


Spring 5-15-2018

Novel Approaches to Studying the Effects of Cis-Regulatory Variants in the Central Nervous System

Susan Shen

Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

 Part of the [Genetics Commons](#), [Molecular Biology Commons](#), and the [Neuroscience and Neurobiology Commons](#)

Recommended Citation

Shen, Susan, "Novel Approaches to Studying the Effects of Cis-Regulatory Variants in the Central Nervous System" (2018). *Arts & Sciences Electronic Theses and Dissertations*. 1578.

https://openscholarship.wustl.edu/art_sci_etds/1578

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics and Genomics

Dissertation Examination Committee:

Joseph Corbo, Chair

Shiming Chen

Donald Conrad

Joseph Dougherty

Justin Fay

Ting Wang

Novel Approaches to Studying the Effects
of *Cis*-Regulatory Variants in the Central Nervous System

by

Susan Qi Shen

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2018
St. Louis, Missouri

© 2018, Susan Shen

TABLE OF CONTENTS

List of Figures and Tables.....	vi
List of Abbreviations.....	ix
Acknowledgements.....	xi
Abstract.....	xiii
Chapter 1: Introduction.....	1
1.1 The significance of <i>cis</i> -regulatory variants in biology and disease.....	3
1.2 Genomic insights into the properties of CREs.....	5
1.3 The retina as a model system for studying <i>cis</i> -regulation.....	8
1.4 Massively parallel reporter assays for functional analysis of <i>cis</i> -regulatory variants.....	11
1.5 Identification of disease-associated variants in the GWAS era.....	13
1.6 The post-GWAS era: convergence of GWAS and functional genomics.....	16
1.7 The brain as a frontier for <i>cis</i> -regulatory biology.....	17
Chapter 2: Hybrid Mice Reveal Parent-of-Origin and <i>Cis</i> - and <i>Trans</i> -Regulatory Effects in the Retina.....	22
2.1 Author contributions.....	23
2.2 Abstract.....	24
2.3 Introduction.....	25
2.4 Results.....	28
2.4.1 Strongly imprinted genes in other tissues show evidence of imprinting in the retina.....	29
2.4.2 One-third of differentially expressed genes between Cast/EiJ and C57BL/6J retinas are associated with photoreceptor CREs.....	32
2.4.3 <i>Cis</i> -regulatory effects account for the bulk of gene regulatory divergence between Cast/EiJ and C57BL/6J retinas.....	33
2.4.4 Higher frequency of variants in photoreceptor CREs correlates with differential expression.....	35
2.4.5 The Cast/EiJ genome harbors both activating and silencing <i>cis</i> -regulatory variants associated with retinal disease genes.....	37
2.4.6 The majority of isolated <i>cis</i> effects and isolated <i>trans</i> effects are tissue- specific.....	38
2.5 Discussion.....	41
2.6 Methods.....	44
2.6.1 Ethics statement.....	44
2.6.2 Animals.....	44
2.6.3 Sample collection and sequencing.....	44
2.6.4 Read alignment and quantification.....	45
2.6.5 Identification of imprinted genes.....	46

2.6.6	Mouse imprinting databases.....	46
2.6.7	Categorization of genes according to <i>cis</i> and <i>trans</i> effects.....	47
2.6.8	Calculation of weighted log fold change.....	48
2.6.9	Assignment of genes to CRX ChIP-seq peaks.....	48
2.6.10	Batch identification of variants.....	49
2.6.11	Identification of variants at individual regions.....	49
2.6.12	RetNet genes.....	49
2.6.13	DNA constructs.....	50
2.6.14	Retinal explant electroporation and quantification of promoter activity...50	
2.7	Data Access.....	51
2.8	Supporting Information.....	51
2.9	Acknowledgements.....	52
Chapter 3: Massively Parallel <i>Cis</i> -Regulatory Analysis in the Mammalian Central Nervous System.....		65
3.1	Author Contributions.....	66
3.2	Abstract.....	67
3.3	Introduction.....	68
3.4	Results.....	70
3.4.1	Identification and characterization of candidate CRE regions.....	70
3.4.2	‘Capture-and-clone’ allows synthesis of targeted <i>cis</i> -regulome libraries.....	71
3.4.3	AAV packaging and delivery preserves CRE-seq library composition....73	
3.4.4	AAV-mediated CRE-seq demonstrates tissue-specific CRE activity of DHSs <i>in vivo</i>	75
3.4.5	Parameters that predict <i>cis</i> -regulatory activity.....	77
3.4.6	Tiling of captured fragments allows for truncation mutation analysis.....80	
3.4.7	Traditional reporter assays confirm that critical bases identified by CRE-seq truncation mutation analysis are required for activity.....83	
3.5	Discussion.....	85
3.6	Methods.....	91
3.6.1	Animals.....	91
3.6.2	Reference genome.....	91
3.6.3	Identification of target tissue-specific DHS peaks.....	91
3.6.4	Capture bait library design and synthesis.....	92
3.6.5	GREAT analysis and Gene Ontology.....	92
3.6.6	Restriction enzymes and PCR reagents.....	93
3.6.7	Preparation of gDNA for capture.....	93
3.6.8	<i>Cis</i> -regulome capture and preparation for cloning.....	94
3.6.9	CRE-seq library construction.....	95
3.6.10	Paired-end sequencing for CRE-barcode correspondence.....	97

3.6.11	Analysis of paired-end sequencing for CRE-barcode correspondence....	97
3.6.12	Retinal explant electroporation and culture for CRE-seq.....	98
3.6.13	Viral production.....	98
3.6.14	Stereotactic cortical injection.....	99
3.6.15	Isolation of RNA and DNA and preparation for sequencing.....	99
3.6.16	Illumina sequencing for CRE-seq barcode abundance.....	100
3.6.17	CRE-seq data analysis.....	100
3.6.18	Histology.....	101
3.6.19	Cluster analysis of biological replicates.....	102
3.6.20	Analysis of TF motif enrichment in low vs. high-expressing DHSs.....	102
3.6.21	Receiver operating characteristic (ROC) curves.....	103
3.6.22	Expression scores for browser screenshots.....	104
3.6.23	Synthesis of individual constructs for validation.....	105
3.6.24	Validation of individual constructs by fluorescent reporter assays.....	105
3.6.25	Comparison with CapSTARR-seq.....	106
3.7	Data Access.....	106
3.8	Supplemental Tables.....	106
3.9	Acknowledgements.....	107
Chapter 4: A Candidate Causal Variant Underlying Both Higher Cognitive Performance and Increased Risk for Bipolar Disorder.....		134
4.1	Author Contributions.....	135
4.2	Abstract.....	136
4.3	Introduction.....	137
4.4	Results.....	140
4.4.1	The <i>MIR2113/POU3F2</i> locus harbors non-coding variants associated with both increased cognitive performance and increased risk for BPD.....	140
4.4.2	Identification of the candidate causal variant rs77910749, a human-specific non-coding variant that falls within a fetal brain-specific open chromatin region.....	141
4.4.3	Mouse epigenomic data suggest that LC1 is an enhancer in the developing brain and reveal that rs77910749 falls within a binding site for Pax6....	144
4.4.4	<i>In silico</i> and <i>in vitro</i> analysis demonstrate modest effects of rs77910749 on Pax6 binding.....	145
4.4.5	Transgenic reporter mice show evidence of LC1 enhancer activity in the developing central nervous system (CNS).....	146
4.4.6	CRE-seq ‘Nano’ measures subtle gain-of-function enhancer activity of rs77910749.....	147
4.4.7	<i>In vivo</i> deletion of LC1 confers region-specific changes in <i>Pou3f2</i> expression.....	149

4.4.8	The novel CpG site created by rs77910749 is methylated at a low frequency in the developing mouse brain.....	150
4.4.9	The effect of rs77910749 on chromatin accessibility in human fetal brain.....	152
4.4.10	LC1 knockout animals have essentially normal behavior.....	152
4.4.11	Humanized rs77910749 knock-in mice have defective sensory gating...153	153
4.5	Discussion.....	153
4.6	Methods.....	157
4.6.1	Animals.....	157
4.6.2	DNase-seq data.....	157
4.6.3	Calculation of linkage disequilibrium (LD).....	158
4.6.4	Analysis of primate genomes.....	158
4.6.5	Motif analysis.....	158
4.6.6	Electrophoretic mobility shift assays (EMSAs).....	159
4.6.7	Generation of transgenic reporter mice.....	160
4.6.8	LacZ staining and histology.....	160
4.6.9	CRE-seq Nano library construction.....	161
4.6.10	Mouse cerebral cortex electroporations.....	162
4.6.11	Human cerebral organoid electroporations.....	163
4.6.12	CRE-seq Nano tissue processing and data analysis.....	165
4.6.13	CRISPR-Cas mice generation.....	165
4.6.14	Allele-specific expression (ASE) analysis.....	166
4.6.15	Allele-specific methylation analysis.....	167
4.6.16	Amplicon-seq.....	168
4.6.17	Allele-specific human fetal brain DNase-seq analysis.....	168
4.6.18	Behavioral assays.....	169
4.7	Acknowledgements.....	170
Chapter 5: Conclusions and Future Directions.....		199
5.1	The utility of hybrid animals for studying <i>cis</i> -regulation and imprinting.....	201
5.2	The future of high-throughput <i>cis</i> -regulatory analysis	202
5.3	Future directions for investigating the <i>MIR2113/POU3F2</i> locus.....	204
5.4	<i>Cis</i> -regulatory biology in the era of clinical whole-genome sequencing.....	206
References.....		208
Appendix 1: Methylation in Photoreceptors During Development.....		233
Appendix 2: The Role of CTCF in the Retina.....		244
Appendix 3: High-coverage CRE-seq libraries tiling the <i>MIR2113/POU3F2</i> locus.....		249

LIST OF FIGURES AND TABLES

Chapter 1

Table 1.1:	Summary of massively parallel reporter assay approaches.....	20
Table 1.2:	Summary of functional studies that have identified the likely causal <i>cis</i> -regulatory variant underlying a GWAS signal.....	21

Chapter 2

Figure 2.1	Study design.....	53
Figure 2.2	Characterization of parent-of-origin effects in the retina.....	54
Figure 2.3	Comparison of differentially expressed and <i>cis</i> -effect genes associated with photoreceptor CREs.....	55
Figure 2.4	Classification of genes by mechanism of gene regulatory divergence.....	56
Figure 2.5	Analysis of variant density in photoreceptor CREs.....	57
Figure 2.6	<i>Cis</i> -effect genes associated with retinal disease and photoreceptor CREs.....	58
Figure 2.7	Comparison of <i>cis</i> effects and <i>trans</i> effects between liver and retina.....	60
Table 2.1:	Agreement between F0 biological replicates.....	62
Table 2.2:	Agreement between F1 biological replicates.....	63
Table 2.3:	Accuracy of X chromosomal read mapping in F1 samples.....	64

Chapter 3

Figure 3.1	‘Capture-and-clone’ allows synthesis of CRE-seq libraries with long CREs.....	108
Figure 3.2	Tiling of captured fragments across target regions.....	109
Figure 3.3	Delivery of capture CRE-seq library into mouse retina <i>ex vivo</i> and cerebral cortex <i>in vivo</i>	111
Figure 3.4	Tissue-specific <i>cis</i> -regulatory activity of DHSs.....	113
Figure 3.5	Parameters that predict CRE activity.....	114
Figure 3.6	Truncation mutation analysis by CRE-seq.....	116
Figure 3.7	Validation of individual loci by fluorescence reporter assays.....	118
Figure 3.S1	Distribution of 4,000 target DHS regions.....	120
Figure 3.S2	Distribution of overlap of captured fragments with target DHS regions.....	122
Figure 3.S3	Co-expression of the library and cellular markers.....	123
Figure 3.S4	Comparison of biological replicates.....	124
Figure 3.S5	CRE activity, DNase-seq signal, GC content, and phylogenetic conservation of assayed DHSs in a 1 kb centered window.....	125
Figure 3.S6	Length of CRE fragments vs. expression.....	127
Figure 3.S7	Distance to nearest TSS vs. expression.....	128
Figure 3.S8	Additional examples of truncation mutation analysis by CRE-seq.....	130
Figure 3.S9	Comparison between enhancer activity of short synthesized CREs and autonomous activity of corresponding captured CRE fragments in the retina...	132

Chapter 4

Figure 4.1	Prioritization of candidate variants at 6q16.1 associated with higher educational attainment, increased cognitive performance, and risk for bipolar disorder.....	171
Figure 4.2	Epigenomic landscape around the orthologous LC1 region in mouse.....	173
Figure 4.3	<i>In silico</i> and <i>in vitro</i> analysis Pax6 binding.....	175
Figure 4.4	Transgenic reporter mice show evidence of LC1 activity in the developing CNS.....	176
Figure 4.5	The variant rs77910749 causes a subtle increase in enhancer activity in developing mouse brain and human cerebral organoids.....	177
Figure 4.6	The effect of LC1 deletion on <i>Pou3f2</i> expression is region-specific.....	179
Figure 4.7	Allele-specific methylation analysis of LC1.....	181
Figure 4.8	Human fetal brain allele-specific DNase-seq analysis.....	183
Figure 4.9	Prepulse inhibition (PPI) is defective in ‘humanized’ rs779710749 knock-in mice.....	184
Figure 4.S1	Phylogenetic conservation of rs13208578 and rs77910749.....	185
Figure 4.S2	Identification of a derived haplotype through construction of a human phylogenetic tree.....	187
Figure 4.S3	Global distribution of rs17814604 and rs77910749 frequencies.....	188
Figure 4.S4	Absence of rs77910749 from non-human primate genomes.....	190
Figure 4.S5	LC1 falls within a conserved topologically associating domain (TAD).....	191
Figure 4.S6	Antibody staining of cerebral organoids.....	192
Figure 4.S7	The methylation landscape of LC1 to LC5 in human primary tissues and cultured cells.....	193
Table 4.1.	Measures of LD among lead SNPs in GWAS studies of educational attainment, cognitive ability, and BPD.....	194
Table 4.2.	Oligonucleotides used in this study.....	195
Table 4.3.	Allele-specific fetal brain DNase-seq analysis.....	197

Appendices

Figure A1.1	The distribution of 5mC and 5hmC in mouse rod photoreceptors during development reflects nuclear architecture.....	238
Figure A1.2	5mC and 5hmC distributions in models of rod-to-cone transdifferentiation.....	240
Figure A1.3	Bisulfite analysis of <i>Rho</i> and <i>Opn1sw</i> promoters in wild-type and <i>Nrl</i> knockout retinas over development.....	242
Figure A1.4	FACS-sorted photoreceptors and bipolar cells reveal cell type-specific methylation patterns at the <i>Rho</i> promoter.....	243
Figure A2.1	Deletion of <i>CTCF</i> in the mouse neural retina results in retinal degeneration.....	247
Figure A2.2	Expression of cellular markers in <i>CTCF</i> knockout retinas.....	248
Figure A3.1	The 100 target regions in the <i>MIR2113/POU3F2</i> locus for the PCR CRE-seq library.....	252

Figure A3.2	Distribution of product lengths in the PCR library and coverage of target DHSs in the <i>MIR2113/POU3F2</i> locus.....	253
Figure A3.3	A BAC library tiling 440 kb of the <i>MIR2113/POU3F2</i> locus at 40X coverage.....	254
Figure A3.4	Distribution of fragment lengths in the BAC library and coverage of the <i>MIR2113/POU3F2</i> locus.....	255

LIST OF ABBREVIATIONS

3C	chromosome conformation capture
AAV	adeno-associated virus
AEI	allelic expression imbalance
AMD	age-related macular degeneration
AUC	area under the curve
BMI	body mass index
BPD	bipolar disorder
CBR	CRX-bound region
ChIP	chromatin immunoprecipitation
CNS	central nervous system
CRE	<i>cis</i> -regulatory element
DAPI	4',6-diamidino-2-phenylindole
DE	differentially expressed
DHS	DNase I hypersensitive site
DNMT	DNA methyltransferase
eQTL	expression quantitative trait locus
EMSA	electrophoretic mobility shift assay
ESC	embryonic stem cell
FACS	fluorescence-activated cell sorting
fl	flox allele
FPKM	fragments per kb of transcripts per million mapped reads
GCL	ganglion cell layer
gDNA	genomic DNA
GFP	green fluorescent protein
GO	Gene Ontology
GRN	gene regulatory network
GS	glutamine synthetase
GWAS	genome-wide association study
H&E	hematoxylin and eosin
IBD	inflammatory bowel disease
INL	inner nuclear layer
iPSC	induced pluripotent stem cell
ITR	inverted terminal repeat
KO	knockout
LD	linkage disequilibrium
MAF	minor allele frequency
MPRA	massively parallel reporter assay
NBL	neuroblast layer

NGS	next-generation sequencing
ONL	outer nuclear layer
OR	odds ratio
PCA	principal component analysis
PPI	prepulse inhibition
ROC	receiver operating characteristic
SD	standard deviation
SEM	standard error of the mean
SLE	systemic lupus erythematosus
SNP	single nucleotide polymorphism
TAD	topologically associating domain
TF	transcription factor
TSS	transcription start site
WGS	whole-genome sequencing
WT	wild-type

ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to my advisor, Joseph Corbo, whose scientific curiosity and intellectual drive inspire me every day. Thank you for your unwavering support, openness to new ideas, and sense of humor. Thank you for training me to be a rigorous scientist, for challenging me to think outside the box, and for exemplifying the role of the physician-scientist. Most of all, thank you for all of the fun, thought-provoking conversations that I hope will continue for many years to come.

Next, I would like to thank my past and present coworkers (Stacy Donovan, Jennifer Enright, Andrew Hughes, Jeongsook Kim-Han, Karen Lawrence, Cynthia Montana, Daniel Murphy, Connie Myers, Matthew Toomey, and Natecia Williams), who make coming into the lab an absolute joy. It has been a true pleasure and honor to work alongside them, learn from them, and grow together. In particular, Cynthia Montana generously shared innumerable protocols, reagents, and samples. Andrew Hughes spent countless hours patiently explaining computational and mathematical concepts. Special thanks to Connie Myers, who has guided me through this journey with patience, clarity, and wit.

I am grateful to my thesis committee (Shiming Chen, Donald Conrad, Joseph Dougherty, Justin Fay, and Ting Wang) for their thoughtful input and sage advice, and to Justin Fay for serving as chair prior to my defense. Thanks to my collaborators (Leah Byrne, John Flannery, Omer Gokcumen, Vladimir Kefalov, Ernest Turro, and Yunlu Sawyer Xue) for being so generous with their time and energy. Thanks to Jamie Kwasnieski, Ilaria Mogno, Michael A. White, and the laboratory of Barak Cohen for sharing CRE-seq protocols and ideas.

Thanks to the Animal Behavior Core (David Wozniak), Center for Genome Sciences and Systems Biology (Jessica Hoisington-Lopez), Department of Ophthalmology (Belinda

McMahan), Genome Engineering and iPSC Center (Shondra Miller), Genome Technology Access Center, Hope Center Animal Surgery Core (Ronald Perez), Hope Center Viral Vectors Core (Mingjie Li), Micro-injection Core (J. Michael White), Mouse Genetics Core (Mia Wallace), and Protein and Nucleic Acid Chemistry Laboratory (Misty Veschak) for their invaluable services.

I am grateful to Tim Schedl and James Skeath, co-directors of the Molecular Genetics and Genomics Program, for their mentorship and support. Furthermore, I am indebted to Wayne Yokoyama and the Washington University Medical Scientist Training Program (MSTP) for investing in me as a trainee and as a person. Thanks to my friends and colleagues in the MSTP for a memorable and enjoyable journey, and to Shuyi Ma for intellectual and emotional support.

I wish to acknowledge my parents, Jiehua Shen and Kejin Wang, whose footsteps I follow while simultaneously forging a path of my own. The depth of their love and sacrifice is beyond description, and I carry their dreams with me in everything I do. I also want to thank my sister Sarah Shen, who was a delightful 11-year-old when I started my PhD and a brilliant 16-year-old when I finished. I am grateful to my in-laws, Tim and Kathy Saylor, for their support. Finally, I thank my husband James Saylor, who has been my significant other ($p < 0.05$) for more than a decade, whose enjoyment of puns rivals my own, and whose presence I hope never to take for granted.

This work was supported by the National Institutes of Health (5T32EY013360).

Susan Shen

Washington University in St. Louis

May 2018

ABSTRACT OF THE DISSERTATION

Novel Approaches to Studying the Effects
of *Cis*-Regulatory Variants in the Central Nervous System

by

Susan Qi Shen

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2018

Professor Joseph C. Corbo, Chair

For decades, studies of the genetic basis of disease have focused on rare coding mutations that disrupt protein function, leading to the identification of hundreds of genes underlying Mendelian diseases. However, many complex diseases are non-Mendelian, and less than 2% of the genome is coding. It is now clear that non-coding variants contribute to disease susceptibility, but the precise underlying mechanisms are generally unknown. *Cis*-regulatory elements (CREs) are transcription factor (TF)-bound genomic regions that regulate gene expression, and variants within CREs can therefore modify gene expression. The putative locations of CREs in a variety of cell types have been identified through genome-wide assays of TF binding and epigenomic signatures, providing a starting point for probing the effects of *cis*-regulatory variants. Unlike coding mutations, which can be interpreted based on the genetic code, the functional consequence of any given *cis*-regulatory variant is difficult to predict even at the molecular level. Therefore, a major bottleneck lies in interpreting the functional significance of these variants.

In the present work, I study the effects of *cis*-regulatory variants in the central nervous system (CNS), specifically in retina and brain. The retina is composed of well-characterized neuronal cell types and an extensively studied transcriptional network, while the brain is the center of human cognition and a target of devastating neuropsychiatric diseases. First, I take advantage of the genetic diversity between two distantly related mouse strains to describe the relationship between *cis*-regulatory variants and differences in retinal gene expression. I identify *cis*- and *trans*-regulatory effects, as well as parent-of-origin effects. Second, I develop a new technology based on an existing massively parallel reporter assay, CRE-seq, to enable the functional study of long CREs in the CNS *in vivo* for the first time. I demonstrate the ability of this approach to measure tissue-specific *cis*-regulatory activity in the brain and to pinpoint DNA bases critical for activity. Finally, I conduct a detailed mechanistic study of a non-coding region containing variants associated with both human cognitive performance and bipolar disorder. This last study illustrates the complexities and challenges of establishing the causal role of non-coding variants in disease.

CHAPTER 1: INTRODUCTION

The following chapter has been adapted from my written qualifying examination, thesis proposal, and grant proposals. The contents of this chapter are unpublished.

“I love gene regulation. I love the process of transcription so much that I regard RNA as an unfortunate by-product of an otherwise elegant process!”

-Michael S. Levine (Levine and Vicente 2015)

“Mientras nuestro cerebro sea un arcano, el Universo, reflejo de su estructura, será también un misterio.” (As long as our brain is a mystery, the universe, a reflection of its structure, will also be one.)

-Santiago Ramón y Cajal (Ramón y Cajal 1922)

1.1 The significance of *cis*-regulatory variants in biology and disease

A fundamental goal of genetics is to understand the phenotypic consequences of specific mutations. For coding mutations, the immediate biochemical consequence of a mutation can be deduced from the DNA sequence alone. From there, the impact on cellular function and organismal phenotype can be investigated. This type of approach has revealed hundreds of genes involved in Mendelian diseases (Hamosh et al. 2005). Often, the pathogenicity of a coding mutation can be predicted based on known structural properties and biological functions of the protein and/or the degree of phylogenetic conservation. However, for non-coding mutations, even the biochemical consequences are unclear, and phylogenetic conservation is an imperfect indicator of functionality (Ng and Henikoff 2006; Cooper and Shendure 2011). Given that over 98% of the genome is non-coding, understanding the impact of non-coding variants is a major challenge. In particular, variants within *cis*-regulatory elements (CREs, e.g., enhancers and promoters) may alter the expression of genes relevant to disease.

CREs are short stretches of genomic DNA that regulate the timing, location, and levels of expression of the gene that they control. They are generally non-coding, although they can overlap coding exons (Mercer et al. 2013; Stergachis et al. 2013). CREs are typically hundreds of base pairs in length, and they are often located thousands of bases away from their target genes (Kulaeva et al. 2012). CREs are the primary determinants of gene expression during development, with cellular environment and epigenetic factors playing secondary roles (Levine and Davidson 2005). By recruiting TFs, CREs allow for fine-tuning of gene expression and serve as important substrates for phenotypic diversity between individuals and between species (Wray 2007; Wittkopp and Kalay 2012; Heinz et al. 2015).

The first detailed mechanism of gene regulation was elucidated for the *lac* operon in *E.*

coli in 1961 by Jacob and Monod (Jacob and Monod 1961). Ten years later, Britten and Davidson speculated that regulatory variants were crucial for phenotypic evolution in eukaryotes (Britten and Davidson 1971). Soon thereafter, King and Wilson suggested that chimpanzees and humans were too similar at the macromolecular level—nucleic acid and protein—to account for inter-species phenotypic differences (King and Wilson 1975). They postulated instead that regulatory variants might account for organismal-level differences, lamenting, “Biologists are still a long way from understanding gene regulation in mammals.” Although much progress has been made in the past decades, the *cis*-regulatory grammar of mammalian cells remains one of the greatest unsolved problems in biology. Furthermore, *cis*-regulatory variants are increasingly recognized as significant contributors to disease.

To illustrate the importance of regulatory variations for both evolution and human health, consider the following examples. Certain single nucleotide polymorphisms (SNPs) upstream of *lactase* (*LCT*) enhance transcription of the gene, allowing for the persistence of the lactase enzyme and the ability to digest milk as adults (Enattah et al. 2002; Olds and Sibley 2003). These SNPs have been under strong positive selection in populations that consume milk into adulthood. *Cis*-regulatory SNPs that decrease gene expression can also confer a selective advantage. The Duffy antigen chemokine receptor (DARC) is a protein required for erythrocyte invasion by certain malarial parasites. A single SNP disrupts binding of GATA1 to the DARC promoter, abolishing DARC expression in erythroid tissues and thereby conferring malarial resistance (Tournamille et al. 1995).

On the other hand, *cis*-regulatory mutations can also cause harm. One such instance is seen in a subset of patients with α -thalassemia: a SNP upstream of the α -globin gene cluster creates a novel promoter that competes with the endogenous promoter, thereby decreasing the

expression of the α -globin gene (De Gobbi et al. 2006). In retinal biology, the importance of CREs became apparent in the study of cone opsins. On the human X chromosome, a locus control region (LCR, a type of CRE) lies upstream of the red and green opsin genes, which are arranged in a tandem array. The LCR is thought to randomly associate with one of the two opsin gene promoters, thereby generating the two alternative cell types (red or green cones) in the retina (Smallwood et al. 2003). Loss-of-function mutations in this LCR causes blue cone monochromacy, a rare condition in which expression of both red and green cone opsin is lost (Nathans et al. 1989).

Distal-acting elements also have critical roles in brain development, as exemplified by *cis*-regulatory mutations that disrupt expression of *SATB2*, a TF important for skeletal development and neuronal specification in the cerebral cortex (Dobрева et al. 2006). For years, coding mutations in *SATB2* were known to underlie a syndrome characterized by craniofacial abnormalities and intellectual disability. More recently, *cis*-regulatory mutations that disrupt *SATB2* expression have also been found to cause this syndrome (Leoyklang et al. 2007; Docker et al. 2014; Rainger et al. 2014; Zarate et al. 2015). These and dozens of other examples highlight the role of CREs in both normal physiology and disease pathogenesis.

1.2 Genomic insights into the properties of CREs

Although studies of individual loci have provided valuable insights into the roles of CREs, in order to fully understand the *cis*-regulatory logic of mammalian cell types, a comprehensive approach is needed. Recent advances in next-generation sequencing (NGS) have enabled large-scale efforts to study DNA in a systematic, genome-wide fashion (Mardis 2011). The ENCODE Project and the NIH Roadmap Epigenomics Consortium have generated an

unprecedented amount of data, ushering in a new era of data-driven biology (The ENCODE Project Consortium 2012; Roadmap Epigenomics et al. 2015). These projects have sought to annotate regulatory regions in a variety of mouse and human cell lines and primary tissues, using a combination of techniques, namely: (1) ChIP-seq for histone marks and TFs, (2) DNase-seq and FAIRE-seq for identifying regions of open chromatin, (3) chromosome conformation capture (3C)-based techniques to examine chromatin looping, and (4) methylation analysis (Appendix 1). These and related studies confirm earlier, smaller studies and also offer new insights, as highlighted below.

Promoters are perhaps the most widely studied type of CRE across all fields of biology. By definition, they are located directly upstream of their target gene. This ease of promoter-gene mapping likely contributes to the observation that promoter activity correlates well with target gene expression. Promoters come in two main varieties: (1) broad-type, CpG-rich promoters that are often associated with housekeeping genes, and (2) narrow-type, TATA box-containing promoters that tend to be associated with highly expressed, tissue-specific genes (Sandelin et al. 2007). Recent studies indicate that promoters and enhancers share many architectural features and functional properties (Kim and Shiekhattar 2015). However, a detailed understanding of how promoter-enhancer compatibility is established is still lacking (van Arensbergen et al. 2014).

Regardless of the type of CRE, transcriptional potential is presumably encoded by the clusters of TF binding sites (TFBS's) that recruit the binding of various TFs. Individual TFs preferentially bind to certain sequence 'motifs,' stretches of typically ~6-12 bp of DNA. Multiple motifs of different affinities, orientations, and relative spacing are thought to act in a combinatorial fashion. Interestingly, TFs are able to recognize and selectively bind certain motifs in the genome, while avoiding other similar motifs in the genome. This suggests the existence of

additional properties within the bound regions that confer added functionality; for instance, the GC content of the region (White et al. 2013; Kwasnieski et al. 2014). To dissect the grammatical rules that govern TF motifs, it is necessary to systematically study the relationship between CRE sequence and CRE activity.

In addition to the complexities of TF binding within a single CRE, the interactions between CREs and target genes add another dimension of complexity. As revealed by a variety of 3C-based methodologies, the physical landscape of gene regulation is highly complex. Many physical looping interactions between a CRE and a target gene occur over a considerable distance (Sanyal et al. 2012). Furthermore, the notion that a CRE has a single target gene is overly simplistic: on average, each TSS interacts with multiple CREs, and a given CRE interacts with multiple TSS's (Thurman et al. 2012). Moreover, even promoters can physically interact and serve as enhancers for each other (Li et al. 2012). Despite all of this seemingly chaotic crosstalk, there is structure and order: studies using Hi-C (another 3C-based approach) have found that topologically associating domains (TADs) are highly conserved not only across cell types but also between species, although subdomains are more specific, presumably due to the action of cell type-specific CREs (Dixon et al. 2016). TADs are thought to be established at least in part by CTCF, a 'master weaver' of 3D genome architecture (Phillips and Corces 2009), but how CTCF establishes the chromatin states of specific cell types is unclear (Appendix 2).

The key question is which of the interactions between CREs and their target genes are physiologically relevant in the context of a particular human disease, and which *cis*-regulatory variants disrupt these interactions. The challenge lies in identifying the disease-relevant tissue and developmental stage. By intersecting genotypic, epigenomic, and transcriptomic information, and by applying powerful machine learning approaches (Libbrecht and Noble

2015), the emergent field of functional genomics has the potential to bioinformatically predict the effect of *cis*-regulatory variants on gene expression. Given its descriptive nature, however, functional genomics is a hypothesis-generating approach that requires alternate means to demonstrate causality.

1.3 The retina as a model system for studying *cis*-regulation

To decode the *cis*-regulatory logic of the mammalian genome, a physiologically relevant system is needed that is genetically tractable as well as amenable to functional testing, but that also harbors the complexities of mammalian gene regulation. Retinal photoreceptors meet these criteria and provide an excellent model system for studying *cis*-regulation. The neural retina is a part of the CNS and is composed of >60 cell types that fall into seven major classes: rod and cone photoreceptors, bipolar cells, amacrine cells, horizontal cells, ganglion cells, and Müller glia (Masland 2012). All of these cell classes have been extensively studied, both with regards to their normal roles in vision, as well as their roles in retinal disease.

Among the retinal cell classes, photoreceptors are by far the most abundant, constituting ~80% of retinal cells in the mouse (Jeon et al. 1998). Moreover, they are arguably the most disease-relevant. Photoreceptors are uniquely susceptible to both Mendelian diseases and complex diseases such as AMD, and nearly 300 retinal disease genes have been identified (RetNet, <http://www.sph.uth.tmc.edu/RetNet/>). Photoreceptor fate specification has been well-studied at the level of gene regulatory networks (GRNs). Although the catalogue of relevant TFs is still incomplete, a hierarchy of TFs is known.

Early in development, OTX2 (orthodenticle homeobox 2) triggers the formation of photoreceptor precursors and turns on another Otx gene family member, CRX (cone-rod

homeobox) (Nishida et al. 2003). As a master regulator of photoreceptor differentiation, CRX activates a large number of downstream photoreceptor genes (Chen et al. 1997; Furukawa et al. 1997; Hsiau et al. 2007). For instance, in conjunction with OTX2 and the ROR β (RAR-related orphan receptor β), CRX activates NRL (neural retina leucine zipper), a key rod TF that activates NR2E3 (Oh et al. 2008; Kautzmann et al. 2011; Montana et al. 2011a). Cone GRNs are not as well understood as rod GRNs, but it is known that TR β 2 (thyroid hormone receptor β 2) regulates the fate decision between the two mouse cone types, blue cones and red/green cones (Roberts et al. 2006).

The transcriptional regulation of photoreceptors has been studied in detail not only in the context of individual TFs, but also on the scale of genome-wide gene expression profiles. For instance, in the retina of the *Nrl*^{-/-} mouse, rods are converted *en masse* into cones. Comparison of *Nrl*^{-/-} retinas to wild-type retinas has enabled identification of cone-enriched and rod-enriched genes through microarray and RNA-seq studies (Corbo et al. 2007; Brooks et al. 2011). While gene expression studies have been valuable for identifying photoreceptor genes, they are particularly informative when combined with ChIP-seq, which profiles the genome-wide occupancy of a TF. Together, ChIP-seq and RNA-seq provide insights about direct and indirect connections within GRNs.

ChIP-seq studies have been conducted in the mouse retina for several photoreceptor TFs, including CRX (Corbo et al. 2010), NRL (Hao et al. 2012), and MEF2D (Andzelm et al. 2015). Several principles emerge from these studies: first, a large fraction of the ChIP-seq peaks are shared among these TFs, reflecting the role of combinatorial inputs in gene regulation. In particular, CRX appears to recruit other TFs and may act as a ‘pioneer factor’ in this regard (Zaret and Carroll 2011). Second, many of the ChIP-seq peaks are *bona fide* CREs that drive

expression (as autonomous elements or as enhancers) in photoreceptors. Third, the binding preferences of TFs, as assessed by motif enrichments within ChIP-seq peaks, agree well with *in vitro* measurements of binding affinity. Fourth, the relationship between TF binding and gene expression is highly complex and can be influenced by interactions at many levels: multiple TFBS's of different affinities, orientations, and relative spacing within a single CRE; multiple TFs cooperating or competing for a given CRE; multiple CREs regulating a given gene; and lastly, multiple negative and positive feedback loops at the GRN level.

In addition to ChIP-seq of photoreceptor TFs, DNase-seq data on mouse retinas at multiple developmental time points (postnatal day 1, day 7, and week 8) have recently become available. By profiling regions of open chromatin, DNase-seq identifies essentially all putative regulatory regions (e.g., enhancers, silencers, promoters, insulators, and LCRs) regardless of the specific TFs bound. Thus, the mouse retina offers the advantage of having comprehensive CRE maps, with data about temporal dynamics (Wilken 2015). Moreover, a newer chromatin accessibility assay, ATAC-seq, provides similar data as DNase-seq but requires far fewer cells, opening the door for not only stage-specific but also cell type-specific profiling (Buenrostro et al. 2013).

Even with a comprehensive CRE map, the mouse retina would not be a powerful system for studying *cis*-regulation if it were not experimentally tractable. Fortunately, the retina is highly amenable to functional testing (Matsuda and Cepko 2008): plasmids can be introduced into explanted retinas via *ex vivo* electroporations, and the explanted retinas can then be grown in culture (Montana et al. 2011b). Alternatively, plasmids can be injected into the eyes of mice and electroporated *in vivo* (de Melo and Blackshaw 2011). The electroporation efficiency of neonatal mouse retinas is high, especially for photoreceptors, rendering these cells particularly suitable for

cis-regulatory analysis.

With a relatively well-defined *cis*-regulatory landscape, and with experimental tools available for functional testing, the mouse retina is an ideal system for investigating the effects of *cis*-regulatory variation. In Chapter 2, I analyze the genome-wide effects of *cis*-regulatory variants on gene expression in the mouse retina.

1.4 Massively parallel reporter assays for functional analysis of *cis*-regulatory variants

The functional effects of *cis*-regulatory variants on transcriptional activity can be experimentally tested with reporter constructs. Plasmid reporters have the advantage of isolating the effects of DNA sequence on transcriptional activity, independent of genomic context. Typically, the CRE of interest is cloned upstream of a promoter and reporter gene (e.g., LacZ, fluorescent protein or luciferase) and then transfected into cultured cells, primary tissues, or even living organisms (Rosenthal 1987; Vesuna et al. 2005). The level of reporter activity serves as a quantitative measure of transcriptional activity. Although theoretically feasible on a genome-wide scale, individually synthesizing and testing CRE plasmid constructs is laborious, costly, and time-consuming.

Recent studies have shown that the challenges of one-at-a-time CRE-reporter analysis can be overcome by engineering massively parallel plasmid reporter assays (MPRAs), which enable efficient large-scale functional testing of *cis*-regulatory variants. The first MPRA was developed in 2009, in which a large library of DNA oligos containing promoter sequences and 3' barcodes were synthesized on oligonucleotide arrays and then transcribed *in vitro* (Patwardhan et al. 2009). The resulting barcoded RNA molecules (i.e., the output of the experiment) were reverse-transcribed into cDNA and sequenced. At the same time, the original DNA oligos (i.e., the input

of the experiment) were also sequenced. The number of barcoded cDNA sequence reads, normalized to the number of barcoded DNA sequence reads, served as a quantitative measure of CRE activity. Using this method, the authors conducted synthetic saturation mutagenesis in three bacteriophage promoters and three mammalian core promoters as a proof-of-principle. They were able to quantify the effects of mutations in known TFBS's, and perhaps more importantly, to identify sites outside of known TFBS's that appeared important for CRE activity.

In 2012, the same group used a similar approach to conduct saturation mutagenesis of three mammalian enhancers *in vivo* (Patwardhan et al. 2012). Instead of *in vitro* transcription of oligos, they constructed plasmids that were introduced into mouse liver via tail vein injection. Most mutations had little or modest effect, but among those with larger effects, many affected conserved binding motifs for known liver-specific tissue factors (Patwardhan et al. 2012). Another group independently developed an MPRA to test the effects of enhancer variants in human cell lines, and they also found mutations in known TFBS's that caused decreased CRE activity (Melnikov et al. 2012).

Around the same time, the Corbo lab and the lab of Barak Cohen collaborated to develop CRE-seq, an MPRA in which barcoded plasmid reporter constructs are introduced into living tissue by electroporation. CRE-seq was used to conduct saturation mutagenesis on a portion of the promoter of *Rhodopsin (Rho)*, a highly expressed gene in rod photoreceptors. This study analyzed ~1000 variants of the *Rho* promoter, including all SNPs within a central 52 bp region, as well as a large number of double-mutant constructs. Unexpectedly, 86% of all single mutations caused a change in CRE activity, even at positions without any known TFBS's. These bases may lie within novel TFBS's, or they may correspond to regions of DNA that do not directly bind to TFs but nonetheless affect CRE activity. Surprisingly, double mutants often

showed unpredictable, non-additive effects on CRE activity (Kwasnieski et al. 2012). Eventually, with a sufficiently deep understanding of *cis*-regulatory grammar, the effect of any given *cis*-regulatory variant on CRE activity should be predictable.

In the past few years, numerous MPRA have been developed by independent groups (reviewed in (Levo and Segal 2014; Shlyueva et al. 2014; White 2015)). These are summarized in Table 1.1. Some notable variations are as follows: in STARR-seq, the CRE serves as its own reporter, such that transcripts of the CRE itself are quantified (Arnold et al. 2013). While most MPRA use non-integrating plasmids, several use targeted (site-specific) integration or random integration. Some MPRA use the fluorescence intensity of individual cells (measured by FACS) instead of transcript levels as the readout. As evidenced by their diversity, MPRA have quickly gained popularity as a potential means for assaying *cis*-regulatory variants.

Thus far, most MPRA have been implemented in cell culture, but it should be possible to implement MPRA in a wide array of disease-relevant tissues. Such methods would be invaluable for understanding the *cis*-regulatory logic of mammalian cells, and for interpreting the significance of the thousands of non-coding variants found in human patients. A number of technical issues remain, such as delivery of MPRA libraries into cell types that are not amenable to transfection, and the difficulty of assaying CREs longer than the ~200 bp limit of oligonucleotide array-synthesized fragments. These issues are addressed in Chapter 3.

1.5 Identification of disease-associated variants in the GWAS era

Even prior to the emergence of high-throughput sequencing technologies, efforts were underway to systemically identify associations between genotype and phenotype through genome-wide association studies (GWAS's). Early GWAS's typically included a few hundred

cases (i.e., individuals with the phenotype of interest) and controls (Visscher et al. 2012). The genotypes of the individuals were determined by SNP arrays, which probed for thousands of known, common (>1% MAF) human SNPs across the genome. Statistical tests were then implemented to identify significant associations between allele frequency and phenotype. The phenotype of interest is typically presence or absence of a disease, but it can also be a continuous, quantitative trait such as height or blood pressure.

SNP arrays serve as a cheaper alternative to whole-genome sequencing (WGS), but because they provide incomplete genotypic information, genotypes are inferred based on knowledge of linkage disequilibrium (LD). LD refers to the tendency for sequence variants to be inherited together and is influenced by recombination rates and other factors (Ardlie et al. 2002; Slatkin 2008). For instance, SNPs in close proximity tend to be in high LD, as measured by an r^2 value near 1. Nowadays, GWAS's often include thousands or even tens of thousands of cases and controls. SNP arrays are still used for genotyping, although this may change in the near future as the cost of WGS continues to decline (Hayden 2014). With improved knowledge of LD architecture, more sophisticated statistical tools, fine-mapping and conditional analysis strategies, and larger sample sizes, GWAS's are increasingly able to detect even weak signals at loci of small effect (Visscher et al. 2012; Yang et al. 2012; Spain and Barrett 2015).

A fundamental assumption underlying GWAS's is that common variants (detectable by SNP arrays) underlie complex traits. At one extreme, a rare variant with a large effect size would cause complete penetrance, resulting in a Mendelian inheritance pattern with the genotype fully predicting the phenotype (Schork et al. 2009). Even for Mendelian traits, however, genetic background and environmental factors modify the phenotype (Dipple and McCabe 2000). At the other extreme, a common variant with a vanishingly small effect size would contribute to disease

susceptibility, albeit almost imperceptibly (Visscher et al. 2012; Loh et al. 2015). Evidence is accumulating that both classes of variants are important (Gibson 2011; Auer and Lettre 2015).

To date, GWAS's have identified thousands of loci associated with hundreds of complex traits, ranging from autoimmune disorders to psychiatric disease (Welter et al. 2014). The vast majority of observed effect sizes are small, with odds ratios rarely above 1.5. One of the few exceptions is age-related macular degeneration (AMD), a common cause of blindness that had previously been assumed to originate in the retina or RPE. In 2005, a landmark GWAS of 96 AMD patients and 50 controls, using ~100,000 SNP markers (paltry numbers by today's standards), identified a risk allele in an intron of *CFH*, a gene encoding a component of the complement cascade (Klein et al. 2005). This allele was associated with ~5-fold higher risk for AMD. Follow-up studies not only confirmed this GWAS result, but also confirmed the central role of the immune system in AMD pathogenesis (Black and Clark 2016).

The success of GWAS for AMD illustrates the potential for GWAS's to reveal novel disease pathways. The number of GWAS's is now so large that meta-analysis of GWAS's has become possible (Evangelou and Ioannidis 2013). For instance, GWAS's of neuropsychiatric disorders have revealed both shared and distinct genetic contributions among bipolar disorder (BPD), major depressive disorder (MDD), and schizophrenia (SCZ), with neuronal pathways and immune pathways playing prominent roles (Cross-Disorder Group of the Psychiatric Genomics et al. 2013; Network and Pathway Analysis Subgroup of Psychiatric Genomics 2015). Another study of GWAS's for 42 traits not only identified genetic associations between seemingly unrelated traits, such as SCZ and inflammatory bowel disease (IBD), but also inferred the causal relationship between associated traits using statistical approaches (Pickrell et al. 2016). However, GWAS's are ultimately descriptive in nature: the lead GWAS SNP at a locus (that is, the most

statistically significant SNP) is not necessarily the ‘causal variant’ that contributes directly to disease pathogenesis. Instead, the lead SNP may serve simply as a tag for the underlying causal variant, which may not have been directly genotyped.

1.6 The post-GWAS era: convergence of GWAS and functional genomics

In recent years, human statistical geneticists and functional genomicists have converged upon the realization that the vast majority of GWAS hits are non-coding, suggesting etiologic roles for underlying causal *cis*-regulatory variants (Maurano et al. 2012a; Schaub et al. 2012). In particular, disease-associated variants are often enriched within DHSs, and the tissue specificity of the DHSs may reflect disease pathogenesis. For instance, variants associated with attention deficit hyperactivity disorder (ADHD) are enriched within fetal brain DHSs (Maurano et al. 2012a). In light of the realization that many GWAS signals are likely due to *cis*-regulatory mechanisms, efforts are now routinely made to intersect GWAS hits with functional genomic annotations and eQTL data (Ward and Kellis 2012b; Edwards et al. 2013).

Thus far, however, there are relatively few examples in which the likely causal *cis*-regulatory variant underlying a non-coding GWAS signal has been identified and experimentally tested. These are summarized in Table 1.2. One particularly interesting example is the *FTO* locus, where intronic variants have been reproducibly associated with body mass index (BMI) and obesity. Two groups have independently demonstrated that this intronic region contains multiple enhancers (or possibly a single superenhancer) that regulate *IRX3* and perhaps other genes (Smemo et al. 2014; Claussnitzer et al. 2015). However, whereas one group argues that adipocytes are the relevant cell type based on experiments in primary human adipocytes (Claussnitzer et al. 2015), the other group argues that the relevant tissue is the hypothalamus

based on a mouse model (Smemo et al. 2014). The *FTO* story illustrates the challenges of demonstrating causality for *cis*-regulatory variants in disease, even when the target gene is known. In Chapter 4, I seek to identify the causal variant underlying a GWAS locus associated with neuropsychiatric phenotypes, specifically human cognition and bipolar disorder.

1.7 The brain as a frontier for *cis*-regulatory biology

The same changes that endowed human with expanded intellectual abilities may also render them susceptible to neuropsychiatric diseases (Somel et al. 2013). Many devastating neuropsychiatric disorders are multifactorial in etiology but are thought to have neurodevelopmental origins. *Cis*-regulatory variants likely contribute substantially to susceptibility for these disorders, via mechanisms that are not well understood.

Compared to the retina, the brain is many orders of magnitude more complex; for instance, there are $\sim 10^8$ vs. $\sim 10^{11}$ neurons in the respective tissues, giving rise to many orders of magnitude more synapses in the brain (Herculano-Houzel 2012; Masland 2012). The brain possesses tremendous cellular diversity of both neuronal and non-neuronal (e.g., glial) cell types. Large-scale efforts to understand the complexities of the brain at the level of gene expression, anatomy, and functional connectivity are now underway (Sunkin et al. 2013; Van Essen et al. 2013). Many of these efforts rely on approaches that were first developed in the retina, reflecting the utility of the retina as a model system for understanding the brain (London et al. 2013).

Numerous studies have sought to map the epigenomic landscape of the brain (Roadmap Epigenomics et al. 2015). In particular, the GRNs of the developing cerebral cortex have been extensively studied (Molyneaux et al. 2007; Nord et al. 2015). One-at-a-time transgenic LacZ reporter assays have also been used to functionally test a subset of candidate CREs in the

developing CNS (Nord et al. 2013; Visel et al. 2013). The cellular complexity of the brain arises from a series of highly overlapping but coordinated developmental programs involving a host of TFs, including TFs such as Pax6 that regulate neurogenesis in both the retina and the brain (Osumi et al. 2008).

Given its anatomical complexity and developmental dynamics, the study of CREs in the brain requires special consideration of the relevant cell types and developmental stage. Methods have been developed for *in vivo* and *ex vivo* electroporation of plasmid reporters into developing mouse brains (Langevin et al. 2007; Nichols et al. 2013). However, due to spatiotemporal gradients during brain development, even small differences in electroporation timing and positioning can dramatically affect what cellular populations are transfected.

Although fundamental developmental programs are conserved between mouse and human brains, there are important differences. For example, the mouse brain is lissencephalic instead of gyrencephalic and lacks the expanded outer subventricular zone of the human brain (Lui et al. 2011). As an alternative to mouse models, human iPSCs can be differentiated into neurons *in vitro* (Denham and Dottori 2011), or fibroblasts can be directly converted into neurons *in vitro* (Yoo et al. 2011). Furthermore, protocols for growing iPSC-derived cerebral organoids have recently been developed (e.g., (Lancaster et al. 2013)). However, the robustness of iPSC-based protocols and the precise properties of the derived cells remain to be fully characterized. Thus, mouse models and iPSC-based systems both have advantages and disadvantages.

With the biological complexities and technical challenges of assaying the brain, the study of *cis*-regulatory variants in the context of CNS disease remains a major frontier. In the

subsequent chapters, I describe my forays into this frontier, with the aid of the retina as a 'window' into the brain.

Table 1.1. Summary of massively parallel reporter assay approaches.

Reference	Method	Distinguishing features	Assayed system	Plasmid or integrated
(Patwardhan et al. 2009)	MPRA	First MPRA proof-of-concept	Uses <i>in vitro</i> transcription	N/A
(Nam et al. 2010; Nam and Davidson 2012)	NanoString	Delivery of library via injection into fertilized egg	Sea urchin embryos	Integrated (random)
(Melnikov et al. 2012)	MPRA	One of the first MPRAs	Human cell line	Plasmid
(Patwardhan et al. 2012)	MPRA	One of the first MPRAs	Mouse liver (hydrodynamic tail vein assay)	Plasmid
(Sharon et al. 2012)	MPRA	One of the first MPRAs; uses fluorescent readout (FACS)	Yeast	Plasmid
(Kwasnieski et al. 2012)	CRE-seq	One of the first MPRAs	Mouse retina (explant electroporation)	Plasmid
(Mogno et al. 2013)			Yeast	Integrated (site-specific)
(Akhtar et al. 2013)	TRIP	Transposase-mediated integration	Mouse ESCs	Integrated (random)
(Arnold et al. 2013)	STARR-seq	CRE serves as its own reporter	Drosophila cell lines, human cell lines	Plasmid
(Gisselbrecht et al. 2013)	enhancer-FACS-seq	Uses phiC31 integrase; fluorescent readout (FACS)	Drosophila embryos	Integrated (site-specific)
(Dickel et al. 2014)	SIF-seq	Targeted integration via homologous recombination; fluorescent readout (FACS)	Human and mouse ESCs	Integrated (site-specific)
(Murtha et al. 2014)	FIREWACH	Uses lentivirus	Mouse ESCs	Integrated (random)
(Vanhille et al. 2015)	CapSTARR-seq	Uses capture-and-clone and STARR-seq	Mouse cell lines	Plasmid
(Shen et al. 2016)*	Capture-and-clone AAV CRE-seq	Uses capture-and-clone for truncation mutation analysis; uses AAV for the first time	Mouse brain (stereotactic injection of AAV)	Non-integrated
(Nguyen et al. 2016)	AAV MPRA	Uses AAV	Primary cultured neurons	Non-integrated
(Verfaillie et al. 2016)	CHEQ-seq	Uses capture-and-clone	Human cell line	Plasmid
(Inoue et al. 2017)	lentiMPRA	Uses lentivirus	Human cell line	Non-integrated (mutant integrase) vs. integrated (random)

*See Chapter 3.

Table 1.2. Summary of functional studies that have identified the likely causal *cis*-regulatory variant underlying a GWAS signal.

Reference	Causal variant(s)	Phenotype(s)	Relevant cell type(s)	Target gene(s)	Disrupted TF motifs	Types of evidence
(Harismendy et al. 2011)	rs10811656, rs10757278	Coronary artery disease, type 2 diabetes	Vascular endothelial cells	Multiple genes	STAT1	Looping, TF binding
(Bauer et al. 2013)	rs1427407	HbF level (sickle cell disease)	Erythroblasts	BCL11A	GATA1, TAL1	Looping, TF binding, gene expression, transgenic reporter mice
(Sakurai et al. 2013)	rs3122605	SLE, IBD, type 1 diabetes	B cells, T cells, monocytes	IL-10	ELK-1	TF binding, gene expression, protein expression
(Zeron-Medina et al. 2013)	rs4590952	Testicular cancer	Various	KITLG	p53	TF binding, enhancer activity, allele-specific expression
(Guenther et al. 2014)	rs12821256	Blond hair color	Hair follicles	KITLG	LEF1	TF binding, enhancer activity, transgenic reporter mice, knock-in mice
(Fogarty et al. 2014)	rs11257655	Type 2 diabetes	Liver/pancreatic islets?	Unknown	FOXA1, FOXA2	TF binding, enhancer activity
(Kulzer et al. 2014)	rs11603334	Type 2 diabetes	Pancreatic beta cells	ARAP1	PAX6, PAX4	TF binding, enhancer activity, gene expression
(Spieler et al. 2014)	rs12469063	Restless legs syndrome	Neuronal progenitors in ganglionic eminence	Meis1	CREB1	TF binding, transgenic reporter zebrafish, transgenic reporter mice
(Visser et al. 2014)	rs12350739	Skin pigmentation	Melanocytes	BNC2	Unknown	Chromatin accessibility, enhancer activity, gene expression
(Claussnitzer et al. 2015) (see also (Smemo et al. 2014))	rs1421085	BMI	Adipocyte precursors (hypothalamus according to (Smemo et al. 2014))	IRX3, IRX5	ARID5B	Looping, TF binding, enhancer activity, gene expression, CRISPR-Cas modification of cells
(Gaulton et al. 2015)	rs10830963	Type 2 diabetes	Liver/pancreatic islets?	Unknown	NEUROD1	TF binding, enhancer activity
(Oldridge et al. 2015)	rs2168101	Neuroblastoma	Neuroblastoma tumor	LMO1	GATA3	TF binding, enhancer activity, allele-specific expression
(Soldner et al. 2016)	rs356168	Parkinson's disease	Neural precursors, neurons	SCNA	EMX2, NKX6-1	TF binding, allele-specific expression, CRISPR-Cas modification of cells

CHAPTER 2:

Hybrid Mice Reveal Parent-of-Origin and *Cis*- and *Trans*-Regulatory Effects in the Retina

2.1 AUTHOR CONTRIBUTIONS

This chapter is adapted from a published manuscript: Shen SQ¹, Turro E^{2,3}, and Corbo JC^{1*}. (2014) “Hybrid mice reveal parent-of-origin and *cis*- and *trans*-regulatory effects in the retina.” *PLoS ONE*. 9:e109382. This work was done in collaboration with Ernest Turro. Joseph Corbo and I conceived the project and designed the experiments. Ernest Turro conducted the allele-specific alignment and quantification. I conducted the experiments and other analyses. Joseph Corbo, Ernest Turro, and I wrote the manuscript.

The experimental design and analytical approaches of this project were inspired by, and rely upon, previous work (Wittkopp et al. 2004; Tirosh et al. 2009; Corbo et al. 2010; Emerson et al. 2010; McManus et al. 2010; Keane et al. 2011; Turro et al. 2011; Goncalves et al. 2012; Turro et al. 2014). Here, I identify *cis*- and *trans*-regulatory effects in the retina and subsequently map *cis*-regulatory variants onto gene expression changes. I also identify parent-of-origin effects (i.e., evidence of imprinting) in the retina for the first time.

¹Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri, United States of America

²Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom

³Department of Haematology, University of Cambridge, National Health Service Blood and Transplant, Cambridge, United Kingdom

*Corresponding author

2.2 ABSTRACT

A fundamental challenge in genomics is to map DNA sequence variants onto changes in gene expression. Gene expression is regulated by *cis*-regulatory elements (CREs, i.e., enhancers, promoters, and silencers) and the *trans* factors (e.g., transcription factors) that act upon them. A powerful approach to dissecting *cis* and *trans* effects is to compare F1 hybrids with F0 homozygotes. Using this approach and taking advantage of the high frequency of polymorphisms in wild-derived inbred Cast/EiJ mice relative to the reference strain C57BL/6J, we conducted allele-specific mRNA-seq analysis in the adult mouse retina, a disease-relevant neural tissue. We found that *cis* effects account for the bulk of gene regulatory divergence in the retina. Many CREs contained functional (i.e., activating or silencing) *cis*-regulatory variants mapping onto altered expression of genes, including genes associated with retinal disease. By comparing our retinal data with previously published liver data, we found that most of the *cis* effects identified were tissue-specific. Lastly, by comparing reciprocal F1 hybrids, we identified evidence of imprinting in the retina for the first time. Our study provides a framework and resource for mapping *cis*-regulatory variants onto changes in gene expression, and underscores the importance of studying *cis*-regulatory variants in the context of retinal disease.

2.3 INTRODUCTION

Photoreceptors mediate vision by converting light into an electrical signal, which is then processed by the inner retina and sent to the brain as visual information. Photoreceptors constitute the vast majority (>70%) of cells in the mouse retina (Young 1985), and they are prominent targets for disease: the majority of more than 200 genetic forms of retinal degeneration affect photoreceptors (SP Daiger 1998). Many of the key transcriptional regulators in photoreceptor development are known, and the transcriptomes of these cells have been profiled over normal development as well as in disease states (Corbo et al. 2007; Hsiao et al. 2007; Swaroop et al. 2010). Furthermore, the regulatory regions of mature photoreceptors in adult mouse retinas have been mapped genome-wide, based on the binding patterns of two key photoreceptor transcription factors, CRX (cone-rod homeobox) (Corbo et al. 2010) and NRL (neural retina leucine zipper) (Hao et al. 2012), as well as the patterns of ENCODE DNaseI hypersensitivity sequencing (DNase-seq) data (The ENCODE Project Consortium 2012). Photoreceptors therefore represent a disease-relevant cell type well-suited for studying the mechanisms of mammalian gene regulation.

Changes in gene expression give rise to cell-type identity, intraspecies variation, and interspecies diversity, thereby acting as the molecular underpinnings for development, disease, and evolution, respectively (Wray 2007; Wittkopp and Kalay 2012). Alterations in gene expression can arise from changes in *cis*-regulatory elements (CREs, i.e., enhancers, promoters, and silencers), or from changes in the *trans* factors (e.g., transcription factors) that interact with CREs. To distinguish between *cis* and *trans* effects, a powerful approach is to compare F1 heterozygous hybrids with F0 homozygotes. In F1 hybrids, both alleles of a gene are contained within the same nucleus and are exposed to the same set of *trans* factors. A *trans*-regulatory

difference (*trans* effect') manifests as conserved expression between the two alleles in the F1 hybrids, despite differential expression of the gene in the F0 homozygotes. In contrast, a *cis*-regulatory difference (*cis* effect') manifests as an allelic expression imbalance (AEI)—i.e., differential expression between the two alleles of a gene in the F1 hybrids, with an allelic ratio that recapitulates the ratio of gene expression levels in the F0 homozygotes. By measuring allele-specific gene expression, the relative contributions of *cis* and *trans* effects can be dissected genome-wide. AEI can also arise from parent-of-origin effects (e.g., imprinting). Importantly, by conducting reciprocal crosses, parent-of-origin effects can be identified and filtered to avoid confounding the analysis of *cis* and *trans* effects.

Prior studies utilizing the F1 hybrid study design in yeast and *Drosophila* have yielded a range of results: earlier pyrosequencing and microarray-based studies found that *cis* effects predominate (Wittkopp et al. 2004; Tirosh et al. 2009), while more recent RNA-seq studies indicate a greater role for *trans* effects (Emerson et al. 2010; McManus et al. 2010). Regardless, all studies acknowledge a high prevalence of *cis* effects. The F1 hybrid study design has been used to investigate gene regulation in one mammalian tissue thus far, the mouse liver (Goncalves et al. 2012). In that study, the authors found that *cis* and *trans* effects often act together in opposite directions, with the net effect of stabilizing gene expression.

Here, we conduct an F1 hybrid study using allele-specific mRNA-seq analysis to chart the regulatory landscape of a portion of the mature mammalian central nervous system, the adult mouse retina. We utilize two distantly related strains of mice, Cast/EiJ and C57BL/6J, whose retinas are known to exhibit phenotypic differences (Haider et al. 2008; Jelcick et al. 2011). The primary goal of our study is to dissect the contributions of *cis* and *trans* effects on gene regulation in photoreceptors. As part of our study, we identify parent-of-origin effects in the

retina, a tissue in which imprinting has not previously been studied. By re-analyzing available liver data (Goncalves et al. 2012) and comparing them to our data from the retina, we assess the degree of tissue specificity of the observed *cis*- and *trans*-regulatory effects. Furthermore, we integrate our gene expression data with knowledge about the location of CREs, thereby providing insights into the effects of *cis*-regulatory variants on gene expression.

2.4 RESULTS

The ancestors of two inbred *Mus musculus* strains, the standard reference strain C57BL/6J and the wild-derived inbred strain Cast/EiJ, diverged ~1 million years ago (Wade et al. 2002). Cast/EiJ harbors ~18 million single nucleotide polymorphisms (SNPs) and ~3 million insertions/deletions (indels) relative to C57BL/6J, involving nearly 1% of the accessible genome (Keane et al. 2011). In addition, Cast/EiJ retinas show substantial phenotypic differences, namely reduced photopic and scotopic electroretinogram amplitudes compared to C57BL/6J retinas (Haider et al. 2008; Jelcick et al. 2011). We reciprocally crossed these two strains to obtain four genotypic classes for analysis (Figure 2.1A): F0 C57BL/6J, F0 Cast/EiJ, F1 B6xCast (resulting from C57BL/6J male x Cast/EiJ female), and F1 CastxB6 (resulting from Cast/EiJ male x C57BL/6J female). For each class, we analyzed three biological replicates, each consisting of a pool of retinas.

We collected retinas from adult mice at age 8 weeks, a time point at which mouse retinal CRX ChIP-seq (Corbo et al. 2010) and ENCODE DNase-seq (The ENCODE Project Consortium 2012) were previously conducted. To control for sex-linked effects and because the X chromosome of Cast/EiJ is preferentially expressed in F1 hybrid females (Chadwick et al. 2006), we used retinas from male mice only and focused our analyses on autosomal genes. We conducted paired-end mRNA-seq and calculated gene expression for F0 samples and allele-specific expression for F1 samples by mapping reads to the C57BL/6J and Cast/EiJ transcriptomes using MMSEQ (Figure 2.1B; see Methods) (Turro et al. 2011).

We verified that biological replicates for each F0 or F1 class exhibited a high degree of agreement for gene expression or allele-specific expression estimates, respectively (Table 2.1 and Table 2.2). We also verified the accuracy of our mapping strategy by examining the X

chromosomal reads in the F1 samples. Since samples derived solely from male retinas, the X chromosomal reads should map exclusively to the maternal genome. Accordingly, X chromosomal reads for F1 B6xCast should map to Cast/EiJ, while those for F1 CastxB6 should map to C57BL/6J. In validation of our mapping strategy, we found high accuracy (>99%) of X chromosomal reads for all F1 samples (Table 2.3). Importantly, the accuracy of mapping to the X chromosome of F1 B6xCast and F1 CastxB6 samples was similar, indicating that there was no substantial read-mapping bias toward the standard reference genome, C57BL/6J, a potential confounding factor in the allele-specific quantification (Degner et al. 2009).

2.4.1 Strongly imprinted genes in other tissues show evidence of imprinting in the retina

To evaluate *cis* and *trans* effects on gene expression in the retina, we first needed to filter genes affected by parent-of-origin effects (e.g., imprinting). Genomic imprinting is an epigenetic phenomenon that causes an imbalance in allelic expression depending on whether the allele is maternally or paternally derived (Reik and Walter 2001). In the extreme case, one allele is completely silenced, rendering the locus functionally monoallelic; for this reason, many mutations in imprinted loci are associated with human disease (Falls et al. 1999). Differential methylation of alleles provides a molecular basis for imprinting, but because methylation can occur in a tissue-specific manner, a gene can be imprinted in one tissue but not another, despite being expressed in both (Prickett and Oakey 2012). Although imprinting has been extensively studied in a number of human and mouse tissues, including brain and placenta (Prickett and Oakey 2012; Xie et al. 2012; Court et al. 2014), it has not previously been studied in the retina.

By analyzing the reciprocal F1 hybrids, we identified autosomal genes that exhibited a significant maternal bias (maternally expressed, paternally silenced) or paternal bias (paternally

expressed, maternally silenced) (Figure 2.2A and 2.2B; Supporting Information S1). To determine whether these genes have been identified as imprinted in other tissues, we searched for known imprinted mouse genes in four databases (see Methods). Using a Bayesian model selection approach implemented in the MMDIFF program (Turro et al. 2014), we ranked genes in our dataset by the probability of imprinting and observed a clear enrichment of known imprinted genes among highly-ranked genes (Figure 2.2C). Among the top-ranked genes, the vast majority were well-characterized imprinted genes listed in multiple imprinting databases (see Methods) and displayed the same allelic bias as previously reported (Figure 2.2D).

We identified 75 genes as highly likely to be imprinted (Bayes factor ≥ 10). Among these, 39 genes were extremely likely to be imprinted (Bayes factor ≥ 30), of which 29/39 (74%) were known imprinted genes. In 27 out of 29 cases, the direction of parental bias that we observed was consistent with that reported in the literature. For instance, *Peg3* (paternally expressed 3) and *Meg3* (maternally expressed 3) were our 2nd and 3rd ranked imprinting genes, respectively. *Igf2* and *Igf2r* were our 30th and 34th ranked imprinted genes, respectively. *Igf2* and its receptor *Igf2r* were the first imprinted mouse genes discovered and remain among the best-characterized, with paternally expressed *Igf2* promoting growth and maternally expressed *Igf2r* inhibiting growth (Moore and Haig 1991; Wilkins and Haig 2003). Consistent with an emerging view of imprinting occurring on a spectrum rather than being an all-or-none event (Goncalves et al. 2012; Xie et al. 2012), we found varying degrees of allelic bias even for well-characterized imprinted genes, ranging from subtle (e.g., <2-fold preference for the maternal over the paternal allele of *Igf2r*) to extreme (e.g., >1000-fold preference for the maternal over the paternal allele of *Rian*).

Rtl1, also known as *Peg11*, is a gene in the *Dlk1-Dio3* imprinted cluster (da Rocha et al. 2008). In our dataset, reads mapped preferentially to the maternal allele at the *Rtl1* locus.

Previous studies in other tissues found that *Rtl1* is expressed from the paternal allele, while an antisense RNA, *anti-Rtl1*, is transcribed from the same locus on the maternal allele and gives rise to two maternally expressed microRNAs (Seitz et al. 2003; da Rocha et al. 2008). Since our RNA-seq was not strand-specific, we could not discern whether *Rtl1* or *anti-Rtl1* is maternally expressed in the adult mouse retina.

Grb10 is unique among imprinted genes in that it exhibits opposite patterns of imprinting depending on the tissue where it is expressed. In adult mice, *Grb10* is maternally expressed in some tissues, such as muscle and adipose, where it plays a role in glucose metabolism (Smith et al. 2007). However, it is paternally expressed in the brain, where it affects social behavior (Garfield et al. 2011). This tissue-specific parent-of-origin effect is associated with usage of a paternal-specific *Grb10* promoter during neural fate commitment (Sanz et al. 2008). Interestingly, in the retina, we found that *Grb10* follows the pattern of muscle and adipose tissue, with preferential expression of the maternal allele. Thus, although the retina belongs to the central nervous system, it does not follow the imprinting pattern observed in the brain for this locus.

Together, these analyses indicate that imprinting occurs in the retina, and that the pattern of imprinting is largely, but not always, concordant between the retina and the brain. Notably, the developing retina expresses the DNA methyltransferase DNMT3A, which is required for the germline methylation of imprinted loci (Kaneda et al. 2004; Nasonkin et al. 2011). Methylation analysis (e.g., bisulfite sequencing) of the retina would confirm whether the parent-of-origin effects identified here correspond to differentially methylated regions (DMRs), as methylation-independent parent-of-origin effects have also been reported (Court et al. 2014; Mott et al. 2014).

2.4.2 One-third of differentially expressed genes between Cast/EiJ and C57BL/6J retinas are associated with photoreceptor CREs

Previous microarray studies have suggested substantial gene expression differences between C57BL/6J and Cast/EiJ retinas (Jelcick et al. 2011). Thus, we surveyed differentially expressed (DE) genes between the adult male F0 Cast/EiJ and F0 C57BL/6J retinas. We identified 3,799 autosomal DE genes between the F0 samples at a false discovery rate (FDR) of 5% using DESeq (Anders and Huber 2010) (Supporting Information S2). Among these, 1,701/3,799 (45%) showed higher expression in Cast/EiJ.

CRX is a key photoreceptor transcription factor required for the expression of many rod and cone genes (Chen et al. 1997; Furukawa et al. 1999). Previous CRX ChIP-seq studies conducted in adult C57BL/6 mouse retinas demonstrated that CRX-bound regions (CBRs) demarcate both known and putative photoreceptor CREs (Corbo et al. 2010). CBRs have a propensity to cluster around genes expressed in photoreceptors, and knowledge of CBR locations has helped pinpoint novel human retinal disease genes (Langmann et al. 2010; Ozgul et al. 2011).

We used available adult mouse retinal CRX ChIP-seq data to determine whether the differentially expressed genes were CBR-associated (Corbo et al. 2010). We found that among all 34,964 autosomal genes, 6,257 (18%) had at least one CBR assigned to them. However, among the 3,799 DE genes between the two strains, 1,275 (34%) were CBR-associated, representing a significant enrichment ($P < 10^{-14}$, hypergeometric distribution). Thus, among all autosomal genes, those that were differentially expressed between Cast/EiJ and C57BL/6J were more likely to be CBR-associated.

Furthermore, differentially expressed CBR-associated genes more often had lower expression in Cast/EiJ than C57BL/6J when compared to differentially expressed non-CBR-

associated genes (Figure 2.3A). This effect was especially pronounced for genes with greater fold change between the two strains. Together, these findings suggest that Cast/EiJ overall has lower expression of photoreceptor genes than C57BL/6J, consistent with a previous microarray analysis (Jelcick et al. 2011). The physiological function of rods, which constitute >97% of the photoreceptors in the mouse retina (Jeon et al. 1998), can be measured by the a-wave of the scotopic electroretinogram (ERG). Interestingly, the gene expression differences may be reflected in the rod photoreceptor physiology of Cast/EiJ, which has a scotopic a-wave amplitude ~40-50% that of C57BL/6J, despite intact morphology (Jelcick et al. 2011; Grubb et al. 2014).

2.4.3 *Cis*-regulatory effects account for the bulk of gene regulatory divergence between Cast/EiJ and C57BL/6J retinas

Next, we determined whether gene expression divergence was attributable to *cis* effects, *trans* effects, or a combination of both. For this analysis, we examined allele-specific expression in the F1 hybrids in conjunction with gene expression in the F0 parents (Figure 2.4A; Supporting Information S3). After excluding 306 genes with an imprinting Bayes factor >3, we were able to classify 11,484 autosomal genes with high confidence (see Methods). Among these, 6,380/11,484 (56%) were best modelled as conserved (i.e., no significant difference), 3,537/11,484 (31%) as divergent due to *cis* effects, 850/11,484 (7%) as divergent due to *trans* effects, and 717/11,484 (6%) as divergent due to a combination of *cis* and *trans* effects. Therefore, *cis*-regulatory effects were the primary mechanism of gene regulatory divergence between Cast/EiJ and C57BL/6J retinas.

We then subcategorized the genes whose divergence was due to a combination of *cis*- and *trans* effects into the following classes: (1) *CIS – trans* (*cis* and *trans* effects acting in opposite

directions, with *cis* effects stronger) had 195/717 (27%) of the genes, (2) *TRANS* – *cis* (*cis* and *trans* effects acting in opposite directions, with *trans* effects stronger) had 327/717 (46%) of the genes, (3) *CIS* + *trans* (*cis* and *trans* effects acting in the same direction, with *cis* effects stronger) had 60/717 (8%) of the genes, and (4) *TRANS* + *cis* (*cis* and *trans* effects acting in the same direction, with *trans* effects stronger) had 135/717 (19%) of the genes (Figure 2.4B).

When *cis* and *trans* effects acted together in the retina, they acted in opposite directions to stabilize gene expression in the majority (522/717 or 73%) of cases, while they acted in the same direction to shift gene expression in a minority (195/717 or 27%) of cases. However, the primary mechanism of gene regulatory divergence was *cis*-regulatory effects acting with little or no contribution from *trans*-regulatory effects, accounting for 3,537/5,104 (69%) of gene regulatory divergence. This suggests that functional *cis*-acting sequence variants in the Cast/EiJ genome often drive altered gene expression.

We further examined the 3,537 *cis*-effect genes, of which 1,751/3,537 (50%) showed higher expression of the Cast/EiJ allele than the C57BL/6J allele, and of which 1,256/3,537 (36%) were CBR-associated. We found that *cis*-effect genes that were CBR-associated more often had lower Cast/EiJ allele expression than *cis*-effect genes that were not CBR-associated, for genes with higher fold change between the two alleles (Figure 2.3B). These results are consistent with the notion that the Cast/EiJ genome overall harbors many *cis*-regulatory variants whose net effect is to diminish photoreceptor gene expression.

Trans effects could arise from differential activity of transcription factors. Therefore, we inspected the rod photoreceptor transcriptional network, whose members include the transcription factors CRX, ROR β , NRL, and NR2E3 (Swaroop et al. 2010). We found that *Crx* and *Nrl* transcript levels were both conserved in the F0 and F1 retinas. *Rorb* was a solely *trans*-

effect gene, with lower expression in Cast/EiJ, suggesting that the upstream regulators of ROR β in the retina (whose identities are unknown) have altered activity in Cast/EiJ. *Nr2e3* was also a solely *trans*-effect gene, with higher expression in Cast/EiJ. Since NR2E3 is known to be regulated by CRX and NRL (Oh et al. 2008), and the mRNA levels of *Crx* and *Nrl* were unaltered, we examined whether CRX or NRL harbored coding mutations that might alter their protein activity. However, we did not find any non-synonymous mutations in *Nrl* or in the best-characterized isoform of *Crx* (Chen et al. 1997; Freund et al. 1997). Thus, we identified differential *trans*-regulation of *Rorb* and *Nr2e3* in Cast/EiJ relative to C57BL/6J, but the reasons for these *trans* effects are unknown.

2.4.4 Higher frequency of variants in photoreceptor CREs correlates with differential expression

Whereas *trans*-regulatory effects are due to the levels or activities of upstream signaling cascades and transcriptional regulators (e.g., transcription factors), *cis*-regulatory effects can arise from functional *cis*-acting variants within CREs. We undertook a survey of Cast/EiJ variants relative to C57BL/6J that fell within CBRs. First, we asked whether CBRs were depleted or enriched for Cast/EiJ variants by tabulating the number of SNPs and indels across the 2 kb region centered on CBRs. We found that the frequency of variants decreased toward the center of CBRs (Figure 2.5A). The bulk of the depletion occurred within the central 300 bp, consistent with the previous finding that phylogenetic conservation, as measured by PhastCons scores (Siepel et al. 2005), is markedly elevated within the central region of CBRs (Corbo et al. 2010). Also consistent with this result, a recent functional analysis of ~1,300 CBRs in the mouse retina demonstrated that short fragments corresponding to the central 84 bp of CBRs possess

substantial *cis*-regulatory activity (White et al. 2013). When we conducted the same analysis of variant depletion for Spret/EiJ, an inbred strain of *Mus spretus* that diverged from *Mus musculus* ~2 million years ago (Dejager et al. 2009), we obtained similar results (Figure 2.5A). Thus, CBRs are functionally constrained and have likely undergone selection in the mouse lineage, particularly in the central-most portion of the CBR.

If *cis* effects are due to altered transcriptional activity driven by *cis*-regulatory variants, we would expect to find a higher frequency of functional variants in the CREs around *cis*-effect genes compared to the *trans*-effect genes. We first observed that the proportion of genes that were CBR-associated was approximately equal across categories: 2,149/6,380 (34%) of conserved genes, 1,256/3,537 (36%) of *cis*-effect genes, 300/850 (35%) of *trans*-effect genes, and 242/717 (34%) of *cis*- and *trans*-effect genes. We then tabulated the Cast/EiJ variants (SNPs and indels) within the central 1 kb centered on the CBRs associated with each gene (Supporting Information S4). For all 10,212 CBRs, we found 86,389 variants, for a frequency of 8.46 variants per kb. When we examined the *cis*-effect genes, we found 21,174 variants in the 2,185 associated CBRs, for a frequency of 9.69 variants per kb. This was significantly higher than the variant frequency in all CBRs ($P < 10^{-14}$, hypergeometric distribution). In contrast, for the *trans*-effect genes, we found 4,068 variants in 487 CBRs, corresponding to a frequency of 8.35 variants per kb, which was not significantly different from the variant frequency in all CBRs ($P = 0.2$, hypergeometric distribution). The tendency for CBRs associated with *cis*-effect genes to have a higher frequency of variants than CBRs associated with *trans*-effect genes is also evident from the distributions of variants across individual CBRs (Figure 2.5B). Collectively, we find that CBRs associated with *cis*-effect genes are enriched for variants, whereas CBRs associated with *trans*-effect genes are not. These results suggest that CBRs contain functional *cis*-regulatory

variants that alter transcriptional activity, but future empirical testing is needed to demonstrate the causality of specific variants.

2.4.5 The Cast/EiJ genome harbors both activating and silencing *cis*-regulatory variants associated with retinal disease genes

Given that hundreds of genes can contribute to retinal disease, we asked whether any of the *cis*-effect genes were associated with human retinopathies (Supporting Information S5). We found 62 *cis*-effect genes with human orthologues that were listed in the RetNet database, an up-to-date and comprehensive compendium of retinal disease genes . Of these, 30/62 (48%) showed higher expression of the Cast/EiJ allele. Therefore, although Cast/EiJ mice have diminished rod and cone ERG responses compared to C57BL/6J, they harbor both activating and silencing *cis*-regulatory variants.

We further focused on the *cis*-effect genes associated with retinal disease that are CBR-associated (Figure 2.6A). Consistent with previous observations that CBRs are enriched around retinal disease genes (Corbo et al. 2010; Langmann et al. 2010; Ozgul et al. 2011), we found that 38/62 (61%) were CBR-associated. Of these CBR-associated genes, 20/38 (53%) showed higher expression of the Cast/EiJ allele.

One of the CBR-associated *cis*-effect genes was *Sag*, which encodes S-arrestin, a protein important for the recovery phase of the phototransduction cascade in rods (Xu et al. 1997; Song et al. 2011). Loss-of-function coding mutations in *Sag* are associated with Oguchi disease, whose clinical features include night blindness and delayed rod adaptation (Fuchs et al. 1995). We found that the Cast/EiJ allele drives ~2-fold higher *Sag* expression than the C57BL/6J allele, suggesting the presence of *cis*-regulatory variants conferring increased activity. Upon inspection

of the *Sag* locus, we identified a CRX ChIP-seq peak located in the promoter/5' UTR region and present in both CRX ChIP-seq biological replicates. This CBR corresponds to a DNaseI-hypersensitivity site (DHS) that is present at three developmental time points and is highly specific to the retina (Figure 2.6B) (The ENCODE Project Consortium 2012).

We hypothesized that *Sag* promoter variants contributed to the differential gene expression between C57BL/6J and Cast/EiJ. To test this hypothesis, we compared the activity of a 0.7 kb promoter region cloned from C57BL/6J genomic DNA ('B6 allele') or from Cast/EiJ genomic DNA ('Cast allele'). This 0.7 kb region encompassed 5 known SNPs and 1 indel (Figure 2.6B). We cloned the 0.7 kb fragment upstream of a reporter gene, DsRed, and conducted a retinal explant electroporation assay to quantify CRE activity based on fluorescence (see Methods) (Montana et al. 2011b).

Consistent with our hypothesis, we found that the Cast allele showed ~22% higher CRE activity than the B6 allele (Figure 2.6C and 2.6D; $P = 0.036$, one-tailed Wilcoxon rank-sum test). Since *Sag* had ~2-fold higher expression in Cast than B6, additional variants beyond this 0.7 kb promoter region likely contribute to the differential gene expression. Three other CBRs besides the promoter region were assigned to the *Sag* gene, containing 37 variants in the 1 kb windows centered on these CBRs (Supporting Information S4). Therefore, the higher expression of *Sag* in Cast compared to B6 likely results from variants in both the assayed region and other regions.

2.4.6 The majority of isolated *cis* effects and isolated *trans* effects are tissue-specific

To determine whether the isolated *cis* effects and isolated *trans* effects we identified were confined to the retina, we compared our data from retina with previously published data from liver (Goncalves et al. 2012) (Supporting Information S6). To ensure uniformity of analysis, we

reprocessed the previously published liver data using our analytic pipeline, beginning with raw reads. After filtering 571 possibly imprinted polymorphic autosomal genes (Bayes factor >3), we were able to classify 9,865 polymorphic autosomal genes with high confidence (Figure 2.7A and 2.7B).

We found 5,494/9,865 (56%) were best modelled as conserved, 2,371/9,865 (24%) were best modelled as divergent due to *cis* effects, 1,495/9,865 (15%) as divergent due to *trans* effects, and 505/9,865 (5%) as divergent due to a combination of *cis* and *trans* effects. For genes in the latter category, 145/505 (29%) were best modelled as *CIS – trans*, 278/505 (55%) as *TRANS – cis*, 26/505 (5%) as *CIS + trans*, and 56/505 (11%) as *TRANS + cis*. Thus, as previously reported for liver, and as we found for retina, when *cis* and *trans* effects act together, they more often act to stabilize (423/505 or 84%) than to destabilize (82/505 or 16%) gene expression (Goncalves et al. 2012).

We then compared the classification of genes between liver and retina. To avoid misattributing tissue-specific gene expression as tissue-specific *cis* or *trans* effects, we restricted our analysis to genes classifiable in both liver and retina. In particular, for comparison of *cis*-effect genes, we required that genes be classified as *cis*-effect in one tissue and conserved in the other tissue, or *cis*-effect in both tissues. Similarly, for the comparison of *trans*-effect genes, we required that genes be classified as *trans*-effect in one tissue and conserved in the other tissue, or *trans*-effect in both tissues. Using these criteria, we found that the vast majority of *cis* effects (1,661/2,242 or 74%) were tissue-specific. Additionally, most *trans* effects (871/976 or 89%) were tissue-specific (Figure 2.7C and 2.7D; Supporting Information S7). Thus, most of the isolated *cis* and isolated *trans* effects identified were tissue-specific.

Recent studies suggest that variants in a given CRE may modulate target gene expression in a tissue-dependent manner; i.e., different tissues may show differential susceptibility to CRE variants (Erceg et al. 2014). To test for tissue-specific variant effects in our system, we examined the 581 genes classified as *cis*-effect in both liver and retina. We found a positive correlation between the expression estimates for the F0 liver and F0 retina samples (Pearson $r = 0.56$, two-tailed $P < 10^{-5}$), and between the expression estimates for the F1 liver and F1 retina samples (Pearson $r = 0.58$, two-tailed $P < 10^{-5}$) (Figure 2.7E). This suggests that there exists differential susceptibility between the liver and retina to CRE variants, but that there is also significant shared susceptibility.

For the 105 genes classified as *trans*-effect in both tissues, we found a positive correlation between the expression estimates for the F0 liver and F0 retina samples (Pearson $r = 0.76$, two-tailed $P < 10^{-5}$) (Figure 2.7F), suggesting that the same *trans*-acting factors regulate many of these *trans*-effect genes in both tissues. In contrast, there was no correlation between the F1 liver and F1 retina samples (Pearson $r = 0.089$, two-tailed $P = 0.37$) for these genes. This is not surprising, since by definition, *trans*-effect genes do not show AEI in F1 hybrids, and hence the \log_2 (Cast allele/B6 allele) ratios are all close to 0. Collectively, these analyses underscore the notion that *cis* effects and *trans* effects are largely tissue-specific, but when they are shared, they tend to have similar effects on gene expression.

2.5 DISCUSSION

Genomic techniques such as ChIP-seq and DNase-seq have greatly expanded our knowledge of *cis*-regulatory regions in various tissues and cell types in recent years (The ENCODE Project Consortium 2012). Concurrently, whole-genome sequencing of thousands of individuals (Genomes Project et al. 2012) and genome-wide association studies (GWAS) have catalogued thousands of disease-associated variants, many of which fall within regulatory regions (Schaub et al. 2012). The next phase of genomic medicine will require mapping of regulatory variants onto disease-relevant phenotypes. Here, we have taken a first step toward understanding the role of regulatory variants in retinal disease by dissecting *cis*- and *trans*-regulatory effects in the mouse retina, a tissue that models many key aspects of human retinal biology (Dalke and Graw 2005).

In contrast to expression quantitative trait loci (eQTL) studies, which are feasible in the human population and are largely powered to detect *cis* effects, the F1 hybrid study approach in model organisms provides tremendous power to detect both *cis* effects and *trans* effects (Gaffney 2013). A major finding in our study is that *cis* effects predominate in the mouse retina. While estimates of the relative contributions of *cis* effects and *trans* effects based on F1 hybrid studies in *Drosophila* and yeast vary (Wittkopp et al. 2004; Tirosh et al. 2009; Emerson et al. 2010; McManus et al. 2010), all studies acknowledge a substantial contribution of *cis* effects. The variability of estimates is likely due at least in part to methodological differences in gene expression estimates and statistical modelling. For instance, when we re-analyzed the raw data from the previously published study of *cis* and *trans* effects in mouse liver (Goncalves et al. 2012), we assigned a greater fraction of gene regulatory divergence to isolated *cis* and isolated *trans* effects than the original study, which assigned a greater fraction of gene regulatory

divergence to combined *cis* and *trans* effects. These differences may be attributable to the fact that in our analysis pipeline, we used an updated reference transcriptome and Bayesian statistical models instead of maximum likelihood estimates (MLE).

Another key finding in our study is that the *cis* effects are largely tissue-specific, with only 26% being shared between liver and retina. Importantly, for this comparison, we included only genes with sufficient power for analysis in both tissues, and hence the observed tissue specificity is not an artifact of tissue-specific expression. Our estimate agrees well with an eQTL study of lymphoblastoid cell lines, skin, and adipose tissue in human twins, which found that 30% of *cis*-eQTLs were shared by the three tissues (Nica et al. 2011).

Predicting the effect of any given regulatory variant is a challenge, even in the face of complete genetic information, and even at the level of a molecular phenotype such as transcription factor binding (Maurano et al. 2012b) or, as in our case, gene expression. Moreover, regulatory variants act in combination, rather than in isolation, to modulate gene expression. Furthermore, gene expression is not always a reliable surrogate for protein levels (Greenbaum et al. 2003; McManus et al. 2014), and the path from protein to organismal phenotype is even more convoluted. With these layers of complexity in mind, we have taken a step toward understanding the links between *cis*-regulatory variants and retinal phenotypes by prioritizing variants within photoreceptor CREs that are associated with *cis*-effect genes.

Our work reveals that *cis*-regulatory effects predominate in the murine retina and are associated with functional *cis*-regulatory variants, with implications for retinal disease. In an approach complementary to eQTL studies, we have demonstrated a strategy for mapping *cis*-regulatory variants onto changes in gene expression by harnessing the power of inbred model organisms. Future empirical testing of such variants in living tissue, e.g., using high-throughput

massively parallel reporter assays (Kwasnieski et al. 2012; Shlyueva et al. 2014), will further elucidate the precise causal effects of specific *cis*-regulatory variants on gene expression.

2.6 METHODS

2.6.1 Ethics statement

All experiments were conducted in strict accordance with the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health (NIH), and were approved by the Washington University in St. Louis Institutional Animal Care and Use Committee (IACUC) (protocol #20110089). Animals were euthanized with CO₂ anesthesia followed by cervical dislocation, and all efforts were made to minimize suffering.

2.6.2. Animals

C57BL/6J (stock #664) and Cast/EiJ (stock #928) mice were purchased from Jackson Laboratory. Mice were maintained on a 12-hour light/dark cycle at ~20-22 °C with free access to food and water. Mating cages were maintained on 5K54 diet (LabDiet) and supplemented with autoclaved shepherd shacks (Shepherd Specialty Papers). Offspring were weaned at age 3 weeks and maintained on 5053 diet (PicoLab) until age 8 weeks, at which point they were sacrificed. Eyes were enucleated immediately after sacrifice. To minimize circadian effects (Storch et al. 2007), samples were collected at approximately the same time of day (late evening).

2.6.3 Sample collection and sequencing

Each biological replicate consisted of a pool of 6-8 retinas from 8 week old male mice. Retinas were dissected in cold sterile HBSS with calcium and magnesium (Gibco) and stored at -80 °C until use. Total RNA was extracted using TRIzol (Invitrogen) and purified using the RNeasy Mini Kit (Qiagen) with on-column DNaseI digestion (Qiagen). Integrity of total RNA was verified on the Agilent 2100 Bioanalyzer. Polyadenylated mRNA was captured from total

RNA using Dynabeads (Invitrogen). The mRNA was fragmented and reverse-transcribed to double-stranded cDNA using random hexamers. The cDNA was blunt-ended and 3'-adenylated before ligation to sequencing adapters. Ligated fragments were amplified for 12 cycles with primers to incorporate unique sample barcodes. Libraries were subjected to paired-end 2x101 bp sequencing on the Illumina HiSeq 2000 at the Genome Technology Access Center at Washington University School of Medicine. One lane of sequencing was conducted for all F0 and F1 samples, and a second lane of sequencing was conducted for the F1 samples only.

2.6.4 Read alignment and quantification

Reads were filtered and trimmed with Trim Galore! v0.2.6 (Krueger) prior to alignment with Bowtie v0.12.9 (Langmead et al. 2009) to a strain-specific reference transcriptome (for F0 data) or a hybrid reference transcriptome (for F1 data). Transcriptomes were constructed using the `mouse_strain_transcriptomes.sh` script within the MMSEQ package (Turro et al. 2011). The reference transcriptomes were based on the Ensembl Release 67 cDNA files and the Wellcome Trust Mouse Genomes Project Release 2 VCF files (which use mm9/NCBI37 as the reference genome) based on November 2012 HiSeq 2x100 bp sequencing with 39x coverage of the Cast/EiJ genome (Keane et al. 2011). MMSEQ v1.0.0 beta was used to estimate gene expression levels for the F0 samples and allele-specific gene expression levels for the F1 samples (Turro et al. 2011). Of the 37,991 Ensembl Release 67 mouse genes, 34,964 were autosomal, of which 29,160 had known exonic polymorphisms between Cast/EiJ and C57BL/6J. Gene-level expression estimates in units roughly equivalent to FPKM (fragments per kb of transcripts per million mapped read pairs) were derived from exponentiation of the log expression estimates.

For differential expression analysis of F0 samples with DESeq v1.10.1 (Anders and Huber 2010), normalized count equivalents were used and a negative binomial test was performed.

2.6.5 Identification of imprinted genes

Using MMDIFF, a null model (no imprinting) was compared to an imprinting model, as recently described (Turro et al. 2014). In brief, the null model assumes that allelic expression differences are the same in F1 B6xCast and F1 CastxB6, while the imprinting model assumes that allelic expression differences have equal magnitude but opposite signs in F1 B6xCast as in F1 CastxB6. Only autosomal genes with known exonic polymorphisms between Cast/EiJ and C57BL/6J were included in this analysis.

2.6.6 Mouse imprinting databases

We examined four online databases that are continually updated with known imprinted mouse genes: WAMIDEX (atlas.genetics.kcl.ac.uk) (Schulz et al. 2008), MouseBook Imprinting Catalog (www.mousebook.org) (Williamson CM 2014), Geneimprint (www.geneimprint.com) (Jirtle 2012), and Catalogue of Parent of Origin Effects (igc.otago.ac.nz) (Morison et al. 2001). For each database, we excluded genes whose imprinting status was listed as ambiguous or disproven. To resolve nomenclature disparities between databases, we converted gene names to Mouse Genome Informatics (MGI) gene names. We combined the gene lists from the four databases into a master gene list of 189 genes, of which 143 had Ensembl Release 67 IDs and 137 were autosomal. After filtering out non-polymorphic genes, we were left with 120 autosomal Ensembl ID's, corresponding to 116 MGI genes. Each Ensembl 67 gene was then assigned a

‘database score’ ranging from 0 to 4, indicating the number of databases that listed the gene as being imprinted (see Supporting Information S1).

2.6.7 Categorization of genes according to *cis* and *trans* effects

A comparison of four models (conserved model, *cis* model, *trans* model, and *cis* and *trans* model) was performed using MMDIFF, as recently described (Turro et al. 2014). In brief, the conserved model assumes there is no differential expression (DE) between the F0’s and no allelic expression imbalance (AEI) in the F1’s. The *cis* model assumes there is DE between the F0’s that is equal to the AEI in the F1’s. The *trans* model assumes there is DE between the F0’s but no AEI in the F1’s. The *cis* and *trans* model assumes that there is DE in the F0’s, but it is unequal to the AEI in the F1’s.

Included in the analysis were the 29,160 autosomal genes polymorphic between C57BL/6J and Cast/EiJ. In our retinal dataset, after excluding 306 possibly imprinted polymorphic autosomal genes (imprinting Bayes factor > 3), we had sufficient statistical power to classify 11,484 genes confidently as conserved, *cis*, *trans*, or *cis* and *trans* based on the following criteria: the winning model must have a posterior probability > 0.5, and the posterior probability of the winning model must be at least twice that of the second-best model, assuming an equal prior probability of 0.25 for each of the four models. In the previously published liver dataset (Goncalves et al. 2012), after excluding 571 possibly imprinted polymorphic autosomal genes (imprinting Bayes factor > 3), we had sufficient statistical power to classify 9,865 genes confidently using these criteria.

Genes best modelled by a combination of *cis* and *trans* effects were then subdivided into the following categories, where x is the weighted log fold change between the strains within the

F1's, and y is the weighted log fold change between the strains within the F0's (Goncalves et al. 2012):

- (1) *CIS – trans* (opposite direction with *cis* stronger than *trans*): $x*y > 0$ and $|x| > |y|$
- (2) *TRANS – cis* (opposite direction with *trans* stronger than *cis*): $x*y < 0$
- (3) *CIS + trans* (same direction with *cis* stronger than *trans*): $x*y > 0$ and $|x| < |y| < |2x|$
- (4) *TRANS + cis* (same direction with *trans* stronger than *cis*): $x*y > 0$ and $|y| > |2x|$

2.6.8 Calculation of weighted log fold change

The weighted log fold change for each gene was calculated by weighting the allele-specific posterior mean of the log expression parameter by the inverse of its posterior variance across biological replicates for each strain and subtracting the results. Let B_1 , B_2 , and B_3 be the log expression parameters for the F0 C57BL/6J samples, and let C_1 , C_2 , and C_3 be the log expression parameters for the F0 Cast/EiJ samples. Then the weighted log fold change between the F0 C57BL/6J samples and the F0 Cast/EiJ samples is given by

$$\frac{\sum_{i=1}^3 B_i / \text{var}(B_i)}{\sum_{i=1}^3 1 / \text{var}(B_i)} - \frac{\sum_{i=1}^3 C_i / \text{var}(C_i)}{\sum_{i=1}^3 1 / \text{var}(C_i)}$$
. The same approach was used to compare the two sets of F1 samples.

2.6.9 Assignment of genes to CRX ChIP-seq peaks

Previously published CRX ChIP-seq data conducted on 8 week old C57BL/6 retinas (Corbo et al. 2010) were used to assign wild-type (WT) CRX-bound regions (CBRs) to genes. CBRs were assigned to all autosomal and sex chromosomal Ensembl Release 67 gene transcripts using custom Perl scripts following a proximity-based algorithm as previously described: if a

CBR was located within a gene, it was assigned to that gene; otherwise, it was assigned to the gene with the nearest transcriptional start site (TSS) (Corbo et al. 2010).

2.6.10 Batch identification of variants

Variant calls (SNPs and indels) were downloaded as Variant Call Format (VCF) files from the Wellcome Trust Sanger Institute's Mouse Genomes Project. These calls (December 2012 release) were based on the latest high-quality, high-coverage HiSeq sequencing data of the strains. The Cast/EiJ variants relative to the reference genome (C57BL/6J NCBI Build 37) were extracted at regions of interest using VCFtools v0.1.10 (Danecek et al. 2011) and BEDtools v2.19.1 (Quinlan and Hall 2010). Only variant sites where the genotype was homozygous were included. The genomic coordinates of CBRs based on NCBI Build 37 were used. Custom Perl scripts were written to tabulate the variants for CBRs associated with Ensembl Release 67 genes.

2.6.11 Identification of variants at individual regions

Individual loci of interest were manually inspected for variants by querying an online database, the Wellcome Trust Sanger Institute's Mouse Genomes Project Mouse SNP Viewer Release 1211 (NCBI Build 37), available at http://www.sanger.ac.uk/sanger/Mouse_SnpViewer/rel-1211.

2.6.12 RetNet genes

Genes associated with human retinal disease in the RetNet database were retrieved. Human gene symbols were converted to Mouse Genome Informatics (MGI) symbols using the MGI Batch Query (Blake et al. 2014).

2.6.13 DNA constructs

Polymerase chain reaction (PCR) with Phusion High-Fidelity DNA Polymerase (New England BioLabs) was used to amplify the 0.7 kb *Sag* promoter region at -558 to +105 (mm9 chr1:89,699,697-89,700,359) relative to the TSS. Genomic DNA purified from C57BL/6J and Cast/EiJ liver tissue was used as the template for the B6 and Cast construct, respectively. The forward primer 5'-TGAGGCAATGACACTTGGTC-3' and reverse primer 5'-GCAGGGAGCTGATTGGATTA-3' with *XhoI* and *EcoRI* restriction enzyme site overhangs, respectively, were used. The fragments were subcloned upstream of DsRed in the no-basal vector (described previously in (Hsiao et al. 2007)) using the *SalI* (compatible with *XhoI*) and *EcoRI* sites. Constructs were confirmed with Sanger sequencing that encompassed the entire 0.7 kb region. We note that based on our high-quality Sanger sequencing of this region, the genomic DNA of our Cast/EiJ mice differed from the reference Cast/EiJ sequence (Keane et al. 2011) by two bases at chr1:89,700,191 (A→C) and chr1:89,700,187 (A→C), as confirmed by Sanger sequencing three different Cast/EiJ mice (representing the three Cast/EiJ RNA-seq biological replicates)

2.6.14 Retinal explant electroporation and quantification of promoter activity

Electroporation and explant culture of mouse retinas were performed as described previously (Montana et al. 2011b). In brief, retinas were dissected from newborn (P0) CD-1 mouse pups and coelectroporated with one of the *Sag* promoter DsRed constructs and a control green fluorescent protein (GFP) reporter that expresses in rod photoreceptors, *Rho*-CBR3-eGFP (Corbo et al. 2010), each at a concentration of 0.5 µg/µL. Retinas were grown in explant culture and harvested 8 days later, whereupon they were fixed and whole-mounted for quantitative

imaging of DsRed fluorescence intensity normalized to GFP fluorescence intensity using a monochromatic camera (Hamamatsu ORCA-AG), as described (Montana et al. 2011b). For each *Sag* promoter construct, 10-11 retinas were quantified. Representative images using a color camera (Olympus DP70) were also taken (see Figure 2.6C). For all retinal imaging, 40X magnification was used, and the exposure times for the red and green channels were consistent across retinas.

2.7 DATA ACCESS

RNA-seq, MMSEQ, and MMDIFF data have been deposited in Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) (accession number GSE60545).

2.8 SUPPORTING INFORMATION

Supporting information is available at:

<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0109382#s5>

2.9 ACKNOWLEDGEMENTS

We thank members of the Corbo lab for input and support during the course of this work, and to Jennifer Enright and Shuyi Ma for critical reading of the manuscript. We thank the Genome Technology Access Center (GTAC) in the Department of Genetics at Washington University School of Medicine for sequencing services.

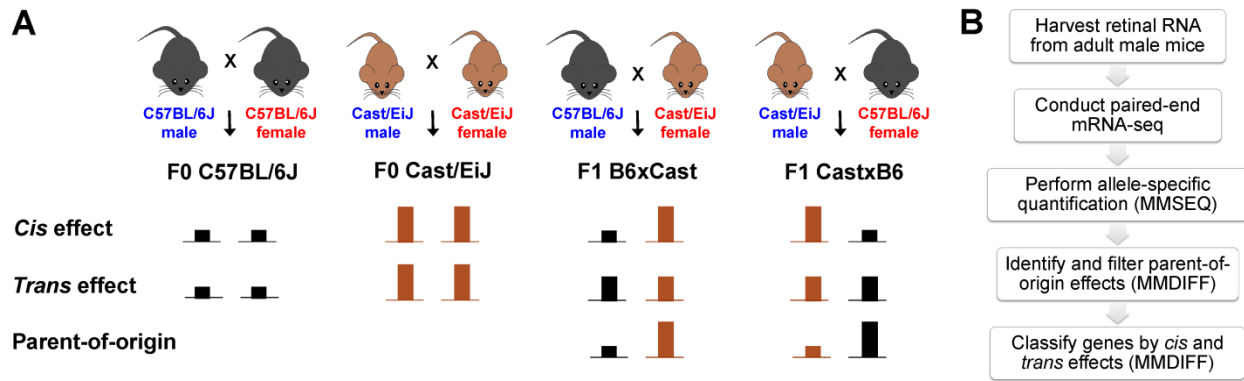


Figure 2.1. Study design. (A) F0 and F1 mice were generated via the depicted crosses. The schematic diagram illustrates example expression patterns for a *cis* effect, *trans* effect, and parent-of-origin effect. For a *cis* effect, in the F1 hybrids, the Cast/EiJ allelic expression relative to the C57BL/6J allelic expression recapitulates the ratio of gene expression levels in the F0 homozygotes. For a *trans* effect, the F1 hybrids express the Cast/EiJ and C57BL/6J alleles equally. For a parent-of-origin effect, there is preferential expression of the maternal allele (as depicted) or the paternal allele, as seen by comparison of the reciprocal F1 hybrids. (B) An overview of the workflow is shown.

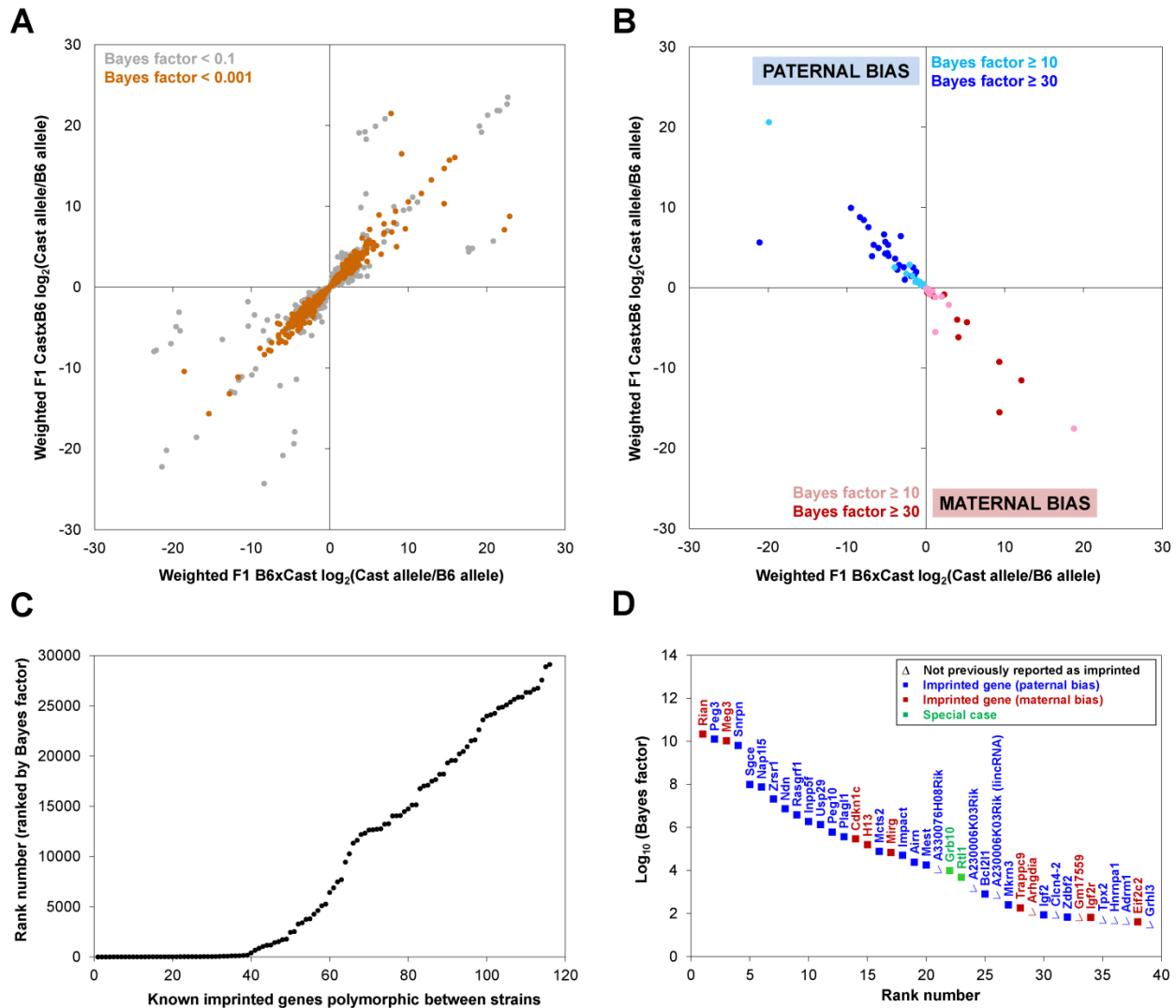


Figure 2.2. Characterization of parent-of-origin effects in the retina. Autosomal genes polymorphic between C57BL/6J and Cast/EiJ were analyzed in retinas from reciprocal F1 hybrids. Higher Bayes factors indicate greater likelihood of imprinting. (A) Non-imprinted genes with Bayes factor <0.1 (gray) and <0.001 (orange) are depicted. (B) Parent-of-origin effects with preferential expression of the paternal (blue) or maternal (red) allele with Bayes factor ≥ 10 (light) and ≥ 30 (dark) are depicted. (C) Top-ranked (low rank number) genes are enriched for known imprinted genes. (D) Genes with strong evidence of imprinting in the retina (Bayes factor ≥ 30) that exhibit preferential expression of the paternal (blue) or maternal (red) allele are depicted. Green, special cases—see text for discussion of *Rtl1* and *Grb10*. Filled squares, genes previously reported as imprinted in other tissues. Empty triangles, not previously reported as imprinted. A230006K03Rik appears twice because it is associated with two Ensembl ID's, a protein-coding gene and a lincRNA.

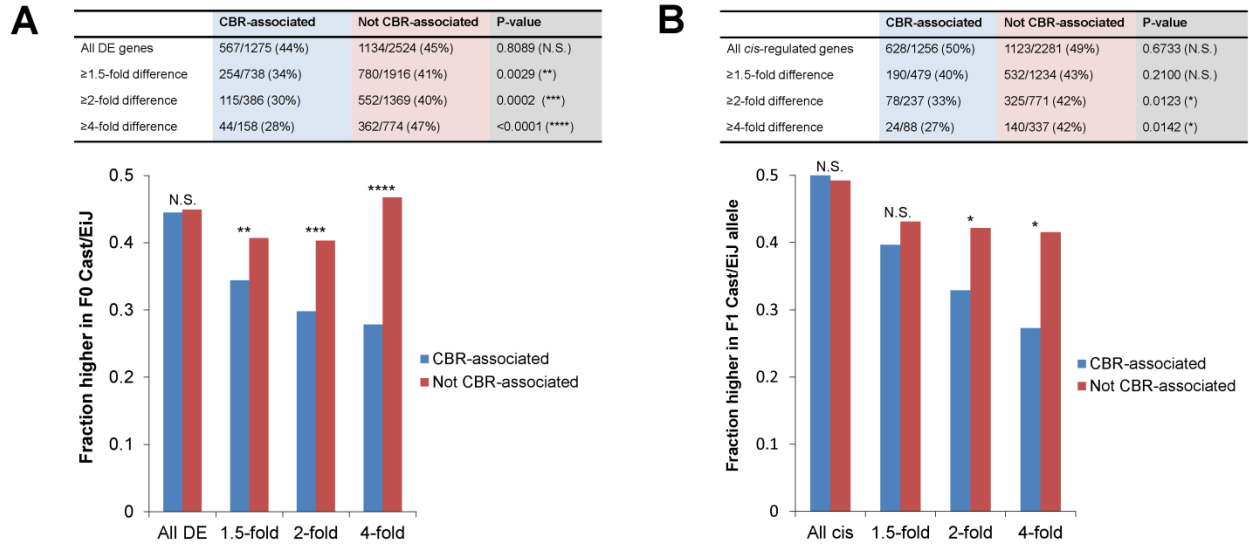


Figure 2.3. Comparison of differentially expressed and *cis*-effect genes associated with photoreceptor CREs. Genes were classified as being associated with CRX ChIP-seq peaks (CBR-associated) or not. (A) Differentially expressed (DE) autosomal genes were identified using DESeq at 5% FDR. The proportions of genes with higher expression in F0 Cast/EiJ than F0 C57BL/6J at various fold changes are shown. (B) *Cis*-effect autosomal genes were identified using MMDIFF. Proportions of genes with higher expression in F1 Cast/EiJ allele than F1 C57BL/6J allele at various fold changes are shown. P-values were calculated with two-tailed Fisher’s exact test. N.S. = not significant, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001.

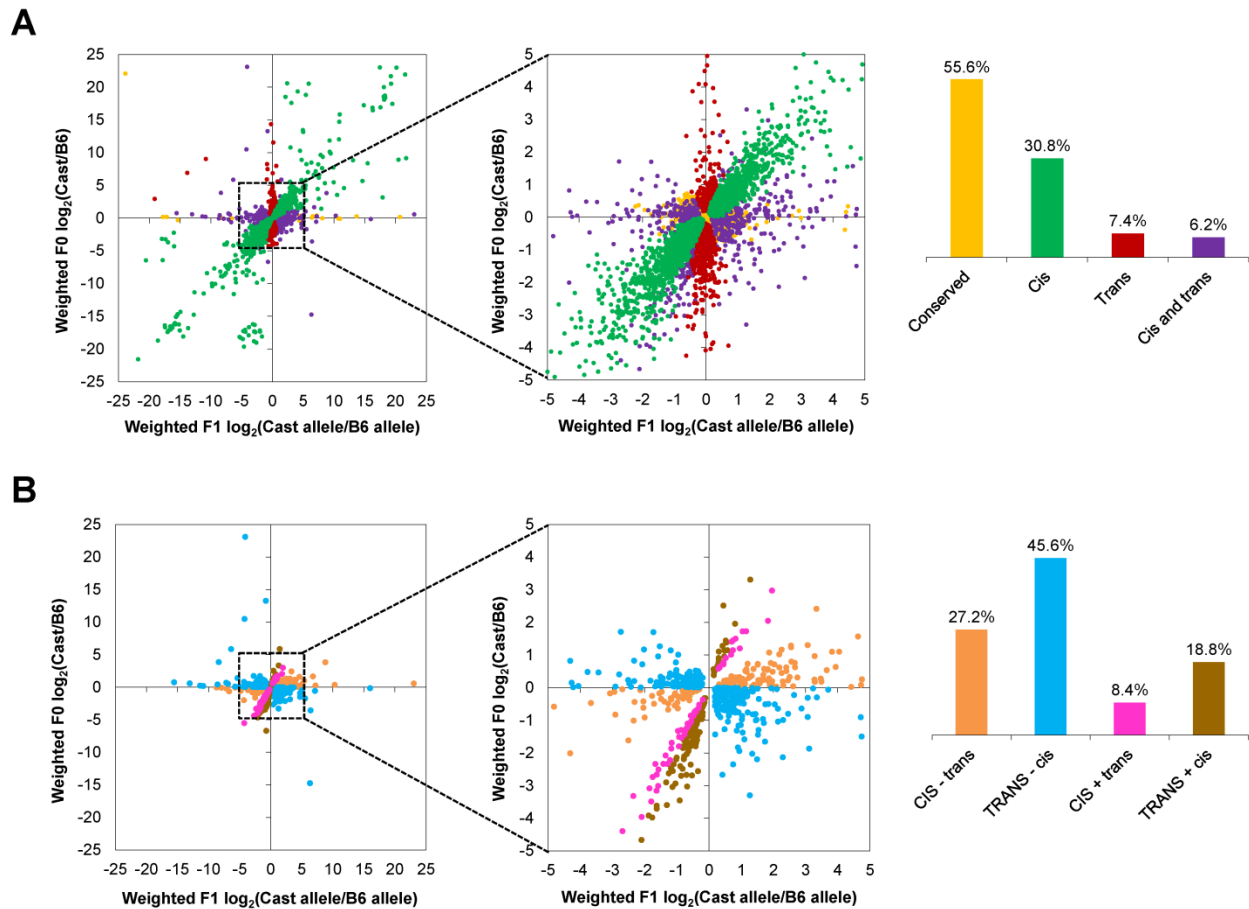


Figure 2.4. Classification of genes by mechanism of gene regulatory divergence. (A) Genes were classified as conserved (yellow; largely obscured), *cis* (green), *trans* (red), or *cis* and *trans* (purple). (B) *Cis*- and *trans*-effect genes were further subcategorized as to whether the *cis* and *trans* effects acted in the same (plus sign; pink and brown) or opposite (minus sign; orange and blue) directions, and whether the *cis* (CAPS; orange and pink) or *trans* (CAPS; blue and brown) effect was stronger. Insets, magnified views.

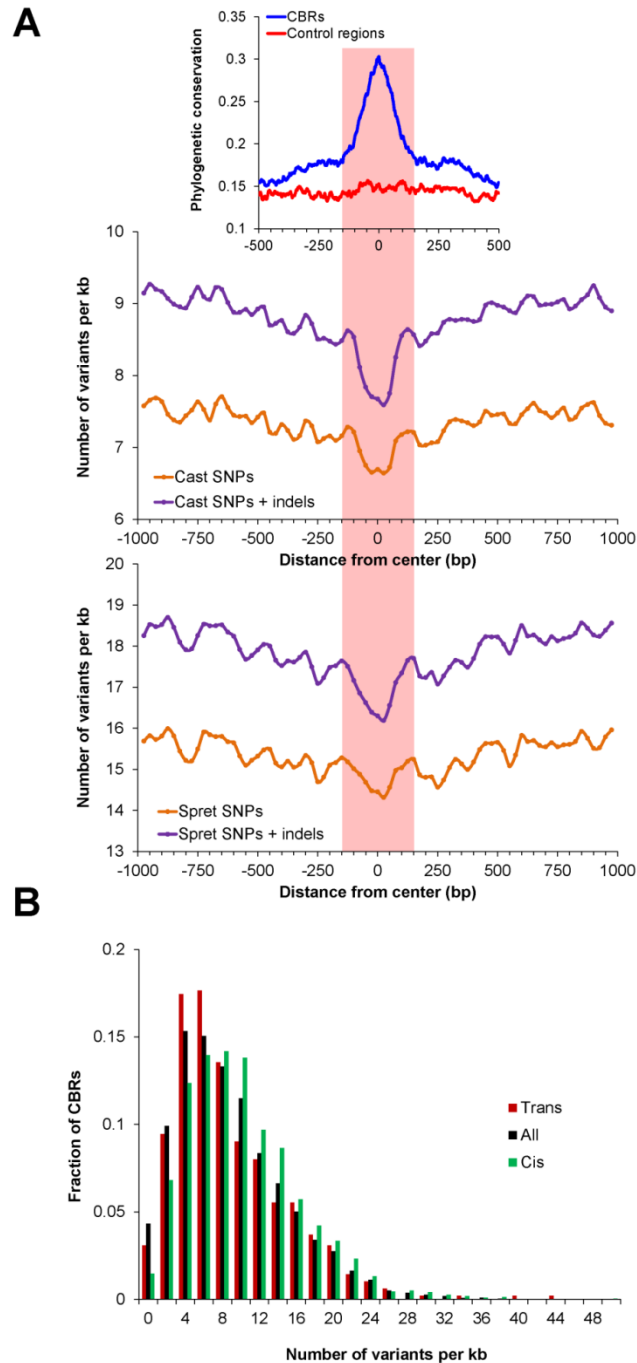


Figure 2.5. Analysis of variant density in photoreceptor CREs. (A) The number of Cast/EiJ (top) or Spret/EiJ (bottom) SNPs and indels relative to C57BL/6J was determined in 50 bp windows (sliding 25 bp at a time) across the 2 kb region centered on CBRs. Phylogenetic conservation for CBRs is based on PhastCons scores as found in (Corbo et al. 2010). The highlighted area corresponds to the central 300 bp region. (B) Histogram showing frequency of variants (SNPs + indels) in the 1 kb region centered on all CBRs (black), CBRs associated with *cis*-effect genes (green), and CBRs associated with *trans*-effect genes (red). Total bar height was normalized to 1 for each category.

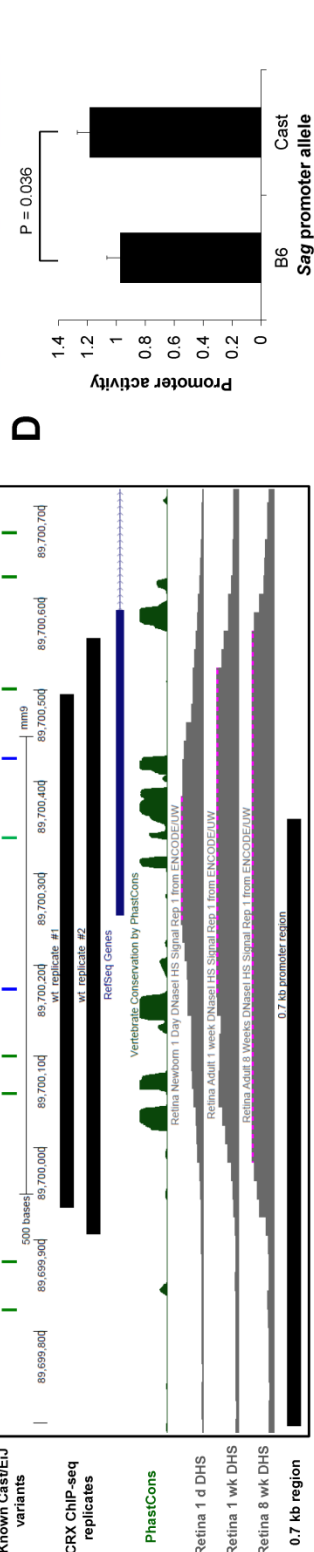
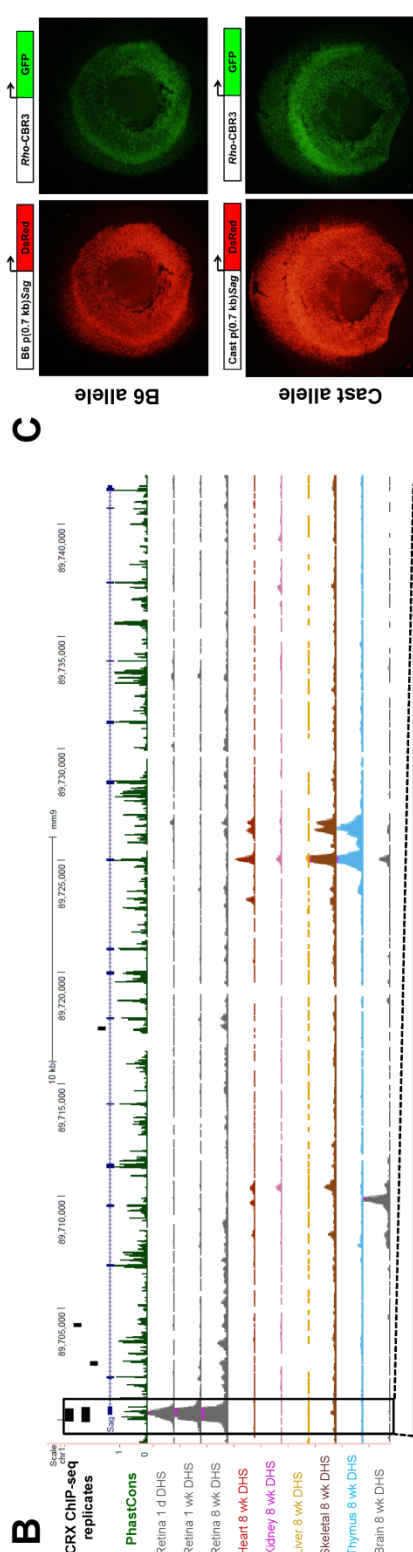
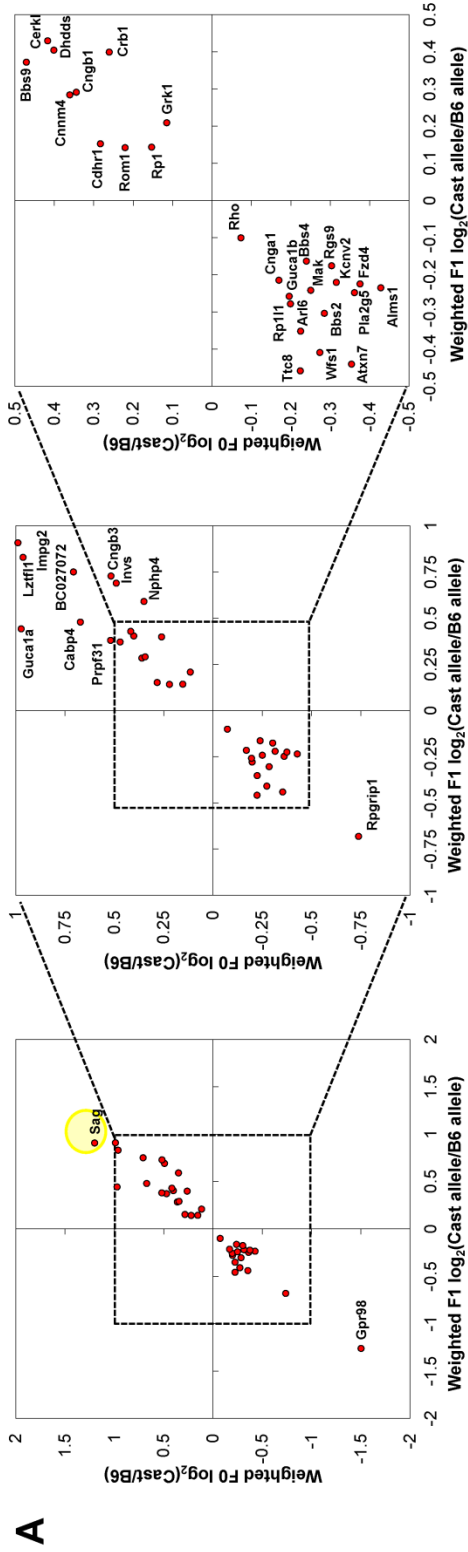


Figure 2.6. *Cis*-effect genes associated with retinal disease and photoreceptor CREs. (A) *Cis*-effect genes associated with CRX ChIP-seq peaks were matched against the RetNet database of retinal disease genes. The yellow circle highlights *Sag*. (B) *Sag* locus, mm9. Top: Screenshot from UCSC Genome Browser (Kent et al. 2002). DNaseI hypersensitivity site (DHS) signals are from ENCODE data (The ENCODE Project Consortium 2012). Bottom: Enlargement of boxed region. The 0.7 kb promoter region is depicted here. Locations of known Cast/EiJ variants (Keane et al. 2011) are depicted as green tic marks (SNPs) or blue tic marks (indels). (C) Retinal explant electroporation was used to assay the activity of the 0.7 kb *Sag* promoter region of B6 and Cast alleles. Representative images are shown here for the B6 (top) and Cast (bottom) 0.7 kb *Sag* promoter constructs driving DsRed expression. *Rho*-CBR3-eGFP served as the loading control. (D) Quantification of the *cis*-regulatory activity measured by the explant electroporation assay. Error bar represents SEM. P-value was calculated with one-tailed Wilcoxon rank-sum test.

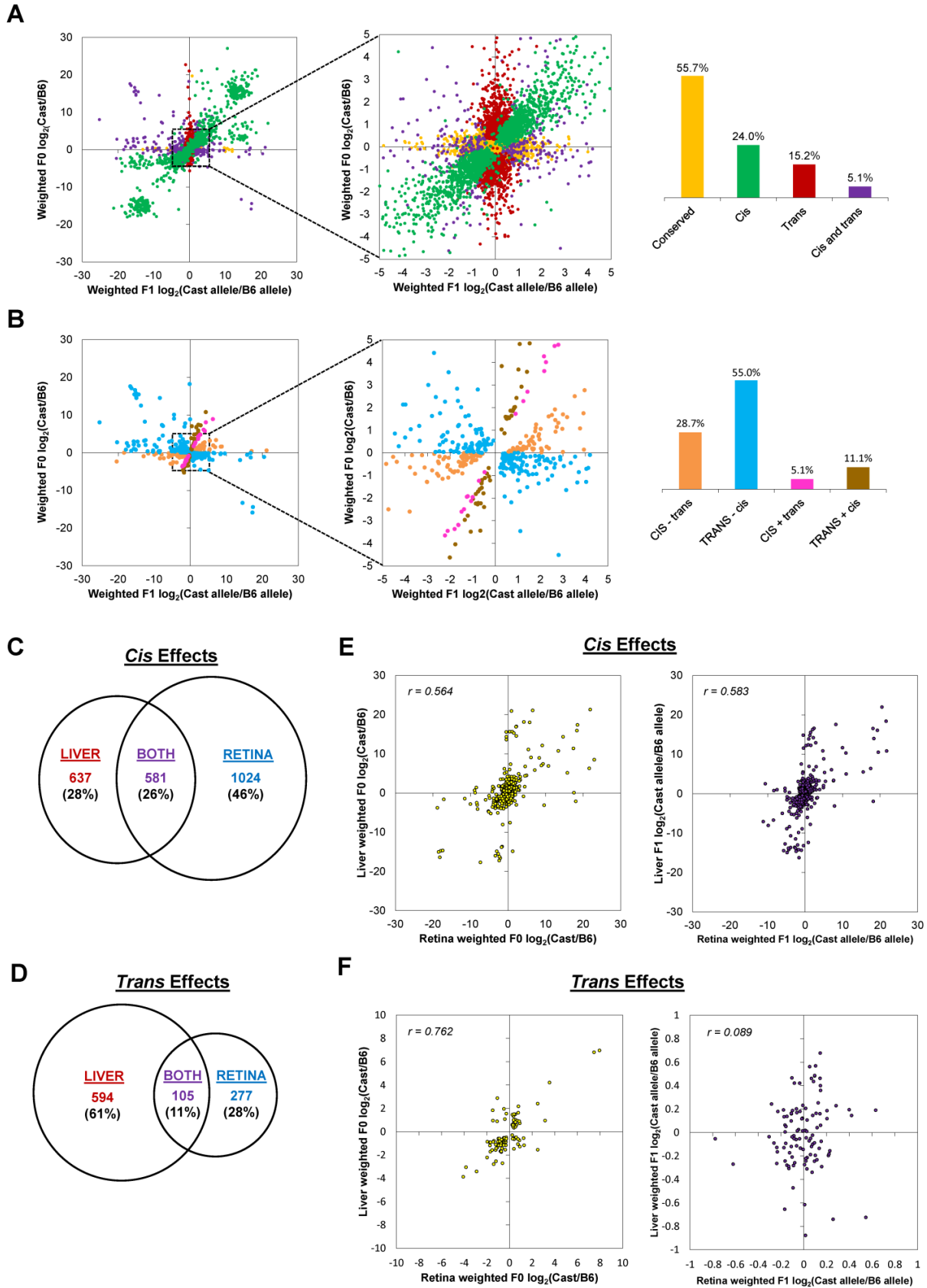


Figure 2.7. Comparison of *cis* effects and *trans* effects between liver and retina. (A) Using the same analytic pipeline as for retina, genes in the liver were classified as conserved (yellow; largely obscured), *cis* (green), *trans* (red), or *cis* and *trans* (purple). (B) *Cis*- and *trans*- regulated genes were further subcategorized as to whether the *cis* and *trans* effects acted in the same (plus sign; pink and brown) or opposite (minus sign; orange and blue) directions, and whether the *cis* (CAPS; orange and pink) or *trans* (CAPS; blue and brown) effect was stronger. (C) Number of genes classified as *cis* in liver and conserved in retina, *cis* in both tissues, or *cis* in retina and conserved in liver. (D) Number of genes classified as *trans* in liver and conserved in retina, *trans* in both tissues, or *trans* in retina and conserved in liver. (E) Correlation between genes classified as *cis* in both tissues. Pearson r values for F0 samples (left) and F1 samples (right) are shown. (F) Correlation between genes classified as *trans* in both tissues. Pearson r values for F0 samples (left) and F1 samples (right) are shown. Insets, magnified view.

Table 2.1. Agreement between F0 biological replicates.

		F0 C57BL/6J			F0 Cast/EiJ		
		R1	R2	R3	R1	R2	R3
F0 C57BL/6J	R1	1					
	R2	0.983	1				
	R3	0.992	0.982	1			
F0 Cast/EiJ	R1	0.873	0.814	0.876	1		
	R2	0.912	0.893	0.919	0.956	1	
	R3	0.907	0.912	0.904	0.897	0.979	1

Pearson r values for FPKM (fragments per kb of transcripts per million mapped read pairs) estimates across F0 samples. Bold denotes comparison between biological replicates (R1, R2, and R3) of the same genotype.

Table 2.2. Agreement between F1 biological replicates.

		F1 B6xCast, B6 allele			F1 B6xCast, Cast allele			F1 CastxB6, B6 allele			F1 CastxB6, Cast allele		
		R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
F1 B6xCast, B6 allele	R1	1											
	R2	0.929	1										
	R3	0.975	0.940	1									
F1 B6xCast, Cast allele	R1	0.949	0.911	0.953	1								
	R2	0.937	0.922	0.946	0.973	1							
	R3	0.923	0.918	0.941	0.977	0.960	1						
F1 CastxB6, B6 allele	R1	0.932	0.971	0.960	0.914	0.924	0.926	1					
	R2	0.927	0.974	0.952	0.911	0.920	0.916	0.987	1				
	R3	0.939	0.957	0.965	0.925	0.924	0.935	0.985	0.975	1			
F1 CastxB6, Cast allele	R1	0.885	0.947	0.923	0.926	0.943	0.939	0.951	0.958	0.950	1		
	R2	0.909	0.909	0.945	0.942	0.940	0.961	0.948	0.928	0.963	0.965	1	
	R3	0.906	0.917	0.947	0.946	0.938	0.959	0.951	0.947	0.948	0.971	0.979	1

Pearson r values for FPKM estimates across F1 samples. Bold denotes comparison between biological replicates (R1, R2, and R3) of the same genotype and for the same allele (B6 allele or Cast allele).

Table 2.3. Accuracy of X chromosomal read mapping in F1 samples.

F1 B6xCast		F1 CastxB6	
Maternal allele: Cast/EiJ		Maternal allele: C57BL/6J	
R1	99.4%	R1	99.5%
R2	99.5%	R2	99.5%
R3	99.5%	R3	99.5%

Percentages of X chromosomal unique hits (i.e., read pairs mapping uniquely to C57BL/6J or Cast/EiJ) that mapped to the correct genome. Since only males were used, reads should derive only from the maternal allele.

CHAPTER 3:

Massively Parallel *Cis*-Regulatory Analysis in the Mammalian Central Nervous System

3.1 AUTHOR CONTRIBUTIONS

This chapter is adapted from a published manuscript: Shen SQ¹, Myers CA¹, Hughes AEO¹, Byrne LC², Flannery JG², and Corbo JC^{1,*}. (2016) “Massively parallel *cis*-regulatory analysis in the mammalian central nervous system.” *Genome Res.* 26:238-55. The experimental work was done in collaboration with Connie Myers, and the bioinformatic analyses were done in collaboration with Andrew Hughes. Joseph Corbo and I conceived the project and designed the experiments. Leah Byrne and John Flannery contributed to the design and construction of the adeno-associated virus (AAV). Joseph Corbo and I wrote the manuscript.

This work builds upon CRE-seq, which was developed by Jamie Kwasnieski, Ilaria Mogno, and Connie Myers in collaboration between the laboratories of Joseph Corbo and Barak Cohen (Kwasnieski et al. 2012).

¹Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri, United States of America

²Helen Wills Neuroscience Institute, University of California, Berkeley, California, United States of America

*Corresponding author

3.2 ABSTRACT

Cis-regulatory elements (CREs, e.g., promoters and enhancers) regulate gene expression, and variants within CREs can modulate disease risk. Next-generation sequencing has enabled the rapid generation of genomic data that predict the locations of CREs, but a bottleneck lies in functionally interpreting these data. To address this issue, massively parallel reporter assays (MPRAs) have emerged, in which barcoded reporter libraries are introduced into cells and the resulting barcoded transcripts are quantified by next-generation sequencing. Thus far, MPRAs have been largely restricted to assaying short CREs in a limited repertoire of cultured cell types. Here, we present two advances that extend the biological relevance and applicability of MPRAs. First, we adapt exome capture technology to instead capture candidate CREs, thereby tiling across the targeted regions and markedly increasing the length of CREs that can be readily assayed. Second, we package the library into adeno-associated virus (AAV), thereby allowing delivery to target organs *in vivo*. As a proof-of-concept, we introduce a capture library of ~46,000 constructs, corresponding to ~3,500 DNase I hypersensitive (DHS) sites, into the mouse retina by *ex vivo* plasmid electroporation and into the mouse cerebral cortex by *in vivo* AAV injection. We demonstrate tissue-specific *cis*-regulatory activity of DHSs and provide examples of high-resolution truncation mutation analysis for multiplex parsing of CREs. Our approach should enable massively parallel functional analysis of a wide range of CREs in any organ or species that can be infected by AAV, such as non-human primates and human stem cell-derived organoids.

3.3 INTRODUCTION

Cis-regulatory elements (CREs, e.g., promoters and enhancers) are DNA regions that regulate gene expression, and variants within CREs can contribute to phenotypic diversity, including disease susceptibility (Wray 2007; Albert and Kruglyak 2015). In the past several years, vast amounts of genomic data have been generated that predict the locations of hundreds of thousands of CREs in cell lines and primary tissues (Shen et al. 2012; The ENCODE Project Consortium 2012; Romanoski et al. 2015). As an avenue for the experimental validation of these predictions, massively parallel reporter assays (MPRAs, e.g., CRE-seq) have been developed, in which barcoded plasmid reporters are introduced into cells. Next-generation sequencing of the resulting barcoded transcripts provides a quantitative measure of CRE activity (Kwasnieski et al. 2012; Melnikov et al. 2012; Patwardhan et al. 2012; Arnold et al. 2013; White et al. 2013; Levo and Segal 2014; Shlyueva et al. 2014). Thus far, MPRAs have been largely restricted to assaying short CRE fragments (<150 bp) synthesized as oligonucleotide libraries on microarrays (Patwardhan et al. 2009; Baker 2011; White et al. 2013) and delivered into select mammalian cells accessible by transfection or electroporation. However, CREs are often hundreds of base pairs in length, and CRE activity depends crucially on the assayed cell type and its particular complement of transcription factors (TFs) (Davidson 2001). Therefore, we sought to expand the biological relevance and applicability of MPRAs by increasing the length of assayed CREs and by widening the repertoire of assayable cell types.

The retina and cerebral cortex are two parts of the central nervous system (CNS) with a shared forebrain origin, whose gene regulatory networks are topics of intense research interest (Swaroop et al. 2010; Wright et al. 2010; Bae et al. 2015; Nord et al. 2015). The genome-wide locations of putative CREs have been mapped in both tissues, using methods such as ChIP-seq

and DNase-seq (Visel et al. 2009; Corbo et al. 2010; The ENCODE Project Consortium 2012; Wilken 2015). Compared to the cortex, the retina is more experimentally amenable to *cis*-regulatory analysis, in part because its cellular composition is more completely understood (Livesey and Cepko 2001; London et al. 2013). Electroporation can be used to efficiently deliver plasmid DNA into rod photoreceptors, which constitute the majority (~80%) of the cells in the retina (Jeon et al. 1998). We previously conducted CRE-seq by electroporating thousands of short CREs into the neonatal mouse retina *ex vivo* (Kwasnieski et al. 2012; White et al. 2013). Although hundreds of putative developmental forebrain enhancers have been assayed with one-at-a-time transgenic mouse reporter assays (Nord et al. 2013; Visel et al. 2013), never before has massively parallel *cis*-regulatory analysis been conducted in the mammalian CNS *in vivo*.

Here, we sought to overcome current technological hurdles by developing a ‘capture-and-clone’ approach for synthesizing CRE-seq libraries with a selectable range of fragment sizes for targeted *cis*-regulome analysis. As a built-in feature, our approach allows for truncation mutation analyses, which can identify regions within CREs that are critical for activity. We furthermore demonstrate the feasibility of conducting *in vivo* CRE-seq in the adult cerebral cortex by AAV-mediated delivery. Our approach provides a framework for the massively parallel functional analysis of CREs in a broad repertoire of organs and species *in vivo*.

3.4 RESULTS

3.4.1 Identification and characterization of candidate CRE regions

The genomic locations of CREs can be predicted by the patterns of phylogenetic conservation, the occurrence of transcription factor binding sites, and the presence of various chromatin features (Levo and Segal 2014; Shlyueva et al. 2014). DNase I hypersensitive (DHS) sites, which demarcate regions of open chromatin, are one of the most informative predictive features of active CREs (Arvey et al. 2012; Natarajan et al. 2012; Kwasnieski et al. 2014). Moreover, DNase-seq data for a variety of primary mouse tissues are available as part of the Mouse ENCODE Project (Yue et al. 2014). To facilitate the direct comparison of a given CRE-seq library in retina and cerebral cortex, we generated a list of tissue-specific candidate CREs based on mouse DNase-seq data, corresponding to 1,000 DHS regions from adult retina and 1,000 DHS regions from adult whole brain. Additionally, we included DHSs from two adult mouse non-neural tissues (1,000 DHSs from heart and 1,000 DHSs from liver) as controls (Supplemental Table S1). Together, this yielded 4,000 target DHS regions.

We first examined the genome-wide distributions of the 4,000 target DHS regions using GREAT and HOMER, two computational tools for annotating coding and non-coding regions (Heinz et al. 2010; McLean et al. 2010). The majority (75%) of the DHS regions were distal elements located more than 10 kb away from the nearest transcriptional start site (TSS) (Figure 3.S1A). Almost all of the DHS regions fell within introns (46%) or intergenic regions (45%) (Figure 3.S1B), similar to the genome-wide distribution of DHS regions in other cell types (Shu et al. 2011). A small number of DHSs (156/4,000 or 4%) were ‘promoter-proximal’, i.e., falling within -1 kb to +100 bp relative to the nearest TSS (Figure 3.S1A). Among these, 77/156 (49%)

were retinal DHSs, consistent with the previous observation that photoreceptor CREs often cluster around TSS's (Corbo et al. 2010).

Tissue-specific CREs are often enriched for the binding of TFs important for cell identity and function (Davidson 2001). Accordingly, we used HOMER (Heinz et al. 2010) to quantify enrichment of TF motifs in the target regions (Supplemental Table S2). For each set of tissue-specific target DHSs, we found strong enrichment of putative binding sites for TFs known to be important in that tissue. For example, among the top statistically significant enrichments for the retina, brain, heart, and liver DHSs were putative motifs for CRX (Chen et al. 1997; Freund et al. 1997), ASCL1 (Kim et al. 2008b), MEF2C (Edmondson et al. 1994), and ONECUT1 (also known as HNF6) (Clotman et al. 2005), respectively.

Since tissue-specific CREs are often associated with genes specifically expressed in the corresponding tissue (Natarajan et al. 2012; Heinz et al. 2015), we also examined the genes associated with the target DHSs based on the nearest TSS (Supplemental Table S1). Gene Ontology (GO) analysis (Carbon et al. 2009) revealed an enrichment for tissue-specific functions that corresponded to the tissue of DHS origin. For instance, among the top significant hits for the retina, brain, heart, and liver target DHSs were 'sensory perception of light stimulus', 'nervous system development', 'cardiovascular system development', and 'organic substance metabolic process', respectively (Supplemental Table S3). Thus, the 4,000 target DHS regions were likely enriched for tissue-specific CREs.

3.4.2 'Capture-and-clone' allows synthesis of targeted *cis*-regulome libraries

To overcome the length restrictions imposed by oligonucleotide array synthesis of CRE fragments (Cleary et al. 2004), we took advantage of DNA capture, a technique routinely used

for exome sequencing. For exome capture, biotinylated RNA baits are designed to selectively hybridize with DNA fragments containing sequences of interest, i.e., exonic regions (Gnirke et al. 2009). Here, we adapted this technology to target our CREs of interest (a subset of the putative ‘*cis*-regulome’) instead of the exome. This approach offers important advantages. First, the input DNA pool can derive from any genomic DNA source. Hence, the *cis*-regulome of any single individual or groups of individuals can be assessed. Second, the input DNA pool can be size-selected for a range of fragment lengths, enabling inclusion of long CREs.

Using mouse (C57BL/6J) genomic DNA that was sheared by sonication and then size-selected to be ~400-500 bp (excluding adapter sequence), we captured with RNA baits tiling the central 300 bp (which is the median size of DHSs (Natarajan et al. 2012)) of the 4,000 target DHS regions. We amplified the captured fragments with primers containing restriction sites for cloning into a barcoded vector library (Figure 3.1A). Since the cloning was non-directional, both orientations were roughly equally represented, as expected (49% and 51% of fragments mapped to the plus and minus strands of the mm9 reference genome, respectively). Paired-end sequencing revealed a distribution of CRE fragment sizes with a median length of 464 bp (SD = 72 bp) (Figure 3.1B). Using two successive rounds of capture, we achieved a very high ‘on-target’ rate: 98.5% of the captured fragments overlapped a target region. The median overlap for on-target fragments was 282 bp out of the 300 bp target, i.e., 94% of the target region length (Figure 3.S2). Overall, 3,483 of the 4,000 (87%) targeted regions were represented, with a median coverage of 8 barcodes per represented region, for a total of 45,670 uniquely barcoded constructs (Figure 3.1C).

The distribution of captured fragments across a representative chromosome is shown in Figure 3.2A. Notably, many loci exhibited a multiplicity of captured fragments corresponding to

a single target region, resulting in a tiling of the DHS peak, as exemplified in Figure 3.2B-E. Hence, the ability to conduct CRE truncation mutation analysis at a given locus is a key built-in feature of our capture-and-clone approach.

3.4.3 AAV packaging and delivery preserves CRE-seq library composition

We next considered how to expand the repertoire of cell types accessible by CRE-seq. Whereas efficient plasmid delivery is limited to mitotic cells amenable to chemical transfection or electroporation (Mortimer et al. 1999; Karra and Dahm 2010), the ideal CRE-seq delivery vehicle would permit access to a variety of tissues, including post-mitotic tissues, and in a range of species. We reasoned that adeno-associated virus (AAV), a non-pathogenic virus commonly used for gene therapy studies, would be suitable for this purpose. AAV causes long-lasting infection in rodents and primates, and its tissue tropism ranges by serotype from promiscuous to cell-type selective (Mingozzi and High 2011). Moreover, unlike DNA delivered by lentivirus, the AAV-delivered DNA remains almost exclusively episomal, thereby permitting *cis*-regulatory analysis without the insertion site effects associated with integration into the host genome (McCarty et al. 2004).

After cloning in a TATA box-containing minimal promoter-green fluorescent protein (GFP) cassette (Figure 3.1A), we transferred the library into a vector with inverted terminal repeats (ITRs), which are necessary for AAV packaging (Yan et al. 2005)). This yielded the final plasmid library (Figure 3.3A). To deliver the library into the retina, we conducted *ex vivo* electroporation of the plasmid library into the neonatal mouse retina, as in our past CRE-seq studies (Kwasnieski et al. 2012; White et al. 2013). We generated three biological replicates, each consisting of multiple electroporated retinas.

To deliver the library into the cerebral cortex, we packaged the plasmid library into AAV9(2YF) and conducted *in vivo* stereotactic injections to infect adult primary motor cortex. AAV9 is a serotype that exhibits broad tissue tropism, and its tyrosine-mutated derivative AAV9(2YF) transduces neurons of the CNS with high efficiency and minimal host-mediated degradation of viral particles (Zhong et al. 2008; Zincarelli et al. 2008; Dalkara et al. 2012; Aschauer et al. 2013). We generated three biological replicates, each consisting of cerebral cortex tissue from a single injected mouse.

As evidence that AAV packaging and stereotactic injection did not adversely affect the composition of the library, we observed a strong correlation (Pearson $r = 0.95$) between the relative abundance of individual barcoded constructs in the retina after delivery of the plasmid CRE-seq library and in the cerebral cortex after infection with the AAV-packaged CRE-seq library (Figure 3.3B). Furthermore, 76% (34,824/45,670) of the on-target barcodes were ‘well-represented’ (i.e., had at least 10 raw DNA reads) in all six biological replicates (three replicates each for retina and cerebral cortex). These 34,824 barcodes covered 97% (3,375/3,483) of the targeted DHS regions that were represented in the initial post-capture library. These results indicated good preservation of barcode abundance and diversity throughout the procedure, from the initial post-capture cloning to the delivery of the library.

We then examined the tissues histologically for evidence of library expression, as visualized by fluorescence microscopy. Upon examination of the electroporated retinas, we observed GFP-positive cells in the outer nuclear layer (ONL) of the retina, where the rod photoreceptor cell bodies reside (Figure 3.3C). Moreover, the GFP-positive cells co-expressed the rod-specific *Rho*-CBR3-DsRed reporter (Corbo et al. 2010) (Figure 3.S3A). These findings

indicated that the GFP-positive cells were rod photoreceptors, which are the predominant cell type assayed by neonatal retinal electroporation.

Upon histological examination of the AAV-injected brains, we observed bilateral GFP-positive regions throughout all layers of the cerebral cortex (Figure 3.3D), corresponding to GFP-expressing cells seen under higher magnification (Figure 3.3E). Many of the GFP-positive cells were morphologically consistent with pyramidal neurons, with an apically oriented primary dendrite and an axon. Furthermore, GFP expression co-localized with RBFOX3 (also known as NeuN) (Mullen et al. 1992), a widely expressed marker of mature neurons (Figure 3.S3B). Interestingly, there were bundles of GFP-positive axons crossing the midline in the corpus callosum (red arrow in Figure 3.3D), indicating that interhemispheric projection neurons were among the cells that expressed the CRE-seq library.

3.4.4 AAV-mediated CRE-seq demonstrates tissue-specific CRE activity of DHSs *in vivo*

Given the histological evidence for expression of the library in both tissues, we next quantified the *cis*-regulatory activity of individual constructs by next-generation sequencing. As quality control measures, we verified that the samples overall clustered by the assayed tissue type (retina vs. cerebral cortex). We also observed that the RNA read counts for individual barcodes were correlated among the three biological replicates for each tissue, although greater variability was observed among the cerebral cortex samples than the retinal samples (Figure 3.S4 and Supplemental Table S4).

Since tissue-specific DHSs are believed to mediate tissue-specific *cis*-regulatory activity (Natarajan et al. 2012; Heinz et al. 2015), we first asked whether this was the case. For this analysis, we assigned the ‘overall’ *cis*-regulatory activity of a given DHS by averaging across

corresponding barcoded constructs (as well as across biological replicates). Here, we included the ~3,000 DHSs with at least two barcoded constructs. When we examined the relationship between the DHS type (i.e., the tissue origin of the DHS) and CRE activity as assayed in the retina, we observed strong enrichment of retinal DHSs among highly expressed DHSs, especially among the top ~20% most highly expressed DHSs in the retina (Figure 3.4A). Since averaging across barcoded constructs may not necessarily be the best metric of *cis*-regulatory activity for a given DHS, we also examined the expression of individual barcoded constructs. This again revealed the strong preference of the retina for expressing retinal DHSs (Figure 3.4B).

Similarly, in the cerebral cortex, there was an enrichment of brain DHSs among highly expressed DHSs, especially among the top ~15% most highly expressed DHSs in the cortex (Figure 3.4A). However, this enrichment was less pronounced than for retina: among the top 15% most highly expressed DHSs in the retina, 79% were retinal DHSs, while among the top 15% most highly expressed DHSs in the cerebral cortex, 42% were brain DHSs ($p < 0.0001$, Fisher's exact test). As seen from the individual barcoded constructs (Figure 3.4B), there was a clear preference for brain DHSs among the most active constructs, but there was overall more promiscuous (less selective) activity of constructs in the cortex. The activity profile of non-brain DHSs in the cortex was right-shifted (increased) and overlapped to a greater extent with the activity profile of brain DHSs in the cortex, compared to the activity profile of non-retinal vs. retinal DHSs in the retina. Overall, these findings indicated that there was tissue-specific *cis*-regulatory activity of DHSs in the retina and the cortex, with the retina exhibiting a stronger preference for retinal DHSs than the cortex exhibited for brain DHSs.

3.4.5 Parameters that predict *cis*-regulatory activity

We next asked whether certain parameters previously found to be associated with *cis*-regulatory activity were predictive of high activity in our assay. For each parameter examined in Figure 3.5A-D, we considered the top 100 and top 200 most highly expressed DHSs for the tissue-appropriate DHS type (i.e., for the retina, we restricted our analysis to retinal DHSs, and for the cerebral cortex, we restricted our analysis to brain DHSs). Corresponding data for the liver and heart DHSs are provided in Figure 3.S5. We first surveyed expression as a function of position relative to the center of the DHS target region, within a 1 kb window (Figure 3.5A). While DNase-seq signals had a relatively narrow peak (~300 bp width) (Figure 3.5B), *cis*-regulatory activity in both the retina and cortex had a much broader peak, plateauing in the central ~500 bp. The breadth of the *cis*-regulatory activity peaks likely reflects the longer length of the captured fragments (median length of 464 bp) and the large extent of overlap with the central 300 bp of the DHS regions (median overlap of 94%). Notably, we did not find a substantial relationship between the length of individual CRE fragments and CRE activity (Figure 3.S6), or between distance from the nearest TSS and CRE activity (Figure 3.S7).

Interestingly, higher DNase-seq scores were significantly associated with higher *cis*-regulatory activity in the retina but not in the cortex (Figure 3.5B). A possible explanation is that the retinal DNase-seq data primarily reflect the chromatin state of rods, since they constitute the vast majority of cells in the retina (Jeon et al. 1998), and that the most strongly expressed DHSs are rod CREs. By comparison, the brain DNase-seq data reflect the chromatin state of a heterogeneous cell population, and the most strongly expressed DHSs in the cortex may be cell type-specific CREs highly active in only a subset of cells.

Next, we investigated GC content, which has been reported to be elevated within CREs. This elevation in GC content is thought to favor nucleosome occupancy in tissues where the CRE is not active, thereby repressing *cis*-regulatory activity in those tissues (Tillo and Hughes 2009; Tillo et al. 2010; Fenouil et al. 2012; Wang et al. 2012; Hughes and Rando 2014). We previously published an enhancer study, in which short (84 bp) synthetic CREs were cloned upstream of a photoreceptor-specific proximal promoter. This study revealed a positive correlation between GC content and enhancer activity in the retina (White et al. 2013). Thus, we were surprised to find that here, the most active retinal DHSs in the retina had significantly lower GC content (Figure 3.5C). However, a recent CRE-seq study using a minimal promoter also found lower GC content in highly active enhancers (Kwasnieski et al. 2014). Therefore, GC content appears to have distinct roles when the CRE acts as an autonomous element with a minimal promoter or as an enhancer with an active proximal promoter. Brain DHSs had a different pattern, with markedly elevated GC content centrally, and further increased GC content was seen among the most active brain DHSs in the cortex (Figure 3.5C). The different effects of GC content in the two tissues may reflect AT-rich vs. GC-rich motifs of tissue-specific TFs, and/or the distinct preferences of tissue-specific TFs for AT-rich vs. GC-rich ‘environments’ surrounding the TF motif (Dror et al. 2015).

An ongoing debate in the field of genomics is the degree to which phylogenetic conservation at the DNA sequence level is an accurate predictor of functional CREs, given that there is rapid turnover of individual TF binding sites in the course of evolution (Dermitzakis and Clark 2002; Vierstra et al. 2014). We observed significantly higher vertebrate conservation (as measured by PhastCons scores (Siepel et al. 2005)) for the most strongly expressed retinal and brain DHSs in the retina and cortex, respectively. This elevated phylogenetic conservation

occurred primarily within the central ~100 bp of DHSs (Figure 3.5D). This distribution of phylogenetic conservation is consistent with the previous observation that highly local (<100 bp) sequences confer substantial CRE activity (White et al. 2013).

We then considered TF motif content, which has been found to be predictive of *cis*-regulatory activity (Kwasnieski et al. 2014; Blatti et al. 2015). Here, we examined the enrichment of TF motifs among the DHSs with the highest or lowest activity in the retina and cortex, regardless of the type of DHS (Figure 3.5E and Supplemental Table S5). In the retina, highly active DHSs were enriched for homeobox, E-box, nuclear receptor (NR), MADS-box, and CCAAT motifs, while in the cerebral cortex, highly active DHSs were enriched for MADS-box, zinc finger (ZF), and helix-turn-helix (HTH) motifs.

To assess the predictive power of these features (DNase-seq scores, GC content, PhastCons scores, and TF motifs), we created logistic regression models and visualized their performance with receiver operating characteristic (ROC) curves, with five-fold cross-validation to control for over-fitting (Figure 3.5F and Supplemental Table S6). All constructs assayed in each tissue were classified as ‘high’ (top ~5% of ~36,000 constructs in retina, or top ~1% of ~39,000 constructs in cerebral cortex) vs. ‘not high’. In the retina, DNase-seq was the single most predictive feature (AUC = 0.921), reflecting the strong tendency for highly active constructs to be retinal DHSs. Retinal CRX ChIP-seq peaks (Corbo et al. 2010) performed nearly as well (AUC = 0.892), likely reflecting the fact that CRX ChIP-seq peaks are essentially a subset of retinal DHSs (Wilken 2015). Interestingly, a model based on 15 TF motifs also performed reasonably well (AUC = 0.785). By comparison, in a prior CRE-seq study conducted in cell lines, a model using 50 TF motifs attained an AUC of 0.80 (Kwasnieski et al. 2014). The predictive values of GC content (AUC = 0.521) and PhastCons (AUC = 0.537) were weak. In the

cerebral cortex, DNase-seq was likewise the single most predictive feature (AUC = 0.778). A model based on 13 TF motifs performed reasonably well (AUC = 0.734), while GC content (AUC = 0.608) and PhastCons (AUC = 0.659) had modest predictive power in the cortex. Notably, in both tissues, the combined model performed only slightly better than DNase-seq alone. Overall, these results reflect the degree of preference of the retina and cerebral cortex for expressing retinal DHSs and brain DHSs, respectively, while underscoring the importance of TF motifs in specifying CRE activity. Furthermore, these results underscore the power of open chromatin mapping techniques such as DNase-seq for identifying functional CREs.

3.4.6 Tiling of captured fragments allows for truncation mutation analysis

The potential for conducting truncation mutation analysis is an attractive and potentially powerful feature of the capture approach. We therefore sought to determine whether the results were comparable to those of a previously published ‘traditional’ one-at-a-time promoter analysis. NRL is a master regulator of rod photoreceptor development, required both for rod fate determination and maintenance (Mears et al. 2001; Swaroop et al. 2010). Past studies of the *Nrl* promoter region identified a 30 bp ‘critical region’ that is absolutely required for promoter activity. This critical region contains TF binding sites for CRX and RORB, both of which are required for *Nrl* expression (Kautzmann et al. 2011; Montana et al. 2011a). Since the *Nrl* promoter contained a retinal DHS that was targeted in our library, we compared the results of CRE-seq and a traditional promoter analysis that used fluorescence as a read-out of *cis*-regulatory activity (Montana et al. 2011a). Since promoters act directionally (Andersson et al. 2014; Duttke et al. 2015), we compared CRE-seq constructs that were oriented in the same direction as the traditional promoter constructs. We found good agreement between the two

assays overall (Figure 3.6A), despite differences in construct design (e.g., the CRE-seq constructs contained a minimal promoter, and the 3' ends of fragments varied). Importantly, both identified the same critical region within a block of phylogenetic conservation (Montana et al. 2011a). Thus, CRE-seq truncation analysis recapitulated the results of a traditional truncation mutation analysis.

Besides the *Nrl* promoter, we found additional instances of novel truncation mutation analyses afforded by the capture approach. As seen in Figure 3.6B, a retinal DHS in the intron of *Rbm20* showed strong activity in the retina and weak activity in the cortex. Intriguingly, our assay revealed a 12 bp critical region containing a predicted binding motif for CRX. This motif, 'CTAATCCT' (on the negative strand) is a near-perfect match to the consensus motif, 'CTAATCCC' (Lee et al. 2010).

Figure 3.6C depicts another truncation mutation analysis, this time for two brain DHSs (labeled '1' and '2') located <0.5 kb apart within an intron of *Bsn* (Bassoon). Bassoon is a presynaptic protein that is important for neurotransmitter release from glutamatergic (excitatory) neurons (Altrock et al. 2003). Both of these brain DHSs contained phylogenetically conserved regions, as observed by PhastCons (Siepel et al. 2005). Interestingly, while both had low *cis*-regulatory activity in the retina, DHS #1 had low activity in the cerebral cortex, whereas DHS #2 had high activity in the cortex. Furthermore, given the extensive tiling of the region, the boundaries of activity could be determined at both the 5' and 3' ends of DHS #2.

Next, we present a brain DHS region with high *cis*-regulatory activity in the cerebral cortex (Figure 3.6D). A critical region of ~150 bp in length was identified that overlapped a block of phylogenetic conservation. Incremental loss of bases in this region resulted in progressive decreases in *cis*-regulatory activity. Within this critical region, two TF motifs were

identified: a consensus E-box motif (recognized by bHLH TFs) (Massari and Murre 2000), immediately next to a motif recognized by basic region leucine zipper (bZIP) proteins of the AP-1 family (Heinz et al. 2010). Like neural bHLH proteins, AP-1 family proteins are known to have important roles in regulating gene expression in the cerebral cortex (Raivich and Behrens 2006; Mongrain et al. 2011).

Additional examples of truncation mutation analysis are presented in Figure 3.S8. Overall, we identified 46 retinal DHSs and 13 brain DHSs with examples of truncation mutation analysis, thus representing 4.6% and 1.3% of the 1000 retinal DHSs and 1000 brain DHSs initially targeted in the library, respectively. We observed that for the loci with truncation mutation analyses, at least 8 barcoded constructs tiled across the DHS. For DHSs with at least 8 assayed barcodes, the fraction of loci with truncation mutation analyses was about 3-fold higher: 46/363 (12.7%) of retinal DHSs and 13/345 (3.8%) of brain DHSs.

Truncation mutation analyses rely on assaying long CRE fragments that tile across CRE regions. Previously, we conducted a CRE-seq enhancer study (White et al. 2013) in which short (84 bp) CREs (synthesized by oligonucleotide array) were assayed upstream of a rod photoreceptor-specific proximal promoter. These short CREs corresponded to retinal CRX ChIP-seq peaks, which are essentially a subset of retinal DHSs (Wilken 2015). Thus, we wondered whether, for a given CRE, our capture-and-clone approach identified active *cis*-regulatory sequences beyond the central region tested by the short CRE. Overall, there were 176 CRE regions in the White et al. library that overlapped with assayed regions in the current library, all of which corresponded to retinal DHSs. Most (141/176 or 80%) regions were more active as short enhancers than as long autonomous elements (Figure 3.S9A). This is not surprising, as it is known that some photoreceptor CREs exhibit strong activity as enhancers but minimal activity

as autonomous elements (Corbo et al. 2010). Interestingly, in a minority (13/176 or 7%) of cases, the long autonomous elements exhibited substantially more activity, likely because they encompassed functional regions (e.g., critical regions and/or phylogenetically conserved regions) that were not found within the short CREs, as illustrated in Figure 3.S9B and 3.S9C. Although the comparison of these two studies is limited by the differences in assay platforms and the small number of shared CREs, these results indicate that the capture-and-clone approach can provide additional *cis*-regulatory information beyond that of short CREs.

Together, these examples illustrate that CRE-seq multiplex truncation mutation analysis can identify both known and novel critical regions. In some cases, the spatial resolution is high enough to pinpoint candidate TF motifs required for activity. Thus, our assay has the ability not only to measure the overall activity of a candidate CRE, but also to demarcate the spatial boundaries of *cis*-regulatory activity.

3.4.7 Traditional reporter assays confirm that critical bases identified by CRE-seq truncation mutation analysis are required for activity

To validate the ability of CRE-seq truncation mutation analysis to identify critical regions *de novo*, we utilized traditional reporter assays. We previously developed a quantitative fluorescence reporter assay in retinal explants that accurately measures CRE activity (Montana et al. 2011b; Kwasnieski et al. 2012). Thus, we selected three retinal DHS loci (including R64, which is the locus depicted in Figure 3.6B) with critical regions identified by CRE-seq truncation mutation analysis to test with the traditional approach (Figure 3.7A). These critical regions contained bioinformatically predicted CRX sites, thus allowing us to test whether these CRX sites were required for *cis*-regulatory activity.

For each locus, we created a ‘long’ construct, a ‘short’ construct missing the critical region, and a ‘mutant’ construct identical to the ‘long’ construct except that a single point mutation was introduced in the predicted CRX site (Figure 3.7A). The point mutation was an adenine-to-cytosine substitution at the fourth position of the CRX motif (thymine-to-guanine in the reverse orientation), which is predicted to inactivate the CRX site (Supplemental Table S7) (Lee et al. 2010; White et al. 2013). The constructs were directionally cloned upstream of the minimal promoter-GFP cassette in a non-AAV vector without barcodes in the 3’ UTR, thus controlling for any effects of orientation, AAV vector sequence, or barcode sequence.

Each construct was individually electroporated into multiple retinas and quantified relative to a loading control, *Rho*-CBR3-DsRed (Figure 3.7B). We observed that in each case, the long construct showed high activity, while the short construct showed extremely low activity. Notably, the mutant construct exhibited a low level of activity comparable to the activity of the short construct (Figure 3.7C). Thus, for all three loci, we not only verified that the critical regions are required for activity, but also that these specific CRX sites are required. These experiments demonstrate that our approach identifies *bona fide* TF binding sites required for activity.

3.5 DISCUSSION

Here, we described an innovative ‘capture-and-clone’ approach for synthesizing CRE-seq libraries. We furthermore demonstrated the feasibility of using AAV-mediated CRE-seq to conduct massively parallel *cis*-regulatory analysis in the cerebral cortex *in vivo*. By comparing retina and cerebral cortex, we showed tissue-specific *cis*-regulatory activity of DHSs. By taking advantage of the truncation mutation analysis afforded by the tiling of captured fragments across targeted loci, we illustrated high-resolution, multiplex functional parsing of CREs.

Previously, high-throughput functional assays of CRE activity had been technologically limited with regards to the length of CREs that could be readily assayed (Levo and Segal 2014; Shlyueva et al. 2014). Our capture-and-clone approach provides a strategy for assaying candidate CREs with lengths of a desired range. Moreover, the capture approach can be used in conjunction with any existing MPRA-like approach, including those that already rely on DNA fragmentation (Dickel et al. 2014; Murtha et al. 2014). For example, STARR-seq (Arnold et al. 2013) has been used to assess long DNA fragments obtained by whole-genome shotgun cloning of the *Drosophila* genome. However, the mouse and human genomes are ~25 times larger than the fly genome. Moreover, only ~5-10% of the mammalian genome is thought to be functionally constrained (Graur et al. 2013; Kellis et al. 2014; Rands et al. 2014). Therefore, whole-genome shotgun cloning of mammalian genomes for *cis*-regulatory analysis is impractical. Instead, capture-and-clone permits targeted *cis*-regulome analysis.

We note that another group has recently coupled capture technology to STARR-seq (i.e., CapSTARR-seq) (Vanhille et al. 2015). Our approach differs from CapSTARR-seq in two key ways (Supplemental Table S8). First, we achieved higher on-target rates of capture (98.5% vs. 14%) due to a rigorous capture protocol to avoid non-specific pull-down of off-target DNA

(Gnirke et al. 2009; Lee et al. 2009). Second, we conducted paired-end sequencing of the input library, whereas CapSTARR-seq mapped only one end of the fragments. Thus, we were able to harness the potential of capture-and-clone for truncation mutation analysis.

Capture-and-clone allows the testing of longer CREs, which presumably harbor more *cis*-regulatory information. However, there was essentially no correlation between fragment length and CRE activity. What accounts for this observation? One consideration is that the size range of assayed CRE fragments was relatively narrow. Another explanation, based on the truncation mutation analyses, is that some long fragments exhibited low activity due to the omission of critical regions. A third possibility is that some long CRE fragments included repressive sequences that decreased activity (Reynolds et al. 2013).

The capture-and-clone approach is particularly well suited for screening thousands of candidate CREs and identifying the most active CREs in a particular tissue of interest, thereby narrowing the list of CREs that may be relevant to a particular phenotype. For instance, genome-wide association studies (GWAS) and whole-genome sequencing studies have generated lists of thousands of disease-associated non-coding variants (Ward and Kellis 2012b; Albert and Kruglyak 2015). To prioritize these lists and thereby accelerate the identification of causal variants, the locations of the candidate variants can be intersected with the locations of putative CREs. The *cis*-regulomes of unaffected and affected individuals can then be screened by capture-and-clone CRE-seq to identify CREs that exhibit the greatest differential activity between the unaffected and affected groups. Capture-and-clone is thus complementary to CRE-by-synthesis, which is better suited to precisely measuring the effects of specific variants (Levo and Segal 2014). Capture-and-clone can be used to assess a broad range of regions in any organism whose DNA and reference genome are available, although certain types of sequences are not amenable

to targeted capture, namely repetitive regions (due to non-specific pull-down) and sequences with very high (>65%) or low (<25%) GC content (Mertes et al. 2011).

Prior to our study, the implementation of MPRA in mammalian cells had been almost exclusively restricted to immortalized cell lines and cultured tissues (Shlyueva et al. 2014). The only mammalian tissue that had been assayed *in vivo* was the mouse liver, due to its ability to take up limited amounts of plasmid DNA via a hydrodynamic tail vein assay (Herweijer and Wolff 2007; Patwardhan et al. 2012). Here, we take a step forward by using AAV to conduct CRE-seq *in vivo* in the mammalian CNS.

One potential drawback of AAV is that packing constraints limit the size of the insert to less than 4.7 kb (Wu et al. 2010). Lentiviruses have greater carrying capacity (Kumar et al. 2001), but their integration into the host genome poses the risk of integration site *cis*-regulatory effects (Clark et al. 1994). By contrast, AAV-mediated CRE-seq measures the *cis*-regulatory potential of elements independent of chromosomal context, thereby interrogating the function of the DNA sequences themselves. Interestingly, there is evidence that despite being episomal, the AAV vector is organized into nucleosomes (Penaud-Budloo et al. 2008). Another limitation of AAV is that the onset of expression is relatively slow, with maximal expression requiring up to several weeks (Day et al. 2014). This delay is due to the required conversion of the genome from single-stranded into double-stranded DNA. Recently, self-complementary AAV (scAAV) serotypes have been developed that exhibit more rapid transgene expression (McCarty 2008). As novel AAV serotypes for gene therapy continue to emerge (Wu et al. 2006; Daya and Berns 2008), AAV-mediated CRE-seq will become increasingly powerful.

Why are some tissue-specific DHSs active and others inactive, even when assayed in the appropriate tissue? One reason is that DHSs demarcate not only active enhancers but also other

types of regulatory elements (e.g., silencers and insulators) (Gross and Garrard 1988; Thurman et al. 2012). Here, we used a TATA-box containing minimal promoter to assay the autonomous *cis*-regulatory activity of the tested elements, rather than a tissue-specific proximal promoter to assay for enhancer/silencer activity (Butler and Kadonaga 2002). Only a minority (~10-20%) of mammalian promoters contain TATA boxes (Sandelin et al. 2007). Future use of tissue-specific proximal promoters may allow for more sensitive assays, especially as enhancer-promoter compatibility and TATA-box vs. DPE-containing promoters become better understood (Sandelin et al. 2007; van Arensbergen et al. 2014; Zabidi et al. 2015). Additionally, since some enhancers become active only in response to particular stimuli (Ostuni et al. 2013; Shlyueva et al. 2014), environmental perturbations may be necessary to unmask their *cis*-regulatory potential. Furthermore, the *cis*-regulatory landscape of a given tissue is dynamic across development, as illustrated by DNase-seq in the developing mouse retina and brain (Wilken 2015). Future CRE-seq experiments at multiple developmental stages will help elucidate the temporal dynamics of CREs. Nonetheless, even with the TATA-box containing minimal promoter assayed in steady-state conditions, we demonstrated tissue-specific CRE activity.

Assaying autonomous activity and assaying enhancer activity are complementary approaches, as they appear to reflect different biological activities and properties of a given CRE. In the current study, we observed that GC content was associated with decreased autonomous CRE activity in the retina. Given the differences in the assays, this finding does not contradict our earlier retinal CRE-seq study (White et al. 2013), in which we observed a positive association between GC content and enhancer activity. In fact, the current result is consistent with a recent CRE-seq study in which GC content was associated with decreased autonomous activity of predicted enhancers in cell culture (Kwasnieski et al. 2014)..

In our study, the retina exhibited a stronger preference for retinal DHSs than the cerebral cortex exhibited for brain DHSs. Several explanations are possible. First, the cellular complexity of the brain is likely a major factor (Wurmbach et al. 2002). A recent DNase-seq study in the mouse brain observed that DHSs could be found around genes expressed in only a small percentage of neurons, such as cortical laminar-specific genes (Wilken 2015). Thus, a given ‘brain DHS’ may actually be a cell type-specific DHS that is active in a small population of cells. When averaged over the entire population of assayed cells, the cell type-specific activity of the DHS may be obscured. For tissues with highly heterogeneous cell populations such as the cerebral cortex, it should be possible to target specific subpopulations by combining AAV-mediated CRE-seq with fluorescence-activated cell sorting (FACS) of defined cell types (Okaty et al. 2011; Gisselbrecht et al. 2013; Dickel et al. 2014). Second, the minimal promoter used in this study contains a possible weak CRX site, whose affinity is predicted to be ~10% that of the CRX consensus motif (Chen and Zack 1996; Lee et al. 2010). Lastly, although DNA barcode representation was similar in the retina and cerebral cortex, the difference in delivery methods for the two tissues may have been a contributing factor.

In summary, we have developed a powerful and efficient strategy for constructing CRE-seq libraries that extends the size range of the CREs that can readily be assayed, using targeted *cis*-regulome capture. At the same time, we have demonstrated the feasibility of conducting CRE-seq *in vivo* in a mammalian tissue using AAV. As new assays for rapidly identifying the locations of putative cell type-specific CREs are developed, e.g., ATAC-seq (Buenrostro et al. 2013), our study sets the stage for the high-throughput functional screening of thousands of candidate CREs in a range of cell types and in a variety of model systems, including non-human

primates and human induced pluripotent stem cell (iPSC)-derived organoids (Lancaster et al. 2013).

3.6 METHODS

3.6.1 Animals

Mice were maintained on a 12-hour light/dark cycle at ~20-22 °C with free access to food and water. Neonatal mice were euthanized by decapitation, and adult animals were euthanized with CO₂ anesthesia followed by cervical dislocation, unless otherwise stated. All experiments were conducted in accordance with the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health, and were approved by the Washington University in St. Louis Institutional Animal Care and Use Committee.

3.6.2 Reference genome

The mouse reference genome used throughout was mm9.

3.6.3 Identification of target tissue-specific DHS peaks

We downloaded DHS data in narrowPeak format from the Mouse ENCODE Project (Yue et al. 2014) for the following tissues (GEO sample accessions are listed): whole brain age E14.5 (GSM1014197, replicate 1), whole brain age E18.5 (GSM1014184, replicate 1), whole brain age 8 weeks (GSM1014151, replicate 1), retina age P1 (GSM1014188), retina age P7 (GSM1014198), retina age 8 weeks (GSM1014175), liver age E14.5 (GSM1014183, replicate 1), liver age 8 weeks (GSM1014195, replicate 1), lung age 8 weeks (GSM1014194, replicate 1), kidney age 8 weeks (GSM1014193, replicate 1), thymus age 8 weeks (GSM1014185, replicate 1), and heart age 8 weeks (GSM1014166, replicate 1). We parsed these data using custom Perl scripts, tallying the number of reads per 150 bp block across the mouse genome to give a DHS ‘score’. We then examined the top ~4,000 tissue-specific peaks each for brain age 8 weeks,

retina age 8 weeks, heart age 8 weeks, and liver age 8 weeks. For a peak to be identified as ‘tissue-specific’, it was required to have a DHS score of >25 in the 8 week tissue of interest and <25 in samples derived from other tissues (but the peak score for samples deriving from different developmental stages of the same tissue type were not required to be <25). For instance, if the score for a retina age 8 weeks peak was >25 and the score for the corresponding retina age P7 peak was >25 , but all non-retinal peaks were <25 , then that peak was called ‘retina-specific’. After removing any tissue-specific peaks that overlapped repetitive genomic sequences (~10% of peaks), we selected the 1,000 peaks with the highest tissue-specific peak scores from each of adult brain, retina, heart, and liver for inclusion as capture targets.

3.6.4 Capture bait library design and synthesis

Baits were synthesized by MYcroarray. For each of the 4,000 target regions, seven 80 bp baits were designed to tile across the 300 bp region (sliding 37 bp at a time), for a total of 1.2 Mb and 28,000 baits. To check for potential off-target bait hybridization, bait candidates were blasted against the mm9 genome, which was masked for the regions from which baits were designed. By definition, T_m is the temperature at which 50% of the molecules are hybridized. Bait candidates were accepted only if no BLAST hits (Altschul et al. 1990) with a predicted $T_m > 40.0$ °C were found.

3.6.5 GREAT analysis and Gene Ontology

GREAT v2.0.2 analysis with mm9 as the reference genome was implemented, using the ‘single nearest gene’ within 1000 kb as the algorithm for associating genomic regions to genes, and using the whole genome as background and excluding the ‘include curated regulatory

domains' option (McLean et al. 2010). The input to the GREAT analysis was the list of 4,000 target DHS regions. Gene Ontology (GO) (Ashburner et al. 2000) enrichment analysis for 'biological process' in *Mus musculus* was implemented using PANTHER (Mi et al. 2005) with AmiGO 2 v2.1.4 (Carbon et al. 2009). The input to the GO analysis was the GREAT-generated list of genes associated with target DHSs ('region-to-gene' associations).

3.6.6 Restriction enzymes and PCR reagents

Unless otherwise indicated, restriction enzymes were from New England Biolabs, and Phusion Hot Start Flex 2X Master Mix (New England Biolabs) was used for PCR. Primer sequences are listed in Supplemental Table S9.

3.6.7 Preparation of gDNA for capture

Genomic DNA was purified from liver tissue of C57BL/6J mice and sonicated with Covaris E210 (duty 10%, intensity 4, cycles/burst 200, time 100 s). The freshly sonicated DNA was end repaired, 3' adenylated, ligated to commercial adapters, and enriched by PCR, using the TruSeq LT or TruSeq Nano Kit (Illumina) according to manufacturer's instructions (1 ug or 200 ng input gDNA, and 10 or 8 cycles of PCR, respectively). For final size selection and purification prior to capture, the samples were gel electrophoresed on 2% low melting point agarose and gel extracted with MinElute (Qiagen). To concentrate the samples in preparation for capture, the samples were speed vacuumed in LoBind tubes (Eppendorf).

3.6.8 *Cis*-regulome capture and preparation for cloning

Capture was conducted in a similar manner as previously described (Gnirke et al. 2009). Two rounds of sequential capture were conducted to achieve high on-target rates (Lee et al. 2009). Briefly, for the first round of capture, a 9 μ L library mix was prepared, consisting of ~300 ng input (TruSeq LT or TruSeq Nano gDNA library), 2.5 μ g human Cot-1 DNA, 2.5 μ g salmon sperm DNA, and 0.6 μ L adapter blocking agent (MYcroarray). This solution was denatured at 95 $^{\circ}$ C for 5 min. Meanwhile, a 36.8 μ L hybridization mix was prepared, consisting of 5 μ L 20X SSPE (instead of the standard 20 μ L), 0.8 μ L 0.5 M EDTA, 8 μ L 50X Denhardt's, 8 μ L 1% SDS, and 15 μ L RNase-free water. This solution was prewarmed at 65 $^{\circ}$ C for 3 min. A 6 μ L capture bait mix was prepared, consisting of 50 ng (instead of the standard 500 ng) biotinylated baits and 1 μ L of SUPERase-In (Ambion). This solution was prewarmed at 65 $^{\circ}$ C for 2 min. Finally, 7 μ L of the library mix, 13 μ L of the hybridization mix, and all 6 μ L of the capture bait mix were incubated at 65 $^{\circ}$ C for ~24 hr. The reaction was then applied to Dynabeads MyOne Streptavidin C1 (Invitrogen) with washing and elution as described (Gnirke et al. 2009). Each capture reaction was purified with MinElute (Qiagen), with an elution volume of 30 μ L. Each eluate was speed vacuumed in a LoBind tube (Eppendorf) down to a volume of 3-4 μ L and used as the library 'input' for a single reaction in the second round of capture. The second round of capture was otherwise identical to the first. No PCR was conducted between the first and second rounds of capture. After the second round of capture, PCR was conducted using Ill_NotI_1XL and Ill_NotI_2XL primers (98 $^{\circ}$ C for 1 min, 14-16 cycles: 98 $^{\circ}$ C for 10 sec, 58 $^{\circ}$ C for 30 sec, 72 $^{\circ}$ C for 1 min, followed by 72 $^{\circ}$ C for 5 min). The samples were PCR purified with MinElute (Qiagen), digested with NotI-high fidelity (HF) , and gel extracted with MinElute (Qiagen). Two

independent pools of capture products were generated, with each pool deriving from multiple capture reactions.

3.6.9 CRE-seq library construction

To minimize the likelihood of cleaving captured fragments, the 8-bp cutters NotI, FseI, and AscI were employed. To create the barcoded vector library for insertion of NotI-ended captured fragments, the *Rho* basal-DsRed construct (Hsiau et al. 2007) was modified with linkers on the 3' end of DsRed to replace a former NotI site with an EagI site and to add NsiI, FseI and AscI sites, and on the 5' end of the *Rho* basal promoter to add a NotI site between XbaI and KpnI sites.

To add 15-mer barcodes, two pools of 30 nmol oligos were synthesized with random 15 bp sequences (Integrated DNA Technologies) as BC_F and BC_R. The two pools were annealed and ligated into the AscI and NsiI sites of the vector. After transformation of 5-alpha chemically competent *E. coli* (New England Biolabs) and overnight growth in liquid culture, a total of $\sim 9.5 \times 10^6$ colonies were harvested (as estimated from plating a small aliquot) and purified with the PureLink HiPure Plasmid Maxiprep Kit (Invitrogen). The barcoded vector library was then digested with EagI-HF and dephosphorylated with alkaline phosphatase (Roche). The captured fragments were digested with NotI-HF and cloned into the EagI site of the vector library with 5-alpha chemically competent *E. coli* (New England Biolabs). A total of $\sim 80,000$ colonies were scraped from LB/ampicillin agar plates, grown for ~ 2 hours in liquid LB/ampicillin culture, and purified with the PureLink HiPure Plasmid Maxiprep Kit (Invitrogen).

After paired-end sequencing to determine the CRE-barcode correspondence (described below), the minimal promoter-eGFP cassette was cloned into the FseI and AscI sites³. The minimal promoter is the previously described ‘*Rho* basal’ minimal promoter, which contains a TATA box (‘CATAA’), and which by itself does not have detectable activity in electroporated retina (Hsiau et al. 2007). The minimal promoter-eGFP cassette was created by replacing DsRed with eGFP (Zhang et al. 1996) in the *Rho* basal-DsRed construct (Hsiau et al. 2007). After transformation with 5-alpha chemically competent *E. coli* (New England Biolabs) and overnight growth in liquid culture, a total of $\sim 2.7 \times 10^6$ colonies were harvested (as estimated by plating a small aliquot) and purified with the PureLink HiPure Plasmid Maxiprep Kit (Invitrogen).

The AAV-ITR vector was prepared by digesting the pAAV2.1-*RHO*-eGFP vector (Allocca et al. 2007) with NheI and XhoI, and replacing the *RHO*-eGFP cassette with a linker containing an EagI site. To transfer the library into the AAV-ITR vector, the entire CRE-minimal promoter-eGFP-polyA cassette was subjected to PCR using 5’ Tak and NotI_polyA_R1 primers (98 °C for 1 min, 10 cycles: 98 °C for 10 sec, 64 °C for 30 sec, 72 °C for 1 min 30 sec, followed by 72 °C for 5 min). The PCR product was digested with NotI-HF (New England Biolabs) and cloned into the EagI site of the AAV-ITR vector. After transformation of 5-alpha chemically competent *E. coli* (New England Biolabs) and overnight growth in liquid culture, a total of $\sim 2.5 \times 10^6$ colonies (as estimated by plating a small aliquot) were harvested and purified with the PureLink HiPure Plasmid Maxiprep Kit (Invitrogen). ITR integrity was verified by restriction digest. Note that the final NotI digestion removes any captured fragments initially cloned in as NotI multimers, leaving only the 3’-most captured fragment.

³Paired-end sequencing was conducted *prior* to insertion of the promoter-reporter cassette so that the barcode and both ends of each CRE fragment would be sequenced with 2x250 bp sequencing.

3.6.10 Paired-end sequencing for CRE-barcode correspondence

Prior to insertion of the promoter-reporter cassette, the library was prepared for paired-end sequencing as follows. PCR amplification was conducted using primers LibPCR_F and LibPCR_R (98 °C for 1 min, 8 cycles: 98 °C for 10 sec, 64 °C for 30 sec, 72 °C for 1 min, followed by 72 °C for 5 min). The product was digested with NotI-HF and SacII, gel purified with MinElute (Qiagen), and ligated to P1_NotI and PE2_SacII adapters with T4 DNA ligase (New England Biolabs), using an equimolar mix of P1_NotI indexed adapters to facilitate nucleotide balance. The ligation products were PCR amplified to enrich for molecules that had both P1 and PE2 adapters, using primers JKP4F and JKP4R (98 °C for 1 min, 14 cycles: 98 °C for 10 sec, 65 °C for 30 sec, 72 °C for 1 min, followed by 72 °C for 5 min). The final product was gel-extracted on 2% low melting point agarose and verified on an Agilent Bioanalyzer. Two lanes of MiSeq 2x250 bp sequencing were run at a loading concentration of 1.6-2 pM and 12-15% spiked-in Phi-X DNA (Illumina).

3.6.11 Analysis of paired-end sequencing for CRE-barcode correspondence

Barcodes and captured fragment sequences were extracted based on flanking bases. Captured fragment sequences were aligned as paired reads to mm9 using Bowtie 2 v2.1.0 (Langmead and Salzberg 2012) with an allowed maximum insert size of 1000 bp ('-X 1000' setting). SAM files were converted to BAM files using SAMtools v0.1.19 (Li et al. 2009) and then to BED files using BEDTools v2.22.1 (Quinlan and Hall 2010). Only paired reads that mapped concordantly were used. Fragments were examined for overlap with the 4,000 target DHS regions (which were each 300 bp). If a fragment overlapped two adjacent target regions, it was assigned to the target region with the most bases of overlap. Barcodes were required to be

14-16 bp in length. Barcodes with multiple CRE fragment associations, and PCR-duplicate CRE fragments associated with multiple barcodes (~1.6% of fragments), were discarded. A list of ‘on-target’ CRE correspondences for 45,670 barcoded constructs (minimum 10 reads) resulted. To determine the ‘off-target’ rate, the number of barcoded constructs that did not overlap a target DHS was found to be 712. Hence, ~98.5% of fragments were on-target.

3.6.12 Retinal explant electroporation and culture for CRE-seq

Electroporation and explant culture of mouse retinas were performed as described previously (Montana et al. 2011b). In brief, retinas were dissected from newborn (P0) CD-1 mouse pups and coelectroporated with 0.5 $\mu\text{g}/\mu\text{L}$ AAV-ITR plasmid CRE-seq library and 0.5 $\mu\text{g}/\mu\text{L}$ *Rho*-CBR3-DsRed, a rod-specific construct for visualizing electroporation efficiency (Corbo et al. 2010). Retinas were grown in explant culture and harvested 8 days later. Five retinas were pooled for each CRE-seq biological replicate.

3.6.13 Viral production

Recombinant AAV9(2YF) was produced and purified as previously described (Grieger et al. 2006). To summarize, HEK293 cells at ~80% confluency were cotransfected with the AAV-ITR plasmid CRE-seq library, p-Helper plasmid, and AAV9(2YF) rep/cap plasmid (Dalkara et al. 2012). Cells were harvested 72 hours after transfection, and the virus was purified by Iodixanol gradient ultracentrifugation, followed by buffer exchange. The viral titer, as determined by dot blot or quantitative PCR, ranged from 5×10^{12} to 1×10^{14} vg/mL (Zolotukhin et al. 2002; Aurnhammer et al. 2012).

3.6.14 Stereotactic cortical injection

Stereotactic cortical injections were performed by the Hope Center Animal Surgery Core at Washington University in a manner similar to that described (Cetin et al. 2006). Briefly, female CD-1 mice (age 4-6 weeks) were anesthetized with isoflurane. Each mouse received bilateral injections. For each injection, a small craniotomy was performed and 1 μ L of AAV9(2YF) CRE-seq library was delivered into the primary motor cortex (stereotactic coordinates: dorsal/ventral axis 0.52 mm, anterior/posterior axis 1 mm, medial/lateral axis 1.5 mm). Animals were harvested 4-5 weeks after injection. The brain was sliced coronally and a fluorescent dissecting scope (Leica MZ16 F) was used to visualize GFP-positive regions, which were isolated by microdissection. Each CRE-seq biological replicate consisted of GFP-positive cortical tissue from a single animal.

3.6.15 Isolation of RNA and DNA and preparation for sequencing

Tissues were rapidly harvested and rinsed in cold sterile HBSS with calcium and magnesium (Gibco) and stored at -80 °C in TRIzol (Invitrogen). Samples were homogenized in TRIzol, and RNA and DNA were isolated according to the manufacturer's instructions. RNA samples were treated with TURBO DNase (Ambion) to remove potential DNA contamination. RNA and DNA were prepared for sequencing essentially as previously described (Kwasnieski et al. 2012). RNA was reverse-transcribed with SuperScript III (Invitrogen) using oligo-dT primers. The resulting first-strand cDNA was treated with RNaseH. Both the cDNA and DNA samples were subjected to PCR to amplify the barcode sequence in the 3' UTR of GFP using the forward primer SSP1F and the reverse primer JKP3R (98 °C for 1 min, 22 cycles for DNA or 26 cycles for cDNA: 98 °C for 10 s, 60 °C for 30 s, 72 °C for 30 s, followed by 72 °C for 5 min). This

resulted in PCR products flanked by EagI and EcoRI restriction enzyme sites. The products were purified with PureLink PCR Purification Kit (Invitrogen) and digested with EagI-HF and EcoRI. After digestion, the samples were gel purified with Qiagen Gel Extraction Kit and ligated to P1_EagI and PE2_EcoRI adapters using T4 DNA ligase (New England Biolabs). To enrich for molecules that had both P1 and PE2 adapters, the ligation products were PCR amplified with primers JKP4F and JKP4R (98 °C for 1 min, 20 cycles: 98 °C for 30 sec, 65 °C for 30 sec, 72 °C for 30 sec, followed by 72 °C for 5 min). The final product was gel purified from 2% low melting point agarose and verified on an Agilent Bioanalyzer.

3.6.16 Illumina sequencing for CRE-seq barcode abundance

For each tissue, the three cDNA samples and three corresponding DNA samples were multiplexed and run on a single lane of Illumina HiSeq 2000 (1x50 bp) at a loading concentration of 8 pM with 10% spiked-in Phi-X DNA.

3.6.17 CRE-seq data analysis

Samples were demultiplexed and the barcode was extracted based on flanking sequences. Reads were tabulated to obtain the raw RNA and DNA counts for each barcode. Only barcodes with at least 10 raw DNA reads in all 3 biological replicates of a tissue were included (36,005 barcodes for retina and 38,826 barcodes for cerebral cortex). For each barcode, the RNA count was normalized to the total RNA counts in the sample, and the DNA count was normalized to the total DNA counts in the sample. The normalized expression was the ratio of the normalized RNA count to the normalized DNA count. A pseudocount of 0.001 was added to the normalized

expression, and the \log_2 was taken. The average of the \log_2 values across biological replicates was the ‘mean expression (\log_2 units)’.

3.6.18 Histology

Retinal explants were rinsed twice with PBS and fixed in 4% paraformaldehyde/PBS for 30-60 min at room temperature, equilibrated in 30% sucrose/PBS, and embedded in Tissue-Tek O.C.T. (Sakura). Retinal cryosections (12-14 μm) were prepared and stored at $-20\text{ }^\circ\text{C}$ until imaging. For stereotactically injected brains, animals were deeply anesthetized with ketamine/xylazine and then transcardially perfused with heparin/PBS followed by 4% paraformaldehyde/PBS. Animals were decapitated and the brains were dissected in PBS and post-fixed in 4% paraformaldehyde/PBS at $4\text{ }^\circ\text{C}$ for at least a day. Vibratome sections (200 μm) were prepared from agarose-embedded brain slices and then optically cleared with glycerol/PBS (Selever et al. 2011). Brain slices were treated with sodium borohydride to minimize autofluorescence (Clancy and Cauller 1998). For anti-RBFOX3 (also known as anti-NeuN) staining of free-floating vibratome sections, the sections were blocked with 4% normal donkey serum (NDS)/0.25% Triton X-100/PBS for at least 1 hr at room temperature with gentle agitation, incubated with rabbit anti-RBFOX3 antibody (ABN78; EMD Millipore) (1:50, diluted in 4% NDS/0.1% Triton X-100/PBS) overnight at $4\text{ }^\circ\text{C}$ with gentle agitation, washed with 0.1% Triton X-100/PBS, incubated with Alexa Fluor 555 donkey anti-rabbit (A-31572; Molecular Probes) (1:800, diluted in 4% NDS/0.1% Triton X-100/PBS) for 1 hr at room temperature with gentle agitation, and washed with 0.1% Triton X-100/PBS. All brain slices were stored in PBS at $4\text{ }^\circ\text{C}$ until imaging. For imaging, tissue was mounted with Vectashield (Vectorlabs) and coverslipped. Confocal imaging was conducted with a laser confocal microscope (Zeiss LSM 700) and ZEN

2009 software (Zeiss). Flat-mount imaging of an untreated brain slice (Figure 3.3D) was conducted with an inverted fluorescent microscope (Nikon Eclipse TE300) and MetaMorph software (Molecular Devices). Images were processed with Adobe Photoshop.

3.6.19 Cluster analysis of biological replicates

Hierarchical clustering and principal component analysis (PCA) were used to assess the underlying structure of CRE expression across retina and brain replicates. For hierarchical clustering, the sample distance was defined as one minus the Pearson correlation coefficient (calculated across the normalized expression of the ~35,000 barcodes with at least 10 DNA reads in all six samples), and clustering was implemented using average linkage. PCA was performed via singular value decomposition on scaled, centered expression data (i.e., zero-centered values with unit variance).

3.6.20 Analysis of TF motif enrichment in low vs. high-expressing DHSs

To compare the motif content of low- and high-expressing constructs (Figure 3.5E), a list of brain and retina TF motifs were obtained as follows. DNase-seq reads for adult brain (GSM1014151, replicate 1) and adult retina (GSM1014175) were downloaded and aligned to mm9 with Bowtie 2 v2.2.3 (Langmead and Salzberg 2012). DNase-seq peaks were then called using MACS2 v2.1.0 (Zhang et al. 2008). For *de novo* motif discovery, peaks were first partitioned by HOMER v4.7 annotations ('promoter,' 'intronic,' and 'intergenic') (Heinz et al. 2010). *De novo* motif discovery was then performed independently for each of these classes of peaks from brain and retina, with the final motif list consisting of all motifs identified at a threshold of $p < 1 \times 10^{-50}$. To compare similar numbers of DHSs in the 'high' and 'low'

categories, individual barcoded constructs were ranked by average expression in each tissue. The highest-expressing constructs that constituted 100 distinct DHS target regions (regardless of DHS tissue origin) were classified as ‘high’ in that tissue, and the lowest-expressing constructs that constituted 100 distinct DHS target regions (regardless of DHS tissue origin) were classified as ‘low’ in that tissue (DNA read count was used to break ties). Finally, overlapping intervals were merged, and the resulting regions were scored for motif enrichment (binomial test, via HOMER) relative to a background of ~50,000 random mm9 sequences matched for size and dinucleotide content.

3.6.21 Receiver operating characteristic (ROC) curves

To quantify the extent to which sequence features and epigenomic data could predict expression (Figure 3.5F), we implemented multiple logistic regression as a means of classifying whether or not individual constructs were among those with the highest expression (similar to the approach described by (Kwasniewski et al. 2014)). Briefly, all assayed constructs (~36,000 constructs for retina and ~39,000 constructs for cerebral cortex) were partitioned by expression into ‘high’ and ‘not high’ expression groups. ‘High’ was defined here as mean expression across replicates (\log_2 units) of >-2 for constructs assayed in the retina (~95th percentile), and >2 for constructs assayed in the cerebral cortex (~99th percentile) (see Figure 3.4B). Our model included terms for GC content (averaged across the CRE fragment), phylogenetic conservation (30-way vertebrate PhastCons, averaged across the CRE fragment) (Siepel et al. 2005), brain or retina DNase-seq data ($\log_2((\text{read depth}+1)/\text{CRE size})$), retina CRX ChIP-seq data ($\log_2((1/2)*(\text{read depth of two WT CRX ChIP-seq replicates} + 1)/\text{CRE size}))$) (Corbo et al. 2010), and individual TF motifs (the number of each motif in each CRE fragment, as identified

by HOMER). CRX ChIP-seq data were only included in the retina model, and distinct TFs were considered for retina and cerebral cortex models. TF motifs for each tissue were identified as described above (17 motifs for retina, and 13 motifs for cerebral cortex; see Supplemental Table S5). Two retinal motifs (YY1 and ZBTB33) were omitted from the model, as they were observed fewer than 100 times across the ~36,000 constructs, and hence 15 motifs were in the retina TF motif model. The performance (AUC) of models was quantified using the ROCR package in R (Sing et al. 2005). Five-fold cross-validation was used to control for over-fitting.

3.6.22 Expression scores for browser screenshots

For Figure 3.6A, the scales for the heat maps are indicated. Elsewhere, heat maps were generated according to the default grayscale on the UCSC Genome Browser (Karolchik et al. 2014), using custom bed tracks that were generated as follows. For each biological replicate, a bed track was created using the useScore=1 attribute for intensity shading of individual barcoded constructs using a 'bed score'. The 'bed score' was obtained by adding 10 to the \log_2 expression and multiplying by 75. For each tissue, an 'average signal' bedGraph track was created by segmenting the tiled regions and averaging the bed scores across replicates and barcodes. A segment was required to be encompassed by at least 2 barcoded constructs to be included in the 'average signal' track. The windowing function was set to 'mean'. A smoothing window function (10 pixels) was applied to the average signal tracks, which were displayed on the following scales: 0 to 1400 for retina, and 300 to 1200 for cortex.

3.6.23 Synthesis of individual constructs for validation

The R28 constructs were cloned as EcoRV/KpnI fragments. To create the long and short R28 constructs, the R28_L/R28_R and R28_S/R28_R primer pairs were used, respectively. To create the mutant R28 construct, R28_MT was ordered as a double-stranded gene block (Integrated DNA Technologies). The R62 constructs were cloned as EcoRI/XbaI fragments. To create the long and short R62 constructs, the R62_L/R62_R and R62_S/R62_R primer pairs were used, respectively. To create the mutant R62 construct, R62_MT was ordered as a double-stranded gene block (Integrated DNA Technologies). The R64 constructs were cloned as EcoRV/KpnI fragments. To create the long, short, and mutant R64 constructs, the R64_L/R64_R, R64_S/R64_R, and R64_MT/R64_R primer pairs were used, respectively. For the PCR reactions, C57BL/6J gDNA was the template. The CREs were digested and cloned upstream of the minimal promoter-eGFP cassette in the *Rho* basal-eGFP vector, which was created from *Rho* basal-DsRed (Hsiau et al. 2007) by replacing DsRed with eGFP at XmaI and NotI sites. Test constructs were confirmed with Sanger sequencing that encompassed the entire CRE.

3.6.24 Validation of individual constructs by fluorescent reporter assays

Electroporation, explant culture, and quantification of fluorescence were performed essentially as previously described (Montana et al. 2011b). In brief, as for CRE-seq, retinas were dissected from newborn (P0) CD-1 mouse pups. Here, they were coelectroporated with 0.5 $\mu\text{g}/\mu\text{L}$ of the test construct and 0.5 $\mu\text{g}/\mu\text{L}$ *Rho*-CBR3-DsRed (Corbo et al. 2010). Retinas were cultured for 8 days, fixed, and then whole mounted for quantitative imaging of fluorescent intensity (GFP intensity normalized to DsRed intensity), using a monochromatic camera (Hamamatsu ORCA-AG) and MetaMorph software (Molecular Devices). For each retina, five

regions were quantified in ImageJ and averaged. SEM was calculated based on normalized fluorescence measurements across retinas (n = 10-12 retinas per test construct). Representative whole mount images using a color camera (Olympus DP70) were also taken.

3.6.25 Comparison with CapSTARR-seq

The raw sequence data for the CapSTARR-seq (Vanhille et al. 2015) input library (GEO accession number GSM1463994) were downloaded and mapped to mm9 with Bowtie 2 v2.1.0 (Langmead and Salzberg 2012).

3.7 DATA ACCESS

The sequence data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE68247. Custom tracks for the UCSC Genome Browser (Karolchik et al. 2014) are provided in Supplemental Table S10.

3.8 SUPPLEMENTAL TABLES

Supplemental tables are available at:

<http://genome.cshlp.org/content/suppl/2015/11/17/gr.193789.115.DC1.html>

3.9 ACKNOWLEDGEMENTS

The authors would like to thank Karen Lawrence and Jennifer Enright for contributing to the design of the barcoded vector library and sequencing adapters, Jean-Marie Rouillard of MYcroarray for capture advice, Ronald Perez of the Animal Surgery Core at the Hope Center for Neurological Disorders for stereotactic cortical injections, Mingjie Li of the Viral Vectors Core at the Hope Center for Neurological Disorders for assistance with viral production, and the Genome Technology Access Center in the Department of Genetics at Washington University School of Medicine for sequencing services. We would also like to thank Michael A. White for helpful discussion and Shuyi Ma for critical reading of the manuscript. This work was supported by the Foundation Fighting Blindness (J.G.F.), Simons Foundation Autism Research Initiative (grant number 275579 to J.C.C.) and the National Institutes of Health (HG006790 and EY018826 to J.C.C., EY022975 to J.G.F., EY024958 to J.C.C. and J.G.F, and 5T32EY013360 to S.Q.S.).

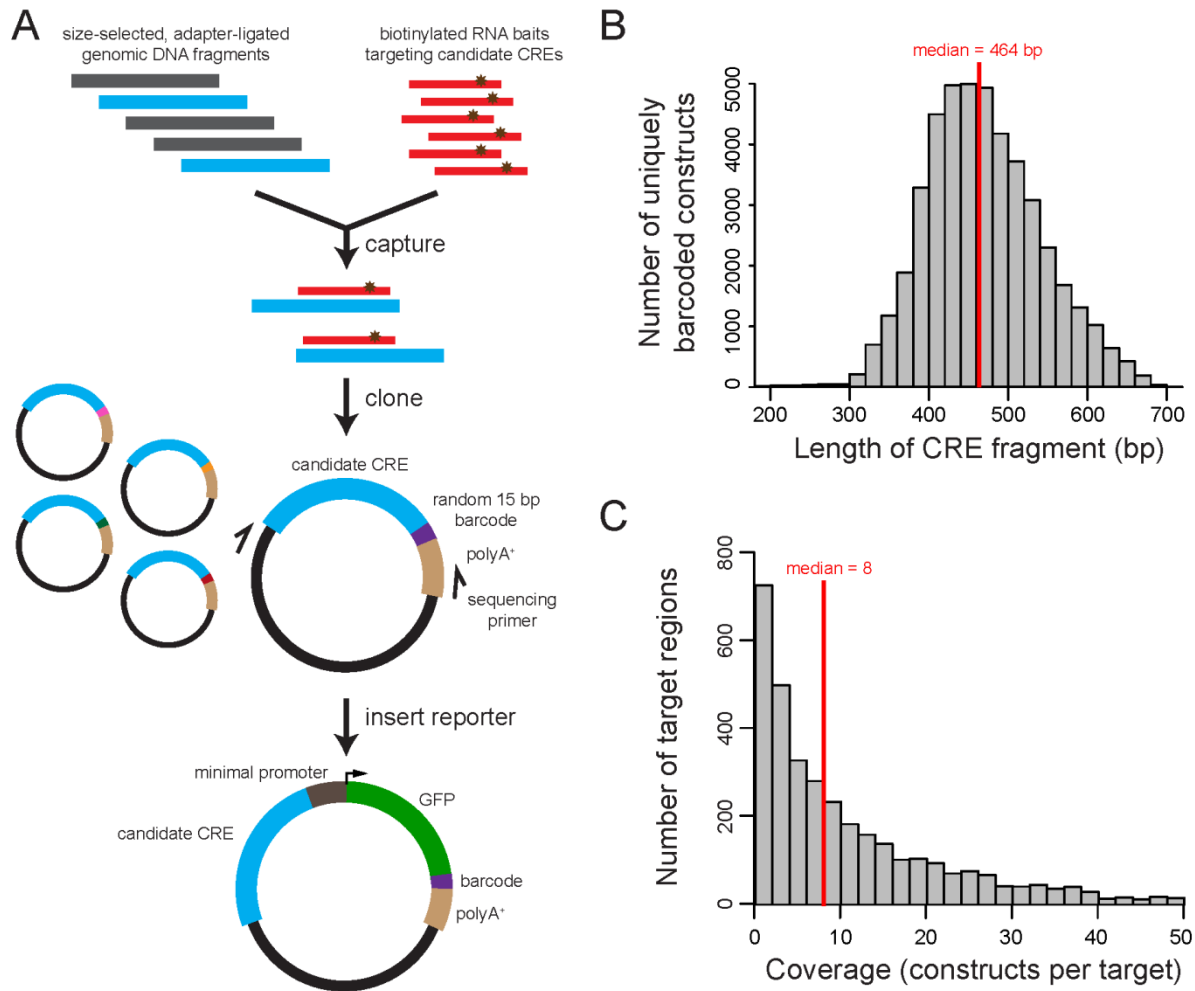


Figure 3.1. ‘Capture-and-clone’ allows synthesis of CRE-seq libraries with long CREs. (A) Schematic of the capture-and-clone approach. Size-selected, adapter-ligated genomic DNA was hybridized to biotinylated RNA baits that tiled across candidate CRE regions of interest. Captured fragments were cloned into a barcoded vector library with unique 15-mer barcodes. Paired-end sequencing revealed the CRE-barcode correspondence. A minimal promoter-GFP reporter cassette was subsequently cloned into the library. (B) Histogram showing the distribution of the lengths of captured fragments that were cloned into the barcoded vector library, based on paired-end sequencing. The median length was 464 bp. (C) Histogram showing the distribution of target coverage, i.e., the number of captured fragments that overlapped a 300 bp target region. Of the 4,000 targeted regions, 3,483 regions were represented by at least one construct. The median coverage among represented regions was 8. Not shown in graph: 517 non-represented regions and 114 target regions with a coverage of >50.

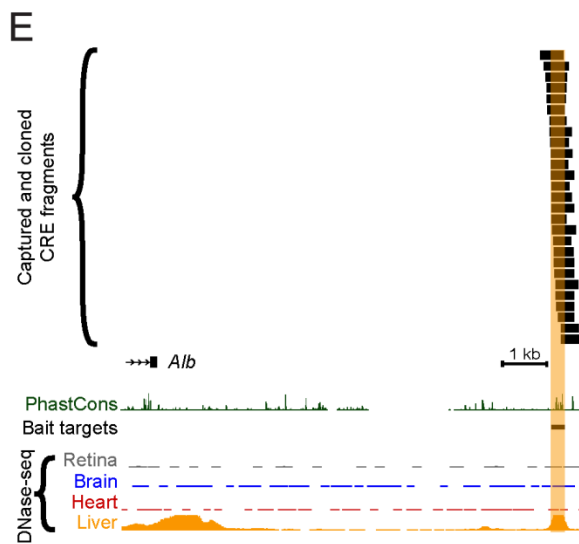
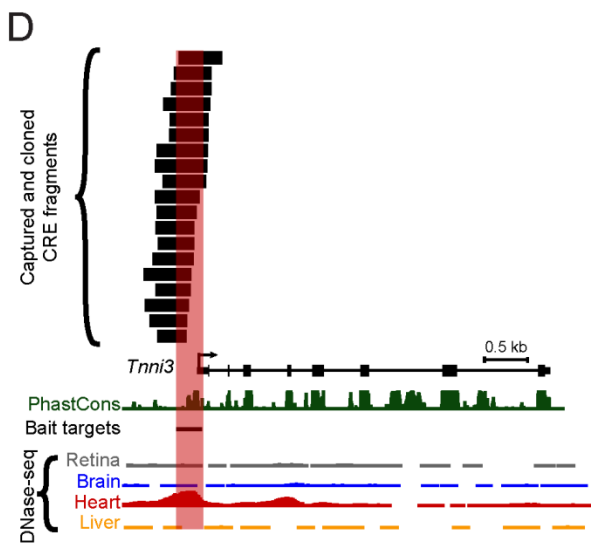
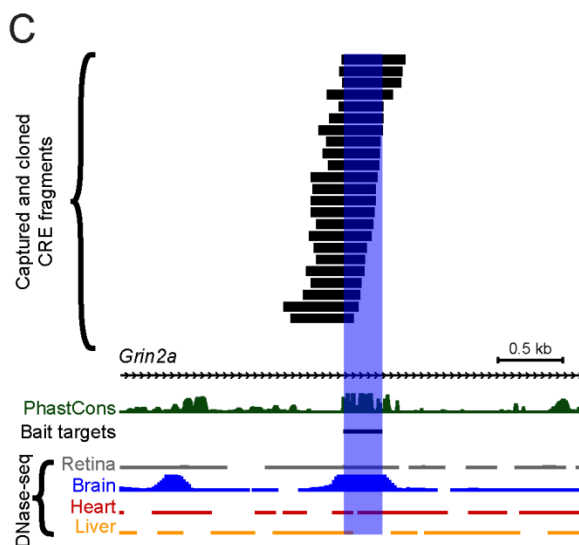
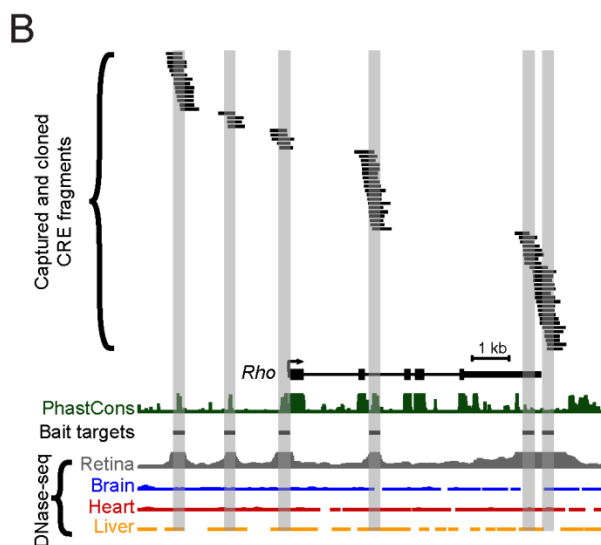
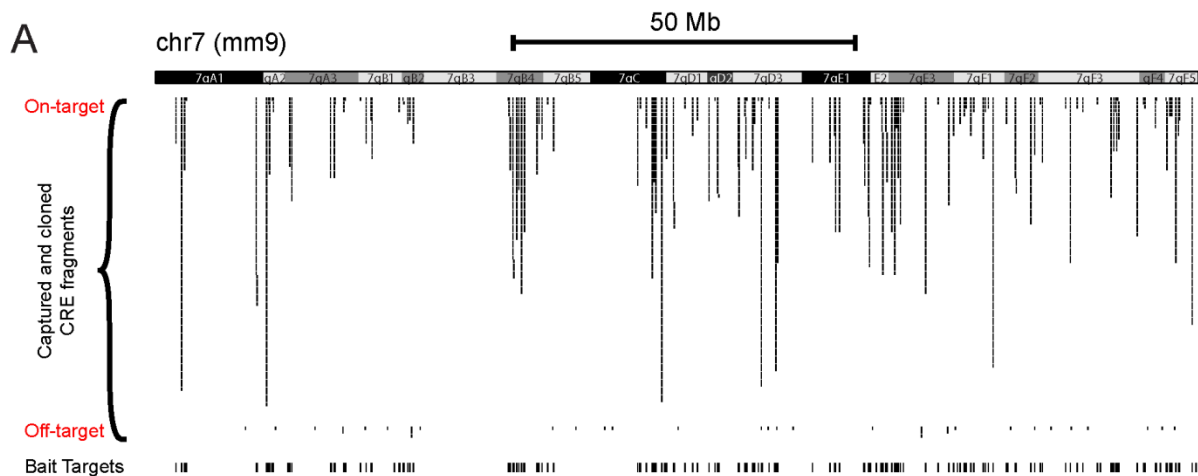


Figure 3.2. Tiling of captured fragments across target regions. Capture baits were designed based on adult (8 week old C57BL/6J) DNase-seq data from Mouse ENCODE (Yue et al. 2014). Paired-end sequencing revealed the locations of individual barcoded, captured-and-cloned fragments. The UCSC Genome Browser (mm9) (Karolchik et al. 2014) screenshots depict: (A) Captured fragments for an entire representative chromosome (chr7). ‘Off-target’ fragments, i.e., those that did not overlap a 300 bp target bait region, are also shown. Examples of captured fragments: (B) around a retina-specific locus, *Rho* (rhodopsin), (C) in an intron of a brain-specific locus, *Grin2a* (glutamate receptor, ionotropic, NMDA2a [epsilon 1]), (D) in the 5’ UTR/promoter region of a heart-specific locus, *Tnni3* (troponin I, cardiac 3), and (E) downstream of a liver-specific locus, *Alb* (albumin). Note that some DNase-seq peaks visible in the screenshots were not included as targets for capture. PhastCons depict 30-way vertebrate phylogenetic conservation (Siepel et al. 2005).

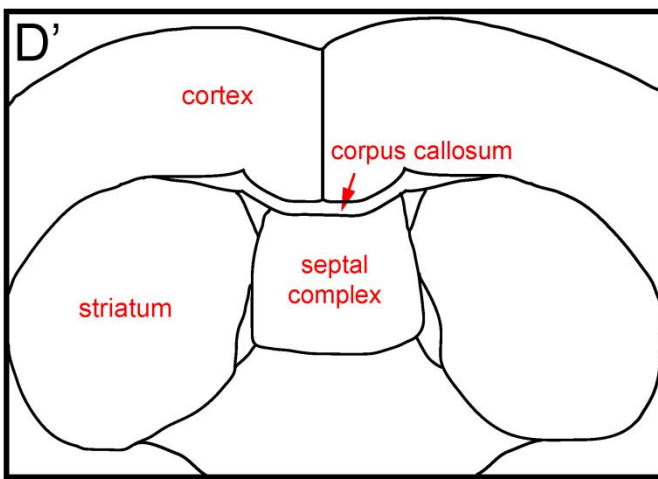
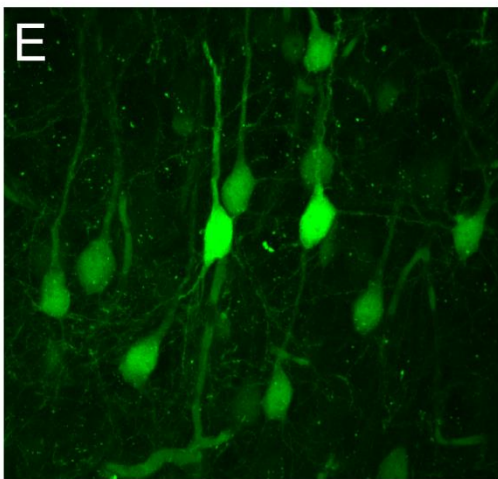
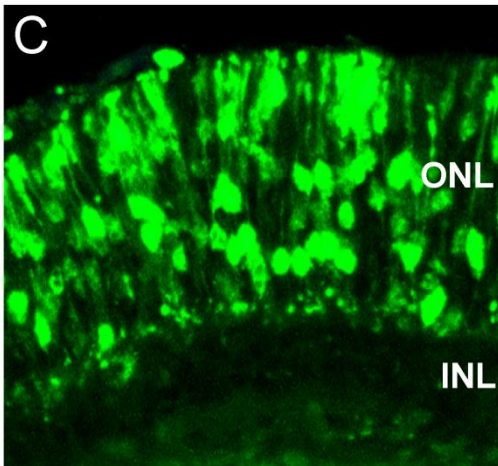
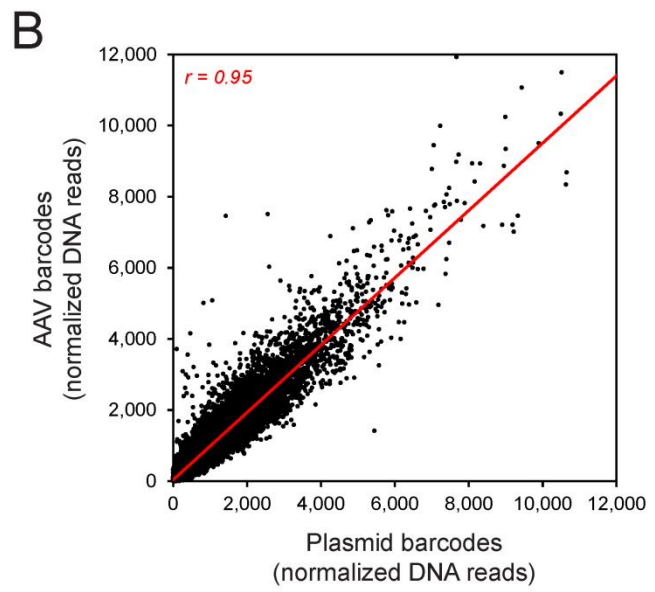
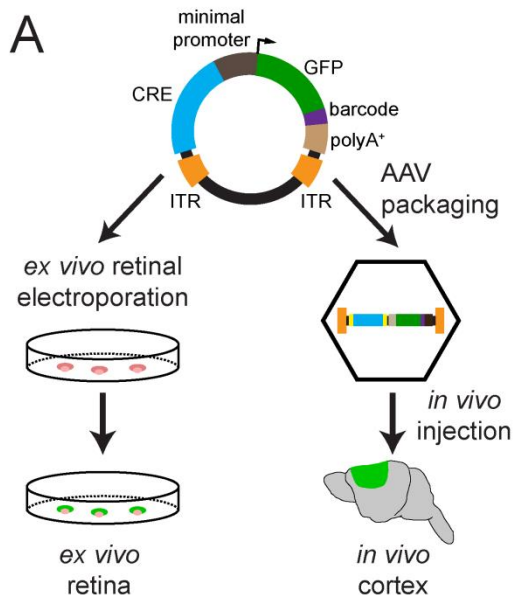


Figure 3.3. Delivery of capture CRE-seq library into mouse retina *ex vivo* and cerebral cortex *in vivo*. (A) Schematic of the CRE-seq library delivery approach. The plasmid library can be directly electroporated into the retina *ex vivo*. Alternatively, the library can be packaged into AAV and delivered via stereotactic injection into the cerebral cortex *in vivo*. (B) Scatterplot comparing the relative abundance of ~45,000 individual barcoded constructs in the plasmid library delivered into the retina, and in the AAV-packaged library delivered into cortex, as measured by barcode DNA reads summed across the three biological replicates for each tissue and then normalized to the total number of barcode DNA reads. Each data point represents a unique barcoded construct. DNA reads were well-correlated (Pearson $r = 0.95$), indicating fidelity of barcode representation after AAV packaging and delivery. Off-target constructs and constructs with 0 reads in all samples were excluded. Not shown: 4 points falling outside the depicted plot range (included in the calculation of Pearson r). Red line, linear regression. (C) Confocal image of a retina that was electroporated with the plasmid library and cryosectioned after 8 days in culture. ONL, outer nuclear layer. INL, inner nuclear layer. (D) Flat-mount image of a coronal slice from a brain injected with the AAV-packaged library bilaterally into the primary motor cortex and harvested ~4 weeks later. (D') Schematic corresponding to the flat-mount image. Note the bilateral GFP-positive regions in the cortex, as well as bundles of GFP-positive axons in the corpus callosum (red arrow). (E) Confocal image of a cortical region infected with the AAV-packaged library.

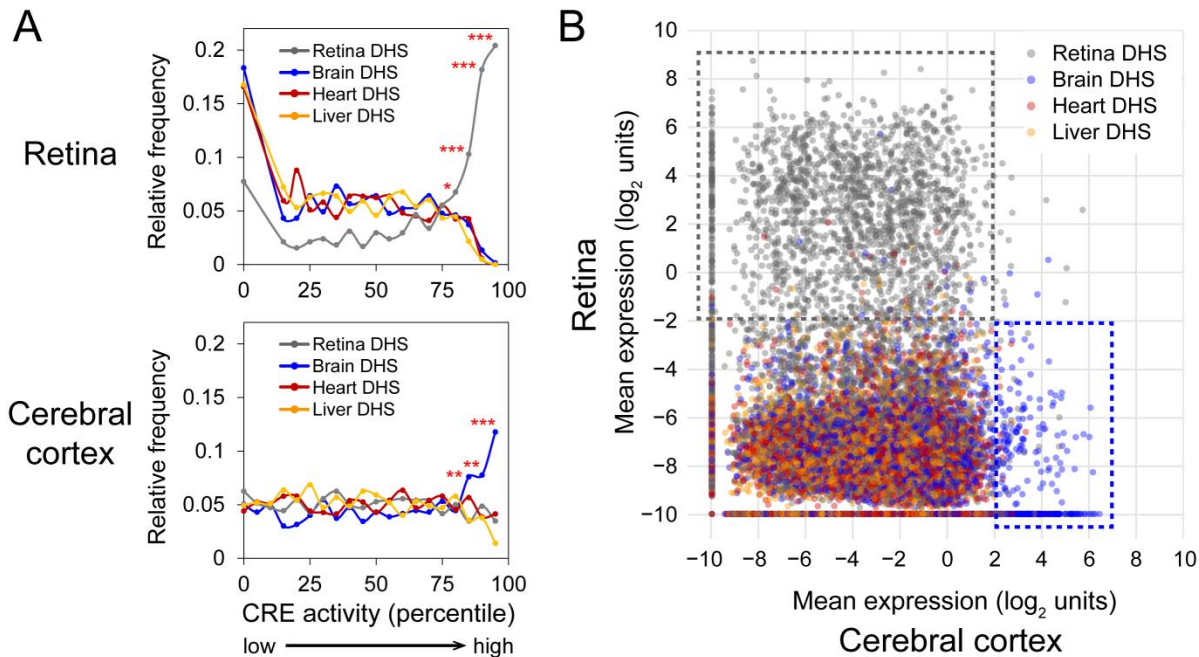


Figure 3.4. Tissue-specific *cis*-regulatory activity of DHSs. (A) Frequency distribution of DHSs ranked by *cis*-regulatory activity (bin size: 5 percentiles) as measured in the retina (top) or cerebral cortex (bottom). In the retina, ~15% DHSs had undetectable activity and hence were binned together. Averages were taken across biological replicates and barcodes for a given target DHS. Only DHSs with at least 2 barcoded constructs were included in this analysis (~3,000 DHSs). Frequencies were normalized to the total number of DHSs in each category. To test for enrichment, chi-squared test was performed (one-tailed): *** $p < 10^{-4}$, ** $p < 0.01$, * $p < 0.05$. (B) Scatterplot showing the expression of individual barcoded constructs as assayed in the cerebral cortex (x-axis) vs. retina (y-axis). Each dot represents an individual construct. For each construct, the average measurement across the three biological replicates for each tissue was taken. The ~35,000 barcodes that were well-represented (at least 10 DNA reads) in all six samples were included in the analysis. Gray, blue, red, and orange dots denote constructs with CRE fragments that overlap retina, brain, heart, and liver DHSs, respectively. The dotted gray box encompasses constructs that are strongly active in the retina, and the dotted blue box encompasses constructs that are strongly active in the cortex.

Figure 3.5. Parameters that predict CRE activity. (A) to (D) Retinal DHSs as assayed in the retina (left) and brain DHSs as assayed in the cerebral cortex (right). Each panel shows a 1 kb centered window. Only DHSs with at least 2 barcodes were included in this analysis, i.e., 710 retinal DHSs in retina (black lines, left) and 696 brain DHSs in cortex (black lines, right). The top 100 (red lines, left) and top 200 (orange lines, left) retinal DHSs expressed in the retina and the top 100 (red lines, right) and top 200 (orange lines, right) brain DHSs expressed in the cortex are shown. To compare the top 100 DHSs vs. the rest of the DHSs in each group, two-tailed student's t-test was calculated for the means within the 1 kb window, except for PhastCons scores, which was calculated within the central 100 bp. *** $p < 0.001$, ** $p < 0.01$, N.S., not significant. (A) *Cis*-regulatory activity, as measured by mean expression in \log_2 units. For each assayed DHS, at each base position across the 1 kb window, the expression values of the individual barcoded constructs whose CREs overlapped the position were averaged across biological replicates. (B) DNase-seq score (Yue et al. 2014). (C) GC content, calculated in 50 bp windows, sliding 25 bp at a time. The fractions denote the proportion of DHSs that were promoter-proximal (i.e., located within -1 kb to +100 bp relative to the nearest TSS) based on GREAT annotations (McLean et al. 2010). (D) Phylogenetic conservation as measured by 30-way vertebrate PhastCons (Siepel et al. 2005). (E) Enrichment for TF motifs among low vs. high-expressing DHSs in each tissue, without restriction on the type of DHS (see Methods). Only significant motifs are shown ($p < 0.05$ in at least one category). For motifs enriched in both tissues, the logo from the tissue with the more significant enrichment is shown. Abbreviations: HD, homeodomain; NR, nuclear receptor; ZF, zinc finger; HTH, helix-turn-helix. (F) Receiver operator characteristic (ROC) curves show the performance of logistic regression models for GC content, PhastCons, TF motifs, retina or brain DNase-seq, or a combined model. A model based on CRX ChIP-seq (Corbo et al. 2010) was included for the retina only. The area under the curve (AUC) for each model is indicated. For cross-validation results, see Supplemental Table S6.

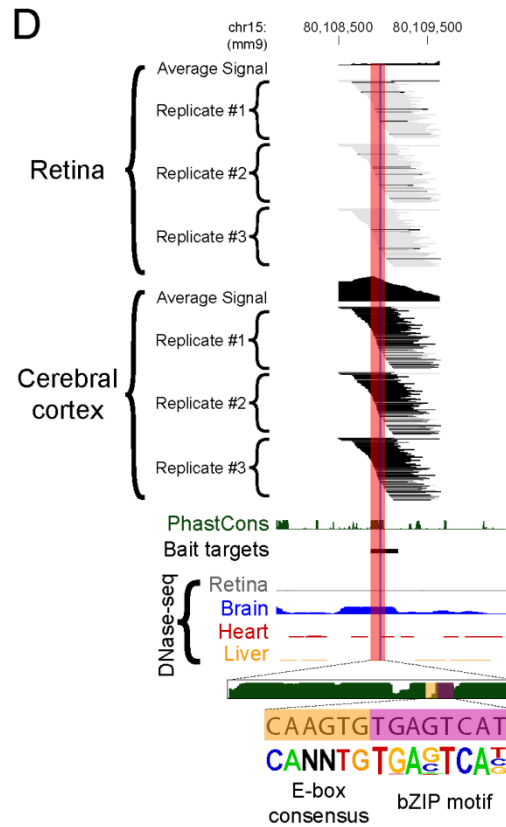
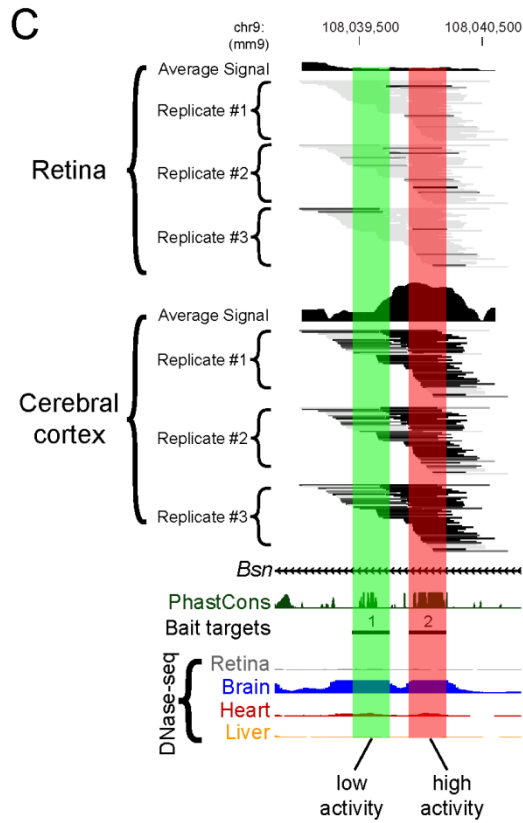
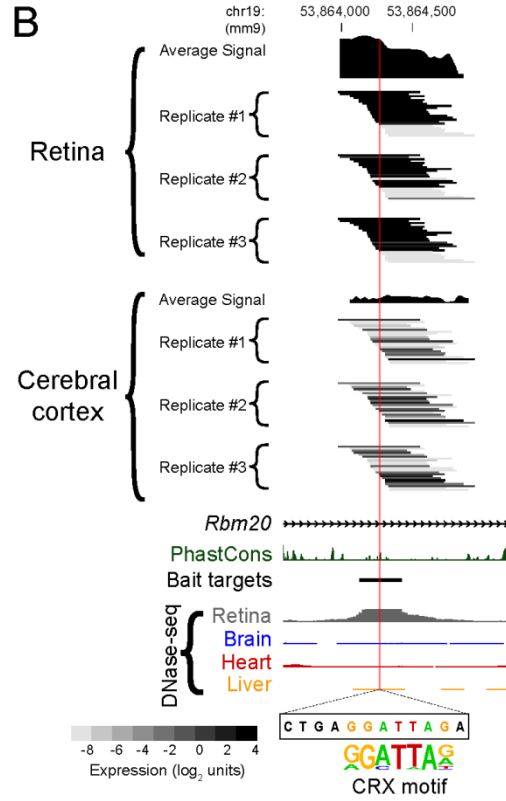
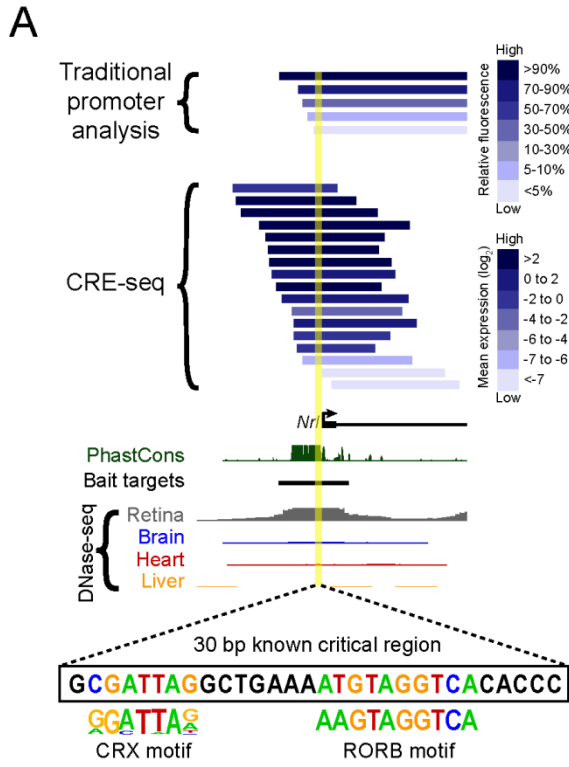


Figure 3.6. Truncation mutation analysis by CRE-seq. (A) Example of a truncation mutation analysis at the *Nrl* promoter via a traditional one-at-a-time reporter assay (Montana et al. 2011b) vs. capture-and-clone CRE-seq. For the traditional reporter constructs, the 3' end extends beyond the window depicted in the figure. For the CRE-seq data, only barcoded constructs in the same orientation as the *Nrl* promoter are shown. The yellow highlighted region corresponds to a known critical region with CRX and RORB motifs (Andre et al. 1998; Montana et al. 2011b). The minus strand of DNA is displayed. In (A) and (B), the CRX motif (from HOMER (Heinz et al. 2010)) is based on CRX ChIP-seq data (Corbo et al. 2010). The reverse orientation of the CRX motif is displayed. Additional examples of CRE-seq truncation mutation analysis: (B) Retinal DHS with retina-specific expression. The critical region identified by CRE-seq (pink) contains a putative CRX motif. (C) Two adjacent brain DHSs in the same intron of *Bsn* exhibit low (DHS #1, green) vs. high (DHS #2, pink) activity in the cortex. (D) Truncation mutation analysis of a brain DHS. A gradual decrease in activity was observed within the ~150 bp critical region (pink), corresponding to a phylogenetically conserved peak. Within this critical region, a smaller region (vertical blue stripe) was identified that contained an E-box consensus motif ('CANNTG') and a motif for a bZIP protein, based on AP-1 ChIP-seq data (Heinz et al. 2010). All browser images are from the UCSC Genome Browser (mm9) (Karolchik et al. 2014). DNase-seq data are from Mouse ENCODE (Yue et al. 2014). PhastCons depict 30-way vertebrate phylogenetic conservation (Siepel et al. 2005). The heat map scale shown in (B) is the same as that used in (C) and (D).

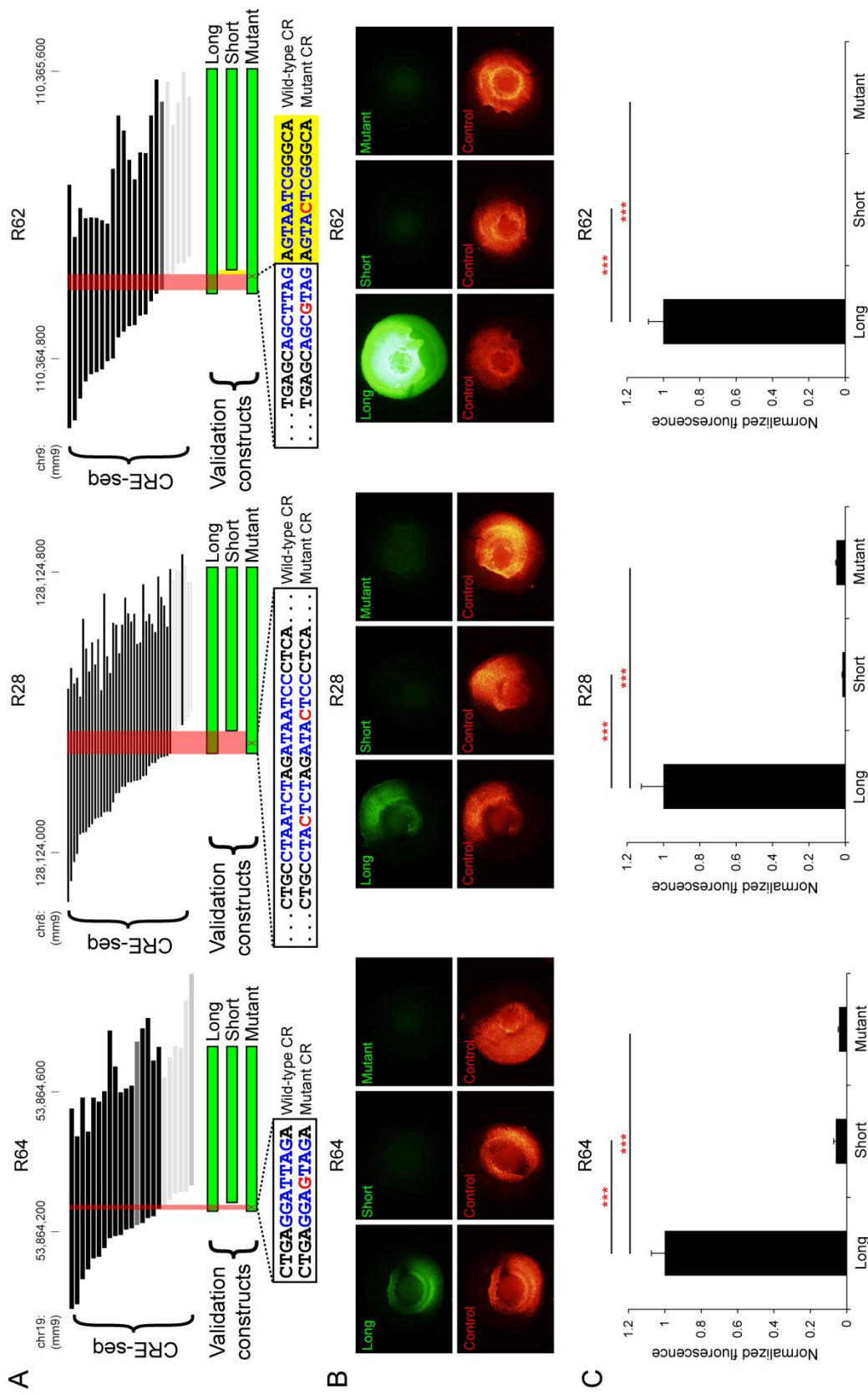


Figure 3.7. Validation of individual loci by fluorescence reporter assays. (A) Critical regions (pink areas) identified by CRE-seq truncation mutation analysis at three retinal DHSs (R64, R28, and R62) were validated by testing of individual constructs with fluorescence reporter assays. Depicted CRE-seq data are based on expression scores averaged across retinal replicates. Note that R64 is the same locus as in Figure 6B. For each locus, a ‘long’ construct containing the critical region (CR), a ‘short’ construct without the critical region, and a ‘mutant’ construct with point mutations (red font) in predicted CRX sites (blue font) were synthesized. Sequences are shown for the plus strand of DNA in all cases. For R62, one CRX site fell within the critical region, and a second CRX site was immediately adjacent (yellow area). Individual test constructs were directionally cloned upstream of the minimal promoter-GFP cassette in a non-AAV vector. The test constructs were coelectroporated into explant retinas with *Rho*-CBR3-DsRed (Corbo et al. 2010) as a loading control. (B) Representative whole mount images of electroporated retinas are shown (exposure times are the same for all images). (C) Quantification of the GFP levels normalized to DsRed levels. Error bar represents SEM (n = 10-12 retinas per test construct). ***P-value < 10⁻⁶ (two-tailed student's t test).

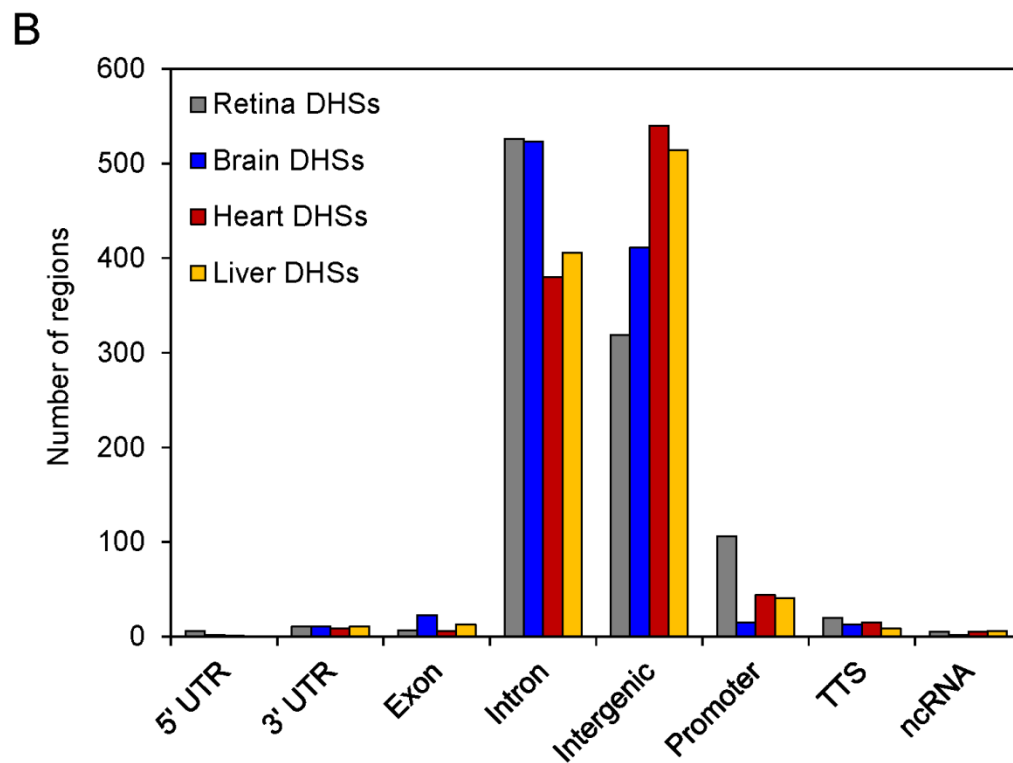
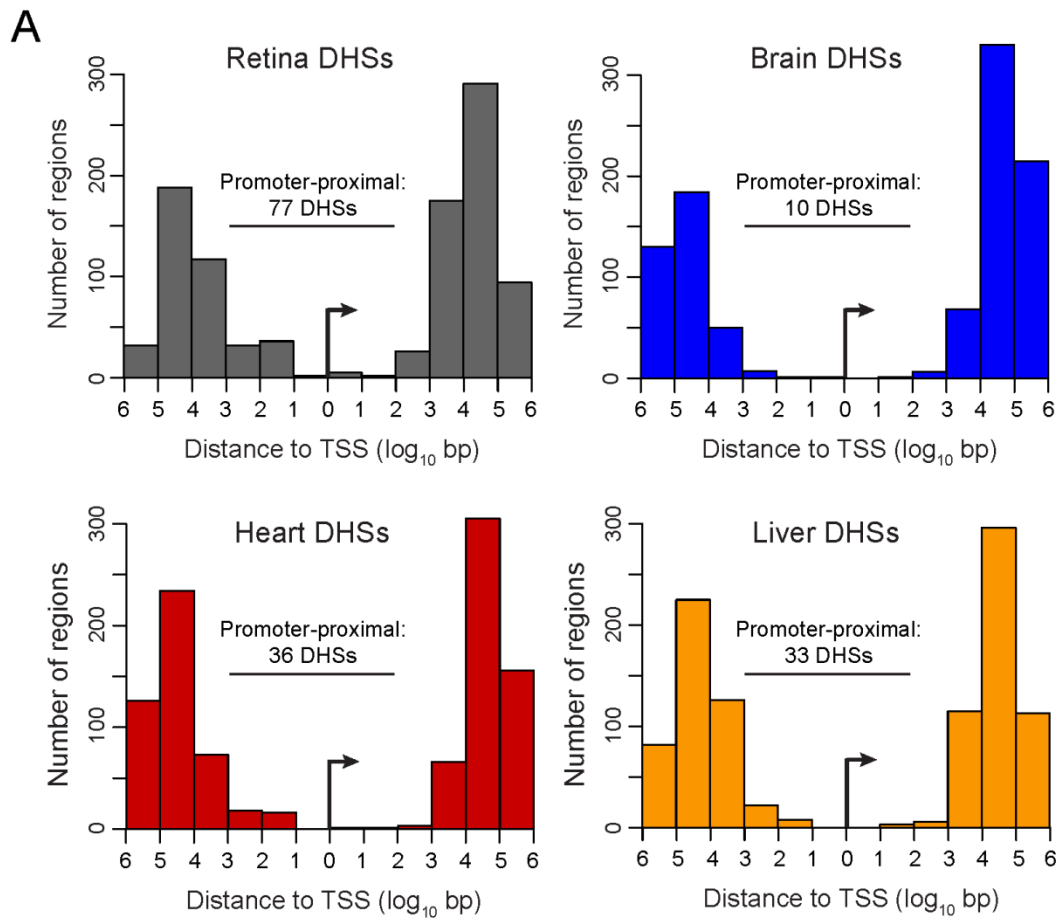


Figure 3.S1. Distribution of 4,000 target DHS regions. (A) Histogram showing the locations of the target regions (up- or downstream) relative to the nearest transcriptional start site (TSS, indicated by arrow) based on GREAT analysis (McLean et al. 2010). The number of ‘promoter-proximal’ DHSs for each group is shown, as defined by DHSs that fell within -1 kb to +100 bp relative to the nearest TSS. (B) Histogram showing the basic annotations for the target regions, based on HOMER (Heinz et al. 2010). Abbreviations: UTR, untranslated region; TTS, transcription termination site; ncRNA, non-coding RNA.

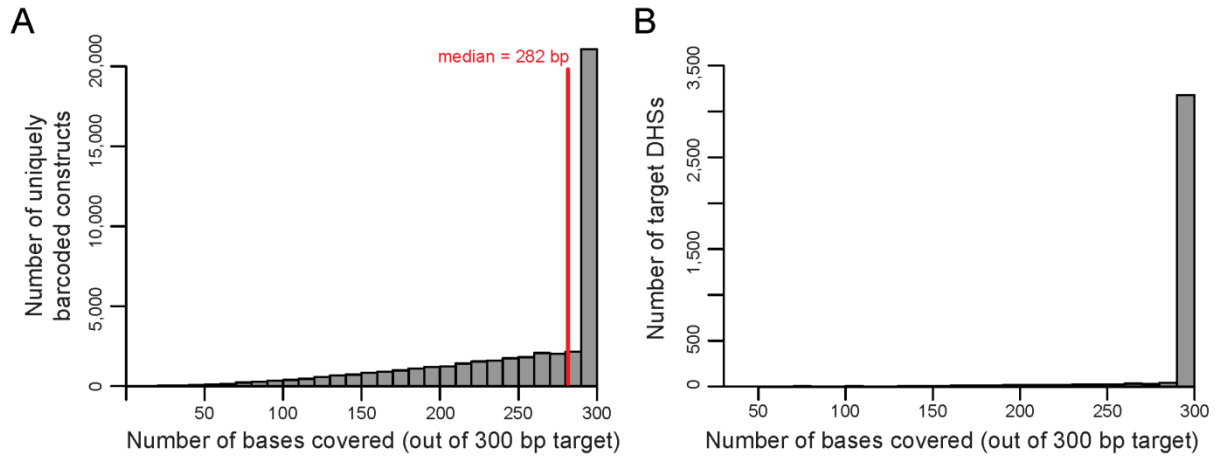


Figure 3.S2. Distribution of overlap of captured fragments with target DHS regions. Each target DHS was 300 bp. (A) Histogram showing the distribution of the overlap between targets and captured fragments for all 45,670 uniquely barcoded constructs. The median number of bases of overlap was 282 bp. (B) Histogram showing the distribution of the overall overlap between all 3,483 represented target regions and the captured fragments, based on the union of the captured fragments. Fragments collectively tiled at least 200 bp out of the 300 bp target for 3,402/3,483 (98%) target regions, and the entire 300 bp target for 3,146/3,483 (90%) target regions.

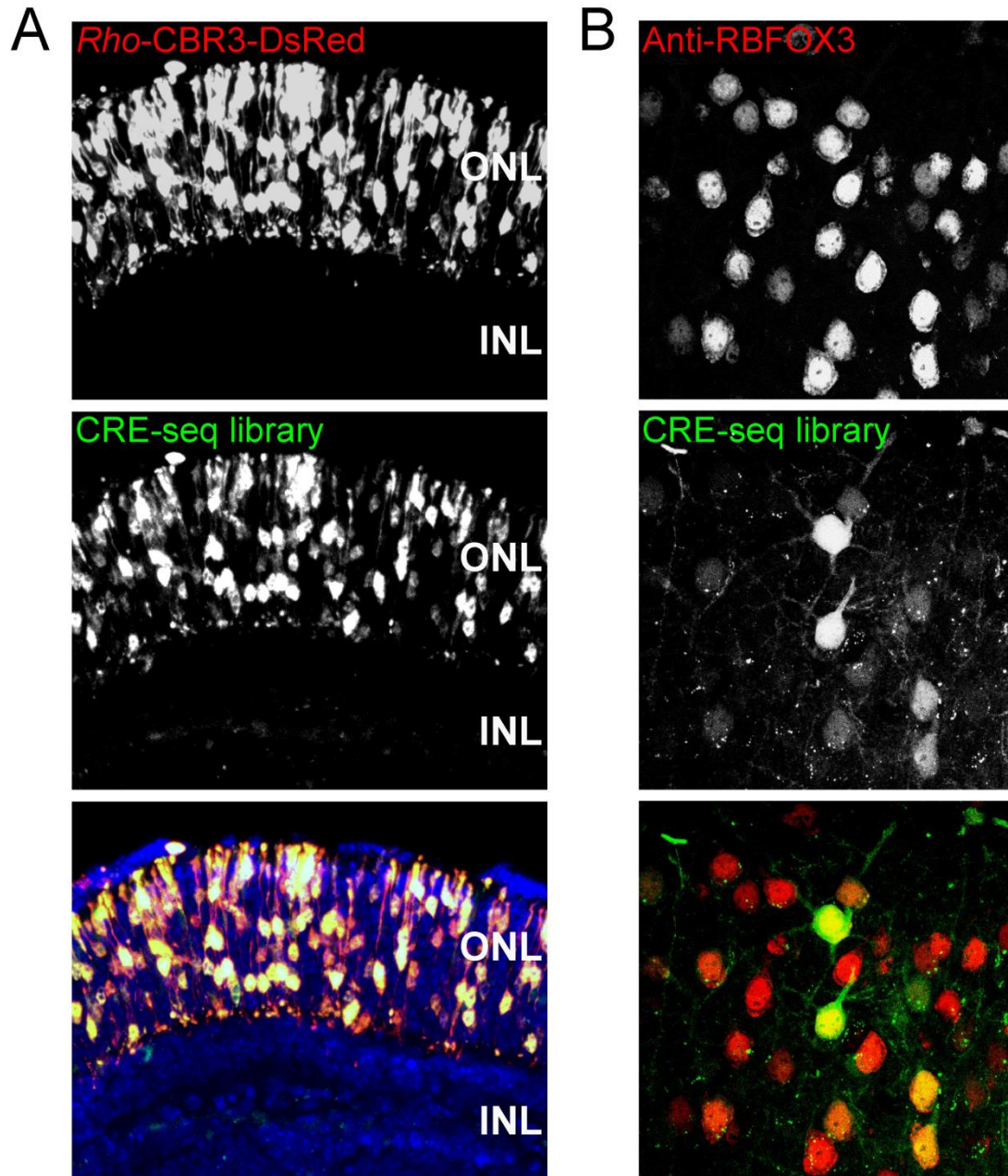


Figure 3.S3. Co-expression of the library and cellular markers. (A) Same retina as in Figure 3C, but a wider field and additional channels are shown. The library contains a GFP reporter. *Rho-CBR3-DsRed* is a rod-specific reporter (Corbo et al. 2010) that was coelectroporated with the library. Colocalization of DsRed and GFP indicates expression of the library in rods. Blue channel in merged image is DAPI, a nuclear counterstain. (B) Antibody staining of the neuronal marker RBFOX3 (also known as NeuN) (red channel) (Mullen et al. 1992) in a region of cerebral cortex that has been infected with the AAV-packaged library. Colocalization of RBFOX3 and GFP indicates expression of the library in neurons.

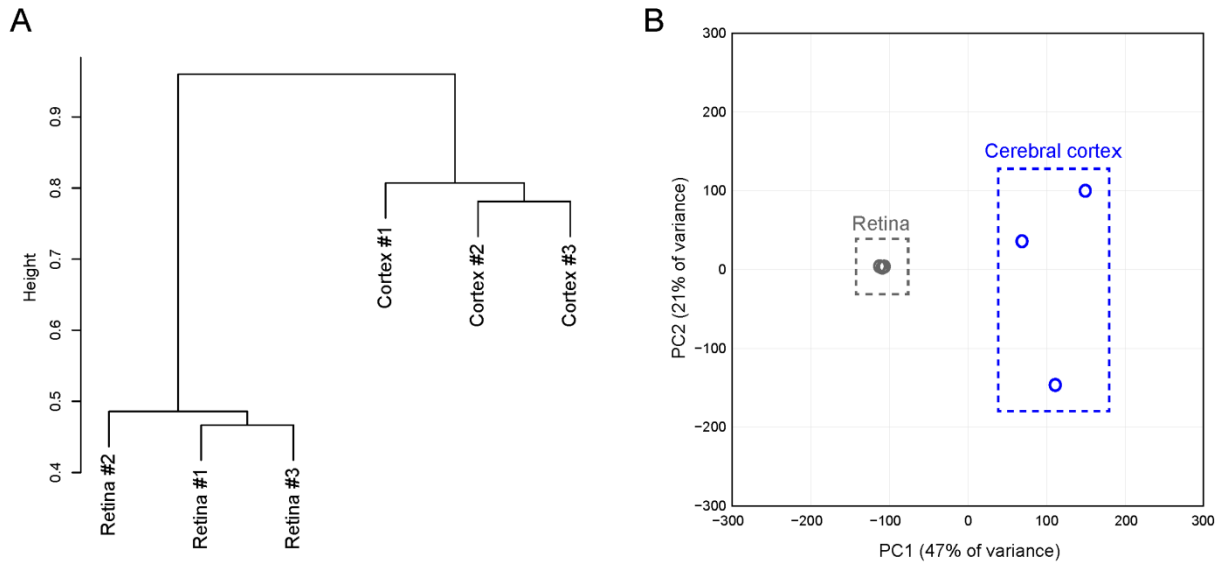


Figure 3.S4. Comparison of biological replicates. (A) Dendrogram showing distance between retinal and cerebral cortex biological replicates. (B) Principal component analysis (PCA) plot showing that PC1, which separates retina vs. cerebral cortex, accounts for the largest fraction of the variance.

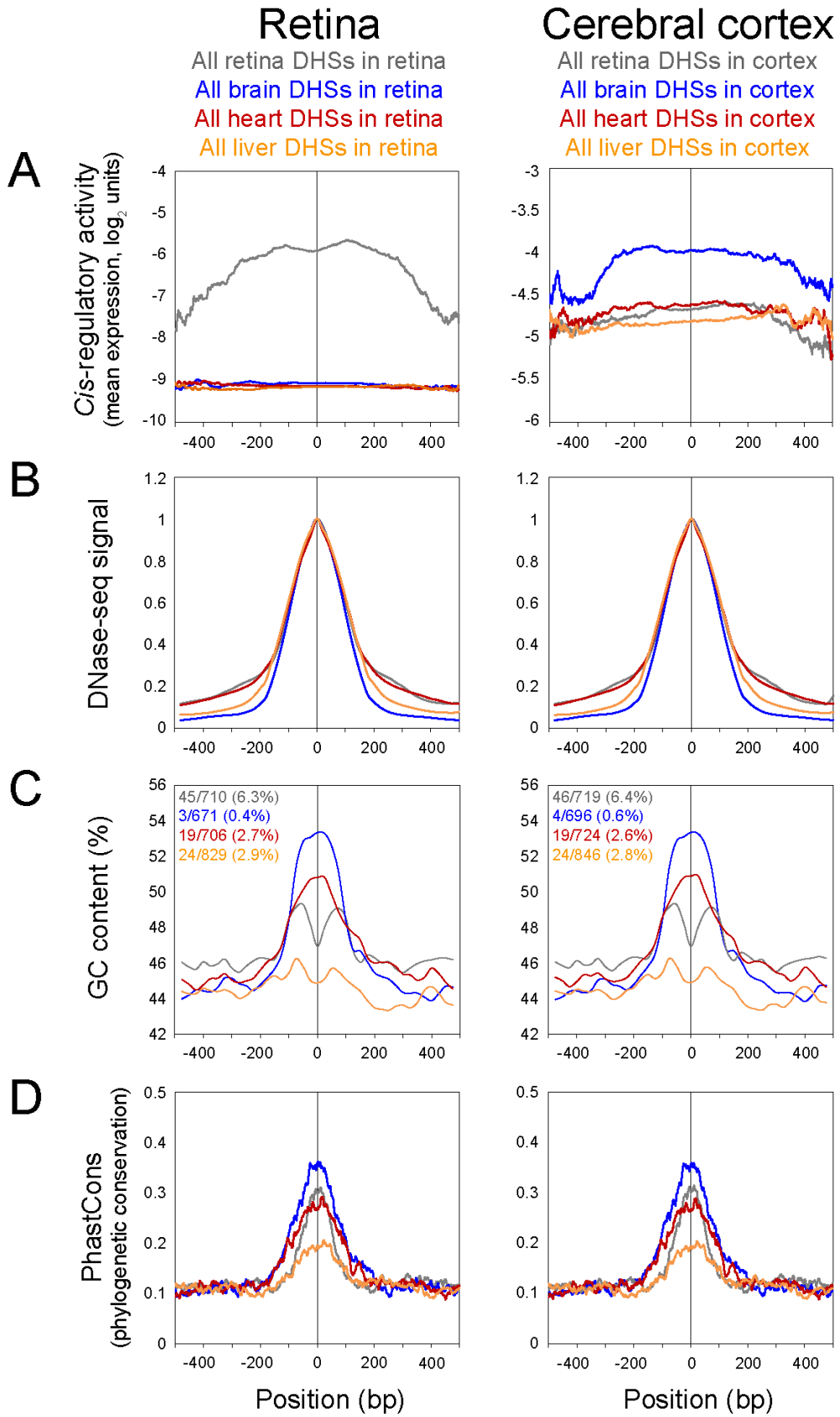


Figure 3.S5. CRE activity, DNase-seq signal, GC content, and phylogenetic conservation of assayed DHSs in a 1 kb centered window. Retina, brain, heart, and liver DHSs were assayed in the retina (left) and cerebral cortex (right). Each panel shows a 1 kb centered window. Only DHSs with at least 2 barcodes were included in this analysis, i.e., in the retina, 710 retinal DHSs, 671 brain DHSs, 706 heart DHSs, and 829 liver DHSs, and in the cerebral cortex, 719 retinal DHSs, 696 brain DHSs, 724 heart DHSs, and 846 liver DHSs. (A) *Cis*-regulatory activity, as measured by mean expression in log₂ units. For each assayed DHS, at each base position across the 1 kb window, the expression values of the individual barcoded constructs whose CREs overlapped the position were averaged across biological replicates. (B) DNase-seq score, normalized to the peak height. (C) GC content, calculated in 50 bp windows, sliding 25 bp at a time. The fractions denote the proportion of DHSs that were promoter-proximal (i.e., located within -1 kb to +100 bp relative to the nearest TSS) based on GREAT annotations (McLean et al. 2010). (D) Phylogenetic conservation as measured by 30-way vertebrate PhastCons (Siepel et al. 2005).

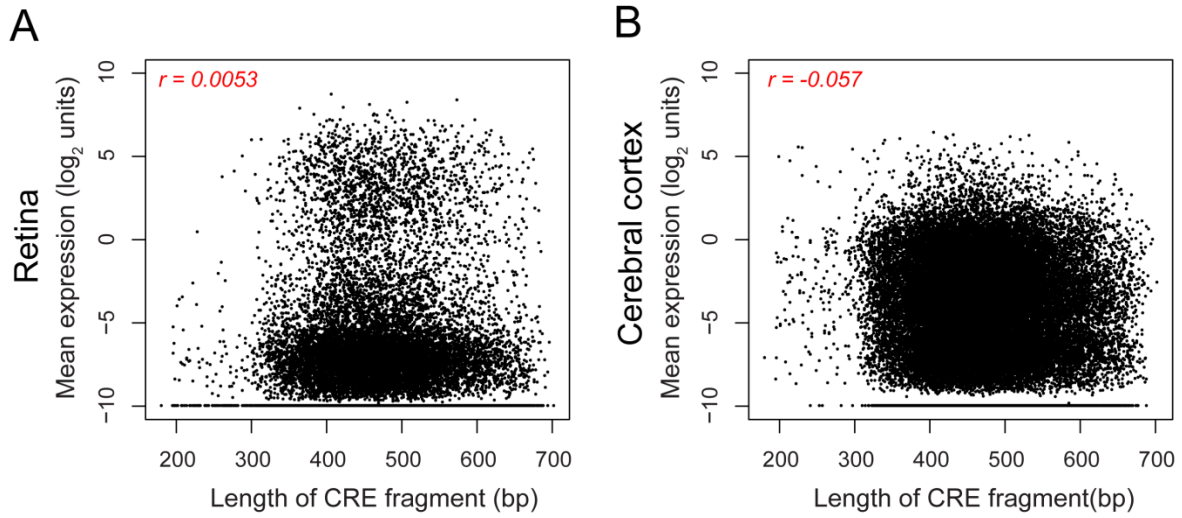


Figure 3.S6. Length of CRE fragments vs. expression. Each dot in the scatterplot represents an individual barcoded construct whose activity was assayed in (A) retina (~36,000 constructs) or (B) cerebral cortex (~39,000 constructs). Expression values were averaged across biological replicates. Pearson correlation values are shown.

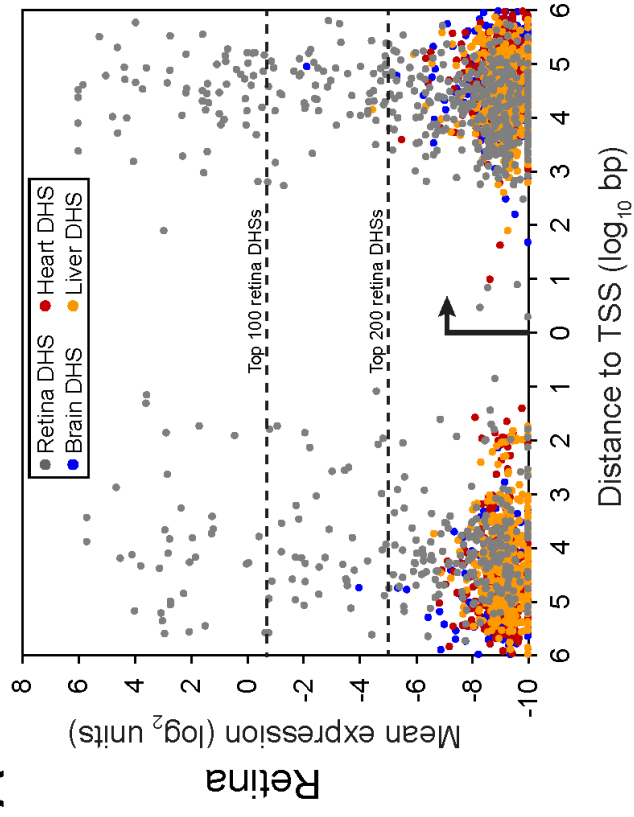
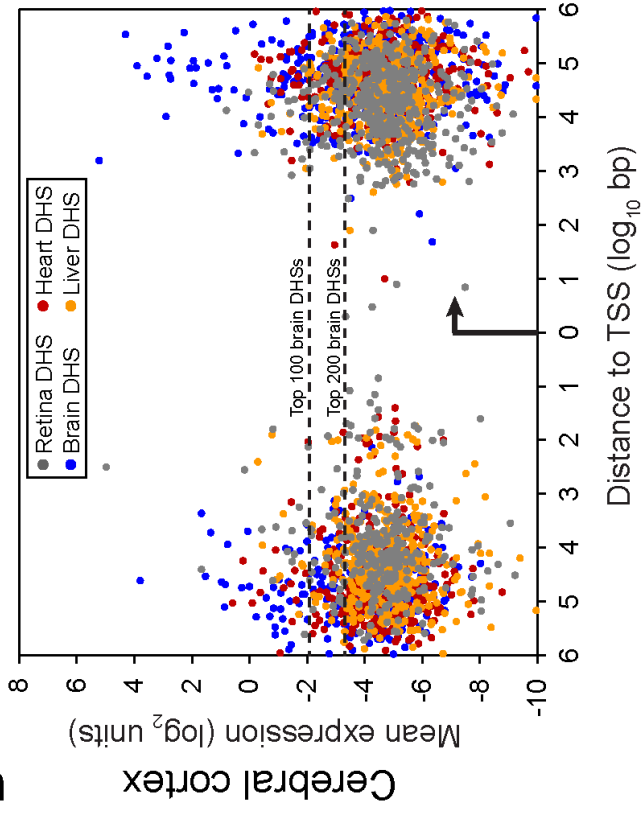
A**B**

Figure 3.S7. Distance to nearest TSS vs. expression. Each dot in the scatterplot represents a DHS whose activity was assayed in (A) retina (~3,000 DHSs) or (B) cerebral cortex (~3,000 DHSs). Expression values were averaged across barcodes and biological replicates, and only DHSs with at least 2 well-represented barcoded constructs were included. Locations of target regions (up- or downstream) relative to the nearest TSS (indicated by arrow) are based on GREAT analysis (McLean et al. 2010). Gray, blue, red, and orange dots denote retina, brain, heart, and liver DHSs, respectively. Dotted lines denote the thresholds for the top 100 and top 200 most active retinal DHSs assayed in the retina, and the top 100 and top 200 most active brain DHSs assayed in the cerebral cortex.

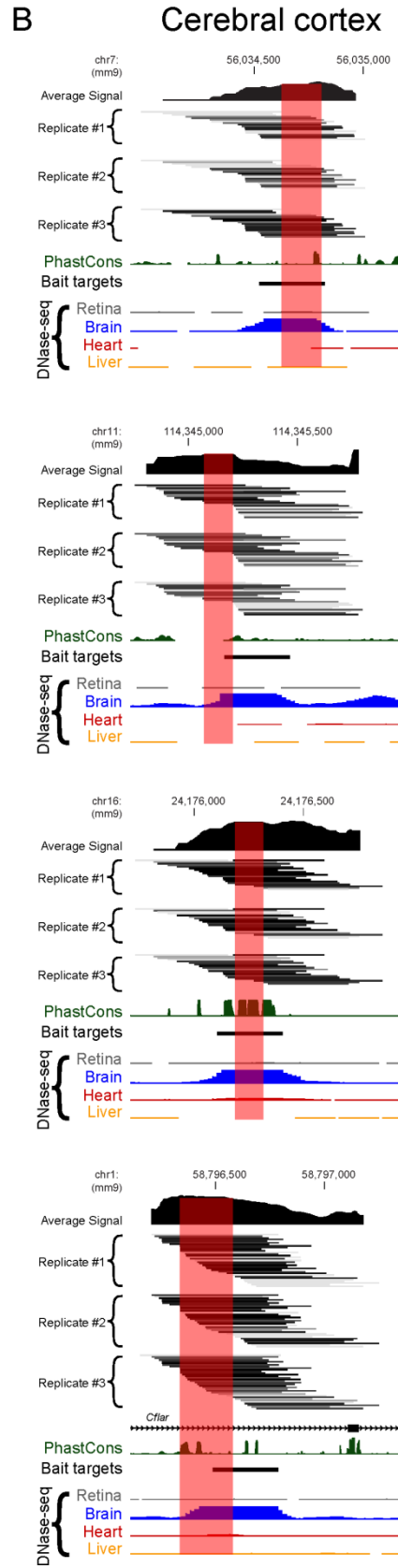
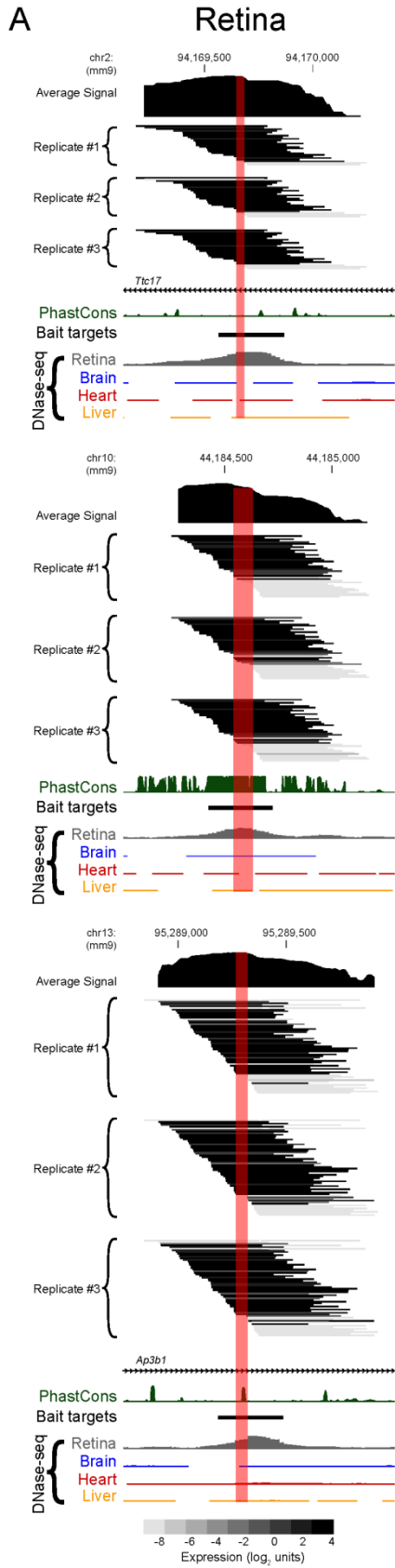


Figure 3.S8. Additional examples of truncation mutation analysis by CRE-seq. Additional examples of CRE-seq truncation mutation analysis for (A) retinal DHSs, based on retinal CRE-seq data, and (B) brain DHSs, based on cerebral cortex CRE-seq data. Individual barcoded constructs are colored by intensity (darker indicates higher expression; the heat map shown at bottom of panel A was used throughout). Critical regions are highlighted in pink. All browser images are from UCSC Genome Browser (mm9) (Karolchik et al. 2014). DNase-seq data are from Mouse ENCODE (Yue et al. 2014). PhastCons depict 30-way vertebrate phylogenetic conservation (Siepel et al. 2005).

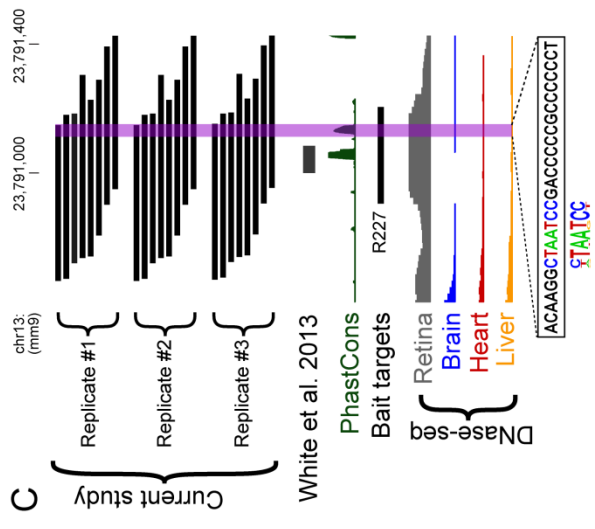
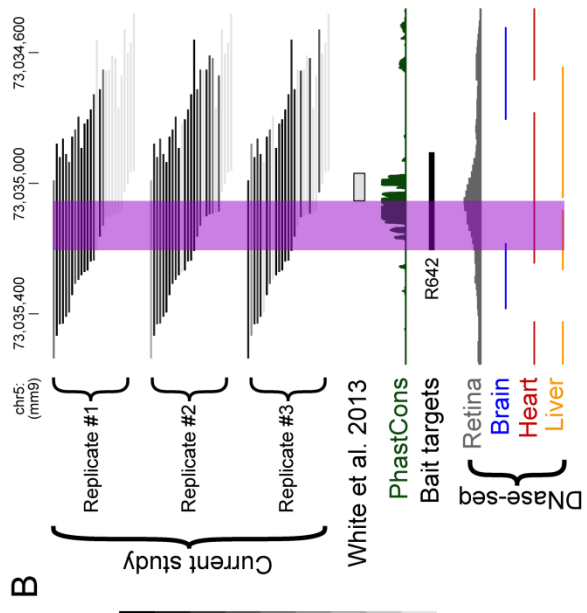
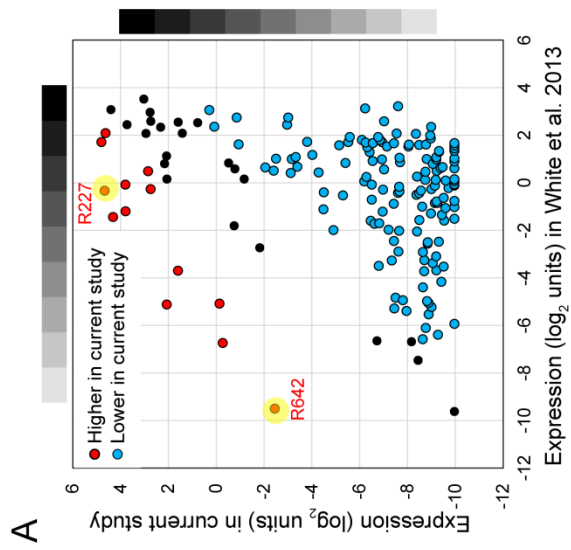


Figure 3.S9. Comparison between enhancer activity of short synthesized CREs and autonomous activity of corresponding captured CRE fragments in the retina. The enhancer activity of short CREs (84 bp in length, synthesized on oligonucleotide arrays), representing the middle of CRX ChIP-seq peaks, was previously assayed in electroporated retinas by CRE-seq using a tissue-specific proximal promoter (White et al. 2013). The current study measured the autonomous activity of captured fragments using a minimal promoter. There were 176 regions (all retinal DHSs) assayed in both studies. (A) Scatterplot comparing the enhancer activity of short CREs (x-axis) with the autonomous activity of corresponding long CREs (y-axis). Each dot represents a DHS region (expression values were averaged across barcoded constructs and retinal replicates). Dots are color-coded based on whether expression was higher by four-fold or more in the current study (red dots) or lower by four-fold or more in the current study (blue dots). Note that R642 and R227 (yellow circles) are examples of constructs with higher activity in the current study. (B) R642 contains a phylogenetically conserved peak that contains a critical region, as identified by truncation mutation analysis in the current study. The short CRE that was tested in the enhancer assay excludes a portion of the phylogenetically conserved peak (purple) (White et al. 2013). The minus strand of DNA is shown. (C) R227 contains two phylogenetically conserved peaks, one of which is encompassed by the short CRE tested in the enhancer assay (White et al. 2013). The other peak (purple) contains a predicted CRX site. The CRX motif (from HOMER (Heinz et al. 2010)) is based on CRX ChIP-seq data (Corbo et al. 2010). Phylogenetic conservation is depicted by 30-way vertebrate PhastCons (Siepel et al. 2005). The heat map scales shown in (A) were consistent between the two studies and also used for (B) and (C).

CHAPTER 4:

A Candidate Causal Variant Underlying Both Higher Cognitive Performance and Increased Risk for Bipolar Disorder

4.1 AUTHOR CONTRIBUTIONS

This project was initially conceived by Joseph Corbo, shortly after the first educational attainment GWAS identifying 6q16.1 was published (Rietveld et al. 2013), as a proof-of-concept for demonstrating the power of CRE-seq for *cis*-regulatory analysis in the brain (Appendix 3). As evidence accumulated for involvement of this locus in human cognition and bipolar disorder (Muhleisen et al. 2014; Davies et al. 2015; Trampush et al. 2015; Hou et al. 2016), Joe and I became more focused on understanding the underlying biological mechanism.

This work was conducted in collaboration with Jeongsook Kim-Han (cortical electroporations), Cheng Lin (EMSA), Omer Gokcumen (phylogenetic analyses), Andrew Hughes (motif analyses), and Connie Myers (cerebral organoid culture). I designed experiments and conducted bioinformatic analyses, EMSAs, CRE-seq, allele-specific experiments, and experiments involving the transgenic, knockout, and knock-in mice. This project is a work in progress, and the contents of this chapter have not yet been published.

4.2 ABSTRACT

Genome-wide association studies (GWAS's) have identified thousands of non-coding regions associated with complex diseases, but few underlying causal variants are known. Multiple GWAS's have identified an intergenic region associated with both cognition and risk for bipolar disorder. This region contains dozens of fetal brain-specific open chromatin peaks and is located ~1 Mb upstream of the neuronal transcription factor *POU3F2*. Using computational approaches, we identified a candidate causal variant that falls within a highly conserved putative enhancer, LC1. This variant, rs77910749, is a single-base deletion that is predicted to be highly deleterious. We hypothesized that rs77910749 alters the enhancer activity of LC1 and thereby alters *POU3F2* expression. First, we created transgenic reporter mice and found evidence of LC1 activity in the developing cerebral cortex and amygdala. To test whether rs77910749 alters LC1 enhancer activity, we implemented CRE-seq in embryonic mouse brain and human iPSC-derived cerebral organoids for the first time, which revealed subtle gain-of-function in enhancer activity. To probe the *in vivo* function of LC1, we deleted the orthologous mouse region and examined resulting allele-specific *Pou3f2* expression, which showed region-specific effects. Lastly, to study the effects of rs77910749 *in vivo*, we knocked the variant into the mouse genome. Overall, modest but significant changes were observed, suggesting that rs77910749 is a variant of small effect and/or exerts a large effect in a small population of cells. Our study provides a framework for establishing the causality of non-coding variants, with particular relevance to neuropsychiatric diseases.

4.3 INTRODUCTION

Genome-wide association studies (GWAS's) have identified thousands of non-coding regions associated with complex diseases, but pinpointing the underlying 'causal variant' contributing to disease pathogenesis is a challenge (Zhang and Lupski 2015). Identifying causal variants and dissecting their molecular effects would not only provide insight into disease pathways but also facilitate the clinical interpretation of non-coding variants. For neuropsychiatric diseases, the functional study of disease-associated variants is particularly challenging, as the etiologically relevant cell type and appropriate experimental model system in which to assay the effects of candidate variants are often unclear.

Bipolar disorder (BPD) is a neuropsychiatric disease characterized by alterations in mood, classically with episodes of both mania and depression (Craddock and Sklar 2013). It affects ~1% of the world population and is associated with high morbidity and mortality (Merikangas et al. 2011; Whiteford et al. 2013). While the disease is highly heritable (~80% heritability), the underlying genes are largely unknown (Craddock and Sklar 2013; Harrison 2016). Furthermore, the etiology of the disease is poorly understood at the level of molecular pathways, neuroanatomy, and neural circuitry, although the amygdala and prefrontal cortex have been strongly implicated (Maletic and Raison 2014). In addition to altered mood, BPD is strongly associated with heightened creativity, substantiating the link between 'madness' and 'genius' that has been speculated for centuries (Srivastava and Ketter 2010).

Recently, several large GWAS's of educational attainment and cognitive performance have reproducibly implicated an intergenic region located at the *MIR2113/POU3F2* locus in chromosome region 6q16.1 (Rietveld et al. 2013; Davies et al. 2015; Trampush et al. 2015). At the same time, two large GWAS's of BPD identified this same region (Muhleisen et al. 2014;

Hou et al. 2016). The lead SNPs in all of these studies are in strong linkage disequilibrium (LD) with each other, suggesting a common underlying causal variant(s). Given that the nearest protein-coding gene, *POU3F2*, is located ~0.7 Mb away, we hypothesized that the underlying ‘causal variant’ affects the activity of a non-coding *cis*-regulatory element (CRE, e.g., enhancer/silencer). Since CREs can act at long distances, *cis*-regulatory variants have the potential to disrupt the expression of distal genes (Kleinjan and van Heyningen 2005).

POU3F2 (also called *BRN-2*) is a transcription factor (TF) known to be important for the development of the hypothalamus and the cerebral cortex. In the cerebral cortex, *POU3F2* acts with *POU3F3* (also called *BRN-1*) to regulate the neurogenesis, maturation, and migration of upper-layer neurons (Nakai et al. 1995; Schonemann et al. 1995; McEvelly et al. 2002; Sugitani et al. 2002; Dominguez et al. 2013). Furthermore, overexpression of *POU3F2* facilitates the direct reprogramming of fibroblasts into neurons (Vierbuchen et al. 2010; Wapinski et al. 2013). In mice, both increased and decreased levels of *Pou3f2* are associated with alterations in neuronal fate (Dominguez et al. 2013; Belinson et al. 2016). In humans, deletions encompassing *POU3F2* have been associated with intellectual disability (Kasher et al. 2016). Thus, *cis*-regulatory changes that alter *POU3F2* dosage levels may perturb brain development.

Here, we used computational and experimental approaches to identify a candidate causal variant, rs77910749, which falls within a putative brain enhancer. To assay for enhancer activity, we generated transgenic reporter mice. We also implemented a massively parallel reporter assay (MPRA), CRE-seq, in the developing mouse brain and human iPSC-derived cerebral organoids for the first time. Finally, to characterize the role of LC1 and the effects of rs77910749 *in vivo*, we used CRISPR-Cas to generate an allelic series of LC1 mutants. We found that rs77910749 had modest but significant effects on transcription factor binding, enhancer activity, and sensory

gating-related behavior. Additionally, we observed region-specific effects, suggesting that rs77910749 may exert a larger effect in a small population of disease-relevant cells.

4.4 RESULTS

4.4.1 The *MIR2113/POU3F2* locus harbors non-coding variants associated with both increased cognitive performance and increased risk for BPD

A GWAS of educational attainment in ~100,000 Caucasian individuals identified a genome-wide significant signal on chromosome 6, in an intergenic region between *MIR2113* and *POU3F2* (Rietveld et al. 2013). This finding was replicated in a subsequent expansion study (Okbay et al. 2016). Further analysis revealed that in this study, educational attainment served as a proxy phenotype for cognitive performance (Rietveld et al. 2014). The lead SNP, rs9320913, was associated with higher verbal and math standardized test scores in children in the ALSPAC study (Ward et al. 2014). Additionally, meta-analysis of GWAS's in the COGENT consortium showed that, while rs9320913 was not directly genotyped, a proxy variant (rs1906252, $r^2 = 0.96$ with rs9320913) was significantly associated with increased general cognitive ability (Trampush et al. 2015). Meta-analysis of GWAS's in the CHARGE consortium also showed a positive association between a proxy variant (rs10457441, $r^2 = 0.91$ with rs9320913) and general cognitive ability (Davies et al. 2015). An earlier GWAS in healthy older adults found an association between rs1906252 and faster information processing as measured by a symbol search task ($P = 2.08 \times 10^{-5}$) (Luciano et al. 2011). Thus, multiple studies have demonstrated a reproducible association between variants at this locus and cognitive performance.

A recent GWAS of BPD in ~10,000 patients and ~14,000 controls identified a novel risk locus in the same region of chromosome 6 between *MIR2113* and *POU3F2* (Muhleisen et al. 2014). The lead SNP, rs12202969, was associated with ~10-20% increased risk for BPD (OR = ~1.1-1.2), which is a typical effect size for GWAS studies (Price et al. 2015). Another independent GWAS of BPD in ~10,000 patients and ~30,000 controls replicated the signal at the

chromosome 6 locus, identifying a genome-wide significant signal at the proxy variant rs1487441 ($r^2 = 0.98$ with rs12202969) with an OR of 1.12 (Hou et al. 2016). We observed that these two BPD GWAS lead SNPs (rs12202969 and rs1487441) are in extremely high LD with the lead SNPs in the GWAS's of educational attainment and cognitive performance (rs9320913, rs1906252, and rs104757441) (pairwise $r^2 = 0.92-0.99$), suggesting a shared genetic basis for cognitive ability and BPD (Supplemental Table 1). In particular, the variants associated with higher cognitive performance were also associated with increased BPD risk. In agreement with an earlier study (Koenen et al. 2009), children in the ALSPAC study with higher IQ scores were more likely to develop manic features of BPD later in life (Smith et al. 2015), further underscoring the potential link between cognition and BPD.

4.4.2 Identification of the candidate causal variant rs77910749, a human-specific non-coding variant that falls within a fetal brain-specific open chromatin region

Since the GWAS's for educational attainment, cognitive performance, and BPD appear to have a shared underlying signal at the *MIR2113/POU3F2* locus, we sought to find candidate causal variants. We first surveyed the epigenomic landscape of the ~0.5 Mb region (Chr6:98.3-98.8 Mb in hg19) identified by the GWAS's (Figure 4.1A, yellow box). This LD block contains dozens of human fetal brain-specific DNase-seq peaks, which are regions of open chromatin that demarcate putative CREs (Bernstein et al. 2010; Roadmap Epigenomics et al. 2015). The lead SNPs (rs9320913, rs1906252, rs10457441, rs12202969, and rs1487441) are located ~0.1 Mb away from *MIR2113* and ~0.7 Mb away from the nearest protein-coding gene, *POU3F2*. This suggested that the underlying causal variant exerts a *cis*-regulatory effect.

We then focused on the ~60 kb region of highest LD, which contains all five of the lead SNPs (Figure 4.1A, purple box). Within this region, we identified six fetal brain-specific DNase I hypersensitive sites (DHSs), termed LC0 and LC5 (henceforth referred to as the ‘local cluster’) (Figure 4.1B). While none of the lead SNPs fall within fetal brain DHSs, four variants in LD with rs9320913 ($r^2 > 0.2$) fell within fetal brain DHSs in the local cluster (Figure 4.1C top panel, blue font): rs77910749 in LC1, rs13208578 in LC2, rs12204181 in LC4, and rs17814604 in LC5.

We next examined these four variants more closely. Since phylogenetic conservation is often a marker of functionality, we hypothesized that the underlying causal variant would fall within a phylogenetically conserved region. LC4 exhibits low conservation, and hence we deemed rs12204181 a less likely candidate. LC2 is highly conserved, but rs13208578 is present in multiple vertebrate species, including primates, suggesting that it is well-tolerated (Figure 4.S1A). Furthermore, LC2 did not exhibit enhancer activity in a transgenic mouse assay (element hs1106 tested at E11.5 in the *pHsp68-LacZ* vector) (Visel et al. 2007). Thus, rs13208578 also appeared less likely to be the causal variant, leaving rs77910749 and rs17814604 as the top candidates. Analysis of variants using CADD, a machine learning-based tool that predicts pathogenicity based on phylogenetic conservation and epigenomic annotations (Kircher et al. 2014), corroborated this result (Figure 4.1C, bottom panel). The scaled CADD scores of rs77910749 and rs17814604 were 27.3 and 34, respectively, placing them in the top 0.2% and 0.04% of all variants (including coding variants) for predicted pathogenicity.

We noticed that the LD of rs17814604 ($r^2 = 0.43$ with rs9320913) was relatively low despite a high D' (0.99), and that rs17814604 was less common than rs9320913. This suggested that individuals with rs17814604 represented a subset of those with rs9320913. Indeed, upon construction of a phylogenetic tree (Figure 4.S2), it became apparent that a ‘derived haplotype’

emerged containing rs77910749, as well as rs13208578 in LC2, rs12204181 in LC4, and the lead SNPs rs10457441 (cognition), rs12202969 (BPD), and rs9320913 (education). This derived haplotype likely arose at least ~60,000-70,000 years ago, after the split of modern humans from Neanderthals and Denisovans. Subsequently, ~30,000 years ago, rs17814604 arose from the derived haplotype block, and thus it is very rare in certain populations, as seen in Figure 4.S3A. In particular, the allele frequency of rs17814604 in East Asians is 0.2% (1000 Genomes Phase 3 (Genomes Project et al. 2015)). A study of 342 Han Chinese individuals found a significant association between rs12202969 ($r^2 = 0.96$ with rs9320913 in Han Chinese) and math ability. Since rs17814604 is nearly absent among Chinese individuals, it is extremely unlikely that the signal at rs12202969 is due to rs17814604 (Zhu et al. 2015). Therefore, rs17814604 is unlikely to be the causal variant.

By contrast, rs77910749 is relatively common across the globe (Figure 4.S3B), with an allele frequency of 51% in Europeans (1000 Genomes Phase 3 (Genomes Project et al. 2015)). It is in strong LD with both rs9320913 ($r^2 = 0.97$) and rs12202969 ($r^2 = 0.98$). Inspection of rs77910749 revealed that it is a single base pair deletion of a ‘T’ in a stretch of ~100 bases that are nearly perfectly conserved among vertebrates down to coelacanth fish (Figure 4.S1B). Based on the phylogenetic conservation of the affected nucleotide and its location within a fetal brain DNase-seq peak, another group also suggested rs77910749 as a candidate causal variant (Trampush et al. 2015). Despite its high frequency among humans, we did not find evidence for a selective sweep, suggesting that this variant does not alter dramatically alter fitness. This is consistent with other studies showing that conserved non-coding regions have undergone relaxed selective constraint in humans, likely due to the small effective population size (Kryukov et al. 2005).

Interestingly, rs77910749 appears to be human-specific, as it is absent from other vertebrate genomes, including all 79 non-human primate individuals (representing five species) that were sequenced in the Great Ape Genome Project (Prado-Martinez et al. 2013) (Figure 4.S4). Therefore, rs77910749 is a common human-specific non-coding variant that is a good candidate causal variant for cognitive performance and BPD.

4.4.3 Mouse epigenomic data suggest that LC1 is an enhancer in the developing brain and reveal that rs77910749 falls within a binding site for Pax6

Since LC1 is highly conserved, we examined the orthologous region in the mouse genome. We observed that LC1 is located between *Mir2113* and *Pou3f2* (~0.1 Mb and 1 Mb away, respectively) in the mouse genome, as well as in other vertebrate genomes, suggesting that LC1 is part of genomic regulatory block (GRB) whose conserved synteny has functional importance (Kikuta et al. 2007). Within a topologically associating domain (TAD), there is a higher frequency of interactions (e.g., enhancer looping) between chromosomal regions. A survey of published Hi-C data (Dixon et al. 2012; Rao et al. 2014; Dixon et al. 2015; Leung et al. 2015) revealed that LC1 falls within a TAD in multiple mouse and human cell types, suggesting that this is an evolutionarily conserved and cell-type invariant TAD (Dixon et al. 2016) (Figure 4.S5). Notably, this TAD encompasses both *Mir2113* and *Pou3f2*.

Next, we examined the epigenomic landscape of LC1 in detail. A time course of DNase-seq across various mouse tissues (The ENCODE Project Consortium 2012) demonstrated that LC1 corresponds to a region of open chromatin specific to the developing mouse brain, with a strong signal at E14.5 and diminished signal by E18.5 (Figure 4.2). ChIP-seq signals in the developing mouse brain for two enhancer marks, the coactivator p300 (Visel et al. 2009; Wenger

et al. 2013) and histone mark H3K27ac (Nord et al. 2013), support the notion that LC1 is a developmentally active brain enhancer at E14.5. Moreover, DNase-seq of mouse retina showed that LC1 is open in early postnatal period but subsequently closes (Figure 4.2), suggesting that LC1 may have a role in neurogenesis in both retina and brain (Wilken 2015). Notably, LC1 does not show a DNase-seq or H3K27ac peak in the adult brain, indicating that LC1 is closed in the majority of cells in the adult brain.

Since *cis*-regulatory variants can alter enhancer activity via disruption of TF binding, we hypothesized that rs77910749 alters TF binding. When we searched for bioinformatically predicted TF motifs using FIMO, we found that rs77910749 falls within a predicted binding site for the paired domain (PD) of Pax6 (Grant et al. 2011) (Figure 4.3A). Pax6 is a TF with numerous critical roles in brain development (reviewed in (Manuel et al. 2015; Ypsilanti and Rubenstein 2016)). In addition, it is likely to be a direct transcriptional regulator of *Pou3f2* (Coutinho et al. 2011; Dominguez et al. 2013; Ninkovic et al. 2013). To determine whether Pax6 binds LC1, we examined published Pax6 ChIP-seq data from E12.5 wild-type mouse forebrain, which revealed that LC1 is strongly bound by Pax6 *in vivo* (80th ranked peak out of 3,536 peaks). Moreover, the predicted Pax6 motif falls in the middle of this peak, suggesting that it is recognized by Pax6 *in vivo* (Sun et al. 2015) (Figure 4.2B and Figure 4.3A). Notably, LC1 was the only prominent ChIP-seq peak in the region.

4.4.4 *In silico* and *in vitro* analysis demonstrate modest effects of rs77910749 on Pax6 binding

Based on *in vitro* binding preferences as determined by SELEX (Jolma et al. 2013), rs77910749 is predicted to cause only a slight (~3%) decrease in Pax6 binding affinity (Figure

4.3A). By comparison, based on *in vivo* Pax6 binding preferences as determined by ChIP-seq in the E12.5 mouse forebrain (Sun et al. 2015), rs77910749 is predicted to decrease binding affinity by ~50%. To directly test the effect of rs77910749 on the binding affinity of the site, we expressed the PD of Pax6 and conducted quantitative electrophoretic mobility shift assays (EMSAs) using fluorescently labeled DNA probes (Man and Stormo 2001) (Figure 4.3B).

We found that PD binds to both the wild-type sequence and the sequence with rs77910749. This binding was specific, as demonstrated by abrogation of the gel shift by cold competition with unlabeled probes. When we quantified the relative affinities of the ‘Ref’ and ‘Var’ probes, we found that rs77910749 confers ~30% decreased binding affinity (Figure 4.3B).

We also examined the binding of PD5a, a splice isoform of Pax6 that is expressed in the brain and contains a 14 amino acid insertion in the PAI domain of PD. The PD5a isoform has a very different DNA binding preference than the canonical PD isoform (Epstein et al. 1994; Kozmik et al. 1997). Neither PD5a nor PD5a-HD bound to either the reference or variant sequence. Together, these results indicate that the canonical but not 5a isoform of Pax6 PD binds to the Pax6 site, and rs77910749 causes a modest decrease in the affinity of Pax6 binding.

4.4.5 Transgenic reporter mice show evidence of LC1 enhancer activity in the developing central nervous system (CNS)

To test whether LC1 is a *bona fide* enhancer and to investigate its spatiotemporal activity pattern, we created transgenic reporter mice, in which human LC1 (~1 kb fragment) was cloned upstream of the minimal *Hsp68* promoter and LacZ (Pennacchio et al. 2006) (Figure 4.4A). Since the DNase-seq signal for the orthologous mouse LC1 appeared strongest at E14.5, we screened ‘transient’ transgenic embryos at age E14.5 (i.e., embryos were F0’s and represented

independent transgenesis events). Among the seven embryos that were genotypically positive for LacZ, five showed LacZ expression (Figure 4.4B). The observed patterns of LacZ staining were consistent with expression in the cerebral cortex (lines #1, 4, 5), amygdala (lines #1, 2, and 3), as well as the skin (line #5).

We also created three independent stable lines (i.e., allowing F0's to produce F1 progeny). Two stable lines showed essentially no enhancer activity in multiple E14.5 embryos that were genotypically positive. The third stable line showed LacZ expression in the developing amygdala (Figure 4.4C). Thus, overall, 6/10 transgenic lines showed LacZ expression in the developing CNS. Among these, 4/6 showed expression in the developing amygdala and 3/6 showed expression in the developing cortex. Additionally, in accordance with DNase-seq data suggesting that LC1 is active in the developing mouse retina (Figure 4.2), 5/6 (all but transient transgenic embryo #5) also expressed LacZ in the retina (Figure 4.4).

Together, our data suggest that LC1 is transcriptionally active in the developing CNS. There was a high degree of variability from line to line, likely reflecting that LC1 is a relatively weak enhancer prone to insertion site effects (Wilson et al. 1990). Nonetheless, in the brain, LC1 is most likely active in the developing amygdala and/or cerebral cortex.

4.4.6 CRE-seq 'Nano' measures subtle gain-of-function enhancer activity of rs77910749

To quantitatively assess whether rs77910749 alters the enhancer activity of LC1, we utilized a multiplexed plasmid reporter assay, CRE-seq (Kwasnieski et al. 2012). In CRE-seq, a library of uniquely barcoded reporter constructs is introduced into cells, and the resulting barcoded transcripts are quantified by RNA-seq. We previously used CRE-seq to measure the activity of thousands of CREs in the early postnatal mouse retina and in the adult cerebral cortex

(Kwasnieski et al. 2012; White et al. 2013; Shen et al. 2016). Here, we adapted CRE-seq to assay a small pool of constructs with high coverage and depth, i.e., ‘CRE-seq Nano.’ We created three types of constructs: wild-type LC1 (‘Ref’), LC1 with rs77910749 (‘Var’), and a promoter-only (no enhancer) control. To increase the sensitivity of our assay, the enhancers were synthesized as multimers (Figure 4.5A). For each of these three construct types, twenty barcoded members were created, for a total of sixty barcoded constructs in the library.

We introduced this library into developing mouse cerebral cortex by *ex vivo* electroporation at E12.5, followed by two days of explant culture (Nichols et al. 2013). Histological sectioning revealed reporter GFP expression in the deeper layers of the cortex (Figure 4.5B). By contrast, p*Dcx*-DsRed (a co-electroporated control construct) expressed in the upper layers of the cerebral cortex as expected (Wang et al. 2007). *Dcx* encodes doublecortin, a microtubule-binding protein that is expressed in the developing cerebral cortex, specifically in post-mitotic neurons undergoing migration (Gleeson et al. 1999). Notably, there was little colocalization of DsRed and GFP, suggesting that in the reporter constructs were active either in progenitors and/or a subset of developing neurons in the cerebral cortex.

As an orthogonal assay system, we also introduced the library into human induced pluripotent stem cell (iPSC)-derived cerebral organoids (Lancaster et al. 2013; Pasca et al. 2015). These organoids expressed Pax6 and Pou3f2, as detected by antibody staining (Figure 4.5C). Seven days after electroporation, live imaging showed electroporated cells expressing a ubiquitous loading control, pCAG-DsRed (Figure 4.5B). A subset of DsRed-expressing cells also expressed GFP, indicating activity of the reporter constructs.

We then quantified the *cis*-regulatory activity of the constructs by sequencing (Figure 4.5C). For both mouse cerebral cortex and human cerebral organoids, we observed enhancer

activity of LC1 multimers (both ‘Ref’ and ‘Var’) compared to the promoter-only control, although the promoter had relatively stronger activity in the human cerebral organoids than in mouse brains. In the mouse cerebral cortex, the LC1 ‘Var’ multimer had ~11% higher activity than ‘Ref’, while in the human cerebral organoids, the LC1 ‘Var’ multimer had ~32% higher activity than ‘Ref’. Thus, rs77910749 confers higher LC1 enhancer activity in these assay systems. Interestingly, this effect is greater in human cerebral organoids than in the mouse cerebral cortex.

4.4.7 *In vivo* deletion of LC1 confers region-specific changes in *Pou3f2* expression

To directly address whether LC1 regulates *Pou3f2* expression and whether rs77910749 affects *Pou3f2* expression, we used CRISPR-Cas to knock out the orthologous mouse LC1 region (~1 kb) (‘LC1 KO’ mice), as well as to knock in the human-specific variant rs77910749 (‘KI’ mice) (Figure 4.6A). We also generated mice with a small deletion (4 bp) in the 3’ UTR of *Pou3f2*, which serves as a molecular barcode for allele-specific expression (ASE) analysis.

To examine the effect of LC1 deletion on *Pou3f2* expression, mice heterozygous for the LC1 deletion (‘LC1 het’) were mated to mice with the 3’ UTR variant (Figure 4.6B). We analyzed the brains of E14.5 mouse embryos that were ‘trans-het’, that is, heterozygous for both the LC1 deletion and the 3’ UTR variant. Importantly, the haplotype phase of trans-het animals is known (i.e., the LC1 KO allele is in *cis* with the wild-type *Pou3f2* 3’ UTR). To control for any effects of the 3’ UTR variant itself, control animals wild-type for LC1 and heterozygous for the 3’ UTR variant were also analyzed. By measuring *Pou3f2* RNA transcripts with or without the 3’ UTR variant, we quantified changes in expression due to the LC1 KO allele relative to the wild-type LC1 allele.

We first examined the whole brain, which revealed no difference in allele-specific *Pou3f2* expression as a result of LC1 KO (Figure 4.6C). Since the LacZ transgenic reporter assays suggested that LC1 is active in the amygdala and cortex (Figure 4.4), we analyzed these two regions separately. No difference was observed in the anterior cortex. Surprisingly, however, in the microdissected amygdala region, the LC1 KO allele was associated with ~8% higher *Pou3f2* expression. This suggests that LC1 normally acts as a silencer in the amygdala, contrary to the expectation that it acts as an enhancer there. Together, these data indicate that LC1 has region-specific effects on *Pou3f2* expression.

We conducted an analogous series of ASE experiments by crossing humanized KI mice to *Pou3f2* 3' UTR variant animals. However, we did not observe any allele-specific changes in *Pou3f2* expression associated with rs77910749 in the whole brain, microdissected amygdala, or microdissected cortex. These data suggest that rs77910749 does not affect *Pou3f2* transcript levels to an extent that is quantifiable by these assays (Figure 4.6C).

4.4.8 The novel CpG site created by rs77910749 is methylated at a low frequency in the developing mouse brain

DNA methylation of enhancers is associated with a decrease in chromatin accessibility and a loss of enhancer activity (Thurman et al. 2012; Plank and Dean 2014), and methylation of a CpG site within a Pax6 binding motif has been associated with decreased *cis*-regulatory activity in one specific instance (Wang et al. 2011). We observed that rs77910749 creates a novel CpG site (Figure 4.3A), raising the possibility that this new CpG is methylated and/or that the methylation status of neighboring CpG's is altered, with possible implications for the activity of this CRE.

First, we surveyed available methylation data from human primary tissues and cell lines (Figure 4.S7). In concordance with chromatin accessibility and other epigenomic data (Figure 4.1), LC1 is essentially unmethylated in the early developing brain and neural progenitors, but methylated in non-neuronal tissues and the adult brain.

To probe for whether rs77910749 affects LC1 methylation in the developing brain, we conducted allele-specific bisulfite sequencing analysis of LC1 in the E14.5 brains of rs77910749 knock-in heterozygous mice. We also examined mice with a small (14 bp) deletion within LC1 ('LC1 Small Indel') (Figure 4.7A). In particular, we analyzed a ~400 bp region that contains endogenous CpG sites (sites #1-5 and site #7), plus the novel CpG site created by rs77910749 (site #6).

Overall, LC1 exhibited very low levels of methylation in the E14.5 brain for all alleles, as expected (Figure 4.7B). In rs77910749 heterozygous animals, site #6 was methylated at a low frequency (2/40 clones) (Figure 4.7B, pink arrows). Thus, the novel CpG site created by rs77910749 is methylated at low frequency and/or in a small population of cells *in vivo*. Additionally, while no dramatic allele-specific differences in methylation were seen across the region, there was lower methylation in the KI allele, particularly at CpG sites #4 and #5, which are physically closest to rs77910749 (Figure 4.8C).

Together, these data show that rs77910749 creates a site that is methylated *in vivo* at low levels, and there may be lower methylation of neighboring CpG sites. However, overall, the methylation status of LC1 is relatively unchanged by the presence of rs77910749. In fact, even with the Pax6 binding motif deleted in the LC1 Small Indel, LC1 methylation is unchanged, suggesting that LC1 methylation is relatively robust to elimination of a key TF binding site.

4.4.9 The effect of rs77910749 on chromatin accessibility in human fetal brain

We wondered whether rs77910749 affects the epigenomic status of LC1 in the developing human brain. To address this question, we conducted allele-specific analysis of chromatin accessibility in fetal human brains using DNase-seq data (Roadmap Epigenomics et al. 2015). We first inferred the genotype of donors (see Methods) and identified six donors heterozygous for rs77910749. Intriguingly, the brains from earlier time points (day 56 and 58) showed a read bias in favor of rs77910749, whereas the brains from later time points (day 96 and later) showed a read bias in favor of the wild-type allele (Figure 4.8 and Table 4.3). This raises the possibility that rs77910749 has stage-specific effects on chromatin accessibility, whereby it promotes chromatin openness early in fetal development and chromatin closure later in fetal brain development. Additional samples are needed to follow-up on these preliminary results, and ideally the samples would be directly genotyped for rs77910749.

4.4.10 LC1 knockout animals have essentially normal behavior

Next, we asked whether deletion of LC1 alters behavior. We subjected adult homozygous LC1 KO mice and wild-type siblings to a locomotion assay and sensorimotor battery, which established that the LC1 KO mice did not have gross abnormalities. We then assayed the animals for the following: spatial learning and memory (Morris water maze), conditioned fear, sensorimotor reactivity and sensory gating (acoustic startle and prepulse inhibition), and anxiety (elevated plus maze and open field test). The LC1 KO animals did not show reproducible deficiencies in any of these domains. Thus, we conclude that mice with deletion of LC1 are essentially normal as measured by standard behavioral assays.

4.4.11 Humanized rs77910749 knock-in mice have defective sensory gating

Lastly, we asked whether rs77910749 alters mouse behavior. In homozygous KI mice and wild-type siblings, no abnormalities in locomotion, sensorimotor battery, Morris water maze, conditioned fear, or elevated plus maze were seen. However, when we subjected the animals to acoustic startle/prepulse inhibition (PPI) testing, we found that the homozygous KI mice had a significant defect in PPI (Figure 4.9). PPI is a measure of sensory gating, and defective PPI is associated with BPD, especially mania (Perry et al. 2001). Thus, the rs77910749 knock-in mice have a specific defect in sensory gating, a psychiatric endophenotype.

4.5 DISCUSSION

In this study, we sought to identify the ‘causal variant’ underlying GWAS signals at the *MIR2113/POU3F2* locus associated with both higher cognitive performance and higher risk for BPD. We computationally identified and then experimentally tested the candidate causal variant rs77910749. We used multiple orthogonal approaches to elucidate the links between rs77910749, enhancer activity, gene expression, and organismal behavior. First, we probed the effect of rs77910749 on TF binding. Second, we assayed the enhancer activity of LC1 with transgenic reporter mice as well as with CRE-seq, implementing the latter assay in developing mouse cerebral cortex and human iPSC-derived cerebral organoids for the first time. Third, we studied the effects of LC1 deletion and rs77910749 knock-in *in vivo*.

Overall, we detected subtle but significant effects of rs77910749 on Pax6 binding and LC1 enhancer activity. This suggests that at the molecular level, rs77910749 exerts a small effect, or a large effect in a small population of cells. Notably, the GWAS signals at this locus had small effect sizes, accounting for several weeks of additional schooling and ~10-20% increased risk for

BPD (Rietveld et al. 2013; Muhleisen et al. 2014; Trampush et al. 2015; Hou et al. 2016; Okbay et al. 2016). These effect sizes are typical for GWAS's of complex diseases, including neuropsychiatric diseases. The relationship between the magnitude of the molecular effect of a causal variant and the magnitude of its phenotypic effect will depend on a range of factors, including gene-environment interaction, genetic modifiers, and the dose sensitivity of the relevant genes. Other non-coding GWAS loci of small effect (which represent the majority of GWAS signals) may reveal underlying causal variants with similarly small effect sizes.

Surprisingly, our transgenic reporter mice suggest activity of LC1 in not only the developing cerebral cortex and retina, but also in the amygdala. Interestingly, *Pax6* has known roles in the development of the cerebral cortex, retina, and amygdala (Warren et al. 1999; Marquardt et al. 2001; Tole et al. 2005). *Pou3f2* has known roles in the cerebral cortex and retina, and a suggested role in the amygdala (McEvelly et al. 2002; Sugitani et al. 2002; Kim et al. 2008a; Garcia-Moreno et al. 2010). The amygdala is one of the most strongly implicated brain regions in BPD, but its development is relatively poorly understood, in part because it is composed of many nuclei of diverse origins (Pabba 2013; Maletic and Raison 2014). Our study underscores the need to better understand amygdala development at the molecular level.

Notably, the regions of LC1 enhancer activity represent only a subset of the spatial pattern of *Pou3f2* expression. Given that the *MIR2113/POU3F2* intergenic region contains many dozens of fetal brain-specific DHSs (and might even be considered a 'superenhancer'), we hypothesize that the full range of *Pou3f2* expression is attained via the action of multiple CREs in this region, possibly in combination, with some degree of functional redundancy among the CREs (i.e., 'shadow enhancers') (Hong et al. 2008; Hnisz et al. 2013). In addition to potential functional redundancy among CREs, *Pou3f2* is functionally redundant with *Pou3f3* in the mouse

cerebral cortex, which may provide additional buffering against the effects of mutations in LCI1 (McEvelly et al. 2002; Sugitani et al. 2002).

Besides rs77910749, what other candidate variants ought to be considered? Since the cognition and BPD GWAS's identified common variants of small effect at 6q16.1, we assumed that the causal variant is also a common variant. Given the reproducibility of the 6q16.1 signal in independent GWAS cohorts, this is the most likely scenario. However, we cannot rule out the possibility that the GWAS signals are attributable to very rare variants with large effects. Indeed, ultra-rare variants in highly constrained genes have been associated with decreased cognition and educational attainment in the general population (Ganna et al. 2016). Interestingly, a rare coding mutation in *RIMS1* is associated with genetic enhancement of cognition (Sisodiya et al. 2007). This raises the possibility that other rare variants, including non-coding variants, confer increased cognitive ability.

In prioritizing candidate causal variants, we assumed that the relevant tissue was the developing brain. However, it is possible that the causal variant exerts its effect in another tissue, such as the immune system, which is increasingly recognized as a major player in complex neuropsychiatric diseases such as schizophrenia and Alzheimer's disease (Muller et al. 2015; Da Mesquita et al. 2016). Additionally, we used phylogenetic conservation as a marker for functionality, but it is possible that the causal variant falls within a functional CRE that has undergone evolutionary modeling, or even within a human-specific CRE (Vierstra et al. 2014). Of course, it is also possible that the causal variant falls outside of a CRE and acts via an altogether different mechanism. Finally, we recognize that multiple variants in a haplotype block may be acting together in non-additive combinations to confer disease risk, such that there may not be a single dominant 'causal variant.'

For neuropsychiatric diseases such as BPD whose etiologies are poorly understood, detection of biologically meaningful effects should provide novel insights into disease pathways, but there are two major bottlenecks. First, the clinical phenotypes of neuropsychiatric diseases are often highly complex and heterogeneous, and the underlying genetics is likely to be as well. Second, while experimental models for neuropsychiatric diseases (including mouse models, non-human primates, iPSCs-derived neurons and organoids) have greatly improved over the past decade, they still have serious limitations. The choice of the experimental assay system is critical: it is possible that certain physiologically relevant deficits will manifest only in certain cell types or species and under certain environmental conditions. For *cis*-regulatory variants, this is a particularly acute issue, given the functional redundancy and buffering that occurs at the level of TF binding, combinatorial action of CREs, and gene regulatory network feedback loops. In some cases, sensitized genetic backgrounds and environmental perturbations may be necessary to unmask disease-relevant effects. As multiplex CRE reporter assays and CRISPR-Cas technologies continue to evolve in parallel with the development of neurobiological experimental models, the functional study of *cis*-regulatory variants relevant to neuropsychiatric disease will continue to accelerate.

4.6 METHODS

4.6.1 Animals

Mice were kept on a 12 hour light/dark cycle at ~20-22 °C with free access to food and water. Pregnant dams were euthanized with CO₂ anesthesia and subsequent cervical dislocation. For timed pregnancies, mating occurred overnight and the next day was considered embryonic day E0.5. All experiments were conducted in accordance with the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health and approved by the Washington University in St. Louis Institutional Animal Care and Use Committee (protocol #20140072).

4.6.2 DNase-seq data

The following human fetal DNase-seq data from Roadmap Epigenomics were visualized in the UCSC Genome Browser (Karolchik et al. 2014; Roadmap Epigenomics et al. 2015). Donor name, age, sex, and GEO accession are listed: fBrain #1 (donor H-23284, 96 day female, GSM595928), fBrain #2 (donor H-22911, 117 day female, GSM595920), fBrain #3 (donor H-22510, 122 day male, GSM530651), fHeart: (donor H-23604, 110 day female, GSM665830), fKidney (donor H-22676, 122 day sex unknown, GSM530655), fLung (donor H-22727, 101 day sex unknown, GSM530662), and fThymus (donor H-23964, 98 day female, GSM701537). The following mouse (C57BL/6) DNase-seq data from ENCODE were visualized (The ENCODE Project Consortium 2012). Mice were 8 weeks old unless otherwise indicated: E14.5 brain (GSM1014197), E18.5 brain (GSM1014184), adult brain (GSM1014151), P1 retina (GSM1014188), P7 retina (GSM1014198), adult retina (GSM1014175), adult heart (GSM1014166), adult kidney (GSM1014193), adult lung (GSM1014194), E14.5 liver (GSM1014183), adult liver (GSM1014195), and adult thymus (GSM1014185).

4.6.3 Calculation of linkage disequilibrium (LD)

Unless otherwise indicated, linkage disequilibrium (r^2 and D') are based on EUR 1000G Phase 1, as calculated by HaploReg V4.1 (Ward and Kellis 2012a).

4.6.4 Analysis of primate genomes

Variant calls (SNPs and indels) for primate genomes (Prado-Martinez et al. 2013) were downloaded as VCF files in hg18 from <https://eichlerlab.gs.washington.edu/greatape/data/VCFs/>. VCFtools v0.1.10 (Prado-Martinez et al. 2013) was used to obtain variants in the interval Chr6:98,673,000-98,674,000 in hg18, which is equivalent to Chr6:98,566,279-98,567,279 in hg19. Variants in this 1 kb window were manually examined for rs77910749, which is at chr6:98,673,228 in hg18 or chr6:98,566,507 in hg 19. The LiftOver tool on the UCSC Genome Browser was used to convert between hg18 and hg19 (Karolchik et al. 2014).

4.6.5 Motif analysis

For ‘SELEX PWM’ scores, FIMO in MEME v4.9.1 (Bailey et al. 2009; Grant et al. 2011) was used with the default p-value threshold (0.0001) to scan for TF motif occurrences. TF motifs used as input were from (Jolma et al. 2013). The Pax6 motif was the only one identified that overlapped with rs77910749. For the endogenous sequence, the following Pax6 motif was found (plus strand of hg19): ‘TTGTCTGCTTGAATGGTCC’. For the variant sequence, the following Pax6 motif was found: ‘TTTGTCGCTTGAATGGTCC’.

For ‘ChIP-seq PWM’ scores, a PWM was generated by aligning the raw Pax6 ChIP-seq data (Sun et al. 2015) (GEO accession GSE66961), aligned to mm9 with Bowtie 2 (v2.2.5) (Langmead and Salzberg 2012), sorted with Picard (v2.1.0) (<http://picard.sourceforge.net/>),

filtered for alignment quality (-q 30) with SAMtools (v1.3) (Li et al. 2009), and PCR duplicates were removed with Picard (v2.1.0). Peaks were called using MACS2 (v2.1.0) (FDR < 0.01) (Zhang et al. 2008). Peak calls were partitioned into TSS-proximal (peak summit within -1kb to +100 bp of an annotated TSS) and TSS-distal sets using HOMER (v4.8) (Heinz et al. 2010). *De novo* motifs were identified using HOMER (200 bp regions centered on TSS-distal peak summits), and the highest scoring *de novo* motif was used.

The logo for the SELEX motif was generated in enoLOGOS using default parameters with *M. musculus* %GC (Workman et al. 2005). The logo for the ChIP-seq motif is from (Sun et al. 2015).

4.6.6 Electrophoretic mobility shift assays (EMSAs)

PAX6 is perfectly conserved at the amino acid level between mouse and human (Ton et al. 1992). PD and PD5a were ordered as gene blocks from Integrated DNA Technologies with *E. coli* codon optimization and cloned as NdeI/NotI fragments into the pET-28a(+) vector. Constructs were confirmed by Sanger sequencing. For protein expression, BL21 cells were transformed and induced with IPTG overnight at 16 °C. His-tagged proteins were purified with HisPur Ni-NTA Resin (Thermo Scientific), dialyzed with PBS, and concentrated with Amicon Ultra-4 10K MWCO (Millipore). Proteins were quantified with the Pierce BCA Assay Kit (Thermo Scientific).

Quantitative EMSAs were conducted essentially as previously described (Man and Stormo 2001; Lee et al. 2010). Plus strand DNA probes were ordered with FAM or ROX fluorophores (Integrated DNA Technologies), and annealed to (unlabeled) minus strand probes. Unlabeled plus and minus strand oligos were annealed and used for cold competition reactions.

DNA binding reactions were conducted light-protected at 4 °C for 1 hr. The binding reaction was conducted in 10 mM Tris pH 7.5, 100 mM KCl, 1 mM β -mercaptoethanol, 2.5 mg/mL BSA, 100 ug/mL poly(dI-dC), and 10% glycerol. Labeled probe concentration was 10 nM in binding reactions (8 nM for cold competition, with 500-fold molar excess of the unlabeled probes), and protein concentration was 1 μ M.

Protein-DNA complexes were separated on 10% TBE gels (Invitrogen) at 100 V for 90 min, light-protected at room temperature. Gels were imaged on a Typhoon Trio Variable Mode Imager with excitation laser at 532 nm, emission filter at 526 nm for FAM, and emission filter at 610 nm for ROX. Band intensities were quantified with ImageQuant.

4.6.7 Generation of transgenic reporter mice

The LC1-*Hsp68*-LacZ construct was synthesized by cloning a 951 bp fragment of LC1 (chr6:98,566,099-98,567,049 in hg19, which was initially obtained by PCR of human gDNA) into the HindIII and PstI sites of *Hsp68*-LacZ Gateway vector (Pennacchio et al. 2006). Sanger sequencing confirmed that LC1 matched the hg19 reference. The construct was linearized with HindIII and purified by gel extraction (Qiagen). The DNA was then microinjected into fertilized eggs of C57BL/6 x CBA hybrid mice and implanted into pseudopregnant dams using standard techniques (Hogan et al. 1994).

4.6.8 LacZ staining and histology

Embryos (age E14.5) were dissected in cold phosphate-buffered saline (PBS). The tail (plus yolk sac for transient transgenics) was saved for PCR genotyping with LacZ primers (Table 4.2). Embryos were rinsed with PBS with 0.1% Tween-20 and then fixed on ice for 90 min with

2% formaldehyde, 0.2% glutaraldehyde, 5 mM EGTA, and 2 mM MgCl₂ in 0.1 M phosphate buffer pH 7.3. After rinsing three times with wash buffer (0.1% sodium deoxycholate, 0.02% NP-40, 2 mM MgCl₂, and 0.5 mg/mL BSA in 0.1 M phosphate buffer pH 7.3), embryos were incubated with X-gal staining solution (5 mM potassium ferricyanide, 5 mM potassium ferrocyanide, 0.1% sodium deoxycholate, 0.02% NP-40, and 2 mM MgCl₂ in 0.1 M phosphate buffer pH 7.3. Incubation conducted at 37 °C overnight (up to several days, with fresh X-gal staining solution added every ~12 hr). Embryos were post-fixed with 4% paraformaldehyde in PBS and stored at 4 °C until whole-mount imaging. For cryosections, embryos were equilibrated in 30% sucrose/PBS and decapitated. The head was embedded in Tissue-Tek OCT (Sakura) and cryosectioned at 20 µm. Sections were rinsed with PBS and counterstained with Nuclear Fast Red (Sigma).

4.6.9 CRE-seq Nano library construction

To create the LC1 multimer constructs, individual 200 bp sequences (centered on the position of rs77910749, which is a deletion of a ‘T’) were obtained by PCR using template DNA with or without rs77910749, with primers to add restriction enzyme sites. These ‘monomers’ were ligated pairwise in two rounds to create the 4X multimer with or without the variant (NotI-LC1-XbaI-LC1-XhoI-LC1-XmaI-LC1-FseI). Note that the LC1 monomer with rs77910749 includes an additional ‘T’ base at the 3’ end, such that the length and base content is the same as the LC1 monomer without rs77910749. Multimer sequences were confirmed by Sanger sequencing. The multimer was then cloned into the NotI/FseI sites of the previously described CRE-seq vector, which has random 15 bp barcodes in the 3’ UTR (Shen et al. 2016). The 3.6 kb *POU3F2* promoter encompassing chr6:99,279,024-99,282,671 (hg19) was obtained by PCR of

human gDNA. The basal *rho* promoter-GFP cassette of the CRE-seq vector was replaced with a 3.6 kb *POU3F2* promoter-GFP cassette between the FseI and AscI sites. The LC1 multimer was cloned into the NotI/FseI site, and individual colonies were picked for Sanger sequencing with Barcode_seq_R to determine barcode sequences (Table 4.2). For the promoter-only control constructs, there was no insert upstream the 3.6 kb *POU3F2* promoter. Twenty barcoded constructs were obtained for each of the LC1 REF multimer, the LC1 VAR multimer, and the promoter-only control. Each pool of twenty constructs was maxipreped (Invitrogen). For electroporations, the maxipreps were pooled in a mass ratio of 1:1:2 of LC1 REF, LC1 VAR, and promoter-only control.

4.6.10 Mouse cerebral cortex electroporations

Ex vivo cerebral cortex electroporation of E12.5 CD-1 mouse embryos (from timed pregnant dams) was conducted essentially as previously described (Nichols et al. 2013). The CRE-seq Nano library (2.5 µg/uL) was pooled with pDcx-DsRed (1 µg/uL) for a total of 3.5 µg/uL DNA. To visualize the injection, ~0.02% Fast Green dye was added. DNA was injected with a pulled glass pipette and Hamilton syringe. Electroporation was conducted with BTX ECM830 (Harvard Apparatus) with the following settings: 33 V, 50 ms pulse duration separated by 950 ms intervals, for 5 pulses. After electroporation, the head was transected just superior to the level of the eye and transferred on ice to explant media (50% DMEM (Gibco), 50% F12 (Gibco), 1X GlutaMAX (Gibco), 100 U/mL penicillin, and 100 µg/mL streptomycin). Up to three heads were arranged on a 25 mm circular Whatman Nucleopore 0.2 µm filter (with the cut surface against the shiny side of the filter), which floated in one well of a 6-well dish containing explant media with supplements (1X B27 (Gibco) and 1X G5 (Gibco)). Explants were incubated

at 37 °C with 5% CO₂. After two days in culture, the electroporated regions were microdissected under a fluorescent microscope (Leica MZ16 F) in cold HBSS with calcium and magnesium and stored in TRIzol (Invitrogen) at -80 °C. Each biological replicate consisted of electroporated tissue from multiple (5-8) cortices.

For histology, tissue was fixed in 4% paraformaldehyde/PBS, embedded in 4% agarose, and vibratome sectioned at 100 µm. Sections were mounted with Vectashield (Vectorlabs), coverslipped, and subjected to laser confocal imaging (Zeiss LSM700) with ZEN 2009 software (Zeiss).

4.6.11 Human cerebral organoid electroporations

Human iPS(IMR90)-4 (WiCell) were cultured in mTeSR1 (STEMCELL Technologies) and passaged every 3-4 days at 1:10 with ReleSR (STEMCELL Technologies) on 6-well plates with Matrigel (Corning). Cells were maintained at 37 °C with 5% CO₂. Cells (passage 64) were differentiated following a protocol similar to (Pasca et al. 2015). On the day of passage (Day 0), cell aggregates were released with ReLeSR™ and allowed to float freely in a 100 mm low-bind Petri dish in neural induction media: Neurobasal (Gibco) supplemented with 1% B27 without vitamin A (Gibco), 1X GlutaMAX (Gibco), 3 µM IWR-1-endo (Wnt antagonist) (Calbiochem), 5 µM SB431542 (TGFβ inhibitor) (Calbiochem), 100 U/mL penicillin, and 100 µg/mL streptomycin. This media was replaced every 3-4 days. On Day 18, media was changed to cerebral growth media similar to that published in (Lancaster et al. 2013): 50% DMEM-F12 (Gibco) and 50% Neurobasal (Gibco) supplemented with 0.5% N2 (Gibco), 1% B27 (with vitamin A) (Gibco), 2.5 µg/mL human insulin (Sigma), 1X GlutaMAX (Gibco), 0.5X MEM-

NEAA (Corning), 25 μ M β -mercaptoethanol, 100 U/mL penicillin, and 100 μ g/mL streptomycin. This media was replaced every 3-4 days.

The CRE-seq Nano library (1 μ g/uL) was co-electroporated with the loading control pCAG-DsRed (1 μ g/uL) (Matsuda and Cepko 2004), for a total of 2 μ g/uL DNA in PBS, into Day 88-109 organoids. The same equipment and similar protocol as for *ex vivo* retinal electroporations was used (Montana et al. 2011b). Four organoids were loaded into an electroporation chamber and allowed to float freely. Electroporation settings were as follows: 35 V, 50 ms pulse duration separated by 950 ms intervals, for five pulses. Organoids were placed back into conditioned cerebral growth media and allowed to float freely. After 7 days in culture, organoids were rinsed with HBSS with calcium and magnesium and stored in TRIzol (Invitrogen) at -80 °C. Each biological replicate consisted of eight electroporated organoids.

Organoids were imaged as live whole mounts with an inverted fluorescent microscope (Nikon Eclipse TE300). For antibody staining, organoids were fixed in 4% paraformaldehyde/PBS for 45 min, equilibrated in 30% sucrose/PBS and embedded in Tissue-Tek OCT (Sakura) for cryosections (12-14 μ m). The following antibodies were used: anti-Pax6 (PRB-278P at 1:300), anti-Pou3f2 (sc-6029 at 1:80), anti-Ki67 (BD Pharmigen 550609 at 1:100). Note that the anti-Pou3f2 antibody recognizes both Pou3f2 and Pou3f3 (Yamanaka et al. 2010). Confocal imaging was conducted on a BX61 WI microscope (Olympus) with a DSU spinning disk and ORCA-ER CCD camera (Hamamatsu). Images were processed with MetaMorph software (Molecular Devices).

4.6.12 CRE-seq Nano tissue processing and data analysis

RNA and DNA were isolated with TRIzol (Invitrogen), treated with TURBO DNase (Ambion), and purified with RNeasy Mini (Qiagen) as previously described (Shen et al. 2016). RNA (~0.5-1 µg) was then reverse-transcribed with SuperScript IV (Invitrogen), and the resulting cDNA was treated with RNaseH. The barcode region of the cDNA and DNA was amplified by PCR with Nano_initial_PCR primers (Table 4.2) using Phusion (New England BioLabs) and as follows: 98 °C for 30 sec, 16-22 cycles of 98 °C for 10 sec, 64 °C for 30 sec, 72 °C for 30 sec, and finally 72 °C for 5 min. The number of PCR cycles was: 16 for DNA, 20 for mouse cortex cDNA, and 22 for organoids cDNA. Samples were then prepared for amplicon-seq (see below).

Data analysis was conducted similarly as in (Shen et al. 2016). Barcode sequences were extracted, requiring a perfect match to one of the sixty known 15 bp barcodes plus 6 bp of flanking sequence on either side (i.e., 27 total bp) in either the forward or reverse direction (since sequencing adapters were ligated non-directionally). The RNA read count was normalized to the DNA read count for each barcode and then averaged across the twenty barcodes to yield the overall activity of a construct type ('Ref', 'Var', or promoter-only) in a given biological replicate.

4.6.13 CRISPR-Cas mice generation

CRISPR/Cas9 reagents were generated at the Genome Engineering and iPSC center at Washington University School of Medicine (St. Louis, MO). For the 'LC1 KO', a pair of guides flanking LC1 was to delete the intervening sequence. Multiple lines were generated with nearly identical deletions, but the line with the deletion chr4:23,438,846-23,439,893 for all 'LC1 KO' experiments. For knock-in of rs77910749, a single-stranded donor oligo centered on the variant

(with ~60 bp of homology on either side) was injected with a central LC1 guide for homologous recombination (Wang et al. 2013). For the ‘*Pou3f2* 3’UTR variant’, a single guide in the 3’ UTR was used to generate a 4 bp deletion (chr4:22,412,587-22,412,590). For the ‘LC1 Small Indel’, the central LC1 guide was used alongside a central LC5 guide, such that this strain carries a 14 bp deletion (chr4:23,439,327-23,439,340) plus an insertion of ‘C’ at the site of the deletion, as well as a 103 bp deletion in LC5 (chr4:23,417,446-23,417,548).

All CRISPR-Cas lines were generated in a C57BL/6J background. For pronuclear microinjections, hormone-primed females were mated to generate embryos (E0.5), which were subjected to pronuclear micro-injection of 2.5 ng/μl guide RNA and 5 ng/μl of Cas9 mRNA. Embryos were transferred to the oviducts of pseudo-pregnant recipient females. CRISPR-Cas guides, knock-in oligo sequence, as well as genotyping information (primer sequences and PCR conditions) are provided in Table 4.2. Founders (F0’s) were outbred to C57BL/6J. After the F1 generation, LC1 KO animals were genotyped by PCR only. Otherwise, CRISPR-Cas alleles were verified by Sanger sequencing of PCR products.

4.6.14 Allele-specific expression (ASE) analysis

E14.5 embryos were harvested in cold HBSS with calcium and magnesium, and brain tissue was rapidly dissected and stored in TRIzol (Invitrogen) at -80 °C. For ‘whole brain’ dissection, the olfactory lobes were left intact, and the brain was transected coronally at the posterior edge of the cortex (i.e., through the midbrain). For ‘amygdala region’ and ‘anterior cortex’ microdissection, the brain was additionally transected coronally approximately at the anterior-posterior level of the middle cerebral artery in the circle of Willis. The anterior tissue (with olfactory lobes removed) was harvested as ‘anterior cortex’. The inferior and lateral

portions of the posterior tissue from the same brain were harvested as ‘amygdala region’. For rs77910749 knock-in ‘whole brain’, only the left half was used for ASE (the right half was harvested for ChIP). Tail tissue was saved from each embryo for genotyping of the LC1 region and *Pou3f2* 3’ UTR region (Table 4.2).

For sequencing, RNA was extracted with TRIzol (Invitrogen), treated with TURBO DNase, and purified with RNeasy Mini (Qiagen). RNA (~1-2 µg) was then reverse-transcribed with SuperScript IV (Invitrogen) and treated with RNase. The 3’ UTR of *Pou3f2* was amplified with Pou3f2_3UTR_F and Pou3f2_3UTR_R primers (Table 4.2) using Phusion (New England BioLabs) as follows: 98 °C for 30 sec, 20-22 cycles of 98 °C for 10 sec, 64 °C for 30 sec, 72 °C for 30 sec, and finally 72 °C for 5 min. The number of PCR cycles was: 20 for whole brain, 21 for half brain, and 22 for microdissected regions. Samples were then prepared for amplicon-seq (see below). Sequence reads containing ‘CGTATATATATGGG’ (wild-type 3’ UTR) or ‘TGCGTATATGGGAT’ (variant 3’ UTR) were tabulated, and the ratio of reads (i.e., allelic bias) was calculated. The 3’ UTR variant itself causes ~10% increased *Pou3f2* mRNA levels compared to the wild-type 3’ UTR sequence, as determined from the animals that were heterozygous for the 3’ UTR variant and wild-type for LC1.

4.6.15 Allele-specific methylation analysis

Sample preparation and analysis conducted following a protocol similar to that previously described (Montana et al. 2013). DNA was extracted from brain tissue with DNeasy (Qiagen). For each biological replicate, ‘whole brain’ was dissected as described above for ASE, and the right half of the brain was used for bisulfite analysis. About 1 µg was bisulfite-converted with EpiTect Bisulfite Kit (Qiagen) and subjected to PCR with LC1_bis_F and LC1_bis_R

primers (Table 4.2). The resulting products were cloned into the pCR2.1 TOPO vector (Invitrogen) and Sanger sequenced with universal M13 reverse primer. Sequence data were analyzed and visualized with BISMAs using default parameters with removal of PCR duplicates (Rohde et al. 2010).

4.6.16 Amplicon-seq

Qubit dsDNA HS Assay (Invitrogen) was used to quantify samples. About 200 ng of cDNA or DNA was end repaired, 3' adenylated, and ligated to MiSeq adapters according to standard protocols (Son and Taylor 2011) (Table 4.2). The product was then amplified with a universal Illumina PCR primer and an indexed primer (Table 4.2) with Phusion as follows: 98 °C for 30 sec, 18 cycles (for ASE) or 20 (for CRE-seq Nano) cycles of 98 °C for 10 sec, 57 °C for 30 sec, 72 °C for 30 sec, and finally 72 °C for 5 min. Products were gel-purified and verified on an Agilent Bioanalyzer. For a given sequencing run, four or six indexed samples were pooled (controls were always processed and sequenced in parallel to the corresponding experimental samples). Samples were loaded at 7-8 pM concentration onto MiSeq for 2x250 bp sequencing as spike-in samples, representing ~10% of reads on a full lane, yielding ~1-2 million reads total per pool of samples. Reads were demultiplexed and checked with FastQC (Andrews 2010).

4.6.17 Allele-specific human fetal brain DNase-seq analysis

Publicly available human fetal brain DNase-seq data from Roadmap Epigenomics and ENCODE were downloaded. Aligned bam files were used when available; otherwise, reads were mapped to hg19 with Bowtie 2 to obtain aligned bam files (Langmead and Salzberg 2012). SAMtools (Li et al. 2009) was used for bam-to-sam conversion. Data were visualized on the

Integrative Genomics Viewer (IGV) (Robinson et al. 2011). To infer donor genotype, reads that overlapped the positions of the following variants in fetal brain DHSs were manually examined (r^2 and D' values are with respect to rs77910749): rs77910749 in LC1, rs13208578 in LC2 ($r^2 = 0.9$, $D' = 0.99$), rs12204181 in LC4 ($r^2 = 0.9$, $D' = 0.99$), and rs17814604 in LC5 ($r^2 = 0.42$, $D' = 0.97$). For donors inferred to be heterozygous for rs77910749, allele-specific read counts at LC1 were tabulated. Additional details are provided in Table 4.3.

4.6.18 Behavioral assays

A total of 10 homozygous LC1 knockout and 10 age-matched, sex-matched wild-type siblings, and a total of 12 homozygous rs77910749 knock-in and 12 age-matched, sex-matched wild-type siblings, were subjected to behavioral testing at the Washington University Animal Behavior Core as previously described (Dougherty et al. 2013). Animals were allowed to habituate in the testing facility for two weeks before testing was initiated at age 10-15 weeks. The following tests were conducted: 1-hour locomotor activity, sensorimotor battery, Morris water maze, conditioned fear, acoustic startle and PPI, elevated plus maze, and open field test. For the acoustic startle and PPI assays in the rs77910749 knock-in animals vs. wild-type animals, two independent cohorts were tested and data from these two cohorts were pooled for 24 homozygous rs77910749 knock-in and 24 wild-type control animals.

4.7 ACKNOWLEDGEMENTS

We would like to thank the Washington University Animal Behavior Core (David Wozniak), Center for Genome Sciences and Systems Biology (Jessica Hoisington-Lopez), Micro-injection Core (J. Michael White), Mouse Genetics Core (Mia Wallace), and Protein and Nucleic Acid Chemistry Laboratory (Misty Veschak). We thank the Alvin J. Siteman Cancer Center for the Genome Engineering and iPSC Center (GEiC), and Shondra Miller at GEiC. The Siteman Cancer Center is supported in part by NCI Cancer Center Support Grant #P30 CA091842, Eberlein, PI. We are grateful to the laboratory of Qiang Lu for the *pDcx-DsRed* construct. We thank Matthew Toomey for guidance with protein expression and Allison Loynd for assistance with methylation studies.

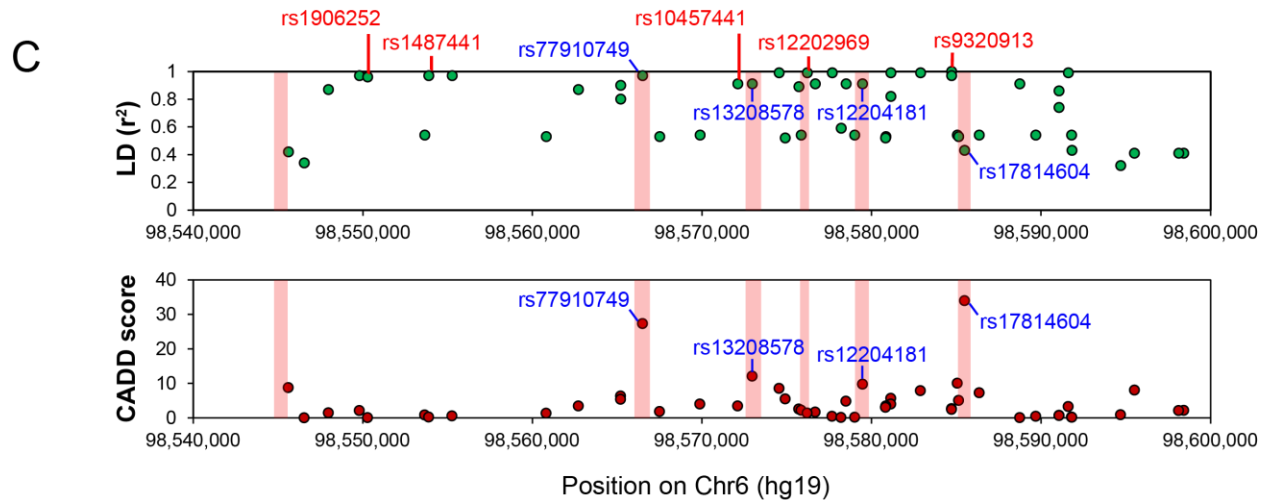
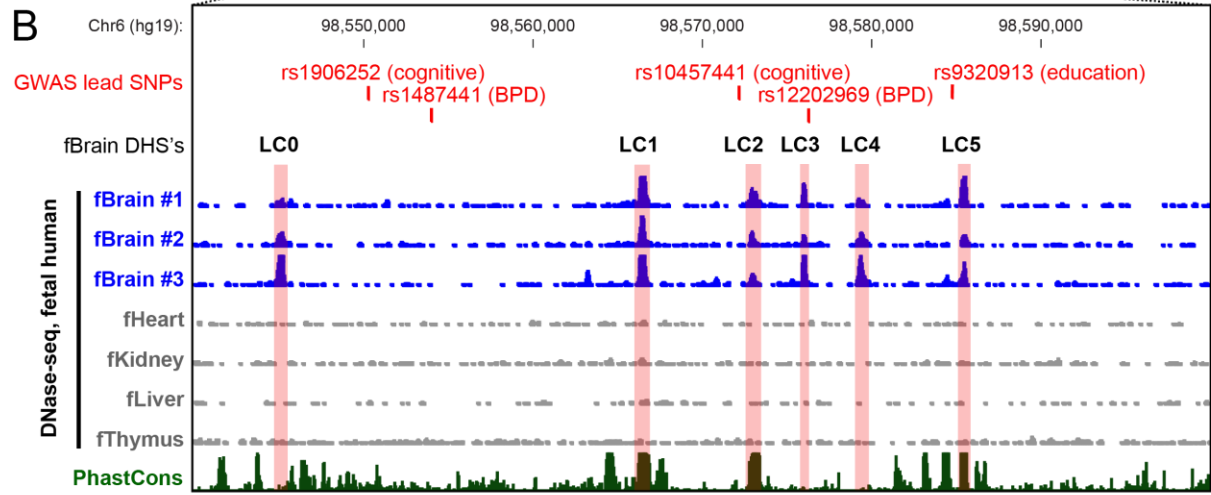
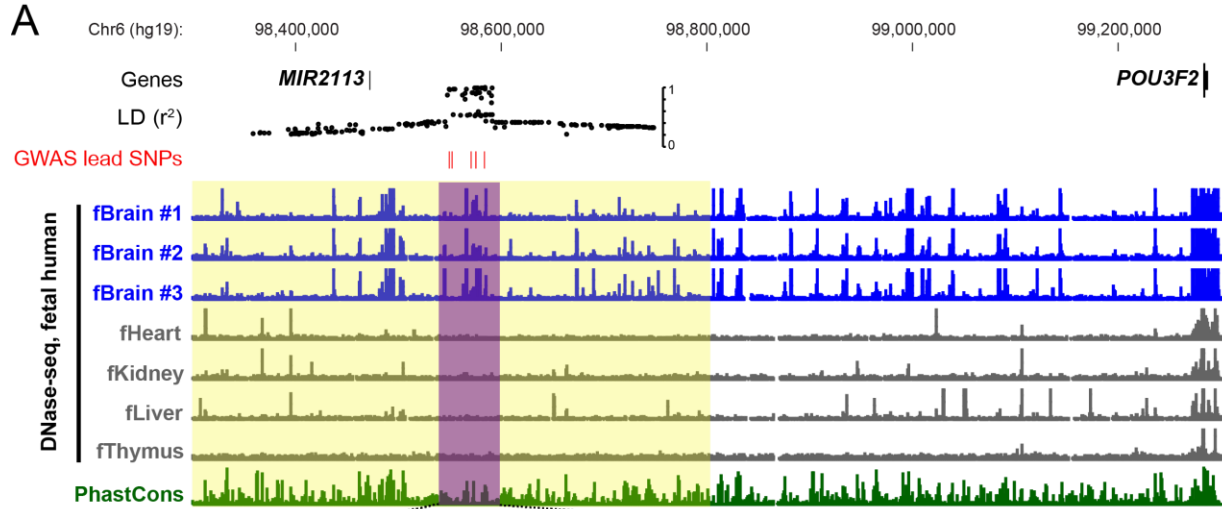


Figure 4.1. Prioritization of candidate variants at 6q16.1 associated with higher educational attainment, increased cognitive performance, and risk for bipolar disorder. (A) Genomic context (hg19, 1 Mb window) of the intergenic locus at 6q16.1 implicated in GWAS studies of educational attainment, cognition, and BPD. The ~0.5 Mb region identified by these studies (highlighted in yellow) contains a ~60 kb ‘local cluster’ region (highlighted in purple) with the highest LD. All variants in LD with rs9320913 ($r^2 > 0.2$) are shown. Note that the nearest protein-coding gene, *POU3F2*, is ~0.7 Mb away. DNase-seq data from three human fetal brains and four other human fetal tissues are shown (Roadmap Epigenomics et al. 2015). PhastCons depict 100-way vertebrate conservation (Siepel et al. 2005). The UCSC Genome Browser was used for visualization (Karolchik et al. 2014). (B) Enlarged view the 60 kb ‘local cluster’. Note the fetal brain (fBrain) DHSs (LC0 to LC5, pink highlight). Lead SNPs (red font)—rs9320913 for educational attainment (Rietveld et al. 2013; Okbay et al. 2016), rs1906252 for cognitive performance (Trampush et al. 2015), rs10457441 for cognitive performance (Davies et al. 2015), rs12202969 for BPD (Muhleisen et al. 2014), and rs1487441 for BPD (Hou et al. 2016)—are depicted. (C) Variants within the local cluster that are in LD with rs9320913 (as defined by $r^2 > 0.2$). Note the five lead SNPs (red font) and four variants that fall within LC1-5 (blue font). The r^2 values are shown (green dots). Phred-scaled CADD scores (blue dots) are from (Kircher et al. 2014).

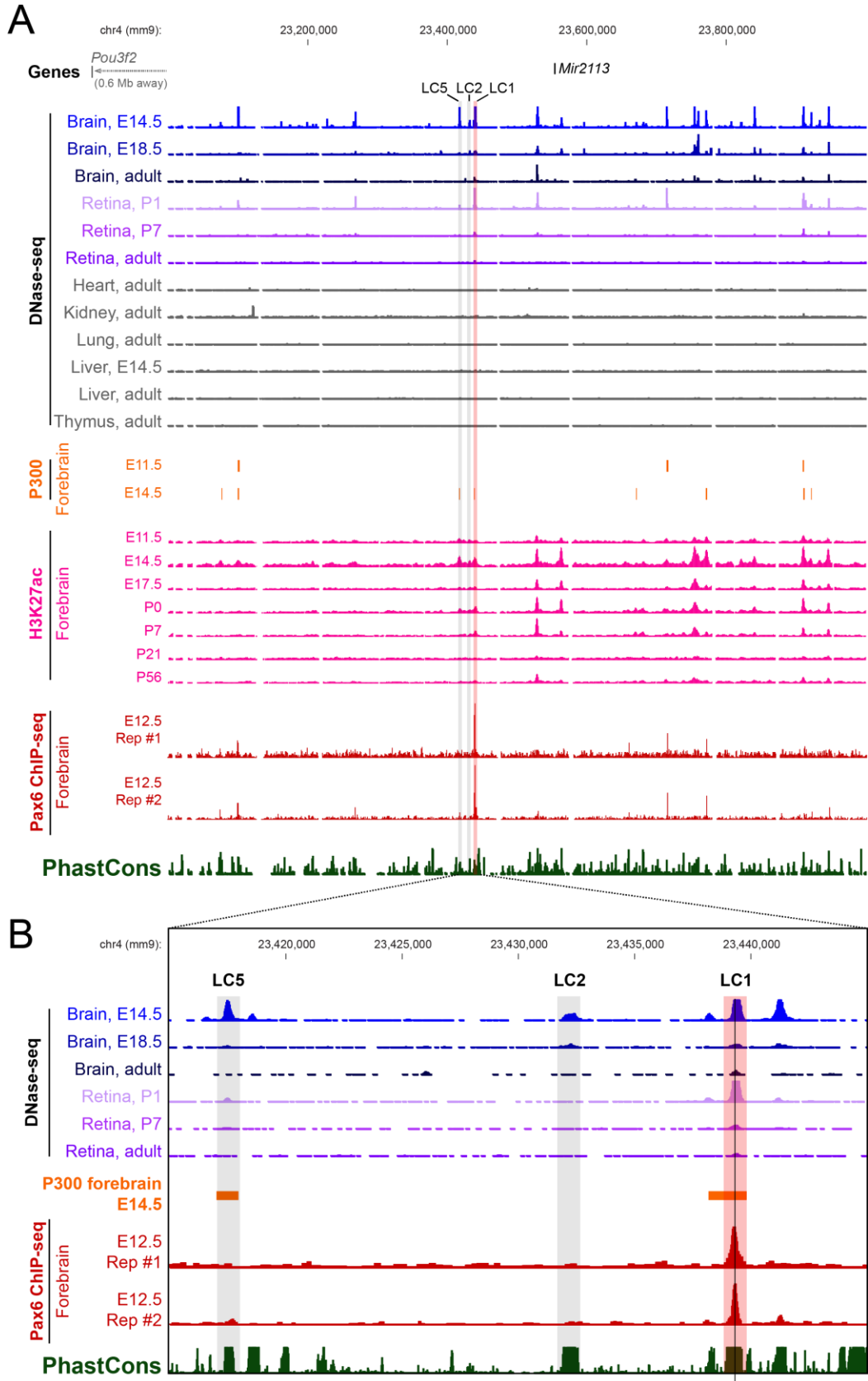


Figure 4.2. Epigenomic landscape around the orthologous LC1 region in mouse. (A) Genomic context (mm9, 1 Mb window) around mouse LC1 (pink highlight). Locations of orthologous LC2 and LC5 are also indicated (gray highlight). *Pou3f2* (gray font) is outside the window at chr4:22,409,242-22,415,513, i.e., ~1 Mb away from LC1, and is transcribed from the minus strand of DNA. *Mir2113* is a non-RefSeq gene identified by homology to the human sequence. DNase-seq data are from (The ENCODE Project Consortium 2012). P300 ChIP-seq data (orange tracks) are from E11.5 forebrain (Visel et al. 2009) and E14.5 forebrain (Wenger et al. 2013). H3K27ac ChIP-seq data (pink tracks) are from forebrain at the indicated ages; for ages with multiple replicates, only the first replicate is shown (Nord et al. 2013). Pax6 ChIP-seq data (dark red tracks) are from E12.5 forebrain; two replicates are shown with y-axis autoscaling (Sun et al. 2015). (B) Enlarged view of the 30 kb mouse ‘local cluster’. Note that LC1 overlaps with E14.5 brain and P1 retina DNase-seq peaks, E14.5 forebrain p300 peak, E14.5 forebrain H3K27ac peak, and E12.5 forebrain Pax6 ChIP-seq peak. The orthologous position of rs77910749 (black vertical line in LC1) falls within the middle of the Pax6 ChIP-seq peak.

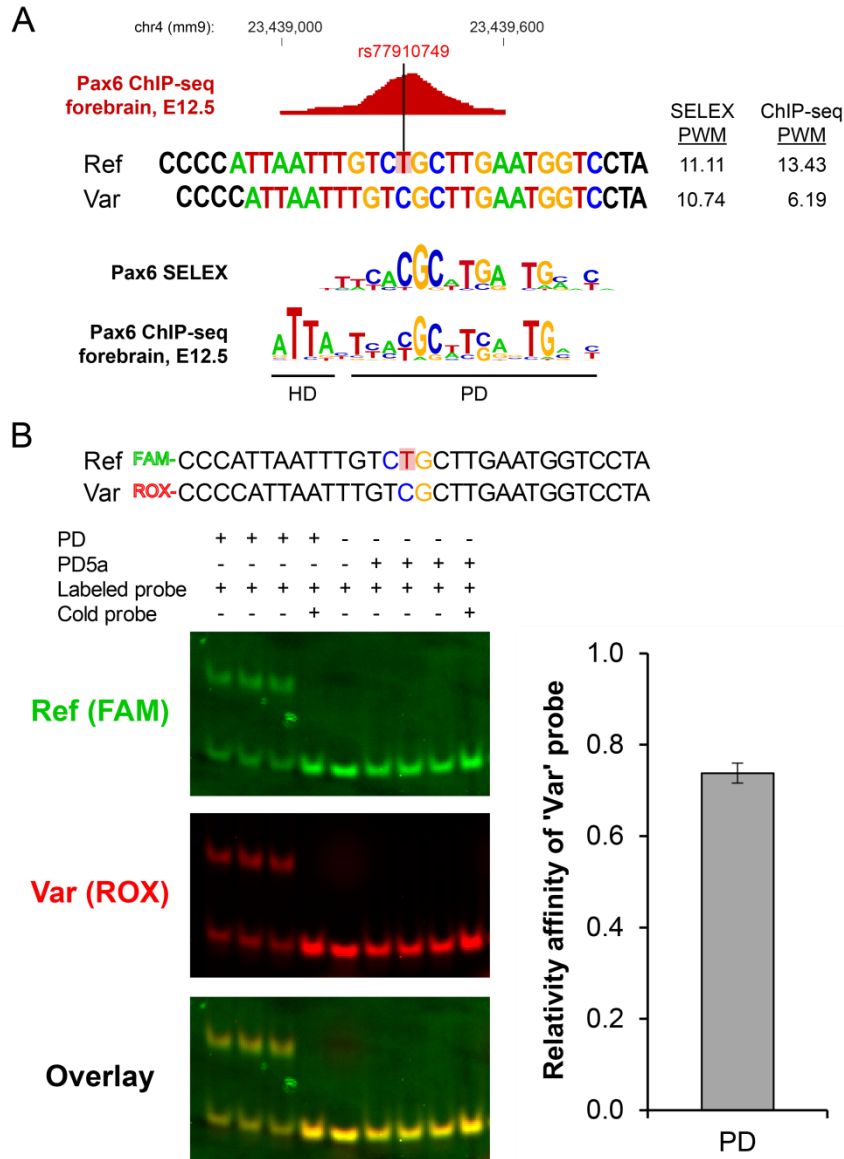


Figure 4.3. *In silico* and *in vitro* analysis Pax6 binding. (A) Comparison of the reference sequence ('Ref'), sequence with rs77910749 ('Var'), and Pax6 consensus motifs. The position of rs77910749 is indicated (red highlighted 'T'). Note that the reference sequence is perfectly conserved between mouse and human, and the minus strand of mm9 is shown. Motifs were scored using Pax6 SELEX and ChIP-seq position weight matrices (PWMs). For SELEX, a protein with Pax6 PD and HD domains was used (Jolma et al. 2013). The logo was generated in enoLOGOS (Workman et al. 2005). The E12.5 Pax6 ChIP-seq motif is based on (Sun et al. 2015), and only the PD PWM was used for FIMO analysis. (B) Quantitative EMSA assay. Reference (FAM) and variant (ROX) probes were fluorescently labeled and incubated with Pax6 PD or PD5a. For the 'cold competition' (lane 4), 500-fold molar excess of cold (i.e., unlabeled) probe was used. The bound and unbound fractions for the PD lanes were quantified and relative binding affinity was calculated according to (Man and Stormo 2001). Error bar represents SD across lanes.

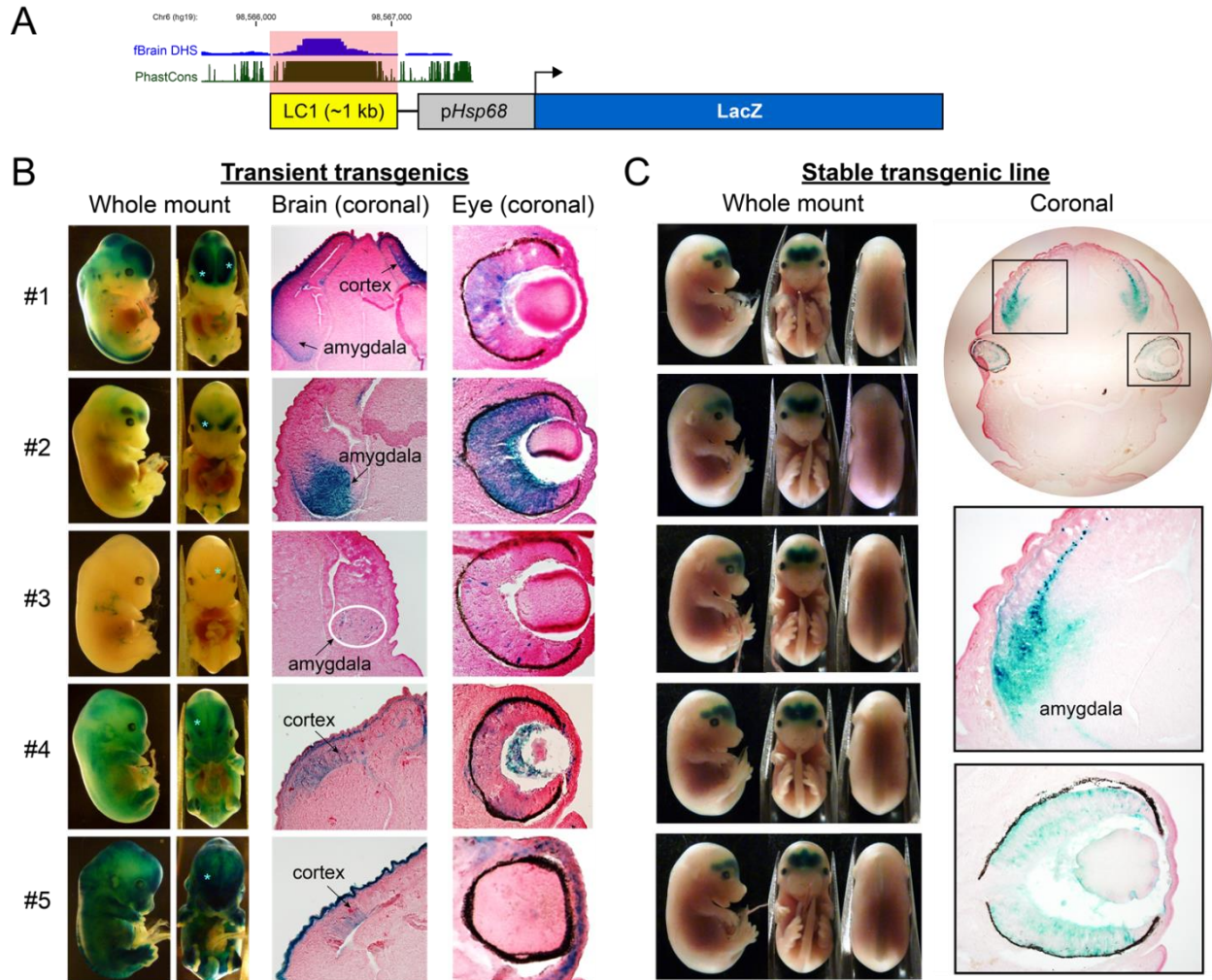


Figure 4.4. Transgenic reporter mice show evidence of LC1 activity in the developing CNS. Mice were generated that carried a reporter construct for wild-type human LC1 (951 bp fragment) on the *Hsp68* promoter, driving the expression of LacZ, which stains blue with X-gal (Pennacchio et al. 2006). (A) Schematic of the reporter construct (drawn to scale). (B) Transient transgenic embryos. Of seven genotypically positive embryos, five (#1-5 shown here) exhibited LacZ staining. Each mouse represents an independent integration event. Whole mount images of side and frontal views; light blue asterisks in the frontal views denote the approximate location of annotated regions in the brain coronal sections. For the brain coronal image of embryo #3, the white oval encircles sparse LacZ-expressing cells. Note that the entire head of the embryo was embedded. Close-up images of the eye are also shown. (C) Embryos from a stable transgenic line. Of three genotypically positive transgenic lines, only this line exhibited LacZ staining. All embryos look essentially identical, as expected for a given line. Side, frontal, and back views are shown (note the staining in the spinal cord, which is part of the CNS). Coronal section of head and corresponding enlarged images of the amygdala and eye are shown. Sections were counterstained with Nuclear Fast Red.

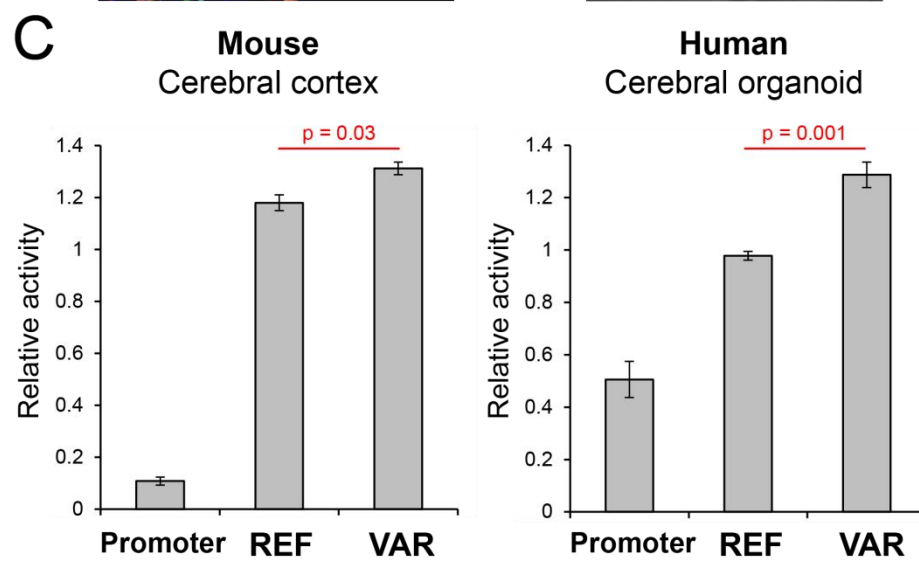
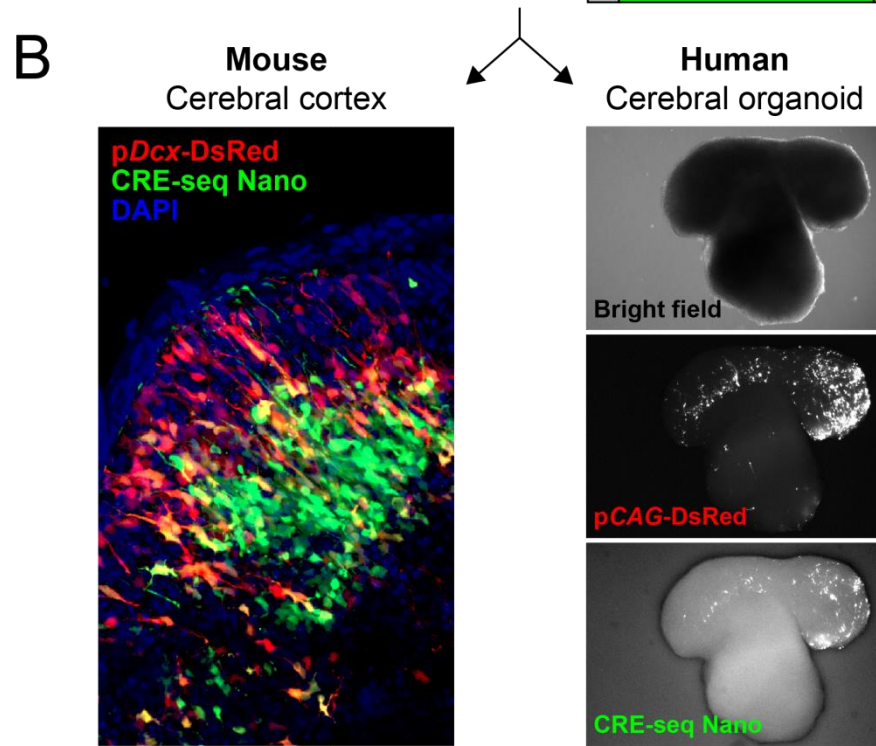
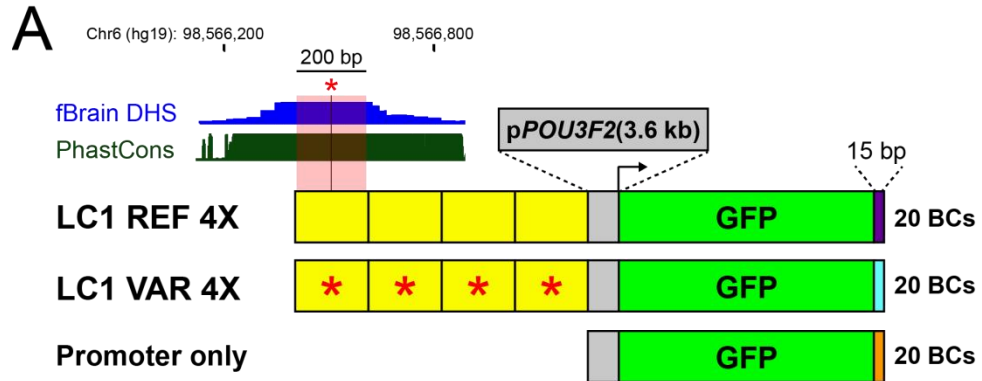


Figure 4.5. The variant rs77910749 causes a subtle increase in enhancer activity in developing mouse brain and human cerebral organoids. (A) Schematic of the CRE-seq Nano experimental design. Multimers (4X) of the central 200 bp of human LC1 were cloned upstream of a 3.6 kb *POU3F2* (human) promoter and GFP with unique 15 bp barcodes (BCs) in the 3' UTR. 'REF' indicates wild-type sequence and 'VAR' indicates the presence of rs77910749 (red asterisk), whose position is indicated by the black vertical line. Twenty barcoded constructs were generated for each of LC1 REF, LC1 VAR, and promoter-only. (B) Delivery of the library. Left: E12.5 mouse cerebral cortex was electroporated and harvested after 2 days *ex vivo*. A vibratome section shows expression of the library (GFP) in the deeper layers of the cerebral cortex. The co-electroporated control construct, *pDcx-DsRed*, is expressed in post-mitotic migrating neurons (Wang et al. 2007). DAPI is a nuclear counterstain. Right: Human iPSC-derived cerebral organoids were electroporated and harvested after 7 days *in vitro*. A whole mount image of a live organoid shows expression of the library (GFP). The co-electroporated control construct, *pCAG-DsRed*, marks electroporated cells. (C) Quantification of *cis*-regulatory activity by CRE-seq. P-values were calculated with two-tailed Student's t-test. Error bars indicate SEM between biological replicates (n = 3 for mouse cerebral cortex, n = 4 for human cerebral organoids).

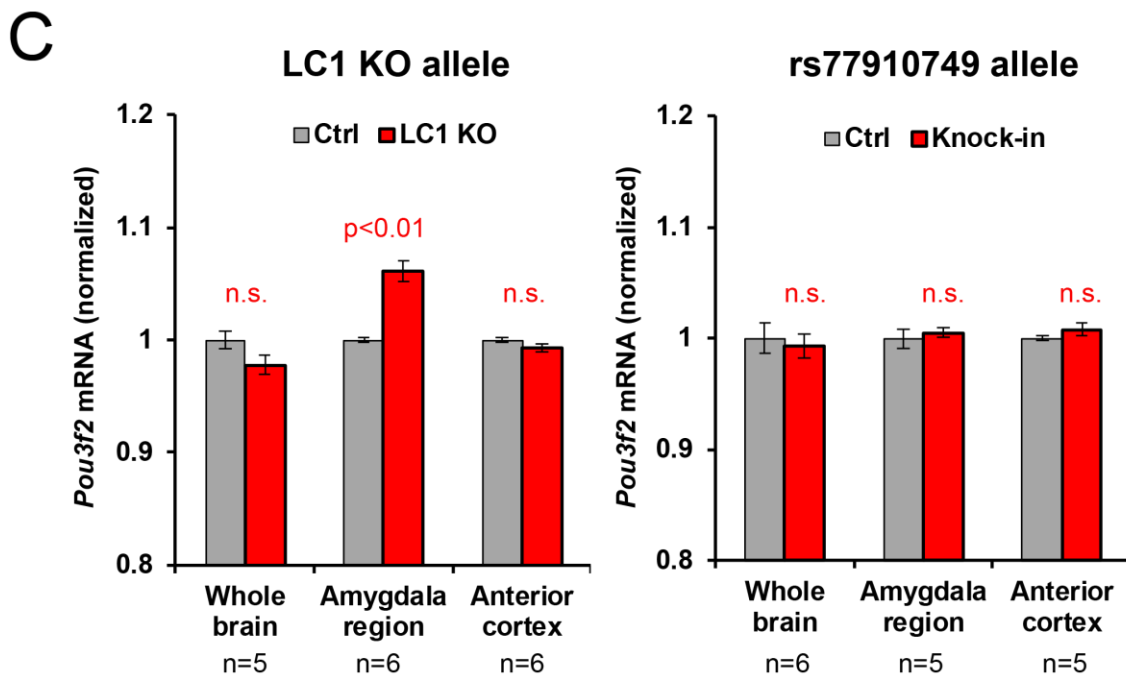
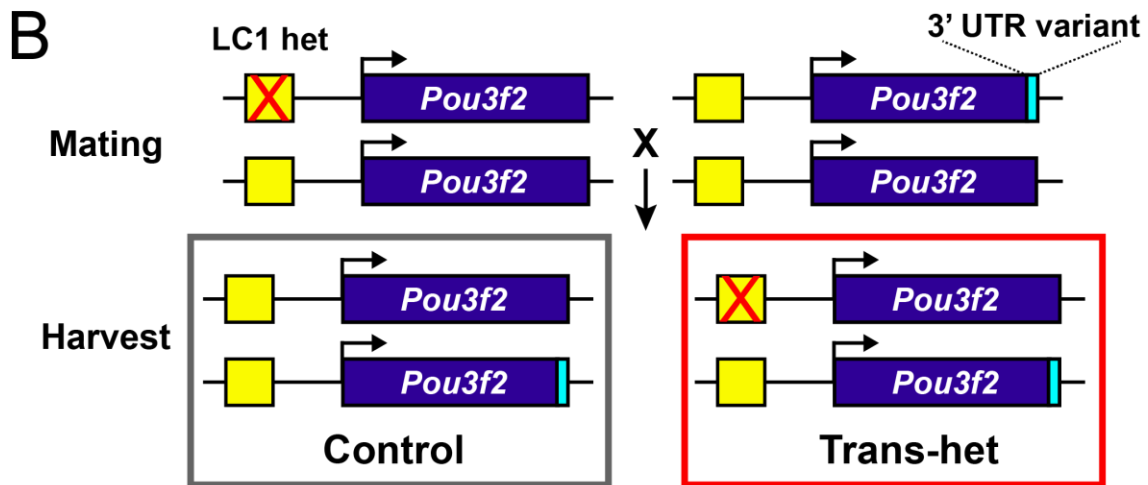
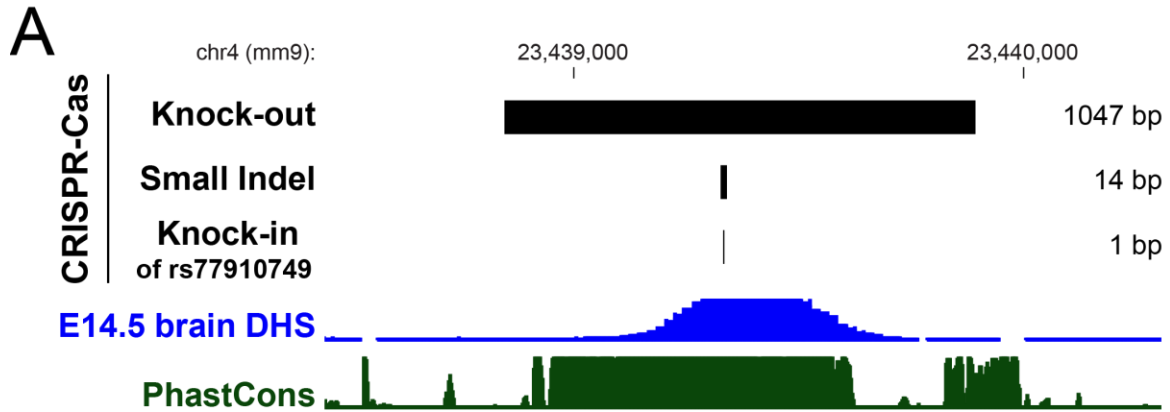


Figure 4.6. The effect of LC1 deletion on *Pou3f2* expression is region-specific. (A) An allelic series of LC1 mutants generated by CRISPR-Cas. Sizes of deletions are indicated. Note that rs77910749 ‘knock in’ is an introduction of a 1 bp deletion. (B) Schematic of ASE experimental design (not to scale). Mice heterozygous for an LC1 mutation were mated (E14.5 timed pregnancies) to mice with a variant in the 3’ UTR of *Pou3f2*, which served as a molecular barcode (light blue rectangle). Resulting trans-heterozygous mice (i.e., heterozygous for both the LC1 mutation and the 3’ UTR variant) were analyzed for allele-specific *Pou3f2* expression. Note the phasing, i.e., the LC1 mutation is in *cis* to the wild-type 3’ UTR. To account for any effects due to the 3’ UTR variant alone, control animals wild-type for LC1 and heterozygous for the 3’ UTR variant were included. (C) E14.5 whole brain, microdissected amygdala region, and microdissected anterior cortex were analyzed for allele-specific *Pou3f2* expression in control and trans-het LC1 KO animals (left panel), and in control and trans-het rs77910749 knock-in animals (right panel). Gray denotes controls, and red denotes trans-het animals. P-values were calculated with two-tailed Student’s t-test. Error bars indicate SEM between biological replicates. Sample size per condition is indicated (trans-het animals and matched controls; amygdala and anterior cortex samples were from the same embryos). Each biological replicate consists of tissue from one brain (amygdala and anterior cortex were harvested from the same brain).

Figure 4.7. Allele-specific methylation analysis of LC1. (A) Region within LC1 analyzed by bisulfite sequencing. The variant rs77910749 is a single bp deletion of ‘T’ (on the minus strand for mm9), creating a novel CpG site (site #6). (B) Bisulfite sequencing of E14.5 brain from mice that were heterozygous for rs77910749 knock-in (KI) allele (n = 4, left panels), or the LC1 Small Indel allele (n = 3, right panels). Each row represents a clone, and each column represents a CpG site. Note the two clones (pink arrow) of the KI allele, in which site #6 is methylated. Methylation was overall slightly higher in the LC1 Small Indel heterozygous animals than in rs77910749 KI heterozygous animals, suggesting a *trans* effect (the LC1 Small Indel allele also has a 103 bp deletion within LC5—see Methods). Red = methylated, blue = unmethylated, white = no data. CpG site #6 is not present in the wild-type allele or the small indel allele. (C) Quantification of methylation at each CpG site. Top: rs77910749 knock-in heterozygotes. Bottom: Small indel heterozygotes. Error bars indicate SEM. P-values were calculated with two-tailed Fisher’s exact test (reads across replicates were combined). N.D., no data.

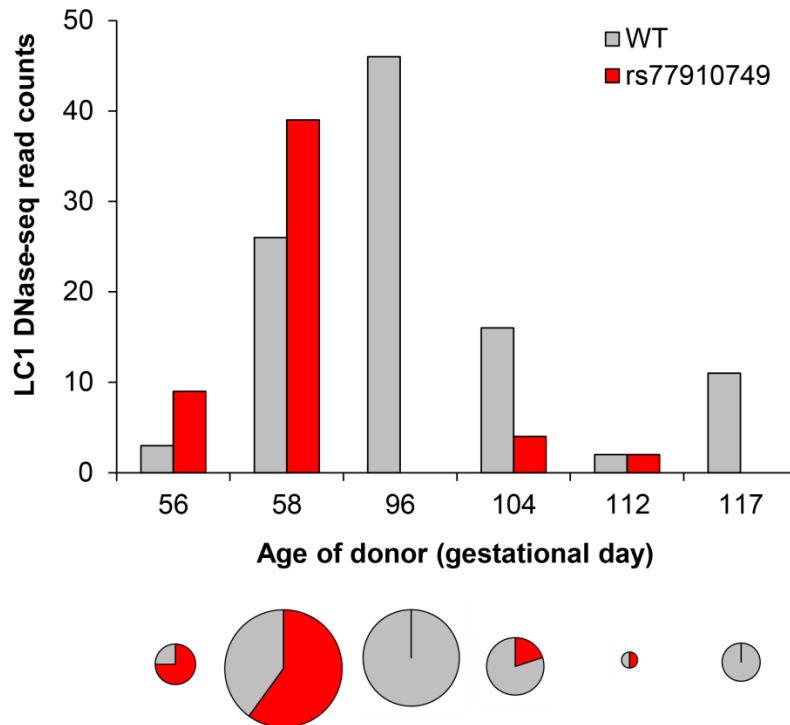


Figure 4.8. Human fetal brain allele-specific DNase-seq analysis. Human fetal brain DNase-seq data (Roadmap Epigenomics et al. 2015) from donors inferred to be heterozygous for rs77910749 (see Methods) were analyzed. Raw read counts are shown in the bar graph (top). Read proportions are shown in the pie charts, whose sizes roughly reflect total read number (bottom).

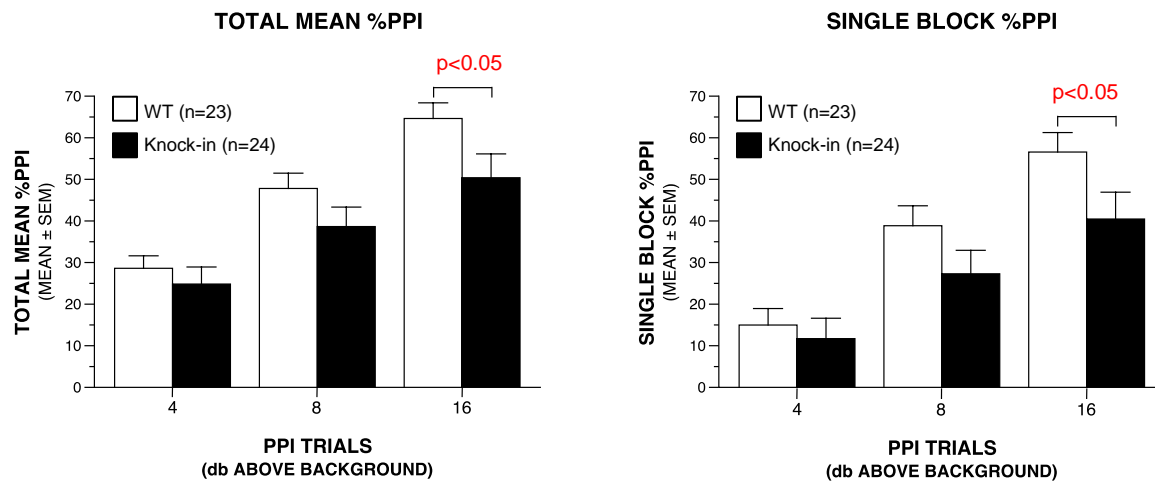


Figure 4.9. Prepulse inhibition (PPI) is defective in ‘humanized’ rs77910749 knock-in mice. Adult mice homozygous for the rs77910749 knock-in allele and wild-type (WT) siblings (age- and sex-matched) underwent acoustic startle testing with prepulse inhibition (PPI) assays. The knock-in (KI) animals showed defective prepulse inhibition that was statistically significant ($p < 0.05$, ANOVA) for the highest decibel (db) tested. One WT animal did not have a startle response at baseline and was excluded from the analysis. PPI measurements were normalized to baseline startle responses. Of note, baseline startle response magnitudes were lower in KI than WT animals ($p = 0.018$).

A

rs13208578



B

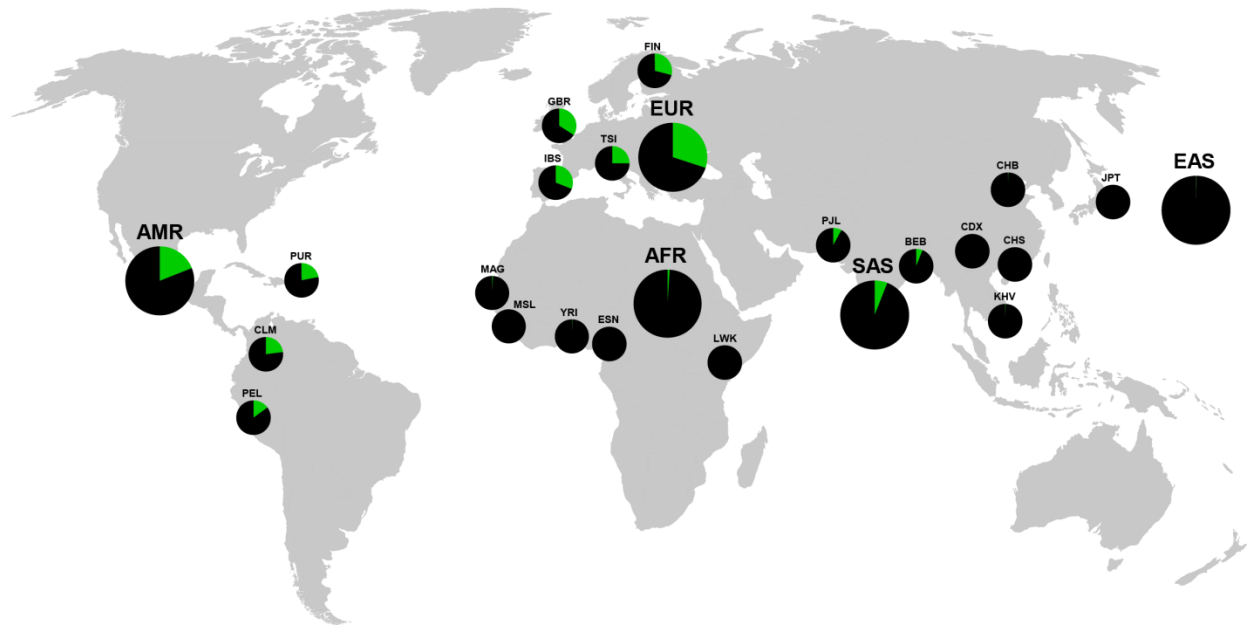
rs77910749



Figure 4.S1. Phylogenetic conservation of rs13208578 and rs77910749. Multiz alignments (100 vertebrates) as viewed on the UCSC Genome Browser (Blanchette et al. 2004; Karolchik et al. 2014). A 150 bp window is shown roughly centered on each variant (position of variant is highlighted in red): (A) rs13208578 (a substitution of ‘C’ to ‘T’), and (B) rs77910749 (a 1 bp deletion of a ‘T’).

A

rs17814604



B

rs77910749

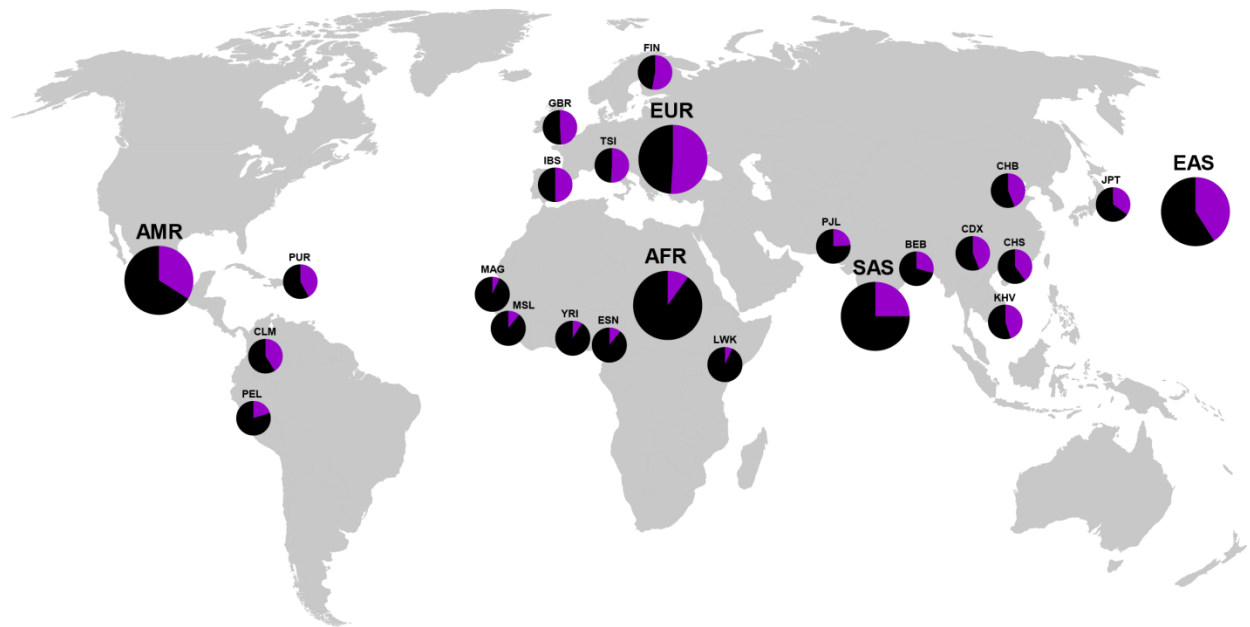


Figure 4.S3. Global distribution of rs17814604 and rs77910749 frequencies. Allele frequencies based on Phase 3 of the 1000 Genomes Project (Genomes Project et al. 2015) are shown for the major populations (large pie charts) as well as for subpopulations (small pie charts), with black indicating the reference allele: (A) rs17814604 (green allele) and (B) rs77910749 (purple allele). Abbreviations: AFR, African; AMR, American; BEB, Bengali in Bangladesh; CDX, Chinese Dai in Xishuangbanna, China; CHB, Han Chinese in Beijing, China; CHS, Southern Han Chinese, China; CLM, Colombian in Medellin, Colombia; EAS, East Asian; ESN, Esan in Nigeria; EUR, European; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian populations in Spain; JPT, Japanese in Tokyo, Japan; KHV, Kinh in Ho Chi Minh City, Vietnam; LWK, Luhya in Webuye, Kenya; MAG, Mandinka in The Gambia; MSL, Mende in Sierra Leone; PEL, Peruvian in Lima, Peru; PJJ, Punjabi in Lahore, Pakistan; PUR, Puerto Rican in Puerto Rico; SAS, South Asian; TSI, Toscani in Italy; YRI, Yoruba in Ibadan, Nigeria.

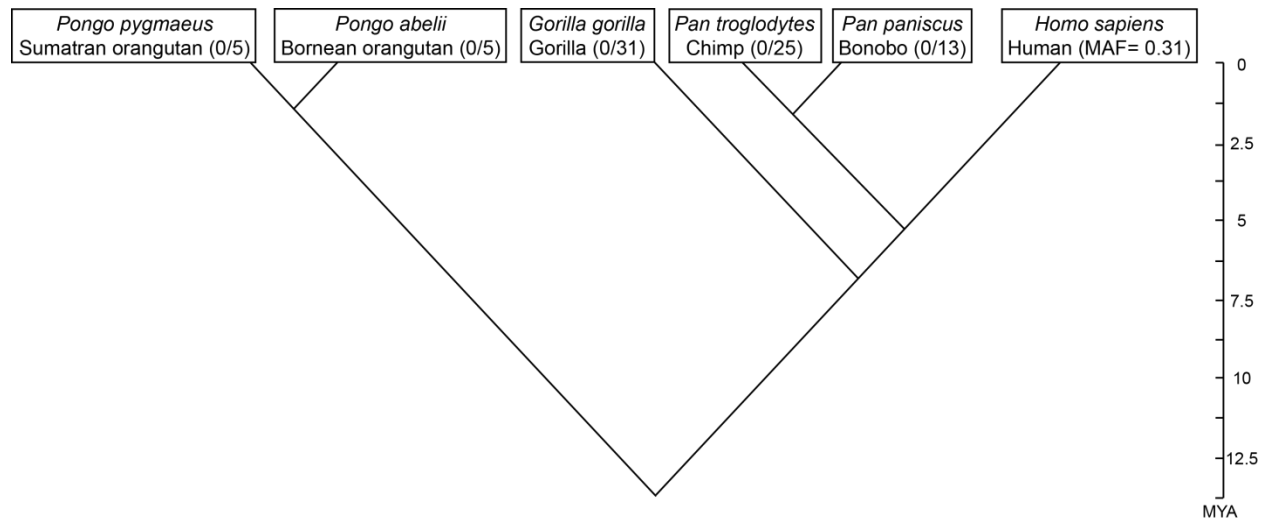


Figure 4.S4. Absence of rs77910749 from non-human primate genomes. The genomes of 79 individuals from five non-human primate species were examined for rs77910749, and none were found to contain this variant. Number of individuals for each species is indicated. Sequences and estimates of divergence times (in millions of years ago, MYA) are from (Prado-Martinez et al. 2013). Minor allele frequency (MAF) in humans is based on aggregate 1000 Genomes Phase 3 data (Genomes Project et al. 2015).

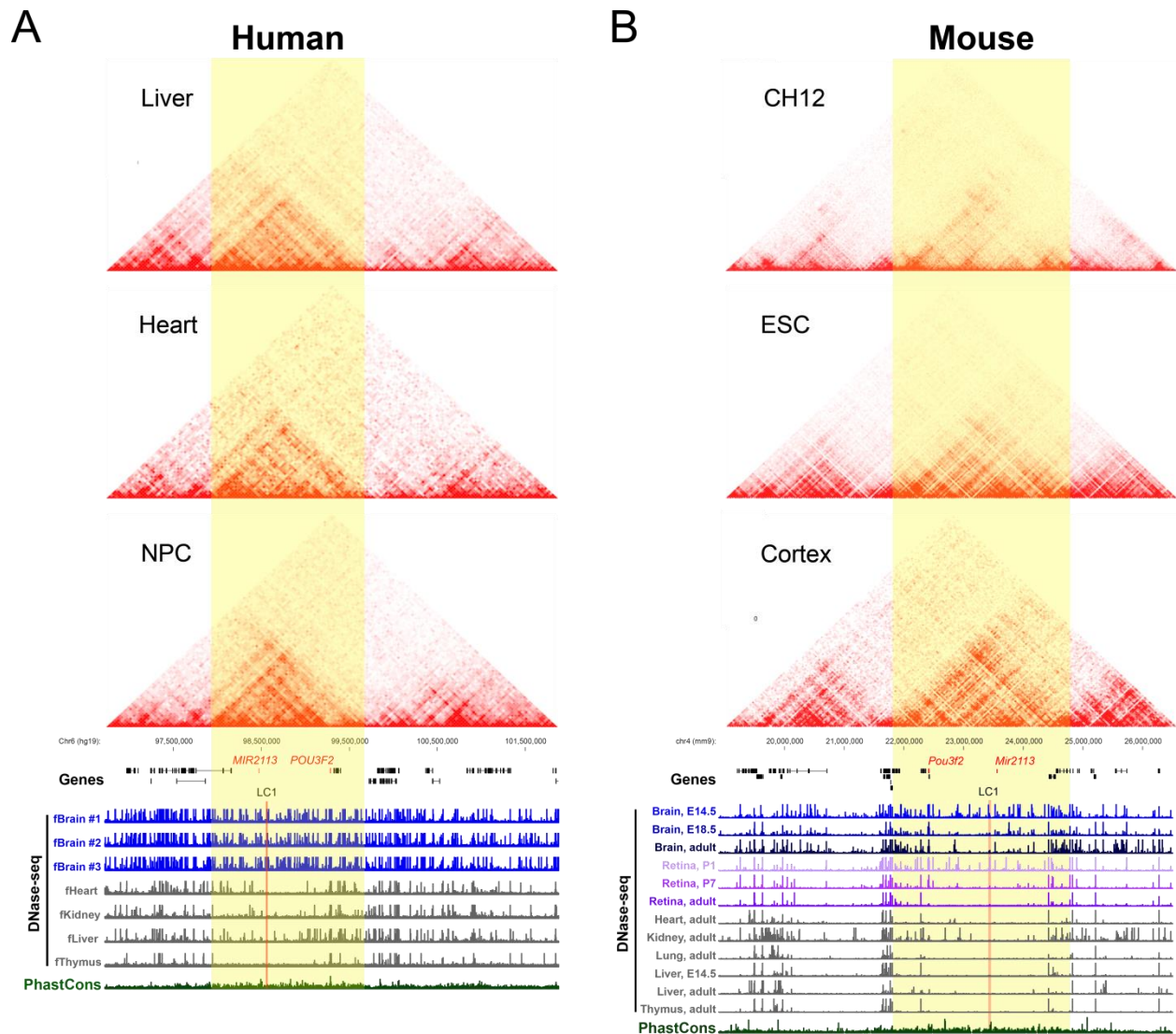


Figure 4.S5. LC1 falls within a conserved topologically associating domain (TAD). Published Hi-C data were visualized with default heat map scaling on the 3D Genome Browser (<http://www.3dgenome.org>). Darker red indicates higher frequency of interactions. Data are shown at 40 kb resolution, except for CH12 (25 kb resolution). The TAD containing LC1 is highlighted in yellow, and LC1 is highlighted in pink. Note the positions of *Mir2113/MIR2113* and *Pou3f2/POU3F2* (red font) within the TADs. In the mouse genome, the *Pou3f2* is transcribed from the minus strand of DNA. In the mouse genome, the region is inverted such that the relative orientation (LC1 upstream of *Pou3f2*) is preserved. Mouse *Mir2113* is a non-RefSeq gene identified by homology with the human sequence. (A) Human Hi-C data for left ventricle of heart, liver, and H1-derived neural progenitor cells (NPC) (Dixon et al. 2015; Leung et al. 2015). (B) Mouse Hi-C data for CH12 (a B cell lymphoma cell line), embryonic stem cells (ESC), and cerebral cortex (Dixon et al. 2012; Rao et al. 2014).

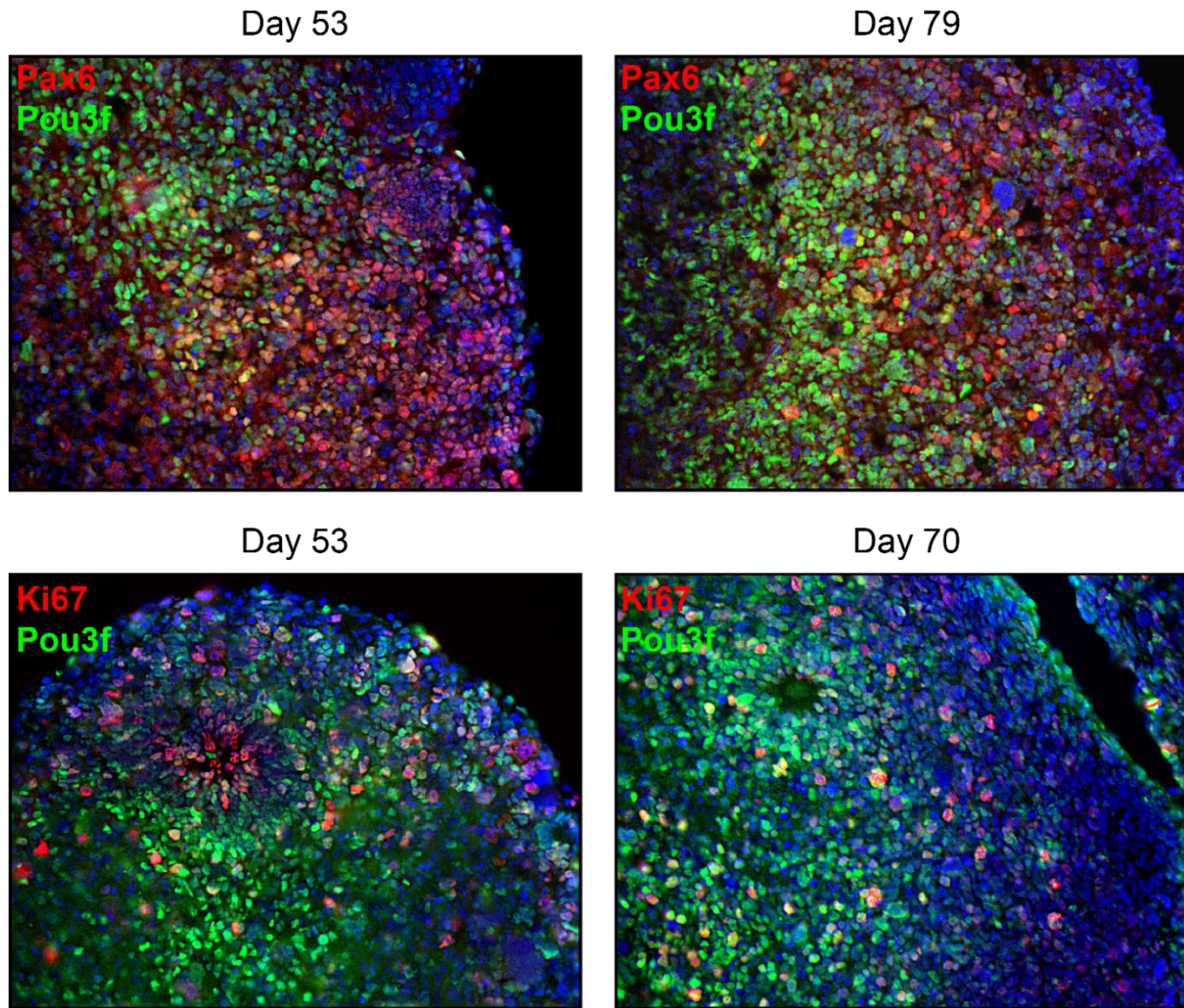


Figure 4.S6. Antibody staining of cerebral organoids. Human iPSCs were differentiated into cerebral organoids and grown in culture for 53 days (left panels) or 70-79 days (right panels) prior to harvest for immunohistochemistry. Cryosections were labeled with anti-Pou3f antibody (green, all panels) and anti-Pax6 antibody (red, top panels) or anti-Ki67 antibody (red, bottom panels). The anti-Pou3f antibody recognizes both Pou3f2 and Pou3f3 (see Methods). Ki67 is a marker of proliferation (Scholzen and Gerdes 2000). Blue, DAPI counterstain.

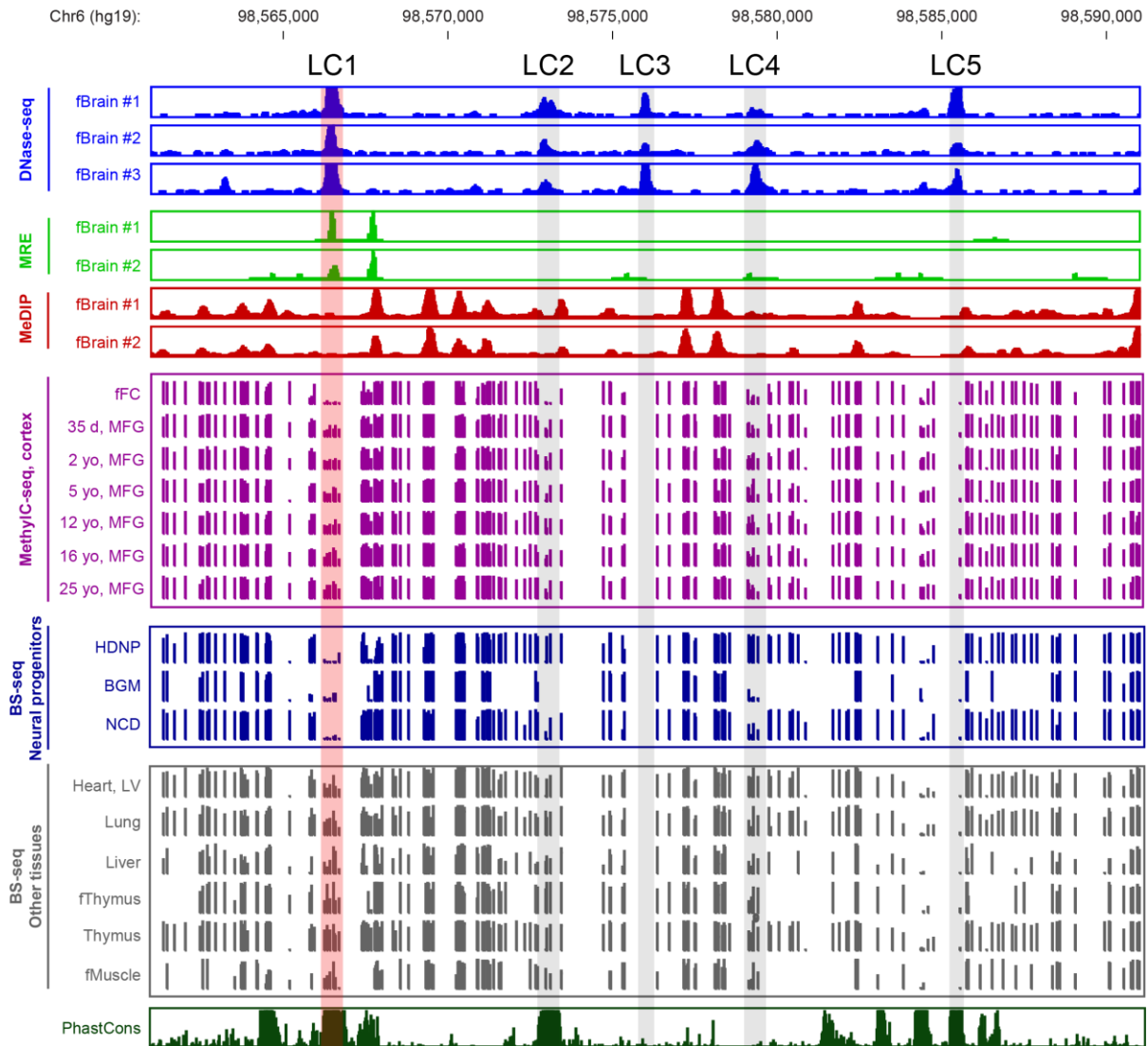


Figure 4.S7. The methylation landscape of LC1 to LC5 in human primary tissues and cultured cells. LC1 is highlighted (pink), and LC2 to LC5 are also shown (gray). LC1 is unmethylated in fetal brain and neural progenitors, and methylated in adult brain and other tissues. MRE enriches for unmethylated regions, while MeDIP enriches for methylated regions. The height of MethylC-seq and bisulfite (BS)-seq signals reflects the degree of methylation at single CpG resolution. The lack of data for LC3 and LC5 is due to the paucity of CpG sites. Data are from (Roadmap Epigenomics et al. 2015) except for MethylC-seq (Lister et al. 2013; Schultz et al. 2015), with the following GEO accessions: MRE (GSM669604 and GSM707015) and the corresponding MeDIP (GSM669614 and GSM707019) of fetal brain (fBrain), MethylC-seq (GSE47966) of fetal frontal cortex (fFC) and middle frontal gyrus (MFG, part of the cortex) at the indicated ages (d = day, yo = years old), H1-derived neuronal progenitor cells (HDNP) (GSM675546), brain germinal matrix (BGM) (GSM941747), NCD neurosphere culture (cortex-derived) (GSM1127118), left ventricle (LV) of heart (GSM1010978), lung (GSM983647), liver (GSM916049), fetal thymus (GSM1172595), thymus (GSM1010979), and fetal muscle (fMuscle) from leg (GSM1172596).

Table 4.1. Measures of LD among lead SNPs in GWAS studies of educational attainment, cognitive ability, and BPD.

r² (EUR, Haploreg v4.1)	rs9320913	rs1906252	rs10457441	rs12202969	rs1487441
rs9320913 - educational attainment	1				
rs1906252 - cognitive ability	0.96	1			
rs10457441 - cognitive ability	0.91	0.88	1		
rs12202969 - BPD	0.99	0.96	0.92	1	
rs1487441 - BPD	0.97	0.98	0.9	0.98	1

D' (EUR, Haploreg v4.1)	rs9320913	rs1906252	rs10457441	rs12202969	rs1487441
rs9320913 - educational attainment	1				
rs1906252 - cognitive ability	0.98	1			
rs10457441 - cognitive ability	1	0.98	1		
rs12202969 - BPD	1	0.98	1	1	
rs1487441 - BPD	0.99	0.99	0.99	0.99	1

The pairwise r^2 and D' values among the following five lead GWAS are shown: rs9320913 for educational attainment (Rietveld et al. 2013; Okbay et al. 2016), rs1906252 for cognitive ability (Trampush et al. 2015), rs10457441 for cognitive ability (Davies et al. 2015), rs12202969 for BPD (Muhleisen et al. 2014), and rs1487441 for BPD (Hou et al. 2016). Values were retrieved from HaploReg V4.1 (Ward and Kellis 2012a) for European populations based on 1000 Genomes Phase 1 (Genomes Project et al. 2012).

Table 4.2. Oligonucleotides used in this study.

Genotyping primers	Purpose	Product size	Polymerase	# PCR cycles	Annealing (°C)	Comment
LacZ_F GTTGCAGTGCACGGCAGATACAC TTGCTGA	For LacZ genotyping	389 bp	GoTaq Flexi (Promega)	30	68	
LacZ_R GCCACTGGTGTGGCCATAATTC AAITTCG	For LacZ genotyping	389 bp	GoTaq Flexi (Promega)	30	68	
Barcode_seq_R GGGCTTCATGATGTCCCGATAA	For Sanger sequencing of barcodes (CRE-seq nano)					
LC1_geno_F TAGGAGCAACACAACTTCAT	For LC1 KO and KI genotyping	WT: 1.6 kb, LC1 KO: 0.6 kb, KI: 1.6 kb	Phusion (New England Biolabs)	30	64	KI: Sanger sequence with LC1_central_right or LC1_central_left
LC1_geno_R ACTGAAACCAGTCAGCAGTGA	For LC1 KO and KI genotyping	WT: 1.6 kb, LC1 KO: 0.6 kb, KI: 1.6 kb	Phusion (New England Biolabs)	30	64	KI: Sanger sequence with LC1_central_right or LC1_central_left
LC1_central_right TTCGAGAGGGGTGCTCAACC	For Sanger sequencing central region of LC1_geno_F/R product					
LC1_central_left GTTACATTTAGCCGTGCAAGC	For Sanger sequencing central region of LC1_geno_F/R product					
Pou3f2_3UTR_F GCAGCAGTGGTTCAACTTTGT	For genotyping of 3'UTR variant	WT: 391 bp, 3' UTR variant: 387 bp	GoTaq Flexi (Promega)	30	60	Sanger sequence with Pou3f2_3UTR_F or Pou3f2_3UTR_R
Pou3f2_3UTR_R TTGTGTGGGAGTAAAGCCAT	For genotyping of 3'UTR variant	WT: 391 bp, 3' UTR variant: 387 bp	GoTaq Flexi (Promega)	30	60	Sanger sequence with Pou3f2_3UTR_F or Pou3f2_3UTR_R

CRE-seq Nano and amplicon-seq	(index sequences are highlighted)	Comment	Product size
Nano initial PCR_F	CAAGTAAAGCCGCCACGTC	Amplifies the barcode region	193 bp
Nano initial_PCR_R	TAGCCAGAGTCAGATGCTCA	Amplifies the barcode region	193 bp
MiSeq adapter plus strand	5' P-GATCGGAGAGCACAGGTCT 3'	Anneal plus and minus strands to form MiSeq adapter	
MiSeq adapter minus strand	5' ACACCTTTCCCTACAGACGGTCTTCCGATCT'	Anneal plus and minus strands to form MiSeq adapter	
Illumina PCR primer_1.0	BAATGATACGGGACCCAGCAGATCTACACTCTTTCCCTA CACGACGGCTCTCCGATCT	Use with indexed primer	
MiSeq indexed primer_TCAAAAGTCT	CAAGCAGAGACGGCATACGAGAT TCAAAAGTCT GTGACT GGAGTTCAGACGCTGTGCTCTTCCGA	Sample index: TCAAAAGTCT	
MiSeq indexed primer_CCTCTGGTC	CAAGCAGAGACGGCATACGAGAT TCTCTGGTC GTGACT GGAGTTCAGACGCTGTGCTCTTCCGA	Sample index: CCTCTGGTC	
MiSeq indexed primer_CCATACTAT	CAAGCAGAGACGGCATACGAGAT CCATACTAT GTGACT GGAGTTCAGACGCTGTGCTCTTCCGA	Sample index: CCATACTAT	
MiSeq indexed primer_TAACAAATTG	CAAGCAGAGACGGCATACGAGAT TAACAAATTG GTGACT GGAGTTCAGACGCTGTGCTCTTCCGA	Sample index: TAACAAATTG	
MiSeq indexed primer_GGTCTTCTC	CAAGCAGAGACGGCATACGAGAT GGTCTTCTC GTGACT GGAGTTCAGACGCTGTGCTCTTCCGA	Sample index: GGTCTTCTC	
MiSeq indexed primer_CGATCGCCG	CAAGCAGAGACGGCATACGAGAT CGATCGCCG GTGACT GGAGTTCAGACGCTGTGCTCTTCCGA	Sample index: CGATCGCCG	
CRISPR-Cas guides (NGG' highlighted)		Coordinates (mm9)	Comment
LC1 flanking left guide	TAGTATAACTTTAGTTCC GGGG	chr4:23438837-23438839	For LC1 KO (~1 kb)
LC1 flanking right guide	AGTTTTAGBBAATGGGAC AGG	chr4:23439883-23439905	For LC1 KO (~1 kb)
KI donor oligo	ACATCTGCCCGGCTCCCGCTAAATTAATCACTCT GGTATTAGGCCAATTCAG CG ACCAAAATTAATGGGAAT BAATAGTGTATCTTTGTTCCAGCTCAGCTTTGCACACCA GACACTCTTGGG	chr4:23439271-23439400	Single-stranded oligo with rs77910749, used with LC1 flanking right guide
Pou3f2_3'UTR_guide	ACTATACCTTCGGTATATATA TGG	chr4:22412585-22412607	For Pou3f2 3'UTR variant
LC1 central guide	TTCAAGCCAGACAAATTAAT GGGG	chr4:23439324-23439346	
LC5 central guide	TCTGGTAGTAAAGAGGGCC CCAGG	chr4:23417496-23417518	
LC1 bisulfite analysis		Coordinates (mm9); prior to bisulfite conversion)	Comment
LC1 bis F	AGGAGTTAAAAATTAATAATGTTTATGTAGT	chr4:23439085-23439114	35 cycles of PCR with GoTaq Flexi (Promega) with 56 °C annealing temperature, 452 bp product
LC1 bis R	CTTTCTTTCCAAAAAATAATCTCTAAC	chr4:23439511-23439536	35 cycles of PCR with GoTaq Flexi (Promega) with 56 °C annealing temperature, 452 bp product

Table 4.3. Allele-specific fetal brain DNase-seq analysis.

GEO accession or URL	Donor ID	Sample name	Gestational day	Sex	Bowtie 2 alignment rate	Inferred het for rs77910749?	rs77910749 (in LC1)	rs13208578 (in LC2)	rs12204181 (in LC4)	rs17814604 (in LC5)
https://www.encodeproject.org/experiments/ENCSCR595CSH/	H-25366	ENCBS539WGT		56 ?	N/A (used aligned bam)	Yes	3T, 9del	8C, 7T	6T, 3C	27A
https://www.encodeproject.org/experiments/ENCSCR595CSH/	H-25606	ENCLB908OXL (library)		58 M	71.52%	Yes	26T, 39del, 1A, 4C, 7G	20C, 25T	3T, 3C	36A, 29G
https://www.encodeproject.org/experiments/ENCSCR475VQD/	H-25636	ENCBS489VFT		72 M	N/A (used aligned bam)		22T	13C	17T	37A
https://www.encodeproject.org/experiments/ENCSCR475VQD/	H-25574	ENCBS980LUR		76 M	N/A (used aligned bam)					
GSM595923	H-23266	DS14718		85 F	97.22%		1del	2T	1C	1A
GSM595922	H-23266	DS14717		85 F	97.62%		2T	1C	1T	2A
GSM595928	H-23284	DS14815		96 F	93.52%	Yes	17T	2C, 3T	1T	20A
GSM595926	H-23284	DS14803		96 F	96.30%	Yes	29T	5C, 1T	1T	16A
GSM878650	H-24279	DS20221		101 M	67.90%		5T	3C	1T	12A
GSM878651	H-24297	DS20226		104 M	68.20%	Yes	16T, 4del, 4C, 1G	4C, 1T	1T	4A, 2G
GSM1027328	H-24510	DS20780		105 M	78.91%		27T	2C	NA	6A
GSM878652	H-24381	DS20231		109 F	59.48%		6T	NA	3T	3A
GSM666804	H-23399	DS15453		112 ?	98.09%	Yes	2T, 2del	1C	NA	1A
GSM595920	H-22911	DS14464		117 F	96.12%	Yes	11T	2C, 2T	2T, 1C	4A
GSM530651	H-22510	DS11872		122 M	79.77%		16T	1C	4T	8A
GSM595913	H-22510	DS11877		122 M	94.60%		12T	3C	1T	12A
GSM666819	H-23548	DS16302		142 F	98.32%	(possibly homozygous)	9del	9T	5C	NA

Data accession codes, donor information, alignment rates, and read counts at positions overlapping SNPs are provided (rs77910749 is highlighted in yellow) (see Methods). Red font indicates non-reference alleles. 'NA' indicates no reads at the position of the SNP. Individuals inferred to be rs77910749 heterozygotes (blue column) were included in the allele-specific analysis of LC1 (Figure 4.8). Note that donors H-23266, H-23284, and H-22510 were each associated with two GEO accessions. For these samples, read counts from the same donor were summed. Donors H-25606 and H-24297 had several reads that had an 'A', 'C,' or 'G' base at rs77910749; these reads were excluded from the analysis.

CHAPTER 5: CONCLUSIONS AND FUTURE DIRECTIONS

“Science is made by scientists, whose creations deeply affect each others’ progress.”

-Eric H. Davidson (Davidson 2006)

“As you know, in most areas of science, there are long periods of beginning before we really make progress.”

-Eric Kandel (Kandel 2012)

In this dissertation, I described: an approach for mapping the effects of *cis*-regulatory variants onto changes in gene expression, which yielded insights into gene regulatory effects (Chapter 2), a method for functionally dissecting large numbers of CREs, which holds promise for studying *cis*-regulatory regulation in a broad repertoire of cell types (Chapter 3), and a mechanistic study of a disease-associated *cis*-regulatory variant, which serves as a blueprint for future studies assessing the causality of non-coding variants (Chapter 4). Below, I discuss recent advances and prospects related to this work.

5.1 The utility of hybrid animals for studying *cis*-regulation and imprinting

In Chapter 2, I described a study in which we mapped the effects of *cis*-regulatory variants onto changes in gene expression genome-wide in the retina by taking advantage of hybrid animals, which serve as unique genetic tools (Shen et al. 2014). Although prior studies also utilized hybrids to identify *cis*- and *trans*-regulatory effects, ours was unique in intersecting knowledge of the locations of CREs, specific sequence variants, and changes in gene expression. Our study is analogous to eQTL studies in human tissues, which probe for statistical associations between variants and changes in gene expression, but with the advantages of complete genetic information and extremely high nucleotide diversity between Cast/EiJ and C57BL/6J alleles. With MPRA, it will be feasible to comprehensively assay the effects of the specific *cis*-regulatory variants identified by hybrid and eQTL studies.

As a byproduct our study, we identified parent-of-origin effects (e.g., imprinted genes) in the retina for the first time. Shortly after the publication of our work, another group published a study that examined *cis*- and *trans*-regulatory effects and parent-of-origin effects in brain, liver, kidney, and lung, based on reciprocal crosses of Cast/EiJ, PWK/PhJ, and WSB/EiJ (Crowley et

al. 2015). Two of the novel imprinted genes identified in our study (A230006K03Rik and A330076H08Rik) were replicated in their study, substantiating the validity and robustness of our approach. Our study establishes the retina as a model system for investigating imprinting and further underscores the value of the retina in studying mechanisms of gene regulation.

5.2 The future of high-throughput *cis*-regulatory analysis

Massively parallel reporter assays (MPRAs) have become the method of choice for assaying *cis*-regulatory variants on a large scale. Our study in Chapter 3 is the first demonstration of high-throughput truncation mutation analysis, the first AAV-mediated MPRA, and the first MPRA in the mammalian brain (Shen et al. 2016). Since the publication of our study, other groups have adopted capture-and-clone and AAV MPRA strategies (Nguyen et al. 2016; Verfaillie et al. 2016). Additionally, others have used data from our study to gain insights into gene regulation (Mo et al. 2016).

In the past, the study of CREs in the brain and most other tissues was limited to laborious one-at-a-time experiments. Due to the difficulty of delivering MPRAs to tissues *in vivo*, most systematic studies of large numbers of CREs were conducted in cell lines, with uncertain relevance to mammalian tissues. Our demonstration of the feasibility of AAV-mediated MPRAs overcomes technological barriers and brings the era of functional genomics to mammalian systems *in vivo*. Virus-based strategies expand the repertoire of cells that could potentially be assayed, and systemic delivery of viruses may allow simultaneous multi-tissue *cis*-regulatory analysis (Inagaki et al. 2006; Zincarelli et al. 2008). Furthermore, since AAV can be designed to target specific cell types (Smith et al. 2000; Michelfelder and Trepel 2009), our study paves the way for refining the study of mammalian *cis*-regulation *in vivo*.

In addition to AAV, lentivirus has also been used to deliver MPRA, with the major difference being that lentiviral constructs integrate into the host genome. It has been suggested that chromosomally integrated reporter constructs recapitulate endogenous CRE activity more faithfully than episomal reporters, but thus far it is unclear whether this is the case (Inoue et al. 2017). In yeast, delivery of 29 reporter constructs as plasmids or as integrated constructs in a specific chromosomal location (i.e., controlling for the site of integration) yielded very similar results (Sharon et al. 2012). However, lentivirus integrates randomly into the host genome, with the potential for undesirable insertion site effects (Murtha et al. 2014; Inoue et al. 2017). Furthermore, lentivirus is a pathogenic retrovirus that elicits a substantial host inflammatory response, whereas AAV is non-pathogenic (Nayak and Herzog 2010). Lentivirus does offer the advantage of a carrying capacity of 8 kb (compared to 4.7 kb for AAV), which would allow the delivery of CRE-seq libraries containing longer promoter-reporter cassettes (Kumar et al. 2001). Thus, AAV and lentivirus both have advantages and disadvantages, and both have utility in future MPRA depending on the experimental goal.

Ideally, one would like to assay CREs in their endogenous context, at their endogenous sites within the genome. How could this be achieved? One possible approach would be to profile enhancer RNAs (eRNAs), non-coding RNAs that are transcribed at active enhancers. While the functional role of eRNAs is still debated (Lam et al. 2014; Kim et al. 2015), there is considerable evidence that levels of eRNAs (specifically, '2D' or bidirectional non-polyadenylated eRNAs) reflect the activity of the corresponding enhancers. Thus, allele-specific eRNA profiling could be a way to detect the effects of enhancer variants *in situ*.

An even more promising approach for studying *cis*-regulatory variants *in situ* is by coupling MPRA with CRISPR-Cas. In the few years since its first implementation in

mammalian cells, CRISPR-Cas has revolutionized molecular biology as a rapid, efficient means of editing DNA (Doudna and Charpentier 2014). CRISPR-Cas has already been utilized for saturation mutagenesis of coding regions (Findlay et al. 2014). It should be possible to use CRISPR-Cas for saturation mutagenesis of CREs, or even for combinations of coding and *cis*-regulatory variants to study their interactions (e.g., epistasis) (Sackton and Hartl 2016).

5.3 Future directions for investigating the *MIR2113/POU3F2* locus

Most GWAS signals fall in non-coding regions and have modest effect sizes, rendering the identification of the underlying causal variant a challenge. The *MIR2113/POU3F2* locus is typical in these regards, but it is exceptionally interesting because it harbors variants associated with both higher cognitive performance and increased risk for bipolar disorder. Unraveling the mechanism underlying this link may not only provide insights into the etiology of BPD, but also elucidate the molecular aspects of human brain development that confer both enhanced cognitive capacities and susceptibility to mental illness.

With this goal in mind, we identified a candidate causal variant, rs77910749, which falls within a highly conserved non-coding region, LC1. Our transgenic mouse lines suggest enhancer activity of LC1 in the developing amygdala and cortex. However, we observed considerable line-to-line variability in transgene expression, likely due to insertion site effects (Wilson et al. 1990). To avoid insertion site effects, LC1 reporter constructs can be integrated into the mouse genome at a specific locus ('safe harbor') with CRISPR-Cas (Lombardo et al. 2011). It would be particularly valuable to generate a targeted transgenic reporter line with LC1 driving the expression of a fluorescent protein (e.g., GFP), which would enable FACS-based isolation of cells with LC1 activity. This may facilitate analysis of subpopulations within nuclei of the

amygdala, potentially enabling the detection of a small population of disease-relevant cells. Another way to isolate amygdala subregions would be with laser-capture microdissection (LCM), although this would not provide single-cell resolution (Zirlinger and Anderson 2003).

In our study, we primarily used mice as the model system. Besides the anatomical differences between mouse and human brains, there are also known species-specific and region-specific differences in gene dosage requirements. For example, humans with mutations in *Dcx* have abnormal neocortical and hippocampal development, whereas mice with mutant *Dcx* have essentially normal neocortices but abnormal hippocampi (Corbo et al. 2002). Furthermore, rs77910749 is a human-specific variant, and we observed differences in LC1 enhancer activity in the developing mouse brain compared to human iPSC-derived cerebral organoids. Many of the experiments that were implemented in mouse (e.g., allele-specific expression analysis, methylation analysis, and allele-specific Pax6 binding) can also be conducted in iPSC-derived neurons or cerebral organoids. In particular, CRISPR-Cas can be used to knock-in rs77910749 (or to revert rs77910749 to the reference allele) in otherwise isogenic cell lines. Furthermore, since *POU3F2* promotes the conversion of differentiated cells into neurons, it would be interesting to measure the efficiency of cellular reprogramming (Vierbuchen et al. 2010; Wapinski et al. 2013).

In order to investigate the effects of *cis*-regulatory variants on organismal phenotype, however, animal models are needed. If LC1 is such a highly conserved region, why does deletion of LC1 in mice produce such modest effects? Previously, other groups deleted several ‘ultra-conserved’ genomic regions in mouse, and no organismal phenotype was found despite extensive assays (Ahituv et al. 2007). To elicit the relevant phenotypes, it may be necessary to stimulate the animals with environmental stressors and/or pharmacological treatments. Additionally,

neuroimaging of mutant animals may provide more sensitive measures for detecting abnormal brain structure and function (Nieman et al. 2007). Ultimately, the study of variants of small effect will likely require highly sensitive assays and large sample sizes.

5.4 *Cis*-regulatory biology in the era of clinical whole-genome sequencing

Routine whole-genome sequencing (WGS) of patients will soon become a reality. Thus far, most instances of clinical WGS have focused on the exome because of the difficulty of interpreting non-coding regions. Hence, the clinical potential of WGS has not been fully realized. Many issues related to medical ethics and healthcare policies remain to be addressed (van El et al. 2013; Howard et al. 2015), but scientifically, one of the biggest bottlenecks is deciphering the functional consequences of the thousands of variants in non-coding regions, which represent 98% of the genome. This need is particularly pressing for neurological and neurodevelopmental disorders, which represent a large fraction of rare diseases with unknown causes (Gahl et al. 2012). Given the complexities of assigning causality to *cis*-regulatory variants in the CNS, this challenge will likely persist for many years.

In the near future, physicians, scientists, and the public alike will grapple with the many uncertainties that accompany modifiable genetic risk. The definition of ‘disease’ will also have to be revisited, as evidence accumulates that disease-associated traits fall along a continuum and manifest in a context-dependent manner (e.g., (Constantino 2011)). The beauty of *cis*-regulatory biology is that it encapsulates many of the dualities pervasive across biology: nature vs. nurture, stochasticity vs. determinism, and flexibility vs. tight control. Likewise, the study of *cis*-regulatory variants should teach us to balance necessary caution with acceptable risk, scientific

curiosity with clinical need, and subjectivity with objectivity. Eventually, though, science for the sake of science is what will drive the next breakthrough in *cis*-regulatory biology.

REFERENCES

- RetNet, <http://www.sph.uth.tmc.edu/RetNet/>.
- Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol* **5**: e234.
- Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, Berns A, Wessels LF, van Lohuizen M, van Steensel B. 2013. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154**: 914-927.
- Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16**: 197-212.
- Allocca M, Mussolino C, Garcia-Hoyos M, Sanges D, Iodice C, Petrillo M, Vandenberghe LH, Wilson JM, Marigo V, Surace EM et al. 2007. Novel adeno-associated virus serotypes efficiently transduce murine photoreceptors. *J Virol* **81**: 11372-11380.
- Altrock WD, tom Dieck S, Sokolov M, Meyer AC, Sigler A, Brakebusch C, Fassler R, Richter K, Boeckers TM, Potschka H et al. 2003. Functional inactivation of a fraction of excitatory synapses in mice deficient for the active zone protein bassoon. *Neuron* **37**: 787-800.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.
- Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome biology* **11**: R106.
- Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455-461.
- Andre E, Gawlas K, Becker-Andre M. 1998. A novel isoform of the orphan nuclear receptor RORbeta is specifically expressed in pineal gland and retina. *Gene* **216**: 277-283.
- Andrews S. 2010. FastQC. p. A quality control tool for high throughput sequence data.
- Andzelm MM, Cherry TJ, Harmin DA, Boeke AC, Lee C, Hemberg M, Pawlyk B, Malik AN, Flavell SW, Sandberg MA et al. 2015. MEF2D drives photoreceptor development through a genome-wide competition for tissue-specific enhancers. *Neuron* **86**: 247-263.
- Ardlie KG, Kruglyak L, Seielstad M. 2002. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3**: 299-309.
- Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074-1077.
- Arvey A, Agius P, Noble WS, Leslie C. 2012. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res* **22**: 1723-1734.
- Aschauer DF, Kreuz S, Rumpel S. 2013. Analysis of transduction efficiency, tropism and axonal transport of AAV serotypes 1, 2, 5, 6, 8 and 9 in the mouse brain. *PLoS One* **8**: e76310.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25-29.
- Auer PL, Lettre G. 2015. Rare variant association studies: considerations, challenges and opportunities. *Genome Med* **7**: 16.
- Aurnhammer C, Haase M, Muether N, Hausl M, Rauschhuber C, Huber I, Nitschko H, Busch U, Sing A, Ehrhardt A et al. 2012. Universal real-time PCR for the detection and quantification of adeno-associated virus serotype 2-derived inverted terminal repeat sequences. *Hum Gene Ther Methods* **23**: 18-28.

- Bae BI, Jayaraman D, Walsh CA. 2015. Genetic Changes Shaping the Human Brain. *Dev Cell* **32**: 423-434.
- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-208.
- Baker M. 2011. Microarrays, megasynthesis. *Nat Meth* **8**: 457-460.
- Bauer DE, Kamran SC, Lessard S, Xu J, Fujiwara Y, Lin C, Shao Z, Canver MC, Smith EC, Pinello L et al. 2013. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**: 253-257.
- Belinson H, Nakatani J, Babineau BA, Birnbaum RY, Ellegood J, Bershteyn M, McEvelly RJ, Long JM, Willert K, Klein OD et al. 2016. Prenatal beta-catenin/Brn2/Tbr2 transcriptional cascade regulates adult social and stereotypic behaviors. *Mol Psychiatry* doi:10.1038/mp.2015.207.
- Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**: 1045-1048.
- Black JR, Clark SJ. 2016. Age-related macular degeneration: genome-wide association studies to translation. *Genet Med* **18**: 283-289.
- Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE, Mouse Genome Database G. 2014. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res* **42**: D810-817.
- Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**: 708-715.
- Blatti C, Kazemian M, Wolfe S, Brodsky M, Sinha S. 2015. Integrating motif, DNA accessibility and gene expression data to build regulatory maps in an organism. *Nucleic Acids Res* **43**: 3998-4012.
- Bock C. 2012. Analysing and interpreting DNA methylation data. *Nat Rev Genet* **13**: 705-719.
- Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, Balasubramanian S. 2012. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**: 934-937.
- Britten RJ, Davidson EH. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol* **46**: 111-138.
- Brooks MJ, Rajasimha HK, Roger JE, Swaroop A. 2011. Next-generation sequencing facilitates quantitative analysis of wild-type and Nrl(-/-) retinal transcriptomes. *Mol Vis* **17**: 3034-3054.
- Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213-1218.
- Butler JE, Kadonaga JT. 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* **16**: 2583-2592.
- Canto-Soler MV, Huang H, Romero MS, Adler R. 2008. Transcription factors CTCF and Pax6 are segregated to different cell types during retinal cell differentiation. *Dev Dyn* **237**: 758-767.

- Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, Ami GOH, Web Presence Working G. 2009. AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**: 288-289.
- Carter-Dawson LD, LaVail MM. 1979. Rods and cones in the mouse retina. I. Structural analysis using light and electron microscopy. *J Comp Neurol* **188**: 245-262.
- Cetin A, Komai S, Eliava M, Seeburg PH, Osten P. 2006. Stereotaxic gene delivery in the rodent brain. *Nat Protoc* **1**: 3166-3173.
- Chadwick LH, Pertz LM, Broman KW, Bartolomei MS, Willard HF. 2006. Genetic control of X chromosome inactivation in mice: definition of the Xce candidate interval. *Genetics* **173**: 2103-2110.
- Chen S, Wang QL, Nie Z, Sun H, Lennon G, Copeland NG, Gilbert DJ, Jenkins NA, Zack DJ. 1997. Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron* **19**: 1017-1030.
- Chen S, Zack DJ. 1996. Ret 4, a positive acting rhodopsin regulatory element identified using a bovine retina in vitro transcription system. *J Biol Chem* **271**: 28549-28557.
- Cheng CL, Djajadi H, Molday RS. 2013. Cell-specific markers for the identification of retinal cells by immunofluorescence microscopy. *Methods Mol Biol* **935**: 185-199.
- Clancy B, Cauller LJ. 1998. Reduction of background autofluorescence in brain sections following immersion in sodium borohydride. *J Neurosci Methods* **83**: 97-102.
- Clark AJ, Bissinger P, Bullock DW, Damak S, Wallace R, Whitelaw CB, Yull F. 1994. Chromosomal position effects and the modulation of transgene expression. *Reprod Fertil Dev* **6**: 589-598.
- Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry JL, Puvion-Vandier V et al. 2015. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* **373**: 895-907.
- Cleary MA, Kilian K, Wang Y, Bradshaw J, Cavet G, Ge W, Kulkarni A, Paddison PJ, Chang K, Sheth N et al. 2004. Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. *Nat Methods* **1**: 241-248.
- Clotman F, Jacquemin P, Plumb-Rudewicz N, Pierreux CE, Van der Smissen P, Dietz HC, Courtoy PJ, Rousseau GG, Lemaigre FP. 2005. Control of liver cell fate decision by a gradient of TGF beta signaling modulated by Onecut transcription factors. *Genes Dev* **19**: 1849-1854.
- Constantino JN. 2011. The quantitative nature of autistic social impairment. *Pediatr Res* **69**: 55R-62R.
- Cooper GM, Shendure J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12**: 628-640.
- Corbo JC, Deuel TA, Long JM, LaPorte P, Tsai E, Wynshaw-Boris A, Walsh CA. 2002. Doublecortin is required in mice for lamination of the hippocampus but not the neocortex. *J Neurosci* **22**: 7548-7557.
- Corbo JC, Lawrence KA, Karlstetter M, Myers CA, Abdelaziz M, Dirkes W, Weigelt K, Seifert M, Benes V, Fritsche LG et al. 2010. CRX CHIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Res* **20**: 1512-1525.
- Corbo JC, Myers CA, Lawrence KA, Jadhav AP, Cepko CL. 2007. A typology of photoreceptor gene expression patterns in the mouse. *Proc Natl Acad Sci U S A* **104**: 12069-12074.
- Court F, Tayama C, Romanelli V, Martin-Trujillo A, Iglesias-Platas I, Okamura K, Sugahara N, Simon C, Moore H, Harness JV et al. 2014. Genome-wide parent-of-origin DNA

- methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome Res* **24**: 554-569.
- Coutinho P, Pavlou S, Bhatia S, Chalmers KJ, Kleinjan DA, van Heyningen V. 2011. Discovery and assessment of conserved Pax6 target genes and enhancers. *Genome Res* **21**: 1349-1359.
- Craddock N, Sklar P. 2013. Genetics of bipolar disorder. *Lancet* **381**: 1654-1662.
- Cross-Disorder Group of the Psychiatric Genomics C Lee SH Ripke S Neale BM Faraone SV Purcell SM Perlis RH Mowry BJ Thapar A Goddard ME et al. 2013. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**: 984-994.
- Crowley JJ, Zhabotynsky V, Sun W, Huang S, Pakatci IK, Kim Y, Wang JR, Morgan AP, Calaway JD, Aylor DL et al. 2015. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nat Genet* **47**: 353-360.
- Da Mesquita S, Ferreira AC, Sousa JC, Correia-Neves M, Sousa N, Marques F. 2016. Insights on the pathophysiology of Alzheimer's disease: The crosstalk between amyloid pathology, neuroinflammation and the peripheral immune system. *Neurosci Biobehav Rev* **68**: 547-562.
- da Rocha ST, Edwards CA, Ito M, Ogata T, Ferguson-Smith AC. 2008. Genomic imprinting at the mammalian Dlk1-Dio3 domain. *Trends Genet* **24**: 306-316.
- Dalkara D, Byrne LC, Lee T, Hoffmann NV, Schaffer DV, Flannery JG. 2012. Enhanced gene delivery to the neonatal retina through systemic administration of tyrosine-mutated AAV9. *Gene Ther* **19**: 176-181.
- Dalke C, Graw J. 2005. Mouse mutants as models for congenital retinal disorders. *Exp Eye Res* **81**: 503-512.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158.
- Davidson EH. 2001. *Genomic regulatory systems : development and evolution*. Academic Press, San Diego.
- Davidson EH. 2006. *The regulatory genome : gene regulatory networks in development and evolution*. Elsevier/Academic Press, Amsterdam ; Boston.
- Davies G Armstrong N Bis JC Bressler J Chouraki V Giddaluru S Hofer E Ibrahim-Verbaas CA Kirin M Lahti J et al. 2015. Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53949). *Mol Psychiatry* **20**: 183-192.
- Day TP, Byrne LC, Schaffer DV, Flannery JG. 2014. Advances in AAV vector development for gene therapy in the retina. *Adv Exp Med Biol* **801**: 687-693.
- Daya S, Berns KI. 2008. Gene therapy using adeno-associated virus vectors. *Clin Microbiol Rev* **21**: 583-593.
- De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, de Jong P et al. 2006. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **312**: 1215-1217.
- de Melo J, Blackshaw S. 2011. In vivo electroporation of developing mouse retina. *J Vis Exp* doi:10.3791/2847.

- Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**: 3207-3212.
- Dejager L, Libert C, Montagutelli X. 2009. Thirty years of *Mus spretus*: a promising future. *Trends Genet* **25**: 234-241.
- Denham M, Dottori M. 2011. Neural differentiation of induced pluripotent stem cells. *Methods Mol Biol* **793**: 99-110.
- Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* **19**: 1114-1121.
- Dickel DE, Zhu Y, Nord AS, Wylie JN, Akiyama JA, Afzal V, Plajzer-Frick I, Kirkpatrick A, Gottgens B, Bruneau BG et al. 2014. Function-based identification of mammalian enhancers using site-specific integration. *Nat Methods* **11**: 566-571.
- Dipple KM, McCabe ER. 2000. Phenotypes of patients with "simple" Mendelian disorders are complex traits: thresholds, modifiers, and systems dynamics. *Am J Hum Genet* **66**: 1729-1735.
- Dixon JR, Gorkin DU, Ren B. 2016. Chromatin Domains: The Unit of Chromosome Organization. *Mol Cell* **62**: 668-680.
- Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W et al. 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**: 331-336.
- Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376-380.
- Dobrev G, Chahrouh M, Dautzenberg M, Chirivella L, Kanzler B, Farinas I, Karsenty G, Grosschedl R. 2006. SATB2 is a multifunctional determinant of craniofacial patterning and osteoblast differentiation. *Cell* **125**: 971-986.
- Docker D, Schubach M, Menzel M, Munz M, Spaich C, Biskup S, Bartholdi D. 2014. Further delineation of the SATB2 phenotype. *Eur J Hum Genet* **22**: 1034-1039.
- Dominguez MH, Ayoub AE, Rakic P. 2013. POU-III transcription factors (Brn1, Brn2, and Oct6) influence neurogenesis, molecular identity, and migratory destination of upper-layer cells of the cerebral cortex. *Cereb Cortex* **23**: 2632-2643.
- Doudna JA, Charpentier E. 2014. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**: 1258096.
- Dougherty JD, Maloney SE, Wozniak DF, Rieger MA, Sonnenblick L, Coppola G, Mahieu NG, Zhang J, Cai J, Patti GJ et al. 2013. The disruption of *Celf6*, a gene identified by translational profiling of serotonergic neurons, results in autism-related behaviors. *J Neurosci* **33**: 2732-2753.
- Dror I, Golan T, Levy C, Rohs R, Mandel-Gutfreund Y. 2015. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res* **25**: 1268-1280.
- Duttke SH, Lacadie SA, Ibrahim MM, Glass CK, Corcoran DL, Benner C, Heinz S, Kadonaga JT, Ohler U. 2015. Human promoters are intrinsically directional. *Mol Cell* **57**: 674-684.
- Edmondson DG, Lyons GE, Martin JF, Olson EN. 1994. *Mef2* gene expression marks the cardiac and skeletal muscle lineages during mouse embryogenesis. *Development* **120**: 1251-1263.

- Edwards SL, Beesley J, French JD, Dunning AM. 2013. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet* **93**: 779-797.
- Emerson JJ, Hsieh LC, Sung HM, Wang TY, Huang CJ, Lu HH, Lu MY, Wu SH, Li WH. 2010. Natural selection on cis and trans regulation in yeasts. *Genome Res* **20**: 826-836.
- Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I. 2002. Identification of a variant associated with adult-type hypolactasia. *Nat Genet* **30**: 233-237.
- Epstein JA, Glaser T, Cai JX, Jepeal L, Walton DS, Maas RL. 1994. 2 Independent and Interactive DNA-Binding Subdomains of the Pax6 Paired Domain Are Regulated by Alternative Splicing. *Genes & Development* **8**: 2022-2034.
- Erceg J, Saunders TE, Girardot C, Devos DP, Hufnagel L, Furlong EE. 2014. Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. *PLoS Genet* **10**: e1004060.
- Evangelou E, Ioannidis JP. 2013. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* **14**: 379-389.
- Falls JG, Pulford DJ, Wylie AA, Jirtle RL. 1999. Genomic imprinting: implications for human disease. *Am J Pathol* **154**: 635-647.
- Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I et al. 2012. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res* **22**: 2399-2408.
- Ficz G, Gribben JG. 2014. Loss of 5-hydroxymethylcytosine in cancer: cause or consequence? *Genomics* **104**: 352-357.
- Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. 2014. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513**: 120-123.
- Fogarty MP, Cannon ME, Vadlamudi S, Gaulton KJ, Mohlke KL. 2014. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. *PLoS Genet* **10**: e1004633.
- Fossat N, Le Greneur C, Beby F, Vincent S, Godement P, Chatelain G, Lamonerie T. 2007. A new GFP-tagged line reveals unexpected Otx2 protein localization in retinal photoreceptors. *BMC Dev Biol* **7**: 122.
- Freund CL, Gregory-Evans CY, Furukawa T, Papaioannou M, Looser J, Ploder L, Bellingham J, Ng D, Herbrick JA, Duncan A et al. 1997. Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. *Cell* **91**: 543-553.
- Fuchs S, Nakazawa M, Maw M, Tamai M, Oguchi Y, Gal A. 1995. A homozygous 1-base pair deletion in the arrestin gene is a frequent cause of Oguchi disease in Japanese. *Nat Genet* **10**: 360-362.
- Furukawa T, Morrow EM, Cepko CL. 1997. Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. *Cell* **91**: 531-541.
- Furukawa T, Morrow EM, Li T, Davis FC, Cepko CL. 1999. Retinopathy and attenuated circadian entrainment in Crx-deficient mice. *Nat Genet* **23**: 466-470.
- Gaffney DJ. 2013. Global properties and functional complexity of human gene regulatory variation. *PLoS Genet* **9**: e1003501.
- Gahl WA, Markello TC, Toro C, Fajardo KF, Sincan M, Gill F, Carlson-Donohoe H, Gropman A, Pierson TM, Golas G et al. 2012. The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genet Med* **14**: 51-59.

- Ganna A, Genovese G, Howrigan DP, Byrnes A, Kurki M, Zekavat SM, Whelan CW, Kals M, Nivard MG, Bloemendal A et al. 2016. Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat Neurosci* **19**: 1563-1565.
- Garcia-Moreno F, Pedraza M, Di Giovannantonio LG, Di Salvio M, Lopez-Mascaraque L, Simeone A, De Carlos JA. 2010. A neuronal migratory pathway crossing from diencephalon to telencephalon populates amygdala nuclei. *Nat Neurosci* **13**: 680-689.
- Garfield AS, Cowley M, Smith FM, Moorwood K, Stewart-Cox JE, Gilroy K, Baker S, Xia J, Dalley JW, Hurst LD et al. 2011. Distinct physiological and behavioural functions for parental alleles of imprinted Grb10. *Nature* **469**: 534-538.
- Gaulton KJ, Ferreira T, Lee Y, Raimondo A, Magi R, Reschen ME, Mahajan A, Locke A, Rayner NW, Robertson N et al. 2015. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat Genet* **47**: 1415-1425.
- Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56-65.
- Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.
- Ghirlando R, Felsenfeld G. 2016. CTCF: making the right connections. *Genes Dev* **30**: 881-891.
- Gibson G. 2011. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**: 135-145.
- Gisselbrecht SS, Barrera LA, Porsch M, Aboukhalil A, Estep PW, 3rd, Vedenko A, Palagi A, Kim Y, Zhu X, Busser BW et al. 2013. Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. *Nat Methods* **10**: 774-780.
- Gleeson JG, Lin PT, Flanagan LA, Walsh CA. 1999. Doublecortin is a microtubule-associated protein and is expressed widely by migrating neurons. *Neuron* **23**: 257-271.
- Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**: 182-189.
- Goncalves A, Leigh-Brown S, Thybert D, Stefflova K, Turro E, Flicek P, Brazma A, Odom DT, Marioni JC. 2012. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res* **22**: 2376-2384.
- Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017-1018.
- Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* **5**: 578-590.
- Greenbaum D, Colangelo C, Williams K, Gerstein M. 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* **4**: 117.
- Grieger JC, Choi VW, Samulski RJ. 2006. Production and characterization of adeno-associated viral vectors. *Nat Protoc* **1**: 1412-1428.
- Gross DS, Garrard WT. 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57**: 159-197.
- Grubb SC, Bult CJ, Bogue MA. 2014. Mouse phenome database. *Nucleic Acids Res* **42**: D825-834.
- Guenther CA, Tasic B, Luo L, Bedell MA, Kingsley DM. 2014. A molecular basis for classic blond hair color in Europeans. *Nat Genet* **46**: 748-752.

- Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al. 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**: 900-910.
- Haider NB, Zhang W, Hurd R, Ikeda A, Nystuen AM, Naggert JK, Nishina PM. 2008. Mapping of genetic modifiers of Nr2e3 rd7/rd7 that suppress retinal degeneration and restore blue cone cells to normal quantity. *Mamm Genome* **19**: 145-154.
- Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**: D514-517.
- Hao H, Kim DS, Klocke B, Johnson KR, Cui K, Gotoh N, Zang C, Gregorski J, Gieser L, Peng W et al. 2012. Transcriptional regulation of rod photoreceptor homeostasis revealed by in vivo NRL targetome analysis. *PLoS Genet* **8**: e1002649.
- Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG et al. 2011. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature* **470**: 264-268.
- Harrison PJ. 2016. Molecular neurobiological clues to the pathogenesis of bipolar disorder. *Curr Opin Neurobiol* **36**: 1-6.
- Hayden EC. 2014. Technology: The \$1,000 genome. *Nature* **507**: 294-295.
- Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**: 576-589.
- Heinz S, Romanoski CE, Benner C, Glass CK. 2015. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol* **16**: 144-154.
- Herculano-Houzel S. 2012. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proc Natl Acad Sci U S A* **109 Suppl 1**: 10661-10668.
- Herweijer H, Wolff JA. 2007. Gene therapy progress and prospects: hydrodynamic gene delivery. *Gene Ther* **14**: 99-107.
- Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA, Young RA. 2013. Super-enhancers in the control of cell identity and disease. *Cell* **155**: 934-947.
- Hogan B, Beddington R, Costantini F, Lacy E. 1994. Manipulating the mouse embryo: a laboratory manual. *Plainview (NY): Cold Spring Harbor Laboratory Press Google Scholar*.
- Hong JW, Hendrix DA, Levine MS. 2008. Shadow enhancers as a source of evolutionary novelty. *Science* **321**: 1314.
- Hou L, Bergen SE, Akula N, Song J, Hultman CM, Landen M, Adli M, Alda M, Arduini R, Arias B et al. 2016. Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. *Hum Mol Genet* **25**: 3383-3394.
- Howard HC, Knoppers BM, Cornel MC, Wright Clayton E, Senecal K, Borry P, European Society of Human G, Platform PGIP, Human Genome O, the PHGF. 2015. Whole-genome sequencing in newborn screening? A statement on the continued importance of targeted approaches in newborn screening programmes. *Eur J Hum Genet* **23**: 1593-1600.
- Hsiao TH, Diaconu C, Myers CA, Lee J, Cepko CL, Corbo JC. 2007. The cis-regulatory logic of the mammalian photoreceptor transcriptional network. *PLoS One* **2**: e643.

- Huangfu D, Maehr R, Guo W, Eijkelenboom A, Snitow M, Chen AE, Melton DA. 2008. Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. *Nat Biotechnol* **26**: 795-797.
- Hughes AL, Rando OJ. 2014. Mechanisms underlying nucleosome positioning in vivo. *Annu Rev Biophys* **43**: 41-63.
- Inagaki K, Fuess S, Storm TA, Gibson GA, McTiernan CF, Kay MA, Nakai H. 2006. Robust systemic transduction with AAV9 vectors in mice: efficient global cardiac gene transfer superior to that of AAV8. *Mol Ther* **14**: 45-53.
- Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27**: 38-52.
- Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**: 318-356.
- Jelcick AS, Yuan Y, Leehy BD, Cox LC, Silveira AC, Qiu F, Schenk S, Sachs AJ, Morrison MA, Nystuen AM et al. 2011. Genetic variations strongly influence phenotypic outcome in the mouse retina. *PLoS One* **6**: e21858.
- Jeon CJ, Strettoi E, Masland RH. 1998. The major cell populations of the mouse retina. *J Neurosci* **18**: 8936-8946.
- Jirtle RL. 2012. Geneimprint.
- Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**: 327-339.
- Joly S, Pernet V, Samardzija M, Grimm C. 2011. Pax6-positive Muller glia cells express cell cycle markers but do not proliferate after photoreceptor injury in the mouse retina. *Glia* **59**: 1033-1046.
- Kandel E. 2012. Art, Mind And Brain Intersect In Kandel's Vienna. In *Science Friday*, (ed. I Flatow).
- Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E, Sasaki H. 2004. Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature* **429**: 900-903.
- Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42**: D764-770.
- Karra D, Dahm R. 2010. Transfection techniques for neuronal cells. *J Neurosci* **30**: 6171-6177.
- Kasher PR, Schertz KE, Thomas M, Jackson A, Annunziata S, Ballesta-Martinez MJ, Campeau PM, Clayton PE, Eaton JL, Granata T et al. 2016. Small 6q16.1 Deletions Encompassing POU3F2 Cause Susceptibility to Obesity and Variable Developmental Delay with Intellectual Disability. *Am J Hum Genet* **98**: 363-372.
- Kautzmann MA, Kim DS, Felder-Schmittbuhl MP, Swaroop A. 2011. Combinatorial regulation of photoreceptor differentiation factor, neural retina leucine zipper gene NRL, revealed by in vivo promoter analysis. *J Biol Chem* **286**: 28247-28255.
- Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**: 289-294.

- Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J et al. 2014. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111**: 6131-6138.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996-1006.
- Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engstrom PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K et al. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res* **17**: 545-555.
- Kim DS, Matsuda T, Cepko CL. 2008a. A core paired-type and POU homeodomain-containing transcription factor program drives retinal bipolar cell gene expression. *J Neurosci* **28**: 7748-7764.
- Kim EJ, Battiste J, Nakagawa Y, Johnson JE. 2008b. Ascl1 (Mash1) lineage cells contribute to discrete cell populations in CNS architecture. *Mol Cell Neurosci* **38**: 595-606.
- Kim TK, Hemberg M, Gray JM. 2015. Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers. *Cold Spring Harb Perspect Biol* **7**: a018622.
- Kim TK, Shiekhattar R. 2015. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* **162**: 948-959.
- Kim YJ, Cecchini KR, Kim TH. 2011. Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. *Proc Natl Acad Sci U S A* **108**: 7391-7396.
- King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107-116.
- Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**: 310-315.
- Kizilyaprak C, Spehner D, Devys D, Schultz P. 2010. In vivo chromatin organization of mouse rod photoreceptors correlates with histone modifications. *PLoS One* **5**: e11039.
- Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**: 385-389.
- Kleinjan DA, van Heyningen V. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet* **76**: 8-32.
- Koenen KC, Moffitt TE, Roberts AL, Martin LT, Kubzansky L, Harrington H, Poulton R, Caspi A. 2009. Childhood IQ and adult mental disorders: a test of the cognitive reserve hypothesis. *Am J Psychiatry* **166**: 50-57.
- Kozmik Z, Czerny T, Busslinger M. 1997. Alternatively spliced insertions in the paired domain restrict the DNA sequence specificity of Pax6 and Pax8. *EMBO J* **16**: 6793-6803.
- Kriaucionis S, Heintz N. 2009. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**: 929-930.
- Krueger F. Trim Galore!
- Kryukov GV, Schmidt S, Sunyaev S. 2005. Small fitness effect of mutations in highly conserved non-coding regions. *Hum Mol Genet* **14**: 2221-2229.
- Kulaeva OI, Nizovtseva EV, Polikanov YS, Ulianov SV, Studitsky VM. 2012. Distant activation of transcription: mechanisms of enhancer action. *Mol Cell Biol* **32**: 4892-4897.

- Kulzer JR, Stitzel ML, Morken MA, Huyghe JR, Fuchsberger C, Kuusisto J, Laakso M, Boehnke M, Collins FS, Mohlke KL. 2014. A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. *Am J Hum Genet* **94**: 186-197.
- Kumar M, Keller B, Makalou N, Sutton RE. 2001. Systematic determination of the packaging limit of lentiviral vectors. *Hum Gene Ther* **12**: 1893-1905.
- Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* **24**: 1595-1602.
- Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci U S A* **109**: 19498-19503.
- Lam MT, Li W, Rosenfeld MG, Glass CK. 2014. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem Sci* **39**: 170-182.
- Lancaster MA, Renner M, Martin CA, Wenzel D, Bicknell LS, Hurles ME, Homfray T, Penninger JM, Jackson AP, Knoblich JA. 2013. Cerebral organoids model human brain development and microcephaly. *Nature* **501**: 373-379.
- Langevin LM, Mattar P, Scardigli R, Roussigne M, Logan C, Blader P, Schuurmans C. 2007. Validating in utero electroporation for the rapid analysis of gene regulatory elements in the murine telencephalon. *Dev Dyn* **236**: 1273-1286.
- Langmann T, Di Gioia SA, Rau I, Stohr H, Maksimovic NS, Corbo JC, Renner AB, Zrenner E, Kumaramanickavel G, Karlstetter M et al. 2010. Nonsense mutations in FAM161A cause RP28-associated recessive retinitis pigmentosa. *Am J Hum Genet* **87**: 376-381.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.
- Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- Lee H, O'Connor BD, Merriman B, Funari VA, Homer N, Chen Z, Cohn DH, Nelson SF. 2009. Improving the efficiency of genomic loci capture using oligonucleotide arrays for high throughput resequencing. *BMC Genomics* **10**: 646.
- Lee J, Myers CA, Williams N, Abdelaziz M, Corbo JC. 2010. Quantitative fine-tuning of photoreceptor cis-regulatory elements through affinity modulation of transcription factor binding sites. *Gene Ther* **17**: 1390-1399.
- Leoyklang P, Suphapeetiporn K, Siriwan P, Desudchit T, Chaowanapanja P, Gahl WA, Shotelersuk V. 2007. Heterozygous nonsense mutation SATB2 associated with cleft palate, osteoporosis, and cognitive defects. *Hum Mutat* **28**: 732-738.
- Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, Yen CA, Lin S, Lin Y, Qiu Y et al. 2015. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**: 350-354.
- Levine M, Davidson EH. 2005. Gene regulatory networks for development. *Proc Natl Acad Sci U S A* **102**: 4936-4942.
- Levine M, Vicente C. 2015. An interview with Mike Levine. *Development* **142**: 3453-3455.
- Levo M, Segal E. 2014. In pursuit of design principles of regulatory sequences. *Nat Rev Genet* **15**: 453-468.
- Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J et al. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**: 84-98.

- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.
- Libbrecht MW, Noble WS. 2015. Machine learning applications in genetics and genomics. *Nat Rev Genet* **16**: 321-332.
- Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD et al. 2013. Global epigenomic reconfiguration during mammalian brain development. *Science* **341**: 1237905.
- Livesey FJ, Cepko CL. 2001. Vertebrate neural cell-fate determination: lessons from the retina. *Nat Rev Neurosci* **2**: 109-118.
- Loh PR, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ, Schizophrenia Working Group of Psychiatric Genomics C, de Candia TR, Lee SH, Wray NR et al. 2015. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet* **47**: 1385-1392.
- Lombardo A, Cesana D, Genovese P, Di Stefano B, Provasi E, Colombo DF, Neri M, Magnani Z, Cantore A, Lo Riso P et al. 2011. Site-specific integration and tailoring of cassette design for sustainable gene transfer. *Nat Methods* **8**: 861-869.
- London A, Benhar I, Schwartz M. 2013. The retina as a window to the brain-from eye research to CNS disorders. *Nat Rev Neurol* **9**: 44-53.
- Luciano M, Hansell NK, Lahti J, Davies G, Medland SE, Raikonen K, Tenesa A, Widen E, McGhee KA, Palotie A et al. 2011. Whole genome association scan for genetic polymorphisms influencing information processing speed. *Biol Psychol* **86**: 193-202.
- Lui JH, Hansen DV, Kriegstein AR. 2011. Development and evolution of the human neocortex. *Cell* **146**: 18-36.
- Maletic V, Raison C. 2014. Integrated neurobiology of bipolar disorder. *Front Psychiatry* **5**: 98.
- Man TK, Stormo GD. 2001. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res* **29**: 2471-2478.
- Manuel MN, Mi D, Mason JO, Price DJ. 2015. Regulation of cerebral cortical neurogenesis by the Pax6 transcription factor. *Front Cell Neurosci* **9**: 70.
- Mardis ER. 2011. A decade's perspective on DNA sequencing technology. *Nature* **470**: 198-203.
- Marquardt T, Ashery-Padan R, Andrejewski N, Scardigli R, Guillemot F, Gruss P. 2001. Pax6 is required for the multipotent state of retinal progenitor cells. *Cell* **105**: 43-55.
- Masland RH. 2012. The neuronal organization of the retina. *Neuron* **76**: 266-280.
- Massari ME, Murre C. 2000. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol Cell Biol* **20**: 429-440.
- Matsuda T, Cepko CL. 2004. Electroporation and RNA interference in the rodent retina in vivo and in vitro. *Proc Natl Acad Sci U S A* **101**: 16-22.
- Matsuda T, Cepko CL. 2008. Analysis of gene function in the retina. *Methods Mol Biol* **423**: 259-278.
- Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J et al. 2012a. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**: 1190-1195.
- Maurano MT, Wang H, Kuttyavin T, Stamatoyannopoulos JA. 2012b. Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genet* **8**: e1002599.

- McCarty DM. 2008. Self-complementary AAV vectors; advances and applications. *Mol Ther* **16**: 1648-1656.
- McCarty DM, Young SM, Jr., Samulski RJ. 2004. Integration of adeno-associated virus (AAV) and recombinant AAV vectors. *Annu Rev Genet* **38**: 819-845.
- McEvelly RJ, de Diaz MO, Schonemann MD, Hooshmand F, Rosenfeld MG. 2002. Transcriptional regulation of cortical neuron migration by POU domain factors. *Science* **295**: 1528-1532.
- McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**: 495-501.
- McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res* **20**: 816-825.
- McManus CJ, May GE, Spealman P, Shteyman A. 2014. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res* **24**: 422-430.
- Mears AJ, Kondo M, Swain PK, Takada Y, Bush RA, Saunders TL, Sieving PA, Swaroop A. 2001. Nrl is required for rod photoreceptor development. *Nat Genet* **29**: 447-452.
- Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Jr., Kinney JB et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**: 271-277.
- Merbs SL, Khan MA, Hackler L, Jr., Oliver VF, Wan J, Qian J, Zack DJ. 2012. Cell-specific DNA methylation patterns of retina-specific genes. *PLoS One* **7**: e32602.
- Mercer TR, Edwards SL, Clark MB, Neph SJ, Wang H, Stergachis AB, John S, Sandstrom R, Li G, Sandhu KS et al. 2013. DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. *Nat Genet* **45**: 852-859.
- Merikangas KR, Jin R, He JP, Kessler RC, Lee S, Sampson NA, Viana MC, Andrade LH, Hu C, Karam EG et al. 2011. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. *Arch Gen Psychiatry* **68**: 241-251.
- Mertes F, Elsharawy A, Sauer S, van Helvoort JM, van der Zaag PJ, Franke A, Nilsson M, Lehrach H, Brookes AJ. 2011. Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief Funct Genomics* **10**: 374-386.
- Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ et al. 2005. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* **33**: D284-288.
- Michelfelder S, Trepel M. 2009. Adeno-associated viral vectors and their redirection to cell-type specific receptors. *Adv Genet* **67**: 29-60.
- Milutinovic S, D'Alessio AC, Detich N, Szyf M. 2007. Valproate induces widespread epigenetic reprogramming which involves demethylation of specific genes. *Carcinogenesis* **28**: 560-571.
- Mingozzi F, High KA. 2011. Therapeutic in vivo gene transfer for genetic disease using AAV: progress and challenges. *Nat Rev Genet* **12**: 341-355.
- Mo A, Luo C, Davis FP, Mukamel EA, Henry GL, Nery JR, Urich MA, Picard S, Lister R, Eddy SR et al. 2016. Epigenomic landscapes of retinal rods and cones. *Elife* **5**.
- Mogno I, Kwasniewski JC, Cohen BA. 2013. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res* **23**: 1908-1915.

- Molyneaux BJ, Arlotta P, Menezes JR, Macklis JD. 2007. Neuronal subtype specification in the cerebral cortex. *Nat Rev Neurosci* **8**: 427-437.
- Mongrain V, La Spada F, Curie T, Franken P. 2011. Sleep loss reduces the DNA-binding of BMAL1, CLOCK, and NPAS2 to specific clock genes in the mouse cerebral cortex. *PLoS One* **6**: e26622.
- Montana CL, Kolesnikov AV, Shen SQ, Myers CA, Kefalov VJ, Corbo JC. 2013. Reprogramming of adult rod photoreceptors prevents retinal degeneration. *Proc Natl Acad Sci U S A* **110**: 1732-1737.
- Montana CL, Lawrence KA, Williams NL, Tran NM, Peng GH, Chen S, Corbo JC. 2011a. Transcriptional regulation of neural retina leucine zipper (Nrl), a photoreceptor cell fate determinant. *J Biol Chem* **286**: 36921-36931.
- Montana CL, Myers CA, Corbo JC. 2011b. Quantifying the activity of cis-regulatory elements in the mouse retina by explant electroporation. *J Vis Exp* doi:10.3791/2821.
- Moore T, Haig D. 1991. Genomic imprinting in mammalian development: a parental tug-of-war. *Trends Genet* **7**: 45-49.
- Morison IM, Paton CJ, Cleverley SD. 2001. The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res* **29**: 275-276.
- Mortimer I, Tam P, MacLachlan I, Graham RW, Saravolac EG, Joshi PB. 1999. Cationic lipid-mediated transfection of cells in culture requires mitotic activity. *Gene Ther* **6**: 403-411.
- Mott R, Yuan W, Kaisaki P, Gan X, Cleak J, Edwards A, Baud A, Flint J. 2014. The architecture of parent-of-origin effects in mice. *Cell* **156**: 332-342.
- Muhleisen TW, Leber M, Schulze TG, Strohmaier J, Degenhardt F, Treutlein J, Mattheisen M, Forstner AJ, Schumacher J, Breuer R et al. 2014. Genome-wide association study reveals two new risk loci for bipolar disorder. *Nat Commun* **5**: 3339.
- Mullen RJ, Buck CR, Smith AM. 1992. NeuN, a neuronal specific nuclear protein in vertebrates. *Development* **116**: 201-211.
- Muller N, Weidinger E, Leitner B, Schwarz MJ. 2015. The role of inflammation in schizophrenia. *Front Neurosci* **9**: 372.
- Murtha M, Tokcaer-Keskin Z, Tang Z, Strino F, Chen X, Wang Y, Xi X, Basilico C, Brown S, Bonneau R et al. 2014. FIREWACH: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat Methods* **11**: 559-565.
- Nakai S, Kawano H, Yudate T, Nishi M, Kuno J, Nagata A, Jishage K, Hamada H, Fujii H, Kawamura K et al. 1995. The POU domain transcription factor Brn-2 is required for the determination of specific neuronal lineages in the hypothalamus of the mouse. *Genes Dev* **9**: 3109-3121.
- Nam J, Davidson EH. 2012. Barcoded DNA-tag reporters for multiplex cis-regulatory analysis. *PLoS One* **7**: e35934.
- Nam JM, Dong P, Tarpine R, Istrail S, Davidson EH. 2010. Functional cis-regulatory genomics for systems biology. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 3930-3935.
- Nasonkin IO, Lazo K, Hambright D, Brooks M, Fariss R, Swaroop A. 2011. Distinct nuclear localization patterns of DNA methyltransferases in developing and mature mammalian retina. *J Comp Neurol* **519**: 1914-1930.
- Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. 2012. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res* **22**: 1711-1722.

- Nathans J, Davenport CM, Maumenee IH, Lewis RA, Hejtmancik JF, Litt M, Lovrien E, Weleber R, Bachynski B, Zwas F et al. 1989. Molecular genetics of human blue cone monochromacy. *Science* **245**: 831-838.
- Nativio R, Wendt KS, Ito Y, Huddleston JE, Uribe-Lewis S, Woodfine K, Krueger C, Reik W, Peters JM, Murrell A. 2009. Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. *PLoS Genet* **5**: e1000739.
- Nayak S, Herzog RW. 2010. Progress and prospects: immune responses to viral vectors. *Gene Ther* **17**: 295-304.
- Network, Pathway Analysis Subgroup of Psychiatric Genomics C. 2015. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nat Neurosci* **18**: 199-209.
- Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* **7**: 61-80.
- Nguyen TA, Jones RD, Snavelly AR, Pfenning AR, Kirchner R, Hemberg M, Gray JM. 2016. High-throughput functional comparison of promoter and enhancer activities. *Genome Res* **26**: 1023-1033.
- Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K et al. 2011. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* **7**: e1002003.
- Nichols AJ, O'Dell RS, Powrozek TA, Olson EC. 2013. Ex utero electroporation and whole hemisphere explants: a simple experimental method for studies of early cortical development. *J Vis Exp* doi:10.3791/50271.
- Nieman BJ, Lerch JP, Bock NA, Chen XJ, Sled JG, Henkelman RM. 2007. Mouse behavioral mutants have neuroimaging abnormalities. *Hum Brain Mapp* **28**: 567-575.
- Ninkovic J, Steiner-Mezzadri A, Jawerka M, Akinci U, Masserdotti G, Petricca S, Fischer J, von Holst A, Beckers J, Lie CD et al. 2013. The BAF complex interacts with Pax6 in adult neural progenitors to establish a neurogenic cross-regulatory transcriptional network. *Cell Stem Cell* **13**: 403-418.
- Nishida A, Furukawa A, Koike C, Tano Y, Aizawa S, Matsuo I, Furukawa T. 2003. Otx2 homeobox gene controls retinal photoreceptor cell fate and pineal gland development. *Nat Neurosci* **6**: 1255-1263.
- Nord AS, Blow MJ, Attanasio C, Akiyama JA, Holt A, Hosseini R, Phouanavong S, Plajzer-Frick I, Shoukry M, Afzal V et al. 2013. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**: 1521-1531.
- Nord AS, Pattabiraman K, Visel A, Rubenstein JL. 2015. Genomic Perspectives of Transcriptional Regulation in Forebrain Development. *Neuron* **85**: 27-47.
- Oh EC, Cheng H, Hao H, Jia L, Khan NW, Swaroop A. 2008. Rod differentiation factor NRL activates the expression of nuclear receptor NR2E3 to suppress the development of cone photoreceptors. *Brain Res* **1236**: 16-29.
- Okaty BW, Sugino K, Nelson SB. 2011. Cell type-specific transcriptomics in the brain. *J Neurosci* **31**: 6939-6943.
- Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA, Turley P, Chen GB, Emilsson V, Meddens SF et al. 2016. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**: 539-542.

- Oldridge DA, Wood AC, Weichert-Leahey N, Crimmins I, Sussman R, Winter C, McDaniel LD, Diamond M, Hart LS, Zhu S et al. 2015. Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. *Nature* **528**: 418-421.
- Olds LC, Sibley E. 2003. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Hum Mol Genet* **12**: 2333-2340.
- Oliver G, Mailhos A, Wehr R, Copeland NG, Jenkins NA, Gruss P. 1995. Six3, a murine homologue of the sine oculis gene, demarcates the most anterior border of the developing neural plate and is expressed during eye development. *Development* **121**: 4045-4055.
- Ostuni R, Piccolo V, Barozzi I, Polletti S, Termanini A, Bonifacio S, Curina A, Prosperini E, Ghisletti S, Natoli G. 2013. Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**: 157-171.
- Osumi N, Shinohara H, Numayama-Tsuruta K, Maekawa M. 2008. Concise review: Pax6 transcription factor contributes to both embryonic and adult neurogenesis as a multifunctional regulator. *Stem Cells* **26**: 1663-1672.
- Ozgul RK, Siemiatkowska AM, Yucel D, Myers CA, Collin RW, Zonneveld MN, Beryozkin A, Banin E, Hoyng CB, van den Born LI et al. 2011. Exome sequencing and cis-regulatory mapping identify mutations in MAK, a gene encoding a regulator of ciliary length, as a cause of retinitis pigmentosa. *Am J Hum Genet* **89**: 253-264.
- Pabba M. 2013. Evolutionary development of the amygdaloid complex. *Front Neuroanat* **7**: 27.
- Pasca AM, Sloan SA, Clarke LE, Tian Y, Makinson CD, Huber N, Kim CH, Park JY, O'Rourke NA, Nguyen KD et al. 2015. Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. *Nat Methods* **12**: 671-678.
- Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**: 265-270.
- Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**: 1173-1175.
- Penaud-Budloo M, Le Guiner C, Nowrouzi A, Toromanoff A, Cherel Y, Chenuaud P, Schmidt M, von Kalle C, Rolling F, Moullier P et al. 2008. Adeno-associated virus vector genomes persist as episomal chromatin in primate muscle. *J Virol* **82**: 7875-7885.
- Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**: 499-502.
- Perry W, Minassian A, Feifel D, Braff DL. 2001. Sensorimotor gating deficits in bipolar disorder patients with acute psychotic mania. *Biol Psychiatry* **50**: 418-424.
- Pfeifer GP, Szabo PE. 2009. 5-hydroxymethylcytosine, a modified mammalian DNA base with a potential regulatory role. *Epigenomics* **1**: 21-22.
- Phillips JE, Corces VG. 2009. CTCF: master weaver of the genome. *Cell* **137**: 1194-1211.
- Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. 2016. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet* **48**: 709-717.
- Plank JL, Dean A. 2014. Enhancer function: mechanistic and genome-wide insights come together. *Mol Cell* **55**: 5-14.
- Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G et al. 2013. Great ape genetic diversity and population history. *Nature* **499**: 471-475.

- Price AL, Spencer CC, Donnelly P. 2015. Progress and promise in understanding the genetic basis of common diseases. *Proc Biol Sci* **282**: 20151684.
- Prickett AR, Oakey RJ. 2012. A survey of tissue-specific genomic imprinting in mammals. *Mol Genet Genomics* **287**: 621-630.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.
- Rainger JK, Bhatia S, Bengani H, Gautier P, Rainger J, Pearson M, Ansari M, Crow J, Mehendale F, Palinkasova B et al. 2014. Disruption of SATB2 or its long-range cis-regulation by SOX9 causes a syndromic form of Pierre Robin sequence. *Hum Mol Genet* **23**: 2569-2579.
- Raivich G, Behrens A. 2006. Role of the AP-1 transcription factor c-Jun in developing, adult and injured brain. *Prog Neurobiol* **78**: 347-363.
- Ramón y Cajal S. 1922. *Charlas de café: Pensamientos, anécdotas y confidencias*. Imprenta de Juan Pueyo, Madrid.
- Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet* **10**: e1004525.
- Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665-1680.
- Reik W, Walter J. 2001. Genomic imprinting: parental influence on the genome. *Nat Rev Genet* **2**: 21-32.
- Reynolds N, O'Shaughnessy A, Hendrich B. 2013. Transcriptional repressors: multifaceted regulators of gene expression. *Development* **140**: 505-512.
- Rietveld CA, Esko T, Davies G, Pers TH, Turley P, Benyamin B, Chabris CF, Emilsson V, Johnson AD, Lee JJ et al. 2014. Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proc Natl Acad Sci U S A* **111**: 13790-13794.
- Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, Westra HJ, Shakhbazov K, Abdellaoui A, Agrawal A et al. 2013. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**: 1467-1471.
- Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317-330.
- Roberts MR, Srinivas M, Forrest D, Morreale de Escobar G, Reh TA. 2006. Making the gradient: thyroid hormone regulates cone opsin expression in the developing mouse retina. *Proc Natl Acad Sci U S A* **103**: 6218-6223.
- Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24-26.
- Roesch K, Jadhav AP, Trimarchi JM, Stadler MB, Roska B, Sun BB, Cepko CL. 2008. The transcriptome of retinal Muller glial cells. *J Comp Neurol* **509**: 225-238.
- Rohde C, Zhang Y, Reinhardt R, Jeltsch A. 2010. BISMAR--fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. *BMC Bioinformatics* **11**: 230.
- Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. 2015. Epigenomics: Roadmap for regulation. *Nature* **518**: 314-316.

- Rosenthal N. 1987. Identification of regulatory elements of cloned genes with functional assays. *Methods Enzymol* **152**: 704-720.
- Rubio ED, Reiss DJ, Welch PL, Distèche CM, Filippova GN, Baliga NS, Aebersold R, Ranish JA, Krumm A. 2008. CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci U S A* **105**: 8309-8314.
- Sackton TB, Hartl DL. 2016. Genotypic Context and Epistasis in Individuals and Populations. *Cell* **166**: 279-287.
- Sakurai D, Zhao J, Deng Y, Kelly JA, Brown EE, Harley JB, Bae SC, Alarcomicronn-Riquelme ME, Biolupus, networks G et al. 2013. Preferential binding to Elk-1 by SLE-associated IL10 risk allele upregulates IL10 expression. *PLoS Genet* **9**: e1003870.
- Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* **8**: 424-436.
- Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. *Nature* **489**: 109-113.
- Sanz LA, Chamberlain S, Sabourin JC, Henckel A, Magnuson T, Hugnot JP, Feil R, Arnaud P. 2008. A mono-allelic bivalent chromatin domain controls tissue-specific imprinting at Grb10. *EMBO J* **27**: 2523-2532.
- Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. 2012. Linking disease associations with regulatory information in the human genome. *Genome Res* **22**: 1748-1759.
- Scholzen T, Gerdes J. 2000. The Ki-67 protein: from the known and the unknown. *J Cell Physiol* **182**: 311-322.
- Schonemann MD, Ryan AK, McEvelly RJ, O'Connell SM, Arias CA, Kalla KA, Li P, Sawchenko PE, Rosenfeld MG. 1995. Development and survival of the endocrine hypothalamus and posterior pituitary gland requires the neuronal POU domain factor Brn-2. *Genes Dev* **9**: 3122-3135.
- Schork NJ, Murray SS, Frazer KA, Topol EJ. 2009. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* **19**: 212-219.
- Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H et al. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**: 212-216.
- Schulz R, Woodfine K, Menhenniott TR, Bourc'his D, Bestor T, Oakey RJ. 2008. WAMIDEX: a web atlas of murine genomic imprinting and differential expression. *Epigenetics* **3**: 89-96.
- Seitz H, Youngson N, Lin SP, Dalbert S, Paulsen M, Bachellerie JP, Ferguson-Smith AC, Cavaille J. 2003. Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. *Nat Genet* **34**: 261-262.
- Selever J, Kong JQ, Arenkiel BR. 2011. A rapid approach to high-resolution fluorescence imaging in semi-thick brain slices. *J Vis Exp* doi:10.3791/2807.
- Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**: 521-530.
- Shen SQ, Myers CA, Hughes AE, Byrne LC, Flannery JG, Corbo JC. 2016. Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res* **26**: 238-255.

- Shen SQ, Turro E, Corbo JC. 2014. Hybrid Mice Reveal Parent-of-Origin and Cis- and Trans-Regulatory Effects in the Retina. *PLoS One* **9**: e109382.
- Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenko VV et al. 2012. A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**: 116-120.
- Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**: 272-286.
- Shu W, Chen H, Bo X, Wang S. 2011. Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. *Nucleic Acids Res* **39**: 7428-7443.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034-1050.
- Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**: 3940-3941.
- Sisodiya SM, Thompson PJ, Need A, Harris SE, Weale ME, Wilkie SE, Michaelides M, Free SL, Walley N, Gumbs C et al. 2007. Genetic enhancement of cognition in a kindred with cone-rod dystrophy due to RIMS1 mutation. *J Med Genet* **44**: 373-380.
- Slatkin M. 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**: 477-485.
- Smallwood PM, Olveczky BP, Williams GL, Jacobs GH, Reese BE, Meister M, Nathans J. 2003. Genetically engineered mice with an additional class of cone photoreceptors: implications for the evolution of color vision. *Proc Natl Acad Sci U S A* **100**: 11706-11711.
- Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF et al. 2014. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**: 371-375.
- Smith DJ, Anderson J, Zammit S, Meyer TD, Pell JP, Mackay D. 2015. Childhood IQ and risk of bipolar disorder in adulthood: prospective birth cohort study. *British Journal of Psychiatry Open* **1**: 74-80.
- Smith FM, Holt LJ, Garfield AS, Charalambous M, Koumanov F, Perry M, Bazzani R, Sheardown SA, Hegarty BD, Lyons RJ et al. 2007. Mice with a disruption of the imprinted Grb10 gene exhibit altered body composition, glucose homeostasis, and insulin signaling during postnatal life. *Mol Cell Biol* **27**: 5871-5886.
- Smith RL, Traul DL, Schaack J, Clayton GH, Staley KJ, Wilcox CL. 2000. Characterization of promoter function and cell-type-specific expression from viral vectors in the nervous system. *J Virol* **74**: 11254-11261.
- Smith ZD, Meissner A. 2013. DNA methylation: roles in mammalian development. *Nat Rev Genet* **14**: 204-220.
- Soldner F, Stelzer Y, Shivalila CS, Abraham BJ, Latourelle JC, Barrasa MI, Goldmann J, Myers RH, Young RA, Jaenisch R. 2016. Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression. *Nature* **533**: 95-99.
- Solovei I, Kreysing M, Lanctot C, Kosem S, Peichl L, Cremer T, Guck J, Joffe B. 2009. Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. *Cell* **137**: 356-368.

- Solovei I, Wang AS, Thanisch K, Schmidt CS, Krebs S, Zwerger M, Cohen TV, Devys D, Foisner R, Peichl L et al. 2013. LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell* **152**: 584-598.
- Somel M, Liu X, Khaitovich P. 2013. Human brain evolution: transcripts, metabolites and their regulators. *Nat Rev Neurosci* **14**: 112-127.
- Son MS, Taylor RK. 2011. Preparing DNA libraries for multiplexed paired-end deep sequencing for Illumina GA sequencers. *Curr Protoc Microbiol* **Chapter 1**: Unit 1E 4.
- Song X, Vishnivetskiy SA, Seo J, Chen J, Gurevich EV, Gurevich VV. 2011. Arrestin-1 expression level in rods: balancing functional performance and photoreceptor health. *Neuroscience* **174**: 37-49.
- SP Daiger BR, J Greenberg, A Christoffels, W Hide. 1998. Data services and software for identifying genes and mutations causing retinal degeneration. *Investigative Ophthalmology and Visual Science* **39**: S295.
- Spain SL, Barrett JC. 2015. Strategies for fine-mapping complex traits. *Hum Mol Genet* **24**: R111-119.
- Spieler D, Kaffe M, Knauf F, Bessa J, Tena JJ, Giesert F, Schormair B, Tilch E, Lee H, Horsch M et al. 2014. Restless legs syndrome-associated intronic common variant in Meis1 alters enhancer function in the developing telencephalon. *Genome Res* **24**: 592-603.
- Srivastava S, Ketter TA. 2010. The link between bipolar disorders and creativity: evidence from personality and temperament studies. *Curr Psychiatry Rep* **12**: 522-530.
- Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, Vernot B, Cheng JB, Thurman RE, Sandstrom R et al. 2013. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**: 888-903.
- Storch KF, Paz C, Signorovitch J, Raviola E, Pawlyk B, Li T, Weitz CJ. 2007. Intrinsic circadian clock of the mammalian retina: importance for retinal processing of visual information. *Cell* **130**: 730-741.
- Sugitani Y, Nakai S, Minowa O, Nishi M, Jishage K, Kawano H, Mori K, Ogawa M, Noda T. 2002. Brn-1 and Brn-2 share crucial roles in the production and positioning of mouse neocortical neurons. *Genes Dev* **16**: 1760-1765.
- Sun J, Rockowitz S, Xie Q, Ashery-Padan R, Zheng D, Cvekl A. 2015. Identification of in vivo DNA-binding mechanisms of Pax6 and reconstruction of Pax6-dependent gene regulatory networks during forebrain and lens development. *Nucleic Acids Res* **43**: 6827-6846.
- Sun W, Zang L, Shu Q, Li X. 2014. From development to diseases: the role of 5hmC in brain. *Genomics* **104**: 347-351.
- Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, Thompson CL, Hawrylycz M, Dang C. 2013. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res* **41**: D996-D1008.
- Swaroop A, Kim D, Forrest D. 2010. Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. *Nat Rev Neurosci* **11**: 563-576.
- Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L et al. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**: 930-935.
- The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

- Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75-82.
- Tillo D, Hughes TR. 2009. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics* **10**: 442.
- Tillo D, Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Field Y, Lieb JD, Widom J, Segal E, Hughes TR. 2010. High nucleosome occupancy is encoded at human regulatory sequences. *PLoS One* **5**: e9129.
- Tirosh I, Reikhav S, Levy AA, Barkai N. 2009. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* **324**: 659-662.
- Tole S, Remedios R, Saha B, Stoykova A. 2005. Selective requirement of Pax6, but not Emx2, in the specification and development of several nuclei of the amygdaloid complex. *J Neurosci* **25**: 2753-2760.
- Ton CC, Miwa H, Saunders GF. 1992. Small eye (Sey): cloning and characterization of the murine homolog of the human aniridia gene. *Genomics* **13**: 251-256.
- Tournamille C, Colin Y, Cartron JP, Le Van Kim C. 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet* **10**: 224-228.
- Trampush JW, Lencz T, Knowles E, Davies G, Guha S, Pe'er I, Liewald DC, Starr JM, Djurovic S, Melle I et al. 2015. Independent evidence for an association between general cognitive ability and a genetic locus for educational attainment. *Am J Med Genet B Neuropsychiatr Genet* **168B**: 363-373.
- Turro E, Astle WJ, Tavaré S. 2014. Flexible analysis of RNA-seq data using mixed effects models. *Bioinformatics* **30**: 180-188.
- Turro E, Su SY, Goncalves A, Coin LJ, Richardson S, Lewin A. 2011. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol* **12**: R13.
- van Arensbergen J, van Steensel B, Bussemaker HJ. 2014. In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol* **24**: 695-702.
- van El CG, Cornel MC, Borry P, Hastings RJ, Fellmann F, Hodgson SV, Howard HC, Cambon-Thomsen A, Knoppers BM, Meijers-Heijboer H et al. 2013. Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics. *Eur J Hum Genet* **21 Suppl 1**: S1-5.
- Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, Consortium WU-MH. 2013. The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**: 62-79.
- Vanhille L, Griffon A, Maqbool MA, Zacarias-Cabeza J, Dao LT, Fernandez N, Ballester B, Andrau JC, Spicuglia S. 2015. High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat Commun* **6**: 6905.
- Verfaillie A, Svetlichnyy D, Imrichova H, Davie K, Fiers M, Kalender Atak Z, Hulselmans G, Christiaens V, Aerts S. 2016. Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic. *Genome Res* **26**: 882-895.
- Vesuna F, Winnard P, Jr., Raman V. 2005. Enhanced green fluorescent protein as an alternative control reporter to Renilla luciferase. *Anal Biochem* **342**: 345-347.
- Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Sudhof TC, Wernig M. 2010. Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**: 1035-1041.

- Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, Stehling-Sun S, Sabo PJ, Byron R, Humbert R et al. 2014. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**: 1007-1012.
- Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**: 854-858.
- Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**: D88-92.
- Visel A, Taher L, Girgis H, May D, Golonzhka O, Hoch RV, McKinsey GL, Pattabiraman K, Silberberg SN, Blow MJ et al. 2013. A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**: 895-908.
- Visser PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am J Hum Genet* **90**: 7-24.
- Visser M, Palstra RJ, Kayser M. 2014. Human skin color is influenced by an intergenic DNA polymorphism regulating transcription of the nearby BNC2 pigmentation gene. *Hum Mol Genet* **23**: 5750-5762.
- Wade CM, Kulbokas EJ, 3rd, Kirby AW, Zody MC, Mullikin JC, Lander ES, Lindblad-Toh K, Daly MJ. 2002. The mosaic structure of variation in the laboratory mouse genome. *Nature* **420**: 574-578.
- Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R. 2013. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**: 910-918.
- Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* **22**: 1798-1812.
- Wang T, Chen M, Liu L, Cheng H, Yan YE, Feng YH, Wang H. 2011. Nicotine induced CpG methylation of Pax6 binding motif in StAR promoter reduces the gene expression and cortisol production. *Toxicol Appl Pharmacol* **257**: 328-337.
- Wang X, Qiu R, Tsark W, Lu Q. 2007. Rapid promoter analysis in developing mouse brain and genetic labeling of young neurons by doublecortin-DsRed-express. *J Neurosci Res* **85**: 3567-3573.
- Wapinski OL, Vierbuchen T, Qu K, Lee QY, Chanda S, Fuentes DR, Giresi PG, Ng YH, Marro S, Neff NF et al. 2013. Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. *Cell* **155**: 621-635.
- Ward LD, Kellis M. 2012a. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**: D930-934.
- Ward LD, Kellis M. 2012b. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**: 1095-1106.
- Ward ME, McMahon G, St Pourcain B, Evans DM, Rietveld CA, Benjamin DJ, Koellinger PD, Cesarini D, Social Science Genetic Association C, Davey Smith G et al. 2014. Genetic variation associated with differential educational attainment in adults has anticipated associations with school performance in children. *PLoS One* **9**: e100248.
- Warren N, Caric D, Pratt T, Clausen JA, Asavaritikrai P, Mason JO, Hill RE, Price DJ. 1999. The transcription factor, Pax6, is required for cell proliferation and differentiation in the developing cerebral cortex. *Cereb Cortex* **9**: 627-635.

- Waterhouse PM, Wang MB, Lough T. 2001. Gene silencing as an adaptive defence against viruses. *Nature* **411**: 834-842.
- Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorf L et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**: D1001-1006.
- Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T et al. 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**: 796-801.
- Wenger AM, Clarke SL, Notwell JH, Chung T, Tuteja G, Guturu H, Schaar BT, Bejerano G. 2013. The enhancer landscape during early neocortical development reveals patterns of dense regulation and co-option. *PLoS Genet* **9**: e1003728.
- White MA. 2015. Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. *Genomics* **106**: 165-170.
- White MA, Myers CA, Corbo JC, Cohen BA. 2013. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc Natl Acad Sci U S A* **110**: 11952-11957.
- Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, Charlson FJ, Norman RE, Flaxman AD, Johns N et al. 2013. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* **382**: 1575-1586.
- Wilken MSB, J.A.; La Torre, A.; Siebenthal, K.; Thurman R.; Sabo, P.; Sandstrom, R.S.; Vierstra, J.; Canfield, T.K.; Hansen, R.S.; Bender, M.A.; Stamatoyannopoulos, J.; Reh, T.A. 2015. DNase I hypersensitivity analysis of the mouse brain and retina identifies region-specific regulatory elements. *Epigenetics & Chromatin* **8**.
- Wilkins JF, Haig D. 2003. What good is genomic imprinting: the function of parent-specific gene expression. *Nat Rev Genet* **4**: 359-368.
- Williamson CM BA, Thomas S, Beechey CV, Hancock J, Cattanauch BM, Peters J. 2014. MouseBook Imprinting Catalog.
- Wilson C, Bellen HJ, Gehring WJ. 1990. Position effects on eukaryotic gene expression. *Annu Rev Cell Biol* **6**: 679-714.
- Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. *Nature* **430**: 85-88.
- Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**: 59-69.
- Workman CT, Yin Y, Corcoran DL, Ideker T, Stormo GD, Benos PV. 2005. enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res* **33**: W389-392.
- Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**: 206-216.
- Wright AF, Chakarova CF, Abd El-Aziz MM, Bhattacharya SS. 2010. Photoreceptor degeneration: genetic and mechanistic dissection of a complex trait. *Nat Rev Genet* **11**: 273-284.
- Wu H, Zhang Y. 2011. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes Dev* **25**: 2436-2452.
- Wu Z, Asokan A, Samulski RJ. 2006. Adeno-associated virus serotypes: vector toolkit for human gene therapy. *Mol Ther* **14**: 316-327.

- Wu Z, Yang H, Colosi P. 2010. Effect of genome size on AAV vector packaging. *Mol Ther* **18**: 80-86.
- Wurmbach E, Gonzalez-Maeso J, Yuen T, Ebersole BJ, Mastaitis JW, Mobbs CV, Sealfon SC. 2002. Validated genomic approach to study differentially expressed genes in complex tissues. *Neurochem Res* **27**: 1027-1033.
- Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B. 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**: 816-831.
- Xu J, Dodd RL, Makino CL, Simon MI, Baylor DA, Chen J. 1997. Prolonged photoresponses in transgenic mouse rods lacking arrestin. *Nature* **389**: 505-509.
- Yamanaka T, Tosaki A, Miyazaki H, Kurosawa M, Furukawa Y, Yamada M, Nukina N. 2010. Mutant huntingtin fragment selectively suppresses Brn-2 POU domain transcription factor to mediate hypothalamic cell dysfunction. *Hum Mol Genet* **19**: 2099-2112.
- Yan Z, Zak R, Zhang Y, Engelhardt JF. 2005. Inverted terminal repeat sequences are important for intermolecular recombination and circularization of adeno-associated virus genomes. *J Virol* **79**: 364-379.
- Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ATC, Replication DIG, Meta-analysis C, Madden PA, Heath AC, Martin NG et al. 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**: 369-375, S361-363.
- Yoo AS, Sun AX, Li L, Shcheglovitov A, Portmann T, Li Y, Lee-Messer C, Dolmetsch RE, Tsien RW, Crabtree GR. 2011. MicroRNA-mediated conversion of human fibroblasts to neurons. *Nature* **476**: 228-231.
- Young RW. 1985. Cell differentiation in the retina of the mouse. *Anat Rec* **212**: 199-205.
- Ypsilanti AR, Rubenstein JL. 2016. Transcriptional and epigenetic mechanisms of early cortical development: An examination of how Pax6 coordinates cortical development. *J Comp Neurol* **524**: 609-629.
- Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**: 355-364.
- Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2015. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**: 556-559.
- Zarate YA, Perry H, Ben-Omran T, Sellars EA, Stein Q, Almureikhi M, Simmons K, Klein O, Fish J, Feingold M et al. 2015. Further supporting evidence for the SATB2-associated syndrome found through whole exome sequencing. *Am J Med Genet A* **167A**: 1026-1032.
- Zaret KS, Carroll JS. 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25**: 2227-2241.
- Zeron-Medina J, Wang X, Repapi E, Campbell MR, Su D, Castro-Giner F, Davies B, Peterse EF, Sacilotto N, Walker GJ et al. 2013. A polymorphic p53 response element in KIT ligand influences cancer risk and has undergone natural selection. *Cell* **155**: 410-422.
- Zhang F, Lupski JR. 2015. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**: R102-110.
- Zhang G, Gurtu V, Kain SR. 1996. An enhanced green fluorescent protein allows sensitive detection of gene transfer in mammalian cells. *Biochem Biophys Res Commun* **227**: 707-711.

- Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.
- Zhong L, Li B, Mah CS, Govindasamy L, Agbandje-McKenna M, Cooper M, Herzog RW, Zolotukhin I, Warrington KH, Jr., Weigel-Van Aken KA et al. 2008. Next generation of adeno-associated virus 2 vectors: point mutations in tyrosines lead to high-efficiency transduction at lower doses. *Proc Natl Acad Sci U S A* **105**: 7827-7832.
- Zhu B, Chen C, Moyzis RK, Dong Q, Lin C. 2015. Educational attainment-related loci identified by GWAS are associated with select personality traits and mathematics and language abilities. *Personality and Individual Differences* **72**: 96-100.
- Zincarelli C, Soltys S, Rengo G, Rabinowitz JE. 2008. Analysis of AAV serotypes 1-9 mediated gene expression and tropism in mice after systemic injection. *Mol Ther* **16**: 1073-1080.
- Zirlinger M, Anderson D. 2003. Molecular dissection of the amygdala and its relevance to autism. *Genes Brain Behav* **2**: 282-294.
- Zolotukhin S, Potter M, Zolotukhin I, Sakai Y, Loiler S, Fraitas TJ, Jr., Chiodo VA, Phillipsberg T, Muzyczka N, Hauswirth WW et al. 2002. Production and purification of serotype 1, 2, and 5 recombinant adeno-associated viral vectors. *Methods* **28**: 158-167.
- Zuin J, Dixon JR, van der Reijden MI, Ye Z, Kolovos P, Brouwer RW, van de Corput MP, van de Werken HJ, Knoch TA, van IWF et al. 2014. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci U S A* **111**: 996-1001.

APPENDIX 1:

DNA Methylation in Photoreceptors During Development

Methylation is a type of DNA modification most often observed at CpG dinucleotides, in which a methyl group is added at the fifth carbon of cytosine (5mC). This modification is generally associated with gene silencing and is thought to have evolved as a host defense mechanism against viral DNA (Waterhouse et al. 2001). In development, methylation and demethylation are dynamic processes mediated by specific enzymes (Smith and Meissner 2013). Since methylation can block the binding of TFs, methylation can alter CRE activity, and methylated DNA is often associated with repressive histone marks and compacted chromatin. Methylation can be assayed with a variety of methods, but bisulfite sequencing provides single base resolution and is the gold standard (Bock 2012). One caveat of associated with this technique is that both 5mC and 5hmC (discussed below) are protected from bisulfite conversion, so they both appear ‘methylated’ in the assay. Newer techniques have been developed to overcome this issue (e.g., (Booth et al. 2012)).

It was recently discovered that in addition to 5mC (the ‘fifth’ DNA base), there is also 5hmC (the ‘sixth’ DNA base), in which the fifth position of cytosine harbors a hydroxymethyl group (Kriaucionis and Heintz 2009; Pfeifer and Szabo 2009; Tahiliani et al. 2009). The conversion of 5mC to 5hmC occurs via the action of TET enzymes, and 5hmC is thought to be an intermediate leading to demethylation of a CpG (Wu and Zhang 2011). Since its discovery, 5hmC has been widely studied in the brain, where it was first discovered (Sun et al. 2014), as well as in cancers (Ficz and Gribben 2014). The retina expresses several DNA methyltransferases (DNMTs) early in development. Later, rods and cones exhibit differential expression of *Dnmt1* (Nasonkin et al. 2011). Therefore, I sought to investigate the potential roles of 5mC and 5hmC in retinal development, with a focus on photoreceptors.

Whereas nearly all mammalian cells exhibit a ‘conventional’ nuclear architecture, with peripheral heterochromatin and central euchromatin, the rods (but not cones) of nocturnal mammals have an ‘inverted’ nuclear architecture. In particular, there is a thin layer of peripheral euchromatin and a single large, central clump of heterochromatin in the rod nucleus of nocturnal mammals, which is thought to act as a lens to concentrate light onto the photosensitive outer segment (Carter-Dawson and LaVail 1979; Solovei et al. 2009). Based on mouse studies, the formation of the central clump of heterochromatin occurs slowly over development, beginning with small spheres that coalesce over the first postnatal month (Solovei et al. 2009). The chromatin structure of the rod nucleus is also reflected by histone marks: from central to peripheral, there is an increasing density of activating histone marks and a decreasing density of repressing histone marks (Kizilyaprak et al. 2010). Recent studies suggest that lamin A/C and lamin B receptor play key roles in the establishment of this rod nuclear architecture, and that methylation may also be involved (Solovei et al. 2013; Mo et al. 2016).

I examined the relationship between 5mC, 5hmC, and nuclear architecture in the developing retina by using antibodies that recognize 5mC-rich or 5hmC-rich DNA (Figure A1.1). At P0 (peak of rod birth), 5mC and 5hmC staining exhibit considerable overlap in cells of the NBL, where presumptive rods reside. As development progresses (P5-P8), 5mC becomes more localized to discrete foci within each nucleus in the ONL (which are nearly all rod nuclei), whereas 5hmC is distributed through the nucleus. By P22, most ONL nuclei have only one or two 5mC foci, and by P35, essentially all ONL cells have a single, central 5mC focus. In contrast, cells of the INL and GCL have similar 5mC and 5hmC staining patterns throughout development, namely one or few 5mC foci located in the nuclear periphery. Thus, it appears that the overall distribution of 5mC and 5hmC reflect heterochromatin and euchromatin, respectively, and mirror

the development of rod nuclear architecture (Figure A1.1C). Interestingly, the rod nuclear architecture is not fully established until ~4-5 weeks after birth, whereas by most other measures (including gene expression and electrophysiology), rods are mature by 3 weeks.

Cones are a relatively rare population in the wild-type mouse retina (~2% of cells) but abundant in the *Nrl*^{-/-} retina, where rods have been developmentally transdifferentiated to cones (Mears et al. 2001). These transdifferentiated photoreceptors have been shown to be largely indistinguishable from native cones at the molecular, morphological, and functional levels. Compared to rods, native cones and *Nrl*^{-/-} cones have a more conventional nuclear architecture, with a lesser degree of central clumping of heterochromatin and an increased amount of peripheral euchromatin (Solovei et al. 2013). In the *Nrl*^{-/-} retina at age P63, the 5mC and 5hmC staining patterns are not qualitatively different from those of the wild-type retina (Figure A1.2). A conditional knockout of *Nrl*, in which mature rods have been partially converted to cones, also appeared normal (Figure A1.2). Additional studies are needed to confirm and clarify these findings.

The promoters of a handful of genes known to be expressed in the retina have been reported to exhibit retina-specific methylation patterns (Merbs et al. 2012). To examine the dynamics of methylation in the retina at the DNA level, I conducted bisulfite sequencing of retinas at multiple ages, focusing on the promoters of two genes: *Rho* (rhodopsin), a canonical rod photoreceptor gene, and *Opn1sw* (S-opsin), a canonical blue cone gene.

These analyses revealed that there is a progressive decrease in methylation over development at the *Rho* promoter in the WT retina but not in the *Nrl*^{-/-} retina (Figure A1.3A). Similarly, there is a progressive decrease in methylation at the *Opn1sw* promoter in the *Nrl*^{-/-} retina but not the WT retina (Figure A1.3B). This suggests that there are waves of demethylation that occur in rods and cones at the *Rho* and *Opn1sw* loci, respectively. Notably, these waves of

demethylation occur somewhat later than the known increase in expression of *Rho* and *Opn1sw* in rods and cones, respectively, although here I did not directly measure expression of these genes.

To confirm the cell type specificity of the effect, I also conducted bisulfite analysis of the *Rho* promoter for FACS-sorted photoreceptors (nearly all of which are rods) and bipolar cells from adult Otx2-GFP mice (Fossat et al. 2007). As expected, the *Rho* promoter was essentially unmethylated in FACS-sorted adult photoreceptors, but heavily methylated in bipolar cells (Figure A1.4).

In summary, methylation in rod photoreceptors is highly dynamic over development, as assessed by 5mC and 5hmC antibody staining as well as by bisulfite sequencing. Changes in methylation are temporally delayed compared to changes in gene expression, suggesting a maintenance rather than causal role. Recent studies in our lab have attempted to directly reprogram rods into cones for therapeutic purposes, but thus far, these efforts have achieved only partial reprogramming, presumably due to epigenetic barriers to transdifferentiation (Montana et al. 2013). It is possible that introducing demethylases or histone deacetylase (HDAC) inhibitors such as valproic acid (Milutinovic et al. 2007; Huangfu et al. 2008) may help overcome epigenetic barriers to transdifferentiation, thereby permitting more efficient conversion of rods into cones.

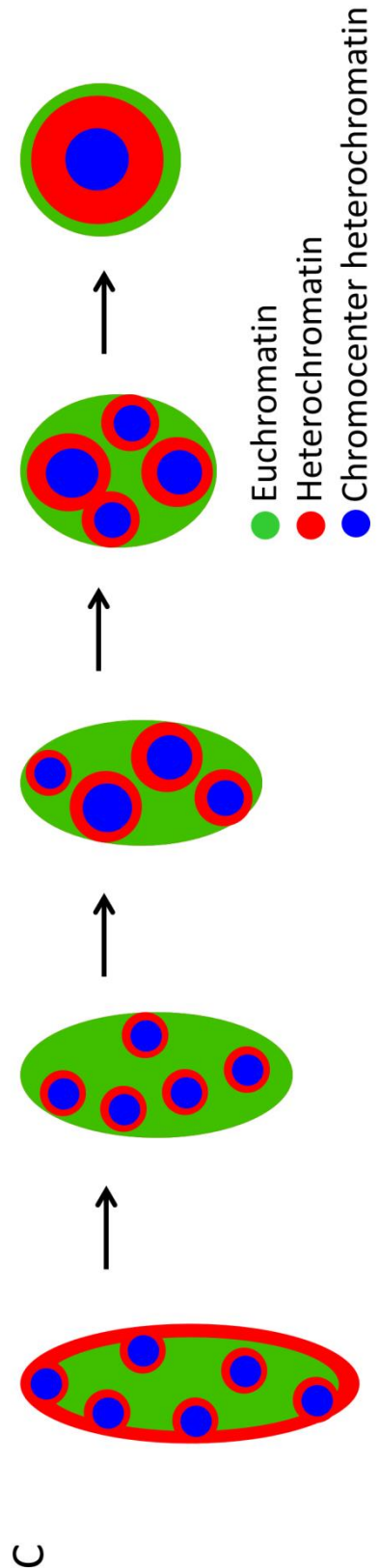
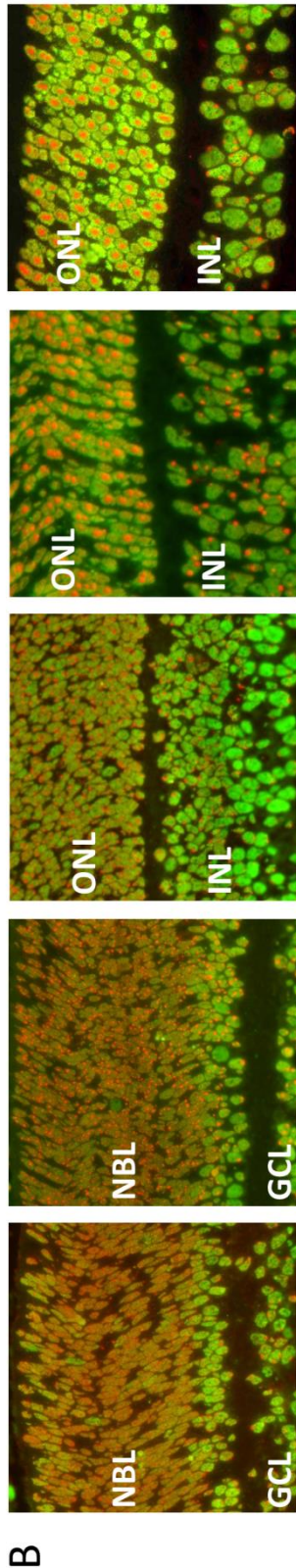
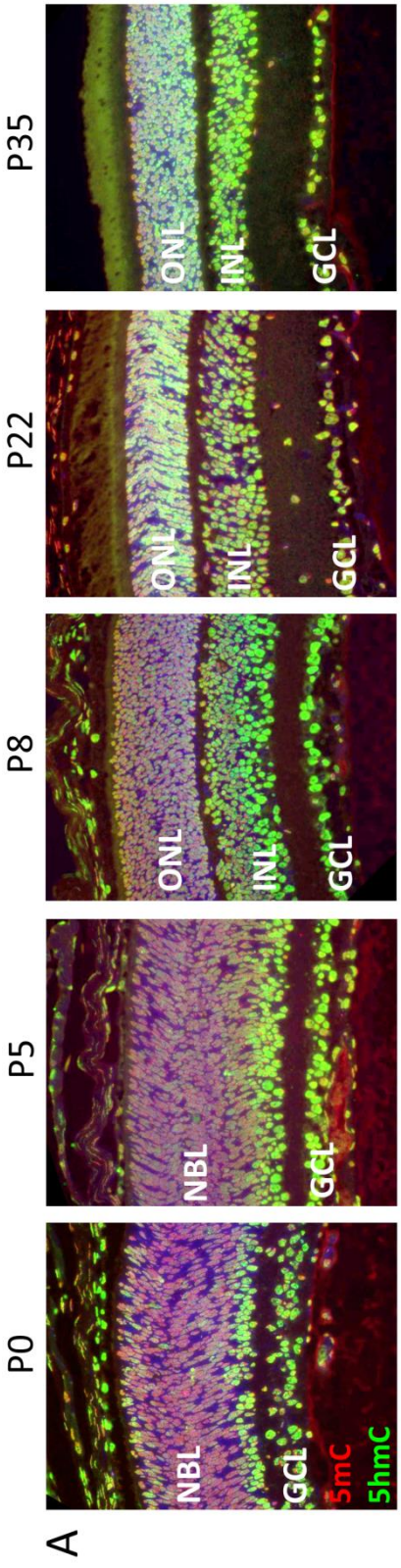
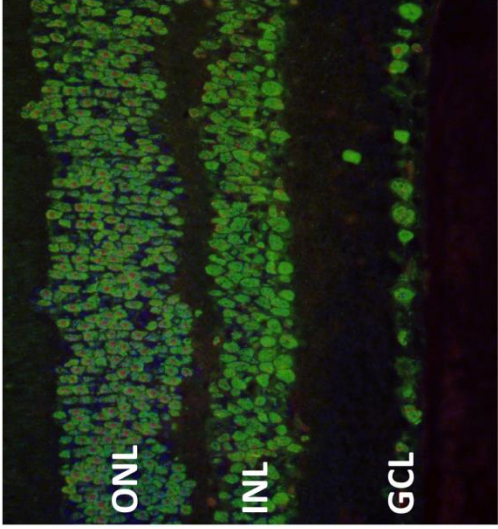
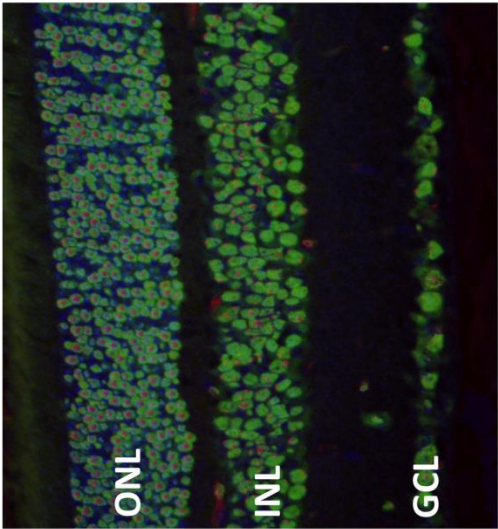


Figure A1.1. The distribution of 5mC and 5hmC in mouse rod photoreceptors during development reflects nuclear architecture. (A) Frozen sections of retinas at the indicated postnatal days were analyzed by immunohistochemistry. Retinas were from controls in CTCF experiments (Appendix 2). CTCF mutant retinas were also analyzed and showed no difference in antibody staining patterns compared to controls (data not shown). Rod nuclei reside in the ONL and constitute most of the cells there. The following antibodies were used: anti-5mC, Eurogentec BI-MECY-9199 (with red secondary); anti-5hmC, ActiveMotif 39769 (with green secondary). Images were taken at 400X magnification. Blue, DAPI stain. NBL, neuroblast layer; ONL, outer nuclear layer; INL, inner nuclear layer; GCL; ganglion cell layer. (B) Enlarged images with DAPI channel removed for clarity. P35 image was taken at 1000X. (C) Model of mouse rod nuclear architecture development, at ages corresponding to those in the images above, based on (Solovei et al. 2009).

Conditional *Nr1* KO



Wild-type



Nr1^{-/-}

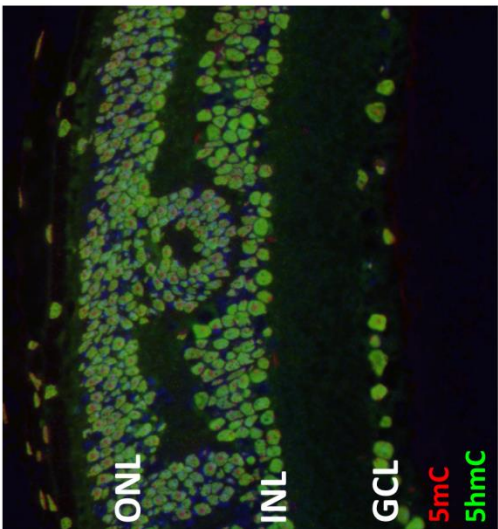


Figure A1.2. 5mC and 5hmC distributions in models of rod-to-cone transdifferentiation. Antibody staining for 5mC and 5hmC were conducted as for Figure A1.1. All mice were age P63. Left panel: *Nrl*^{-/-} retina, which contains cones only (no rods) for photoreceptors. Note the rosettes typical for this mutant. Middle panel: Control retina (CAG-Cre-ERT;*Nrl*^{fl/+}) treated with tamoxifen daily at P42-P44. Right panel: Conditional *Nrl* knockout retina (CAG-Cre-ERT;*Nrl*^{fl/fl}) treated with tamoxifen daily at P42-P44. Slides were a gift from Cynthia Montana—see (Montana et al. 2013) for information. All images were taken at 400X magnification. ONL, outer nuclear layer; INL, inner nuclear layer; GCL; ganglion cell layer.

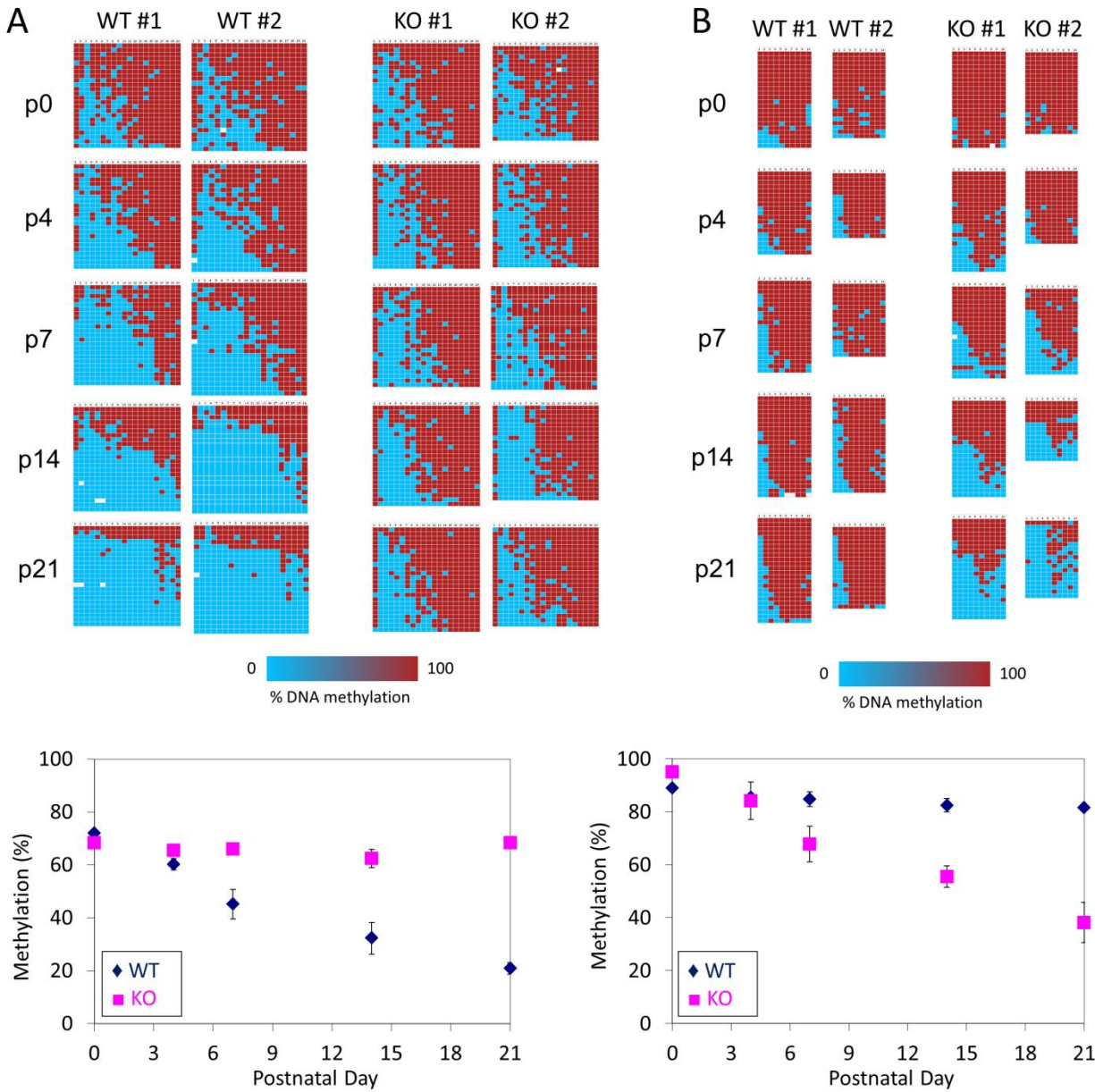


Figure A1.3. Bisulfite analysis of *Rho* and *Opn1sw* promoters in wild-type and *Nrl*^{-/-} retinas over development. Bisulfite treatment, PCR, cloning, and analysis were conducted described in (Rohde et al. 2010; Montana et al. 2013) for these two loci. Two replicates (each consisting of multiple retinas) for each time point were harvested. Top: Data for individual CpG sites are shown. Each row represents an analyzed cell. Red = methylated, blue = unmethylated, white = no data. Bottom: Quantification of methylation levels (averaged over the analyzed region). Error bars indicate SEM between the two replicates. (A) *Rho* promoter. (B) *Opn1sw* promoter.

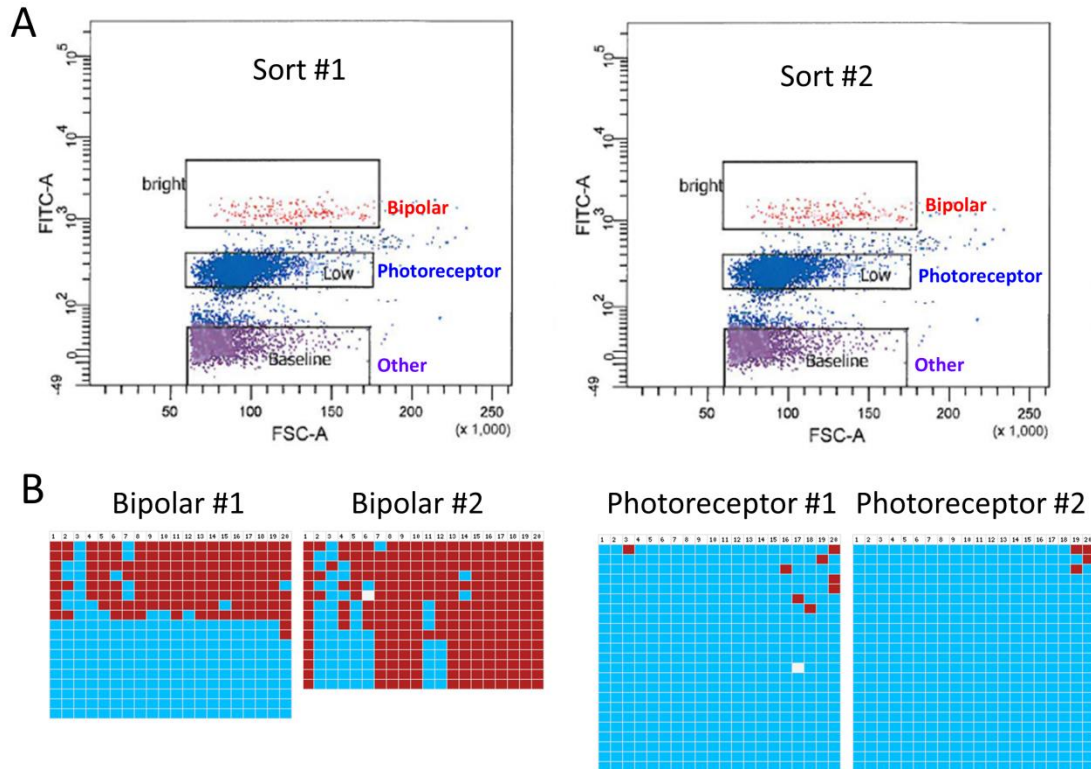


Figure A1.4. FACS-sorted photoreceptors and bipolar cells reveal cell type-specific methylation patterns at the *Rho* promoter. (A) FACS plot depicting sorting of bipolar cells and photoreceptor cells (most of which are rods). The retinas of mice (age 3 months) heterozygous for the *Otx2*-GFP transgene were dissociated. Cells with high GFP levels (bipolar cells) and cells with low GFP levels (photoreceptors) were collected. As a control, retinas from wild-type littermates were dissociated and sorted to establish baseline levels of fluorescence. Two independent sorts were conducted, resulting in two biological replicates each of bipolar cells and photoreceptors. X-axis (FSC-A), forward scatter. Y-axis (FITC-A), GFP levels. (B) Bisulfite analysis of the *Rho* promoter. Data for individual CpG sites are shown. Each row represents an analyzed cell. Red = methylated, blue = unmethylated, white = no data. The difference in methylation levels between the two bipolar replicates may be due to contamination of first bipolar replicate with rod photoreceptors.

APPENDIX 2:

The role of CTCF in the retina

CTCF (CCCTC-binding factor) is a ubiquitously expressed transcription factor (TF) that facilitates the establishment of 3D genome architecture by forming topologically associating domains (TADs). TADs are separated by DNA elements called insulators, which contain motifs bound by CTCF (Ghirlando and Felsenfeld 2016). For many years, CTCF was widely touted as the ‘master weaver’ of the genome (Phillips and Corces 2009). This was most convincingly demonstrated by CRISPR-Cas mediated mutation of CTCF sites, which caused changes in gene expression and chromatin looping (Guo et al. 2015). In its role as a mediator of chromatin looping, CTCF is thought to act in concert with cohesin (Rubio et al. 2008; Wendt et al. 2008; Nativio et al. 2009). However, cohesin-independent effects of CTCF have been reported (Kim et al. 2011; Zuin et al. 2014). Additionally, it is unclear how CTCF establishes cell type-specific chromatin architecture.

In the avian retina, *CTCF* and *Pax6* are initially coexpressed in early development but then segregate, such that photoreceptors are CTCF+, Pax6- and amacrine cells are Pax6+, CTCF- (Canto-Soler et al. 2008). In that study, it was suggested the *CTCF* represses *Pax6* expression and thereby indirectly promotes photoreceptor fate. To clarify the role of CTCF in the mammalian retina, I histologically characterized *CTCF* knockout mouse retinas in collaboration with Connie Myers.

To knock out *CTCF* in the developing retina, we recombined a floxed allele of *CTCF* using a Cre recombinase driven by the *Six3* promoter. Since *Six3* is widely expressed in retinal progenitors (as well as in other parts of the CNS) by E11.5, *Six3Cre⁺;CTCF^{fl/fl}* retinas should be essentially CTCF-deficient (Oliver et al. 1995). We found morphological evidence of retinal degeneration in the *CTCF* mutants, with rapidly progressive thinning of all retinal layers (Figure A2.1). Despite this degeneration, multiple cell types could still be identified in the *CTCF*

mutants by antibody staining against M-opsin (M-cones), S-opsin (S-cones), PKCa (bipolar cells), anti-Pax6 (amacrine cells), and glutamine synthetase (Müller glia) (Figure A2.2). These findings suggest that while CTCF may be important for the maintenance of these retinal cell types, it is not required for their formation. It is possible that CTCF has subtle effects on specific cell subpopulations or on the relative proportions of cell types. Also, rods and horizontal cells were not analyzed, although well-characterized markers for these cells, and additional markers for the other cell types, are available and should be used in future studies (Cheng et al. 2013).

Notably, in the mutants but not in the controls, there appeared to be a greater degree of co-localization of Pax6 and glutamine synthetase (GS) expression in the cells of the INL, both at P28 (Figure A2.2C, white arrowheads) and at P10 (data not shown). This suggests that either amacrine cells (typically Pax6⁺ and GS⁻) have gained GS expression, or Müller glia (generally assumed to be GS⁺ and Pax6⁻) have gained Pax6 expression. It is now known that Müller glia can express *Pax6* and their nuclei can migrate in response to injury (Roesch et al. 2008; Joly et al. 2011). It is also possible that, since *CTCF* normally represses *Pax6* expression, the deletion of *CTCF* may directly lead to the derepression of *Pax6* in Müller glia. Further experiments are needed to verify the initial observation and to distinguish these scenarios. Also, the distribution of *CTCF* expression in the wild-type retina, and the extent of *CTCF* knockout in the mutant, should be assessed in the future.

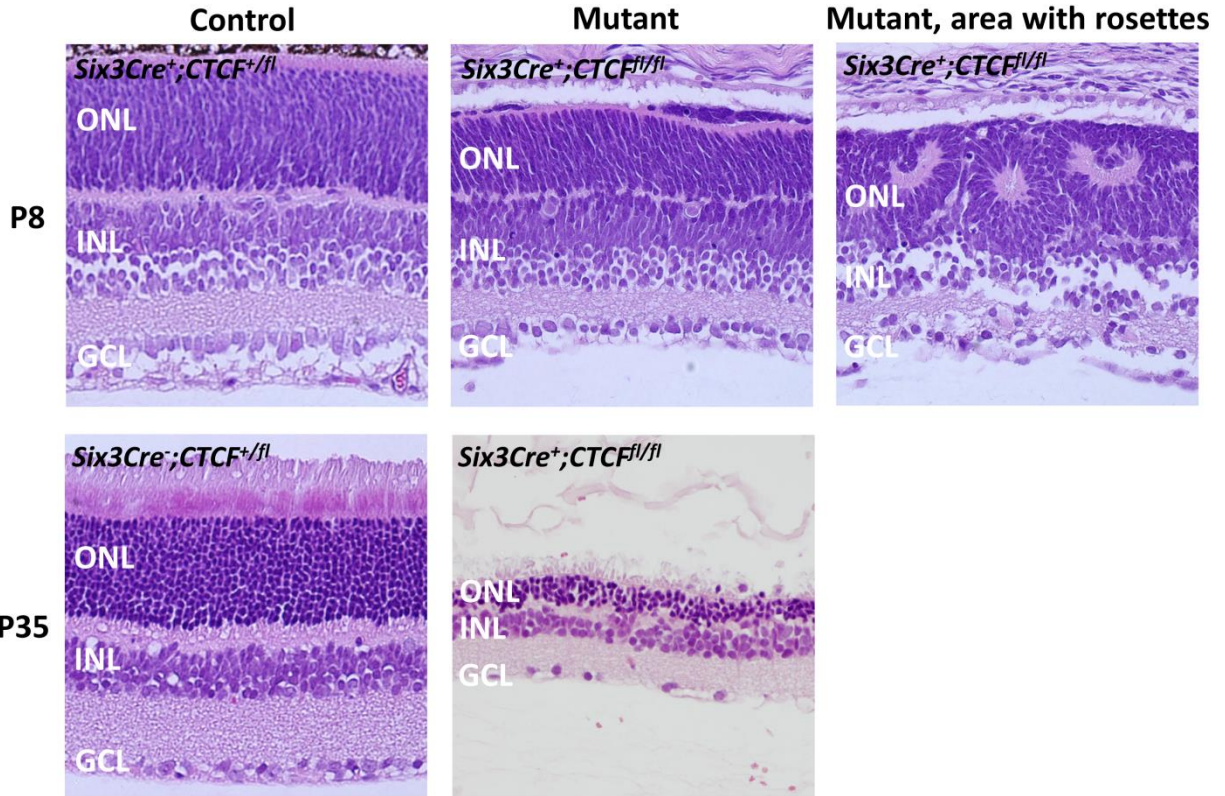


Figure A2.1. Deletion of *CTCF* in the mouse neural retina results in retinal degeneration. Control and mutant eyes were examined by H&E staining of paraffin sections. Ages and genotypes are indicated. For P8 mutant, rosettes were observed in some regions. Images were taken at 400X magnification. ONL, outer nuclear layer; INL, inner nuclear layer; GCL, ganglion cell layer.

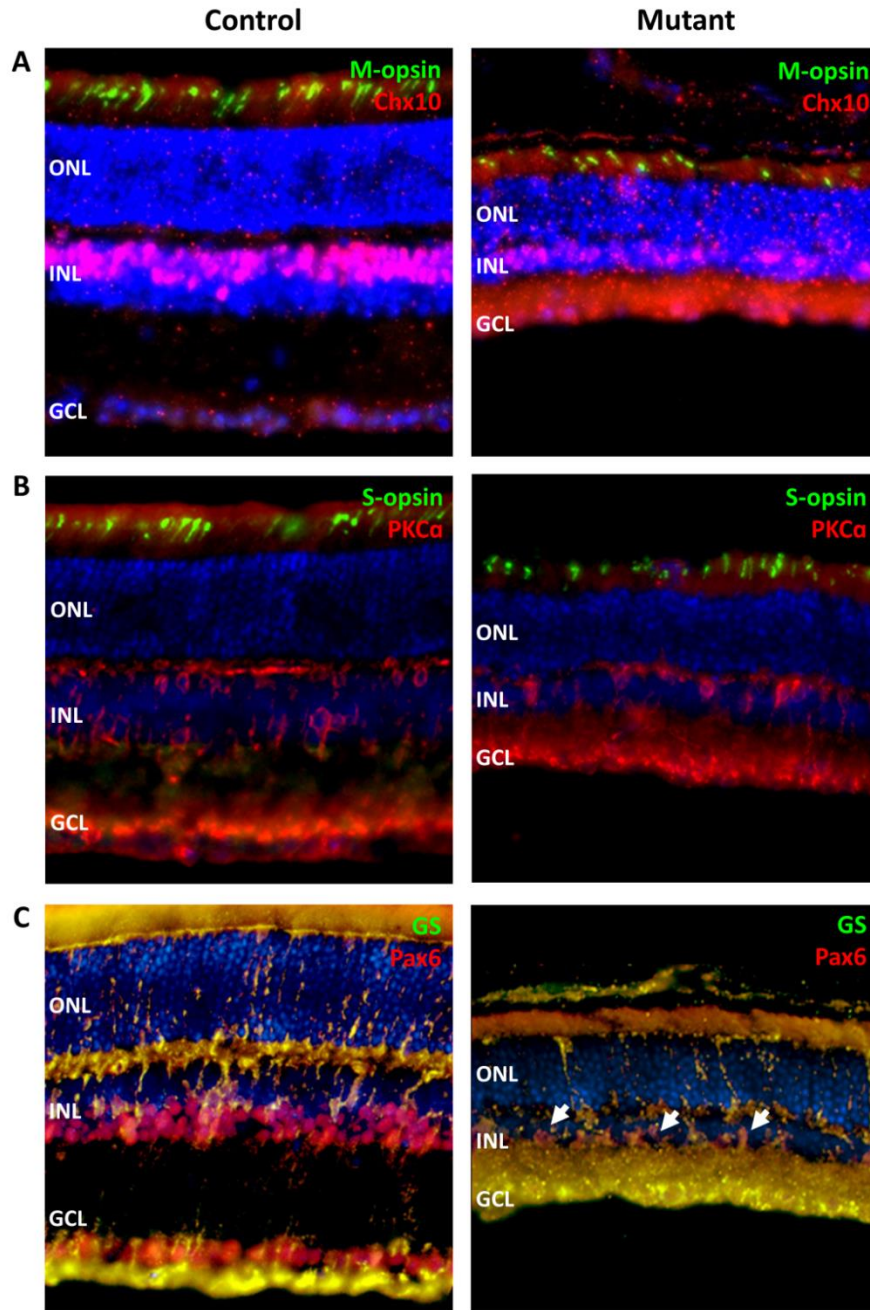


Figure A2.2. Expression of cellular markers in CTCF knockout retinas. Frozen sections of control ($Six3Cre^+;CTCF^{+/fl}$) and mutant ($Six3Cre^+;CTCF^{fl/fl}$) P28 retinas were analyzed with antibodies. (A) anti-M-opsin (red/green cones), Millipore AB5405; anti-Chx10 (bipolar cells), Exalpha Biologicals X1180; (B) anti-S-opsin (blue cones), Millipore AB5407; anti-PKCa (bipolars), Millipore 05-154; (C) anti-Pax6 (amacrine cells), Developmental Studies Hybridoma Bank; anti-glutamine synthetase (Müller glia), BD 610517. White arrowheads are described in the text. Images were taken at 200X magnification. Blue, DAPI stain. GS, glutamine synthetase; ONL, outer nuclear layer; INL, inner nuclear layer; GCL, ganglion cell layer.

APPENDIX 3

High-coverage CRE-seq libraries tiling the *MIR2113/POU3F2* locus

In our study of the *MIR2113/POU3F2* locus, we pursued a candidate causal variant (rs77910749) that fell within a fetal brain DHS (LC1). However, the locus contains many dozens of fetal brain-specific DHS peaks, suggesting that multiple CREs within this region act in combination to regulate *POU3F2* and/or other target genes. To systematically and comprehensively assay these fetal brain-specific DHSs for *cis*-regulatory activity, I synthesized two types of CRE-seq libraries: (1) a PCR-based library of candidate CREs within a 1.5 Mb window, and (2) a bacterial artificial chromosome (BAC)-based library of elements that tiled across 440 kb. These two libraries are targeted and unbiased strategies, respectively, that complement each other.

For the PCR library, I selected 100 fetal brain-specific DHSs in a 1.5 Mb window (Chr6:97.8-99.3 Mb in hg19), designed primers, and conducted individual PCR reactions, using commercially available human gDNA as the template (Figure A3.1). Next, I cloned each PCR product (~0.5-2 kb in length, with an average of ~1 kb) as a NotI fragment into a barcoded CRE-seq vector (described in Chapter 3) and picked individual colonies for Sanger sequencing to determine the barcode sequence. A total of 799 barcoded constructs representing 97 (out of the targeted 100) DHS's were obtained (Figure A3.2). Notably, since the template DNA for PCR came from a pool of individuals, variants were represented in this library.

I made two versions of the PCR library: one with *Rho* basal-GFP (described in Chapter 3), and another with the 3.6 kb *POU3F2* promoter (described in Chapter 4) driving DsRed. The promoter-reporter cassette was cloned into the FseI/AscI sites of the vector. These libraries are ready for CRE-seq by transfection or electroporation. Alternatively, the libraries can be transferred into the AAV vector for packaging and delivery as AAV libraries (described in Chapter 3). Preliminary studies (using the 3.6 kb *POU3F2*-DsRed version of the library) showed

minimal DsRed expression in *ex vivo* electroporated developing mouse cerebral cortex, as seen under a dissecting fluorescent microscope. This suggests that most elements in this library are inactive and/or incompatible with this promoter.

To generate a library that would tile the locus in a relatively unbiased manner, I created a CRE-seq library with three BAC constructs (RP11-640D17, RP11-13H22, RP11-71E9) that encompass a region of ~440 kb (Chr6:98,378,700-98,821,999 in hg19). After purifying the BACs with the Qiagen Large-Construct Kit, I sonicated the DNA to a target fragment size of ~600-700 bp. After end repair, I cloned the fragments into the NotI site of the barcoded CRE-seq vector. I determined the correspondence between the BAC fragments and barcode sequences using paired-end sequencing (described in Chapter 3). Overall, I obtained 20,867 barcodes with a median BAC fragment size of ~630 bp and 40X median coverage (Figure A3.3 and Figure A3.4). The *Rho* basal-GFP cassette has been cloned into this library.

Together, the PCR library and BAC library should be valuable tools for screening the *cis*-regulatory potential of regions within the *MIR2113/POU3F2* intergenic locus. In each case, an alternate promoter-reporter cassette can be cloned into the FseI/AscI sites. The choice of the promoter is an important consideration, because detection of enhancer activity may require a compatible proximal promoter with some level of basal activity. The choice of the assayed cell type is another important consideration. Given that both of these libraries are composed of human DNA elements, it may be valuable to test them in both developing mouse cerebral cortex and iPSC-derived cerebral organoids (as in Chapter 4).

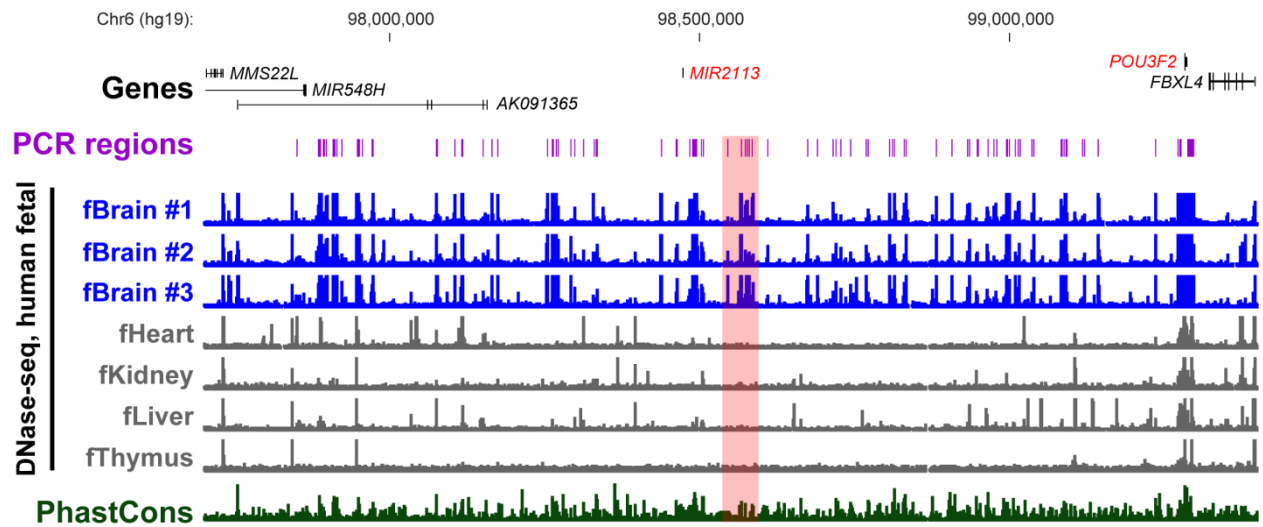


Figure A3.1. The 100 target regions in the *MIR2113/POU3F2* locus for the PCR CRE-seq library. One hundred human fetal brain-specific DHSs within a 1.5 Mb window (roughly centered on the ‘local cluster’, highlighted in pink) were selected for PCR and cloning (purple regions). Note the locations of *MIR2113* and *POU3F2* (red font).

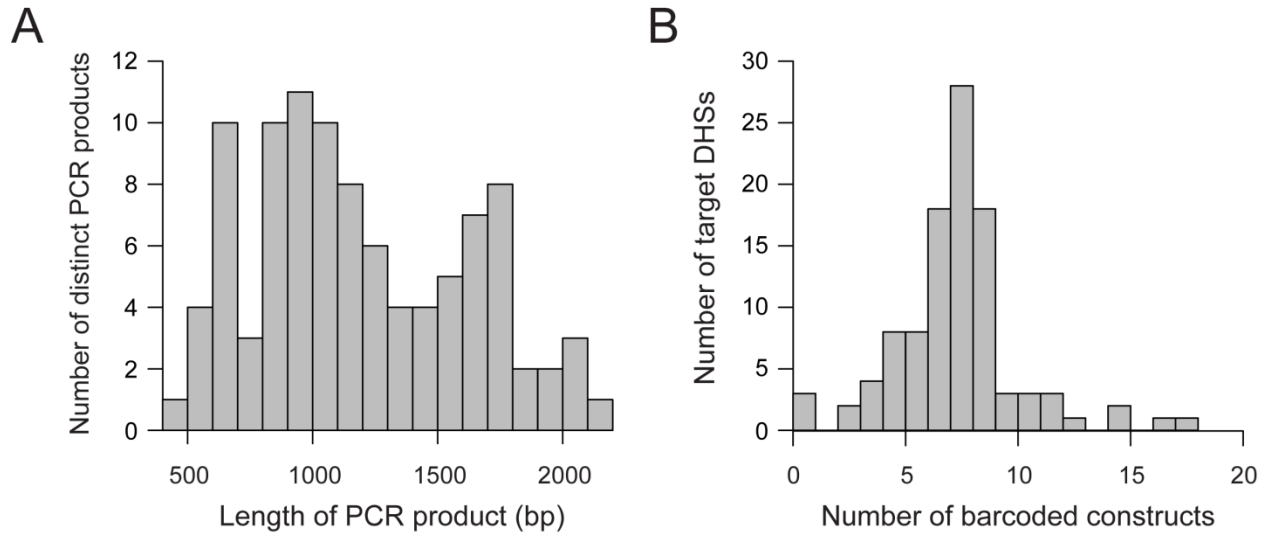


Figure A3.2. Distribution of product lengths in the PCR library and coverage of target DHSs in the *MIR2113/POU3F2* locus. (A) Distribution of the lengths of the PCR products. (B) Coverage of target DHS's. Of the 100 targeted DHSs, 97 were successfully cloned with a total of 799 barcoded constructs. The median coverage was 8X.

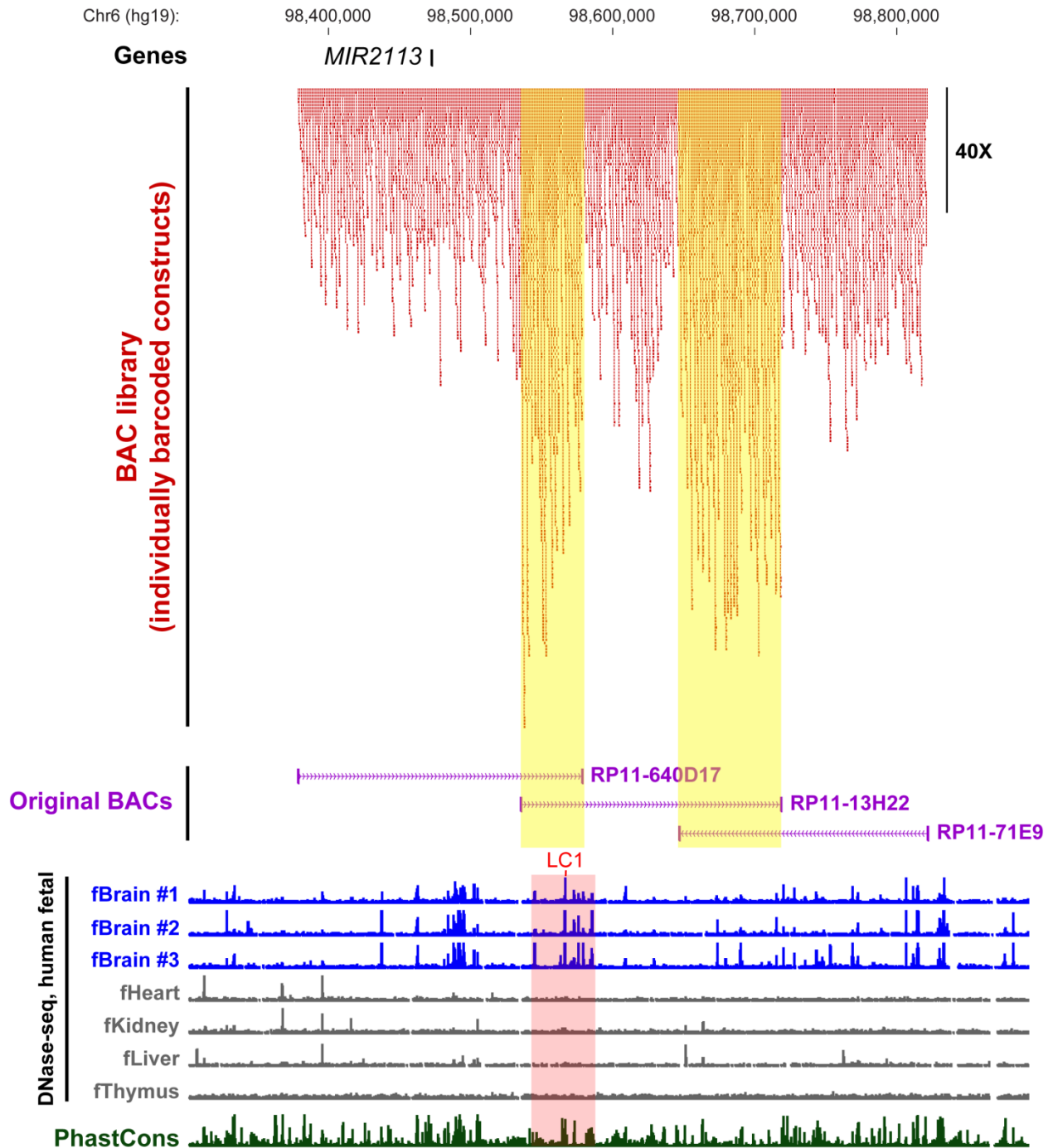


Figure A3.3. A BAC library tiling 440 kb of the *MIR2113/POU3F2* locus at 40X coverage. A total of 20,867 barcoded constructs were obtained, with the individual BAC fragments shown in red. The vertical scale for 40X coverage is indicated. Note the overlap of the three original BACs where there is higher coverage in the library as expected (yellow highlighted regions). Also note the location of LC1 (red font) within the ‘local cluster’ (pink highlighted region) (see Chapter 4).

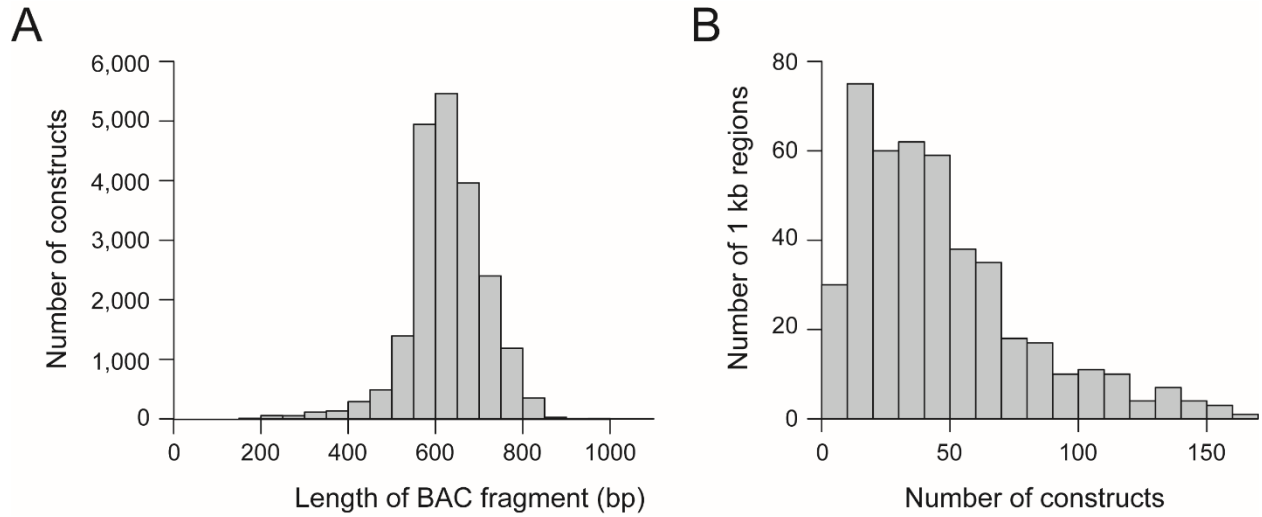


Figure A3.4. Distribution of fragment lengths in the BAC library and coverage of the *MIR2113/POU3F2* locus. Three BACs covering a ~440 kb region (Chr6:98,378,700-98,821,999 in hg19) were sonicated, cloned into a barcoded CRE-seq vector, and subjected to paired-end sequencing. (A) Distribution of the lengths of BAC fragments cloned into the library. The average length was 629 bp (SD = 87 bp). (B) Coverage of the region (split into 1 kb windows, i.e., 440 regions). The median coverage was 40X.