Transcriptional Changes and Regulation of Voltage Gated Calcium Channels Across Human Stem Cell Differentiation



Alejandro De Los Angeles

St. John's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

2023

Acknowledgements

As I reflect on my academic journey and express gratitude in these thesis acknowledgements, I feel a sense of melancholy at the thought of departing from this chapter of my life. My time at Oxford has been some of the most rewarding and fulfilling years of my life, and I am forever grateful for the privilege to study here. I recall the day when I interviewed for Oxford and received the remarkable news from one of my esteemed DPhil advisors that I had been admitted and awarded full scholarships for my studies. The memory of that day remains vivid and serves as a testament to the immense support and encouragement that I have received throughout my academic journey. While the prospect of moving forward brings mixed emotions, I am proud of the achievements attained during my DPhil program and eager for what the future holds.

I would like to express my deep gratitude to my supervisory team, Professor Paul Harrison and Dr. Elizabeth Tunbridge for their valuable guidance and generous support throughout my journey at the University of Oxford, which was challenging given the unprecedented coronavirus pandemic. They have been an inspiration, and one cannot thank my team enough for their patience. Having the opportunity to be mentored by such outstanding individuals was truly an honor. Without them, I would not be where I am today.

I would like to acknowledge Professor Wilfried Haerty at the Earlham Institute for his mentorship and his wisdom in helping to supervise my work. The coronavirus pandemic deeply affected my DPhil as I lost access to the laboratory for conducting experimental work. He, along with Dr. Tunbridge, helped design a computational project for analyzing data from the Human Induced Pluripotent Stem Cell Initiative that later served as a foundation for my DPhil.

My gratitude goes out to each of my colleagues at the Department of Psychiatry, especially Syed Husain, Nicola Hall, Arne Mould, and Hami Lee. Thank you for helping me

to navigate the world of calcium channels, alternative splicing, and transcription factors.

I would also like to acknowledge Professor Katherine Pollard at UCSF for assistance analyzing evolutionary conservation at RBP binding sites and interpreting PhyloP scores at repeat regions, specifically the HNRNPL binding site at the CACNA2D2 locus; Mette Bentsen for advice and support regarding TOBIAS; and Richard Koche, Director of Computational Epigenetics at Memorial Sloan Kettering Cancer Center for reviewing the thesis. I reserve great thanks to the MRC WIMM Centre for Computational Biology (CCB) for providing me with the computational resources and support I needed to carry out the studies in this thesis.

Finally, I would like to express my sincere appreciation to the friends I have made during my time at Oxford, particularly those at St. John's College, the Clarendon Scholarship Program, and the Oxford Philippines Society, who have provided me with unwavering support and encouragement throughout the four years. It was my privilege to have served as President of the Oxford Philippines Society. I would like to extend my heartfelt gratitude to Greg Henderson, Andy Libertun, and Adel Reyes. Lastly, I would like to acknowledge my partner, Marra Balascopo, for her constant love and support.

Most of all, I would like to thank my parents and siblings for their boundless love and support. I am especially grateful to my brother, Carlo, for his support. I would also like to acknowledge my sister Paula and my nieces Poppy and Lilly for their emotional encouragement throughout the journey of completing the thesis. To my parents, Dr. Servando De Los Angeles and Dr. Sandra Ang, for teaching me the importance of hard work and perseverance as well as to my Tita Baby and Tito Tito for their love and constant support. You have been my pillars throughout my life. Without you, I would not have achieved my dreams.

Declarations

This dissertation is the result of my own work, and all collaborations are specified in the text. I declare that no substantial part of this dissertation has been submitted or is currently being submitted for any such degree, diploma, or other qualification at the University of Oxford or another university or similar institution.

In accordance with the guidelines, this thesis does not exceed 50,000 words, and it contains less than 150 figures.

Abstract

Human neural development comprises a phase during which a risk for psychiatric conditions such as schizophrenia is established. Robust evidence has linked voltage-gated calcium channels (VGCCs) to psychiatric illness. However, the biology of VGCCs in the context of human neural development, and particularly the prenatal stages, has not been sufficiently studied. This thesis examines the transcriptional regulation of VGCCs in a model system of neural development, the differentiation of human induced pluripotent stem cells (hiPSCs) into neurons. By applying bioinformatics tools to hiPSC and hiPSC-derived neuron data from dozens of donors, I characterized changes in the abundance and splicing of VGCCs across hiPSC differentiation into neurons. Additionally, motif-based bioinformatics tools were applied to identify candidate regulators of VGCC splicing and abundance. The analyses revealed that as hiPSCs differentiate into neurons, VGCC transcript levels robustly increase. The alternative splicing of CACNA1C, CACNA2D1, and CACNA2D2, including a cell state-specific switch in the utilization of Exons 21 and 22 in CACNA1C, were identified. Analyzing the regions of interest using two different bioinformatics tools identified a binding site for BRUNOL4 and BRUNOL5 upstream of CACNA1C Exon 21, which showed neither evolutionary conservation nor acceleration; and a binding site for HNRNPL near CACNA2D2 Exon 26 showing evolutionary acceleration in primates. Finally, transcription factor (TF) footprinting analyses of ATAC-seq data identified candidate TFs that might regulate multiple VGCC loci, several of which are expressed more highly in human fetal compared with postnatal brain development, including KLF7 and TCF4. Altogether, the results provide a foundation for future research investigating the functions of the identified RBPs and TFs in VGCC transcriptional regulation. The findings also indicate that hiPSCs possess the potential for providing insights into the changes that occur in VGCC transcription and putative mechanisms of VGCC regulation in human neural development and different cell states.

Contents

Acknowledgements **i**

Declarations **iii**

Abstract **iv**

Abbreviations **xvii**

1 Introduction **1**

1.1 Neurodevelopmental origins of psychiatric illness 1

1.2 Introduction to VGCCs 3

1.2.1 Canonical and Physiological Functions of VGCCs 6

1.2.2 VGCCs and Gene Expression Regulation 8

1.2.3 Non-Canonical Functions of VGCCs 8

1.3 VGCCs and Neural Development 9

1.3.1 VGCCs and Neural Induction from Early Embryonic Cells 9

1.3.2 VGCCs and Later Stages of Neural Development 10

1.4 VGCCs and Brain Development 10

1.4.1 VGCCs, Neural Progenitors and Neurogenesis 11

1.4.2 VGCCs, Synapses and Developing Neural Networks 11

1.5 VGCCs and Psychiatric Disorders 13

1.5.1 Common Variation in VGCCs and Psychiatric Disorders 13

1.5.2 Rare Variation in VGCCs and Brain Diseases 15

1.5.3 Summary 17

1.6 Splicing 17

1.6.1 Introns 18

1.6.2 Spliceosome 18

1.6.3 Alternative Splicing 18

1.6.4 Splicing Regulation 20
1.6.4.1 Trans-Regulation of Splicing (RBPs) 20
1.6.4.2 Cis-Regulation of Alternative Splicing 21
1.6.4.3 Epigenetic and Transcriptional Regulation of Alternative Splicing 21
1.7 VGCCs and Alternative Splicing 22
1.7.1 Effects of Alternative Splicing on VGCC Function 23
1.7.2 Regulation of VGCC Alternative Splicing 25
1.8 Transcriptional Regulation 25
1.8.1 Transcription Factors 25
1.8.2 Transcription Factor-Mediated Regulation of VGCC Abundance . . 26
1.9 Human iPSCs 27
1.10 Neural Cells and Tissues from hiPSCs 30
1.11 Aims 31

2 Transcriptional Changes of VGCCs Across Neural Differentiation of hiPSCs 33

2.1 Introduction 33
2.2 Methods 35
2.2.1 Datasets 37
2.2.2 Programming Code 38
2.2.3 Gene Expression Analyses 39
2.2.4 Alternative Splicing Analyses 39
2.2.4.1 DEXSEQ 41
2.2.4.2 MAJIQ 43
2.2.5 Statistics 45
2.2.6 Protein Sequence Alignment 46
2.3 Results 47
2.3.1 Gene Expression Changes Across hiPSC Differentiation to Neurons . 47
2.3.1.1 Transcriptome-Wide Comparison of hiPSCs and hiPSC-derived Neurons 47
2.3.1.2 Changes in the Abundance of VGCC Subunits Across hiPSC Differentiation 49

2.3.2 Changes in Alternative Splicing in VGCCs Across hiPSC Differentiation . . . 52
2.3.2.1 Changes in Differential Exon Usage Across hiPSC Differentiation . . . 52
2.3.2.2 Changes in Differential Splice Junction Usage of VGCC Subunits Across hiPSC Differentiation . . . 57
2.3.2.3 Predicted Impact of Alternative Splicing on CACNA2D1 and CACNA2D2 Function . . . 58
2.3.3 Results Summary . . . 61
2.4 Discussion . . . 61
2.4.1 Limitations . . . 63
2.4.2 Future Directions . . . 64
2.4.3 Conclusion . . . 67

3 Identification of Candidate RNA-Binding Proteins Regulating Splicing Events of Interest 68

3.1 Introduction . . . 68
3.2 Methods . . . 71
3.2.1 FIMO . . . 72
3.2.1.1 Sensitivity Analysis for FIMO . . . 77
3.2.2 HOMER . . . 78
3.2.2.1 HOMER Global . . . 78
3.2.2.2 HOMER Local . . . 81
3.2.2.3 Common Downstream Processing Steps of the HOMER Pipeline for HOMER Global and HOMER Local . . . 82
3.2.3 Intersecting FIMO and HOMER Results . . . 83
3.2.4 Evolutionary Conservation Analysis . . . 83
3.2.5 Gene Expression Analysis . . . 84
3.2.5.1 DESEQ2 . . . 84
3.2.5.2 Gene Expression Across Human Neural Development (Inspection of BrainSpan Atlas data) . . . 85
3.2.6 Statistics . . . 85
3.3 Results . . . 86
3.3.1 Assessing Sensitivity . . . 86

3.3.2 Intersection of FIMO and HOMER Global Results 89
3.3.3 HOMER Searches Using Targeted Foreground Sequence (HOMER Local) 92
3.3.4 Evolutionary Conservation of RBP Motif Occurrences 94
3.3.5 Gene Expression Analyses of FIMO and HOMER-Identified RBPs . 100
3.3.5.1 Differential Gene Expression Between hiPSCs and hiPSC-Derived Neurons 100
3.3.5.2 Gene Expression Patterns During Human Neural Development 103
3.4 Discussion 106
3.4.1 Identified RBPs 107
3.4.1.1 BRUNOL4 and BRUNOL5 107
3.4.1.2 HNRNPL 108
3.4.1.3 U2AF2 109
3.4.1.4 RBMS3 110
3.4.2 Evolutionary Conservation 111
3.4.3 Limitations 113
3.4.4 Next Steps 115
3.4.5 Summary 116

4 Identification of Candidate Transcription Factors That Regulate VGCC Abundance 117

4.1 Introduction 117
4.2 Methods 119
4.2.1 Datasets 121
4.2.2 Programming Code 121
4.2.3 Pre-Processing of ATAC-seq Data 122
4.2.4 TF Footprinting Using TOBIAS 124
4.2.5 Gene Expression Analyses 125
4.2.6 Inspection of Publicly Available ChIP-seq Data 125
4.2.7 Statistics 125
4.3 Results 126
4.3.1 In-house Application of the TOBIAS Computational Pipeline 126
4.3.1.1 Replication of Published Figures 126

4.3.1.2 Sensitivity and Specificity of TOBIAS . . . 131
4.3.2 TF Footprinting Analyses of hiPSC-Derived Neurons . . . 148
4.3.2.1 Accessible Chromatin Near VGCC TSS . . . 148
4.3.2.2 Bound TF Sites Near VGCC TSS . . . 152
4.3.2.3 Identifying Potential Regulators by Differential Gene Expression . . . 152
4.3.2.4 Searching for Binding Events For TFs Whose Footprints Were Found in VGCC Loci in Publicly Available ChIP-seq Datasets . . . 168
4.3.2.5 TF Expression Across Human Neural Development . . . 170
4.3.3 Summary . . . 173
4.4 Discussion . . . 174
4.4.1 Identification of Candidate TFs . . . 174
4.4.1.1 KLF7 . . . 174
4.4.1.2 TCF4 . . . 175
4.4.1.3 ZIC2 . . . 176
4.4.1.4 REST . . . 176
4.4.1.5 EBF1 . . . 177
4.4.2 Limitations . . . 177
4.4.3 Future Directions . . . 180
4.4.4 Conclusion . . . 181

5 Discussion 182
5.1 Introduction . . . 182
5.2 Links Between VGCC Biology and Cell State . . . 182
5.3 Interpreting Data From Motif-Based Bioinformatics Tools . . . 184
5.4 Transcriptional and Epigenetic Regulation of VGCCs . . . 187
5.5 Human iPSCs as a Model System . . . 187
5.6 Conclusions and Perspectives . . . 189

Appendix I Supplementary Material for Chapter 2 191

Appendix II Supplementary Material for Chapter 3 193

Appendix III Supplementary Material for Chapter 4 205

References **223**

List of Figures

1.1 Structure of VGCCs 4
1.2 Molecular structure of the alpha-2 delta subunits. 6
1.3 Different types of alternative splicing 19
1.4 Alternative splicing regulation by cis mRNA elements and trans-acting factors. 22
1.5 Overview of transcriptional regulation 26
1.6 Neural differentiation of hiPSCs. 29

2.1 Analyzing short-read RNA-seq data. 34
2.2 Overall strategy for identifying changes in VGCC abundance and alternative splicing across hiPSC differentiation. 36
2.3 Overview of the DESEQ2 pipeline. 40
2.4 Identifying alternative splicing events in short-read sequencing data. 41
2.5 Overview of the DEXSEQ pipeline. 42
2.6 Exons versus exonic parts. 43
2.7 Overview of the MAJIQ pipeline. 44
2.8 What is an LSV? 44
2.9 Principal component analysis (PCA) for datasets from Schwartzentruber et al., 2018. 48
2.10 Differential gene expression analysis of the CACNA1C subunit. 52
2.11 Testing for DEU in CACNA1C using DEXSEQ between hiPSCs and hiPSC-derived neurons 53
2.12 Summary of DEU across different VGCC alpha-1 subunits (regions where threshold |log2FoldChange|>1.5) 55
2.13 Summary of DEU across different VGCC alpha-1 subunits (heatmap including regions where |log2FoldChange|<1.5). 56
2.14 Differential splice junction usage in CACNA2D1 and CACNA2D2. 58

2.15 Mapping alternative splicing events onto CACNA2D1 and CACNA2D2 protein models. 60

3.1 Outline of strategy to identify candidate regulators of VGCC splicing events of interest. 71
3.2 Pipeline for conducting searches for RBP motifs in given nucleotide sequences using FIMO. 73
3.3 Definition of a PWM. 73
3.4 Scanning for motif occurrences using FIMO (general principles of motif scanning). 77
3.5 Pipeline for conducting searches for RBP motifs genome-wide using HOMER (HOMER global). 79
3.6 Representative example of a MOTIF file. 79
3.7 Pipeline for conducting searches for RBP motifs at VGCC regions of interest using HOMER. 81
3.8 Detection of FIMO motif occurrences in eCLIP datasets. 87
3.9 Detection of HOMER global motif occurrences in eCLIP datasets 88
3.10 RBP motif occurrences identified by both FIMO and HOMER global in the CACNA1C, CACNA2D1, and CACNA2D2 regions of interest. 90
3.11 Identification of BRUNOL4 and BRUNOL5 binding motif occurrence by FIMO, HOMER global, and HOMER local in the vicinity of CACNA1C Exon 21 and 22. 93
3.12 Examining the HNRNPL binding site. 98
3.13 Lack of evolutionary conservation of the BRUNOL4 / BRUNOL5 binding site relative to the intronic sequence between CACNA1C Exon 20 and Exon 21. 99
3.14 Expression of candidate RBPs during human neural development. 104
3.15 Visual representation of identified RBP motifs that contribute to alternative splicing of CACNA1C Exons 21 and Exon 22. 108
3.16 Visual representation of identified RBP motifs contributing to alternative splicing of CACNA2D2 Exon 26. 110

4.1 Overview of ATAC-seq. 118

4.2 Overall strategy to identify candidate TFs that regulate VGCC abundance across hiPSC differentiation. 120
4.3 Overview of the ATAC-seq data pre-processing pipeline. 123
4.4 TOBIAS footprinting pipeline. 124
4.5 In-house reproduction of TOBIAS analyses published in Bentsen et al. (2020). 130
4.6 Assessing the sensitivity of TOBIAS by studying mouse ESCs genetically engineered to repress Oct4 expression upon doxycycline treatment. 133
4.7 Assessing the specificity of TOBIAS by studying mouse ESCs genetically engineered to repress Oct4 expression upon doxycycline treatment. 136
4.8 Assessing TOBIAS capacity to detect the enrichment of known TFs in a well-characterized cell state transition 140
4.9 Assessing whether TOBIAS can detect a known TF binding event at the NANOG locus. 144
4.10 Assessing whether TOBIAS can detect TF binding when a TF is not expressed. 147
4.11 Accessible chromatin at the TSS of top hiPSC-derived neuron-enriched genes. 149
4.12 CACNA1C accessible chromatin regions. 151
4.13 Correlation between frequencies of TF footprints at VGCC loci and at randomly selected genes. 168
4.14 Overlapping ChIP-seq KLF7 binding events in HEK293 cells with TOBIAS analyses from hiPSC-derived neurons. 169
4.15 Expression of candidate TFs during human neural development 171

Supplementary Figure 1.1 Differential gene expression analysis of individual VGCC subunits. 191
Supplementary Figure 2.1. HOMER global search for BRUNOL4 binding motif occurrences at known BRUNOL4 targets. 195
Supplementary Figure 2.2 HOMER global search for ELAVL2 binding motif occurrences at DDX6 locus. 195
Supplementary Figure 2.3 RBP Motifs enriched at CACNA1C, CACNA2D1, and CACNA2D2 regions of interest. 197
Supplementary Figure 2.4 HOMER global search at CACNA1C region of interest Exons 21 and 22. 199
Supplementary Figure 2.5 HOMER global search at CACNA2D1 region of interest. 199

Supplementary Figure 2.6 HOMER global search at the CACNA2D2 region of interest. 200
Supplementary Figure 3.1. Assessing sensitivity and specificity of TOBIAS in the context of TF upregulation 206
Supplementary Figure 3.2. CACNA1A accessible chromatin region 209
Supplementary Figure 3.3. CACNA1B accessible chromatin region. 210
Supplementary Figure 3.4. CACNA1D accessible chromatin region. 211
Supplementary Figure 3.5. CACNA1E accessible chromatin region. 212
Supplementary Figure 3.6. CACNA1G accessible chromatin region. 213
Supplementary Figure 3.7. CACNA1H accessible chromatin region. 214
Supplementary Figure 3.8. CACNA1I accessible chromatin region. 215
Supplementary Figure 3.9. CACNA2D1 accessible chromatin region. 216
Supplementary Figure 3.10. CACNA2D2 accessible chromatin region. 217
Supplementary Figure 3.11. CACNA2D3 accessible chromatin region. 218
Supplementary Figure 3.12. CACNB1 accessible chromatin region. 219
Supplementary Figure 3.13. CACNB2 accessible chromatin region. 220
Supplementary Figure 3.14. CACNB3 accessible chromatin region. 221
Supplementary Figure 3.15. CACNB4 accessible chromatin region. 222

List of Tables

1.1 VGCC nomenclature. 5

1.2 VGCCs and psychiatric disease associations. 13

2.1 Demographic information of hiPSC donors. 38

2.2 Definitions of DESEQ2/DEXSEQ output. 46

2.3 Top 10 hiPSC-enriched genes. 49

2.4 Top 10 hiPSC-derived neuron-enriched genes. 49

2.5 DESEQ2 and DEXSEQ analyses of differential gene expression of individual VGCC subunits. 51

2.6 Summary of alternative splicing analyses findings. 57

3.1 Summary of sensitivity analyses for FIMO and HOMER. 88

3.2 Genomic coordinates of RBP motif occurrences identified by FIMO and HOMER global in the vicinity of CACNA1C Exons 21 and 22. 91

3.3 Genomic coordinates of RBP motif occurrences identified by FIMO and HOMER global in the vicinity of CACNA2D1 exon 28. 91

3.4 Genomic coordinates of RBP motif occurrences identified by FIMO and HOMER global in the vicinity of CACNA2D2 exon 26. 91

3.5 Genomic coordinates of BRUNOL4 and BRUNOL5 motif occurrence identified by FIMO, HOMER global, and HOMER local in the vicinity of CACNA1C Exon 21 and Exon 22. 92

3.6 Evolutionary conservation of RBP motif occurrences identified by FIMO and HOMER global relative to the intronic regions where they are located in (comparison of 100 vertebrate species). 95

3.7 Evolutionary conservation of RBP motif occurrences identified by FIMO and HOMER global relative to the intronic regions where they are located (comparison of 17 primate species). 96

3.8 Differential gene expression of RBPs with motifs enriched in the CACNA1C region of interest as identified by FIMO and HOMER global across hiPSC differentiation. . . . 101
3.9 Differential gene expression of RBPs with motifs enriched in CACNA2D1 region of interest as identified by FIMO and HOMER global across hiPSC differentiation. . . . 101
3.10 Differential gene expression of RBPs with motifs enriched in the CACNA2D2 region of interest as identified by FIMO and HOMER global across hiPSC differentiation. . . . 102
3.11 Expression levels of RBPs that are more highly expressed in fetal relative to postnatal brain development (frontal cortex). . . . 105
3.12 Comparison of RBP expression levels between early and late fetal development. . . . 106
3.13 Summary of results. . . . 107

4.1 Demographic information of hiPSC donors. . . . 122
4.2 Summary of TOBIAS TF footprinting analyses. . . . 152
4.3 Gene expression analyses of TFs with footprints that were detected at VGCC TSS and with an expression detected as differentially expressed between hiPSCs and hiPSC-derived neurons. . . . 165
4.4 TFs and the number of TF footprints at various VGCC loci. . . . 166
4.5 Prevalence of TF footprints in randomly selected genes. . . . 167
4.6 Inspection of published ChIP-seq datasets partially validates TOBIAS footprinting results. . . . 170
4.7 Expression levels of TFs in fetal relative to postnatal brain development (frontal cortex). . . . 172
4.8 Comparison of TF expression levels between early and late fetal development. 173

Supplementary Table 2.1 100 vertebrate species analyzed in 100 species evolutionary conservation analysis. . . . 203
Supplementary Table 2.2 17 primate species analyzed in 100 species evolutionary conservation analysis. . . . 203
Supplementary Table 2.3 Brain regions profiled by RNA sequencing for BrainSpan Developmental Transcriptome analyses. . . . 203

Abbreviations

ASCL1 Achaete-Schute Homolog 1

ATAC-seq Assay of Transposase Accessible Chromatin

BRUNOL4 Bruno-like 4

BRUNOL5 Bruno-like 5

CELF4 CUGBP Elav-like family member 4

CELF5 CUGBP Elav-like family member 5

ChIP-seq Chromatin immunoprecipitation followed by high-throughput DNA sequencing

DEU Differential exon usage

DEXSEQ Differential Exon and Transcript analysis for RNA-seq

eCLIP enhanced Cross-Linking and Immunoprecipitation

ELAVL2 Embryonic lethal abnormal vision like 2

FIMO Find Individual Motif Occurrences

hiPSC human induced pluripotent stem cell

HOMER Hypergeometric Optimization of Motif EnRichment

iPSC induced pluripotent stem cell

KLF Kruppel-like factor

LSV local splice variation

MAJIQ Modeling Alternative Junction Inclusion Quantification

Oct4 Octamer-binding transcription factor 4

PhyloP Phylogenetic P-values

PSSM Position-specific score matrix

POU5F1 POU domain, class 5, transcription factor 1

PWM Position weight matrix

RBMS3 RNA-binding motif, single-stranded-interacting protein 3

RBP RNA-binding protein

RNA-seq RNA sequencing

SOX2 SRY (sex determining region Y)-box 2

STAR Spliced alignment to a reference

TCF4 Transcription Factor 4

TF Transcription factor

TOBIAS Transcription factor Occupancy prediction By Investigation of ATAC-seq Signal

U2AF2 U2 auxiliary factor 2

VGCC Voltage-gated calcium channel

Chapter 1

Introduction

Many genomic loci identified through genome-wide association studies (GWAS) have been associated with psychiatric disorders including bipolar disorder and schizophrenia. One of the most reproducible associations with psychiatric disorders is the family of voltage-gated calcium channels (VGCCs), which regulate the entry of calcium into cells. Neural development comprises a phase during which the risk of later psychiatric disease can be established (Birnbaum and Weinberger, 2017). However, our understanding of VGCC function and regulation during development remains primitive. This thesis will investigate the transcriptional regulation of VGCCs during neural differentiation of human induced pluripotent stem cells (hiPSCs), a model system for human neural development. A deeper understanding of the transcriptional regulation of VGCCs could lay a foundation for future investigation of VGCC biology during human neural development.

1.1 Neurodevelopmental origins of psychiatric illness

Psychiatric conditions such as schizophrenia and bipolar disorder are thought to be caused by genetic and environmental factors, and many insights have been gained into risk factors. GWAS studies have identified both common and rare risk variants for these conditions, but in most cases, their relevance to neurobiology is unclear, and detailed functional characterization is lacking.

The risk of psychiatric conditions, such as schizophrenia, might, in part, be established during neural development (Birnbaum and Weinberger, 2017). The neurodevelopmental model puts forward the view that abnormal brain development causes mental illness through the influences of genetic and epigenetic risk factors on early neural development, resulting in a perturbed developmental trajectory that eventually culminates in the man-

ifestation of symptoms and illness in adulthood. In this model, a disease risk-increasing process happens early in neural development, decades before symptoms appear (Birnbaum and Weinberger, 2017). The disease risk-increasing process is subtle and can be compensated for early in life. Later, decompensation occurs, psychiatric symptoms manifest, and brain abnormalities that developed earlier during development are unveiled (Birnbaum and Weinberger, 2017).

Much of the early data for the neurodevelopmental model comes from epidemiological studies of schizophrenia (Birnbaum and Weinberger, 2017). Cohort studies found that a prenatal insult, such as starvation, can increase schizophrenia risk in offspring (Susser and Lin, 1992; St Clair et al., 2005). Other public health studies found that in utero or prenatal maternal infections, such as influenza and rubella, among others, are linked to an increased incidence of schizophrenia (Brown et al., 2010). At later stages of neural development, other studies have identified links to perinatal obstetric complications and childhood viral infections (Cannon et al., 2002; Nielsen et al., 2014; Khandaker et al., 2012).

More evidence for the neurodevelopmental model comes from recent genomic analyses (Birnbaum and Weinberger, 2017). Many genes associated with psychiatric conditions, such as schizophrenia, have been linked to early developmental processes such as neuronal differentiation and maturation. Transcriptomic analyses of post-mortem brains across human development have found that the genetic risk of schizophrenia is played out in development (Birnbaum and Weinberger, 2017). Data from Human Brain Transcriptome microarray or BrainSpan RNA-sequencing (RNA-seq) has found that rare coding variants linked to schizophrenia risk whose expression is enriched during fetal development (Xu et al., 2012; Gulsuner et al., 2013; Gilman et al., 2012). Autism spectrum disorder-associated genes also tend to be expressed fetally (Willsey et al, 2013; Miller et al, 2014). Another study interrogating prefrontal cortex transcriptomes found that the prenatal stage possesses an enriched expression of GWAS-positive loci for schizophrenia (Jaffe et al., 2015). Finally, alternative splicing has been linked to neural development-mediated risk of illness. Investigations of BrainSeq prefrontal cortex samples have found that isoform shifts from fetal to postnatal stages are enriched for schizophrenia risk loci genes (Jaffe et al., 2018). In summary, analyses suggest genetic contributions to illness risk is in part established during neural development.

In terms of specific loci, some disease-associated variants appear to manifest during

neural development but not during adulthood (Birnbaum and Weinberger, 2017). For instance, for ZNF804A, a transcription factor (TF) linked to schizophrenia risk, the risk allele decreases ZNF804A expression in the second-trimester fetal brain but has no effect on the adult brain (Hill and Bray, 2012). Another study implicating developmentally regulated alternative splicing reported an alternatively spliced, truncated ZNF804A transcript found an intronic risk variant associated with reduced expression in the fetal brain, but with no link with full-length ZNF804A in the fetal or adult brain (Tao et al., 2014).

While evidence supports the neurodevelopmental model of mental illness, our understanding of how VGCCs, a family of proteins reproducibly linked to psychiatric disease, function and are regulated during human neural development is less understood. In mice, VGCCs play a role in regulating neurite extension and radial migration during neural development (Kamijo et al., 2018). Given the links between VGCCs and psychiatric diseases, it would be of interest to characterize VGCC function and regulation during human neural development.

1.2 Introduction to VGCCs

Calcium signaling is involved in many cellular processes. VGCCs are a type of ion channels found on the cell membrane of different cell types, including neurons, muscle, and endocrine cells. They enable the selective passage of calcium ions across cell membranes in response to changes in the electrical potential across the cell membrane. The opening and closing of VGCCs are regulated by changes in the membrane potential. When the membrane potential becomes more positive (i.e. when the cell depolarizes), VGCCs open, causing calcium ions to enter the cell.

There exist different types of VGCCs: L-type, P/Q-type, N-type, R-type, and T-type. L-type calcium channels are prevalent in muscle cells and neurons, enabling muscle contraction and hormone secretion. P/Q-type and N-type channels, primarily located in the brain, play essential roles in neurotransmitter release at synapses. R-type channels, also found in neuronal tissues, contribute to neurotransmitter release as well. T-type channels, known for their transient activity, help regulate pacemaking activities in the heart and neurons.

VGCCs are comprised of multiple subunits that function together as a channel. The multi-subunit complexes are typically comprised of alpha-1, alpha-2, beta, and delta subunits (**Figure 1.1**).

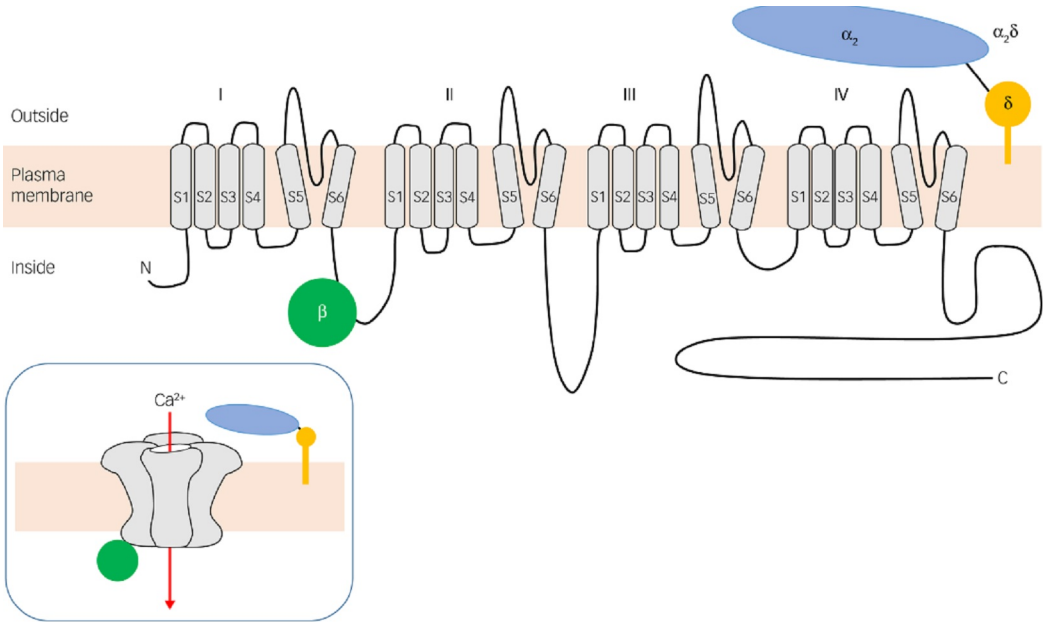


Figure 1.1: Structure of VGCCs

Schematic of VGCC topology. The alpha-1 subunit, embedded in the cell membrane contains four voltage-sensing domains (I-IV), each of which is comprised of six helices (S1-S6) that form the channel pore. Also indicated are the alpha2-delta subunit and the beta subunit. The alpha2 subunit is extracellular and the delta subunit is anchored to the membrane. The inset shows how the protein folds to produce a pore through which calcium channels pass (taken from Harrison et al., 2020).

The alpha-1 subunit forms the main pore and is comprised of four repeat domains (I-IV) each containing six transmembrane segments (S1-S6). Both amino and carboxyl terminals are intracellular. The alpha-1 subunit forms the calcium-selective pore and contains the voltage sensor and regulatory binding sites. The S4 segment of each domain behaves as a voltage sensor and positive charges on the S4 segment shift in response to changes in membrane potential. When membrane depolarization occurs, S4 segments move, which results in the opening of the channel pore, allowing calcium to enter the cell.

The function of the alpha-1 subunits is regulated by protein interaction partners, including alpha-2-delta and beta auxiliary subunits. An overview of VGCC nomenclature is provided in **Table 1.1** .

Subunit type	Subunit name	Gene symbol	Channel name	Channel Type
Alpha (α_1)	$\alpha_l S$	CACNAlS	Cav1.1	L-type
	$\alpha_l C$	CACNA1C	Cav1.2	
	$\alpha_l D$	CACNA1D	Cav1.3	
	$\alpha_l F$	CACNAlF	Cav1.4	
	$\alpha_l A$	CACNA1A	Cav2.1	P-/Q-type
	$\alpha_l B$	CACNA1B	Cav2.2	N-type
	$\alpha_l E$	CACNAlE	Cav2.3	R-type
	$\alpha_l G$	CACNA1G	Cav3.1	T-type
	$\alpha_l H$	CACNA1H	Cav3.2	
	$\alpha_1 l$	CACNAll	Cav3.3	
Beta (β)	β_1	CACNB1		
	β_2	CACNB2		
	β_3	CACNB3		
	β_4	CACNB4		
Alpha2delta ($\alpha_2\delta$)	$\alpha_2\delta_1$	CACNA2D1		
	$\alpha_2\delta_2$	CACNA2D2		
	$\alpha_2\delta_3$	CACNA2D3		
	$\alpha_2\delta_4$	CACNA2D4		

Table 1.1: VGCC nomenclature.

Subunit type, subunit name, gene symbol, channel name, and channel type. (Adapted from Harrison et al, 2018).

The alpha-2-delta subunit is an auxiliary subunit associated with VGCCs. It is comprised of two subunits (alpha-2 and delta) encoded by a single gene. After translation, the alpha-2-delta subunit is cleaved into the alpha-2 and delta subunits, which are linked via a disulphide bond. Both the alpha-2 and delta subunits reside extracellularly, although only the delta subunit is attached to the plasma membrane (with almost no intracellular residues). The functions of the alpha-2-delta subunit include regulating the trafficking and localization of VGCCs to the plasma membrane, modulating the gating properties of VGCCs, and influencing the kinetics of channel activation or inactivation. The alpha-2-delta subunits possess different protein-interacting domains. Two notable examples include a Von Willebrand factor-A (VWA) domain that shares similarities with integrin extracellular matrix binding regions (Whittaker and Hynes, 2002), and a CACHE domain (Anantharaman and Aravind, 2000), which is less well-studied. The molecular structure of an alpha-2 delta subunit is presented in **Figure 1.2**.

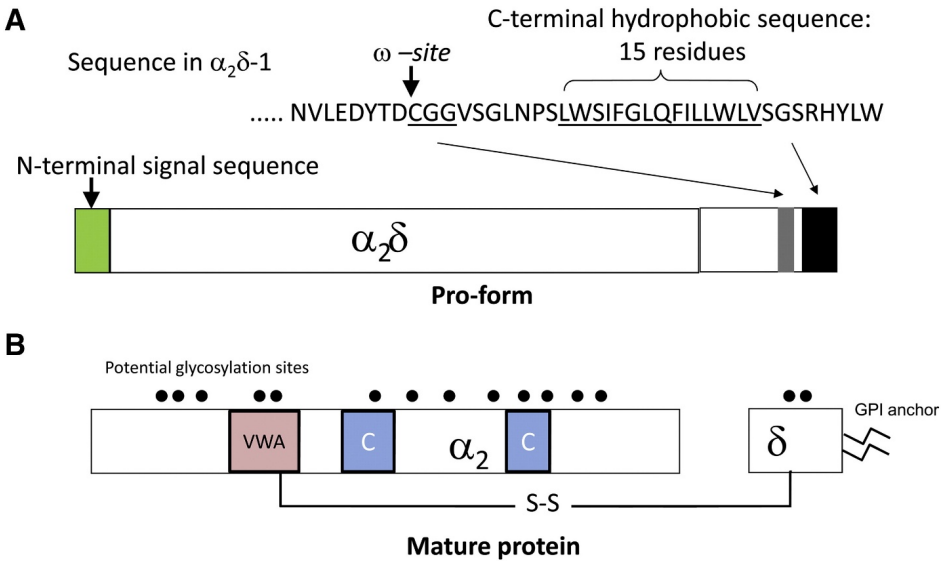


Figure 1.2: Molecular structure of the alpha-2 delta subunits.

The alpha-2-delta unit-1 is shown. A) Pro-protein with key features. B) Mature alpha-2 delta-1 protein cleaved into alpha-2 and delta subunits. The positions of the VWA domain and the two chemosensory domains are presented. The inter-subunit disulfide bond between alpha-2 and delta is indicated. Glycosylation sites (•) are indicated as well (taken from Dolphin, 2013).

The beta subunits are another type of auxiliary subunit associated with VGCCs. Their functions include regulating the gating properties of the alpha-1 subunit and influencing the kinetics of channel activation and inactivation. This can be achieved by modulating the voltage sensitivity and speed of activation or inactivation of the alpha-1-subunit (Lacerda et al., 1991; Josephson and Varadi, 1996). Additionally, the beta subunits also regulate their trafficking to the plasma membrane via interactions with the alpha-1 subunits (Buraei and Yang, 2010).

1.2.1 Canonical and Physiological Functions of VGCCs

VGCCs are found on the plasma membrane of excitable cells and are permeable to calcium. At resting membrane potential, VGCCs are typically closed. When the cell is depolarized, VGCCs are activated, allowing the flow of calcium along its electrochemical gradient. Typically, at resting membrane potential, the concentration of calcium is several orders of magnitude higher outside compared with inside the cell and so calcium flows inwards. The resulting calcium influx into the cell causes different phenotypes depending on the cell type, including muscle contraction, neuronal excitation, gene expression changes and/or hormone or neurotransmitter release.

VGCCs possess myriad functions in the nervous system. All but one type of calcium channel – the skeletal muscle-specific CACNA1S (Cav1.1) channels – are expressed in the

nervous system (Zamponi, 2015). CACNA1C (Cav1.2) channels are present in most neuronal types and are thought to be key for the activation of calcium-dependent enzymes (Hutchinson et al., 2015), calcium-activated potassium channels (Ramirez-Latorre, 2012), and for calcium-dependent gene expression changes (as described below; Dolmetsch et al., 2001). CACNA1D (Cav1.3) channels are present in dopaminergic neurons whose death causes Parkinson's disease (Chan et al., 2007). The dependence of dopaminergic neurons on CACNA1D (Cav1.3) thought to drive their pacemaking is thought to make them susceptible to stress that can cause Parkinson's disease (Chan et al., 2007). CACNA1D is also present in cochlear hair cells where they function in auditory transmission (Brandt et al., 2005). Transgenic Cacna1d knockout mice and human patients with null mutations in CACNA1D are deaf (Platzer et al., 2000; Baig et al., 2011). CACNA1F (Cav1.4) channels are expressed in rod cells where they mediate glutamate release. Mice lacking Cacna1f and the loss of CACNA1F expression in human patients result in night blindness (Lodha et al., 2010; Lodha et al., 2012).

CACNA1A (Cav2.1) and CACNA1B (Cav2.2) are present at pre-synapses and regulate neurotransmitter release (Wheeler et al., 1994; Zamponi, 2003). Cacna1a mice tend to have seizures and cerebellar ataxia, whereas Cacna1b knockout mice are hyposensitive to pain (Hatakeyama, 2001; Jun et al., 1999). Ataxia refers to poor muscle control that can lead to clumsy voluntary movements, among other symptoms related to poor muscle coordination. Human patients with gain-of-function mutations in CACNA1A have familial hemiplegic migraine, whereas loss-of-function mutations in CACNA1A cause episodic ataxia type II and spinocerebellar ataxia 6 (Saegusa et al., 2007; Nachbauer et al., 2014; Vecchia et al., 2015). Familial hemiplegic migraine refers to a genetically transmitted form of migraine that includes weakness of half the body. Episodic ataxia type II is a neurological disorder that manifests with attacks of imbalance, vertigo and ataxia. Spinocerebellar ataxia type 6 is distinguished by slowly progressive ataxia, dysarthria (difficult articulation of speech) and nystagmus (involuntary movement of the eyes). CACNA1B mutations have been associated with a myoclonus dystonia-like syndrome (Groen et al., 2015). Myoclonus-dystonia is a movement disorder that involves twitching (myoclonus) and involuntary tensing of muscles (dystonia).

Cav3 family members (CACNA1G, CACNA1H, and CACNA1I) control neuronal firing behavior (Molineux et al., 2006). Altogether VGCCs have been linked to a wide variety of functions in the mammalian nervous system, and their genetic perturbations have been

linked to various phenotypic manifestations resembling neurological conditions (Zamponi, 2015).

1.2.2 VGCCs and Gene Expression Regulation

VGCCs can also influence gene regulation through different mechanisms. One way is by the activation of calcium-dependent signaling pathways. Calcium entering through VGCCs binds to various intracellular signaling molecules, eventually activating transcription factors (TFs) such as CREB2 and MEF2 (Dolmetsch et al., 2001). Calcium-calmodulin binding to VGCCs is essential to turn on the MAPK pathways, which eventually leads to the activation of CREB and MEF2. Two different mechanisms lead to the activation of CREB. First, calcium itself enters the nucleus and turns on kinases that activate CREB by phosphorylation (Dolmetsch et al., 2001). Second, calcium acts at the membrane to produce a second messenger signal that turns on nuclear kinases (Dolmetsch et al., 2011). In addition to CREB and MEF2 activation, it is likely that VGCCs can regulate gene expression by other mechanisms as a result of calcium influx.

1.2.3 Non-Canonical Functions of VGCCs

VGCC subunits have functions that are unrelated to their role in forming ion channels. For example, certain subunits may regulate genes by acting as TFs. Gomez-Ospina and colleagues showed that a C-terminal fragment of Cacna1c, termed calcium channel-associated transcriptional regulator (CCAT), possesses a nuclear localization signal and acts as a TF, regulating genes important for neuronal signaling (Gomez-Ospina et al., 2006).

Some VGCC auxiliary subunits also possess non-canonical functions. Alpha-2-delta subunits can act as a thrombospondin receptor to modulate synapse formation, a function unrelated to VGCC modulation (Eroglu et al., 2009; Kurshan et al., 2009). Eroglu and colleagues demonstrated that the VWA domain of CACNA2D1 interacts with EGF-like repeats found in thrombospondins. CACNA2D1 is involved in excitatory synaptogenesis (Eroglu et al., 2009). Gabapentin, a neuropathic pain medication that can bind CACNA2D1 with high affinity, strongly inhibits thrombospondin/astrocyte-induced synapse formation (Eroglu et al., 2009). Alternatively, Cacna2d3 expression promotes synaptic morphogenesis and cytoskeletal organization in a manner unrelated to its role in modulating alpha-1 subunit biology (Kurshan et al., 2009).

Additionally, a short VGCC beta4 subunit can ameliorate the repressor function of a

chromatin protein, HP1 (Hibino et al., 2003). Moreover, beta3 can inhibit the transcriptional activity of TF Pax6 in vitro (Zhang et al., 2009).

Altogether, evidence suggests certain VGCC subunits possess non-canonical functions that are unrelated to calcium channel activity.

1.3 VGCCs and Neural Development

This thesis investigates the regulation of VGCCs in a model system of human neural development. As a caveat, VGCCs may play a role in psychiatric disorder etiology independently of their developmental role. Nonetheless, the prenatal stages comprise a phase when a risk for psychiatric disorders plays out, making it of interest to explore VGCC biology during human neural development. VGCCs have various functions in different stages of neural development, including during the earliest stages of neural induction from early embryonic cells to later stages of differentiation of neural precursors into terminally differentiated derivatives.

1.3.1 VGCCs and Neural Induction from Early Embryonic Cells

VGCCs play an instrumental role in the early neural induction of the vertebrate embryo (Moreau et al., 2008). Development of the vertebrate embryo involves the induction of the three germ layers, ectoderm, mesoderm and endoderm. In particular, the ectoderm lineage has the choice of two fates: neural progenitors and epiderm. Neural fate is in part specified by calcium signaling (Moreau et al, 2008). Calcium is required for the induction of neural identity, whereas the alternative epidermal fate is specified when calcium signaling is inactivated (Moreau et al, 2008). Several decades ago, Barth and Barth suggested that calcium is required to induce neural fate in Rana pipiens embryos (Barth and Barth, 1964; Moreau et al., 2008). Embryo dissociation in a calcium-free and magnesium-free medium triggers neural induction by simulating an increase in intracellular calcium (Grunz and Tacke, 1989; Leclerc et al, 2001; Moreau et al, 2008). Neural induction is inhibited when embryo cells are treated with calcium chelator BAPTA (Leclerc et al., 2001). Taken together, these studies indicate that calcium is necessary for the induction of neural fate in vertebrate embryo species (Moreau et al., 2008).

Functional experiments support a role for VGCCs in neural induction. VGCC blockade inhibits neural induction (Leclerc et al., 1997). Small molecule inhibition of LTCCs by nifedipine or nimodipine blocks calcium transients and decreases calcium levels, demon-

strating that calcium changes are mediated by VGCCs. Small molecule inhibition of VGCCs reduces the expression of neural master regulators Zic3 and geminin and triggers anterior nervous system abnormalities (Leclerc et al., 2000; Leclerc et al., 2003). Alternatively, embryos treated with VGCC agonists, such as S(-)Bay K 8644, generate an increase in intracellular calcium to induce neural markers and to generate neurons and glial cells (Moreau et al., 1994). Taken together, studies in amphibians indicate neural induction requires VGCCs.

While the preponderance of the evidence of the role of VGCCs in neural induction arises from analyses of amphibian embryos, it is possible that VGCCs also play a role in human neural induction. One potentially interesting line of research is to determine whether VGCCs participate in the induction of neural fate in humans.

1.3.2 VGCCs and Later Stages of Neural Development

VGCCs also play roles in later stages of neural development, including neurogenesis, the emergence of neuronal network activity, and the differentiation of other neural cell types. Below, studies that have reported a role for VGCC function in neural differentiation will be reviewed.

1.4 VGCCs and Brain Development

While VGCCs might affect brain function through development-dependent and development-independent modalities, this thesis will focus on brain development. In this regard, Cacna1c has been shown to have a function in brain development and adult behavior in mice. A recent study has shown that the genetic disruption of Cacna1c in a forebrain-specific manner (by mating Cre-Emx1 mice with loxp-Cacna1c mice) can disrupt spontaneous Ca2+ activity in neural progenitors, affect cerebral cortical development, and cause anxiety (Smedler et al., 2022). These Cacna1c knockout mice showed abnormal brain anatomy but normal cortical layering. The knockout mice manifested no significant changes in total brain volume, but significant increases were noted in the volume of the periaqueductal gray, internal capsule, striatum, midbrain, globus pallidus and, mammillary body. Additionally, decreases in the volume of the hippocampus, neocortex, and trigeminal tract were observed. However, no significant changes in the thickness of layers were noted. Key issues with this study's report of regional brain volumes could include limitations arising from the MRI and histochemistry methods used to measure neocortex

volume and layer formation. However, if such results are indeed robust, genetic ablation of Cacna1c might affect cortical development without perturbing cortical layering architecture.

Alterations in local brain volume could be owing to changes in mechanisms underlying proliferation, differentiation, migration, and apoptosis (Smedler et al., 2022). Indeed, it has been observed that mice harboring a gain-of-function mutation in Timothy Syndrome (TS) showed calcium-dependent differentiation defects that can affect the number of neural progenitors and impair the radial migration of layer 2/3 excitatory neurons (Kamijo et al., 2018; Panagiotakos et al., 2019). Moreover, genetic alterations in Cacna1c can affect oligodendrocyte function as well. Knockout of Cacna1c can cause hypomyelination, and gain-of-function TS mutation triggers more complex morphology and elevated levels of myelin protein expression (Cheli et al., 2016; Cheli et al., 2018).

1.4.1 VGCCs, Neural Progenitors and Neurogenesis

A key function of VGCCs is to regulate neurogenesis in adult neural stem/progenitor cells. In a key study by Deisseroth et al., the authors demonstrate that VGCCs CACNA1C and CACNA1D mediate excitatory stimuli-induced promotion of adult neurogenesis (Deisseroth et al., 2004). In their model, excitatory stimuli act directly on hippocampal neural progenitor cells to enhance neuron production. Excitation is detected through CACNA1C/CACNA1D VGCCs on proliferating precursors. The application of a CACNA1C/CACNA1D channel agonist enhances neurogenesis by inhibiting the expression of glial TFs Hes1 and Id2 and promoting the expression of NeuroD, a master neuronal TF. CACNA1C shifts the glia-to-neuron ratio in favor of neurons. Altogether, this study suggests that VGCC activity may promote hippocampal neurogenesis and can particularly bias neural progenitor differentiation in favor of neuron over glial fate.

1.4.2 VGCCs, Synapses and Developing Neural Networks

A key function of VGCCs involves mediating calcium entry through presynaptic VGCCs, which can lead to the release of neurotransmitters (Catterall et al, 2008). Regulation of such presynaptic channels can affect synaptic strength. Neural activity regulates the ability of VGCCs to open, close or, inactivate after membrane depolarization. Channel regulation can powerfully affect neurotransmission (Catterall et al, 2008).

Another function of VGCCs involves the regulation of cortical neuronal networks.

Plumbly et al. (2019) observed the regulation of synchronized network activity in hiPSC-derived neural cultures due to L-type VGCCs. The authors first observed the dose-dependent effects of calcium channel blockers (CCBs) on synchronous firing patterns in hiPSC-derived neural networks. Application of the CCBs diltiazem and nifedipine triggered a dose-dependent increase in the number of synchronized bursts (SBs) and a reduction in the interval between SBs. VGCCs appear to play a role in aiding the conversion of neural cells from spontaneously active cells to more mature cells with characteristics seen in rodent primary neurons.

Spontaneous opening of VGCCs plays a key function in neural development and corticogenesis (Kamijo et al., 2018). Kamijo et al. found that VGCCs help to determine the calcium signaling activity in immature mouse neurons. Knocking out of VGCCs in immature neurons also causes aberrant neuritic growth and cortical migration (Kamijo et al., 2018). Altogether, disrupted cortical wiring could be one way in which dysregulated Cav1.2 activity deleteriously affects development.

Interestingly, the auxiliary subunits of VGCCs have also been identified as playing roles in the development of neuronal networks. In one study, the authors demonstrated that CACNA2D1 overexpression promotes spontaneous neuronal network activity in differentiating hippocampal neurons (Bikbaev et al., 2020). Overexpression of CACNA2D3 increased neuronal firing in less differentiated neurons but inhibited neuronal activity later in differentiation. These findings indicate that CACNA2D1 and CACNA2D3 have different but supporting functions in the generation of neuronal network connectivity (Bikbaev et al., 2020). Altogether, changes in the expression of auxiliary subunits during differentiation may affect excitatory-to-inhibitory balance and the development of neuronal networks.

Finally, triple knockout/knockdown of alpha-2-delta subunits (Cacna2d1, Cacna2d2, and Cacna2d3) has shown that pre-synaptic alpha-2-delta subunits are crucial for the appropriate differentiation of glutamatergic synapses (Schopf et al., 2021). Presynaptic alpha-2-delta deficiency has been associated with a decrease in calcium currents, presynaptic clustering of N-type and P/Q-type VGCCs, and active zone size (Schopf et al, 2021).

1.5 VGCCs and Psychiatric Disorders

Robust genomic evidence linking VGCC subunit genes with psychiatric disorders has been unveiled by GWAS and experimental systems (Harrison et al., 2020; Harrison et al., 2022). A table of VGCC genetic associations with psychiatric disorders is provided in **Table 1.2** (Harrison et al., 2020).

Gene symbol	GWAS association	Rare variants
CACNAlS	Cross-diagnostic	SCZ
CACNA1C	BD, SCZ, ASD, cross-diagnostic	ASD, BD, SCZ
CACNA1D	Cross-diagnostic	ASD, BD
CACNA1F		
CACNA1A		
CACNA1B		BD, SCZ
CACNAlE	MDD, cross-diagnostic	
CACNA1G		
CACNA1H		ASD, SCZ
CACNA1I	ASD, SCZ	
CACNB1		
CACNB2	SCZ	ASD
CACNB3		
CACNB4		SCZ
CACNA2D1	MDD	SCZ
CACNA2D2	Cross-diagnostic	SCZ
CACNA2D3		ASD
CACNA2D4	Cross-diagnostic	SCZ

Table 1.2: VGCCs and psychiatric disease associations.

GWAS associations (common variants) and rare variants are indicated. BD, bipolar disorder; SCZ, schizophrenia; ASD, autism spectrum disorder; MDD, major depressive disorder; Cross-diagnostic (SCZ/BD/MDD/ASD/attention-deficit hyperactivity disorder). (adapted from Harrison et al. 2020).

1.5.1 Common Variation in VGCCs and Psychiatric Disorders

Among the first genomic associations made between a VGCC locus and psychiatric illness was CACNA1C for bipolar disorder (Ferreira et al., 2008; Sklar et al., 2008). However, the link between VGCCs and psychiatric diseases extends beyond CACNA1C and bipolar disorder. GWAS studies have also linked CACNA1C and other VGCC loci to the risk of schizophrenia and major depression (Trubetskoy et al., 2022), and genomic studies have identified other VGCC subunits in bipolar disorder including CACNA1D and CACNA1B. Additionally, CACNB2 has been linked to several psychiatric disorders. Altogether, GWAS data provide robust evidence that VGCC subunit genes, particularly CACNA1C, are a key component of the genetic risk underlying severe psychiatric disease (Casamassima et al., 2010; Bhat et al., 2012; Heyes et al., 2015; Harrison et al., 2022).

However, despite the strong genomic evidence implicating VGCC loci, some questions remain.

First, the mechanistic basis for each genetic association needs to be clarified (Harrison et al., 2022). The genomic hits for VGCCs are frequently non-coding, suggesting that each genetic association possibly affects transcriptional regulation, particularly either abundance or splicing. Although different SNPs have been identified in CACNA1C, one well-reproduced hit involves the risk SNP rs1006737, located within the large intron 3 (300 kb) of CACNA1C. SNP rs1006737 has been linked to bipolar disorder, major depressive disorders, and schizophrenia (Ferreira et al., 2008; Green et al., 1010; Nyegaard et al., 2010). There is a lack of clarity as to whether risk SNPs such as rs1006737 are associated with increases or decreases in CACNA1C mRNA transcript levels (Bigos et al., 2010; Gershon et al., 2014; Yoshimizu et al., 2015; Jaffe et al., 2020). While a few studies have reported higher CACNA1C expression in patients, or in cells from patients with the risk allele (rs1006737) (Bigos et al., 2010; Yoshimizu et al., 2015), another study observed opposite changes (Gershon et al., 2014). In this regard, the allele effect may vary temporally or spatially (Harrison et al., 2022). Of particular relevance for this thesis, this variation could pertain to changes in allele effect during neurodevelopment. For example, it is possible that sequences in Intron 3, where the disease risk-associated variant rs1006737 is located, might be involved in the regulation of CACNA1C expression through the presence of different cis-regulatory elements that are active in different cell types at different developmental stages (Yoshimizu et al., 2015).

Alternatively, since the GWAS SNP rs1006737 is intronic, it is possible that this GWAS hit affects alternative splicing and the generation of different VGCC isoforms (Harrison et al., 2022). Although the possibility that GWAS risk alleles affect splicing has not yet been demonstrated and no study has linked the bipolar risk SNP to splicing; splicing irregularities have notably been associated with other psychosis-linked genes such as DRD2 and ZNF804A (Kleinman et al., 2011; Tao et al., 2014; Xiao et al., 2017; Gandal et al., 2018; Zhang et al., 2022). For example, SNPs at the DRD2 locus show genome-wide associations with schizophrenia. DRD2 is transcribed as two isoforms, D2-short receptor (D2SR) and D2-long receptor (D2LR). The schizophrenia genetic risk allele (rs1076560) likely regulates the relative levels of D2SR and D2LR, causing changes in D2 receptor-mediated signaling (Zhang et al., 2022). Another study that reported a novel, alternatively spliced and truncated ZNF804A transcript also identified an intronic risk variant (rs1344706) that

reduces the spliced ZNF804A transcript's expression in the fetal brain but has no link with full-length ZNF804A in the fetal or adult brain (Tao et al., 2014). Altogether, a better understanding of alternative splicing might identify the putative mechanisms by which a certain SNP causes an increase in disease risk in development.

Second, the phenotypic consequences of disease-associated genetic variation need to be determined (Harrison et al., 2022). Some evidence indicates that genotype can influence CACNA1C channel properties (Yoshimizu et al., 2015). For example, in a study of induced neurons from individuals harboring the bipolar risk-associated SNP rs1006737, L-type VGCC current density was increased in induced neurons homozygous for the risk genotype, compared to non-risk genotypes (Yoshimizu et al., 2015). It would be of great interest to determine whether there exist convergent downstream effects of different disease-associated variants (Harrison et al., 2022). Moreover, how VGCC disease variants affect calcium signaling remains to be clarified.

Finally, genomic studies might also provide insights into how to facilitate precision treatment by examining how expression of CACNA1C is affected in disease. A transcriptome wide association study (TWAS) found that CACNA1C was less active in schizophrenia, suggesting that repurposing a drug that increases its activity (like Ibutilide) might help treat the disorder (Reay et al, 2022).

1.5.2 Rare Variation in VGCCs and Brain Diseases

In addition to common variations conferring risk to psychiatric conditions, rare variants in VGCCs have also been linked to brain diseases. One example that illustrates this association is the connection between CACNA1C and Timothy Syndrome (TS). TS is an autosomal dominant syndrome characterized by neurological features akin to autism spectrum disorders. In 2004, a mutation in CACNA1C was identified as the molecular cause (Splawski et al., 2004). Since the discovery of CACNA1C being linked to TS, two types of TS have been discerned: TS Type I and TS Type II. Both TS Type I and TS Type II are related to mutations in the alternatively spliced Exons 8 and 8A of CACNA1C. The mutation identified in TS type I is located in Exon 8A of the CACNA1C gene, leading to a change of the amino acid Gly406 to Arg (G406R), which disrupts the 6th transmembrane helix in domain I of CACNA1C (Splawski et al., 2004; Bauer et al., 2021). A separate mutation at the same exact location (G406R) but in Exon 8 characterizes TS type II (Splawski et al, 2005). Both of these mutations result in the production of gain-

of-function channels characterized by delayed channel closing, resulting in an increase in cellular excitability. Additionally, an alternative G402S mutation has also been associated with TS Type II.

Alternative splicing of mutually expressed Exons 8 and 8A might explain differences in the clinical presentation of different types of TS. It has been suggested that the tissue-specific pattern of Exon 8/8A splicing leads to differences in symptoms between patients with TS Type I or TS Type II mutations (Splawski et al, 2005). Notably, a TS type II patient (G406R mutation in Exon 8) presented with more severe CNS defects compared to TS Type I patients (Splawski et al, 2005). The reported severity of CNS defects in a TS Type II patient might be related to the fact that Exon 8 represents approximately 80% of Cav1.2 mRNAs in the brain (Splawski et al, 2005). It should be noted that other TS Type II patients with the G402S mutation did not show neurological defects (Frohler et al, 2014; Hiippala et al, 2015), but this might be related to mosaicism or different properties of G406R versus G402S mutant channels (Dick et al, 2016).

Although VGCCs such as CACNA1C are typically thought to be associated with excitable cells, as noted above, altered intracellular calcium levels have been observed in other cell types, including platelets (Harrison et al., 2021). Indeed, the non-excitable cellular phenotypes observed in TS, including syndactyly (webbing of fingers and toes), support a role for CACNA1C function in non-excitable cells as well (Pitt et al., 2021).

A recent study has shown that the splicing of Exons 8 and 8A of CACNA1C is developmentally regulated in both mouse and humans, with Exon 8A being expressed lowly in earlier stages; whereas Exon 8 increases as development proceeds. The TS mutation in Exon 8A affects this regulation (Panagiotakos et al, 2019). The TS mutation is in the splice acceptor of Exon 8A, thus showing a tendency to be selected over the splice acceptor of Exon 8 (Panagiotakos et al, 2019). Indeed, Timothy Syndrome NPCs and neurons showed favored inclusion of mutant Exon 8A when compared to control hiPSCs. The upregulation of Exon 8 observed in mouse and human development was not present in Timothy Syndrome hiPSC-derived neurons (Panagiotakos et al, 2019). Additionally, a previous study showed that neural precursor cells and neurons from Timothy Syndrome iPSCs had differentiation abnormalities, including decreased expression of genes that are expressed in lower cortical layers and in callosal projection neurons (Pasca et al, 2011). Finally, another study reported that cortical interneurons in organoids from Timothy Syndrome hiPSCs showed an interneuron migration defect (Birey et al, 2017). More direct

causative evidence for how splicing can affect differentiation might be gleamed from future experiments correcting Timothy Syndrome mutations or finding alternative ways of manipulating the Exon 8 / Exon 8A ratio.

Rare mutations in auxiliary VGCC subunits are also associated with brain disease (Dahimene et al., 2022). Various studies have uncovered a role of CACNA2D proteins in neurological disease (reviewed in Ablinger, 2020). Cacna2d2 mutant mice show epilepsy and cerebellar ataxia (Barclay et al., 2001). In humans, patients harboring CACNA2D2 variants manifest different degrees of neurological disease, from congenital ataxia with cerebellar vermian atrophy to cerebellar atrophy with developmental and epileptic encephalopathies (Edvardson, et al., 2013; Pippucci et al., 2013; Valence et al., 2019). A recent study by Dolphin and colleagues showed that patients with CACNA2D1 mutations show developmental and epileptic encephalopathy (Dahimene et al., 2022). The study also revealed that these disease-causing mutations reduce the trafficking of CACNA2D1 to the cell membrane and inhibit the trafficking and function of the Cav2 channels (Dahimeme et al., 2022). Taken together, Dolphin and colleagues' findings shed some insight into CACNA2D1 function during human neural development (Dahimene et al., 2022).

In summary, rare variants in different VGCC subunits, including CACNA1C, CACNA2D1 and, CACNA2D2, have been linked to neuropsychiatric conditions such as TS, developmental and epileptic encephalopathy, and other neurological disease manifestations.

1.5.3 Summary

Altogether, a body of both genomic and experimental evidence supports the postulate that VGCC signaling is a critical pathway in different neurological and psychiatric diseases. However, the regulation (and specifically transcriptional regulation) of VGCCs during human neural development, a phase when the risk of neurological and psychiatric diseases is established, has not been thoroughly investigated.

1.6 Splicing

RNA splicing refers to a phenomenon where a newly synthesized precursor messenger RNA (pre-mRNA) is converted into messenger mRNA. During this process, introns (non-coding regions of RNA) are removed and exons (coding regions) are joined together. Splicing is mediated by the spliceosome, a complex of proteins and RNAs.

1.6.1 Introns

Introns refer to nucleotide sequences between two exons. They can either be the DNA sequence within a gene or an RNA sequence in pre-mRNA. Introns contain key features that are needed for splicing, including 1) the donor 5' splice site, 2) the acceptor 3' splice site, 3) the branch site and 4) the polypyrimidine tract (PPT).

Splice sites comprise short, well-conserved consensus sequences (Krebs et al., 2021). The highest conservation is observed within the intron at exon-intron junctions. The sequence of most introns is GU ..... AG (Krebs et al., 2021). The GU-AG rule refers to the presence of the dinucleotides GU and AG at the beginning and end of introns of pre-mRNAs, respectively. Additionally, a smaller set of introns can follow other rules (e.g. the AU-AC rule) with different consensus sequences at exon-intron boundaries.

Further upstream from the AG is a PPT (a sequence with a high concentration of pyrimidines C and U). Upstream from the PPT is the branch site, which contains an adenine nucleotide that mediates the formation of a branched intermediate (i.e. lariat) during the splicing reaction (Krebs et al., 2021).

1.6.2 Spliceosome

Splicing is mediated by the spliceosome, a complex of proteins and RNAs that mediates intron removal and exon joining (Wang and Burge, 2008). The spliceosome is a sophisticated complex comprising more than 100 core proteins and five small nuclear RNAs (snRNAs U1, U2, U4, U5, and U6) (Wang and Burge, 2008). Spliceosomes can be classified into two types: major and minor, which differ in their constituent proteins and RNAs and the target introns they splice. The major spliceosome removes introns that harbor consensus splice site sequences, or signals that indicate intron-exon boundaries. The minor spliceosome removes a different type of intron.

1.6.3 Alternative Splicing

As noted above, splicing can create different proteins by changing the exons present in the same pre-mRNA. This process is referred to as alternative splicing. According to this process, specific exons or exonic sequences can be included or excluded by using different splicing sites. There exist different types of alternative splicing, and these are indicated in **Figure 1.3** (Krebs et al., 2021).

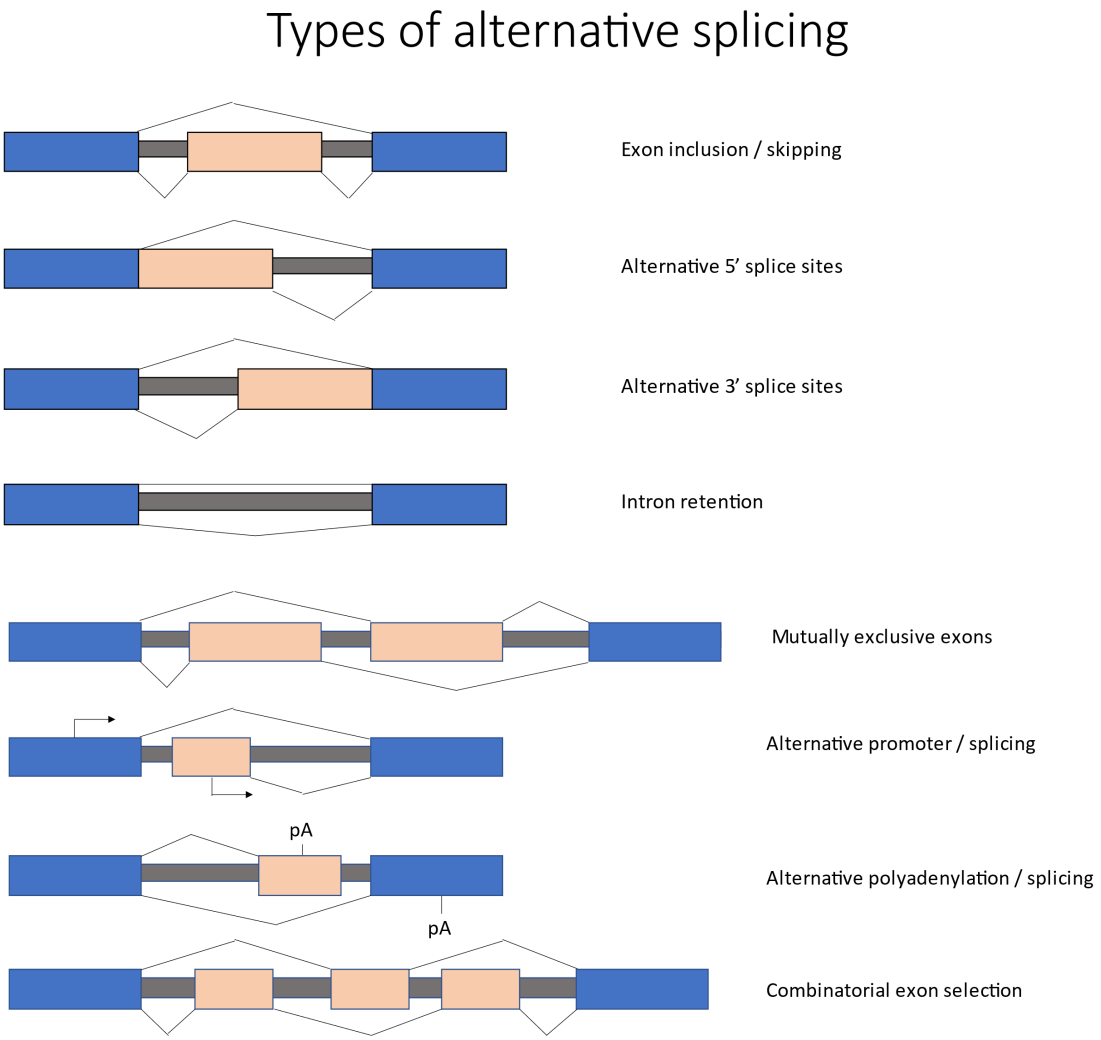


Figure 1.3: Different types of alternative splicing

The different types include 1) exon inclusion or skipping, which occurs when a particular exon is included or skipped during splicing, resulting in a protein that is including or missing a particular exon; 2) usage of an alternative 5' splice site, which occurs when different splice sites are used at the 5' end of the intron; 3) usage of an alternative 3' splice site, which occurs when different splice sites are used at the 3' end of the intron; 4) intron retention, which occurs when an intron is not spliced out of the mRNA; 5) mutually exclusive exons, which occurs when two or more exons are mutually exclusive, meaning that only one of the exons is included in the final mRNA; 6) alternative promoters, which occurs when different promoters are used to initiate transcription; and 7) alternative polyadenylation, which is when different polyadenylation sites are used (redrawn from Krebs et al., 2021).

Most human genes undergo alternative splicing. The consequences of alternative splicing can have significant effects, for example generating soluble versus membrane-bound forms of the Fas receptor with different effects on apoptosis; or generating different isoforms of the Drosophila fruitless protein which guides sexual orientation (Cascino et al., 1995; Demir and Dickson, 2005).

1.6.4 Splicing Regulation

As noted above, the key splicing signals include the 5'-splice- and 3'-splice site (at the 5 and 3'ends of the intron), the PPT upstream of the 3' splice site, and the branch site upstream of the PPT. Aside from these key sequences, regulatory intersections between 1) trans-acting RNA-binding proteins (RBPs) or splicing factors and 2) cis-acting splicing regulatory elements (SREs) on the pre-mRNA regulate alternative splicing.

1.6.4.1 Trans-Regulation of Splicing (RBPs)

RBPs are proteins that bind to RNA molecules and form ribonucleoprotein (RNP) complexes (Khanal, 2019). RBPs and RNP complexes regulate RNA metabolism and contribute to a plethora of cellular functions, including gene expression, post-transcriptional gene regulation, and regulation of translation and transcription control (Glisovic et al., 2008; Beckmann et al., 2015; Khanal, 2019). Additionally, several diseases have been associated with abnormal RBP expression and function (Castello et al., 2012; Khanal, 2019). These studies underscore the need to identify RBPs. While this thesis focuses on known RBPs, it should be noted that it is not possible to explain all of the splicing using just RBPs. For example, certain TFs have been identified to act as RBPs, as is the case with ZNF804A, one of the first risk genes linked to schizophrenia (Chapman et al., 2019).

The human genome is thought to possess a total of 424 known and predicted RBPs. RBPs bind short, single-stranded RNA sequences and also structured RNAs (Cook et al., 2011; Gabut et al., 2008; Auweter et al., 2006). RBPs possess RNA binding domains (RBDs), including RNA recognition motifs, zinc fingers, heterogeneous nuclear ribonucleoproteins (hnRNP) K homology domains, and another domain that regulates the splicing of its target. The biological role of the protein is specific to each domain. For example, the RBD is responsible for binding affinity, whereas other domains activate or repress splicing.

Many studies have been conducted, and different experimental and computational approaches have been invented to identify RBPs. Early attempts to study RBPs nearly fifty years ago involved the isolation of RNPs from UV-irradiated polysomal fractions, UV-irradiated cells, and untreated cells (Lindberg and Sundquist, 1974; Greenberg, 1979; Wagenmakers et al., 1980). More recently, numerous RBPs have been identified using HeLa cells, embryonic kidney cells, and mouse embryonic stem cells using various UV crosslinking methods (Baltz et al. 2012; Castello et al., 2012; Kwon et al., 2013; Mitchell et al., 2013). Nonetheless, experiments can be expensive and labor-intensive.

There is great interest in developing large-scale approaches to study RBPs. One such in vitro approach is RNAcompete (Ray et al., 2009). RNAcompete involves a competitive binding reaction in which an RBP is incubated with an excess of a complex pool of RNAs. The RBP is then retrieved and associated RNAs are investigated using computational analysis. By using custom-designed RNA pools containing all possible 9 base nucleotide sequences, one study identified the sequence preferences for 207 different RBPs (Ray et al., 2013). By using information about the sequence preferences for different RBPs and bioinformatics tools such as Find Individual Motif Occurrences (FIMO) and Hypergeometric Optimization of Motif EnRichment (HOMER) (Heinz et al., 2010; Grant et al., 2011), it is possible to infer the identity of putative regulators of a splicing event of interest by interrogating pre-mRNA sequences near the splicing event of interest.

1.6.4.2 Cis-Regulation of Alternative Splicing

More generally, splicing regulatory elements (SREs) influence the binding of splicing factors to pre-mRNA and act as activators or inhibitors of splicing by influencing spliceosome assembly and the usage of particular splice sites (**Figure 1.4**). The nucleotide sequences in the pre-mRNA that regulate splicing involve four types: exonic splicing enhancers (ESE), intronic splicing enhancers (ISE), exonic splicing silences (ESS), and intronic splicing silences (ISS) (Wang and Burge, 2008). The function of an SRE is also context-dependent. The location of a splicing regulatory element is relevant because it can affect the activities of SREs. For instance, G triplets repress splicing when located in an exon but may conversely promote splicing when located near an exon (McCullough and Berget 1997). Alternatively, the ability of splicing factors (such as SR proteins) to enhance splicing can be affected by the distance between the ESE and the splice site (Graveley et al., 1998). SR proteins represent a family of splicing factors that promote splicing. Another example of RBPs that regulate splicing is the class of hnRNPs, which serve as splicing repressors (Han et al, 2010). A diagram of how interactions between trans-acting factors and SREs regulate splicing is indicated in **Figure 1.4**.

1.6.4.3 Epigenetic and Transcriptional Regulation of Alternative Splicing

Polymerase speed and chromatin state are two parameters that can impact alternative splicing (Haque and Oberdoerffer, 2014). First, the speed of RNA polymerase II (Pol II) during transcription has been implicated in alternative splicing regulation. It has been

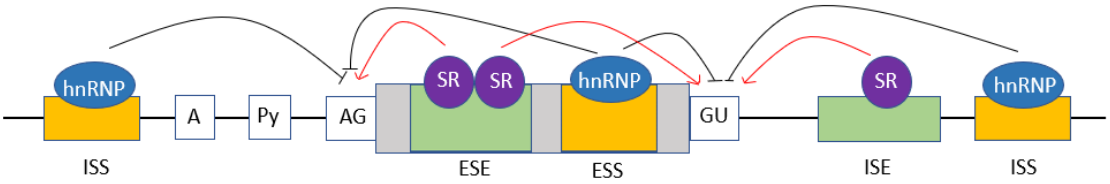


Figure 1.4: Alternative splicing regulation by cis mRNA elements and trans-acting factors.

Sequence elements that are involved mark the borders between exons and introns (such as 5' and 3' splice sites), GU-AG, the polypyrimidine (Py) tract, and the branch point sequence (BPS) (Sune-Pou et al., 2017). Enhancer and silencer regulatory elements (i.e. ISS: intronic splicing silencer; ESE: exonic splicing enhancer; ESS: exonic splicing silencer; ISE: intronic splicing enhancer), help regulate whether alternative splicing occurs at a particular location (Sune-Pou et al., 2017). Splicing factors, including SR proteins and hnRNPs, bind to cis-regulatory elements and direct the spliceosome (Smith et al., 2000; Singh et al., 2005; Izquierdo et al., 2006). (figure is redrawn from Sune-Pou et al. 2017).

demonstrated that the rate of transcription elongation impacts splicing decisions such that weak exons are more likely to be excluded due to a rapid elongation rate (de la Mata et al, 2003).

Chromatin state has been shown to influence splicing decisions. It has been suggested that this might occur by at least two mechanisms: first, by influencing Pol II elongation rate (kinetic mechanism) and second, by recruiting RBPs via interactions with chromatin-binding proteins (adaptor mechanism) (Haque and Oberdoerffer, 2014). Given that exons possess a higher rate of nucleosome occupancy, it was suggested that nucleosomes might promote exon definition by promoting pausing of Pol II and spliceosome assembly. Additionally, evidence for chromatin modifications serving as adaptors for RBP recruitment comes from studies identifying CHD1 as an H3K4me3 interacting protein and CHD1 was identified to interact with spliceosomal proteins U2 snRNP (Sims et al, 2007). Down-regulation of CHD1 or H3K4me3 reduced U2 association with chromatin and hindered pre-mRNA splicing efficiency (Sims et al, 2007).

1.7 VGCCs and Alternative Splicing

There are many different possible coding transcripts of VGCCs, as VGCC transcripts are long (mRNAs >14 kb) and include a large number of exons, allowing for an enormous amount of transcriptional diversity. Because VGCC SNPs identified in GWAS studies are non-coding and intronic, the possibility that such SNPs (or genomic loci linked to such SNPs) cause increased disease risk by regulating transcript diversity is tantalizing. Indeed, the production of different splice variants from VGCCs has been implicated in diseases

with neurodevelopmental phenotypes, which as described earlier, include TS (see section 1.5.2) (Splawski et al., 2004; Bauer et al., 2021). While remaining to be definitively proven, the presence of variants in non-coding regions of CACNA1C has led to speculation that alternative splicing might be affected (Harrison et al, 2022).

1.7.1 Effects of Alternative Splicing on VGCC Function

Alternative splicing can affect VGCC function and alter the sensitivity of VGCCs to drugs (Striessnig et al., 2014, Lipscombe and Andrade, 2015). Different alpha-1 subunit isoforms exhibit differences in inactivation kinetics, voltage-dependence and sensitivity to calcium channel blockers.

For example, Cacna1c possesses different alternatively spliced portions that can affect different sections of the subunit. Exons 8a and 8 are spliced mutually exclusively and encode different sections of the Cacna1c pore region, specifically the IS6 segment (Welling et al., 1997). Analyses of the Exon 8 and 8a Cacna1c variants indicate that, while they possess nearly identical electrophysiological properties, they possess different sensitivity to dihydropyridine (DHP) drugs (Welling et al., 1997; Zuhlke et al., 1998).

Alternative splicing of Exons 21/22 and 31/32 and constitutive splicing of Exons 33 and 45 have also been reported (Soldatov et al., 1992). Exons 21 and 22 exhibit mutually exclusive expression patterns (either Exon 21 or Exon 22 is used, but not both) and encode the IIIS2 transmembrane segment and part of the linker between IIIS1 and IIIS2. They are part of the voltage sensing domain, and although DHP drugs bind elsewhere, Exon 21/22 alternative splicing affects the voltage-dependent effect of the drug isradipine (Soldatov et al., 1995). In a non-neural system, switching usage of CACNA1C from Exon 21 to Exon 22 is a molecular hallmark of CACNA1C-mediated changes of the vascular smooth muscle cell to a pathological proliferating state in atherosclerosis (Soldatov, 2013). Additionally, the aberrant generation of a CACNA1C splice variant containing both Exon 21 and Exon 22 is non-functional and unable to conduct any calcium influx (Hu et al., 2016). Therefore, alternative splicing of Exons 21 and 22 can affect CACNA1C function (Soldatov et al., 1995; Tiwari et al., 2006).

A potential function for alternative splicing of Exons 31/32, which are also mutually exclusively utilized, has been suggested. Levels of Exon 32 utilization are significantly higher in failing human hearts, whereas Exon 31 utilization is higher in non-failing hearts (Yang et al., 2000). More studies of the roles and mechanisms by which each CACNA1C

variant plays in heart failure and especially the brain are needed (Yang et al., 2000).

More generally, alternative splicing of VGCCs such as CACNA1C is more complex than previously realized. A recent study investigated CACNA1C transcripts in different regions of the human brain (Clark et al., 2020). For the first time, long-read sequencing elucidated the complete exonic structure of CACNA1C, identifying 38 novel exons and 90 isoforms. Interestingly, this effort was only able to identify seven out of 40 previously identified isoforms, suggesting the possible presence of false positives in previous studies. Importantly, many of the 241 novel transcripts were predicted to encode channels with altered functions (Clark et al., 2020). There exist questions regarding the data generated by Clark et al. It is unclear which transcripts are translated and where they are subcellularly located (Harrison et al., 2022). Moreover, the functional and pharmacological properties of different isoforms need to be tested (Harrison et al., 2022).

The application of new technologies, such as long-read sequencing, to studies of alternative splicing will help to uncover the true nature of VGCC transcripts in different tissue types. Of particular relevance is the feasibility of understanding the transcriptional dynamics of VGCC splicing during human neural development.

Alternative splicing has also been found for the alpha-2-delta subunits as well (Dolphin, 2013). Studies that first identified splice variants in alpha-2-delta subunits found that the cDNA sequence of CACNA2D1 in rat brains was different from that in skeletal muscle (Kim et al., 1992). Alternative splicing of CACNA2D2, CACNA2D3, and CACNA2D4 has also been reported (Klugbauer et al., 1999; Barclay and Rees, 2000; Qin et al., 2002). Indeed, different isoforms of CACNA2D1 have different affinities for gabapentin, as has been demonstrated in of a study a CACNA2D1 splice variant with lower affinity for gabapentin after sensory nerve injury (Beatrice et al., 2014).

Splice variants have also been identified for VGCC beta subunits (Buraei and Yang, 2010). These splicing events have been identified in the N- and C-termini, as well as in the HOOK region. Different isoforms of beta subunits show different expressions in different tissues. Different beta subunit isoforms also show alternative functions. For example, a beta-4 splice variant acts as a transcription regulator, whereas another beta-4 subunit will not (Hibino et al., 2003).

1.7.2 Regulation of VGCC Alternative Splicing

A key question is how alternative splicing of VGCCs is regulated. VGCCs undergo extensive alternative splicing in different cell types, including in the setting of cellular differentiation. Two examples of this concern the developmental control of Cacna1c splicing by Fox proteins and PPT-binding protein (Tang et al., 2009; Tang et al., 2011).

In the former, Exon 9* and Exon 33 are regulated by Fox splicing regulators (Tang et al., 2009). These exons affect the electrophysiological properties and show opposing expression patterns during the development of the cortex. Therefore, Fox proteins can fine-tune the properties of Cacna1c during neural differentiation. Regarding the latter case, PTB regulates a flip from splicing of Exon 8 to Exon 8a (Tang et al., 2011). During neural differentiation, PTB is downregulated, and this change is accompanied by an elevation in levels of the PTB neuronal homolog, nPTB (Tang et al., 2011). This results in Exon 8a being at low levels in the developing mouse brain, but its expression is activated during neuronal differentiation as PTB levels are depleted.

These two examples suggest that the relative levels of different cell type-specific splicing regulators can influence the splicing, and therefore the properties, of VGCCs during cellular differentiation.

1.8 Transcriptional Regulation

Transcription is the initial step of gene expression. It is the process of synthesizing an RNA molecule from a DNA template. Like splicing, its regulation involves both cis-regulatory elements and trans-acting factors. Such cis-regulatory elements include core promoter and promoter-proximal elements placed near the transcription start sites (TSS) of genes, as well as more distally located enhancers. Regarding trans-acting factors, TFs possess a leading role in initiating gene transcription, which will be elaborated upon below. A brief overview of transcriptional regulation is featured in **Figure 1.5**.

1.8.1 Transcription Factors

TFs are proteins that control gene expression in different cell types. There exist about 1,640 human TFs (Lambert et al., 2018). For cell state transitions, TFs can specify cell identity through differentiation and drive reprogramming (i.e. de-differentiation) and transdifferentiation (Takahashi and Yamanaka, 2006). Mutations in TFs and TF-binding

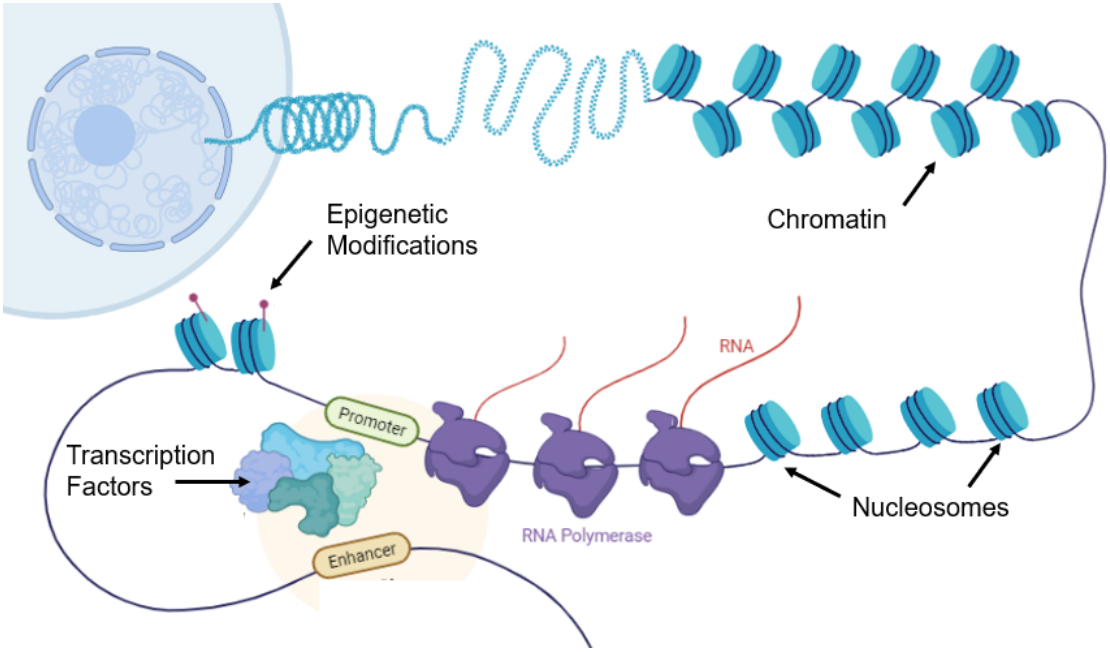


Figure 1.5: Overview of transcriptional regulation

DNA is tightly packaged in chromatin, a blend of DNA and proteins that form the chromosomes. Chromatin is comprised of nucleosomes, which are sections of DNA that are encased around histone proteins. Histones can possess epigenetic modifications that are associated with increased or reduced transcription. A combination of cis-regulatory sequences (long-distance enhancers and proximal promoters) and trans-acting regulators (TFs) regulate the initiation of transcription. Transcription occurs when DNA is transcribed into RNA by the protein RNA polymerase II. (taken from biorender.com).

sites are linked to human diseases. The Human Phenotype Ontology reports that a large number of mutations at TF loci are associated with human diseases (Kohler et al., 2014).

For decades, the term "TF" referred to any protein involved in the regulation of gene expression levels (Lambert et al., 2018). Currently, however, the term TF refers to proteins that can bind DNA with sequence-specificity and modulate transcription. A key to understanding TF biology is to identify the DNA sequences to which they bind. TF binding motifs are short stretches of DNA highly associated with certain TFs. Identifying the DNA-binding motif allows for the identification of putative TF binding sites and serves as a prelude to deeper characterizations.

TFs can regulate transcription in a number of ways following binding to DNA (Lambert et al., 2018). Most eukaryotic TFs recruit cofactors such as coactivators or corepressors, which are usually multi-subunit protein complexes. These complexes bind chromatin, induce nucleosome remodeling, or covalently modify histones or other proteins.

1.8.2 Transcription Factor-Mediated Regulation of VGCC Abundance

While much attention has focused on the post-transcriptional regulation of VGCCs (for instance, by alternative splicing), little is known about the transcriptional regulation of

VGCCs (Gonzalez-Ramirez and Felix, 2018). In particular, the identity of TFs that regulate their expression in different cell types is not well-understood.

A study reported that Egr1 can regulate the expression of Cacna2d2 (Gonzalez-Ramirez et al, 2018). Overexpression and knockdown showed changes in the transcript levels of Cacna2d2. Moreover Egr1 knockdown reduced calcium currents in DRG neurons, whereas overexpression of Egr1 increased calcium currents in a neuroblastoma cell line. Therefore, TF levels can regulate both VGCC transcript levels and calcium signaling.

Another study reported that Sp1 can regulate Cacna1g (Gonzalez-Ramirez et al, 2014). Overexpression of Sp1 enhanced Cacna1g promoter activity and knockdown of Sp1 decreased Cav3.1 protein and the amplitude of whole-cell T-type Ca2+ currents in a neuroblastoma cell line. These data also corroborate the notion that changes in TF abundance not only alter VGCC transcript and protein levels but can also affect calcium signaling.

Initial investigations of VGCC gene promoters have revealed that all of the VGCC genes start transcription from a TATA-less promoter (Gonzalez-Ramirez and Felix, 2018). CpG islands (GC-rich areas) are commonly observed and Sp1 elements appear to influence gene expression. It is tempting to speculate that other TFs with zinc-finger motifs similar to those of Sp1, such as Sp1-like proteins such as KLFs, might also regulate VGCC expression as well (Kaczynski et al, 2003; Lomberk and Urrutia, 2005). Additionally, some promoters possess silencer elements, which cause VGCCs to be expressed specifically in neurons.

Altogether, our understanding of the transcriptional regulation of VGCCs as well as the identity of TFs involved is limited especially in the context of human neural development, which motivates the studies carried out in this thesis.

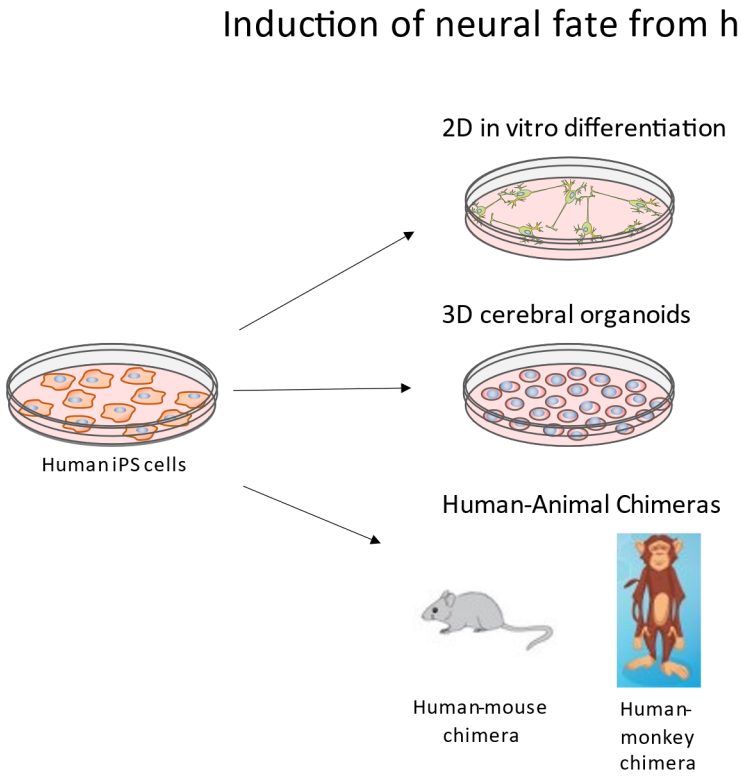
1.9 Human iPSCs

Probing VGCC function for neuropsychiatric research will require access to relevant model systems. A number of different experimental systems have been used to probe VGCC function, including mice, rats, human post-mortem brain tissue, and human cell lines. Studies of neural circuits in mice in vivo, in brain slices, or in isolated cultured neurons have been foundational for neuroscience for decades. While mice have served as a useful model to dissect VGCC function in a mammalian system, mice and humans exhibit dramatic differences in brain structure and function that raise issues regarding whether observations made in mice can be extrapolated to humans. Notably, transcriptional analyses of the

mouse and human cerebral cortex have identified differences regarding isoform diversity and alternative splicing (Leung et al., 2021), underscoring the potential importance of conducting VGCC studies using human-relevant systems. To address this limitation, studies of VGCC function and regulation in a human-specific context can be conducted using human post-mortem brain tissue and human cell lines. While studies of human post-mortem brain tissue can provide some insights into VGCC expression, it would be useful to be able to dissect VGCC function experimentally in live cells. Alternatively, the physiological relevance of observations made in human cancer cell lines, which are transformed and genetically abnormal, remains unclear.

Human pluripotent stem cells (PSCs) represent a promising model system for the dissection of VGCC function using a human-relevant system. The striking discovery that human somatic cells can be reprogrammed to pluripotency by the transient expression of the reprogramming factors OCT4, SOX2, KLF4, and c MYC has revolutionized studies of human gene function and human development (Takahashi et al., 2007). Human induced pluripotent stem cells (hiPSCs) can self-renew indefinitely and form all cell types. hiPSCs comprise a limitless source of any human cell type, including neural cells, and from any genetic background (**Figure 1.6**). By avoiding the use of embryos, human iPSC generation has overcome the ethical limitations associated with human embryonic stem cells. Moreover, human iPSCs are euploid and while still exhibiting some genetic instability, show decreased genetic instability when compared to other transformed human cancer cell lines historically used to carry out some genetic studies in human cells, permitting studies in a more physiologically relevant context.

Moreover, as noted, induced pluripotent stem cell (iPSC) lines can be generated from any genetic background, including patients harboring disease-associated mutations, enabling practitioners to potentially model diseases or disease risk in a dish (Soldner and Jaenisch, 2018). The fact that human iPSCs can be derived from an unlimited number of genetic backgrounds means that, for the first time, analyses involving high numbers of cell lines from different genetic backgrounds can be carried out.



Pros	Cons
• Standard method	• Maturation only to fetal stages
• Reduced variability	• Inability to study cell -cell interactions

Pros	Cons
• Recapitulates many human brain features	• Maturation only to fetal stages
• Studies of cell -cell interactions	• Variability

Pros	Cons
• Study human cells in in vivo environment	• Very low contributions of human cells to animal hosts (lack of success to date)
• Possible maturation to adult stage	• Ethical concerns

Figure 1.6: Neural differentiation of hiPSCs.

(left) Diagram showing that hiPSCs can be differentiated into neural fate by three different methods: 2D in vitro differentiation, 3D culture systems including the production of cerebral organoids, and via human-animal chimeras. In the human-animal chimera approach, hiPSCs are introduced into animal preimplantation embryos and the resulting chimeric embryos are transferred into foster mothers. (right) The pros and cons of each approach are shown. Note: the methods are not mutually exclusive (e.g. organoids can be transplanted into mice to make human-animal interspecies chimeras as well).

1.10 Neural Cells and Tissues from hiPSCs

The induction of neural fate from hiPSCs uses cocktails of exogenous factors (Chambers et al., 2009). hiPSC-neurons are electrophysiologically active (Ebert et al., 2009; Lafaille et al., 2012). The availability of living human neural cells from any patient to conduct functional studies has the potential to influence psychiatry research. **Figure 1.6** provides a brief overview of different strategies to induce neural fate.

Historically, the production of neural cells from hiPSCs has occurred in two-dimensional monolayer culture systems. In recent years, studies have demonstrated the feasibility of generating human neural cells using alternative methods. Cerebral organoids are 3D multicellular structures produced from human PSCs that can self-organize into aggregates that recapitulate features of different brain regions. Single-cell RNA-seq, chromatin analyses (assay of transposase accessible chromatin (ATAC-seq)), and electrophysiological measurements in organoids can be applied to assess transcriptional differences and neuronal activity in a model of human neural development (Xiang et al., 2017; Birey et al., 2022). Of particular relevance for this thesis, organoid models have revealed the complex role of VGCC function in human cortical interneuron migration and have unveiled possible ways to treat disease manifestations associated with TS (Birey et al., 2022).

Additionally, the transplantation and development of human neural cells into mouse models allow for the study of human neural cells in an in vivo context. In contrast to 2D systems, the development of cerebral organoids and interspecies chimeras aims to improve the fidelity of differentiation (Lancaster et al., 2013, Wu et al., 2017).

To better understand the roles and regulation of VGCCs during human neural development, it would also be desirable to use an experimental system that recapitulates key features of human neural development. Additionally, for an experimental dissection of VGCC alternative splicing using hiPSCs, it would be desirable to conduct highly powered studies using high numbers of isogenic hiPSC lines and neurons. While 3D organoid systems can promisingly model different characteristics of human brain development in vitro, cerebral organoid methods are currently highly variable. Alternatively, the efficiency of interspecies chimera formation is too low for high-powered analyses. Given these limitations, it has been surmised that 2D cultures comprise the most suitable system for dissecting VGCC function owing to scalability.

Currently available differentiation protocols tend to produce relatively immature neural

cells that share features with fetal, but not adult development. While this suggests caution is warranted in interpreting hiPSC findings, this characteristic highlights the unique opportunities to study previously inaccessible stages of human brain development. It should be noted that, aside from producing immature cells, more generally, hiPSCs do possess limitations (see **Chapter 5**). Therefore, alternative model systems remain crucial for VGCC studies. Nonetheless, this thesis investigates the transcriptional regulation of VGCCs in a cell state transition of interest to neuroscientists, the differentiation of hiPSCs into neurons, a putative model of human neural development.

1.11 Aims

Transcriptional regulation of VGCCs plays an important role in channel function. For example, alternative splicing of VGCCs has been shown to modulate their properties, including their electrophysiological properties and responsiveness to DHPs (Soldatov et al., 1995), and is involved with different pathophysiological conditions, including TS.

While it has historically been challenging to identify an experimental system that allows access to human brain development, hiPSCs will enable the investigation of gene function in a culture system that serves as a model for human neural development. For example, it is possible to conduct loss-of-function or gain-of-function studies with different VGCC subunits in hiPSC and their derivative neural cells to analyze gene function.

To investigate the regulation of VGCCs during human neural development, I have analyzed data from the Human Induced Pluripotent Stem Cell Initiative (HIPSCI), the largest collection of hiPSC lines in the world (Streeter et al., 2017). The HIPSCI has differentiated a large number of hiPSCs into neurons, offering opportunities to study the transcriptional regulations of VGCCs in different genetic backgrounds (Kilpinen et al., 2017; Streeter et al., 2017; Schwartzentruber et al., 2018). The opportunity to study data from multiple donors and isogenic pairs of hiPSCs and neurons allows one to address the confounding variable of genetic background (Kilpinen et al., 2017). For RNA-seq analyses, I employed 40 isogenic pairs of hiPSCs and hiPSC-neuron datasets. For analyses of chromatin accessibility data, I will study data from 23 individuals.

The objective of this thesis can be distilled into three aims:

1. In **Chapter 2**, I analyzed changes in the abundance and splicing of VGCCs during the differentiation of hiPSCs into neurons.

2. In **Chapter 3**, I identified candidate RBPs that may regulate the alternative splicing of different VGCC regions of interest using publicly available bioinformatics tools. Specifically, I focus on the pre-mRNA sequences in close proximity to the splicing events identified in **Chapter 2**.

3. In **Chapter 4**, I identified candidate TFs that may regulate the changes in VGCC abundance characterized in **Chapter 2**. This was achieved by studying chromatin accessibility data of hiPSC-derived neurons from the HIPSCI. I identified open chromatin regions at the TSS of VGCC loci, followed by TF footprinting analyses to identify proteins potentially occupying these open chromatin regions in hiPSC-derived neurons. I then determined which of the identified TFs show differential expression between hiPSCs and neurons. The list of candidate TFs was further refined by identifying TFs that bind to several VGCC loci.

4. **Chapter 5** will discuss the findings made in **Chapters 2 − 4** and place them within the larger context of the fields of psychiatry, stem cells, and transcriptional regulation research.

Altogether, the insights gained from studying the transcriptional regulation of VGCCs during neural differentiation of hiPSCs serve as a foundation for future efforts to investigate the regulation of VGCCs during human neural development.

Chapter 2

Transcriptional Changes of VGCCs Across Neural Differentiation of hiPSCs

2.1 Introduction

VGCCs undergo complex transcriptional regulation across different states and tissues that can affect their function. Changes in VGCC abundance can affect their function through different mechanisms. First, at a global level, changes in transcript levels can lead to changes in the abundance of channels on the cell surface that can affect calcium ion influx and downstream signaling (Catterall, 2011). Second, at a more local level, subcellular localization and distribution can affect synaptic plasticity and modulate learning and memory (Hoppa et al., 2012; Jung et al., 2017). Alternative splicing, a process by which different mRNAs are generated from a single pre-mRNA transcript, can affect VGCC function as revealed by altering the sensitivity of VGCCs to drugs including CCBs (Striessnig et al., 2014, Lipscombe and Andrade, 2015). Little is known about the nature and regulation of VGCC abundance and splicing across human neural development, which would have implications for channel function and disease pathology.

The development of next-generation sequencing methods has revolutionized the study of gene expression. Next-generation sequencing of RNA involves the preparation of a cDNA library to produce short DNA molecules ("reads") with customized adapters attached to the end(s) of each molecule. Sequencing can occur either at one end of each molecule, known as single-read sequencing, or at both ends (the 5'- and 3'-ends) of each molecule, known as paired-end sequencing (Bennett, 2004; Slatko et al., 2018). Sequenced reads are mapped onto a reference genome or transcriptome (Norton, 2019). When multiple reads are derived from the same genomic region, a representation of the genomic region

can be inferred from the number of reads mapping to it (Norton, 2019). This produces a 'count' that enables quantification of the expression of genomic loci (**Figure 2.1**).

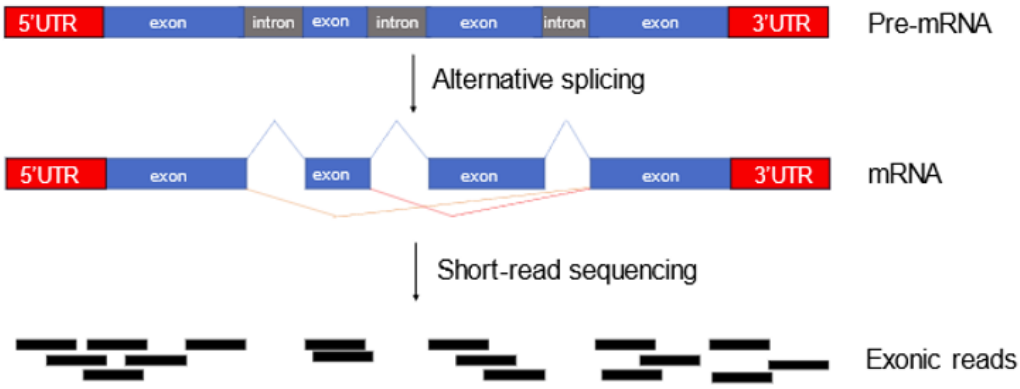


Figure 2.1: Analyzing short-read RNA-seq data.

. During the preparation of libraries for next-generation sequencing, mRNA is reverse transcribed into cDNA and the resulting cDNA is fragmented and sequenced, generating "reads". During bioinformatic analyses, the sequenced reads are mapped onto a reference genome or transcriptome using computational tools such as HISAT2 (Kim et al., 2015) or Spliced Transcripts Alignment to a Reference (STAR) (Dobin et al., 2013). For comparing mRNA abundance between hiPSCs and hiPSC-derived neurons, all exonic reads will be quantified using DESEQ2.

To our knowledge, transcriptomic analyses of the abundance and alternative splicing of the VGCC subunit family across cellular differentiation have not been previously reported. Carrying out such analyses would prove useful. Accumulating evidence indicates that the alternative splicing of VGCCs plays an important role in neural development by generating VGCC isoforms with different features (Tang et al., 2009; Panagiotakos et al., 2019; Panagiotakos and Pasca, 2021). For example, CACNA1C undergoes a switch in exon usage from Exon 8A to Exon 8 as neural progenitors differentiate into neurons (Panagiotakos and Pasca, 2021). The TS mutation in CACNA1C inhibits the switch from Exon 8A to Exon 8, resulting in a disruption of neural differentiation (Panagiotakos et al., 2019). These recent observations highlight the possible importance of the developmentally distinct roles of VGCCs.

Probing VGCC transcriptional regulation and function for neuropsychiatric research will require access to relevant model systems. hiPSCs represent a potential human-relevant model system (Takahashi et al., 2007). To explore VGCC biology using hiPSCs, it would be desirable to carry out studies that control for genetic background and laboratory batch effects (Kilpinen et al., 2017). In this regard, the datasets generated by HIPSCI – the largest collection of hiPSC lines in the world – are of particular interest (Streeter et al., 2017). The HIPSCI has differentiated a large number of hiPSCs into neurons in a 2D

system, offering unprecedented opportunities to carry out studies of cellular differentiation that are controlled for genetic background and laboratory batch effects (Kilpinen et al., 2017; Streeter et al., 2017; Schwartzentruber et al., 2018). Moreover, several bioinformatics packages have been developed for characterizing alternative splicing in short-read RNA-seq data, enabling analyses of alternative splicing in the HIPSCI datasets (Anders et al., 2012; Vaquero-Garcia et al., 2016; Mehmood et al., 2020). In particular, packages such as DEXSEQ and Modeling Alternative Junction Inclusion Quantification (MAJIQ) allow for the quantification of differential exon usage (DEU) and splice junction utilization, respectively.

Given the reasons described above, a systematic bioinformatics analysis of the abundance and splicing of the different VGCC family members across the differentiation of hiPSCs into neurons was undertaken to characterize changes in abundance and splicing across a cell state transition.

2.2 Methods

The aim of this Chapter is to characterize the changes in VGCC abundance and splicing that occur across cellular differentiation of hiPSCs to neurons.See **Figure 2.2** for the strategy to identify changes in abundance and splicing of VGCCs across hiPSC differentiation.

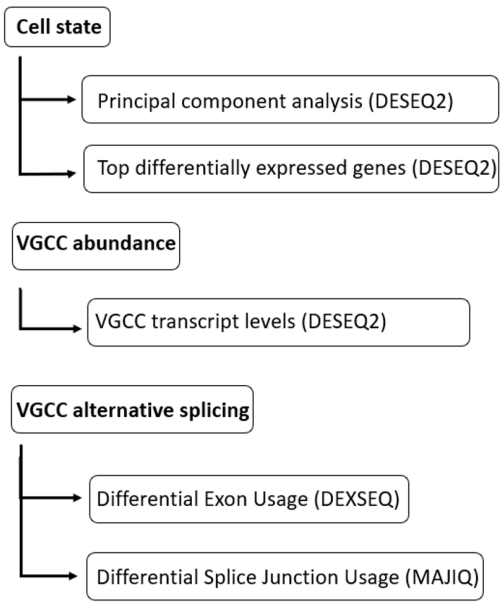


Figure 2.2: Overall strategy for identifying changes in VGCC abundance and alternative splicing across hiPSC differentiation.

First, a brief survey of cell state was carried out by applying DESEQ2 for principal component analysis and differential gene expression identification. Second, VGCC transcript abundance was characterized using DESEQ2. Third, VGCC alternative splicing events were identified using DEXSEQ (for differential exon usage) and MAJIQ (for differential splice junction usage).

To achieve this, short-read RNA-seq data derived from 40 isogenic pairs of hiPSCs and hiPSC-derived neurons were analyzed. For a survey of cell state and analysis of VGCC abundance, a computer program called DESEQ2 was used (Love et al., 2014). Analyzing cell state involved principal component analyses (PCA) and identification of top differentially expressed genes. Identifying changes in VGCC abundance included measurements of VGCC transcript levels using DESEQ2. For analyses of alternative splicing events, two types of alternative splicing events were studied: differential exon usage (DEU) and differential splice junction usage. DEXSEQ was used to identify DEU and MAJIQ was deployed to study differential splice junction usage (Anders et al., 2012; Vaquero-Garcia et al., 2016). DEXSEQ and MAJIQ have performed well in analyses of short-read RNA-seq data (Mehmood et al., 2020). As outlined further below, DEXSEQ and MAJIQ provide complementary information to allow for a more complete picture of the changes in alternative splicing occurring during cellular differentiation. The provenance of the cells from which the datasets are derived, along with the computational analyses that have been conducted will be described below.

2.2.1 Datasets

To analyze VGCC expression levels and splicing, a publicly available dataset from the Sanger Institute's HIPSCI Consortium was accessed (Streeter et al., 2017; Kilpinen et al., 2017; Schwartzentruber et al., 2018). Short-read RNA-seq data were analyzed from 40 isogenic pairs of hiPSCs and hiPSC-derived neurons.

The hiPSCs reported in the neuronal differentiation study were generated from unrelated, healthy individuals (Kilpinen et al., 2017). Metadata regarding the sex and age of donors is provided in **Table 2.1.**

Although not directly relevant to the bioinformatics methods used in this Chapter, it is worth commenting on the cell populations being studied. The hiPSC-derived neurons studied here possess a sensory neuronal (nociceptor) phenotype (Schwartzentruber et al., 2018). The sampling of hiPSC-derived neurons for RNA-seq analyses occurred on day 28 of differentiation from hiPSCs (Schwartzentruber et al., 2018). The neurons were generated by treating hiPSCs using a small molecule protocol involving chemicals that target different signaling pathways (Young et al., 2014). Key sensory neuronal marker genes are upregulated in hiPSC-derived neurons, and their capacity to respond to sodium ion channel modulators has been confirmed (Schwartzentruber et al., 2018). For example, the expression of key neuron-specific markers – such as ISL1, DRGX, NGFR, ion channel SCN9A, and peptide precursor TAC1 – has been reported in two studies using this small molecule protocol (Young et al., 2014; Schwartzentruber et al., 2018). Additionally, hiPSC-derived neurons express 204 of the 387 ion channels present in the human genome and 141 out of the 168 ion channel genes present in human dorsal root ganglia samples (Young et al., 2014). As the presence of mRNA expression might not sufficiently indicate channel function per se (as, for example, ion channels need to be appropriately translated and targeted to the correct cellular location), the presence and function of different ion channels, including GABA(A)R, HCN, and KCNQ ion channels have also been confirmed using electrophysiological assays (Young et al., 2014). Altogether, these data provide evidence that hiPSC-derived neurons possess features observed in dorsal root ganglia neurons (Young et al., 2014). Although the developmental maturity of HIPSCI neurons was not characterized, similar differentiations of hiPSCs have produced cells with the highest transcriptional similarity to first-trimester fetal brain tissue (Brennand et al., 2015).

Donor id	Sex	Age
HPSI0214i-wibj_2	Female	55-59
HPSI0114i-fikt_3	Male	50-54
HPSI0314i-qonc_2	Female	65-69
HPSI0314i-qaqx_1	Female	60-64
HPSI0114i-eipl_1	Female	40-44
HPSI1013i-wuye_2	Female	30-34
HPSI0314i-cuhk_2	Female	45-49
HPSI0913i-eika_2	Male	45-49
HPSI1013i-yemz_1	Male	70-74
HPSI0214i-kucg_2	Male	65-69
HPSI1213i-hehd_2	Female	60-64
HPSI0214i-kehc_2	Female	65-69
HPSI0913i-lise_1	Female	45-49
HPSI0114i-oevr_3	Male	70-74
HPSI1113i-uofv_1	Male	60-64
HPSI1113i-dons_1	Female	55-59
HPSI0214i-feec_3	Male	60-64
HPSI0314i-bipt_1	Male	70-74
HPSI0214i-pelm_3	Female	40-44
HPSI1113i-hayt_1	Male	70-74
HPSI0314i-fafq_1	Male	60-64
HPSI1213i-babk_2	Female	50-54
HPSI1013i-cups_3	Female	45-49
HPSI1213i-nekd_1	Male	60-64
HPSI0314i-xugn_1	Male	65-69
HPSI0114i-rozh_5	Female	65-69
HPSI1013i-sebz_1	Female	55-59
HPSI0114i-vabj_3	Female	50-54
HPSI1013i-kuxp_1	Male	25-29
HPSI0114i-vass_1	Female	30-34
HPSI1013i-hiaf_2	Male	65-69
HPSI0114i-joxm_1	Female	45-49
HPSI1213i-pahc_4	Female	55-59
HPSI0114i-kolf_3	Male	55-59
HPSI1013i-pamv_3	Male	65-69
HPSI1113i-podx_1	Female	65-69
HPSI0913i-oapg_5	Female	40-44
HPSI0114i-bezi_1	Female	60-64
HPSI1113i-qolg_1	Male	35-39
HPSI0214i-eiwy_1	Female	65-69

Table 2.1: Demographic information of hiPSC donors.

The sex and age of donors from which hiPSCs were generated are indicated.

2.2.2 Programming Code

All programming code for this Chapter was deposited in Github at `github.com/adelosan521/chapter2`. Raw sequencing data were obtained from the European Nucleotide Archive (Project: PRJEB7388; PRJEB18630). For differential gene expression analyses using DESEQ2, the code in the "DESEQ2_initial_processing.sh" script was used for pre-processing of raw sequencing data and in the "DESEQ_analyses.R" script was used to carry out DESEQ2 analyses ("DESEQ2_initial_processing.sh" script was generated by

thesis author; "DESEQ2_analyses.R" script was generated with assistance from Wilfried Haerty). For DEU analyses using DEXSEQ, the code in the "DEXSEQ_initial_ processing" script was used for pre-processing of raw sequencing data and in the "DEXSEQ_ analyses" script was used to carry out DEXSEQ analyses ("DEXSEQ_initial_processing.sh" was generated by thesis author; "DEXSEQ_analyses.R" script was largely generated by Wilfried Haerty). For the identification of local splice variations (LSVs) using MAJIQ, bam files were generated and sorted. The script called "majiq.sh" was used for MAJIQ (in-house implementation of published MAJIQ pipeline). A summary of the scripts and how they relate to one another is provided in the "README.md".

2.2.3 Gene Expression Analyses

To investigate changes in VGCC abundance that occur in hiPSCs and in hiPSC-derived neurons, a computer program called DESEQ2 (version 1.39.0) within R/Bioconductor was applied to the short-read sequence RNA-seq datasets obtained from the HIPSCI (Love et al., 2014).

Figure 2.3 shows a scheme for the pipeline used for DESEQ analyses. Raw sequencing data were downloaded from the European Nucleotide Archive (ebi.ac.uk/ena/browser/home). As the preparation of next-generation sequencing libraries involves adapter ligation, adapters need to be computationally trimmed from the sequenced reads. The quality of next-generation sequencing libraries generated from the HIPSCI study was assessed by FASTQC (version 0.10.1) (Andrews). If per base sequencing quality, per sequence quality scores, per sequence GC content, per base N content, and overrepresented sequences were not indicated as problematic by FASTQC, the dataset was included in subsequent analyses. Adapter sequences were computationally trimmed from sequencing data using Trim_Galore (version 0.3.1). Trimmed adapter files were aligned to a reference genome (GRCh38) using STAR (RNA-STAR version 2.6.1d) (Dobin et al., 2013). HTSEQ was used to generate count files. A merged count file was used as input for DESEQ2. Additionally, the genes that showed the greatest differential expression between hiPSCs and hiPSC-derived neurons were identified using DESEQ2.

2.2.4 Alternative Splicing Analyses

Alternative splicing of different VGCC subunits was investigated by analyzing DEU and local splicing variation (differential splice junction utilization) **(Figure 2.4)**. Two pro-

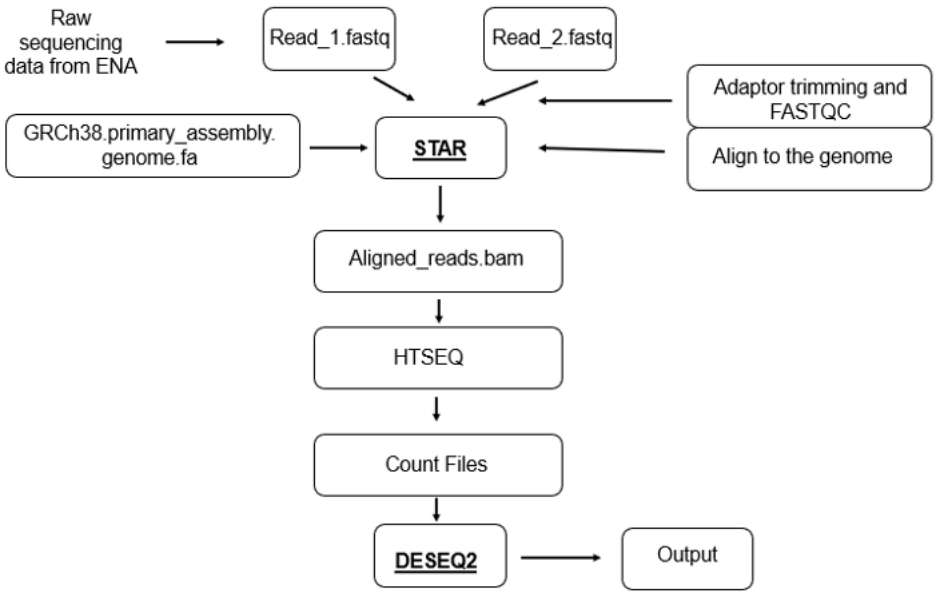


Figure 2.3: Overview of the DESEQ2 pipeline.

Scheme of the analyses pipeline for differential gene expression analysis using DESEQ2. A description of the pipeline is provided in the main text.

grams within R/Bioconductor, DEXSEQ and MAJIQ were used. DEXSEQ measures DEU in RNA-seq data (Anders et al., 2012). LSVs differentially utilized between the two cell types were identified using MAJIQ (Vaquero-Garcia et al., 2016).

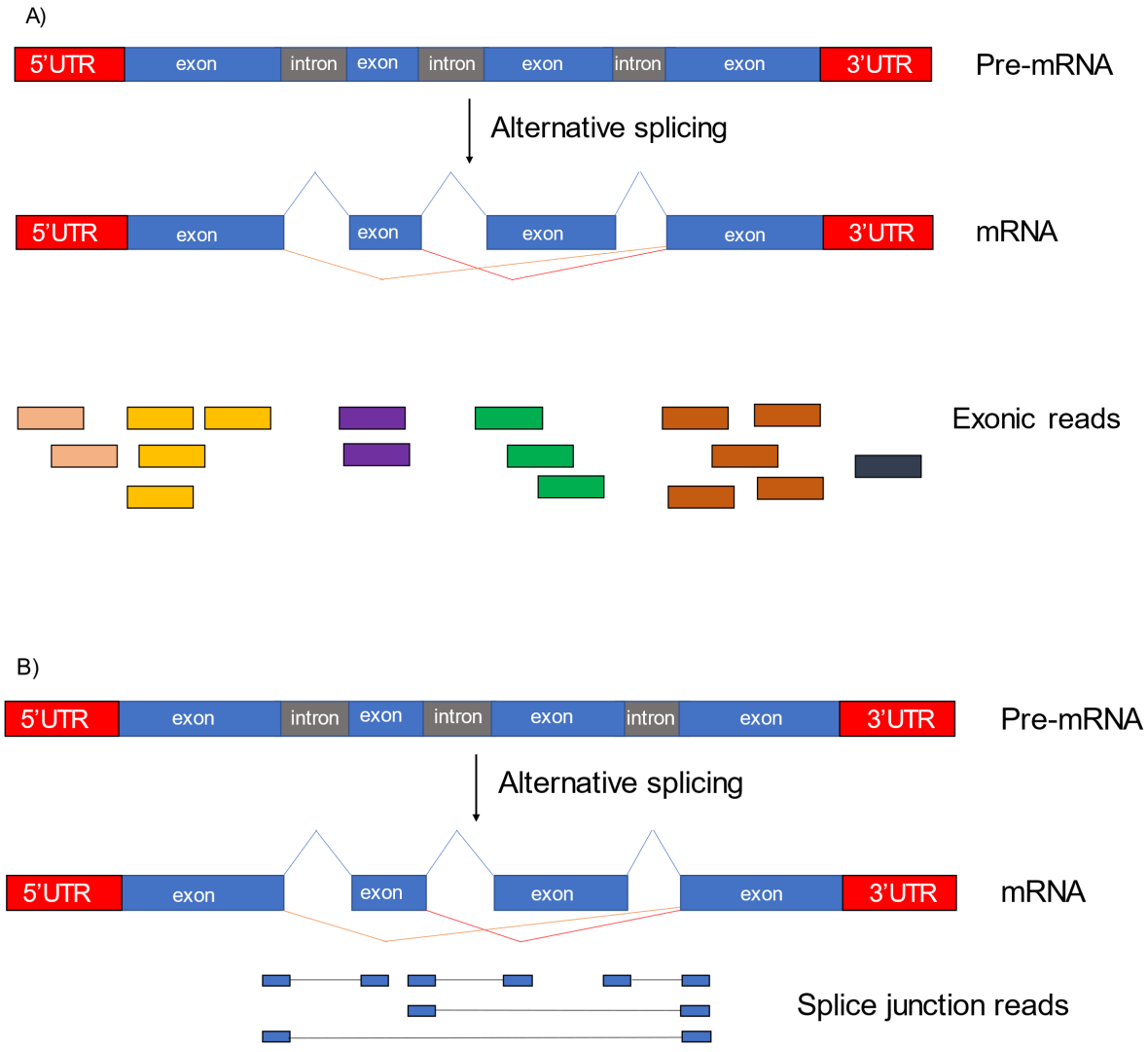


Figure 2.4: Identifying alternative splicing events in short-read sequencing data.

To compare alternative splicing, exonic reads were quantified for DEU using DEXSEQ (A), and splice junction reads (reads that span a splice junction) were quantified using MAJIQ (B). Relative usage of an exon is defined as (the number of transcripts from the gene that contain this exon / number of all transcripts from the gene). Splice junction reads comprise the minority of reads, which is likely to have significant implications for the relative statistical power of the two approaches. This predicts that it would be more challenging to detect differential splice junction utilization relative to DEU. Note: although the schematic presents exons as part of the coding sequence, exons are also present in the 5' and 3' UTRs as well.

2.2.4.1 DEXSEQ

Analyses of DEU involved the use of DEXSEQ (version 1.46.0) (Anders et al., 2012). DEU refers to alterations in the relative usage of different exons that manifest when comparing two different conditions (in the present study hiPSCs and hiPSC-derived neurons) (Reyes et al., 2021). The relative usage of an exon is the number of transcripts in a certain gene that contains a particular exon divided by the number of all transcripts from a certain gene (Reyes et al., 2021). In other words, DEXSEQ determines the number of reads that

map to a certain exon and the number of reads that map to other exons of a specific gene (Reyes et al., 2021). DEXSEQ then determines how this ratio changes across different experimental conditions (Reyes et al., 2021). DEXSEQ attempts to identify alterations in relative exon usage that are not just the effects of upregulation or downregulation of a gene (Reyes et al., 2021). In addition to alternative splicing of the open reading frame, DEU can also capture changes in the utilization of alternative transcript start sites and polyadenylation sites (i.e. differential usage of exons at the 5'- and 3'- ends of transcripts) (Reyes et al., 2021).

Figure 2.5 shows the DEXSEQ analysis pipeline.

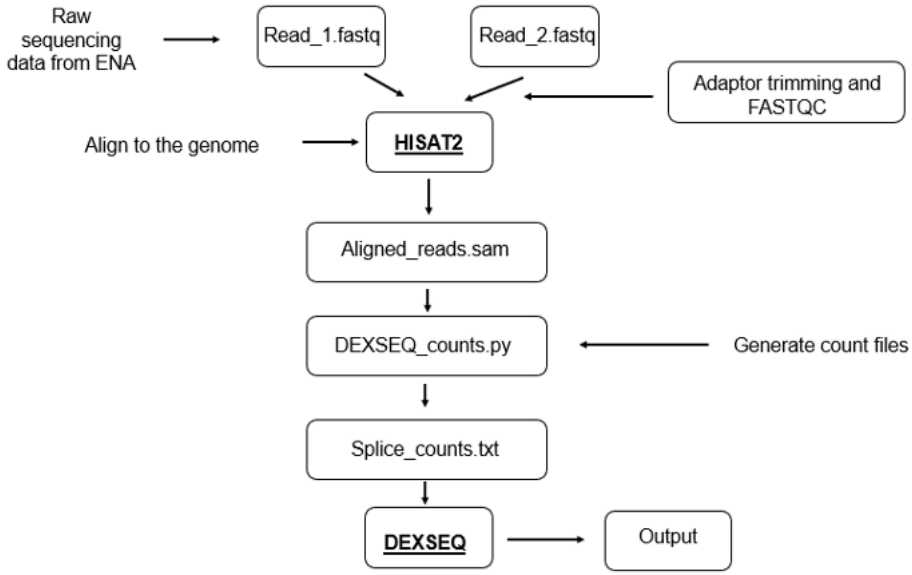


Figure 2.5: Overview of the DEXSEQ pipeline.

Scheme of an analyses pipeline for DEU analysis using DEXSEQ. A description of the pipeline is provided in the main text.

To generate DEXSEQ analyses, first, the quality of next-generation sequencing libraries generated from the HIPSCI study was assessed by FASTQC similarly to how quality control was conducted for DESEQ analyses. All of the raw sequencing data files passed the QC analyses from FASTQC. Adapter sequences were computationally trimmed from sequencing data using Trim_Galore. Trimmed adapter files were aligned to the human genome using HISAT2, and count files were generated and analyzed using DEXSEQ. When interpreting DEXSEQ data, it is important to note that DEXSEQ relies on an annotation file containing coordinates for exonic parts, which map to exons for all genes. However, exonic parts are either an entire exon or part of an actual exon-they do not necessarily correspond to whole exons (**Figure 2.6**). The annotation file was derived from the human annotation (Gencode v38). The annotation was modified to include the exons identified

for CACNA1C as part of Clark et al, 2020. Similarly to DESEQ2, reads are mapped onto a reference genome and when multiple reads are derived from the same genomic region, a representation of the exonic region can be inferred from the number of reads mapping to it. This produces a count that enables quantification of the expression levels of exonic parts.

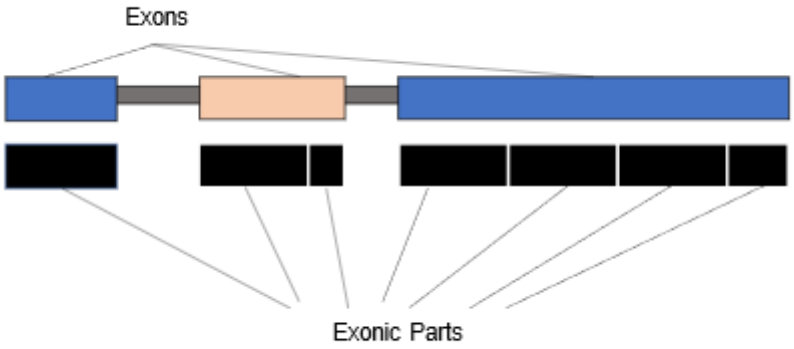


Figure 2.6: Exons versus exonic parts.

DEXSEQ relies on an annotation file containing coordinates for exonic parts, which map to exons. Exonic parts can, but do not necessarily, correspond to whole exons.

2.2.4.2 MAJIQ

The analyses of differential splice junction utilization used MAJIQ and was visualized using VOILA as shown in **Figure 2.7** (MAJIQ version 2.3). MAJIQ is a software program used to quantify and visualize LSVs in short-read RNA-seq data (MAJIQ; Welcome, 2021).

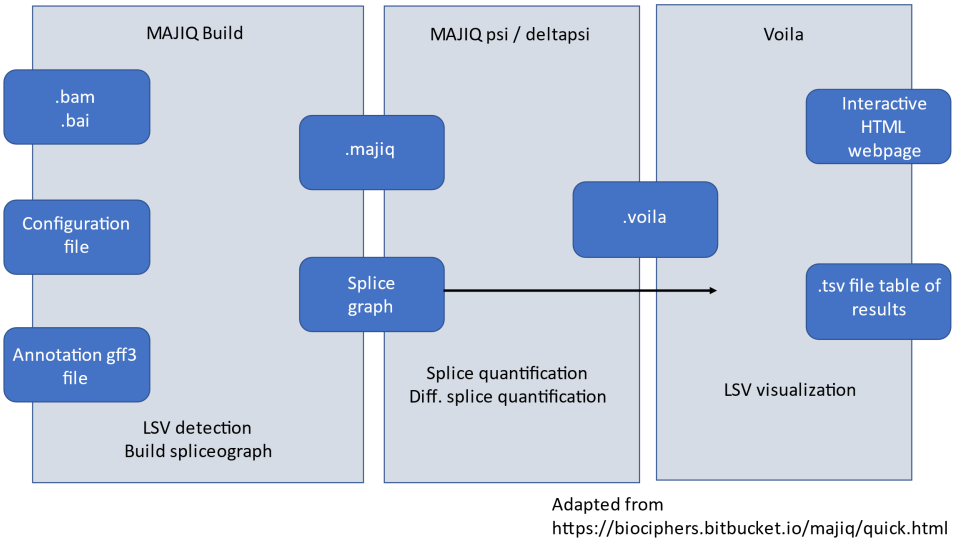


Figure 2.7: Overview of the MAJIQ pipeline.

Scheme of an analyses pipeline for differential splice junction usage analysis using MAJIQ. The input files and intermediate files are indicated. For example, "MAJIQ Build" receives input.bam, .bai, configuration file, and annotation file as input and produces .majiq files and a splice graph as output. (taken and adapted from MAJIQ: Welcome, 2021).

Figure 2.8 depicts a definition of LSVs (Vaquero-Garcia et al., 2016)**.** LSVs include both canonical types of alternative splicing (exon skipping, mutually exclusive exons, intron retention, and alternative 5'- or 3'- splice sites) and non-canonical types of splicing that do not fall into the canonical types listed above (Vaquero-Garcia et al., 2016).

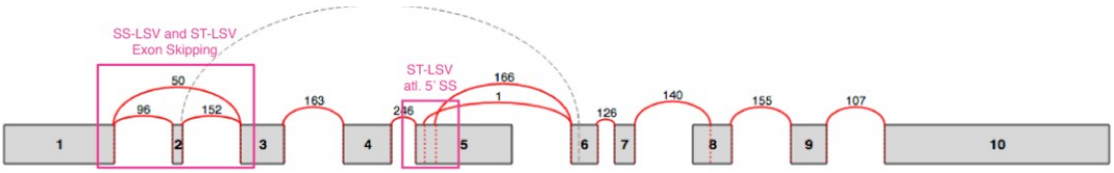


Figure 2.8: What is an LSV?

Rectangles are exons, junctions are arcs. The number of reads spanning two exons is indicated. SS-LSV stands for single source LSV (where splits originate), or ST-LSV stands for single target LSV where different junctions converge. Common alternative splicing events include exon skipping, alternative 5' splice site usage, or alternative 3' splice site usage. (figure taken and legend adapted from (What is an LSV, 2021) website).

As this study involves a comparison between two conditions (hiPSCs versus hiPSC-derived neurons), the program VOILA reports delta PSI (percent spliced in) quantifications. The delta PSI quantification provides information about the extent to which a particular splice junction is used (Vaquero-Garcia, 2016).

MAJIQ possesses limitations and cannot perform a number of different RNA-seq analyses relevant to the characterization of alternative splicing (Vaquero-Garcia, 2016). This is because MAJIQ focuses entirely on the differential usage of splice junctions. The limitations that MAJIQ cannot characterize include gene/isoform expression estimation, relative isoform abundance, novel gene/non-coding RNA detection, alternative transcription start/end, and the identification/quantification of alternative polyadenylation.

Thus, it is important to note that DEXSEQ and MAJIQ analyses investigate different aspects of alternative splicing. Whereas DEXSEQ is exon-based, MAJIQ is event-based (i.e. DEU versus differential splice junction usage). A limitation of DEXSEQ is that it does not identify which type of splicing event occurs in a gene but rather differentially expressed exonic parts between two different experimental conditions (Mehmood et al., 2020). It is also likely that MAJIQ analyses are insufficiently powered to detect LSVs in genes with limited transcript abundance. Nonetheless, the two methods complement each other and together provide a more complete picture of alternative splicing in RNA-seq data.

Finally, as mentioned above, the alignment of RNA-seq data to the genome was conducted using STAR for DESEQ2 analyses.

2.2.5 Statistics

For transcriptome-wide analyses, first, the similarity of datasets to each other was analyzed by principal component analyses (PCA) using DESEQ2 software. PCA is a standard statistical technique for reducing the dimensionality of complex datasets and is used to identify strong patterns in a dataset. PCA was used to determine whether different datasets would cluster according to their cell type of origin.

DESEQ2 and DEXSEQ each produced several output results from the raw-sequencing data, including baseMean, log2FoldChange, lfcSE, p-value, and adjusted p-value (defined in **Table 2.2**).

Some program-specific nuances in the present analyses are listed below. For DESEQ2 analyses, VGCC subunits that exhibited low expression levels were excluded from subsequent analyses (baseMean values <100). A baseMean value <100 has been used to filter out genes with lower expression levels in DESEQ2 analyses (Chiou et al, 2020; Jansson-Fritzberg et al, 2023). For DEXSEQ, a threshold of log2FoldChange <-1.5 or >1.5 was scored as a DEU event (i.e. a log2 fold change of 1.5 means that gene expression in-

Data point	Definition
baseMean	Average of normalized count values, divided by size factors, taken over all samples
log2FoldChange	Effect size estimate. This value indicates the extent to which a transcript's expression changed between two different groups. A positive value indicates higher expression in hiPSC-derived neurons whereas a negative value reflects higher expression in hiPSCs
lfcSE	log2FoldChange standard error
p-value	p-value of the Wald test for a gene or transcript
Adjusted p-value	Adjusted p-value for multiple testing for the gene or transcript (Benjamini-Hochberg Procedure)

Table 2.2: Definitions of DESEQ2/DEXSEQ output.

Definitions of baseMean, log2FoldChange, lfcSE, p-value, and adjusted p-value are presented.

creased by approximately 2.82 fold). A log2FoldChange of 1.5 as a threshold has been used in a previous study (Nazarov et al, 2017). For both DESEQ2 and DEXSEQ, the log2FoldChange value indicates the extent to which a transcript's or exonic part's expression changed between two different groups, respectively.

A p-value is a measurement of the likelihood of an observed difference being due to chance. A smaller p-value provides more convincing evidence to reject the null hypothesis. The null hypothesis, tested by DESEQ2 and DEXSEQ, would be that gene expression levels/DEUs in hiPSCs and hiPSC-derived neurons do not differ. The p-values produced by DESEQ2 and DEXSEQ refer to the p-value of the test for a gene or transcript or exonic part obtained by the Wald test. For DESEQ2 and DEXSEQ, the adjusted p-value refers to the adjusted p-value for multiple testing for the gene or transcript (the Benjamini-Hochberg Procedure). The Benjamini-Hochberg procedure is a method for decreasing the false discovery rate (i.e. a method for avoiding type I errors or false positives) (Benjamini and Hochberg, 1995).

2.2.6 Protein Sequence Alignment

Clustal Omega, a well-established bioinformatics tool, was used for multiple sequence alignment of the protein sequences of CACNA2D1 and CACNA2D2. (https://www.ebi.ac.uk/Tools/msa/clustalo/).This approach was used to determine whether DEUs in both CACNA2D1 and CACNA2D2 were found to occur in the same orthologous region of the respective proteins.

2.3 Results

The primary aim of the analyses conducted in this Chapter is to characterize changes in the transcriptional regulation of VGCCs as hiPSCs differentiate into neurons. The results can be broadly divided into two sections: one about changes in gene expression (i.e. changes in transcript abundance) and one about changes in alternative splicing.

2.3.1 Gene Expression Changes Across hiPSC Differentiation to Neurons

In this section, efforts focused on characterizing changes in the expression levels of VGCCs across hiPSC differentiation. Changes at the global level were analyzed first, followed by paying closer attention to VGCCs.

2.3.1.1 Transcriptome-Wide Comparison of hiPSCs and hiPSC-derived Neurons

As noted in the Methods, all of the raw sequencing data files passed FASTQC quality control. PCA was conducted on the full dataset containing two groups of hiPSCs and hiPSC-derived neurons (**Figure 2.9**). PC 1 revealed two groups of samples: datasets from iPSCs formed one cluster and datasets from iPSC-neurons formed a second cluster. Compared with hiPSCs, hiPSC-derived neurons showed a greater spread along PC2.

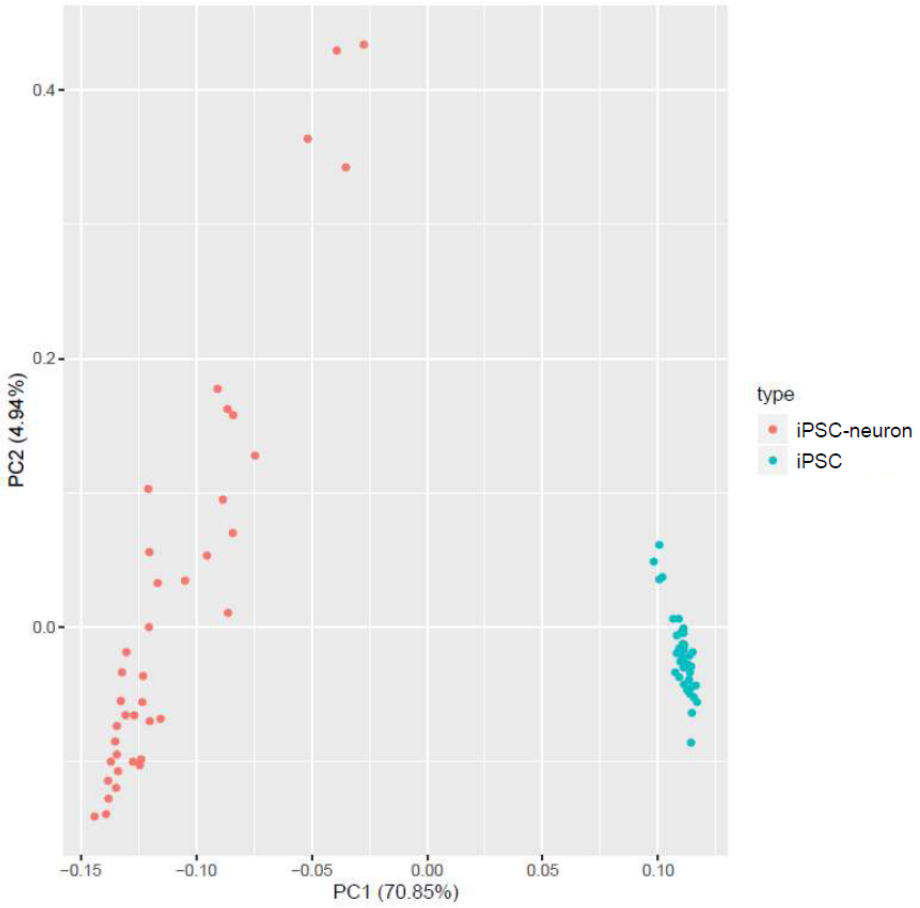


Figure 2.9: Principal component analysis (PCA) for datasets from Schwartzentruber et al., 2018.

hiPSCs (iPSCs) (teal); hiPSC-derived neuron (iPSC-neuron) (red). Each sample is represented with a dot.

Differential gene expression analyses via DESEQ2 found 42 67 genes more highly expressed in hiPSCs and 6 303 genes more highly expressed in hiPSC-derived neurons (|log2FoldChange|>1.5; q-value <0.05). The genes that showed the greatest differential expression (i.e. the highest magnitude of log2FoldChange) between hiPSCs and hiPSC-derived neurons were identified (see **Table 2.3** and **Table 2.4**). Notably, OCT4 was the top hiPSC-enriched gene. OCT4 is a TF associated with the pluripotent state (**Table 2.3**) (Takahashi and Yamanaka, 2006; Takahashi et al., 2007). NANOG is another key TF linked with pluripotency (Yu et al., 2007).

With regard to the 6 303 genes more highly expressed in hiPSC-derived neurons, the list includes factors identified as having a role in neural biology, including DRGX (also known as DRG11), a TF important in sensory neurons (**Table 2.4**) (Chen et al., 2001; Rebelo et al., 2006); TRIM67, which regulates brain development and behavior (Boyer et al., 2018); and ISL1, which is expressed in motor neurons and is necessary for their

development (Liang et al., 2011). The higher expression of neuron-specific regulators in hiPSC-derived neurons, therefore, confirms the successful differentiation of hiPSCs into a neuron-like identity.

Gene	baseMean	log2FoldChange	Padj
OCT4	7604.63	-15.101	4.08 E-53
TDGF1	3654.42	-14.930	3.69 E-33
VRTN	1365.89	-13.362	1.34 E-25
ESRG	4727.88	-13.238	1.00 E-19
ZSCAN10	1338.756	-12.450	2.08 E-60
LINC00678	661.55	-12.316	2.86E-231
LCK	455.349	-11.703	2.70E-260
NANOG	355.08	-11.418	2.86E-231
FOXH1	438.656	-11.205	4.14E-234
MIR302CHG	1431.295	-11.033	5.30 E-53

Table 2.3: Top 10 hiPSC-enriched genes.

The table contains the results baseMean, log2FoldChange, and adjusted p-value (Padj). See **Table 2.2** for interpretation of results.

Gene	baseMean	log2FoldChange	Padj
DRGX	2155.077	14.311	2.69E-224
LINC02735	2357.138	14.069	5.75E-161
TRIM67	1405.91	13.473	2.59 E-36
CARTPT	643.387	13.085	8.17E-101
VIP	222.532	12.457	9.91E-45
PRPH	5819.707	12.389	6.82 E-36
ONECUT3	1374.229	11.152	3.41 E-63
NSG2	4298.743	11.147	1.29 E-36
GRIA2	720.874	11.109	1.15E-231
CTD-2314G24.2	549.840	11.020	5.48E-178

Table 2.4: Top 10 hiPSC-derived neuron-enriched genes.

The table contains the results baseMean, log2FoldChange, and adjusted p-value. See **Table 2.2** for interpretation.

2.3.1.2 Changes in the Abundance of VGCC Subunits Across hiPSC Differentiation

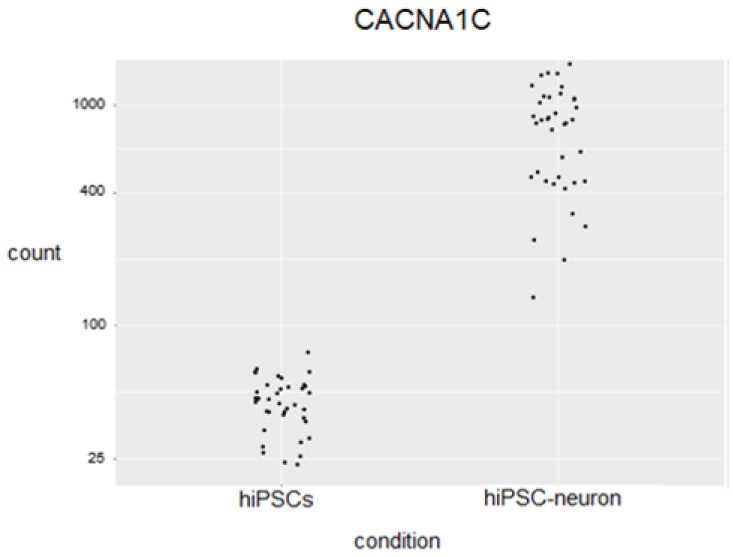
The results of differential gene expression analyses are summarized in **Table 2.5**. Generally, VGCC expression increased across hiPSC differentiation to neurons. Among VGCC subunits, CACNA1A, CACNA1B, CACNA1C, CACNA1D, CACNA1E and, CACNA1I were more highly expressed in hiPSC-derived neurons compared to hiPSCs. The calcium channel subunits similarly expressed between hiPSCs and hiPSC-derived neurons were CACNA1G and CACNA1H (**Figure 2.10** shows data for CACNA1C; graphs for the other alpha-1 subunit data are not shown). CACNA1F and CACNA1S were excluded from subsequent analyses based on low expression levels (DESEQ2 baseMean < 100).

Among the auxiliary subunits, CACNA2D4 was expressed more highly in hiPSCs. In contrast, CACNA2D1, CACNA2D3, CACNB1, CACNB3, and CACNB4 were more highly expressed in hiPSC-derived neurons (see **Supplementary Figure 1.1**; other auxiliary subunits are not shown). CACNA2D2 and CACNB2 were similarly expressed between hiPSCs and hiPSC-derived neurons. CACNA2D4 was excluded based on low expression levels (DESEQ2 baseMean < 100).

Gene	ENSG	baseMean	Log2FoldChange	lfcSE	P-value	Padj	hiPSCs?	hiPSC-neuron?	# of exonic parts changed
CACNA1A	ENSG00000141837	653.29	1.845	0.126	1.543E-48	9.44E-48		Higher	9
CACNA1B	ENSG00000148408	1384.19	3.95	0.19	9.74E-94	1.22E-92		Higher	0
CACNA1C	ENSG00000151067	430.21	4.17	0.14	6.05E-188	2.42E-186		Higher	3
CACNA1D	ENSG00000157388	148.92	4.23	0.14	1.45E-213	7.63E-212		Higher	19
CACNA1E	ENSG00000198216	648.15	5.63	0.16	4.86E-272	4.50E-270		Higher	0
CACNA1F	ENSG00000102001	25.33	-6.55	0.39	7.35E-63	5.71E-62	Higher		N/A
CACNA1G	ENSG00000006283	108.16	-0.17	0.20	0.39	0.44	Higher		N/A
CACNA1H	ENSG00000196557	1307.51	0.54	0.088	8.37E-10	1.97E-09		Higher	0
CACNA1I	ENSG00000100346	142.16	1.74	0.16	3.00E-26	1.160E-25		Higher	16
CACNA1S	ENSG00000081248	2.30	-1.33	0.30	0.0000116	0.0000220	Higher		N/A
CACNA2D1	ENSG00000153956	2104.96	2.57	0.131	6.20E-85	6.82E-84		Higher	3
CACNA2D2	ENSG00000007402	958.89	0.119	0.0942	0.206	0.250		Higher	1
CACNA2D3	ENSG00000157445	480.42	4.30	0.166	1.78E-148	4.81E-147		Higher	1
CACNA2D4	ENSG00000151062	7.01	-2.53	0.318	1.59E-15	4.59E-15	Higher		68
CACNB1	ENSG00000067191	735.63	3.35	0.107	3.66	1.93E-212		Higher	0
CACNB2	ENSG00000165995	100.29	-0.591	0.117	4.298E-07	8.89E-07		Higher	N/A
CACNB3	ENSG00000167535	918.67	-3.12	0.111	1.31E-173	4.51E-172		Higher	3
CACNB4	ENSG00000182389	273.37	-2.25	0.12	6.64E-75	6.21E-74		Higher	N/A

Table 2.5: DESEQ2 and DEXSEQ analyses of differential gene expression of individual VGCC subunits.

Examining differential gene expression and DEU for VGCC subunits across cellular differentiation. The table contains the results baseMean, log2FoldChange, lfcSE (standard error of log2FoldChange), p-value, adjusted p-value, columns indicating whether the gene is expressed more highly in hiPSCs or hiPSC-derived neurons, and columns indicating the number of exonic parts changed for DEXSEQ data. For DEXSEQ analyses, the number of exonic parts scored as changed between hiPSCs and hiPSC-derived neurons for each VGCC subunit is indicated in the final column. Italics font indicates VGCC subunits that were excluded from subsequent analyses owed to low baseMean values (<100).



(A)

Figure 2.10: Differential gene expression analysis of the CACNA1C subunit.

The figure presents dot plots containing normalized counts for CACNA1C subunits in either hiPSCs or hiPSC-derived neurons. For other VGCC subunits, see **Supplementary Figure 1.1** or data not shown. Each sample is represented with a dot.

2.3.2 Changes in Alternative Splicing in VGCCs Across hiPSC Differentiation

In this section, changes in alternative splicing in VGCCs across hiPSC differentiation have been characterized. Both changes in DEU and differential splice junction utilization that were identified will be described.

2.3.2.1 Changes in Differential Exon Usage Across hiPSC Differentiation

DEXSEQ identified DEU in different VGCC subunits (**Table 2.5**). An example of the DEXSEQ output plot (for CACNA1C) is shown in **Figure 2.11.** Three exonic parts showed significant differential usage (defined here as log2FoldChange <-1.5 or > 1.5; adjusted p-value < 0.05).

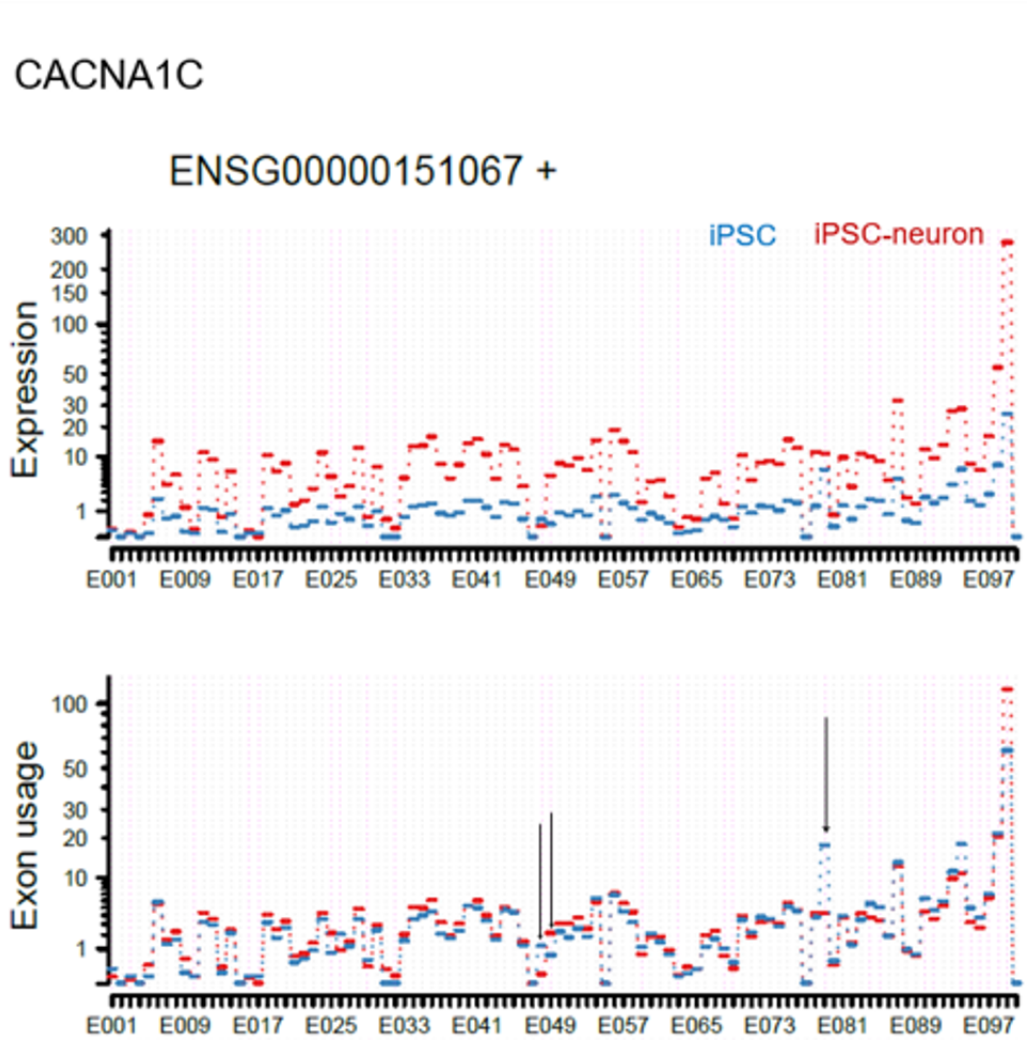


Figure 2.11: Testing for DEU in CACNA1C using DEXSEQ between hiPSCs and hiPSC-derived neurons

Expression level values (top) and, and the exon usage values (bottom) of different individual VGCC subunits in hiPSCs and hiPSC-derive neurons are shown. Numbers along the x-axes of the top and middle plots indicate the exonic part number. Blue: iPSC; red: hiPSC-derived neuron (iPSC-neuron). Black arrows in the bottom plot indicate exonic parts that are differentially utilized between the two cell states. It is important to note DEXSEQ identifies changes in exon usage that are not merely a result of changes in overall expression levels.

To better visualize the results and to identify whether differential expression targeted consistent parts of the VGCCs, DEU events were mapped onto the generic protein models of their respective VGCC subunits **(Figure 2.12; Figure 2.13)**. The S2 region of domain III showed changes in L-type VGCCs CACNA1C and CACNA1D, indicating that it could be a potential candidate hotspot for alternative splicing, at least in L-type VGCCs **(Figure 2.12; Figure 2.13)**. Changes in the S2 region of domain III were not observed in non-L-type VGCC subunits CACNA1A and CACNA1I. Changes were also observed in the C-terminus of the CACNA1C and CACNA1D proteins, along with CACNA1A, as well as those in the 3'UTR of CACNA1A and CACNA1D.

Finally, the protein sequence alignment of CACNA2D1 and CACNA2D2 revealed that DEUs identified in CACNA2D1 and CACNA2D2 mapped to the same orthologous region of both proteins. However, this region did not map either to the VWA or to the CACHE domains, which are important regulatory regions of auxiliary alpha-2-delta subunits.

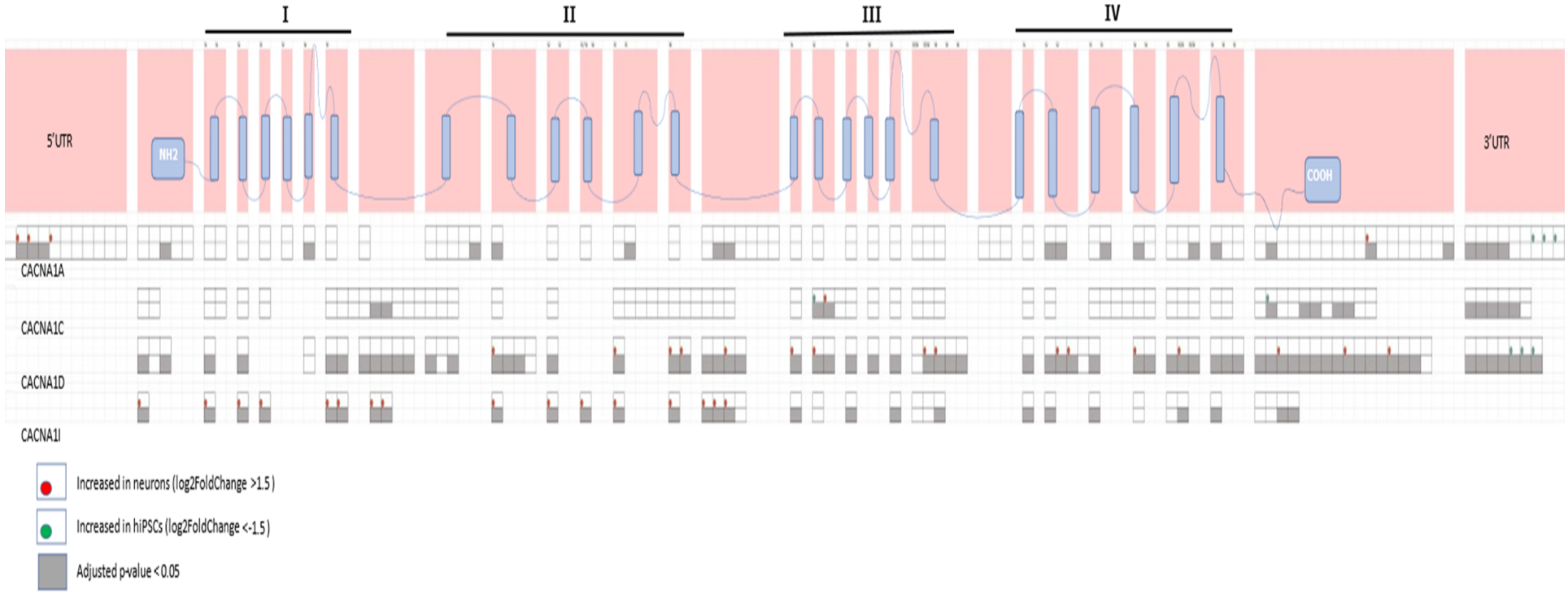


Figure 2.12: Summary of DEU across different VGCC alpha-1 subunits (regions where threshold |log2FoldChange|>1.5)

The exonic parts differentially utilized between hiPSCs and hiPSC-derived neurons were mapped onto a generic VGCC alpha-1 subunit protein model. Each rectangle contains a top portion (magnitude of DEU) and a bottom portion (adjusted p-value) of an exonic part. For the magnitude of DEU, a box with a red circle indicates an increase in neurons, and a box with a green circle indicates an increase in hiPSCs. The threshold for differential exon usage was log2FoldChange <-1.5 or >1.5. For the bottom portion, gray indicates an adjusted p-value <0.05. Note some exonic parts may exhibit a low magnitude of log2FoldChange but may still show an adjusted p-value <0.05. An event was scored as an actual DEU event if both the magnitude of change (log2Foldchange) was greater than 1.5 or less than -1.5 and the adjusted p-value was less than 0.05. Only VGCC alpha-1 subunits with significant expression levels and statistically significant DEU changes are mapped here. The alpha-1 subunits not listed here either showed low expression or no significant changes in DEU. No hotspots were observed in auxiliary VGCC subunits. Note the DEXSEQ annotation file did not have exonic parts that mapped to the 5'UTRs CACNA1C, CACNA1D, and CACNA1I.

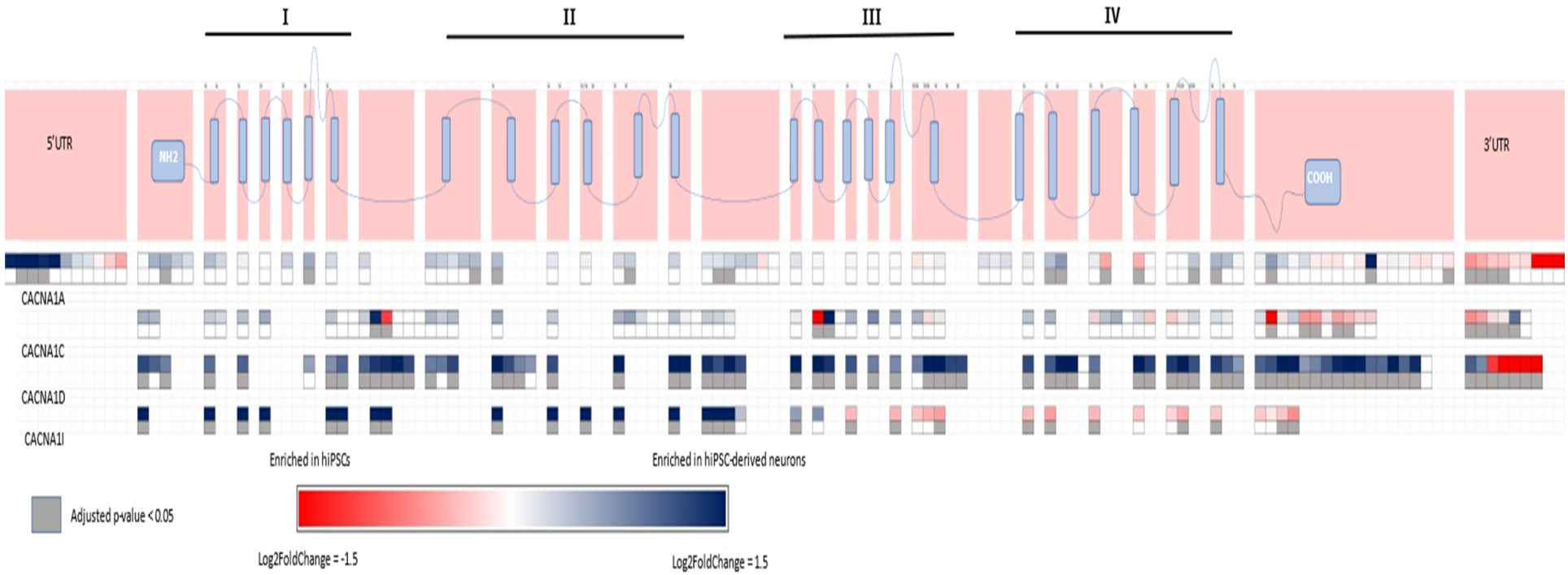


Figure 2.13: Summary of DEU across different VGCC alpha-1 subunits (heatmap including regions where |log2FoldChange|<1.5).

The exonic parts differentially utilized between hiPSCs and hiPSC-derived neurons were mapped onto a generic VGCC alpha-1 subunit protein model, similarly to **Figure 2.12.** Each rectangle contains a top portion (magnitude of DEU) and a bottom portion (adjusted p-value) of an exonic part. For the magnitude of DEU in B), the magnitude of DEU is represented as a heatmap. For the bottom portion, gray indicates an adjusted p-value <0.05. Note some exonic parts may exhibit a low magnitude of log2FoldChange but may still show an adjusted p-value <0.05. **Figure 2.12.** and **Figure 2.13.** are similar representations of DEXSEQ data, but **Figure 2.12.** shows the exonic parts with differential expression and adjusted p-values that met the threshold, and **Figure 2.13.** also shows (via heatmap representation) DEU of log2foldchange greater than -1.5 and less than 1.5. Only VGCC alpha-1 subunits with significant expression levels and statistically significant DEU changes are mapped here. The alpha-1 subunits not listed here either showed low expression or no significant changes in DEU. No hotspots were observed in auxiliary VGCC subunits. Note the DEXSEQ annotation file did not have exonic parts that mapped to the 5'UTRs CACNA1C, CACNA1D, and CACNA1I.

2.3.2.2 Changes in Differential Splice Junction Usage of VGCC Subunits Across hiPSC Differentiation

The previous section focused on DEU across the undifferentiated and differentiated cell states. Next, the differential utilization of splice junctions across the two cellular states was investigated.

The use of different bioinformatics tools identified a number of different alternative splicing events (summary of findings in **Table 2.6**). For example, DEXSEQ identified a number of differentially expressed exonic parts in hiPSCs and hiPSC-derived neurons. In contrast, MAJIQ identified a limited number of alternative splicing events.

	DEXSEQ		**MAJIQ**	
Gene name	**Number of exonic parts**	**Number of differentially expressed exonic parts**	**Number of differentially expressed junctions (LSVs)**	**Number of differentially expressed novel junctions (LSVs)**
CACNA1A	165	9	0	n/a
CACNA1B	60	0	0	n/a
CACNA1C	100	3	0	n/a
CACNA1D	113	19	0	n/a
CACNA1E	69	0	0	n/a
CACNA1F	*N/A*	*N/A*	*0*	*n/a*
CACNA1G	N/A	N/A	0	n/a
CACNA1H	54	0	0	n/a
CACNA1I	41	16	0	n/a
CACNA1S	*N/A*	*N/A*	*0*	*n/a*
CACNA2D1	59	3	4	2
CACNA2D2	53	1	2	0
CACNA2D3	59	3	0	n/a
CACNA2D4	*93*	*68*	*0*	*n/a*
CACNB1	34	0	0	n/a
CACNB2	N/A	N/A	0	n/a
CACNB3	64	3	0	n/a
CACNB4	N/A	N/A	0	n/a

Table 2.6: Summary of alternative splicing analyses findings.

DESEQ2 (gene expression levels), DEXSEQ (DEU) and MAJIQ (differential splice junction usage) were used to characterize transcript abundance and to identify alternative splicing events in this Chapter. The findings for the different VGCC subunits are summarized. Italics indicate VGCC subunits that were excluded from analyses due to low expression levels (DESEQ2 baseMean <100).

MAJIQ was conducted on all VGCC subunit genes but only detected significant changes in CACNA2D1 and CACNA2D2 genes (**Figure 2.14)** (in other words, MAJIQ did not detect changes in alpha-1 subunits). MAJIQ analysis identified four LSVs in CACNA2D1 and two LSVs in CACNA2D2. Interestingly, two novel splice junctions were identified among the CACNA2D1 LSVs.

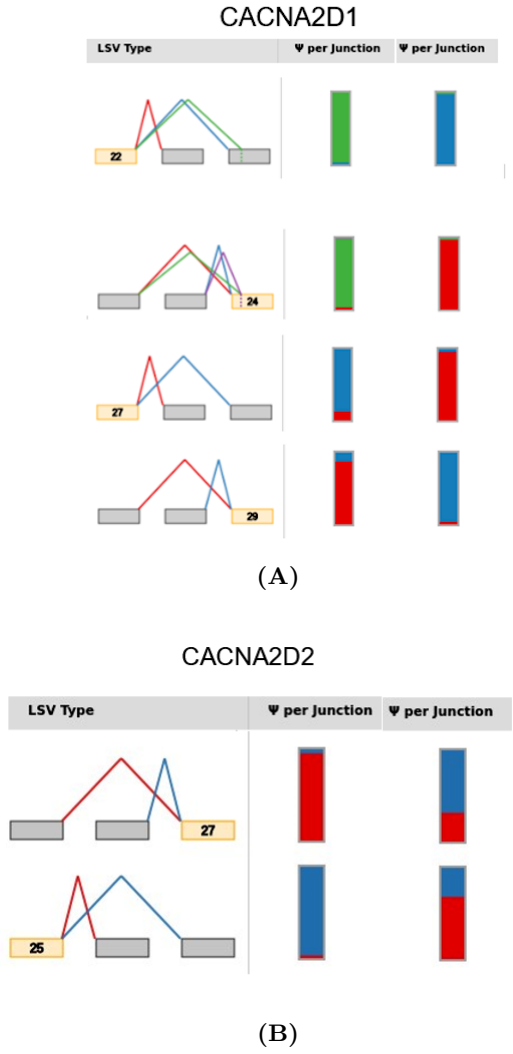


Figure 2.14: Differential splice junction usage in CACNA2D1 and CACNA2D2.

LSV type: type of LSV observed (i.e. deletions, differential inclusions of mutually exclusive exons, etc.). (left) (LSV type): diagram indicating differential splice junction usage shown. Ψ refers to the percent spliced in (PSI); (middle) (first Ψ per junction): PSI in hiPSCs; (right) (second Ψ per junction): PSI in hiPSC-derived neurons. (A) CACNA2D1; (B) CACNA2D2. For example, in (B), the top middle figure indicates that the red splicing event (skipping of exon 26) occurs more frequently in hiPSCs, whereas the top right figure indicates that the blue splicing event (inclusion of exon 26) occurs more frequently in hiPSC-derived neurons.

2.3.2.3 Predicted Impact of Alternative Splicing on CACNA2D1 and CACNA2D2 Function

DEXSEQ identified DEU events in CACNA2D1 and MAJIQ identified differential splice junction usage in CACNA2D1. Given the presence of both DEXSEQ and MAJIQ findings for CACNA2D1, the consequences of the identified splicing events on the CACNA2D1 gene model were examined. It should be noted that exon notations were different between MAJIQ and DEXSEQ, so exons will be referred to according to their respective annotations. DEXSEQ identified DEU events in the alpha-2 moiety and the calcium channels

and chemotaxis receptor(CACHE) domain. The identified DEU events did not shift the open reading frame. MAJIQ identified splicing events in the CACHE domain. Reassuringly, the two MAJIQ events co-occurred at the same location with DEUs identified by DEXSEQ.

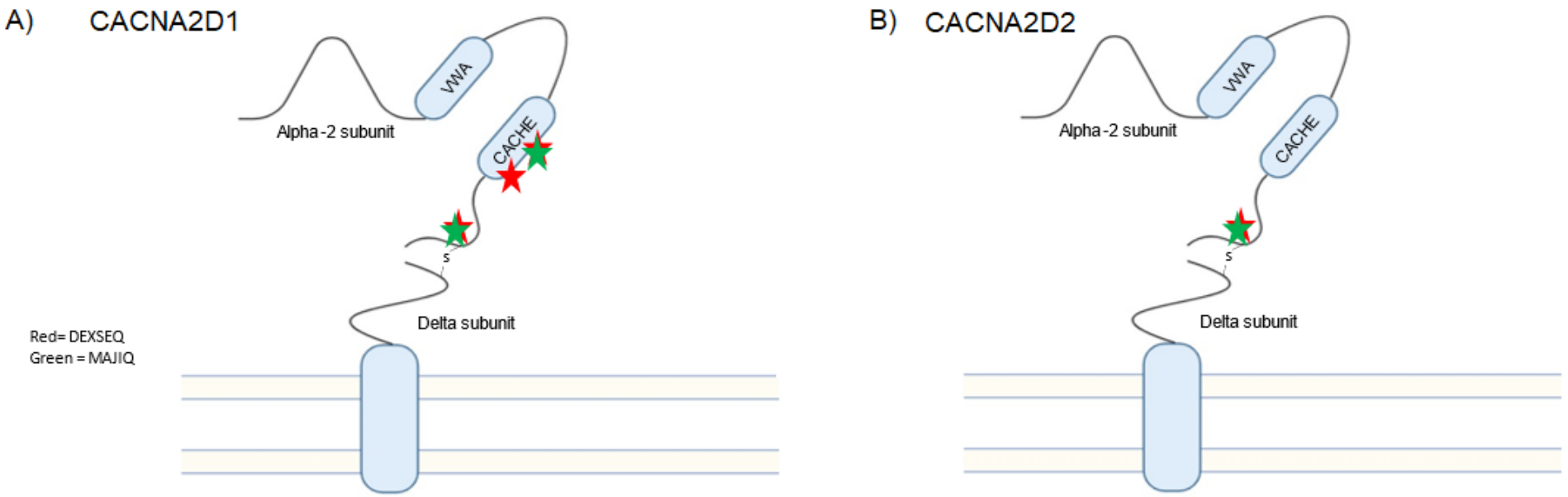


Figure 2.15: Mapping alternative splicing events onto CACNA2D1 and CACNA2D2 protein models.

Alternative splicing events between hiPSCs and hiPSC-derived neurons were mapped onto (A) CACNA2D1 and (B) CACNA2D2 protein models. Stars mark splicing events and might indicate events in multiple exons. "S" indicates linkage by a disulfide bond (Calderon-Rivera et al., 2012). The red stars indicate DEXSEQ findings, and a green star indicates MAJIQ findings. The overlapping red and green stars indicate DEXSEQ and MAJIQ findings at the same location (generated with biorender.com).

Regarding CACNA2D2, MAJIQ identified two LSVs in CACNA2D2 involving MAJIQ Exons 25 – 27 that do not map either to the VWA or to CACHE domains**.** DEXSEQ analyses of CACNA2D2 identified a DEU event that colocalized to the same region as the MAJIQ finding. The DEU event did not shift the open reading frame, indicating that alternative splicing likely produces full-length proteins.

Mappings of MAJIQ and DEXSEQ events onto the CACNA2D1 and CACNA2D2 protein model are provided in **Figure 2.15.**

Finally, the Clustal Omega protein sequence alignment of CACNA2D1 and CACNA2D2 revealed that DEUs identified by DEXSEQ in CACNA2D1 and CACNA2D2 occurred in the same orthologous region of both proteins. MAJIQ analyses identified the inclusion of this region in hiPSC-derived neurons and deletion in hiPSCs. This region did not map either to the VWFA or to CACHE domains.

2.3.3 Results Summary

Summaries of all results from the analyses of transcript abundance and alternative splicing events are presented in **Table 2.5** (abundance) and **Table 2.6** (splicing)**.** Overall expression levels (DESEQ2), DEU (DEXSEQ), and differential splice junction utilization (MAJIQ) are presented.

Altogether, these analyses of the HIPSCI RNA-seq data revealed changes in the abundance and splicing of VGCCs as hiPSCs differentiate into neurons. More detailed analyses identified alternative splicing events of interest in CACNA1C, CACNA2D1, and CACNA2D2.

2.4 Discussion

Here, transcriptomic analyses of VGCCs were conducted in 40 isogenic pairs of hiPSCs and hiPSC-derived neurons from the HIPSCI. The present analyses found changes in the abundance and splicing of VGCCs in the transition from hiPSCs to hiPSC-derived neurons, information that is critical to advance hiPSC-derived neurons as a model system for studying VGCC biology.

Our findings demonstrate the dynamic regulation of VGCCs across distinct cellular states. Analyses revealed that VGCC abundance tends to increase across differentiation, consistent with the emerging electrophysiological activity of hiPSC-derived neurons. Moreover, the data identified changes in the splicing of VGCCs across cell states. L-type

VGCCs CACNA1C and CACNA1D showed DEU in hiPSCs and hiPSC-derived neurons, including in the domain III repeat 2 region and at the C-terminus. For CACNA1C in particular, changes in Exon 21/22 splicing of CACNA1C were observed across hiPSC differentiation, with DEXSEQ analyses showing a shift from Exon 21 utilization in hiPSCs to Exon 22 utilization in hiPSC-derived neurons. Exon 22 is also preferentially utilized in the adult human brain compared with the heart and aorta, indicating that hiPSC-derived neurons replicate brain-specific splicing patterns in CACNA1C, validating hiPSC-derived neurons as a model (N. Hall, unpublished observations). Channels containing Exon 21 or Exon 22 exhibit differential sensitivity to DHP inhibition, particularly in the voltage dependence of DHP action (Soldatov et al., 1995). The C-terminus is the site of multiple binding domains that connect CACNA1C signaling to intracellular signaling cascades and regulate calcium-dependent inactivation. Given the role of the C-terminus in the regulation of channel inactivation and intracellular signaling, splicing might produce isoforms with altered inactivation properties or coupling to second messenger systems.

As mentioned earlier, a |log2FoldChange|>1.5 and q-value <0.05 were used to score positive DEU events for DEXSEQ. DEXSEQ found that Exon 8 is upregulated in hiPSC-derived neurons (log2FC = 1.426; q-value = 0.00186) and Exon 8A is upregulated in hiPSCs (log2FC = -1.109; q-value = 6.88E-05). Because the |log2FoldChange| was less than 1.5, this was not scored as a change. These results raise questions about whether the chosen DEXSEQ threshold may exclude potentially biologically relevant changes. Their statistical significance suggests they are differentially expressed. Future studies might reconsider effect size when interpreting results. Additionally, the direction of changes are consistent with a previous report comparing control hiPSC-derived neural progenitor cells (NPCs) to hiPSC-derived neurons (Panagiotakos et al, 2019).

Regions of auxiliary subunits CACNA2D1 and CACNA2D2 also showed changes in alternative splicing. Both DEXSEQ and MAJIQ identified DEU and differential splice junction usage at Exon 23 in CACNA2D1, which maps neither to the VWA nor CACHE domain (**Figure 2.15**). In rats, loss of Exon 23 in the alpha-2 moiety is associated with a lower binding affinity for gabapentin compared to variants that contain Exon 23 (Lana et al., 2014). Therefore, as observed in rats, one could surmise that the gabapentin binding affinity of CACNA2D1 could also be altered in hiPSCs and hiPSC-derived neurons. Since DEXSEQ identified increased usage of Exon 23 in hiPSC-derived neurons, it is possible that hiPSC-derived neurons gain a higher binding affinity for gabapentin as hiPSCs differ-

entiate into neurons. DEXSEQ and MAJIQ also identified splicing changes in the CACHE domain of CACNA2D1 (**Figure 2.15**).The chemosensory-like domains are implicated in the trafficking of alpha-2-delta subunits and are also speculated to be implicated in binding to gabapentinoid drugs (Dolphin, 2018). The fact that two different computational methods assessing different types of splicing events identified such events at the same locations provides confidence in the present analyses. Analyses also identified differential splicing of CACNA2D2. Notably, protein sequence alignment of CACNA2D1 and CACNA2D2 revealed that DEUs identified by DEXSEQ in CACNA2D1 and CACNA2D2 occurred in the same orthologous region of both proteins. Altogether, analyzing the differentiation of hiPSCs to neurons opens up new opportunities to probe the role of auxiliary subunits in the emergence of the neuronal phenotype and neuronal networks.

Finally, the use of different bioinformatic methods is a complementary way of analyzing abundance and splicing changes across hiPSC differentiation (**Table 2.5**).

2.4.1 Limitations

DEXSEQ and MAJIQ possess limitations that are important to consider. Both DEXSEQ and MAJIQ rely on high-quality transcriptome annotation. Inaccurate or incomplete annotations can affect the overall analysis. In this regard, DEXSEQ can only analyze known exons, meaning that novel exons or alternative splicing events not present in the reference annotation will not be detected. A similar limitation pertains to MAJIQ, as it relies on the reference annotation to build infer splice junction usage. DEU detection can be affected by the read length and sequencing depth of the RNA-seq data. Short read lengths can result in ambiguous mappings, while insufficient sequencing depth may limit the detection of lowly expressed exons or rare splicing events. For instance, MAJIQ may struggle to accurately quantify splicing events in lowly expressed genes due to limited read coverage, potentially leading to false negative results.

For some studies, biases introduced during library preparation and sequencing can affect the accuracy of analyses. Unwanted technical variation, such as differences in sample preparation of sequencing runs, can introduce batch effects that may confound the results. Moreover, a sufficient number of biological replicates is needed for robust statistical analyses. The ability to study a sizable (at least, to the extent that has been reported to date) set of hiPSCs and hiPSC-derived neurons from the same study and laboratory aims to address these issues, but it cannot be fully excluded that these factors still affect the

quality of the analyses conducted in this Chapter.

Additional key limitations concern study design. The fact that only two cell states are considered, and particularly that no other differentiated cultures were examined, means that it is impossible to disentangle whether observations can be considered as specific to neurons or related simply towards being a non-pluripotent, differentiated cell type. It will be crucial to analyze hiPSC derivatives from alternative lineages, such as cardiomyocytes or smooth muscle. Reassuringly, the pattern of Exon 21 and Exon 22 DEU in hiPSC-derived neurons recapitulates DEU in the adult human brain (Soldatov et al., 1995), indicating that hiPSC-derived neurons, at least in this instance, recapitulate what is observed in authentic human neural tissue.

If the present findings are indeed specific to neurons, one should also consider the inherent features of the hiPSC-derived neuron cultures analyzed in the present studies with regard to heterogeneity and the level of fidelity to the adult human brain and heterogeneity. PCA confirm the presence of heterogeneity in the hiPSC-derived neuron datasets. Heterogeneity could be caused by sampling at slightly different stages of differentiation, suboptimal conditions for differentiation, or even the culture condition of starting hiPSCs. Interestingly, all the hiPSC datasets sampled appeared to cluster closely at the gene expression level, as revealed by PCA, regardless of culture condition. The effects of different cultural conditions are an area of future investigation. Population RNA-seq analyses showed greater similarity of hiPSC-derived neurons to hiPSCs compared with primary dorsal root ganglia samples, which was postulated to indicate a lack of maturity in hiPSC-derived neurons, a commonly reported problem with hiPSC-derived derivatives (Schwartzentruber et al., 2018). Additionally, as hiPSCs and hiPSC-derived neurons were derived from the same individuals, the increased similarity between hiPSCs and hiPSC-neurons might also be attributed to donor genetic background (Schwartzentruber et al., 2018). It is also important to note that hiPSC-derived neuron datasets were characterized as having greater heterogeneity when compared with dorsal root ganglia samples (Schwartzentruber et al., 2018). HiPSC-derived neuron cultures were also reported to contain subpopulations of fibroblast-like cells (Schwartzentruber et al., 2018).

2.4.2 Future Directions

Beyond the computational tools DEXSEQ and MAJIQ used in this Chapter, short-read sequencing data analyses of alternative splicing events more generally possess limitations

that could be overcome by emerging technologies. Usually, the analyses of short-read sequencing data centers around individual exons (DEXSEQ) or small numbers of exons (MAJIQ) but cannot capture full-length transcripts where alternative exons might exist. A potential future direction involves the application of long-read sequencing approaches towards studies of VGCC splicing. Indeed, the advent of long-read sequencing has identified new splicing events in CACNA1C and demonstrated that current transcriptome annotations, which relied upon for analyses of splicing in short-read sequencing data, are not complete (Clark et al., 2020).

The functional significance of the newly identified splicing events remains to be determined, but they suggest that VGCCs might play different functional roles in hiPSCs compared with neurons. While some VGCC subunits were excluded from analyses due to low baseMean values, it is worth noting that higher expression levels of CACNA1F, CACNA1G, CACNA1S, and CACNA2D4 were observed in hiPSCs (**Table 2.5**). Surprisingly little is known about the roles that calcium, and more specifically VGCCs, play in the pluripotency and differentiation of hiPSCs. Given the role of VGCCs in regulating the neural induction of early amphibian embryos, and the observation that increased intracellular calcium can be seen in both amphibian and mammalian embryos, it could be interesting to determine whether VGCCs also regulate neural differentiation of hiPSCs. One experiment is to inhibit VGCCs during the neural induction of hiPSCs and to determine whether this affects neural differentiation.

The expression of VGCCs in hiPSCs, which do not fire action potentials, raises the interesting question of what the function of VGCCs have in non-excitable cells (Pitt et al., 2021). Historically, this function of VGCCs has been less investigated because of the lack of a known mechanism by which cells without the ability to fire an action potential could regulate VGCCs (Pitt et al., 2021). However, dysregulated limb development, craniofacial abnormalities, and immunodeficiencies observed in patients that harbor TS mutations in CACNA1C suggest VGCCs do have a function in non-excitable cells (Pitt et al., 2021). It has been suggested that VGCCs in non-excitable cells can be activated by mechanisms that do not require action potentials (Pitt et al., 2021). For example, osteocytes express a mechanosensitive channel Piezo1, which provides an inward depolarizing current, including calcium, when induced by stretch and mechanical forces (Pitt et al., 2021). Additionally, VGCCs in non-excitable cells could be turned on by membrane depolarization by transient receptor potential cation channels (TRPCs). If VGCCs possess a function in hiPSCs, it

would be of great interest to determine whether VGCCs would be activated in hiPSCs, and how. It is worth noting that as all cells have a membrane potential, the presence of VGCCs in non-excitable cells could be contributing to homeostasis.

A body of evidence demonstrates that VGCCs (such as CACNA1C) that have well-understood functions in mature neurons possess previously unappreciated, different roles in less differentiated cells, suggesting possible functions in neural development (Panagiotakos and Pasca, 2021). For example, CACNA1C undergoes a switch in exon usage from Exon 8A to Exon 8 as neural progenitors differentiate into neurons (Panagiotakos and Pasca, 2021). The TS mutation in CACNA1C inhibits the switch from 8A to 8, disrupting neural differentiation (Panagiotakos et al., 2019). Therefore, alternative splicing modulates the developmental functions of VGCCs by creating isoforms with different features (Tang et al., 2009; Panagiotakos and Pasca, 2021). hiPSCs represent a human-relevant system that enables the dissection of the developmentally distinct and cell-specific roles of VGCCs, as well as of their regulation by alternative splicing.

Another potential future direction is to carry out more studies of differentiation-associated changes in VGCCs at the protein level. It will be of interest to confirm increases in VGCC protein levels as hiPSCs differentiate, as has been observed at the transcript level, as studies have shown that RNA levels may not be sufficient to predict protein levels (Liu et al., 2016). Additionally, one could study whether the isoforms identified in the present analyses possess different characteristics, such as electrophysiological properties, by generating hiPSC-derived neurons harboring transgenic overexpression of different isoforms and characterizing their properties. Finally, a key limitation is the lack of available antibodies that are specific to different VGCC isoforms. However, if such antibodies were to become available, one potential line of research that could prove fruitful would involve conducting mass spectrometry and identifying VGCC interaction partners. Altogether, more studies at the protein level are key to unraveling VGCC biology.

A key area of future study will be to understand whether these transcriptional changes change how VGCCs regulate gene expression, and – conversely – how VGCCs are regulated by changes in gene expression. **Chapter 3** will describe efforts to identify putative splicing regulators that regulate changes in the VGCC splicing of CACNA1C, CACNA2D1, and CACNA2D2 in the transition from hiPSCs to neurons. There will be a particular emphasis on CACNA1C because of its link to psychiatric conditions, as well as on CACNA2D1 and CACNA2D2 because alternative splicing of alpha-2-delta subunits has been under-

studied. Indeed, the potential to target these VGCC subunits with drugs such as CCBs and gabapentinoids heightened interest.

2.4.3 Conclusion

Altogether, these analyses provide insights into the nature of the changes that occur in VGCCs as electrophysiological activity emerges during hiPSC differentiation. The present analyses predict that, as hiPSCs differentiate into neurons, alternative splicing would likely affect several different types of VGCC properties, possibly including VGCC pharmacological properties, channel-related functions, and channel-independent functions. The findings have implications for identifying the potential roles of VGCCs during human development as well as for advancing the use of hiPSCs to experimentally dissect VGCC function for psychiatric drug development. As well as being a promising system for conducting studies, these data suggest that hiPSC-based systems are suitable for studying the mechanisms, e.g. switches in transcription and splicing factors, underlying changes in gene expression, and splicing across cell state transitions.

Chapter 3

Identification of Candidate RNA-Binding Proteins Regulating Splicing Events of Interest

3.1 Introduction

Alternative splicing, the process by which different mRNAs with differing exon and intron compositions are generated from pre-mRNA, is regulated by proteins that bind RNA. One class of proteins with well-characterized RNA-binding properties is RNA-binding proteins (RBPs). More generally, RBPs influence the maturation and fate of RNAs, affecting not only pre-mRNA splicing, but also cleavage, polyadenylation, RNA stability, RNA localization, RNA editing, and translation (Van Nostrand et al., 2020).

RNA splicing plays a key role in the regulation of human neural development. Several RBPs regulating neuronal phenotypes and neural development have been identified. These include FMRP, Smaug2, Nanos1, Rbfox, and Ptb/Ptbp1 (Fogel et al., 2012; Amadei et al., 2015; Doll and Broadie, 2015; Linares et al., 2015). For example, RBPs can regulate the activity of neural TFs, as is the case in Ptb/Ptbp1 regulation of Pbx1 during neuronal differentiation (Linares et al., 2015). Altogether, RBPs can be regulators of cellular identity, especially in the nervous system (Fogel et al., 2012; Amadei et al., 2015; Doll and Broadie, 2015; Linares et al., 2015).

Cell-type specific splicing patterns are regulated by the absolute and relative expression levels of different RBPs. For example, the knockdown of splicing factors can cause changes in alternative splicing patterns (Papasaikas et al., 2015; Tejedor et al., 2015). Additionally, the absolute and relative expression levels of different RBPs can also affect cell-type-

specific splicing patterns of VGCCs. Two examples of this concern the developmental control of Cav1.2 splicing by Fox proteins and PPT-binding protein (Tang et al., 2009; Tang et al., 2011).

Two exons in CACNA1C are regulated by members of the Fox family of RBPs (Tang et al., 2009). The differential inclusion of Exon 9* and Exon 33 affects the electrophysiological functions of Cav1.2 and shows opposite expression levels during cortical development. As development proceeds, the inclusion of Exon 9* in CACNA1C mRNA decreases, whereas inclusion of Exon 33 increases. Fox protein expression is upregulated in neurons in concordance with alterations in CACNA1C splicing, and both Exon 9* and Exon 33 possess Fox protein binding elements in adjacent introns. Accordingly, knockdown or overexpression of Fox proteins in cultured cells has opposing influences on the inclusion or exclusion of Exon 9* and Exon 33, with Fox protein downregulation increasing Exon 9* splicing and reducing Exon 33; and overexpression of Fox family members reducing Exon 9* and increasing Exon 33 splicing. Therefore, the absolute expression levels of RBPs in different cell types can fine-tune the properties of Cacna1c during neural differentiation (Tang et al., 2009).

Another instance of how differential expression of RBPs in different cell states can influence the splicing of VGCCs involves the PTB and nPTB splicing factors (Tang et al., 2011). PTB/PTBP1 regulates a switch from the splicing of Exon 8 to Exon 8a in CACNA1C mRNA. During neural differentiation, PTB/PTBP1 is downregulated, and this change is accompanied by an elevation in the levels of the PTB/PTBP1 neuronal homolog, nPTB. This change results in Exon 8a being largely absent at low levels in the developing mouse brain, but its expression is activated during neuronal differentiation as PTB levels are depleted.

These two examples suggest that both the absolute and relative levels of different cell-type specific splicing regulators can influence the splicing of VGCCs during neural development and cellular differentiation.

Many experimental and computational methods have been developed to identify RBPs as well as their binding sites (Li et al., 2014; Ramanathan et al., 2019). To complement experimental methods, various computational tools have been developed to identify putative regulators of alternative splicing using only sequence information. These methods rely on earlier studies that have identified the sequence preferences for different RBPs (Ray et al., 2009; Ray et al., 2013). For example, an in vitro experimental approach called RNA-

compete has been used to identify the sequence preferences of 207 different RBPs (Ray et al., 2013). Well-established bioinformatics methods rely on these experimentally determined sequence preferences to identify the presence of RBP binding motif occurrences. For example, FIMO scans specific sequences for individual matches to provided motifs, and HOMER can be used to predict motif sites across the entire human genome (Heinz et al., 2010; Grant et al., 2011).

As a final note, other proteins such as TFs have been identified as being able to bind RNA and potentially regulate their targets, but their interactions with RNA are not as well-characterized (Saldana-Meyer et al., 2019). It remains to be clarified how widespread the ability to interact with RNA is among TFs more generally. Moreover, it is still unclear whether TFs have the same sequence-specific preferences to bind RNA as they do for DNA (in other words, whether there is an RNA-binding motif for TFs) (Oksuz et al., 2022). Therefore, this Chapter will focus on RBPs, whose interactions with RNA are more well-characterized and whose motifs have been identified. However, TFs still possess key functions in the transcriptional regulation of VGCCs, which will be characterized in more detail in **Chapter 4**.

Chapter 2 identified changes in the splicing patterns of VGCC subunits as hiPSCs differentiate into neurons. Analyses of DEU and differential splice junction usage identified changes in CACNA1C, CACNA2D1, and CACNA2D2 because these subunits showed marked changes in splicing patterns across hiPSC differentiation (see **Chapter 2**). Additionally, the analyses focused on CACNA1C because of its link to psychiatric conditions and on CACNA2D1 and CACNA2D2 as the alternative splicing of auxiliary and particularly alpha-2-delta subunits has been less well-characterized. Moreover, the potential to target these VGCC subunits with drugs such as CCBs and gabapentinoids also heightened interest.

In this Chapter, well-established computational tools Find Individual Motif Occurrences (FIMO) and Hypergeometric Optimization of Motif EnRichment (HOMER) are applied to the HIPSCI dataset to identify putative regulators of differential splicing of VGCCs across hiPSC differentiation to neurons. Attention focused on three splicing events of interest concerning the alternative splicing of Exon 21 and 22 in CACNA1C; alternative splicing of Exon 28 in CACNA2D1; and alternative splicing of Exon 26 in CACNA2D2. Identifying candidate regulators will help guide future experiments to functionally validate regulators of the VGCC splicing events.

3.2 Methods

To identify candidate splicing regulators, searches for RBP motif occurrences at the selected VGCC regions delineated above were conducted using three different versions of searches from two different bioinformatics tools FIMO (Version 5.5.2) and HOMER (version 4.11)(Heinz et al., 2010; Grant et al., 2011). An outline of the strategy to identify RBPs and RBP binding sites is provided in **Figure 3.1**.

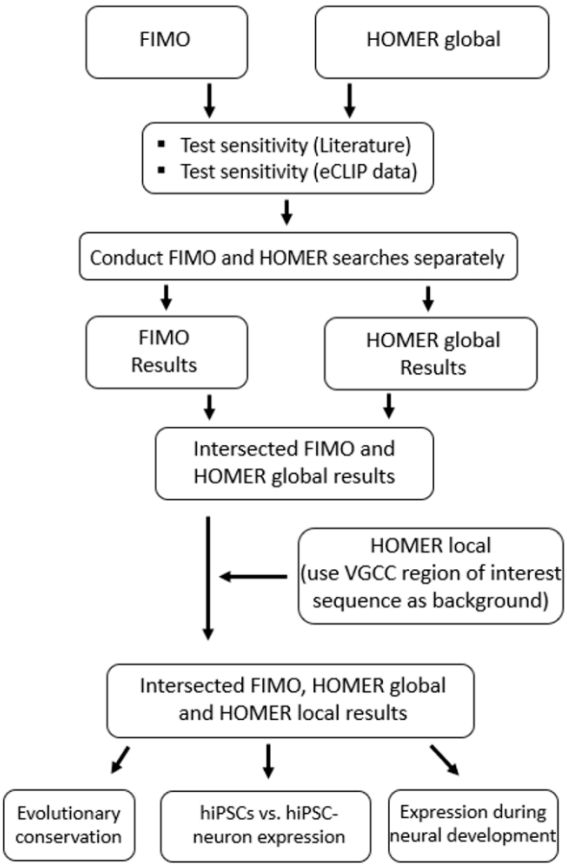


Figure 3.1: Outline of strategy to identify candidate regulators of VGCC splicing events of interest.

See main text for more detailed description of strategy. FIMO and HOMER are two bioinformatics tools that will be deployed to conduct RBP motif occurrences in this Chapter.

FIMO and HOMER are two bioinformatics tools that were deployed to conduct RBP motif occurrences in this Chapter. FIMO and HOMER represent well-established tools among motif-based bioinformatics tools. The original paper reporting FIMO has been cited over 3200 times (Grant et al, 2011). The original study describing HOMER has been cited over 9600 times (Heinz et al, 2010). Initially, the sensitivity of FIMO and HOMER was assessed by determining whether these tools can identify motif occurrences in loci known to be targets from the literature and at loci where eCLIP binding events

occur (more details in **3.3.1** and **Appendix II**). In the context of this chapter, sensitivity refers to the ability to detect motif occurrences. For example, if FIMO or HOMER possess sensitivity, they will possess a low number of false negative results. Next, VGCC regions of interest were separately scanned using FIMO and HOMER (for scans conducted separately, see **Appendix II**). First, the VGCC regions were scanned for RBP motifs using FIMO (Grant et al., 2011). Next, the instances of motifs from FIMO-identified RBPs were identified across the whole genome using HOMER (termed "HOMER global") (Heinz et al., 2010). After identifying RBP motif occurrences through separate searches, the results from FIMO and HOMER local were intersected (more details in **3.3.2**). HOMER global uses the human genome as an input – as the nucleotide frequencies of the VGCC regions of interest deviate from genome-wide nucleotide frequencies, it was surmised that a more stringent type of search using HOMER would involve using the actual genomic regions of interest to control for nucleotide frequency (an approach termed HOMER local) (Heinz et al., 2010) (see **3.3.3**). The results from HOMER local were intersected with those from FIMO and HOMER global to identify another list of motif instances. Finally, additional analyses of the identified RBP motif occurrences included 1) investigating the evolutionary conservation of the identified RBP sites (see **3.3.4**), 2) assessing the expression levels of the RBPs identified using both methods in hiPSCs, and hiPSC-derived neurons using DESEQ2 (see **3.3.5.1**), and 3) studying expression patterns of the identified RBPs in human neural development using the BrainSpan Atlas dataset (see **3.3.5.2**).

3.2.1 FIMO

Sequences of interest were scanned for individual matches to motifs using FIMO from the MEME Suite (Grant et al., 2011). FIMO searches sequences of interest for the occurrence of motifs that are provided to the software. The regions scanned were MAJIQ Exons 20 – 23 of CACNA1C; MAJIQ Exons 26 – 30 of CACNA2D1; and MAJIQ Exons 24 – 28 of CACNA2D2. The pipeline for FIMO is presented in **Figure 3.2**.

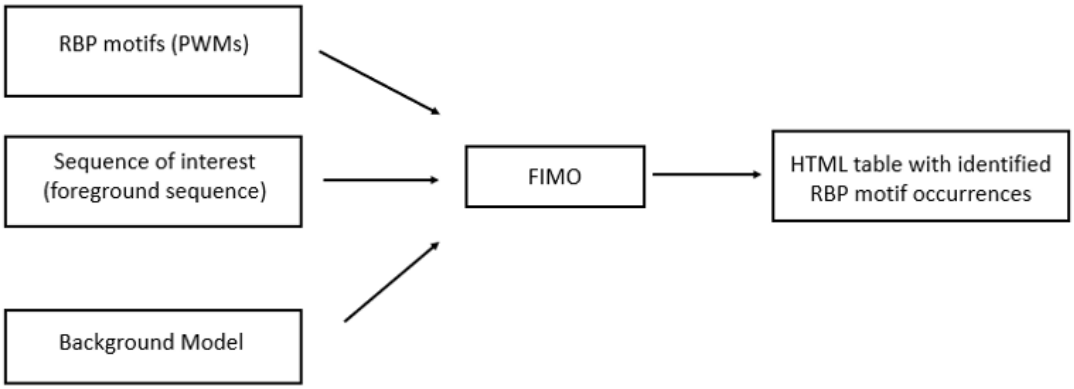


Figure 3.2: Pipeline for conducting searches for RBP motifs in given nucleotide sequences using FIMO.

FIMO requires three inputs: a database of RBP motifs (containing RBP position weight matrices or PWMs), a sequence of interest (in this case, the VGCC region of interest), and a background model. The principles of how FIMO works are described later in this Methods section. The output for FIMO is an HTML table with RBP motif occurrences. FIMO is a program developed by Noble and colleagues (Grant et al, 2011).

Regarding the RBP motifs input, the motif database containing position weight matrices (PWMs) for the different RBPs that was used was the CisBP-RNA database for Homo sapiens (Ray et al., 2009; Ray et al., 2013). A schematic of a PWM is provided in **Figure 3.3.**

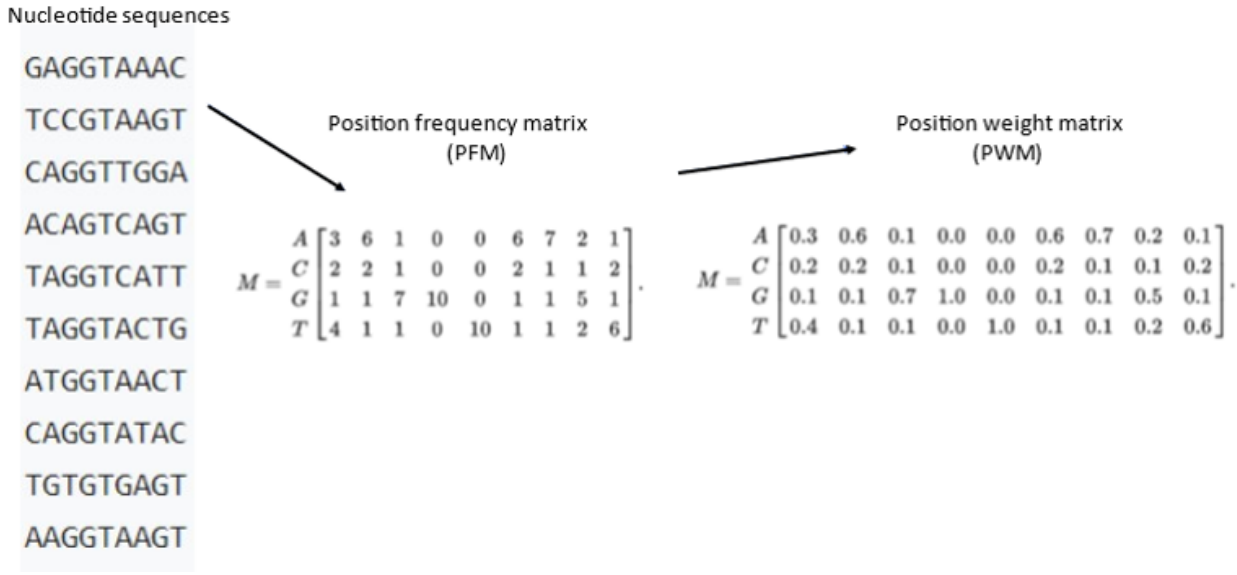


Figure 3.3: Definition of a PWM.

The first step in generating a PWM is to convert nucleotide sequences to a position frequency matrix. Each column in the position frequency matrix represents a position in the set of nucleotide sequences. Given the set of nucleotide sequences on the left, the corresponding position frequency matrix is the center matrix. From the position frequency matrix (PFM), a PWM (or a position probability matrix, PPM) can be derived by dividing the nucleotide count at each position (in the PFM) by the number of nucleotide sequences, creating normalized values (each column sums up to 1). A pseudocount can be added to the PFM to eliminate 0 values before generating the PWM (this was not done in the above example). The corresponding PWM or PPM is derived from the PFM (Note this figure is adapted from Guigo, 2003).

The RBP motifs in the CisBP-RNA database were derived using a method called RNAcompete (Ray et al., 2009). RNAcompete involves a competitive binding reaction in which an RBP is incubated with an excess of a complex pool of RNAs. The RBP is then retrieved and associated RNAs are investigated using computational analysis. Hughes and colleagues identified the sequence preferences for 207 different RBPs (Ray et al., 2013).

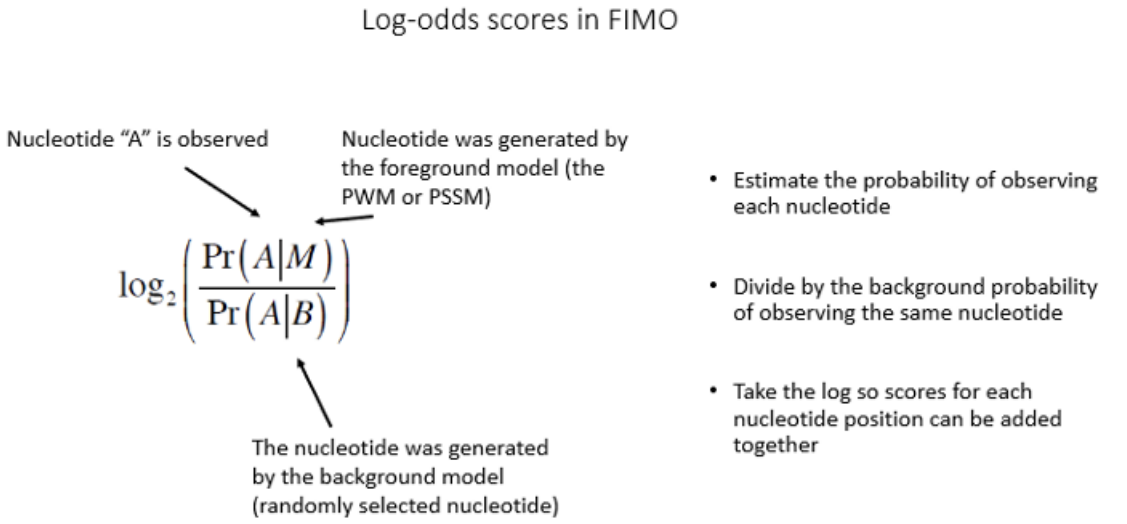
Regarding the sequence of interest, FIMO identifies motifs by comparing a foreground sequence with a background model (i.e. nucleotide frequencies in the input region of interest). For the foreground sequences, the regions scanned were Exons 20 – 23 of CACNA1C; Exons 26 – 30 of CACNA2D1; and Exons 24 – 28 of CACNA2D2. These regions were analyzed because they were identified as exhibiting either DEU or differential splice junction usage across hiPSC differentiation to neurons by DEXSEQ or MAJIQ, respectively (see **Chapter 2**). Given the different composition of intronic sequences to exonic sequences (exons are GC-rich compared to introns), intronic and exonic sequences spanning the exons flanking the exons of interest were scanned separately. For example, as Exon 21 and Exon 22 in CACNA1C are of interest, introns and exons spanning from Exon 20 and Exon 23 were separately scanned. As pre-mRNA is single-stranded, the reverse complement of each input sequence was not scanned.

The background model for each FIMO search was obtained by running the command fasta-get-markov on the nucleotide sequence comprising each search region. The background model comprises the nucleotide frequencies of A, G, C, and T in a given input nucleotide sequence.

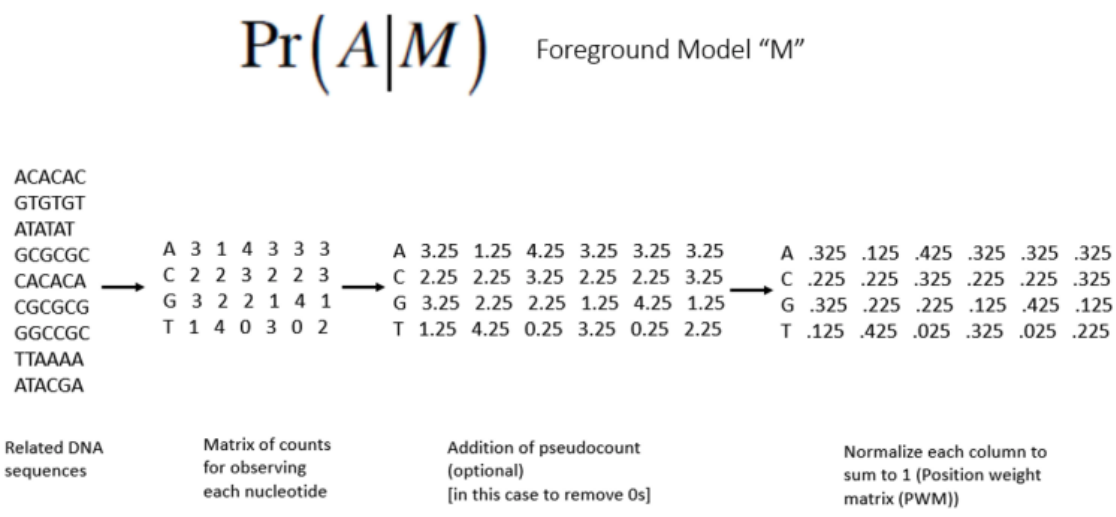
FIMO assesses whether the presence of a given set of nucleotides in an input foreground sequence is due to chance. FIMO scans DNA sequences with motifs described in the provided input PWMs (**Figure 3.4.**). The program calculates a log-odds ratio score for each nucleotide position and then sums the log-odds score based on the motif length (general principles regarding how log-odds scores are calculated and how scanning occurs are presented in **Figure 3.4.**) (Grant et al., 2011). The formula for how the log-odds scores are calculated indicates that a sequence's score for a given motif represents how well the sequence matches the motif while considering motif length and the expected background (**Figure 3.4.**). FIMO uses a dynamic programming algorithm to transform the log-odds scores into p-values (Grant et al., 2011). The p-value of motif occurrence is the probability of a random sequence of the same length of the motif matching the position of the sequence with a good or better score. Q-values are then generated from p-values for

a motif occurrence using the Benjamini-Hochberg method. The purpose of the statistical correction is to reduce the rate of false positives (i.e. the algorithm indicates that a certain position is a binding site when actually it is not). A q-value can be defined as the minimal false discovery rate for which a motif occurrence can be considered statistically meaningful (Grant et al., 2011). A q-value of 1.0 indicates that matches at that level of significance are due to chance, whereas a very low value indicates that the occurrence of a certain sequence is not likely to be due to chance. For this Chapter, a q-value of less than 0.05 was used.

Figure 3.4. presents how the log-odds scores are calculated**.**



(A)



(B)

$\Pr(A|B)$ Background Model

A	0.25
C	0.25
G	0.25
T	0.25

Above presumes equal probabilities for A, C, G, T
(not the case for VGCC regions of interest)

For FIMO, calculate the frequency in input nucleotide sequences but discard positions

(C)

Assembling PSSM (position-specific scoring matrix)

A	.325	.125	.425	.325	.325	.325
C	.225	.225	.325	.225	.225	.325
G	.325	.225	.225	.125	.425	.125
T	.125	.425	.025	.325	.025	.225

$\Pr(A|M)$

A	1.3	0.5	1.7	1.3	1.3	1.3
C	0.9	0.9	1.3	0.9	0.9	1.3
G	1.3	0.9	0.9	0.5	1.7	0.5
T	0.5	1.7	0.1	1.3	0.1	0.9

A	0.379	-1.000	0.766	0.379	0.379	0.379
C	-0.125	-0.125	0.379	-0.125	-0.125	0.379
G	0.379	-0.125	-0.125	-1.000	0.766	-1.000
T	-1.000	0.766	-3.322	0.379	-3.322	-0.125

A	0.25
C	0.25
G	0.25
T	0.25

$\Pr(A|B)$

$\dfrac{\Pr(A|M)}{\Pr(A|B)}$

$\log_2\left(\dfrac{\Pr(A|M)}{\Pr(A|B)}\right)$

(D)

Scanning for motif occurrences

DNA motif represented by PSSM

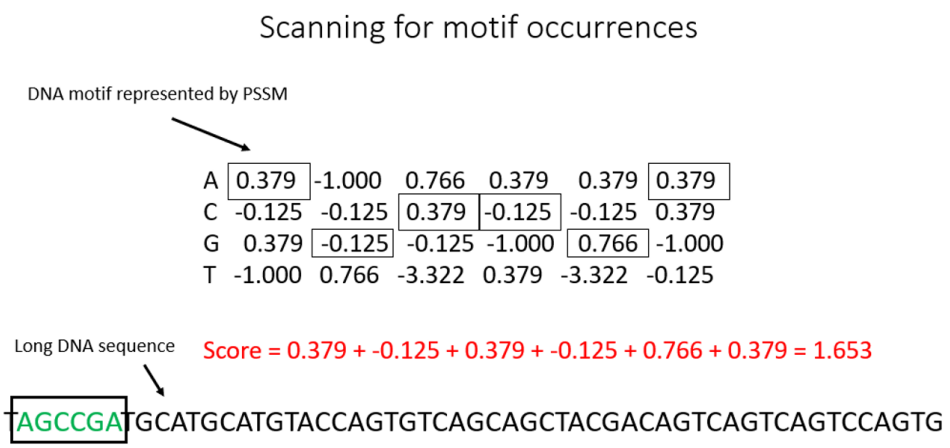
A	0.379	-1.000	0.766	0.379	0.379	0.379
C	-0.125	-0.125	0.379	-0.125	-0.125	0.379
G	0.379	-0.125	-0.125	-1.000	0.766	-1.000
T	-1.000	0.766	-3.322	0.379	-3.322	-0.125

Long DNA sequence

Score = -1.000 + -1.000 + -0.125 + -0.125 + -0.125 + -1.000 = -3.375

TAGCCGATGCATGCATGTACCAGTGTCAGCAGCTACGACAGTCAGTCAGTCCAGTG

(E)



(F)

Figure 3.4: Scanning for motif occurrences using FIMO (general principles of motif scanning).

A) Formula for calculating log-odds scores; estimate the probability of observing each nucleotide and divide by the background probability of observing the same nucleotide. B) Generating the foreground model "M": the figure presents how to generate a PWM from a set of related DNA sequences. C) Background model that has information for the nucleotide frequencies in each input nucleotide sequence: the figure presents nucleotide frequencies presuming equal probabilities for A, C, G, and T. For FIMO, a background model is generated where the frequency of each nucleotide in input nucleotide sequences is calculated and positions are discarded. Although what is presented in (C) is equal probabilities for A, C, G, and T in the figure; the actual FIMO searches used background models that accounted for the unique nucleotide frequency for each inputted genomic region for the searches. D) Assembling a position-specific score matrix (PSSM): the PWM is divided by the background model. The log (base 2) is calculated from the product. E) Scanning for motif occurrences in a long DNA sequence (panel 1): six nucleotides in the long DNA sequence are considered (TAGCCG) and the corresponding log-odds scores are added for those six positions. The negative value for the score indicates a poor match. F) Scanning for motif occurrences in a long DNA sequence (panel 2): the next six nucleotides in the long DNA sequence are considered (AGCCGA) and the corresponding log-odds scores are added for the next six positions. Note that the higher score value indicates a better match compared with the previous six nucleotides. In the example above (E and F), the maximum score for the motif ATATA/TAGA/C is 0.379 + 0.766 + 0.766 + 0.379 + 0.766 + 0.379 = 3.435. (adapted from https://www.csb.pitt.edu/ComputationalGenomics/Lectures/Lec21.pdf). Note: these figures are adapted from published material.

3.2.1.1 Sensitivity Analysis for FIMO

To assess the sensitivity of FIMO, searches in both known literature targets (see **Appendix II**) and experimental data were conducted. For searches using experimental data, enhanced cross-linking and immunoprecipitation (eCLIP) data for ELAVL1, KHDRBS1, TARDBP, PABPC4, and FXR1 were downloaded from ENCODE. eCLIP is a method for detecting RBP-RNA interactions (Van Nostrand et al, 2016). ELAVL1, KHDRBS1, TARDBP, PABPC4, and FXR1 were randomly selected from the RBPs that had eCLIP data available on the ENCODE website (https://www.encodeproject.org/encore-matrix/?type=Experiment&status=released&internal_tags=ENCORE). The corresponding bigwig files were uploaded into UCSC Genome Browser. 30 randomly selected eCLIP peaks were identified using both a random gene selector script ("randomgene.R"

script at www.github.com/adelosan521/chapter3) and a random number generator script ("randomnumber.R" script at www.github.com/adelosan521/chapter3). A list of randomly selected genes was derived and the genomic regions surrounding the eCLIP peaks (where present) were analyzed. If greater than three eCLIP peaks were present at a randomly selected gene, three eCLIP peaks were randomly selected from that gene using a random number generator. For each eCLIP peak, the eCLIP peak sequence, along with the 600 nucleotide sequence upstream and 600 nucleotide sequence downstream were inputted into FIMO. The justification for using approximately 1200 nucleotides flanking the eCLIP peak is because the size of the intron between CACNA1C Exons 20 and 21, which is in the vicinity of the CACNA1C splicing of interest, is also 1200 nucleotides.

3.2.2 HOMER

A second bioinformatics tool, HOMER, was used to identify motif occurrences across the human genome ("HOMER global", see section **3.2.2.1**) or at the VGCC regions of interest ("HOMER local", see section **3.2.2.2**) (Heinz et al, 2010). Both have similar downstream processing steps (Section **3.2.2.3**).

3.2.2.1 HOMER Global

First, HOMER was applied to identify motif instances across the whole genome (an approach henceforth termed "HOMER global") (Heinz et al., 2010; Cesana et al., 2018, Caldwell et al., 2020; Cheng et al., 2021; Zhou et al., 2022). The scanMotifGenomeWide.pl tool was used with hg38 as the input genome. In this approach, the entire human genome is used as the foreground. With the HOMER global approach, the background sequence is not an explicitly defined genomic sequence of interest, as background HOMER will randomly select sequences from the genome that are matched to the input sequence for GC% content. A number of studies have conducted RBP motif occurrences using whole genomes as the foreground in HOMER global, and a similar approach has been adopted for this Chapter (Cesana et al., 2018; Caldwell et al., 2020; Cheng et al., 2021; Zhou et al., 2022). The pipeline used for HOMER global is presented in **Figure 3.5**.

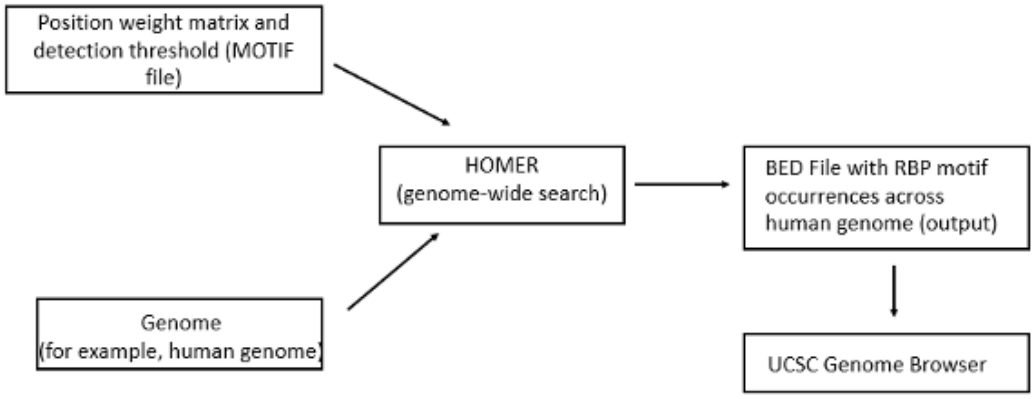


Figure 3.5: Pipeline for conducting searches for RBP motifs genome-wide using HOMER (HOMER global).

HOMER requires two inputs: a MOTIF file and a genome (in this case, human genome hg38). After HOMER conducts its genome-wide search a BED file with RBP motif occurrences across the human genome is generated that can be uploaded to the UCSC Genome Browser. The components of the MOTIF file are described in the next figure.

PWMs for the custom MOTIF files for the RBPs of interest were obtained from the CisBP-RNA database (http://cisbp-rna.ccbr.utoronto.ca) (Ray et al., 2009; Ray et al., 2013), similarly to FIMO. A diagram of a MOTIF file is presented in **Figure 3.6.**

RBP → BRUNOL4 Motif score → 7.54

0.0851	0.0851	0.176	0.654
0.0130	0.0130	0.777	0.197
0.0130	0.0130	0.0130	0.961
0.0130	0.0130	0.765	0.209
0.0130	0.0130	0.105	0.870
0.0130	0.0130	0.667	0.307
0.0831	0.0831	0.265	0.570

Figure 3.6: Representative example of a MOTIF file.

The figure shows the MOTIF file for BRUNOL4. The MOTIF file has two components: a motif score and the PWM. The motif score is used to set stringency (i.e. what HOMER counts as an RBP motif occurrence in the human genome). When calculating the motif score for HOMER global, the motif score considers the percentage of each nucleotide in the human genome. The different columns represent position-specific probabilities: the first column represents A, the second column represents C, the third column represents G, and the fourth column represents U. The table enclosed within the red box indicates the PWM for BRUNOL4.

For search parameters, a motif score was calculated by adding the log-odds probabilities for each nucleotide. While the calculation of the motif score has some similarity with calculating the log-odds scores for FIMO, the purpose of the HOMER motif score

is that it is inputted into HOMER to set stringency, reduce false positive hits (that is establish the parameters of the search). For example, for BRUNOL4, a stringent motif score for search parameters would be calculated by adding the log-odds probabilities for the most frequent nucleotides in each position of the PWM. For instance, considering the PWM in **Figure 3.6**, the log-odds probabilities for the sequence TGTGTGT would be added together to generate a putative motif score. Assuming nucleotide frequencies of 0.25 for each A, C, G, and T, the formula below would be used (as described in: http://homer.ucsd.edu/homer/motif/creatingCustomMotifs.html):

$$\text{Motif score} = \ln(\text{pT1}/0.25) + \ln(\text{pG2}/0.25) + \ln(\text{pT3}/0.25) + \ln(\text{pG4}/0.25) + \ln(\text{pT5}/0.25) + \ln(\text{pG6}/0.25) + \ln(\text{pT7}/0.25)$$

However, the GC genomic content of the human genome is 40.9% (Piovesan et al., 2019). Accounting for the estimated nucleotide frequencies in the human genome and substituting the PWM values in **Figure 3.6**, the revised formula for TGTGTGT below would be used:

$$\text{Motif score} = \ln(0.654/0.2955) + \ln(0.777/0.2045) + \ln(0.961/0.2955) + \ln(0.765/0.2045) + \ln(0.869/0.2955) + \ln(0.667/0.2045) + \ln(0.569/0.2955)$$

This results in a score of a value of 7.54, which would comprise a "perfect match" to the PWM. "Perfect match" motif scores were used to conduct global searches. A caveat to the approach of using the "perfect match" motif scores would be that HOMER would be likely not to accept a given sequence as a motif if there is a mismatch in one position, even though a motif imperfectly matching the PWM could still be biologically functional.

Because nucleotide frequency is used in calculating the motif score, the nucleotide frequency in the background can significantly influence results. This is relevant for genomic regions with nucleotide frequencies that deviate significantly from the estimated frequencies for A, G, C, and T in the human genome. The next section will discuss running HOMER while accounting for actual nucleotide frequency in the target regions of interest.

3.2.2.1.1 Sensitivity Analysis of HOMER

Similar methods to those used to assess the sensitivity of FIMO were used for HOMER. To assess the sensitivity of HOMER, searches in both known literature targets (see **Appendix II**) and experimental data were conducted. For searches using experimental data, enhanced cross-linking and immunoprecipitation (eCLIP) data for ELAVL1, KHDRBS1, TARDBP, PABPC4, and FXR1 were downloaded from ENCODE. The corresponding

bigwig files were uploaded into the UCSC Genome Browser.

HOMER global data for ELAVL1, KHDRBS1, TARDBP, PABPC4, and FXR1 were also uploaded into UCSC Genome Browser and visualized alongside the eCLIP data. The same 30 randomly selected eCLIP peaks identified using both a random gene selector script and a random number generator script for FIMO were analyzed. If greater than three eCLIP peaks were present at a randomly selected gene, three eCLIP peaks were randomly selected from that gene using a random number generator. Overlap between eCLIP peaks and motif occurrences was assessed at the genomic regions where randomly selected eCLIP peaks occurred. To allow for a comparison of FIMO and HOMER, the same PWMs were used for both FIMO and HOMER searches.

3.2.2.2 HOMER Local

Nucleotide frequencies can vary across the genome possibly affecting results. HOMER scans were conducted accounting for the nucleotide frequencies in the CACNA1C, CACNA2D1, and CACNA2D2 regions of interest (the pipeline is shown in **Figure 3.7**). The scanMotifGenomeWide.pl tool was used with the CACNA1C, CACNA2D1, and CACNA2D2 regions of interest as the custom input genomes.

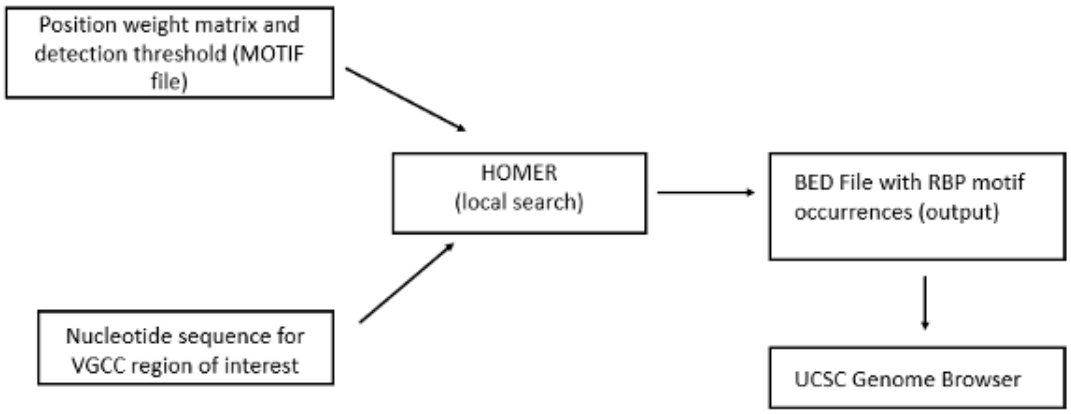


Figure 3.7: Pipeline for conducting searches for RBP motifs at VGCC regions of interest using HOMER.

HOMER requires two inputs: a MOTIF file and a genome (in this case, the nucleotide sequence for the VGCC region of interest). After HOMER conducts its search a BED file with RBP motif occurrences in the VGCC regions of interest is generated that can be uploaded to the UCSC Genome Browser.

The nucleotide frequencies in the CACNA1C, CACNA2D1, and CACNA2D2 regions of interest were calculated using a script that calculates nucleotide frequencies in a given nucleotide sequence ("nucleotide_calculator.R" script at www.github.com/adelosan521/chapter3).

Varying nucleotide frequencies for each of the VGCC regions of interest that deviated from the estimated genome-wide frequencies were noted. For the CACNA1C region of interest, the nucleotide frequencies were A (22%), T (27%), G (24%), and C (27%). For the CACNA2D1 region of interest, the nucleotide frequencies were A(32%), T(36%), G(16%), and C(16%). For the CACNA2D2 region of interest, the nucleotide frequencies were A(18%), T(23%), G(24%), and C(35%). Accordingly, HOMER local-specific MOTIF files for each candidate RBP identified from the intersections of FIMO and HOMER global that accounted for the local nucleotide frequencies were generated for each VGCC region. HOMER local searches were then conducted using the input nucleotide sequence of the VGCC region of interest and the new RBP MOTIF file. As noted previously, motif searching tools such as HOMER generally rely on comparisons in nucleotide frequency between foreground and background sequences. For background, HOMER randomly selects sequences from the genome matched for GC% content.

3.2.2.3 Common Downstream Processing Steps of the HOMER Pipeline for HOMER Global and HOMER Local

Sections **3.2.2.1** and **3.2.2.2** described the parameters inputted into the HOMER bioinformatics tool. In this section, the actual HOMER bioinformatics function will be described.

Searches for the binding motif occurrences of RBPs of interest in the entire human genome (HOMER global) and targeted searches (HOMER local) of CACNA1C, CACNA2D1, and CACNA2D2, were conducted via the scanMotifGenomeWide function in HOMER using the hg38 reference genome (Cesana et al., 2018) or the VGCC regions of interest as custom input genomes. HOMER employs ZOOPS scoring (zero or one occurrence per sequence) and hypergeometric enrichment calculations (or binomial) to determine motif enrichment (Heinz et al. 2010). "Stringent" motif scores were maintained throughout the analyses conducted in this chapter. For example, for the MOTIF file (PWM) featured in **Figure 3.6**, the most stringent highest scores were those in part calculated from summing the highest score for each nucleotide position. If searches failed, the RBP was discarded from subsequent scanMotifGenomeWide motif analyses. A search was defined as having failed if the HOMER search failed to yield a BED file with any genomic coordinates for a perfect motif occurrence.

The output from HOMER produces a BED file containing the binding motif occurrences (i.e. their genomic coordinates) across the human genome (HOMER global) or

the VGCC regions of interest (HOMER local). The appropriate output BED files were uploaded to UCSC Genome Browser for visualization.

3.2.3 Intersecting FIMO and HOMER Results

To identify the overlap between FIMO and HOMER global results, the FIMO results containing coordinates of motif occurrences were first converted into a BED file. Binding motif occurrences that were observed as co-occurring from the FIMO BED file and the HOMER global BED file were identified using "intersectBed" from the BEDTools package (version 2.29.2) to identify overlapping binding sites. The code for intersecting BED files is contained in the "intersection.sh" script at www.github.com/adelosan521/chapter3. The intersection was confirmed by manual inspection on UCSC Genome Browser. Next, the FIMO BED and HOMER global BED results were intersected with the HOMER local BED results.

3.2.4 Evolutionary Conservation Analysis

To infer the functionality of identified RBP binding sites, evolutionary conservation was assessed using phylogenetic P-values (PhyloP) (Pollard et al, 2010). If a particular region of the genome is conserved, it suggests potential biological function (Haerty and Ponting, 2014). PhyloP models sequence differences between species, per nucleotide site, under a null hypothesis of neutral evolution. PhyloP was identified as the best approach available for identifying individual sites under selection (Sadri et al., 2011). The absolute values of PhyloP scores represent -log(10)p-values based on alignment and under a null hypothesis of neutral evolution. A positive score indicates conservation and a negative score suggests acceleration. A positive score indicates that changes at that position may have a detrimental effect on the organism's fitness. In contrast, a negative PhyloP score indicates that a given nucleotide is evolving faster than expected under neutral evolution. The PhyloP scores were derived from a comparison of 100 vertebrate species and 17 primate species (listed in **Supplementary Table 2.1** and **Supplementary Table 2.2**). The 100 vertebrate species comparison allows one to search for deeply conserved elements whereas the 17 primate species comparison enables one to study primate-specific evolutionary changes. If a site is conserved in both the 100 vertebrate species and 17 primate species comparisons it is hypothesized to possess biological function. If a site shows conservation among 100 vertebrate species but not among 17 primate species, it is possible that the site is

functional in vertebrates more broadly but has lost constraint and/or function in the primate lineage. Conversely, if a site shows no conservation among 100 vertebrate species but conservation or acceleration in the primate species, it is possible that the site gained function in the primate lineage.

To examine the conservation of candidate RBP binding sites, random sampling of the intronic sequence where the RBP binding sites were located was conducted by sampling the same number of nucleotides as the size of the RBP binding site and nucleotides of the same composition as the respective binding site. The code that was used to achieve this is in the "sampling.R" script at www.github.com/adelosan521/chapter3. For example, in the case of BRUNOL4 and BRUNOL5, because the binding site was GTGTGTGT, 4Gs, and 4Ts were randomly sampled from the intronic region. The median of the PhyloP conservation scores from the eight sampled nucleotides was taken and this process was repeated. This iterative process was repeated 1000 times and generated a distribution of scores. The 95% confidence interval was also obtained (2.5% to 97.5%). The median PhyloP value for the RBP binding site was compared to the distribution of "simulated" scores and a 95% confidence interval. The sampling script describes the code for the BRUNOL4 and BRUNOL5 binding site within the intron between CACNA1C Exon 20 and 21 but the script was adapted accordingly to sample other RBP binding sites within the other introns.

3.2.5 Gene Expression Analysis

To gain insights into the possible functions of the identified RBPs, gene expression analyses were conducted. This included analyses of the gene expression levels across hiPSC differentiation. Higher expression in hiPSC-derived neurons might suggest function in human neural cells. The gene expression analyses also involved identifying when the RBPs were expressed in fetal and postnatal brain development (BrainSpan Atlas data). Understanding when RBPs were identified are expressed in the brain might provide clarity into possible function. For example, if a given RBP was expressed more highly fetal relative to postnatal brain development, it could have a function in fetal development.

3.2.5.1 DESEQ2

DESEQ2 was employed for differential gene expression analyses of the HIPSCI data, as described in **Chapter 2**.

3.2.5.2 Gene Expression Across Human Neural Development (Inspection of BrainSpan Atlas data)

As hiPSC-derived neurons share similarities with fetal rather than adult-stage neurons, attempts were made to study the expression of identified RBPs in fetal brain tissue (Brennand et al., 2015). The BrainSpan Atlas is a resource for studying gene expression changes during both fetal and postnatal human brain development. Developmental transcriptome data were inspected (www.brainspan.org) comprising RNA-seq data capturing 16 regions of the brain across fetal and postnatal human brain development (see **Supplementary Table 2.3.** for more information about the different brain regions sampled). For inspection of BrainSpan Atlas data, the heatmap was generated using the graphical user interface on the website. For statistical analyses of gene expression of fetal vs postnatal development as well as early fetal vs late fetal development, processed RNA-seq files were downloaded from www.brainspan.org. The expression levels of RBPs of interest in prenatal and postnatal frontal cortex samples, as well as early fetal and late fetal frontal cortex samples, were compared statistically (see section 3.2.**6**) The code for the statistical analyses is in the "brainspan_stats.R" (fetal versus postnatal) and "brainspan_stats2.R" (early fetal versus late fetal) scripts at www.github.com/adelosan521/chapter3.

3.2.6 Statistics

To compare FIMO and HOMER sensitivity, Fisher's exact test was used. Fisher's exact test was chosen given the low counts observed for motif occurrences at eCLIP peaks (as noted later, FIMO failed to detect any motif occurrences in 30 randomly selected eCLIP peaks from FXR1 data). The code for this comparison is featured at the "fisher.R" script at www.github.com/adelosan521/chapter3.

To quantify transcript levels for RBPs from the BrainSpan Atlas for comparing fetal versus postnatal stages as well as early fetal versus late fetal stages; first, normality was checked using the Shapiro-Wilk test. Because not all datasets displayed a normal distribution, the median and interquartile range were calculated. The Mann-Whitney U test was used to determine if there was a significant difference between the groups. If the p-value was less than 0.05, the difference between the groups was considered significant. The code for the above comparisons is featured in the "brainspan_stats.R" and "brainspan_stats2.R" scripts at www.github.com/adelosan521/chapter3.

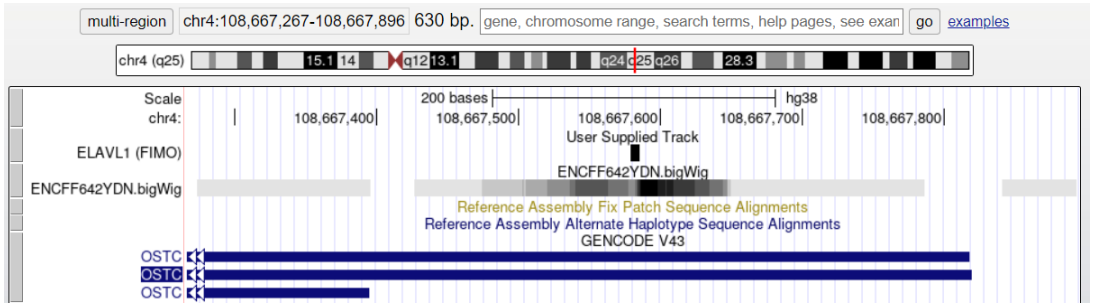
3.3 Results

To derive a list of potential regulators of splicing events in VGCCs, RBP binding motifs were first identified in the close vicinity of the VGCC regions of interest using two different bioinformatics tools separately, and the results were then intersected. The evolutionary conservation of the identified RBP binding sites was also assessed. Subsequently, the gene expression levels of the identified RBPs in hiPSCs and hiPSC-derived neurons were compared. Their expression patterns in adult tissues and during human neural development were analyzed.

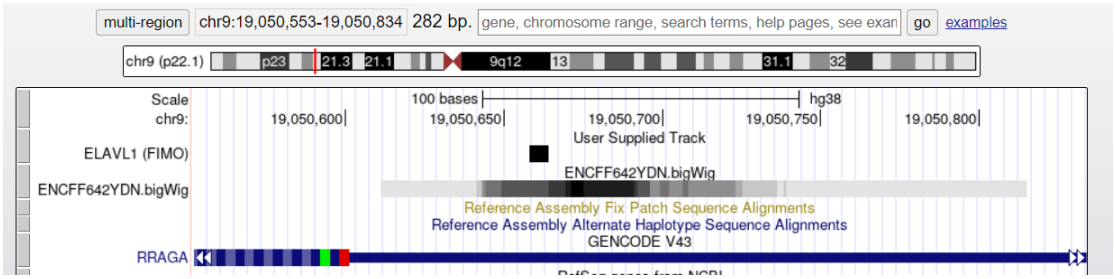
3.3.1 Assessing Sensitivity

The sensitivity of both FIMO and HOMER to detect motif occurrences of RBPs in known RBP targets was assessed below (see also **Appendix II**). First, the capacity of FIMO to detect RBP motif occurrences in known targets from the literature was determined. FIMO successfully identified motif occurrences for BRUNOL4 in three out of four known BRUNOL4 targets (**Appendix II**). HOMER successfully identified the appropriate motif occurrences in five out of five known BRUNOL4 targets and one out of one ELAVL2 targets (**Appendix II**; **Supplementary Figure 2.1**, **Supplementary Figure 2.2**). These findings suggest FIMO and HOMER are capable of identifying motif occurrences across entire loci in RBP targets that are well-established in the literature.

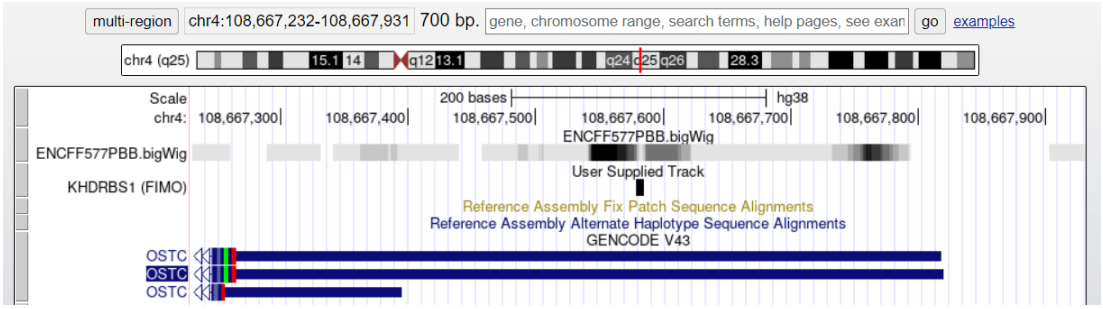
The capacity of these tools to detect motif occurrences in specific regions where RBP-RNA interactions have been experimentally determined to occur was examined next. eCLIP (experiments for determining RBP-RNA interactions) datasets were inspected. Attempts were made to determine whether FIMO and HOMER could detect motif occurrences at 30 randomly selected eCLIP peaks (see **Methods** for the description of how sensitivity analyses were conducted). The eCLIP data from datasets for 5 RBPs; ELAVL1, KHDRBS1, TARDBP, PABPC4, and FXR1 were inspected. Representative images of FIMO motif occurrences coinciding with eCLIP peaks are presented in **Figure 3.8** and HOMER motif occurrences coinciding with eCLIP peaks are presented in **Figure 3.9**.



(A)



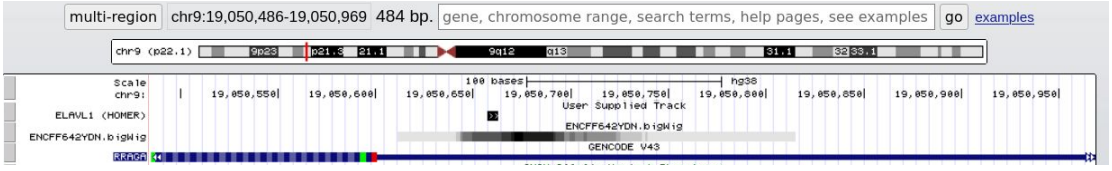
(B)



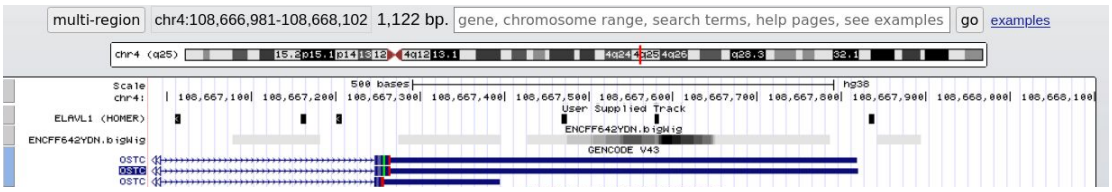
(C)

Figure 3.8: Detection of FIMO motif occurrences in eCLIP datasets.

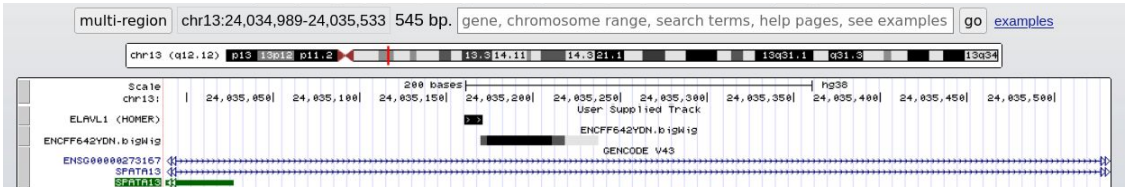
Snapshot of UCSC Genome Browser. Overlay of FIMO results for ELAVL1 (the ELAVL1 (FIMO) track) and ELAVL1 eCLIP data (ENCFF642YDN.bigWig track) at the (A) OSTC and (B) RRAGA loci. Overlay of FIMO results for KHDRBS1 (the KHDRBS1 (FIMO) track) and KHDRBS1 eCLIP data at the (C) OSTC locus. The black bar for FIMO tracks indicates binding motif occurrences. Gray to black bar (black bar represents the peak of the eCLIP signal) in the eCLIP data track indicates experimentally determined RBP-RNA interaction. eCLIP data were obtained from ENCODE and are from K562 cells.



(A)



(B)



(C)

Figure 3.9: Detection of HOMER global motif occurrences in eCLIP datasets

Snapshot of UCSC Genome Browser. Overlay of HOMER global results for ELAVL1 (the ELAVL1 (HOMER) track) and ELAVL1 eCLIP data (ENCFF642YDN.bigWig track) at (A) RRAGA locus; (B) OSTC locus; and (C) SPATA13. (ENCFF642YDN.bigWig Black bar in ELAVL1 (HOMER) track indicates binding motif occurrence. The gray to black bar (black bar represents the peak of the eCLIP signal) in the eCLIP data track indicates experimentally determined RBP-RNA interaction. eCLIP data were obtained from ENCODE and are from K562 cells.

A summary of the results for testing sensitivity is presented in **Table 3.1**. Whether there was a statistical difference between FIMO and HOMER was tested using Fisher's exact test. This statistical test was deemed suitable given that some frequencies were low (in the case of FXR1 with FIMO, it was zero; KHDRBS1 with FIMO was just two). In four out of five RBPs (except ELAVL1), FIMO and HOMER did not show differences. In summary, FIMO and HOMER can detect motif occurrences at some, but not most, known regions where eCLIP peaks occur, suggesting a predisposition for false negatives.

RBP	FIMO frequency	HOMER frequency	FishersExact	P-value
ELAVL1	6/30	20/30	0.130	0.000574
KHDRBS1	2/30	5/30	0.363	0.424
TARDBP	5/30	11/30	0.352	0.143
PABPC4	7/30	11/30	0.531	0.399
FXR1	0/30	1/30	0	1

Table 3.1: Summary of sensitivity analyses for FIMO and HOMER.

Searches were conducted to test whether FIMO or HOMER would detect motif occurrences at 30 randomly selected eCLIP peaks. For example, FIMO identified 6 ELAVL1 motif occurrences out of 30 eCLIP peaks; whereas HOMER identified 20 ELAVL1 motif occurrences out of 30 eCLIP peaks. First column: RBP. Second column: frequency at which FIMO identified motif occurrences in random eCLIP peaks. Third column: frequency at which HOMER identified motif occurrences in random eCLIP peaks. Whether there was a difference between FIMO and HOMER was tested using Fisher's exact test (value provided in fourth column). Fifth column: p-value. Alpha value is set at 0.05.

Altogether, inspecting eCLIP data confirmed that FIMO and HOMER can detect motif occurrences at genomic regions where eCLIP peaks were observed, suggesting their suitability for identifying RBP binding motif occurrences. Both overlap and non-overlap between the two methods was also noted. For ELAVL1, FIMO and HOMER identified 6 overlapping motif occurrences (where FIMO identified 6 motif occurrences and HOMER detected 20), and 14 events specific to HOMER global. For KHDRBS1, FIMO and HOMER identified no overlapping motif occurrences, with 4 motif occurrences specific

to HOMER and 2 events specific to FIMO. For TARDBP, FIMO and HOMER identified 5 overlapping motif occurrences (where FIMO identified 5 motif occurrences and HOMER detected 11), and 6 events specific to HOMER global. For PABPC4, FIMO and HOMER identified 6 overlapping motif occurrences (where FIMO detected 7 and HOMER detected 11), with one motif occurrence specific to FIMO and 5 specific to HOMER. For FXR1, FIMO and HOMER identified zero overlapping motif occurrences (where FIMO detected zero motif occurrences and HOMER detected one).

Finally, the frequency of HOMER binding motif occurrences in the above-searched genes raised questions about the prevalence of RBP binding motif occurrences more generally. To assess specificity, HOMER global searches for BRUNOL4 and ELAVL2 binding sites in ten randomly selected genes were conducted using a random gene selection script. In this context, a specific bioinformatics tool will not give many false positive results. HOMER global identified BRUNOL4 motif occurrences in seven out of ten random genes, and ELAVL2 motif occurrences for six out of ten random genes (**Appendix II**). In all genes where motif occurrences were found, the number of HOMER motif occurrences outnumbered the number of eCLIP peaks. These findings indicate that RBP binding sites are common in the genome and/or a possible predisposition of HOMER global to produce false positive results.

Therefore, instances where both methods failed to identify motif occurrences where eCLIP peaks occurred, as well as identifying HOMER motif occurrences where eCLIP peaks did not occur, indicate caution is warranted in interpreting results (see **Discussion**).

3.3.2 Intersection of FIMO and HOMER Global Results

To derive a first list of candidate regulators, search hits from two different bioinformatics tools –FIMO and HOMER– were intersected (Heinz et al., 2010; Grant et al., 2011). As FIMO and HOMER global rely on distinct motif discovery algorithms, it was surmised that intersecting FIMO and HOMER global results would generate more robust findings as such findings would arise from two independent methods. Before intersecting results, first, searches using FIMO and HOMER global were conducted separately. FIMO and HOMER global each identified motif occurrences at the VGCC regions of interest (as FIMO and HOMER showed reduced sensitivity, results for the separately conducted searches are provided in **Appendix II**, see **Supplementary Figures 2.3 -2.6**) . Motif occurrence results that overlapped between both FIMO and HOMER were identified. Both FIMO and

HOMER global found binding motif occurrences in CACNA1C (six RBPs), CACNA2D1 (one RBP), and CACNA2D2 (14 RBPs), whose positions are presented in **Figure 3.10**. Tables of the binding motif occurrences are presented in **Tables 3.2-3.4**.

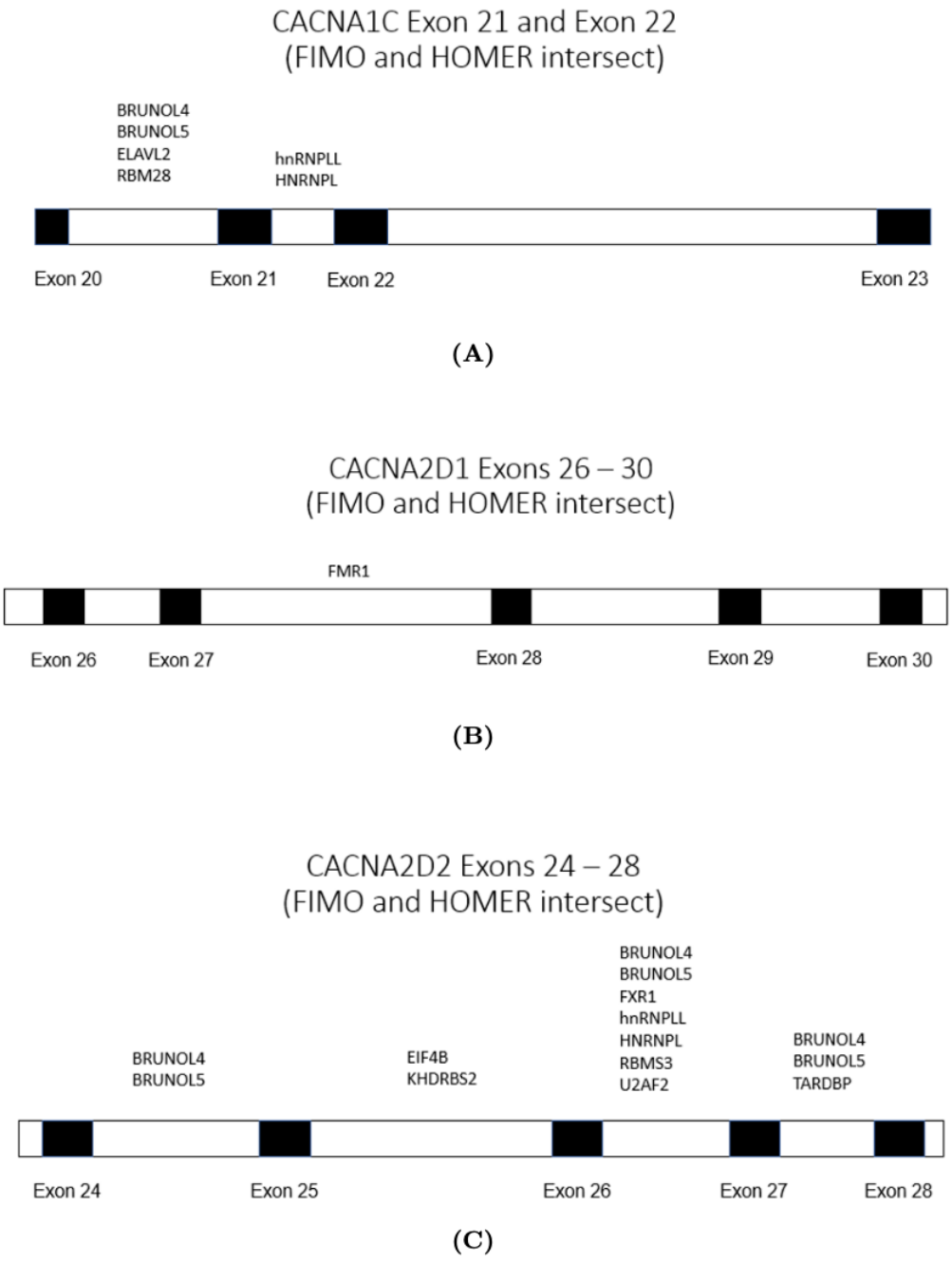


Figure 3.10: RBP motif occurrences identified by both FIMO and HOMER global in the CACNA1C, CACNA2D1, and CACNA2D2 regions of interest.

RBP motif occurrences in the vicinity of (A) CACNA1C Exon 21 and Exon 22; (B) CACNA2D1 Exon 28; and (C) CACNA2D2 Exon 26. Black boxes indicate exons; white boxes indicate introns. Exons and introns are not drawn to scale.

Intron between Exons 20 and 21			
chr12	2597212	2597225	ELAVL2
chr12	2596425	2596431	RBM28
chr12	2597053	2597059	BRUNOL5
chr12	2597055	2597061	BRUNOL5
chr12	2597053	2597059	BRUNOL4
chr12	2597055	2597061	BRUNOL4
Intron between Exons 21 and 22			
chr12	2597365	2597371	HNRNPL

Table 3.2: Genomic coordinates of RBP motif occurrences identified by FIMO and HOMER global in the vicinity of CACNA1C Exons 21 and 22.

First column: chromosome number (e.g. chr12); second column: starting coordinate of binding motif occurrence (e.g. 2597212); third column: end coordinate of binding motif occurrence (e.g. 2597225); fourth column: name of RBP.

Intron between Exons 27 and 28			
chr7	82004971	82004977	FMR1

Table 3.3: Genomic coordinates of RBP motif occurrences identified by FIMO and HOMER global in the vicinity of CACNA2D1 exon 28.

See the legend to **Table 3.2** for interpretation.

Intron between Exons 24 and 25			
chr3	50375347	50375353	BRUNOL5
Intron between Exons 25 and 26			
chr3	50373186	50373192	KHDRBS2
chr3	50374424	50374430	EIF4B
Intron between Exons 26 and 27			
chr3	50371826	50371833	FXR1
chr3	50370739	50370745	RBMS3
chr3	50370737	50370743	RBMS3
chr3	50370794	50370800	U2AF2
chr3	50370726	50370732	BRUNOL5
chr3	50370712	50370718	BRUNOL5
chr3	50370726	50370732	BRUNOL4
chr3	50370712	50370718	BRUNOL4
chr3	50371777	50371783	U2AF2
chr3	50371363	50371369	HNRNPL
chr3	50371361	50371367	HNRNPL
Intron between Exons 27 and 28			
chr3	50369264	50369271	TARDBP
chr3	50368950	50368957	TARDBP
chr3	50370230	50370236	BRUNOL5
chr3	50370230	50370236	BRUNOL4

Table 3.4: Genomic coordinates of RBP motif occurrences identified by FIMO and HOMER global in the vicinity of CACNA2D2 exon 26.

See the legend to **Table 3.2** for interpretation.

3.3.3 HOMER Searches Using Targeted Foreground Sequence (HOMER Local)

For a more stringent analysis, an additional search type to the motif occurrence analyses was added. HOMER searches were conducted in a more local fashion where HOMER did not use the whole human genome as a foreground; rather, the CACNA1C, CACNA2D1, and CACNA2D2 sequences were used as custom "genomes" for the HOMER input (an approach termed "HOMER local").

The HOMER local approach was considered because the nucleotide frequency of the input genome could influence search results. Indeed, the nucleotide frequencies at the VGCC regions of interest diverged significantly from the estimated genome-wide nucleotide frequencies (see **Methods**).

By applying the HOMER local approach to all RBPs identified from the FIMO and HOMER global intersections, HOMER local only identified a single binding motif occurrence for BRUNOL4 and BRUNOL5 in the intron between Exon 20 and Exon 21 of CACNA1C (**Table 3.5; Figure 3.11**). No HOMER local hits were found in CACNA2D1 and CACNA2D2, indicating possible issues with HOMER local sensitivity or false positive results with the other methods (see **Discussion**). Given that the BRUNOL4 and BRUNOL5 binding motif occurrence in CACNA1C was identified by all three search types (FIMO, HOMER global, and HOMER local), this binding site was considered a potentially more robust observation.

chr12	2597055	2597061	BRUNOL4
chr12	2597055	2597061	BRUNOL5

Table 3.5: Genomic coordinates of BRUNOL4 and BRUNOL5 motif occurrence identified by FIMO, HOMER global, and HOMER local in the vicinity of CACNA1C Exon 21 and Exon 22.

See the legend to **Table 3.2** for interpretation.

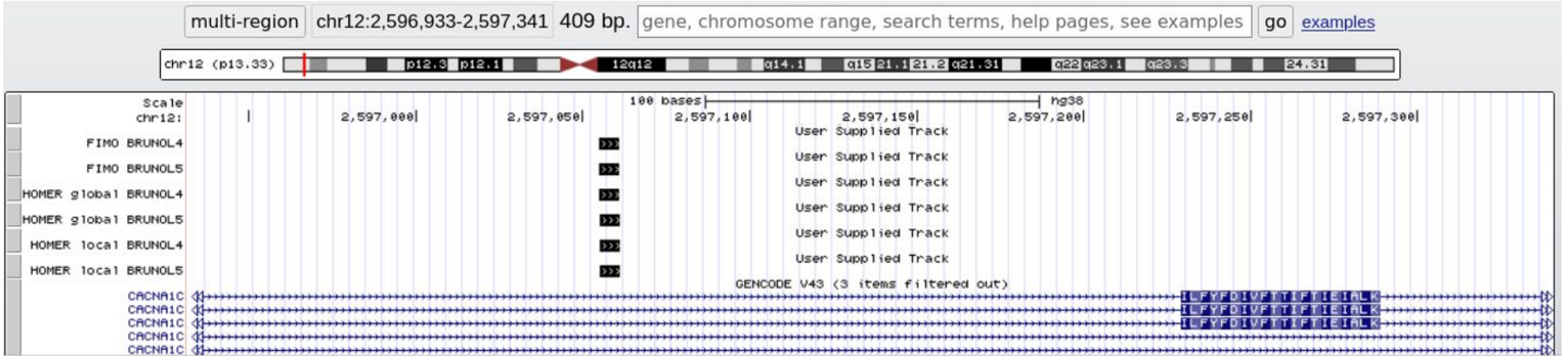


Figure 3.11: Identification of BRUNOL4 and BRUNOL5 binding motif occurrence by FIMO, HOMER global, and HOMER local in the vicinity of CACNA1C Exon 21 and 22.

The thin black bars indicate binding motif occurrences. The thick blue boxes indicate Exon 21 and Exon 22. The figure presents FIMO, HOMER global, and HOMER local BED file tracks for BRUNOL4 and BRUNOL5.

3.3.4 Evolutionary Conservation of RBP Motif Occurrences

Evolutionary conservation of the sequences across different species at the binding site could provide possible insights into whether identified binding sites are functional (Haerty and Ponting, 2014). The evolutionary conservation of the binding sites identified by double intersection of FIMO and HOMER global; as well as triple intersection of FIMO, HOMER global and HOMER local, was assessed next using PhyloP (Pollard et al, 2010). Evolutionary conservation scores from background sequences from the same genomic location as the binding site were examined and compared to the evolutionary conservation score of the candidate RBP binding site.

The results for analyses across 100 vertebrate species and 17 primate species (see **Supplementary Table 2.1** and **Supplementary Table 2.2**) for the RBPs identified by FIMO and HOMER Global are presented in **Table 3.6** and **Table 3.7**. For the 100 vertebrate species comparison (**Table 3.6**), RBP binding motif occurrences with PhyloP median scores higher than 97.5% of randomly resampled scores in the local intronic region were: RBMS3 in the intron between CACNA2D2 Exon 26 and Exon 27; U2AF2 in the intron between CACNA2D2 Exon 26 and Exon 27; BRUNOL4 and BRUNOL5 in the intron between CACNA2D2 Exon 26 and Exon 27, and BRUNOL4 and BRUNOL5 in the intron between CACNA2D2 Exon 27 and Exon 28. For the 17 primate species analyses (**Table 3.7**), the binding sites with higher than 97.5 percentile PhyloP scores in the 100 vertebrate species comparison no longer had higher than 97.5 percentile PhyloP scores. Therefore, for those sites, the null hypothesis, which was that the RBP median PhyloP score is not different than a randomly sampled intronic median PhyloP score, could not be rejected. However, in the CACNA2D2 intron between Exon 26 and Exon 27, there was an HNRNPL binding motif occurrence with a PhyloP median score less than 2.5% of random resampled scores, suggesting this site was undergoing accelerated evolution.

RBP	Chromosome	Start	End	Motif match	Binding site median	Median of intronic region (2.5 percentile value, 97.5 percentile value) [95% CI]
CACNA1C intron between Exons 20 and 21						
RBM28	Chr12	2596425	2596431	GAGTAGA	0.463	0.078 (-0.445, 0.698)
BRUNOL4/5	Chr12	2597053	2597059	TGTGTGT	-0.1	0.078 (-0.432, 0.568)
BRUNOL4/5	Chr12	2597055	2597061	TGTGTGT	0.033	0.078 (-0.432, 0.568)
ELAVL2	Chr12	2597212	2597225	TTCCTTTTTGTTTT	-0.0555	N/A
CACNA1C intron between Exons 21 and 22						
HNRNPL	Chr12	2597365	2597371	ACACACA	-0.161	0.014 (-0.1865, -0.678)
CACNA2D1 intron between Exons 27 and 28						
FMR1	Chr7	82004971	82004977	GGACAGG	-0.059	0.02 (-0.49, 0.69715)
CACNA2D2 intron between Exons 24 and 25						
BRUNOL4/5	Chr3	50375347	50375353	TGTGTGT	0.185	0.208 (0.185, 0.208)
CACNA2D2 intron between Exons 25 and 26						
KHDRBS2	Chr3	50373186	50373192	AATAAAA	0.153	0.186 (0.137, 0.208)
EIF4B	Chr3	50374424	50374430	CTTGGAA	0.208	0.186 (0.137, 0.208)
CACNA2D2 intron between Exons 26 and 27						
FXR1	Chr3	50371826	50371833	AATGACAA	0.137	0.137 (0.136, 0.16)
RBMS3	**Chr3**	**50370739**	**50370745**	**ATATATA**	**0.277**	**0.137 (0.136, 0.16)**
RBMS3	**Chr3**	**50370737**	**50370743**	**ATATATA**	**0.277**	**0.137 (0.136, 0.16)**
U2AF2	**Chr3**	**50370794**	**50370800**	**TTTTTTC**	**0.311**	**0.137 (0.136, 0.16)**
BRUNOL4/5	Chr3	50370726	50370732	TGTGTGT	0.153	0.137 (0.136, 0.16)
BRUNOL4/5	**Chr3**	**50370712**	**50370718**	**TGTGTGT**	**0.299**	**0.137 (0.136, 0.16)**
U2AF2	**Chr3**	**50371777**	**50371783**	**TTTTTCC**	**0.311**	**0.137 (0.136, 0.16)**
HNRNPL	Chr3	50371363	50371369	ACACACA	0.137	0.137 (0.136, 0.16)
HNRNPL	Chr3	50371361	50371367	ACACACA	0.137	0.137 (0.136, 0.16)
CACNA2D2 intron between Exons 27 and 28						
TARDBP	Chr3	50369264	50369271	TGAATGAA	0.34	0.274 (-0.222, 0.375)
TARDBP	Chr3	50368950	50368957	TGAATGAG	0.261	0.274 (-0.222, 0.375)
BRUNOL4/5	**Chr3**	**50370230**	**50370236**	**TGTGTGT**	**0.75**	**0.274 (-0.389, 0.437)**

Table 3.6: Evolutionary conservation of RBP motif occurrences identified by FIMO and HOMER global relative to the intronic regions where they are located in (comparison of 100 vertebrate species).

First column: RBP; second column: chromosome; third column: start of the binding site coordinate; fourth column: end of the binding site coordinate; fifth column: actual nucleotide sequence FIMO and HOMER global identified as a match; sixth column: median PhyloP score for the respective RBP binding site. For the seventh, eighth, and ninth columns: the median of PhyloP scores for randomly sampled nucleotides (same composition and size as the corresponding RBP site being tested) was obtained 1,000 times for the intron where the RBP motif occurrence was located in. This generated a distribution of "simulated" scores. The seventh table column corresponds to the median "simulated" score. The eighth column corresponds to the 2.5 percentile of the simulated scores; the ninth column corresponds to the 97.5 percentile of the simulated scores. The phyloP scores were conservation scores for multiple alignments of 99 vertebrate genomes to the human genome (comparison of 100 species). Bold indicates evolutionary conservation relative to local intronic sequence (i.e. greater than 97.5% of simulated scores). The ELAVL2 binding site near CACNA1C Exon 21 was located within the 50 nucleotides that were truncated at the 5'- and 3'- end of the intronic region for the random resampling analysis. The median value, the 2.5 percentile value, and the 97.5 percentile value are indicated as N/A accordingly.

RBP	Chromosome	Start	End	Motif match	Binding site median	Median of intronic region (2.5 percentile value, 97.5 percentile value) [95% CI]
CACNA1C intron between Exons 20 and 21						
RBM28	Chr12	2596425	2596431	GAGTAGA	-0.281	0.373 (-0.339, 0.56)
BRUNOL4/5	Chr12	2597053	2597059	TGTGTGT	0.333	0.373 (-0.319, 0.541)
BRUNOL4/5	Chr12	2597055	2597061	TGTGTGT	0.366	0.373 (-0.310, 0.541)
ELAVL2	Chr12	2597212	2597225	TTCCTTTTTGTTTT	0.1695	N/A
CACNA1C intron between Exons 21 and 22						
HNRNPL	Chr12	2597365	2597371	ACACACA	-0.247	0.497 (-0.296, 0.497)
CACNA2D1 intron between Exons 27 and 28						
FMR1	Chr7	82004971	82004977	GGACAGG	0.618	0.567 (-0.202, 0.686)
CACNA2D2 intron between Exons 24 and 25						
BRUNOL4/5	Chr3	50375347	50375353	TGTGTGT	0.676	0.599 (-0.203, 0.676)
CACNA2D2 intron between Exons 25 and 26						
KHDRBS2	Chr3	50373186	50373192	AATAAAA	0.549	0.442 (-0.272, 0.618)
EIF4B	Chr3	50374424	50374430	CTTGGAA	0.609	0.442 (-0.253, 0.618)
CACNA2D2 intron between Exons 26 and 27						
FXR1	Chr3	50371826	50371833	AATGACAA	0.098	0.207 (-0.821, 0.561)
RBMS3	Chr3	50370739	50370745	ATATATA	0.587	0.207 (-0.629, 0.587)
RBMS3	Chr3	50370737	50370743	ATATATA	-0.203	0.207 (-0.629, 0.587)
U2AF2	Chr3	50370794	50370800	TTTTTTC	0.587	0.207 (-0.842, 0.587)
BRUNOL4/5	Chr3	50370726	50370732	TGTGTGT	0.53	0.207 (-0.811, 0.587)
BRUNOL4/5	Chr3	50370712	50370718	TGTGTGT	0.53	0.207 (-0.811, 0.587)
U2AF2	Chr3	50371777	50371783	TTTTTCC	0.587	0.207 (-0.855, 0.596)
HNRNPL	**Chr3**	**50371363**	**50371369**	**ACACACA**	**-0.975**	**0.207 (-0.790, 0.561)**
HNRNPL	**Chr3**	**50371361**	**50371367**	**ACACACA**	**-1.071**	**0.207 (-0.790, 0.561)**
CACNA2D2 intron between Exons 27 and 28						
TARDBP	Chr3	50369264	50369271	TGAATGAA	0.25	0.25 (-0.222, 0.315)
TARDBP	Chr3	50368950	50368957	TGAATGAG	0.25	0.25 (-0.203, 0.326)
BRUNOL4/5	Chr3	50370230	50370236	TGTGTGT	0.25	0.25 (-0.360, 0.329)

Table 3.7: Evolutionary conservation of RBP motif occurrences identified by FIMO and HOMER global relative to the intronic regions where they are located (comparison of 17 primate species).

First column: RBP; second column: chromosome; third column: start of the binding site coordinate; fourth column: end of the binding site coordinate; fifth column: actual nucleotide sequence FIMO and HOMER global identified as a match; sixth column: median PhyloP score for the respective RBP binding site. For the seventh, eighth, and ninth columns: the median of PhyloP scores for randomly sampled nucleotides (same composition and size as the corresponding RBP site being tested) was obtained 1,000 times for the intron where the RBP motif occurrence was located. This generated a distribution of "simulated" scores. The seventh table column corresponds to the median "simulated" score; the eighth column corresponds to the 2.5 percentile of the simulated scores; and the ninth column corresponds to the 97.5 percentile of the simulated scores. The phyloP scores were conservation scores for multiple alignments of 16 primate genomes to the human genome (comparison of 100 species). Bold indicates evolutionary conservation relative to local intronic sequence (i.e. greater than 97.5% of simulated scores) or accelerated evolution relative to local intronic sequence (i.e. less than 2.5% of simulated scores). The ELAVL2 binding site near CACNA1C Exon 21 was located within the 50 nucleotides that were truncated at the 5' and 3' end of the intronic region for the random resampling analysis. The median value, 2.5 percentile value, and 97.5 percentile value are indicated as N/A accordingly.

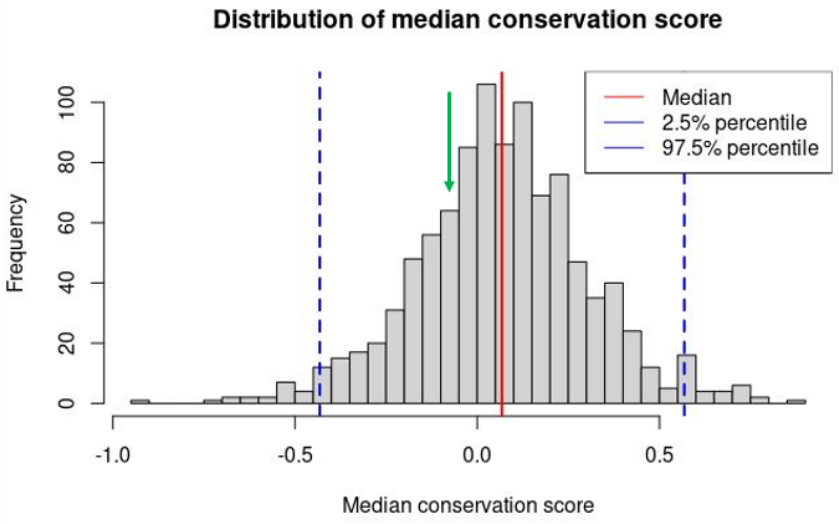
Because of the low PhyloP median score in the 17 primate species comparison, a closer examination of the HNRNPL region was carried out (**Table 3.7**; **Figure 3.12**). As the RBP site sequence on the pre-mRNA was considered as "ACACACACA" and the CACNA2D2 transcript runs in the opposite direction on the UCSC Genome Browser, the corresponding genomic sequence was considered as "TGTGTGTGT". Investigating the region of the binding site (chr3:50371361-50371369) found partial deletions of the binding site (chr3:50371366-50371369) in many primates, but not in humans and chimpanzees. Exploring the upstream region further revealed that the human sequence including the binding site and upstream of the binding site is a repeat of multiple "TG" dinucleotides. The binding site and its upstream region comprise a repetitive element that varies in its copy number across species. In this regard, it appears that the primate deletion of nucleotides chr3:50371366-50371369 (whose sequence is GTGT) does not destroy the motif, because there are more TG nucleotides upstream of chr3:50371361 (for interpretation, see **Discussion**). Additionally, regarding human genetic variation at the HNRNPL binding site, the binding site overlaps an indel that is polymorphic in humans (rs35155126) in which humans differ in how many TG dinucleotides they have.



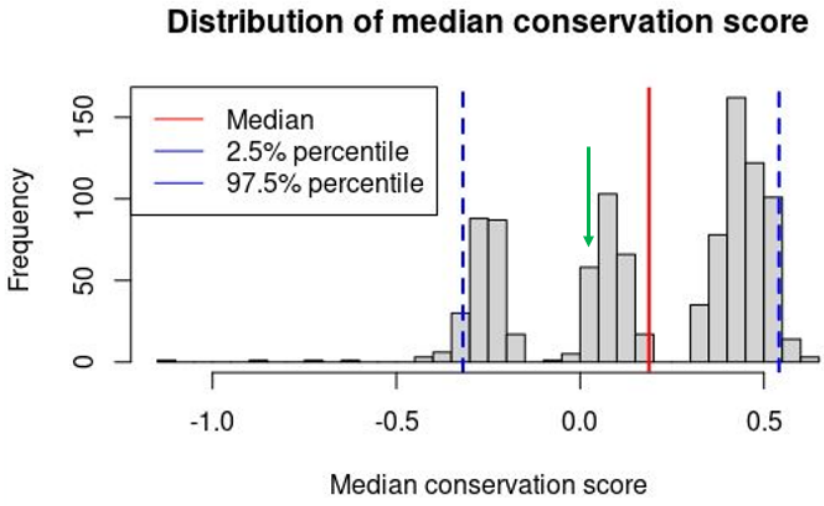
Figure 3.12: Examining the HNRNPL binding site.

(top) MultiZ alignment of the HNRNPL binding site in the intron between CACNA2D2 Exon 26 and 27. Red box indicates binding site (genomic coordinates: chr3:50371361-50371369). Genomic sequence is indicated in blue letters (for example, human binding site sequence is TGTGTGTGT). Alignment of humans is with other primates, mammals, and vertebrate species. Lack of letters indicates deletion (for example, gorilla binding site sequence is TGTGT). (bottom) Indicated is an indel that is polymorphic in humans (rs35155126) that spans the binding site and the upstream region. The red box indicates the binding site.

The BRUNOL4 and BRUNOL5 binding site upstream of Exon 21 of CACNA1C was of particular interest as it was identified using three different search methods. The results for the BRUNOL4 and BRUNOL5 binding sites are presented in **Figure 3.13.** The findings reveal a relative lack of conservation for this binding site.



(A)



(B)

Figure 3.13: Lack of evolutionary conservation of the BRUNOL4 / BRUNOL5 binding site relative to the intronic sequence between CACNA1C Exon 20 and Exon 21.

The median of PhyloP scores for four randomly sampled Gs and four randomly sampled Ts was obtained 1,000 times from the intronic sequence between CACNA1C Exon 20 and Exon 21. For (A), the phyloP scores were conservation scores for multiple alignments of 99 vertebrate genomes to the human genome (comparison of 100 species). The median PhyloP score of the intronic region is 0.078 and is indicated with a solid red line. The 2.5 percentile and 97.5 percentile values are -0.432 and 0.568 respectively and are indicated by blue dashed lines. The median score of the RBP binding site is -0.10075, falling within the 95% confidence interval. For (B), the PhyloP scores were conservation scores for multiple alignments of 16 primate genomes to the human genome (comparison of 17 species). The median PhyloP score of the intronic region is 0.373 and is also indicated with a solid red line. The median PhyloP score of the RBP binding site is 0.0325, following within the 95% confidence interval. Green arrows indicate the location of the RBP binding site median. The 2.5 percentile and 97.5 percentile values are -0.319 and 0.5410125. The median score of the RBP binding site is 0.0325.

Altogether, examining the RBP motif occurrences identified by two of the three search types identified some RBP motif occurrences showing evolutionary conservation relative to the local intronic sequence in the CACNA2D2 region of interest, but only in the 100 vertebrate species comparison and not the 17 primate species comparison. Conversely, an RBP motif occurrence for HNRNPL showed evolutionary acceleration in the 17 primate species comparison but not in the 100 vertebrate species comparison. Finally, investigating the evolutionary conservation of the RBP motif occurrences found a lack of relative evolutionary conservation for the BRUNOL4 and BRUNOL5 binding site identified by all three search types.

Random resampling of median PhyloP scores in the 17 primate species comparison identified three different distributions. This pattern of resampled median scores could suggest that this intronic region might be under different levels of evolutionary pressure or selective constraints in different subregions. The more conserved distribution (PhyloP score close to 0.5) indicates selective pressure is acting against change in this region.

The faster evolving distribution (PhyloP score close to -0.5) shows more rapid changes than expected over time, indicating possible positive selection for changes. The neutrally evolving distribution (PhyloP score closer to zero) show a rate of change expected under neutral evolution, meaning that changes have neither been significantly advantageous nor disadvantageous. While the cutoff for a PhyloP score that indicates conservation can vary, that the BRUNOL4 / BRUNOL5 RBP motif occurrence's median PhyloP score occurs close to the mode of the middle peak (with a PhyloP score closer to zero) favors the interpretation that this motif occurrence is neither evolutionarily conserved nor undergoing acceleration.

3.3.5 Gene Expression Analyses of FIMO and HOMER-Identified RBPs

3.3.5.1 Differential Gene Expression Between hiPSCs and hiPSC-Derived Neurons

The absolute and relative expression levels of RBPs can affect splicing patterns (Tang et al., 2009; Tang et al., 2011). Next, the differential gene expression levels of FIMO and HOMER global-identified RBPs were evaluated. The results for the CACNA1C, CACNA2D1, and CACNA2D2 regions of interest are presented in **Table 3.8**, **Table 3.9**, and **Table 3.10,** respectively.

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	p-adj	hiPSCs?	hiPSC-derived neurons?
BRUNOL4	ENSG00000101489	407.91	4.46	0.16	2.67E-174	9.30E-173		Higher
BRUNOL5	ENSG00000161082	1057.14	3.04	0.11	9.17E-159	2.65E-157		Higher
ELAVL2	ENSG00000107105	2144.36	3.90	0.12	1.85E-236	1.23E-234		Higher
HNRNPL	ENSG00000104824	4534.98	-1.90	0.074	2.27E-146	5.66E-145	Higher	
RBM28	ENSG00000106344	657.75	-0.74	0.040	6.52E-76	6.21E-75	Higher	

Table 3.8: Differential gene expression of RBPs with motifs enriched in the CACNA1C region of interest as identified by FIMO and HOMER global across hiPSC differentiation.

The table contains the results baseMean, log2FoldChange, lfcSE (standard error of log2FoldChange), p-value, adjusted p-value, and columns indicating whether the gene is expressed more highly in hiPSCs or hiPSC-derived neurons. Interpretation of DESEQ2 data points is described in **Chapter 2.**

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	p-adj	hiPSCs?	hiPSC-derived neurons?
FMR1	ENSG00000102081	839.48	-0.33	0.060	2.05E-08	4.52E-08	Higher	

Table 3.9: Differential gene expression of RBPs with motifs enriched in CACNA2D1 region of interest as identified by FIMO and HOMER global across hiPSC differentiation.

Interpretations of the results baseMean, log2FoldChange, lfcSE (standard error of log2FoldChange), p-value, and adjusted p-value; columns are the same as described in Table 3.8.

Gene	ENSG	baseMean	Log2FoldChange	lfcSE	p-value	p-adj	hiPSCs?	hiPSC-derived neurons?
BRUNOL4	ENSG00000101489	407.91	4.46	0.16	2.67E-174	9.30E-173		Higher
BRUNOL5	ENSG00000161082	1057.14	3.05	0.11	9.17E-159	2.66E-157		Higher
EIF4B	ENSG00000063046	8072.83	-1.39	0.081	5.01E-65	4.03E-64	Higher	
FXR1	ENSG00000114416	3425.40	-1.19	0.05	3.86E-125	7.30E-124	Higher	
HNRNPL	ENSG00000104824	4534.98	-1.90	0.074	2.27E-146	5.66E-145	Higher	
KHDRBS2	ENSG00000112232	81.29	0.25	0.28	0.366	0.42		Higher
RBMS3	ENSG00000144642	652.48	4.50	0.17	8.48E-153	2.28E-151		Higher
TARDBP	ENSG00000120948	1978.29	-1.17	0.054	1.31E-104	1.91E-103	Higher	
U2AF2	ENSG00000063244	2531.02	-1.03	0.087	3.58E-33	1.62E-32	Higher	

Table 3.10: Differential gene expression of RBPs with motifs enriched in the CACNA2D2 region of interest as identified by FIMO and HOMER global across hiPSC differentiation.

Interpretations of the results baseMean, log2FoldChange, lfcSE (standard error of log2FoldChange), p-value, adjusted p-value, and columns are the same as described in Table 3.8.

3.3.5.2 Gene Expression Patterns During Human Neural Development

Neural cells derived from hiPSCs tend to possess characteristics of fetal, rather than adult stages of development (Brennand et al., 2015). It was hypothesized that the identified RBPs could be more highly expressed during fetal, rather than adult, stages of human brain development. To gain insight into their expression patterns during human brain development, data from the Brainspan Atlas of the Developing Brain (www.brainspan.org) were inspected. The heatmap data for the expression of RBPs across fetal and postnatal development are presented in **Figure 3.14.**

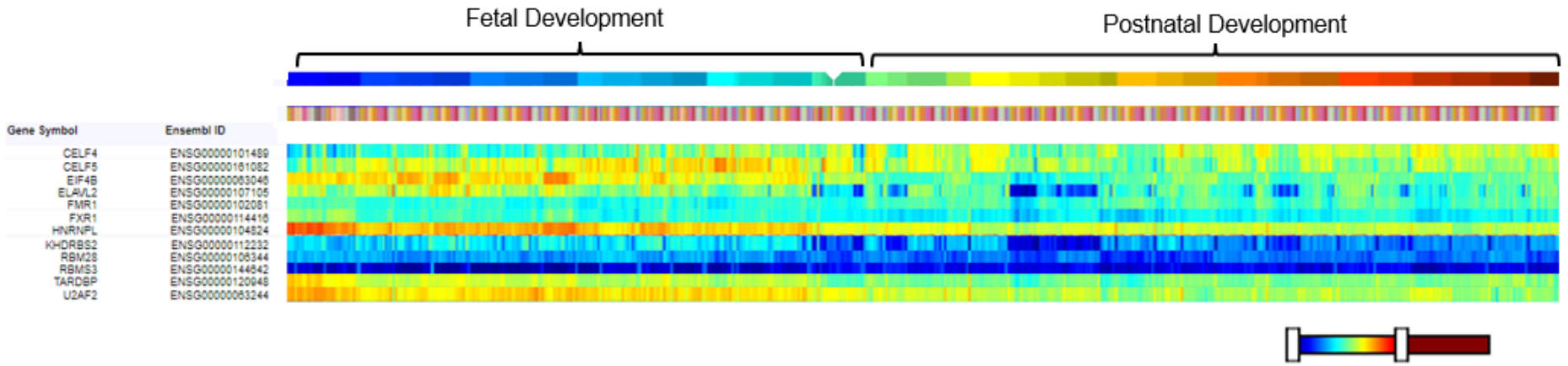


Figure 3.14: Expression of candidate RBPs during human neural development.

Heatmap of RBP expression during human neural development (both fetal and postnatal stages). The top row with different colors represents individual donors. The second row represents brain structure, where each color represents a different brain region. The bottom rows (the bulk of the figure) show a heatmap of expression levels for each RBP. Heatmap legend: red indicates higher expression levels, and blue indicates low or absent expression levels. CELF4 and CELF5 are also known as BRUNOL4 and BRUNOL5, respectively. (data from www.brainspan.org; RNA-seq data derived from bulk samples). Information about different sampled brain regions is in **Supplementary Table 2.3**. Scale represents relative expression levels (for more rigorous quantification, see **Table 3.11)**.

A higher-resolution view of the RBPs that showed higher fetal expression in the heatmap or evolutionarily conserved / accelerated binding sites in the VGCC regions of interest was pursued next. The quantification of transcript levels across fetal and postnatal development of frontal cortex is presented in **Table 3.11**. Data from the frontal cortex was analyzed because of its relevance to psychiatry. RPKM values were available for samples of fetal stages between week 8 and week 37 and postnatal stages between 4 months and 40 years. First, normality was checked using the Shapiro-Wilk test. Because not all datasets displayed a normal distribution, the median and interquartile range were calculated. The Mann-Whitney U test was used to determine if there was a significant difference between the groups. The analyses revealed statistically significant differences in gene expression among RBPs with regard to fetal versus postnatal development.

RBP	Fetal FC median	Fetal FC IQR	Postnatal FC median	Postnatal FC IQR	U statistic	p-value	Significant?
BRUNOL4	12.313	7.432	30.308	16.125	447	2.843E-14	Yes
BRUNOL5	36.988	28.946	27.840	18.917	2532	0.0415	Yes
ELAVL2	28.140	16.846	13.283	9.623	3764	9.982E-15	Yes
EIF4B	53.208	24.379	14.131	7.621	4071	5.183E-20	Yes
HNRNPL	64.642	25.200	32.123	8.336	4091	2.189E-20	Yes
KHDRBS2	8.061	2.754	4.053	2.750	3734	2.946E-14	Yes
RBMS3	0.227	0.315	0.697	0.333	673	5.40E-11	Yes
RBM28	4.733	1.557	2.864	1.392	3742	2.21E-14	Yes
U2AF2	53.631	15.258	26.774	13.314	4049	1.32E-19	Yes

Table 3.11: Expression levels of RBPs that are more highly expressed in fetal relative to postnatal brain development (frontal cortex).

Indicated in the first column: RBP; second column: median of fetal samples; third column: interquartile range for fetal samples; fourth column: postnatal median; and fifth column: postnatal interquartile range. The sixth column represents the U statistic; seventh column represents p-value. In the eighth column, whether the differences are considered statistically significant is presented (p-value <0.05). Data from 133 samples of frontal cortex (51 fetal samples and 82 postnatal samples). Values represent RPKM. All the RBPs above are more highly expressed in the frontal cortex of fetal brain except BRUNOL4 and RBMS3. FC: frontal cortex.

It was surmised that studying gene expression during fetal neural development could provide insight into RBP expression patterns during in vivo neural differentiation. Cortical neurogenesis occurs between 45 days (6 weeks) and 125 days (17 weeks) of fetal development (Rakic, 1988). As noted earlier, RPKM values were available starting at week 8 and until week 37 of fetal development. It was speculated that a comparison of gene expression patterns during fetal neurogenesis and after neurogenesis had occurred would shed light on RBP expression patterns in "pre-differentiated" and differentiated in vivo cellular states (for caveats to this interpretation, see **Discussion** of this Chapter). Accordingly, samples collected between weeks 8 and 17 inclusive were considered "early fetal" samples and samples collected after week 17 up to week 37 were considered "late fetal" samples. The analyses identified statistically significant changes in several RBP expression levels

between the period where neurogenesis is occurring and afterwards (**Table 3.12**).

RBP	Early Fetal FC median	Early Fetal FC IQR	Late Fetal FC median	Late Fetal FC IQR	U statistic	p-value	Significant?
BRUNOL4	10.416	4.049	19.748	11.198	44	5.392E-08	Yes
BRUNOL5	30.830	17.291	53.672	12.005	77	6.256E-06	Yes
EIF4B	56.854	20.695	36.681	4.099	528	1.460E-07	Yes
ELAVL2	33.247	16.99	22.688	5.832	398	0.0291	Yes
HNRNPL	72.429	37.858	58.659	12.542	453	0.000756	Yes
KHDRBS2	7.548	2.944	8.355	2.572	224	0.200	No
RBMS3	0.180	0.259	0.345	0.479	154	0.00632	Yes
RBM28	4.833	1.636	4.731	1.536	259	0.559	No
U2AF2	54.281	18.881	53.631	8.687	267	0.671	No

Table 3.12: Comparison of RBP expression levels between early and late fetal development.

Statistical analysis was conducted similarly as to **Table 3.11**. For interpretation of statistics, see **Table 3.11**. The number of samples in early fetal stages was 34 and the number of samples in late fetal was 17.

Altogether, the data provide insights into the relative expression of RBPs in fetal vs postnatal development as well as early fetal vs late fetal stages. For the fetal versus postnatal comparison – BRUNOL4 and RBMS3 showed higher expression in postnatal relative to fetal stages; whereas BRUNOL5, EIF4B, ELAVL2, KHDRBS2, HNRNPL, RBM28, and U2AF2 manifested higher expression in fetal stages relative to later postnatal stages. For the early versus later fetal stage comparison – EIF4B, ELAVL2 and HNRNPL showed higher expression in early fetal relative to late fetal samples. BRUNOL4, BRUNOL5, and RBMS3 showed higher expression in the late fetal stages relative to early fetal stages. KHDRBS2, RBM28, and U2AF2 showed no differences or higher expression in late fetal stages. Altogether significant changes were observed in RBP expression between fetal and postnatal development, as well as the early and later fetal stages of development.

3.4 Discussion

In this Chapter, candidate regulators of splicing events that occur during hiPSC differentiation in VGCCs of interest including CACNA1C, CACNA2D1, and CACNA2D2 were identified. A summary of the RBP results is presented in **Table 3.13**.

Identified RBPs		
FIMO and HOMER global intersection		
CACNA1C	**CACNA2D1**	**CACNA2D2**
HNRNPL, hnRNPLL, RBM28, BRUNOL4, BRUNOL5, ELAVL2	FMR1	EIF4B, FXR1, HNRNPL, TARDBP, U2AF2, BRUNOL4, BRUNOL5, KHDRBS2, RBMS3
FIMO, HOMER global, and HOMER local intersection		
CACNA1C	**CACNA2D1**	**CACNA2D2**
BRUNOL4, BRUNOL5	N/A	N/A

Table 3.13: Summary of results.

Indicated are RBPs identified in the CACNA1C, CACNA2D1, and CACNA2D2 regions of interest using FIMO and HOMER. RBPs were identified via two intersections: 1) FIMO and HOMER global; and 2) FIMO, HOMER global, and HOMER local. Red indicates upregulated in hiPSCs; blue indicates upregulated in hiPSC-derived neurons. Underlined indicates the presence of a binding site showing either evolutionary conservation or acceleration in either the 100 vertebrate species or 17 primate species comparison.

Altogether, these analyses identify putative RBPs and motif occurrences that might regulate alternative splicing of VGCCs.

3.4.1 Identified RBPs

This section considers the RBPs identified by the triple and double intersection of different searches, their evolutionary conservation, and expression patterns during human neural development.

3.4.1.1 BRUNOL4 and BRUNOL5

BRUNOL4/CELF4 and BRUNOL5/CELF5 RBP motifs were identified in the analyses conducted in this Chapter. The stringent intersection of FIMO, HOMER global, and HOMER local identified only a single binding motif occurrence in CACNA1C, CACNA2D1, and CACNA2D2 – a binding motif occurrence upstream of Exon 21 in CACNA1C. Examination of publicly available data showed that BRUNOL4/CELF4 and BRUNOL5/CELF5 are upregulated in hiPSC-derived neurons compared with hiPSCs. BRUNOL4 showed higher expression in postnatal relative to fetal brain and higher expression in late fetal stage relative to early fetal stage samples. BRUNOL4 expression appears generally more correlated with later stage cells and tissues. BRUNOL4 is more associated with more dif-

ferentiated states. BRUNOL5 was more highly expressed in fetal brain vs. postnatal brain, and higher in late fetal stage samples relative to early fetal stages samples. BRUNOL5 expression appears more correlated with the later stages of fetal neural development.

Regarding VGCCs, murine BRUNOL4 has been reported to affect the splicing of Cacna2d2 as revealed in studies of Brunol4-deficient mice (Wagnon et al., 2012). BRUNOL/CELF proteins have been implicated in neuronal function and biology as well as in various neurological disorders (Gallo and Spickett, 2010). For example, mice lacking BRUNOL4/CELF4 have seizures ascribed to increased neuronal excitability (Yang et al., 2007; Wagnon et al., 2011). Moreover, mutations in unc-75, an orthologue of CELF3, CELF4, and CELF5 caused sprouting defects in subpopulations of motor neurons with a locomotor phenotype (Loria et al., 2003). Taken together, these data support a potential role for BRUNOL4/CELF4 and/or BRUNOL5/CELF5 in neural tissues. A visual representation of how BRUNOL4 and BRUNOL5 might regulate the splicing of CACNA1C Exon 21 and Exon 22 is presented in **Figure 3.15**. In such a model, considering that BRUNOL4 and BRUNOL5 are highly expressed in hiPSC-derived neurons relative to hiPSCs, BRUNOL4 and BRUNOL5 would bind to the intron upstream of CACNA1C Exon 21, promoting the exclusion of Exon 21 and the inclusion of Exon 22.

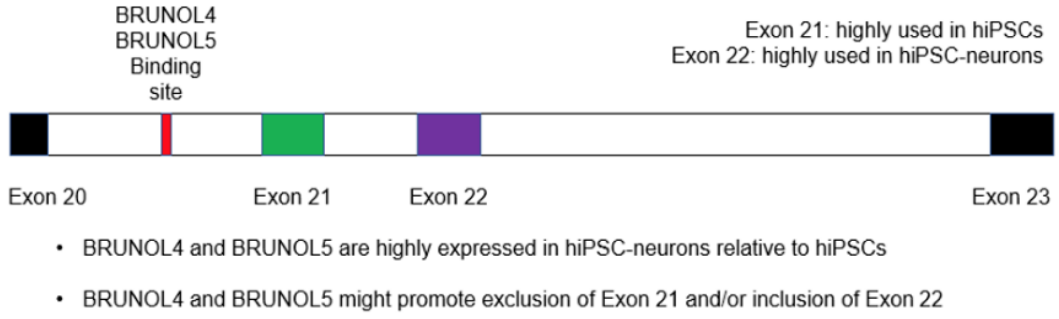


Figure 3.15: Visual representation of identified RBP motifs that contribute to alternative splicing of CACNA1C Exons 21 and Exon 22.

Identifying RBP motifs based on a stringent triple intersection of FIMO, HOMER global, and HOMER local. BRUNOL4 and BRUNOL5 are highly expressed in hiPSC-derived neurons relative to hiPSCs. Exon 21 is highly used in hiPSCs, and Exon 22 is highly used in hiPSC-derived neurons. The binding of BRUNOL4 and BRUNOL5 to their binding site in the intron between Exon 20 and Exon 21 might promote the exclusion of Exon 21 and the inclusion of Exon 22 in CACNA1C transcripts. Thin red box: BRUNOL4/5 binding site. White indicates introns. Green indicates Exon 21. Purple indicates Exon 22. Black indicates exons that are not Exon 20 and Exon 23. Introns and exons are not drawn to scale.

3.4.1.2 HNRNPL

HNRNPL was an additional RBP identified by FIMO and HOMER global. Differential gene expression found that HNRNPL expression is higher in hiPSCs relative to hiPSC-

derived neurons. HNRNPL expression is higher in early fetal brain samples relative to late fetal brain samples. Thus, one might surmise that HNRNPL expression is more associated with a less differentiated state. A binding site for HNRNPL in the intron between CACNA2D2 Exon 26 and Exon 27 showed evolutionary acceleration in the 17 primate species comparison, but neither conservation nor acceleration among 100 vertebrate species. This suggests that this binding site has undergone significant evolutionary changes in primates but has not been conserved across a broader range of vertebrates. **Figure 3.16** presents a visual representation of the RBP binding site, identified based on evolutionary conservation analysis, involved in alternative splicing of CACNA2D2 Exon 26. Heterogeneous nuclear ribonucleoproteins (hnRNPs) comprise a significantly sized family of RBPs that regulate different aspects of the RNA life cycle, including alternative splicing (Geuens et al., 2016). Several hnRNPs have been associated with different neurodegenerative conditions, including spinal muscular atrophy, amyotrophic lateral sclerosis, and Alzheimer's disease.

3.4.1.3 U2AF2

U2 small nuclear RNA auxiliary factor 2 (U2AF2) plays an important role in pre-mRNA splicing. Specifically, it is involved in recognizing the PPT, a cis-acting sequence that directs intron removal and recruits other splicing factors to form the spliceosome complex (Coolidge et al., 1997). FIMO and HOMER global identified U2AF2 binding motif occurrences near the CACNA2D2 region of interest that showed evolutionary conservation across 100 vertebrate species. Analyses across 17 primate species found that a U2AF2 motif occurrence had a median PhyloP score that was not greater than the 97.5 percentile, indicating that the null hypothesis in primate species could not be rejected. If functional, the higher expression of U2AF2 in hiPSCs relative to hiPSC-derived neurons is consistent with the hypothesis that its binding may prevent the skipping of CACNA2D2 Exon 26. U2AF2 showed higher expression levels in fetal relative to postnatal stages of brain development but did not show statistically significant differences between early and late stages of fetal development. Similarly to HNRNPL, it appears to be associated with less differentiated and more immature stages of development, although no difference was observed between early and late fetal stages.

In terms of function associated with neural tissues, mutations in U2AF2 are associated with neurodevelopmental diseases. A de novo U2AF2 variant has been associated with

global developmental delay, brain malformations, and epilepsy (Hiraide et al., 2021). A possible variant in U2AF2 causes exon skipping in two patients with U2AF2-associated neurodevelopmental illness (Wang et al., 2023). U2AF2 has also been identified as being elevated in glioma and is associated with poor patient survival (Jiang et al., 2020). Altogether, U2AF2 likely possesses a function during human neural development.

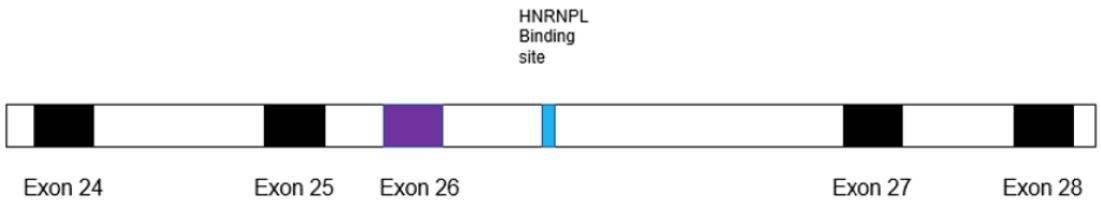


Figure 3.16: Visual representation of identified RBP motifs contributing to alternative splicing of CACNA2D2 Exon 26.

Identifying RBP motifs based on evolutionary conservation analyses. HNRNPL is expressed higher in hiPSCs relative to hiPSC-derived neurons. Exon 26 is preferentially included in hiPSC-derived neurons and skipped in hiPSCs. In hiPSCs, binding of HNRNPL in the intron between Exon 26 and Exon 27 might promote the skipping of Exon 26. Thin teal box: HNRNPL binding site. Note, the HNRNPL binding site shows evolutionary acceleration in the 17 primate species comparison. White indicates introns. Purple indicates Exon 26. Black indicates exons that are not Exon 26. Introns and exons are not drawn to scale.

3.4.1.4 RBMS3

RBMS3 binding motif occurrences were identified by FIMO and HOMER global in the vicinity of the CACNA2D2 region of interest. Analysis of PhyloP scores revealed evolutionary conservation of a binding site among 100 vertebrate species, but not among 17 primate species. Similarly to U2AF2, analyses across 17 primate species found an RBMS3 motif occurrence with a median PhyloP score that was not conserved. Differential gene expression found that RBMS3 expression levels are higher in hiPSC-derived neurons relative to hiPSCs. Inspection of BrainSpan Atlas data showed that RBMS3 is more highly expressed in postnatal relative to fetal stages of brain development, with some statistically significant differences between early and late fetal stages of brain development. RBMS3's role in the nervous system is little studied, but RBMS3 is expressed in the brain (Lu et al., 2012; Penkov et al., 2000) in different GABAergic neuron subclasses (Furlanis et al., 2019; Tasic et al., 2018) and RBMS3 deletions have been linked to autism spectrum disorders (Leblond et al., 2019; O'Roak et al., 2011). In light of the findings in this Chapter, more research into RBMS3, particularly in the human nervous system is warranted.

3.4.2 Evolutionary Conservation

One way to infer the possible functionality of an RBP binding site is to assess its evolutionary conservation across different species. Identifying functional non-coding regions of the human genome is a major challenge in the genomics field (Sadri et al., 2011; Haerty and Ponting, 2014). By evaluating how a particular region has evolved over time, it becomes possible to identify whether a negative or positive selection is occurring, perhaps indicating that a given region is functional (Sadri et al., 2011). Comparisons of the sequences of protein-coding genes across different species have proven useful for inferring functionality, for example, helping to facilitate the understanding of gene function, or assisting the identification of protein domains or functional amino acid residues (Ponting et al., 2000; Gaucher et al., 2002; Lichtarget and Sowa, 2002; Asthana et al., 2007; Haerty and Ponting, 2014). A similar approach was considered for the conservation of RBP binding sites in this Chapter.

Across the 100 vertebrate species comparison, evolutionarily conserved binding sites for BRUNOL4, BRUNOL5, RBMS3 and U2AF2 were identified in the CACNA2D2 intron between Exon 26 and Exon 27. Additionally, a BRUNOL4 and BRUNOL5 binding site in the intron between CACNA2D2 Exon 27 and 28 showed conservation. Due to their evolutionary conservation, these RBP binding sites might possess function. However, the PhyloP score medians for these binding sites across 17 primates showed a lack of conservation or acceleration. This could potentially support the hypothesis that these RBP binding sites have an important function in vertebrates more broadly, but not in primates.

While evolutionary conservation likely suggests function, accelerated evolution could also indicate functionality as well (Capra et al., 2013; Holloway et al., 2016). Primate accelerated regions (PARs) are genomic regions that have undergone rapid evolution in primate species compared to other species (Lindblad-Toh et al., 2011). If a PhyloP score shows evolutionary acceleration among 17 primates but no conservation or acceleration among 100 vertebrate species, it could suggest a genomic region has undergone significant evolutionary changes in primates but has not been conserved among a broader range of vertebrates. In this regard, the low PhyloP scores for the HNRNPL binding site in the intron between CACNA2D2 exon 26 and 27 indicate that this site might possess biological function. This could indicate that the binding site has evolved rapidly in primates due to positive selection or other selective pressures. PARs are thought to be involved in the

evolution of primate-specific traits, such as brain development, sensory perception, and immunity. One might speculate that this binding site has a specific function or role in primates that is different from its function in other vertebrates, possibly related to brain development. Human accelerated regions (HARs), which are genomic regions that have undergone rapid evolutionary change in the human lineage compared to other species, serve as developmental enhancers (Capra et al., 2013; Holloway et al., 2016); and also have been identified as having functional effects on human neurodevelopment (Whalen and Pollard, 2022; Whalen et al., 2023). It will be of interest to examine whether any of the binding sites identified by the double intersection of FIMO and HOMER represent HARs.

There are additional points to note when considering the HNRNPL binding site with a low PhyloP median score in the comparison of 17 primates (genomic coordinate chr3:50371361-50371369) (K. Pollard, personal communications). The binding site (sequence "TGTGTGTGT") and the region surrounding the binding site comprise a "simple repeat", as an inspection of the human sequence revealed that the region contains a repeat of multiple "TG" nucleotides. A comparison of the human sequence with other species showed that this repetitive element varies in its copy number across species. In many primate species, a portion of the binding site was deleted (chr3:50371366-50371369; TGTGT**GTGT**, where bolded nucleotides indicate primate deleted portion), which accounts for the low PhyloP score. The presence of additional TG nucleotides upstream of chr3:50371361 indicates that the binding site is not destroyed in primate species with the GTGT deletion. Nonetheless, the deleted portion of the binding site could still be functional, if changes in TG number affect RNA structure, which could in turn affect RBP binding affinity or specificity. For example, there could be a creation or disruption of a stem-loop structure that could affect the accessibility or stability of the binding site for HNRNPL. Alternatively, there could exist alterations in the spacing of the binding site relative to other RBP motifs that could affect cooperative or competitive binding of RBPs. Although the speculation above regards variation in binding site sequence among primates, it is interesting to note that the binding site and surrounding sequence overlaps with an indel that is polymorphic in humans (rs35155126), suggesting that among humans, there exist differences in how many TG dinucleotides different individuals possess at this locus.

Finally, it is important to note that PhyloP scores can be difficult to interpret in repeat regions (K. Pollard, personal communications). The issue is the MultiZ alignment that

the PhyloP scores are based on. It is challenging to generate optimal multiple sequence alignments in repeat regions. For instance, a shift in the deletion a few base pairs so that it no longer overlaps the HNRNPL binding site would change the PhyloP scores for the binding site (the conservation signal would increase) and the region adjacent to the binding site (the conservation score would decrease). Altogether, while the HNRNPL binding site's PhyloP scores indicate acceleration, a consideration of genomic context suggests some caution is warranted in interpreting the scores.

The analysis of the BRUNOL4 and BRUNOL5 binding sites identified by the triple intersection revealed a lack of evolutionary conservation or acceleration of the motif occurrence relative to the CACNA1C intron where it was found, in both 100 vertebrate species and 17 primate species comparisons. It is possible that this finding indicates that the motif occurrence is non-functional. Functional assays involving experiments, potentially using hiPSCs and genome editing, which test the effect of altering the RBP binding site sequence can help determine whether an RBP binding sequence has a role in splicing.

3.4.3 Limitations

Over the course of the studies carried out in this Chapter, two key caveats concerning the motif-based FIMO and HOMER tools became apparent. A key limitation pertains to potential "noise" in the binding site predictions. Noise in this context refers to the predisposition for false negative and false positive results.

The sensitivities of both FIMO and HOMER were assessed in the analyses conducted in this Chapter. Experimental methods like RIP-seq and CLIP-seq can provide direct evidence of RBP binding sites in vivo. The sensitivity analysis revealed that FIMO and HOMER are capable of detecting motif occurrences where some experimentally determined eCLIP peaks were located, but not in a majority of instances. For the four out of five RBPs assessed there was no statistical difference between FIMO and HOMER. The generation of false negative results could be related to the fact that FIMO and HOMER, and potentially RBP motif searching tools more generally, do not consider the context-dependent binding of RBPs. Aside from the primary sequence of the pre-mRNA, RBP binding can be influenced by additional factors including RNA secondary structure, tertiary structure, and post-transcriptional modifications. An alternative explanation for the absence of FIMO and HOMER motif occurrences where eCLIP peaks do occur is possible artifacts and biases in the experimental eCLIP data, which for example can emerge from

cross-linking of RNA and protein during eCLIP. Therefore, false negative results could represent true failure of FIMO and HOMER to detect motif occurrences where RBP-RNA interactions occur or they could also be explained by limitations in experimental methods.

Specificity, the ability to distinguish true from false positive results, is another factor. One way of evaluating specificity occurred by searching randomly selected genes for motif occurrences for a given RBP (see **Appendix II**). HOMER detected RBP motifs in nearly every random gene surveyed, suggesting RBP motifs are widely prevalent in the human genome. The wide prevalence of RBP motif occurrences raises questions. One interpretation is that FIMO and HOMER global could be prone to generating false positive predictions. This could be related to the limitations of RBP motif searching tools more generally, which often rely on relatively short and degenerate sequence motifs, contributing to a high false positive rate.

The widespread occurrence of RBP motifs could also indicate that alternative splicing occurs not by individual RBP binding events but via combinatorial regulation mediated by different combinations of RBPs. Regarding combinatorial regulation of alternative splicing, it is worth carefully considering whether BRUNOL4 / 5 binding or HNRNPL binding alone is sufficient to explain alternative splicing at CACNA1C Exon 21 and 22 and CACNA2D2 Exon 26; or whether there are other unidentified RBP binding events that are also involved in alternative splicing of those exons.

Altogether, while there have been significant advances in the identification and characterization of RBP motifs, there are still challenges in accurately predicting these motifs and interpreting the results. As previously noted, the original studies reporting FIMO and HOMER have been cited thousands of times. A brief survey of the literature of FIMO and HOMER papers revealed that experimental validation of RBP motif occurrences identified by FIMO or HOMER is generally lacking. One representative example involves the use of FIMO to identify RBP binding sites in SARS-CoV-2 with no experimental corroboration (Srivastava et al, 2020). However, there exist examples in the literature where individual motif occurrences identified by FIMO or HOMER have been experimentally validated (Cesana et al, 2018). Nonetheless, to my knowledge, a systematic evaluation of whether FIMO and HOMER can detect motif occurrences in a sizable sample of experimentally determined RBP-RNA interaction data has not been carried out until now. The analyses in this Chapter indicate that practitioners should be cautious when interpreting RBP binding predictions and requires the use of complementary methods, such as eCLIP, to

validate findings.

Finally, it is worth briefly considering how to interpret the BrainSpan Developmental Transcriptome data. If a gene is expressed more highly in fetal relative to adult brain development, it could suggest that the gene might play a significant role in the developing brain (applies for both RBPs in this chapter and TFs in **Chapter 4**). Additionally, analyses of relative expression levels during "early" fetal development (between 8 and 17 weeks) and "late" fetal development (17 and 37 weeks), during which neurogenesis is occurring or after neurogenesis occurred, respectively were carried out (Raskic, 1988). It was speculated that partitioning data in this way could shed insight into RBP expression patterns as neural cells become differentiated neurons in vivo, possibly reflecting "pre-differentiated" versus more differentiated states in the fetal brain. However, there exist caveats to this interpretation. One is that neurogenesis is a multi-faceted process that involves the proliferation of neural progenitors, migration of cells into different brain regions, among other processes, and an interpretation of the early fetal development stages as "pre-differentiated" does not capture this complexity. Another limitation of this analysis is that it remains unclear how in vitro differentiation of hiPSCs maps onto in vivo fetal development. The datasets do not appear available but the in vivo analogs of hiPSCs and in vitro differentiated neurons would likely be epiblast tissue from early-stage human postimplantation embryos and first/second trimester fetal brain samples, depending on what stages hiPSC-derived neurons correspond to. Nonetheless, while imperfect, the analyses capture changes in RBP expression between early and later stages of fetal brain development. Altogether, gene expression patterns can provide insight into whether an RBP possesses a function in neural tissue or a particular developmental stage, but there exist caveats to consider for their interpretation.

3.4.4 Next Steps

The limitations of available bioinformatics tools highlight the importance of conducting experiments. As noted previously, experiments in the laboratory using hiPSCs and their differentiated cells could potentially resolve this uncertainty. Such experiments could involve manipulating the expression levels of target RBPs in hiPSCs or differentiated neural derivatives by transgenic overexpression, knockdown or CRISPR-mediated knockout and assess effects on VGCC alternative splicing, e.g. for short or long read sequencing. Such genetic perturbations could be accompanied by the mapping of in vivo RNA-protein in-

teractions. It might be necessary to disrupt multiple members of the same RBP family (for example, double disruption of BRUNOL4/CELF4 and BRUNOL5/CELF5) to fully unveil the role of certain RBPs in regulating certain splicing events, given the potential for redundancy.

3.4.5 Summary

In summary, a list of possible RBPs that regulate alternative splicing of CACNA1C, CACNA2D1, and CACNA2D2 was generated by intersecting the results from two different bioinformatics tools: FIMO and HOMER. These RBPs were differentially expressed between hiPSCs and hiPSC-derived neurons. The double intersection of FIMO and HOMER global identified evolutionarily conserved binding sites for BRUNOL4, BRUNOL5, RBMS3, and U2AF2 near the CACNA2D2 region of interest; as well as a HNRNPL site that showed evolutionary acceleration. Different evolutionary conservation scores across the 100 vertebrate species and 17 primate species comparisons revealed changes in the primate lineage. A more stringent intersection of FIMO, HOMER global, and HOMER local identified only BRUNOL4 and BRUNOL5 at the CACNA1C region of interest. The analysis of the evolutionary conservation of this BRUNOL4 and BRUNOL5 binding site revealed a lack of conservation relative to the intron where it was located. The transcripts of some of the identified RBPs, including BRUNOL5 and U2AF2, were more highly expressed in human fetal brain samples relative to postnatal brain samples. These findings serve as a prelude to future experimental studies in the laboratory to dissect the mechanisms underlying the alternative splicing of VGCCs.

Chapter 4

Identification of Candidate Transcription Factors That Regulate VGCC Abundance

4.1 Introduction

In **Chapter 2**, transcriptional changes in VGCCs were characterized across hiPSC differentiation, a model experimental system for human neural development, in the abundance and splicing of VGCCs. In **Chapter 3**, candidate proteins that might regulate the VGCC splicing events of interest in CACNA1C, CACNA2D1, and CACNA2D2 were identified. However, the transcriptional regulation of VGCC gene expression, and in particular, the identity of candidate proteins that regulate VGCC abundance, are less well understood.

TFs regulate gene expression programs across different cell types. There exist nearly 1,500 TFs constituting a large diverse group of proteins (Vaquerizas et al., 2009; Lambert et al., 2018). The relative levels and combinations of expressed TFs control cellular identity and cell type-specific gene expression. Indeed, combinatorial overexpression of TFs can induce dramatic changes in cell state, as demonstrated by the fact that overexpressing pluripotency TFs Oct4, Sox2, Klf4, and c-Myc can restore pluripotency in somatic cells (Takahashi and Yamanaka, 2006). TFs typically contain DNA-binding domains and effector domains. The DNA-binding domains recognize short sequences of DNA (TF motifs) and combinations of TFs bind to enhancers and promoters of genes to regulate their expression (Panne et al., 2007; Avsec et al., 2021). The effector domains can regulate gene expression through the recruitment of coactivator or corepressor proteins (Arnold et al., 2018; Boija et al., 2018; Soto et al., 2022). TF-coactivator or TF-corepressor complexes regulate genes by various means, including mobilizing nucleosomes, altering chromatin-associated proteins, affecting genome architecture, and finally, recruiting the transcription

apparatus (including RNA polymerase II and associated proteins) and regulating transcription initiation and elongation (Vos, 2021; Richter et al., 2022).

Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) identifies regions of accessible chromatin at a genome-wide scale, allowing concurrent examination of nucleosome occupancy and factor occupancy in regulatory sites (Buenrostro et al., 2013). ATAC-seq involves direct in vitro transposition of sequencing adaptors by Tn5 transposase into accessible chromatin, and steric blockage of transposition in closed chromatin makes this less likely (overview in **Figure 4.1**). (Buenrostro et al., 2013). Consequently, the resulting DNA fragments for next-generation sequencing are enriched for open chromatin regions. While open chromatin regions are enriched for ATAC-seq signal, in ATAC-seq DNA sequences occupied by DNA-binding proteins such as TFs are protected from transposition. A resulting "footprint" demonstrates the presence of a DNA-binding protein (Buenrostro et al., 2013). Using this method, it has been possible to use ATAC-seq data to infer the identity of TFs that regulate transitions in cell states.

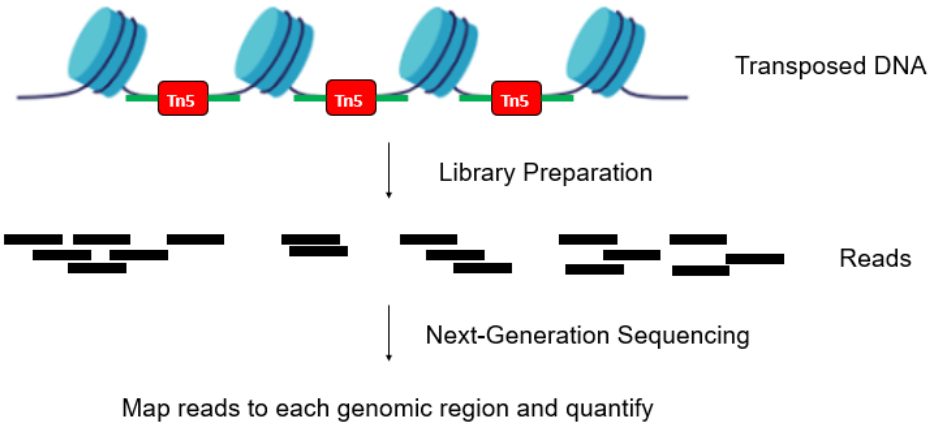


Figure 4.1: Overview of ATAC-seq.

ATAC-seq allows regions of accessible chromatin to be surveyed by using hyperactive mutant Tn5 transposase to insert sequencing adapters into regions of the genome that show increased chromatin accessibility (Buenrostro et al., 2013). During "tagmentation", Tn5 transposase cleaves and tags double-stranded DNA with sequence adaptors (depicted above by green lines attached to the red Tn5 enzyme). DNA fragments are then subjected to next-generation sequencing, and sequenced reads are then mapped to the relevant genome. Peaks of sequencing reads, where many reads map to the same location, indicate regions of high accessibility. Chromatin cartoon from biorender.com

The primary objective of this Chapter is to identify putative regulators of VGCC abundance by examining chromatin accessibility data from hiPSC-derived neurons. I hypothesized that TFs with footprints at open chromatin regions at VGCC loci that are also differentially expressed between hiPSCs and hiPSC-derived neurons would be candidate regulators of VGCC abundance. A motif-based bioinformatics tool called Transcription

factor Occupancy prediction by Investigation of ATAC-seq Signal (TOBIAS) was deployed to conduct TF footprinting analyses of ATAC-seq data (Bentsen et al., 2020). This type of analysis allows one to analyze the genome-wide occupancy of multiple TFs in parallel from a single experiment. Previously, TF footprinting was complicated by the finding that enzymes used to study chromatin accessibility such as DNase I or Tn5 transposase are biased towards certain sequences, which has been reported as an obstacle for identifying true footprints (Sung et al, 2016; Li et al, 2019). Among TOBIAS's unique features is its capability to correct for Tn5 transposase bias, which the authors demonstrate improves the signal-to-noise ratio (Bentsen et al, 2020). Additionally, Bentsen et al. show that TOBIAS performs better than most state-of-the-art tools, except msCentipede, to which TOBIAS performs similarly but at a significantly reduced proportion of computational cost and running time.

4.2 Methods

The overall strategy for identifying candidate TFs that regulate VGCC abundance across hiPSC differentiation is presented in **Figure 4.2**.

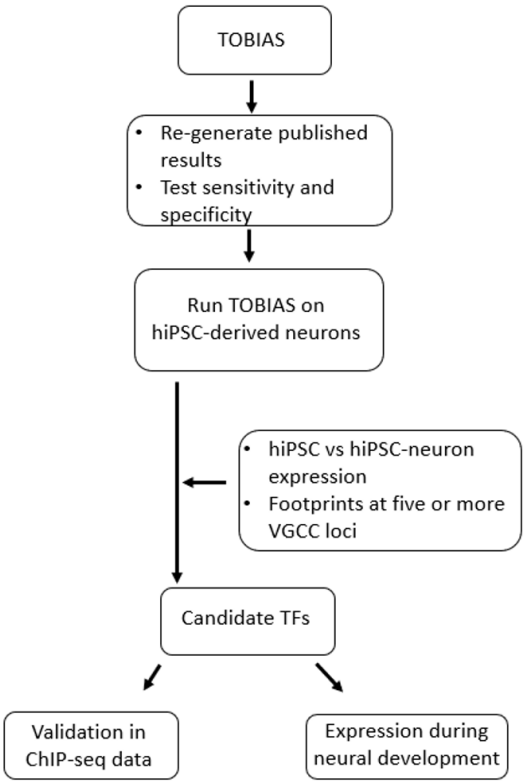


Figure 4.2: Overall strategy to identify candidate TFs that regulate VGCC abundance across hiPSC differentiation.

See main text for more details.

First, the computational pipeline was validated by replicating published studies' TOBIAS analyses, which involved re-generating the results described in Bentsen et al, 2020 (see **4.3.1.1**). Next, the sensitivity and specificity of the in-house application of the published TOBIAS computational pipeline were assessed (see **4.3.1.2**). In the context of this Chapter, the definition of sensitivity slightly differs from that in **Chapter 3** – here, it refers to detecting known changes in motif enrichment and individual TF occupancy events. Specificity refers to the capacity to distinguish between true and false positive results. Testing the capabilities and limitations of TOBIAS, including sensitivity and specificity, was necessary because TOBIAS was only published in 2020.

Next, TOBIAS was applied to hiPSC-derived neuron ATAC-seq data (see **4.3.2**). TF footprints were then identified by TOBIAS at the VGCC loci TSS. The list of TFs was refined by identifying which TFs underwent changes in expression levels across hiPSC differentiation and which TFs whose footprints were found at five or more VGCC loci. These analyses resulted in a list of candidate regulators of VGCC abundance. Finally, attempts were made to validate TOBIAS findings in published experimental chromatin immunoprecipitation sequencing (ChIP-seq) data (see **4.3.2.4**) and gene expression pat-

terns during human neural development were studied by analyses of BrainSpan Atlas data (see **4.3.2.5**).

4.2.1 Datasets

To assess the suitability of using TOBIAS to analyze VGCCs in hiPSC-derived neurons, datasets from three studies were re-analyzed (Wu et al., 2018; Yang et al., 2019; Xiong et al., 2022). The first study, which described TOBIAS for the first time (Bentsen et al., 2020), used a chromatin accessibility dataset of cells from different stages of human preimplantation development, as well as human naïve and primed PSCs (accession number: GSE101571) (Wu et al., 2018). The second study involves chromatin accessibility data derived from mouse embryonic stem cells (ESCs) engineered to repress expression of the TF Oct4 upon doxycycline treatment (accession number: GSE174774) (Niwa et al., 2000; Xiong et al., 2022). The third study investigated chromatin accessibility data from mouse ESCs exiting naïve pluripotency and entering a formative phase of pluripotency (ArrayExpress: E-MTAB-7207) (Yang et al., 2019).

For the analyses of hiPSC-derived neurons, ATAC-seq data from the HIPSCI were downloaded from the European Nucleotide Archive (Project: PRJEB18630) (Schwartzentruber et al., 2018). All available sequencing files were downloaded that included data from 23 different individual donors. The different donors were unrelated, healthy individuals (Kilpinen et al., 2017). As previously described in **Chapter 2**, the hiPSC-derived neurons from which the ATAC-seq data were derived possess a sensory neuronal (nociceptor) phenotype (Schwartzentruber et al., 2018). ATAC-seq data were generated from hiPSC-derived neurons at day 28 of hiPSC differentiation. Because ATAC-seq data for parental hiPSCs were not available, only hiPSC-derived neurons were analyzed.

Metadata regarding the sex and age of the donors is provided in **Table 4.1.**

4.2.2 Programming Code

The programming code for this Chapter was deposited in Github at `github.com/adelosan521/chapter4`. Raw sequencing data were obtained from the European Nucleotide Archive (Project: PRJEB18630). The code for pre-processing the raw ATAC-seq data files is in the "atac_preprocessing.sh" script. To replicate the analyses in Bentsen et al., 2020; the code in the "embryo.sh" script was used. For assessing the sensitivity and specificity of TOBIAS by re-analyzing the data in Xiong et al., 2022; the code in the "oct4.sh" script

Donor id	Sex	Age
HPSI1013i-cups_3	Female	45-49
HPSI1013i-wuye_2	Female	30-34
HPSI0214i-feec_3	Male	60-64
HPSI0214i-kucg_2	Male	65-69
HPSI0134i-hoik_1	Female	40-44
HPSI0314i-qonc_2	Female	65-69
HPSI1113i-hayt_1	Male	70-74
HPSI0114i-rozh_5	Female	65-69
HPSI0214i-rayr_1	Male	75-79
HPSI0314i-xugn_1	Male	65-69
HPSI1013i-kuxp_1	Male	25-29
HPSI0913i-lise_1	Female	45-49
HPSI0114i-vabj_3	Female	50-54
HPSI0114i-eipl_1	Female	40-44
HPSI0214i-pelm_3	Female	40-44
HPSI0214i-kehc_2	Female	65-69
HPSI1013i-pamv_3	Male	65-69
HPSI0214i-eiwy_1	Female	65-69
HPSI1113i-podx_1	Female	65-69
HPSI1213i-xuja_2	Female	45-49
HPSI1213i-hehd_2	Female	60-64
HPSI0314i-fafq_1	Male	60-64
HPSI1113i-qolg_1	Male	35-39

Table 4.1: Demographic information of hiPSC donors.

The sex and age of donors from which hiPSCs were generated are indicated.

was used. For further assessment of sensitivity of TOBIAS by re-analyzing the data in Yang et al., 2019; the code in the "naïve_formative.sh" script was used. To conduct TF footprinting analyses, the code in the "hipsci.sh" script was used. A summary of the scripts and how they relate to one another is provided in "README.md". For statistical analysis of footprinting frequency, the "scatterplot.R" script was used. For statistical analyses of BrainSpan Developmental Transcriptome data, the "brainspan_stats.R" and "brainspan_stats2.R" scripts were used.

4.2.3 Pre-Processing of ATAC-seq Data

For TF footprinting analyses, the computational pipeline described by Bentsen et al, 2020, was adapted for in-house use. For pre-processing of ATAC-seq data, methods were used similar to those used for quality control of RNA-seq analyses (**Chapter 2**). In order to replicate the methods reported by Bentsen et al as closely as possible, the authors of the original TOBIAS were contacted and information about how quality control was conducted was requested (M. Bentsen, personal communications). The ATAC-seq pre-processing pipeline is presented in **Figure 4.3.**

Briefly, adapter sequences were computationally trimmed from sequenced reads using

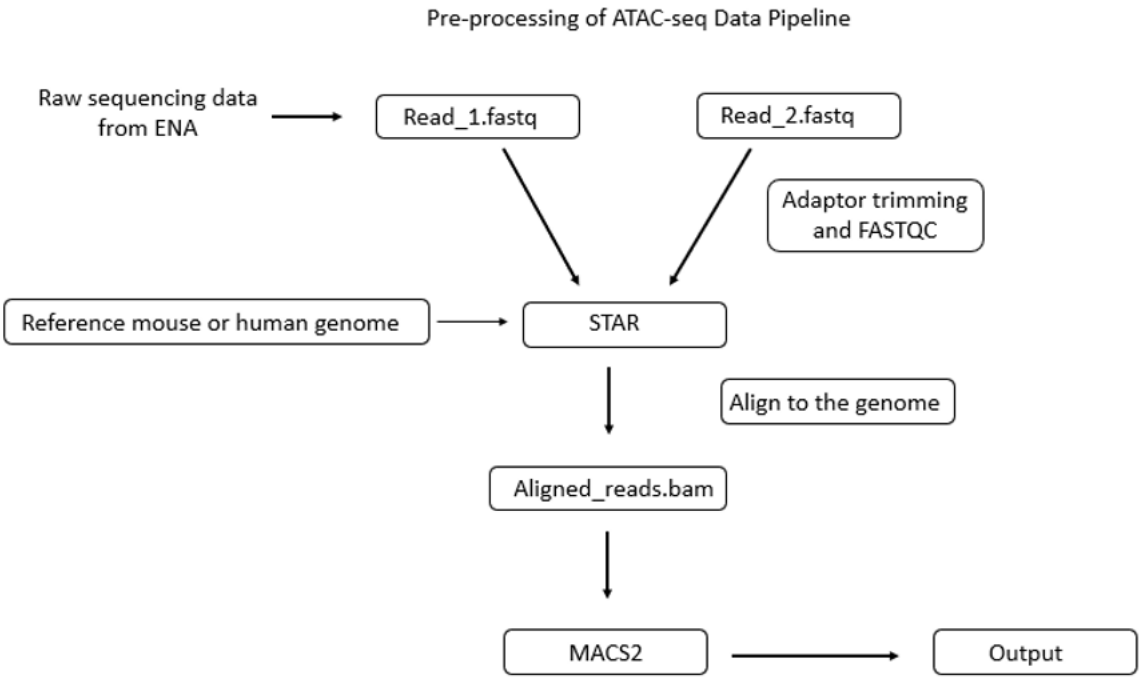


Figure 4.3: Overview of the ATAC-seq data pre-processing pipeline.

The figure shows the scheme of the pre-processing pipeline for ATAC-seq data. A description of the pipeline is provided in the main text.

Trim_galore (version 0.6.5). The quality of the trimmed raw sequencing FASTQ files was evaluated using FASTQC (version 0.10.1) (Andrews). If per base sequencing quality, per sequence quality scores, per sequence GC content, per base N content, and overrepresented sequences were not flagged as problematic FASTQC, the dataset was included in subsequent analyses. None of the sampled datasets failed quality control by these criteria (data not shown).

For the analyses of human preimplantation embryo and human ESCs (to validate TOBIAS), as well as hiPSC-derived neurons (the dataset of interest), trimmed adapter files were aligned to a reference human genome (GRCh38) using STAR (version 2.6.0c) (Dobin et al., 2013). For mouse ESC data (to validate TOBIAS), trimmed adapter files were aligned to a reference mouse genome (GRCm39). STAR parameters were identical to those listed in Bentsen et al, 2020. The samples were checked for a sufficient mapping percentage and adjustment for GC bias did not occur (M. Bentsen, personal communications). The PCR duplicates in the output bam files were removed using Picard (version 2.27.5). Next, the resulting bam files for different biological replicates were merged using samtools (version 1.16.1) (Li et al., 2009). Next, filtering out of chrM reads (mitochondrial reads) occurred by samtools, and removal of all peaks overlapping blacklisted regions was carried out by bedtools (version 2.29.2), as the TOBIAS authors had done for their work

(M. Bentsen et al, 2020) (M. Bentsen, personal communications). Accessible regions were identified by peak calling for each sample using MACS2 (version 2.2.7.1). For analyses of human preimplantation embryos and mouse ESCs, peaks from each sample were merged across all conditions using bedtools (version 2.29.2).

Pre-processing of hiPSC-derived neuron data occurred similarly to as above, with the following modifications. After bam files were generated for each hiPSC-derived neuron dataset, removal of PCR duplicates was conducted using Picard. The resulting bam files for the different hiPSC-derived neuron datasets were then merged using samtools (Li et al., 2009). Similarly to above, samtools was used to filter out chrM reads, and bedtools was deployed to remove all peaks overlapping blacklisted regions. As a single merged bam file was already generated for hiPSC-derived neurons, there was no need to merge peaks for hiPSC-derived neuron analyses. Accessible regions were identified by peak calling of the resulting merged bam file using MACS2.

4.2.4 TF Footprinting Using TOBIAS

For TF footprinting analyses, TOBIAS version 0.14.0 was used (released in October 2022; and installed on the Oxford CCB computer cluster in November 2022) (Bentsen et al, 2020). The footprinting pipeline for TOBIAS is presented in **Figure 4.4**.

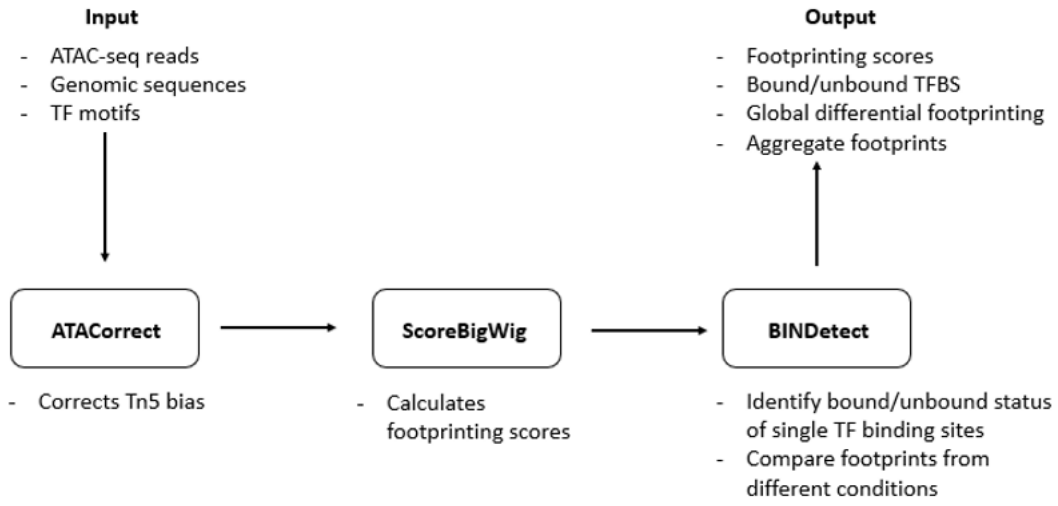


Figure 4.4: TOBIAS footprinting pipeline.

The figure shows the in-house application of the TOBIAS computational pipeline from Bentsen et al., 2020. It depicts input into and output from TOBIAS, along with TOBIAS's different functions. To begin the TOBIAS analysis, ATAC-seq reads, genomic sequences and TF motifs are inputted into ATACorrect, which corrects for Tn5 bias. The output of ATACorrect is then fed into ScoreBigWig, which calculates footprinting scores, which reflect the probability that a TF is occupying a certain genomic position. Finally, the output of ScoreBigWig is fed into BINDetect, which identifies bound and unbound TF binding sites and allows for the comparison of footprints from different conditions. The net output of TOBIAS includes footprinting scores, bound and unbound TF binding sites, global differential footprinting, and visualization of aggregate footprints.

4.2.5 Gene Expression Analyses

To infer the possible functional relevance of TFs, gene expression analyses were conducted. TF levels across hiPSC differentiation and gene expression patterns across human neural development were investigated. The data for expression levels across hiPSC neural differentiation was from the HIPSCI data (the same dataset analyzed in **Chapter 2** and **Chapter 3**). The data for expression levels across human neural development was from the BrainSpan Atlas (Developmental Transcriptome), as described in **Chapter 3**. The code for the statistical analyses of TF expression in fetal vs postnatal samples, as well as early fetal vs late fetal samples, is in the "brainspan_stats.R" and "brainspan_stats2.R" scripts at www.github.com/adelosan521/chapter4.

4.2.6 Inspection of Publicly Available ChIP-seq Data

For inspection of publicly available ChIP-seq data for candidate TFs, processed data files were obtained from publicly available repositories. For the inspection of EBF1, KLF7, REST, TCF4, and ZIC2 binding at VGCCs loci; processed ChIP-seq data files (bigBed file format) were downloaded from the ENCODE portal. Both file types were uploaded into the Integrative Genome Browser (IGV) for visualization.

4.2.7 Statistics

To determine whether TF footprints in VGCC loci occurred at a frequency greater than that which is observed at random genes, a chi-square test was conducted in R. The randomly selected genes were chosen using a script that randomly selects genes from the human genome ("randomgene.R" script at www.github.com/adelosan/chapter3). The code for the Chi-square test in R is the "chisquare.R" script at www.github.com/adelosan521/chapter4. The significance level was set at 0.05. To confirm the results generated from the "chisquare.R" script, an alternative chi-square test calculator was used (https://www.socscistatistics.com/tests/chisquare/default2.aspx). To assess the relationship between TF frequency at VGCC loci and at randomly selected genes, analyses were conducted using the "scatterplot.R" script that calculates the Spearman correlation coefficient, p-value, and generates a scatterplot.

To quantify transcript levels from the BrainSpan Atlas for comparing fetal versus postnatal stages and early versus late fetal stages, similar methods to those described in **Chapter 3** were adopted. First, normality was checked using the Shapiro-Wilk test.

Because not all datasets displayed a normal distribution, the median and interquartile range were calculated. The Mann-Whitney U test was used to determine if there was a significant difference between the groups. If the p-value was less than 0.05, the difference between the groups was considered significant. The code for the above analyses is in the "brainspan_stats.R" and "brainspan_stats2.R" scripts at www.github.com/adelosan521/chapter4.

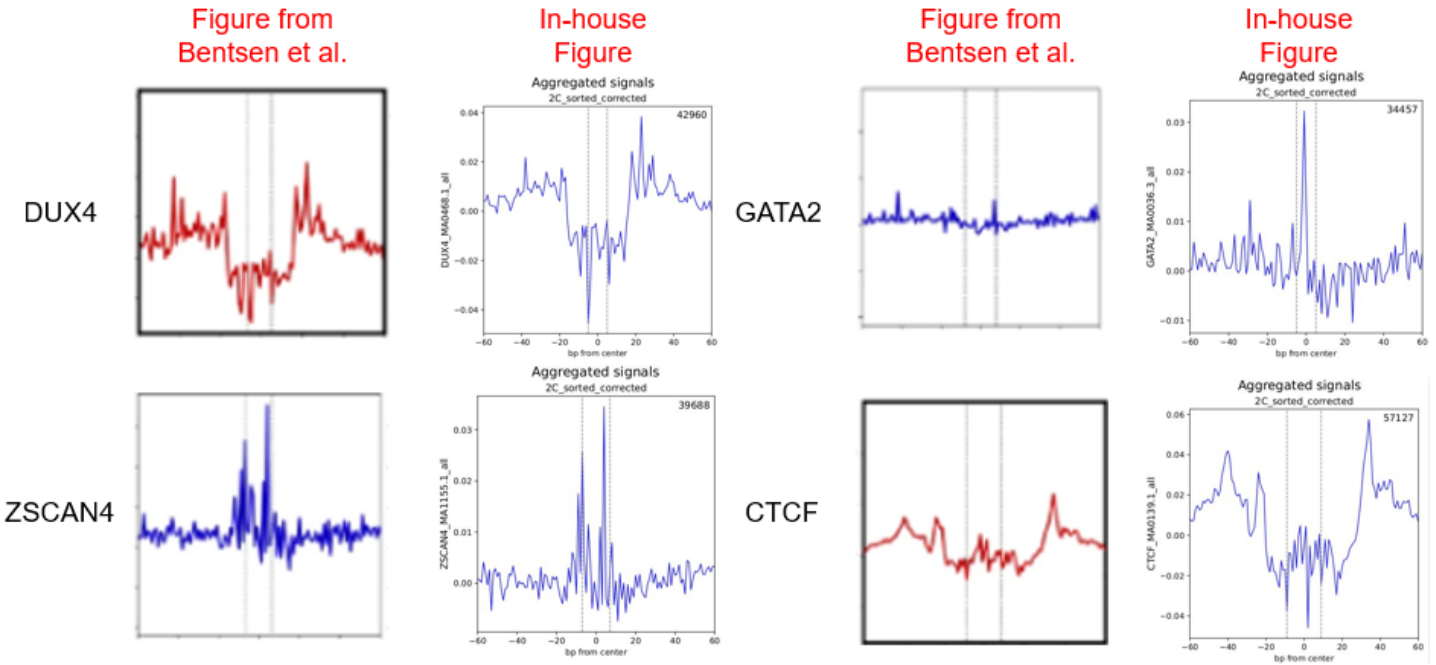
4.3 Results

The aim of the analyses conducted in this Chapter was to identify candidate TFs that regulate VGCC abundance. The results will be divided into two sections. Section **4.3.1** will focus on efforts to test the capabilities and limitations of TOBIAS by replicating published analyses and assessing sensitivity and specificity. As noted earlier, this was necessary as TOBIAS had only been reported in 2020 (Bentsen et al, 2020). **4.3.2**. will describe the results from TF footprinting analyses of hiPSC-derived neurons, along with a validation of TF footprint analyses in ChIP-seq data, differential gene expression analyses, and assessment of candidate TF expression during human neural development.

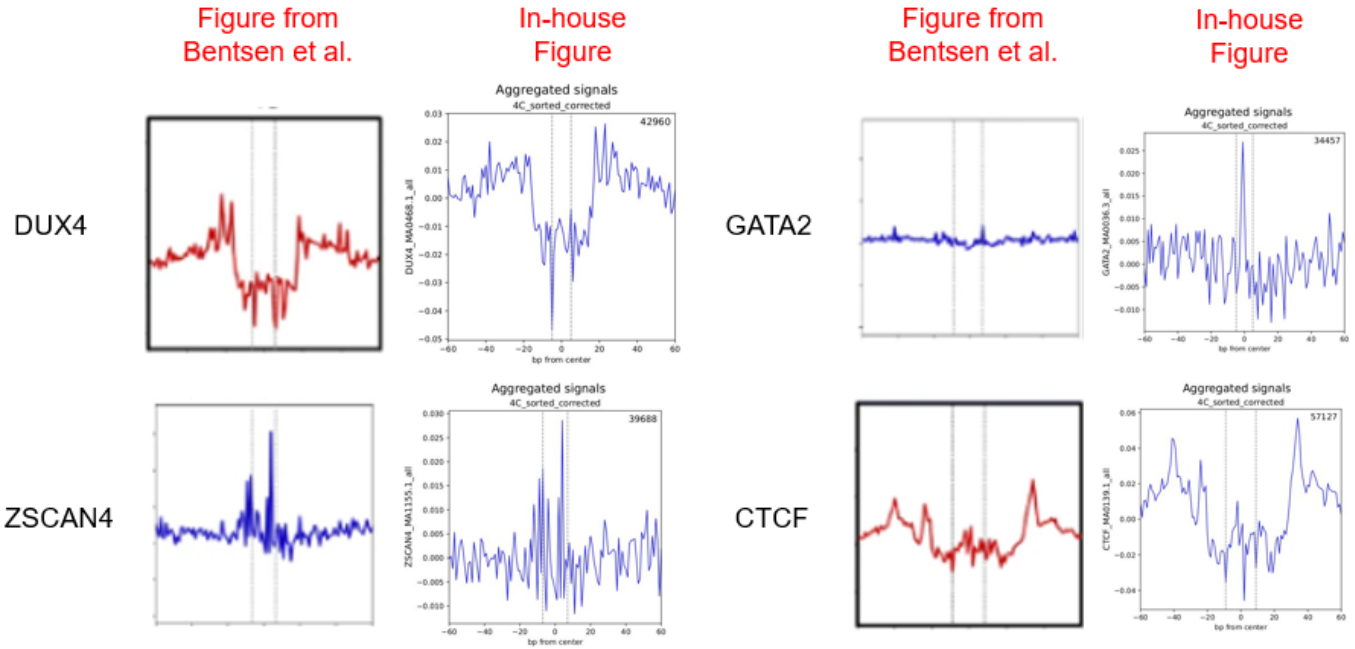
4.3.1 In-house Application of the TOBIAS Computational Pipeline

4.3.1.1 Replication of Published Figures

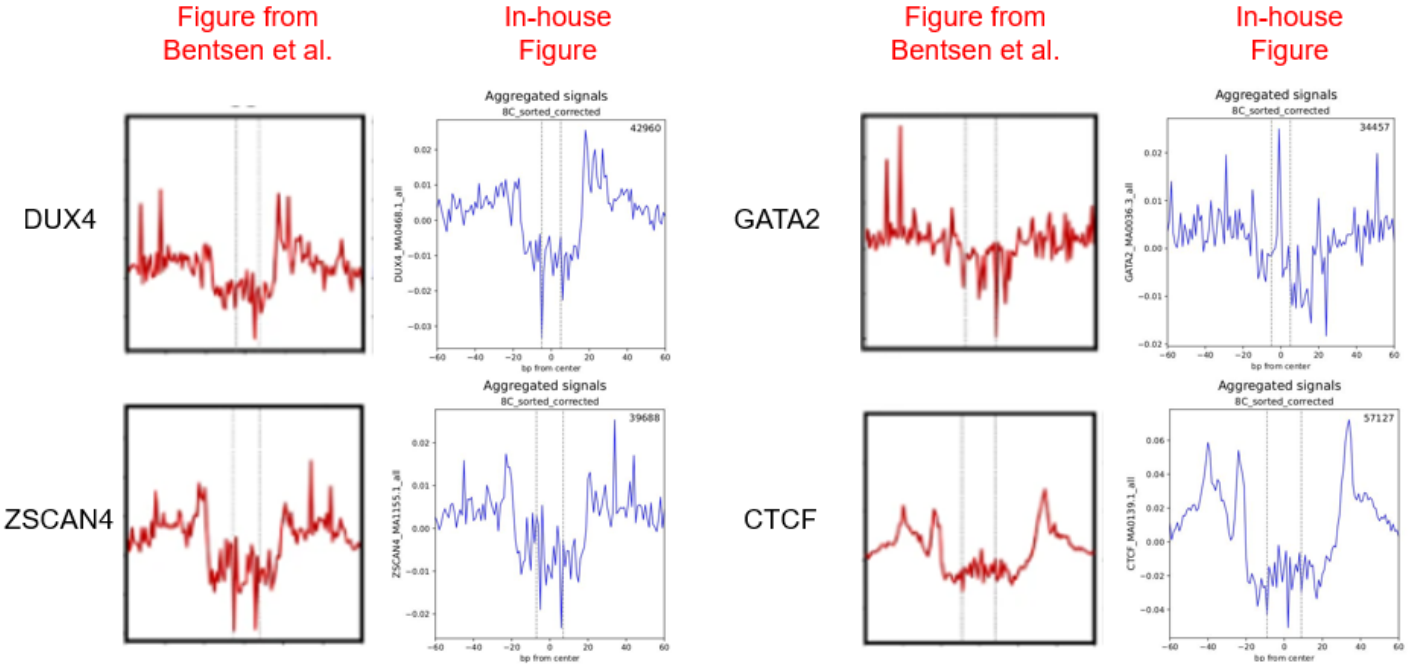
Efforts were first directed at determining whether the published TOBIAS computational pipeline was working appropriately in-house. Attempts were made to re-generate the TOBIAS analyses presented in the original study (Bentsen et al., 2020) when applied to the same dataset (Wu et al., 2018). Attempts to replicate the analyses in Figure 2C from Bentsen et al. 2020 (**Figure 4.5**) were made, including TF aggregate footprints and numbers of binding sites in the genome. The TF aggregate footprints, for instance in the case of the DUX4 aggregate TF footprint, represent all the peaks that harbor a DUX4 motif across the whole genome.


Figure from Bentsen et al.
In-house Figure
Figure from Bentsen et al.
In-house Figure
DUX4
GATA2
ZSCAN4
CTCF
Aggregated signals
2C_sorted_corrected
42960
34457
39688
57127
bp from center

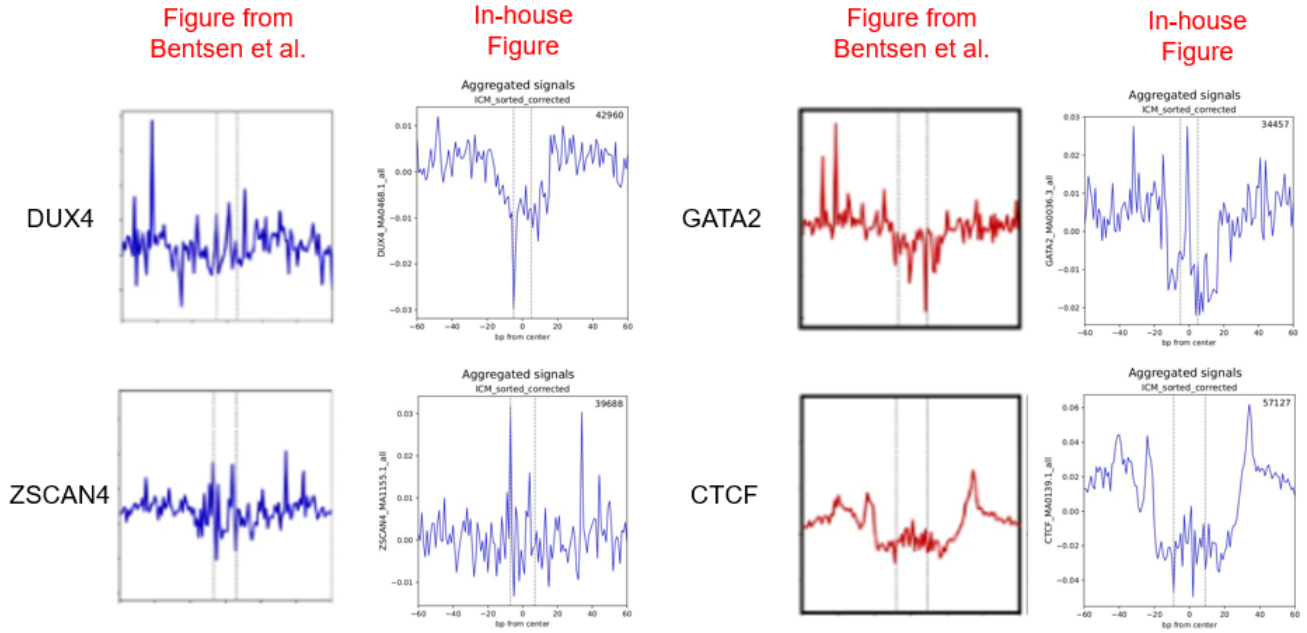
(A)


Figure from Bentsen et al.
In-house Figure
Figure from Bentsen et al.
In-house Figure
DUX4
GATA2
ZSCAN4
CTCF
Aggregated signals
4C_sorted_corrected
42960
34457
39688
57127
bp from center

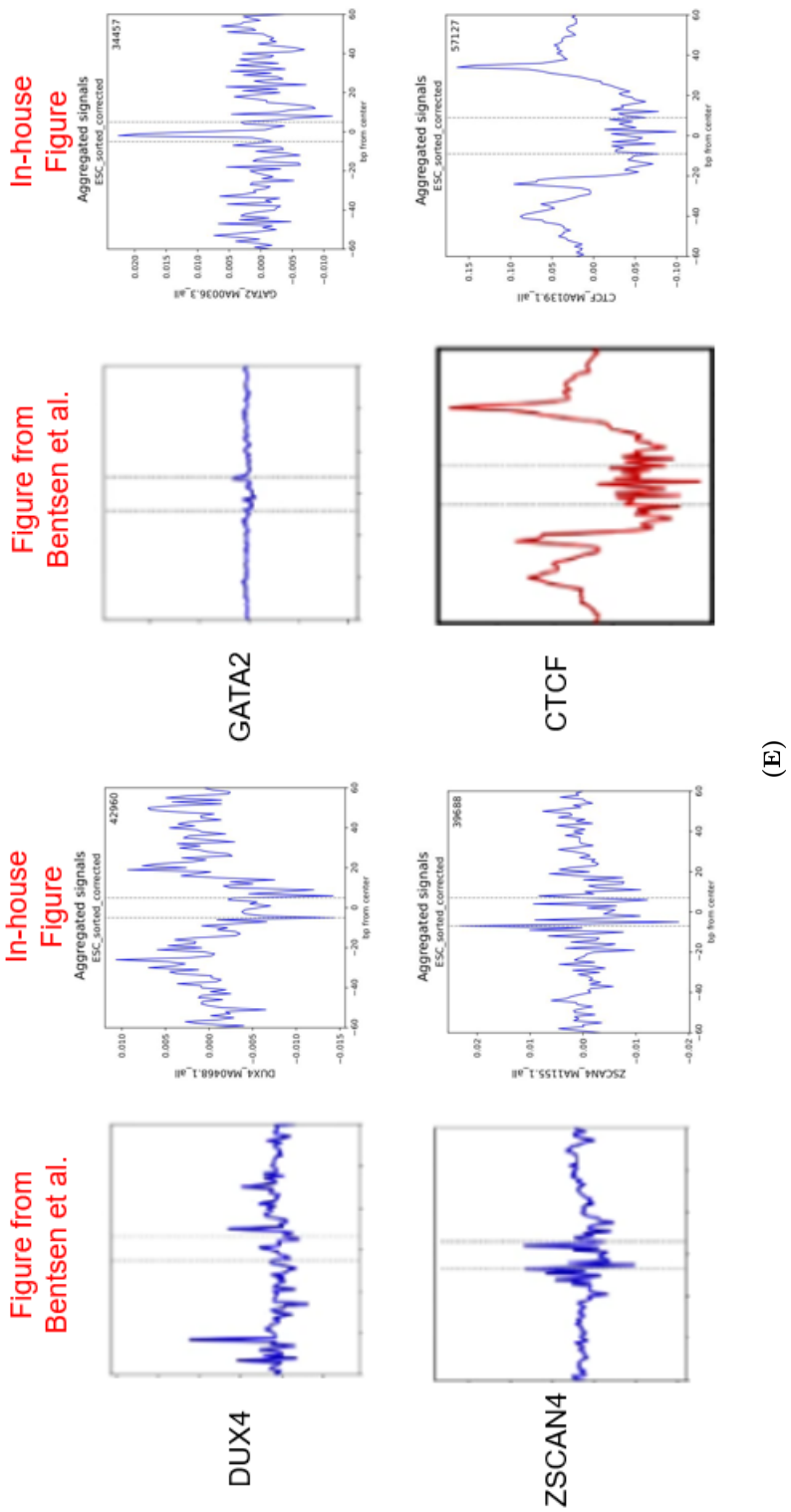
(B)


Figure from Bentsen et al.
In-house Figure
Figure from Bentsen et al.
In-house Figure
DUX4
ZSCAN4
GATA2
CTCF
Aggregated signals
8C_sorted_corrected
42960
39688
34457
57127
bp from center

(C)


Figure from Bentsen et al.
In-house Figure
Figure from Bentsen et al.
In-house Figure
DUX4
ZSCAN4
GATA2
CTCF
Aggregated signals
ICM_sorted_corrected
42960
39688
34457
57127
bp from center

(D)


In-house Figure
Figure from Bentsen et al.
GATA2
CTCF
DUX4
ZSCAN4
Aggregated signals
ESC_sorted_corrected
bp from center

(E)

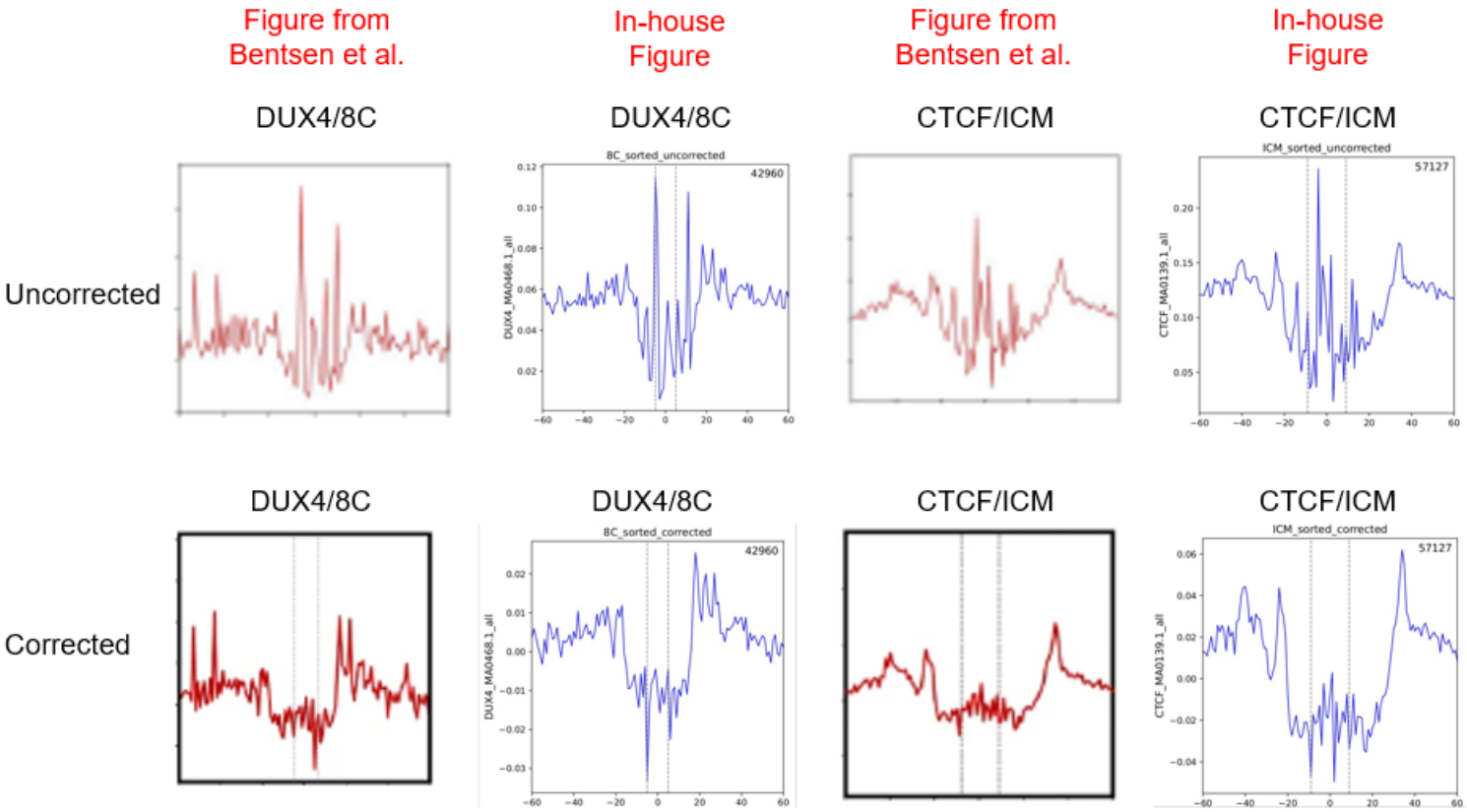


Figure 4.5: In-house reproduction of TOBIAS analyses published in Bentsen et al. (2020).

Attempted replication of Bentsen et al. (2020) figures through in-house application of the published TOBIAS computational pipeline. For (A), (B), (C), (D), and I, TF footprints for DUX4, ZSCAN4, GATA2, and CTCF during human preimplantation development are from Bentsen et al. (2020). Figures from Bentsen et al. and those generated in-house are indicated. (A) TF aggregate footprints from 2C stage embryos; (B) TF aggregate footprints from 4C stage embryos; (C) TF aggregate footprints from 8C stage embryos; (D) TF aggregate footprints from ICM stage embryo; (E) TF aggregate footprints from ESCs; and (F) replicating uncorrected and Tn5 corrected footprint figures from Bentsen et al. TOBIAS involves a correction step for Tn5 bias to improve the visualization of TF footprints. The aim in (F) was to assess whether the pre- and post-correction TF footprint curves could be replicated in-house. Visualizing both uncorrected (top) and corrected aggregate (bottom) footprints for DUX4 at the 8C stage and CTCF at the ICM stage. Individual plots are centered around binding motifs. The numbers in the top right of in-house figures relate to the number of binding sites. The scales for the Bentsen et al. figures were not available, so the similarity of their curve to the in-house footprint curves was assessed. Regarding binding sites, the number of DUX4 binding sites was 42,628 (Bentsen et al.) and 42,960 (in-house); the number of ZSCAN binding sites was 52,777 (Bentsen et al.) and 39,688 (in-house); the number of GATA2 binding sites was 47,168 (Bentsen et al.) and 34,457 (in-house); and the number of CTCF binding sites was 85,490 (Bentsen et al.) and 57,127 (in-house).

While many of the aggregate footprints were similar to Bentsen et al. (2020) (19 similar footprint plots out of 24 total footprint plots), some aggregate footprints appeared dissimilar, such as in the case of GATA2 footprint plots (5 dissimilar footprint plots out of 24 total footprint plots). Additionally, the number of binding sites that were analyzed was similar in the case of DUX4 (1 out of 4 genes) but dissimilar in others (3 out of 4 genes). These differences between the figures generated by Bentsen et al. and the figures generated in-house could be caused by differences in the versions of tools used between Bentsen et al. and in-house application of the TOBIAS pipeline, such as the use of different versions of TOBIAS (M. Bentsen, personal communications). TOBIAS has been updated 31 times since its original publication in 2020 (M. Bentsen, personal communications). Unfortunately, the older version of TOBIAS was no longer available, so an exact replication of the Bentsen et al. method is not possible. Altogether, while unable to entirely replicate the published results, it was possible to adapt the published TOBIAS computational pipeline for in-house use.

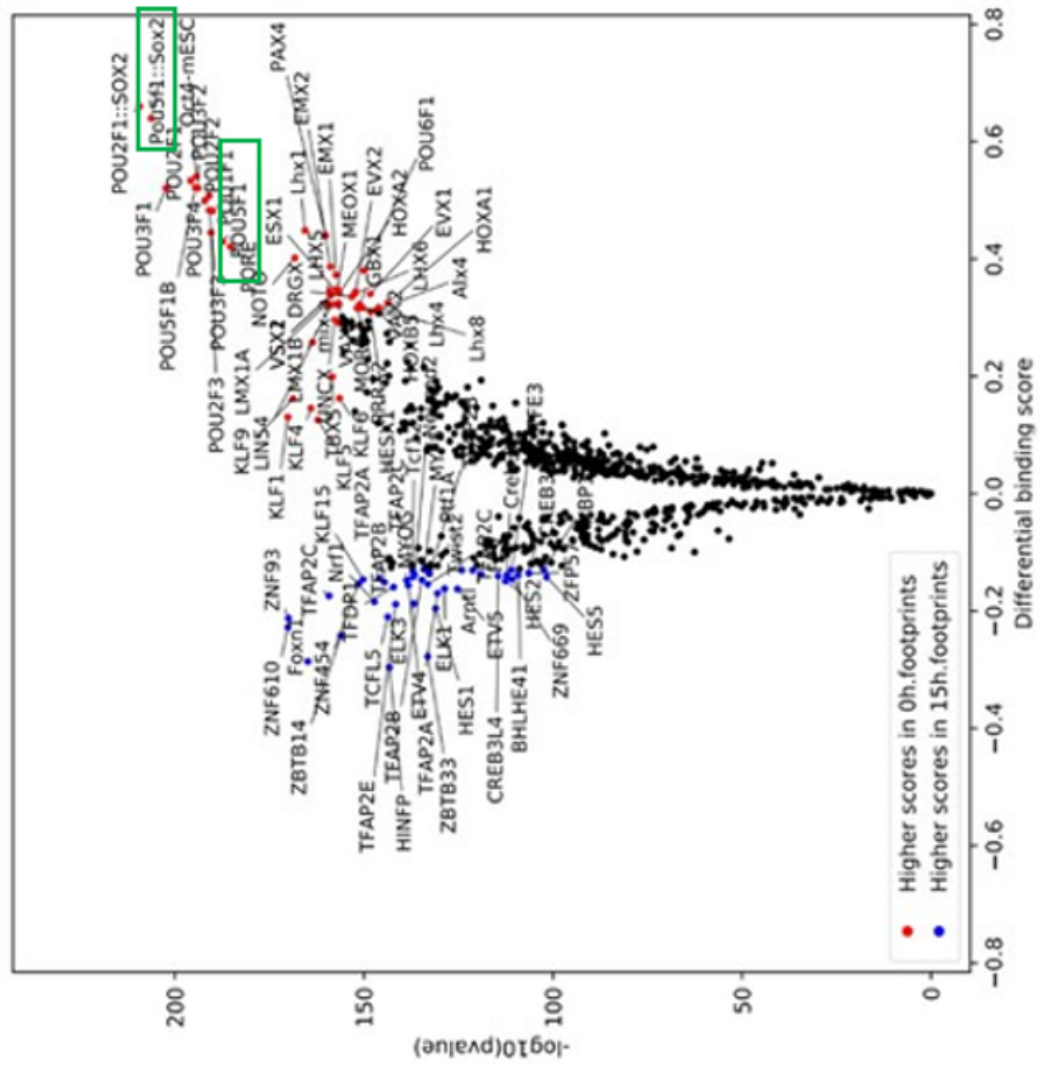
4.3.1.2 Sensitivity and Specificity of TOBIAS

Next, the sensitivity (the ability to detect a change when a known perturbation has occurred) and specificity (a lack of false positive results) of TOBIAS were assessed.

For sensitivity, the capacity was assessed of TOBIAS to detect motif enrichments of a single TF (and related TFs) after its expression was experimentally modulated. I attempted to replicate the findings of a previous study (Xiong et al., 2022), in which ATAC-seq data were generated from an ESC line, ZHBTc4, which had been genetically engineered to repress the expression of the TF Oct4 (also known as Pou5f1) upon doxycycline treatment. In Xiong et al. 2022, the addition of doxycycline for 15 hours was sufficient to cause a reduction in Oct4 protein levels.

Oct4 forms a heterodimer with Sox2 (Rodda et al., 2005); thus, the ability to detect a loss of Oct4 and Oct4/Sox2 motif enrichment upon Oct4 depletion was assessed. As observed in **Figure 4.6**, when comparing 0-hour (untreated) and 15-hour (Oct4-depleted) ATAC-seq data, TOBIAS detected enrichment of Oct4 and Oct4/Sox2 motifs in the 0-hour samples relative to the 15-hour samples. This indicates TOBIAS is sensitive enough to detect changes in motif enrichment of a single TF and its heterodimer upon downregulation of that single TF. Moreover, the figures generated are similar to those published by Xiong et al. (2022).

Figure from Xiong et al.



(A)

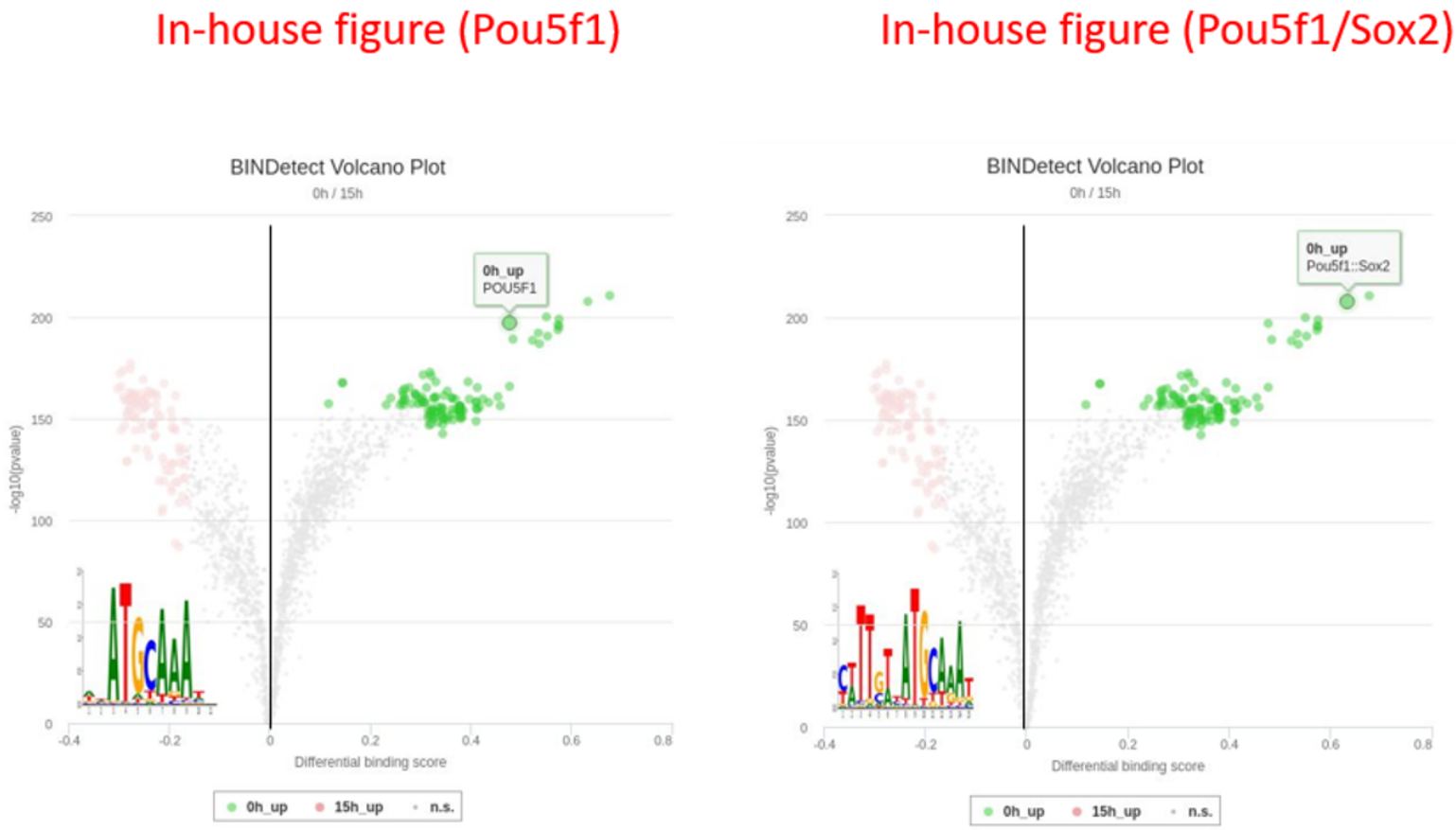
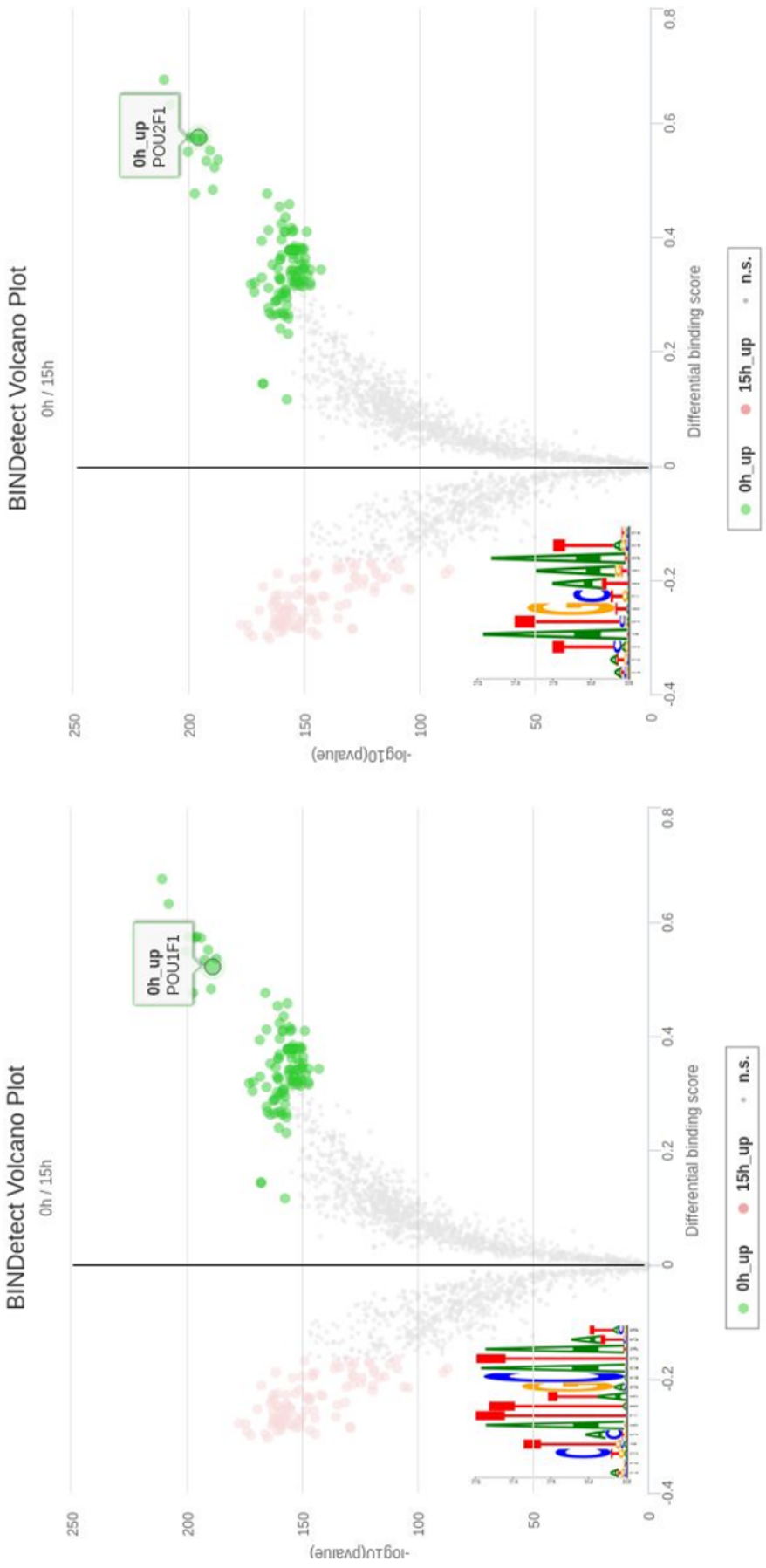


Figure 4.6: Assessing the sensitivity of TOBIAS by studying mouse ESCs genetically engineered to repress Oct4 expression upon doxycycline treatment.

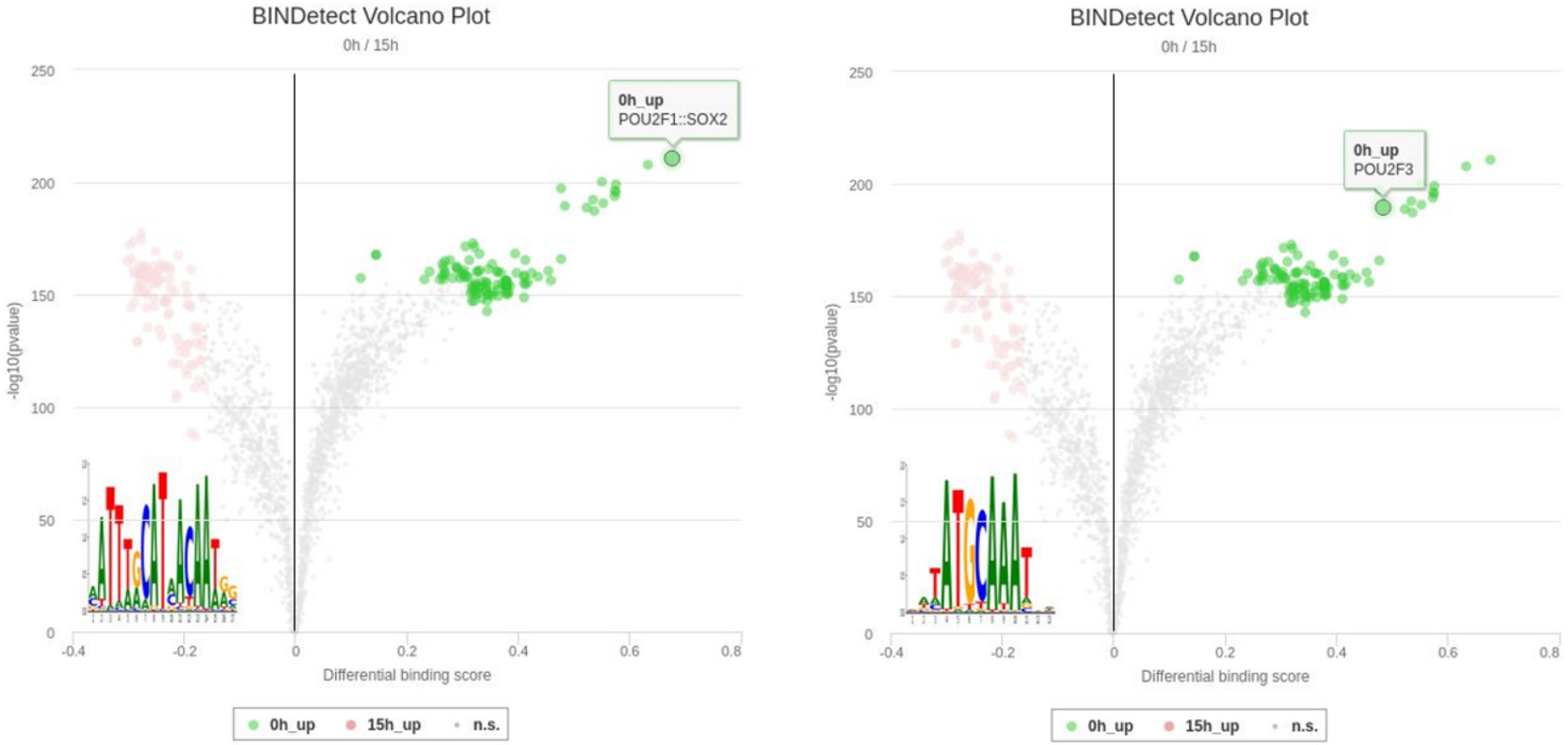
BINDetect plots represent the pairwise comparison of TF binding between different conditions. The volcano plots show differential binding activity against -log10(p-value) of investigated TF motifs; each dot represents one motif. Comparison of data from 0-hour (0h) (untreated) and 15-hour (15h) doxycycline-treated ZHBTc4 mouse ESCs (Xiong et al., 2022). (A) BINDetect figure from Xiong et al. (2022); (B) in-house figures generated using Xiong et al. data and in-house application of the TOBIAS computational pipeline. (B) (left) Differential TF motif enrichment between 0h and 15h samples. Oct4/Pou5f1 enrichment in 0h samples is indicated. (B) (right) Differential TF motif enrichment between 0h and 15h samples. Pou51/Sox2 motif enrichment in 0h samples is indicated.

The specificity of TOBIAS was assessed next. Oct4 belongs to a family of TFs containing POU domains. In addition to the enrichment of Oct4 and Oct4/Sox2 motifs, TOBIAS also reported 0-hour enrichment for other POU domain-containing factors, including Pou1f1, Pou2f1, Pou2f2, Pou2f3, Pou3f1, Pou3f2, among others, including the Pou2f1/Sox2 heterodimer (**Figure 4.7**). Footprints for TFs that were not expressed were detected. Aside from Oct4, only Pou2f1 (also known as Oct1) was expressed in mouse ESCs. These data indicate that TOBIAS has limited specificity in discriminating between TFs with high similarity (i.e. different TFs from the same TF family). However, the knowledge of the context (i.e. Oct4 depletion) suggests that TOBIAS is detecting changes in the motif enrichment of the relevant TF.

Withdrawal of doxycycline results in the recovery of Oct4 protein levels (Xiong et al., 2022). Whether TOBIAS would be sensitive enough to detect changes in motif enrichment upon upregulation of a single TF was next investigated. The capacity of TOBIAS to detect gain of Oct4 and Oct4/Sox2 motif enrichment upon upregulation of Oct4 was assessed. TOBIAS detected pronounced enrichment of Oct4 and Oct4/Sox2 motifs in the 4.5-hour recovery (Oct4-mid) samples relative to the 15-hour treated (Oct4-low) samples (**Supplementary Figure 3.1** in **Appendix III**). The in-house figures generated in **Supplementary Figure 3.1.** also resemble the corresponding ATAC-seq figure in Xiong et al. 2022. However, in line with a potentially reduced ability of TOBIAS to discriminate between similar TFs, gain of other POU domain factors was also observed, including Pou1f1, Pou2f2, Pou3f1, and Pou3f2, among others (**Supplementary Figure 3.1**).


BINDetect Volcano Plot
0h / 15h
0h_up
POU2F1
-log10(pvalue)
Differential binding score
0h_up
15h_up
n.s.
BINDetect Volcano Plot
0h / 15h
0h_up
POU1F1
-log10(pvalue)
Differential binding score
0h_up
15h_up
n.s.

(A)



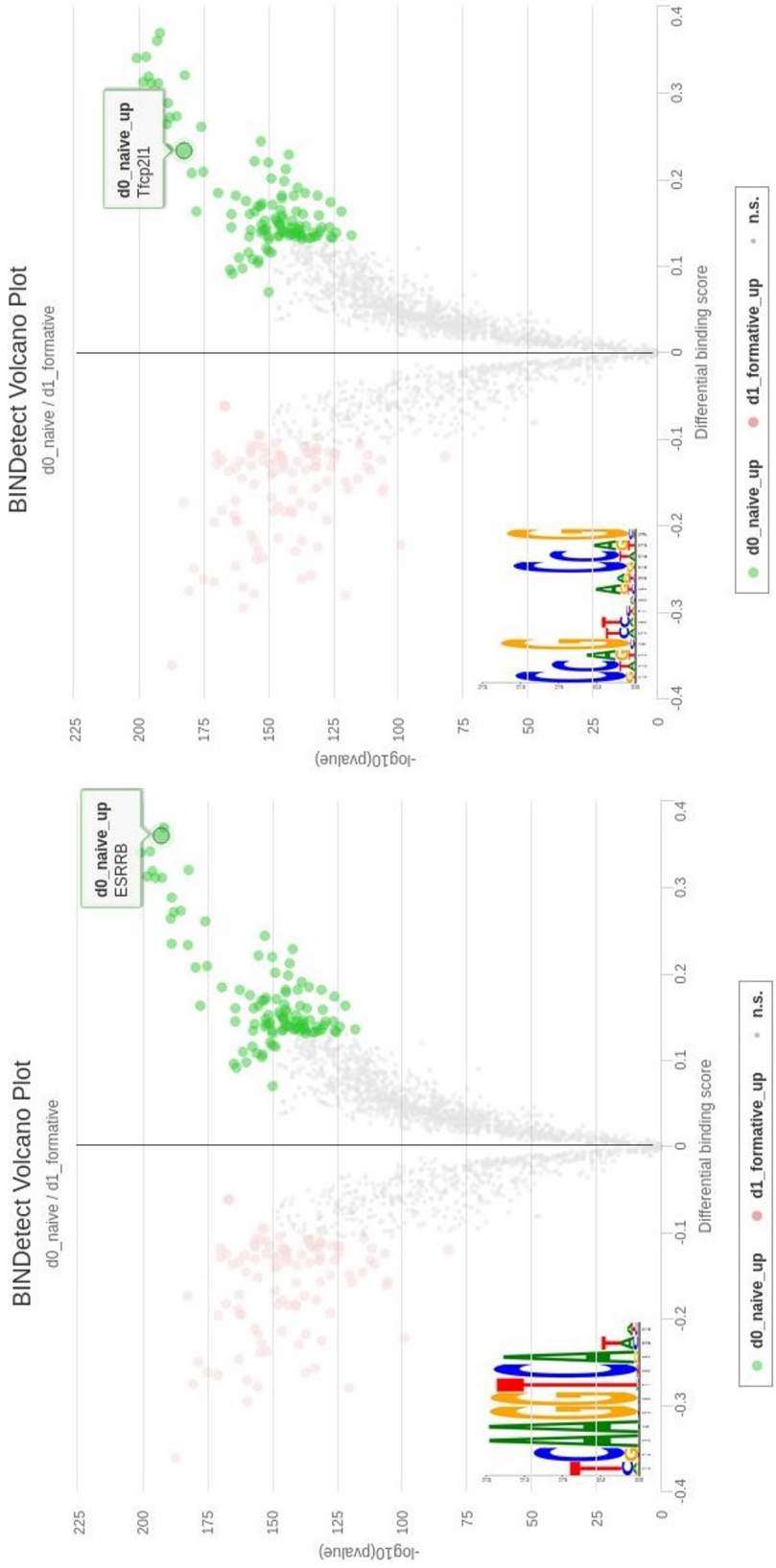
(B)

Figure 4.7: Assessing the specificity of TOBIAS by studying mouse ESCs genetically engineered to repress Oct4 expression upon doxycycline treatment.

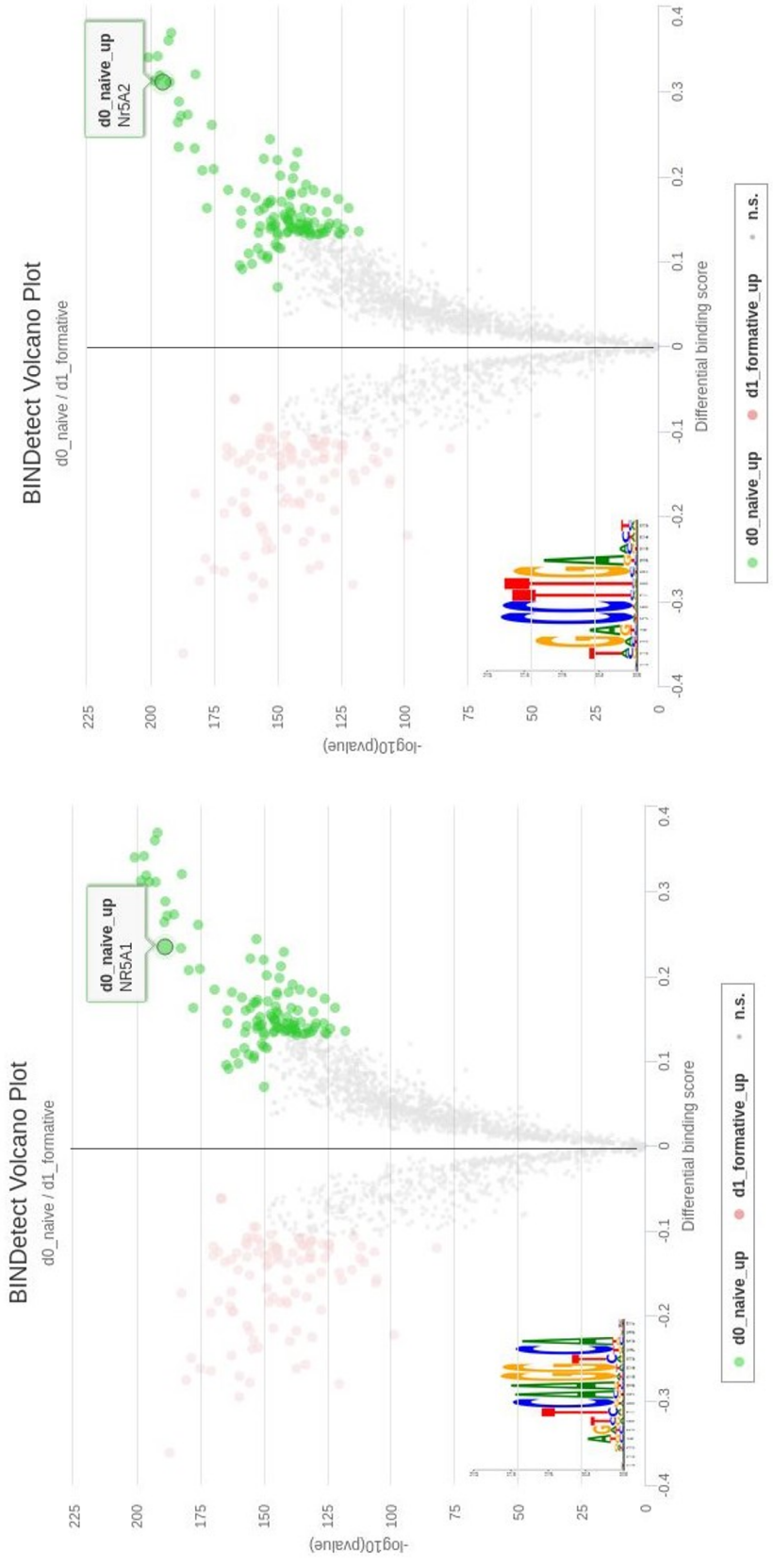
See **Figure 4.6** for the interpretation of BINDetect plots. Comparison of data from 0-hour (0h) treated (untreated) and 15-hour (15h) doxycycline-treated ZHBTc4 mouse ESCs (Xiong et al., 2022). (left) BINDetect figure from Xiong et al. (2022). A and B show differential TF motif enrichment between 0h and 15h samples. In (A) Pou1f1 (left) and Pou2f1 (right) motif enrichment are shown. In (B) heterodimer Pou2f1/Sox2 (left) and Pou2f3 (right) motif enrichment are shown.

The capacity of TOBIAS to identify expected TF motif enrichment changes in a well-characterized cell state transition was also assessed. Chromatin accessibility data of mouse ESCs exiting from the naïve pluripotent state were re-analyzed (Yang et al., 2019; MacCarthy et al., 2022). Similarly to MacCarthy et al., 2022, two time points were compared: day 0 of differentiation, when naïve pluripotency gene expression is active; and day 1, when stem cells are leaving naïve pluripotency and entering a "formative" pluripotency state. The TFs that are specifically and functionally linked to naïve pluripotency have been reported in the literature (Dunn et al., 2014). The capacity of TOBIAS to detect the enrichment of TFs published as linked to naïve pluripotency in day 0 samples relative to day 1 samples was tested. TOBIAS detected enrichment of footprints for Esrrb, Tfcp2l1, Nr5a2, Nr5a1, Kl5, and Klf4 in day 0 samples relative to day 1 samples (**Figure 4.8**). All six TFs have been linked to naïve pluripotency (i.e. the day 0 cell state) (Jiang et al., 2008; Hall et al., 2009; Guo and Smith, 2010; Martello et al., 2012; Martello et al., 2013). Therefore, it appears TOBIAS is able, in an unbiased manner, to detect TF motif enrichment changes that would be expected given published literature (MacCarthy et al., 2022).

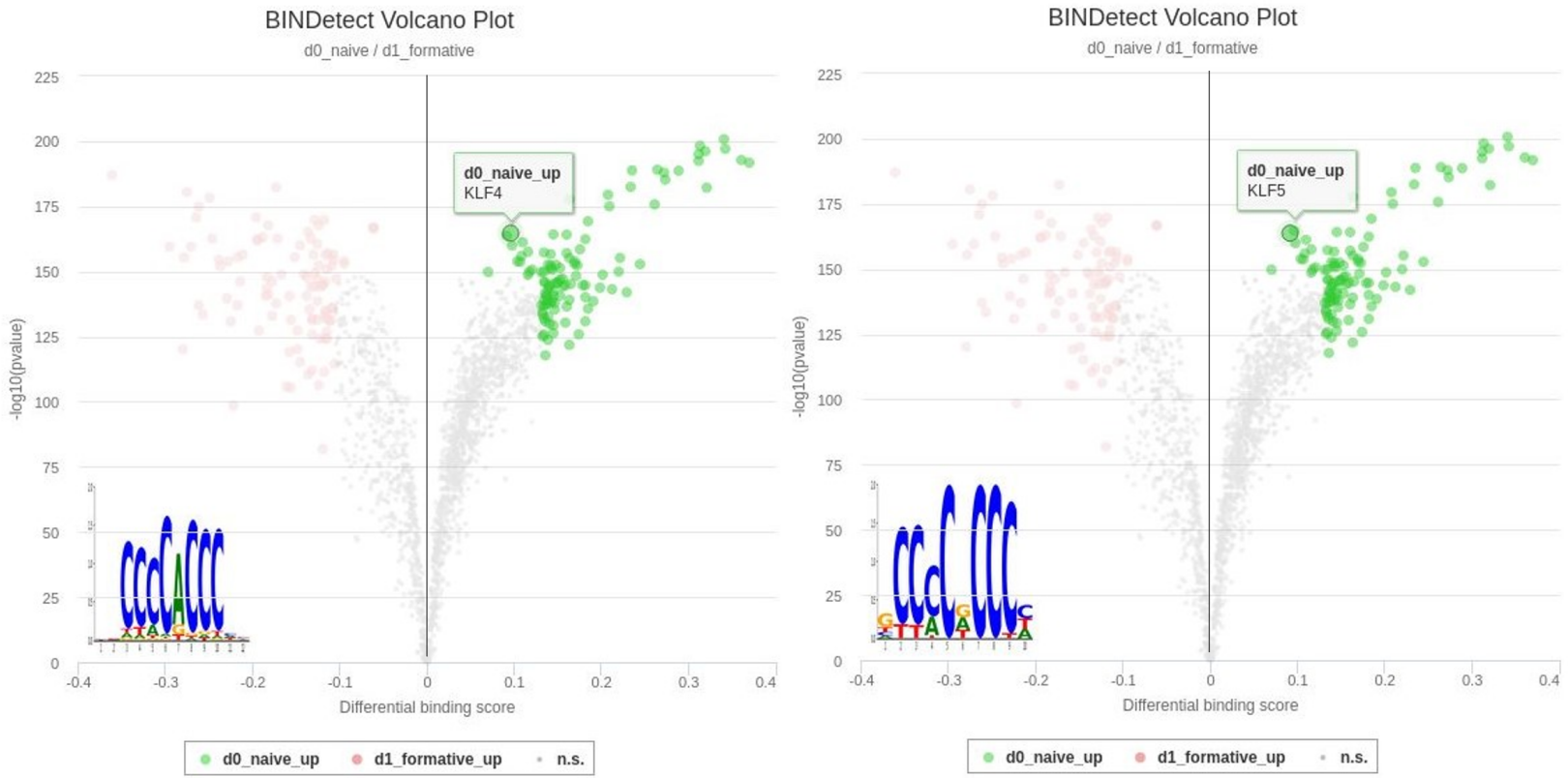
Altogether, these analyses suggest that TOBIAS is sensitive enough to detect changes in motif enrichment upon the experimental modulation of expression levels of a single TF. Moreover, TOBIAS was able to detect a number of expected changes in TF motif enrichment in a well-characterized cell state transition (Yang et al., 2019). However, TOBIAS appears to lack some specificity in being able to discriminate between TFs from the same family.


BINDetect Volcano Plot
d0_naive / d1_formative
d0_naive_up
Tfcp2l1
-log10(pvalue)
Differential binding score
d0_naive_up
d1_formative_up
n.s.
BINDetect Volcano Plot
d0_naive / d1_formative
d0_naive_up
ESRRB
-log10(pvalue)
Differential binding score
d0_naive_up
d1_formative_up
n.s.

(A)


BINDetect Volcano Plot
d0_naive / d1_formative
d0_naive_up
NR5A1
-log10(pvalue)
Differential binding score
d0_naive_up
d1_formative_up
n.s.
BINDetect Volcano Plot
d0_naive / d1_formative
d0_naive_up
Nr5A2
-log10(pvalue)
Differential binding score
d0_naive_up
d1_formative_up
n.s.

(B)



(C)

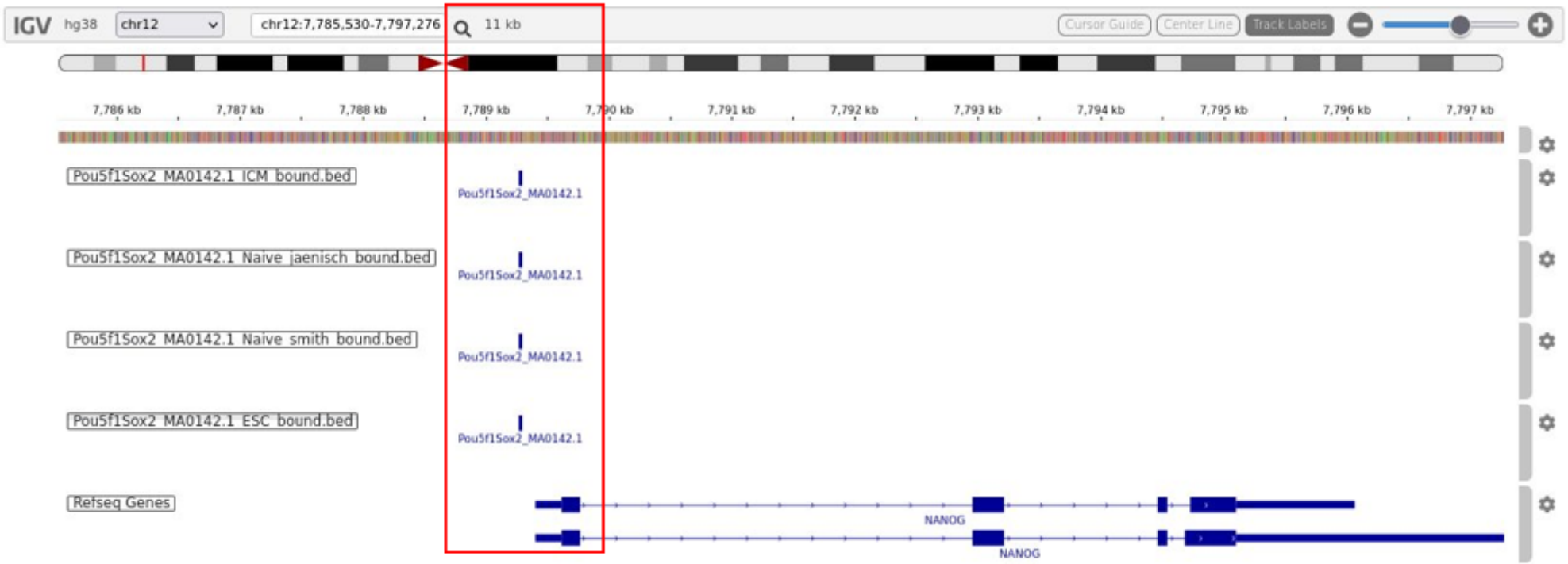
Figure 4.8: Assessing TOBIAS capacity to detect the enrichment of known TFs in a well-characterized cell state transition

See **Figure 4.6** for an explanation of how to interpret BINDetect plots. Comparison between naïve ("d0_naive") and formative ("d1_formative") chromatin accessibility data. (A) BINDetect volcano plot indicating differential TF motif enrichment between d0_naive and d1_formative samples. Esrrb (left) and Tfcp2l1 (right) motif enrichments in d0_naive samples are indicated. (B) BINDetect volcano plot indicating differential TF motif enrichment between d0_naive and d1_formative samples. Nr5a1 (left) and Nr5a2 (right) motif enrichments in d0_naive samples are indicated. C) BINDetect volcano pIot indicating differential TF motif enrichment between d0_naive and d1_formative samples. Klf4 (left) and Klf5 (right) motif enrichments in d0_naive samples are indicated.

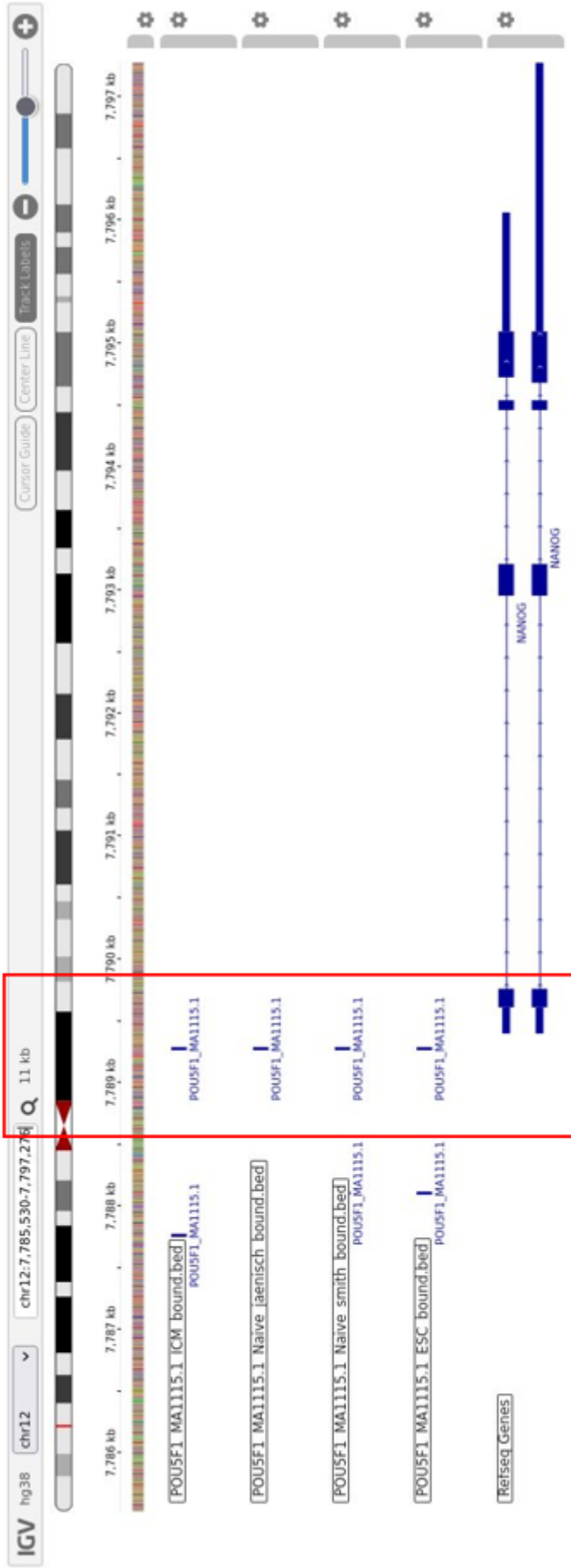
4.3.1.2.1 Sensitivity and Specificity for Detecting Individual Binding Events

The sensitivity of TOBIAS for identifying individual TF binding events was evaluated next. To achieve this, whether TOBIAS could detect known binding events in pluripotent cells that are well-established in the literature.

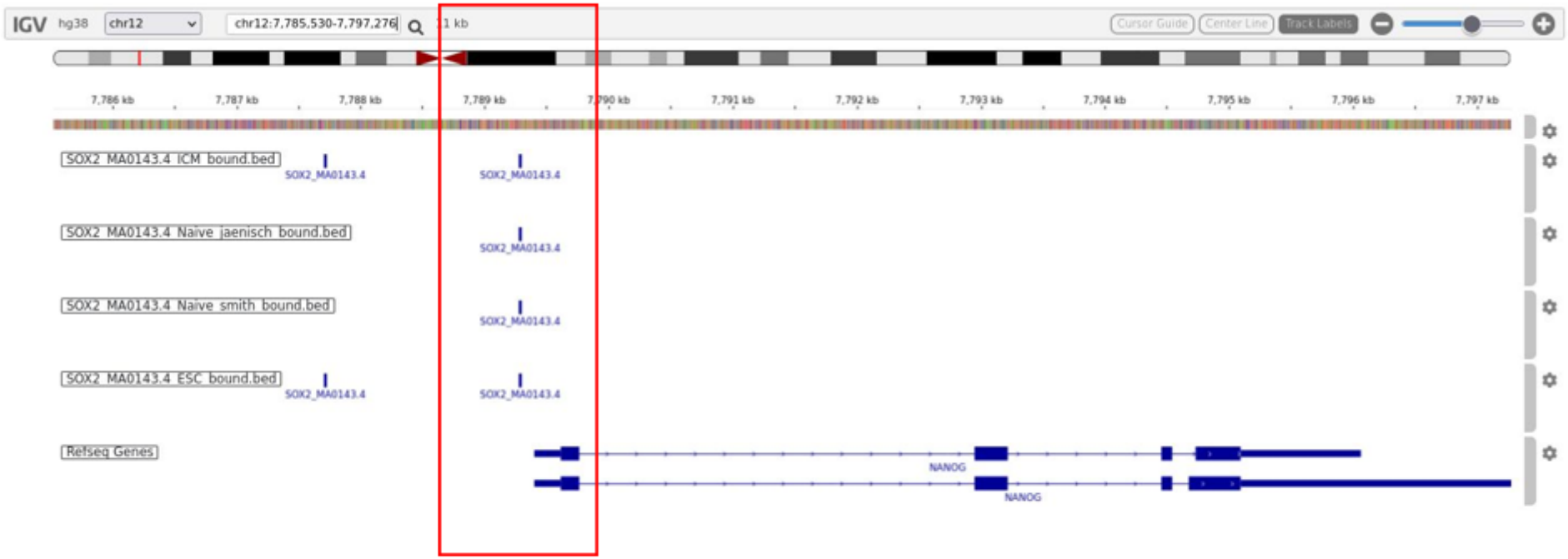
Through ChIP assays, OCT4 and SOX2 have been identified as binding regulatory regions at the promoter of NANOG, another master regulator of pluripotency, in PSCs (Rodda et al., 2005). OCT4/SOX2 occupancy at the NANOG genomic locus was additionally assessed in the different samples of human PSCs (**Figure 4.9**) (Wu et al., 2018)**.** TOBIAS identified OCT4/SOX2 occupancy near the NANOG TSS in all of the different human pluripotent samples (Wu et al., 2018). TOBIAS also identified occupancy of OCT4 and SOX2 as "monomers" near the NANOG TSS. These data appear consistent with the report of Oct4/Sox2 binding at the Nanog promoter in PSCs (Rodda et al., 2005). A close examination of the OCT4/SOX2 binding site within 200 bp upstream of the NANOG TSS elaborated TF occupancy at the same, precise location where Oct4/Sox2 motif site had been reported in different mammalian species and within the genomic region where it was possible to detect OCT4 and SOX2 binding by ChIP quantitative PCR in mouse and human ESCs (Rodda et al., 2005).


IGV
hg38
chr12
chr12:7,785,530-7,797,276
11 kb
Cursor Guide
Center Line
Track Labels
7,786 kb
7,787 kb
7,788 kb
7,789 kb
7,790 kb
7,791 kb
7,792 kb
7,793 kb
7,794 kb
7,795 kb
7,796 kb
7,797 kb
Pou5f1Sox2_MA0142.1_ICM_bound.bed
Pou5f1Sox2_MA0142.1
Pou5f1Sox2_MA0142.1_Naive_jaenisch_bound.bed
Pou5f1Sox2_MA0142.1
Pou5f1Sox2_MA0142.1_Naive_smith_bound.bed
Pou5f1Sox2_MA0142.1
Pou5f1Sox2_MA0142.1_ESC_bound.bed
Pou5f1Sox2_MA0142.1
Refseq Genes
NANOG
NANOG

(A)


IGV
hg38
chr12
chr12:7,785,530-7,797,276
11 kb
Cursor Guide
Center Line
Track Labels
7,786 kb
7,787 kb
7,788 kb
7,789 kb
7,790 kb
7,791 kb
7,792 kb
7,793 kb
7,794 kb
7,795 kb
7,796 kb
7,797 kb
POU5F1_MA1115.1 ICM bound.bed
POU5F1_MA1115.1
POU5F1_MA1115.1 Naive jaenisch bound.bed
POU5F1_MA1115.1
POU5F1_MA1115.1 Naive smith bound.bed
POU5F1_MA1115.1
POU5F1_MA1115.1 ESC bound.bed
POU5F1_MA1115.1
Refseq Genes
NANOG
NANOG

(B)



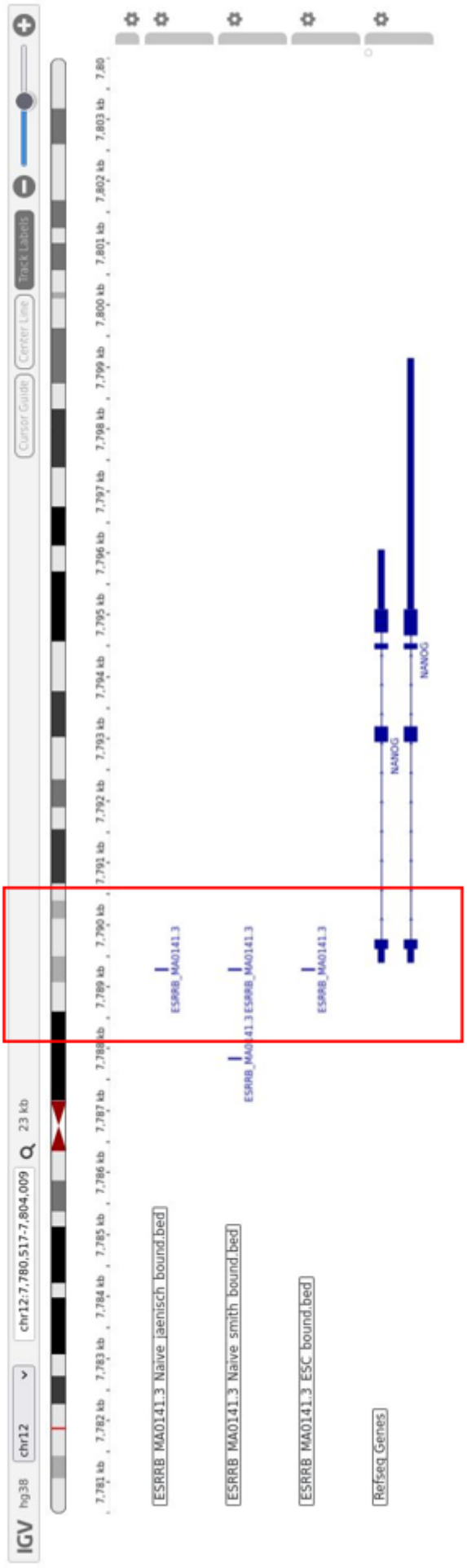
(C)

Figure 4.9: Assessing whether TOBIAS can detect a known TF binding event at the NANOG locus.

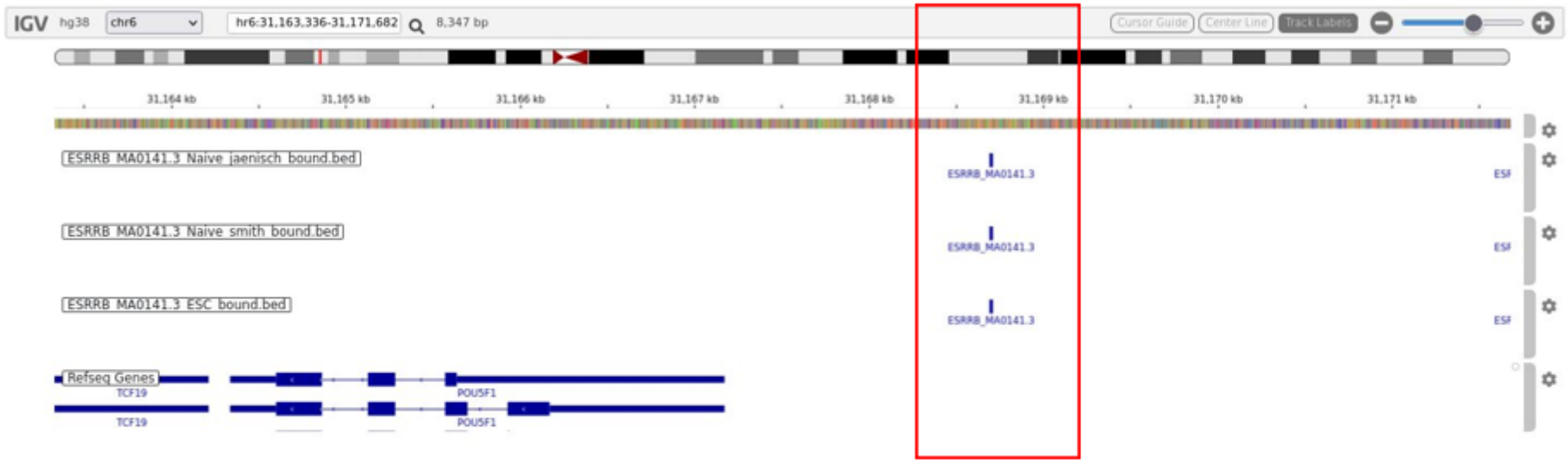
Integrative genome browser view of TOBIAS footprinting analysis in the vicinity of the NANOG locus. The top track shows human ICM samples; the middle two tracks are from human naïve PSCs cultured in two different laboratories; and the bottom track shows conventional ESCs. Pou5f1Sox2_MA0142.1 indicates Pou5f1/Sox2 motif; POU5F1_MA1115.1 indicates OCT4/POU5F1 motif. The thin blue bars indicate sites of TF occupancy. (A) OCT4/SOX2 occupancy is upstream of NANOG TSS in four different human pluripotent cell samples. (B) OCT4 occupancy upstream of NANOG TSS in four different human pluripotent cell samples. (C) SOX2 occupancy upstream of NANOG TSS in four different human pluripotent cell samples. The red box in (A), (B), and (C) highlights TF footprints.

The publicly available version of TOBIAS does not readily allow for identifying all TF footprints at a given open chromatin region, so a minor modification was made to enable efficient and systematic identification of all TF footprints at a given accessible chromatin region. The modification was a single line of code and it is described at www.github.com/adelosan521/chapter4/embryo.sh (lines 25 - 27). Applying this modification, several TFs were identified with footprints found at the open chromatin peak at the human OCT4 distal enhancer in human naïve PSCs. In mouse naïve PSCs, the Oct4 distal enhancer is a genomic region bound by naïve pluripotency TFs (Chen et al., 2008). This analysis of human naïve PSCs identified binding of OCT4/SOX2, which had already been identified above based on prior assumptions from the literature, and a number of other TFs, including TFAP2C, a known regulator of human naïve pluripotency (Pastor et al., 2018). Therefore, the slightly modified TOBIAS computational pipeline is suitable for identifying putative TF regulators in an unbiased manner.

As TOBIAS proved sensitive to individual TF occupancy events, the specificity of TOBIAS for individual TF occupancy events was assessed, by investigating whether it would detect a TF footprint for a lowly or non-expressed TF. ESRRB is a master regulator of pluripotency in mice and regulates Oct4 and Nanog through binding to their regulatory regions (Feng et al., 2009). However, it is not expressed in naïve or conventional human PSCs (Takashima et al., 2014). However, TOBIAS still detected ESRRB footprints at human OCT4 and NANOG loci near where Esrrb binds at the murine Oct4 and Nanog loci at the NANOG promoter and near the OCT4 distal enhancer (**Figure 4.10**). These data indicate that ESRRB might still be occupying the genome at very low expression levels in human PSCs or that potentially another TF is occupying the location where TOBIAS is detecting an ESRRB footprint (for example, another ESRR family member). These data suggest a limited specificity of TOBIAS for individual TF occupancy.


IGV
hg38
chr12
chr12:7,780,517-7,804,009
23 kb
Cursor Guide
Center Line
Track Labels
ESRRB MA0141.3 Naive jaenisch bound.bed
ESRRB MA0141.3 Naive smith bound.bed
ESRRB MA0141.3 ESC bound.bed
Refseq Genes
NANOG
NANOG

(A)



(B)

Figure 4.10: Assessing whether TOBIAS can detect TF binding when a TF is not expressed.

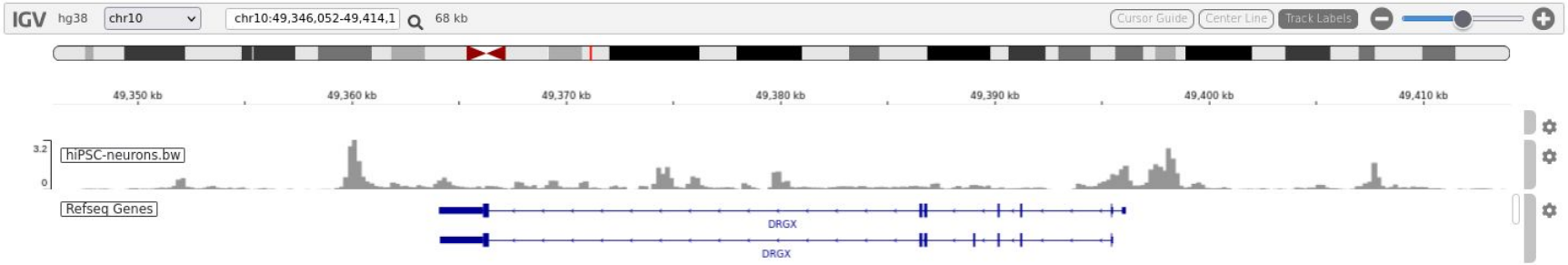
Integrative genome browser view of TOBIAS footprinting analysis in the vicinity of the POU5F1 and NANOG loci locus. The top two tracks from human naïve PSCs cultured in two different laboratories. The bottom track shows conventional ESCs. ESRRB_MA0101.43 Indicates ESRRB motif. The thin blue bars indicate sites of TF occupancy. (A) ESRRB "occupancy" upstream of NANOG TSS in three different human pluripotent cell samples. (B) ESRRB "occupancy" upstream of OCT4 TSS in three different human pluripotent cell samples. The red box highlights TF footprints.

4.3.2 TF Footprinting Analyses of hiPSC-Derived Neurons

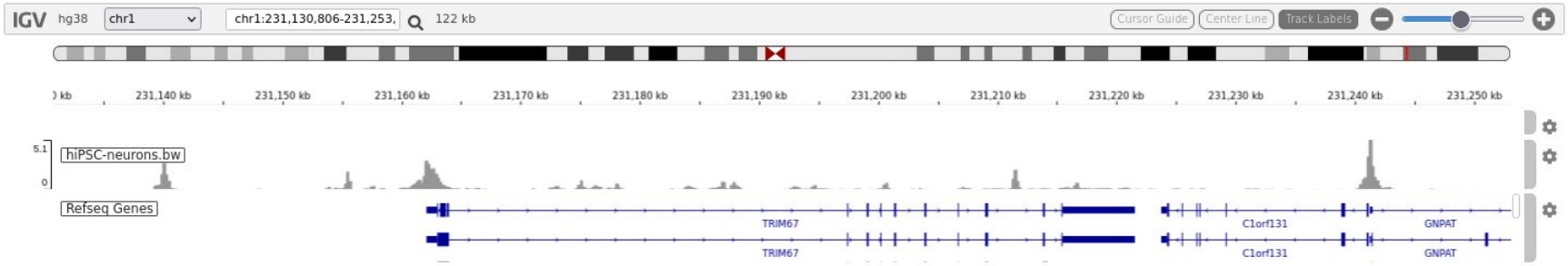
The second part of the results section of this Chapter will describe the identification of candidate TFs that regulate VGCC abundance. Having validated its suitability, TOBIAS was applied to the HIPSCI ATAC-seq data from hiPSC-derived neurons. ATAC-seq data from 23 different individual donors were merged and analyzed (i.e. peaks were called). Open chromatin peaks were analyzed for the presence of bound TF footprints. Open chromatin peaks occurring within the VGCC locus TSS were identified and analyzed. TFs which showed footprints at VGCC loci and differential expression between hiPSCs and hiPSC-derived neurons were identified. To search for potential master regulators of VGCC loci and to reduce the number of putative regulators further, attention was focused on TFs that bound to five or more VGCC loci.

4.3.2.1 Accessible Chromatin Near VGCC TSS

As a quality control for ATAC-seq, an open chromatin peak is expected at the TSS of expressed genes (Yan et al., 2020). The TSS was identified based on UCSC gene annotation for the longest transcript. The top genes more highly expressed in hiPSC-derived neurons were surveyed (DRGX, LINC02735, TRIM67, CARTPT, VIP, PRPH, ONECUT3, NSG2, and GRIA2) (identified in **Chapter 2**), and all the surveyed loci showed accumulation of ATAC-seq signal at their TSS (**Figure 4.11**, data not shown). Representative illustrations at the DRGX and TRIM67 loci are shown in **Figure 4.11**.



(A) DRGX TSS

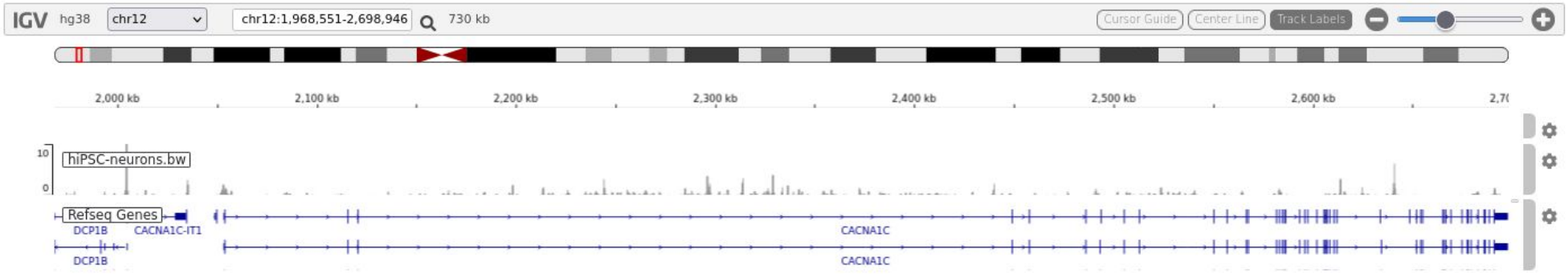


(B) TRIM67 TSS

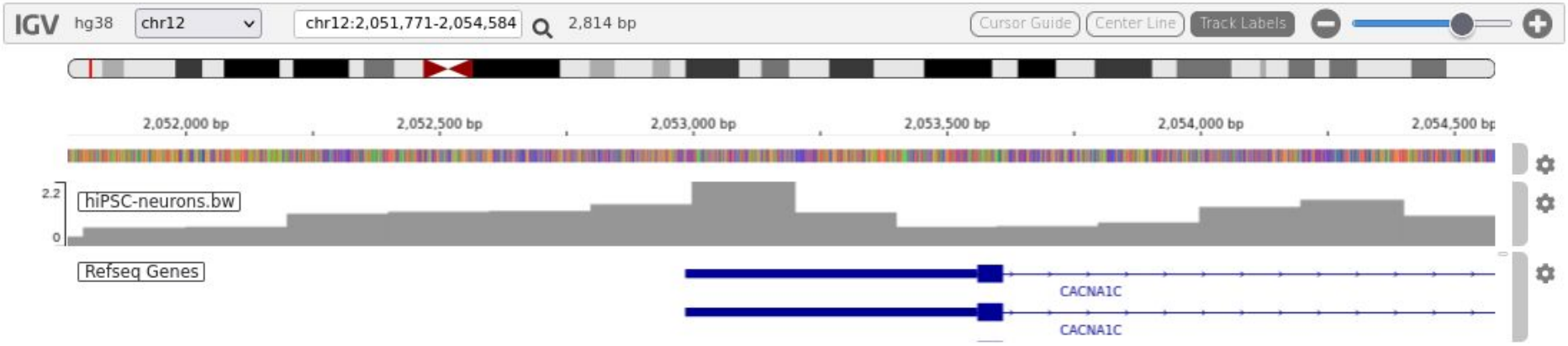
Figure 4.11: Accessible chromatin at the TSS of top hiPSC-derived neuron-enriched genes.

IGV view of ATAC-seq signal at entire loci of genes more highly expressed in hiPSC-derived neurons. (A) DRGX, (B) TRIM67. hiPSC-neurons.bw indicates the ATAC-seq signal from hiPSC-derived neurons (".bw" indicates bigwig file).

The presence of an open chromatin peak at the TSS of VGCC loci was also confirmed. A representative example is CACNA1C (**Figure 4.12).** Open chromatin regions at the TSS were also present at the other VGCC loci (Supplementary Figures 3.2 - 3.15).



(A) CACNA1C locus



(B) CACNA1C TSS

Figure 4.12: CACNA1C accessible chromatin regions.

(A) IGV view of ATAC-seq signal at the entire CACNA1C locus. (B) IGV view of ATAC-seq signal at CACNA1C TSS (chromatin peak 79071).

4.3.2.2 Bound TF Sites Near VGCC TSS

TF footprints at the TSS of VGCC loci were identified (**Table 4.2**). TOBIAS was able to identify bound TF footprints at all VGCC loci except CACNA1A, CACNA1H, CACNA1I and CACNA2D4. To better understand why TOBIAS might fail to identify TFs at certain loci, the first author of the original TOBIAS paper was contacted and she said this is related to signal strength at those loci (M. Bentsen, personal communications).

4.3.2.3 Identifying Potential Regulators by Differential Gene Expression

To refine the list of TFs that might regulate VGCCs, a search was conducted for those that were differentially expressed between hiPSCs and hiPSC-derived neurons (using DESEQ2 data generated in **Chapter 2**). The results of intersecting the TOBIAS footprinting analysis with the DESEQ2 differential gene expression analyses are presented in **Table 4.2** and **Table 4.3**.

VGCC subunit	Number of bound footprints identified	Number of human TFs	Number of differentially expressed TFs
CACNA1B	419	114	21
CACNA1C	850	135	26
CACNA1D	316	67	15
CACNA1E	91	20	5
CACNA1G	625	111	18
CACNA2D1	1590	185	46
CACNA2D2	125	12	3
CACNA2D3	92	44	7
CACNB1	37	25	0
CACNB2	1058	152	24
CACNB3	1114	175	31
CACNB4	591	215	35

Table 4.2: Summary of TOBIAS TF footprinting analyses.

The table indicates the number of bound footprints identified, human TFs, and differentially expressed TFs for each VGCC subunit. Differential expression was scored as a DESEQ2 baseMean greater than 100 and a log2FoldChange less than -1.5 or greater than 1.5. TOBIAS did not generate any bound TF footprints for CACNA1A, CACNA1H, CACNA1I, or CACNA2D4.

TOBIAS detected thousands of footprints at 12 VGCC loci. After focusing on the human TFs that differentially expressed, 80 unique TFs were identified. Notably, despite the detection of TF footprints, no TFs were detected as differentially expressed at CACNB1.

A) CACNA1C (log2FoldChange 4.17 [higher in hiPSC-derived neurons])

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	padj	hiPSCs	hiPSC-neuron
ASCL1	ENSG00000139352	106.402	5.810	0.440	9.65E-40	4.25E-39		Higher
CAMTA1	ENSG00000171735	1786.490	1.987	0.555	0.0003443	0.000528		Higher
CAMTA2	ENSG00000108509	558.093	2.827	0.120	9.74E-124	1.53E-122		Higher
ELK4	ENSG00000158711	567.521	-1.8134	0.129	1.24E-44	5.96E-44	Higher	
KLF15	ENSG00000163884	160.404	-3.022	0.172	1.60E-68	1.14E-67	Higher	
KLF7	ENSG00000118263	1436.37	2.700	0.111	5.94E-131	1.03E-129		Higher
LEF1	ENSG00000138795	297.372	3.662	0.211	1.35E-67	9.52E-67		Higher
MAFA	ENSG00000182759	787.447	6.279	0.301	6.61E-97	7.33E-96		Higher
MEIS1	ENSG00000143995	166.046	6.302	0.409	1.93E-53	1.09E-52		Higher
MEIS2	ENSG00000134138	1270.242	6.984	0.422	1.57E-61	1.01E-60		Higher
MEIS3	ENSG00000105419	1557.049	2.831	0.115	8.42E-134	1.52E-132		Higher
NR2F1	ENSG00000175745	672.143	6.280	0.397	2.76E-56	1.63E-55		Higher
PRDM5	ENSG00000138738	225.829	-2.712	0.128	4.75E-99	5.44E-98	Higher	
SOX5	ENSG00000134532	154.071	2.358	0.215	5.91E-28	2.00E-27		Higher
SP1	ENSG00000185591	1398.115	-1.941	0.101	5.78E-83	5.21E-82	Higher	
SP8	ENSG00000164651	151.621	-1.815	0.630	.00397	0.005	Higher	
TCF3	ENSG00000071564	2707.850	-2.302	0.181	4.43E-37	1.84E-36	Higher	
TCF4	ENSG00000196628	1269.598	-1.535	0.259	2.63E-09	5.22E-09	Higher	
TCF7L1	ENSG00000152284	1249.047	-3.246	0.305	2.10E-26	6.86E-26	Higher	
TCF7L2	ENSG00000148737	730.074	-2.574	0.200	5.26E-38	2.24E-37	Higher	
TFAP2A	ENSG00000137203	251.937	5.305	0.334	1.15E-56	6.81E-56		Higher
TFAP2C	ENSG0000087510	121.161	-2.533	0.326	7.73E-15	1.86E-14	Higher	
ZBTB7B	ENSG00000160685	241.070	-2.213	0.247	3.69E-19	1.00E-18	Higher	
ZIC5	ENSG00000139800	837.856	-5.007	0.446	3.14E-29	1.10E-28	Higher	
ZNF324	ENSG00000083812	155.615	2.214	0.072	3.01E-209	1.30E-207		Higher
ZNF93	ENSG00000184635	390.593	-2.474	0.219	1.08E-29	3.82E-29	Higher	

B) CACNA1B (log2FoldChange 3.95 [higher in hiPSC-derived neurons])

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	Padj	hiPSCs	hiPSC-neuron
EBF1	ENSG00000164330	729.426	8.79	0.209	0	0		Higher
EBF3	ENSG00000108001	432.284	6.058	0.152	0	0		Higher
ELK4	ENSG00000158711	567.521	-1.8134	0.129	1.24E-44	5.96E-44	Higher	
HEY1	ENSG00000164683	369.332	4.570	0.204	2.14E-110	2.86E-109		Higher
HEY2	ENSG00000135547	112.981	-2.795	0.240	3.20E-31	1.17E-30	Higher	
INSM1	ENSG00000173404	110.160	2.675	0.353	3.44E-14	8.13E-14		Higher
KLF15	ENSG00000163884	160.404	-3.022	0.172	1.60E-68	1.14E-67	Higher	
KLF7	ENSG00000118263	1436.37	2.700	0.111	5.94E-131	1.03E-129		Higher
MLXIPL	ENSG00000009950	117.211	-1.971	0.215	4.30E-20	1.20E-19	Higher	
PRDM5	ENSG00000138738	225.829	-2.712	0.128	4.75E-99	5.44E-98	Higher	
REST	ENSG00000084093	916.440	-2.065	0.142	6.37E-48	3.26E-47	Higher	
SNAI2	ENSG00000019549	101.190	2.924	0.302	3.08E-22	9.06E-22		Higher
SP1	ENSG00000185591	1398.115	-1.941	0.101	5.78E-83	5.21E-82	Higher	
SP8	ENSG00000164651	151.621	-1.815	0.630	.00397	0.005	Higher	
TCF3	ENSG00000071564	2707.850	-2.302	0.181	4.43E-37	1.84E-36	Higher	
TCF4	ENSG00000196628	1269.598	-1.535	0.259	2.63E-09	5.22E-09	Higher	
TFAP2A	ENSG00000137203	251.937	5.305	0.334	1.15E-56	6.81E-56		Higher
TFAP2C	ENSG00000087510	121.161	-2.533	0.326	7.73E-15	1.86E-14	Higher	
TP53	ENSG00000141510	1360.724	-1.735	0.200	4.03E-20	1.12E-19	Higher	
ZEB1	ENSG00000148516	614.388	4.083	0.250	8.64E-60	5.38E-59		Higher
ZNF93	ENSG00000184635	390.593	-2.474	0.219	1.08E-29	3.82E-29	Higher	

C) CACNA1D (log2FoldChange 4.23 [higher in hiPSC-derived neurons])

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	Padj	hiPSCs	hiPSC-neuron
ASCL1	ENSG00000139352	106.402	5.810	0.440	9.65E-40	4.25E-39		Higher
EBF1	ENSG00000164330	729.426	8.79	0.209	0	0		Higher
EBF3	ENSG00000108001	432.284	6.058	0.152	0	0		Higher
KLF15	ENSG00000163884	160.404	-3.022	0.172	1.60E-68	1.14E-67	Higher	
KLF7	ENSG00000118263	1436.37	2.700	0.111	5.94E-131	1.03E-129		Higher
MEIS2	ENSG00000134138	1270.242	6.984	0.422	1.57E-61	1.01E-60		Higher
MEIS3	ENSG00000105419	1557.049	2.831	0.115	8.42E-134	1.52E-132		Higher
SOX2	ENSG00000181449	2792.861	-3.430	0.394	2.99E-18	7.92E-18	Higher	
SP1	ENSG00000185591	1398.115	-1.941	0.101	5.78E-83	5.21E-82	Higher	
TCF3	ENSG00000071564	2707.850	-2.302	0.181	4.43E-37	1.84E-36	Higher	
TCF4	ENSG00000196628	1269.598	-1.535	0.259	2.63E-09	5.22E-09	Higher	
TFAP2A	ENSG00000137203	251.937	5.305	0.334	1.15E-56	6.81E-56		Higher
TFAP2C	ENSG00000087510	121.161	-2.533	0.326	7.73E-15	1.86E-14	Higher	
ZIC5	ENSG00000139800	837.856	-5.007	0.446	3.14E-29	1.10E-28	Higher	
ZNF93	ENSG00000184635	390.593	-2.474	0.219	1.08E-29	3.82E-29	Higher	

D) CACNA1E (log2FoldChange 5.63 [higher in hiPSC-derived neurons])

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	Padj	hiPSCs	hiPSC-neuron
BHLHE40	ENSG00000134107	400.686	2.865	0.217	1.13E-39	4.96E-39		Higher
CAMTA1	ENSG00000171735	1786.490	1.987	0.555	0.0003443	0.000528		Higher
REST	ENSG00000084093	916.440	-2.065	0.142	6.37E-48	3.26E-47	Higher	
TFAP2A	ENSG00000137203	251.937	5.305	0.334	1.15E-56	6.81E-56		Higher
TFAP2C	ENSG00000087510	121.161	-2.533	0.326	7.73E-15	1.86E-14	Higher	

E) CACNA1G (log2Fold2Change -0.17 [higher in hiPSCs])

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	Padj	hiPSCs	hiPSC-neuron
BRN2	ENSG00000184486	444.863	5.111	0.480	1.59E-26	5.19E-26		Higher
EBF1	ENSG00000164330	729.426	8.79	0.209	0	0		Higher
EBF3	ENSG00000108001	432.284	6.058	0.152	0	0		Higher
INSM1	ENSG00000173404	110.160	2.675	0.353	3.44E-14	8.13E-14		Higher
KLF15	ENSG00000163884	160.404	-3.022	0.172	1.60E-68	1.14E-67	Higher	
KLF7	ENSG00000118263	1436.37	2.700	0.111	5.94E-131	1.03E-129		Higher
NR6A1	ENSG00000148200	2916.850	-3.495	0.169	2.80E-95	3.04E-94	Higher	
PLAG1	ENSG00000181690	322.561	-2.284	0.244	7.82E-21	2.22E-20	Higher	
REST	ENSG00000084093	916.440	-2.065	0.142	6.37E-48	3.26E-47	Higher	
RREB1	ENSG00000124782	457.261	-2.367	0.211	3.84E-29	1.34E-28	Higher	
RUNX1	ENSG00000159216	211.086	4.694	0.288	6.89E-60	4.29E-59		Higher
SP1	ENSG00000185591	1398.115	-1.941	0.101	5.78E-83	5.21E-82	Higher	
SP8	ENSG00000164651	151.621	-1.815	0.630	.00397	0.005	Higher	
TFAP2A	ENSG00000137203	251.937	5.305	0.334	1.15E-56	6.81E-56		Higher
TFAP2C	ENSG00000087510	121.161	-2.533	0.326	7.73E-15	1.86E-14	Higher	
ZIC2	ENSG00000043355	2094.364	-4.554	0.373	2.61E-34	1.02E-33	Higher	
ZIC3	ENSG00000156925	799.440	-3.959	0.579	7.81E-12	1.70E-11	Higher	
ZNF93	ENSG00000184635	390.593	-2.474	0.219	1.08E-29	3.82E-29	Higher	

F) CACNA2D1 (log2FoldChange 2.57 [higher in hiPSC-derived neurons])

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	Padj	hiPSCs	hiPSC-neuron
ASCL1	ENSG00000139352	106.402	5.810	0.440	9.65E-40	4.25E-39		Higher
BRN2	ENSG00000184486	444.863	5.111	0.480	1.59E-26	5.19E-26		Higher
EBF1	ENSG00000164330	729.426	8.79	0.209	0	0		Higher
EBF3	ENSG00000108001	432.284	6.058	0.152	0	0		Higher
GLI1	ENSG00000111087	131.201	-3.372	0.300	3.08E-29	1.08E-28	Higher	
HEY2	ENSG00000135547	112.981	-2.795	0.240	3.20E-31	1.17E-30	Higher	
IRF3	ENSG00000126456	536.501	-2.301	0.123	4.63E-78	3.88E-77	Higher	
KLF15	ENSG00000163884	160.404	-3.022	0.172	1.60E-68	1.14E-67	Higher	
KLF7	ENSG00000118263	1436.37	2.700	0.111	5.94E-131	1.03E-129		Higher
MEF2C	ENSG00000081189	450.320	5.845	0.230	2.55E-142	5.11E-141		Higher
MEIS1	ENSG00000143995	166.046	6.302	0.409	1.93E-53	1.09E-52		Higher
MEIS2	ENSG00000134138	1270.242	6.984	0.422	1.57E-61	1.01E-60		Higher
NFATC1	ENSG00000131196	112.798	-1.727	0.249	3.90E-12	8.61E-12	Higher	
NFIC	ENSG00000141905	575.96	3.825	0.164	3.90E-120	5.87E-119		Higher
NR2F1	ENSG00000175745	672.143	6.280	0.397	2.76E-56	1.63E-55		Higher
NR6A1	ENSG00000148200	2916.850	-3.495	0.169	2.80E-95	3.04E-94	Higher	
NRG1	ENSG00000157168	1138.507	5.662	0.148	0	0		Higher
PLAG1	ENSG00000181690	322.561	-2.284	0.244	7.82E-21	2.22E-20	Higher	
PRDM14	ENSG00000147596	278.725	-9.440	0.337	1.51E-172	4.36E-171	Higher	
PRDM4	ENSG00000110851	964.481	1.880	0.0546	4.44E-260	3.09E-258		Higher
PRDM5	ENSG00000138738	225.829	-2.712	0.128	4.75E-99	5.44E-98	Higher	
RARA	ENSG00000131759	400.718	1.501	0.161	9.45E-21	2.68E-20		Higher

RARB	ENSG00000077092	146.023	5.444	0.274	3.69E-88	3.61E-87		Higher
REST	ENSG00000084093	916.440	-2.065	0.142	6.37E-48	3.26E-47	Higher	
RREB1	ENSG00000124782	457.261	-2.367	0.211	3.84E-29	1.34E-28	Higher	
RUNX1	ENSG00000159216	211.086	4.694	0.288	6.89E-60	4.29E-59		Higher
SCRT1	ENSG00000261678	268.566	7.917	0.179	0	0		Higher
SCRT2	ENSG00000215397	292.159	5.575	0.165	1.21E-251	7.80E-250		Higher
SNAI2	ENSG00000019549	101.190	2.924	0.302	3.08E-22	9.06E-22		Higher
SOX10	ENSG00000100146	982.729	4.772	0.393	7.25E-34	2.82E-33		Higher
SP1	ENSG00000185591	1398.115	-1.941	0.101	5.78E-83	5.21E-82	Higher	
SP8	ENSG00000164651	151.621	-1.815	0.630	.00397	0.005	Higher	
TCF3	ENSG00000071564	2707.850	-2.302	0.181	4.43E-37	1.84E-36	Higher	
TCF4	ENSG00000196628	1269.598	-1.535	0.259	2.63E-09	5.22E-09	Higher	
TEAD2	ENSG00000074219	1120.903	-2.736	0.276	3.73E-23	1.12E-22	Higher	
TEAD3	ENSG00000007866	428.275	-2.075	0.213	2.16E-22	6.38E-22	Higher	
TEAD4	ENSG00000197905	324.361	-3.715	0.206	1.98E-72	1.51E-71	Higher	
TFAP2A	ENSG00000137203	251.937	5.305	0.334	1.15E-56	6.81E-56		Higher
TFAP2C	ENSG00000087510	121.161	-2.533	0.326	7.73E-15	1.86E-14	Higher	
ZBTB7B	ENSG00000160685	241.070	-2.213	0.247	3.69E-19	1.00E-18	Higher	
ZEB1	ENSG00000148516	614.388	4.083	0.250	8.64E-60	5.38E-59		Higher
ZFP42	ENSG00000179059	1159.133	-9.524	0.264	3.57E-284	3.07E-282	Higher	
ZIC2	ENSG00000043355	2094.364	-4.554	0.373	2.61E-34	1.02E-33	Higher	
ZIC3	ENSG00000156925	799.440	-3.959	0.579	7.81E-12	1.70E-11	Higher	
ZIC5	ENSG00000139800	837.856	-5.007	0.446	3.14E-29	1.10E-28	Higher	
ZNF93	ENSG00000184635	390.593	-2.474	0.219	1.08E-29	3.82E-29	Higher	

G) CACNA2D2 (log2FoldChange 0.119 [higher in hiPSC-derived neurons])

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	Padj	hiPSCs?	hiPSC-neuron?
KLF15	ENSG00000163884	160.404	-3.022	0.172	1.60E-68	1.14E-67	Higher	
KLF7	ENSG00000118263	1436.37	2.700	0.111	5.94E-131	1.03E-129		Higher
ZNF93	ENSG00000184635	390.593	-2.474	0.219	1.08E-29	3.82E-29	Higher	

H) CACNA2D3 (log2FoldChange 4.30 [higher in hiPSC-derived neurons])

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	Padj	hiPSCs?	hiPSC-neuron?
ASCL1	ENSG00000139352	106.402	5.810	0.440	9.65E-40	4.25E-39		Higher
EBF1	ENSG00000164330	729.426	8.79	0.209	0	0		Higher
KLF15	ENSG00000163884	160.404	-3.022	0.172	1.60E-68	1.14E-67	Higher	
TCF3	ENSG00000071564	2707.850	-2.302	0.181	4.43E-37	1.84E-36	Higher	
TCF4	ENSG00000196628	1269.598	-1.535	0.259	2.63E-09	5.22E-09	Higher	
ZIC5	ENSG00000139800	837.856	-5.007	0.446	3.14E-29	1.10E-28	Higher	
ZNF93	ENSG00000184635	390.593	-2.474	0.219	1.08E-29	3.82E-29	Higher	

I) CACNB2 (log2FoldChange 0.591 [higher in hiPSC-derived neurons])

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	Padj	hiPSCs?	hiPSC-neuron?
CAMTA1	ENSG00000171735	1786.490	1.987	0.555	0.0003443	0.000528		Higher
CAMTA2	ENSG00000108509	558.093	2.827	0.120	9.74E-124	1.53E-122		Higher
EBF3	ENSG00000108001	432.284	6.058	0.152	0	0		Higher
GLI1	ENSG00000111087	131.201	-3.372	0.300	3.08E-29	1.08E-28	Higher	
INSM1	ENSG00000173404	110.160	2.675	0.353	3.44E-14	8.13E-14		Higher
KLF15	ENSG00000163884	160.404	-3.022	0.172	1.60E-68	1.14E-67	Higher	
KLF7	ENSG00000118263	1436.37	2.700	0.111	5.94E-131	1.03E-129		Higher
MEIS1	ENSG00000143995	166.046	6.302	0.409	1.93E-53	1.09E-52		Higher
MYCN	ENSG00000134323	998.843	-1.707	0.119	8.01E-47	4.02E-46	Higher	
NR2F1	ENSG00000175745	672.143	6.280	0.397	2.76E-56	1.63E-55		Higher
PLAG1	ENSG00000181690	322.561	-2.284	0.244	7.82E-21	2.22E-20	Higher	
PRDM5	ENSG00000138738	225.829	-2.712	0.128	4.75E-99	5.44E-98	Higher	
REST	ENSG00000084093	916.440	-2.065	0.142	6.37E-48	3.26E-47	Higher	
RUNX1	ENSG00000159216	211.086	4.694	0.288	6.89E-60	4.29E-59		Higher
SP1	ENSG00000185591	1398.115	-1.941	0.101	5.78E-83	5.21E-82	Higher	
SP8	ENSG00000164651	151.621	-1.815	0.630	.00397	0.005	Higher	
TBX3	ENSG00000135111	1053.845	8.377	0.412	4.43E-92	4.58E-91		Higher
TFAP2A	ENSG00000137203	251.937	5.305	0.334	1.15E-56	6.81E-56		Higher
TFAP2C	ENSG00000087510	121.161	-2.533	0.326	7.73E-15	1.86E-14	Higher	
ZIC2	ENSG00000043355	2094.364	-4.554	0.373	2.61E-34	1.02E-33	Higher	
ZIC3	ENSG00000156925	799.440	-3.959	0.579	7.81E-12	1.70E-11	Higher	
ZIC5	ENSG00000139800	837.856	-5.007	0.446	3.14E-29	1.10E-28	Higher	
ZNF331	ENSG00000130844	626.710	-1.615	0.0671	3.81E-128	6.33E-127	Higher	
ZNF93	ENSG00000184635	390.593	-2.474	0.219	1.08E-29	3.82E-29	Higher	

J) CACNB3 (log2FoldChange 3.12 [higher in hiPSC-derived neurons])

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	Padj	hiPSCs?	hiPSC-neuron?
CAMTA1	ENSG00000171735	1786.490	1.987	0.555	0.0003443	0.000528		Higher
CAMTA2	ENSG00000108509	558.093	2.827	0.120	9.74E-124	1.53E-122		Higher
E2F1	ENSG00000101412	353.788	-2.114	0.187	1.08E-29	3.83E-29	Higher	
E2F2	ENSG00000007968	93.543	-1.717	0.336	3.35E-07	6.07E-07	Higher	
EBF1	ENSG00000164330	729.426	8.79	0.209	0	0		Higher
EBF3	ENSG00000108001	432.284	6.058	0.152	0	0		Higher
ELK4	ENSG00000158711	567.521	-1.8134	0.129	1.24E-44	5.96E-44	Higher	
HEY1	ENSG00000164683	369.332	4.570	0.204	2.14E-110	2.86E-109		Higher
HEY2	ENSG00000135547	112.981	-2.795	0.240	3.20E-31	1.17E-30	Higher	
KLF15	ENSG00000163884	160.404	-3.022	0.172	1.60E-68	1.14E-67	Higher	
KLF7	ENSG00000118263	1436.37	2.700	0.111	5.94E-131	1.03E-129		Higher
NFIC	ENSG00000141905	575.96	3.825	0.164	3.90E-120	5.87E-119		Higher
NR6A1	ENSG00000148200	2916.850	-3.495	0.169	2.80E-95	3.04E-94	Higher	
NRG1	ENSG00000157168	1138.507	5.662	0.148	0	0		Higher
PAX6	ENSG00000007372	364.997	5.316	0.500	9.80E-27	3.22E-26		Higher
PIF1	ENSG00000140451	532.182	-4.573	0.298	4.99E-53	2.79E-52	Higher	
PLAG1	ENSG00000181690	322.561	-2.284	0.244	7.82E-21	2.22E-20	Higher	
RARG	ENSG00000172819	525.507	-2.910	0.269	3.10E-27	1.03E-26	Higher	
REST	ENSG00000084093	916.440	-2.065	0.142	6.37E-48	3.26E-47	Higher	
SP1	ENSG00000185591	1398.115	-1.941	0.101	5.78E-83	5.21E-82	Higher	
SP8	ENSG00000164651	151.621	-1.815	0.630	.00397	0.005	Higher	

TCF3	ENSG00000071564	2707.850	-2.302	0.181	4.43E-37	1.84E-36	Higher	
TCF4	ENSG00000196628	1269.598	-1.535	0.259	2.63E-09	5.22E-09	Higher	
TFAP2A	ENSG00000137203	251.937	5.305	0.334	1.15E-56	6.81E-56		Higher
TFAP2C	ENSG00000087510	121.161	-2.533	0.326	7.73E-15	1.86E-14	Higher	
ZEB1	ENSG00000148516	614.388	4.083	0.250	8.64E-60	5.38E-59		Higher
ZIC2	ENSG00000043355	2094.364	-4.554	0.373	2.61E-34	1.02E-33	Higher	
ZIC3	ENSG00000156925	799.440	-3.959	0.579	7.81E-12	1.70E-11	Higher	
ZIC5	ENSG00000139800	837.856	-5.007	0.446	3.14E-29	1.10E-28	Higher	
ZNF331	ENSG00000130844	626.710	-1.615	0.0671	3.81E-128	6.33E-127	Higher	
ZNF93	ENSG00000184635	390.593	-2.474	0.219	1.08E-29	3.82E-29	Higher	

K) CACNB4 (log2FoldChange 2.25 [higher in hiPSC-derived neurons])

Gene	ENSG	baseMean	log2FoldChange	lfcSE	p-value	Padj	hiPSCs?	hiPSC-neuron?
ARNT2	ENSG00000172379	1296.213	2.964	0.0807	1.01E-295	9.86E-294		Higher
ASCL1	ENSG00000139352	106.402	5.810	0.440	9.65E-40	4.25E-39		Higher
BACH2	ENSG00000112182	696.018	3.003	0.0844	2.27E-277	1.85E-275		Higher
BCL6	ENSG00000113916	303.832	1.706	0.176	2.93E-22	8.64E-22		Higher
BHLHE40	ENSG00000134107	400.686	2.865	0.217	1.13E-39	4.96E-39		Higher
EBF1	ENSG00000164330	729.426	8.79	0.209	0	0		Higher
FOXA3	ENSG00000170608	156.464	-8.623	0.295	1.91E-187	6.53E-186	Higher	
FOXK1	ENSG00000164916	592.187	1.706	0.102	1.62E-62	1.05E-61		Higher
FOXO1	ENSG00000150907	607.609	-2.839	0.239	1.40E-32	5.30E-32	Higher	
JDP2	ENSG00000140044	185.002	-1.600	0.156	1.35E-24	4.21E-24	Higher	
JUNB	ENSG00000171223	290.225	2.375	0.196	1.14E-33	4.40E-33		Higher
KLF15	ENSG00000163884	160.404	-3.022	0.172	1.60E-68	1.14E-67	Higher	
KLF7	ENSG00000118263	1436.37	2.700	0.111	5.94E-131	1.03E-129		Higher
NRG1	ENSG00000157168	1138.507	5.662	0.148	0	0		Higher
PLAG1	ENSG00000181690	322.561	-2.284	0.244	7.82E-21	2.22E-20	Higher	
RARA	ENSG00000131759	400.718	1.501	0.161	9.45E-21	2.68E-20		Higher
RARB	ENSG00000077092	146.023	5.444	0.274	3.69E-88	3.61E-87		Higher
RARG	ENSG00000172819	525.507	-2.910	0.269	3.10E-27	1.03E-26	Higher	
RUNX1	ENSG00000159216	211.086	4.694	0.288	6.89E-60	4.29E-59		Higher
SOX10	ENSG00000100146	982.729	4.772	0.393	7.25E-34	2.82E-33		Higher
SOX21	ENSG00000125285	231.388	-1.685	0.660	0.0106	0.0145	Higher	
SP1	ENSG00000185591	1398.115	-1.941	0.101	5.78E-83	5.21E-82	Higher	
TBX2	ENSG00000121068	337.339	7.506	0.453	1.45E-61	9.33E-61		Higher

TBX3	ENSG00000135111	1053.845	8.377	0.412	4.43E-92	4.58E-91		Higher
TCF4	ENSG00000196628	1269.598	-1.535	0.259	2.63E-09	5.22E-09	Higher	
TCF7L1	ENSG00000152284	1249.047	-3.246	0.305	2.10E-26	6.86E-26	Higher	
TEAD2	ENSG00000074219	1120.903	-2.736	0.276	3.73E-23	1.12E-22	Higher	
TEAD3	ENSG00000007866	428.275	-2.075	0.213	2.16E-22	6.38E-22	Higher	
TEAD4	ENSG00000197905	324.361	-3.715	0.206	1.98E-72	1.51E-71	Higher	
TFAP2A	ENSG00000137203	251.937	5.305	0.334	1.15E-56	6.81E-56		Higher
TFAP2C	ENSG00000087510	121.161	-2.533	0.326	7.73E-15	1.86E-14	Higher	
TFAP4	ENSG00000090447	256.396	-1.704	0.132	3.37E-38	1.44E-37	Higher	
TP53	ENSG00000141510	1360.724	-1.735	0.200	4.03E-20	1.12E-19	Higher	
ZIC2	ENSG00000043355	2094.364	-4.554	0.373	2.61E-34	1.02E-33	Higher	
ZIC3	ENSG00000156925	799.440	-3.959	0.579	7.81E-12	1.70E-11	Higher	

Table 4.3: Gene expression analyses of TFs with footprints that were detected at VGCC TSS and with an expression detected as differentially expressed between hiPSCs and hiPSC-derived neurons.

The differential expression for TFs was examined across cellular differentiation. The table contains the results for baseMean, log2FoldChange, and adjusted p-value (Padj). baseMean is the average of normalized count values, divided by size factors, taken over all samples. Log2FoldChange is the effect size estimate. This value indicates the extent to which a transcript's expression changes between two different groups. A positive value indicates enrichment in hiPSC-derived neurons, whereas a negative value reflects higher expression in hiPSCs. The adjusted p-value (padj) refers to the adjusted p-value for multiple testing for the gene or transcript (the Benjamini-Hochberg Procedure). The presented table features only TFs with footprints that were detected at VGCC TSS and with a differential expression was a log2FoldChange less than -1.5 or greater than 1.5.

Table 4.4 shows the 80 individual TFs with footprints found at the TSS of the 12 VGCC loci.

TF	# of VGCC loci	TF	# of VGCC loci
ARNT2	1	**PLAG1**	**5**
ASCL1	**5**	PRDM14	1
BACH2	1	PRDM4	1
BHLHE40	2	PRDM5	4
BRN2	2	RARA	2
BCL6	1	RARB	2
CAMTA1	4	RARG	2
CAMTA2	3	**REST**	**6**
EBF1	**7**	RREB1	2
EBF3	**6**	RUNX1	4
ELK4	3	SCRT1	1
E2F1	1	SCRT2	1
E2F2	1	SNAI2	2
FOXA3	1	SOX2	1
FOXK1	1	SOX5	1
FOXO1	1	SOX10	2
GLI1	2	SOX21	1
HEY1	2	**SP1**	**8**
HEY2	3	**SP8**	**6**
INSM1	3	TBX2	1
IRF3	1	TBX3	2
JDP2	1	**TCF3**	**6**
JUNB	1	**TCF4**	**7**
KLF15	**10**	TCF7L1	2
KLF7	**9**	TCF7L2	1
LEF1	1	TEAD2	2
MAFA	1	TEAD3	2
MEF2C	1	**TFAP2A**	**9**
MEIS1	3	**TFAP2C**	**9**
MEIS2	3	TFAP4	1
MEIS3	2	TP53	2
MLXIPL	1	ZBTB7B	2
MYCN	1	ZEB1	3
NFATC1	1	ZFP42	1
NFIC	2	**ZIC2**	**5**
NRG	3	**ZIC3**	**5**
NR2F1	3	**ZIC5**	**6**
NR6A1	3	ZNF324	1
PAX6	1	ZNF331	2
PIF1	1	**ZNF93**	**9**

Table 4.4: TFs and the number of TF footprints at various VGCC loci.

The table indicates TFs and the number of VGCC loci that had footprints for those particular TFs. Bold font indicates TFs that bind at five or more VGCC loci.

TFs that were found at five or more VGCC loci (referred to here as candidate TFs) received further attention. The hypothesis was that TFs that bind multiple VGCC loci would comprise "master regulators" that coordinately regulate multiple VGCC loci. Precedence for coordinate regulation of different gene family members by a single TF has been reported in mouse PSCs (Hall et al., 2009).

I next considered whether the detection of footprints for certain TFs is specific to VGCCs or if it reflects the prevalence in the human genome. A random gene selector script was used to generate a list of random genes and the prevalence of the footprints of candidate TFs was assessed in 48 random genes (**Table 4.5**; data not shown).

Candidate TF	Frequency in VGCC loci	Frequency in random genes	Chi-square statistic	P-value
ASCL1	5 / 12	27 / 48	0.820	0.365
EBF1	7 / 12	21 / 48	0.820	0.365
EBF3	6 / 12	19 / 48	0.429	0.513
KLF15	10 / 12	42 / 48	0.144	0.704
KLF7	9 / 12	42 / 48	1.177	0.278
PLAG1	5 / 12	29 / 48	1.374	0.241
REST	6 / 12	28 / 48	0.272	0.602
SP1	8 / 12	39 / 48	1.203	0.273
SP8	6 / 12	32 / 48	1.148	0.284
TCF3	6 / 12	26 / 48	0.067	0.796
TCF4	7 / 12	23 / 48	0.417	0.519
TFAP2A	9 / 12	37 / 48	0.0233	0.879
TFAP2C	9 / 12	37 / 48	0.0233	0.879
ZIC2	5 / 12	28 / 48	1.077	0.299
ZIC3	5 / 12	30 / 48	1.714	0.190
ZIC5	6 / 12	31 / 48	0.864	0.353
ZNF93	9 / 12	40 / 48	0.445	0.505

Table 4.5: Prevalence of TF footprints in randomly selected genes.

Searches were conducted for the presence of candidate TF footprints in the TSS of the randomly selected genes. The second column shows the frequency at VGCC loci; the third column shows the frequency at random loci. The fourth column represents the chi-square statistic, and the fifth column represents the p-value. Alpha was set at 0.05.

There were no statistically significant differences between the frequency of a candidate TF footprint at a VGCC locus versus that of the randomly selected gene. A scatterplot showing the correlation between TF footprint frequencies at VGCC loci and randomly selected genes is presented in **Figure 4.13.** First, normality of the data was checked using the Shapiro-Wilk test. This test revealed that the data did not follow a normal distribution. The Spearman correlation coefficient was 0.61 (with a p-value of 0.0094), indicating that the frequency of TF footprints at VGCC loci is correlated with TF frequency at randomly selected genes. The fact that the p-value is less than 0.05 indicates there is evidence to reject the null hypothesis there is no correlation between the variables, allowing one to conclude that there is a correlation for Spearman's correlation test.

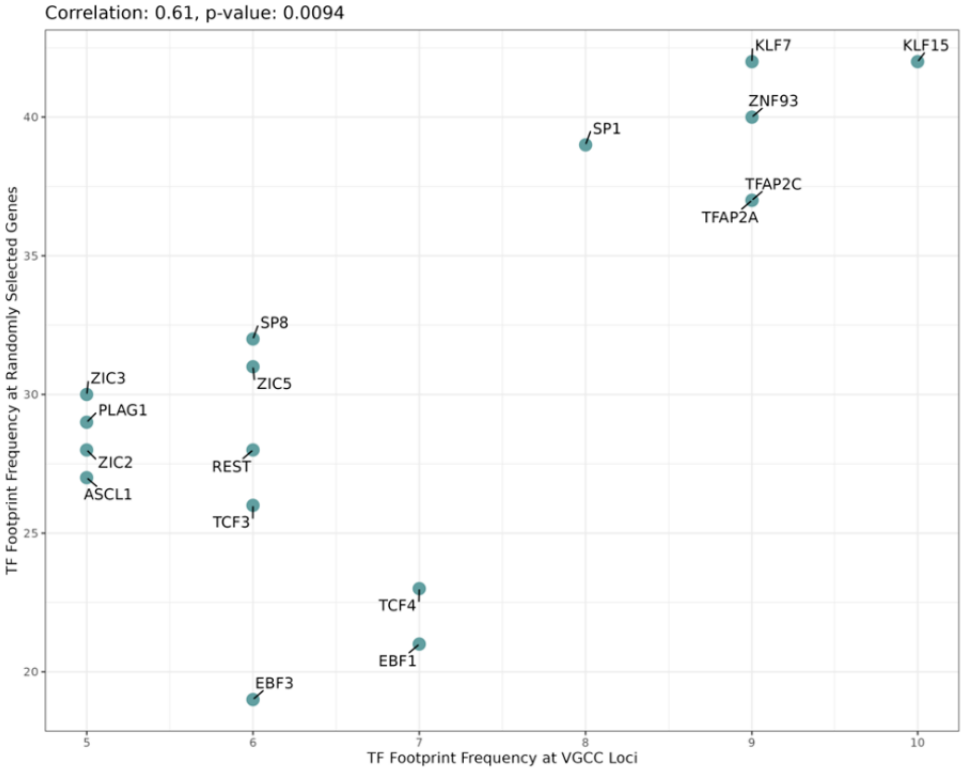


Figure 4.13: Correlation between frequencies of TF footprints at VGCC loci and at randomly selected genes.

The scatterplot shows the relationship between the frequency of TF footprints at VGCC loci and at randomly selected genes. Based on data in **Table 4.5**. Each point in the plot represents the frequency of a specific TF at VGCC loci (x-axis) and its frequency at randomly selected genes (y-axis). Numbers refer to the actual number of events (i.e. EBF3 footprints were found at 6 loci and at 19 randomly selected genes). As noted in the main text, first, normality was checked using the Shapiro-Wilk test. Owed to lack of normality, Spearman's correlation was used rather than Pearson's correlation. Correlation in the figure refers to Spearman's correlation coefficient (p-value is 0.0094). TF that corresponds to each point on the plot is indicated. Both TFAP2A and TFAP2C had the same footprint frequency at VGCC loci and randomly selected genes.

While TF motif enrichment among VGCCs would have been a good indicator that VGCCs are regulated by the candidate TFs, the absence of enrichment does not mean that VGCCs are not also likely to be regulated by these TFs, if the corresponding footprint is present. Therefore, the list of candidate TFs was examined further, and inspection of ChIP-seq data indicating binding to VGCC genes along with more characterizations of their gene expression patterns are described below.

4.3.2.4 Searching for Binding Events For TFs Whose Footprints Were Found in VGCC Loci in Publicly Available ChIP-seq Datasets

Efforts were next made to validate the TF footprints identified by TOBIAS in published ChIP-seq datasets. ChIP-Seq data from ENCODE for EBF1, KLF7, REST, TCF4, and ZIC2 were investigated. Unfortunately, ChIP-seq data for the other TFs of interest were

not available. The analyses identified TF binding for these five TFs at different, but not all VGCC loci. A representative illustration of the ENCODE ChIP-seq data inspection is shown in **Figure 4.14** (showing KLF7 binding at the TSS of CACNB)

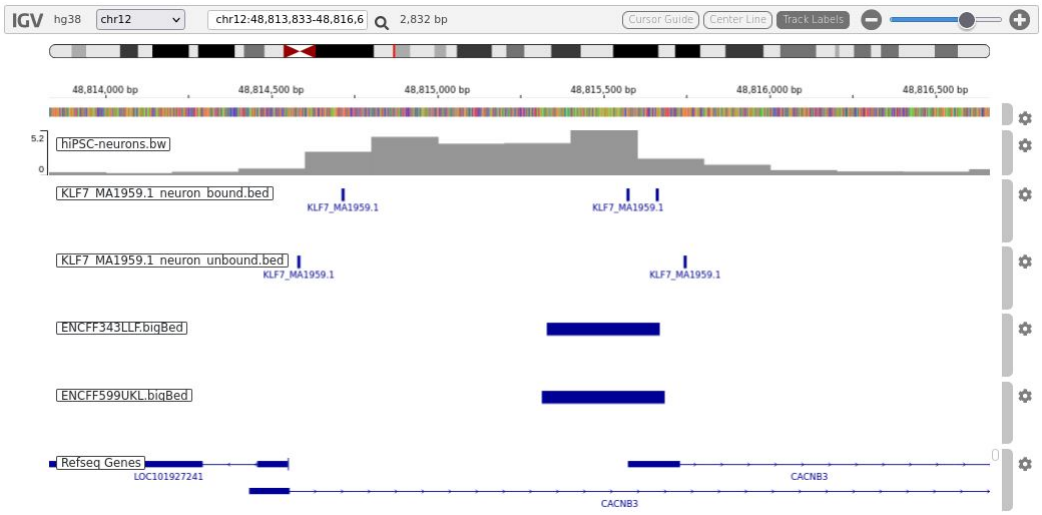


Figure 4.14: Overlapping ChIP-seq KLF7 binding events in HEK293 cells with TOBIAS analyses from hiPSC-derived neurons.

HEK293 cells were genetically modified using site-specific recombination targeting human KLF7. The figure shows views from the Integrative Genome Browser featuring an overlay of eGFP-KLF7 ChIP-seq from ENCODE with HIPSCI hiPSC-derived neuron ATAC-seq data and TOBIAS TF footprinting analysis for KLF7. (A) High-magnification view of the promoter of CACNB3 (TSS). hiPSC-neurons.bw track indicates ATAC-seq data of hiPSC-derived neurons from the HIPSCI. hiPSC_neurons.bw indicates ATAC-seq data signal from HIPSCI neurons. KLF7 _MA1959.1_neuron_bound.bed indicates bound TOBIAS footprints. KLF7_MA1959.1_neuron_unbound.bed indicates unbound TOBIAS footprints. ENCFF343LFF.bigBed and ENCFF599UKL.bigBed indicate tracks for eGFP-KLF7 ChIP-seq data.

A summary of the efforts to inspect ChIP-seq data from ENCODE is provided in **Table 4.6**. Altogether, inspection of ChIP-seq data partially validates the TOBIAS TF footprinting analyses.

	EBF1 (GM127878)	**KLF7 (HEK293)**	**REST (K562)**	**TCF4 (SK-N-SH)**	**ZIC2 (HEK293)**
CACNA1B	Absent	Absent	**Present**	Absent	Absent
CACNA1C	Absent	Absent	Absent	Absent	**Present**
CACNA1D	Absent	Absent	Absent	Absent	**Present**
CACNA1E	Absent	Absent	Absent	Absent	Absent
CACNA1G	Absent	Absent	Absent	Absent	Absent
CACNA2D1	**Present**	Absent	Absent	Absent	**Present**
CACNA2D2	Absent	Absent	Absent	Absent	**Present**
CACNA2D3	Absent	Absent	Absent	**Present**	**Present**
CACNB1	Absent	Absent	Absent	Absent	**Present**
CACNB2	Absent	Absent	Absent	Absent	**Present**
CACNB3	Absent	**Present**	**Present**	Absent	**Present**
CACNB4	Absent	Absent	Absent	Absent	Absent

Table 4.6: Inspection of published ChIP-seq datasets partially validates TOBIAS footprinting results.

ChIP-seq data for EBF1, KLF7, REST, TCF4, and ZIC2 from ENCODE were downloaded and uploaded to IGV for inspection. The TOBIAS footprint results and ChIP-seq results for the corresponding TF were uploaded to the Integrative Genome Browser. The intersection of the TOBIAS footprinting results (from hiPSC-derived neurons) with the ChIP-seq data results was carried out by inspecting whether a ChIP-seq binding event occurred at the TSS of a VGCC locus. "Present" in bold font indicates the existence of ChIP-seq binding event at noted VGCC locus; absent indicates non-existence of ChIP-seq binding event.

4.3.2.5 TF Expression Across Human Neural Development

To gain an insight into the expression patterns of the candidate TFs during human brain development, data were collated from the BrainSpan Atlas of the Developing Brain (www.brainspan.org) and were inspected. A heatmap showing expression levels of the different candidate TFs during both fetal and postnatal development is presented in **Figure 4.15**. Several TFs showed higher expression during human fetal brain development relative to postnatal brain development, including KLF7, and TCF4.

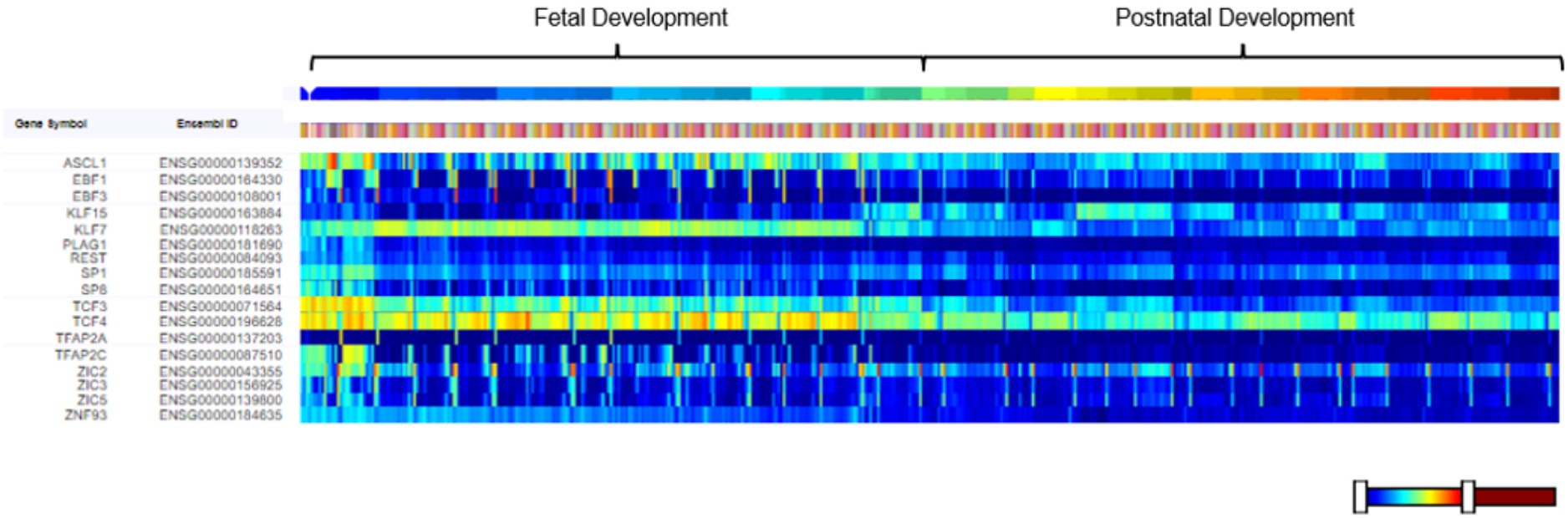


Figure 4.15: Expression of candidate TFs during human neural development

The figure shows a heatmap of TF expression during human neural development (at both the fetal and postnatal stages). The top row with different colors shows the donors. Individual donors contributed to different stages of human neural development. Each color shift represents another donor. The second row with different colors shows brain structure. Individual colors represent a sample from a different brain region. The bottom rows (the bulk of figure) show a heatmap of expression levels. Rows indicating expression levels from each TF are indicated. Heatmap legend: red indicates higher expression levels and blue indicates low or absent expression levels. (data from www.brainspan.org). Information about different sampled brain regions can be found in **Supplementary Table 2.3**. Scale represents relative expression (for more rigorous quantification, see **Table 4.7**).

For higher-resolution analyses, the BrainSpan expression data were downloaded and statistically analyzed. A quantification of the frontal cortex transcript levels for the TFs identified as being possibly higher by heatmap and/or with partial validation of TOBIAS findings by inspection of ChIP-seq data is presented in **Table 4.7**. Data from the frontal cortex were analyzed because of its relevance to psychiatry. To quantify transcript levels for TFs from the BrainSpan Atlas for comparing fetal and postnatal brain development, similar methods to those used in **Chapter 3** were used. First, normality was checked using the Shapiro-Wilk test. Because not all datasets displayed a normal distribution, the median and interquartile range were calculated. The Mann-Whitney U test was used to determine if there was a significant difference between the groups. If the p-value was less than 0.05, the difference between the groups was considered significant. Several TFs showed higher expression in fetal brain samples compared with postnatal brain samples, illustrated by ASCL1, KLF7, PLAG1, REST, SP1, SP8, TCF3, TCF4 and TFAP2C. EBF1 showed higher expression in postnatal brain samples compared with fetal brain samples. Other TFs either showed lower expression throughout fetal and postnatal tissues; sporadic expression in the human fetal brain, or no specific enrichment in fetal brain tissues (data not shown).

TF	Fetal FC median	Fetal FC IQR	Postnatal FC median	Postnatal FC IQR	U statistic	p-value	Significant?
ASCL1	6.487	10.450	4.305	2.567	2599	0.0189	Yes
EBF1	0.296	0.366	0.642	0.575	979	2.697E-07	Yes
KLF7	22.064	12.810	3.833	2.558	3993	1.378E-18	Yes
PLAG1	0.593	0.568	0.224	0.194	3451	3.152E-10	Yes
REST	1.295	1.032	0.756	0.543	3001	2.568E-05	Yes
SP1	3.148	1.910	2.432	1.567	2894	0.000204	Yes
SP8	1.620	2.081	0.444	0.396	3863	2.452E-16	Yes
TCF3	11.908	8.835	2.918	3.450	3953	7.050E-18	Yes
TCF4	36.872	16.452	14.899	9.837	3965	4.336E-18	Yes
TFAP2C	0.860	4.462	0.122	0.167	3297	2.426E-08	Yes
ZIC2	2.164	4.287	2.135	1.833	2016	0.730	No

Table 4.7: Expression levels of TFs in fetal relative to postnatal brain development (frontal cortex).

The first column indicates TF; the second column indicates the median of fetal samples; the third column indicates the interquartile range for fetal samples; the fourth column indicates the postnatal median; the fifth column indicates postnatal interquartile range. The sixth column represents U statistic; seventh column represents p-value. In the eighth column, whether the differences are considered statistically significant is presented (p-value <0.05). Data from 133 samples of frontal cortex tissue (51 fetal samples and 82 postnatal samples). Values represent RPKM. All of the TFs above except EBF1 and ZIC2 were more highly expressed in the fetal brain.

Analyses of TF expression levels across early fetal and late fetal development were carried out identically to analyses for RBPs described in **Chapter 3**. As previously noted, a comparison of gene expression during neurogenesis and after neurogenesis had occurred would shed insight into TF expression patterns into the pre-differentiated and

differentiated in vivo cellular states (see **Chapter 3 Discussion** for interpretation). A comparison of the early and late fetal RPKM values found that PLAG1, REST, and SP1 were expressed more highly in early fetal stages; whereas ASCL1, EBF1, KLF7, SP8, TCF4, and TFAP2C did not show statistically significant differences between early and late fetal stages (**Table 4.8**).

TF	Early Fetal FC median	Early Fetal FC IQR	Late Fetal FC median	Late Fetal FC IQR	U statistic	p-value	Significant?
ASCL1	4.412	15.235	7.471	2.930	224	0.200	No
EBF1	0.274	0.347	0.403	0.375	277	0.820	No
KLF7	22.416	14.918	22.064	8.671	257	0.533	No
PLAG1	0.696	2.207	0.331	0.202	523	3.180E-07	Yes
REST	1.384	1.238	0.837	0.772	448	0.00113	Yes
SP1	3.591	2.722	2.649	1.072	388	0.0483	Yes
SP8	1.444	3.046	1.994	1.239	254	0.494	No
TCF3	12.522	13.886	11.343	7.067	350	0.229	No
TCF4	33.515	18.397	38.363	9.274	228	0.229	No
TFAP2C	0.966	8.284	0.824	3.201	337	0.346	No
ZIC2	1.973	4.863	2.164	3.655	275	0.789	No

Table 4.8: Comparison of TF expression levels between early and late fetal development.

Statistical analysis was conducted similarly to **Table 4.7**. For interpretation of statistics, see **Table 4.7**. The early fetal samples were samples aged 8 pcw to 17 pcw and the late fetal samples were aged after 17 pcw up to 37 pcw. The number of samples in early fetal stages was 34 and the number of samples in late fetal stages was 17. Most TFs did not show differential expression between early fetal and late fetal stages.

Altogether, an inspection of BrainSpan Developmental Transcriptome data revealed that several candidate TFs showed relatively higher expression during fetal brain development relative to postnatal brain development, as well as TFs that were more highly expressed in early fetal vs late fetal stages of development. It appears that there were more TFs that were differentially expressed between fetal vs postnatal stages compared with early fetal stages vs late fetal stages of development.

4.3.3 Summary

In this Chapter, a TOBIAS pipeline to detect TF footprints in ATAC-seq data of chromatin accessibility was applied to the HIPSCI dataset (Bentsen et al., 2020). A list of candidate TFs was generated based on the presence of TF footprints at five or more loci and differential gene expression across hiPSC differentiation. Publicly available experimental ChIP-seq data identified some TF binding events at VGCC loci partially validated TOBIAS analyses.

4.4 Discussion

In this Chapter, candidate TFs that might regulate the abundance of VGCCs in hiPSC-derived neurons were identified. First, attempts were made to reproduce the figures in the original TOBIAS paper and to assess the sensitivity and specificity of TOBIAS in PSCs. Next, TOBIAS analyses of VGCC loci found 80 unique human TFs with footprints found at the TSS of different VGCC loci. This long list was narrowed to 17 TFs by focusing on those that mapped to 5 or more VGCC loci. It was hypothesized that TFs that bind to five or more VGCC loci would be coordinately regulating multiple VGCC loci. Precedence for coordinate regulation of gene families by individual TFs has been reported in mouse PSCs (Hall et al., 2009). However, no statistical difference was observed between candidate TF footprint frequency at VGCC loci versus randomly selected genes. Nonetheless, an inspection of experimental ChIP-seq datasets for several of these TFs (EBF1, KLF7, REST, TCF4, and ZIC2) validated some binding events at VGCC loci. The expression of several TFs (ASCL1, KLF7, PLAG1, REST, SP1, SP8, TCF3, TCF4 and TFAP2C) was also enriched in human fetal brain tissues relative to postnatal brain tissues.

Some of the candidate TFs have known functions in a neural context, as summarized later in this Discussion. The candidate TFs that have been partially validated by inspection of ChIP-seq data will be considered below in more detail. This could not only shed light on their ability to maintain VGCC expression in the differentiated state but also possibly on the capacity to program the expression of VGCCs across cellular differentiation.

4.4.1 Identification of Candidate TFs

4.4.1.1 KLF7

Another TF that emerged as potentially being of interest was KLF7. KLF7 transcripts were more highly expressed in hiPSC-derived neurons, as well as more highly expressed in fetal stages relative to postnatal stages of brain development. There was no statistically significant difference between early and late stages of fetal brain development. The analyses identified KLF7 footprints at nearly all VGCC loci, but these results need to be interpreted with caution, as KLF7 footprints were also frequently identified among randomly selected genes, indicating the possible prevalence of KLF7 motifs across the human genome. However, an inspection of KLF7 ChIP-seq data showed that KLF7 can bind a VGCC locus (CACNB3) in HEK293 cells.

KLF7 is key for neuronal morphogenesis, as it plays a key role in neurite outgrowth and neuronal survival (Laub et al., 2005). An analysis of differentiating olfactory sensory neurons in Klf7 knockout mice shows that Klf7 regulates genes involved in olfactory neuron differentiation, axonal growth, cytoskeletal dynamics, cell adhesion, and synaptogenesis (Kajimura et al., 2007). Regarding VGCCs, Klf7 knockout mice show lower expression of Cacna1b relative to wild-type mice (Kajimura et al., 2007). Knockdown of KLF7 during human PSC differentiation into retinal ganglion cells not only affects retinal ganglion cell maturation but also a perturbed expression of CACNA2D1 and CACNB4 (Gautam et al., 2021). These data possibly suggest that VGCCs might be regulated by KLF7 more generally.

Regarding human biology, KLF7 has also been suggested as a possible risk gene for autism spectrum disorder (Tian et al., 2022). Patients with a deletion in 2q33.3q34, where KLF7 is located, possess autistic features (Courtens et al., 1997; Pescucci et al., 2003; Bisgaard et al., 2006; Brandau et al., 2008; Jang et al., 2015; Tian et al., 2022). Additionally, individuals with de novo mutations in KLF7 possess neuropsychiatric symptoms, including developmental delay and intellectual disability, psychiatric features, feeding/swallowing problems, and neuromuscular issues (Powis et al., 2017; Tian et al., 2022).

KLF7 is also a TF of interest because different members of the Kruppel-like factor family have been established as master regulators of cellular identity. The most notable example is KLF4, one of the four Yamanaka TFs that when overexpressed can reprogram somatic cells to pluripotency in combination with OCT4, SOX2, and c MYC overexpression (Takahashi and Yamanaka, 2006). The ability of KLFs, such as KLF4, to act as reprogramming factors has been linked to their being pioneer factors – that is, factors that can bind to closed chromatin (Soufi et al., 2015). Similarly to KLF4, KLF7 has also been identified as a pioneer factor (Grossman et al., 2018).

4.4.1.2 TCF4

Another TF that emerged as a gene of interest was TCF4. TCF4 was more highly expressed in hiPSCs relative to hiPSC-derived neurons, and also in fetal stages relative to postnatal stages of development. Therefore, like KLF7, TCF4 appears linked to more immature stages of development more generally. However, no statistically significant differences between expression levels in early and late fetal stages of development were observed. An inspection of TCF4 ChIP-seq data showed binding of TCF4 at CACNA2D3,

indicating possible regulation of VGCCs. TCF4 is expressed in many bodily tissues and also during human neural development (Teixeira et al., 2021). Mutations in TCF4 are associated with Pitt-Hopkins syndrome, a type of autism spectrum disorder distinguished by intellectual disability, absence of speech, and delayed cognitive development (Amiel et al., 2007; Brockschmidt et al., 2007; Zweier et al., 2007). Polymorphisms in TCF4 have also been linked to schizophrenia, bipolar disorder, post-traumatic stress disorder, and major depressive disorder. Most polymorphisms have been found in non-coding regions. While additional studies are warranted, preliminary evidence indicates that TCF4 expression was higher in hiPSC-derived neurons from patients compared with non-schizophrenic individuals (Brennand et al., 2011). Tcf4 overexpression in transgenic mice produces issues with fear memory and sensorimotor gating (Brzozka et al., 2010; Brzozka and Rossner, 2013. TCF/LEF family members are associated with acute transcriptional activation but lack pioneer factor properties (Wallmen et al., 2012). Altogether, the analyses potentially link TCF4, a gene associated with psychiatric conditions, with the potential regulation of VGCC expression in a model system of human neural development.

4.4.1.3 ZIC2

Inspection of ZIC2 ChIP-seq data collected in HEK293 cells showed ZIC2 binding to various VGCC loci, which included CACNA1C, CACNA1D, CACNA2D1, CACNA2D2, CACNA2D3, CACNB1, CACNB2, and CACNB3. Among the different TFs with publicly available ChIP-seq data, ZIC2 had the highest number of ChIP-seq binding events at VGCC loci. ZIC2 was more highly expressed in hiPSCs relative to hiPSC-derived neurons. However, its expression levels were not observed as statistically different in the fetal versus postnatal tissue comparison or in early fetal relative to late fetal stages of development. In humans, ZIC2 mutations have been associated with approximately 3-4% of cases of holoprosencephaly, a type of brain malformation (Nakayama et al., 2016).

4.4.1.4 REST

REST is a TF whose footprints were identified at six VGCC loci and whose expression was enriched in hiPSCs. However, it was expressed more highly in fetal vs. postnatal neural development and additionally expressed more highly in early fetal compared with late fetal neural development. Based on these expression patterns, it is plausible that REST is linked to less differentiated states. It should be noted that the expression levels

of REST were relatively low (median RPKM <2), but it has been implicated in neural biology. It prevents the expression of neuronal genes in non-neuronal cells, such as in the case of sodium channel gene expression (Chong et al., 1995). Dysregulation of REST has also been linked to neurodegenerative diseases (Hwang and Zukin, 2018).

4.4.1.5 EBF1

EBF1 was identified as an hiPSC-enriched TF that had footprints at VGCC loci. Inspection of EBF1 ChIP-seq data from ENCODE identified a ChIP-seq binding event at CACNA2D1. EBF1 was more highly expressed in postnatal brain tissue relative to fetal brain tissue but showed no differences between early fetal stage tissues relative to late fetal stage tissues. The expression levels of EBF1 across all stages of neural development were low (median RPKM <1). In a human context, EBF1 enhances neuronal differentiation and maturation in human embryonic stem cell differentiation (Faedo et al., 2017). EBF1 can bind and open chromatin (Boller et al., 2016). Indeed, in pro-B cells, EBF1 has been described as being continuously required for stabilizing local chromatin accessibility. (Zolotarev et al., 2022).

4.4.2 Limitations

A primary limitation of the studies carried out in this Chapter pertains to the limitations of TOBIAS. Although TOBIAS analysis of motif enrichment and binding events in PSCs validated its sensitivity, TOBIAS also appeared to lack some specificity in discriminating between TFs in the same TF family. This could best be reflected in the observations that TOBIAS would detect the footprints of factors in cell types where they are not expressed, such as the demonstrated examples of observing changes in the enrichment of non-pluripotency associated POU factors in Oct4 depletion experiments in mouse pluripotent stem cells and ESRRB footprints in human pluripotent stem cells. This indicates that TOBIAS is unable to discriminate between TFs with high sequence similarity and produce false positive results. However, one can still infer which TF is relevant based on context. In the case of Oct4 depletion, in spite of changes in the motif enrichment for different POU factors, it would be known that the relevant factor is Oct4 (Xiong et al., 2022). For analysis of my dataset of interest, hiPSC-derived neurons from the HIPSCI, knowledge of differential gene expression levels of the different TFs between the undifferentiated and differentiated states can help guide one towards identifying which TFs are the relevant

TFs.

Another important caveat to the TF footprinting analysis was the observation of TF footprints (for example KLF motifs) in randomly selected genes. One might need to carefully consider how to interpret footprint detection where the TF motif is not complex and is a repeat (as is in the case of KLF7, whose motif is GGGCGGGG). Indeed, the detection of KLF7 motifs might reflect the fact that many promoters are GC-rich (possessing what is known as CpG islands). One could more easily interpret footprints for TFs whose motifs are more complex as true hits (as is in the case of TCF4). As stated earlier, the prevalence of candidate TF footprints in randomly selected genes argues against the presence of a VGCC-specific regulator and suggests the prevalence of TF footprints in the human genome (including potentially false positive sites) or that the identified TFs might bind to many sites in the human genome. Moreover, it is possible that combinatorial regulation by multiple TFs could be in play for the regulation of VGCC expression. Finally, while motif enrichment would have been a good indicator that VGCCs are regulated by these TFs, the absence of enrichment does not mean that VGCCs are not also likely to be regulated by these TFs if the corresponding footprint is present. Many motifs are ubiquitous. On a related matter, another observation worth discussing was the range of number of footprints detected at the different VGCC loci. It remains unclear why certain loci had many footprints but others did not, but it is possible that this might be related to the signal strength (i.e. how open the chromatin peak is) at a particular locus (M. Bentsen, personal communications).

A potential key caveat to the analyses described herein is the incomplete replication of the results described in the original TOBIAS paper (Bentsen et al., 2020). For example, the curves of the aggregate footprints for different TFs implicated in human preimplantation development, also appeared not to be identical to those from the original paper, although they were similar. There were also differences in the number of binding sites serving as the basis for the aggregate footprints. The authors of the original TOBIAS paper were contacted, and they replied that these differences could be related to the differing versions of the tools used in the TOBIAS pipeline (M. Bentsen, personal communications). For example, the TOBIAS software has been updated 31 times since the original 2020 paper was published (M. Bentsen, personal communications). The version used in the original 2020 paper was 0.6.9 and the currently used version was 0.14.0 (M. Bentsen, personal communications). Unfortunately, the older version of TOBIAS is no longer available

to test this explanation. Nonetheless, the demonstrated sensitivity of TOBIAS to detect changes in differential motif enrichment at the global level as well as known binding events in PSCs suggest its utility for TF footprinting analyses.

There exist limitations inherent to TOBIAS and TF footprinting methods more generally. For example, TF footprinting methods rely on assumptions about DNA sequence preferences of TFs. While it is known that TFs possess sequence preferences, some TF binding patterns might not be accurately predicted based on DNA sequence alone. Some TFs may bind to DNA in a context-dependent manner, meaning that their binding preferences may change depending on the presence of other DNA-binding proteins. Moreover, TF footprinting methods may not take into account the dynamic nature of TF binding, which could be influenced by factors beyond chromatin accessibility, including DNA methylation, post-translational modifications and interactions with other proteins. While chromatin accessibility is important, such parameters likely affect the capacity of TFs to access and act at genomic sites and the stability of TF-DNA complexes. While TOBIAS is also a footprinting method, regarding limitations of TOBIAS more specifically, a recent pre-print has suggested TOBIAS is subject to false positive footprint results that are possibly related to limitations in how TOBIAS accounts for the Tn5 transposase sequence preference bias (Hu et al, 2023). In that pre-print, a machine learning approach was used to account for Tn5 bias correction, achieving improved results with less false positive footprints. Practitioners should proceed with caution when interpreting TF binding predictions and strongly consider the use of complementary methods to validate TF occupancy.

In this regard, an important question is how to define a "ground truth" for TF occupancy in the genome. ChIP-seq for genome-wide localization of TF targets is another key assay for determining TF occupancy. Inspection of ChIP-seq data from ENCODE suggests that human KLF7 is able to bind VGCC loci such as CACNB3 in alternative cell types such as HEK293 cells. Why KLF7 is not bound at the other VGCC loci could relate to the possibility that the chromatin accessibility state of VGCC loci in HEK293 cells might differ from hiPSC-derived neurons. Expression levels of KLF7 might also differ between the HEK293 cells overexpressing KLF7 and hiPSC-derived neurons that could affect genome-wide occupancy. Conducting ChIP-seq analyses for endogenous KLF7 and the other TFs of interest in the relevant cell type (i.e. hiPSC-derived neurons) would represent an experiment closer to ground truth.

It is important to note that generating gold standard datasets using ChIP is not always possible. For example, the lack of a high-quality antibody precludes the generation of high-quality ChIP data. To circumvent this, CRISPR/CAS9-mediated insertion of a biotin tag into the endogenous KLF7 locus, followed by streptavidin capture and sequencing could comprise a feasible alternative for analyzing KLF7 targets in the setting of normal, endogenous expression levels (Kim et al., 2008). However, it still cannot be fully excluded that the insertion of a biotin tag into the N- or C-terminus of KLF7 could affect the genome-wide targeting of KLF7. Indeed, the KLF7, REST, TCF4, and ZIC2 ChIP-seq datasets that were inspected involved ChIP for eGFP-TF fusion proteins that could affect results, such as failure to detect binding at other VGCC loci. Altogether, while ideally experimental data would represent a ground truth, the lack of gold standard datasets, such as those in the relevant cell type of interest, means inferring TF occupancy at a genomic site is best achieved by integrating insights from multiple assays. For example in this chapter, validation of TF footprinting with inspection of publicly available ChIP-seq data was conducted.

Nonetheless, the modified TOBIAS computational pipeline described here, which possesses the ability to identify putative TFs in an unbiased manner, could in theory be applied to any sufficiently powered dataset to identify the potential regulators of any gene.

4.4.3 Future Directions

In the future, it is advised to complement any bioinformatic analyses with experimental validation. In this regard, and as stated earlier, key experiments involve using drug-inducible expression of TFs or genetic ablation using CRISPR/CAS9 genome editing in hiPSCs to study TF function during differentiation or in the differentiated state. Such experiments would be accompanied by conducting ChIP-seq to assess genome-wide localization and RNA-seq analysis of the perturbed cells.

It will be of interest to gain an understanding of intermediate cellular states generated during the differentiation process. The exact number and identity of intermediate cellular states being generated during the differentiation process to a particular cell type are unknown (Fowler et al., 2020). It is not known when VGCCs are first upregulated during the differentiation process, and it is possible that the TFs that regulate VGCC expression might change between the different stages of the differentiation process. The

activation and stabilization of VGCC expression during differentiation and development could be a multi-step process involving different sets of TFs, as has been observed in the reprogramming of pluripotency-associated loci during iPSC generation (Li et al., 2010; Samvarchi-Tehrani et al., 2010; Golipour et al., 2012).

4.4.4 Conclusion

Here, I tested the capabilities of TOBIAS and applied it to identify a number of TFs that have footprints at the different VGCC loci in hiPSC-derived neurons. The list of TFs was refined based on differential gene expression analyses and on the number of VGCC loci where TF footprints were found. A search found these TF footprints in randomly selected genes, suggesting either the prevalence of such motifs in the genome and/or that the TFs might bind to many targets in the genome. Nonetheless, inspection of publicly available ChIP-seq datasets for some candidate TFs partially validated TOBIAS analyses, suggesting that some of these TFs could potentially regulate VGCC expression. The refined candidate TF list included KLF7 and TCF4. While numerous studies to date have focused on the functions and properties of VGCCs, as well as their post-transcriptional regulation, the transcriptional regulation of VGCCs has been less thoroughly explored. The findings described here have potential implications for understanding the regulation of VGCCs in the context of human neural development.

Chapter 5

Discussion

5.1 Introduction

This thesis has examined the transcriptional regulation of VGCCs in a putative model system of human neural development, the differentiation of hiPSCs into neurons. By applying publicly available bioinformatics tools to a set of hiPSC and hiPSC-derived neuron data, I characterized changes in the abundance and splicing of VGCCs during hiPSC differentiation (see **Chapter 2**). Additionally, I applied motif-based bioinformatics tools to identify candidate regulators of VGCC splicing and abundance (see **Chapter 3** and **Chapter 4**, respectively).

This Chapter will present a discussion of the results and an exploration of the implications in the context of psychiatry research. The discussion can be broadly divided into three different topics: links between VGCC biology and cell state, findings regarding identifying candidate RBPs and TFs, and hiPSCs as a model system. This Chapter will also suggest ways in which the work can be improved and suggest future lines of research to address questions resulting from the data described in this thesis.

5.2 Links Between VGCC Biology and Cell State

This thesis explored changes in abundance and splicing of VGCCs in a cell state transition of interest for neuroscience research, the differentiation of hiPSCs into neurons. **Chapter 2** analyzed a set of isogenic hiPSC and hiPSC-derived neuron data from 40 different individuals from the HIPSCI (Schwartzentruber et al., 2018). The analyses revealed robust changes in the abundance and the splicing of VGCC transcripts across differentiation into neurons. Generally, the transcript levels of different VGCC subunits substantially

increased as hiPSCs became neurons, consistent with emerging electrical activity. Moreover, alternative splicing events in CACNA1C (Exon 21 and 22), CACNA2D1 (Exon 28), and CACNA2D2 (Exon 26) were identified.

In the future, it would be important to relate these findings to the protein level. Regarding abundance, it is known that changes in mRNA levels do not always predict alterations in protein levels (Liu et al., 2016). Therefore, it will be key to confirm that VGCC protein levels also increase during differentiation by, for example, immunoblotting using high-quality antibodies. Additionally, it will be of interest to characterize the functional properties of alternatively spliced channels. My analyses identified a cell-state-specific switch in the usage of CACNA1C Exons 21 and Exon 22 in hiPSC-derived neurons. Differential utilization of Exon 21 and 22 can result in changes in the properties of the resulting channels, such as sensitivity to CCBs (Soldatov et al., 1995; Tiwari et al., 2006). Functional analyses of VGCC proteins could be achieved by studying hiPSC-derived neurons overexpressing specific isoforms of interest followed by investigations of electrophysiological and pharmacological properties.

A key limitation of the studies carried out is the fact that only two different cellular states (hiPSCs and differentiated neurons) were analyzed. This limitation in the study design makes it impossible to discern whether findings are specific to neuronal cells or related to being a non-pluripotent, differentiated cell type. Therefore, future experiments should involve analyzing cells and tissues, potentially from hiPSCs, of alternative lineages such as cardiomyocytes and smooth muscle cells. Carrying out analyses of different types of differentiated cells from the same parental hiPSCs would allow one to reduce the influence genetic background has on analyses (Kilpinen et al., 2017). These potential lines of research have translational relevance. There has been interest in repurposing CCBs currently used in clinical medicine for treating hypertension, and for the treatment of psychiatric conditions (Harrison et al., 2022). Characterizing and understanding differences in the alternative splicing of VGCCs between the brain and other tissues, such as cardiac and smooth muscle tissues, could lead to the development of novel, brain-selective CCBs that could avoid unwanted peripherally mediated side effects.

In addition to a lack of analysis of cells from various lineages, the studies are also limited by examining only two states of differentiation. Sampling differentiation cultures at different stages would allow one to examine if and how VGCC regulation changes as cells transition from a pluripotent to a differentiated neuron and potentially even different states

of neuronal maturation. It should be noted that the lack of maturity in iPS cell-derived neural derivatives, which retain fetal stage characteristics, remains a challenge in the field. Further optimization of differentiation protocols, co-culture systems, and potentially direct conversion techniques will allow practitioners to access even more mature, postnatal-like cellular states.

Different hiPSC neural differentiations produce neurons with features associated with different subtypes. A limitation of the studies reported in this thesis pertains to the fact that the hiPSC-derived neurons in the HIPSCI studies are sensory neurons with a nociceptor phenotype. In this regard, the observations with alpha-2-delta subunits and binding affinity for gabapentin might have relevance for the use of these in vitro-derived sensory neurons for studying clinical pain states. It is possible that the findings described in this thesis may not be readily extrapolated to cells and tissues of central nervous system identity, which have higher relevance to psychiatry. Reassuringly, the differential splicing of Exon 21 and Exon 22 in CACNA1C observed in hiPSC-derived neurons recapitulates the splicing pattern observed in human brain tissue, partially validating the hiPSC-derived neuron model (N. Hall, unpublished observations). The identification of putative TFs that regulate VGCC expression in hiPSC-derived neurons and their expression in fetal brain tissues raise the question of whether such TFs also perform any function in the development of the human brain. The computational pipelines described in this thesis could be applied to a more detailed analysis of RNA-seq and ATAC-seq data from the human fetal brain (Ziffra et al., 2021).

5.3 Interpreting Data From Motif-Based Bioinformatics Tools

This thesis pursued the identification of candidate proteins involved in the transcriptional regulation of VGCCs across hiPSC differentiation. To achieve this, publicly available motif-based bioinformatics tools were deployed. Motif-based bioinformatics tools such as FIMO, HOMER and to a lesser extent, TOBIAS, have been widely used to identify RBP and TF binding sites in RNA and DNA, respectively. The analyses conducted in this thesis suggest that these three tools, and perhaps motif-based bioinformatics tools more generally, possess limitations that should be considered when interpreting their predictions. One of these limitations is the potential for "noise" in the predictions. As stated previously, in this context, noise refers to the tendency to produce false positive or false negative results.

The noise in predictions made by motif-based bioinformatics tools can emerge from different sources. One such source is the degeneracy of sequence motifs that can produce a high rate of false positive results, as motifs may occur by chance in genomic regions being scanned. Indeed, an assessment of HOMER global's specificity found RBP motif occurrences in randomly selected genes. Additionally, an appreciable frequency of TOBIAS-identified TF footprints was observed at randomly selected genomic loci. Another potential source is the fact that these bioinformatics tools do not always consider the larger context in which RBPs or TFs operate, contributing to potentially inaccurate results, as RBP or TF binding is affected by factors beyond the sequence motif, such as RNA secondary structure, DNA methylation, among other factors. The advent of motif-based bioinformatics tool that consider context would prove useful for reducing noise. In this thesis, the use of chromatin accessibility data by TF footprinting methods, such as TOBIAS, to inform the identification of TF occupancy marks a step forward in terms of considering context when making binding site predictions. Very recent advances involving the development of TF footprinting tools that correct Tn5 sequence bias using deep learning and that can be applied to single-cell data could further reduce false positive results (Hu et al, 2023 pre-print).

The sensitivity analyses conducted in **Chapter 3**revealed that FIMO and HOMER cannot detect motif occurrences at most eCLIP peaks. As noted previously, the fact that widely-established tools such as FIMO and HOMER lack sensitivity was unexpected and highlights the importance of empirical validation. Nonetheless, it is challenging to determine whether the failure to detect a motif occurrence or footprint where an RBP-RNA or TF-DNA interaction has been experimentally observed represents an authentic false negative result. In this regard, what might be interpreted as noise in analyses can be made worse by the lack of high-quality empirical validation and gold standard datasets. In this thesis, inspection of data from eCLIP or ChIP-seq experiments was used to validate bioinformatics analyses. RBP eCLIP data were able to corroborate some FIMO and HOMER motif occurrences (see **Chapter 3**). Moreover, ChIP-seq data verified some of the TOBIAS footprinting predictions (see **Chapter 4**). However, these experimental assays have their own limitations. As noted earlier, the cross-linking used in eCLIP can introduce artifacts and biases. High-quality antibodies are not always available for ChIP to produce premium datasets. Even if such challenges could be overcome, datasets, particularly for experiments conducted in the relevant cell type of interest, are often not

available. Indeed, the failure to detect ChIP-seq binding events at different VGCC loci in the inspected ChIP-seq datasets where TOBIAS footprints were observed could reflect cell-type specific differences. For example, differences in epigenetic state between HEK293 cells and hiPSC-derived neurons could explain why footprints identified in hiPSC-derived neurons could not be validated in HEK293 cells.

Therefore, while motif-based bioinformatics tools can be powerful with regard to deriving lists of candidate regulators, the studies in this thesis indicate that researchers should integrate additional information beyond what can be gleaned from the sequence motif. I pursued this strategy in **Chapter 3**, where I intersected the results from different bioinformatics tools, assessed evolutionary conservation, evaluated differential gene expression across hiPSCs and hiPSC-derived neurons, and investigated expression patterns in adult human brain tissues, as well as the putative in vivo analog to hiPSC-derived neurons, the fetal brain. I employed a similar approach in Chapter 4, where I determined which TFs whose footprints were found at VGCC loci also exhibited differential gene expression, attempted to validate TF occupancy predictions by inspecting ChIP-seq data, and also interrogated expression patterns in fetal, postnatal and adult brain tissues. Additional evidence for the potential function for the candidate RBPs and TFs was also gained by surveying the literature. Altogether, by combining additional analyses with motif-based bioinformatics tools, it becomes possible to better discern what is a true regulator of VGCCs.

Ultimately, a limitation of the studies carried out in this thesis is the inability to infer causality. Without conducting experiments, it is not possible to distinguish between correlation and causation. Carrying out experiments in the laboratory would help to clarify the biological function of such RBPs, TFs, and their motif occurrences or footprints. Such studies involve experiments combining hiPSC-based experiments with genome editing or manipulation of expression levels of a key regulator. One type of experiment involves deleting the BRUNOL4 / BRUNOL5 binding site upstream of CACNA1C Exon 21 identified in Chapter 3. To determine the function of a candidate TF (identified in **Chapter 4**), transgenic overexpression or RNA interference by shRNA or more recent CasRx technology followed by RNA-seq could provide insights into targets regulated by a certain TF (Konermann et al., 2018).

5.4 Transcriptional and Epigenetic Regulation of VGCCs

While splicing of VGCCs has been studied, the transcriptional regulation of VGCCs, and particularly the identity of TFs that regulate VGCC expression, has been less well-explored. This thesis attempted to advance knowledge regarding the TF-mediated regulation of VGCCs (**Chapter 4**). **Chapter 4** applied the TF footprinting tool TOBIAS to identify TF footprints at VGCC loci. For some of the identified candidate TFs, a survey of the published literature confirmed that these TFs have functions related to mammalian neural biology. Interestingly, some TFs (KLF7, REST, TCF3, and TCF4) showed higher expression in fetal stages of development relative to postnatal stages. Future experiments involve overexpressing or knocking down a candidate TF and assessing for changes in VGCC expression levels. Additionally, it would be of interest to map the protein interaction network of KLF7 and TCF4 in hiPSC-derived neurons. This could potentially identify chromatin-remodeling enzymes that could also be involved in the regulation of VGCC expression.

Notably, some TFs associated with pioneer factor properties were found. It has been speculated that "pioneering" is a general strategy by which TFs program gene expression during cell state transitions, including both reprogramming and stem cell differentiation (Soufi et al., 2015; Meers et al., 2019). An understanding of the ability of TFs to access VGCC genomic loci in different cell types could be improved by studies of the VGCC loci chromatin state. While I have inspected chromatin accessibility data in this thesis, little is known about other epigenetic characteristics such as histone modifications and DNA methylation. Future experiments could be aimed at defining the epigenetic state of different VGCC loci in different cell types. Such experiments would involve generating datasets for various histone modifications, such as H3K4me3 (active), H3K27me3 (repressive), and H3K9me3 (repressive). Characterizing the DNA methylation status of VGCC loci will also be informative, as an antagonistic relationship has been reported between TF binding and DNA methylation.

5.5 Human iPSCs as a Model System

Owing to their capacity to self-renew indefinitely and make all cell types, hiPSCs are predicted to contribute significantly to neuroscience research. This thesis involved the study of human iPSC differentiation into neurons. Because neural cells derived from hiPSCs

often exhibit a fetal, rather than an adult, phenotype (Brennand et al., 2015), hiPSCs offer opportunities to study previously inaccessible stages of human neural development through the study of cells with a fetal phenotype. Below, I will briefly discuss hiPSC advances, the current limitations of using hiPSCs, along with alternative non-hiPSC model systems.

Recent developments are advancing the use of hiPSCs as models to study neural gene function. The advent of cerebral organoids illustrates this. Cerebral organoids are iPSC-derived, self-organizing three-dimensional cultures of cells with a neural identity that model the human brain (Lancaster et al., 2013). Such organoids have attracted significant interest as a novel model system to investigate diseases (Quadrato et al. 2017; Velasco et al., 2019). In addition, they have modeled various brain conditions, including autism spectrum disorder, microcephaly, Zika virus infection and of relevance for this thesis, TS (Lancaster et al., 2013; Mariani et al., 2015; Garcez et al., 2016; Birey et al., 2017). It should be noted that questions have arisen regarding the extent to which these organoids recreate neural development (Bhaduri et al., 2020). Regarding VGCC biology, experiments involving the transplantation of organoids into mice unveiled neuronal defects in cells derived from TS patients (Revah et al, 2022). In these experiments, cerebral organoids were generated from three TS patients carrying the common substitution (pG406R) and from three control patients. Following organoid transplantation, TS neurons possessed an altered dendritic morphology, with an increase in primary dendrite number and a reduction in dendrite length (Revah et al., 2022). Further, an increase in synaptic spine density and a higher occurrence of spontaneous EPSCs were observed in TS samples. These experiments demonstrate that in vivo organoid experiments are useful for unveiling circuit-level phenotypes (Revah et al., 2022).

While hiPSCs comprise a model system, there do exist certain obstacles when one considers the use of hiPSCs. These limitations originate from the inter-cell line variability inherent to human pluripotent stem cells and events that could originate from the reprogramming process. Three such limitations involve: variable efficiency of differentiation between cell lines, genetic instability, and the possibility of epigenetic memory. The efficiencies across different human PSC lines for differentiating into specific neuronal or glial cell types may vary, potentially affecting the reproducibility of results (Osafune et al., 2008). Single-cell RNA-seq and single-cell ATAC seq data have revealed substantial heterogeneity during hiPSC differentiation (Fowler et al., 2020). Studies have reported that

the process of reprogramming somatic cells in hiPSCs can introduce genetic abnormalities, which could affect the functionality of derived neurons and glial cells (Martins-Taylor et al., 2011; Taapken et al., 2011). Moreover, the reprogramming process can generate "intermediate" cell types that while retaining some characteristics of iPSC lines have not been fully reprogrammed (Mikkelsen et al., 2008). A possible strategy to overcome such limitations could involve conducting reprogramming or maintaining hiPSCs in specialized culture conditions that retain cells in a state of naïve pluripotency, but these experimental conditions still possess a predisposition for genetic instability (Theunissen et al., 2014).

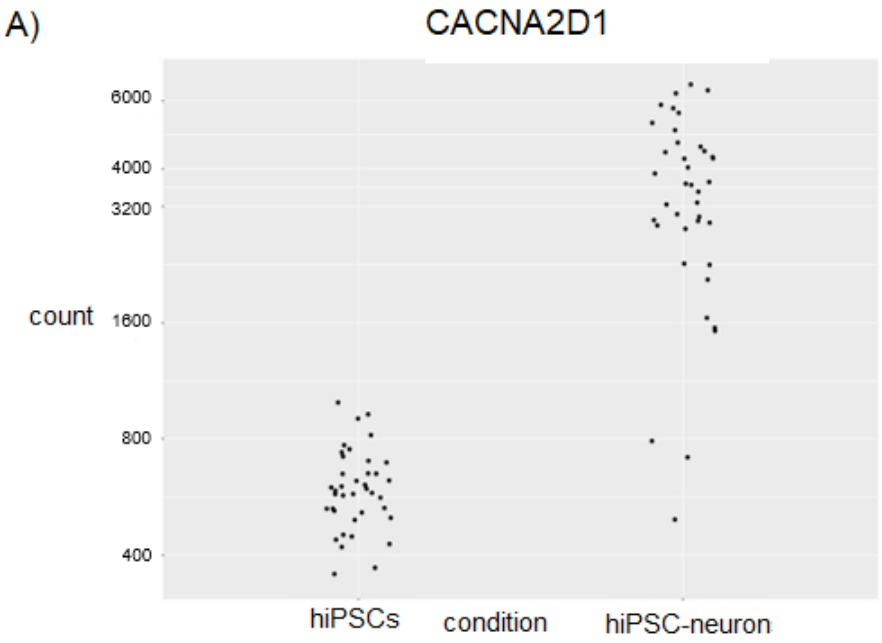
To overcome some of the limitations associated with hiPSCs, it is also worth considering there exist alternatives to hiPSCs for studying the transcriptional regulation of VGCCs. Three such experimental systems include: primary neuronal cultures, organotypic brain slices, and in vivo animal models. Brain slices and animal models provide a model system that maintains the complex architecture of the brain. These systems can provide insights into VGCC function in the context of neural circuits and synapses that may not be possible in "simpler" model systems, such as hiPSCs differentiated into neurons. Moreover, animal models allow practitioners to study VGCC regulation in different physiological and pathological states – as well as use them as a testbed for potential therapies for VGCC-related diseases. Two disadvantages of these alternative approaches include: first, technical challenges associated with the preparation of primary neuronal cultures and brain slices; and second, ethical considerations surrounding the use of animal models.

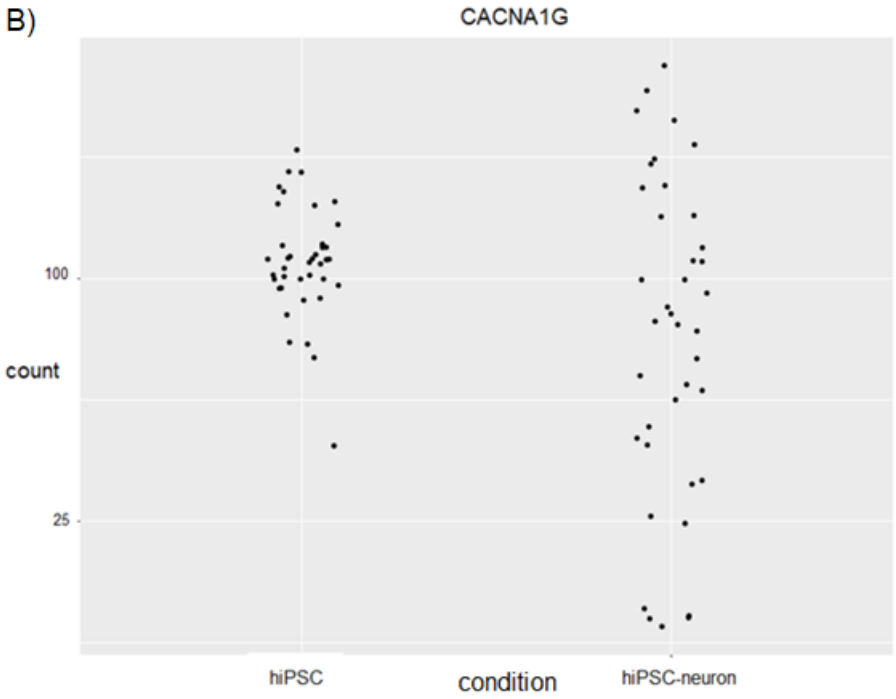
5.6 Conclusions and Perspectives

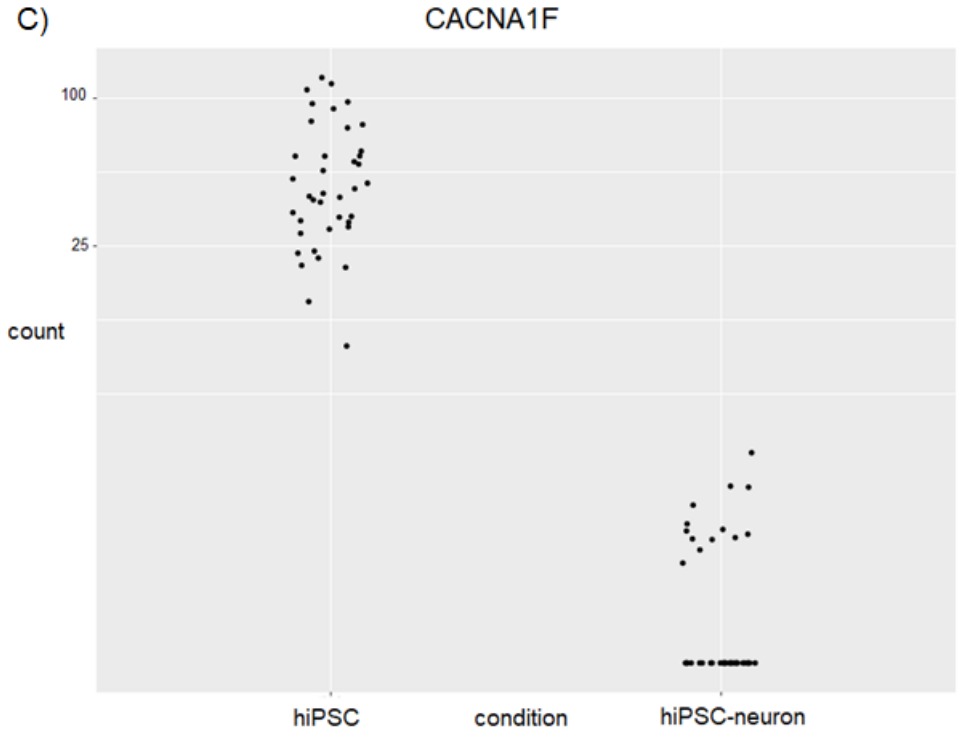
Human neural development, particularly in the fetal stages, is a phase where genetic risk of certain psychiatric conditions might be played out. While VGCCs have been linked to psychiatric illness, the regulation of VGCCs during human neural development has not been sufficiently characterized. In this thesis, I studied the transcriptional regulation of VGCCs during hiPSC differentiation into neurons, a putative model system for human neural development. I identified changes in the abundance and splicing of VGCCs across differentiation. The splicing events of interest occurred in CACNA1C, CACNA2D1, and CACNA2D2. A key cell state-associated switch in the usage of Exon 21 and Exon 22 of CACNA1C was found, which, as suggested by bioinformatic analyses, could be caused by the binding of BRUNOL4 and/or BRUNOL5 to a site upstream of Exon 21. Additionally,

analyses identified an HNRNPL binding site in the CACNA2D2 splicing region of interest that showed evolutionary acceleration, but caution is warranted in interpreting evolutionary conservation scores in repeat regions. The study also identified putative regulators of VGCC abundance using TF footprinting. The list of resulting candidate TFs included KLF7 and TCF4, which have all been linked to neural biology and central nervous system function. Altogether, the results indicate that hiPSCs possess the potential for helping to uncover mechanisms of VGCC regulation in a model system of human neural development. The complete picture of the function and regulation of VGCCs during human neural development is yet to be painted, but the discovery of transcriptional changes and putative regulators described here can help address the gaps in our knowledge.

Appendix I Supplementary Material for Chapter 2







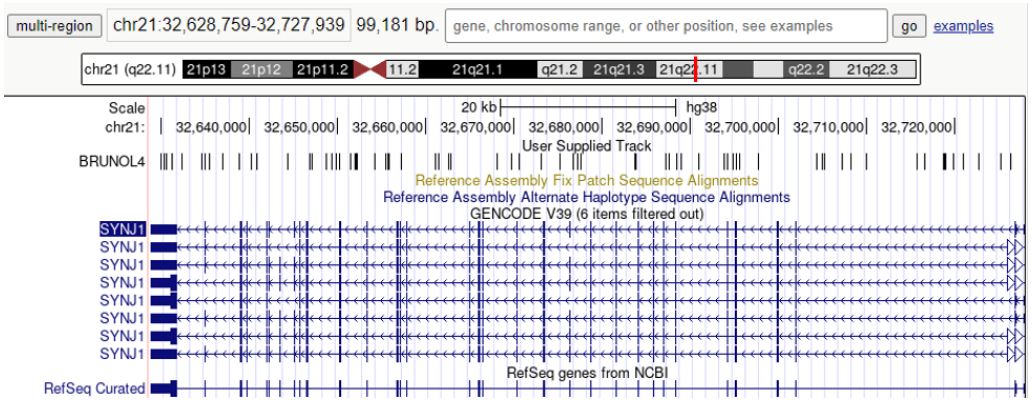
Supplementary Figure 1.1 Differential gene expression analyses of individual VGCC subunits.

Presented are dot plots containing normalized read counts for three individual VGCC subunits in either hiPSCs or hiPSC-derived neurons (a) CACNA1C, (a) CACNA2D1; unchanged between hiPSCs and hiPSC-neurons: (b) CACNA1G; and upregulated in hiPSCs: (c) CACNA1F. For other VGCC subunits, data not shown. Each sample is represented with a dot.

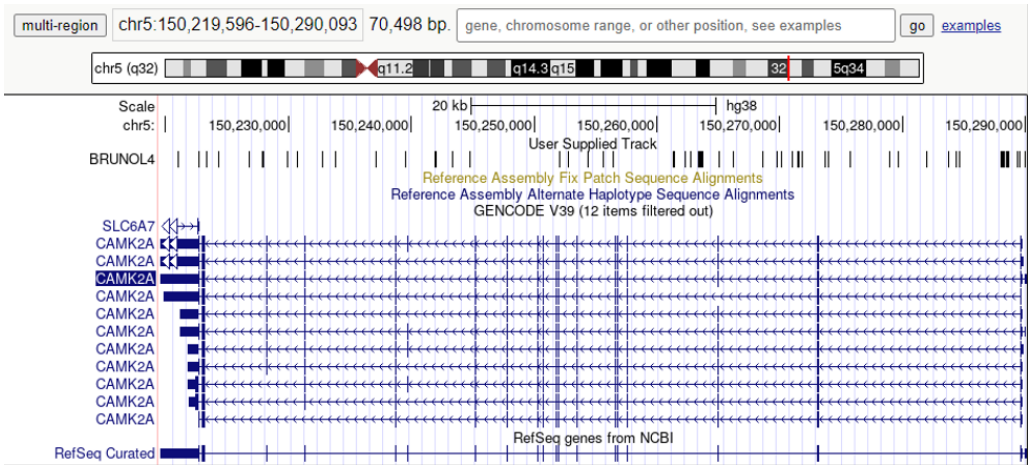
Appendix II Supplementary Material for Chapter 3

Additional Information Regarding FIMO and HOMER sensitivity

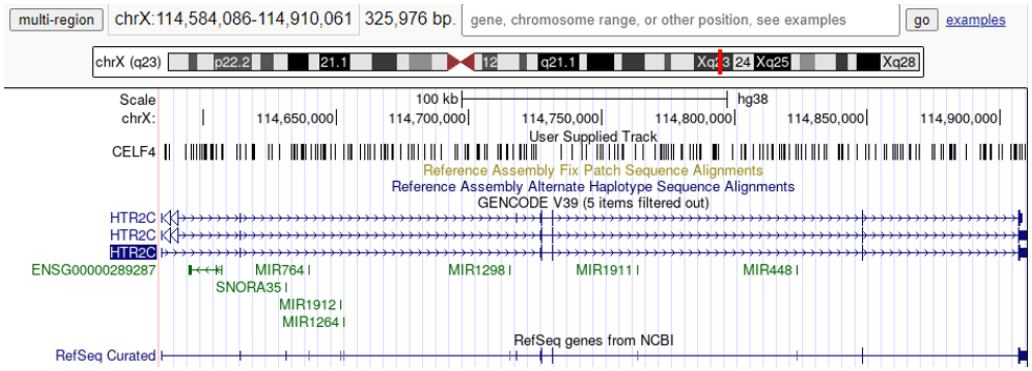
In Section **3.3.1.1**., the sensitivity of FIMO was assessed by testing the capacity of FIMO to detect RBP motif occurrences in known targets from the literature or experimental eCLIP data. The findings regarding literature targets are provided here. Regarding known targets, FIMO successfully identified motif occurrences in three out of four known BRUNOL4 targets (SYNJ1, CAMK2A, and SNCA), but failed to successfully identify motif occurrences in a fourth known BRUNOL4 target NSF (data not shown) (Wagnon et al, 2012). The sensitivity of HOMER was assessed in **3.3.1.1** by searching for motif occurrences in known targets or eCLIP data. For searches in known targets, the presence of motif occurrences of BRUNOL4 in BRUNOL4 targets known in the literature (SYNJ1, CAMK2A, HTR2C, SNCA, and NSF) was confirmed (**Supplementary Figure 2.1**). Additionally, the presence of motif occurrences of ELAVL2 in ELAVL2 target DDX6 was confirmed (**Supplementary Figure 2.2**). A more detailed assessment of FIMO's and HOMER's sensitivity to detect motif occurrences at eCLIP peaks is presented in **3.3.1.1.**


multi-region
chr21:32,628,759-32,727,939
99,181 bp.
gene, chromosome range, or other position, see examples
go
examples
chr21 (q22.11)
21p13
21p12
21p11.2
11.2
21q21.1
q21.2
21q21.3
21q22.11
q22.2
21q22.3
Scale
20 kb
hg38
chr21:
32,640,000
32,650,000
32,660,000
32,670,000
32,680,000
32,690,000
32,700,000
32,710,000
32,720,000
User Supplied Track
BRUNOL4
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V39 (6 items filtered out)
SYNJ1
RefSeq genes from NCBI
RefSeq Curated

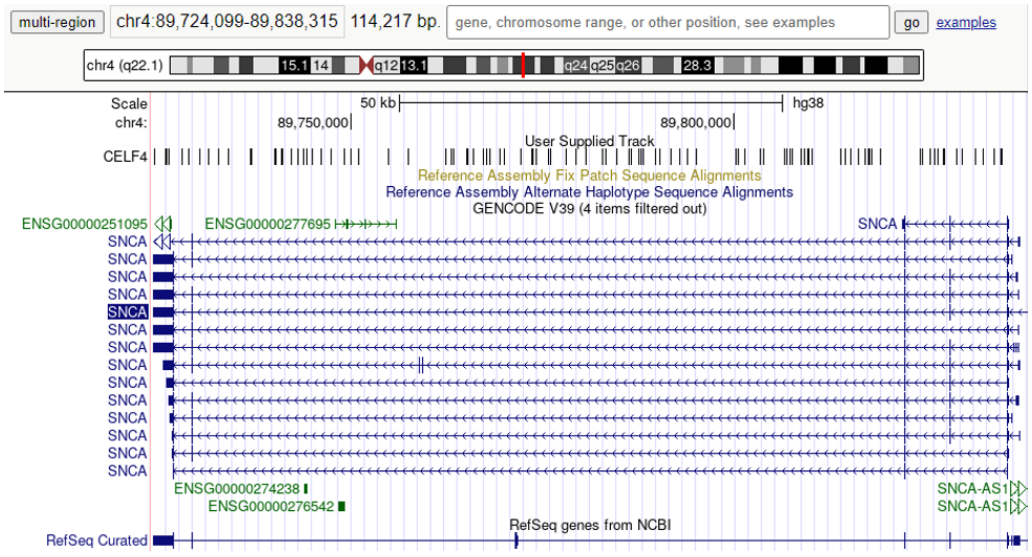
(A)


multi-region
chr5:150,219,596-150,290,093
70,498 bp.
gene, chromosome range, or other position, see examples
go
examples
chr5 (q32)
q11.2
q14.3
q15
32
5q34
Scale
20 kb
hg38
chr5:
150,230,000
150,240,000
150,250,000
150,260,000
150,270,000
150,280,000
150,290,000
User Supplied Track
BRUNOL4
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V39 (12 items filtered out)
SLC6A7
CAMK2A
RefSeq genes from NCBI
RefSeq Curated

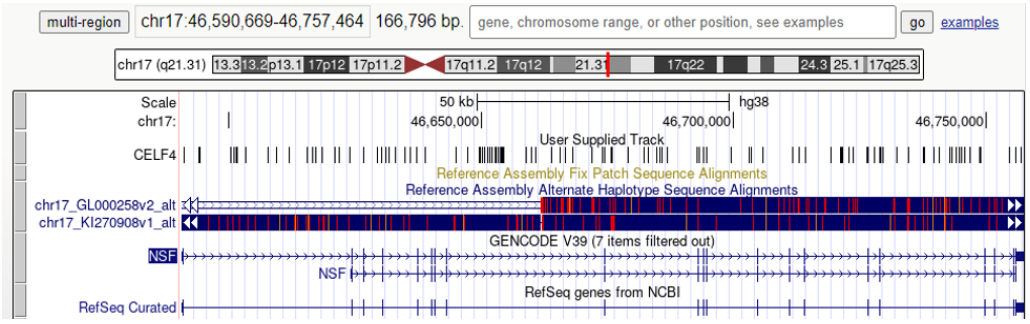
(B)


multi-region
chrX:114,584,086-114,910,061
325,976 bp.
gene, chromosome range, or other position, see examples
go
examples
chrX (q23)
p22.2
21.1
12
q21.1
Xq24
Xq25
Xq28
Scale
100 kb
hg38
chrX:
114,650,000
114,700,000
114,750,000
114,800,000
114,850,000
114,900,000
User Supplied Track
CELF4
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V39 (5 items filtered out)
HTR2C
ENSG00000289287
MIR764
SNORA35
MIR1912
MIR1264
MIR1298
MIR1911
MIR448
RefSeq genes from NCBI
RefSeq Curated

(C)


multi-region
chr4:89,724,099-89,838,315
114,217 bp.
gene, chromosome range, or other position, see examples
go
examples
chr4 (q22.1)
15.1
14
q12
13.1
q24
q25
q26
28.3
Scale
50 kb
hg38
chr4:
89,750,000
89,800,000
User Supplied Track
CELF4
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V39 (4 items filtered out)
ENSG00000251095
ENSG00000277695
SNCA
ENSG00000274238
ENSG00000276542
SNCA-AS1
RefSeq genes from NCBI
RefSeq Curated

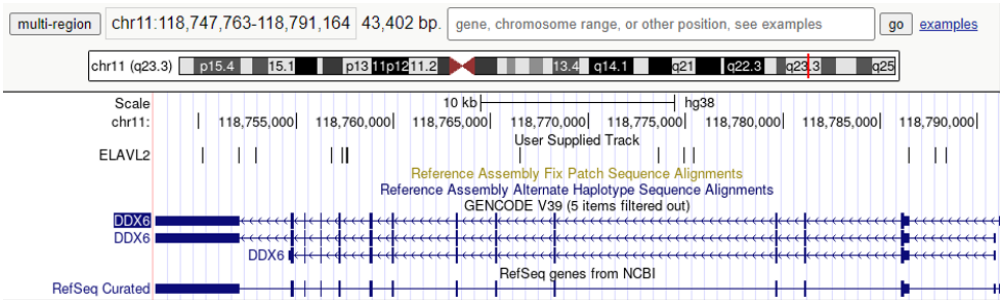
(D)



(E)

Supplementary Figure 2.1. HOMER global search for BRUNOL4 binding motif occurrences at known BRUNOL4 targets.

(A) SYNJ1, (B) CAMK2A, (C) HTR2C, (D) SNCA, and I NSF (Wagnon et al., 2012). Searches were also conducted for motif occurrences of a second RBP, in known ELAVL2 targets. ELAVL2 binding motif occurrences could be detected at the locus of DDX6, a known ELAVL2 target (Kato et al., 2019) (**Supplementary Figure 2.3**).



Supplementary Figure 2.2 HOMER global search for ELAVL2 binding motif occurrences at DDX6 locus.

HOMER global search for ELAVL2 binding motif occurrences at DDX6, a known ELAVL2 target (Kato et al., 2019).

The frequency of BRUNOL4 binding sites in the above-selected genes raised questions about HOMER's specificity and the prevalence of BRUNOL4 binding motif occurrences more generally. Searches for BRUNOL4 binding sites in ten randomly selected genes were conducted using an in-house generated random gene selector script. The randomly selected genes were RRAGA, OSTC, ISM1, KCNQ1, CDRT15P3, SPATA13, LENG9, OR4A8, NAPA-AS1, and TRAJ61.

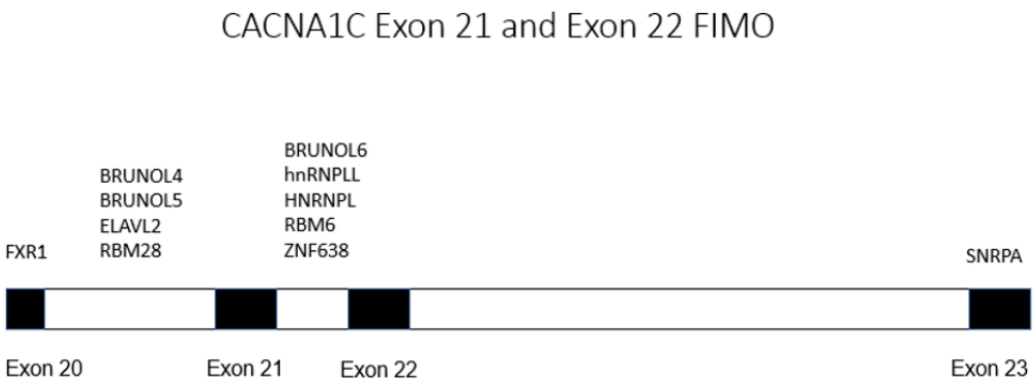
For BRUNOL4: in RRAGA, OSTC, KCNQ1, CDRT15P3, SPATA13, LENG9, and NAPA-AS1, binding sites were observed. In ISM1, OR4A8, and TRAJ61, no binding sites were observed. For ELAVL2: in OSTC, ISM1, KCNQ1, SPATA13, OR4A8, and NAPA-AS1, binding sites were observed. In RRAGA, CDRT15P3, LENG9, and TRAJ1, no binding sites were observed.

Altogether, HOMER global was able to observe motif occurrences for BRUNOL4 and ELAVL2 in known targets from the literature. However, HOMER global also identified

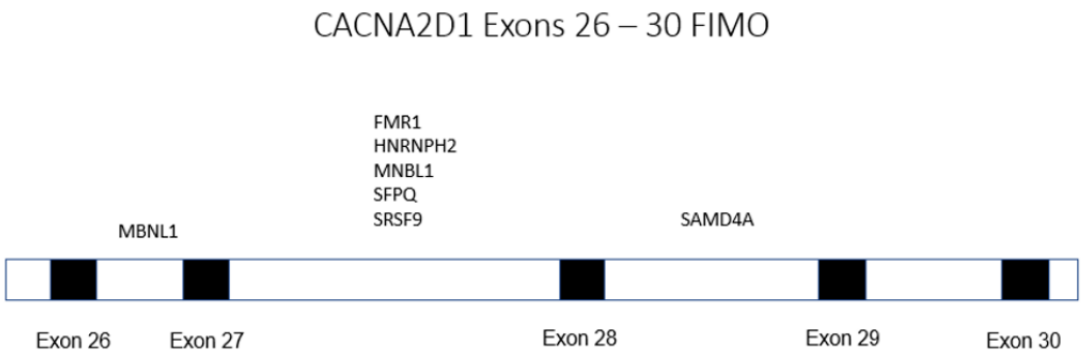
a significant number of BRUNOL4 and ELAVL2 motif occurrences in randomly selected genes. This suggests that BRUNOL4 and ELAVL2 binding sites are common in the genome and a possible predisposition of HOMER to produce false positive results.

FIMO Searches for RBP Motif Occurrences in VGCC Regions of Interest

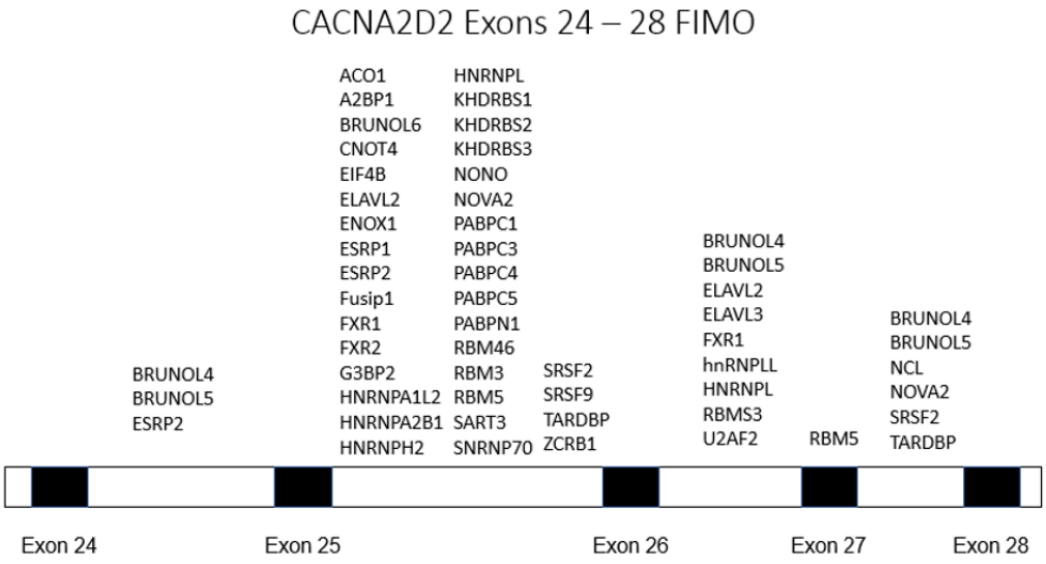
As noted in **3.3.2.**, results from separate scans using FIMO and HOMER were intersected. For separate searches using FIMO, To identify candidate splicing regulators, sequences in the regions of interest in VGCC subunits CACNA1C, CACNA2D1, and CACNA2D2 were first scanned for motif instances. The positions of the identified RBP binding sites in CACNA1C, CACNA2D1, and CACNA2D2 are indicated in **Supplementary Figure 2.3.**



(A)



(B)



(C)

Supplementary Figure 2.3 RBP Motifs enriched at CACNA1C, CACNA2D1, and CACNA2D2 regions of interest.

(A) CACNA1C exon 21 and exon 22 region identified by FIMO. Introns and exons were separately input into FIMO spanning the 5' flanking exon (in this case exon 20) to the 3' flanking exon (in this case exon 23). (B) RBP Motifs enriched at the CACNA2D1 region of interest identified by FIMO. Introns and exons were separately input into FIMO spanning the 5' flanking exon (in this case exon 26) to the 3' flanking exon (in this case exon 30). (C) RBP Motifs enriched at the CACNA2D2 region of interest identified by FIMO. Introns and exons were separately input into FIMO spanning the 5' flanking exon (in this case exon 24) to the 3' flanking exon (in this case exon 28). Black indicates exons and white indicates introns. Exons and introns are not drawn to scale.

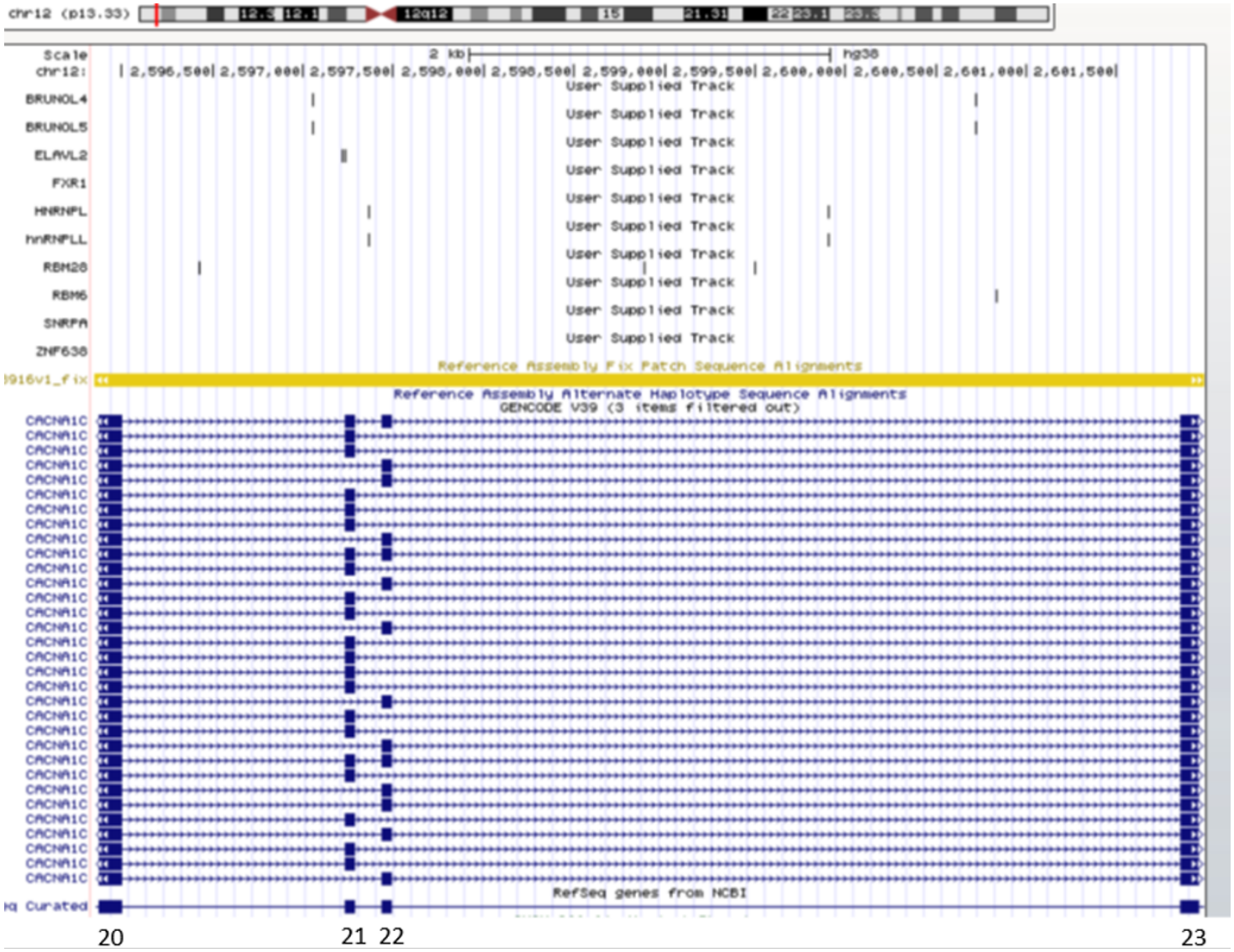
Altogether, FIMO identified RBP binding motifs at VGCC regions of interest nearly exclusively in introns, with the exception of a few RBP binding motif occurrences identified in exons in CACNA1C and CACNA2D2.

HOMER Global Searches for Motif Occurrences in VGCC Regions of Interest

As noted in **3.3.2.**, separate scans for FIMO and HOMER were conducted. Next, the human genome was scanned for binding motif occurrences of the RBPs identified by FIMO analyses using HOMER global (Heinz et al., 2010). Subsequently, RBP motif occurrences at the VGCC regions of interest were investigated in greater detail (i.e. in CACNA1C, CACNA2D1, and CACNA2D2).

HOMER Global Results at CACNA1C

A search of the CACNA1C region of interest is presented in **Supplementary Figure 2.4.**



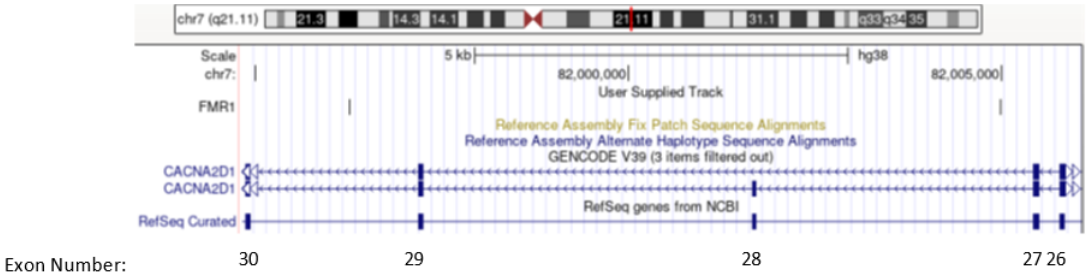
Supplementary Figure 2.4 HOMER global search at CACNA1C region of interest Exons 21 and 22.

Shown are UCSC genome browser tracks between Exons 20 – 23. Thin bars represent motif occurrences for the indicated RBP. Exon numbers are indicated.

A closer, higher-magnification examination of the CACNA1C exons 21 and 22 regions identified BRUNOL4, BRUNOL5, and HNRNPL motif occurrences in the close vicinity of exon 21.

HOMER Global Results at CACNA2D1

The HOMER global search data for the CACNA2D1 region of interest is presented In **Supplementary Figure 2.5.** HOMER global failed to generate BED files for HNRNPH2, MBNL1, SAMD4A, SFPQ, and SRSF9 RBPs that had been identified by FIMO. In contrast, HOMER global identified FMR1 binding motif occurrences in the introns between 27 and 28; and 29 and 30.



Supplementary Figure 2.5 HOMER global search at CACNA2D1 region of interest.

Exon 28 is differentially spliced between hiPSCs and hiPSC-derived neurons. Shown above are UCSC genome browser tracks between Exons 26 and 30. Thin bars represent motif occurrences for the indicated RBP. Exon numbers are indicated.

HOMER Global Results at CACNA2D2

The HOMER global search results for the CACNA2D2 region of interest are presented in **Supplementary Figure 2.6**. A visual inspection of the UCSC Genome Browser tracks indicated that while HOMER identified motifs from a number of RBPs identified by FIMO at CACNA2D2, there were also a significant proportion of FIMO-identified RBPs not identified by HOMER global.



Supplementary Figure 2.6 HOMER global search at the CACNA2D2 region of interest.

Exon 26 is differentially spliced between hiPSCs and hiPSC-derived neurons. HOMER global search for FIMO-identified RBPs at CACNA2D2 region of interest (CACNA2D2 Exons 24-28). Thin bars represent motif occurrences for the indicated RBP. Exon numbers are indicated.

Species	Scientific Name
Primate	
Human	Homo sapiens
Chimp	Pan troglodytes
Gorilla	Gorilla
Orangutan	Pongo pygmaeus abelii
Gibbon	Nomascus leucogenys
Rhesus	Macaca mulatta
Crab-eating macaque	Macaca fascicularis
Baboon	Papio hamadryas
Green monkey	Chlorocebus sabaeus
Marmoset	Callithrix jacchus
Squirrel monkey	Saimiri boliviensis
Bushbaby	Otolemur garnettii
Euarchontoglires	
Chinese tree shrew	Tupaia chinensis
Squirrel	Spermophilus
Lesser Egyptian jerboa	tridecemlineatus
Prairie vole	Jaculus jaculus
Chinese hamster	Microtus ochrogaster
Golden hamster	Cricetulus griseus
Mouse	Mesocricetus auratus
Rat	Mus musculus
Naked mole-rat	Rattus norvegicus
Guinea pig	Heterocephalus glaber
Chinchilla	Cavia porcellus
Brush-tailed rat	Chinchilla lanigera
Rabbit	Octodon degus
Pika	Oryctolagus cuniculus
	Ochotona princeps
Laurasiatheria	
Pig	Sus scrofa
Alpaca	Vicugna pacos
Bactrian camel	Camelus ferus
Dolphin	Tursiops truncatus
Killer whale	Ornicus orca
Tibetan antelope	Pantholops hodgsonii
Cow	Bos taurus
Sheep	Ovis aries
Domestic goat	Capra hircus
Horse	Equus caballus
White rhinoceros	Ceratotherium simum
Cat	Felis catus
Dog	Canis lupus
Ferret	Mustela putorius
Panda	Ailuropoda melanoleuca
Pacific walrus	Odobenus rosmarus
Weddell seal	Leptonychotes weddellii
Black flying-fox	Pteropus alecto
Megabat	Pteropus vampyrus
Davi's myotis bat	Myotis davidii
Microbat	Myotis lucifugus
Big brown bat	Eptesicus fuscus
Hedgehog	Erinaceus europaeus
Shrew	Sorex araneus
Star-nosed mole	Condylura cristata

Afrotheria	
Elephant	Loxodonta africana
Cape elephant shrew	Elephantulus edwardii
Manatee	Trichechus manatus
Cape golden mole	latirostris
Tenrec	Chrysochloris asiatica
Aardvark	Echinops telfairi
	Orycteropus afer afer
Mammal	
Armadillo	Dasypus novemcinctus
Opossum	Monodelphis domestica
Tasmanian devil	Sarcophilus harrisii
Wallaby	Macropus eugenii
Platypus	Ornithorhynchus anatinus
Aves	
Saker falcon	Falco cherrug
Peregrine falcon	Falco peregrinus
Collared flycatcher	Ficedula albicollis
White-throated	Zonotrichia albicollis
sparrow	Geospiza fortis
Medium ground finch	Taeniopygia guttata
Zebra finch	Pseudopodoces humilis
Tibetan ground jay	Melopsittacus undulatus
Budgerigar	Amazona vittata
Parrot	Ara macao
Scarlet macaw	Columba livia
Rock pigeon	Anas platyrhynchos
Mallard duck	Gallus gallus
Chicken	Meleagris gallopavo
Turkey	
Sarcopterygii	
American alligator	Alligator mississippiensis
Green seaturtle	Chelonia mydas
Painted turtle	Chrysemys picta bellii
Chinese softshell turtle	Pelodiscus sinensis
Spiny softshell turtle	Apalone spinifera
Lizard	Anolis carolinensis
X. tropicalis	Xenopus tropicalis
Coelacanth	Latimeria chalumnae
Fish	
Tetraodon	Tetraodon nigroviridis
Fugu	Takifugu rubripes
Yellowbelly pufferfish	Takifugu flavidus
Nile tilapia	Oreochromis niloticus
Princess of Burundi	Neolamprologus brichardi
Burto's mouthbreeder	Haplochromis burtoni
Zebra mbuna	Maylandia zebra
Pundamilia nyererei	Pundamilia nyererei
Medaka	Oryzias latipes
Southern platyfish	Xiphophorus maculatus
Stickleback	Gasterosteus aculeatus
Atlantic cod	Gadus morhua
Zebrafish	Danio rerio
Mexican tetra	Astyanax mexicanus
(cavefish)	Lepisoteus oculatus
Spotted gar	Petromyzon marinus
Lamprey	

Supplementary Table 2.1 100 vertebrate species analyzed in 100 species evolutionary conservation analysis.

PhyloP scores from 100 species were analyzed for 100 species evolutionary conservation analysis. Left: species name. Right: scientific name. Taken from the UCSC Genome Browser website.

Species	Scientific name
Human	Homo sapiens
Baboon	Papio anubis
Bushbaby	Otolemur garnettii
Bonobo	Pan paniscus
Chimp	Pan troglodytes
Crab-eating macaque	Macaca fascicularis
Gibbon	Nomascus leucogenys
Golden snub-nosed monkey	Rhinopithecus roxellana
monkey	
Gorilla	Gorilla gorilla
Green monkey	Chlorocebus savaeus
Marmoset	Callithrix jacchus
Mouse lemur	Microcebus murinus
Orangutan	Pongo pygmaeus
Proboscis monkey	Nasalis larvatus
Rhesus	Macaca mulatta
Squirrel monkey	Saimiri boliviensis
Tarsier	Tarsius syrichta

Supplementary Table 2.2 17 primate species analyzed in 100 species evolutionary conservation analysis.

PhyloP scores from 17 species were analyzed for evolutionary conservation analysis. Left: species name. Right: scientific name. Taken from the UCSC Genome Browser website.

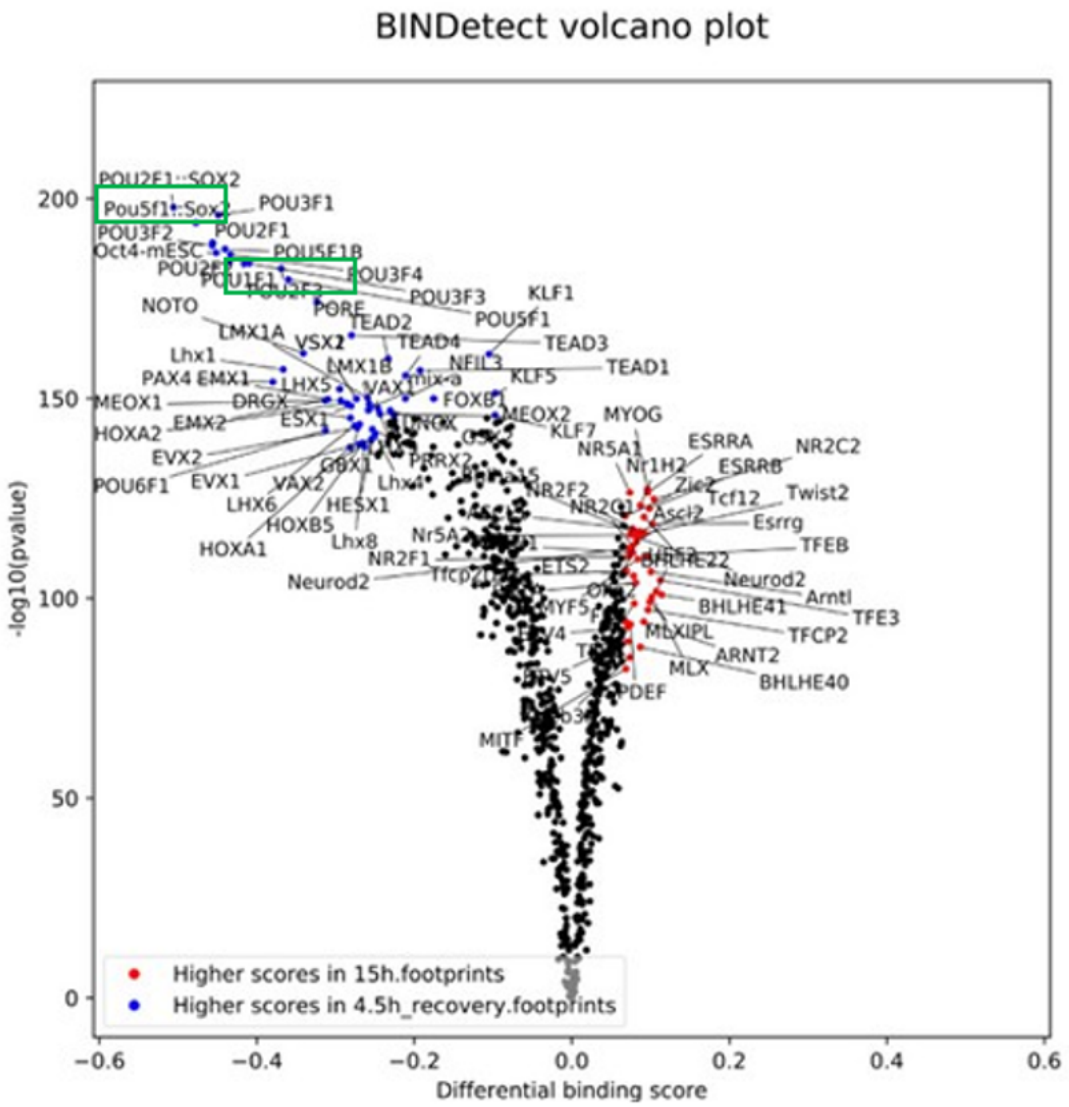
Stage 1	Stage 2A	Stages 2B-11
Frontal (Anterior) cortical wall	Frontal cortical wall	Dorsolateral prefrontal cortex (FC)
		Ventrolateral FC
		Medial FC
		Orbital FC
		Primary motor cortex
Parietal (Middle dorsal) cortical wall	Parietal cortical wall	Primary somatosensory cortex
		Posterior inferior parietal cortex
Temporal (Postero-lateral) cortical wall	Temporal cortical wall	Primary auditory temporal cortex
		Posterior superior temporal cortex
		Inferior temporal cortex
Occipital (Posterior) cortical wall	Occipital cortical wall	Primary visual cortex (V1)
Hippocampal anlage	Hippocampal anlage	Hippocampus
		Amygdala
Ventral forebrain	Medial ganglionic eminence	Striatum
	Lateral ganglionic eminence	
	Caudal ganglionic eminence	
Dorsal diencephalon	Dorsal thalamic anlage	Mediodorsal nucleus of thalamus
Upper rhombic lip	Upper rhombic lip	Cerebellar cortex

Supplementary Table 2.3 Brain regions profiled by RNA sequencing for BrainSpan Developmental Transcriptome analyses.

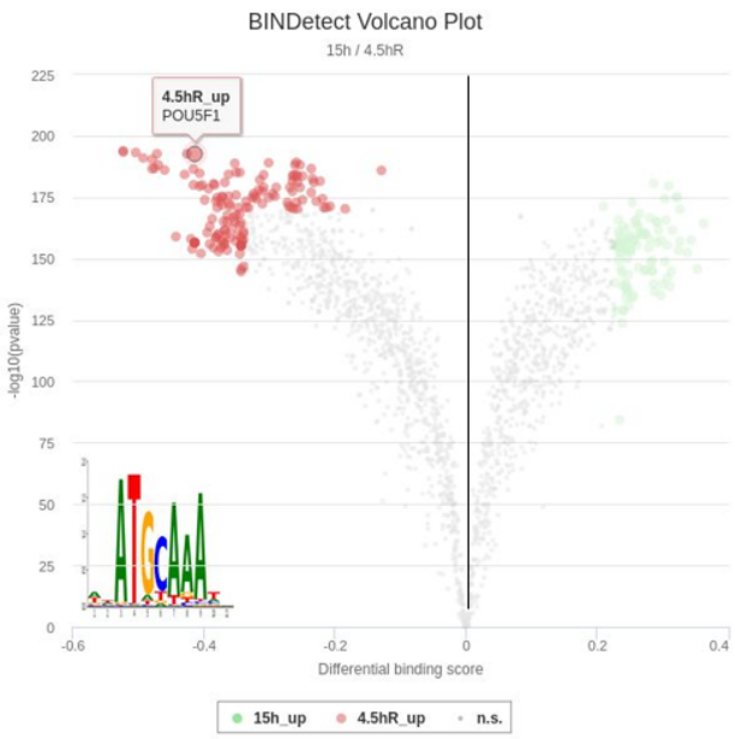
Indicated are the different regions of the fetal and postnatal brains from which sample data were derived. Taken from www.brainspan.org. For heatmap data in **Chapter 3** and **Chapter 4**, data from Stage 1, Stage 2A, and Stages 2B-11 samples are plotted (includes data from all brain regions). For statistical comparisons of fetal versus postnatal expression levels, only frontal cortex (not all brain regions) samples from Stage 2A and Stages 2B-11 were analyzed (as RPKM values for Stage 1 samples were not available). For the fetal versus postnatal comparison, the fetal stages comprised stage 2A to stage 5; and the postnatal samples comprise stage 6 to stage 11. Analysis of early fetal versus late fetal expression levels involved only frontal cortex (not all brain regions) samples from Stage 2A and Stage 2B to Stage 5. For the early fetal versus late fetal comparison, the early fetal stage samples comprised stages 2A to stage 3B samples and the late fetal stage samples comprised stage 4 samples to stage 5 samples. Stage 1: 4-7 pcw; stage 2A: 8-9 pcw; stage 2B: 10-12 pcw; stage 3A: 13-15 pcw; stage 3B: 16-17 pcw; stage 4: 19-24 pcw; stage 5: 25-38 pcw; stage 6: birth-5 months; stage 7: 6-18 months; stage 8: 19 months – 5 years; stage 9: 6-11 years; stage 10: 12-19 years; stage 11: 20-60+ years.

Appendix III Supplementary Material for Chapter 4

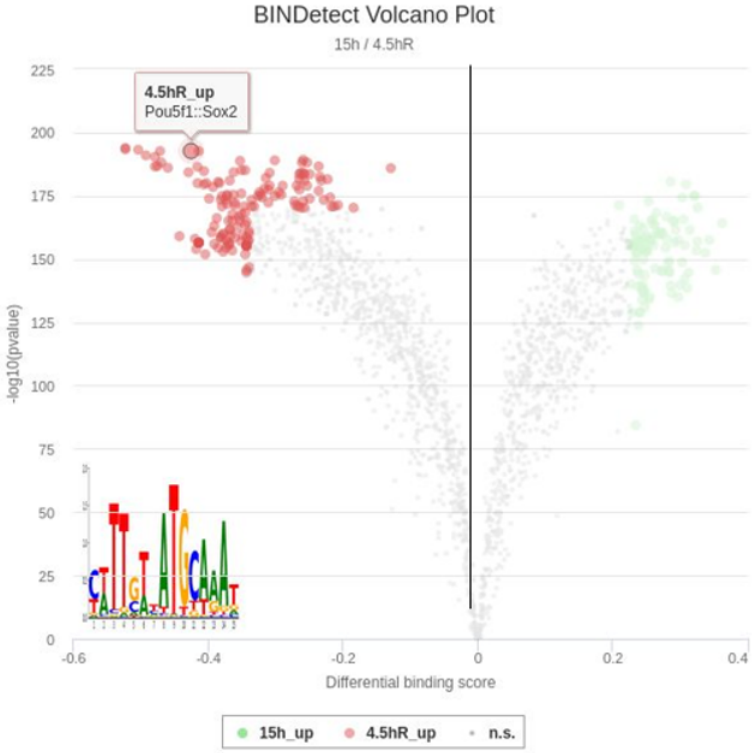
Figure from Xiong et al.



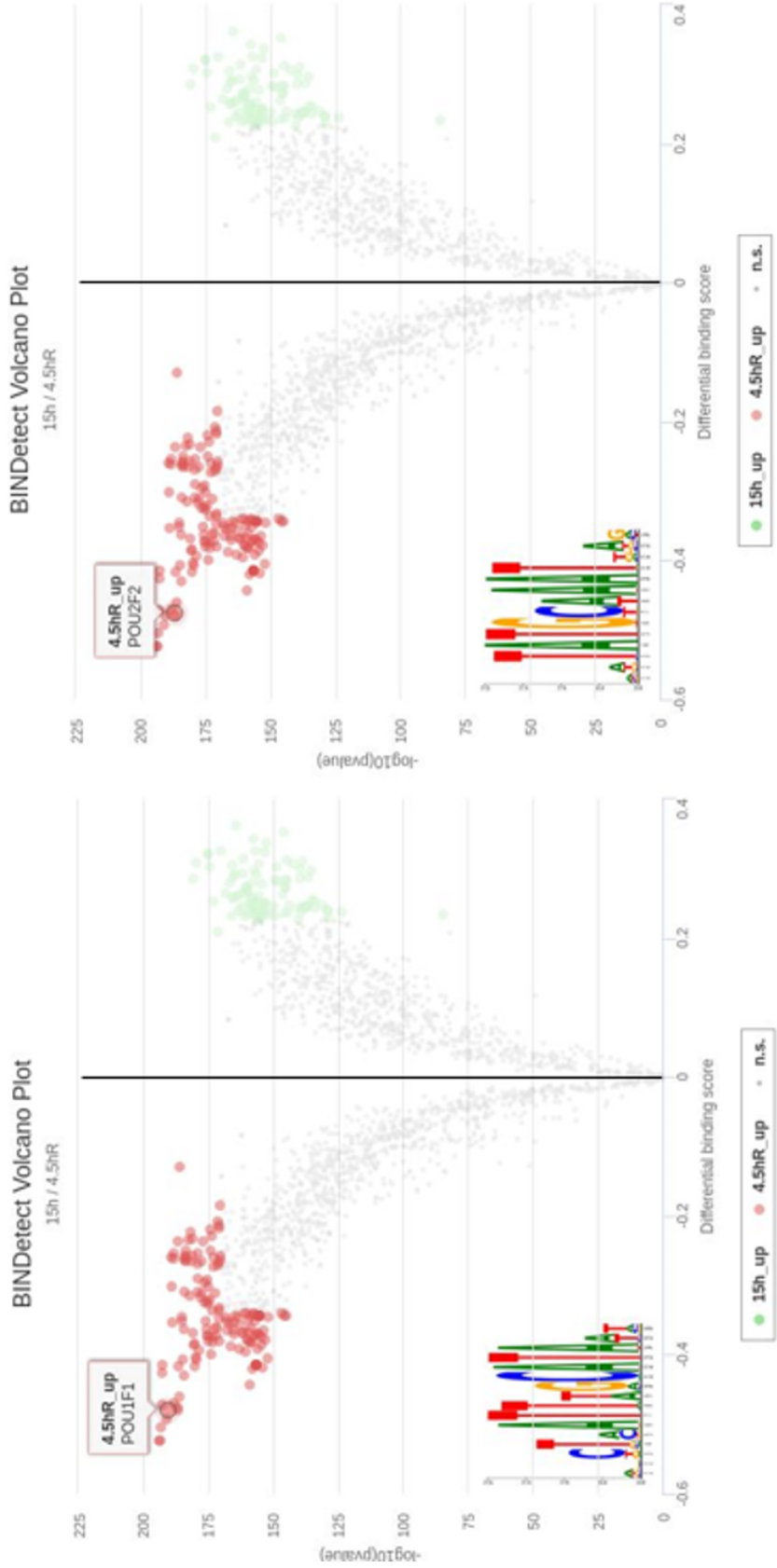
In-house figure (Pou5f1 motif enrichment)



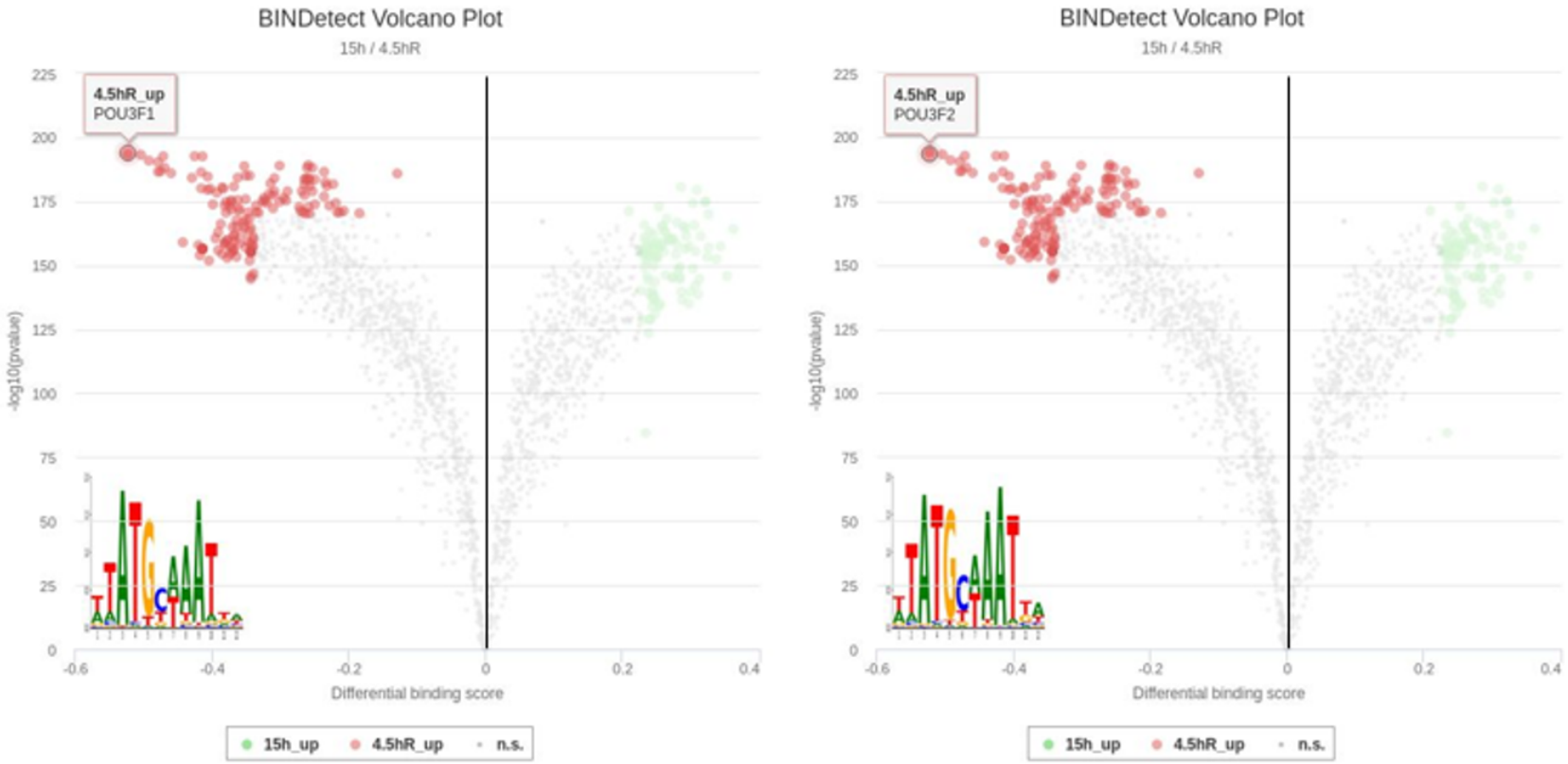
In-house figure (Pou5f1/Sox2 motif enrichment)



(A)


BINDetect Volcano Plot
15h / 4.5hR
4.5hR_up
POU2F2
4.5hR_up
POU1F1
-log10(pvalue)
Differential binding score
15h_up
4.5hR_up
n.s.

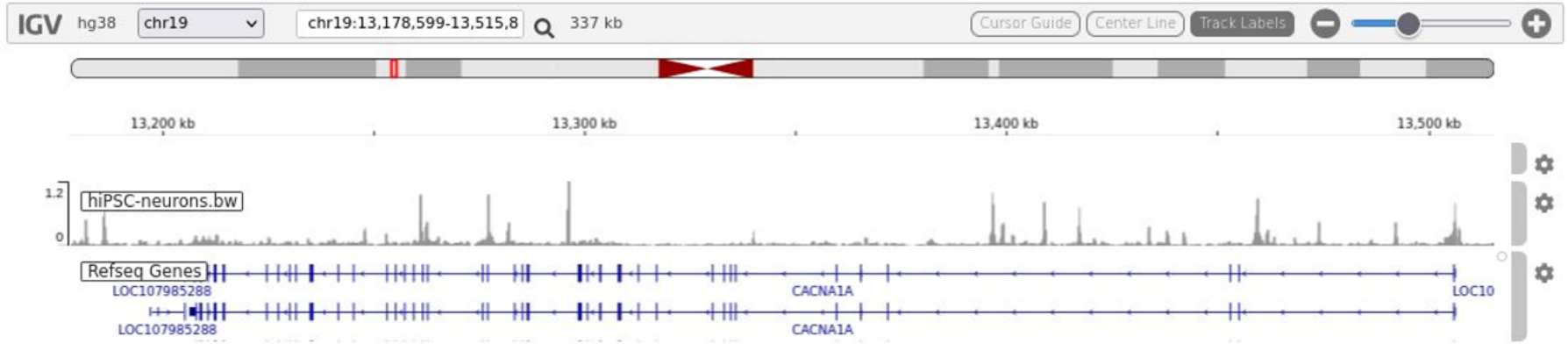
(B)



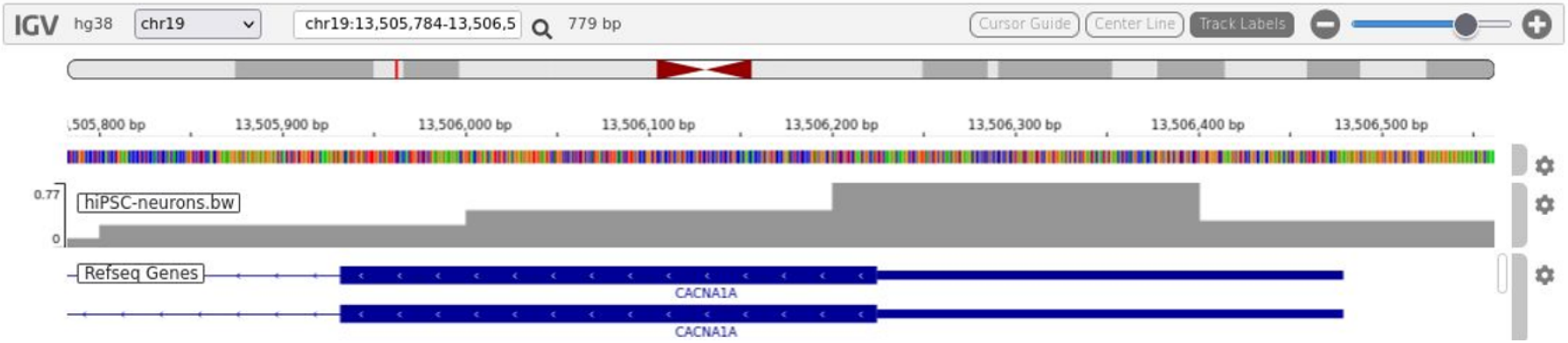
(C)

Supplementary Figure 3.1. Assessing sensitivity and specificity of TOBIAS in the context of TF upregulation.

Re-analyses of data from mouse embryonic stem cells genetically engineered to repress Oct4 expression upon doxycycline treatment. See **Figure 4.6** for an explanation of how to interpret BINDetect plots. (A) Comparison of data from 15-hour (15h) treated (untreated) and 4.5-hour doxycycline washout (4.5h recovery, or 4.5hR) ZHBTc4 mouse embryonic stem cells (Xiong et al., 2022). Top: BINDetect figure from Xiong et al., 2022; Bottom left and right: in-house figures generated using Xiong et al. data and in-house computational pipeline. Bottom left: Differential TF motif enrichment between 15h and 4.5hR samples. Oct4/Pou5f1 motif enrichment in 4.5hR samples is indicated. Bottom right: Differential TF motif enrichment between 15h and 4.5hR samples. Pou51/Sox2 motif enrichment in 4.5hR samples is indicated. For (B) and (C) shown is differential TF motif enrichment for non-Oct4 POU factors. For (B) and (C), differential motif enrichment between 15h and 4.5hR samples is shown. (B) Pou1f1 (left) and Pou2f2 (right) motif enrichment are shown. (C) Pou3f1 (left) and Pou3f2 (right) motif enrichment are shown.



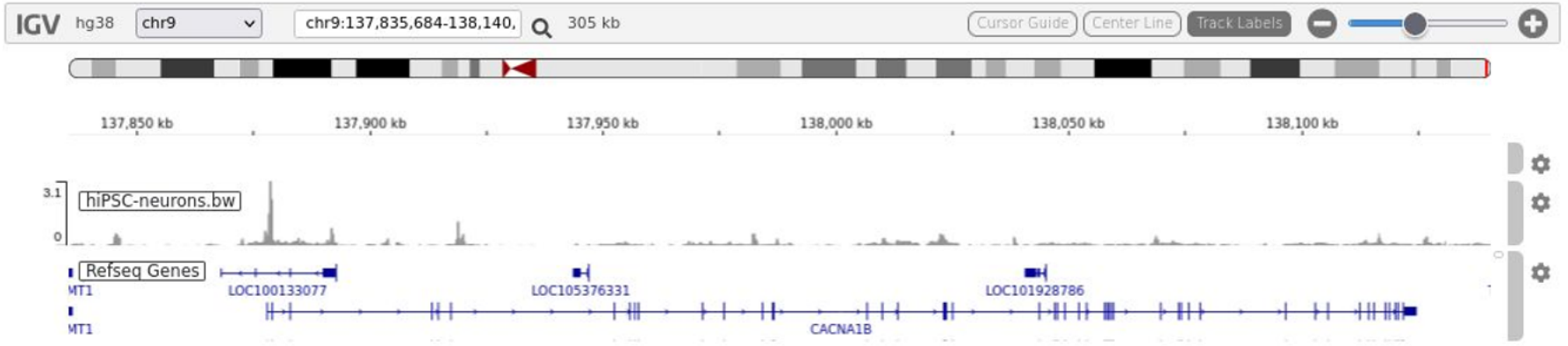
(A)



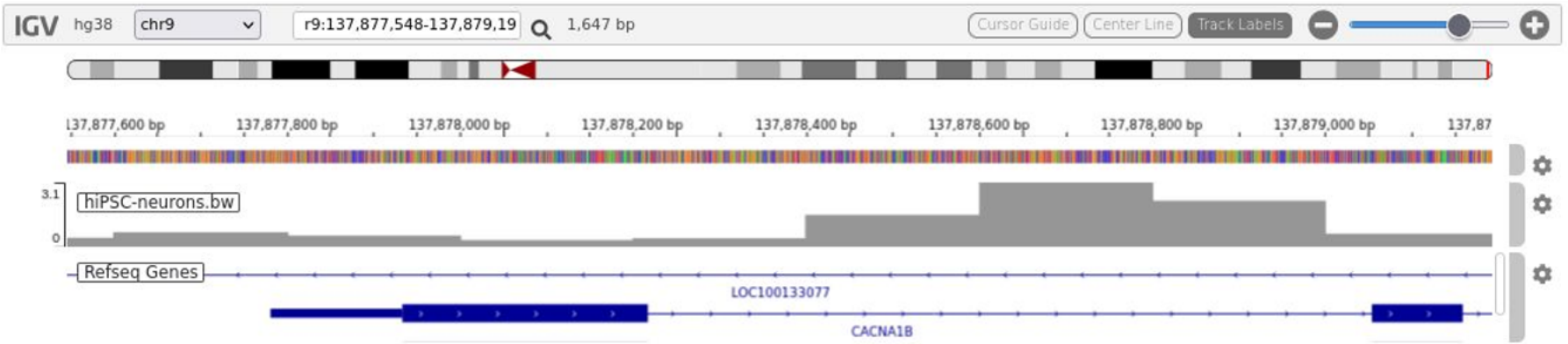
(B)

Supplementary Figure 3.2. CACNA1A accessible chromatin region

(A) IGV view of ATAC-seq signal at entire CACNA1A locus. (B) IGV view of ATAC-seq signal at CACNA1A TSS (chromatin peak 176257).



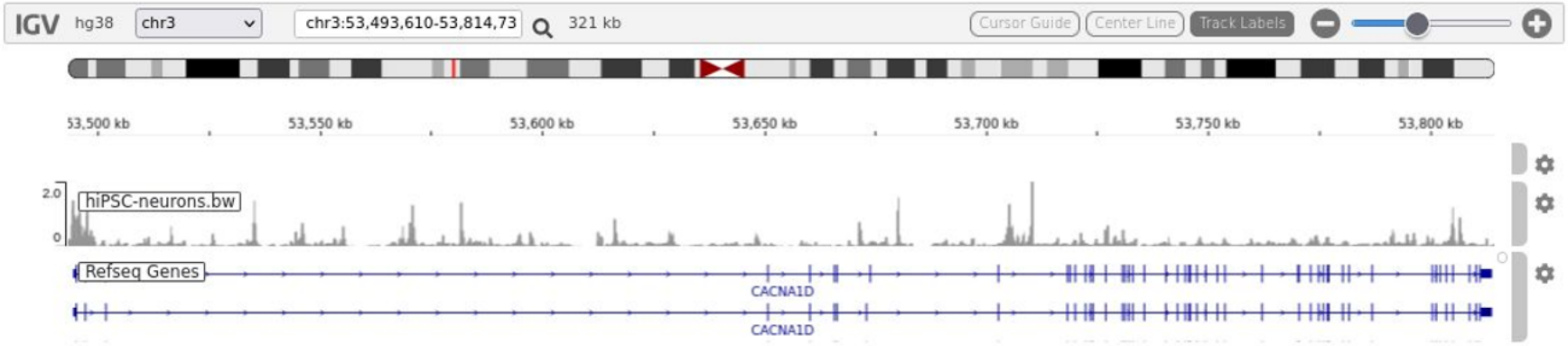
(A)



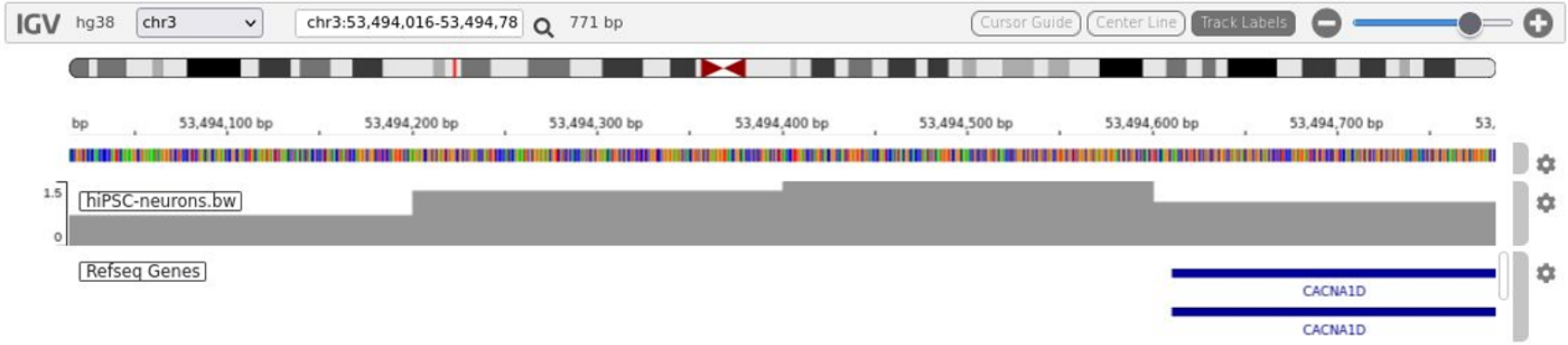
(B)

Supplementary Figure 3.3. CACNA1B accessible chromatin region.

(A) IGV view of ATAC-seq signal at entire CACNA1B locus. (B) IGV view of ATAC-seq signal at CACNA1B TSS (chromatin peak 438022).



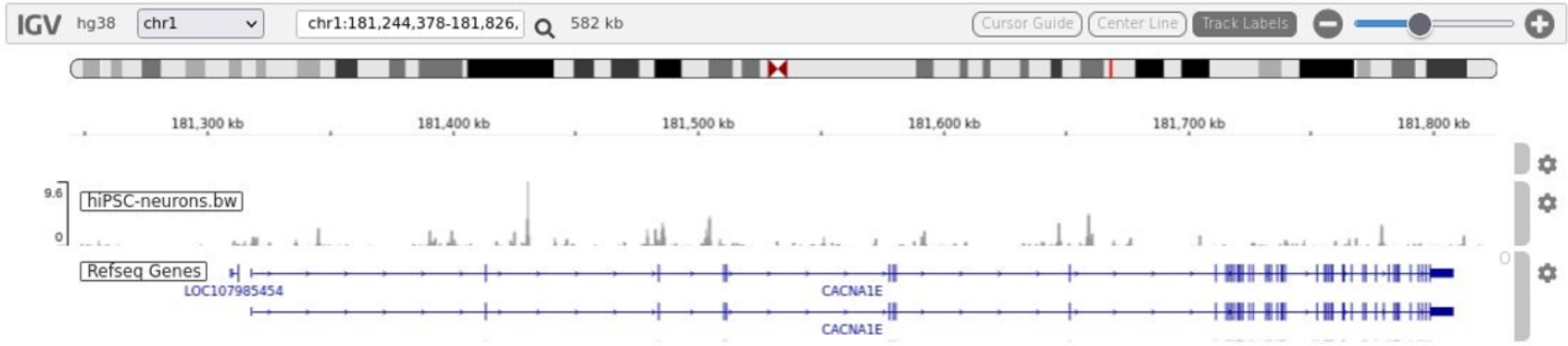
(A)



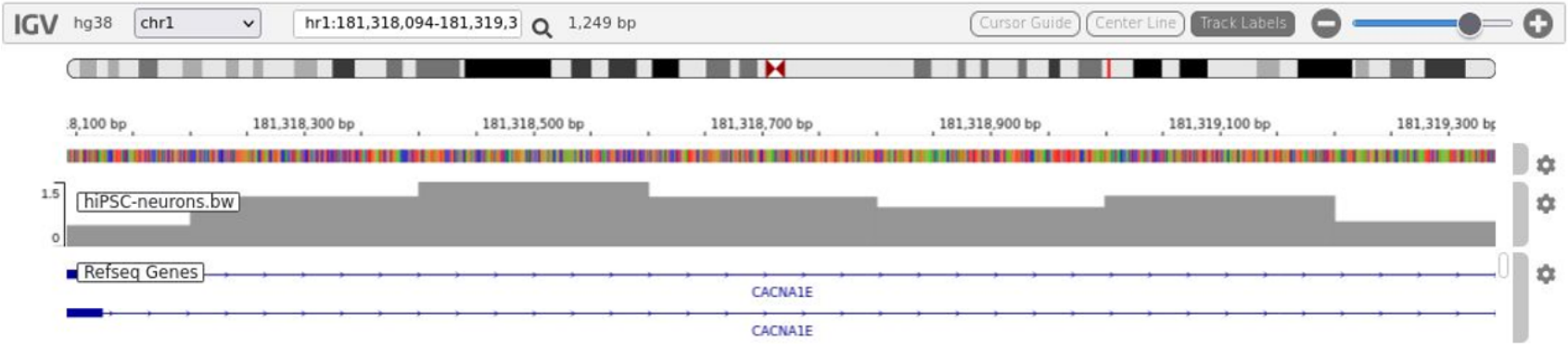
(B)

Supplementary Figure 3.4. CACNA1D accessible chromatin region.

(A) IGV view of ATAC-seq signal at entire CACNA1D locus. (B) IGV view of ATAC-seq signal at CACNA1D TSS (chromatin peak 247816).



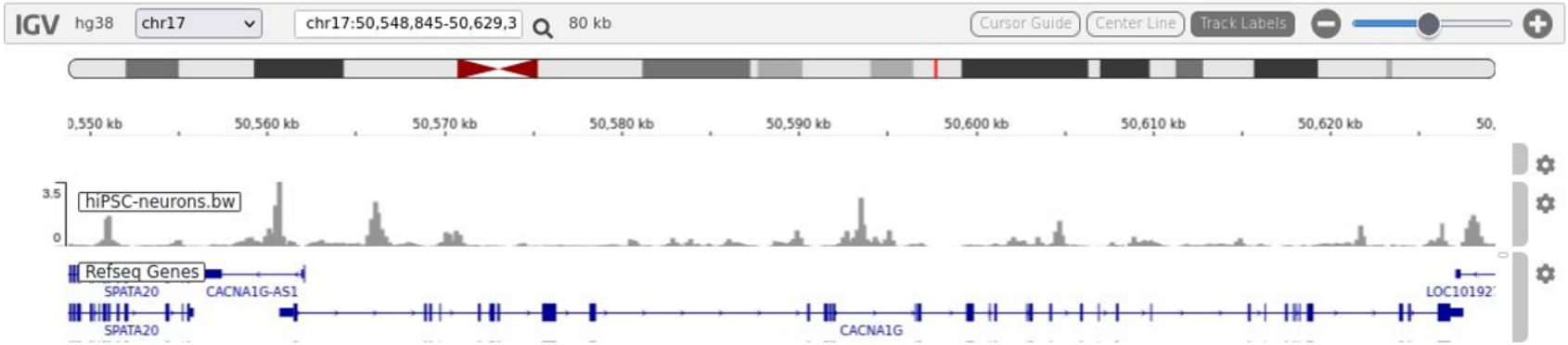
(A)



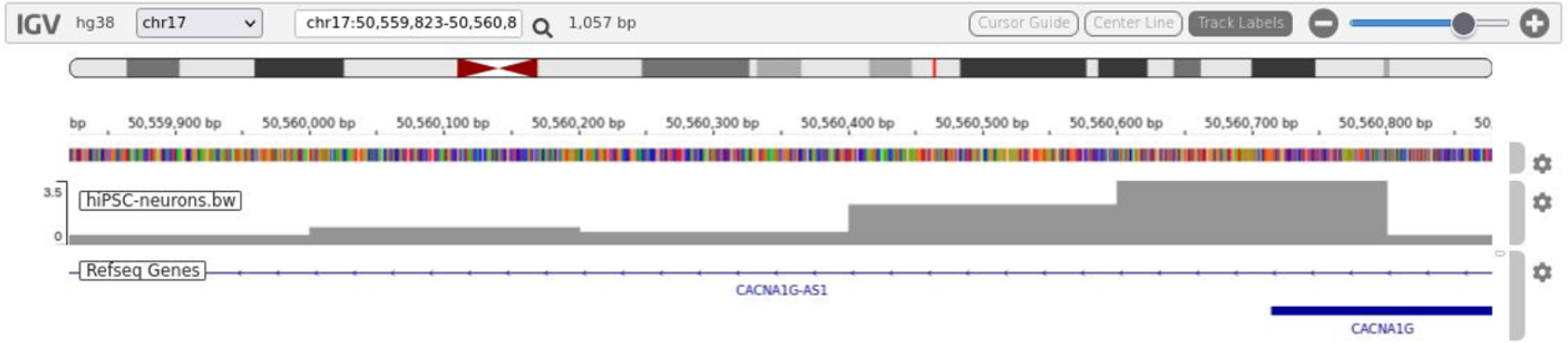
(B)

Supplementary Figure 3.5. CACNA1E accessible chromatin region.

(A) IGV view of ATAC-seq signal at entire CACNA1E locus. (B) IGV view of ATAC-seq signal at CACNA1E TSS (chromatin peak 25372).



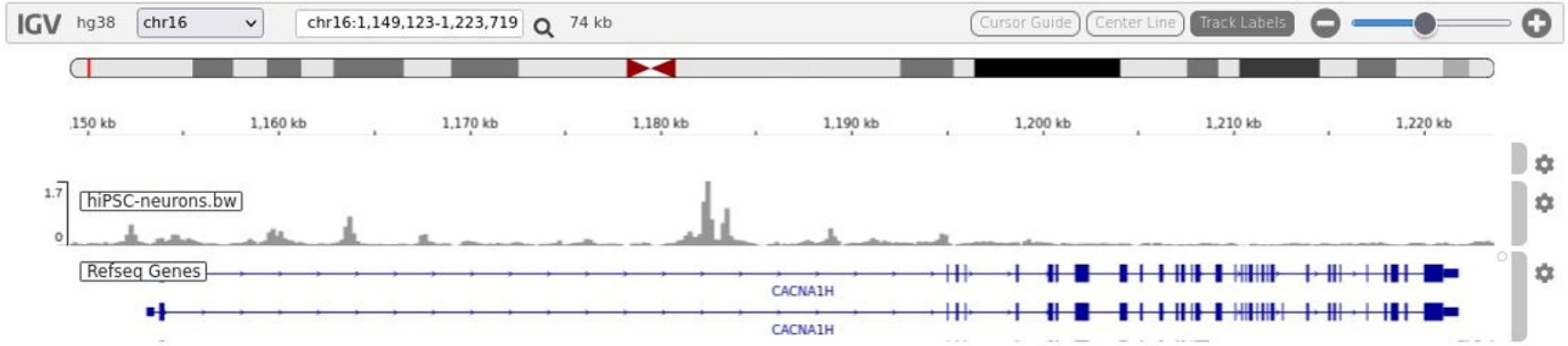
(A)



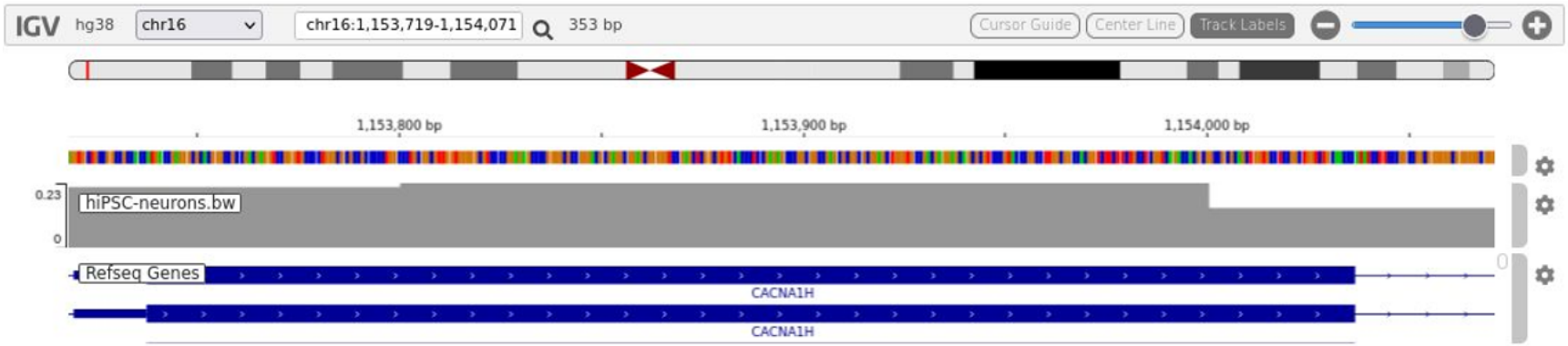
(B)

Supplementary Figure 3.6. CACNA1G accessible chromatin region.

(A) IGV view of ATAC-seq signal at entire CACNA1G locus. (B) IGV view of ATAC-seq signal at CACNA1G TSS (chromatin peak 157060).



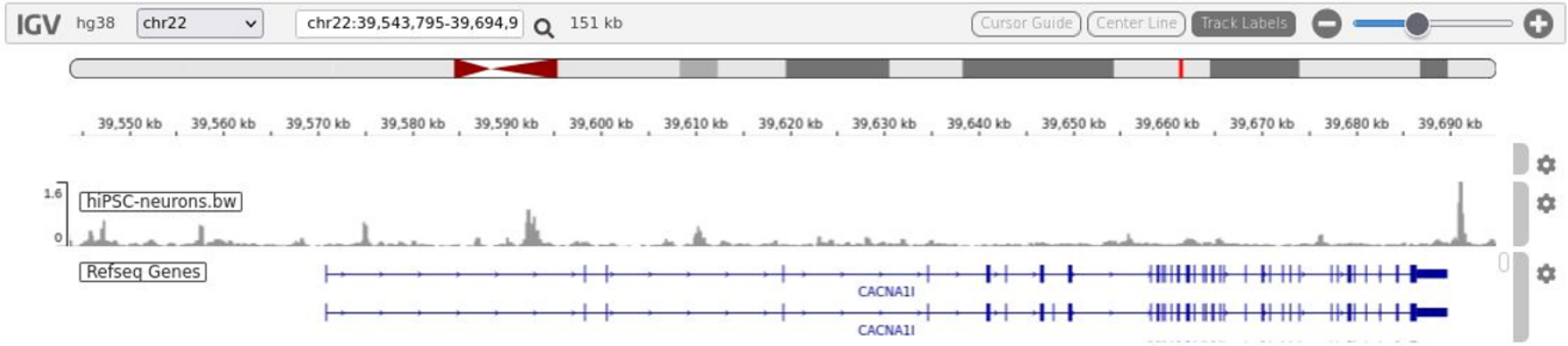
(A)



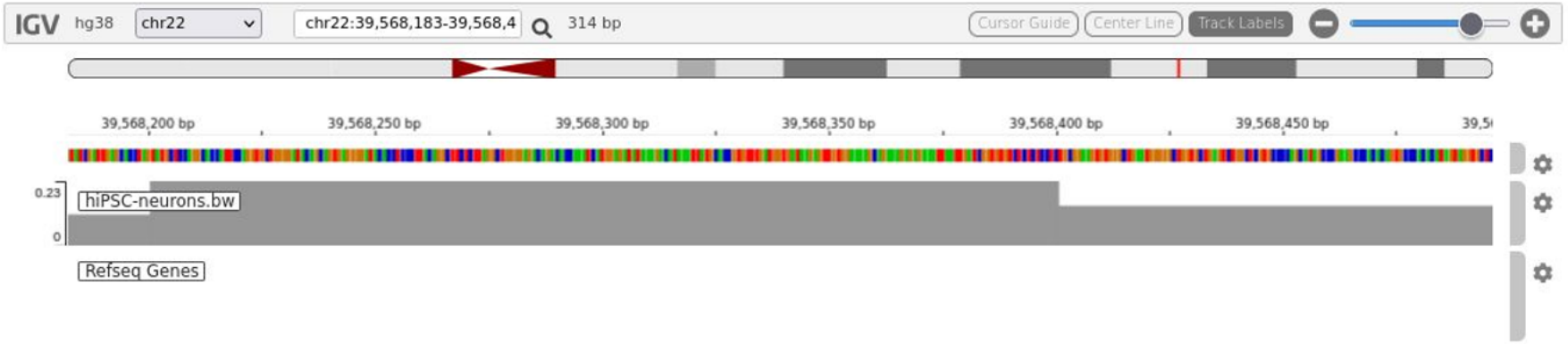
(B)

Supplementary Figure 3.7. CACNA1H accessible chromatin region.

(A) IGV view of ATAC-seq signal at entire CACNA1H locus. (B) IGV view of ATAC-seq signal at CACNA1H TSS (chromatin peak 141589). A)



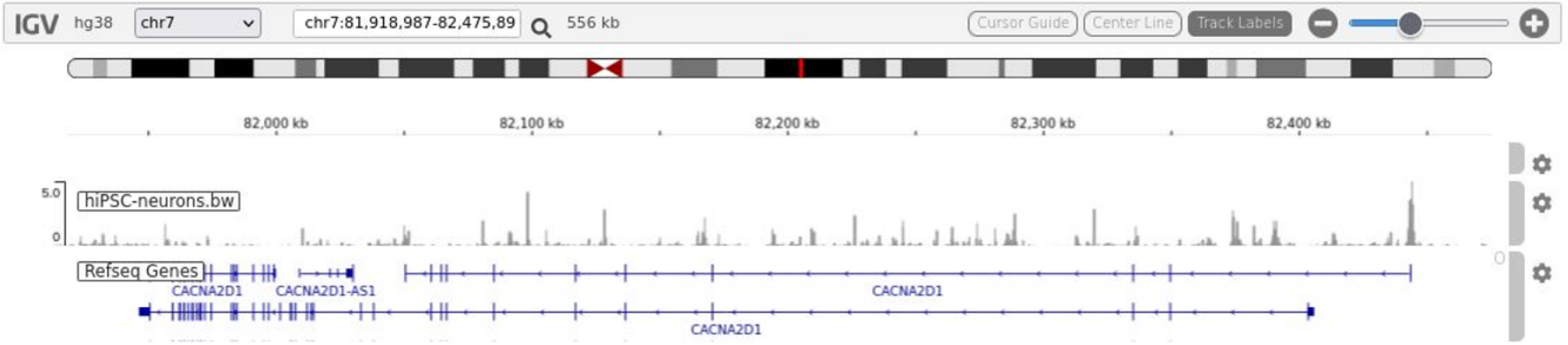
(A)



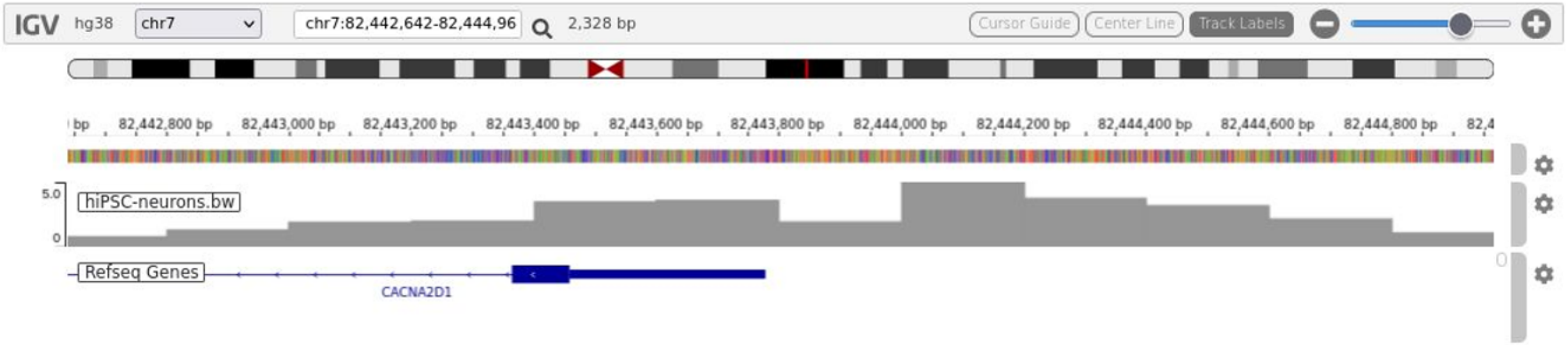
(B)

Supplementary Figure 3.8. CACNA1I accessible chromatin region.

(A) IGV view of ATAC-seq signal at entire CACNA1I locus. (B) IGV view of ATAC-seq signal at CACNA1I TSS (chromatin peak 237721).



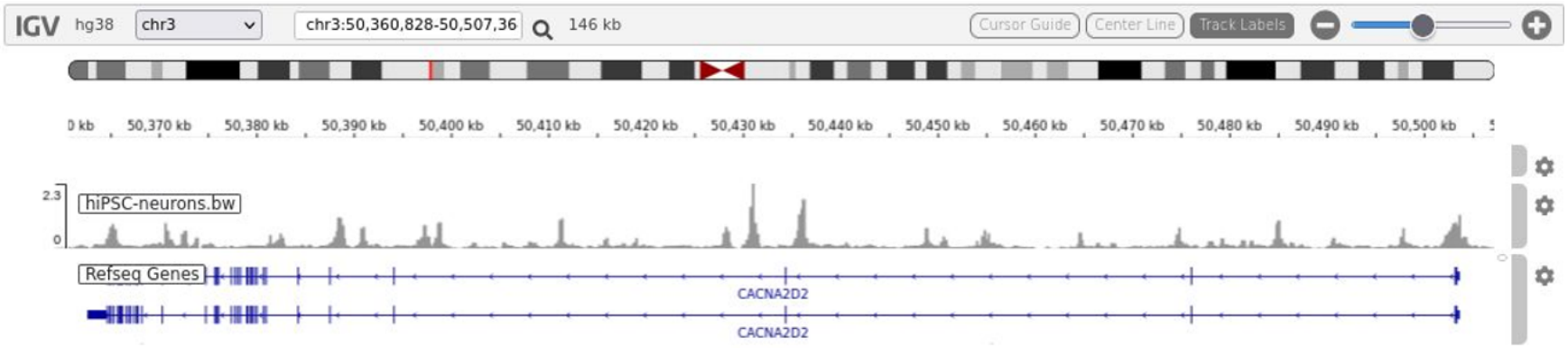
(A)



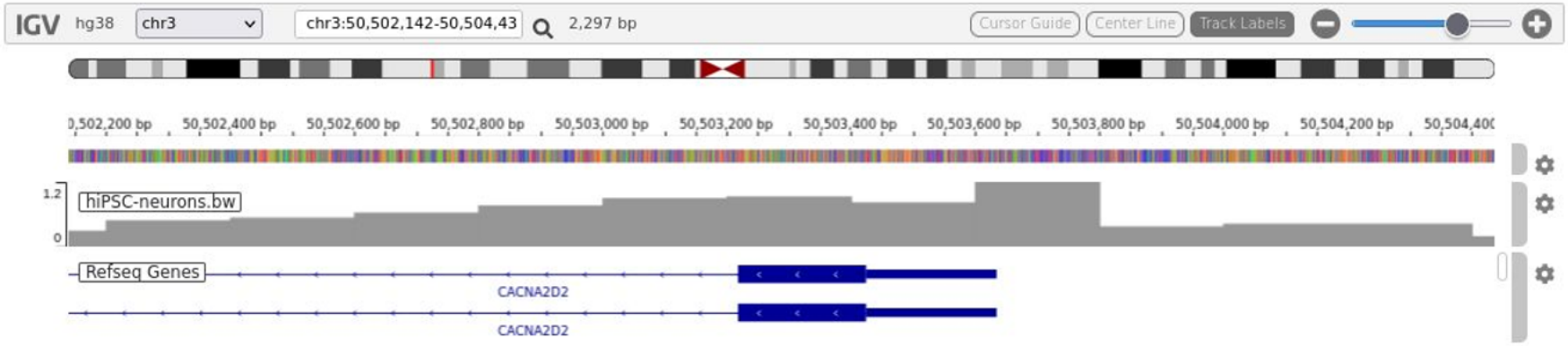
(B)

Supplementary Figure 3.9. CACNA2D1 accessible chromatin region.

(A) IGV view of ATAC-seq signal at CACNA2D1 locus. (B) IGV view of ATAC-seq signal at CACNA2D1 TSS (chromatin peak 379901).



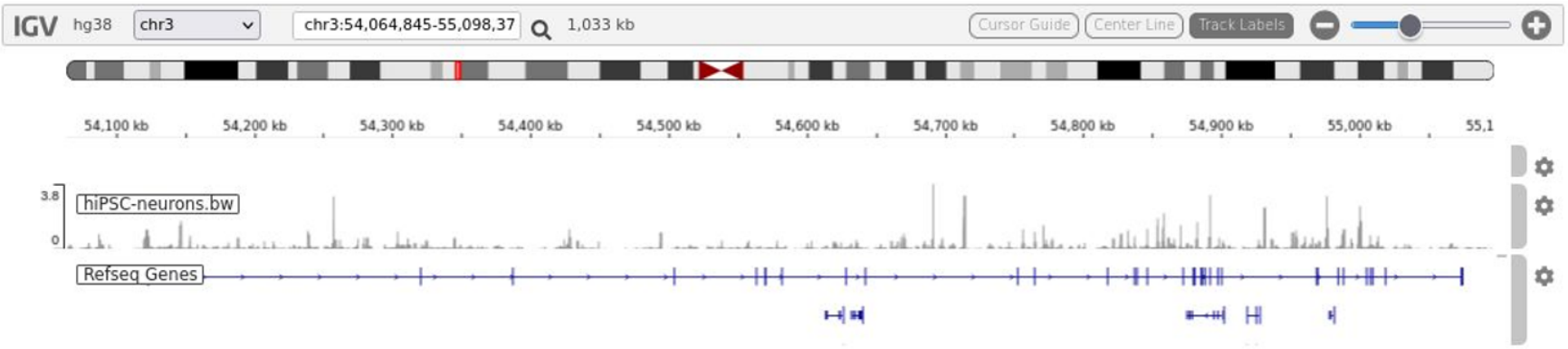
(A)



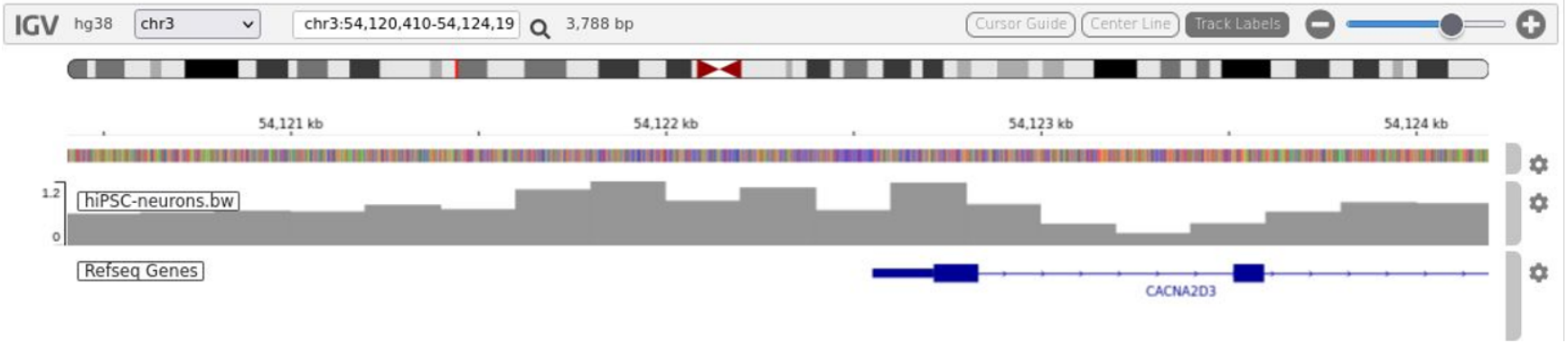
(B)

Supplementary Figure 3.10. CACNA2D2 accessible chromatin region.

(A) IGV view of ATAC-seq signal at entire CACNA2D2 locus. (B) IGV view of ATAC-seq signal at CACNA1C TSS (chromatin peak 247523).



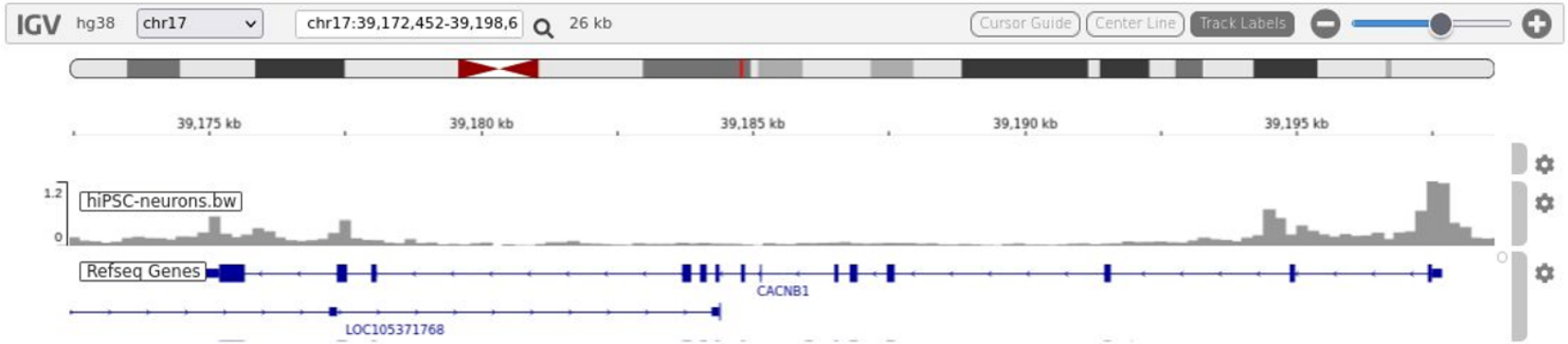
(A)



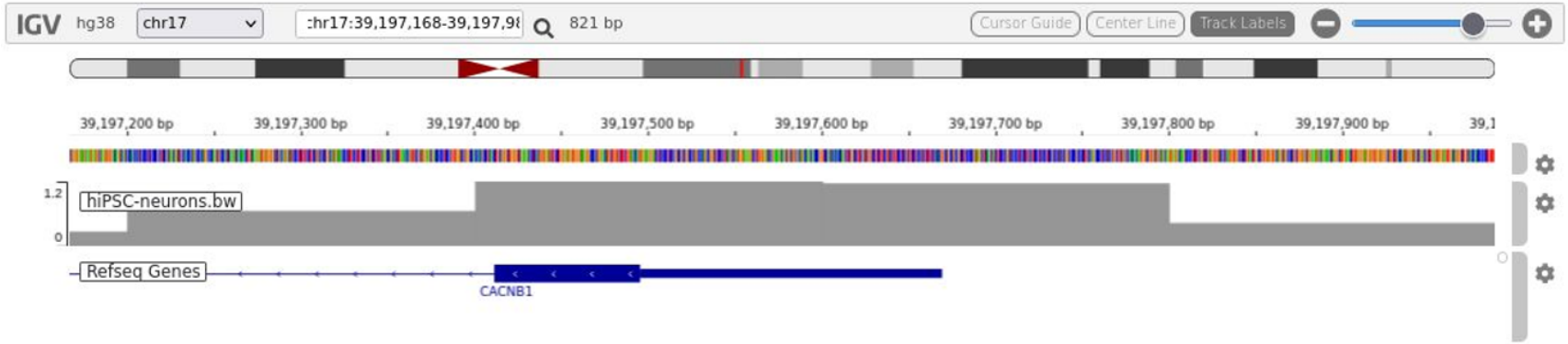
(B)

Supplementary Figure 3.11. CACNA2D3 accessible chromatin region.

(A) IGV view of ATAC-seq signal at entire CACNA2D3 locus. (B) IGV view of ATAC-seq signal at CACNA2D3 TSS (chromatin peak 247921).



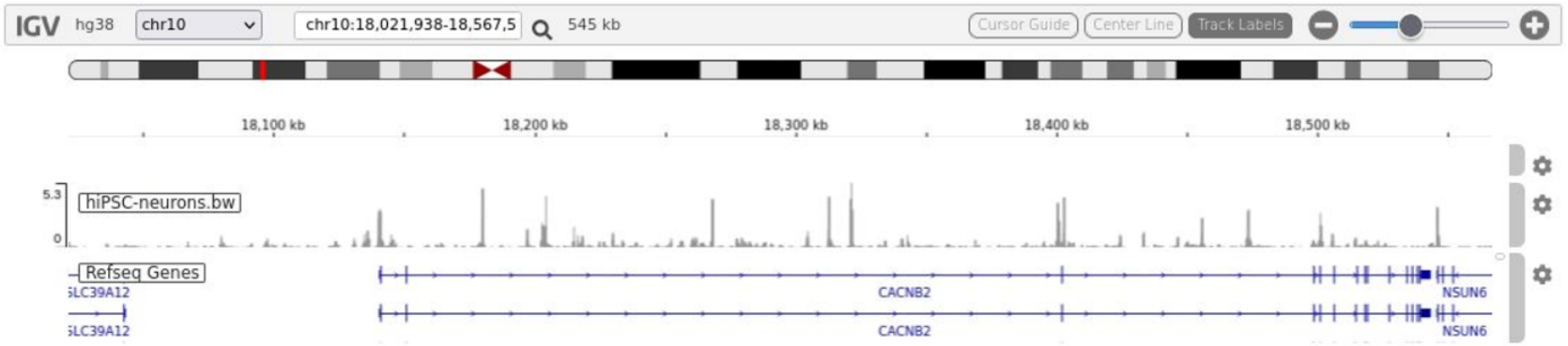
(A)



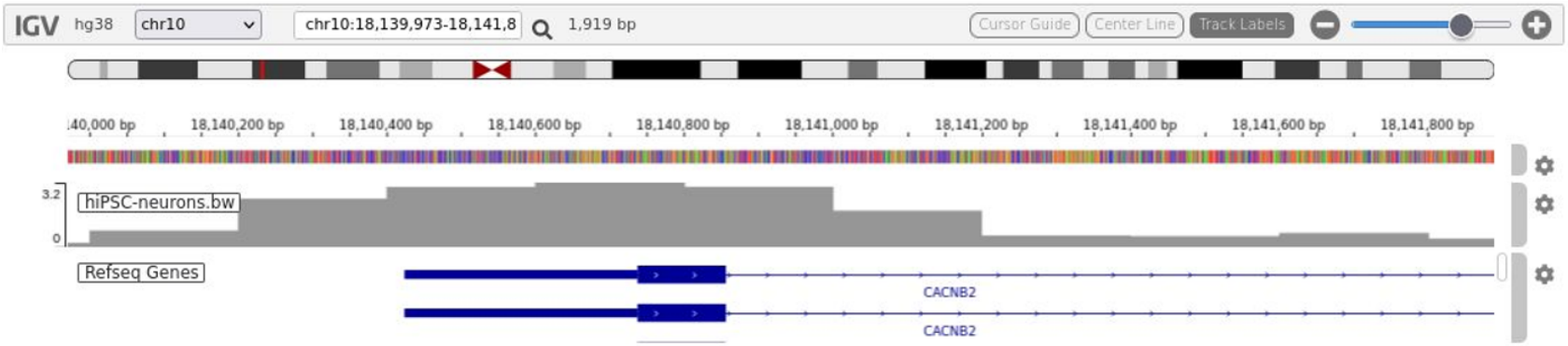
(B)

Supplementary Figure 3.12. CACNB1 accessible chromatin region.

(A) IGV view of ATAC-seq signal at entire CACNB1 locus. (B) IGV view of ATAC-seq signal at CACNB1 TSS (chromatin peak 155809)



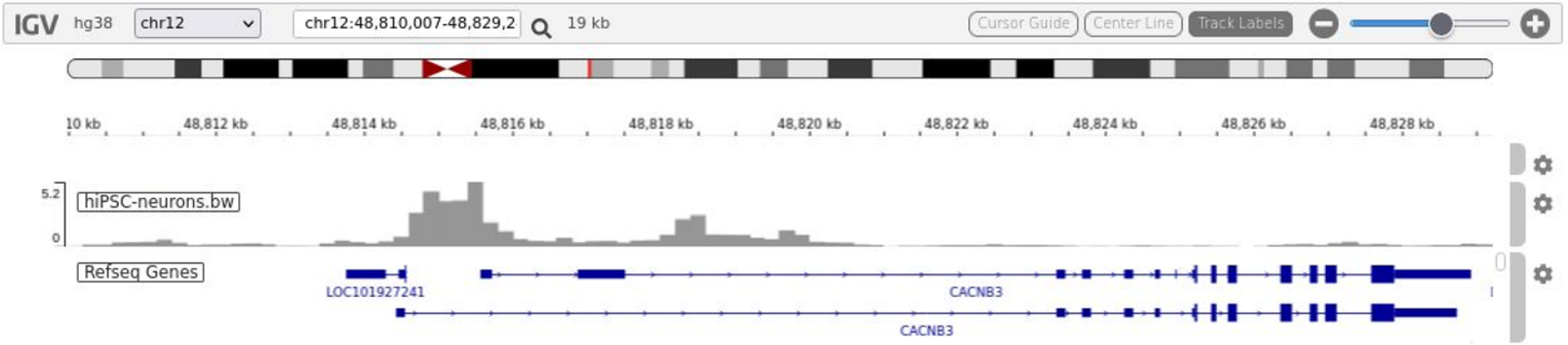
(A)



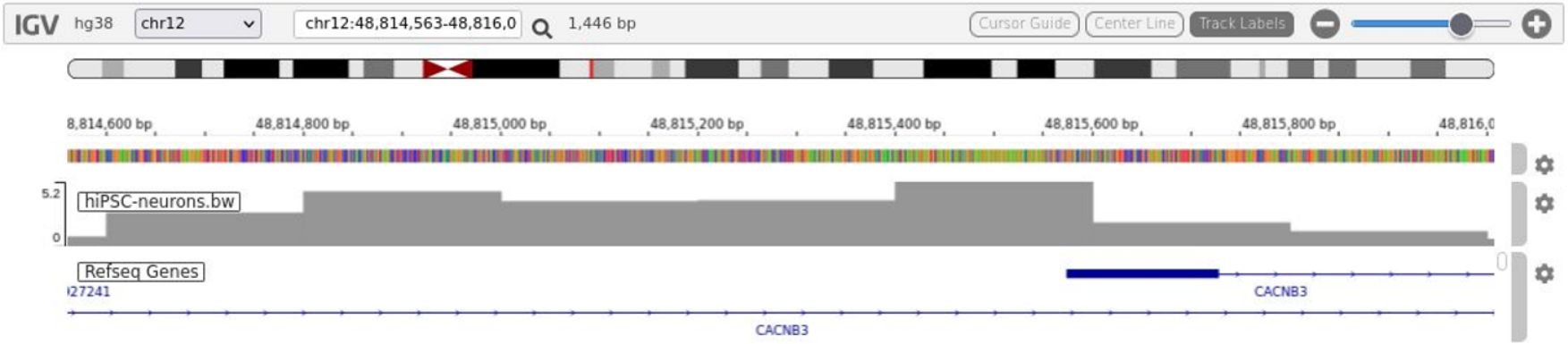
(B)

Supplementary Figure 3.13. CACNB2 accessible chromatin region.

(A) IGV view of ATAC-seq signal at entire CACNB2 locus. (B) IGV view of ATAC-seq signal at CACNB2 TSS (chromatin peak 39988).



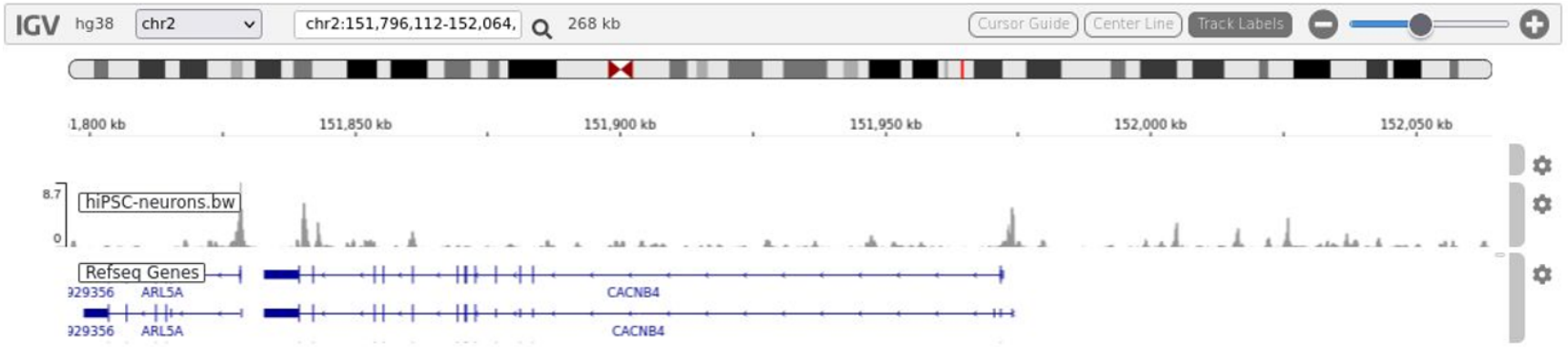
(A)



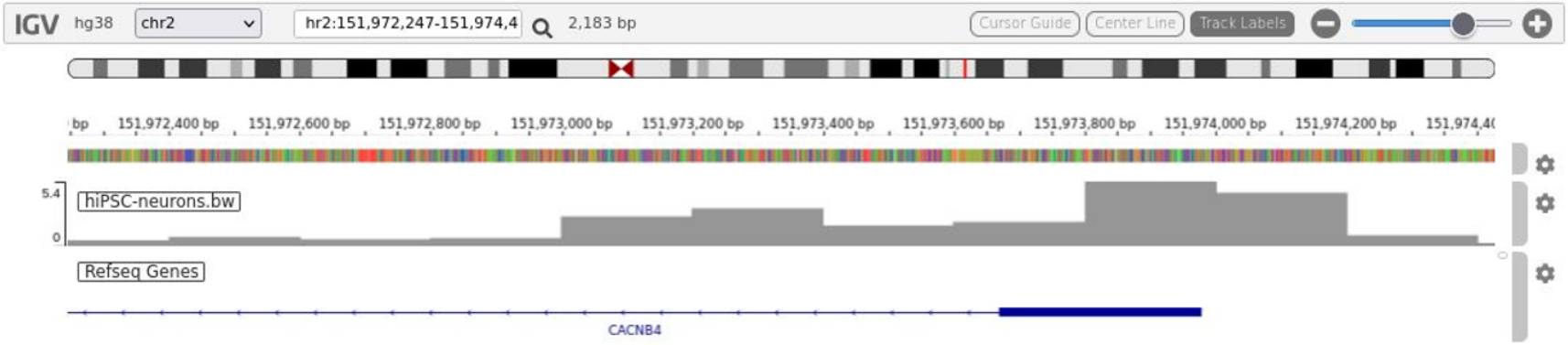
(B)

Supplementary Figure 3.14. CACNB3 accessible chromatin region.

(A) IGV view of ATAC-seq signal at entire CACNB3 locus. (B) IGV view of ATAC-seq signal at CACNB3 TSS.



(A)



(B)

Supplementary Figure 3.15. CACNB4 accessible chromatin region.

(A) IGV view of ATAC-seq signal at entire CACNB4 locus. (B) IGV view of ATAC-seq signal at CACNB4 TSS (chromatin peak 204971).

References

Ablinger C, Geisler SM, Stanika RI, et al. Neuronal alpha-2-delta proteins and brain disorders. Pflugers Arch 472: 845 – 863 (2020).

Aguilera-Castrejon A, et al. Ex utero embryogenesis from pre-gastrulation to late organogenesis. Nature 539: 119 – 124 (2021).

Alzu'bi A, Clowry GJ. Expression of ventral telecephalon transcription factors ASCL1 and DLX in the early fetal human cerebral cortex. J Anat 235: 555 – 568 (2019).

Amiel J, et al. Mutations in TCF4, encoding a Class I basic helix-loop-helix transcription factor, are responsible for Pitt-Hopkins syndrome, a severe epileptic encephalopathy associated with autonomic dysfunction. Am J Hum Genet 80: 988 – 993 (2007).

Anantharaman V, Aravind L. Cache – a signaling domain common to animal Ca(2+)-channel subunits and a class of prokaryotic chemotaxis receptors. Trends Biochem Sci 25: 535 – 537 (2000).

Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Research 22: 2008-17 (2012).

Andrews, S. "FASTQC: A Quality Control Tool for High Throughput Sequence Data". Babraham Bioinformatics – FASTQC a Quality Control Tool for High Throughput Sequence Data, https://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Arnold CD, et al. A high-throughput method to identify trans-activation domains within transcription factor sequences. EMBO J 37: e98896 (2018).

Arrikath J, Campbell KP. Auxiliary subunits: essential components of the voltage-gated calcium complex. Current Opinion in Neurobiology 13: 298 – 307 (2003).

Arteaga-Bracho EE, et al. Postnatal and adult consequences of loss of huntingtin during development: implications for Huntington's disease. Neurobiol Dis 96: 144 – 155 (2016).

Atkinson LZ, Colbourne L, Smith A, et al. The Oxford study of calcium channel antagonism, cognition, mood instability and sleep (OxCaMS): study protocol for a randomized controlled, experimental medicine study. Trials 20: 120 (2019).

Auweter SD et al. Sequence-specific binding of single-stranded RNA: is there a code for recognition? Nucleic Acids Research 34: 4943 – 4959 (2006).

Avsec Z, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nature Genetics 53: 354 – 366 (2021).

Bading H. Nuclear calcium signaling in the regulation of brain function. Nat Rev Neurosci 14: 593 – 608 (2013).

Baez-Nietro D, et al. Analyzing an allelic series of rare missense variants of CACNA1I in a Swedish schizophrenia cohort. Brain 145: 1839 – 1855 (2022).

Baig SM, et al. Loss of Cav1.3 (CACNA1D) function in a human channelopathy with bradycardia and congenital deafness. Nature Neuroscience 14: 77 – 84 (2011).

Bain JM, et al. Variants in HNRNPH2 on the X chromosome are associated with a neurodevelopmental disorder in females. Am J Hum Genet 99: 728 – 734 (2016).

Baker KD, et al. The role of intracellular calcium stores in synaptic plasticity and memory consolidation. Neurosci Biobehav Rev 37: 1211- 39 (2013).

Baltz AG, et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Molecular Cell 46: 674 – 690 (2012).

Barclay J, et al. Ducky mouse phenotype of epilepsy and ataxia is associated with mutations in the Cacna2d2 gene and decreased calcium channel current in cerebellar Purkinje cells. J Neurosci 21: 6095 – 6104 (2001).

Barclay J, Rees M. Genomic organization of the mouse and human alpha2-delta2 voltage-dependent calcium channel subunit genes. Mamm Genome 11: 1142 – 1144 (2000).

Barnat M, et al. Huntington's disease alters human neurodevelopment. Science 369: 787 – 793 (2020).

Barth LG, Barth LJ. Sequential induction of the presumptive epidermis of the Rana pipiens gastrula. Biol Bull 127: 413 – 427 (1964).

Bauer R, et al. Update on the molecular genetics of Timothy Syndrome. Front Pediatr 9: 668546 (2021).

Beatrice L, et al. Differential upregulation in DRG neurons of an alpha2-delta 1 splice variant with a lower affinity for gabapentin after peripheral sensory nerve industry. Pain 155: 522 – 533 (2014).

Beckmann BM, et al. The RNA-binding proteomes from yeast to man harbour conserved RBPs. Nature Communications 6 (2015).

Benjamini Y, Hochberg Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 57: 289 – 300 (1995).

Bennett S. Solexa ltd. Pharmacogenomics 5: 433 – 438 (2004).

Bentsen M, et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nature Communications 11: 4267 (2020).

Berto S, et al. ELAVL2-regulated transcriptional and splicing networks in human neurons link neurodevelopment and autism. Hum Mol Genetics 25: 2451 -2464 (2016).

Bertrand N, et al. Proneural genes and the specification of neural cell types. Nature Review Neurosciences 3: 517 – 530 (2002).

Betschinger J, et al. Exit from pluripotency is gated by intracellular redistribution of the bHLH transcription factor Tfe3. Cell 153: 335 – 347 (2013).

Bhaduri A, et al. Cell stress in cortical organoids impairs molecular subtype specification. Nature 578: 142 – 148 (2020).

Bhat S, et al. CACNA1C (Cav1.2) in the pathophysiology of psychiatric disease. Prog Neurobiol 99 1-14 (2012).

Bigos KL, et al. Genetic variation in CACNA1C affects brain circuitries related to mental illness. Arch Gen Psychiatry 67: 939 – 45 (2010).

Bikbaev A, et al. Auxiliary alpha2-delta1 and alpha2-delta3 subunits of calcium channels drive excitatory and inhibitory neuronal network development. J Neurosci 40: 4824 – 4841 (2020).

Birey F, et al. Assembly of functionally integrate human forebrain spheroids. Nature 545: 54-59 (2017).

Birey F, et al. Dissecting the molecular basis of human interneuron migration in forebrain assembloids from Timothy Syndrome. Cell Stem Cell 29: 248 – 264 (2022).

Birnbaum R, et al. Investigation of the prenatal expression patterns of 108 schizophrenia-associated genetic loci. Biol Psychiatry 77: e43 – e51 (2015).

Birnbaum R, et al. Prenatal expression patterns of genes associated with neuropsychiatric disorders. Am J Psychiatry 171: 758 – 767 (2014).

Birnbaum R, Weinberger DR. Genetic insights into the neurodevelopmental origins of schizophrenia. Nature Reviews Neuroscience 18: 727 – 740 (2017).

Bisgaard AM, et al. Additional chromosomal abnormalities in patients with a previously detected abnormal karyotype, mental retardation and dysmorphic features. Am J Med Genet 140: 2180 – 7 (2006).

Bitzenhofer SH, et al. A transient developmental increase in prefrontal activity alters network maturation and causes cognitive dysfunction in adult mice. Neuron 109: 1350 – 1364 (2021).

Blackmore MG, et al. Kruppel-like Factor 7 engineered for transcriptional activation promotes axon regeneration in the adult corticospinal tract. Proc Natl Acad Sci USA 109: 7517 – 7522 (2012).

Blanchette M, et al. Aligning multiple genomic sequences with the threaded blockset al.igner. Genome Research 14: 708 – 15 (2004).

Bocchi R, et al. Perturbed Wnt signaling leads to neuronal migration delay, altered interhemispheric connections and impaired social behavior. Nature Communications 8: 1158 (2017).

Boija A, et al. Transcription factors activate genes through the phase-separation capacity of their activation domains. Cell 175: 1842 – 1855 (2018).

Boller S, et al. Pioneering activity of the C-terminal domain of EBF1 shapes the chromatin landscape for B cell programming. Immunity 44: 527 – 541 (2016).

Boyer LA, et al. Core transcriptional regulatory circuitry in human embryonic stem cells.

Cell 122: 947 – 956 (2005).

Boyer NP, Monkiewicz C, Menon S, et al. Mammalian TRIM67 functions in brain development and behavior. eNeuro 5: ENEURO.0186-18.2018 (2018).

Brandau DT, et al. Autistic and dysmorphic features associated with a submicroscopic 2q33.3-q34 interstitial deletion detected by array comparative genomic hybridization. Am J Med Genet A 146A(4): 521 – 4 (2008).

Brandt A, et al. few Cav1.3 channels regulate the exocytosis of a synaptic vesicle at the hair cell ribbon synapse. J Neurosci 25: 11577 – 11585 (2005).

Braz BY, et al. Treating early postnatal circuit defect delays Huntington's disease onset and pathology in mice. Science 377: 377 (2022).

Brennand K, et al. Phenotypic differences in hiPSC NPCs derived from patients with schizophrenia. Molecular Psychiatry 20: 361 – 368 (2015).

Brennand KJ, et al. Modeling schizophrenia using human induced pluripotent stem cells. Nature 473: 221 – 225 (2011).

Brockschmidt A, et al. Severe mental retardation with breathing abnormalities (Pitt-Hopkins syndrome) is caused by haploinsufficiency of the neuronal bHLH transcription factor TCF4. Hum Mol Genet 16: 1488 – 1494 (2007).

Brown AS, Derkits EJ. Prenatal infection and schizophrenia: a review of epidemiologic and translational studies. Am J Psychiatry 167: 261 – 280 (2010).

Brzozka MM, et al. Cognitive and sensorimotor gating impairments in transgenic mice overexpressing the schizophrenia susceptibility gene Tcf4 in the brain. Biol Psychiatry 68: 33 – 40 (2010).

Brzozka MM, Rossner MJ. Deficits in trace fear memory in a mouse model of the schizophrenia risk gene TCF4. Behav Brain Res 237: 348 – 356 (2013).

Buckley NJ, et al. Is REST a regulator of pluripotency? Nature 457: E5-6 (2009).

Buckley NJ, et al. The role of REST in transcriptional and epigenetic dysregulation in Huntington's disease. Neurobiol Dis 39: 28 – 39 (2010).

Buenrostro JD, et al. ATAC-seq: a method for assaying chromatin accessibility genome-

wide. Curr Protoc Mol Biol. 109: 21.29.1-21.29.9 (2015).

Buenrostro JD, et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins, and nucleosome position. Nature Methods 10: 1213-8 (2013).

Buraei Z, Yang J. Structure and function of the beta subunit of voltage-gated calcium channels. Biochimica et Biophysica Acta – Biomembranes. 128: 1530 – 1540 (2013).

Buraei Z, Yang J. The beta subunit of voltage-gated Ca2+ channels. Physiol Rev 90: 1461 – 1506 (2010).

Calderon-Rivera A, Andrade A, Hernandez-Hernandez O, et al. Identification of a disulfide bridge essential for structure and function of the voltage-gated calcium channel alpha(2)delta-1 auxiliary subunit. Cell Calcium 51: 22- 30 (2012).

Caldwell AB, et al. Dedifferentiation and neuronal repression define familial Alzheimer's disease. Sci Advances 6: eaba5933 (2020).

Cannon M, et al. Obstretic complications and schizophrenia: historical and meta-analytic review. Am J Psychiatry 159: 1080 – 1092 (2002).

Capra JA, et al. Many human accelerated regions are developmental enhancers. Philosophical Transactions of the Royal Society 368: 20130025 (2013).

Casamassima f, et al. L-type calcium channels and psychiatric disorders: a brief review. Am J Med Genet 153B: 1373 – 1390 (2010).

Cascino I, et al. Three functional soluble forms of the human apoptosis-inducing Fas molecule are produced by alternative splicing. J Immunol 154: 2706 – 2713 (1995).

Castello A, et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell 149: 1393 – 1406 (2012).

Castro DS, et al. A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets. Genes Dev 25: 930 – 45 (2011).

Catterall WA, Few AP. Calcium channel regulation and presynaptic plasticity. Neuron 59: 882 – 901 (2008).

Catterall WA. Voltage-gated calcium channels. Cold Spring Harbor Perspectives in Biol-

ogy 3: a003947 (2011).

Cesana M, et al. A CLK3-HMGA2 alternative splicing axis impacts human hematopoietic stem cell molecular identity throughout development. Cell Stem Cell 22: 575 – 588 (2018).

Chambers SM, Fasano CA, Papapetrou EP, et al. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nature Biotechnology 27: 275 – 80 (2009).

Chan CS, et al. 'Rejuvenation' protects neurons in mouse models of Parkinson's disease. Nature 447: 1081 – 1086 (2007).

Chanda S, et al. Generation of induced neuronal cells by the single reprogramming factor ASCL1. Stem Cell Reports 3: 282 – 96 (2014).

Chao H-T, et al. A syndromic neurodevelopmental order caused by de novo variants in EBF3. American Journal of Human Genetics 100: 128 – 137 (2017).

Chapman RM, et al. Convergent evidence that ZNF804A is a regulator of pre-messenger RNA processing and gene expression. Schizophr Bull 45: 1267 – 1278 (2019).

Charles CE, et al. FIMO: Scanning for occurrences of a given motif. Bioinformatics 27: 1017 – 1018 (2011).

Cheli VT et al. Conditional deletion of the L-type calcium channel Cav1.2 in oligodendrocyte progenitor cells affects postnatal myelination in mice. J Neurosci 36: 10853 – 10869 (2016).

Cheli VT et al. Enhanced oligodendrocyte maturation and myelination in a mouse model of Timothy syndrome. Glia 66: 2324 – 2339 (2018).

Chen X, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133: 1106 – 17 (2008).

Chen ZF, Rebelo S, White F, et al. The paired homeodomain protein DRG11 is required for the projection of cutaneous sensory afferent fibers to the dorsal spinal cord. Neuron 31: 59-73 (2001).

Cheng Z, et al. Cis-regulatory mutations with driver hallmarks in major cancers. iScience 4: 102144 (2021).

Chew J-L, et al. Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. Mol Cell Biol 25: 6031 – 46 (2005).

Chia NY, Chan YS, Feng B, et al. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature 468: 316-20 (2010).

Chiou Y-Y, et al. A sextuple knockout cell line system to study the differential roles of CRY, PER, and NR1D in the transcription-translation feedback loop of the circadian clock. Frontiers in Neuroscience 14: 1 – 13 (2020).

Chong JA, et al. REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. Cell 80: 949 – 57 (1995).

Cipriani A, Saunders K, Attenburrow MJ, et al. A systematic review of calcium channel antagonists in bipolar disorder and some considerations for their future development. Mol Psychiatry 21: 1324 – 32 (2016).

Clapcote SJ, et al. Behavioral phenotypes of Disc1 missense mutations in mice. Neuron 54: 387 – 402 (2007).

Clark MB, et al. Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain. Mol Psychiatry 25: 37 – 47 (2020).

Cook KB, et al. RBPDB: a database of RNA-binding specificities. Nucleic Acids Research 39: D301 – D308 (2011).

Coolidge CJ, et al. Functional anlayis of the polypyrimidine tract in pre-mRNA splicing. Nucleic Acids Research 25: 888 – 896 (1997).

Cornacchia D, Zhang C, Zimmer B, et al. Lipid deprivation induces a stable, naïve-to-primed intermediate state of pluripotency in human PSCs. Cell Stem Cell 25: 120 – 36 (2019).

Courtens W, et al. Interstitial deletion 2q33.3-q34 in a boy with a phenotype resembling the Seckel syndrome. Am J Med Genet 71: 479 – 85 (1997).

Danesi C, et al. Increased calcium influx through L-type calcium channels in human and mouse neural progenitors lacking fragile X mental retardation protein. Stem Cell Reports 11: 1449 – 1461 (2018).

De Donato M, et al. KLF7: a new candidat biomarker and therapeutic target for high-

grade serous ovarian cancer. Journal of Experimental & Clinical Cancer Research 39 (2020).

De Jongh KS, Warner C, Catterall WA. Subunits of purified calcium channels. Alpha 2 and delta are encoded by the same gene. J Biol Chem 265: 14738 – 41 (1990).

De la Mata M, et al. A slow RNA polymerase II affects alternative splicing in vivo. Mol Cell 12: 525-532 (2003).

De Los Angeles A et al. Hallmarks of pluripotency. Nature 525: 469 – 78 (2015).

De Los Angeles A, et al. Induced pluripotent stem cells in psychiatry: an overview and critical perspective. Biol Psychiatry 90: 362 – 372 (2021).

Deisseroth K, et al. Excitation-neurogenesis coupling in adult neural stem/progenitor cells. Neuron 42: 535 – 52 (2004).

Demir E, Dckson BJ. Fruitless splicing specifies male courtship behavior in Drosophila. Cell 121: 785 – 794 (2005).

Denny SK, et al. Nfib promotes metastasis through a widespread increase in chromatin accessibility. Cell 166: 328 – 342 (2016).

Dewey C, et al. Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. Genome Research 14: 661 – 666 (2004).

Dick IE, et al. Arrhythmogenesis in Timothy Syndrome is associated with defects in Ca2+-dependent inactivation. Nature Communications 7: 10370 (2016).

Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21 (2013).

Doll CA, Broadie K. Activity-dependent FMRP requirements in development of the neural circuitry of learning and memory. Development 142: 1346 – 1356 (2015).

Dolmetsch RE, et al. Signaling to the nucleus by an L-type calcium channel-calmodulin complex through the MAP kinase pathway. Science 294: 333- 339 (2001).

Dolphin AC. Calcium channel auxiliary alpha2-delta and beta subunits: trafficking and one step beyond. Nature Reviews Neuroscience 13: 542 – 555 (2012).

Dolphin AC. The alpha2-delta subunits of voltage-gated calcium channels. Biochimica et

Biophysica Acta – Biomembranes. 1828: 1541 – 1549 (2013).

Dolphin AC. Voltage-gated calcium channel alpha2-delta subunits: an assessment of proposed novel roles. F1000res 7F:1000 (2018).

Dowell R, Eddy S. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. BMC Bioinformatics 7: 40 (2006).

Dubovsky SL, Franks RD. Intracellular calcium ions in affective disorders: a review and an hypothesis. Biological Psychiatry 18: 781 – 97 (1983).

Dunn S-J, et al. Defining an essential transcription factor program for naïve pluripotency. Science 344: 1156 – 1160 (2014).

Ebert AD, Yu J, Rose Jr FF, et al. Induced pluripotent stem cells from a spinal muscular atrophy patient. Nature 457: 277 – 80 (2009).

Edvardson S, et al. Early infantile epileptic encephalopathy associated with a high voltage-gated calcium channelopathy. J Med Genet 50: 118 – 123 (2013).

Eroglu C, et al. Gabapentin receptor alpha2-delta-1 is a neuronal thrombospondin receptor for excitatory CNS synaptogenesis. Cell 139: 380 – 392 (2009).

Ertel EA, et al. Nomenclature of voltage-gated calcium channels. Neuron 25: 533 – 5 (2000).

Faedo A, et al. Differentiation of human telencephalic progenitor cells into MSNs by inducible expression of Gsx2 and Ebf1. Proc Natl Acad Sci USA 114: E1234 – 1242 (2017).

Feng B, et al. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. Nature Cell Biology 11: 197 – 203 (2009).

Ferreira MA, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nature Genetics 40: 1056 – 8 (2008).

Fogel BL, et al. RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. Hum Mol Genet 21: 4171 – 86 (2012).

Forrest et al. Open chromatin profiling in hiPSC-derived neurons prioritizes functional noncoding psychiatric risk variants and highlights neurodevelopmental loci. Cell Stem Cell 21: 305 – 318 (2017).

Fowler JL, et al. A critical look: challenges in differentiating human pluripotent stem cells into desired cell-types and organoids. Wiley Interdiscp Rev Dev Biol 9: e368 (2020).

Frohler S, et al. Exome sequencing helped the fine diagnosis of two siblings afflicted with atypical Timothy Syndrome (TS2). BMC Medical Genetics 15 (2014).

Furlanis E, Scheiffele P. Regulation of neuronal differentiation, function and plasticity by alternative splicing. Annu Rev Cell Dev Biol 34: 451 – 469 (2018).

Gabut M, et al. The splicing regulatory machinery. Cell 133: 192.e1 (2008).

Gallo J-M, et al. The role of CELF proteins in neurological disorders. RNA Biol 7: 474 – 479 (2010).

Gandal MJ, et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science 362 (2018).

Garcez PP, et al. Zika virus impairs growth in human neurospheres and brain organoids. Science 352: 816 – 818 (2016).

Gaucher EA, et al. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci 27: 315 – 321 (2002).

Gautam P, et al. Multi-species single-cell transcriptomic analysis of ocular compartment regulons. Nature Communications 12: 5675 (2021).

Geisler S, Schopf CL, Stanika R, et al. Presynaptic alpha-2-delta-2 calcium channel subunits regulate postsynaptic GABA(A) receptor abundance and axonal wiring. Journal of Neuroscience 39: 2581 – 2605 (2019).

Gershon ES, et al. A rare mutation of CACNA1C in a patient with bipolar disorder, and decreased gene expression associated with a bipolar-associated common SNP of CACNA1C in brain. Mol Psychiatry 19: 890 – 4 (2014).

Geuens T et al. The hnRNP family: insights into their role in health and disease. Hum Genet 135: 851 – 867 (2016).

Gilman SR et al. Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. Nature Neuroscience 15: 1723 – 1728 (2012).

Glen AI. Lithium prophylaxis of recurrent affective disorders. J Affect Disord 8: 259 –

265 (1985).

Glisovic T, et al. RNA-binding proteins and post-transcriptional gene regulation. FEBS Letters 582 (2008).

Golipour A, et al. A late transition in somatic cell reprogramming requires regulators distinct from the pluripotency network. Cell Stem Cell 11: 769 – 82 (2012).

Gomez-Ospina N, et al. The C terminus of the L-type voltage-gated calcium channel Ca(V)1.2 encodes a transcription factor. Cell 127: 591 – 606 (2006).

Gonzalez-Ramirez R, Felix R. Transcriptional regulation of voltage-gated Ca2+ channels. Acta Physiologica 222 (2018).

Gonzalez-Ramirez R, et al. Regulation of the voltage-gated Ca2+ channel Cacna2d1 subunit expression by the transcription factor Egr-1. Neuroscience Letters 673: 136 – 141 (2018).

Gonzalez-Ramirez R, et al. Transcriptoin factor Sp1 regulates T-type Ca(2+) channel Cav 3.1 gene expression. J Cell Physiol 229: 551 – 60 (2014).

Grant CE, et al. FIMO: scanning for occurrences of a given motif. Bioinformatics 27: 1017 – 1018 (2011).

Graveley BR, et al. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. EMBO J 17: 6747 – 6756 (1998).

Green EK, et al. The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. Mol Psychiatry 15: 1016 – 1022 (2010).

Greenberg JR. Ultraviolet light-induced crosslinking of mRNA to proteins. Nucleic Acids Research 6: 715 – 732 (1979).

Greer PL, Greenberg ME. From synapse to nucleus: calcium-dependent gene transcription in the control of synapse development and function. Neuron 59: 846 – 60 (2008).

Groen JL, et al. CACNA1B mutation is linked to unique myoclonus-dystonia syndrome. Hum Mol Genet 24: 987 – 993 (2015).

Gross S, Brent M. Using multiple alignments to improve gene prediction. J Comput Biol 13: 379 – 379 (2006).

Grossman SR, et al. Positional specificity of different transcription factor classes within enhancers. Proc Natl Acad Sci USA 115: E7222-E7230 (2018).

Grunz H, Tacke L. Neural differentiation of Xenopus laevis ectoderm takes place after disaggregation and delayed reaggregation without inducer. Cell Differ Dev 28: 211 – 217 (1989).

Guigo, Roderic. "An Introduction to Position Specific Scoring Matrices". Bioinformatica Principal, 4 Jan 2003, https://bioinformaticaupf.crg.eu/T12/MakeProfile.html

Guillemot F, et al. Mammalian achaete-schute homolog 1 is required for the early development of olfactory and autonomic neurons. Cell 75: 463 – 476 (1993).

Gulsuner S, et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell 154: 518 – 529 (2013).

Guo G, Smith A. A genome-wide screen in EpiSCs identifies Nr5a nuclear receptors as potent inducers of ground state pluripotency. Development 137: 3185 – 92 (2010).

Haerty W, Ponting CP. No gene in the genome makes sense except in the light of evolution. Annu Rev Genomics Hum Genet 15: 71 – 92 (2014).

Halgren C, et al. Haploinsufficiency of CELF4 at 18q12.2 is associated with developmental and behavioral disorders, seizures, eye manifestations, and obesity. Eur J Hum Genet doi: 10.1038/ejhg.2012.1092 (2012).

Hall J, et al. Oct4 and LIF/Stat3 additively induce Kruppel factors to sustain embryonic stem cell self-renewal. Cell Stem Cell 5: 597 – 609 (2009).

Han SP, et al. Functional diversity of the hnRNPs: past, present and perspectives. Biochm J 430: 379-392 (2010).

Haque N, Oberdoerffer S. Chromatin and splicing. Methods Mol Biol 1126: 97-113 (2014).

Harnqvist SE, et al. Variables influencing differences in sequence conservation in the fission yeast Schizosaccharomyces pombe. Journal of Molecular Evolution 89: 601 – 610 (2021).

Harrison PJ, et al. CACNA1C (Cav1.2) and other L-type calcium channels in the pathophysiology and treatment of psychiatric disorders: Advances from functional genomics and pharmacoepidemiology. Neuropharmacology. 220: 109262 (2022).

Harrison PJ, et al. Cellular calcium in bipolar disorder: systematic review and meta-analysis. Molecular Psychiatry 26: 4106 – 4116 (2021).

Harrison PJ, et al. Shorter Oxford Textbook of Psychiatry. Oxford University Press, 2018.

Harrison PJ, et al. Voltage-gated calcium channel blockers for psychiatric disorders: genomic reappraisal. Br J Psychiatry 216: 250 – 253 (2020).

Harrison PJ, Geddes JR, Tunbridge EM. The emerging neurobiology of bipolar disorder. Trends Neurosci 41: 18 – 30 (2018).

Harrison PJ, Tunbridge EM, Dolphin AC, et al. Voltage-gated calcium channel blockers for psychiatric disorders: genomic reappraisal. Br J Psychiatry 216: 250 – 253 (2020).

Harrison PJ. Molecular neurobiological clues to the pathogenesis of bipolar disorder. Current Opinion in Neurobiology 36: 1 – 6 (2016).

Harrms FL et al. Mutations in EBF3 disturb transcriptional profiles and cause intellectual disability, ataxia, and facial dysmorphism. American Journal of Human Genetics 100: 117 – 127 (2017).

Harvard Chan Bioinformatics Core. (n.d.) *Introduction to DGE – ARCHIVED.* https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Hatakeyama S, et al. Differential nociceptive responses in mice lacking the alpha1-B subunit of N-type Ca2+ channels. Neuroreport 12: 2423 – 2427 (2001).

Heinz S, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38: 576 – 589 (2010).

Helmeste DM, Tang SW. The role of calcium in the etiology of the affective disorders. Jpn J Pharmacol 77: 107 – 16 (1998).

Heyes S, et al. Genetic disruption of voltage-gated calcium channels in psychiatric and neurological disorders. Prog Neurobiol 134: 36 – 54 (2015).

Hibino H, et al. Direct interaction with a nuclear protein and regulation of gene silencing by a variant of the Ca2+-channel beta4 subunit. Proc Natl Acad Sci USA 100: 307 – 312 (2003).

Hiippala A, et al. Expanding the phenotype of Timothy syndrome type 2: an adolescent with ventricular fibrillation but normal development. Am J Med Genet A 167A: 629-34 (2015).

Hikida T, et al. Dominant-negative DISC1 transgenic mice display schizophrenia-associated phenotypes detected by measures translatable to humans. Proc Natl Acad Sci USA 104: 14501 – 6 (2007).

Hill MJ, Bray NJ. Evidence that schizophrenia risk variation in the ZNF804A gene exerts its effects during fetal brain development. Am J Psychiatry 169: 1301 – 1302 (2012).

Hiraide T, et al. Global developmental delay, systemic dysmorphism and epilepsy in a patient with a de novo U2AF2 variant. Journal of Human Genetics 66: 1185 – 1187 (2021).

Holloway AK, et al. Accelerated evolution of enhancer hotspots in the mammal ancestor. Molecular Biology and Evolution 33: 1008 – 1018 (2016).

Hoppa MB, et al. Alpha-2-delta expression sets presynaptic calcium channel abundance and release probability. Nature 486: 122 – 125 (2012).

Hor CHH, Goh ELK. Rab23 regulates radial migration of projection neurons via N-cadherin. Cereb Cortex 28: 1516 – 1531 (2018).

Hu Y, et al. Single-cell multi-scale footprinting reveals the modular organization of DNA regulatory elements. bioRxiv https://doi.org/10.1101/2023.03.28.533945.

Hu Z et al. Aberrant splicing promotes proteasomal degradation of L-type Cav1.2 calcium channels by competitive binding for CAvbeta subunits in cardiac hypertrophy. Sci Reports 6: 35247 (2016).

Huang X, et al. OCT4 cooperates with distinct ATP-dependent chromatin remodelers in naïve and primed pluripotent states in human. Nat Comun 12: 5123 (2021).

Hutchinson TE, et al. L-type calcium channels contribute to 5-HT3-receptor-evoked CaMKIIalpha and ERK activation and induction of emesis in the least shrew (Crptotis parva). Eur J Pharmacol 755: 110 – 118 (2015).

Hwang J-Y, Zukin RS. REST, a master transcriptional regulator in neurodegenerative disease. Current Opinion in Neurobiology 48: 193 – 200 (2018).

Hyun I. Stem cell ethics and policy: What's old is new again. Med (NY) 3: 820 – 823

(2022).

Ince-Dunn G, et al. Neuronal eval-like (Hu) proteins regulate RNA splicing and abundance to control glutamate levels and neuronal excitability. Neuron 75: P1067 – 1080 (2012).

Izquierdo JM, Valcarcel J. A simple principle to explain the evolution of pre-mRNA splicing. Genes Dev 20: 1679 – 1684 (2006).

Jaffe AE, et al. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. Nature Neuroscience 21: 1117 – 1125 (2018).

Jaffe AE, et al. Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. Nature Neuroscience 18: 154 – 161 (2015).

Jaffe AE, et al. Profiling gene expression in human dentate gyrus granule cel layer reveals insights into schizophrenia and its genetic risk. Nature Neuroscience 23: 510 – 519 (2020).

Jang DH, et al. Autistic and Rett-like features associated with 2q33.3-q34 interstitial deletion. Am J Med Genet A 167A: 2213 – 8 (2015).

Jannson-Fritzberg LI, et al. DNMT1 inhibition by pUG-fold quadruplex RNA. RNA 29: 346 – 360 (2023).

Jeon H, et al. Comprehensive identification of Kruppel-like Factor family members contributing to the self-renewal of mouse embryonic stem cells and cellular reprogramming. PLoS One 11: e0150715 (2016).

Jiang Y, et al. The U2AF2 / circRNA ARF1 / miR-342-3p / ISL2 feedback loop regulates angiogenesis in glioma stem cells. Journal of Experimental and Clinical Cancer Research 39 (2020).

Jorgensen HF, et al. Is REST required for ESC pluripotency? Nature 457: E4-5 (2009).

Josephson IR, Varadi G. The beta subunit increases Ca2+ currents and gating charge movements of human cardiac L-type Ca2+ channels. Biophys J 70: 1285 – 1293 (1996).

Juan-Mateu J, et al. Pancreatic microexons regulate islet function and glucose homeostasis. Nature Metabolism (2023).

Jun K, et al. Ablation of P/Q-typeCa2+ channel currents, altered synaptic transmission,

and progressive ataxia in mice lacking the alpha1A-subunit. Proc Natl Acad Sci USA 96: 15245 – 15250 (1999).

Jung S, et al. Role of dendritic voltage-gated calcium channels of pyramidal neurons in regulating synaptic plasticity. Neural Plasticity 2947902 (2017).

Kaczynski J, et al. Sp1- and Kruppel-like transcription factors. Genome Biol 4: 206 (2003).

Kajimura D, et al. Identification of genes regulated by transcription factor Klf7 in differentiating olfactory sensory neurons. Gene 388: 34 – 42 (2007).

Kamijo S, et al. A critical neurodevelopmental role for L-type voltage-gated calcium channels in neurite extension and radial migration. Journal of Neuroscience 38: 5551-5566 (2018).

Kantarci H, et al. Tfap2a promotes specification and maturation of neurons in the inner ear through modulation of Bmp, Fgf, and notch signaling. PloS Genet 17: e1005037 (2015).

Kato Y, et al. ELAVL2-directed RNA regulatory network drives the formation of quiescent primordial follicles. EMBO Reports 20: e48251 (2019).

Kay RF, et al. Anthropoid origins. Science 275: 797 – 804 (1997).

Khanal R. Identification of RNA binding proteins and RNA binding residues using effective machine learning techniques. Senior Honor Theses https://scholarworks.uno.edu/honors_theses/128 (2019).

Khandaker GM, et al. Childhood infection and adult schizophrenia: a meta-analysis of population-based studies. Schizophr. Res. 139: 161 – 168 (2012).

Kilpinen H, et al. Common genetic variation drives molecular heterogeneity in human iPSCs. Nature 546: 370 – 375 (2017).

Kilpinen H, Goncalves A, Leha A, et al. Common genetic variation drives molecular heterogeneity in human iPSCs. Nature 546: 370 – 375 (2017).

Kim D, et al. HISAT: a fast spliced aligner with low memory requirements. Nature Methodds 12:357 (2015).

Kim H-L, et al. Rat brain expresses an alternatively spliced form of the dihydropyridine-sensitive L-type calcium channel alpha2 subunit. Proc Natl Acad Sci USA 89: 3251 – 3255 (1992).

Kim J, et al. An extended transcriptional network for pluripotency in embryonic stem cells. Cell 132: 1049 – 1061 (2008).

Kim JY, et al. Interplay between DISC1 and GABA signaling regulaltes neurogenesis in mice and risk for schizophrenia. Cell 148: 1051 – 64 (2012).

Kirischuk S, et al. Modulation of neocortical development by early neuronal activity: physiology and pathophysiology. Front Cell Neurosci 11: 379 (2017).

Kleinman JE, et al. Genetic neuropathology of schizophrenia: new approaches to an old uestion and new uses for postmorterm human brains. Biological Psychiatry 69: 140-5 (2011).

Kloiber S, et al. Neurodevelopmental pathways in bipolar disorder. Neurosci Biobehav Rev 112: 213 – 226 (2020).

Klugbauer N, Lacinova L, Marais E, Hobom M, Hofmann F. Molecular diversity of the calcium channel alpha2-delta subunit. J NEurosci 19: 684 – 691 (1999).

Kohler S, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Research 42: D966 – D974 (2014).

Konermann S, et al. Transcriptome engineering with RNA-targeting Type VI-D CRISPR effectors. Cell 173: 665 – 676 (2018).

Krebs J, et al. Genes (4^{th} ed). Jones & Bartlett Learning, LLC. (2021).

Kurshan PT, et al. Presynaptic alpha2-delta-3 is required for synaptic morphogenesis independent of its Ca2+-channel functions. Nature Neurosci 12: 1415- 1423 (2009).

Kwon SC, et al. The RNA-binding protein repertoire of embryonic stem cells. Nature Structural and Molecular Biology 20: 1122 – 1130 (2013).

Lacerda AE, et al. Normalization of current kinetics by interaction between the alpha-1 and beta subunits of the skeletal muscle dihydropyridine-sensitive Ca2+ channel. Nature 352: 527 – 530 (1991).

Lacinova L, Klugbauer N, Hofmann F. State- and isoform-dependent interaction of isradipine with the alpha(1C) L-type calcium channel. Pflugers Arch 440: 50 – 60 (2000).

Lafaille FG, Pessach IM, Zhang S-Y et al. Impaired intrinsic immunity to HSV-1 in human iPSC-derived TLR3-deficient CNS cells. Nature 491: 769 – 73 (2012).

Lambert SA, et al. The human transcription factors. Cell 172: 650 – 665 (2018).

Lana B, Schlick B, Martin S, et al. Differential upregulation in DRG neurons of an alpha2-delta-1 splice variant with lower affinity for gabapentin after peripheral sensory neuron injury. Pain 155: 522 – 533 (2014).

Lancaster MA, Renner M, Martin C-A, et al. Cerebral organoids model human brain development and microcephaly. Nature 501: 373 – 9 (2013).

Larson ED, et al. Pioneering the developmental frontier. Mol Cell 81: 1640 – 1650 (2021).

Laub F, et al. Transcription factor KLF7 is important for neuronal morphogenesis in selected regions of the nervous system. Mol Cell Biol 25: 5699 – 5711 (2005).

Leblond CS, et al. Both rare and common genetic variants contribute to autism in the Faroe Islands. NPJ Genom Med 4: 1 (2019).

Leclerc C, et al. An elevation of internal calcium occurring via L-type channels mediate neural induction in the amphibian embryo. In Organisation of the early vertebrate (eds N.Zagris, A.M. Duprat, and J. Durston) pp. 209 – 226. New York, NY: Plenum Press (1995).

Leclerc C, et al. Calcium transients triggered by planar signals induce the expression of ZIC3 gene during neural induction in Xenopus. Developmental Biology 261: 381 – 390 (2003).

Leclerc C, et al. Imaging patterns of calcium transients during neural induction in Xenopus laevis embryos. J Cell Sci 113: 3519 – 3529 (2000).

Leclerc C, et al. L-type calcum channel activation contros the in vivo transduction of the neutralizing signal in amphibian embryos. Mech Dev 64: 105 – 110 (1997).

Leclerc C, et al. Neural determination in Xenopus laevis embryos: control of early neural gene expression by calcium. J Soc Biol 195: 327 – 337 (2001).

Lelli KM, et al. Disentangling the many layers of eukaryotic transcriptional regulation. Annu Rev Geneti. 46: 43 – 68 (2012).

Leung SK, et al. Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. Cell Reports 37: 110022 (2021).

Li H, et al. The sequence alignment/map format and SAMtools. Bioionformatics 25: 2078 – 2079 (2009).

Li Z, et al. Identification of transcription factor binding sites using ATA-seq. Genome Biol 20: 45 – 45 (2019).

Li R, et al. A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. Cell Stem Cell 7: 51 – 63 (2010).

Li X, et al. Finding the target sites of RNA-binding proteins. Wiley Interdiscip Rev RNA 5: 111 – 130 (2014).

Li YI, et al. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in the human brain transcriptome. Genome Research 25: 1 – 13 (2015).

Liang X, Song M-R, Xu Z, et al. Isl1 is required for multiple aspects of motor neuron development. Mol Cell Neurosci 47: 215 – 222 (2011).

Liao P, Yu D, Lu S, et al. Smooth muscle-selective alternatively spliced exon generates functional variation in Cav1.2 calcium channels. J Biol Chem 279: 50329 – 50335 (2004).

Liao P, Zhang Hy, Soong TW. Alternative splicing of voltage-gated calcium channels: from molecular biology to disease. Pflugers Arch 458: 481 – 487 (2009).

Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 12: 21 – 27 (2002).

Linares AJ, et al. The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. Elife 4: e09268 (2015).

Lindberg U, Sundquist B. Isolation of messenger ribonucleoproteins from mammalian cells. Journal of Molecular Biology 86: 451 – 468 (1974).

Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29

mammals. Nature 478: 476 – 482 (2011).

Lipscombe D, Andrade A. Calcium channel Cava1 splice isoforms – tissue specificity and drug action. Curr Mol Pharmacol 8: 22-31 (2015).

Lipscombe D, Andrade A. Calcium channel CaVa1 splice isoforms – tissue specificity and drug action. Curr Mol Pharmacol 8: 22 – 31 (2015).

Liu N-N, et al. The absence of NIPA2 enhances neural excitability through BK (big potassium) channels. CNS Neurosci Ther 25: 865 – 875 (2019).

Liu Q, et al. Interleukin-17 inhibits adult hippocampal neurogenesis. Scientific Reports 4: 7554 (2014).

Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell 165: 535 – 550 (2016).

Lodha N, et al. Congenital stationary night blindness: mutation update and clinical variability. Adv Exp Med Biol 723: 371 – 379 (2012).

Lodha N, et al. Congenital stationary night blindness in mice – a tale of two Cacna1f mutants. Adv Exp Med Biol 664: 549 – 558 (2010).

Lomberk G, Urrutia R. The family feud: turning off Sp1 by Sp1-like KLF proteins. Biochem J 392: 1 – 11 (2005).

Loots GG, Ovcharenko I. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Research 32: 217 – 221 (2004).

Loria PM, et al. Two neuronal, nuclear-localized RNA binding proteins involved in synaptic transmission. Curr Biol 13: 1317 – 1323 (2003).

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESEQ2. Genome Biology 15: 550 (2014).

Lu CK, et al. Rbms3, an RNA-binding protein, mediates the expression of Ptf1a by binding to its 3'UTR during mouse pancreas development. DNA Cell Biol 31: 1245 – 1251 (2012).

MacCarthy CM, et al. Enhancing Sox/Oct cooperativity induces higher-grade developmental reset. bioRxiv doi: https://doi.org/10.1101/2022.09.23.509242

Mao Y, et al. Disrupted in schizophrenia 1 regulates neuronal progenitor proliferation via modulation of GSK3beta/beta-catenin signaling. Cell 136: 1017 – 31 (2009).

Mariani J et al. FOXG1-dependent dysregulation of GABA/glutamate neuron differentiation in autism spectrum disorders. Cell 162: 375 – 390 (2015).

Martello G, et al. Esrrb is a pivotal target of the Gsk3/Tcf3 axis regulating embryonic stem cell self-renewal. Cell Stem Cell 11: 491 – 504 (2012).

Martello G, et al. Identification of the missing pluripotency mediator downstream of leukaemia inhibitory factor. EMBO J 32: 2561 – 74 (2013).

Martins-Taylor K, et al. Recurrent copy number variations in human induced pluripotent stem cells. Nature Biotechnology 29: 488 – 491 (2011).

McConnell BB, Yang VW. Mammalian Kruppel-like Factors in health and diseases. Physiol Rev 90: 1337 – 1381 (2010).

McCullough AJ, Berget SM. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol Cell Biol 17: 4562 – 4571 (1997).

Meers MP, et al. Pioneeer factor-nucleosome binding events during differentiation are motif encoded. Mol Cell 75: 562 – 575 (2019).

Mehmood A, Laiho A, et al. Systematic evaluation of differential splicing tools for RNA-seq studies. Briefings in Bioinformatics 21: 2052 – 2065 (2020).

Mikkelsen TS, et al. Dissecting direct reprogramming through integrative genomic analysis. Nature 454: 49 – 55 (2008).

Miller G. Is pharma running out of brainy ideas? Science 329: 502 – 504 (2010).

Miller JA, et al. Transcriptional landscape of the prenatal human brain. Nature 508: 199 – 206 (2014).

Mitchell SF, et al. Global analysis of Yeast mRNPs. Nature Structural and Molecular Biology 20: 127 – 133 (2013).

Mitchell SF, et al. Global analysis of yeast mRNPs. Nature Structural and Molecular Biology 20: 127 – 133 (2013).

Molero AE, et al. Selective expression of mutant huntingtin during development recapitulates characteristic feature of Huntington's disease. Proc Natl Acad Sci USA 113: 5736 – 5741 (2016).

Molineux ML, et al. Specific T-type calcium channel isoforms are associated with distinct burst phenotypes in deep cerebellar nuclear neurons. Proc Natl Acad Sci USA 103: 5555-5560 (2006).

Moore DL, et al. Kruppel-like transcription factors in the nervous system: novel players in neurite outgrowth and axon regeneration. Mol Cell Neurosci 47: 233 – 243 (2011).

Moreau M, et al. Calcium signalling during neural induction in Xenopus laevis embryos. Philos Trans R Soc Lond B Biol Sci 363: 1371 – 5 (2008).

Moreau M, et al. Increased internal Ca2+ mediates neural induction in the amphibian embryo. Proc Natl Acad Sci USA 91: 12639 – 12643 (1994).

Morris-Rosendahl, DJ. Neurodevelopmental disorders – the history and future of a diagnostic concept. Dialogues Clin Neurosci 22: 65 – 72 (2020).

Moses A, et al. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding sitespecific evolutionary model. Genome Biol 5: R9 (2004).

Moskvina V, et al. Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. Mol Psychiatry 14: 252 – 260 (2009).

Nachbauer W, et al. Episodic ataxia type 2: phenotype characteristics of a novel CACNA1A mutation and review of the literature. J Neurol 261: 983 – 991 (2014).

Nagai T, et al. Zic2 regulates the kinetics of neurulation. PNAS 97: 1618 – 1623 (2000).

Nakayama J, et al. Monozygotic twins with de novo ZIC2 gene mutations discordant from holoprosencephaly. Neurology 86: 15 (2016).

Nanou E, Catterall WA. Calcium channels, synaptic plasticity, and neuropsychiatric disease. Neuron 98: 466 – 81 (2018).

Nazarov PV, et al. RNA sequencing and transcriptome arrays analyses show opposing results for alternative splicing in patient derived samples. BMC Genomics 18 (2017).

Nielsen PR et al. Hospital contacts with infection and risk of schizophrenia: a population-based cohort study with linkage of Danish national registers. Schizophr. Bull 40: 1526 – 1532 (2014).

Niwa H, et al. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nature Genetics 24: 372 – 376 (2000).

Nordhoff V, et al. Comparative analysis of human, bovine, and murine Oct-4 upstream promoter sequences. Mammalian Genome 12: 309 – 317 (2001).

Norton S "Methods for Robust Quantification of RNA Alternative Splicing I Heterogeneous RNA-seq Datasets" (2019). Publicly Accessible Pen Dissertations. 3460. https://repository.upenn.edu/edissertations/3460

Nystoriak MA, Murakami K, Penar PL, et al. Ca(v)1.2 splice variant with exon 9* is critical for regulation of cerebral artery diameter. Am J Physiol Heart Circ Physiol 297: H1820-H1828 (2009).

O'Roak BJ, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet 43: 585 – 589 (2011).

Oksuz O, et al. Transcription factors interact with RNA to regulate genes. bioRxiv https://doi.org/10.1101/2022.09.27.509776

Oldak B, et al. Recent insights into mammalian natural and synthetic ex utero embryogenesis. Curr Opin Genet Dev 77: 101988 (2022).

Osafune K, et al. Marked differences in differentiation propensity among human embryonic stem cells. Nature Biotechnology 26: 313 – 5 (2008).

Otteson DC, et al. Kruppel-like factor 15, a zinc-finger transcriptional regulator, represses the rhodopsin and interphotoreceptor retinoid-binding protein promoters. Invest Ophthalmol Vis Sci 45: 2522 – 30 (2004).

Owen MJ, et al. Neurodevelopmental hypothesis of schizophrenia. Br J Psychiatry 198: 173 – 175 (2011).

Owen MJ, O'Donovan MC. Schizophrenia and the neurodevelopmental continuum: evidence from genomics. World Psychiatry 16: 227 – 235 (2017).

Panagiotakos G, et al. Aberrant calcium channel splicing drives defects in cortical differ-

entiation in Timothy Syndrome. ELife 8: e51037 (2019).

Panagiotakos G, Pasca SP. A matter of space and time: emerging roles of disease-associated proteins in neural development. Neuron (2021).

Pang Z, et al. Induction of human neuronal cells by defined transcription factors. Nature 476: 220 – 3 (2011).

Panne D, et al. An atomic model of the interferon-beta enhanceosome. Cell 129: 1111-1123 (2007).

Papasaikas et al. Functional splicing network reveals extensive regulatory potential of core spliceosomal machinery. Molecular Cell 57: 7 -22 (2015).

Pasca SP, et al. Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy Syndrome. Nature Medicine 17: 1657-62 (2011).

Pastor WA, et al. TFAP2C regulates transcription in human naïve pluripotency by opening enhancers. Nature Cell Biology 20: 553 – 564 (2018).

Paun O, et al. ASCL1 interacts with mSWI/SNF at distal regulatory elements to regulate neural differentiation. bioRxiv https://doi.org/10.1101/2022.10.09.510609

Pedersen J, et al. Identification and classification of conserved RNA secondary structures in the human genome. PLOS Computat. Biol 2: e3 (2006).

Penkov D, et al. Cloning of a human gene closely related to the genes coding for the c-myc single-strand binding proteins. Gene 243: 27 – 36 (2000).

Pescucci C, et al. Chromosome 2 deletion encompassing the MAP2 gene in a patient with autism and Rett-like features. Clin Genet 64: 497 – 501 (2003).

Piovesan A, et al. On the length, weight, and GC content of the human genome. BMC Research Notes 12: 106 (2019).

Pippucci T, et al. A novel null homozygous mutation confirms CACNA2D2 as a gene mutated in epileptic encephalopathy. PLoS ONE 8: e82154 (2013).

Pitt GS, et al. Voltage-gated calcium channels in nonexcitable tissues. Annu Rev Physiol 83: 183 – 203 (2021).

Pitt GS, Matsui M, Cao C. Voltage-gated calcium channels in nonexcitable tissues. Annu

Rev Physiol 83: 183 – 203 (2021).

Platzer J, et al. Congenital deafness and sinoatrial node dysfunction in mice lacking class D L-type Ca2+ channels. Cell 102: 89 – 97 (2000).

Plumbly W, et al. L-type voltage-gated calcium channel regulation of in vitro human cortical neuronal networks. Scientific Reports 9 (2019).

Pollard K, et al. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research 20: 110 – 121 (2010).

Polo JM, et al. Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. Nature Biotechnology 28: 848 – 55 (2010).

Ponting CP, et al. Evolution of domain families. Adv Protein Chem 54: 185 – 244 (2000).

Powis Z, et al. Clinical diagnostic exome evaluation for an infant with a lethal disorder: genetic diagnosis of TARP syndrome and expansion of the phenotype in a patient with a newly reported RBM10 alteration. BMC Med Genet 18: 60 (2017).

Qin N, et al. Molecular cloning and characterization of the human voltage-gated calcium channel alpha2-delta 4 subunit. Mol Pharmacol 62: 485 – 496 (2002).

Quadrato G, et al. Cell diversity and network dynamics in photosensitive human brain organoids. Nature 545: 48 – 53 (2017).

Rakic P. Specification of cerebral cortical areas. Science 241: 170 – 175 (1988).

Ramanathan M, et al. Methods to study RNA-protein interactions. Nature Methods 16: 225 – 234 (2019).

Ramirez-Latorre JA. Functional upregulation of Ca2+ activated K+ channels in the development of substantia nigra dopamine neurons. PLoS ONE 7: e51610 (2012).

Ray D, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature 499: 172 – 177 (2013).

Ray D, et al. Rapid and systematic analysis of RNA recognition specificities of RNA-binding proteins. Nature Biotechnology 27: 667 – 670 (2009).

Reay WR, et al. Genetics-informed precision treatment formulation in schizophrenia and bipolar disorder. American Journal of Human Genetics 109: 1620-1637 (2022).

Rebelo S, Chen ZF, Anderson DJ, et al. Involvement of DRG11 in the development of the primary afferent nociceptive system. Mol Cell Neurosci 33: 236-246 (2006).

Reiner A, et al. Wild-type huntingtin plays a role in brain development and neuronal survival. Mol Neurobiol 28: 259 – 275 (2003).

Revah O, et al. Maturation and circuit integration of transplanted human cortical organoids. Nature 610: 319 – 326 (2022).

Reyes, A., Anders, S. and Huber, W., 2021. *Inferring differential exon usage in RNA-Seq data with the DEXSeq package.* [online] Bioconductor.org. Available at: <https://bioconductor.org/packages/devel/bioc/vignettes/DEXSeq/inst/doc/DEXSeq.html> [Accessed 15 December 2021].

Richter WF, et al. The mediator complex as a master regulator of transcription by RNA polymerase II. Nat Rev Mol Cell Biol 1-18 (2022).

Rieger MA, et al. CLIP and massively parallel functional analysis of CELF6 reveal a role in destabilizing synaptic gene mRNAs through interaction with 3'UTR elements. Cell Reports 33: 1-15 (2020).

Rodda DJ, et al. Transcriptional regulation of nanog by OCT4 and SOX2. J Biol Chem 280: 24731-7 (2005).

Rothstein M, Simoes-Costa M. Heterodimerization of TFAP2 pioneer factors drives epigenomic remodeling during neural crest specification. Genome Research 30: 35 – 48 (2020).

Sadri J, et al. Predicting site-specific human selective pressure using evolutionary signatures. Bioinformatics 27: i266 – i274 (2011).

Saegusa H, et al. Properties of human Cav2.1 channel with spinocerebellar ataxia type 6 mutation expressed in Purkinje cells. Mol Cell Neurosci 34: 261 – 270 (2007).

Saldana-Meyer R et al. RNA interactions are essential for CTCF-mediated genome organization. Molecular Cell 76: 412 – 422 (2019).

Samavarchi-Tehrani P, et al. Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. Cell Stem Cell 7: 64 – 77 (2010).

Samavarchi-Tehrani P, et al. Functional genomics reveals a BMP-driven mesenchymal-to-

epithelial transition in the initiation of somatic cell reprogramming. Cell Stem Cell 7: 64 – 77 (2010).

Schopf CL, et al. Presynaptic alpha2-delta subunits are key organizers of glutamatergic synapses. Proc Natl Acad Sci USA 118: e1920827118 (2021).

Schwartzentruber J, Foskolou S, Kilpinen H, et al. Molecular and functional variation in iPSC-derived sensory neurons. Nature Genetics 50: 54-61 (2018).

Seitter H, Koschak A. Relevance of tissue specific subunit expression in channelopathies. Neuropharmacology 132: 58 – 70 (2018).

Shao J, et al. Identification of lncRNA expression profiles and ceRNA analysis in the spinal cord of morphine-tolerant rats. Molecular Brain 11(2018).

Sherwood RI, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling Dnase profile magnitude and shape. Nature Biotechnology 32: 171 – 178 (2014).

Sims RJ 3rd, et al. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol Cell 28: 665-676 (2007).

Singh R, et al. Building specificity with nonspecific RNA-binding proteins. Nature Structural Molecular Biology 12: 645 – 653 (2005).

Singh SK, et al. REST maintains self-renewal and pluripotency of embryonic stem cells. Nature 453: 223 – 7 (2008).

Singh T, et al. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. Nature Genetics 49: 1167 – 1173 (2017).

Sklar P, et al. Whole-genome association study of bipolar disorder. Mol Psychiatry 13: 558 – 69 (2008).

Slatko BE et al. Overview of next-generation sequencing technologies. Current Protocols in Molecular Biology 122: e59 (2018).

Sleven H, et al. De novo mutations in EBF3 cause a neurodevelopmental syndrome. American Journal of Human Genetics 100: 138 – 150 (2017).

Smedler E, et al. Disrupted Cacna1c gene expression perturbs spontaneous Ca2+ activity causing abnormal brain development and increased anxiety. Proc Natl Acad Sci USA 119: e2108768119 (2022).

Smith CW, et al. Alternative pre-mRNA splicing: the logic of combinatorial control. Trends Biochem Sci 25: 381 – 388 (2000).

Soldatov NM, Bouron A, Reuter H. Different voltage-dependent inhibition by dihydropyridines of human calcium channel splice variants. J Biol Chem 270: 10540 – 10543 (1995).

Soldatov NM, et al. Different voltage-dependent inhibition by dihydropyridines of human Ca2+ channel splice variants. Journal of Biological Chemistry 270: 10540 – 1543 (1995).

Soldatov NM, Oz M, O'Brien KA, et al. Molecular determinants of L-type calcium channel inactivation. Segment exchange analysis of the carboxyl-terminal cytoplasmic motif encoded by exons 40-42 of the human alpha1C subunit gene. J Biol Chem 273 957 – 963 (1998).

Soldatov NM, Zuhlke RD, Bouron A, et al. Molecular structures involved in L-type calcium channel inactivation: role of the carboxyl-terminal region encoded by exons 40-42 in alpha(1C) subunit in the kinetics and calcium-dependence of inactivation. J Biol Chem 272: 3560-3566 (1997).

Soldatov NM. Cav1.2, cell proliferation, and new target in atherosclerosis. ISRN Biochem 2013: 463527 (2013).

Soldatov NM. Molecular diversity of L-type Ca2+ channel transcripts in human fibroblasts. Proceedings of the National Academy of Sciences of the United States of America 89: 4628 – 4632 (1992).

Soldner F, Jaenisch R. Stem cells, genome editing, and the path to translational medicine. Cell 175: 615 – 632 (2018).

Soto LF, et al. Compendium of human transcription factor effector domains. Molecular Cell 82: 514 – 526 (2022).

Soufi A, et al. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell 151: 994 – 1004 (2012).

Soufi A, et al. Pioneer transcription factors target partial DNA motifs on nucleosomes to

initiative reprogramming. Cell 161: 555 – 568 (2015).

Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 13: 613 – 626 (2012).

Splawski I, et al. CaV1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. Cell 119: 19 – 31 (2004).

Splawski I, et al. Severe arrhythmia disorder caused by cardiac L-type calcium channel mutations. Proc Natl Acad Sci USA 102: 8089 – 96 (2005).

Srivastava R, et al. Role of SARS-Cov-2 in altering the RNA-binding protein and miRNA-directed post-transcriptional regulatory networks in humans. Int J Mol Sci 21: 7090 (2020).

St Clair D, et al. Rates of adult schizophrenia following prenatal exposure to the Chinese famine of 1959 – 1961. JAMA 294: 557 – 562 (2005).

Stephens MC, et al. The developmental roots of neurodegeneration. Neuron 110: 1 – 3 (2022).

Stotz SC, Jarvis SE, Zamponi GW, et al. Functional roles of cytoplasmic loops and pore lining transmembrane helices in the voltage-dependent inactivation of HVA calcium channels. J Physiol 554: 263-273 (2004).

Streeter I, Harrison PW, Faulconbridge A, et al. The human-induced pluripotent stem cell initiative-data resources for cellular genetics. Nucleic cids Research 45: D691-D697 (2017).

Striessnig J, Pinggera A, Kaur G, et al. L-type Ca2+ channels in heart and brain. Wiley Interdiscp Rev Membr Transp Signal 3: 15-38 (2014).

Striessnig J. Pharmacology, structure, and function of cardiac L-type calcium channels. Cell Physiol Biochem 9: 242 – 269 (1999).

Sun W, Larson MJ, Kiyoshi CM, et al. Gabapentinoid treatment promotes corticospinal plasticity and regeneration following murine spinal cord injury. J Clin Invest 130: 345-358 (2020).

Sune-Pou M, et al. Targeting splicing in the treatment of human disease. Genes 8:37 (2017).

Sung MH, et al. Genome-wide footprinting: ready for prime time? Nature Methods 13: 222 – 228 (2016).

Susser ES, Lin SP. Schizophrenia after prenatal exposure to the Dutch Hunger Winter of 1944 – 1945. Arch Gen Psychiatry 49: 983 – 988 (1992).

Taapken SM, et al. Karyotypic abnormalities in human induced pluripotent stem cells and embryonic stem cells. Nature Biotechnology 29: 313 – 314 (2011).

Takahashi K, Tanabe K, Ohnuki M, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131: 861 – 72 (2007).

Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126: 663 – 676 (2006).

Takashima Y, et al. Resetting transcription factor control circuitry toward ground-state pluripotency in human. Cell 158: 1254 – 1269 (2014).

Tang ZZ, et al. Regulation of the mutually exclusive exons 8a and 8 in the Cav1.2 calcium channel transcript by polypyrimidine tract-binding protein. J Biol Chem 286: 10007 – 10016 (2011).

Tang ZZ, Liang MC, Lu S, et al. Transcript scanning reveals novel and extensive splice variations in human l-type voltage-gated calcium channel, Cav1.2 alpha1 subunit. J Biol Chem 279: 44335- 44343 (2004).

Tang ZZ, Zheng S, et al. Developmental control of Cav1.2 L-type calcium channel splicing by Fox proteins. Mol. Cell. Biol 29: 4757-4765 (2009).

Tao R, et al. Expression of ZNf804A in human brain and alterations in schizophrenia, bipolar disorder, and major depressive disorder: a novel transcript fetally regulated by the psychosis risk variant rs1344706. JAMA Psychiatry 71: 1112-20 (2014).

Tarazi S, et al. Post-gastrulation synthetic embryos generated ex utero from mouse naïve ESCs. Cell 185: 3290 – 3306 (2022).

Tasic B, et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature 563: 72 – 78 (2018).

Tedeschi A, Dupraz S, Laskowski CJ, et al. The calcium channel subunit alpha2delta2 suppresses axon regeneration in the adult CNS. Neuron 92: 419 – 434 (2016).

Teixeira JR, et al. Transcription factor 4 and its association with psychiatric disorders. Translational Psychiatry 11 (2021).

Tejedor JR, et al. Genome-wide identification of Fas/CD95 alternative splicing regulators reveals links with iron homeostasis. Molecular Cell 57: 23 – 38 (2015).

Tesar PJ, et al. New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature 448: 196 – 9 (2007).

Theunissen TW, et al. Systematic identification of culture conditions for induction and maintenance of naïve human pluripotency. Cell Stem Cell 15: 471 – 487 (2014).

Tian H, et al. Kruppel-like factor 7 deficiency causes autistic-like behavior in mice via regulating Clock gene. Cell & Bioscience 12 (2022).

Tiwari S, et al. Artherosclerosis-related molecular alteration of the human Cav1.2 calcium channel alpha1-C subunit. Proc Natl Acad Sci USA. 103: 17024 – 17029 (2006).

Trubetskoy V, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 604:502 – 508 (2022).

Valence S, et al. Exome sequencing in congenital ataxia identifies two new candidate genes and highlights a pathophysiological link between some congenital ataxias and early infantile epileptic encephalopathies. Genet Med 21: 553 – 563 (2019).

Van Nostrand EL, et al. Robust transcriptome-wide discovery of RNA-binding protein sites with enhanced CLIP (eCLIP). Nature Methods 13: 508 – 514 (2016).

Van Nostrand EL, et al. A large-scale binding and functional map of human RNA-binding proteins. Nature 583: 711 – 719 (2020).

Vaquerizas JM, et al. A census of human transcription factors: function, expression and evolution. Nature Reviews Genetics 10: 252 – 263 (2009).

Vaquero-Garcia J, Barrera A, Gazzara MR, et al. A new view of transcriptome complexity and regulation through the lens of local splicing variations. Elife 5:e11752 (2016).

Vecchia D, et al. Abnormal cortical synaptic transmission in Cav2.1 knockin mice with the S218L missense mutation which causes a severe familial hemiplegic migraine syndrome in humans. Front Cell Neurosci 9: 8 (2015).

Velasco S, et al. Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. Nature 570: 523 – 528 (2019).

Vierbuchen T, et al. Direct conversion of fibroblasts to functional neurons by defined factors. Nature 463: 1053 – 41 (2010).

Vos SM. Understanding transcription across scales: from base pairs to chromosomes. Molecular Cell 81: 1601 – 1616 (2021).

Vue TY, et al. ASCL1 regulates neurodevelopmental transcription factors and cell cycle genes in brain tumors of glioma mouse models. Glia 68: 2613 – 2630 (2020).

Wagenmakers et al. Cross-linking of mRNA to proteins by irradiation of intact cells with ultraviolet light. European Journal of Biochemistry 112 (1980).

Wagnon JL, et al. CELF4 regulates translation and local abundance of a vast set of mRNAs, including genes associated with regulation of synaptic function. PLoS Genetics 8: e1003067 (2012).

Wagnon JL, et al. Etiology of a genetically complex seizure disorder in Celf4 mutant mice. Genes Brain Behav 10: 765 – 777 (2011).

Wallmen B, et al. Intrinsic properties of Tcf1 and Tcf4 splice variants determine cell-type-specific Wnt/Beta-catenin target expression. Nucleic Acids Research 19: 9455 – 9469 (2012).

Wang J, et al. A protein interaction network for pluripotency of embryonic stem cells. Nature 444: 364 – 8 (2006).

Wang J, et al. Tandem affinity purification of protein complexes in mouse embryonic stem cells using in vivo biotinylation. Curr Protoc Stem Cell Biol Chapter 1: Unit1B.5 (2009).

Wang X, et al. A presumed missense variant in the U2AF2 gene causes exon skipping in neurodevelopmental diseases. Journal of Human Genetics (2023).

Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA 14: 802 – 813 (2008).

Welling A, Ludwig A, Zimmer S, et al. Alternatively spliced IS6 segments of the alpha 1C gene determine the tissue-specific dihydropyridine sensitivity of cardiac and vascular smooth muscle L-type calcium channels. Circ Res 81: 526-532 (1997).

Whalen S, et al. Machine learning dissection of human accelerated regions in primate neurodevelopment. Neuron 111: 857 – 873 (2023).

Whalen S, Pollard KS. Enhancer function and evolutionary roles of human accelerated regions. Annu Rev Genet 56: 423 – 439 (2022).

Wheeler DB, et al. Roles of N-type and Q-type Ca2+ channels in supporting hippocampal synaptic transmission. Science 264: 107 – 111 (1994).

Whiffin N, et al. Characterizing the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. Nature Communications 11: 2523 (2020).

Whittaker CA, Hynes RO. Distribution and evolution of von Willebrand/integrin A domain: dispersed domains with roles in cell adhesion and elsewhere. Mol Biol Cell 13: 3369 – 3387 (2002).

Willsey AJ, et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell 155: 997 – 1007 (2013).

World Health Organization (2001). The World Health Report 2001: Mental Disorders affect one in four people. https://www.who.int/news/item/28-09-2001-the-world-health-report-2001-mental-disorders-affect-one-in-four-people

Wu J, et al. Chromatin analysis in human early development reveals epigenetic transition during ZGA. Nature 557: 256 – 260 (2018).

Wu J, et al. Interspecies chimerism with mammalian pluripotent stem cells. Cell 168: 473 – 486 (2017).

Wu J, et al. The landscape of accessible chromatin in mammalian preimplantation embryos. Nature 534: 652 – 657 (2016).

Xiang Y, et al. Fusion of regionally specified hPSC-derived organoids models human brain development and interneuron migration. Cell Stem Cell 21: 383 – 398 (2017).

Xiao X, et al. Molecular mechanisms underlying noncoding risk variations in psychiatric gene studies. Molecular Psychiatry 22: 497-511 (2017).

Xiong L et al. Oct4 differentially regulates chromatin opening and enhancer transcription in pluripotent stem cells. Elife 11:e71533 (2022).

Xu B, et al. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nature Genetics 44: 1365 – 1369 (2012).

Yamada K, et al. Genome-wide association study of schizophrenia in Japanese population. PLoS One 6: e20468 (2011).

Yan F, et al. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biology 21 (2020).

Yang S-H, et al. ZIC3 controls the transition from naïve to primed pluripotency. Cell Reports 27: 3215 – 3227 (2019).

Yang Y, et al. Complex seizure disorder caused by Brunol4 deficiency in mice. PLoS Genet 3: e124 (2007).

Yang Y, et al. L-type Ca2+ channel a1c subunit isoform switching in failing human ventricular myocardium. J Mol Cell Cardiol 32: 973 – 984 (2000).

Yeo G, e al. Variation in alternative splicing across human tissues. Genome Biol 5: R74 (2004).

Yoshimizu T, et al. functional implications of a psychiatric risk variant within CACNA1C in induced human neurons. Mol Psychiatry 20: 162 – 9 (2015).

Young GT, et al. Characterizing human stem cell-derived sensory neurons at the single-cell level reveals their ion channel expression and utility in pain research. Mol Ther 22: 1530 – 43 (2014).

Yu J, Vodyanik MA, Smuga-Otto K, et al. Induced pluripotent stem cell lines derived from human somatic cells. Science 318: 1917-20 (2007).

Zainolabidin N, et al. Distinct activities of Tfap2A and Tfap2B in the specification of GABAergic interneurons in the developing cerebellum. Front Mol Neurosci 10: 281 (2017).

Zamponi GW. Regulation of presynaptic calcium channels by synaptic proteins. J Pharmacol Sci 92: 79 – 83 (2003).

Zamponi GW. Targeting voltage-gated calcium channels in neurological and psychiatric diseases. Nature Reviews Drug Discovery 15: 19 – 34 (2016).

Zaret KS. Pioneer transcription factors initiating gene network changes. Annu Rev Genet

54: 367 – 385 (2020).

Zhang CY, et al. An alternative splicing hypothesis for neuropathology of schizophrenia: evidence from studies on historical candidate genes and multi-omics data. Molecular Psychiatry 27: 95 – 112 (2022).

Zhang S, et al. OCT4 and PAX6 determine the dual function of SOX2 in human ESCs as a pluripotent or neural factor. Stem Cell Res Ther 10: 122 (2019).

Zhang Y, et al. Rapid single-step induction of functional neurons from human pluripotent stem cells. Neuron 78: 785 – 698 (2013).

Zhang Y, et al. The {beta} subunit of voltage-gated Ca2+ channels interacts with and regulates the activity of a novel isoform of Pax6. J Biol Chem 285: 2527 – 36 (2010).

Zhou M, et al. The CLASSY family controls tissue-specific DNA methylation patterns in Arabidopsis. Nature Communications 13: 244 (2022).

Ziffra RS, et al. Single-cell epigenomics reveals mechanisms of human corticial development. Nature 598: 205 – 213 (2021).

Ziller MJ, et al. Dissecting neural differentiation regulatory networks through epigenetic footprinting. Nature 518: 35 – 359 (2015).

Zolotarev N, et al. EBF1 is continuously required for stabilizing local chromatin accessibility in pro-B cells. Immunity 119: e2210595119 (2022).

Zuhlke RD, et al. Ca2+ channel sensitivity towards the blocker isradipine is affected by alternative splicing of the human alpha1C subunit gene. FEBS Lett 427: 220 – 4 (1998).

Zuhlke RD, Reuter H. Calcium-sensitive inactivation of L-type calcium channels depends on multiple cytoplasmic amino acid sequences of the alpha(1C) subunit. Proc Natl Acad Sci USA 95: 3287 – 3294 (1998).

Zweier C, et al. Haploinsufficiency of TCF4 causes syndromal mental retardation with intermittent hyperventilation (Pitt-Hopkins Syndrome). Am J Hum Genet. 80: 994 – 1001 (2007).