Title

Genomic and Transcriptomic Examination of Functional Elements and Absent Sequences

Author

Chan, Candace S.Y

Publication Date

2024

Peer reviewed|Thesis/dissertation

Genomic and transcriptomic examination of functional elements and absent sequences

by
Candace S.Y. Chan

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biochemistry and Molecular Biology

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Hao Li
FBE217337D9A45A...

Hao Li
Chair

DocuSigned by:
Nadav Ahituv
Signed by:
DAE34F4...

Nadav Ahituv

Martin Hemberg
DocuSigned by:
44B...

Martin Hemberg

DocuSigned by:
Hani Goodarzi
FDD44359FCC6487...

Hani Goodarzi

Committee Members

ACKNOWLEDGEMENTS

I would first like to express my gratitude to my thesis advisors, Professor Nadav Ahituv and Professor Martin Hemberg. Their unwavering support and commitment to my growth as a scientist have been invaluable. Beyond their scientific guidance, they have been my advocates throughout this journey, always encouraging me while thoughtfully challenging me to reach my fullest potential. I am thankful for their mentorship and for the impact they have had on my academic and personal development. I have always been in awe of both mentors' passion for their work and their ability to balance so much of science and life. Their drive has inspired and will continue to motivate me past graduate school. Thank you, Nadav and Martin, for everything.

I would also like to express my appreciation for my thesis committee members, Professor Hao Li and Professor Hani Goodarzi. Thank you both for your support, invaluable insights, and encouragement in broadening my knowledge. I also want to sincerely thank previous mentors, Professor Sui Wang and Professor Henrik Scheller. Their mentorship provided me with the foundation to embark on this scientific path. Through their extensive support, I was able to pursue and cultivate a career in science. Their influence and encouragement have left a lasting impact on my development as a researcher, and for that, I am profoundly grateful.

Starting graduate school during a global pandemic was not an easy feat, but I was blessed to join the UCSF community during that time. Thank you to my lab mates in the Ahituv lab, who patiently taught me from so many areas of expertise. I am particularly thankful for Ofer Yizhar-Barnea's mentorship and continue to be inspired by the positivity he brought to every interaction and his determination. I am immensely grateful for Dianne Laboy Cintrón, JingJing Zhao, and Zhe Liu for the endless laughs and support they provided, even during the most challenging times. Thank you to Yelena Guttman and Yarden Golan Maor for sharing wisdom and quality

time together through salad lunches. Thank you to Peter Lu and Toni Hurley for guiding me through every step of graduate school.

I am deeply grateful to my parents for their countless sacrifices, which opened the doors for me to embark on this journey. Though they often said they could not guide me through life in a foreign land, they instilled in me an unwavering belief in the value of education—a belief that has become the most important guide in my life. Without their dedication and sacrifices, I would not have been able to pursue a PhD. I am also immensely grateful to have had my sister's support throughout this journey, continually motivating me to stay healthy and happy. Growing up, we never knew we could take this path. How grateful I am that I was able to share that with you to this day.

Friends near and far have been an invaluable source of support and joy throughout this journey. To the friends I met at UCSF—especially Jenny, Simon, Chi Yun, and Wilhelm—thank you for the unforgettable memories. Your friendship has meant the world to me and has made this journey all the more special. Victor, thank you for your positivity and encouragement, which inspires me to conquer the hills we hike and the metaphorical mountains. To my friends from the court, thank you for the camaraderie courtside. Amy, Winnie, Ben, and Erwin, thank you for keeping me grounded and connected. Byron, you've been my dearest friend for most of my life, standing by me through thick and thin - thank you.

To the broader scientific community: my work stems from the collective knowledge and dedication to discovery of countless others. It is a privilege to be able to pursue a doctorate, and impossible without the tireless efforts of mentors, collaborators, and supporters. I am grateful for the opportunity to contribute, even in a small way, to the well of human knowledge.

Finally, I am profoundly grateful to Ilias Georgakopoulos-Soares. I struggle to put into words the depth of gratitude I feel. I am forever grateful to Ilias for teaching me to stay focused on what I now hope is a long road ahead. Thank you, Ilias, for taking a chance on me, for all your patience and humor, and for being my greatest supporter through all the challenges of graduate training and life these past few years.

CONTRIBUTIONS

Chapter 2 is adapted from an unpublished manuscript with the following authors: Candace S.Y Chan, Ioannis Mouratidis, Austin Montgomery, Georgios Christos Tsiatsianis, Nikol Chantzi, Martin Hemberg, Nadav Ahituv, Ilias Georgakopoulos-Soares. Detailed contributions are in Section 2.6.

Chapter 3 is adapted from an unpublished manuscript with the following authors: Candace S.Y Chan, Hai Nguyen, Seungbyn Baek, Dianne Laboy Cintron, Rory Sheng, Lana Harshman, Mai Nobuhara, Aki Ushiki, Cassidy Biellak, Kelly An, Gracie Gordon, Francois Mifsud, Abbie Diener, Aristides Diamant, Insuk Lee, Eric Huang, Martin Hemberg, Christian Vaisse, Nadav Ahituv. Detailed contributions are in section 3.7.

Genomic and Transcriptomic Examination of Functional Elements and Absent Sequences

Candace S.Y. Chan

ABSTRACT

Background: Genome-wide association studies have identified numerous disease-associated variants, but a vast majority are located in non-coding regions, making it challenging to understand their functional impact. This complexity necessitates new techniques to identify causal variants in non-coding regions and elucidate their specific cellular contexts and mechanisms of action. Here we present work i) examining mutations that create nullomers in the human genome to explore its potential utility in identifying pathogenic mutations and ii) a single-cell multi-omic study identifying the transcriptome and regulome of the human and mouse hypothalamus to identify regulatory regions of obesity-associated variants.

Methods: (i) We generated all possible mutations of the human genome that can lead to emergence of a nullomer and examine where in the genome they emerge. (ii) We apply single-cell RNA and ATAC sequencing to adult hypothalamus samples.

Results and Conclusions: (i) Our findings highlight CpG hypermutability and methylated cytosines as key elements leading to resurfacing of nullomers in individuals. We also showcase that nullomers can have applications in disease annotation and pathogenic variant identification. (ii) We identified regulatory elements of hypothalamus cell types and mapped obesity-associated variants to cell-type specific peaks. We validated these regions to be enhancers using CRISPR editing and CRISPRi.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

The first human reference genome was published in 2001 (Lander et al. 2001); since then, genomics has experienced a transformation driven by advancements in high-throughput sequencing technologies and the accumulation of datasets. These technological breakthroughs have reduced the time and cost of genome sequencing, enabling scientists to generate genomic data from diverse populations and species at an accelerated pace. The dramatic increase in genomic data has revolutionized our ability to investigate genetic variation, gene regulation, and disease mechanisms. These technologies have accelerated progress and paves the way for personalized medicine and precision healthcare approaches.

Large-scale sequencing studies enable us to understand how the genome translates to common phenotypes and diseases. To date, tens of thousands of genome-wide association studies (GWAS) have identified numerous variants associated with disease risk (Sollis et al., 2023). While GWAS can be a powerful tool towards associating variants to traits, a significant challenge remains: over 90% of disease-associated variants are located outside of protein-coding regions (Chatterjee and Ahituv, 2017). This raises questions of how these variants contribute to phenotypic outcomes, and some of them are most likely regulatory as they are found within regulatory elements in the non-coding genome (Maurano et al., 2012). Enhancers are a type of *cis*-regulatory element that can regulate gene expression in a cell-type-specific manner and can exert their influence over considerable genomic distances, sometimes up to several megabases (Song et al. 2019). This complexity makes it challenging to link these enhancers, and the variants they contain, to their target genes and to determine the specific cellular context in which they are active. Additionally, while a minority of diseases are Mendelian, i.e. caused by coding mutations

in single genes, complex diseases arise from multiple variants that jointly contribute, with each variant only explaining a small proportion of trait heritability. Linkage disequilibrium further complicates the task of determining variants' functional impact, as causal variants may also be in close proximity to non-causal variants (Gaulton, Preissl, and Ren 2023). Therefore, there is a pressing need for techniques to identify causal variants in non-coding regions and elucidate the contexts in which it functions.

Nullomers offer a unique approach to the challenge of prioritizing variants. Nullomers are short nucleotide kmers that are absent from a genome (Hampikian and Andersen 2007). Nullomers are between 11-18 base pairs, and absent in the genome from the set of all possible DNA kmers which is equal in size to 4^k. It was first hypothesized that DNA sequences were not missing by chance, but due to negative selection (Hampikian and Andersen 2007). Building on this idea, we hypothesize that genetic variants introducing nullomer sequences into the genome may disrupt biological function.

The absence of nullomers, and the proteomic equivalent nullpeptides, have been leveraged towards multiple applications in biotechnology. Hampikian and Andersen first described its potential as a therapeutic target against pathogenic species versus its host (Hampikian and Andersen 2007), and introducing nullomers into synthetic constructs to be used as identifiers. Goswami et al. introduced the use of nullomers as a barcode in forensic science, relying on the absence of nullomers to differentiate DNA samples from evidence (Goswami et al. 2013). The cytotoxic effects of nullpeptides 9R and 9S1R across cancer cells and mouse models demonstrated its potential as a cancer drug (Alileche et al. 2012; Alileche and Hampikian 2017; Ali et al. 2024). The emergence of nullomers and nullpeptides from somatic mutations during

cancer development was leveraged as biomarkers for detection of cancer subtypes (I. Georgakopoulos-Soares et al. 2021; Montgomery et al. 2023; Tsiatsianis et al. 2024).

Although nullomers have many promising applications, the reasons for their absence in genomes are not yet fully understood (Acquisti et al. 2007; Vergni and Santoni 2016; Hampikian and Andersen 2007; Ilias Georgakopoulos-Soares et al. 2021; Koulouras and Frith 2021). In the work described in chapter 2, we utilize a set of simulated mutations that lead to emergence of nullomers in the human genome to understand whether nullomers are absent by chance or whether it has deleterious properties. We find that nullomer-emerging mutations are more likely to lead to pathogenic mutations, expanding the application of nullomers towards identification of disease-causing mutations.

Another challenge towards understanding how non-coding variants contribute to disease is finding the cell-type specific context in which they function. Conventionally, enhancers were identified by using bulk sequencing methods to filter for functional regions, such as ChIP-seq for marks of active promoters and enhancers (Johnson et al. 2007) and ATAC-seq (Buenrostro et al. 2015), which identifies open chromatin regions. However, for complex systems such as the hypothalamus, bulk processing and averaged signals leads to loss of cell-type specific signals. Furthermore, the target genes of enhancers still remain unknown using bulk sequencing methods. This renders it difficult to identify context-specific activity of enhancers or to link non-coding genetic variants to their effects in tissues with specialized cell populations such as the hypothalamus.

Single-cell approaches allow for the resolution of signals across different cell types and contexts. Recent advancements made in single-cell technologies enable multiple modalities to be simultaneously interrogated in the same cells (Cao et al. 2019; Stuart and Satija 2019). These

advances have provided a unique opportunity to apply single-cell multiomic methods towards identifying regulatory elements in the non-coding region of the genome, thus enabling identification of cell-type-specific regulatory networks, which are crucial for understanding complex biological processes and disease mechanisms.

The hypothalamus is a complex system whose dysfunction can result in a wide range of diseases, making it an ideal target for single-cell multiomic approaches. It is a region of the brain located above the pituitary gland that connects the nervous and endocrine system to regulate essential physiological processes such as hunger, temperature regulation, and emotional responses. The hypothalamus is organized into distinct regions called nuclei that each control a distinct function. The most well-studied region is the arcuate nucleus, where the leptin-melanocortin pathway is regulated by the orexigenic agouti-related peptide/neuropeptide Y-expressing neurons and the anorexigenic pro-opiomelanocortin/cocaine-amphetamine related transcript expression neurons (Schwartz et al. 2000). These neurons act as agonists and antagonists for the melanocortin 4 receptor (*MC4R*)-expressing neurons in the paraventricular nucleus to promote or reduce food intake depending on circulating levels of leptin, insulin, and ghrelin (Baldini and Phelan 2019). The hypothalamus is central to regulation of appetite and energy expenditure, and thus its malfunction has strong effects on body weight.

The hypothalamus' role in regulating appetite and energy expenditure is influenced by sex differences, contributing to sexual dimorphism in susceptibility to obesity (Palmer and Clegg 2015). For instance, silencing of estrogen receptors in the ventromedial nucleus of the hypothalamus in mice led to obesity in females due to reduced energy expenditure (Xu et al. 2011; Musatov et al. 2007; Heine et al. 2000). Another example highlighting sex-dimorphism in the hypothalamus was found in POMC neurons, which exhibited higher neural activity in female

mice, indicating that POMC may be responsible for relatively lower body weight in female animals (Wang et al. 2018). These studies serve to highlight the complex interaction between sex, hypothalamic function, and obesity. Further research into sex differences in the hypothalamus can serve to develop sex-specific treatments for metabolic disorders.

Obesity accounts for a large proportion of worldwide chronic diseases such as type 2 diabetes, cardiovascular disease, and cancer (Loos and Yeo 2022). Obesity has a strong genetic component, with twin and family studies estimating its heritability to be between 40-70% (Elks et al. 2012). Monogenic forms of obesity, most commonly caused by coding mutations in *MC4R*, accounts for a small proportion of obesity cases, while the majority of cases are polygenic obesity. GWAS have also found that many obesity-associated genes such as the *MC4R* locus are strongly associated with polygenic obesity. However, identifying causal variants have proven difficult due to their position in noncoding regions, as well as obesity-associated genes only being expressed in subpopulations of the central nervous system, where annotation and characterization of regulatory elements have not been achieved.

Previous studies have used single-cell sequencing to survey the celltypes within the hypothalamus, revealing distinct gene expression profiles of each population (Steuernagel et al. 2022; Tadross et al. 2023). Building on this work, we applied single-cell multiome technology to the hypothalamus, enabling us to simultaneously profile both chromatin accessibility and gene expression in the same cell in the adult mouse and human hypothalamus. This integrated approach allows us to correlate chromatin accessibility with gene expression data within subpopulations. In this work, described in chapter 3, we mapped hypothalamus cell-type specific putative genes, and regulatory elements such as promoters and enhancers. Finally, we utilized

these cell-type-specific elements to explore the functional impact of non-coding obesity-associated variants identified through GWAS.

References

Acquisti, Claudia, George Poste, David Curtiss, and Sudhir Kumar. 2007. "Nullomers: Really a Matter of Natural Selection?" *PloS One* 2 (10): e1022.

Alileche, Abdelkrim, Jayita Goswami, William Bourland, Michael Davis, and Greg Hampikian. 2012. "Nullomer Derived Anticancer Peptides (NulloPs): Differential Lethal Effects on Normal and Cancer Cells in Vitro." *Peptides*.

Alileche, Abdelkrim, and Greg Hampikian. 2017. "The Effect of Nullomer-Derived Peptides 9R, 9S1R and 124R on the NCI-60 Panel and Normal Cell Lines." *BMC Cancer* 17 (1): 533.

Ali, Nilufar, Cody Wolf, Swarna Kanchan, Shivakumar R. Veerabhadraiah, Laura Bond, Matthew W. Turner, Cheryl L. Jorcyk, and Greg Hampikian. 2024. "9S1R Nullomer Peptide Induces Mitochondrial Pathology, Metabolic Suppression, and Enhanced Immune Cell Infiltration, in Triple-Negative Breast Cancer Mouse Model." *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie* 170 (January):115997.

Bailey, Timothy L., Nadya Williams, Chris Misleh, and Wilfred W. Li. 2006. "MEME: Discovering and Analyzing DNA and Protein Sequence Motifs." *Nucleic Acids Research* 34 (Web Server issue): W369–73.

Baldini, Giulia, and Kevin D. Phelan. 2019. "The Melanocortin Pathway and Control of Appetite-Progress and Therapeutic Implications." *The Journal of Endocrinology* 241 (1): R1–33.

Breitwieser, Florian P., Jennifer Lu, and Steven L. Salzberg. 2019. "A Review of Methods and Databases for Metagenomic Classification and Assembly." *Briefings in Bioinformatics* 20 (4): 1125–36.

Buenrostro, Jason D., Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales,

Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. 2015. “Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation.” *Nature* 523 (7561): 486–90.

Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, et al. 2019. “The Single-Cell Transcriptional Landscape of Mammalian Organogenesis.” *Nature* 566 (7745): 496–502.

Eilbeck, Karen, Aaron Quinlan, and Mark Yandell. 2017. “Settling the Score: Variant Prioritization and Mendelian Disease.” *Nature Reviews. Genetics* 18 (10): 599–612.

Elks, Cathy E., Marcel den Hoed, Jing Hua Zhao, Stephen J. Sharp, Nicholas J. Wareham, Ruth J. F. Loos, and Ken K. Ong. 2012. “Variability in the Heritability of Body Mass Index: A Systematic Review and Meta-Regression.” *Frontiers in Endocrinology* 3 (February):29.

Gaulton, Kyle J., Sebastian Preissl, and Bing Ren. 2023. “Interpreting Non-Coding Disease-Associated Human Variants Using Single-Cell Epigenomics.” *Nature Reviews. Genetics* 24 (8): 516–34.

Georgakopoulos-Soares, I., O. Barnea, I. Mouratidis, R. Bradley, R. Easterlin, C. Chan, E. Chen, J. Witte, M. Hemberg, and N. Ahituv. 2021. “Leveraging Sequences Missing from the Human Genome to Diagnose Cancer.” *medRxiv*, August. https://doi.org/10.1101/2021.08.15.21261805.

Georgakopoulos-Soares, Ilias, Ofer Yizhar-Barnea, Ioannis Mouratidis, Martin Hemberg, and Nadav Ahituv. 2021. “Absent from DNA and Protein: Genomic Characterization of Nullomers and Nullpeptides across Functional Categories and Evolution.” *Genome Biology* 22 (1): 245.

Goswami, Jayita, Michael C. Davis, Tim Andersen, Abdelkrim Alileche, and Greg Hampikian.

2013. “Safeguarding Forensic DNA Reference Samples with Nullomer Barcodes.” *Journal of Forensic and Legal Medicine* 20 (5): 513–19.

Hampikian, Greg, and Tim Andersen. 2007. “Absent Sequences: Nullomers and Primes.” *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 355–66.

Heine, P. A., J. A. Taylor, G. A. Iwamoto, D. B. Lubahn, and P. S. Cooke. 2000. “Increased Adipose Tissue in Male and Female Estrogen Receptor-Alpha Knockout Mice.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (23): 12729–34.

Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. 2007. “Genome-Wide Mapping of in Vivo Protein-DNA Interactions.” *Science* 316 (5830): 1497–1502.

Koulouras, Grigorios, and Martin C. Frith. 2021. “Significant Non-Existence of Sequences in Genomes and Proteomes.” *Nucleic Acids Research* 49 (6): 3139–55.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921.

Loos, Ruth J. F., and Giles S. H. Yeo. 2022. “The Genetics of Obesity: From Discovery to Biology.” *Nature Reviews. Genetics* 23 (2): 120–33.

Maurano, Matthew T., Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, et al. 2012. “Systematic Localization of Common Disease-Associated Variation in Regulatory DNA.” *Science* 337 (6099): 1190–95.

Montgomery, Austin, Georgios Christos Tsiatsianis, Ioannis Mouratidis, Candace S. Y. Chan, Maria Athanasiou, Anastasios D. Papanastasiou, Verena Kantere, et al. 2023. “Utilizing Nullomers in Cell-Free RNA for Early Cancer Detection.”

Musatov, Sergei, Walter Chen, Donald W. Pfaff, Charles V. Mobbs, Xue-Jun Yang, Deborah J. Clegg, Michael G. Kaplitt, and Sonoko Ogawa. 2007. “Silencing of Estrogen Receptor Alpha in the Ventromedial Nucleus of Hypothalamus Leads to Metabolic Syndrome.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (7): 2501–6.

Nute, Michael, and Tandy Warnow. 2016. “Scaling Statistical Multiple Sequence Alignment to Large Datasets.” *BMC Genomics* 17 (Suppl 10): 764.

Palmer, Biff F., and Deborah J. Clegg. 2015. “The Sexual Dimorphism of Obesity.” *Molecular and Cellular Endocrinology* 402 (February):113–19.

Schoenfelder, Stefan, and Peter Fraser. 2019. “Long-Range Enhancer-Promoter Contacts in Gene Expression Control.” *Nature Reviews. Genetics* 20 (8): 437–55.

Schwartz, M. W., S. C. Woods, D. Porte Jr, R. J. Seeley, and D. G. Baskin. 2000. “Central Nervous System Control of Food Intake.” *Nature* 404 (6778): 661–71.

Song, Michael, Xiaoyu Yang, Xingjie Ren, Lenka Maliskova, Bingkun Li, Ian R. Jones, Chao Wang, et al. 2019. “Mapping Cis-Regulatory Chromatin Contacts in Neural Cells Links Neuropsychiatric Disorder Risk Variants to Target Genes.” *Nature Genetics* 51 (8): 1252–62.

Steuernagel, Lukas, Brian Y. H. Lam, Paul Klemm, Georgina K. C. Dowsett, Corinna A. Bauder, John A. Tadross, Tamara Sotelo Hitschfeld, et al. 2022. “HypoMap-a Unified Single-Cell Gene Expression Atlas of the Murine Hypothalamus.” *Nature Metabolism* 4 (10): 1402–19.

Stuart, Tim, and Rahul Satija. 2019. “Integrative Single-Cell Analysis.” *Nature Reviews. Genetics* 20 (5): 257–72.

Tadross, John A., Lukas Steuernagel, Georgina K. C. Dowsett, Katherine A. Kentistou, Sofia Lundh, Marta Porniece-Kumar, Paul Klemm, et al. 2023. “Human HYPOMAP: A Comprehensive Spatio-Cellular Map of the Human Hypothalamus.” *bioRxiv*. https://doi.org/10.1101/2023.09.15.557967.

Tsiatsianis, Georgios Christos, Candace S. Y. Chan, Ioannis Mouratidis, Nikol Chantzi, Anna Maria Tsiatsiani, Nelson S. Yee, Apostolos Zaravinos, Verena Kantere, and Ilias Georgakopoulos-Soares. 2024. “Peptide Absent Sequences Emerging in Human Cancers.” *European Journal of Cancer* 196 (January):113421.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, and Tuuli Lappalainen. 2021. “Genome-Wide Association Studies.” *Nature Reviews Methods Primers* 1 (1): 1–21.

Vergni, Davide, and Daniele Santoni. 2016. “Nullomers and High Order Nullomers in Genomic Sequences.” *PloS One* 11 (12): e0164540.

Wang, Chunmei, Yanlin He, Pingwen Xu, Yongjie Yang, Kenji Saito, Yan Xia, Xiaofeng Yan, et al. 2018. “TAp63 Contributes to Sexual Dimorphism in POMC Neuron Functions and Energy Homeostasis.” *Nature Communications* 9 (1): 1544.

Xu, Yong, Thekkethil P. Nedungadi, Liangru Zhu, Nasim Sobhani, Boman G. Irani, Kathryn E. Davis, Xiaorui Zhang, et al. 2011. “Distinct Hypothalamic Neurons Mediate Estrogenic Effects on Energy Homeostasis and Reproduction.” *Cell Metabolism* 14 (4): 453–65.

CHAPTER 2

2. 1 Abstract

Nullomers are short DNA sequences (11-18 base pairs) that are absent from a genome; however, they can emerge due to mutations. Here, we characterize all possible putative human nullomer-emerging single base pair mutations, population variants and disease-causing mutations. We find that the primary determinants of nullomer emergence in the human genome are the presence of CpG dinucleotides and methylated cytosines. Putative nullomer-emerging mutations are enriched at specific genomic elements, including transcription start and end sites, splice sites and transcription factor binding sites. We also observe that putative nullomer-emerging mutations are more frequent in highly conserved regions and show preferential location at nucleosomes. Among repeat elements, Alu repeats exhibit pronounced enrichment for putative nullomer-emerging mutations at specific positions. Finally, we find that disease-associated pathogenic mutations are significantly more likely to cause emergence of nullomers than their benign counterparts.

2. 2 Introduction

Nullomers are the shortest absent sequences from a genome (Hampikian and Andersen 2007). In the human genome, the first nullomers appear at eleven base pairs (bps) and the number of nullomers exponentiates with kmer length. Even though nullomers are absent from the reference genome, they can be present in the genomes of other individuals. Germline mutations can be linked with the presence of nullomers, originally absent from the reference human genome (Georgakopoulos-Soares, Yizhar-Barnea, Mouratidis, Hemberg, et al. 2021), including rare

variants (Koulouras and Frith 2021). Mutations that have arisen in the life of a person include private somatic mutations and clonal mutations, such as those that appear during cancer development and those can give rise to nullomers (Georgakopoulos-Soares, Yizhar-Barnea, Mouratidis, Bradley, et al. 2021; Tsiatsianis et al. 2024; Montgomery et al. 2023). For cancer, presence of nullomers has been shown to be a useful biomarker for the early detection of cancer, using liquid biopsies (Georgakopoulos-Soares, Yizhar-Barnea, Mouratidis, Bradley, et al. 2021; Tsiatsianis et al. 2024; Montgomery et al. 2023). Other applications involving nullomers have been described, including cancer cell killing and drug discovery targets (Alileche and Hampikian 2017; Alileche et al. 2012; Ali et al. 2024), forensic applications (Goswami et al. 2013), pathogen detection and surveillance (Pratas and Silva 2021; Silva et al. 2015; Mouratidis, Chan, et al. 2023; Mouratidis, Konnaris, et al. 2023) and immunogenic compounds (Patel et al. 2012). Putative nullomer-emerging mutations have been only examined in a single study to date (Georgakopoulos-Soares, Yizhar-Barnea, Mouratidis, Hemberg, et al. 2021). In that study, all possible single base pair substitutions and single base pair insertions and deletions that can cause the emergence of nullomers were examined. It was shown that nullomers generated from putative nullomer emerging mutations are enriched in promoters and coding regions, while the subset of nullomers that could emerge at hundreds of genomic loci were primarily found at Alu repeats.

The reasons why nullomers are absent from the human genome is an active area of research. Increased likelihood of mutagenesis at specific contexts, including at CpG sites, has been proposed (Acquisti et al. 2007), along with negative selection (Koulouras and Frith 2021; Georgakopoulos-Soares, Yizhar-Barnea, Mouratidis, Hemberg, et al. 2021). One such example

was the identification of restriction site nullomers at viral genomes, which was suggested to be a mechanism safeguarding them against bacterial endonucleases (Koulouras and Frith 2021). In addition, a subset of nullomers was found to be shared between different organismal genomes, which could reflect stronger selection constraints (Georgakopoulos-Soares, Yizhar-Barnea, Mouratidis, Hemberg, et al. 2021; Chantzi et al. 2023; Mouratidis, Baltoumas, et al. 2023; Hampikian and Andersen 2007). Nevertheless, a thorough association between nullomer-emerging and functional genomic elements, or between nullomer-emerging and pathogenicity has yet to be investigated.

Here, we perform a systematic examination of nullomer-emerging mutations across the human genome. We analyze all possible one-base pair mutations, which we term putative nullomer emerging mutations throughout the manuscript. We find that putative nullomer-emerging mutations are enriched at early-replicating regions and at specific genomic sites, including transcription factor binding sites (TFBSs), CpG islands and relative to transcription start sites (TSSs) and splice sites. They are also preferentially positioned relative to nucleosomes. The strongest enrichment patterns for putative nullomer-emerging mutations are observed at CpG methylation sites, consistent with the increased mutation rate at these loci. Finally, we show their clinical relevance using disease causing mutations and pathogenic mutation sites, which are significantly more likely to cause nullomer emergence. In summary, we provide evidence for the mechanisms that cause nullomer formation and the selection constraints against them.

2.3 Results

For our analyses we used all human nullomers between the lengths of 11 and 13 bps, as described previously (Georgakopoulos-Soares, Yizhar-Barnea, Mouratidis, Hemberg, et al. 2021). The shortest putative nullomer emerging length studied was the minimal length at which nullomers appear (11 bps). The upper length (13 bp) was selected as the largest kmer length for which the number of putative nullomer emerging mutations is less than the number of bps of the human genome, which is 13bps. For larger lengths, the excess of putative nullomer emerging mutations makes it harder to characterize the subset of putative nullomer-emerging mutations that are biologically relevant. The thirteen bp length limit was selected because for longer kmer lengths the majority of loci generate nullomer-emerging mutations and therefore stochastic effects increase (40.09 nullomer mutations/ 1kb). We examined every possible substitution and one base-pair insertion or deletion for their ability to cause the emergence of nullomers. For 13 bps, there were a total of 271,432,758 mutations, which were subdivided into insertions, deletions, and the six possible substitution types (**Figure 1a-b**). On average, each putative nullomer-emerging mutation resulted in 2.03, 2.36 and 3.43 nullomers for 11bp, 12bp and 13bp respectively. For each putative nullomer-emerging mutation, we also generated a simulated mutation, controlling for trinucleotide context and proximity (within <1kb from the original mutation), against which we performed multiple comparisons to determine statistical significance.

Putative nullomer-emerging mutations are enriched in early replicating, genic regions and in *cis*-regulatory elements

To investigate the degree of putative nullomer-emerging mutation clustering in close genomic proximity to each other, we examined the distance between consecutive putative nullomer-emerging mutations relative to the expected distance from the simulations. We found a significantly different inter-mutation distance across nullomer lengths and mutation types, with putative nullomer-emerging mutations showing a 5-fold higher degree of clustering than expected by chance (p-value=0, Mann-Whitney U-test, **Figure 1d**, **Supplementary Figure 1a-b, Supplementary Table 1**). This result suggests the presence of putative nullomer-emerging-mutation clusters in the human genome.

We split the genome into 1kB or 50kB or 500kB bins and calculated the number of unique k-mers per bin as well as the number of putative nullomer emerging mutations. We observe that in all cases, bins with nullomer emerging mutations have more unique k-mers than in bins without (p-value=0, Mann-Whitney U-test, **Supplementary Figure 1c**, **Supplementary Table 2**). We conclude that genomic loci with putative nullomer emerging mutations are more likely to be information-rich sequences of the human genome.

Next, we examined if putative mutations that cause the emergence of nullomers are differentially distributed within the human genome and within functional genomic elements. We separated mutations into coding and non-coding and observed a higher density of coding relative to non-coding putative nullomer-emerging mutations across kmer lengths (p-value=0, Mann-Whitney U-test, **Figure 1c, Supplementary Figure 1d, Supplementary Table 3**) and across mutation

categories (**Supplementary Figure 1e**). Importantly, the mutational density was 8.59-fold higher in coding relative to non-coding regions, indicating that putative nullomer-emerging mutations are preferentially located at coding sites.

Replication timing stratifies multiple genomic features, which include gene organization, histone modifications, DNA methylation, heterochromatization and likelihood of mutagenesis (Aran et al. 2011; Suzuki et al. 2011; Stamatoyannopoulos et al. 2009). Repli-Seq is a method used to infer the replication timing of different regions across the genome in a cell type. We used Repli-Seq data from fourteen human cell lines (BG02ES, BJ, GM06990, GM12801, GM12812, GM12813, GM12878, HeLaS3, HepG2, HUVEC, IMR90, K562, MCF7 and NHEK cell lines) (ENCODE Project Consortium 2012) to study the distribution of putative nullomer-emerging mutations relative to replication timing. We separated the data into deciles, based on the replication timing of genomic regions and examined the density of putative nullomer-emerging mutations in each decile. We observed that early replicating regions had an excess of putative nullomer-emerging mutations (**Figure 1e,** Pearson correlation r = 0.97, p-value = 1.94E-06, **Supplementary Table 4**). Separation of putative nullomer-emerging mutations by mutation category revealed that the largest difference between early and late replicating regions in mutation density was observed for insertions and substitutions relative to deletions (**Supplementary Figure 1f**). We also examined how the mutational density of each substitution type for putative nullomer-emerging mutations changed across the replication timing deciles and found that G>C mutations (or equivalently C>G) showed the most pronounced differences across replication timing deciles (**Supplementary Figure 1g**). These results indicate that the

emergence of nullomers is more likely to occur in early replicating and coding regions, genomic regions with higher GC content which are under stronger selection constraints.

Putative nullomer-emerging mutations are enriched at functional genic sites in transcribed regions

We next examined the distribution of putative nullomer-emerging mutations relative to functional genomic sites at base-pair resolution. Promoter sequences are composed of specific regulatory elements such as the TATA-box, the INR element and TFBSs, which tend to be under evolutionary constraint. We therefore reasoned that putative nullomer-emerging mutations, which could impair transcriptional activity, would be enriched relative to transcription start sites (TSSs). Indeed, we found an enrichment of 1.79-fold immediately upstream of the TSS for mutations that cause the emergence of 13bp nullomers (**Figure 1f**). We also adjusted for the simulated mutations, finding an enrichment of 1.53-fold of putative nullomer-emerging mutations around the TSS (Kolmogorov-Smirnov test, p-value=4.07E-59, **Supplementary Figure 2a,** see **Methods**). In particular, the enrichment levels for substitutions, insertions and deletions were 1.79-fold, 1.76-fold and 1.99-fold (**Supplementary Figure 2b**), with the substitution type and indel type influencing the likelihood of putative nullomer emergence. Similar results were obtained for 11bp and 12bp putative nullomer-emerging mutations (**Supplementary Figure 2a-c**) and when adjusting for the simulated mutation controls. These results indicate that putative nullomer-emerging mutations are enriched relative to the TSSs.

We replicated this methodology relative to 3’ splice sites (3’ss) and 5’ splice sites (5’ss) as well as relative to Transcription End Sites (TESs). We found that the enrichment levels ranged across

these elements with 3'ss, 5'ss and TES showing enrichments of 2.07-fold, 2.10-fold and 1.23-fold respectively relative to surrounding regions (**Figure 1f-g**). We also adjusted for the simulated mutations, finding adjusted enrichments of 1.50-fold, 1.72-fold and 1.05-fold at the 3'ss, 5'ss and TES respectively (**Supplementary Table 5**). The 3'ss and 5'ss mutation types with lowest and highest adjusted enrichments were insertions and substitutions with 1.70-fold and 1.56-fold enrichments and were replicated for 11bp and 12bp with high consistency (**Supplementary Figure 2d-f**). The enrichment of putative nullomer-emerging mutations at TSS and splice sites indicates that these mutations are prone to affect transcriptional activity.

We also investigated if putative nullomer-emerging mutations are enriched for specific epigenetic modifications, including histone modifications and open chromatin marks. To that end, we analyzed H3K4me3, H3K27ac, H3K4me1 and DNaseI data across four human cell lines (**Figure 1h, Supplementary Figure 2g-h**). We find consistently that H3K27ac and H3K4me3 are most enriched for putative nullomer-emerging mutations (mean enrichments of 8.2-fold and 7.1-fold). We conclude that epigenetic marks are linked to differences in putative nullomer emergence frequencies.

Transcription factor binding sites show an excess of putative nullomer-emerging mutations

We next examined whether putative nullomer-emerging mutations are associated with certain TFBSs. First, we used collapsed consensus motifs from genome-wide DNase footprinting data generated from 263 cell and tissue types (Vierstra et al. 2020)(see **Methods**), which have been previously shown to reflect bound TFBSs (Galas and Schmitz 1978). We observed that 39.89% of DNase footprints overlapped one or more putative nullomer-emerging mutations representing

a 2.13-fold enrichment (**Supplementary Figure 3a**). When adjusting based on our simulated controls we find an enrichment of putative nullomer-emerging mutations of 1.56-fold, suggesting that nucleotide composition accounts for a subset of the pattern. We observed consistent patterns when we replicated this analysis across nullomer lengths and separating by mutation category (**Supplementary Figure 3a-c**), with deletions observed to have the strongest enrichment (1.70-fold) and substitutions (1.55-fold) showing the weakest enrichment.

To verify these findings, we used a collection of TFBSs derived from the UniBind database (Puig et al. 2021). In this dataset, for every ChIP-seq peak the corresponding TFBS is inferred and therefore it reflects high-confidence transcription factor bound TFBSs. We analyzed this dataset to examine potential enrichment of putative nullomer-emerging mutations within actively-bound TFBSs. We found that the results obtained were highly consistent with those obtained using DNase footprinting (**Figure 2a, Supplementary Figure 3d-f**), with 40.68% of TFBSs overlapping one or more putative nullomer-emerging mutation, representing an overall enrichment of 1.18-fold over background rates and providing additional support that nullomer emergence occurs more frequently at TFBSs. We were also interested in investigating potential differences in the enrichment of putative nullomer-emerging mutations by transcription factor category. We examined the density of putative nullomer-emerging mutations across the TFBSs of individual transcription factors; we observed that a subset of transcription factors had a high nullomer-emerging mutation density, with the highest densities occurring with ZBED1 and GMEB2 among others (**Figure 2b**). These findings suggest an enrichment of putative nullomer-emerging mutations at TFBSs across the human genome.

Putative nullomer-emerging mutations are preferentially positioned relative to nucleosomes

We reasoned that transcription factor binding site accessibility during transcriptional activity can be influenced by chromatin organization. Thus, we examined if nucleosome positioning influences the likelihood of nullomer emergence. Micrococcal Nuclease (MNase) data are generated by MNase digestion, in which exposed DNA regions are digested which in turns enables the derivation of nucleosome positioning. We used available MNase data for K562 and GM12878 cell lines from the ENCODE Consortium (ENCODE Project Consortium et al. 2020) to identify whether putative nullomer-emerging mutations show a preference for nucleosome core or linker regions. We found that 13bp putative nullomer-emerging mutations show an 1.08-fold enrichment at nucleosome core sequences, with a periodicity that is approximately the size of the inter-nucleosome distance (Sasaki et al. 2009) (**Figure 2c**). The results were also replicated for 11bp and 12bp mutations with very similar results obtained (**Supplementary Figure 4a**). However, when we separated by mutation type and repeated the same analysis we found significant differences. Overall, between substitutions, insertions and deletions we found largely consistent results with nucleosome cores displaying an enrichment for putative nullomer-emerging mutations (**Supplementary Figure 4b**); however when we separated by substitution type we observed that G>A (and C>T) and G>T (and C>A) mutations were more likely to be found at the linker regions, whereas all other substitution types were enriched for the nucleosome core sequence (**Supplementary Figure 4c**). Our findings indicate that nullomer emergence, as examined using all putative mutations, is influenced by nucleosome positioning and by the mutation type.

CpG sites are the primary determinants of nullomer emergence in the human genome

Previous reports have suggested that CpG sites are hypermutable; we therefore examined the association between nullomer emergence and presence of CpG dinucleotides (Sved and Bird 1990; Vergni and Santoni 2016). We found that putative nullomer-emerging mutations are highly enriched in CpG dinucleotides, with 100% (104), 100% (44,287) and 99.996% (2,347,572) of them harboring one or more CpG dinucleotides for 11bp, 12bp and 13bp respectively. This is particularly unexpected given that the percentage of kmers that harbor CpG dinucleotides in all possible 11,12, and 13bp kmers are 69%, 72%, and 75%. This result suggests that the primary driver of nullomer generation is presence of CpG dinucleotides. We also examined if putative nullomer-emerging mutations are enriched at CpG islands (binomial test, p-value = 0, **Figure 2d**, **Supplementary Table 6,** see **Methods**). We found that 22.4%, 95.8% and 96.9% of CpG islands contain one or more putative nullomer-emerging mutation for 11bp, 12bp and 13bp nullomers respectively, with 1.09-fold, 1.10-fold and 2.33-fold enrichments at substitutions, insertions and deletions respectively when compared to simulated mutations (**Supplementary Figure 4d-f**).

Next, we used whole-genome bisulfite sequencing (WGBS) from the ENCODE Consortium (ENCODE Project Consortium 2012) on six different tissues (adrenal gland, esophagus squamous epithelium, gastroesophageal sphincter, stomach, small intestine, spleen) to examine methylation of DNA at the fifth position in cytosine (5mC), across putative nullomer-emerging mutation sites. Across all putative nullomer-emerging mutations, we observed an enrichment of 2.31-fold, directly at the 5mC sites (Mann Whitney U-test, p-value=0, **Figure 2e, Supplementary Figure 3g, Supplementary Table 5**), findings that were consistent across

mutation categories (**Supplementary Figure 3h-i**). These results indicate that putative nullomer-emerging mutations occur at CpG sites, which are the most frequently mutated dinucleotides in the human genome (Fryxell and Moon 2005); therefore these enrichments likely reflect hyper-mutation at these sequences. We conclude that a driver of nullomer generation in the human genome is the CpG mutation rate.

Alu repeats display positional enrichment for putative nullomer-emerging mutations

We examined the distribution of putative nullomer-emerging mutations across transposable elements. We analyzed emerging nullomers in Long interspersed nuclear elements (LINE), Short interspersed nuclear elements (SINE) and Long Terminal Repeats (LTR) transposable element families and found that the highest nullomer-emerging density from all putative one bp mutations was observed at SINE repeats with median of 60 mutations per kB (**Figure 3a**). In particular, we observed this for insertions relative to substitutions and deletions and the findings were consistent across kmer lengths (**Figure 3a; Supplementary Figure 5**). We also separated the transposable repeat element families into sub-types and found that the most recently active Alu repeats in the human genome, including AluSx, AluJ and AluY repeats displayed the highest putative nullomer-emerging mutational density (**Figure 3b**), and specifically for insertions (**Figure 3c**). The repeat elements with the highest putative nullomer-emerging mutation density were AluYe5, AluYk12 and AluYb9 (mean densities 155.4, 145.0, 139.5 elements per kb, **Figure 3c**).

We next examined the subset of nullomers that had the highest density of putative nullomer-emerging mutations for single base-pair substitutions, insertions and deletions in the human

genome. We find that the recurrent putative nullomer-emerging mutations can be explained by their presence at recently evolved Alu repeats, particularly for insertions (**Figure 3d**). Additionally, we observe that the mutations are inhomogeneously distributed across the Alu repeat elements (**Figure 3e**), with specific hotspots, notably at regions with AluYa5. These findings provide support for putative nullomer-emerging mutational hotspots in the human genome, likely reflecting selection constraints and silencing mechanisms that have been operative during recent human evolution.

Nullomer-emergence in human population variants

Previous work has showcased the utility of nullomers in forensic applications (Goswami et al. 2013). We have also previously shown an association between population variants and nullomer emergence (Georgakopoulos-Soares, Yizhar-Barnea, Mouratidis, Hemberg, et al. 2021). Here, we examined the likelihood of nullomer emergence due to population variants, to be set as a background against which to examine the nullomer pathogenicity. We first examined the likelihood of nullomer emergence due to human population variants. Interestingly, we observe that the SNP variant allele frequency is anti-correlated with the likelihood of nullomer emergence (**Figure 4a**). We find that nullomer-emerging germline mutations were more likely to be found in the rarest population variants (MAF <0.01). To understand whether rare germline mutations that can lead to nullomers emergence may have a deleterious effect, we then examined the pathogenicity of population variants. Using scores derived from CADD (Rentzsch et al. 2019) and population variants derived from gnomAD (Karczewski et al. 2020), we examined the predicted pathogenicity of each mutation collected from the human genome. We find that across mutation types, nullomer-emerging germline mutations found in gnomAD yielded a higher

CADD score, with insertions yielding the strongest enrichment (**Figure 4b,** 1.58-fold**,** Mann-Whitney U test, p-value<0.01, **Supplementary Figure 6a, Supplementary Table 7)**. These findings suggest that nullomer emerging germline mutations are rare in the human population and associated with pathogenicity.

Predicted pathogenic mutations are more likely to result in nullomer emergence

Assuming that some of the causes of nullomer absence include selection constraints and pathogenicity, we examined the pathogenicity of nullomer-emerging mutations. We analyzed the deleteriousness of all possible putative nullomer-emerging substitutions relative to all putative substitutions that do not cause nullomer emergence, throughout the human genome. We found that putative nullomer-emerging substitutions on average display a higher CADD score for 13bp nullomer emerging mutations compared to controls (1.11-fold, Mann-Whitney U, p-value<0.01, **Supplementary Table 7**). We also separated putative substitutions into ten quantiles based on the deleteriousness of each substitution. We found that the most pathogenic putative substitution mutations are also the most likely to cause nullomer-emergence (**Figure 4c**, Spearman correlation = 0.92, p-value = 1.86E-5). Furthermore, when separating putative nullomer-emerging mutations and mutations that do not cause nullomer emergence into deciles based on the CADD score, across the deciles putative nullomer-emerging mutations have a higher pathogenicity (**Figure 4d**).

We next separated the mutation types based on their effect into: i) 3'UTR, ii) 5'UTR, iii) canonical splice, iv) stop gained, v) stop loss, vi) synonymous and vii) non-synonymous. We examined which of these showed the largest discrepancy in pathogenicity when they caused

nullomer-emerging versus when they did not cause nullomer-emerging. We observed that putative mutations in 5'UTR showed the strongest enrichment (1.17-fold) while the 3'UTR were under-enriched (0.97-fold) (**Figure 4e**, **Supplementary Table 7**). We obtained consistent results across nullomer lengths (**Supplementary Figure 6b**).

We then examined if putative nullomer-emerging mutations are enriched in amino acid substitutions that are predicted to affect protein function. Using SIFT scores we show that putative nullomer-emerging mutations are slightly more likely to affect protein function (**Figure 4f,** 1.02-fold, Mann-Whitney U, p-value=0, **Supplementary Table 8**). These findings provide evidence for the higher likelihood of pathogenicity for putative nullomer-emerging substitutions across mutation types, for mutations that cause protein sequence changes and across functional genomic compartments.

Pathogenic mutations cause the emergence of nullomers in the human genome

ClinVar is a database that encompasses manually curated mutations that have been annotated relative to their pathogenicity status as "Benign" or "Likely Benign" and "Pathogenic" or "Likely Pathogenic" (Landrum et al. 2016). We examined if clinically relevant nullomer-emerging mutations show an enrichment for Pathogenic or Likely Pathogenic relative to Benign or Likely Benign mutations. We find that across nullomer lengths, pathogenic mutations are more likely to cause the emergence of nullomers (**Supplementary Figure 6c,** chi-square test, p-value=3.26E-243, **Supplementary Table 9**). We find that pathogenic clinical variants show a 1.34-fold enrichment in causing 13bp nullomer emergence over the frequency of benign and likely benign clinical variants, in particular for deletions and insertions relative to substitutions

(**Figure 4g**). Across substitution types, T>C mutations are the most enriched type (**Supplementary Figure 6d**). We also examined the frequency of nullomer-emerging mutations in expression quantitative trait loci (eQTL) relative to common SNPs and found a 1.18-fold enrichment at eQTLs for 13bp nullomers (**Figure 4h, Supplementary Table 10**). This was replicated across mutation types (**Supplementary Figure 6e).** These results indicate that nullomer emergence is associated with pathogenicity and human disease.

2.4 Discussion

Here, we performed a thorough genomic characterization of nullomer-emerging mutations across the human genome and provided new evidence that increased mutation rate in methylated cytosines (Mugal and Ellegren 2011) are the primary driver of nullomer emergence. We find a remarkable 16.12-fold enrichment of nullomer-emerging mutations in 5-methylcytosines. We also find almost every nullomer encompassing one or more CpG dinucleotides, suggesting that hypermutation of CpG loci is the principal force of nullomer formation, consistent with the proposition by (Acquisti et al. 2007).

We also find that putative nullomer-emerging mutations are inhomogeneously distributed across the human genome, with clear enrichment patterns across functional genetic elements and *cis*-regulatory sequences. Therefore, putative nullomer-emerging mutations provide evidence for negative selection constraints within those elements, which is the second force driving nullomer formation. Selection constraints against nullomers are highest at TFBSs, splice sites and the TSS. Previously, we and others have shown that selection constraint can be detected for nullomers in

humans as well as in other species (Koulouras and Frith 2021; Georgakopoulos-Soares, Yizhar-Barnea, Mouratidis, Hemberg, et al. 2021).

Selection constraints against nullomer emergence are evident at functional genomic sites. In addition, the intriguing putative nullomer-emerging mutation enrichment at the most recently evolved Alu repeats, might indicate repeat silencing mechanisms. Specifically, we find that only the most recently evolved Alu repeats, including the small subset of Alu repeat sub-families that are still active in humans (Bennett et al. 2008; Häsler and Strub 2006), show the pronounced nullomer emergence hotspots. The mutation type we observed at these hotspots is primarily insertions, suggesting that the silencing mechanisms which are operative could likely involve deletion events. Therefore, future work is required to investigate the hypothesis that nullomer emergence at these sites can cause an increased Alu repeat activity.

Finally, we showcase how disease-causing and pathogenic mutations are more likely to create nullomer-emerging mutations. In addition, rare germline mutations, which are more likely to be pathogenic, are also more likely to cause the emergence of nullomers. As a result, a better understanding of nullomers and their emergence, a topic which has been severely understudied, can provide breakthroughs in the understanding of human diseases. Nullomer emergence could be used to find loci that are more likely to be associated with human diseases and could be particularly useful in other species, for which disease annotations and identification of pathogenic variants is still lacking. Future work is required to showcase the mechanisms of pathogenicity at individual nullomer-emerging hotspots and to deconvolute them from the increased mutagenicity of nullomers.

2.5 Materials and Methods

Nullomer extraction and putative nullomer-emerging mutation map generation. Nullomer extraction and putative nullomer-emerging mutation map generation was performed as described previously (Georgakopoulos-Soares, Yizhar-Barnea, Mouratidis, Hemberg, et al. 2021) for kmer lengths of eleven to thirteen base pairs. Briefly, at each base pair, each nucleotide is changed to all three other possibilities and the resulting kmer is compared to a list of nullomers. Mutation types were separated in substitutions, insertions and deletions. Substitutions were subdivided based on the reference and alternate allele in six subtypes.

Simulated mutations For each putative nullomer-emerging mutation, we identified a position with a distance less than 1,000 bp away that had the same trinucleotide context but did not lead to nullomer emergence. Simulated mutations were generated using a custom Python script that is found in (Georgakopoulos-Soares et al. 2018). Using this methodology we created a set of simulated mutations that matched the nullomer-emerging mutations, for each nullomer length.

K-mer analysis across genomic bins To investigate the association between k-mer diversity and nullomer emerging mutations we split the genome in 1kB or 50kB or 500kB bins and calculated the number of unique k-mers per bin and the number of putative nullomer emerging mutations in each bin. From the analysis, we removed the first and last 50kB regions. To estimate statistical significance we performed Mann Whitney U tests.

Genomic and genic annotation data. The reference human genome assembly GRCh38 (hg38) was used. Genic analyses were performed using the GENCODE v40 annotation, for which coding and non-coding regions, TSS, TES, 3'ss and 5'ss annotations were derived. Locations of CpG islands were obtained from the UCSC genome browser. The enrichment was calculated as the number of occurrences at a position over the mean number of occurrences across the window of 1kB. The corrected enrichment was calculated as the ratio of the real enrichment over the background enrichment of simulated mutations.

Repli-seq data. Repli-seq data for fourteen cell lines were derived from (ENCODE Project Consortium 2012) and analyzed as previously described in (Morganella et al. 2016). Repli-seq data were binned into deciles relative to early and late replicating regions. The density per 1kB window of nullomer-emerging mutations was calculated at each decile across mutation categories for all cell lines. Pearson correlation was calculated between the decile number and the mean mutational density at each decile.

TFBS datasets. TFBSs at DNase footprints from 243 human cell and tissue types and states were obtained from (Vierstra et al. 2020) and ChIP-seq bound TFBSs were obtained from UniBind (Puig et al. 2021). We measured the distribution of nullomer or simulated mutations across 1kB windows. The enrichment was calculated as the number of occurrences at a position over the mean number of occurrences across the window of 1kB. The corrected enrichment was calculated as the ratio of the real enrichment over the background enrichment of simulated mutations. The density of TFBS motifs was calculated as the number of motif occurrences over the total number of base pairs.

MNase datasets. MNase-seq data were downloaded from the ENCODE (ENCODE Project Consortium 2012) portal for GM12878 and K562. Significance of difference in nucleosome density signal was calculated using scores from nullomer or control mutations extracted with bedtools map function, followed by Mann-Whitney U test. Nucleosome signal was calculated as mean score at each loci in the 1kB window of nullomer-emerging mutation and simulated control mutation.

ClinVar, dbSNP mutation and eQTL datasets. Population variants were derived from dbSNP (Sherry, Ward, and Sirotkin 1999) and gnomAD (Karczewski et al. 2020). Clinical variants were derived from the ClinVar database (Landrum et al. 2016) and were subdivided based on pathogenicity in "Benign", "Likely Benign", "Likely Pathogenic" and "Pathogenic". The frequency of nullomer-emergence was compared between mutations of different pathogenicity. eQTLs were derived from GTEx Portal (GTEx_Analysis_v8_eQTL.tar)(Lonsdale et al. 2013).

Measurement of deleteriousness using CADD. The CADD tool was used for scoring the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome. All single nucleotide variants across the human genome were derived from: https://krishna.gs.washington.edu/download/CADD/v1.6/GRCh38/whole_genome_SNVs.tsv.gz. All gnomAD substitutions were derived from: https://krishna.gs.washington.edu/download/CADD/v1.6/GRCh38/gnomad.genomes.r3.0.snv.tsv.gz. All gnomAD insertion and deletion mutations were derived from: https://krishna.gs.washington.edu/download/CADD/v1.6/GRCh38/gnomad.genomes.r3.0.indel.tsv.gz. All SIFT scores were derived from:

https://krishna.gs.washington.edu/download/CADD/v1.6/GRCh38/whole_genome_SNVs_inclAnno.tsv.gz. Nullomer extraction was performed for each mutation for nullomers of lengths between eleven and thirteen base pairs. Mutations were separated into those that caused nullomer emergence and those that did not and the significance of difference in deleteriousness between the two groups was calculated using Mann-Whitney U tests.

Whole Genome Bisulfite Sequencing Analysis. WGBS data were downloaded from ENCODE for six different human tissues, the adrenal gland (ENCFF524MTO), esophagus squamous epithelium (ENCFF283YAZ), the gastroesophageal sphincter (ENCFF441OSB), stomach (ENCFF896GOF), small intestine (ENCFF537NCQ) and spleen (ENCFF865OXJ) tissues. Methylation signal was calculated as the mean score at each loci in the 1kB window of nullomer-emerging mutation and simulated control mutation.

Code Availability

All code to perform case study analysis is provided at https://github.com/Georgakopoulos-Soares-lab/nullomer_topography.

2.6 Author contributions

C.C. N.A., and I.G.S. conceived the study. C.C, I.M. A.M. and I.G.S., wrote the code, C.C., A.M, and I.G.S., performed the analyses and generated the visualizations. N.A. and I.G.S. supervised the research, I.G.S., and C.C wrote the manuscript with input from all authors.

2.7 Acknowledgements

I.G.S, I.M, A.M were funded by the startup funds from the Penn State College of Medicine. C.S.Y.C was funded in part by UCSF Hillblom Center for the Biology of Aging and Bakar Aging Research Institute Graduate Fellowship.

2.8 References

Acquisti, Claudia, George Poste, David Curtiss, and Sudhir Kumar. 2007. "Nullomers: Really a Matter of Natural Selection?" *PloS One* 2 (10): e1022.

Alileche, Abdelkrim, Jayita Goswami, William Bourland, Michael Davis, and Greg Hampikian. 2012. "Nullomer Derived Anticancer Peptides (NulloPs): Differential Lethal Effects on Normal and Cancer Cells in Vitro." *Peptides*. https://doi.org/10.1016/j.peptides.2012.09.015.

Alileche, Abdelkrim, and Greg Hampikian. 2017. "The Effect of Nullomer-Derived Peptides 9R, 9S1R and 124R on the NCI-60 Panel and Normal Cell Lines." *BMC Cancer* 17 (1): 533.

Ali, Nilufar, Cody Wolf, Swarna Kanchan, Shivakumar R. Veerabhadraiah, Laura Bond, Matthew W. Turner, Cheryl L. Jorcyk, and Greg Hampikian. 2024. "9S1R Nullomer Peptide Induces Mitochondrial Pathology, Metabolic Suppression, and Enhanced Immune Cell Infiltration, in Triple-Negative Breast Cancer Mouse Model." *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie* 170 (January): 115997.

Aran, Dvir, Gidon Toperoff, Michael Rosenberg, and Asaf Hellman. 2011. "Replication Timing-Related and Gene Body-Specific Methylation of Active Human Genes." *Human Molecular Genetics* 20 (4): 670–80.

Bennett, E. Andrew, Heiko Keller, Ryan E. Mills, Steffen Schmidt, John V. Moran, Oliver Weichenrieder, and Scott E. Devine. 2008. "Active Alu Retrotransposons in the Human Genome." *Genome Research* 18 (12): 1875–83.

Chantzi, Nikol, Ioannis Mouratidis, Manvita Mareboina, Maxwell A. Konnaris, Austin Montgomery, and Ilias Georgakopoulos-Soares. 2023. "The Determinants of the Rarity of Nucleic and Peptide Short Sequences in Nature." *bioRxiv*.

https://doi.org/10.1101/2023.09.24.559219.

ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, et al. 2020. “Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes.” *Nature* 583 (7818): 699–710.

Fryxell, Karl J., and Won-Jong Moon. 2005. “CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content.” *Molecular Biology and Evolution* 22 (3): 650–58.

Galas, D. J., and A. Schmitz. 1978. “DNAse Footprinting: A Simple Method for the Detection of Protein-DNA Binding Specificity.” *Nucleic Acids Research* 5 (9): 3157–70.

Georgakopoulos-Soares, Ilias, Sandro Morganella, Naman Jain, Martin Hemberg, and Serena Nik-Zainal. 2018. “Noncanonical Secondary Structures Arising from Non-B DNA Motifs Are Determinants of Mutagenesis.” *Genome Research* 28 (9): 1264–71.

Georgakopoulos-Soares, Ilias, Ofer Yizhar-Barnea, Ioannis Mouratidis, Rachael Bradley, Ryder Easterlin, Candace Chan, Emmalyn Chen, John S. Witte, Martin Hemberg, and Nadav Ahituv. 2021. “Leveraging Sequences Missing from the Human Genome to Diagnose Cancer.” *medRxiv*.

Georgakopoulos-Soares, Ilias, Ofer Yizhar-Barnea, Ioannis Mouratidis, Martin Hemberg, and Nadav Ahituv. 2021. “Absent from DNA and Protein: Genomic Characterization of Nullomers and Nullpeptides across Functional Categories and Evolution.” *Genome Biology* 22 (1): 245.

Goswami, Jayita, Michael C. Davis, Tim Andersen, Abdelkrim Alileche, and Greg Hampikian.

2013. “Safeguarding Forensic DNA Reference Samples with Nullomer Barcodes.” *Journal of Forensic and Legal Medicine* 20 (5): 513–19.

Hampikian, Greg, and Tim Andersen. 2007. “Absent Sequences: Nullomers and Primes.” *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 355–66.

Häsler, Julien, and Katharina Strub. 2006. “Alu Elements as Regulators of Gene Expression.” *Nucleic Acids Research* 34 (19): 5491–97.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43.

Koulouras, Grigorios, and Martin C. Frith. 2021. “Significant Non-Existence of Sequences in Genomes and Proteomes.” *Nucleic Acids Research* 49 (6): 3139–55.

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2016. “ClinVar: Public Archive of Interpretations of Clinically Relevant Variants.” *Nucleic Acids Research* 44 (D1): D862–68.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature Genetics* 45 (6): 580–85.

Montgomery, Austin, Georgios Christos Tsiatsianis, Ioannis Mouratidis, Candace S. Y. Chan, Maria Athanasiou, Anastasios D. Papanastasiou, Verena Kantere, et al. 2023. “Utilizing Nullomers in Cell-Free RNA for Early Cancer Detection.” *medRxiv*. https://doi.org/10.1101/2023.06.10.23291228.

Morganella, Sandro, Ludmil B. Alexandrov, Dominik Glodzik, Xueqing Zou, Helen Davies, Johan Staaf, Anieta M. Sieuwerts, et al. 2016. “The Topography of Mutational Processes

in Breast Cancer Genomes." *Nature Communications* 7 (May): 11383.

Mouratidis, Ioannis, Fotis A. Baltoumas, Nikol Chantzi, Candace S. Y. Chan, Austin Montgomery, Maxwell A. Konnaris, George C. Georgakopoulos, et al. 2023. "kmerDB: A Database Encompassing the Set of Genomic and Proteomic Sequence Information for Each Species." *bioRxiv*. https://doi.org/10.1101/2023.11.13.566926.

Mouratidis, Ioannis, Candace S. Y. Chan, Nikol Chantzi, Georgios Christos Tsiatsianis, Martin Hemberg, Nadav Ahituv, and Ilias Georgakopoulos-Soares. 2023. "Quasi-Prime Peptides: Identification of the Shortest Peptide Sequences Unique to a Species." *NAR Genomics and Bioinformatics* 5 (2): lqad039.

Mouratidis, Ioannis, Maxwell A. Konnaris, Nikol Chantzi, Candace S. Y. Chan, Austin Montgomery, Fotis A. Baltoumas, Michail Patsakis, et al. 2023. "Nucleic Quasi-Primes: Identification of the Shortest Unique Oligonucleotide Sequences in a Species." *bioRxiv*. https://doi.org/10.1101/2023.12.12.571240.

Mugal, Carina F., and Hans Ellegren. 2011. "Substitution Rate Variation at Human CpG Sites Correlates with Non-CpG Divergence, Methylation Level and GC Content." *Genome Biology* 12 (6): R58.

Patel, Ami, Jessica C. Dong, Brett Trost, Jason S. Richardson, Sarah Tohme, Shawn Babiuk, Anthony Kusalik, Sam K. P. Kung, and Gary P. Kobinger. 2012. "Pentamers Not Found in the Universal Proteome Can Enhance Antigen Specific Immune Responses and Adjuvant Vaccines." *PloS One* 7 (8): e43802.

Pratas, Diogo, and Jorge M. Silva. 2021. "Persistent Minimal Sequences of SARS-CoV-2." *Bioinformatics* 36 (21): 5129–32.

Puig, Rafael Riudavets, Paul Boddie, Aziz Khan, Jaime Abraham Castro-Mondragon, and

Anthony Mathelier. 2021. “UniBind: Maps of High-Confidence Direct TF-DNA Interactions across Nine Species.” *BMC Genomics* 22 (1): 482.

Rentzsch, Philipp, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. 2019. “CADD: Predicting the Deleteriousness of Variants throughout the Human Genome.” *Nucleic Acids Research* 47 (D1): D886–94.

Sasaki, Shin, Cecilia C. Mello, Atsuko Shimada, Yoichiro Nakatani, Shin-Ichi Hashimoto, Masako Ogawa, Kouji Matsushima, et al. 2009. “Chromatin-Associated Periodicity in Genetic Variation Downstream of Transcriptional Start Sites.” *Science* 323 (5912): 401–4.

Sherry, S. T., M. Ward, and K. Sirotkin. 1999. “dbSNP-Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation.” *Genome Research* 9 (8): 677–79.

Silva, Raquel M., Diogo Pratas, Luísa Castro, Armando J. Pinho, and Paulo J. S. G. Ferreira. 2015. “Three Minimal Sequences Found in Ebola Virus Genomes and Absent from Human DNA.” *Bioinformatics* 31 (15): 2421–25.

Stamatoyannopoulos, John A., Ivan Adzhubei, Robert E. Thurman, Gregory V. Kryukov, Sergei M. Mirkin, and Shamil R. Sunyaev. 2009. “Human Mutation Rate Associated with DNA Replication Timing.” *Nature Genetics* 41 (4): 393–95.

Suzuki, Masako, Mayumi Oda, María-Paz Ramos, Marién Pascual, Kevin Lau, Edyta Stasiek, Frederick Agyiri, et al. 2011. “Late-Replicating Heterochromatin Is Characterized by Decreased Cytosine Methylation in the Human Genome.” *Genome Research* 21 (11): 1833–40.

Sved, J., and A. Bird. 1990. “The Expected Equilibrium of the CpG Dinucleotide in Vertebrate

Genomes under a Mutation Model." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.87.12.4692.

Tsiatsianis, Georgios Christos, Candace S. Y. Chan, Ioannis Mouratidis, Nikol Chantzi, Anna Maria Tsiatsiani, Nelson S. Yee, Apostolos Zaravinos, Verena Kantere, and Ilias Georgakopoulos-Soares. 2024. "Peptide Absent Sequences Emerging in Human Cancers." *European Journal of Cancer* 196 (January): 113421.

Vergni, Davide, and Daniele Santoni. 2016. "Nullomers and High Order Nullomers in Genomic Sequences." *PloS One* 11 (12): e0164540.

Vierstra, Jeff, John Lazar, Richard Sandstrom, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, et al. 2020. "Global Reference Mapping of Human Transcription Factor Footprints." *Nature* 583 (7818): 729–36.

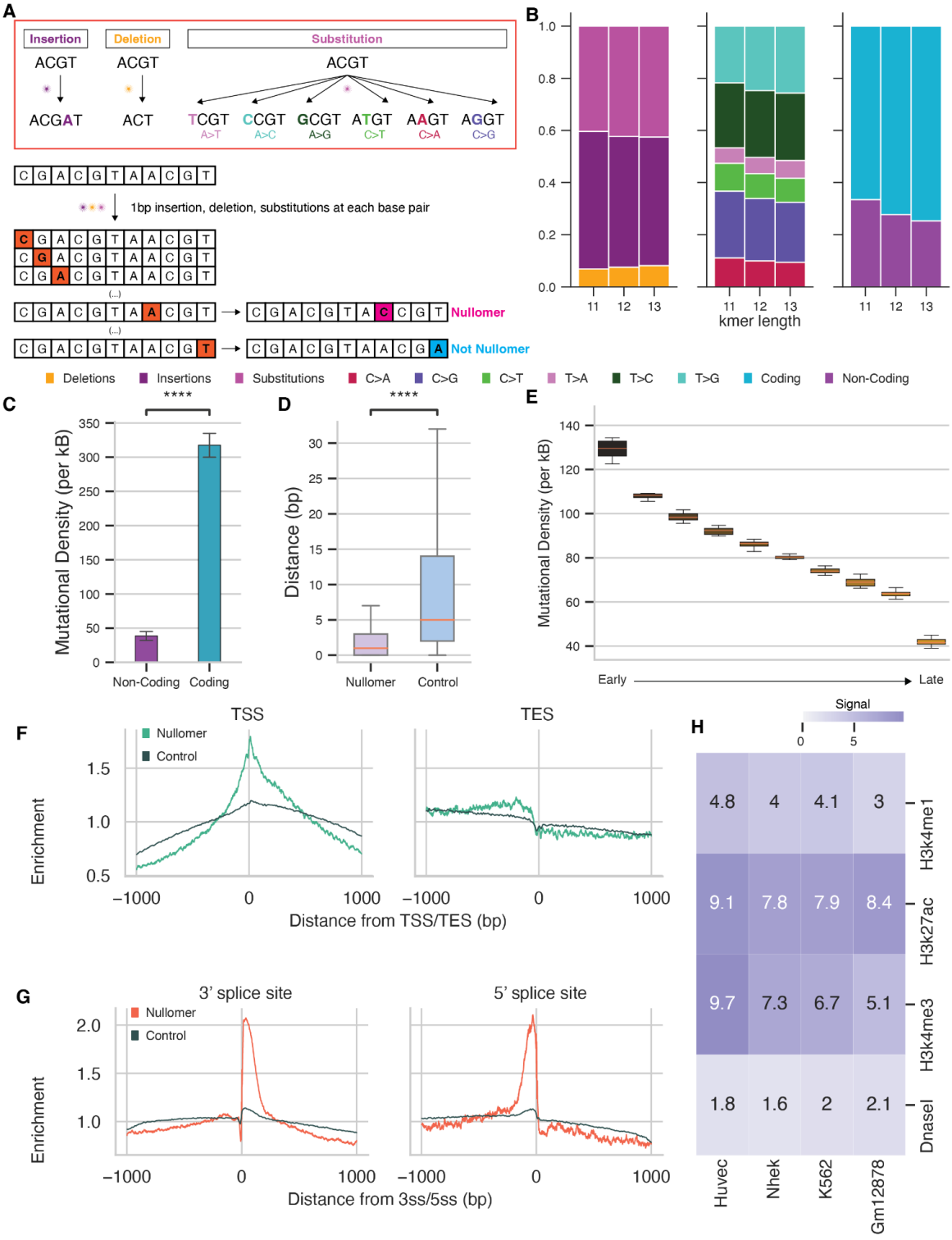

Figure 2.1 Putative nullomer-emerging mutations show clustering patterns and are enriched in early-replicating regions, promoters and coding sequences.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

a) Schematic of simulated mutations to generate putative nullomer-emerging mutations and control mutations. b) Proportion of putative nullomer-emerging mutations by mutation type for kmer lengths of 11bp, 12bp and 13bp. c) Distance distribution between consecutive putative nullomer-emerging mutations and simulated putative nullomer-emerging mutations. d) putative nullomer-emerging mutational density at coding and non-coding regions. e) Density of putative nullomer-emerging mutations across replication timing deciles. Early replicating regions display a higher nullomer emergence density (Pearson correlation r = 0.97, p-value = 2.4e-06). Mean mutational density across all fourteen cell lines are shown. Error bars indicate standard deviation of mutational density between cell lines. f-h) Enrichment of putative nullomer-emerging mutations in : f) TSS, TES and g) 3'ss, 5'ss, and h) at DNase footprinting sites and histone modifications. In f-g, the fold enrichment for putative nullomer-emerging mutations was calculated as the ratio of the number of mutations found in a given position over the mean number of mutations across the whole window. The dark grey lines in f-g represent the distribution of putative nullomer-emerging mutations for simulated controls.

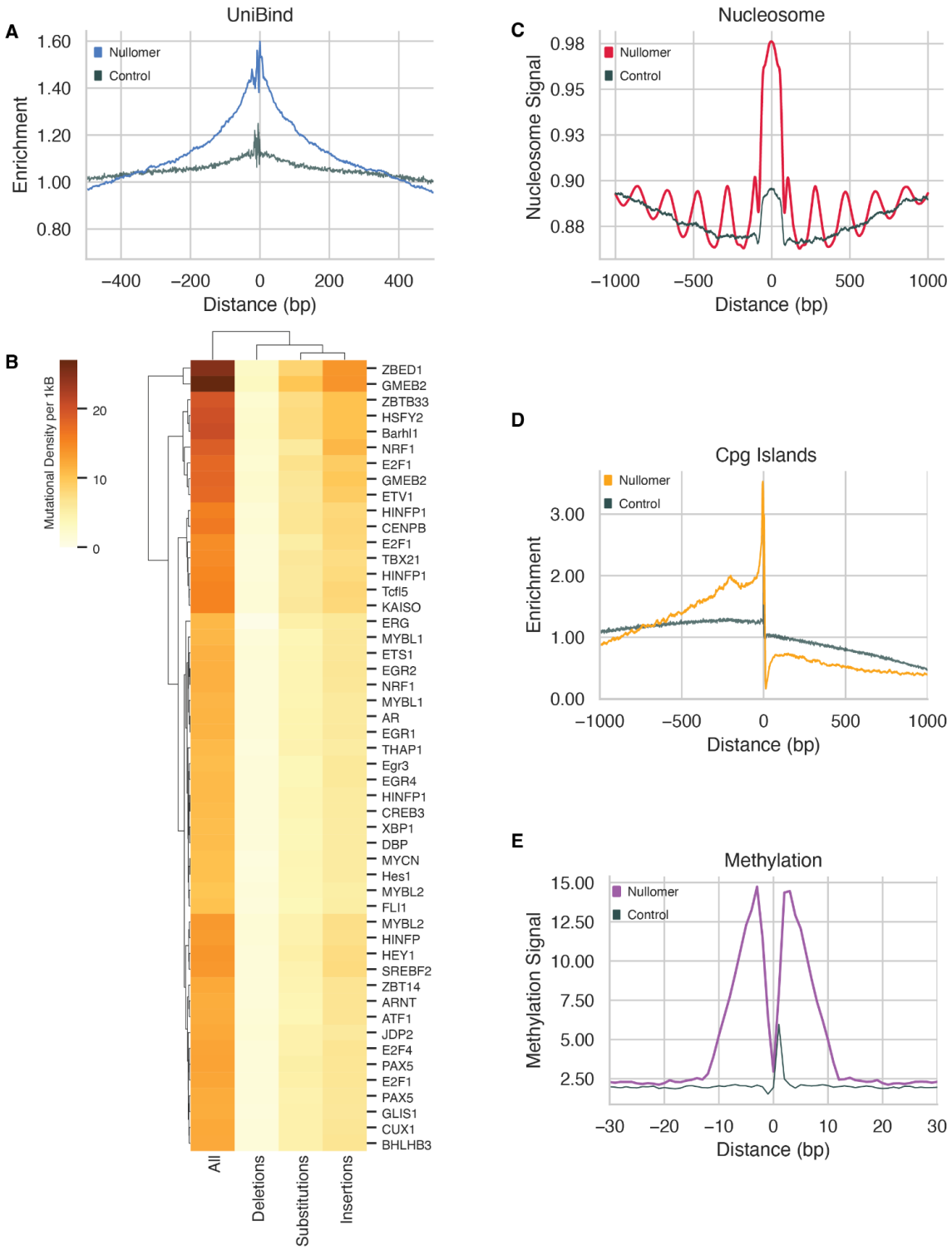

Figure 2.2 Association between putative nullomer-emerging mutations and open epigenetic marks and methylation.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

Putative nullomer-emerging mutation sites are enriched at a) TF-DNA interaction sites based on UniBind inferred data, b) transcription factor binding sites, c) nucleosome core positions, d) CpG islands, and e) methylation sites from WGBS.

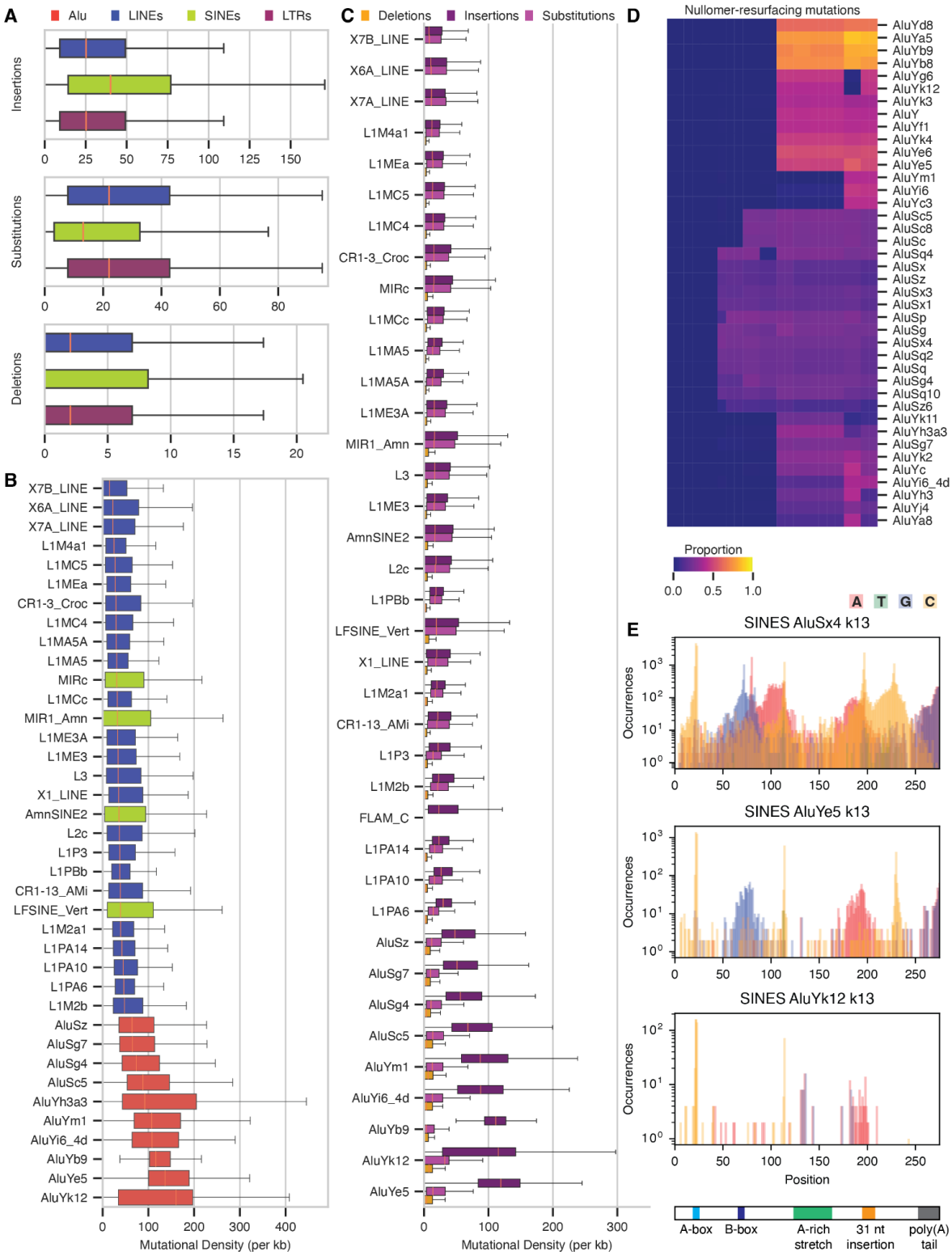

Figure 2.3 Nullomer emergence is pronounced at Alu repeat elements.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
a) Mutational density at LINE, SINE and LTR repeats across mutation types. b) Mutational density at transposable element repeat sub-families. c) Mutational density at transposable element repeat sub-families for the mutation subtypes. d) Heatmap of proportion of putative nullomer-emerging mutations appearing at Alu repeat elements. e) Distribution of putative nullomer-emerging mutations occurrences across Alu repeat elements AluYe5, AluYk12, AluSx4.

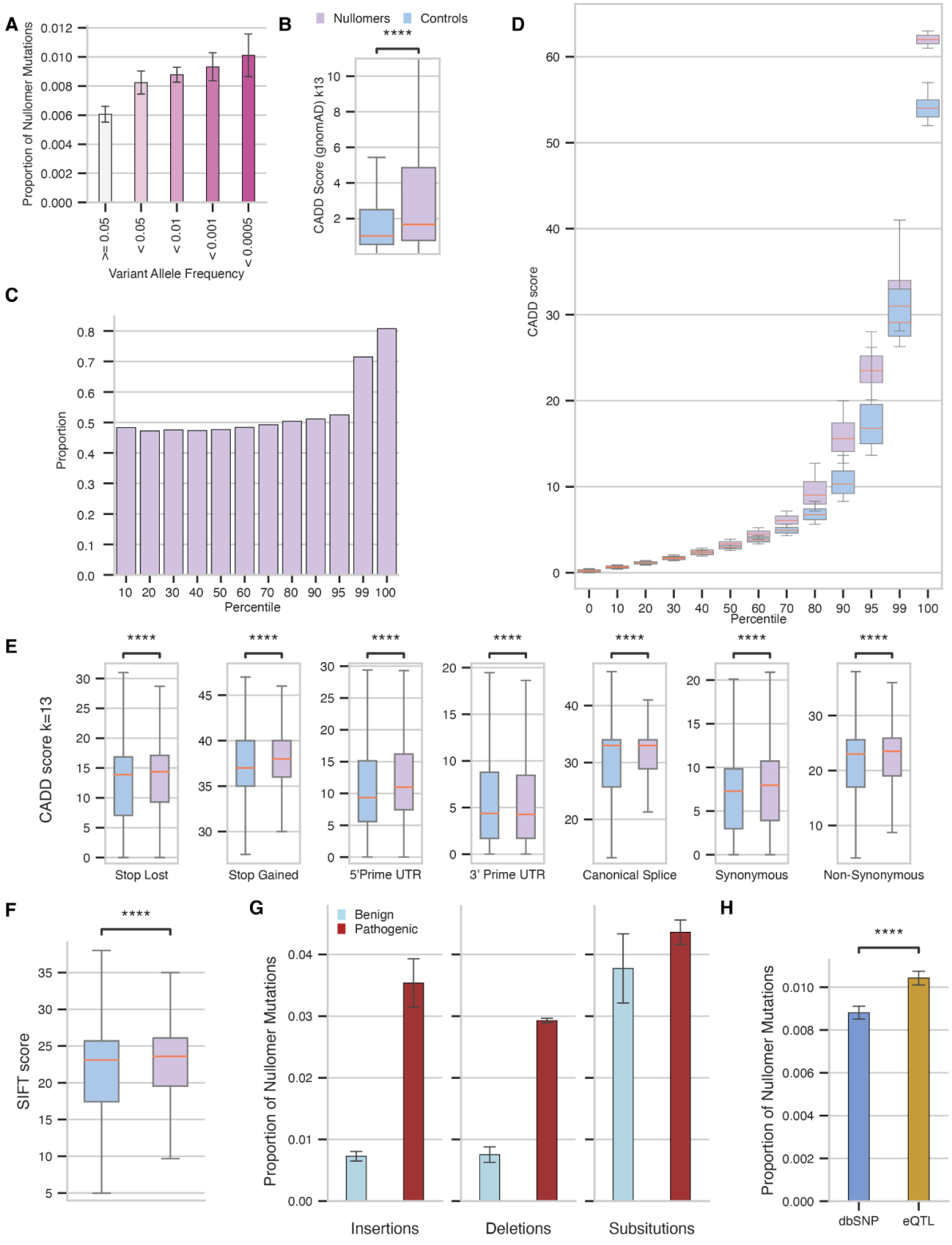

Figure 2.4 Pathogenicity of nullomer-emerging sequences in the human genome.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

a) Variant allele frequencies of nullomer-emerging population variants. b) CADD scores of mutations that cause or not cause the emergence of nullomers in population variants from gnomAD. c) Proportion of nullomer-emerging mutations across CADD score percentiles. d) Distribution of CADD scores across percentiles in nullomer-emerging mutations and non-nullomer emerging mutations. e) Association of CADD score and nullomer emergence for mutations at the stop codon mutation loss, stop stop codon mutation gain, 5'UTR, 3'UTR, canonical splice sites, synonymous and nonsynonymous mutations. f) SIFT scores from nullomer emerging mutations and non-nullomer emerging mutations. g) ClinVar pathogenic mutations are more likely to cause nullomer emergence than their benign counterparts. h) Enrichment of eQTLs relative to common SNPs for nullomer-emerging mutations. In panels a-f, controls are based on simulations controlling for trinucleotide context and proximity to the original nullomer emerging mutation.

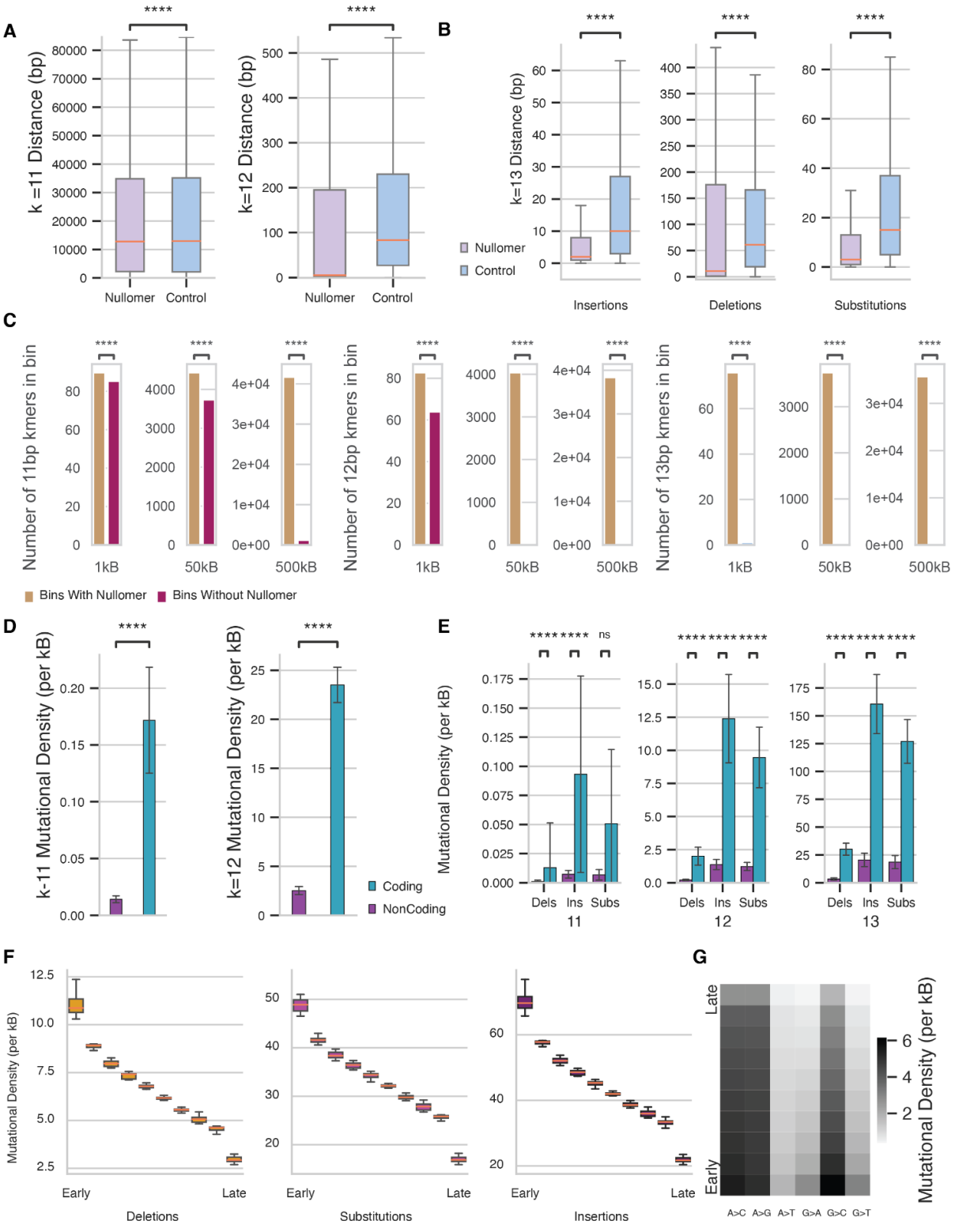

Figure S2.1 Enrichment across genome of putative nullomer-emerging mutations

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

Distance distribution between consecutive nullomer-emerging mutations and simulated nullomer-emerging mutations for a) k=11, k=12 and b) across mutation types for k=13. c) Barplot of the number of unique kmers in 1kb, 50kb, and 500kb bins of the genome with or without putative nullomer-emerging mutations d) Nullomer-emerging mutational density at coding and non-coding regions for k=11, 12, and e) and across mutation types for k=13. f) Mean mutational density of nullomer-emerging mutations for k=13 across replication stages in deletions, substitutions, insertions and g) substitution subtypes.

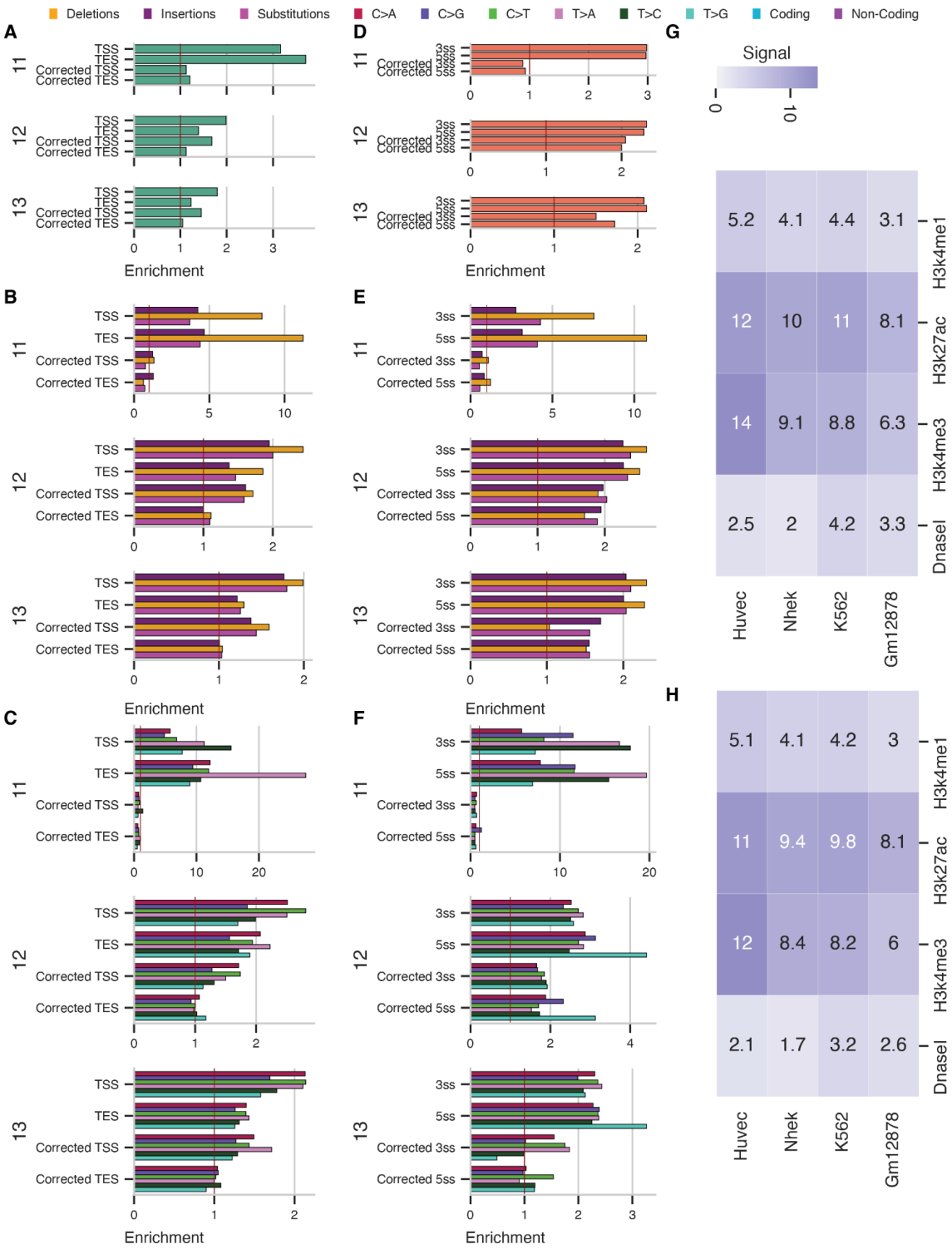

Figure S2.2 Enrichment of putative nullomer-emerging mutations in functional regions of genomes across kmers and mutation subtypes

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

Enrichment of nullomer-emerging mutations in TSS and TES a) in k=11, 12, 13, b) stratified by mutation type and c) substitution subtypes. Enrichment of nullomer-emerging mutations in 3' and 5' splice sites in d) k=11,12,13 e) stratified by mutation type and f) substitution subtypes. Enrichment of nullomer-emerging mutations in DNase footprinting sites and histone modification sites in g) k=11 and h) k=12.

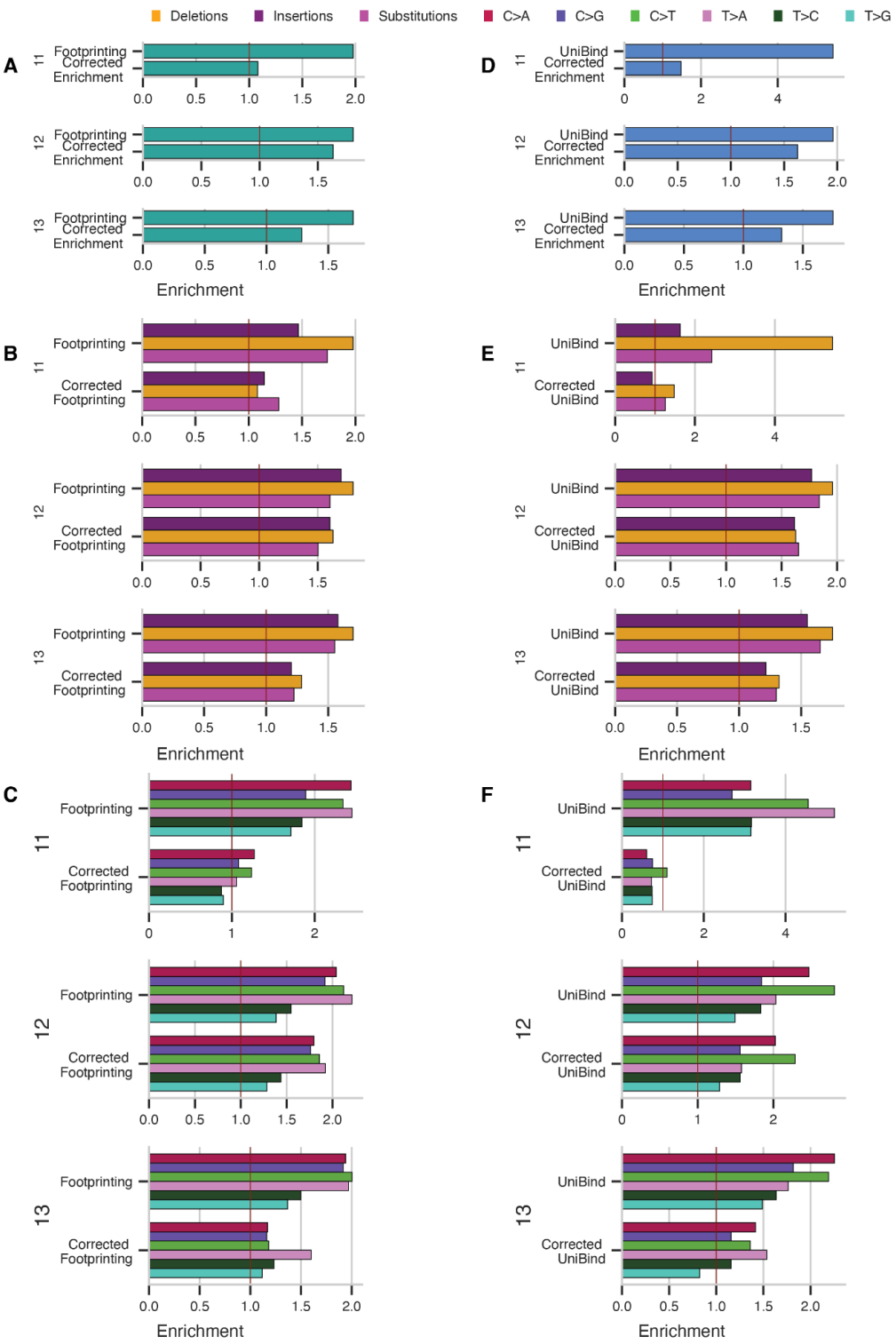

Figure S2.3 Enrichment of putative nullomer-emerging mutations and TF-DNA interaction sites across kmers and mutation subtypes

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

Enrichment of nullomer-emerging mutations across k=11,12,13, mutation subtypes, and substitution subtypes for TF-DNA interaction sites based on a-c) DNase footprinting d-f) and sites derived from UniBind.

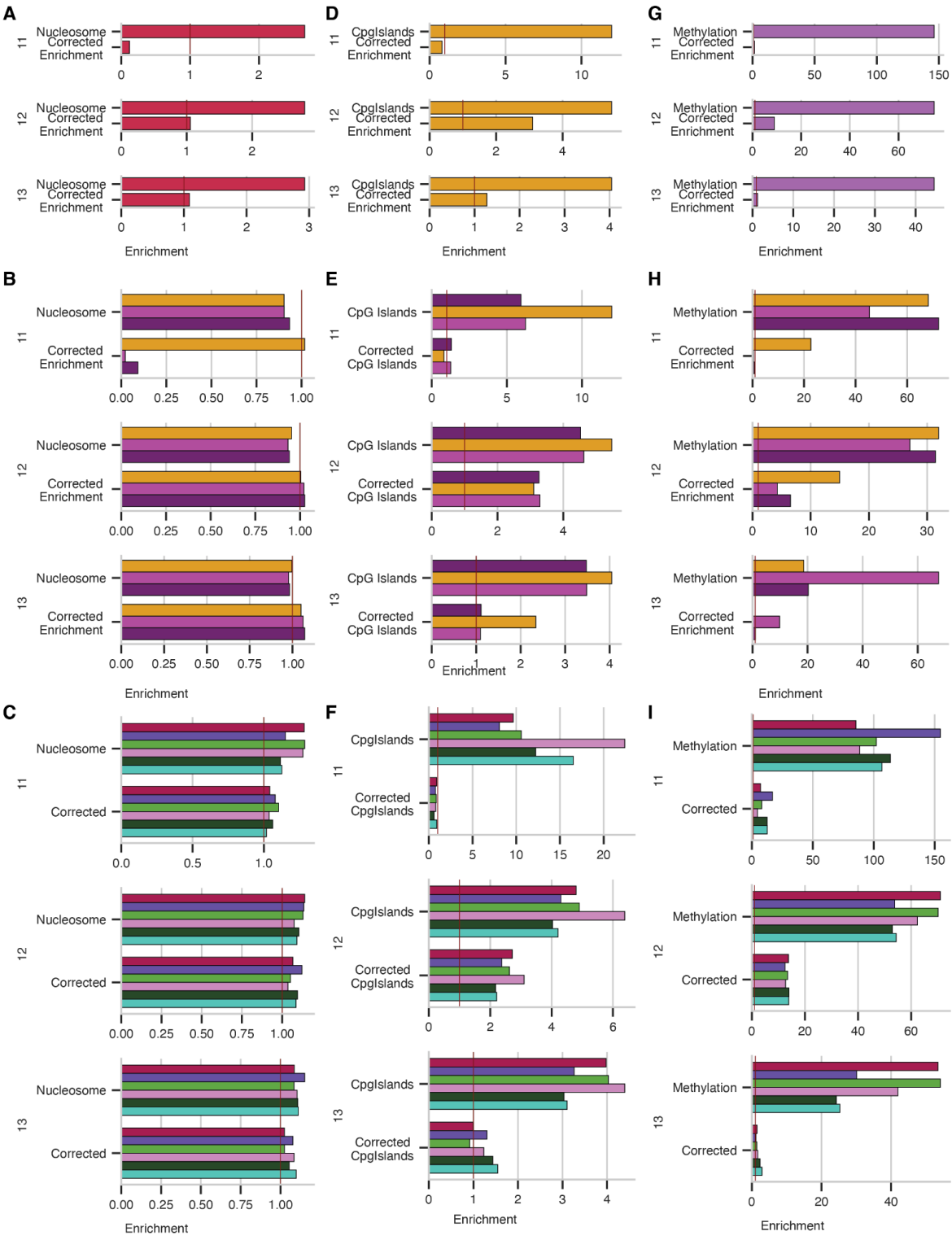

Figure S2.4 Putative nullomer-emerging mutations in nucleosome core positions, CpG islands, and methylation sites across kmers and mutation subtypes

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

Enrichment of nullomer-emerging mutations across k=11,12,13, mutation subtypes, and substitution subtypes for a-c) nucleosome core positions, d-f) CpG islands, g-i) methylation sites.

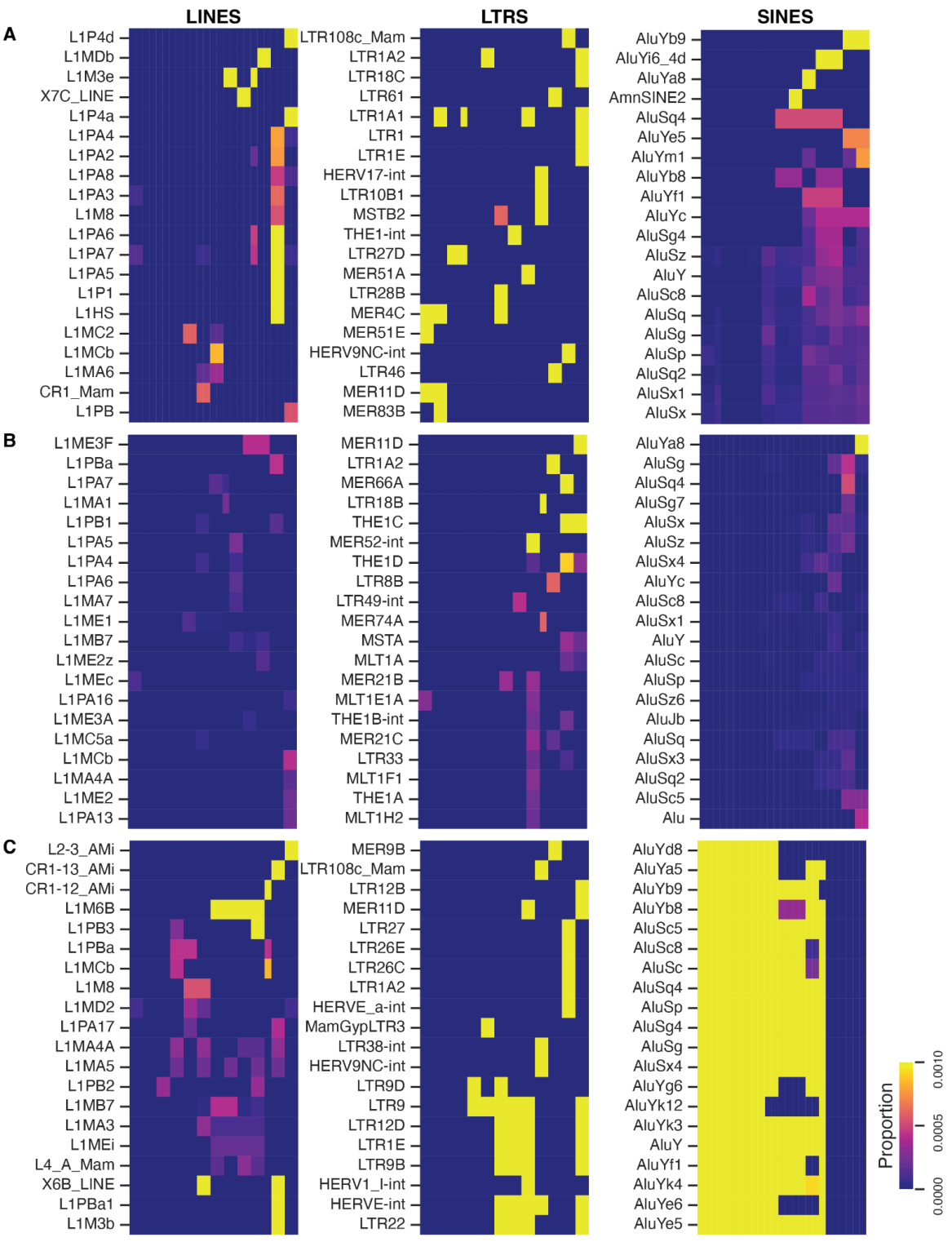

Figure S2.5 Association between putative nullomer-emerging mutations and Alu elements across kmers and mutation subtypes

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

Heatmap of proportion of nullomer-emerging mutations appearing at LINES, LTRs, and SINES for k=13 a) substitutions, b) deletions and c) insertions.

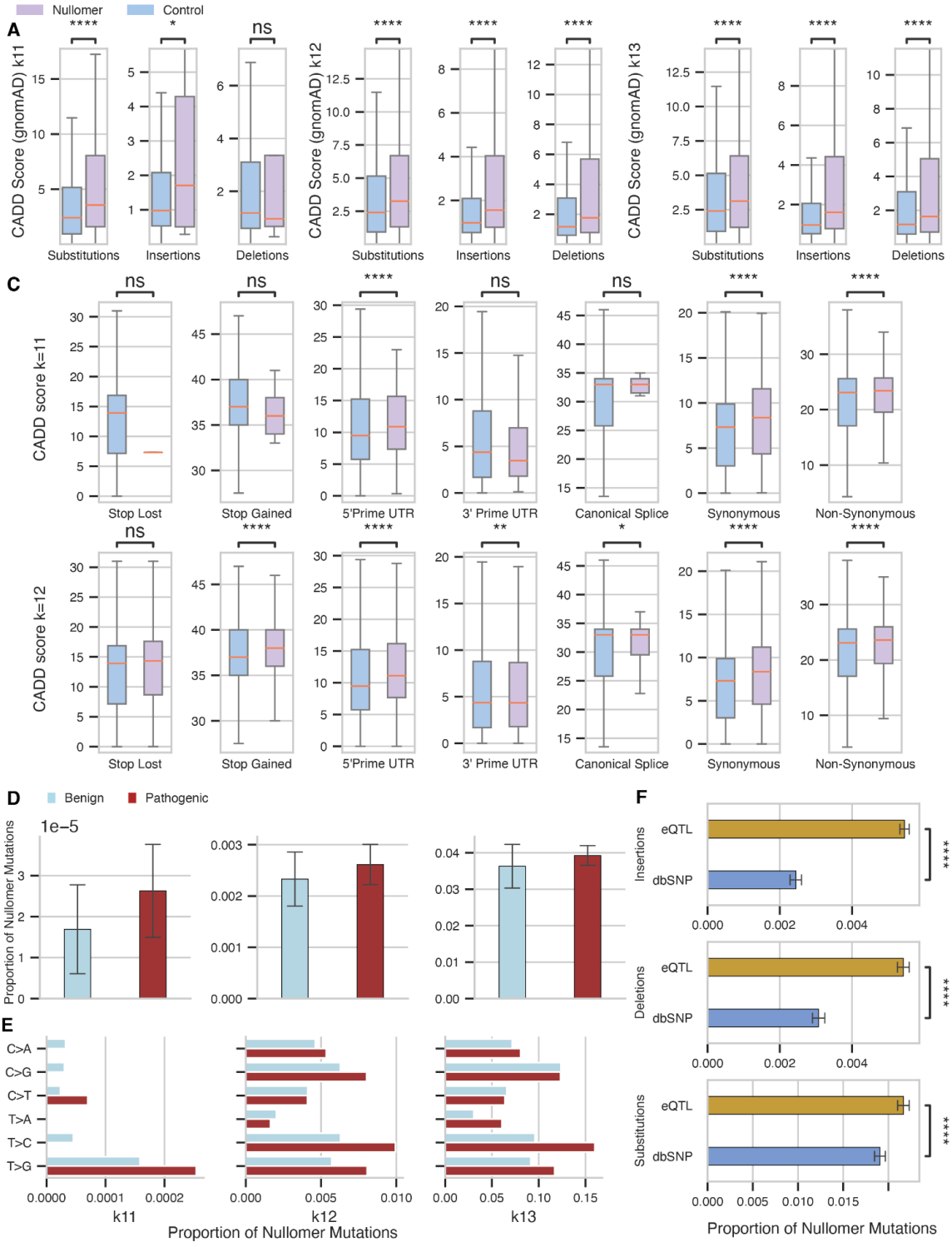

Figure S2.6 Pathogenicity of putative nullomer-emerging mutations across kmers and mutation subtypes

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

a) CADD scores of nullomer-emerging and simulated mutations in population variants from gnomAD for k=11,12,13 subdivided by mutation type. b) Association of CADD score and nullomer emergence for mutations at the stop codon mutation loss, stop stop codon mutation gain, 5'UTR, 3'UTR, canonical splice sites, synonymous and nonsynonymous mutations. c) ClinVar pathogenic and benign nullomer-emerging mutations across k=11,12,13. d) ClinVar pathogenic and benign mutations for k=13 across substitution subtypes. e) Enrichment of eQTLs relative to common SNPs for nullomer emerging mutations in insertions, deletions, and substitutions.

CHAPTER 3

3.1 Abstract

Over 500 noncoding genomic loci are associated with obesity. The majority of these loci reside near genes that are active in the hypothalamus in specific neuronal subpopulations that regulate food intake, hindering the ability to characterize them. Here, we carried out multi-omic analysis (RNA/ATAC-seq) on both mouse and human male and female hypothalamus, identified over 30 different cell populations and characterized several sex-specific differentially expressed genes and regulatory elements. By overlapping cell-specific ATAC peaks with obesity GWAS variants, we identified obesity-associated gene regulatory elements. Utilizing reporter assays and CRISPR editing, we show that many of these sequences, including the top two obesity-associated loci (*FTO* and *MC4R*), regulate known obesity genes and that their activity is altered due to the obesity-associated variant. Combined, our work provides a catalog of genes and regulatory elements in hypothalamus cell subpopulations and shows how multi-omic single-cell sequencing can identify functional variants associated with obesity.

3.2 Introduction

Obesity dramatically increases the risk of morbidity from hypertension, dyslipidemia, diabetes, and cardiovascular diseases and is a major public health concern(Berrington de Gonzalez et al. 2010; Hales et al. 2018). Environmental and genetic factors are both involved in the onset and progression of weight gain(Swinburn et al. 2011), with genetic factors being a major component(Silventoinen et al. 2010; Stunkard, Foch, and Hrubec 1986). Both rare, monogenic obesity causing variants, and common variants, detected through large obesity genome wide association studies (GWAS), have been found to predispose to obesity. While rare gene

mutations causing obesity have outlined a neuro-physiological pathway involved in body weight regulation, the majority of which are part of the hypothalamic leptin-melanocortin pathway(Ranadive and Vaisse 2008), how common variants influence body weight remains mostly poorly understood. GWAS have identified over 500 common loci associated with obesity(Speliotes et al. 2010; Thorleifsson et al. 2009; Loos 2018). The majority of obesity GWAS loci represent clusters of common variants in noncoding regions that likely alter gene expression and have been found to reside primarily near genes involved in central nervous system (CNS)-related processes(Loos 2018; Ghosh and Bouchard 2017). In addition, a number of loci appear to directly influence neurons of the hypothalamic leptin-melanocortin system(Ghosh and Bouchard 2017). Combined, these studies single out neuronal subpopulations in the hypothalamus that are involved in food intake as the major cause of this condition.

The hypothalamus is one of the most complex regions in the brain, composed of diverse neurons and glial cells that are spatially distributed in different compartments, called nuclei(Burbridge, Stewart, and Placzek 2016; Daniel 1976; Flament-Durand 1965). Extensive studies on the hypothalamus have demonstrated that individual nuclei have distinctive functions in regulating different homeostatic processes, such as thermoregulation, mating motivation, sweat, blood pressure, reward, and hunger(Bligh 1966; Gao et al. 2021; Horio and Liberles 2021; Siemian et al. 2021; Tan and Knight 2018; S. X. Zhang et al. 2021; Atasoy et al. 2012). Most of these processes are controlled by neuronal subtypes in each nucleus. In addition, sex-dimorphism in the hypothalamus has been characterized(Heck and Handa 2019; Seale et al. 2004, 2005), with the medial preotic nucleus being a major sex-dimorphic region, releasing different hormones depending on sex(Allen et al. 1989; Bailey and Silver 2014; Stamatiades and Kaiser 2018).

However, the understanding of sex-differential gene expression and transcriptional regulation in the hypothalamus remains largely uncharacterized.

With advances in single-cell sequencing technologies, several studies have set out to characterize diverse cell populations in the whole hypothalamus or specific nuclei using single-cell RNA sequencing (scRNA-seq)(Bai et al. 2019; Chen et al. 2017; Huang et al. 2021; Khrameeva et al. 2020; D. W. Kim et al. 2020; Mickelsen et al. 2019; Moffitt et al. 2018; Ren et al. 2019; Romanov et al. 2020; Wen et al. 2020; Yu, Rubinstein, and Low 2022; Zhou et al. 2022; Steuernagel et al. 2022). These studies characterized various hypothalamus neuronal populations, such as GABAergic and glutamatergic neurons, non-neuronal cell types, including microglia, oligodendrocytes, and astrocytes. While these studies were able to identify various cell populations based on their transcriptional profiles, they lack the capacity to identify the regulatory elements driving these cell type-specific subpopulations, where many of the obesity GWAS variants or other hypothalamus-associated phenotypes could reside.

Multiomic single-cell sequencing utilizes RNA-seq to identify the transcriptome and ATAC-seq to characterize the regulome in the same cell. Here, we utilized this technique to generate a combined scRNA- and sc-ATAC seq atlas of the adult female and male mouse and human hypothalamus. This atlas, composed of more than 50,000 cells, identified numerous non-neuronal and neuronal cells that play key roles in controlling various hypothalamus-associated physiological processes, including blood pressure, mating, nursing, and feeding. We also identified cell-type specific regulatory elements and linked them to their target genes in several cell clusters. Moreover, we uncover cell cluster-specific transcription factor (TF) activity and

their regulated genes and identify cell cluster-specific core TFs (regulatory machinery) shared between mice and humans. In addition, we annotated sex-associated gene expression and regulatory element differences, including hormones/neuropeptides in different neuronal subtypes that are regulated by TFs expressed in a sex-dimorphic manner. Overlapping our cell-type specific regulatory elements with obesity GWAS variants, we identified eighteen potential enhancers for several obesity associated loci, including *FTO* and *MC4R*, the top two associated obesity variants(Speliotes et al. 2010; Thorleifsson et al. 2009; Loos 2018). Enhancer assays in mouse hypothalamus cells, identify ten of these to be functional enhancers and five to alter enhancer activity due to the obesity-associated variant. Removal or CRISPRi of two of these enhancers in cultured human neurons or astrocytes leads to a reduction of gene expression of their target genes. Taken together, this work provides a catalog of genes and regulatory elements in different cellular subpopulations of the hypothalamus in both males and females in mice and humans which can be utilized to identify causative sequences that lead to hypothalamus-associated phenotypes, such as obesity.

3.3 Results

Combined sc-RNA and ATAC profiling of the mouse hypothalamus

To characterize both the genes and regulatory elements of hypothalamic cell types, we utilized the 10X Chromium multiome platform. We initially focused on mouse, utilizing three male and three female hypothalami dissected from 12-week old C57BL/6J mice. We isolated nuclei via FACS and subjected them to 10X Chromium Multiome following established protocols (see Methods). Overall, we sequenced 72,504 cells. After quality control filtering (see Methods), we

retained 13,455 from female mice and 13,329 from males, amounting to 26,784 mouse hypothalamus cells in our dataset (Extended Data Fig.1a). We next set out to characterize the various hypothalamus cell populations using our scRNA-seq data. Utilizing a previous atlas of the mouse hypothalamus, we first determined the various cell-type specific clusters(Steuernagel et al. 2022; Bai et al. 2019; Chen et al. 2017; Huang et al. 2021; Khrameeva et al. 2020; D. W. Kim et al. 2020; Mickelsen et al. 2019; Moffitt et al. 2018; Ren et al. 2019; Romanov et al. 2020; Yu, Rubinstein, and Low 2022; Zhou et al. 2022). Using HypoMap(Steuernagel et al. 2022), we identified 25 distinctive cell populations, including 16 different types of neurons and oligodendrocytes, astrocytes, and immune cell clusters (Fig.1b, Extended Data Fig. 1b, Supplementary Table 1).

For neuronal cells, we found 16 different neuronal cell populations, including glutamatergic, GABAergic, and specialized neuron clusters. We found 9 glutamatergic neuron populations that showed high expression levels of the glutamatergic marker, *Slc17a6*, which encodes the vesicular glutamate transporter 2 (*Vglut2*) gene. We also identified nine clusters of GABAergic neurons, expressing GABAergic marker *Slc32a*, which encodes the vesicular GABA transporter (*Vgat*) gene (Fig.1c). The GABAergic neurons also expressed the synthetic enzyme for GABA, *Gad1*[4] (Extended Data Fig.1c).

Furthermore, we found other specialized neurons secreting various neuronal peptides and hormones that regulate whole-body endocrine and physiological homeostasis. Among these cells, we identified Pomc/Cart neurons (GLU-5) that distinctly express both proopiomelanocortin (*Pomc*) and cocaine and amphetamine regulated transcript (*Cartpt*) and Npy/Argp neurons

(GABA-4), highly expressing neuropeptide Y (*Npy*) and agouti-related peptide (*Agrp*) (Fig.1c, Extended Data Fig. 1d). Both cell types are known to control satiety by regulating melanocortin 4 receptor (*Mc4r*) neurons(Y. Wang et al. 2021) (Extended Data Fig.1c). These neurons were identified in various nuclei, including paraventricular nucleus, accurate nucleus, and intermediolateral nucleus(Govaerts et al. 2005; Lubrano-Berthelier et al. 2003). Similar to previous reports(Lin, Storlien, and Huang 2000; Wauman and Tavernier 2011), we found that Pomc/Cart and Npy/Agrp neurons express leptin receptor (*Lepr*) (Extended Data Fig.1c). In addition, we identified oxytocin (*Oxt*)- and pro-melanin concentrating hormone (*Pmch*)-secreting neurons (GLU-6, GLU-7 clusters) (Extended Data Fig.1d). Oxytocin is known to be involved in female reproduction, breastfeeding, and childbirth(Magon and Kalra 2011), and pro-melanin concentrating hormone, is known to control skin pigmentation and is also an important peptide implicated in the control of motivated behaviors, such as feeding, drinking, and mating(Diniz and Bittencourt 2017).

We also identified dopamine-producing neurons (GLU-9) which show high expression levels of dopamine synthetic enzymes, such as tyrosine hydroxylase (*Th*) and dopa decarboxylase (*Ddc*) (Extended Data Fig.1c-d). We annotated hypocretin neuropeptide (*Hcrt,* GLU-7) or orexin -expressing neurons, which are reported to regulate arousal, wakefulness, and appetite(Ganjavi and Shapiro 2007; Mignot 2004). Additionally, we found growth hormone-releasing hormone (*Ghrh*)- specific neurons (GABA-3) (Fig.1c), which regulate growth hormone production in the anterior pituitary gland(Grossman, Savage, and Besser 1986; García-Tornadu et al. 2010). When examining other neuron peptides, we found oxytocin neurons expressed *Avp* (vasopressin), regulating blood pressure and kidney function(Bartter 1981) (Extended Data Fig.1d). We found

that Ghrh neurons highly expressed *Gal* (Galanin-like peptide), which is known to regulate feeding and reproduction(Bartter 1981; Shiba et al. 2010) (Extended Data Fig.1d). In addition, GABAergic neurons widely expressed *Adcyap1*, encoding for pituitary adenylate cyclase-activating polypeptide binding to its receptor in intestinal cells and associated with post-traumatic stress disorder (PTSD)(Bartter 1981; Shiba et al. 2010; Ressler et al. 2011) (Fig.1c).

We also identified several non-neuronal cell populations. These include oligodendrocytes (ODC), oligodendrocyte precursor cells (OPC), which show high expression of suppression of tumorigenicity 18 (*St18*) and proline-rich 5 like (*Prr5l*) genes and mature oligodendrocytes, expressing platelet-derived growth factor receptor A (*Pdgfra*) and oligodendrocyte transcription factor 1 (*Olig1*) (Extended Data Fig.1e). In addition, we found astrocytes that show high levels of astrocyte gene markers, such as angiotensinogen (*Agt*) expression and immune cells that express the immune marker cathepsin S (*Ctss*) (Extended Data Fig.1e). Additionally, strong expression of decorin (*Dcn)* was found in the fibroblast cluster (Extended Data Fig.1e).

We next examined our scATAC-seq data, isolated from cells that were jointly profiled with RNA-seq. Dimensional reduction was performed using latent semantic indexing and batch correction using Harmony(Korsunsky et al. 2019) (see Methods). Peak calling on each cell-type from joint clustering recovered 414,747 accessible peaks. We observed 166,521 (39%) peaks that were found in only one cluster, indicating robustness of clustering and annotation. We found that many of the cell type-specific ATAC peaks were located near cell type-specific gene markers (average 39.8%, Supplementary Table 2). For example, we found increased accessibility

near *Slc176* in glutamatergic neuron cell types, *Slc32a6* in the GABAergic neuron populations, and *Pmch* in the Pmch neurons (Extended Data Fig.2a).

Combined scATAC and scRNA sequencing allows to identify potential *cis*-regulatory elements by annotating peaks where chromatin accessibility correlates with gene expression (see Methods). We found 930 peaks that were linked to the expression of 49 marker genes. Among the specialized neurons, *Pomc* expression was linked to two scATAC-seq peaks that were highly specific to Pomc/Cart neurons cells (Fig.1e). We also identified cell type-specific scATAC-seq peaks that were linked to *Cartpt* expression in neuronal populations, such as Pomc/Cart and *Pmch* neurons that highly express *Cartp* (Fig.1e). In addition, we annotated three *Npy* linked scATAC-seq peaks that were highly specific to *Npy*/*Argp* neurons. We also found a specific peak that was linked to *Th* expression in dopamine neurons (Extended Data Fig. 2b). Moreover, there were multiple peaks that were linked to *Slc17a6* expression in GABAergic neurons (Extended Data Fig. 2b). Combined, these results show the ability of multiome single-cell analyses to identify specific hypothalamus cell subpopulations and link potential gene regulatory elements with their putative target genes.

Combined sc-RNA and ATAC profiling of the human hypothalamus

We next conducted a similar multiome single-cell experiment on three males and three female adult human hypothalami. We collected post-mortem hypothalami from patients ages 43-82 and a post-mortem interval (PMI) between 4-12 hours (Supplementary Table 3, Extended Data Fig.3a) and carried out scRNA/ATAC-seq, as described for mice. Due to the longer PMI of these samples, compared to mice, we initially obtained a lower number of cell populations. We thus

carried out an additional scRNA/ATAC-seq experiment for the same samples subjecting them to NeuN (neuronal marker) FACS sorting to increase for neuronal populations. Furthermore, in a third multiome experiment on the same samples, we also utilized the Chromium Nuclei Isolation Kit (10X Genomics) to improve the recovery of nuclei from neurons. We combined all experiments and sequenced a total of 113,854 cells, which after filtering for high quality cells retained 36,593 cells (see Methods).

Joint analysis of scATAC and scRNA identified five broad clusters, which we annotated as mature oligodendrocytes, oligodendrocyte progenitor cells, astrocytes, microglia, and neurons (Extended Data Fig.3b, Extended Data Fig.3c). Using differentially expressed genes, we found that neuronal clusters showed high expression for excitatory neuronal markers, including *GAD2*, which encodes for glutamate decarboxylase 2 and *NRG1*, encoding neuregulin 1 (Extended Data Fig.3d). They also expressed other neuronal markers, including stathmin 2 (*STMN2*) and synaptotagmin (*STY1*). Astrocytes expressed astrocyte gene markers, such as glial fibrillary acidic protein (*GFAP*) and *AGT* (Extended Data Fig.3d). Similar to mice, OPC expressed oligodendrocyte markers, such as *PDGFRa* and chondroitin sulfate proteoglycan 4 (*CSPG4*) (Extended Data Fig.3d). More mature oligodendrocytes expressed proteolipid protein 1 (*PLP1*) and myelin oligodendrocyte glycoprotein (*MOG*) (Extended Data Fig.3d). We also found microglia cells with high expression of the pan-immune marker *PTPRC* (CD45), complement C1q B chain (*C1QB*), integrin alpha M (*ITGAM*), and integrin subunit alpha X (*ITGAX*) (Extended Data Fig.3d).

Despite the neuron top-off experiment, we found that the majority of cells (69.9%) isolated from human samples consisted of mature oligodendrocytes. This low neuronal population recovery is consistent with other multiomic single-cell assays on post-mortem human midbrain or frontal lobe(Brase et al. 2022; Adams et al. 2022). To increase the resolution of neuronal cells and other smaller populations, we performed reclustering without the mature ODC, retaining 11,012 cells. Removal of the ODC cluster and re-clustering led to the identification of GABAergic and glutamatergic neuronal clusters, comprising 77 cells and 386 cells respectively. In addition, we found eight astrocyte clusters, four microglia clusters, four myelinating oligodendrocytes (MODC) and three OPC clusters originating from the OPC population (Fig.2c, Supplementary Table 4). We found that 32.4% of peaks were specific to one cluster.

We found three neuronal populations, highly expressing neuron marker, *RBFOX3* (NeuN) (Fig.2b). Among the neuronal populations, glutamatergic neurons had high expression levels of *SLC17A6* while GABAergic neurons showed high *GAD2* expression (Fig.2b). In glutamatergic neurons, we found oxytocin-expressing neurons with expression of *OXT* and *AVP* (Extended Data Fig.3e). We also found *POMC*- and *CARTPT*- expressing neurons as well as *NPY*-expressing cells (Extended Data Fig. 3e). There were also *PMCH*, *HCRT*, and *GHRH* neurons (Extended Data Fig. 3e). GABAergic neurons showed expression of *COL1A1, GRIP2*, and hydroxytryptamine receptor 2c (*HTR2C*) (Extended Data Fig. 3e). In the GABAergic neuron population, we also found neurons expressing neuropeptides, such as vasoactive intestinal peptide (*VIP*), proenkephalin (*PENK*) which plays a role in pain perception and response to stress(Moeller et al. 2015), tachykinin (*TAC1*) which is a ghrelin target and is involved in energy balance and food intake(Trivedi et al. 2015), and thyrotropin releasing hormone (*TRH*) which

stimulates the release of thyroid stimulating hormone (*TSH*) and prolactin from the anterior pituitary(Daimon et al. 2013) (Extended Data Fig. 3e). Interestingly, we identified neuronal clusters that showed high expression of the neural progenitor gene marker, Vimentin (*VIM*) and stem cell/progenitor marker, transcription factor *HES1* (Extended Data Fig. 3e). Furthermore, we identified an ependymal cluster with a highly expressed ependymal marker, *CCDC153(Chen et al. 2017)*. These cells occupy dorsal walls of the third ventricle. We also found a tanycyte cell population, expressing tanycyte markers, peroxiredoxin 6 (*PRDX6*) and glycoprotein *CD59(Chen et al. 2017)*. These cells have important roles in regulating energy homeostasis(Chen et al. 2017)(Goodman and Hajihosseini 2015).

Using markers from previous single-cell studies on oligodendrocytes, we identified three OPC, including OPC, cycling OPC, and committed OPC, and five ODC, including a newly-formed ODC cluster and four mature ODC (MODC) populations. We found that OPC clusters had high expression levels of *PDGFRA*, protocadherin related 15 (*PCDH15*) and versican (*VCAN*) (Fig.2b, Extended Data Fig.4a). As OPC differentiate to myelin oligodendrocytes, they expressed G protein-coupled receptor 17 (*GPR17*), a Gi-coupled GPCR, that acts as an intrinsic timer of oligodendrocyte differentiation and myelination, and transcription factor 7 like 2 (*TCF4*)(Fumagalli et al. 2011). The MODC population showed high expression of oligodendrocytic myelin paranodal and inner loop protein (*OPALIN*) (Fig.2b, Extended Data Fig. 4a). We also identified eight astrocyte clusters with high expression of *AGT* and *GFAP*. Gene set enrichment analysis of differentially expressed genes for each of the astrocyte clusters revealed that each had a unique function (Extended Data Fig.5a). Among the non-neuronal cell types, we also identified four microglia populations with high *ITGAM* and *ITGAX* expression (Fig.2b,

Extended Data Fig.4a). In addition, we found that *ITGAM*, *PTPRC*, and *CD86*, gene markers for active/proinflammatory microglia, were expressed most highly in the microglia 1 and 2 clusters, and less expressed in microglia 3 and 4 clusters. Microglia cluster 3 showed high expression for *CD163* and *CD200R1*, markers for quiescent/anti-inflammatory microglia(Grieve et al. 1990; Jurga, Paleczna, and Kuter 2020) (Extended Data Fig.5b).

We observed that our cell populations had unique gene and chromatin profiles in a cell-specific manner (Fig.2c). We next examined our scATAC-seq data and found many cell type-specific ATAC peaks located near cell type-specific gene markers. For example, we found stronger ATAC peaks near *GAD2* and *OXT* in specialized neurons and neural progenitors, and *GFAP* in the astrocyte populations (Extended Data Fig.6a, Supplementary Table 5). We next analyzed 46 cell marker genes to see whether chromatin accessibility in their locus correlated with gene expression. We examined chromatin accessibility up to one million base pairs away from the marker gene, which is roughly the median length of topologically associated domains (TADs)(McArthur and Capra 2021). Amongst the 46 marker genes analyzed, we identified 629 regions where accessibility correlated with gene expression. These regions were on average 326,444 base pairs away from the transcription start site (TSS) of their postulated target gene. For example, we found that the expression of *SNAP25*, a neuronal marker, was linked to four scATAC-seq peaks in neurons (Extended Data Fig.4b). We also found five peaks linked to the expression of RNA binding fox-1 homolog 3 (*RBFOX3*) in neurons. In addition, in the OPC cells, we found the expression of platelet-derived growth factor receptor alpha (*PDGFRA*), which regulates OPC proliferation and survival(Zhu et al. 2014), linked to many scATAC-seq peaks (Fig.2d). Similar to our mouse data, using the integrative single-cell analyses allowed us to

identify specific cell subpopulations in the human hypothalamus and potential gene regulatory elements with their putative target genes.

Motif enrichment analyses and gene regulatory network inference identifies cell-type specific regulatory pathways

We next set out to identify the major transcription factors (TFs) that govern the various hypothalamus cell types, initially starting with mice. We identified motifs enriched in accessible regions for each cell-type, finding distinct profiles of TF binding sites (TFBS) for each cell cluster. For example, for the glutamatergic neuron populations, we found significant TFBS enrichment of hypothalamic transcription factor *MEF2A*, a known regulator of neurite retraction in hypothalamic neurons(Meyer et al. 2018) (Fig.3a). We also found enrichment for the forkhead box P1 (FOXP1) motif and the NeuroD family of basic helix-loop-helix 1 (NEUROD1) motif, both known to regulate glutamatergic neurogenesis(Hisaoka et al. 2010; Tutukova, Tarabykin, and Hernandez-Miranda 2021) (Fig.3a). Dopamine neurons showed enrichment for FOXA motifs, known to be associated with dopamine biosynthesis(Pristerà et al. 2015). For non-neuronal populations, we identified binding motif enrichment of known cluster-specific TFs, such as *BHLHE22*, enriched in OPC populations, *NFIB* for astrocytes(Yeon et al. 2021), and Spi-B transcription factor (Spi-1/PU.1 related; *SPIB*) known to regulate microglia/macrophages(Stolt et al. 2003; Cakir et al. 2022) (Fig.3a).

In humans, similar to mice, we identified unique TFBS profiles in a cell type-specific manner. We found that both mouse and human neuronal populations have highly enriched motifs for regulatory factor binding to the X-box (RFX) family, including *RFX2* and *RFX4* (Fig.3b). *SOX9*

TFBS and its family members were enriched in OPC and MODC populations (Fig.3b). We also found enrichment for transcription factor 3 (*TCF3*), which is known to maintain OPC(S. Kim et al. 2011). In addition, we found *BHLH21* TFBSs were highly enriched in astrocytes along with nuclear factor IA (NIFA), which is known to induce astrocyte formation(Tchieu et al. 2019) (Fig.3b). Similar to the mouse data, microglia cells showed an enrichment for *SPIB* TFBSs (Fig.3b).

We next compared the gene regulatory networks (GRN) that govern the diversity of hypothalamus cell-types. We used Pando(Fleck et al. 2023), which jointly uses scRNA-seq and scATAC-seq data to model the relationships between transcription factor binding sites (TFBS) and target gene expression. We identified the GRNs that differentiate excitatory and inhibitory neuronal populations in humans and mice. In mice, we found gene regulatory networks centered around transcription factors and genes such as *Bnc2* and *Slc6a3* that were reported to play different roles in GABAergic neuron development and function(Sulistio et al. 2024; Romanov et al. 2020) (Fig 3c, Extended Data Fig. 7). We found a set of gene regulatory networks that were distinctive to glutamatergic neurons, including *Tcf712*, *Lef1*, and *Synop2*. In humans, we found a set of TFs with high specificity for GABAergic neurons, such as *RREB1*, *E2F6*, and *ETV5*, that are involved in GABAergic neuron development(Kherrouche, De Launoit, and Monte 2004; Farley et al. 2018; Yang Liu and Zhang 2019) (Fig 3d, Extended Data Fig.8). We did not find many TFs that were specific to glutamatergic neurons (Fig 3d). We also found solute carrier family 1 member 3 (*SLC1A3*), which enables glutamate uptake, to have high centrality in the glutamatergic neurons but not GABAergic neurons (Fig 3d).

Characterization of sex-differential gene and regulatory activity

The hypothalamus is an endocrine tissue that releases various hormones in a sex-specific manner. As we sequenced both male and female human and mouse hypothalamus, we next characterized sex-differential expression across various cell populations. Specific examination of all neuropeptides found various genes to be differentially expressed in neuronal clusters between sexes. We found a number of neuropeptide human genes to be more expressed in males than females, including *AVP*, *HCRT*, PENK, and *TRH*. In contrast, we found *VIP* expression to be higher in females than males (Fig.4a). Several of our findings are supported by past studies. For example, *AVP*, which regulates various social behaviors including aggression and mating in males, was shown to be highly expressed in hypothalamic GABAergic and glutamatergic neurons in males versus females(Aulino and Caldwell 2020; Tong, Abdulai-Saiku, and Vyas 2021). *VIP* which regulates the release of prolactin, stimulating milk production in the mammary gland(Falsetti et al. 1988; Y. Kim et al. 2017), was previously found to show higher expression in the female brain than males(Y. Kim et al. 2017; Goodwill et al. 2018).

We first identified 54 sex-differentially expressed genes that had sex-differentially accessible promoter regions. For example, in GABAergic neurons, we observed increased promoter accessibility and gene expression in males for the contactin-associated protein-like 2 (*CNTNAP2*) (Fig.4b), a neurexin protein localized in axons and important for neuronal development and synapse formation(Moncini et al. 2016). In addition, adenosine deaminase RNA specific B2 *ADARB2* had stronger expression in female glutamatergic neurons, accompanied by higher accessibility at its promoter (Fig.4b). We also found *OLIG2*, an

oligodendrocyte gene marker, to have higher expression in males and increased accessibility in its promoter in male ODC compared to females (Fig.4b).

Next, to identify putative regulatory elements that regulate sex-differential expression, we identified several sex-differentially accessible peaks linked to sex-differentially expressed genes. For example, in oligodendrocyte progenitor cells, a scATAC peak enriched in males was linked to OLIG1 (Fig.4c). In astrocyte populations, *CD44*, a marker of astrocyte differentiation, showed higher expression and stronger scATAC peaks in females versus male(Ying Liu et al. 2004) (Fig.4c). In addition, we found an open chromatin region that was linked to the expression of the glutamate receptor ionotropic, NMBDA 2b (*GRIN2B*), which plays an important role in cell growth and division, in the female glutamatergic neuron population (Fig.4c).

Similarly, we also found sex-differentially expressed genes in the mouse hypothalamus. We found the expression tachykinin (*Tac2*), a neuropeptide regulating stress processing(Zelikowsky et al. 2018), to be higher in the Glu3 cluster of males than females and its expression was linked to a male-specific scATAC peak (Fig.4d). In addition, blocking *Tac2* pathways have shown inhibition in fear expression in male mice(Florido et al. 2021). Similarly, we found a strong scATAC peak in males linked to the expression of somatostatin (*Sst*), which inhibits hormone production, to be higher in male Glu3 cells. We also identified another scATAC peak differentially accessible to males that is linked to the expression of *Cartpt*, which is involved in regulating body weight, stress response and reward and addiction, that was more highly expressed in male Glu3 cells compared to females (Fig.4d), in line with previous mouse studies that showed higher male expression of this gene in the mouse hypothalamus(Xu et al. 2012).

Taken together, our multiomic single cell analysis was able to identify not only sex-specific gene expression differences in the hypothalamus but also potential regulatory elements that could be driving these differences.

Obesity GWAS prioritization

We next set out to utilize our single-cell catalog of human hypothalamus cell-type specific genes and regulatory elements, to characterize obesity associated GWAS noncoding variants. We compiled a list of 508 index SNPs associated with obesity from the NHGRI-EBI GWAS catalog(Buniello et al. 2019). We performed linkage disequilibrium (LD) expansion on all the lead SNPs to include nearby variants with high probability of coinheritance (r^2 >0.8). The majority of these SNPs (>97%) localize to noncoding regions of the genome. We then checked for overlap of these SNPs with scATAC-seq peaks (Fig.5a), finding 102 SNPs that overlap 55 different peaks, revealing an enrichment of obesity-associated SNPs in chromatin-accessible regions of hypothalamus (p-value=3.8e-18, binomial test, compared to the distribution of autism SNPs(K. Wang et al. 2009)). These SNPs comprised some of the most highly associated loci with obesity, including the top two loci (*FTO* and *MC4R*)(Poveda, Ibáñez, and Rebato 2014; Yohn et al. 2018). Among the 55 peaks, we had 2 scATAC peaks linked to *FAIM2*, 2 scATAC peaks linked to *FTO*, and 2 scATAC peaks in the *MC4R locus*. We did not observe a specific cell-type that was more strongly associated with obesity, having overlapping scATAC peaks from various cell types including oligodendrocytes, oligodendrocyte precursor cells, astrocytes and neurons.

We next tested the enhancer activity of 18 obesity-associated scATAC-seq peaks that overlapped with the strongest obesity-associated SNPs. We cloned these sequences into an enhancer assay vector that contains a minimal promoter followed by the luciferase reporter gene and transfected them into mouse hypothalamus Pomc neuronal cells. We found 10 of 18 (55.5%) of these sequences to have luciferase activity that is significantly higher than the empty vector negative control (Fig. 5b). These include sequences near the genes *ADCY3, FAIM2, FTO, GIPR, GNAT2, GPRC5B, LEMD2, MC4R*, *SDCCAG8, SH2B1, SULT1A1, TDH,* and *VPS45.* (Fig. 5b). Using a similar assay, we also tested the enhancer activity of sequences from scATAC astrocyte peaks (*FTO*, *SDCCAG8*, *LPP*) in primary human astrocytes, obtaining similar results to the neuronal cells (Extended Data Fig.9). We next used site-directed mutagenesis to introduce the obesity-associated SNPs to the sequences that showed enhancer activity and tested them for differential activity compared to the unassociated variants. We found that 5 out of the 10 sequences, near the genes *FAIM2*, *FTO*, *GIPR*, *MC4R* and *VPS45* showed significant differential enhancer activity with the obesity-associated variant/s compared to the reference variant (Fig. 5c).

The SNP that showed differential enhancer activity for *MC4R*, is the second top obesity-associated GWAS SNP and also happens to be the lead SNP, rs17782313. It is located 187k bp from *MC4R* in a sequence that is not conserved in mice. Transcription factor binding site (TFBS) analysis using TRANSFAC(Matys et al. 2006), PROMO(Messeguer et al. 2002), and JASPAR(Messeguer et al. 2002; Castro-Mondragon et al. 2022) of the risk allele found it to potentially disrupt binding of the GATA binding protein 1 (GATA1) (Fig.5d), which is known to also bind to the 5'UTR of the *MC4R* gene(Shishay et al. 2019). To characterize its function and test whether it regulates *MC4R*, we set out to knockout this sequence in neurons using the

CRISPR-Cas9 system. As *MC4R* is not expressed in many cell types, we first differentiated WTC11-NGN2 iPSC cells(C. Wang et al. 2017) to functional glutamatergic neurons (see Methods) and tested whether it is expressed in these cells. We found *MC4R* to be significantly expressed in these cells (~8-fold compared to iPSC cells; Fig.5d). We then transfected these iPSC cells with two gRNAs targeting both ends of the postulated *MC4R* enhancer along with Cas9 protein. Single-cell colonies were FACS-isolated and screened by genotyping for homozygous cells, finding two colonies. These colonies along with wild-type cells were differentiated to neurons and analyzed for *MC4R* expression via qRT-PCR. We found that the two independent *MC4R* enhancer knockout cell lines had significantly lower *MC4R* expression (60%) compared to wild-type cells (Fig.5d). Since there are no other genes in the *MC4R* TAD boundary, we did not measure expression for other genes. Taken together, these results showcase that the obesity associated *MC4R* SNP reduces enhancer activity, and that removal of this enhancer leads to reduced *MC4R* expression, suggesting that it regulates this gene.

We next examined the function of the *FTO* enhancer. This sequence was previously shown to regulate its neighboring genes, Iroquois homeobox 3 and 5 (*IRX3*/*5*), in the hypothalamus(Smemo et al. 2014) and adipocytes(Claussnitzer et al. 2015). TFBS analysis found the obesity-associated SNP, rs8043757, to decrease the binding affinity for the high mobility group forkhead box P3 (*FOXP3*) (Fig.5e), which was shown to regulate gonadotropin expression in the mouse pituitary(Jung et al. 2012). To further characterize its target genes in a cell-type specific manner, we used CRISPR inactivation (CRISPRi). We also attempted to ablate the *FTO* enhancer in these cells, however, they failed to proliferate and form colonies, likely due to the known role of *FTO* in the proliferation of neuronal cells(Cao et al. 2020). Since the associated

SNP overlaps with scATAC peaks in astrocytes, we designed four gRNAs targeting the identified enhancer and tested them for their ability to downregulate expression by co-transfecting human astrocytes with a nuclease deficient Cas9 (dCas) fused to the KRAB transcriptional repressor. Examination by qRT-PCR found *FTO* to have significantly lower expression for three of the four gRNAs, with the lowest reduction for gRNA-4 (Fig.5g). We also analyzed the expression of *FTO* enhancer target genes, *IRX3* and *IRX5*, finding both to have decreased expression levels with all gRNAs, with the highest impact for gRNA-4 (Fig.5g). In line with the enhancer being specific to astrocytes, we found that both *IRX3* and *IRX5* were highly expressed in astrocyte populations (Extended Data Fig.10b). Combined, this work showcases the ability of multiomic scRNA and scATAC-seq to identify gene regulatory elements whose alteration is associated with hypothalamus related phenotypes, such as obesity.

3.4 Discussion

In this study, we utilized multiomic single-cell sequencing to characterize mouse and human hypothalamus. Multiomic single-cell analysis allowed us to identify various neuronal and non-neuronal cell types, uncover cell-specific regulatory elements and their target genes, and elucidate the transcriptional regulation networks of these cells in the hypothalamus. This atlas identified numerous non-neuronal and neuronal cells that play key roles in controlling various hypothalamus-associated physiological processes. By using both female and male hypothalamus, we were able to identify many sex-differential expressed genes and regulatory elements. This integrative data also allowed us to identify and validate regulatory elements that encompass obesity-associated SNPs, including *FTO* and *MC4R*, which alter their enhancer activity. Deletion

of the obesity associated *MC4R* enhancer or CRISPRi targeting of the *FTO* enhancer in astrocytes validated their target genes.

We identified many neuronal cell populations in mice (N=33), including glutamatergic, GABAergic, specialized neuron clusters, OPC, ODC, astrocytes, and microglia, with the majority of cells (90%) being neurons. In contrast, in the human hypothalamus, the majority of cells were oligodendrocyte associated. Several studies in which multiomic single-cell sequencing was performed using post-mortem human midbrain or frontal lobe, showed similar recovery of neuronal and oligodendrocyte populations(Brase et al. 2022; Adams et al. 2022). Similarly, a study by Siletti et al(Siletti et al. 2023),. in which ~3.4 million nuclei from 106 sections of three human adult brains were sequenced, observed a large number of oligodendrocytes (~600K) despite the fact that they isolated neurons by FACS with NeuN (similar to what was done here) aiming to collect 90% neurons and 10% non-neuronal cells. This might be due to human neurons being sensitive to nuclear permeabilization or samples being collected post-mortem when nucleus neurons are the most sensitive to hypoxia(Yoshida, Sasa, and Takaori 1988) as well as the age of donors (Supplemental Table 1). Our analysis did not find significant differences in cell-type proportions across samples except in GABAergic neurons, which was significantly higher in one of the female samples.

Tissue- or cell-specific gene regulatory networks control cell fate specification and drive dynamic processes, such as cell differentiation. Multiomic single-cell sequencing allowed us to identify regulatory elements and link them to their regulated genes in specific clusters and further map TFs activity in a cell type-specific manner. For example, in the mouse hypothalamus, we

identified two regulatory elements linked to *Pomc* expression only in the *Pomc*/*Cartpt* neurons which have been reported to be enhancers of this gene(de Souza et al. 2005). In the human hypothalamus, we identified four regulatory elements linked to *GAD1* expression, one of which is located ~50kb from its promoter and has already been characterized as its enhancer(Bharadwaj et al. 2013). Furthermore, by performing TFBS enrichment analyses on scATAC peaks, we identified many cell type-specific networks of TF activity. Comparison of cell-types across mice and human hypothalamus allowed us to identify shared core TFs that might serve as cell cluster-specific regulatory machinery. We also found genes that are regulated by these core TFs and had cell type-specific functions, further confirming that each cell type requires distinctive gene regulatory networks to regulate its function and activities.

We identified numerous genes that are sex-differentially expressed across various cell clusters in mouse and human hypothalamus. We found that sex effects are small, but ubiquitous across cell populations. We found several genes that were sex-differentially expressed. Some of these differences were previously reported. For example, we found *CNTNAP2* to be differentially expressed between male and female GABAergic neurons, while previous studies have shown sex-dimorphic effects of *CNTNAP2* in social behavior(Dawson et al. 2023). Sex-dimorphic expression of some of these key genes might be involved in sex-differential processes, such as reproductive and social behaviors, and physiological responses to environment cues.

Our multiome single-cell datasets allowed the identification of novel genes, regulatory elements and pathways associated with obesity. Numerous obesity GWAS loci reside near hypothalamus expressed genes(Loos 2018; Ghosh and Bouchard 2017), several of which are associated with

the hypothalamic leptin-melanocortin system(Ghosh and Bouchard 2017), and have been difficult to detect due to the inability to identify gene regulatory elements in distinct neuronal subpopulations. We were able to identify multiple hypothalamus cell type specific scATAC peaks that overlap with obesity-associated SNPs, including the top two obesity associated loci, *FTO* and *MC4R*. This allowed us to pinpoint the cell types in which these potential obesity-associated regulatory elements are involved, finding them to be evenly distributed in various cell populations, including OPC, ODC, astrocyte, and neurons. Interestingly, we found no potential obesity-associated enhancers in the microglia population. Utilizing enhancer assays, we found that more than half of these overlapping scATAC peaks have enhancer activity in either neurons or astrocytes. Additional cell types and or primary cells could also be used to dissect the function of these sequences. Amongst the active sequences, we found that half of them altered enhancer activity. Interestingly, in all these cases this led to a reduction in enhancer activity and not an increase. Our CRISPR experiments for *MC4R* and *FTO* in iPSC-derived neurons and primary astrocytes further confirmed the target genes of these enhancers, showcasing the utility of these datasets.

3.5 Methods

Nuclei isolation and library preparation

For mice, the whole hypothalamus of three male and three female 12-week-old *Mc4r*$^{t2aCre/t2aCre}$ *x* *Rosa26*$^{Ai14/Ai14}$ C57BL/6J mice generated by Dr. Vaisse(Y. Wang et al. 2021) were subjected to nuclei isolation following the 10X Genomics established protocol (CG000375-Rev A). In brief, flash-frozen hypothalamus samples were lysed and homogenized in 0.1% NP40 lysis buffer supplemented with RNase inhibitor (Sigma, 3335402001). The nuclei mixture was filtered through a 70um strainer and then stained with 7-AAD (7-aminoactinomycin D) in PBS with 1% BSA and RNase inhibitor. Nuclei were FACS-sorted using the ARIA Fusion Cell Sorter with a 100um nozzle. Nuclei were washed and pre-metallized with 0.01% digitonin lysis buffer supplemented with RNase inhibitor. For each sample, 20,000 nuclei were subjected for GEX and ATAC libraries prepared using Chromium Next GEM Single Cell Multiome ATAC + Gene Expression (CG000338 Rev B). Both libraries were sequenced using HiSeq 4000 with 25,000 paired reads per nucleus. For human samples, three male and three female hypothalami were collected post-mortem at UCSF medical center following the UCSF human research protection program institutional review board protocol number 21-34261. The donors' age ranged from 43-82 (Supplementary Table 1). The human samples were subjected to the same protocols as mouse hypothalamus. Due to the large size of the human sample, the hypothalamus was split in half and subjected to the lysis step and pooled together prior to FACS. 40,000 nuclei per sample were used for the library prep.

Data preprocessing and quality control

Demultiplexed scRNA- and scATAC-seq fastq files were generated using CellRanger Arc mkfastq (10x Genomics, 7.0.1). scRNA and scATAC-seq data were aligned to pre-built reference genomes GRCh38 and GRCm38 from 10x Genomics. Barcoded count matrices of gene expression and ATAC data were generated using CellRanger ARC pipeline (version 2.0.0) (10X Genomics). Count and peak matrices and fragment files from CellRanger were analyzed using the 'Seurat' package (version 4.1.1)(Stuart et al. 2019). Quality cells were selected using the following quality control metrics: TSS enrichment greater than 1; nucleosome signal greater than 2; fraction of mitochondrial genes less than 30%; between 100 and 7500 genes detected in each cell; and between 1000 and 30000 peaks with at least one readcount detected in each cell. For the human dataset, contaminating ambient RNA was detected using SoupX and the corrected count matrix used for downstream analysis.

Data analysis

scRNA-seq data of nuclei were analyzed using Seurat (version 5.01)(Stuart et al. 2019) after quality control filtering. Gene expression data were log-normalized and multiplied by 10,000 using the NormalizeData and ScaleData functions. The top 3,000 variable features were identified using FindVariableFeatures function. Batch-correction was performed using Harmony(Korsunsky et al. 2019). Dimensionality reduction was performed with PCA and the top 50 principal components were kept. Nearest neighbors used Harmony embeddings, and clustering was performed using resolution = 0.6. Differentially expressed genes were identified using(Stuart et al. 2021) Seurat's 'FindAllMarkers' function(Stuart et al. 2019), using Wilcoxon

Rank Sum test, filtering for a minimum log2 fold change threshold of 1, and minimum fraction of 0.1 cells in the population.

Cluster annotations for the mouse hypothalamus from HypoMap were performed as described in(Steuernagel et al. 2022). scATAC-seq data were processed using 'Signac' R package (version 1.7.0). We apply TFIDF normalization to ATAC peaks, followed by feature selection and dimension reduction using singular value decomposition. Batch-correction was performed using Harmony(Korsunsky et al. 2019). 50 dimensions for ATAC and 50 dimensions RNA PCA dimensions were used to construct the weighted nearest neighbor graph. Peak calling was performed using MACS2(Y. Zhang et al. 2008), and peaks not found on standard chromosomes and within blacklist regions were pruned from the peak set. Differentially accessible regions were identified using Seurat's 'FindAllMarkers' function, using likelihood ratio test with log2 fold change threshold > 0.05. Linked peaks were identified using Signac's 'LinkPeaks' function using peaks found in at least 3 cells, distance cutoff of 1 million bp, p-value cutoff of 0.05. Motif analysis was performed on accessible peaks using chromVar(A. N. Schep et al. 2017). Motif position weight matrices were downloaded from the JASPAR 2020 database. Differential testing on chromVar z-scores was performed using the 'FindAllMarkers' function (Wilcoxon Rank Sum test) in Signac. Peaks containing motifs were identified using the `motifmatchr` package(A. Schep 2022). *motifmatchr: Fast Motif Matching in R* using R package version 1.18.0. Gene regulatory network analysis was performed using Pando (version 1.0.5) as described previously (Fleck et al. 2023).

SNPs selection and overlapped with cell type-specific ATAC peaks

Indexed SNPs associated with obesity were collected from the NHGRI-EBI GWAS Catalog(Buniello et al. 2019). LD expansion was then performed using LDLink(Machiela and Chanock 2015) to identify LD linked SNPs with r2>=0.8. LD linked SNPs were filtered to exclude variants in coding regions based on annotations in dbSNP to obtain a final set of noncoding SNPs. The coordinates of these SNPs were intersected with differentially accessible ATAC peaks. P-values for SNP enrichment in differentially-accessible ATAC peaks were computed using the binomial test as implemented in SciPy. We set the number of successes as SNPs that overlapped the ATAC peaks. We set the number of trials as number of SNPs from obesity collected as described above. We set the probability of success as number of SNPs from autism GWAS (from GWAScentral(Beck et al. 2023)) that overlapped the ATAC peaks. 'motifbreakR' package(Coetzee, Coetzee, and Hazelett 2015) was used to predict effects of SNPs on motifs on its surrounding region.

Luciferase assay

The scATAC-seq peaks that overlapped obesity-associated SNPs were PCR amplified from human genomic DNA (see primers in Supplementary Table 8) and cloned into the pGL4.23 plasmid (Promega, E84111). The associated SNPs were then introduced into these plasmids by PCR amplification with primers containing the associated variants. The PCR products were then treated with DpnI to remove WT and the constructs with the associated SNPs were confirmed with Sanger sequencing. These unassociated and associated constructs (200ng) along with Renilla luciferase (20ng), to correct for the transfection efficiency, were transfected into mouse hypothalamus Pomc neuronal cells (mHypoA-POMC/GFP-1 from Sanbio #CLU500) or human

astrocytes (CCF-STTG, ATCC) grown in 96-well plates using X-tremeGene (Sigma, 6366244001) following the manufacturer's protocol. An empty pGL4.23 vector was used as negative control and pGL4.13 (Promega, E668A), which has an SV40 early enhancer, as a positive control. Forty-eight hours post transfection, cells were lysed, and luciferase activity was measured using the Dual-Luciferase Reporter Assay System (Promega, E1910). Six technical replicates were performed for each condition.

Neuronal differentiation

WTC11-NGN2 iPSC cells(C. Wang et al. 2017) (a generous gift from Li Gan (Gladstone Institute)) were maintained in Matrigel-coated plates with mTeSR1 basal media (StemCell, 85850). Cells were subcultured using Accutase and plated with mTeSR1 media supplemented with 10uM Rock inhibitor (Fisher Scientific, NC1286855). To differentiate these cells to mature neurons, we followed the previously published protocol in(C. Wang et al. 2017). In brief, cells were plated in Matrigel-coated 6-well plate in pre-differentiation media (see recipe in Supplementary Table 9) and Rock inhibitor. Media was changed every day for the next 2 days. On day 4, cells were then sub-cultured into poly-L-Ornithine-coated plates (Sigma, P3665) in maturation media (Supplementary Table 9) supplemented with 2ug/mL doxycycline (Sigma, D3447). Half of the media was then replaced with maturation media after 7 days. Cells were collected at day 14 of differentiation.

MC4R enhancer knockout cells were generated by transfecting WTC11-NGN2 iPSC cells in 6-well plates with 6.25ug Cas9 protein (Fisher Scientific, A36498) and 800ng sgRNAs (IDT), and 0.5ug GFP plasmid (Addgene, 13031) using Lipofectamine CRISPRMax Cas9 transfection reagent (Fisher Scientific, CMAX00-003) following the manufacturer's protocol. After 48 hours, GFP+ cells were isolated into 96 well-plates into single clones using FACS (BD FACSAria

Fusion) (see gRNAs in Supplementary Table 8). These colonies were then genotyped by PCR with primers flanking 200bp the deletion site. The knockout genotype has 400 bp PCR product while WT has 3500 bp PCR product (see primers in Supplementary Table 8). Cells were then subjected to the neuron differentiation protocol described above.

RNA Isolation, cDNA synthesis and qRT-PCR

Total RNA was extracted from cells using RNeasy Plus kit (Qiagen, 74106). Reverse transcription was performed with 1µg of total RNA using qScript cDNA Synthesis Kit (Quantabio, 95047) following the manufacturer's protocol. qRT-PCR was performed on QuantStudio 6 Real Time PCR system (ThermoFisher) using Sso Fast (Biorad, 1705205). Statistical analysis was done using ddct method with GAPDH primers as control (see primer sequences in Supplementary Table 8). Gene expression results were generated using mean values for over 4-6 biological replicates.

CRISPRi study

gRNAs targeting the *FTO* enhancer were designed using CRISPick (Broad Institute)(Doench et al. 2016) (see gRNAs in Supplementary Table 8). gRNAs were cloned into an AAV vector (pAAV-U6-sasgRNA-CMV-mCherry-WPREpA) and co-transfected into human astrocyte cells (CCF-STTG1, ATCC) along with dCas9-KRAB (kind gift from Dr. Alejandro Lomniczi OHSU). After two days, cells were lysed and RNA and cDNA were prepared as mentioned above.

Data deposited in the NCBI short read archive (SRA) as Bioprojects PRJNA899089 (mouse) and PRJNA902416 (human).

3.6 Acknowledgments

We would like to thank Dr. Alejandro Lomniczi for kindly providing the pCVM-sadCas9-KRAB vector. This work was funded in part by the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK) R01DK116738 (C.V. and N.A), the California Institute for Regenerative Medicine (CIRM) postdoctoral fellowship (H.P.N.) and the UCSF Hillblom Center for the Biology of Aging and Bakar Aging Research Institute Graduate Fellowship (C.S.Y.C).

3.7 Author contributions

H.P.N, N.A. designed the study. Nuclear isolation and preparation was performed by H.P.N. H.P.N, D.L.C, R.S, L.H, M.N, A.U, C.B, K.A performed the luciferase assay. H.P.N.,C.S.Y.C, S.B, performed data analysis. F.M, A.D, and C.V provided mice, E.H. provided human samples, H.P.N, C.S.Y.C and N.A. wrote the manuscript with contributions from all authors. C.V, and N.A. acquired funding, M.H, I.L, C.V, and N.A. supervised research.

3.9 References

Adams, Levi, Min Kyung Song, Yoshiaki Tanaka, and Yoon-Seong Kim. 2022. "Single-Nuclei Paired Multiomic Analysis of Young, Aged, and Parkinson's Disease Human Midbrain Reveals Age- and Disease-Associated Glial Changes and Their Contribution to Parkinson's Disease." *medRxiv*. https://doi.org/10.1101/2022.01.18.22269350.

Allen, L. S., M. Hines, J. E. Shryne, and R. A. Gorski. 1989. "Two Sexually Dimorphic Cell Groups in the Human Brain." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 9 (2): 497–506.

Atasoy, Deniz, J. Nicholas Betley, Helen H. Su, and Scott M. Sternson. 2012. "Deconstruction of a Neural Circuit for Hunger." *Nature* 488 (7410): 172–77.

Aulino, Elizabeth A., and Heather K. Caldwell. 2020. "Subtle Sex Differences in Vasopressin mRNA Expression in the Embryonic Mouse Brain." *Journal of Neuroendocrinology* 32 (2): e12835.

Bailey, Matthew, and Rae Silver. 2014. "Sex Differences in Circadian Timing Systems: Implications for Disease." *Frontiers in Neuroendocrinology* 35 (1): 111–39.

Bai, Ling, Sheyda Mesgarzadeh, Karthik S. Ramesh, Erica L. Huey, Yin Liu, Lindsay A. Gray, Tara J. Aitken, et al. 2019. "Genetic Identification of Vagal Sensory Neurons That Control Feeding." *Cell* 179 (5): 1129–43.e23.

Bartter, F. C. 1981. "Vasopressin and Blood Pressure." *The New England Journal of Medicine* 304 (18): 1097–98.

Beck, Tim, Thomas Rowlands, Tom Shorter, and Anthony J. Brookes. 2023. "GWAS Central: An Expanding Resource for Finding and Visualising Genotype and Phenotype Data from Genome-Wide Association Studies." *Nucleic Acids Research* 51 (D1): D986–93.

Berrington de Gonzalez, Amy, Patricia Hartge, James R. Cerhan, Alan J. Flint, Lindsay Hannan,

Robert J. MacInnis, Steven C. Moore, et al. 2010. “Body-Mass Index and Mortality among 1.46 Million White Adults.” *The New England Journal of Medicine* 363 (23): 2211–19.

Bharadwaj, Rahul, Yan Jiang, Wenjie Mao, Mira Jakovcevski, Aslihan Dincer, Winfried Krueger, Krassimira Garbett, et al. 2013. “Conserved Chromosome 2q31 Conformations Are Associated with Transcriptional Regulation of GAD1 GABA Synthesis Enzyme and Altered in Prefrontal Cortex of Subjects with Schizophrenia.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 33 (29): 11839–51.

Bligh, J. 1966. “The Thermosensitivity of the Hypothalamus and Thermoregulation in Mammals.” *Biological Reviews of the Cambridge Philosophical Society* 41 (3): 317–68.

Brase, Logan, Shih-Feng You, Ricardo D’oliveira Albanus, Jorge L. Del-Aguila, Yaoyi Dai, Brenna C. Novotny, Carolina Soriano-Tarraga, et al. 2022. “A Landscape of the Genetic and Cellular Heterogeneity in Alzheimer Disease.” *medRxiv*. https://doi.org/10.1101/2021.11.30.21267072.

Buniello, Annalisa, Jacqueline A. L. MacArthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. “The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019.” *Nucleic Acids Research* 47 (D1): D1005–12.

Burbridge, Sarah, Iain Stewart, and Marysia Placzek. 2016. “Development of the Neuroendocrine Hypothalamus.” *Comprehensive Physiology* 6 (2): 623–43.

Cakir, Bilal, Yoshiaki Tanaka, Ferdi Ridvan Kiral, Yangfei Xiang, Onur Dagliyan, Juan Wang, Maria Lee, et al. 2022. “Expression of the Transcription Factor PU.1 Induces the Generation of Microglia-like Cells in Human Cortical Organoids.” *Nature*

Communications 13 (1): 430.

Cao, Yuhang, Yingliang Zhuang, Junchen Chen, Weize Xu, Yikai Shou, Xiaoli Huang, Qiang Shu, and Xuekun Li. 2020. "Dynamic Effects of Fto in Regulating the Proliferation and Differentiation of Adult Neural Stem Cells of Mice." *Human Molecular Genetics* 29 (5): 727–35.

Castro-Mondragon, Jaime A., Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, et al. 2022. "JASPAR 2022: The 9th Release of the Open-Access Database of Transcription Factor Binding Profiles." *Nucleic Acids Research* 50 (D1): D165–73.

Chen, Renchao, Xiaoji Wu, Lan Jiang, and Yi Zhang. 2017. "Single-Cell RNA-Seq Reveals Hypothalamic Cell Diversity." *Cell Reports* 18 (13): 3227–41.

Claussnitzer, Melina, Simon N. Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, et al. 2015. "FTO Obesity Variant Circuitry and Adipocyte Browning in Humans." *The New England Journal of Medicine* 373 (10): 895–907.

Coetzee, Simon G., Gerhard A. Coetzee, and Dennis J. Hazelett. 2015. "motifbreakR: An R/Bioconductor Package for Predicting Variant Effects at Transcription Factor Binding Sites." *Bioinformatics* 31 (23): 3847–49.

Daimon, Caitlin M., Patrick Chirdon, Stuart Maudsley, and Bronwen Martin. 2013. "The Role of Thyrotropin Releasing Hormone in Aging and Neurodegenerative Diseases." *American Journal of Alzheimer's Disease* 1 (1). https://doi.org/10.7726/ajad.2013.1003.

Daniel, P. M. 1976. "Anatomy of the Hypothalamus and Pituitary Gland." *Journal of Clinical Pathology. Supplement* 7:1–7.

Dawson, Matt S., Kevin Gordon-Fleet, Lingxin Yan, Vera Tardos, Huanying He, Kwong Mui, Smriti Nawani, et al. 2023. “Sexual Dimorphism in the Social Behaviour of Cntnap2-Null Mice Correlates with Disrupted Synaptic Connectivity and Increased Microglial Activity in the Anterior Cingulate Cortex.” *Communications Biology* 6 (1): 846.

Diniz, Giovanne B., and Jackson C. Bittencourt. 2017. “The Melanin-Concentrating Hormone as an Integrative Peptide Driving Motivated Behaviors.” *Frontiers in Systems Neuroscience* 11 (May):32.

Doench, John G., Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W. Vaimberg, Katherine F. Donovan, Ian Smith, et al. 2016. “Optimized sgRNA Design to Maximize Activity and Minimize off-Target Effects of CRISPR-Cas9.” *Nature Biotechnology* 34 (2): 184–91.

Falsetti, L., V. Zanagnolo, A. Gastaldi, M. Memo, C. Missale, and P. F. Spano. 1988. “Vasoactive Intestinal Polypeptide (VIP) Selectively Stimulates Prolactin Release in Healthy Women.” *Gynecological Endocrinology: The Official Journal of the International Society of Gynecological Endocrinology* 2 (1): 11–18.

Farley, Jonathan E., Thomas C. Burdett, Romina Barria, Lukas J. Neukomm, Kevin P. Kenna, John E. Landers, and Marc R. Freeman. 2018. “Transcription Factor Pebbled/RREB1 Regulates Injury-Induced Axon Degeneration.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (6): 1358–63.

Flament-Durand, J. 1965. “Observations on Pituitary Transplants into the Hypothalamus of the Rat.” *Endocrinology* 77 (3): 446–54.

Fleck, Jonas Simon, Sophie Martina Johanna Jansen, Damian Wollny, Fides Zenk, Makiko Seimiya, Akanksha Jain, Ryoko Okamoto, et al. 2023. “Inferring and Perturbing Cell Fate Regulomes in Human Brain Organoids.” *Nature* 621 (7978): 365–72.

Florido, A., E. R. Velasco, C. M. Soto-Faguás, A. Gomez-Gomez, L. Perez-Caballero, P. Molina, R. Nadal, O. J. Pozo, C. A. Saura, and R. Andero. 2021. "Sex Differences in Fear Memory Consolidation via Tac2 Signaling in Mice." *Nature Communications* 12 (1): 2496.

Fumagalli, Marta, Simona Daniele, Davide Lecca, Philip R. Lee, Chiara Parravicini, R. Douglas Fields, Patrizia Rosa, et al. 2011. "Phenotypic Changes, Signaling Pathway, and Functional Correlates of GPR17-Expressing Neural Precursor Cells during Oligodendrocyte Differentiation*." *The Journal of Biological Chemistry* 286 (12): 10593–604.

Ganjavi, Hooman, and Colin M. Shapiro. 2007. "Hypocretin/Orexin: A Molecular Link between Sleep, Energy Regulation, and Pleasure." *The Journal of Neuropsychiatry and Clinical Neurosciences* 19 (4): 413–19.

Gao, Hong-Li, Xiao-Jing Yu, Yan Zhang, Chen-Long Wang, Yi-Ming Lei, Jia-Yue Yu, Dong-Miao Zong, et al. 2021. "Astaxanthin Ameliorates Blood Pressure in Salt-Induced Prehypertensive Rats Through ROS/MAPK/NF-κB Pathways in the Hypothalamic Paraventricular Nucleus." *Cardiovascular Toxicology* 21 (12): 1045–57.

García-Tornadu, Isabel, Gabriela Risso, Maria Ines Perez-Millan, Daniela Noain, Graciela Diaz-Torga, Malcolm J. Low, Marcelo Rubinstein, and Damasia Becu-Villalobos. 2010. "Neurotransmitter Modulation of the GHRH-GH Axis." *Frontiers of Hormone Research* 38 (July):59–69.

Ghosh, Sujoy, and Claude Bouchard. 2017. "Convergence between Biological, Behavioural and Genetic Determinants of Obesity." *Nature Reviews. Genetics* 18 (12): 731–48.

Goodman, Timothy, and Mohammad K. Hajihosseini. 2015. "Hypothalamic Tanycytes-Masters

and Servants of Metabolic, Neuroendocrine, and Neurogenic Functions." *Frontiers in Neuroscience* 9 (October):387.

Goodwill, Haley L., Gabriela Manzano-Nieves, Patrick LaChance, Sana Teramoto, Shirley Lin, Chelsea Lopez, Rachel J. Stevenson, et al. 2018. "Early Life Stress Drives Sex-Selective Impairment in Reversal Learning by Affecting Parvalbumin Interneurons in Orbitofrontal Cortex of Mice." *Cell Reports* 25 (9): 2299–2307.e4.

Govaerts, Cedric, Supriya Srinivasan, Astrid Shapiro, Sumei Zhang, Franck Picard, Karine Clement, Cecile Lubrano-Berthelier, and Christian Vaisse. 2005. "Obesity-Associated Mutations in the Melanocortin 4 Receptor Provide Novel Insights into Its Function." *Peptides* 26 (10): 1909–19.

Grieve, D. A., S. C. Chen, P. H. Chapuis, and R. Bradbury. 1990. "Streptococcus Bovis Bacteraemia: Its Significance for the Colorectal Surgeon." *The Australian and New Zealand Journal of Surgery* 60 (7): 550–52.

Grossman, A., M. O. Savage, and G. M. Besser. 1986. "Growth Hormone Releasing Hormone." *Clinics in Endocrinology and Metabolism* 15 (3): 607–27.

Hales, Craig M., Cheryl D. Fryar, Margaret D. Carroll, David S. Freedman, and Cynthia L. Ogden. 2018. "Trends in Obesity and Severe Obesity Prevalence in US Youth and Adults by Sex and Age, 2007-2008 to 2015-2016." *JAMA: The Journal of the American Medical Association* 319 (16): 1723–25.

Heck, Ashley L., and Robert J. Handa. 2019. "Sex Differences in the Hypothalamic-Pituitary-Adrenal Axis' Response to Stress: An Important Role for Gonadal Hormones." *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology* 44 (1): 45–58.

Hisaoka, T., Y. Nakamura, E. Senba, and Y. Morikawa. 2010. “The Forkhead Transcription Factors, Foxp1 and Foxp2, Identify Different Subpopulations of Projection Neurons in the Mouse Cerebral Cortex.” *Neuroscience* 166 (2): 551–63.

Horio, Nao, and Stephen D. Liberles. 2021. “Hunger Enhances Food-Odour Attraction through a Neuropeptide Y Spotlight.” *Nature* 592 (7853): 262–66.

Huang, Wei-Kai, Samuel Zheng Hao Wong, Sarshan R. Pather, Phuong T. T. Nguyen, Feng Zhang, Daniel Y. Zhang, Zhijian Zhang, et al. 2021. “Generation of Hypothalamic Arcuate Organoids from Human Induced Pluripotent Stem Cells.” *Cell Stem Cell* 28 (9): 1657–70.e10.

Jung, Deborah O., Jake S. Jasurda, Noboru Egashira, and Buffy S. Ellsworth. 2012. “The Forkhead Transcription Factor, FOXP3, Is Required for Normal Pituitary Gonadotropin Expression in Mice.” *Biology of Reproduction* 86 (5): 144, 1–9.

Jurga, Agnieszka M., Martyna Paleczna, and Katarzyna Z. Kuter. 2020. “Overview of General and Discriminating Markers of Differential Microglia Phenotypes.” *Frontiers in Cellular Neuroscience* 14 (August):198.

Kherrouche, Zoulika, Yvan De Launoit, and Didier Monte. 2004. “The NRF-1/alpha-PAL Transcription Factor Regulates Human E2F6 Promoter Activity.” *Biochemical Journal* 383 (Pt. 3): 529–36.

Khrameeva, Ekaterina, Ilia Kurochkin, Dingding Han, Patricia Guijarro, Sabina Kanton, Malgorzata Santel, Zhengzong Qian, et al. 2020. “Single-Cell-Resolution Transcriptome Map of Human, Chimpanzee, Bonobo, and Macaque Brains.” *Genome Research* 30 (5): 776–89.

Kim, Dong Won, Parris Whitney Washington, Zoe Qianyi Wang, Sonia Hao Lin, Changyu Sun,

Basma Taleb Ismail, Hong Wang, Lizhi Jiang, and Seth Blackshaw. 2020. “The Cellular and Molecular Landscape of Hypothalamic Patterning and Differentiation from Embryonic to Late Postnatal Development.” *Nature Communications* 11 (1): 4360.

Kim, Suhyun, Ah-Young Chung, Dohyun Kim, Young-Seop Kim, Hyung-Seok Kim, Hyung-Wook Kwon, Tae-Lin Huh, and Hae-Chul Park. 2011. “Tcf3 Function Is Required for the Inhibition of Oligodendroglial Fate Specification in the Spinal Cord of Zebrafish Embryos.” *Molecules and Cells* 32 (4): 383–88.

Kim, Yongsoo, Guangyu Robert Yang, Kith Pradhan, Kannan Umadevi Venkataraju, Mihail Bota, Luis Carlos García Del Molino, Greg Fitzgerald, et al. 2017. “Brain-Wide Maps Reveal Stereotyped Cell-Type-Based Cortical Architecture and Subcortical Sexual Dimorphism.” *Cell* 171 (2): 456–69.e22.

Korsunsky, Ilya, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-Ru Loh, and Soumya Raychaudhuri. 2019. “Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony.” *Nature Methods* 16 (12): 1289–96.

Lin, S., L. H. Storlien, and X. F. Huang. 2000. “Leptin Receptor, NPY, POMC mRNA Expression in the Diet-Induced Obese Mouse Brain.” *Brain Research* 875 (1-2): 89–95.

Liu, Yang, and Yuanyuan Zhang. 2019. “ETV5 Is Essential for Neuronal Differentiation of Human Neural Progenitor Cells by Repressing NEUROG2 Expression.” *Stem Cell Reviews and Reports* 15 (5): 703–16.

Liu, Ying, Steve S. W. Han, Yuanyuan Wu, Therese M. F. Tuohy, Haipeng Xue, Jingli Cai, Stephen A. Back, Larry S. Sherman, Itzhak Fischer, and Mahendra S. Rao. 2004. “CD44 Expression Identifies Astrocyte-Restricted Precursor Cells.” *Developmental Biology* 276

(1): 31–46.

Loos, Ruth Jf. 2018. “The Genetics of Adiposity.” *Current Opinion in Genetics & Development* 50 (June):86–95.

Lubrano-Berthelier, Cecile, Martha Cavazos, Beatrice Dubern, Astrid Shapiro, Catherine L. E. Stunff, Sumei Zhang, Franck Picart, et al. 2003. “Molecular Genetics of Human Obesity-Associated MC4R Mutations.” *Annals of the New York Academy of Sciences* 994 (June):49–57.

Machiela, M. J., and S. J. Chanock. 2015. “LDlink: A Web-Based Application for Exploring Population-Specific Haplotype Structure and Linking Correlated Alleles of Possible Functional Variants.” *Bioinformatics* 31 (21): 3555–57. doi: 10.1093/bioinformatics/btv402. Epub 2015 Jul 2.

Magon, Navneet, and Sanjay Kalra. 2011. “The Orgasmic History of Oxytocin: Love, Lust, and Labor.” *Indian Journal of Endocrinology and Metabolism* 15 Suppl 3 (Suppl3): S156–61.

Matys, V., O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, et al. 2006. “TRANSFAC and Its Module TRANSCompel: Transcriptional Gene Regulation in Eukaryotes.” *Nucleic Acids Research* 34 (Database issue): D108–10.

McArthur, Evonne, and John A. Capra. 2021. “Topologically Associating Domain Boundaries That Are Stable across Diverse Cell Types Are Evolutionarily Constrained and Enriched for Heritability.” *American Journal of Human Genetics* 108 (2): 269–83.

Messeguer, Xavier, Ruth Escudero, Domènec Farré, Oscar Núñez, Javier Martínez, and M. Mar Albà. 2002. “PROMO: Detection of Known Transcription Regulatory Elements Using Species-Tailored Searches.” *Bioinformatics* 18 (2): 333–34.

Meyer, Magdalena, Ilona Berger, Julia Winter, and Benjamin Jurek. 2018. “Oxytocin Alters the

Morphology of Hypothalamic Neurons via the Transcription Factor Myocyte Enhancer Factor 2A (MEF-2A)." *Molecular and Cellular Endocrinology* 477 (December):156–62.

Mickelsen, Laura E., Mohan Bolisetty, Brock R. Chimileski, Akie Fujita, Eric J. Beltrami, James T. Costanzo, Jacob R. Naparstek, Paul Robson, and Alexander C. Jackson. 2019. "Single-Cell Transcriptomic Analysis of the Lateral Hypothalamic Area Reveals Molecularly Distinct Populations of Inhibitory and Excitatory Neurons." *Nature Neuroscience* 22 (4): 642–56.

Mignot, Emmanuel. 2004. "Sleep, Sleep Disorders and Hypocretin (orexin)." *Sleep Medicine* 5 Suppl 1 (June):S2–8.

Moeller, Scott J., Nicasia Beebe-Wang, Kristin E. Schneider, Anna B. Konova, Muhammad A. Parvaz, Nelly Alia-Klein, Yasmin L. Hurd, and Rita Z. Goldstein. 2015. "Effects of an Opioid (proenkephalin) Polymorphism on Neural Response to Errors in Health and Cocaine Use Disorder." *Behavioural Brain Research* 293 (October):18–26.

Moffitt, Jeffrey R., Dhananjay Bambah-Mukku, Stephen W. Eichhorn, Eric Vaughn, Karthik Shekhar, Julio D. Perez, Nimrod D. Rubinstein, et al. 2018. "Molecular, Spatial, and Functional Single-Cell Profiling of the Hypothalamic Preoptic Region." *Science* 362 (6416). https://doi.org/10.1126/science.aau5324.

Moncini, Silvia, Paola Castronovo, Alessandra Murgia, Silvia Russo, Maria Francesca Bedeschi, Marta Lunghi, Angelo Selicorni, Maria Teresa Bonati, Paola Riva, and Marco Venturin. 2016. "Functional Characterization of CDK5 and CDK5R1 Mutations Identified in Patients with Non-Syndromic Intellectual Disability." *Journal of Human Genetics* 61 (4): 283–93.

Poveda, Alaitz, María Eugenia Ibáñez, and Esther Rebato. 2014. "Common Variants in BDNF,

FAIM2, FTO, MC4R, NEGR1, and SH2B1 Show Association with Obesity-Related Variables in Spanish Roma Population." *American Journal of Human Biology: The Official Journal of the Human Biology Council* 26 (5): 660–69.

Pristerà, Alessandro, Wei Lin, Anna-Kristin Kaufmann, Katherine R. Brimblecombe, Sarah Threlfell, Paul D. Dodson, Peter J. Magill, Cathy Fernandes, Stephanie J. Cragg, and Siew-Lan Ang. 2015. "Transcription Factors FOXA1 and FOXA2 Maintain Dopaminergic Neuronal Properties and Control Feeding Behavior in Adult Mice." *Proceedings of the National Academy of Sciences of the United States of America* 112 (35): E4929–38.

Ranadive, Sayali A., and Christian Vaisse. 2008. "Lessons from Extreme Human Obesity: Monogenic Disorders." *Endocrinology and Metabolism Clinics of North America* 37 (3): 733–51, x.

Ren, Jing, Alina Isakova, Drew Friedmann, Jiawei Zeng, Sophie M. Grutzner, Albert Pun, Grace Q. Zhao, et al. 2019. "Single-Cell Transcriptomes and Whole-Brain Projections of Serotonin Neurons in the Mouse Dorsal and Median Raphe Nuclei." *eLife* 8 (October). https://doi.org/10.7554/eLife.49424.

Ressler, Kerry J., Kristina B. Mercer, Bekh Bradley, Tanja Jovanovic, Amy Mahan, Kimberly Kerley, Seth D. Norrholm, et al. 2011. "Post-Traumatic Stress Disorder Is Associated with PACAP and the PAC1 Receptor." *Nature* 470 (7335): 492–97.

Romanov, Roman A., Evgenii O. Tretiakov, Maria Eleni Kastriti, Maja Zupancic, Martin Häring, Solomiia Korchynska, Konstantin Popadin, et al. 2020. "Molecular Design of Hypothalamus Development." *Nature* 582 (7811): 246–52.

Schep, Alicia. 2022. "Motifmatchr: Fast Motif Matching in R." *R Package Version* 1 (0).

Schep, A. N., B. Wu, J. D. Buenrostro, and W. J. Greenleaf. 2017. "chromVAR: Inferring Transcription-Factor-Associated Accessibility from Single-Cell Epigenomic Data." *Nature Methods* 14 (10): 975–78. doi: 10.1038/nmeth.4401. Epub 2017 Aug 21.

Seale, J. V., S. A. Wood, H. C. Atkinson, M. S. Harbuz, and S. L. Lightman. 2004. "Gonadal Steroid Replacement Reverses Gonadectomy-Induced Changes in the Corticosterone Pulse Profile and Stress-Induced Hypothalamic-Pituitary-Adrenal Axis Activity of Male and Female Rats." *Journal of Neuroendocrinology* 16 (12): 989–98.

Seale, J. V., S. A. Wood, H. C. Atkinson, S. L. Lightman, and M. S. Harbuz. 2005. "Organizational Role for Testosterone and Estrogen on Adult Hypothalamic-Pituitary-Adrenal Axis Activity in the Male Rat." *Endocrinology* 146 (4): 1973–82.

Shiba, Kanako, Haruaki Kageyama, Fumiko Takenoya, and Seiji Shioda. 2010. "Galanin-like Peptide and the Regulation of Feeding Behavior and Energy Metabolism." *The FEBS Journal* 277 (24): 5006–13.

Shishay, Girmay, Guiqiong Liu, Xunping Jiang, Yun Yu, Wassie Teketay, Dandan Du, Huang Jing, and Chenghui Liu. 2019. "Variation in the Promoter Region of the Gene Elucidates the Association of Body Measurement Traits in Hu Sheep." *International Journal of Molecular Sciences* 20 (2). https://doi.org/10.3390/ijms20020240.

Siemian, Justin N., Miguel A. Arenivar, Sarah Sarsfield, and Yeka Aponte. 2021. "Hypothalamic Control of Interoceptive Hunger." *Current Biology: CB* 31 (17): 3797–3809.e5.

Siletti, Kimberly, Rebecca Hodge, Alejandro Mossi Albiach, Ka Wai Lee, Song-Lin Ding, Lijuan Hu, Peter Lönnerberg, et al. 2023. "Transcriptomic Diversity of Cell Types across the Adult Human Brain." *Science*, October. https://doi.org/10.1126/science.add7046.

Silventoinen, K., B. Rokholm, J. Kaprio, and T. I. A. Sørensen. 2010. "The Genetic and

Environmental Influences on Childhood Obesity: A Systematic Review of Twin and Adoption Studies." *International Journal of Obesity* 34 (1): 29–40.

Smemo, Scott, Juan J. Tena, Kyoung-Han Kim, Eric R. Gamazon, Noboru J. Sakabe, Carlos Gómez-Marín, Ivy Aneas, et al. 2014. "Obesity-Associated Variants within FTO Form Long-Range Functional Connections with IRX3." *Nature* 507 (7492): 371–75.

Souza, Flávio S. J. de, Andrea M. Santangelo, Viviana Bumaschny, María Elena Avale, James L. Smart, Malcolm J. Low, and Marcelo Rubinstein. 2005. "Identification of Neuronal Enhancers of the Proopiomelanocortin Gene by Transgenic Mouse Analysis and Phylogenetic Footprinting." *Molecular and Cellular Biology* 25 (8): 3076–86.

Speliotes, Elizabeth K., Cristen J. Willer, Sonja I. Berndt, Keri L. Monda, Gudmar Thorleifsson, Anne U. Jackson, Hana Lango Allen, et al. 2010. "Association Analyses of 249,796 Individuals Reveal 18 New Loci Associated with Body Mass Index." *Nature Genetics* 42 (11): 937–48.

Stamatiades, George A., and Ursula B. Kaiser. 2018. "Gonadotropin Regulation by Pulsatile GnRH: Signaling and Gene Expression." *Molecular and Cellular Endocrinology* 463 (March):131–41.

Steuernagel, Lukas, Brian Y. H. Lam, Paul Klemm, Georgina K. C. Dowsett, Corinna A. Bauder, John A. Tadross, Tamara Sotelo Hitschfeld, et al. 2022. "HypoMap-a Unified Single-Cell Gene Expression Atlas of the Murine Hypothalamus." *Nature Metabolism* 4 (10): 1402–19.

Stolt, C. Claus, Petra Lommes, Elisabeth Sock, Marie-Christine Chaboissier, Andreas Schedl, and Michael Wegner. 2003. "The Sox9 Transcription Factor Determines Glial Fate Choice in the Developing Spinal Cord." *Genes & Development* 17 (13): 1677–89.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. "Comprehensive Integration of Single-Cell Data." *Cell* 177 (7): 1888–1902.e21.

Stuart, Tim, Avi Srivastava, Shaista Madad, Caleb A. Lareau, and Rahul Satija. 2021. "Single-Cell Chromatin State Analysis with Signac." *Nature Methods* 18 (11): 1333–41.

Stunkard, A. J., T. T. Foch, and Z. Hrubec. 1986. "A Twin Study of Human Obesity." *JAMA: The Journal of the American Medical Association* 256 (1): 51–54.

Sulistio, Yanuar Alan, Yuna Lee, Kelvin Pieknell, Sebin Hong, Jumi Kim, Min Jong Seok, Na-Kyung Lee, et al. 2024. "Hypothalamus-Specific NSCs Derived from hPSCs Ameliorate Age-Associated Dysfunction upon Transplantation into Aged Mouse Hypothalamus." *bioRxiv*. https://doi.org/10.1101/2024.05.23.595504.

Swinburn, Boyd A., Gary Sacks, Kevin D. Hall, Klim McPherson, Diane T. Finegood, Marjory L. Moodie, and Steven L. Gortmaker. 2011. "The Global Obesity Pandemic: Shaped by Global Drivers and Local Environments." *Lancet*. Elsevier BV.

Tan, Chan Lek, and Zachary A. Knight. 2018. "Regulation of Body Temperature by the Nervous System." *Neuron* 98 (1): 31–48.

Tchieu, Jason, Elizabeth L. Calder, Sudha R. Guttikonda, Eveline M. Gutzwiller, Kelly A. Aromolaran, Julius A. Steinbeck, Peter A. Goldstein, and Lorenz Studer. 2019. "NFIA Is a Gliogenic Switch Enabling Rapid Derivation of Functional Human Astrocytes from Pluripotent Stem Cells." *Nature Biotechnology* 37 (3): 267–75.

Thorleifsson, Gudmar, G. Bragi Walters, Daniel F. Gudbjartsson, Valgerdur Steinthorsdottir, Patrick Sulem, Anna Helgadottir, Unnur Styrkarsdottir, et al. 2009. "Genome-Wide Association Yields New Sequence Variants at Seven Loci That Associate with Measures

of Obesity." *Nature Genetics* 41 (1): 18–24.

Tong, Wen Han, Samira Abdulai-Saiku, and Ajai Vyas. 2021. "Arginine Vasopressin in the Medial Amygdala Causes Greater Post-Stress Recruitment of Hypothalamic Vasopressin Neurons." *Molecular Brain* 14 (1): 141.

Trivedi, Chitrang, Xiaoye Shan, Yi-Chun Loraine Tung, Dhiraj Kabra, Jenna Holland, Sarah Amburgy, Kristy Heppner, Henriette Kirchner, Giles S. H. Yeo, and Diego Perez-Tilve. 2015. "Tachykinin-1 in the Central Nervous System Regulates Adiposity in Rodents." *Endocrinology* 156 (5): 1714–23.

Tutukova, Svetlana, Victor Tarabykin, and Luis R. Hernandez-Miranda. 2021. "The Role of Neurod Genes in Brain Development, Function, and Disease." *Frontiers in Molecular Neuroscience* 14 (June):662774.

Wang, Chao, Michael E. Ward, Robert Chen, Kai Liu, Tara E. Tracy, Xu Chen, Min Xie, et al. 2017. "Scalable Production of iPSC-Derived Human Neurons to Identify Tau-Lowering Compounds by High-Content Screening." *Stem Cell Reports* 9 (4): 1221–33.

Wang, Kai, Haitao Zhang, Deqiong Ma, Maja Bucan, Joseph T. Glessner, Brett S. Abrahams, Daria Salyakina, et al. 2009. "Common Genetic Variants on 5p14.1 Associate with Autism Spectrum Disorders." *Nature* 459 (7246): 528–33.

Wang, Yi, Adelaide Bernard, Fanny Comblain, Xinyu Yue, Christophe Paillart, Sumei Zhang, Jeremy F. Reiter, and Christian Vaisse. 2021. "Melanocortin 4 Receptor Signals at the Neuronal Primary Cilium to Control Food Intake and Body Weight." *The Journal of Clinical Investigation* 131 (9). https://doi.org/10.1172/JCI142064.

Wauman, Joris, and Jan Tavernier. 2011. "Leptin Receptor Signaling: Pathways to Leptin Resistance." *Frontiers in Bioscience* 16 (7): 2771–93.

Wen, Shao 'ang, Danyi Ma, Meng Zhao, Lucheng Xie, Qingqin Wu, Lingfeng Gou, Chuanzhen Zhu, Yuqi Fan, Haifang Wang, and Jun Yan. 2020. "Spatiotemporal Single-Cell Analysis of Gene Expression in the Mouse Suprachiasmatic Nucleus." *Nature Neuroscience* 23 (3): 456–67.

Xu, Xiaohong, Jennifer K. Coats, Cindy F. Yang, Amy Wang, Osama M. Ahmed, Maricruz Alvarado, Tetsuro Izumi, and Nirao M. Shah. 2012. "Modular Genetic Control of Sexually Dimorphic Behaviors." *Cell* 148 (3): 596–607.

Yeon, Gyu-Bum, Won-Ho Shin, Seo Hyun Yoo, Dongyun Kim, Byeong-Min Jeon, Won-Ung Park, Yeonju Bae, et al. 2021. "NFIB Induces Functional Astrocytes from Human Pluripotent Stem Cell-Derived Neural Precursor Cells Mimicking in Vivo Astrogliogenesis." *Journal of Cellular Physiology* 236 (11): 7625–41.

Yohn, Christine N., Amanda B. Leithead, Julian Ford, Alexander Gill, and Elizabeth A. Becker. 2018. "Paternal Care Impacts Oxytocin Expression in California Mouse Offspring and Basal Testosterone in Female, but Not Male Pups." *Frontiers in Behavioral Neuroscience* 12 (August):181.

Yoshida, S., M. Sasa, and S. Takaori. 1988. "Different Sensitivity to Hypoxia in Neuronal Activities of Lateral Vestibular and Spinal Trigeminal Nuclei." *Stroke; a Journal of Cerebral Circulation* 19 (3): 357–64.

Yu, Hui, Marcelo Rubinstein, and Malcolm J. Low. 2022. "Developmental Single-Cell Transcriptomics of Hypothalamic POMC Neurons Reveal the Genetic Trajectories of Multiple Neuropeptidergic Phenotypes." *eLife* 11 (January). https://doi.org/10.7554/eLife.72883.

Zelikowsky, Moriel, May Hui, Tomomi Karigo, Andrea Choe, Bin Yang, Mario R. Blanco,

Keith Beadle, Viviana Gradinaru, Benjamin E. Deverman, and David J. Anderson. 2018. "The Neuropeptide Tac2 Controls a Distributed Brain State Induced by Chronic Social Isolation Stress." *Cell* 173 (5): 1265–79.e19.

Zhang, Stephen X., Andrew Lutas, Shang Yang, Adriana Diaz, Hugo Fluhr, Georg Nagel, Shiqiang Gao, and Mark L. Andermann. 2021. "Hypothalamic Dopamine Neurons Motivate Mating through Persistent cAMP Signalling." *Nature* 597 (7875): 245–49.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9): R137.

Zhou, Xin, Yufeng Lu, Fangqi Zhao, Ji Dong, Wenji Ma, Suijuan Zhong, Mengdi Wang, et al. 2022. "Deciphering the Spatial-Temporal Transcriptional Landscape of Human Hypothalamus Development." *Cell Stem Cell* 29 (2): 328–43.e5.

Zhu, Qiang, Xiaofeng Zhao, Kang Zheng, Hong Li, Hao Huang, Zunyi Zhang, Teresa Mastracci, et al. 2014. "Genetic Evidence That Nkx2.2 and Pdgfra Are Major Determinants of the Timing of Oligodendrocyte Differentiation in the Developing CNS." *Development* 141 (3): 5482–55.

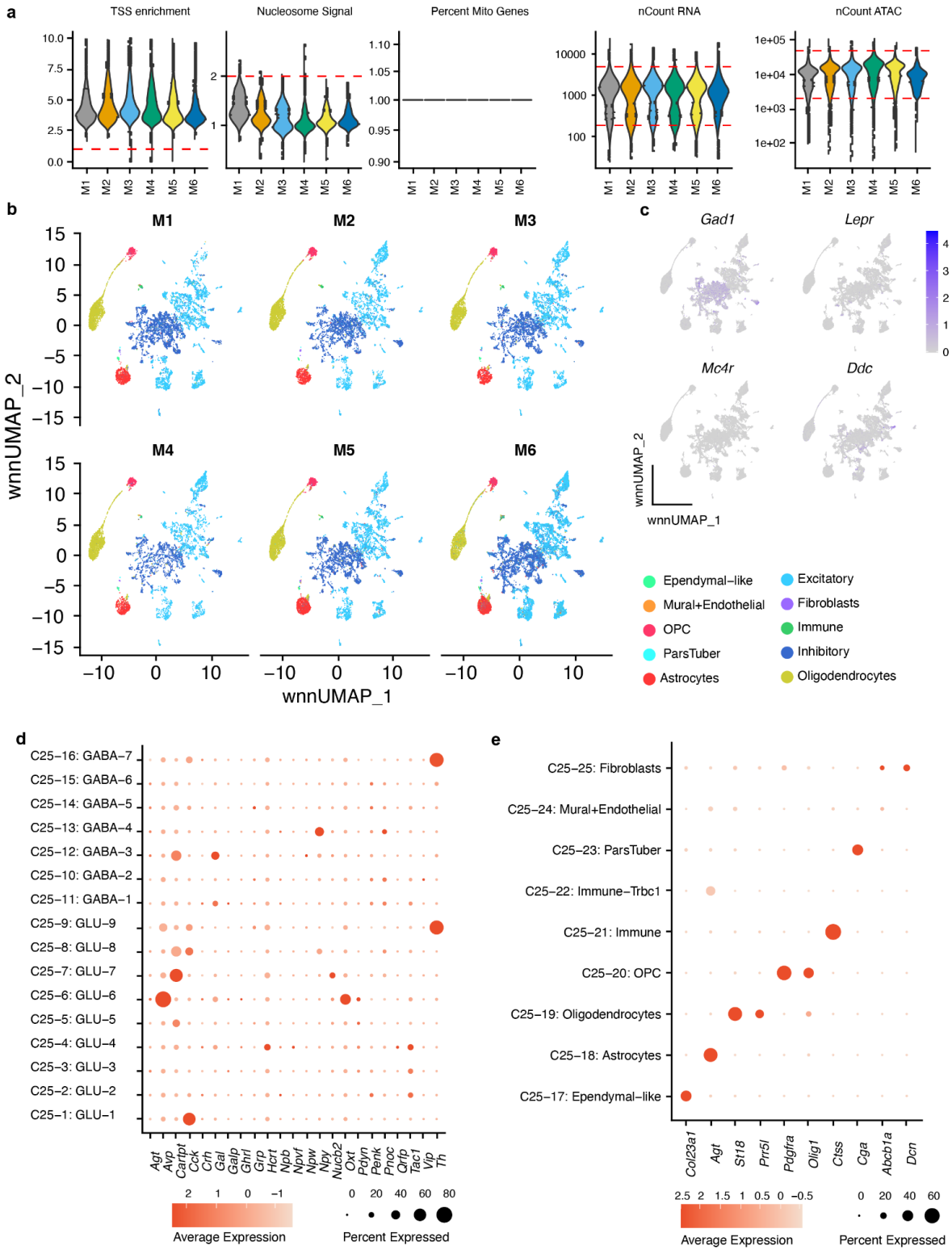

Figure S3.1 Combined scRNA and ATAC profiling of the mouse hypothalamus.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

a) Violin plot of quality control metrics: TSS enrichment scores, nucleosome signal, percentage of mitochondrial genes, number of features in RNA data, and number of features in ATAC data. b) UMAP of cell-type annotations across each mouse sample. c) UMAP showing cells expressing Gad1, Lepr, Mc4r, and Ddc colored in purple. d) Dot plot of neuropeptide gene expression across neuronal cell types in mouse hypothalamus. e) Dot plot of gene markers of non-neuronal cell types.

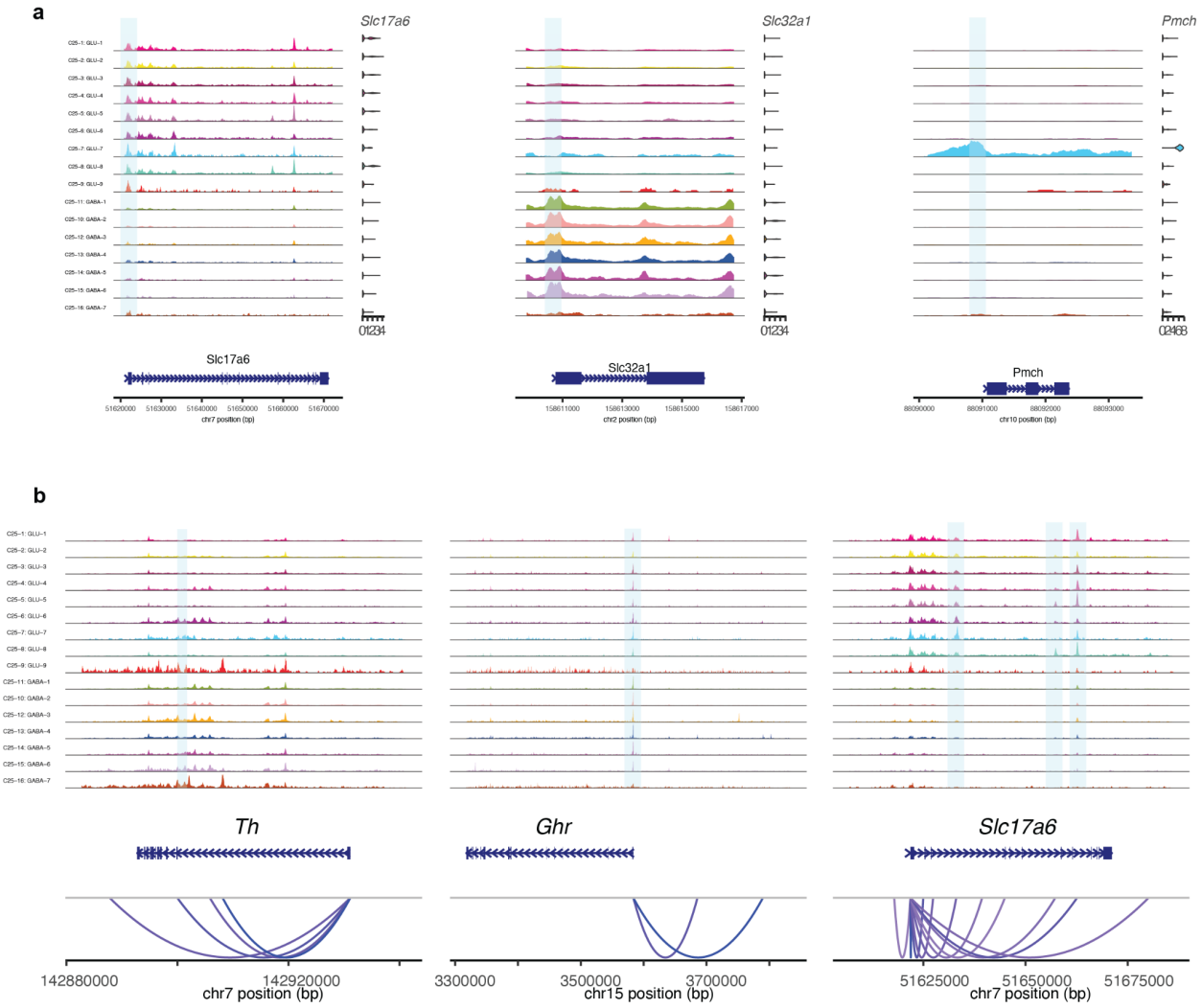

Figure S3.2 scATAC-seq coverage plots of the mouse hypothalamus.

a) Coverage plot of genes, Slc17a6, Slc32a1, Pmch showing scATAC peaks highlighted in blue around the gene body. b) scATAC peaks to gene linkage plots for Th, Ghr and Slc17a6 in the neuronal populations with the genome tracks. The linked scATAC peaks to gene promoters are highlighted in blue.

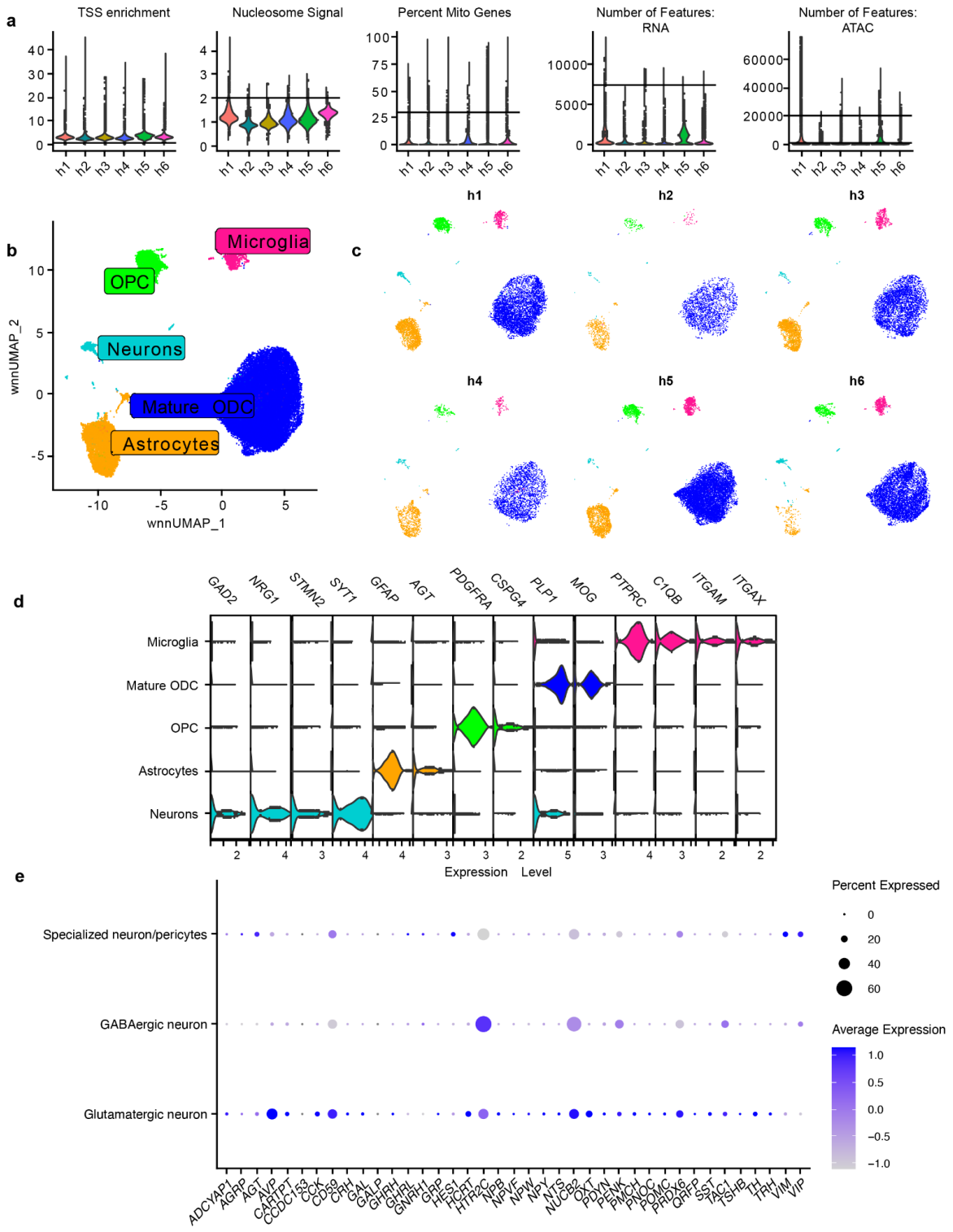

Figure S3.3 Combined scRNA and ATAC profiling of the human hypothalamus.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

a) Violin plot of quality control metrics: TSS enrichment scores, nucleosome signal, percentage of mitochondrial genes, number of features in RNA data, and number of features in ATAC data. b) UMAP of integrated scRNA and scATAC-seq data. c) UMAP of cell-type annotations across human samples. d) Violin plot of gene markers used to identify cell clusters. e) Dot plot of neuropeptide gene expression across neuronal cells.

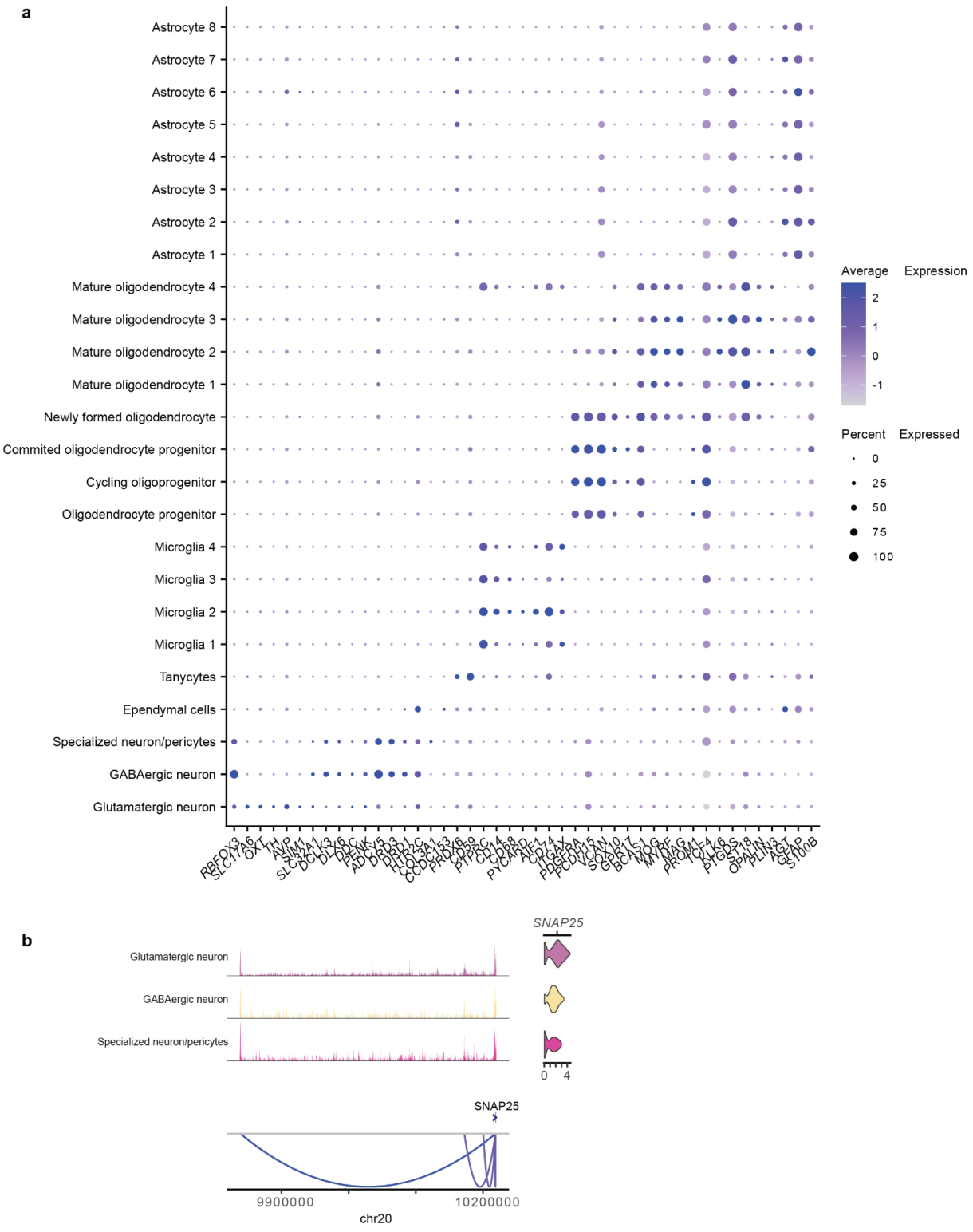

Figure S3.4 Expression of cell-type markers and sc-ATAC coverage plots of human hypothalamus.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

a) Dot plots showing expression of gene markers used to identify cell clusters. b) Coverage plot showing sc-ATAC peaks linked to expression of SNAP25.

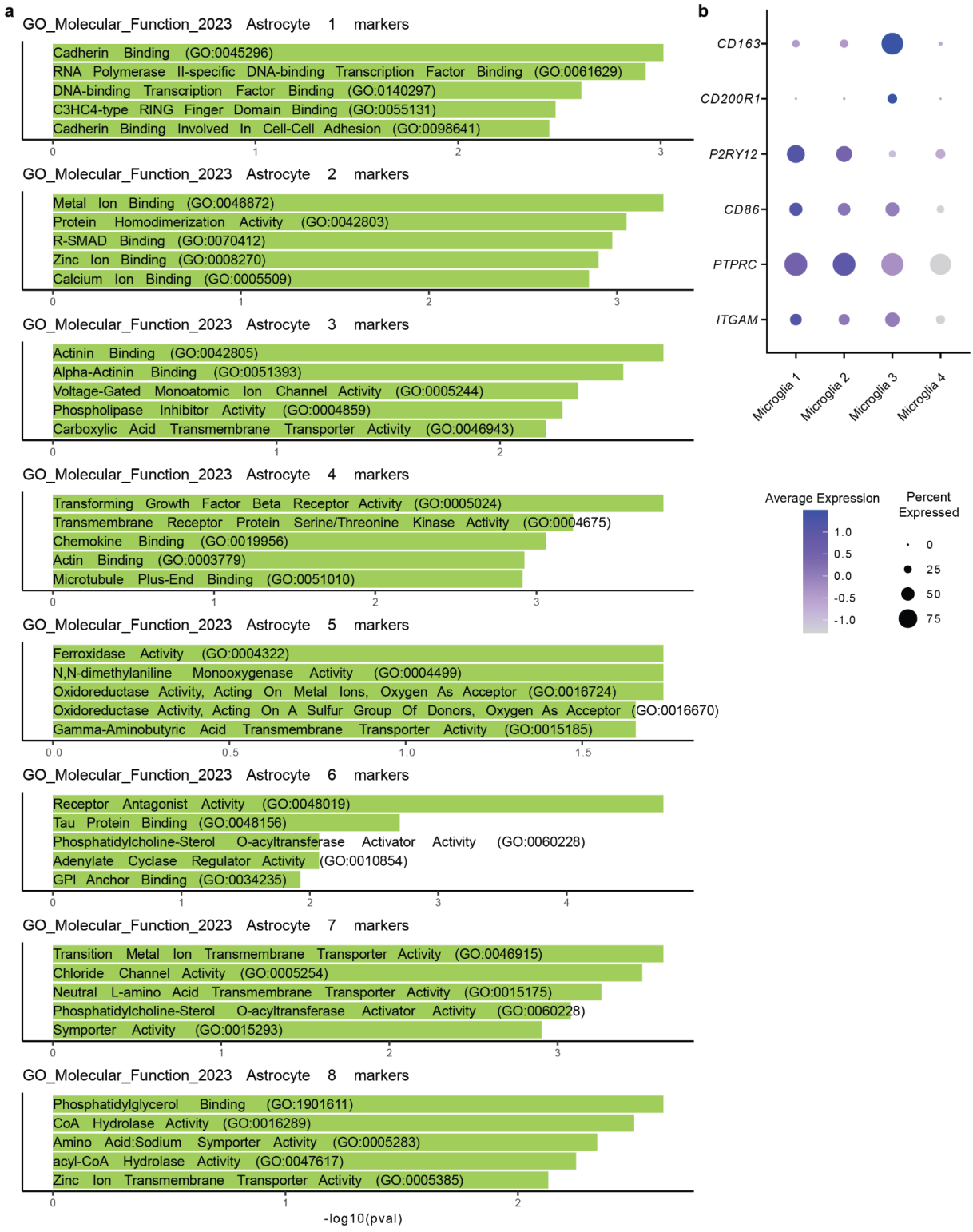

Figure S3.5 Gene set enrichment analysis and microglia gene expression profile in human hypothalamus.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

a) Barplots showing GO Molecular Function 2023 pathways enriched in differentially expressed genes of eight astrocyte cell clusters of human hypothalamus. b) Dot plot of CD163, CD200R1, P2RY12 , CD86, PTPRC , and ITGAM expression in human hypothalamus microglia populations.

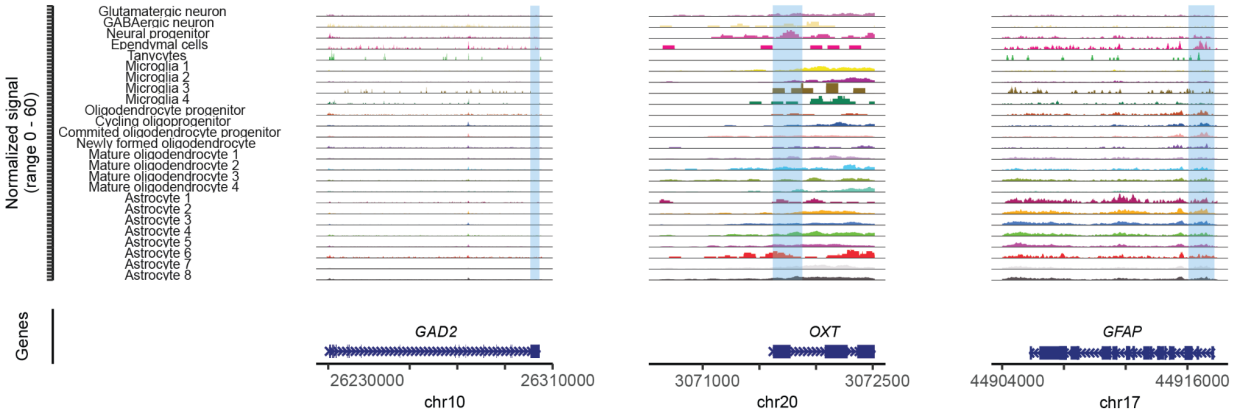

Figure S3.6 Differentially-expressed genes and differentially-accessible regions of human hypothalamus.

a) Coverage plot of differentially-expressed genes GAD2, OXT, and GFAP, with differentially-accessible scATAC peaks highlighted in blue.

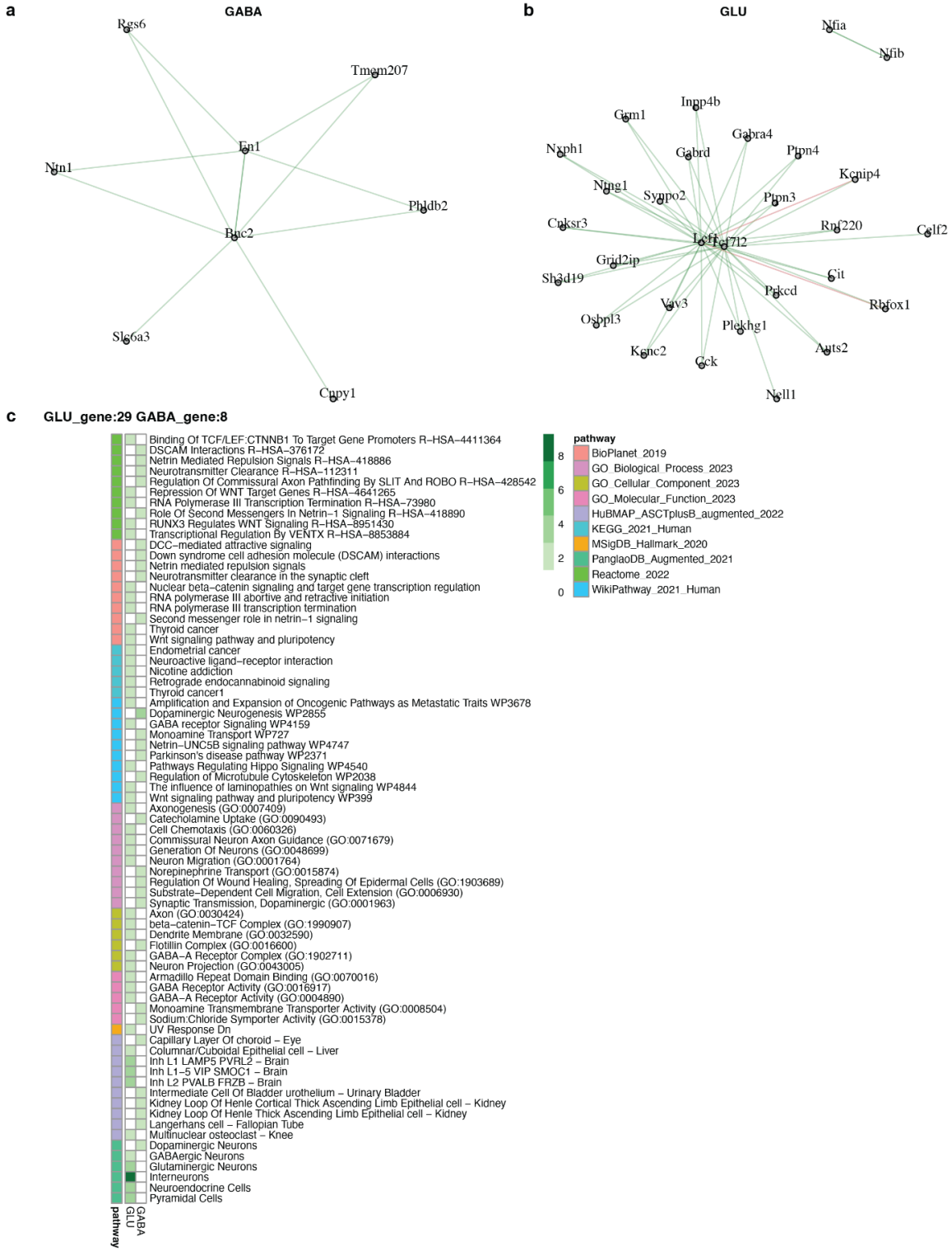

Figure S3.7 Multi-omics GRN analyses of mouse hypothalamus.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

a) GABAergic and b) glutamatergic GRNs of human hypothalamus determined by Pando. c) Heatmap of enriched pathways in human hypothalamus GABAergic and glutamatergic populations.

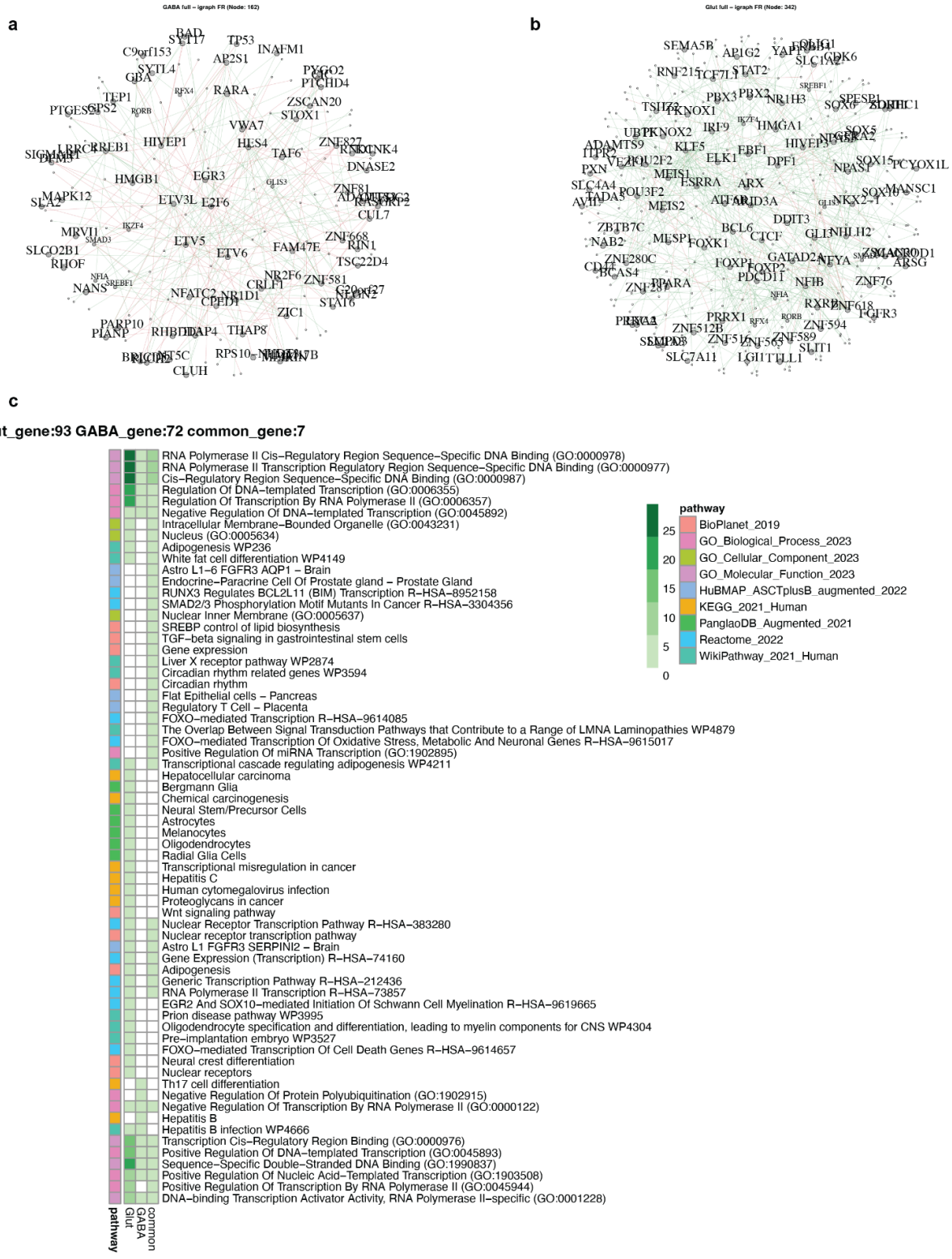

Figure S3.8 Multi-omics GRN analyses of human hypothalamus.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

a) GABAergic and b) glutamatergic GRNs of mouse hypothalamus determined by Pando. c) Heatmap of enriched pathways in mouse hypothalamus GABAergic and glutamatergic populations.

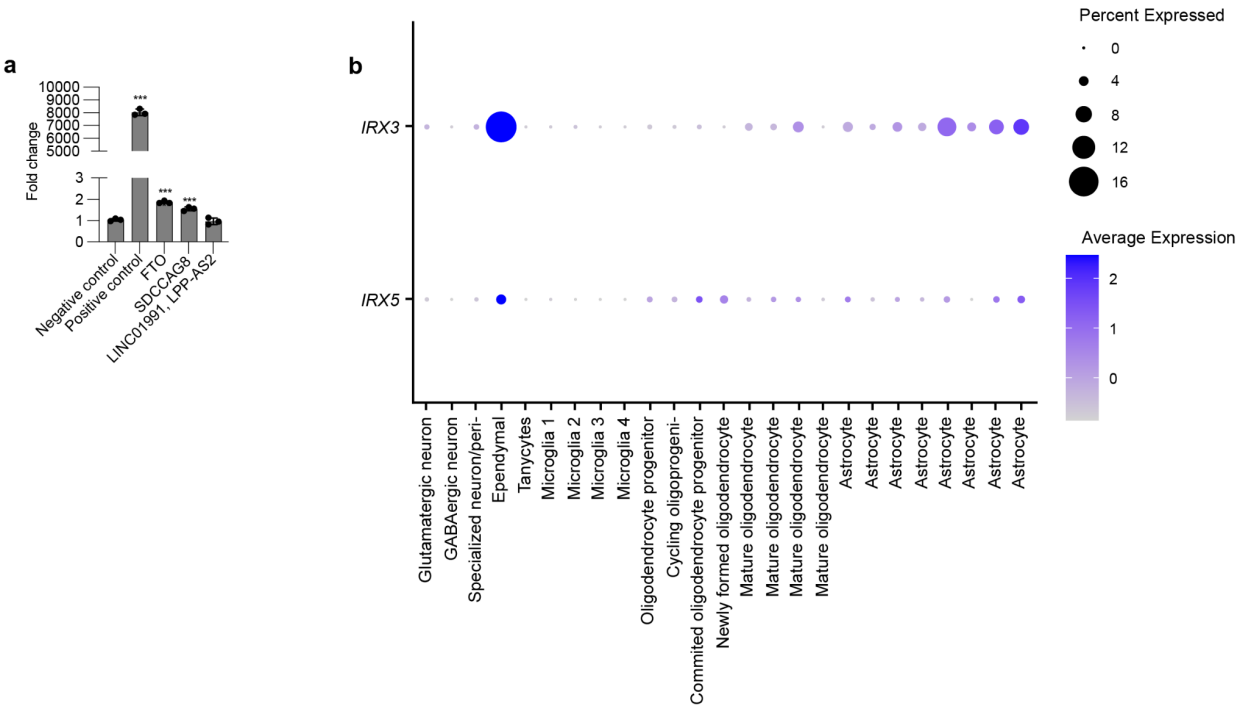

Figure S3.9 Luciferase assays and IRX3 and IRX5 gene expression in various hypothalamus cell populations.

a) Luciferase assays in human astrocytes for astrocyte-specific scATAC-seq peaks overlapping obesity-associated SNPs. b) Dot plots of IRX3 and IRX5 gene expression across cell populations in the human hypothalamus.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

Signed by:
Candace S.Y. Chan
5C7FBF7FA07547F...
Author Signature

8/30/2024
Date