchmark in quantitatively assessing the evolution of neuronal transcriptional profiles.

The innovation of our approach lies in its scale and depth, leveraging cell taxonomies from single-cell atlases to provide a nuanced understanding of MYT1L's specific influence on deep layer excitatory neurons, which exhibit a notably immature transcriptomic state due to MYT1L haploinsufficiency. This nuanced analysis has uncovered the critical nature of MYT1L in orchestrating the complex symphony of gene expression programs that guide neuronal maturation, emphasizing the sensitivity of neurodevelopmental processes to even minor fluctuations in key transcriptional regulators.

Our findings not only pinpoint MYT1L's repressive action on genes steering axon guidance, neuron migration, and fate determination but also reveal its less stringent control over synaptic function genes, demonstrating a sophisticated balance of gene regulation crucial for healthy neurodevelopment. This work also advances our understanding of the intricate relationships between transcriptional regulation and epigenetic mechanisms, proposing that MYT1L's impact extends through both direct and indirect pathways that evolve across developmental stages.

This study enhances our fundamental knowledge of neuronal development, offers potential pathways for investigating neurodevelopmental disorders, and provides a robust methodological framework for future research in genomic medicine. By detailing the specific and stage-dependent roles of MYT1L, this research paves the way for targeted therapeutic strategies and enriches our toolkit for deciphering the molecular intricacies of brain development.

Single-nucleus Calling Cards with combinatorial indexing

Integration of combinatorial indexing with Calling Cards technology has resulted in significantly improved sequencing efficiency as well as a cost-effective strategy to scale the number of assayed cells or nuclei. The goal of this pilot was to demonstrate feasibility of the method and to identify if loss of MYT1L can impact BRD4 enhancer usage and potential insights into the regulatory mechanisms governing the altered neuronal maturation trajectory. The combinatorial barcoding permits the inclusion of multiple biological replicates, allowing for more robust statistical analyses that can identify nuanced changes, thus setting a new standard for understanding the genomic underpinnings of cellular functions.

A notable challenge with single-cell/nucleus Calling Cards is data sparsity. For example, Moudgil et al. (2020b) demonstrated initial proof-of-principle by using postnatal mouse cortex which yielded an average of 3.1 insertions per cell, and an experiment with K562 cells demonstrated a much higher efficiency, averaging 15.3 insertions per cell. This discrepancy underscores the influence of transgene delivery methods on insertion density. Electroporation in vitro introduces multiple vector copies per cell, whereas viral vector delivery in vivo potentially leads to sparser data due to fewer vector copies per cell.

To enhance our ability to detect regions with differential insertion densities, increasing the total number of insertions per cell type is critical. Several strategies could be explored to achieve this. First, adjusting the transposase to SRT ratio might increase insertion numbers, given that the SRT is thought to be the limiting reagent. Additionally, focusing on targeted delivery of Calling Cards, such as using stereotaxic injections to administer reagents directly into specific brain regions, might improve reagent concentration within targeted cells compared to more generalized, dispersed delivery methods. Should these methods still result in low insertion rates per cell, increasing the number of analyzed cells could serve as a workaround, especially as the ability to scale the combinatorial barcoding is feasible. This brute-force approach could leverage economies of scale to compensate for data sparsity, potentially enabling more robust statistical analyses and insights.

Building Calling Cards on Scale Bioscience’s combinatorial indexing platform marks a significant stride forward, enriching our capacity to assay many more cells in a cost-effective manner. This will allow us to link enhancer activity with gene expression on a single-cell scale and across a broad population. This approach is instrumental in dissecting cellular heterogeneity and the molecular networks that regulate neurodevelopmental pathways. Future enhancements in this technology, paired with sophisticated analytical methods, are poised to uncover the intricacies of cellular identity and functioning, offering profound insights into the genetic and epigenetic dynamics that underlie biological complexity and disease.

7.3 Final thoughts

In this dissertation, I harnessed Calling Cards technology to illuminate the nuanced mechanisms of gene regulation that underpin brain development and its disorders. This research extends beyond traditional genomic methodologies, offering a novel perspective of retrospective analysis of retrospective analysis of gene regulatory elements and their long-term impacts on cellular fate and function, particularly in the context of neurodevelopmental disorders.

My journey into this field is fueled by a fascination with the brain's cellular complexity and a passion for technology development. Brain development and its dysfunctions are still largely not understood. The necessity to bridge the gap in our understanding of how transient molecular interactions influence cellular outcomes has driven this work, underscoring its potential to redefine our approach to studying cellular phenotypes and their evolution. Through the development and application of Calling Cards, this research not only advances our methodological toolkit but also enriches our conceptual grasp of cellular differentiation and disease pathology. The technique's deployment in analyzing brain masculinization and its pertinence to neurodevelopmental conditions exemplifies its potential, offering fresh vistas into the epigenetic narratives that shape our neural architecture and its variances.

Looking ahead, I envision the methodologies described here to catalyze further explorations into the genetic and epigenetic mechanisms of biological processes and disease, not just those relating to neurodevelopment. The groundwork laid through this dissertation is poised to stimulate a broader adoption and adaptation of Calling Cards technology, enabling researchers to investigate enhancers that dictate cellular identity and state across diverse biological landscapes.

In conclusion, this body of work represents a significant milestone in my academic journey and contributes to the fields of genomic science and neurodevelopment. By presenting new findings and innovative tools, I hope that this work will inform future research into the complex intricacies of biological systems. My aspiration is that elements of my dissertation will not only facilitate but also inspire further explorations into the molecular mechanisms that underlie biological functions and processes.

References

Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, Menashe I, Wadkins T, Banerjee-Basu S, Packer A (2013) SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol Autism 4:36.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Alexopoulou AN, Couchman JR, Whiteford JR (2008) The CMV early enhancer/chicken β actin (CAG) promoter can be used to drive transgene expression during the differentiation of murine embryonic stem cells into vascular progenitors. BMC Cell Biol 9:2.

Allard CB (2012) Dorsolateral Prefrontal Cortex Activation During Emotional Anticipation and Neuropsychological Performance in Posttraumatic Stress Disorder. Arch Gen Psychiatry 69:360.

Almazan G, Lefebvre DL, Zingg HH (1989) Ontogeny of hypothalamic vasopressin, oxytocin and somatostatin gene expression. Dev Brain Res 45:69–75.

Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet 23:185–188.

Aschauer DF, Kreuz S, Rumpel S (2013) Analysis of Transduction Efficiency, Tropism and Axonal Transport of AAV Serotypes 1, 2, 5, 6, 8 and 9 in the Mouse Brain. PLoS ONE 8:e76310.

Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium et al. (2019) Identification of common genetic risk variants for autism spectrum disorder. Nat Genet 51:431–444.

Azevedo FAC, Carvalho LRB, Grinberg LT, Farfel JM, Ferretti REL, Leite REP, Filho WJ, Lent R, Herculano-Houzel S (2009) Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. J Comp Neurol 513:532–541.

Babaian A, Mager DL (2016) Endogenous retroviral promoter exaptation in human cancer. Mob DNA 7:24.

Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME Suite. Nucleic Acids Res 43:W39–W49.

Bakken TE et al. (2018) Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. PLOS ONE 13:e0209648.

Banerjee-Basu S, Packer A (2010) SFARI Gene: an evolving database for the autism research community. Dis Model Mech 3:133–135.

Bartosovic M, Kabbe M, Castelo-Branco G (2021) Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. Nat Biotechnol 39:825–835.

Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ (2020) Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol 38:1408–1414.

Berry GE, Asokan A (2016) Cellular transduction mechanisms of adeno-associated viral vectors. Curr Opin Virol 21:54–60.

Betschinger J, Knoblich JA (2004) Dare to Be Different: Asymmetric Cell Division in Drosophila, C. elegans and Vertebrates. Curr Biol 14:R674–R685.

Bhagwat AS, Roe J-S, Mok BYL, Hohmann AF, Shi J, Vakoc CR (2016) BET Bromodomain Inhibition Releases the Mediator Complex from Select cis -Regulatory Elements. Cell Rep 15:519–530.

Blanchet P et al. (2017) MYT1L mutations cause intellectual disability and variable obesity by dysregulating gene expression and development of the neuroendocrine hypothalamus. PLOS Genet 13:e1006957.

Bramble MS, Roach L, Lipson A, Vashist N, Eskin A, Ngun T, Gosschalk JE, Klein S, Barseghyan H, Arboleda VA, Vilain E (2016) Sex-Specific Effects of Testosterone on the Sexually Dimorphic Transcriptome and Epigenome of Embryonic Neural Stem/Progenitor Cells. Sci Rep 6:36916.

Cammack AJ, Moudgil A, Chen J, Vasek MJ, Shabsovich M, McCullough K, Yen A, Lagunas T, Maloney SE, He J, Chen X, Hooda M, Wilkinson MN, Miller TM, Mitra RD, Dougherty JD (2020) A viral toolkit for recording transcription factor–DNA interactions in live mouse tissues. Proc Natl Acad Sci 117:10003–10014.

Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, Daza RM, McFaline-Figueroa JL, Packer JS, Christiansen L, Steemers FJ, Adey AC, Trapnell C, Shendure J (2018) Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science 361:1380–1385.

Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, Adey A, Waterston RH, Trapnell C, Shendure J (2017) Comprehensive single-cell transcriptional profiling of a multicellular organism. Science 357:661–667.

Challis RC, Ravindra Kumar S, Chan KY, Challis C, Beadle K, Jang MJ, Kim HM, Rajendran PS, Tompkins JD, Shivkumar K, Deverman BE, Gradinaru V (2019) Systemic AAV vectors for widespread and targeted gene delivery in rodents. Nat Protoc 14:379–414.

Chan MM, Smith ZD, Grosswendt S, Kretzmer H, Norman TM, Adamson B, Jost M, Quinn JJ, Yang D, Jones MG, Khodaverdian A, Yosef N, Meissner A, Weissman JS (2019) Molecular recording of mammalian embryogenesis. Nature 570:77–82.

Chatterjee S, Ahituv N (2017) Gene Regulatory Elements, Major Drivers of Human Disease. Annu Rev Genomics Hum Genet 18:45–63.

Chen J, Fuhler NA, Noguchi KK, Dougherty JD (2023) MYT1L is required for suppressing earlier neuronal development programs in the adult mouse brain. Genome Res genome;gr.277413.122v1.

Chen J, Lambo ME, Ge X, Dearborn JT, Liu Y, McCullough KB, Swift RG, Tabachnick DR, Tian L, Noguchi K, Garbow JR, Constantino JN, Gabel HW, Hengen KB, Maloney SE, Dougherty JD (2021) A MYT1L syndrome mouse model recapitulates patient phenotypes and reveals altered brain development due to disrupted neuronal maturation. Neuron S0896627321006814.

Chen S, Lake BB, Zhang K (2019) High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol 37:1452–1457.

Clevers H (2005) Stem cells, asymmetric division and cancer. Nat Genet 37:1027–1028.

Coursimault J et al. (2021) MYT1L-associated neurodevelopmental disorder: description of 40 new cases and literature review of clinical and molecular aspects. Hum Genet.

Crews D (2011) Epigenetic modifications of brain and behavior: Theory and practice. Horm Behav 59:393–398.

Crump NT, Ballabio E, Godfrey L, Thorne R, Repapi E, Kerry J, Tapia M, Hua P, Lagerholm C, Filippakopoulos P, Davies JOJ, Milne TA (2021) BET inhibition disrupts transcription but retains enhancer-promoter contact. Nat Commun 12:223.

Czech B, Hannon GJ (2016) One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends Biochem Sci 41:324–337.

Davey RA, Grossmann M (2016) Androgen Receptor Structure, Function and Biology: From Bench to Bedside. Clin Biochem Rev 37:3–15.

de Ligt J, Willemsen MH, van Bon BWM, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, del Rosario M, Hoischen A, Scheffer H, de Vries BBA, Brunner HG, Veltman JA, Vissers LELM (2012) Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. N Engl J Med 367:1921–1929.

Deans C, Maggert KA (2015) What Do You Mean, “Epigenetic”? Genetics 199:887–896.

Deegan DF, Karbalaei R, Madzo J, Kulathinal RJ, Engel N (2019) The developmental origins of sex-biased expression in cardiac development. Biol Sex Differ 10:46.

Deliu E, Arecco N, Morandell J, Dotter CP, Contreras X, Girardot C, Käsper E-L, Kozlova A, Kishi K, Chiaradia I, Noh K-M, Novarino G (2018) Haploinsufficiency of the intellectual disability gene SETD5 disturbs developmental gene expression and cognition. Nat Neurosci 21:1717–1727.

Dey A, Chitsaz F, Abbasi A, Misteli T, Ozato K (2003) The double bromodomain protein Brd4 binds to acetylated chromatin during interphase and mitosis. Proc Natl Acad Sci 100:8758–8763.

Dhakal P, Kelleher AM, Behura SK, Spencer TE (2020) Sexually dimorphic effects of forkhead box a2 (FOXA2) and uterine glands on decidualization and fetoplacental development. Proc Natl Acad Sci 117:23952–23959.

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 (2021) Molecular logic of cellular diversification in the mouse cerebral cortex. Nature 595:554–559.

Di Nisio E, Lupo G, Licursi V, Negri R (2021) The Role of Histone Lysine Methylation in the Response of Mammalian Cells to Ionizing Radiation. Front Genet 12:639602.

Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI, Zhao K, Young RA (2014) Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. Cell 159:374–387.

Doyle GA, Crist RC, Karatas ET, Hammond MJ, Ewing AD, Ferraro TN, Hahn C-G, Berrettini WH (2017) Analysis of LINE-1 Elements in DNA from Postmortem Brains of Individuals with Schizophrenia. Neuropsychopharmacology 42:2602–2611.

Ernst C, Odom DT, Kutter C (2017) The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. Nat Commun 8:1411.

Ewels PA, Peltzer A, Filinger S, Alneberg J, Wilm A, Garcia MU, Tommaso PD, Nahnsen S (2020) The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol 38:271–271.

Fass SB, Mulvey B, Yang W, Selmanovic D, Chaturvedi S, Tycksen E, Weiss LA, Dougherty JD (2023) Relationship between sex biases in gene expression and sex biases in autism and Alzheimer’s disease. medRxiv.

Filippakopoulos P et al. (2010) Selective inhibition of BET bromodomains. Nature 468:1067–1073.

Fu JM et al. (2022) Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat Genet 54:1320–1331.

Gandal MJ et al. (2022) Broad transcriptomic dysregulation occurs across the cerebral cortex in ASD. Nature 611:532–539.

Gegenhuber B, Wu MV, Bronstein R, Tollkuhn J (2022) Gene regulation by gonadal hormone receptors underlies brain sex differences. Nature 606:153–159.

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, Smith SM, Van Essen DC (2016) A multi-modal parcellation of human cerebral cortex. Nature 536:171–178.

Goldstein RZ, Volkow ND (2011) Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. Nat Rev Neurosci 12:652–669.

Gray SJ, Choi VW, Asokan A, Haberman RA, McCown TJ, Samulski RJ (2011) Production of Recombinant Adeno-Associated Viral Vectors and Use in In Vitro and In Vivo Administration. Curr Protoc Neurosci 57.

Grimm S, Beck J, Schuepbach D, Hell D, Boesiger P, Bermpohl F, Niehaus L, Boeker H, Northoff G (2008) Imbalance between Left and Right Dorsolateral Prefrontal Cortex in Major Depression Is Linked to Negative Emotional Judgment: An fMRI Study in Severe Major Depressive Disorder. Biol Psychiatry 63:369–376.

Guo H, Zhu P, Wu X, Li X, Wen L, Tang F (2013) Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res 23:2126–2135.

Guo J, Zhang W, Chen X, Yen A, Chen L, Shively CA, Li D, Wang T, Dougherty JD, Mitra RD (2024) Pycallingcards: an integrated environment for visualizing, analyzing, and interpreting Calling Cards data. Bioinformatics 40:btae070.

Han D, Liu G, Oh Y, Oh S, Yang S, Mandjikian L, Rani N, Almeida MC, Kosik KS, Jang J (2023) ZBTB12 is a molecular barrier to dedifferentiation in human pluripotent stem cells. Nat Commun 14:632.

Heiman M, Kulicke R, Fenster RJ, Greengard P, Heintz N (2014) Cell type–specific mRNA purification by translating ribosome affinity purification (TRAP). Nat Protoc 9:1282–1291.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 38:576–589.

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

Holguera I, Desplan C (2018) Neuronal specification in space and time. Science 362:176–180.

Hu J, Ho AL, Yuan L, Hu B, Hua S, Hwang SS, Zhang J, Hu T, Zheng H, Gan B, Wu G, Wang YA, Chin L, DePinho RA (2013) Neutralization of terminal differentiation in gliomagenesis. Proc Natl Acad Sci 110:14520–14527.

Hu Y, Huang K, An Q, Du G, Hu G, Xue J, Zhu X, Wang C-Y, Xue Z, Fan G (2016) Simultaneous profiling of transcriptome and DNA methylome from a single cell. Genome Biol 17:88.

Husmann D, Gozani O (2019) Histone lysine methyltransferases in biology and disease. Nat Struct Mol Biol 26:880–889.

International Human Genome Sequencing Consortium et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921.

Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, Zhang D, Li D, Xing X, Kim S, O’Donnell D, Gordon JI, Wang T (2019) Transposable elements drive widespread expression of oncogenes in human cancers. Nat Genet 51:611–617.

Jang MK, Mochizuki K, Zhou M, Jeong H-S, Brady JN, Ozato K (2005) The Bromodomain Protein Brd4 Is a Positive Regulatory Component of P-TEFb and Stimulates RNA Polymerase II-Dependent Transcription. Mol Cell 19:523–534.

Jiang Y, Yu VC, Buchholz F, O’Connell S, Rhodes SJ, Candeloro C, Xia Y-R, Lusis AJ, Rosenfeld MG (1996) A Novel Family of Cys-Cys, His-Cys Zinc Finger Transcription Factors Expressed in Developing Nervous System and Pituitary Gland. J Biol Chem 271:10723–10730.

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316:1497–1502.

Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13:484–492.

Jumper J et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589.

Kameyama T, Matsushita F, Kadokawa Y, Marunouchi T (2011) Myt/NZF family transcription factors regulate neuronal differentiation of P19 cells. Neurosci Lett 497:74–79.

Kanno T, Kanno Y, LeRoy G, Campos E, Sun H-W, Brooks SR, Vahedi G, Heightman TD, Garcia BA, Reinberg D, Siebenlist U, O’Shea JJ, Ozato K (2014) BRD4 assists elongation of both coding and enhancer RNAs by interacting with acetylated histones. Nat Struct Mol Biol 21:1047–1057.

Karczewski KJ et al. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581:434–443.

Katayama Y, Nishiyama M, Shoji H, Ohkawa Y, Kawamura A, Sato T, Suyama M, Takumi T, Miyakawa T, Nakayama KI (2016) CHD8 haploinsufficiency results in autistic-like phenotypes in mice. Nature 537:675–679.

Kaya-Okur HS, Janssens DH, Henikoff JG, Ahmad K, Henikoff S (2020) Efficient low-cost chromatin profiling with CUT&Tag. Nat Protoc 15:3264–3283.

Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, Ahmad K, Henikoff S (2019) CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Commun 10:1930.

Kepa A et al. (2017) Associations of the Intellectual Disability Gene MYT1L with Helix–Loop–Helix Gene Expression, Hippocampus Volume and Hippocampus Activation During Memory Retrieval. Neuropsychopharmacology 42:2516–2526.

Kfoury N, Qi Z, Prager BC, Wilkinson MN, Broestl L, Berrett KC, Moudgil A, Sankararaman S, Chen X, Gertz J, Rich JN, Mitra RD, Rubin JB (2021) Brd4-bound enhancers drive cell-intrinsic sex differences in glioblastoma. Proc Natl Acad Sci 118:e2017148118.

Kim JG, Armstrong RC, Agoston D v., Robinsky A, Wiese C, Nagle J, Hudson LD (1997) Myelin transcription factor 1 (Myt1) of the oligodendrocyte lineage, along with a closely related CCHC zinc finger, is expressed in developing neurons in the mammalian central nervous system. J Neurosci Res 50:272–290.

Kim S, Oh H, Choi SH, Yoo Y-E, Noh YW, Cho Y, Im GH, Lee C, Oh Y, Yang E, Kim G, Chung W-S, Kim H, Kang H, Bae Y, Kim S-G, Kim E (2022) Postnatal age-differential ASD-like transcriptomic, synaptic, and behavioral deficits in Myt1l-mutant mice. Cell Rep 40:111398.

Kochan J, Wawro M, Kasza A (2015) Simultaneous detection of mRNA and protein in single cells using immunofluorescence-combined single-molecule RNA FISH. BioTechniques 59:209–221.

Koenigs M, Grafman J (2009) The functional neuroanatomy of depression: Distinct roles for ventromedial and dorsolateral prefrontal cortex. Behav Brain Res 201:239–243.

Korb E, Herre M, Zucker-Scharff I, Darnell RB, Allis CD (2015) BET protein Brd4 activates transcription in neurons and BET inhibitor Jq1 blocks memory in mice. Nat Neurosci 18:1464–1473.

La Manno G et al. (2018) RNA velocity of single cells. Nature 560:494–498.

La Manno G, Siletti K, Furlan A, Gyllborg D, Vinsland E, Mossi Albiach A, Mattsson Langseth C, Khven I, Lederer AR, Dratva LM, Johnsson A, Nilsson M, Lönnerberg P, Linnarsson S (2021) Molecular architecture of the developing mouse brain. Nature 596:92–96.

Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma’ayan A (2018) Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun 9:1366.

Lagunas T, Plassmeyer SP, Fischer AD, Friedman RZ, Rieger MA, Selmanovic D, Sarafinovska S, Sol YK, Kasper MJ, Fass SB, Aguilar Lucero AF, An J-Y, Sanders SJ, Cohen BA, Dougherty JD (2023) A Cre-dependent massively parallel reporter assay allows for cell-type specific assessment of the functional effects of non-coding elements in vivo. Commun Biol 6:1151.

Lalli M, Yen A, Thopte U, Dong F, Moudgil A, Chen X, Milbrandt J, Dougherty JD, Mitra RD (2022) Measuring transcription factor binding and gene expression using barcoded self-reporting transposon calling cards and transcriptomes. NAR Genomics Bioinforma 4:lqac061.

Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, Lickert H, Ansari M, Schniering J, Schiller HB, Pe’er D, Theis FJ (2022) CellRank for directed single-cell fate mapping. Nat Methods.

Langouët M, Glatt-Deeley HR, Chung MS, Dupont-Thibert CM, Mathieux E, Banda EC, Stoddard CE, Crandall L, Lalande M (2018) Zinc finger protein 274 regulates imprinted expression of transcripts in Prader-Willi syndrome neurons. Hum Mol Genet 27:505–515.

Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11:204–220.

Lee J-E, Park Y-K, Park S, Jang Y, Waring N, Dey A, Ozato K, Lai B, Peng W, Ge K (2017) Brd4 binds to active enhancers to control cell identity gene induction in adipogenesis and myogenesis. Nat Commun 8:2217.

Lee S, Lee S-K (2010) Crucial roles of histone-modifying enzymes in mediating neural cell-type specification. Curr Opin Neurobiol 20:29–36.

Lee TI, Young RA (2013) Transcriptional Regulation and Its Misregulation in Disease. Cell 152:1237–1251.

LeRoy G, Chepelev I, DiMaggio PA, Blanco MA, Zee BM, Zhao K, Garcia BA (2012) Proteogenomic characterization and mapping of nucleosomes decoded by Brd and HP1 proteins. Genome Biol 13:R68.

Li D, Hsu S, Purushotham D, Sears RL, Wang T (2019) WashU Epigenome Browser update 2019. Nucleic Acids Res 47:W158–W165.

Li X, Harrell RA, Handler AM, Beam T, Hennessy K, Fraser MJ (2005) piggyBac internal sequences are necessary for efficient transformation of target genomes. Insect Mol Biol 14:17–30.

Liang L, Cao C, Ji L, Cai Z, Wang D, Ye R, Chen J, Yu X, Zhou J, Bai Z, Wang R, Yang X, Zhu P, Xue Y (2023) Complementary Alu sequences mediate enhancer–promoter selectivity. Nature 619:868–875.

Liu H et al. (2023) Single-cell DNA Methylome and 3D Multi-omic Atlas of the Adult Mouse Brain. Nature 624:366–377.

Liu S, Du T, Liu Z, Shen Y, Xiu J, Xu Q (2016) Inverse changes in L1 retrotransposons between blood and brain in major depressive disorder. Sci Rep 6:37530.

Loid P, Mäkitie R, Costantini A, Viljakainen H, Pekkinen M, Mäkitie O (2018) A novel MYT1L mutation in a patient with severe early-onset obesity and intellectual disability. Am J Med Genet A 176:1972–1975.

Lonsdale J et al. (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45:580–585.

Lord C, Elsabbagh M, Baird G, Veenstra-Vanderweele J (2018) Autism spectrum disorder. The Lancet 392:508–520.

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

Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA (2013) Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. Cell 153:320–334.

Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, Theis FJ (2022) Benchmarking atlas-level data integration in single-cell genomics. Nat Methods 19:41–50.

Macaulay IC et al. (2015) G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods 12:519–522.

Macaulay IC, Teng MJ, Haerty W, Kumar P, Ponting CP, Voet T (2016) Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. Nat Protoc 11:2081–2103.

Mall M et al. (2017) Myt1l safeguards neuronal identity by actively repressing many non-neuronal fates. Nature 544:245–249.

Mansfield P, Constantino JN, Baldridge D (2020) MYT1L A systematic review of genetic variation encompassing schizophrenia and autism. Am J Med Genet B Neuropsychiatr Genet 183:227–233.

Manukyan A, Kowalczyk I, Melhuish TA, Lemiesz A, Wotton D (2018) Analysis of transcriptional activity by the Myt1 and Myt1l transcription factors. J Cell Biochem 119:4644–4655.

Martin BK, Qiu C, Nichols E, Phung M, Green-Gladden R, Srivatsan S, Blecher-Gonen R, Beliveau BJ, Trapnell C, Cao J, Shendure J (2023) Optimized single-nucleus transcriptional profiling by combinatorial indexing. Nat Protoc 18:188–207.

Matsumoto T, Honda S, Harada N (2003) Alteration in Sex-Specific Behaviors in Male Mice Lacking the Aromatase Gene. Neuroendocrinology 77:416–424.

Matsushita F, Kameyama T, Kadokawa Y, Marunouchi T (2014) Spatiotemporal expression pattern of Myt/NZF family zinc finger transcription factors during mouse nervous system development: Expression of *NZF* S in Neural Development. Dev Dyn 243:588–600.

McBurney MW, Sutherland LC, Adra CN, Leclair B, Rudnicki MA, Jardine K (1991) The mouse Pgk-1 gene promoter contains an upstream activator sequence. Nucleic Acids Res 19:5755–5761.

McCarthy MM (2008) Estradiol and the Developing Brain. Physiol Rev 88:91–134.

McGinnis CS, Murrow LM, Gartner ZJ (2019) DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst 8:329-337.e4.

Melhuish TA, Kowalczyk I, Manukyan A, Zhang Y, Shah A, Abounader R, Wotton D (2018) Myt1 and Myt1l transcription factors limit proliferation in GBM cells by repressing YAP1 expression. Biochim Biophys Acta Gene Regul Mech 1861:983–995.

Moreau-Fauvarque C, Kumanogoh A, Camand E, Jaillard C, Barbin G, Boquet I, Love C, Jones EY, Kikutani H, Lubetzki C, Dusart I, Chédotal A (2003) The Transmembrane Semaphorin Sema4D/CD100, an Inhibitor of Axonal Growth, Is Expressed on Oligodendrocytes and Upregulated after CNS Lesion. J Neurosci 23:9229–9239.

Morellet N, Li X, Wieninger SA, Taylor JL, Bischerour J, Moriau S, Lescop E, Bardiaux B, Mathy N, Assrir N, Bétermier M, Nilges M, Hickman AB, Dyda F, Craig NL, Guittet E (2018) Sequence-specific DNA binding activity of the cross-brace zinc finger motif of the piggyBac transposase. Nucleic Acids Res 46:2660–2677.

Moudgil A, Li D, Hsu S, Purushotham D, Wang T (2020a) The qBED track: a novel genome browser visualization for point processes. Bioinformatics 3.

Moudgil A, Wilkinson MN, Chen X, He J, Cammack AJ, Vasek MJ, Lagunas T, Qi Z, Lalli MA, Guo C, Morris SA, Dougherty JD, Mitra RD (2020b) Self-Reporting Transposons Enable Simultaneous Readout of Gene Expression and Transcription Factor Binding in Single Cells. Cell 182:1–17.

Murao N, Noguchi H, Nakashima K (2016) Epigenetic regulation of neural stem cell property from embryo to adult. Neuroepigenetics 5:1–10.

Nakazawa Y, Huye LE, Dotti G, Foster AE, Vera JF, Manuri PR, June CH, Rooney CM, Wilson MH (2009) Optimization of the PiggyBac Transposon System for the Sustained Genetic Modification of Human T Lymphocytes. J Immunother 32:826–836.

Ogawa S, Chester AE, Hewitt SC, Walker VR, Gustafsson J-Å, Smithies O, Korach KS, Pfaff DW (2000) Abolition of male sexual behaviors in mice lacking estrogen receptors α and β (αβERKO). Proc Natl Acad Sci 97:14737–14741.

Ouwenga RL, Dougherty J (2015) Fmrp targets or not: long, highly brain-expressed genes tend to be implicated in autism and brain disorders. Mol Autism 6:16.

Philippe C, Vargas-Landin DB, Doucet AJ, Van Essen D, Vera-Otarola J, Kuciak M, Corbin A, Nigumann P, Cristofari G (2016) Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. eLife 5:e13926.

Phoenix CH, Goy RW, Gerall AA, Young WC (1959) Organizing action of prenatally administered testosterone proprionate on the tissues mediating mating behavior in the female guinea pig. Endocrinology 65:369–382.

Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R (2014) Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc 9:171–181.

Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, Calero-Nieto FJ, Mulas C, Ibarra-Soria X, Tyser RCV, Ho DLL, Reik W, Srinivas S, Simons BD, Nichols J, Marioni JC, Göttgens B (2019) A single-cell molecular map of mouse gastrulation and early organogenesis. Nature 566:490–495.

Preissl S, Fang R, Huang H, Zhao Y, Raviram R, Gorkin DU, Zhang Y, Sos BC, Afzal V, Dickel DE, Kuan S, Visel A, Pennacchio LA, Zhang K, Ren B (2018) Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. Nat Neurosci 21:432–439.

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

Rad R et al. (2015) A conditional piggyBac transposition system for genetic screening in mice identifies oncogenic networks in pancreatic cancer. Nat Genet 47:47–56.

Ramaswami G, Won H, Gandal MJ, Haney J, Wang JC, Wong CCY, Sun W, Prabhakar S, Mill J, Geschwind DH (2020) Integrative genomics identifies a convergent molecular subtype that links epigenomic with transcriptomic differences in autism. Nat Commun 11:4873.

Ramón y Cajal S (1954) Neuron Theory or Reticular Theory? Objective Evidence of the Anatomical Unity of Nerve Cells, 1st ed. Madrid: Consejo Superior de Investigaciones Cientificas, Instituto Ramon y Cajal.

Ramón y Cajal S (1888) Estructura de los centros nerviosos de las aves, 1st ed. Rev. Trim. Histol. Norm. Pat.

Rang FJ, De Luca KL, De Vries SS, Valdes-Quezada C, Boele E, Nguyen PD, Guerreiro I, Sato Y, Kimura H, Bakkers J, Kind J (2022) Single-cell profiling of transcriptome and histone modifications with EpiDamID. Mol Cell 82:1956-1970.e14.

Romm E, Nielsen JA, Kim JG, Hudson LD (2005) Myt1 family recruits histone deacetylase to regulate neural transcription. J Neurochem 93:1444–1453.

Sanders SJ (2015) First glimpses of the neurobiology of autism spectrum disorder. Curr Opin Genet Dev, Molecular and genetic bases of disease 33:80–92.

Sanders SJ et al. (2015) Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron 87:1215–1233.

Sartor GC, Powell SK, Brothers SP, Wahlestedt C (2015) Epigenetic Readers of Lysine Acetylation Regulate Cocaine-Induced Plasticity. J Neurosci 35:15062–15072.

Satterstrom FK et al. (2020) Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell 180:568-584.e23.

Shah NM, Jang HJ, Liang Y, Maeng JH, Tzeng S-C, Wu A, Basri NL, Qu X, Fan C, Li A, Katz B, Li D, Xing X, Evans BS, Wang T (2023) Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. Nat Genet 55:631–639.

Shi L, Zhang Z, Su B (2016) Sex Biased Gene Expression Profiling of Human Brains at Major Developmental Stages. Sci Rep 6:21181.

Skene PJ, Henikoff JG, Henikoff S (2018) Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nat Protoc 13:1006–1019.

Skene PJ, Henikoff S (2017) An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife 6:e21856.

Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods 11:817–820.

Smucny J, Dienel SJ, Lewis DA, Carter CS (2022) Mechanisms underlying dorsolateral prefrontal cortex contributions to cognitive dysfunction in schizophrenia. Neuropsychopharmacology 47:292–308.

Soh KT, Tario Jr. JD, Colligan S, Maguire O, Pan D, Minderman H, Wallace PK (2016) Simultaneous Single-Cell Measurement of Messenger RNA Cell Surface Proteins and.pdf. Curr Protoc Cytom 7:1–33.

Solodushko V, Bitko V, Fouty B (2014) Minimal piggyBac vectors for chromatin integration. Gene Ther 21:1–9.

Sullivan JM, Badimon A, Schaefer U, Ayata P, Gray J, Chung C, Von Schimmelmann M, Zhang F, Garton N, Smithers N, Lewis H, Tarakhovsky A, Prinjha RK, Schaefer A (2015) Autism-like syndrome is induced by pharmacological suppression of BET proteins in young mice. J Exp Med 212:1771–1781.

Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, von Mering C (2023) The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res 51:D638–D646.

Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Curr Protoc Bioinforma 25.

The DDD Study et al. (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515:209–215.

Thrupp N, Sala Frigerio C, Wolfs L, Skene NG, Fattorelli N, Poovathingal S, Fourne Y, Matthews PM, Theys T, Mancuso R, De Strooper B, Fiers M (2020) Single-Nucleus RNA-Seq Is Not Suitable for Detection of Microglial Activation Genes in Humans. Cell Rep 32:108189.

Tomaz DMR (2016) Insights on the function of MyT1L in Ascl1 mediated neuronal reprogramming.

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

Varadi M et al. (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50:D439–D444.

Velmeshev D, Perez Y, Yan Z, Valencia JE, Castaneda-Castellanos DR, Wang L, Schirmer L, Mayer S, Wick B, Wang S, Nowakowski TJ, Paredes M, Huang EJ, Kriegstein AR (2023) Single-cell analysis of prenatal and postnatal human cortical development. Science 382:eadf0834.

Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Südhof TC, Wernig M (2010) Direct conversion of fibroblasts to functional neurons by defined factors. Nature 463:1035–1041.

Vogel MJ, Peric-Hupkes D, van Steensel B (2007) Detection of in vivo protein–DNA interactions using DamID in mammalian cells. Nat Protoc 2:1467–1478.

Wang C, Lin H (2021) Roles of piRNAs in transposon and pseudogene regulation of germline mRNAs and lncRNAs. Genome Biol 22:27.

Wang H, Heinz ME, Crosby SD, Johnston M, Mitra RD (2008) “Calling Cards” method for high-throughput identification of targets of yeast DNA-binding proteins. Nat Protoc 3:1569–1577.

Wang H, Johnston M, Mitra RD (2007) Calling cards for DNA-binding proteins. Genome Res 17:1202–1209.

Wang H, Mayhew D, Chen X, Johnston M, Mitra RD (2012) “Calling Cards” for DNA-Binding Proteins in Mammalian Cells. Genetics 190:941–949.

Wang T et al. (2016) De novo genic mutations among a Chinese autism spectrum disorder cohort. Nat Commun 7:13316.

Wang X, Xu Z, Tian Z, Zhang X, Xu D, Li Q, Zhang J, Wang T (2017) The EF-1α promoter maintains high-level transgene expression from episomal vectors in transfected CHO-K1 cells. J Cell Mol Med 21:3044–3054.

Wang Y, Pryputniewicz-Dobrinska D, Nagy EÉ, Kaufman CD, Singh M, Yant S, Wang J, Dalda A, Kay MA, Ivics Z, Izsvák Z (2017) Regulated complex assembly safeguards the fidelity of *Sleeping Beauty* transposition. Nucleic Acids Res 45:311–326.

Wapinski OL et al. (2013) Hierarchical Mechanisms for Direct Reprogramming of Fibroblasts to Neurons. Cell 155:621–635.

Weigel B et al. (2023) MYT1L haploinsufficiency in human neurons and mice causes autism-associated phenotypes that can be reversed by genetic and pharmacologic intervention. Mol Psychiatry 28:2122–2135.

Wen F, Tang X, Xu L, Qu H (2022) Comparison of single-nucleus and single-cell transcriptomes in hepatocellular carcinoma tissue. Mol Med Rep 26:339.

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA (2013) Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell 153:307–319.

Wilhelm D, Bernard P (Eds.) (2016) Non-coding RNA and the Reproductive System, Advances in Experimental Medicine and Biology. Dordrecht: Springer Netherlands.

Willsey HR, Willsey AJ, Wang B, State MW (2022) Genomics, convergent neuroscience and progress in understanding autism spectrum disorder. Nat Rev Neurosci 23:323–341.

Windheuser IC et al. (2020) Nine newly identified individuals refine the phenotype associated with MYT1L mutations. Am J Med Genet A 182:1021–1031.

Wöhr M, Fong WM, Janas JA, Mall M, Thome C, Vangipuram M, Meng L, Südhof TC, Wernig M (2022) Myt1l haploinsufficiency leads to obesity and multifaceted behavioral alterations in mice. Mol Autism 13:19.

Wu H, Kirita Y, Donnelly EL, Humphreys BD (2019) Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. J Am Soc Nephrol 30:23–32.

Wu J, Cai Y, Wu X, Ying Y, Tai Y, He M (2021) Transcardiac Perfusion of the Mouse for Brain Tissue Dissection and Fixation. BIO-Protoc 11.

Wu SC-Y, Meir Y-JJ, Coates CJ, Handler AM, Pelczar P, Moisyadi S, Kaminski JM (2006) *piggyBac* is a flexible and highly active transposon as compared to *Sleeping Beauty* , *Tol2* , and *Mos1* in mammalian cells. Proc Natl Acad Sci 103:15008–15013.

Wu S-Y, Chiang C-M (2007) The Double Bromodomain-containing Chromatin Adaptor Brd4 and Transcriptional Regulation. J Biol Chem 282:13141–13145.

Wu Z, Yang H, Colosi P (2010) Effect of Genome Size on AAV Vector Packaging. Mol Ther 18:80–86.

Yamaguchi W, Tamai R, Kageura M, Furuyama T, Inagaki S (2012) Sema4D as an inhibitory regulator in oligodendrocyte development. Mol Cell Neurosci 49:290–299.

Yamashita YM, Fuller MT, Jones DL (2005) Signaling in stem cell niches: lessons from the *Drosophila* germline. J Cell Sci 118:665–672.

Yang S, Zhou X, Li R, Fu X, Sun P (2017) Optimized PEI-based Transfection Method for Transient Transfection and Lentiviral Production: PEI-based Transfection for Lentiviral Production. Curr Protoc Chem Biol 9:147–157.

Yang Z, Yik JHN, Chen R, He N, Jang MK, Ozato K, Zhou Q (2005) Recruitment of P-TEFb for Stimulation of Transcriptional Elongation by the Bromodomain Protein Brd4. Mol Cell 19:535–545.

Yao Z et al. (2023) A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature 624:317–332.

Yen A, Mateusiak C, Sarafinovska S, Gachechiladze MA, Guo J, Chen X, Moudgil A, Cammack AJ, Hoisington-Lopez J, Crosby M, Brent MR, Mitra RD, Dougherty JD (2023) Calling Cards: A Customizable Platform to Longitudinally Record Protein-DNA Interactions. Curr Protoc 3.

Yen L, Svendsen J, Lee J-S, Gray JT, Magnier M, Baba T, D'Amato RJ, Mulligan RC (2004) Exogenous control of mammalian gene expression through modulation of RNA self-cleavage. Nature 431:471–476.

Yen M, Qi Z, Chen X, Cooper JA, Mitra RD, Onken MD (2018) Transposase mapping identifies the genomic targets of BAP1 in uveal melanoma. BMC Med Genomics 11:97.

Yin Y et al. (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science 356:eaaj2239.

Yoshida J, Akagi K, Misawa R, Kokubu C, Takeda J, Horie K (2017) Chromatin states shape insertion profiles of the piggyBac, Tol2 and Sleeping Beauty transposons and murine leukemia virus. Sci Rep 7:43613.

Yusa K, Zhou L, Li MA, Bradley A, Craig NL (2011) A hyperactive piggyBac transposase for mammalian applications. Proc Natl Acad Sci 108:1531–1536.

Zappia L, Oshlack A (2018) Clustering trees: a visualization for evaluating clusterings at multiple resolutions. GigaScience 7.

Zhang H-L, Wang J, Tang L (2014) Sema4D Knockdown in Oligodendrocytes Promotes Functional Recovery After Spinal Cord Injury. Cell Biochem Biophys 68:489–496.

Zhang M, Pan X, Jung W, Halpern AR, Eichhorn SW, Lei Z, Cohen L, Smith KA, Tasic B, Yao Z, Zeng H, Zhuang X (2023) Molecularly defined and spatially resolved cell atlas of the whole mouse brain. Nature 624:343–354.

Zhang M, Zhao J, Lv Y, Wang W, Feng C, Zou W, Su L, Jiao J (2020) Histone Variants and Histone Modifications in Neurogenesis. Trends Cell Biol 30:869–880.

Zu S et al. (2023) Single-cell analysis of chromatin accessibility in the adult mouse brain. Nature 624:378–389.